FORESIGHT

Fall 2020

Issue 59

THE INTERNATIONAL JOURNAL OF APPLIED FORECASTING

VIRTUAL FORESIGHT PRACTITIONER CONFERENCE

December 2, 2020

Integrated Business Planning and Forecasting: Innovations to Drive Profitable Growth

Many organizations encounter problems because their Integrated Business Planning (IBP) approach is tactical. Tactical IBP leaves capability gaps that perpetuate silos, erode value, and can't support profitable growth.

The virtual FPC 2020 will explore strategic IBP, which nurtures operational excellence, strategic alignment, and competitive advantage.

Who should tune in?

- Forecasters and planners
- Senior executives and stakeholders in finance, supply chain, sales and marketing, and product development
- Academics, researchers, and data scientists
- Anyone who seeks to find the right approach to strategic IBP

Speakers and topics

- **Jerry Bendiner**, Technologix
  Supply Chain Sustainability, A Critical Component of Strategic IBP
- **Stefan de Kok**, Wahupa
  Forecasting and Analytics in Mature IBP Processes
- **Dean Sorensen**, *Foresight* IBP Editor
  Strategic IBP: Redefining What Fully Integrated Processes Entail
- **Chris Turner**, StrataBridge
  Serious About Growth? Forget S&OP and IBP – Start with Leadership and Strategy

Registration is now open!

IIF Member Price
Free

Non-member Price
$125, which includes a one-year IIF membership with four issues of *Foresight*

Student Price
$25, which includes a one-year Premium IIF membership

FORESIGHT

foresight.forecasters.org/2020-conference

contents

> *"Knowledge of truth is always more than theoretical and intellectual. It is the product of activity as well as its cause. Scholarly reflection therefore must grow out of real problems, and not be the mere invention of professional scholars."*
>
> John Dewey, University of Vermont

forecasting practice

book review

artificial intelligence

election forecasting

forecasting and planning perspectives

Article Coding: Managers (MGR), Modelers (MOD), Planners (PLN), General Audience (GEN)

Foresight, an official publication of the International Institute of Forecasters, seeks to advance the practice of forecasting. To this end, it will publish high-quality, peer-reviewed articles, and ensure that these are written in a concise, accessible style for forecasting analysts, managers, and students.

Topics include:

- Design and Management of Forecasting Processes
- Forecast Model Building: The Practical Issues
- Forecasting Methods Tutorials
- Forecasting Principles and Practices
- S&OP and Collaborative Forecasting
- Forecasting Books, Software and Other Technology
- Applications in Political, Climate and Media Forecasting
- Long-Range and Strategic Forecasts
- Case Studies

Contributors of articles include:

- Analysts and managers, examining the processes of forecasting within their organizations
- Scholars, writing on the practical implications of their research
- Consultants and vendors, reporting on forecasting challenges and potential solutions

All invited and submitted papers will be subject to a blind editorial review. Accepted papers will be edited for clarity and style.

Foresight welcomes advertising. Journal content, however, is the responsibility of, and solely at the discretion of, the editors. The journal will adhere to the highest standards of objectivity. Where an article describes the use of commercially available software or a licensed procedure, we will require the author to disclose any interest in the product, financial or otherwise. Moreover, we will discourage articles whose principal purpose is to promote a commercial product or service.

Foresight is published by the International Institute of Forecasters, Business Office: 53 Tesla Avenue, Medford, MA 02155 USA

FORESIGHT

Editor Len Tashman
lentashman@forecasters.org

Associate Editors
Stephan Kolassa
stephan.kolassa@sap.com

Mike Gilliland
Mike.Gilliland@sas.com

Column Editors
Financial Forecasting Roy Batchelor
r.a.batchelor@city.ac.uk

Forecasting Support Systems
Fotios Petropoulos
f.petropoulos@bath.ac.uk

Hot New Research Paul Goodwin
mnspg@management.bath.ac.uk

Integrated Business Planning Dean Sorensen
deanbsorensen@ibpcollaborative.org

Long-Range Forecasting Ira Sohn
imsfinc@gmail.com

Supply-Chain Forecasting Aris Syntetos
SyntetosA@cardiff.ac.uk

Foresight Staff
Design and Production Liza Woodruff
liza@lizawoodruff.com

Marketing and Sponsorship
Pam Stroud
pamstroud@forecasters.org

Michelle Kitchens
michelle.kitchens@forecasters.org

Manuscript and Copy Editing
Ralph Culver
letter_perfect@fast-email.com

Mary Ellen Bridge
me.foresight@comcast.net

Advisory Board and News Notes
Holly Lancaster
hollylancaster@forecasters.org

Memberships Pam Stroud, IIF Business Director
forecasters@forecasters.org

note from the editor

The Fall 2020 issue of *Foresight*—number 59 since inception in 2005—features the final installment of a three-part article on the forecasting system and practices at the Target Corporation. The senior author of the series is **Phillip Yelland**, Principal Data Scientist at Target. The first two contributions described the architecture and design of the system and recounted lessons learned in the development process. This last segment, **A Modern Retail Forecasting System in Production**, explores the challenges that arise and steps to be taken when a forecasting system such as Target's is actually deployed to provide forecasts for users.

This third installment is followed by a Commentary from **Simon Clarke**, who argues that **It's the Soft Problems that Are Hard to Overcome**, and in turn by a response from the Target team.

Our latest book review from Long-Range Forecasting Editor **Ira Sohn** is of *After Shock*, in which ***The World's Foremost Futurists Reflect on 50 Years of* Future Shock**. The volume is a collection of essays and commentaries that look back upon Alvin Toffler's original best-selling opus from 1970, including in the fields of AI, economics, health, technology, and academia.

Speaking of AI, **John Wood** and **Nada Sanders** issue stern warnings in this issue against the insidious threat of *deepfakes*—the term being a combination of "deep learning" and "fake." Their article **Dealing with "Deepfakes": How Synthetic Media Will Distort Reality, Corrupt Data, and Impact Forecasts** reports that

> *Machine-learning capabilities are escalating the technology's sophistication, making deepfakes ever more realistic and increasingly resistant to detection. The implications for communication, data integrity, forecasting, and decision making are vast and unequivocally grim.*

With the looming November elections in the U.S., vote forecasting is again in high gear. A new and very sophisticated methodological entry comes from *The Economist*. Here, **Colin** and **Michael Lewis-Beck** examine the strengths and weaknesses of ***The Economist* Model**, and provide their perspectives on the various types of election-forecasting models.

Earlier this year, a group of practitioners and academics began discussions about the practical challenges facing the forecasting field and the need to learn why many organizations have not exploited advances in forecasting knowledge and technology. This fall issue concludes with **The Benefits of Systematic Forecasting for Organizations: The UFO Project**, the group's initial assessment of the problem and its plan to better understand what it will take to improve the *Usage of Forecasting in Organizations (UFO)*.

NEW FAB CHAIRMAN

Foresight welcomes the new Chairman of our advisory board, **Jim Hoover**. Following a career in OR for the U.S. Navy, Jim became Managing Director of Accenture Federal Services, with a focus on supply-chain analytics. Jim then received his Doctorate of Business Administration from the University of Florida in 2017 and joined the faculty there in 2019, where he is now Director of the Business Analytics Program. Jim initially served *Foresight* as Software Editor, authored several articles on tracking forecast performance, and is now a member of the aforementioned UFO project team.

Our deep appreciation and thanks go to **Jeff Hunt**, who served as FAB Chairman for the past seven years. Jeff was a prime mover of upgrades made to *Foresight* subscription and promotion practices.

THE 2020 *FORESIGHT* VIRTUAL PRACTITIONER CONFERENCE

Reserve December 2, 2020 for our virtual conference on *Integrated Business Planning and Forecasting: Innovations to Drive Profitable Growth*. It's free to IIF members and $125 for non-members, the price of a one-year membership. See our announcement on the inside front cover and review program and registration details at ***https://foresight.forecasters.org/2020-conference/.***

VIRTUAL FORESIGHT
PRACTITIONER CONFERENCE
December 2, 2020
Integrated Business Planning and Forecasting:
Innovations to Drive Profitable Growth

A Modern Retail Forecasting System in Production

PHILLIP YELLAND

PREVIEW *This is the third and final installment of an article documenting the large-scale demand-forecasting system developed by U.S. retailer Target. The first two articles in the series described the architecture and design of the system and recounted lessons learned in the development process. This last piece explores the issues that arise when a forecasting system such as Target's is actually deployed to provide forecasts for users and recommends steps that can be taken to address those issues.*

INTRODUCTION

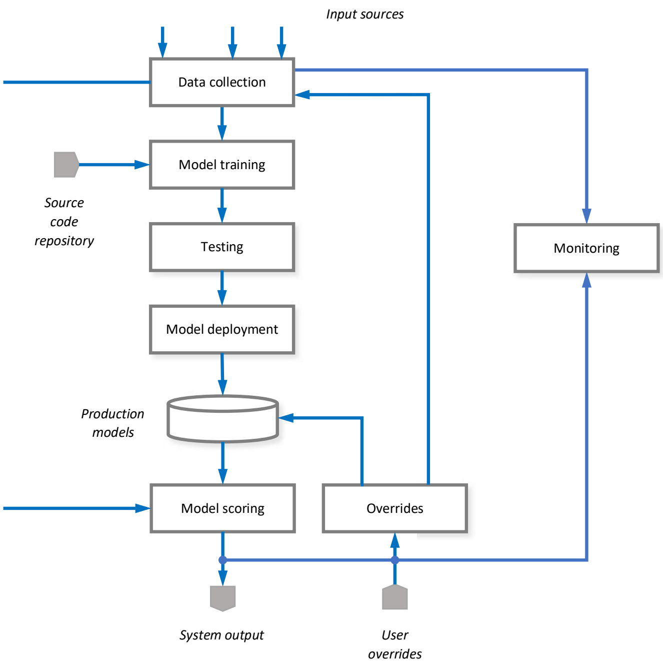
This is our third and final installment of a series for *Foresight* in which we document the Demand Forecasting Engine (DFE) project for Target. The DFE project's goal was to develop and implement a comprehensive forecasting system that could meet the need to provide nearly a billion *item* x *store* x *week* forecasts per week. Part one (Yelland and colleagues, 2019) described the overall architecture of the DFE system, while the second (Yelland and Erkin Baz, 2020) recounted some of the more important management and organizational lessons we learned in the development process.

Here we explore issues that arise when an automated forecasting system of the scale and complexity of the DFE is deployed into production—i.e., set up to provide forecasts on a sustained basis for users across the organization. The DFE system is at once a large-scale forecasting application, a complex software-engineering application with a significant statistical/machine-learning (ML) component, and an element of mission-critical business processes, so its deployment touches on a broad spectrum of topics.

We'll begin with an overview of the system's operation, presenting the main processes involved; these are explored in greater detail in the remainder of the article. Though essentially a description of the DFE system, the article also tries to draw general lessons and to make recommendations that might be applied to the production deployment of forecasting systems of similar scale and scope.

In fact, many of the topics discussed apply to the deployment of most modern, large-scale machine-learning systems. Where possible, therefore, we employ the vernacular of machine learning: in particular, we use the term *model* to refer to a collection of estimated parameters for a particular machine-learning model form, which may be stored and retrieved as required. A *model specification* provides a procedure for constructing a new untrained model. *Model training* is the

Figure 1. Production Schematic



process of model parameter estimation, and *model scoring* uses the parameter estimates in a model to produce forecasts.

Key Points

- Work on a large-scale, mission-critical demand forecasting system does not end with the development of an effective forecasting model: The system needs to be deployed *into production*, so that it can provide forecasts reliably on demand for users across the organization.
- A large-scale commercial ML project like the DFE relies on a wide range of data inputs, which usually originate in upstream reporting systems. Such input data must be procured and ingested reliably, week in and week out, with little manual attention. Automated monitoring of the quality of the input data is vital and failures in upstream processes must be noted, as well as anomalies such as missing data, out-of-range data, or changes in data distributions.
- For a production forecasting system, proper provision must also be made for user interaction with the system. Users should be able to obtain an understandable description of how any particular statistical forecast is derived by the system. If users can see which sales drivers figured into the system's calculation of a forecast, and the effect each driver had on the result, they are better placed to decide if any influences were omitted by the system, or if any effects should be increased or reduced.
- Since deployment is often complex and tedious, try to automate as much of the deployment process as possible. If practicable, try to automate estimation and testing, too, although human participation may be required to appraise the test results of a new model before its deployment.

SYSTEM OVERVIEW

Figure 1 is a schematic of the operation of the DFE system. The blocks in the figure represent processes carried out to make the system's forecasts available to its users. Arrows indicate that data or other artifacts output by one process comprise inputs to one or more other processes. The production models represent the results of the model training process and, after suitable testing, they are deployed in the scoring process to produce forecasts required from the system. Inputs and outputs of the system as a whole are denoted by flowchart-style connectors. The diagram is somewhat stylized, as the actual operation of the DFE in production is more complicated; nonetheless, the representation is sufficiently accurate for the purposes of our discussion below.

In the following sections, I examine each component process, exposing the challenges that arise in production and providing approaches to their resolution.

DATA COLLECTION

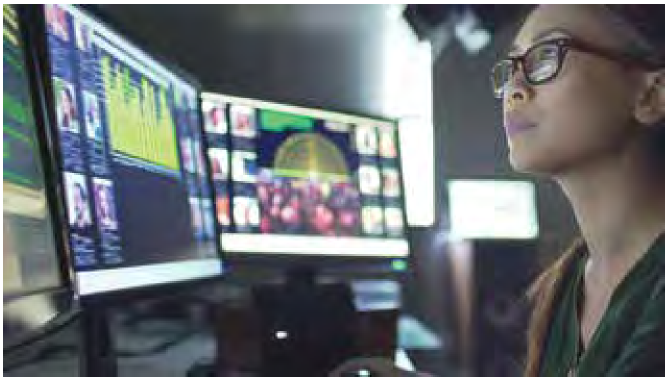
A large-scale commercial ML project like the DFE relies on a wide range of data inputs, which usually originate in upstream reporting systems. Point-of-sales data, for example, is used to compute unit sales and selling prices, stock-keeping data to track stock-outs, and the outputs of one or more promotion management systems for records of past and future promotions. In production, such input data must be procured and ingested reliably, week in and week out, with little manual attention.

Quality Assurance

In production, automated monitoring of the quality of the input data is vital. Failures in upstream processes must be noted, as well as anomalies such as missing data, out-of-range data, or changes in data distributions. We need to dispatch appropriate alerts when upstream problems are detected. (I offer further details of input-data monitoring in the *Monitoring* section below.)

The upstream systems may themselves produce alerts if they experience problems. These alerts should be propagated by the system's monitoring process if they have the potential to impact the system's functioning or performance. Also, consumers of the upstream data must be made aware of any notifications put out by the groups responsible for the corresponding reporting systems regarding breakages, scheduled downtime, schema changes, and the like.

As Sculley and colleagues (2015) have written, the responses of ML systems to changes in input data may be complex and counterintuitive. For example, an ML system may learn to compensate for a noisy or malformed input feature to the extent that improving the feature may cause the system to over-compensate and actually lead to a deterioration in its predictions. It is vital, therefore, to be aware of input changes as soon as possible, so that remedial actions such as model retraining may be taken.



Coverage

A particular problem facing the DFE is ensuring that the required forecasts are produced consistently. The DFE needs to forecast sales of products expected to be sold in the company's stores in given weeks. Unfortunately, the set of *item* x *store* x *week* combinations for which forecasts are required is by no means static: new items are introduced continuously, and items already selling in a subset of stores may begin selling in other stores. In addition, items may be eliminated from the product line or withdrawn from the selection at particular stores. Furthermore, the set of combinations must be determined not only in the current week but for all future weeks covered by the forecasts.

The process of determining the correct set of combinations normally involves reconciling several disparate upstream IT systems. Ensuring that the exercise can be reliably repeated over time in production can be a fearsome challenge. When embarking on a mission-critical ML project of a scale and scope comparable to that of the DFE, you can expect to devote similar effort to seemingly trivial data-related problems.

Training/Scoring Skew

Machine-learning applications are particularly vulnerable to the input data-related problem known as the *training/scoring skew*, in which input data used for training differs in some statistical sense from that used for scoring (forecasting).

For example, it is natural to use features representing promotions to predict sales; however, often, for a given product, several promotions of different types (a price reduction, for example, and an advertising placement) occur simultaneously. If historical records are accurate, therefore, the corresponding promotion features will be highly correlated and such *collinearity* can make the parameter estimates in a regression model unstable. Even so, the predictions from such a model should be reasonably accurate, provided that the features continue to be correlated in the same way as in the historical records: errors in the parameter estimates for the two features essentially cancel each other out.

To generate forecasts, however, data about *planned* promotions will be required. Such plans generally originate from different IT systems than the historical records, and there is a risk that the promotional signals do not co-occur in the plans as they do in the historical records. For example, plans for one type of promotion may be entered some time after those for another. Applied to *planned* promotions, therefore, the regression model of the previous paragraphs is liable to produce wildly inaccurate predictions.

The most straightforward remedy for the training/scoring skew is to bring the data sources for training and scoring into alignment. Unfortunately, this may be challenging, given the span of IT systems involved. More technically complicated approaches center on model training, such as the use of penalized/regularized regression methods (Zou and Hastie, 2005) to ameliorate problems with collinearity.

Feature Curation

When selecting features for a machine-learning system such as the DFE, the costs of data provision in production must be borne in mind. A feature that provides little improvement in the system's predictive accuracy should be eliminated if providing it in production is costly or complicated. Moreover, a feature's importance should ideally be considered in conjunction with others used by the system. One method of achieving this is to rank prospective features in order of "procurement cost" (often groups of features will share a rank as they are obtained as a group), looking at the additional accuracy conferred by increasingly expensive features.

MODEL TRAINING

The process of model training is central to the DFE, as it is to any machine-learning system. As illustrated in Figure 1, data used in the model training process is provided by the data collection process—but, additionally, inputs for model training include input specifications for the models to be constructed and trained. In our DFE, these specifications take the form of R scripts that construct definitions of *generalized additive mixed models*, which we discussed in the first installment of this article (2019).

Managing Model Specifications

Even after a major ML application is deployed, we should expect that model specifications will evolve over time in pursuit of improved forecast performance: new specifications will be added and old ones retired to accommodate changes in the business (new types of product, for example, or retirements of existing product lines). Given this likelihood of change and the pivotal role of models and model training, model specifications should be treated like any other vital software component of an evolving system: they should be properly documented, subject to code review, unit tested (see section Testing below) and committed to source-code control.

> **Statistical model development is subject to rapidly diminishing marginal returns in terms of performance. Therefore, a substantial improvement over a well-developed, mature model is unlikely to be feasible, or at the very least will require significant investment.**

Investment in Model Development

Given that refinement, extension, and replacement of model specifications is likely to be an ongoing activity, it is important to prioritize investment in model development. Several considerations enter into such decisions:

- *The value of a good forecast to the business.* The archetypal approach to assessing the value of a product-demand forecast, of course, is ABC inventory analysis (Wild, 2017, ch. 3): products in the high-value "A" class generally merit more accurate forecasts than those in classes "B" and "C," and those in class "B" more accurate forecasts than in the low-value "C" class. Assuming increased forecast performance is actually feasible, therefore (see bullet point below), additional model development effort may be merited for the higher-value classes. Performing an ABC analysis of the product line—and keeping it current in the face of product-line revisions—thus provides a foundation for the prioritization of model development. It may be the case that one forecast model has acceptable accuracy when applied to products in all classes, and so special model development effort is not required. In this case, consider the computing costs associated with training and scoring that model (rather than a simpler one) for lower-value products (which are usually much more numerous than the high-value ones), and the effort that adjusting lower-value product forecasts would require of forecast analysts.

- *The likely return on development effort.* As a rule, statistical model development is subject to rapidly diminishing marginal returns in terms of performance (Makridakis and colleagues, 2018). Therefore, a substantial improvement over a well-developed, mature model is unlikely to be feasible, or at the very least will require significant investment.
- *The cost in production.* As well as being more challenging to develop, debug and explain, complex models usually require more compute time and/or memory in production to train and score than simpler models, and more storage space, too. Also, if a new or revised model needs an additional feature, the cost of obtaining that feature should be considered (see *Feature Curation*, above).

In weighing decisions concerning model development, bear in mind that improved input data processing or a better understanding of the relevant business processes may yield greater value for an equal or even lesser effort.

Model Metadata

When errant forecasts are observed in production, it is often necessary to conduct *root-cause analysis*; i.e., to trace the ultimate causes of the faulty prediction. To support this analysis, since forecasts are derived from trained models, it is useful to store *metadata* with the models. A model's metadata set records its salient attributes, such as the source-code version of its specification, the author of the specification, an identifier for the data set used to train the model, and so on. A number of recently introduced general purpose frameworks such as MLFlow (Databricks, 2020) provide mechanisms for recording model metadata (although as work began on the DFE project when such frameworks were largely immature, we found it more expedient to develop a custom facility).

TESTING

Testing plays an essential role in the operation, maintenance, and enhancement of a machine-learning application in production. In the DFE project, we have found it useful to distinguish two types of tests:

- *Functional* tests establish that the system operates correctly, in that it maps inputs to outputs without errors or crashes.
- *Forecast performance* tests, on the other hand, attempt to estimate the accuracy of the system's predictions or inferences.

I've used the term *forecast performance* to distinguish the testing discussed in this section from testing that analyzes a system's *compute* performance—it's responsiveness, run-time, memory requirements, etc. Compute performance is an important concern with any large-scale software system (Molyneaux, 2014) like the DFE. For the most part, however, compute-performance tests feature less prominently in our testing infrastructure than functional and forecast performance tests, as concerns of this sort apply most particularly to the *Model Scoring* subsystem (see below).

Functional Tests

Functional tests may take the form of *unit tests*, carried out on isolated components of the system, or *integration tests*, which test several components in combination, ranging from small assemblies of two or three components up to the entire system.

Good software engineering practice recommends that unit tests be formulated for as many of the system's components as is practicable. Techniques by which components can be unit tested in isolation are well established (Koskela, 2013).

As with many ML systems, much of our DFE software is structured as pipelines, with constituent stages performing data extraction and transformation, model estimation, scoring, and so on. By developing a mechanism for inserting data into and extracting intermediate results from a pipeline stage, it is straightforward to put together integration tests for sequences of pipeline stages.

Since the quality of predictions is not an issue in functional tests, realistic data is not generally required, so "fake data" that conforms with input requirements can be used. Using methods from fuzz testing (Godefroid and colleagues, 2008), such fake data can be produced automatically given a suitable specification.

If users are interested in forecast performance under certain conditions, it is often possible to prepare a test data set that exemplifies those conditions using actual historical data, synthetic data, or a combination. Users may inspect the results of tests on these data sets to gauge prospective system performance under their specified conditions.

Forecast Performance Tests

Our performance tests for the DFE system use historical data to perform *rolling window cross-validation* (variously known as *sliding window validation* or *walk-forward validation*). We train models on historical records up to a chosen past date and then score those models using the records after that date (the "holdout" records). Forecast-error metrics are then calculated by comparing the forecasts with the corresponding observed sales in the holdout records. Metrics are computed for horizons up to 52 weeks, and forecast horizon is among the dimensions along which metrics are "sliced and diced" for appraisal (see below). This procedure is repeated several times, advancing the chosen date by a week before each successive iteration, generating a series of forecast metrics.

We keep a standard set of tests available for model validation, based on selected historical data sets of various sizes. Our practice is to run tests with data sets of increasing size, with the smallest completing in no more than 10 minutes on a single CPU and the largest comprising a substantial proportion of the (historical) production data. This way, if a new model or other system change results in markedly inferior forecasts, we can detect this with only modest testing effort using the smaller data sets. Conversely, the larger data sets, though they require greater resources, allow for more refined performance measurements.

We have found it particularly useful to produce metrics comparing the performance of prospective models with simple baseline models or legacy forecasting systems (where available). We use versions of the *mean absolute scaled error metric* of Hyndman (2006) for comparisons with baseline "naïve" models (which simply use selected historical sales as forecasts), and *relative absolute error* measures (Fildes, 1992) for legacy comparisons. Check with your stakeholders, however, before selecting user-facing metrics (see section *Metrics to Monitor Inputs and Forecast* below); it may be that a graphical comparison of the new model's performance with that of the baseline or legacy system is more accessible to users than metrics that may be unfamiliar to them.

Test Appraisal

As emphasized in the first two articles in this series, evaluating forecast performance is often a complicated process—rarely is it a simple matter of appraising one- or two-figure summaries. Models superior in certain circumstances may be inferior in others, and some users may be more sensitive to forecast performance at particular times of the year (holiday season or during promotions). Often, therefore, it is necessary to have stakeholders participate in the appraisal of the metrics produced by pre-deployment tests. This can be facilitated by providing visualizations of the metrics and allowing them to be "sliced and diced" by the stakeholders—see *Monitoring*, below, for further discussion.

To further stakeholder participation, we have found it helpful to prepare test data sets that specifically reflect user concerns. If users are interested in forecast performance under certain conditions, it is often possible to prepare a test data set that exemplifies those conditions using actual historical data, synthetic data, or a combination. Users may inspect the results of tests on these data sets to gauge prospective system performance under their specified conditions.

MODEL DEPLOYMENT

Once models have been estimated, tested, and deemed suitable for providing forecasts, they are *deployed*—i.e., made available to the system's scoring process (see next section), which in turn uses them to provide the forecasts to users. Deployment is a central concern of the IT discipline *DevOps* (Bass and colleagues, 2015), and a number of recommended deployment practices have been developed in that field. For our DFE project, we found the following practices particularly useful:

- Since it is often complex and tedious, try to automate as much of the deployment process as possible. While moving newly trained models manually to scoring might be entertained on a one-off basis, it is far less practical to do so continuously in production. If practicable, try to automate estimation and testing, too, although human participation may be required to appraise the test results of a new model before its deployment.
- In a mission-critical setting, it pays to be cautious when deploying new models. DevOps practices that help defuse the risk that inevitably attaches to new models include *tiered releases*, in which new models are released into scoring environments that mirror the one actually serving users, but in which requested forecasts are delivered only to monitoring, not to users; *canary releases*, which use new models to serve only a fraction of the requested forecasts (requests can be partitioned by product, by user, location, etc., and the fraction may grow over time); and *rollbacks*, which reserve old models so that they may be restored in case of problems with the new ones.

MODEL SCORING

In common with many production ML systems, our DFE separates model training (the calibration of models using historical data) from model scoring (producing forecasts with trained models). Doing so introduces architectural complications, but it has the critical benefit that model training—frequently an expensive and relatively time-consuming operation—does not impinge upon the user's experience. This is because the user's forecast requirements can be fulfilled in the scoring process using pretrained models, and the latter can be carried out much faster and more cheaply than training those models in the first place.

Speedy Scoring

If the benefits of separate training and scoring are to be fully realized, the scoring process should be as efficient as practicable. This may mean using a different programming language for scoring than that used for training. In the DFE, for example, we use R for training models, as it offers extensive facilities supporting our chosen model form. Scoring, however, is carried out in Scala—an efficient language well matched to the Apache Spark platform we use for large-scale data processing. Though beneficial, this approach is not without costs: it can make the system's code base less accessible to programmers, requires that objects be stored in a language-independent format, and can lead to functionality reimplemented in different languages.

Graceful Degradation

Since the scoring process is the crux of the user's experience, it is imperative that scoring continues to provide forecasts (if less-than-perfect ones) under a wide variety of error conditions—a quality known as *graceful degradation* (Herlihy and Wing, 1991).

Consider producing a forecast for a given item, store, and week, where the week in question is at forecast horizon h. The scoring process might employ a succession of fallback strategies like the following, where each successive strategy is tried should all the preceding ones fail:

- Use the most recently trained, most applicable model to produce the forecast, based on the most recent scoring data (promotion plans, etc.).
- If an older version of the model is available (see the discussion above of rollbacks), consider using it in conjunction with the most recent scoring data.
- If a forecast is available that was prepared earlier for the requested item, store, and week, but with a horizon greater than h, then consider reusing it.
- If a coarser-grained model or forecast is available that—with disaggregation, perhaps—may be used to provide the forecast required, consider using it.
- Use a simple "baseline" forecast, such as sales of the item during the corresponding period last year.

Metadata

As with model training, the scoring process should deliver metadata along with the forecasts it produces. The metadata might include the time and date on which the forecast was computed, identifiers for the model and data used to produce it, which fallback mechanism—if any—was invoked, and so on. We have found such information indispensable in diagnosing and correcting any problems detected in monitoring or by the users of the DFE system.

JUDGMENTAL ADJUSTMENTS

The literature concerning the overall effectiveness of judgmental adjustments to statistical forecast is somewhat equivocal (Lawrence and colleagues, 2006), but there does appear to be a consensus that judgment-based adjustments to statistical forecasts can be very valuable on occasion, particularly when those people making the adjustments possess information about determinants of future sales that are not accounted for by the statistical model. Such information might comprise novel events such as recent natural disasters or other sales drivers for which reliable signals are unavailable, or which have not been incorporated into the model.

The Need for Adjustments

Manual inputs are also occasionally required in cases of forecast failure, when—due to a problem in the system that is inadequately addressed by the monitoring and recovery mechanisms—some forecasts produced by the system are clearly inaccurate.

Human intervention is helpful, too, in removing or down-weighting historical data which there is good reason to believe is not representative of likely future sales patterns. For example, the occurrence of a natural disaster may require not only adjustment to forecasts to account for unanticipated shifts in sales, but also removal of those aberrant sales from future training data, as it is unlikely that they will repeat in the normal course of events.

Managing Adjustments

While manual adjustments are arguably important to the success of a large-scale forecasting system, such adjustments should be carefully controlled and managed (Fildes and Goodwin, 2007). Otherwise, human cognitive distortions, political pressures, or a simple desire to contribute may result in interventions that do more harm than good.

We have found several measures effective in moderating manual adjustments:

- Most importantly, users should be able to obtain an understandable description of how any particular statistical forecast is derived by the system. In particular, if users can see which sales drivers figured into the system's calculation of a forecast, and the effect each driver had on the result, they are better placed to decide if any influences were omitted by the system, or if any effects should be increased or reduced. In the first installment, we discuss the importance of explicability to the design of the DFE, and the facilities provided to allow user inspection of the system's forecasts.

- Determined efforts should be made to educate users in effective forecast adjustment. They can be taught techniques for avoiding cognitive biases (Harvey, 2001), the dangers of over-adjustment, and so on. And as Fildes and colleagues (2009) observe, small judgmental adjustments tend to decrease forecast accuracy, so users should be encouraged to avoid them. It is often advisable to restrict permission to adjust system forecasts to a select group of users with the requisite training.
- Track user overrides and measure their effect on forecast accuracy. Records and summaries of their past interventions often provide valuable feedback to users. Be aware, however, that some users may perceive such information as a potentially prejudicial appraisal of their job performance and it should be handled with sensitivity—see the discussion in Fildes and colleagues, 2009.

Monitoring user interactions with the system is also advisable—recording adjustments made to forecasts is particularly valuable, but recording other interactions, such as bug reports, use of any user interface components, API calls, and so on, is also helpful.

Other Feedback

Feedback from users and stakeholders about the system should not be restricted to forecast adjustments alone. Develop formal channels (ideally supported by software) for other forms of feedback, such as bug reports (which may comprise reports of errant forecasts), requests for expanded explanations, requests for enhancement (particularly for forecast improvements in specific circumstances), etc.

MONITORING

Continuous monitoring is vital to the proper functioning of a production forecasting system. As suggested in the section on *Data Collection* above, monitoring needs to be applied to both the inputs of the system and its outputs. Monitoring user interactions with the system is also advisable—recording adjustments made to forecasts is particularly valuable, but recording other interactions, such as bug reports, use of any user interface components, API calls, and so on, is also helpful. In all cases, suitable metrics should be computed and displayed, and alerts are raised when specified conditions occur.

Production monitoring of the type described in this section has long been a focus of study in the field of *statistical process control* (SPC), and a recent article by Katz (2020) provides a very apposite discussion of the use of SPC techniques for monitoring forecast systems.

Metrics to Monitor Inputs and Forecast

Choosing metrics for monitoring input data is reasonably straightforward. For each individual input feature, they might comprise: missingness counts; for numerical features, selected quantiles including maximum, minimum, and median, together with mean and variance; for categorical features, counts of distinct values, maximum and minimum relative frequencies. Low-order combinations of features can be summarized by counting distinct combinations of categorical features (along, with measures of association, such as conditional entropy), and correlations of numerical ones. Look for training/scoring skew (see *Data Collection*) by comparing corresponding metric values for training and scoring data sets.

Choosing metrics to monitor forecasts is less clear-cut. A vast array of forecast accuracy metrics appears in the literature (Wallström and Segerstedt, 2010 and Shcherbakov and colleagues, 2013), each with differing degrees of familiarity, intuitive appeal, and technical merit. Ideally, metrics should be chosen in consultation with the stakeholders—aim to select a small set that combines accessibility and technical quality. While it is conventional to compute aggregate accuracy measures using averages (usually arithmetic or geometric) of individual measures, other summaries such as selected quantiles should also be calculated.

In addition to computing metrics for a particular period, compare corresponding metric values across time periods to detect any deterioration in forecast performance or increase in the fraction of missing values in a feature. As with metrics in testing, it is also useful to allow users to view metrics by slice—forecast performance for products by supplier, for example, or stores by region.

Alerts are commonly generated by comparing metrics or anomaly scores derived from metrics with fixed thresholds; these thresholds should be chosen to keep the number of alerts generated within acceptable bounds.

Operational Metrics

In addition to metrics for the input and output of the system, metrics providing information about the operational state and trajectory of the system itself are essential. They might include the age of estimated models in the system, the frequency of input data refreshes, counts of error conditions encountered during training and fallbacks during scoring. More exotic metrics like the relative importance of input features can also shed light on the functioning of the system.

Testing

There are obvious parallels between the testing and monitoring processes in the DFE; in many respects, testing seeks to provide a preview of the monitoring results for a model before its deployment. This means that much of the software infrastructure used for monitoring can be shared with testing, so that—for example—metric calculations are implemented only once. It is also a good idea to check regularly that testing results prior to deployment correlate appropriately with monitoring results afterwards.

Anomalies

Anomaly detection—the identification of data points which deviate from the norm—plays a prominent role in monitoring a system like the DFE, both to detect problems in the system's input and output data and to identify service problems reflected in the operational metrics.

Fortunately, anomaly detection is an area of extensive historical and ongoing research, and many techniques can be applied more or less off-the-shelf (Chandola and colleagues, 2009). If development time and resources are limited, a basic univariate anomaly detector (ideally one capable of handling time series) will provide the greatest return on investment.

Anomaly detection for metrics facilitates management by exception—the principle that operator intervention should be required only under exceptional circumstances—all but essential for a system as large and complex as the DFE. In essence, this means that operators are alerted only when anomalies are detected in metrics.

Regarding anomalies in input and output data, the principle of graceful degradation highlighted above should apply; where possible, consider correcting data anomalies as they are detected. In input data, this normally means removing the anomalous observation; in output data (and sometimes in input data, too), reasonable corrections can be made by fitting a simple model to the data and using it to impute a value.

Dashboards and Alerts

There are two main conduits for monitoring output: *dashboards*, which provide graphical displays of summary metrics, and *alerts*, which inform operators of conditions that merit attention. Both types are common in modern large-scale software systems, and there are many products—both open-source and commercial—that facilitate their implementation; for a popular example, see Grafana (2020).

Since alerts consume operator attention—a limited resource, and one apt to be diminished by unnecessary calls on it—they merit careful management. Alerts are commonly generated by comparing metrics or anomaly scores derived from metrics with fixed thresholds; these thresholds should be chosen to keep the number of alerts generated within acceptable bounds.

CONCLUSION

This third part of the article has described the machinery of the Demand Forecasting Engine in production and highlighted issues associated with its ongoing operation and maintenance. As observed in the introduction, while the discussion has centered on the DFE, many of the same issues would arise when operating any forecasting system with similar size and organizational role.

A system like the DFE is intended to be a critical, long-lived asset of an organization, evolving with the organization's needs. As such, it is never actually "finished": in addition to operation and maintenance, the system will likely be extended, revised and (at least to some extent) rearchitected even after entering production—in all probability, by substantially the same team responsible for its development. Hence the lessons (architectural, organizational, and operational) recounted in all the articles in this series should remain relevant throughout the life of the system.

REFERENCES

Bass, L., Weber, I. & Zhu, L. (2015). *DevOps: A Software Architect's Perspective*, Addison-Wesley Professional.

Chandola, V., Banerjee, A. & Kumar, V. (2009). Anomaly Detection: A Survey, *ACM Computing Surveys (CSUR)*, 41(3), 1-58.

Databricks Inc. (2020). *MLflow Documentation*, **https://mlflow.org/docs/latest/index.html**

Fildes, R. (1992). The Evaluation of Extrapolative Forecasting Methods, *International Journal of Forecasting*, 8(1), 81-98.

Fildes, R., Goodwin, P., Lawrence, M. & Nikolopoulos, K. (2009). Effective Forecasting and Judgmental Adjustments: An Empirical Evaluation and Strategies for Improvement in Supply-Chain Planning, *International Journal of Forecasting*, 25, 3-23.

Godefroid, P., Levin, M.Y & Molnar, D.A. (2008). Automated Whitebox Fuzz Testing, In *Network and Distributed System Security Symposium*, 8, 151-166.

Grafana (2020). Grafana Labs website, **https://grafana.com/**.

Harvey, N. (2001). Improving Judgment in Forecasting, In *Principles of Forecasting*, Springer, Boston, MA, 59-80.

Herlihy, M.P. & Wing, J.M. (1991). Specifying Graceful Degradation, *IEEE Transactions on Parallel and Distributed Systems*, 2(1), 93-104.

Hyndman, R.J. (2006). Another Look at Forecast-Accuracy Metrics for Intermittent Demand, *Foresight*, Issue 4, 43-46.

Katz, J.H. (2020). Monitoring Forecast Models Using Control Charts, *Foresight*, Issue 56, 20-25.

Koskela, L. (2013). *Effective Unit Testing*, Manning.

Lawrence, M., Goodwin, P., O'Connor, M. & Önkal, D. (2006). Judgmental Forecasting: A Review of Progress over the Last 25 Years, *International Journal of Forecasting*, 22, 493-518.

Makridakis, S., Spiliotis, E. & Assimakopoulos, V. (2018). Statistical and Machine Learning Forecasting Methods: Concerns and Ways Forward, *PloS one*, 13(3).

Molyneaux, I. (2014). *The Art of Application Performance Testing: From Strategy to Tools*, O'Reilly Media, Inc.

Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.F. & Dennison, D. (2015). Hidden Technical Debt in Machine Learning Systems, In *Advances in Neural Information Processing Systems*, 2503-2511.

Shcherbakov, M.V., Brebels, A., Shcherbakova, N.L., Tyukov, A.P., Janovsky, T.A. & Kamaev, V.A.E. (2013). A Survey of Forecast Error Measures, *World Applied Sciences Journal*, 24, 171-176.

Wallström, P. & Segerstedt, A. (2010). Evaluation of Forecasting Error Measurements and Techniques for Intermittent Demand, *International Journal of Production Economics*, 128(2), 625-636.

Wild, T. (2017). *Best Practice in Inventory Management*, Routledge.

Yelland, P., Erkin Baz, Z. & Serafini, D. (2019). Forecasting at Scale: The Architecture of a Modern Retail Forecasting System, *Foresight*, Issue 55, 10-18.

Yelland, P. & Erkin Baz, Z. (2020). Developing a Modern Retail Forecasting System: People and Processes, *Foresight*, Issue 57, 27-38.

Zou, H. & Hastie, T. (2005). Regularization and Variable Selection via the Elastic Net, *Journal of the Royal Statistical Society*, Series B. 67(2), 301-320.

Phillip Yelland is Principal Data Scientist at Target Corporation. See his "Forecaster in the Field" interview in our Fall 2019 issue.

Phillip.Yelland@target.com

Commentary: It′s the Soft Problems that Are Hard to Overcome

SIMON CLARKE

The challenge of creating a large-scale demand forecasting system capable of developing hundreds of millions of forecasts is brought to life in the excellent contributions by Phillip Yelland, Zeynep Erkin Baz, and their team at Target. While I agree with all of the important lessons, it does strike me that the most difficult and persistent challenges to overcome relate to "soft" problems, such as achieving organizational support for an initiative or obtaining a clear brief on what is required of a solution. I feel more confident that "hard" problems, such as data cleanliness and availability, will be either mitigated or improved with advances in data engineering and application of self-healing data-handling routines.

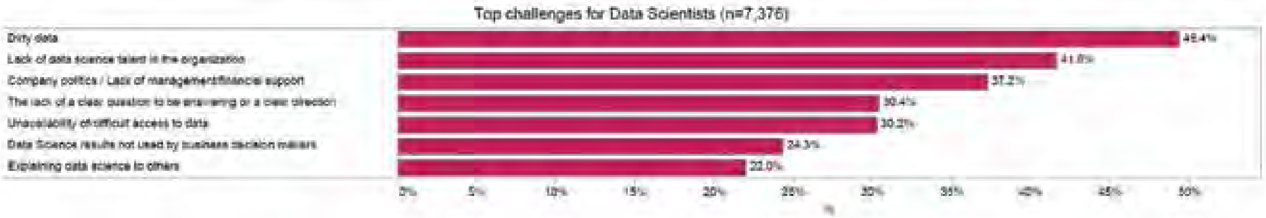
Soft problems are commonly more difficult to solve, in part because they are pluralistic. How to achieve the end state includes many different options that can involve negotiation, experimentation, and ongoing learning. They are difficult to solve because they involve people with their in-built biases and preferences. In the data-science domain, I fear without fundamental change these issues will linger and will be a brake on the development of the practice.

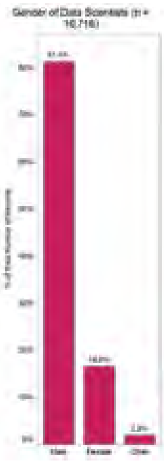
The findings of the 2017 Kaggle "State of Machine Learning and Data Science Survey" echo many of the challenges described in the Target project. Of the 7,376 data scientists who were asked to list their biggest challenges at work, a large number pointed to soft problems. These included getting clear work briefs, support for the work itself (through funding and/or talent acquisition) and adoption. In fact, of the top eight issues, only two related to data, albeit significant ones. The themes that seem to emerge are as follows.

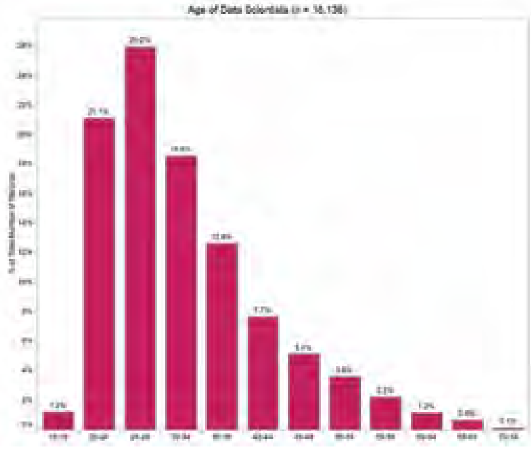
ORGANIZATIONAL SUPPORT

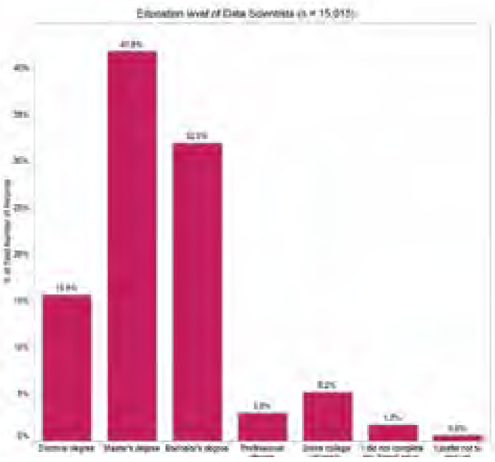
While many organizations understand its importance, data science remains something of an esoteric activity. As a result, it doesn't command the organizational support that other more established functional areas have. It shouldn't be a surprise that practitioners point toward inadequate funding, shortfalls in talent, and a lack of tangible commitment from their employers.

We hear of organizational frustration that the data-science team doesn't understand the business and fails to appreciate subtleties of how things work. In addition, there is the reality that data scientists are expensive—the 2019 Kaggle survey revealed that, in the U.S., the majority of data scientists earn between $100k and $200k per annum. Meanwhile, their business benefit is often more intangible than the investment in something physical, like a piece of equipment. Their value proposition is made even more tenuous with each piece of work that is not adopted by business leaders.









STAKEHOLDER ENGAGEMENT

In a data-science project, a crisp problem statement is essential—too much ambiguity, and the solution stands a high chance of missing the mark. The development of the problem statement requires stakeholders to be very specific about why a simple solution has not already succeeded. Both the Target project and Kaggle survey responses suggest that getting that engagement is elusive.

A common perception is that data scientists don't speak the language of the business community. This may be because they haven't a detailed understanding of the business or fail to make their points understandable to stakeholders. Data scientists have great pride in their hard-earned skills and their grasp of often very complicated topics. Unfortunately, some also want to demonstrate them in full. This can be disastrous in interaction with stakeholders.

ADOPTION

Despite the investment in technology and people capability, many tools are either rejected or adopted halfheartedly, creating a vicious circle. Organizational support is hard to come by because of a lack of adoption—but for solutions to be fully adopted, organizational support is required.

One big challenge is in explaining how solutions work, what assumptions were made, and what uncertainties remain. A business leader may find the answers provided to be technically accurate but fail to address the underlying concern. Some data scientists fear that simplification could undervalue their contribution and gloss over features that they believe are critical to understanding.

So, how should these issues be addressed? The Target team makes make some excellent suggestions, but I think there are more deep-seated issues.

Broaden the Diversity of the Data-Science Community

The 2017 Kaggle study reveals that the composition of the more than 16,000 respondents from around the world is overwhelmingly male, youthful, and highly educated.

Over 80% are male and over 65% are under 34 years old. Over 55% have either a master's or doctoral degree. The age of data scientists perhaps shouldn't be a great surprise given the recent growth of the practice, but what is alarming is the male bias. It is unlikely most businesses will be as dominated by young, educated males.

It is important for the data-science community to understand how gender, age, and education affect how it communicates and engages with stakeholders. If we are to bridge the gap with the business community, every effort should be made to broaden diversity in the field. Firms should work to identify, recruit, and train talent from within the organization, creating teams that integrate business and data-science skills

Clearly Define the Role of the Data-Science Team

Leaders often find it difficult to distinguish a data-science team from resources in other functional areas working to solve problems through data analysis. This can lead to a competition for resources ("dueling analytics"), a lack of effective leverage of expertise, and ambiguity over budget and other ongoing responsibilities.

The mission statement must clearly define the difference between data science and business analytics, and where the fault lines exist between them. For example, the data-science team may be responsible for analytics that have enterprise-wide, cross-functional reach, including both structured and unstructured data, while the business analytics teams may be assigned to work with structured-data sources in a more narrowly defined context (such as price strategy or assortment planning). Potential project overlap should be pinpointed, and governance put in place to avoid duplication and optimize resource allocation. There should be no ambiguity around who has budgetary responsibility for the ongoing ownership and maintenance of tools and insights.

Organizations having significant challenges in bridging the divide between the data-science and business communities should consider an embedded or deployed model. These structures can cement the accountability of business leaders to the successful adoption and delivery of projects.

Data science is a means to an end, with the role to support the business and focus on the development of working solutions that deliver value. Productivity should be prioritized over perfection.

Consider the Most Effective Organizational Design

Organizational designs include the centralized, embedded, and deployed models.

Centralized, the most common design, has the data-science team organized within a single structure, and the leadership of this group determines what projects to prioritize and resources to apply. It offers the benefit of scale, flexibility to adjust as demands change, and motivation from the variety of team members. Its weakness is that the team is removed from the business unit, making stakeholder engagement challenging.

Embedded data-science teams are fully entrenched in business units, this client proximity promoting responsiveness to the needs of the business and increasing the likelihood that the deliverable is accepted and adopted. The risk is that the work is compromised and lacks objectivity. There may also be a tendency to focus on suboptimal "pet projects."

Deployed is a hybrid of the centralized and embedded models. Members report to a centralized data-science leader but are embedded in the business units, potentially combining the advantages of the two prior models. This matrix organization, however, can be problematic if team members are unsure to whom they are accountable and thus be caught in organizational power struggles.

Organizations having significant challenges in bridging the divide between the data-science and business communities should consider an embedded or deployed model. These structures can cement the accountability of business leaders to the successful adoption and delivery of projects.

Create the Right Mix of Roles on the Team

Data scientists require a combination of domain expertise with math and computer-science skills. Domain knowledge provides an understanding of the context in which the problem resides, mathematics provides a theoretical foundation for how problems are examined, and computer science produces an understanding of the data products and solutions that are available. All these skills are hard to find in any one individual, so there is a need for specialized roles within data-science teams.

The Target team emphasizes the importance of agile data science in which each team has a *project manager* who is highly skilled at managing in "gray" space where constant readjustments are required in the face of unexpected discoveries and analytical dead ends. *Data engineers* are specialists needed to unify data found in a variety of formats, possibly in multiple databases. *Data analysts* should focus on identifying relevant data sources, preparing analyses, and building the business case for action. Target has elevated this last specialty to "Polymath Pioneer"—a team member with command of both data science (mathematics) and software engineering (computer science). Finally, the team needs a business integrator to clarify the specific objective of the data-science team and navigate the operational realities that will make the results actionable.

Focus on Communication

Messaging from the data-science team must be tailored in ways that educate but do not bury business users under obscure terminology and technical intricacies. Conveying the message effectively can be achieved by creating narratives, diagrammatic representations, and visualizations, all of which can humanize what could be inaccessible to those without mathematical or computer-science backgrounds.

Build Credibility

The objective of the data-science teams should be on delivering quick wins for the organization and building the base of work to larger and more ambitious projects. Providing confidence that value can be derived from data science is the first step to securing additional resources and organizational commitment. It can also help to reduce the likelihood that functional areas "go rogue," working on their own data projects without collaboration with the data-science team.

Simon Clarke is a Principal of Crimson & Co North America and formerly Group Director of Forecasting at Coca-Cola.

simon.clarke@crimsonandco.com

Response to Commentary of Simon Clarke

PHILLIP YELLAND AND ZEYNEP ERKIN BAZ

Our sincere appreciation goes to Simon Clarke for his thoughtful commentary on our articles. We agree wholeheartedly with his basic thesis: that social and organizational problems often dominate technical issues in determining the success or failure of data-science efforts. We also endorse the remedies he suggests and indeed have used a subset of them in our own work.

Our only caveat is that, as with most "soft" problems, the choice of an effective solution is contingent on the particularities of the problem's context, and that effective solutions bring costs with them. The need for clear expression of these considerations is precisely what drove us to develop the pattern language approach we employed in Part 2 of our article (Spring 2020 issue). We have now taken the liberty of restating Simon's recommendations in this form—**Table 1**.

RECRUITING

Staff for Success

Context A data-science team considering staffing decisions.

Problem Hiring exclusively for "data science" skills (mathematics/statistics, computer science, etc.) may be expensive and frequently leaves a team without capabilities vital for success.

Table 1. Pattern Summary

Problem area	Problem	Pattern
Recruiting	Hiring data scientists exclusively can be expensive and results in skill gaps.	*Staff for Success*
	The demographics of a data-science team may be radically different from those of the wider organization.	*Busting the Brotherhood*
Organization	Providing data-science services economically.	*Centralized Data Science*
	Enhancing interactions between data scientists and business.	*Embedded Data Science*
	Balancing economy against interaction with the business.	*Deployed Data Science*
	Fitting a data-science team into the business organization.	*Fuzzy Boundaries*
Operations	Overcoming skepticism about the potential of data science to deliver value to the business.	*Show Me First*
	Difficulties in communication between data scientists and the business.	*Failure to Communicate*
	Data scientists pursue complexity for its own sake.	*Galloping Complexity*

Forces

Comprehensive capabilities: Data-science skills are obviously indispensable, but to be truly productive the team needs other skills (project management expertise, business knowledge, and so on).

Hen's teeth: Insisting that all team members have the full spectrum of hard-to-find skills increases the cost and difficulty of recruiting.

Solution

Staff roles that are complementary to the data scientists on the team. Important examples include: project manager, responsible for drafting and maintaining specifications, planning work, etc.; data engineer, helping to scale solutions by the team and put them into production; data analyst, identifying data sources, preparing data analyses, assisting in building business cases for data-science efforts, and so on; business integrator, bridging the gap between data scientists and the business, preparing the initial problem statement and defining project goals for the team.

Consequences

Benefits

- Team possesses skills that are often pivotal in delivering effective solutions.
- Non-data-science-related activities of the team receive appropriate attention and effort.
- Easier hiring.

Liabilities

- Data science is clearly central to a team, and too much emphasis on ancillary roles may result in a lack of focus.
- Assigning responsibility for certain activities to specific roles may cause other team members to avoid involvement, even tangentially. Worse, if a designated role has not been filled, these activities may be neglected altogether.

See also Failure to Communicate

Busting the Brotherhood

Context A data-science team working with a broader business organization.

Problem The tendency for the demographic makeup of teams to skew young and male, posing an impediment to working with more diverse organizations.

Forces

Tech's imbalance: Situated within the panoply of IT-related professions, data science suffers from the demographic imbalance that affects these professions(workforces predominantly young, male, with advanced degrees).

Communication barriers: Lack of diversity in teams can constitute an impediment to communication and interaction with the broader business, whose demographic makeup may differ materially.

Hen's teeth: Insisting that all team members have the full spectrum of hard-to-find skills will increase costs and difficulty of recruiting.

Solution Consider adding recruits from the business organization itself to the data-science team to increase the latter's diversity.

Consequences

Benefits

- A data-science team that more closely resembles the organization it works in enhances trust and communication.
- Recruits from the business are likely to have useful operational knowledge.
- In-house recruits often retain links to their former groups, which may also foster enhanced communication.
- A more diverse team has the potential to produce more creative and effective data-science solutions.

Liabilities

- Effectively coordinating a diverse team with a range of backgrounds and professional experience can be challenging.

- Other functions in the business may resent what they perceive as poaching of staff.

See also Failure to Communicate (regarding business integrator)

ORGANIZATION

Centralized Data Science

Context Assembling data scientists to work with a broader business organization.

Problem Seeking a straightforward and economical means of providing data-science services.

Solution Set up a single team within the organization. Leadership of this group chooses projects to pursue and allocates resources accordingly.

Consequences

Benefits

- This is a fairly simple organizational structure to institute.
- Concentrating all data-science resources together may yield economies of scale and scope.
- An independent data-science business function may be able to prioritize efforts that deliver value to the whole organization, rather than catering exclusively to particular units.
- Centralized decision making can be quick to respond to changing demands.
- The variety of work undertaken by a large group can provide motivation for the scientists within it.

Liabilities

- Interaction with the business can be challenging from inside a large data-science team.
- Business leaders may resent feeling they are competing for the attention of a single team.

Embedded Data Science

Context Assembling data scientists to work with a broader business organization.

Problem How can interactions between such scientists and the business units be promoted?

Solution Embed data scientists within the business units, reporting to the leadership of those units.

Consequences

Benefits

- Data scientists who are actually members of the business units are likely to be more aware of the business's requirements.
- Business units may be more receptive to the efforts of data scientists who work for them.

Liabilities

- If data scientists identify too closely with the business units they work for, their work may be compromised and lack objectivity.
- Data scientists may be focused on "pet projects" of business units that are of dubious wider value.

Deployed Data Science

Context Assembling data scientists to work with a broader business organization.

Problem Is there a way to balance independence and efficiency against interaction with the business units?

Solution In a hybridized arrangement of Centralized Data Science and Embedded Data Science, embed data scientists with the business units, having them report to a centralized data-science leadership.

Consequences

Benefits

- Can combine some of the positive aspects of Centralized Data Science and Embedded Data Science: independent and efficient data-science decision making, together with increased rapport with the business units.

Liabilities

- The pathologies of "matrix management": confused and ambiguous accountability, and proliferating power struggles.

Fuzzy Boundaries

Context A business organization trying to accommodate a data-science team productively.

Problem It can be difficult to fit a data-science team into established business organizations.

Forces

Esoterica: Business people may perceive the subject matter of data science (mathematics, statistics, advanced algorithms, etc.) as esoteric and irrelevant to their needs.

Turf battles: Central preoccupations of data science—software development and business analytics—frequently overlap the responsibilities of existing business departments. This can lead to duplication of effort, budget and headcount conflicts, and even "rogue" efforts to preempt or undermine data-science initiatives.

Production orphans: Organizations may fail to properly assign responsibility for production deployment and maintenance of systems developed by the data-science team.

Solution

Devote time and effort to clearly delimiting the role and responsibilities of the data-science team. Distinguish clearly between data science and business analytics; the former's bailiwick generally spans the enterprise and deals with structured and unstructured data, while the latter concentrates on structured data in specific business contexts. Define the division of responsibility between the data-science team and existing IT groups regarding production deployment and postproduction maintenance of data-science systems.

Consequences

Benefits

- Tackling issues of functional areas up front helps minimize the likelihood that data-science efforts will need to be abandoned in the face of organizational resistance or lack of support.
- A careful examination of roles and responsibilities may uncover opportunities for cooperation between the data-science team and other organizational groups.

Liabilities

- It may be difficult to reach agreement on the proper allocation of authority and responsibility and to ensure compliance after agreement has been reached.
- Even with painstaking effort, areas of functional overlap may persist and must be handled with sensitivity.
- Too great a sensitivity to possible encroachment on others' areas of responsibility may result in excessive and unwarranted caution and stifled team initiative.

OPERATIONS

Show Me First

Context A business organization seeking to derive value from a newly constituted data-science team.

Problem Organizations may be skeptical about the value of data science, and such expectations can be self-fulfilling.

Forces

New kid on the block: Data science is a late arrival to most businesses, so it may be difficult to see how and where it can deliver value (see Fuzzy Boundaries).

Esoterica: See Fuzzy Boundaries above

A vicious cycle: Skepticism on the part of the business may lead to inadequate support for data-science efforts. This can result in disappointing output from data-science teams, reinforcing the skepticism.

Solution

Build credibility and scale: at the outset, the data-science team should seek to deliver quick wins, building a base of work for larger, more ambitious projects. The objective should be to demonstrate that business value can indeed be derived from data science.

Consequences

Benefits

- Delivery of real value to the business.
- Additional credibility for data science should increase their traction with the business, affording them the opportunity to embark on more ambitious projects in the future.
- Opportunities for both parties to learn how to work together effectively.

Liabilities

- Work on bigger and potentially more valuable projects is necessarily postponed.
- The business may form the impression that all data-science projects can be completed quickly and with minor effort.
- Data scientists may feel frustrated and underutilized if their work is restricted to short-term deliverables only.

Failure to Communicate

Context A business organization seeking to work productively with data scientists.

Problem Delivery of value by the data-science team may be impeded by ineffective communication between data scientists and businesspeople.

Forces

Esoterica: See Fuzzy Boundaries above

Business illiteracy: Data scientists—especially more junior ones—may not understand business in general. Furthermore, the particularities of a given business application domain may be initially inscrutable to the data-science team.

Blind 'em with science: Data scientists may seek to signal their expertise by using excessively technical language in their interactions with businesspeople.

Solution

Both business and data science need to participate in the solution: businesspeople should try to express their requirements as crisply, clearly, and unambiguously as possible. Correspondingly, data scientists should take time to explain how their solutions work, spell out their assumptions, the uncertainty associated with their inferences or predictions, and any attendant risks. They should communicate in terms understandable by the business and use narratives, diagrams, and visualizations to convey their ideas and results. Communication may be facilitated by recruiting a business integrator to mediate (see Staffing for Success).

Consequences

Benefits

- More effective communication between data scientists and their colleagues in the business, leading to increased productivity and trust between the parties.
- Data scientists are less likely to deliver "the right answer to the wrong question."

Liabilities

- Lack of agility: in an attempt to capture their requirements crisply and unambiguously, businesspeople may produce specifications that are too rigid to allow for the development of an effective solution. For possible remedies, see Agile Data Science in Part 1 of the article in the Fall 2019 issue.
- Simplified explanations of data-science solutions may gloss over critical features, increasing the risks of future problems.
- Data scientists may feel that simplified accounts of their work understate their contributions.

See also Busting the Brotherhood, Staffing for Success (Business integrator)

Galloping Complexity

Context A data-science team working with a broader business organization.

Problem Data scientists may be tempted to pursue technical complexity for its own sake, preventing them from delivering effective solutions to the business.

Forces

Academic norms: Many data scientists have postgraduate education and employment in academia. As such, their professional aspirations may be shaped by academia, which values publication of technically compelling work, somewhat regardless of its practical application.

The allure of mastery: For many data scientists, the development of an advanced model, algorithm, or software system can be an attractive undertaking in its own right—it provides challenge and the opportunity to demonstrate proficiency to oneself or others.

The unreasonable effectiveness of simple solutions: More often than one might expect, business problems are susceptible to (humdrum) solutions with limited technical complexity.

Solution

Make the purpose and mission of the data-science team clear from the outset. Their primary responsibility is to support the business, which may mean prioritizing productivity over perfection.

Consequences

Benefits

- Data-science solutions that are delivered in a timely manner and are better fitted to the needs of the business.
- A simple solution may form the basis for the development of a more sophisticated one—see Agile Data Science (in Part 1 of the article) for further details.

Liabilities

- Not all business problems are amenable to simple solutions, though discerning this may require experimentation and/or experience.
- Data scientists may feel frustrated if they are instructed to moderate their ambitions, particularly if they feel the promised rewards for doing so are not forthcoming.

Phillip Yelland is Principal Data Scientist and **Zeynep Erkin Baz** Director of AI Science at Target Corporation.

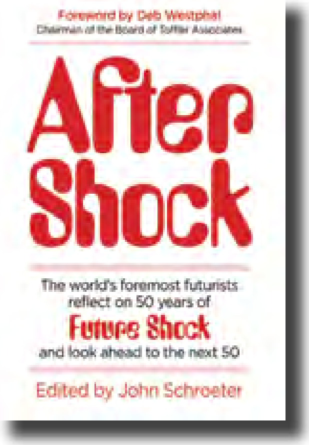
I'm grateful that my observations and suggestions have been adopted into the model proposed by Phillip and Zeynep. I think it is also worth making the point that data science remains largely in its embryonic stage of development and evidence of what works (and what does not) is still emerging. Contributions to this understanding are most welcome and will help accelerate the delivery of business value in organizations that adopt the key principles recommended.

—Simon Clarke

PLAN
FOR THE
RESILIENT
ENTERPRISE

After Shock: The World's Foremost Futurists Reflect on 50 Years of Future Shock

REVIEWED BY IRA SOHN


Foreword by Deb Westphal
Chairman of the Board of Toffler Associates
After Shock
The world's foremost futurists reflect on 50 years of
Future Shock
and look ahead to the next 50
Edited by John Schroeter

FIFTY YEARS FROM *FUTURE SHOCK* TO *AFTER SHOCK*

This year we're celebrating the 50th anniversary of a number of important events: 2020 is 50 years after the world observed the first Earth Day, recognized by many as the birth of environmentalism; it is the golden anniversary of the design of the first commercially viable liquid-crystal-display (LCD) technology, ubiquitously used beginning with the earliest digital watches and portable calculators to today's wide-screen televisions and smartphones; it marks five decades since the the flight of Apollo 13; and it has been a half-century since the publication of the sensational bestseller *Future Shock* by Alvin Toffler and his wife and collaborator Heidi Toffler—a publishing phenomenon, with many millions of books sold across numerous languages.

We now have *After Shock*, a collection of essays and commentaries that reflect upon the Tofflers' original opus. The editor of *After Shock* is John Schroeter, the Executive Director of the Abundant World Institute, a society comprised of technologists, futurists, and entrepreneurs. Schroeter assembled over a hundred of his colleagues and acquaintances to produce this Festschrift to the Tofflers (Alvin died in 2016, Heidi in 2019) for their contributions to futures studies and forecasting. Despite *After Shock*'s 2020 publication date, the book went to press prior to the eruption of the COVID-19 pandemic.

THE CONTRIBUTING AUTHORS

Almost half of *After Shock*'s contributory dramatis personae identify as futurists, about 5% consider their principal area of work to be related to artificial intelligence (AI), and the remaining contributors are evenly distributed in the fields of economics, health, technology, and academia. According to Schroeter, about 75% of the contributors are baby boomers born between 1946-64 and Gen Xers born between 1965-80, with millennials (those born between 1981-96) accounting for the remaining 25%. Many are either "graduates" of or affiliated with Santa Clara-based Singularity University, founded in 2008 by Peter Diamandis and Ray Kurzweil as a "global learning and innovation community using exponential technologies to tackle the world's biggest challenges and build a better future for all." Jerome Glenn, one of the contributors, supplies a working definition of the term "futurist":

> *Futurists systematically look at future possibilities, consequences, and, given all that, figure out what we should do to improve our prospects* (page 421).

Some of the 100-plus essays in *After Shock* provide only a look back to 1970 and the environment that provoked the theme of *Future Shock*, which Toffler defined as "a time phenomenon, a product of the accelerated rate of change in society arising from the superimposition of a new culture on an old one." Others have interpreted Toffler's concept as "a disease: the disease of cultural change that is happening too quickly for human adaptation" (page 442). Various contributors were forward looking, focusing on the next 20-50 years, but also invoking Toffler's warning that technological change often advances so quickly that it can overwhelm society's ability to adapt to it.

Many of the contributors are on the lecture circuit as keynote speakers at corporate lunches and dinners, engage in considerable moonlighting as business consultants, and serve on advisory boards of corporations, government commissions, and nonprofit organizations. Quite a few cannot resist the opportunity to toot their own horn by citing their own papers, books, and lectures, affiliations, and awards. Some contributions are merely cameo appearances, for example those of Kurzweil and Newt Gingrich.

THE TIMES THEY ARE A-CHANGIN'

The years of the late-1960s, when Toffler planted the seeds of *Future Shock*, were one of the most turbulent periods in United States history, characterized by "the Vietnam War protests, spiritual movements, women's rights demonstrations, the civil rights movement, black liberation movements, and student protests and rebellions" (page 406). In 1972, Congress formed its own futurist think tank, the Office of Technology Assessment (OTA), "to examine issues involving new or expanding technologies, to assess their impacts, to analyze alternative policies to avert crises, and for scientific expertise to match that of the executive branch" (***https://www.gao.gov/products/103962***).

The 1970s should rightly be remembered as a heyday for the development of large, computerized economic models of national and global scope, with forecasts extending out decades. Many of these models were equipped with detailed representations of natural resources sectors—including energy, non-fuel minerals (such as copper and steel), and major agricultural resources (such as grains, root crops, and livestock). Part of this flurry of activity can be attributed to the powerful advances occurring in information technology (IT), the development of which afforded opportunities for economists, natural scientists, engineers, and mathematicians to store and process large amounts of data about the Earth's inventory of resources and the demands made upon those resources.

But times and attitudes have changed. According to *After Shock* contributing author Paul Saffo, futures research went into decline during the mid-1980s and 1990s while long-range thinking became unfashionable during the Reagan/Thatcher era. Futurism was being ridiculed in Washington. Many of the futures institutions that were created a decade earlier dropped out of existence, and the few that remained struggled to survive in what amounted to a "futurist winter." In 1995 the Office of Technology Assessment was abolished.

The arrival of the World Wide Web on the eve of the new millennium spawned a new but short-lived dawn for futures research. It was quickly extinguished by a suffocating combination of events and influences including the bursting of the dot-com bubble, the terrorist attacks on September 11, 2001, the wars in Afghanistan and Iraq, and ultimately the economic crises of 2007-09. Saffo asserts that "futures thinking today remains longer on show than on substance. Serious futures work continues, but both its scale and impact still fall far short of what the Tofflers and others hoped for half a century ago" (page 78).

THE FORECASTS FROM THESE FUTURISTS

Most of the contributions seem to be heavily invested in computing, data mining, connectivity, machine learning, digitization, blockchain, and the like. But their focus fails to address how these technologies will benefit the average person. This reader has the impression that these futurists can't see the forest for the trees. And they raise issues without proffering solutions.

There are exceptions, such as the contribution by Maciej Kranz, a Vice President for Strategic Innovation at Cisco, who provides examples of the application of the IOT, AI, blockchain, and big data to agriculture. John Petersen, President of the Arlington Institute, speculates on increasing life expectancy, the end of traditional manufacturing and employment, rapid climate change, and new energy, food, and transport systems.

If we accept the mandate of futurists defined by Jerome Glenn and the criticism of the state of today's futures studies by Paul Saffo, there's work to be done to prepare for the next half-century. As this review is being written—while the economy is in COVID-19 lockdown—several contributors, including Martin Rees and Rick Sax, presciently cite the risk of global pandemics, food insecurity, and increased international migration triggered by the lack of security and opportunity.

Here are a few excerpts offering a taste (frequently "mushy") of the hopes and worries embedded in futurist thinking.

We can awaken our educational system from its "silence about tomorrow" by charging students with the responsibility for thinking about the future, for the simple reason that they are going to spend the rest of their lives there. In other words, if they are going to be the ones who imagine, invent, create, and safeguard the future, they must first begin by thinking about it.—Jack Uldrich

What happens to facts, information, knowledge, and history when seeing is no longer believing, and we literally cannot trust our senses anymore? Toffler conceived of information becoming kinetic in this manner, and in the space of half a century we have witnessed the shift to information becoming not only hyperkinetic, but also ephemeral. By extrapolation, facts, information, knowledge, and history could become increasingly perishable, and the current catch-phrase, "a post-fact society" may live up to that moniker. —Tanya Accone

Many of the world's problems are, in fact, caused by slowing down, rather than speeding up. The rate of population growth has been slowing for almost 30 years; economic growth has been slowing for more than 50 years, and productivity growth for much of that time; even the rate of digital innovation is slowing, in the face of mature and saturated markets and the end of Moore's Law. Of the big generational drivers of change, only the rate of environmental change is accelerating, catastrophically.—Andrew Curry

In a world of accelerating disruption driven by exponential technological change, reacting quickly has less strategic value every year. It is now an imperative to learn a new competency—how to accurately anticipate the future. —Daniel Burrus

The irony is that our scientific triumphs over the 50 years since the publication of Future Shock *continue to be subverted by our own human cravings, folly, hubris, and evolutionary hardwiring. We are still a long way from achieving Toffler's grandest rose-colored aspiration, that the "super-industrial revolution" could "erase hunger, disease, ignorance and brutality."* —Rick Sax

John Schroeter is to be congratulated for assembling this group of futurists under one roof, who collectively have defined the likely forecasting to-do list for the next half-century. Perhaps a subset of these futurists will agree to discuss their forecasts in greater detail in a follow-up volume, along with the likely consequences of those predictions.

After Shock: The World's Foremost Futurists Reflect on 50 Years of Future Shock, edited by John Schroeter, Abundant World Institute, an imprint of John August Media, LLC, 2020; ISBN Print: 978-0-9997364-4-9; ISBN eBook: 978-0-9997364-4-6; 542 pages (no index)

Ira Sohn is Professor of Economics at Montclair State University in New Jersey and *Foresight*'s Editor for Long-Range Forecasting.

imsfinc@gmail.com

Dealing with "Deepfakes": How Synthetic Media Will Distort Reality, Corrupt Data, and Impact Forecasts

JOHN WOOD AND NADA SANDERS

PREVIEW AND KEY POINTS *from the authors: Distorted data are nothing new. However,* deepfake *technology—the term is a combination of "deep learning" and "fake"— has created the ability to distort reality in new and alarming ways. This technology is capable of fabricating audio, video, and even text files that are almost indistinguishable from authentic documentation. Machine-learning capabilities are escalating the technology's sophistication, making deepfakes ever more realistic and increasingly resistant to detection. The implications for communication, data integrity, forecasting, and decision making are vast and unequivocally grim.*

Our best hope for dealing with deepfakes may lie with the creative problem solving of the data-science community, sponsored and supported by corporate leadership.

AN EMERGING THREAT

Imagine a video picked up by cable news of a prominent CEO announcing her resignation in anticipation of a corporate internal fraud investigation. Another video goes viral online, featuring a prominent politician trashing his own party and praising the competition in an apparently behind-closed-doors meeting. Now imagine that these videos do not depict real events but are, in fact, the eerily realistic manipulations of audio and image data. Synthetic media in the form of deepfakes can easily cause irreversible and nearly instantaneous reputational harm and inject visual, audio, and textual falsehoods into the collective consciousness that is the globally connected internet. This could have vast consequences: stock-market losses, character assassination, mistrust in institutions, and undermined public policy. In an age of big data, deepfakes corrupt data, resulting in misleading forecasts requiring diversion of time and analytical resources to identify and correct. But the damage can be done long before the deepfake is detected and removed.

These finely crafted falsifications of visual and audio media will create a paradigm shift in our trust of data. Well-known human biases such as the self-serving bias and confirmation bias can be especially pernicious when combined with a microtargeted deepfake. As law professors Robert Chesney and Danielle Citron (2019) observe,

> *The marketplace of ideas already suffers from truth decay as our networked information environment interacts in toxic ways with our cognitive biases. Deepfakes will exacerbate this problem significantly. ... Soon it will be impossible to know if what your eyes and ears are telling you is true.*

The human eye is perhaps the most sophisticated information processing instrument known, developed over billions of years of selective evolutionary pressure. Yet with deepfakes, we can no longer rely on our eyes to provide an accurate interpretation of the real world, insofar as we are looking at the screen of a television, computer, or mobile device. With the ubiquity of these screens and our widespread reliance on social media for news, we are massively exposed and vulnerable.

And the harm done remains even if the victim of a deepfake is compensated by the perpetrator, because the record may remain on the internet where others can redistribute it while the victim may not succeed in having the content removed.

Deepfakes are by no means confined to audio and visual formulations. With the public release of OpenAI's GPT-3, they now extend to text: think bot-generated tweets and user-submitted comments on online polls and forums. The internet can easily become saturated with AI-produced text that is "shockingly human-sounding," and since textfakes are easy to produce in high volume and practically impossible to discern from human-generated text, it will be all too easy "to stitch a blanket of pervasive lies" (DiResta, 2020).

CORRUPTED DATA AND CORPORATE AND POLITICAL SABOTAGE

Deepfakes can significantly corrupt data that are the backbone of forecasting and decision making. They are especially disruptive to big-data ecosystems which serve as a basis for analytics algorithms. Historically, lack of data had been a key problem for forecasters; now, however, deepfakes are creating another problem: questioning the integrity of the data itself.

To date, cybersecurity has primarily focused on the unauthorized access of data, but the motivations behind attacks have changed. Instead of stealing information, the modern hacker now attempts to modify data *while leaving it in place*, which can create significantly more damage. Corrupted data can serve cybercriminals better than stolen information, including everything from financial gain to election fraud, and could completely sabotage a firm's reporting, ruining the company's relationship with customers, partners, and investors. It could create small intrusions that undercut data integrity but have the potential to be powerful in pushing forth misinformation. Small errors can snowball. Just as protocols exist for data backup and encryption, we need security protocols to ensure data authenticity and validation. Otherwise, a company could be unwittingly paying to move, store, and distribute deepfakes, driving up the cost and fallibility of its decision-making processes.

In addition to harming business, deepfakes can interfere in political outcomes, including democratic elections, particularly those taking place during "choke-points," narrow windows of time during which irrevocable decisions are made. Releasing prejudicial information about a candidate in the month prior to a U.S. presidential election is so common, we have a term for it: the "October surprise." We should expect these surprises to come in the form of deepfakes as increasingly potent weapons of political disinformation.

Even if deepfakes can be flagged and punished appropriately, their existence creates what Chesney and Citron call the *liar's dividend*, realized, for example, when someone who is dastardly in real life can pass off a "hot mic" moment as just another pesky deepfake, when in fact they were caught in flagrante delicto. They can easily claim a real scandal is simply a deepfake and avoid immediate repercussions by hiding behind the resulting public uncertainty.

If our tone seems hyperbolic, we should not be dismissed as fearmongering. An audio deepfake has already been used to perpetrate white-collar crime "in a remarkable case that some researchers are calling… the world's first publicly reported artificial-intelligence heist" (Harwell, 2019). According to the insurance company investigating the incident, "Thieves used voice-mimicking software to imitate a company executive's speech and dupe his subordinate into sending hundreds of thousands of dollars to a secret account."

GENERATIVE ADVERSARIAL NETWORKS

Deepfakes are powered by Generative Adversarial Networks (GANs), a species of artificial intelligence (AI) invented by researchers at Google. GANs work by pitting deep neural networks against each other in a competition, similar to the game of "fox and hound." It begins with one network, called a *generator*, making a product based on a real version of an audio or video. The product is shared with another network, called a *discriminator*, that determines if the image is real or fake. As the images are created and judged, the facsimile's realism improves proportionate to the discriminator's decreasing ability to determine what is real or not (Littell, 2019). GANs follow a process of iterative improvement, similar to when algorithms adjust forecasts based on error.

Like many forecasting algorithms, a GAN trains itself to improve over time. In this case, however, "improvement" means becoming increasingly adept at deception. The result is a false simulacrum based on authentic video and audio data. This technology is only in its infancy, yet is already potentially dangerous and will become more so as it improves over time. Sophisticated computer-generated imagery (CGI) is no longer the exclusive domain of Hollywood but is now readily available to practically anyone.

Open-source software is contributing to the distribution of deepfakes, allowing the author to share that software with others while preventing those other users from excluding anyone else from using the code. Whereas traditional software is behind a paywall that's often cost-prohibitive to startups and individuals, open-source software is usually free—and can proliferate like kudzu. According to *MIT Technology Review* (Knight, 2018), "GANs are a double-edged sword. Open-source software and cloud-based machine-learning platforms have granted liberal access to the AI programs that can be used for these purposes, such as OpenFaceSwap or Paperspace."

DETERRENCE PROBLEMS AND THE LAW'S LIMITED REACH

It is unlikely the U.S. legal system will address the risk of deepfakes in a meaningful way.

Civil and Criminal Law

Civil tort law is inapt for the nature of this risk since its remedies come limping along far after the injury has occurred. The saying from trial lawyers goes, you can't un-ring a bell. Even if the victim can meet the burden of proof to establish a prima facie case, who is the defendant? Determining the actual culprit is not at all straightforward in a cyber world where bad actors hide behind anonymity. Victims of deepfakes are unlikely to find relief in court.

Criminal law may serve as a deterrent to the fainthearted, but not to those most likely to weaponize deepfakes: white-collar criminals, fraudsters, saboteurs, hostile foreign states, and terrorists. Insider-trading laws could stop some from posting deepfake for financial gain through carefully timed stock sales, but what about outside investment firms? Short-sellers tired of losing bets against a target company, for example, might resort to creating deepfakes about that company, its products, its leaders, or other areas sensitive to investor confidence.

Self-Policing

And the private sector is not offering protection. As noted recently in *Slate*, we have "a total misalignment of incentives," with social-media platforms pressured to promote deepfakes because of the engagement generated thereby (Pangburn, 2019). Comments, likes, and shares seem to multiply when content is rage-inducing. Some platform self-policing is all we are likely to see by way of throttling the damage done by deepfakes.

A powerful federal law, Section 230 of the Communications Decency Act, shields social platforms from civil liability stemming from users' posts. The Electronic Frontier Foundation calls Section 230 "one of the most valuable tools for protecting freedom of expression and innovation on the Internet, because... companies that lack the resources to challenge lawsuits based on posts by their users rely on this protection, and many could not exist without it" (Brown, 2019). If industry self-regulation is to be our only source of meaningful protection, in a very real sense the fox is guarding the henhouse.

Legislation and the First Amendment

Are there legislative steps that can be taken? Could governments pass laws making it illegal to create or disseminate them?

A blanket deepfake ban has constitutional and practical challenges. Unlike self-policing by social media, limitations inherent to government narrow the legislative options. The First Amendment provides that Congress shall make no law abridging the freedom of speech. Creating a deepfake could reasonably be characterized as a creative speech act, like political cartooning. Whereas public policy should strongly favor expression, how do we draw the line between protected political satire that uses the likeness of a public figure and a malicious deepfake designed to cause economic disruption?

We are extremely doubtful that government can get this right. Consider proposed legislation in the 116th Congress; you know Congress means business when it busts out a legislative acronym: the Defending Each and Every Person from False Appearances by Keeping Exploitation Subject to Accountability Act of 2019 ("DEEPFAKES" Accountability Act). This law would require mandatory watermarks and clear labeling on all deepfakes, "a step that is likely to be ignored by those whose entire purpose is to weaponize a deepfake" (Brown, 2019).

Further, the law defines deepfakes as any media that falsely appears to authentically depict the speech or conduct of a person and that is produced substantially by technical means. This clumsy definition renders the legislation vulnerable to First Amendment attack. The law would exempt officials of the U.S. government who claim they created a deepfake "in furtherance of public safety or national security." We can all easily imagine a White House press secretary arguing the president created a deepfake of his political opponent "in furtherance of public safety or national security," simply by stating the president's opponent would make the country unsafe or insecure. That is a gaping loophole and makes it liable to be weaponized by public officials against political rivals, with public safety or national security as a pretext for the deception.

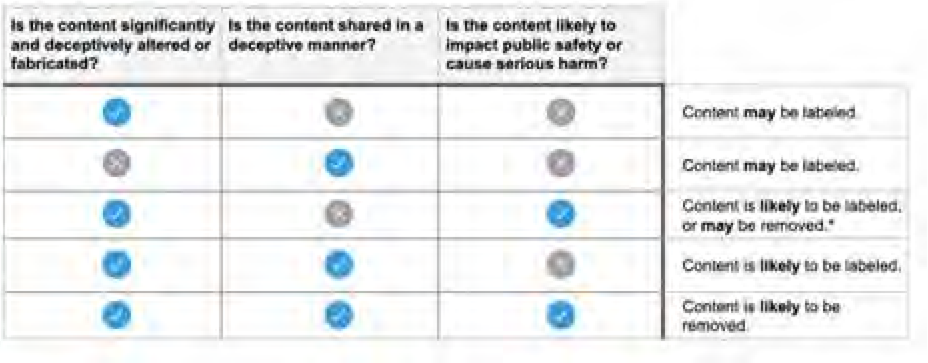
The First Amendment enshrines the freedom of expression and any law restricting online content, particularly political content, risks running up against these constitutional protections. Furthermore, deepfake bans will be difficult to enforce due to the internet's anonymity.

In summary, unfortunately, neither civil nor criminal law provides adequate relief to victims of deepfakes, and constitutional protections do not help. Society simply isn't ready for the tribulation that deepfakes can unleash. Pandora's box is wide open. So where can we find help?

TWITTER TO THE RESCUE?

The first meaningful source of deterrence comes from what may seem a surprising player: the social-media giant Twitter. Twitter solicited public comment on whether and to what extent it should police misleading content. Options

Table 1. Twitter's Deepfake Criteria



Is the content significantly and deceptively altered or fabricated?	Is the content shared in a deceptive manner?	Is the content likely to impact public safety or cause serious harm?	
✓	☒	☒	Content **may** be labeled.
☒	✓	☒	Content **may** be labeled.
✓	☒	✓	Content is **likely** to be labeled, or **may** be removed.*
✓	✓	☒	Content is **likely** to be labeled.
✓	✓	✓	Content is **likely** to be removed.

included flagging deepfakes, removing deepfakes, warning users before allowing them to post deepfakes, and banning users who repeatedly violate these terms. On February 4, 2020, Twitter Safety announced its rules and policies: "You may not deceptively share synthetic or manipulated media that are likely to cause harm. In addition, we may label tweets containing synthetic and manipulated media to help people understand their authenticity and to provide additional context." ***https://help.twitter.com/en/rules-and-policies/twitter-rules***

Deepfakes posted that meet all three of Twitter's criteria may not be shared on Twitter and are subject to removal as shown in **Table 1**.

Contrasted with Facebook, which has refused to take down false political speech, Twitter's posture has been more proactive. Still, the notion that industry self-regulation is our primary defense against widespread deception inspires confidence in none of us.

A CALL TO DATA SCIENTISTS

Although its concern is growing, the technology community lacks consensus on how best to deal with the detection and policing of deepfakes. "A number of different techniques are being researched and tested. One team investigated digital watermarking of footage.... Another team is using blockchain technology to establish trust, which is one of its strengths. And yet another team is identifying deepfakes by using the very same deep learning techniques that created them in the first place" (Pangburn, 2019).

New Software

Given that deepfakes are based on AI in the first place, one option lies in the creativity of data science. Researchers have already built systems for detection of deepfakes that assess lighting, shadows, facial movement, and other features that can flag fabricated images. Another approach adds a filter to an original image that makes it impossible to use that image to generate a deepfake. A handful of startups have emerged that offer software to defend against deepfakes, including Truepic and Deeptrace.

Biometric Signatures

Another promising solution is the use of biometric signatures. "Every person has their own unique facial tics—raised brows, lip movements, hand movements—that function as personal signatures of sorts. The basic idea is we can build these soft biometric models of various world leaders, such as 2020 presidential candidates.... Companies could offer soft biometric signatures for [executive] identity verification purposes in the future. Such a signature could be something as well-known as eye scans or a full body scan" (Pangburn, 2019).

The problem here is that this approach requires the VIP to surrender actual biometric signatures to a third party. If that third party is hacked, the biometric signatures can be misappropriated. Remember, we cannot control or alter our biometric signatures so the technique is risky for the very reason it would be effective. If you can hack into the biometric signature bank, then you've stolen (or made copies of) the keys to the kingdom, and those locks can never be changed.

Detection Filters

It is far easier to create a deepfake than to detect one. But assume we can design a filter that uses an automated deepfake detection tool, prescreening every tweet, Facebook post, blog post, and other forms of social media. The platforms could, in theory, adopt an automation system to scrub their sites from GAN-generated synthetic media. However, even once detected, there would have to be an evaluation phase that determined whether the deepfake was "designed to mislead" (a la Twitter), or for some malicious intent (a la tort law), or not fair use (a la copyright law), or whatever other substantive standard we choose to apply. This post-identification analysis takes time. While it is true that this process may lead to an endless cat-and-mouse dynamic, similar to what exists in cybersecurity today, we are confident that data scientists can come up with breakthroughs on deepfake detection. Perhaps, the open-source ethos that made the spread of deepfakes possible could make its remedy more likely to come about.

CONCLUSION

Deepfakes are here and we're dangerously underprepared. An essential first step is to increase awareness of the possibilities and dangers, from misinformation to data corruption. True information is the only available treatment against the infections caused by viral disinformation. We must all be willing to suspend belief and ask for verification when confronted with video and audio that appears too good, or too bad, to be true. If the content seems excessively useful to a certain powerful interest group, we may be better off assuming it's a deepfake until proven otherwise. We must cultivate widespread skepticism toward unverified things we see and hear on social media.

In the short term, an effective solution may come from major tech platforms, as we have seen with Twitter. In the long term, however, we believe it will be up to the creative genius of the AI and data-science research community to develop solutions to debug corrupted data sets and detect deepfakes. For now, by raising awareness of the risks posed by deepfakes, perhaps journalists, forecasters, analysts, business leaders, and politicians can anticipate the certain inevitable impacts of this risk and adjust our behaviors appropriately, mitigating their disruptive effects.

REFERENCES

Brown, N.I. (2019). Deepfakes Are Frightening, But So Is Congress' Rush to Regulate Them, *Slate* (July 15).

Chesney, R. & Citron, D.K. (2019). Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security, *California Law Review*, Vol 107: 1753 – 1820.

DiResta, R. (2020). AI-Generated Text Is the Scariest Deepfake of All. *Wired* (July 31), ***https://www.wired.com/story/ai-generated-text-is-the-scariest-deepfake-of-all/.***

Harwell, D. (2019). An Artificial-Intelligence First: Voice-Mimicking Software Reportedly Used in a Major Theft, *The Washington Post* (September 4).

https://www.technologyreview.com/10-breakthrough-technologies/2018/

Knight, W. (2018). Technology Is Threatening Our Democracy. How Do We Save It? *MIT Technology Review*, 121(5):37, 2018.

Littell, J. (2019). Don't Believe Your Eyes (or Ears): The Weaponization of Artificial Intelligence, Machine Learning, and Deepfakes, *War on the Rocks* (October 7).

Pangburn, D.J. (2019). You've Been Warned: Full Body Deepfakes are the Next Step in AI-Based Human Mimicry, *Fast Company* (September 21).

John D. Wood, Esq. is Partner at the Law Firm of Green, Klein & Wood, and co-founder of The Humachine LLC, providing consulting on technology, strategy and ethics. He is author of multiple law review articles on human nature and risk management and co-author of two books, *Foundations of Sustainable Business: Theory, Function and Strategy* (Wiley 2e, 2019) and *The Humachine: Humankind, Machines, and the Future of Enterprise* (Routledge, 2020).

john@johnwoodlawfirm.com

Nada R. Sanders is the Distinguished Professor of Supply Chain Management at the D'Amore-McKim School of Business at Northeastern University. An internationally recognized thought leader in forecasting and supply-chain management, she is author of the books *Supply Chain Analytics* (Prospect Press, 2019), *Foundations of Sustainable Business* (Wiley 2e, 2019), and *Supply Chain Management: A Global Perspective* (Wiley 3e, 2020) and co-author of the book *Operations Management* (Wiley 7e, 2019) and, most recently *The Humachine: Humankind, Machines and the Future of Enterprise* (Routledge, 2020). Nada is a Fellow of the Decision Sciences Institute and has served on the Board of Directors of the International Institute of Forecasters (IIF), Decision Sciences Institute (DSI), and the Production Operations Management Society (POMS), where she is immediate Past-President.

n.sanders@northeastern.edu

U.S. Presidential Election Forecasting: *The Economist* Model

COLIN LEWIS-BECK AND MICHAEL LEWIS-BECK

PREVIEW *In June of this year,* The Economist *began publishing regular forecasts of the outcome of the 2020 U.S. presidential election. In this article, Colin and Michael Lewis-Beck describe the model used, evaluate its potential strengths and weaknesses, and provide many perspectives on election forecasting models in general. They conclude with forecasts of the results of the vote in the upcoming November 3 U.S. presidential election.*

INTRODUCTION: SYNTHETIC MODELING

The Economist has begun fielding a presidential election model developed by Andrew Gelman and Merli Heidermanns, political scientists at Columbia University. The community of election forecasting scholars should welcome this addition to the enterprise, especially given the widely perceived failure of the polls and, more seriously, poll-driven models to correctly forecast the 2016 victory of Donald Trump. In this essay, we briefly explicate the idea of election forecasting models, then assess the strengths and weaknesses of this worthy effort, which, for clarity here, we label the "E-model." We conclude with a remark on its prospects for calling out the 2020 winner.

With respect to forecasting American elections, there are two basic modeling approaches. The dominant one in the academic literature consists of structural modeling, whereby the researcher posits a substantive explanation of the vote choice, to be tested in a regression equation estimated well in advance of the election. Characteristically, such models are based on theories of how fundamental political and economic issues shape the presidential vote.

By way of contrast, the dominant approach in the news media consists of poll "modeling" that eschews substantive explanation of the vote choice in favor of probing public-opinion data on respondent answers about their voting intentions. An emerging strategy that combines the structural and polling approaches calls itself "synthetic" modeling (Lewis-Beck and Dassonneville, 2015). A synthetic model aims to capture the long-term theoretical potency of the former approach and the short-term flexibility of the latter, with the goal of achieving more accuracy. The E-model, which we unfold a bit below, can be classified as a synthetic model.

THE E-MODEL

In outline, the E-model begins with a prediction of the national popular vote, as a function of national polls and political and economic fundamentals, such as stock-market performance and real disposable income. The power of various possible predictors is evaluated empirically, via different variable combinations and different election samples, following some standard "out-of-sample" robustness techniques. The model yields impressive ex post forecasts of election outcomes, from 1948 to 2016.

Characterizing the structural component of their model as a prior belief about the final election results, they update their structural forecast daily, modifying it on the basis of national and state-level polls. As election day approaches—and the number of pre-election surveys increases—the model down-weights earlier forecasts in favor of more recent polling data. The final output provides daily forecasts (with measures of uncertainty) of popular-vote shares at the state and national level.

Electoral College

Of course, the national popular-vote outcome, important as it is, does not ultimately determine the presidential winner; that prize goes to the victor in the Electoral College, as decided by the vote distribution in the states. Therefore, the researchers move on to prediction of the Electoral College vote, state by state. (No account is apparently taken of the 15 states whose electoral vote goes to the winner of the national popular vote.) That is, they essentially follow the above analytical steps, but change the dependent variable from the absolute party popular-vote share to the state's "partisan lean" measured by several local predictors, among them results from the past two presidential elections, the home states of the presidential candidates and, importantly, the predicted national popular vote for the upcoming presidential election.

Bayesian Methodology

A notable aspect of the E-model is its Bayesian methodology, which differs from the classical (frequentist) approach of almost all the structural election forecasting models. In the context of the expected vote share for, say, the Democratic party in state X, the researchers make a "prior" prediction before systematically looking at the available state polling data. For the E-model, the initial prior prediction derives from their structural model, made months before election day. Then, as state polling data comes in over the course of the campaign, the prior prediction is updated, making it a "posterior" prediction, which serves as the current forecast. This Bayesian updating continues (each day's posterior becoming the next day's prior, etc.) until the day before the election, when a final posterior prediction is produced.

Hierarchical Structure

In addition to its Bayesian estimation approach, the model incorporates a hierarchical structure that extends the 2008 presidential forecast model of Linzer (2013). The hierarchical specification allows the E-model to borrow information across states and time. This is especially useful for modeling state polling data due to variation in the frequency of polls across the U.S.

The E-model incorporates a correlational structure based on nine variables, such as racial demographics, median age of all residents, and population density, to account for similarities in political preferences between states. Modeling this dependence allows polling data to be shared across similar states (e.g., Iowa and Minnesota), which produces more precise and stable estimates. By allowing for temporal dependence, the E-model generates steady forecasts even when daily polling data are highly variable or unavailable.

Forecast Uncertainty

Of course, these forecasts are not certain; first, there's the inevitable error arising when sampling from a population; second, they are subject to non-sampling error, foremost being the problem of knowing who among those sampled will in fact vote. The authors struggle valiantly to adjust for the various types of bias arising from this unknown. Their final model contains bias corrections for partisan nonresponse, the type of survey, and the survey population. However, the model makes no allowance for possible voter suppression efforts, such as postal service breakdowns in mail-vote deliveries.

Having built their model with adjustments for polling bias, they use a Markov Chain Monte Carlo (MCMC) algorithm to simulate possible outcomes. For example, how might the predicted election outcome change if it is assumed the calculated effect of telephone polling underestimates the Democratic vote by six percentage points? There are many such possibilities, depending on the assumptions imposed. In all, they generate 20,000 possible paths to the White House and quantify the probability of various election results.

As of August 14, for instance, the distribution of the simulated paths showed Biden the very likely winner (88% probability) of the Electoral College. Every day after additional polling information becomes available, the MCMC algorithm is run again and a new set of updated paths are produced. To their credit, the authors are modest about their forecast, observing it is "not guaranteed" and even if it misses they "will welcome the opportunity to learn."

EVALUATING AN ELECTION FORECASTING MODEL

Having laid out the bare bones of the E-model, we now turn to evaluating it as a forecasting instrument.

A long-standing approach to model evaluation focuses on four characteristics: accuracy, lead-time, parsimony, and transparency:

- Accuracy concerns how close the prediction is to the actual result. Usually, accuracy serves as the sole criterion for evaluation.
- Lead time, i.e., distance in days (weeks, months) before the election itself, plays a crucial role. After all, without lead time there is no forecast; there is just ex post curve fitting.
- The desire for parsimony rests on the venerable principle of Ockham's razor, which argues that a few theoretically strong predictor variables should be favored over multiple variables that have a questionable place in the explanation.
- Finally, the goal of transparency: to be convincing, other researchers should be able to access the same variables and reproduce the same forecast.

Illustrative Example of Model Quality

Application of these criteria can be appreciated readily via a simple demonstration from the Political Economy model utilized to forecast the 2016 U.S. presidential election (Lewis-Beck and Tien, 2016). A verbal statement of the model reads as follows:

Presidential Vote = Presidential Popularity + Economic Growth

with the Presidential Vote equal to the two-party (Democrat and Republican) share of the national popular vote for the president's party, Economic Growth equal to the gross national product (GNP) growth in the first two quarters of the election year, and Presidential Popularity equal to the job approval rating for the president in the July Gallup Poll.

The Political Economy model derives from a referendum theory of elections, which assumes the electorate will reward or punish the White House party at the ballot box, depending on how well the president has handled economic and noneconomic issues. When the model is estimated with ordinary least squares regression across the post-World War II period (17 elections, 1948-2012), these results were obtained:

$$\text{Vote} = \underset{(14.83)}{37.50} + \underset{(4.4)}{0.26^*}\,\text{Popularity} + \underset{(2.04)}{1.17^*}\text{Growth}$$

Adj. R-squared = 0.73 Root Mean Squared Error = 2.84
Durbin-Watson = 2.36 Figures in ( ) = t-ratios
***statistically significant at 0.05, one-tail.**

In order to forecast the 2016 presidential election, the values of Popularity and Growth were inserted into the equation; as of August 26, 2016, Popularity = 51 and Growth = 0.20 (non-annualized) yielding the point estimate of 51.0 percent of the popular two-party vote for Hillary Clinton.

As a pedagogical example, we offer an evaluation of the quality of this forecasting instrument. With respect to accuracy, it did exceedingly well; the actual national two-party popular vote share for Hillary Clinton was 51.1 percent, meaning negligible error. The lead time, too, is nontrivial, i.e., forecast was made over two months before the election itself. The parsimony of the model would appear exemplary, as it contains only two predictor variables and they have strong theoretical pull. Finally, the model has considerable transparency, as the variables are simply defined and readily available. The E-model recognizes the reliability of this structural modeling approach as well as saying, "We were surprised by both the size of the fundamental model's advantage over national polls early in the election cycles, and by how long that gap persisted."

Polling Models

Across the series of elections, the predictions of the Political Economy model do not always achieve this 2016 level of precision. We offer this illustration merely to provide a ready frame for thinking about the quality of a forecasting model. While the Political Economy model takes a simple form, as do most of the structural models, poll-driven models—which rely heavily on vote-intention measures—are more complex but still can be evaluated.

Take the recent efforts by data journalist teams to forecast the 2016 race. Leading media poll aggregators, such as *The New York Times* Upshot, FiveThirtyEight, The Huffington Post, and RealClear Politics, all gave Clinton at least a 70 percent probability of winning. The Princeton Election Consortium even gave her more than a 99 percent certainty of winning. On the criterion of accuracy, then, these forecasts returned a poor performance. And they accomplished that with little lead time, little transparency, and an absence of parsimony.

The culprit for the 2016 forecasting debacle appears to be excessive reliance on woefully inaccurate vote-intention polls. Take the RealClearPolitics daily vote-intention averages, which many observers followed because of their relative transparency and unadorned math. Over the campaign period, from June 16 to November 8, fully 178 out of their 180 observations showed Clinton leading. Small wonder so many believed she had it in the bag the whole way. Moreover, the "final" national polls from eleven leading firms consistently showed Clinton with an absolute lead, with most of these point estimates falling outside the traditional "margin of error." Finally, the situation was even worse with state polling, which played a critical role in the calculation of the Electoral College vote. In over 35 states, the average final vote intention for Trump was an underestimate. Additionally, on average, these state polls generally overestimated Clinton's support by about five percentage points (see Jackson and colleagues, 2020 for details on these polling shortcomings).

THE E-MODEL EVALUATION

A big question for the 2020 presidential election is: Will the complex, poll-driven models do better than they did in 2016? In particular, will the E-model, with its poll-driven component, do better? It has many apparent strengths: big data, sophisticated estimation, dynamic modeling, attention to uncertainty, and corrections for apparent biases affecting state polls. Its daily updating provides both current information of voter preferences as well as a forecast of presidential vote share on election day. However, there are weaknesses in the areas of theory, measurement, process, and inference. Forecasting requires more than curve fitting: it wants good theory as well.

Subjectivity

The E-model does give a nod to *fundamentals* but that search appears data-driven, attending more to the empirical exploration of many measures, rather than to careful model specification informed by the spirit of parsimony. For example, they built an "economic index" that "used a blend of the changes." Such a blend goes against the usual specification of structural models, where almost all focus on a measure of economic growth.

It is not surprising that researchers must sometimes employ subjective judgment in deciding the way to go. A telling example comes from their adjustment of the economic index to incorporate the impact of the coronavirus. Because of the extreme economic values accompanying the virus, they felt the historical series itself was an inadequate guide. Thus, as a working assumption, they judged that the economic impact of the virus would be 40 percent worse than that of the Great Recession. That estimation undoubtedly runs in the right direction; however, it would seem to be open to competition from rival, perhaps larger, estimates. Such subjectivity is, of necessity, enhanced by the Bayesian approach, which, in an explicit fashion, alerts us to the incorporation of prior information into the model.

Timing

The fact the election event occurs uniquely in time means that every daily election forecast the E-model makes can be regarded simply as a snapshot forecast of what would occur on that day, if the election were in fact held. More boldly, it can suggest a forecast for the unique, perhaps quite distant, event: the election outcome itself. Of course, the real election result on November 3 is what we would most like to know, but that task can sometimes appear daunting. This helps explain why some forecasters, including many structural modelers, prefer to seek out the optimal date across the time series for forecasting the true election result. As it turns out, recent research has shown that the optimal date for U.S. presidential election forecasting may well be two or three months before the election itself (Jennings and colleagues, 2020). Such a finding offers further justification for the value of a nontrivial lead time in forecasting.

Sampling Corrections

A further challenge the E-model faces, along with other models that rely heavily on vote-intention polls, concerns the difficulty of inference. All the pre-election polls, state or national, seek to sample a currently nonexistent population, i.e., the voters on election day. The "problem of the missing population" almost guarantees polling inaccuracy.

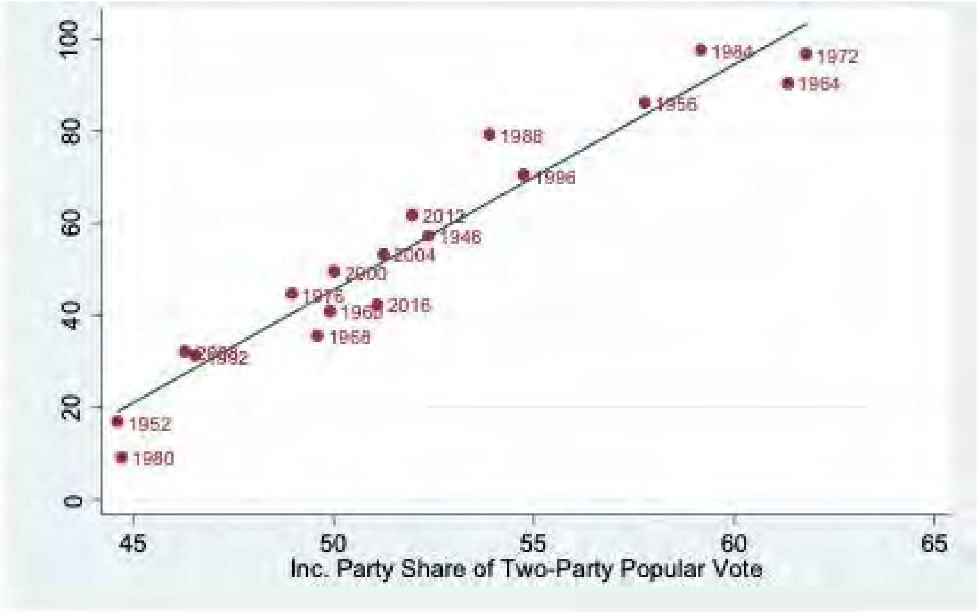
Moreover, the common quick fix of weighting the data falls short. The canons of probability sampling demand that every respondent be selected randomly from the relevant population. That has become expensive to do, not to say impossible. The sampling frames that pollsters actually use—e.g., from random digit dialing or opt-in online panels—are nonprobability samples using quota methods. This implies the necessity of considerable guesswork and assumptions about the quality of the polls. The E-model includes all state polls, which provide additional information; however, it also requires more model assumptions and additional parameters. It would be interesting to see the sensitivity of the E-model forecasts to its numerous bias-correction parameters, in order to get a better idea of the benefits of adjusting for so many potential sources of error. At first blush, one might imagine a comparison to Electoral-Vote.com since it relies on aggregation of state polls; however, this would be a risky proposition, given the high degree of error that infects most state polling.

Popular to Electoral Vote

A last challenge to the E-model, which the authors fully accept, comes from the need to forecast the Electoral College winner. In this effort, they take the sensible path, focusing on prediction of the Electoral College outcome in each of the states. As they mention, one of the key independent variables in that calculation is the predicted popular-vote share, state by state. However, because of the poor quality of most state vote-intention polls, this could be a fraught exercise, generating a fair number of errors. One alternative to the state-level forecasting of the Electoral College outcome involves the use of national-level data. In other words, the unit of analysis becomes the nation, rather than the state, with the national popular vote share used to predict the national Electoral College outcome.

Such an assessment can be seen in **Figure 1**, where the Electoral College vote share

Figure 1. Electoral College Share (%) and Popular Vote Share (%), 1948-2016



nationwide (in percent) is regressed on the two-party popular vote, for the elections from 1948 to 2016 (Jackson and colleagues, 2020). The election results fall very close to the prediction line, with an almost perfect statistical fit (R-squared = 0.93). Knowing the popular vote share at the national level remains an extremely good predictor, in the sense that the point estimates generally fall very close to the true value, especially by the usual standards of observational social science. However, the tight races of 2016 and, especially, 2000 underscore its limits.

The observations from 2000 on tend to hug the line (including 2016, which, unfortunately for Clinton, was just on the wrong side of the line), which says that the E-model might wish to incorporate this national-level relationship in its calculations, rather than relying solely on state-level predictions. (The E-model was run in 2008, 2012, and 2016 ex post. In all three cases a Democratic victory was predicted. The 2020 election is the first where the E-model is run in real time.)

PERSPECTIVE

We conclude with a bit of historical perspective on the enterprise of U.S. presidential election forecasting and its strengths and weaknesses. In his careful review of the book *Forecasting Elections*, Professor Andrew Gelman (1993, p. 119) observes that Lewis-Beck and Rice (1992) "do a lot better in predictions, impressively showing how objective analysis of a few columns of numbers can regularly outperform pundits who use inside knowledge."

This comment points to the question he and his colleague are trying to answer for readers of *The Economist*; namely, can their current scientific efforts do better than those of past commentators and researchers? As of September 8. 2020, the E-model predicts Biden will receive 53.7% of the national vote share compared to 46.3% for Trump. In terms of the Electoral College, the model forecasts Biden will get 334 electoral votes to 204 for Trump and gives Biden an 88% chance of winning the election. (For comparison, the Political Economy model, as of the same date, forecasts a stronger defeat for Trump, with 43 percent of the popular vote and 68 electoral votes.) We, along with many other keen observers of U.S. presidential election forecasting, are watching the E-model with great interest, as they unfold their highly crafted predictions for this monumental race.

REFERENCES

Gelman, A. (1993). Review of Forecasting Elections by Michael S. Lewis-Beck & Tom W. Rice, *Public Opinion Quarterly*, 57(1), 119-121.

Jackson, N., Tien, C. & Lewis-Beck, M.S. (2020). Pollster Problems in the 2016 US Presidential Election: Vote Intention, Vote Prediction, *Italian Journal of Electoral Studies*, 83(1), 17-28.

Jennings, W., Lewis-Beck, M.S. & Wlezien, C. (2020). Election Forecasting: Too Far Out? *International Journal of Forecasting*, 36(3), 947-962.

Lewis-Beck, M.S. & Dassonneville, R. (2015). Comparative Election Forecasting: Further Insights from Synthetic Models, *Electoral Studies*, 39, 275-283.

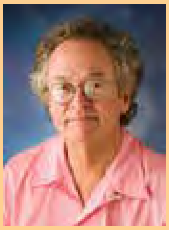
Lewis-Beck, M.S. & Rice, T.W. (1992., *Forecasting Elections*, Washington D.C., Congressional Quarterly Press.

Lewis-Beck, M.S. & Tien, C. (2016). The Political Economy Model: 2016 US Election Forecasts, *PS: Political Science & Politics*, 49(4), 661-663.

Linzer, D. (2013). Dynamic Bayesian Forecasting of Presidential Elections in the States, *Journal of the American Statistical Association*, 108(501), 124-134.

Colin Lewis-Beck is a Visiting Assistant Professor in the Department of Statistics and Actuarial Science at the University of Iowa. His interests include hierarchical modeling, reliability, economics and elections, and public policy.

colin-lewis-beck@uiowa.edu



Michael S. Lewis-Beck is F. Wendell Miller Distinguished Professor of Political Science at the University of Iowa and author or co-author of numerous books and articles on election forecasting. He has served as Editor of the *American Journal of Political Science, Electoral Studies, the Sage QASS* series in quantitative methods, Associate Editor of *International Journal of Forecasting* and current Associate Editor of *French Politics*.

michael-lewis-beck@uiowa.edu

The Benefits of Systematic Forecasting for Organizations: The UFO Project

SPYROS MAKRIDAKIS, ELLEN BONNELL, SIMON CLARKE, ROBERT FILDES, MIKE GILLILAND, JIM HOOVER, AND LEN TASHMAN

INTRODUCTION

The purpose of this paper is to provide a realistic assessment of the potential benefits to business organizations that derive from applying systematic forecasting methods, particularly with respect to operational and tactical forecasting problems. Our overall goal is to improve the usage of forecasting in organizations—UFO—while incentivizing the adoption of systematic forecasting in organizations that now employ only ad hoc methods.

We define *systematic forecasting* as the use of appropriate quantitative methods when suitable data are available, while allowing for judgmental inputs and adjustments that are supported by a documented and defensible rationale. Where little or no data are available, such as with new products, our definition encompasses structured management judgment including use of intention surveys, decision aids, Delphi procedures, and others.

The genesis of the UFO project lies in a series of discussions within a group of practitioners and academics about the challenges facing the forecasting field and the need to learn why many organizations do not exploit what have grown to be remarkable advances in forecasting knowledge and technology.

The article seeks to present the advantages as well as the limitations of systematic forecasting methods. We do so to set fair, reasonable expectations of what can and cannot be achieved, considering the uncertainty associated with all predictions. Realistic expectations are key to establishing good forecasting practice.

We also explore the obstacles encountered by companies in the implementation and improvement of their forecasting processes and provide our understanding of how to overcome resistance to process improvement. And for organizations at "ground zero," we offer guideposts on how to get started utilizing systematic forecasting procedures.

We begin with an assessment of the accomplishments achieved in quantitative forecasting methods. As we note below, the many firms that still lack systematic forecasting need to realize that these approaches, whether simple or complex, have enormous potential benefits for their bottom lines and competitive positions.

The many firms that still lack systematic forecasting need to realize that these approaches, whether simple or complex, have enormous potential benefits for their bottom lines and competitive positions.

THE FORECASTING FIELD TODAY

It has been more than 60 years since Robert Brown's pioneering book *Statistical Forecasting for Inventory Control* (1959), which essentially founded the field of business forecasting. Brown's exponential-smoothing methods were simple but effective for forecasting large numbers of items, many down to the SKU/location level, such as those characterizing inventory demand. Yet many statisticians, engineers, and econometricians decried the lack of a theoretical underpinning or statistical/mathematical elegance of these methods, failing to realize their value as practical forecasting tools. Instead, they touted more sophisticated/complex methods. And while there was evidence that the more complex methods proved superior in tracking historical data (the same data used to make the forecasts), there were doubts that they improved the accuracy of forecasting future data (post-sample time periods), at least until the wider utilization of machine-learning (ML) methods.

What distinguished forecasting, however, from other empirical sciences (especially statistics) was and continues to be its emphasis on testing the post-sample accuracy of forecasting methods. In a paper published in the *Journal of the Royal Statistical Society* (JRSS), Makridakis and Hibon (1979) reported two highly surprising findings concerning post-sample forecast accuracy:

- Among the two-dozen methods put to the test, the most accurate results were found using Brown's simple exponential smoothing adjusted for seasonality—a very straightforward, uncomplicated method.
- Second, averaging the forecasts of more than one method improved overall accuracy.

These findings were not well received by the statistical community of that time (Hyndman, 2020), which—taking steady aim at the messengers—often blamed incompetence for the results. In defense, Makridakis organized a study using 1,001 time series (Makridakis and colleagues, 1982). This time, however, anyone could submit forecasts, making this the first *true* forecasting competition.

This first M-competition and the additional competitions and empirical studies to follow provided the forecasting field with the equivalent of the controlled experimentation used in the physical sciences. This fundamentally changed the field of forecasting, separating facts from opinions and folklore, guiding academic research, and abetting the selection and usage of forecasting methods in practice (Hyndman, 2020).

The results of the first M-competition mirrored the findings that statistically sophisticated methods did not produce more accurate forecasts than simpler ones and that combining forecasts would on average improve forecast accuracy. These conclusions, now replicated through other competitions and individual studies, have at last been well accepted by the academic community (Armstrong, 2006).

Armstrong (1978) had concluded that time-series forecasting methods, based only upon the history of the items being forecast, were often more accurate than models using explanatory variables, a counterintuitive finding. In a more recent forecasting competition regarding tourism, Athanasopoulos and colleagues (2011) argued that explanatory variables can be useful, but only under two specific conditions: (1) when the future values of the explanatory variables are known or can be accurately forecast; and (2) when the measured impacts of the explanatory variables are likely to continue into the forecast period. Sometimes both conditions can be satisfied, such as for forecasting electricity demand when temperatures for a few days ahead are predictable, or when certain variables such as promotional activities in retail sales can be controlled. However, neither condition is always satisfied for tourism demand or many other areas of business forecasting.

Recent competitions have upgraded the potential value of sophisticated methods applied to large collections of data (Salinas and colleagues, 2017). The M4 Competition (2018) showed that those sophisticated methods incorporating machine learning (ML) were often more accurate than simple counterparts.

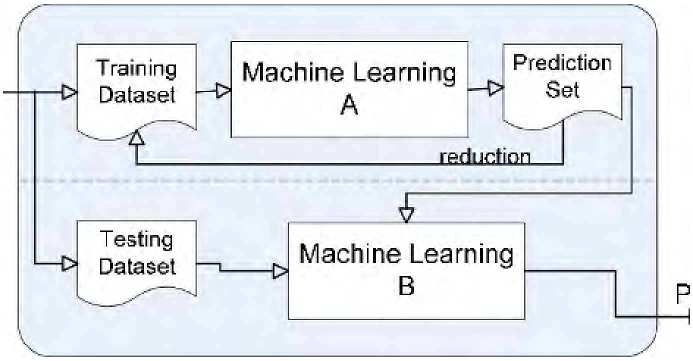
Thus ended a long "forecasting winter chill" against model complexity. The forecasting spring began with the M4 Competition, where a complex hybrid approach combining statistical and ML elements came in first place, while on average the top 16 methods were almost 5% more accurate than that of a common benchmark (Makridakis and Petropoulos, 2020). The top two methods, both hybrids of ML and time-series models, also achieved unforeseen success in estimating the degree of uncertainty in the forecasts, something normally underestimated.

Another important basis for the relative success of ML (in combination with time-series) models is their ability to learn from pooled data. This *cross-learning* results when data from multiple time series are linked in model estimation; for example, modeling groups of products or stores that share common elements of behavior. The top performers in many recent forecasting competitions used cross-learning to improve forecast accuracy over local univariate methods (Boger and Meldgaard, 2020). Applying ML to mobile payment data, Ma and Fildes (2020) report that "by capitalizing on the commonalities in the data across participating retailers, customer flow forecasting based on a large pool of stores from a variety of categories can generate forecasts that are more accurate than those generated by methods based on individual stores" (p.756), a result subsequently confirmed with promotional retail data.

Despite their impressive potential (Li and colleagues, 2020), *ensembles* of different methods remain broadly unadopted. Portfolio segmentation is becoming more widely used, but this has been glacially slow to emerge as a standard way to contend with the challenge of large numbers of time series to forecast with limited resources.

While the findings from the M4 and other recent competitions have elevated the promise of sustained improvements in forecasting methodology, especially through refinements in ML and hybrid models, the accuracy gains come at a huge cost (Gilliland, 2020) in model development and computation time. Gains in forecast accuracy, therefore, must be weighed against the increased costs, knowing that simple methods can achieve respectable levels of accuracy at a small fraction of the resources and effort.

In sum, there has been a slow diffusion of new ideas and approaches among practitioners. We expect the future to deliver further improvements in both accuracy and estimation of uncertainty, hopefully along with processing efficiencies that increase the value and usefulness of forecasting in organizations and hence receptivity to promising new methods.



FUNDAMENTAL UNDERSTANDINGS

How can we convince decision makers of the benefits of systematic forecasting, while avoiding the formation of unreasonable expectations of what forecasting can deliver?

All organizational leaders know that forecasts are necessary for future-oriented decisions, including budgeting and planning activities. But their first option is often an ad hoc approach, delivering judgmental forecasts when required. Pure judgmental methods are too often aspirational—driven by optimism and the desire to achieve future goals versus assessing the objective reality of what is most likely to happen (Lawrence and colleagues, 2006).

The more challenging alternative is to establish a systematic set of procedures that produces forecasts from proven quantitative methods. Extensive research has shown that usage of systematic forecasting methods results in forecasts less susceptible to bias and superior in accuracy. In addition, some firms, including Johnson & Johnson, have reported major payoffs to improved forecast accuracy; see ***https://www.capgemini.com/us-en/client-story/johnson-johnson-transforms-its-demand-planning-and-external-manufacturing-processes/*** .

Makridakis and Petropolous (2021) summarize the fundamental understandings that decision makers must have.

Executives must understand:

- The forecast is an estimation of a future situation. It is not a target. It is not a plan. Nor is it an inventory decision. We may expect sales of 500 units (a forecast), but decide to stock 600 to minimize the risk of a stock-out to an acceptable level (a decision).
- Forecasting is not crystal-ball gazing. Forecasting methods, from the simplest to the most sophisticated, do not possess prophetic powers. Their predictions are based on identifying and estimating past patterns and/or relationships that are then extrapolated to forecast the future.
- All forecasts come with an error. All forecasts are uncertain with the only certainty being the existence of uncertainty.
- The most important advantage of systematic forecasting is its objectivity. It seeks to (a) identify past patterns and relationships to predict the future in a mathematically optimal manner, and (b) base estimates of the uncertainty in the forecasts on the volatility (variance) in the observed patterns/relationships.
- Forecasting accuracy and uncertainty can be estimated consistently in usual, everyday situations when established patterns/relationships remain fairly constant and can be extrapolated reasonably well.
- During periods of recessions/crises, or when unusual events occur, forecasting accuracy deteriorates—often significantly—while the level of uncertainty increases exponentially and sometimes cannot be measured quantitatively.
- There are trade-offs between the achieved forecasting performance and the respective resources needed (such as data availability, computational cost, and personnel time), so companies need to carefully choose an appropriate balance.
- When possible, forecasts should be assessed in terms of their utility (such as the decrease in the holding cost) instead of their forecasting accuracy.
- Lastly, while technology can substantially improve forecasting accuracy and our understanding of uncertainty, we cannot ignore the value of human judgment in the overall forecasting effort.

With all the focus on technology for the statistical modeling side of forecasting, there is also great opportunity for the augmentation of human judgment using artificial intelligence, ML, and even simple logic and business rules (Van Hove, 2020). For example, ML has shown promise in assisting the demand planner by identifying those forecasts most likely to benefit from adjustment, while also suggesting their direction and magnitude (Chase, 2019).

In the next section, we argue that setting proper expectations about forecast accuracy requires an understanding of the conditions that determine *forecastability*, including the distinction between "normal" vs. extreme behaviors and the inherent element of randomness in all behavior.

Extensive research has shown that usage of systematic forecasting methods results in forecasts less susceptible to bias and superior in accuracy.

AVOIDING UNREASONABLE EXPECTATIONS

Unreasonable expectations lead to disappointment and frustration when unexpected errors are blamed on the inadequacy of forecasting methods and processes. Some of these errors certainly result from the way the forecasts have been generated, but also from the inherent unpredictability of the forces being projected. Even if models could have forecast that the COVID-19 pandemic would occur, it would not have been possible to predict its exact timing and destructive economic impact (Osterholm, 2005), such as skyrocketing unemployment rates and the scarcity of bathroom tissue.

Normal vs Extreme Behavior

Let's consider that toilet-paper shortage. On March 12, 2020, U.S. bathroom-tissue sales ballooned 734% compared with the same day the previous year, becoming the top-selling product at grocery stores by dollars spent. Clearly, forecasts did not predict the huge surge in demand that created panic buying, demand exaggerated once photos of empty store shelves began circulating on social and mass media. Worse, the scarcity lasted for several months even as manufacturers rushed to produce and ship more paper. While this story reveals an aspect of our phenomenal failures during the pandemic, it also confirms the notable success of forecasts for all those normal time periods when toilet paper has been available to buy. We must distinguish the rare from the usual.

Analogy to Forecasting Time to Commute

Consider the challenge of forecasting the time it takes to travel to work in the morning. Most of us know very well how much time to allow to travel from home to work and back and realize that such time varies depending on different factors, such as the day of the week and the time one leaves. It is also evident that the commuting time varies even for the same day of the week and when leaving at the same time for any number of *uncontrollable* factors: a major road accident, highway roadwork, a sudden snowstorm, and so forth.

In the absence of these uncontrollable factors, deviations from the average time it takes to go to work are well behaved, most of the time being small, less often larger, and in rare cases substantial. We usually assume that these deviations follow a normal distribution, allowing the measurement of the variations or uncertainty around the average time it usually takes to travel to work each morning.

The extra time, however, that it will take to get to work in case of accidents, roadwork, and inclement weather is vastly different from the usual commute: it is not only highly uncertain, but cannot be expected to follow normal behavior. Rather, it creates a fat-tail distribution (Taleb, 2020) in which extreme travel times become more frequent, and forecasts cannot be found by simple extrapolation of past patterns, partly because we lack sufficient data on unusual events and also because we can't know whether they will recur and what impacts that will have. The 2020 coronavirus pandemic, which combines health and economic crises, presents a worst-case scenario in terms of the uniqueness of the lockdowns and their economic implications, making forecasting extremely difficult and uncertainty impossible to assess.

Distinguishing the normal from the extreme is of particular importance in how businesses set service levels and safety stock. In normal time, the risk of running short of product versus overstocking can be balanced by considering the costs of lost sales versus the cost of carrying extra inventories. However, during a pandemic or other period of major upheaval, the rules change: the degree of uncertainty is magnified and becomes difficult to assess. Consumers and suppliers aggravate the problem if they overreact, as with toilet paper during COVID-19.

Inherent Randomness

Even in normal time, forecast accuracy is limited by the extent of randomness in behavior. Using the previous example, your travel time to work may also vary with "subliminal factors" such as how tired you feel—you had a bad night's sleep—luck in just missing the change from a green to red light, a call that comes while driving, and so on. The degree of randomness in a variable determines its *forecastability*, and the quality of a forecasting method must be judged in light of the variable's forecastability.

> *If the nature of the demand is so gracious as to allow us to forecast it with 90% accuracy, then with good people, systems, and processes, we should be able to achieve that level of accuracy. But if the nature of the demand does not permit it to be forecast with 90% accuracy, then we never will ... no matter how much time and money and effort and sophistication we apply* (Gilliland, 2010).

There are ways we can reduce randomness, such as by aggregating data into more forecastable groups (e.g. using monthly rather than weekly data) or taking moving averages of volatile variables. Beyond a certain point, however, randomness cannot be reduced further, setting a limit to improvements in accuracy and lowering uncertainty. This is the notion of "unavoidable error" expressed by Morlidge (2013).

Hype

Forecasters are often the target of serious and, at times, legitimate complaints from forecasting users. Some of these surely come from negative experiences in the past and unrealistic expectations of what forecasting can achieve. We frequently hear arguments that if a forecast fails to achieve at least 90% accuracy, either the forecaster or the method used is not believable, this notwithstanding the margins for error reported in the forecast. Alas, consultants and software vendors are prone to exaggeration about the effectiveness of their forecasting toolbox. This is particularly the case with AI solutions and their brethren, which overpromise substantial accuracy improvements and problem-free implementation.

BARRIERS TO FORECAST IMPROVEMENT

There are many practical barriers to forecast usage, both in the initial phases of adoption and post-implementation, that can either stunt progress or render the process redundant. While our principal goal is to foster adoption of systematic forecasting in organizations where it is absent, we can build upon our knowledge of the impediments other organizations have faced to achieving their forecasting goals.

Challenging Preconditions

Preconditions for a systematic forecast methodology include having data sources (such as sales) that are at the appropriate levels, that don't suffer from latency, and that require minimal manipulation to eliminate erroneous or missing values. For some firms, the absence of such data creates a hurdle that requires cross-functional support and investment. Ideally, these data sources should be aligned to the master data used across the organization to provide a bridge to adoption in functional areas outside the one responsible for forecasting.

The 2020 coronavirus pandemic, which combines health and economic crises, presents a worst-case scenario in terms of the uniqueness of the lockdowns and their economic implications, making forecasting extremely difficult and uncertainty impossible to assess.

In the initial phases of adoption, there is often a lack of clear definition of how the forecast will be used—to support an operational process that consumes the data at a high level of frequency and detail, or to support a process that requires output at a higher level of aggregation, perhaps with a longer time horizon? Understanding the specific purposes of the forecasts is a key ingredient in process design. Too little attention can be paid to the units of measure, the time buckets to be used, and the hierarchy elements to include. There is limited understanding of supporting methods, such as clustering, to group similar hierarchical elements to provide the right balance of detail versus scale.

Process Design

Process design is often difficult because it requires cross-functional participation and engagement. It's often far simpler to design a process within a function, but this frequently fails to realize understanding, trust, and ultimately adoption by partners. In many instances, the lack of understanding of which inputs add value and which do not is a major cause of unsatisfactory outcomes.

There can be organizational anxiety about which function "owns" the forecast. In many organizations with an operational forecast output, ownership is in the supply-chain function. This is not to say that it can't also thrive in Sales or Finance. Much attention is paid to this, but little is given to decision rights (Gray, 2019). Who has the final say on the consensus forecast? If not thoughtfully considered, it can render systematic forecasting efforts redundant.

Forecasting Support Systems (FSS)

With the large number of FSS available, numerous selection considerations arise (Entrup and Goetjes, 2018). For those with existing ERP systems, should the tool be an advanced planning tool extension of that? Perhaps a best-in-breed solution is more appropriate? Many fail to follow a structured process of software selection, favoring what is suggested by the IT organization, often with little consideration of gaps or results from a proof-of-concept. Failure to consider these can lead to an unhappy partnership coupled with unfulfilled expectations.

Resistance can also come from the cost and difficulty of implementing an FSS. This is especially true for small and medium-sized firms. SMEs are unlikely to have the skilled staff to implement systematic forecasting nor the databases that these systems rely on. While there are cheap software products designed for such businesses, an additional barrier is that the use of a system may not match the way operations and tactical forecasting are carried out. Costly consultants may be required to ensure proper implementation and to train users.

Finding that small and medium-sized enterprises have lagged behind their larger counterparts in the adoption of suitable forecasting support systems, Matthias Luetke Entrup and Dennis Goetjes (2018) set out a structured process for the SME to identify, select, and implement an FSS that meets the organization's goals.

Metrics

Even when good designs and forecasting support systems are implemented, sustaining success and improvement can be elusive. Managing performance through the "right" metrics and applying improvement efforts specifically against those KPIs is a recipe for success. Too often, however, improvement efforts are applied against the biggest misses without consideration of what improvement is possible, considering the inherent unpredictability of the data.

Organizational Politics

Another source of resistance relates to human nature in the overall forecasting process. Forecasting can be a highly politicized process, with many human touch points. Each touch point becomes an opportunity for bias and personal agendas to contaminate what should be an objective, dispassionate process. Research has repeatedly shown that the more strategic the forecasts, even down to the annual budgeting cycle, the more senior (and inexpert) executives introduce bias and unnecessary inaccuracies.

The key questions then are how firms can achieve the most benefit from systematic forecasting, given that there are a wide selection of methods to choose from, many options for implementation, and a range of considerations in assigning responsibilities for the forecasting function.

GETTING STARTED FROM GROUND ZERO

Need for Historical Data

To initiate a systematic forecasting process, firms must recognize the necessity of developing a historical database. Doing so may require little or no monetary outlay. Initially there will be no need for consultants or expensive software. Instead, they would need to keep detailed information of the number of units sold at each time period of interest. Such data will allow firms to identify and exploit seasonality that contributes the most in improving forecasting accuracy. Later, they can record information about additional factors such as price, advertising, and promotions. These data can be also used for the objective estimation of budgets and cash-flow analysis.

Data should be captured at the most granular level (such as Item/Store for a retailer, or Item/Ship-to Location for a manufacturer, aggregated to days or weeks) and stored indefinitely (or aiming at least for 5+ years). Granular data can always be aggregated to higher levels based on product, location, or time hierarchies. Orders, sales/shipments, stockouts, and back orders would all be useful variables for constructing a time series approximating "true" customer demand.

For causal models, historical data on potential explanatory variables and other data features such as promotions, sales, and coupons need to be recorded. Implementation of ML algorithms benefits from such features as well as from data on related products.

Even when good designs and forecasting support systems are implemented, sustaining success and improvement can be elusive.

Exploring Naïve Methods and Developing Benchmarks

To evaluate forecasts, a firm needs benchmarks that put bounds on what can be achieved from historical data. A good start to generating benchmark forecasts is to explore several "naïve" forecast methods. The Naïve 1 and seasonal Naïve are two examples: for monthly sales forecasting, Naïve 1 uses the most recent month's sales as the forecast for the next month, while the sNaïve uses the sales of the same month of the previous year as the forecast for the current month. Analogous naïve forecasts can be calculated for data on daily, weekly, quarterly, or any other periodicity. The projections from a Naïve 1 reveal the future of sales if there is no change that increases or decreases sales from the most recent period. The projections from an sNaive extrapolate the seasonal pattern of sales from that in the most recent seasonal cycle.

Many other naïve variations are possible with simple arithmetic extrapolations of the data (e.g. the overall historical mean or median), testing their forecasting accuracy versus those produced within the firm. These simple benchmarks deliver a further benefit: when evaluating a more complex method (such as those proposed by a software house) they show how much of an improvement, if any, could be achieved from a potentially expensive new forecasting method. All too often they may reveal the inadequacy of the in-house forecasting processes: failure to beat the naïve is a damning indictment (Morlidge, 2014b).

When the scale of the data (number of time series) is relatively small, an inexpensive and ubiquitous tool like Excel could allow comparison of naïve forecasts to the internal judgmental or other projections made by the firm. (Larger firms with more time series would require more scalable data management like SAS.) Moving on from monthly forecasts, data can be also collected for weekly and daily sales figures to expand and benefit from the improved accuracy of systematic forecasting methods and the increasing need to plan on a shorter-term basis. Firms can also explore application of the methods to different periodicities (time buckets) such as weekly, monthly, and quarterly. There is good potential in averaging forecasts made from different time buckets (Petropoulos and Korentzes, 2014).

What we want to emphasize in this section is that a simple systematic forecasting system should be introduced step-by-step to test its value before more expensive solutions are adopted. Relative performance is best evaluated in relation to benchmarks, which will often highlight the need to adopt a more formal process of forecasting and evaluation. A common approach is that of calculating forecast value added, or FVA (Gilliland, 2013).

Stepping into Forecasting Software

Almost every software package—including spreadsheet add-ins—will offer a set of forecasting procedures known as exponential smoothing. This family of procedure extends the naïve methods by utilizing weighted averages of the most recent historical data. For example, while a Naïve 1 forecast for June would be the actual sales in May, the simplest exponential-smoothing procedure would forecast June sales as a weighted average of May, April, March, and continuing back in time, giving less weight to each month the farther back it is in time. More sophisticated members of the exponential-smoothing family would similarly measure and project any trend and seasonal pattern in the historical data. See Stellwagen (2012) for an introductory tutorial on exponential smoothing.

In addition to spreadsheet add-ins, there are inexpensive commercial packages, most requiring little training to begin usage. Fildes and colleagues (2020) have recently provided a survey of commercial software and their features. An increasingly popular solution that allows the usage of all popular forecasting methods is the free R library (Hyndman, 2019), although some learning effort is required to use it effectively. An alternative, Forecasting-as-a-Service (FaaS), is an emerging approach that some vendors are offering, which delivers cheap access to a variety of methods. We see software vendors increasingly offering ML methods—so, in principle, these advanced methods are becoming readily available, even though they cannot be used "out of the box."

As comfort with the software grows (and the historical database lengthens), the firm can begin experimenting with more advanced methods, comparing their effectiveness (and explainability and scalability) to the simpler methods. Available forecasting methods range from the extremely simple, such as single exponential smoothing, to the highly sophisticated, such as deep learning (DL), which require specialized knowledge and substantial computer resources to run. Both types of methods could be useful; the first when large numbers of forecasts are needed and there are constraints on time and resources to create them, and the second when even small improvements in accuracy/uncertainty are important to save large amounts of money by improving decision making in critical business areas. There are also methods of intermediate complexity. These can be considered by balancing accuracy/uncertainty versus interpretability and ease of use, as well as the computer time required to obtain the forecasts and measures of uncertainty.

For firms initiating a forecasting process, applying free and inexpensive software would allow them to see how well systematic forecasting fulfills their forecasting needs and how it can complement their managerial expertise.

There is a considerable body of knowledge to be found, including on the Web, in the many forecasting books, and in journals such as this one. These resources show how various methods work, when they work well, and when they seem to fail.

For firms initiating a forecasting process, applying free and inexpensive software would allow them to see how well systematic forecasting fulfills their forecasting needs and how it can complement their managerial expertise. Many organizations will find that shifting from purely judgmental to systematic methods of forecasting will provide a more reliable basis for their operational decisions.

Judgmentally Adjusting Statistical Forecasts

Forecasting methods are accurate if established patterns/relationships do not change during the forecasting period. This means that any changes such as a large order from a new customer, a major new promotional campaign, a significant price reduction, or a competitor going out of business will not be included in the forecast model, and thus will have to be incorporated into the final predictions judgmentally. A novel promotion would probably justify judgmental intervention, but in some cases we may have a sufficient record of the effectiveness of past promotions or price reductions to justify statistical modeling of their effects. Equally importantly, we should not let the optimism about the potential success of the promotion unduly influence its forecast.

Judgmental adjustments present a major management challenge. Advice in the forecasting literature on how to manage adjustments include:

- Avoid small adjustments to the forecast—even if directionally correct, they have at best a small impact on forecast accuracy and have little effect on decision making. Rather, concentrate on large adjustments that will impact the future by requiring changes to existing plans.
- Recognize and attempt to minimize optimistic biases in judgmental adjustments of statistical forecasts.
- Keep track of and document the reasons for the adjustments. Doing so reduces gratuitous adjustments and enables us to determine their forecast value added (Gilliland, 2013)—which adjustments are justified and which aren't.

Judgmental adjustment of statistical forecasts is attractive to executives for many reasons; it offers the forecaster control and allows the incorporation of myriad factors not included in the model. Particularly with complex ML methods, managers are "algorithm averse": they prefer to rely on their own judgments rather than on incomprehensible models. While research has shown the need to improve the effective incorporation of judgment into the statistical forecasts, for many companies this has proved difficult. The online Appendix [***https://foresight.forecasters.org/wp-content/uploads/UFOAPPENDIX_Aug26-2020.pdf***] summarizes key studies about the desirability and impact of judgmental adjustments and the manner in which they should be implemented.

OFFERING GUIDELINES

To organizations endeavoring to create systematic forecasting, we have few guidelines at present to offer that demonstrate an awareness and understanding of what constitutes best practices in the field. Some attempts at such guidelines include Morlidge (2010) and Smith and Clark (2011). Lacking such guidelines, companies may seek role models in other firms, and surveys of similar size organizations that have successful forecasting functions should be valuable. A useful preliminary to these surveys is holding direct interviews to identify successful firms and understand how they are utilizing forecasting.

We need also to conduct interviews with firms that do not use formal forecasting, to determine what information and motivation they would require to initiate a systematic forecasting process. To support this initiative, the Makridakis Open Forecasting Center (MOFC) at the University of Nicosia will sponsor a project of interviews and questionnaires, with *Foresight* serving as co-sponsor and forum for publication of results. Producing a set of guidelines for proper forecasting usage, as well as an inventory of best practices, will provide a valuable service to the field and increase the use of systematic forecasting. It may also help to identify "bad" practices, make firms aware of their negative consequences, and offer recommendations on how to do better.

CONCLUSIONS

The field of forecasting has advanced a great deal in recent years, while data availability and computer power have seen spectacular increases. The more apparent benefits of systematic forecasting should make adoption of such a process much more advantageous to organizations that have not yet "seen the light."

Producing a set of guidelines for proper forecasting usage, as well as an inventory of best practices, will provide a valuable service to the field and increase the use of systematic forecasting.

A key challenge is that of persuading more organizations of the considerable benefits from systematic forecasting. The central argument is the gain in business efficiency, accountability, and profitability that firms stand to realize utilizing systematic forecasting methods versus those with ad hoc judgment. Ultimately, the challenge is how to demonstrate to skeptics that a scientific/statistical approach to forecasting, while imperfect, still works better than the alternatives.

While we have focused our remarks on operational and tactical forecasting, even with strategic analysis some major components will depend on analytical methods. To establish credibility, this requires acknowledging to practitioners—and skeptical management—that a scientific/statistical approach often does not work very well because of the inherent limitations on forecastability. It also requires recognition, by all parties, of the difficult and challenging dilemma in which the forecaster is placed: having to show confidence about his or her predictions to management, while at the same time providing management with what can amount to a wide range of uncertainty around the forecasts.

Of one thing we are certain, however: forecasting skeptics are so used to the hype and overpromises of consultants and vendors that they are reluctant to believe anything. This can only be addressed, and must be addressed, with a refreshing dose of candor.

REFERENCES

Armstrong, J.S. (2006). Findings from Evidence-Based Forecasting: Methods for Reducing Forecast Error, *International Journal of Forecasting*, 22(3), 583–598.

Armstrong, J.S. (1978). Forecasting with Econometric Methods: Folklore versus Fact, *Journal of Business*, 51(4), 549-564.

Athanasopoulos, G., Hyndman, R.J., Song, H. & Wu, D.C. (2011). The Tourism Forecasting Competition, *International Journal of Forecasting*, 27(3), 822–844.

Boger, C. & Meldgaard, J. (2020). The M5: A Preview from Prior Competitions, *Foresight*, Issue 58 (Summer), 17-23.

Brown, R.G. (1959). Statistical *Forecasting for Inventory Control*, McGraw-Hill, New York.

Chase, C. (2019). Assisted Demand Planning Using Machine Learning for CPG and Retail, SAS whitepaper.

Entrup, M. & Goetjes, D. (2018). A Blueprint for Selecting and Implementing a Forecasting Support System, *Foresight*, Issue 50, 10-18 & Issue 51, 8-15.

Fildes, R. & Goodwin, P. (2007a). Against Your Better Judgment? How Organizations Can Improve Their Use of Management Judgment in Forecasting, *Interfaces*, 37(6), 570–576.

Fildes, R. & Goodwin, P. (2007b). Good and Bad Judgment in Forecasting: Lessons from Four Companies, *Foresight*, Issue 8 (Fall) , 5-10.

Gilliland, M., Tashman, L. & Sglavo, U. *Business Forecasting (Vol 2): The Emerging Role of AI and Machine Learning* (Wiley).

Gilliland, M. (2020). The M4 Forecasting Competition – Takeaways for the Practitioner, *Foresight*, Issue 57 (Spring), 5-10.

Gilliland, M. (2013). FVA: A Reality Check on Forecasting Practices, *Foresight*, Issue 29 (Spring), 14-18.

Gilliland, M. (2010). *The Business Forecasting Deal*, Hoboken, NJ: John Wiley & Sons.

Gray. C. (2019). Why Is It So Hard to Hold Anyone Accountable for the Sales Forecast? *Foresight*, Issue 54 (Summer), 38-43.

Hyndman R.J. (2020). A Brief History of Forecasting Competitions, *International Journal of Forecasting*, Issue 36 (1), 7-14.

Hyndman, R.J. (2019). Forecasting Functions for Time Series and Linear Models, *RDocumentation* ***https://www.rdocumentation.org/packages/forecast/versions/8.10***

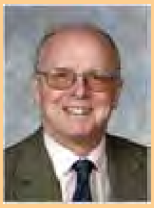
Hyndman, R.J. & Khandakar, Y. (2008). Automatic Time Series Forecasting: The Forecast Package for R, *Journal of Statistical Software*, 27(3), 1–22.

Spyros Makridakis is Professor and Director of the Institute for the Future at the University of Nicosia.

Ellen Bonnell is a consultant and author of "How to Get Good Forecasts from Bad Data" (*Foresight* Issue 7, Summer 2007).

Simon Clarke is a Principal of Crimson & Co and formerly Group Director of Forecasting at Coca-Cola.

Robert Fildes is Distinguished Professor of Management Science at Lancaster University and Director of the Lancaster Centre for Forecasting.



Mike Gilliland is Marketing Manager for SAS forecasting software and *Foresight* Associate Editor.

Jim Hoover is Director of the Business Analytics program at the University of Florida and Chairman of the *Foresight* Advisory Board.

Len Tashman is *Foresight* Editor.

Lawrence, M., Goodwin, P., O'Connor, M. & Onkal, D. (2006). Judgmental Forecasting: A Review of Progress Over the Last 25 Years, *International Journal of Forecasting*, 22(3), 493-518. **https://doi.org/10.1016/j.ijforecast.2006.03.007**

Li, Y., Berry, D. & Lee, J. (2020). How to Choose Among Three Forecasting Methods: Machine Learning, Statistical Models, and Judgmental Forecasts, *Foresight*, Issue 58 (Summer), 7-14.

Ma, S. & Fildes, R. (2020). Forecasting Third-Party Mobile Payments with Implications for Customer Flow Prediction, *International Journal of Forecasting*, 36:3, 739-760.

Makridakis, S. & Petropoulos, F. (2020). The M4 competition: Conclusions, *International Journal of Forecasting*, 36(1), 224–227.

Makridakis, S., Spiliotis, E. & Assimakopoulos, V. (2020). The M4 Competition: 100,000 Time Series and 61 Forecasting Methods, *International Journal of Forecasting* **http://dx.doi.org/10.1016/j.ijforecast.2019.04.014.**

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., et al. (1982). The Accuracy of Extrapolation (Time Series) Methods: Results of a Forecasting Competition, *Journal of Forecasting*, 1(2), 111–153.

Makridakis, S.G. & Hibon, M. (1979). Accuracy of Forecasting: An Empirical Investigation (with Discussion), *Journal of the Royal Statistical Society*, Series A, 142, 97–145.

Morlidge, S. (2014a). Do Forecasting Methods Reduce Avoidable Error? Evidence from Forecasting Competitions, *Foresight*, Issue 32 (Winter), 34–39.

Morlidge, S. (2014b). Forecast Quality in the Supply Chain, *Foresight*, Issue 33 (Spring), 26–31.

Osterholm, M.T. (2005). Preparing for the Next Pandemic, *Foreign Affairs* **https://www.foreignaffairs.com/articles/2005-07-01/preparing-next-pandemic**

Petropoulos, F. & Kourentzes, N. (2014). Improving Forecasting via Multiple Temporal Aggregation, *Foresight*, Issue 34, 12-17.

Salinas, D., Flunkert, V., & Gasthaus, J. (2017). DeepAR: Probabilistic Forecasting with Autoregressive Recurrent Networks, arXiv preprint arXiv:1704.04110.

Smith, J. & Clarke, S. (2011). Who Should Own the Business Forecasting Function? *Foresight*, Issue 20, 4-7.

Stellwagen, E. (2012). Exponential Smoothing: The Workhorse of Business Forecasting, *Foresight*, Issue 27 (Fall), 23-28.

Taleb, N.N. (2020). Statistical Consequences of Fat Tails: Real World Preasymptotics, Epistemology, and Applications, **https://arxiv.org/abs/2001.10488**

Van Hove, N. (2020). Technology Support in Business Planning: Automation, Augmentation, and Human Centricity, *Foresight*, Issue 58 (Summer), 43-48.

Virtual

FORESIGHT

Business Office: 53 Tesla Avenue
Medford, Massachusetts 02155 USA
https://foresight.forecasters.org

LAST ISSUE ALERT?

If it says **Last Issue Alert** above your name, take a couple of minutes to renew your membership now and keep *Foresight* coming your way.

Renew or start your IIF membership:
https://forecasters.org/membership/join/

Email our Business Director at
forecasters@forecasters.org