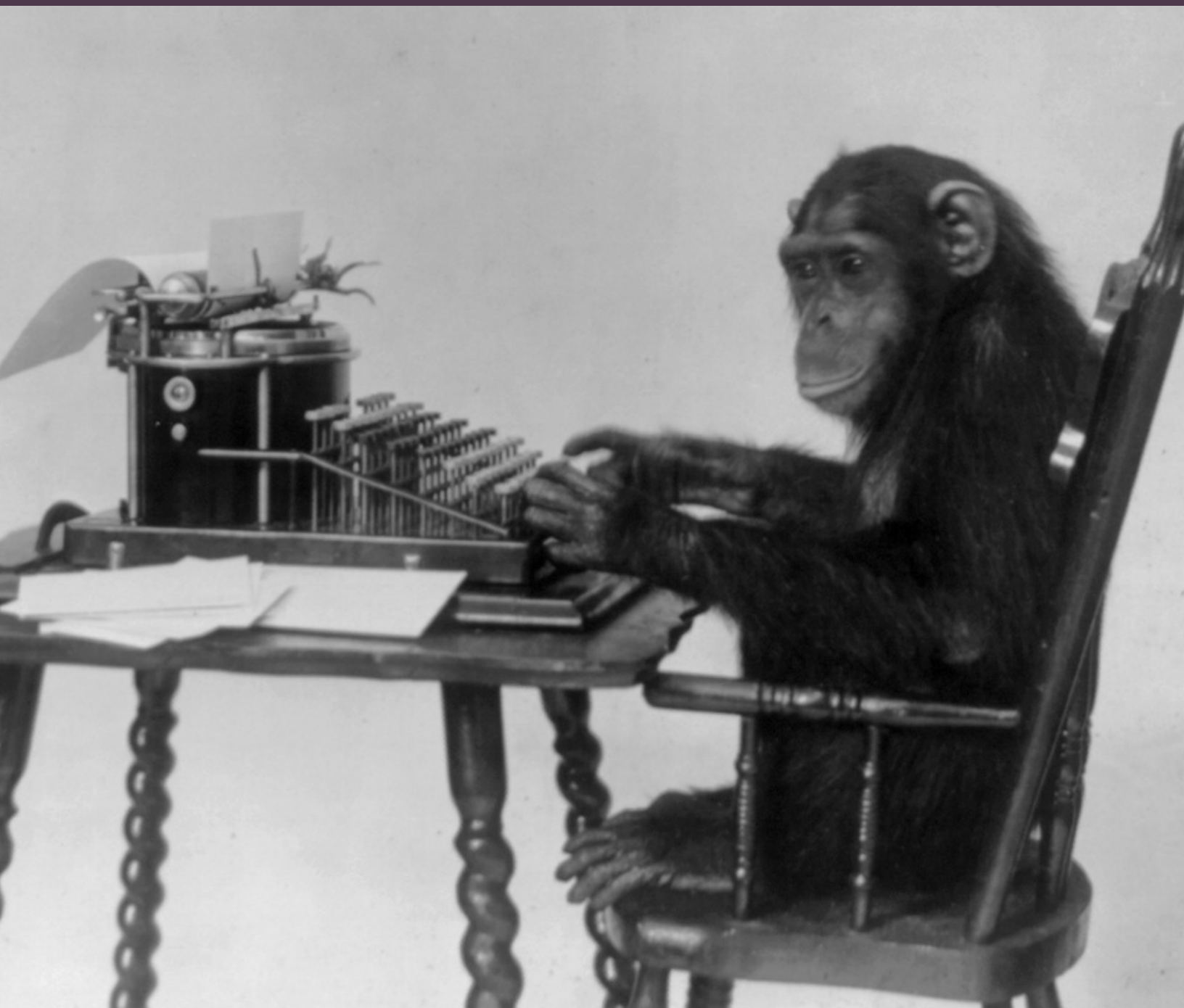
HOW TO BE LESS WRONG AND MORE USEFUL

THINKING CRITICALLY ABOUT DECISION MODELS



This guide was written by Kyle Novak.
The views expressed in this publication are solely those of the author.

Cover photo: New York Zoological Society | LC-USZ62-51619.

INTRODUCTION

Mathematical models are basic tools of quantitative decision making. But they are frequently misunderstood and misinterpreted, because many who must make decisions about them don't always know the right questions to ask. While this guide is born out of my experience working with development practitioners and the challenges that they face, this guide provides a broad framework for thinking critically about decision models.

A mathematical model is a representation of the real world using mathematical concepts or equations. Models are used to help simplify and explain complex systems. They can help managers prioritize how to best use limited resources. They can help policy makers test new strategies. And they can help decision makers ask probing what-if questions. For example, a life insurance model helps determine how much you pay in premiums, and a weather model helps you decide whether or not you should pack an umbrella. Mathematical models are used to better understand any number of wide ranging development challenges including HIV prevention,[1] poverty dynamics,[2] food security,[3] and family planning.[4]

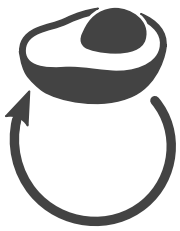
Statistician George E. P. Box once quipped "All models are wrong; some models are useful." As we know even the best weather model may miss a freak summer afternoon thunderstorm. On the other hand, it would be challenging for many of us to make an umbrella decision if all we are given is a complex digital map of evolving isopleths. While no model can be perfect, a worthwhile goal is to make decision support models less wrong and more useful. Complicated models like deep learning models and expansive multi-compartment models are virtual black boxes. It is important that the decision maker understand and trust what the black box is telling them. Because of this development donors and practitioners, proposal and grant reviewers, and contracting officer representatives should have a basic framework for thinking critically about mathematical decision-making models. This guide provides such a framework of best practices to make decision models less wrong and more useful. While the guide highlights specific decision models, the best practices apply to a much wider range of tools.

GERHARD BÖGNER | PIXABAY

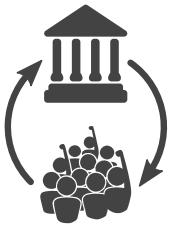
AN OVERVIEW OF MATHEMATICAL DECISION MODELS

There are many important mathematical models used in decision making. This section provides a brief overview of a few of them, along with specific applications.

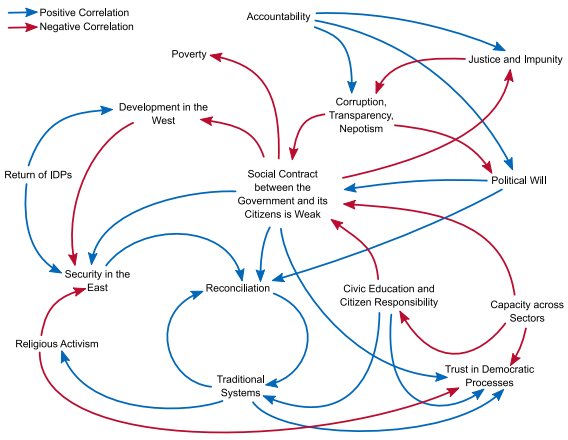
A system is an assembly of interacting components. For example, the human body is made up of a number of organ systems—the immune system, the digestive system, the cardiovascular system, and so on. Goods traded between industries to meet consumer demand is another system. And an ecosystem is a community of biological organisms and their physical environment. We can mathematically model a system as a network by assigning numerical values or functions to each component of the system and to each of the connections between those components, allowing us to make calculations about that system.



One of the simplest but most powerful systems is a feedback loop. A feedback loop is a system in which the output is routed back into the input, often resulting in growth or decay within that system. That avocado that seems to suddenly go from rock-hard unripe to mushy brown overnight—you can thank positive feedback loops. When an avocado ripens it releases a fruit-ripening plant hormone called ethylene. The avocado ripens slowly at first, releasing a small amount of ethylene. The ethylene causes fruit to ripen faster, producing more ethylene and causing the fruit to ripen even faster, producing even more ethylene and...you see where this goes. Another example, Moore's law of semiconductor doubling responsible for the digital revolution is the result of a financial-reinvestment-in-research feedback loop.[5] The dynamics that help keep a bicycle balanced upright and maintain economic law of supply and demand are both examples of negative feedback loops.



The feedback loop between a transparent, responsive government and an engaged civil society may be positive or negative, virtuous or viscous. The Millennium Challenge Corporation (MCC) in its Threshold Program with Kosovo recognized that civil society did not constructively engage with the Government of Kosovo due to a lack of publicly available data that led to a perception of poor government performance.[6] Therefore, MCC supported a series of open data challenges focusing on labor and gender, the environment, and judicial transparency to help foster productive partnerships between the Government of Kosovo, private sector, and civil society and strengthen positive, virtuous digital feedback loops based on data-informed decision making.

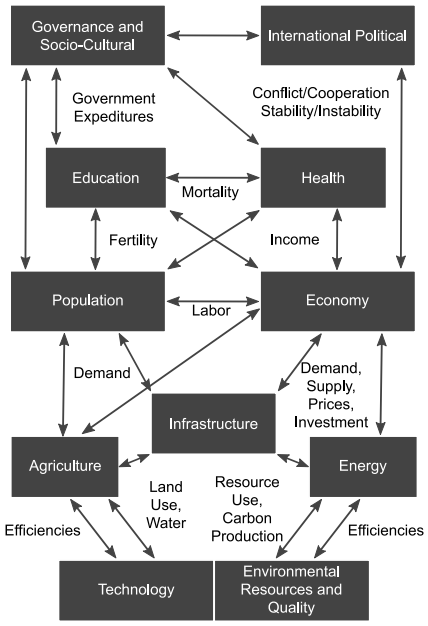


Causal loop diagram for a theory of change.[7]

An international development project begins with a development hypothesis and a theory of change, which describes how and why a result is expected

to be achieved in a particular context.[8] A theory of change can be depicted mathematically as a logic model illustrating the connection between what the project will do and what it hopes to achieve. One type of logic model, the causal loop diagram (CLD), is a visualization of the variables and cause-and-effect connections between any two variables.[9] Important features of CLDs include reinforcing loops (positive feedback loops) and balancing or self-correcting loops (negative feedback loops).[10]

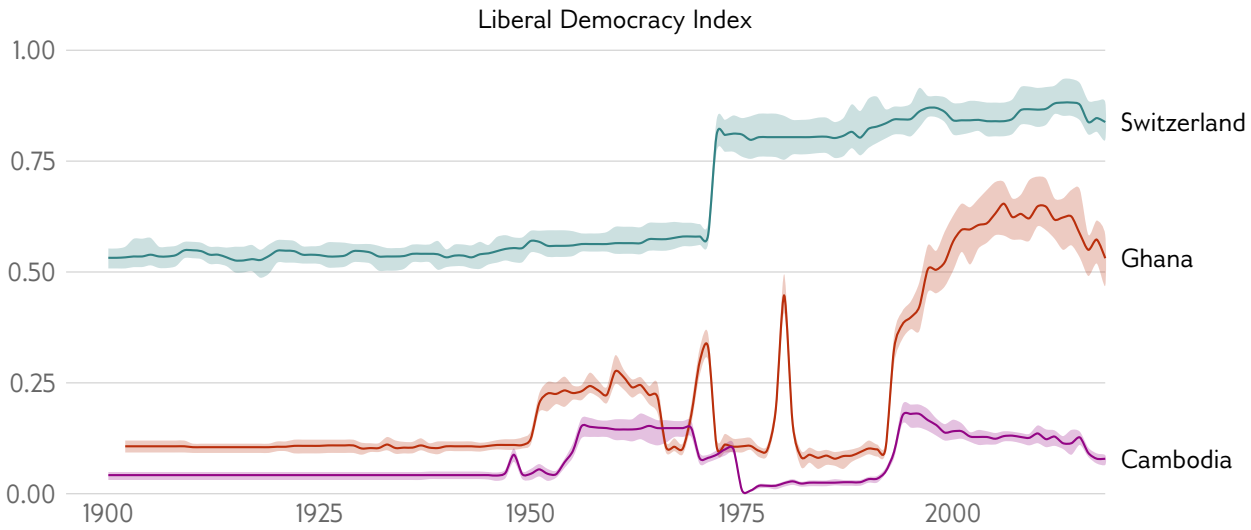
The logical framework, or simply LogFrame, is another logic model.[11] In a logical framework, activities implemented causally lead to outputs delivered, that causally lead to a purpose achieved, that causally leads to a goal achieved. With each step in the framework there are critical assumptions for the success of the development project. "*If* we plant seeds and we train farmers to cultivate them, *then* crops will grow *assuming* the right amount of rain." Reality is not quite this simple or certain. There may be several compounding factors and several options for intervention each with uncertain effectiveness. Climate change is a significant source of uncertainty as the cycle of extreme drought and flooding becomes more unpredictable in many regions of the world. Counterfeit seed is another source of uncertainty, which has prompted the government of Kenya to help farmers verify the authenticity of seed using mobile phones.[12] Because of the limitations of traditional, rational planning methods in a complex environment, the US Agency for International Development (USAID) has embraced adaptive management as "an intentional approach to making decisions and adjustments in response to new information and changes in context."[13] Adaptive management helps decision makers to not only change a system but to also learn about the system in the process and use that new information to improve future outcomes. In the context of adaptive management, a logical framework can be generalized using a Bayesian network, also called a Bayes net, a mathematical model for a set of variables along with their conditional dependencies. When applied to a logical framework, Bayesian networks help us examine conditional dependencies and estimate the likelihood that any one of several possible activities attributed to a given outcome. For example,"if crops did not grow, was it because of bad seeds, ineffective training, or too little or too much rain?" Bayesian networks are important decision support tool in adaptive management of natural resources but have not yet been adopted in adaptive management for international development. USAID's Learning Lab provides additional resources on developing CLDs and LogFrames.[14]



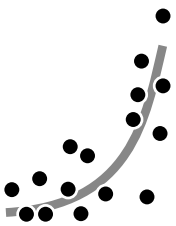
System-of-systems diagram of the IFs model.[15]

System dynamics (SD) is a method of modeling the behavior of complex systems over time using stock-and-flow, feedback loops, and statistical models. The International Futures (IFs) model, a system-of-systems model developed by the Pardee Center for International Futures at the University of Denver, helps identify long-term global trends and conduct strategic planning. The IFs model provides a country-level forecast (for 186 countries) by linking together several submodels for key global systems such as population, agriculture, health, climate, education, and economy. The model uses dynamic recursion, a system of several interconnected feedback loops along with statistical models, with annual time steps through the year 2100.[16]

An index is a composite statistic that aggregates multiple indicators providing a simple way to summarize, compare, and rank different objects and see change over time. For example, the Human Development Index (HDI) combines life expectancy, education, and income to rank countries into four tiers of human development. The USAID Journey to Self-Reliance (JSR) Country Roadmaps consist of seventeen indexes, called Self-Reliance Metrics, for low- and middle-income countries.[17] The Liberal Democracy Index (LDI), one of the Self-Reliance Metrics, is sourced directly from the Varieties of Democracy (V-Dem) Institute of the University of Gothenburg. It is itself a combination of two sub-indexes, which together combine eight indices, ultimately composed of 61 individual indicators.[18] The LDI measures the extent to which liberal democracy is achieved through measures such as constitutionally protected civil liberties, strong rule of law, an independent judiciary, and effective checks and balances, taking the level of electoral democracy into account. The LDI simplifies a complex sixty-one dimensional comparison down to only one dimension—0 (not democratic) to 1 (fully democratic). Using it, for example, we can quickly compare the liberal democracy of Cambodia (0.08) to that of Ghana (0.53) and Switzerland (0.84). However, without digging back down into individual indicators, it may be difficult for us to understand why Cambodia has such a low ranking one or Switzerland such as high ranking one.



Liberal Democracy Index comparison of Cambodia, Ghana, and Switzerland from 1900 to present.



A statistical model is a mathematical model that applies statistical assumptions or rules to observations. The assumption could be linearity—that we can fit the data with a straight line. It could be a distribution assumption. For example, wealth disparity is sometimes modeled using a Pareto distribution, also known as the 80/20 rule following the rule-of-thumb observation that 80 percent of property is owned by 20 percent of the people. Application of the generalized Pareto distribution leads to the Gini coefficient, a common gauge of income inequality within a country. The Famine Early Warning Systems Network, or FEWS NET, helps decision makers plan for food emergencies by using statistical models of crops, food prices, rainfall, etc. Indexes can also be constructed as statistical models where observations are combined using rules such as arithmetic or geometric means.

Similar to statistical models, machine learning (ML) models are a broad class of decision models that use historical observations or collected data about a smaller group to predict future behavior or behavior of a larger group. They can be used to simplify complex data to a manageable number of components or classify the data in a natural way. Machine learning models can also be used to assist automated decision making, or artificial intelligence (AI). Machine learning models use statistical models such as linear regression, logistic regression for binary classification, or Bayesian networks. Artificial neural networks—or simply neural networks—are an important class of machine learning models that extend the idea behind logistic regression to more complex classification. For an in-depth discussion of machine learning see the USAID report *Reflecting the Past, Shaping the Future: Making AI Work for International Development.*[19]

SUMMARY OF SOME MATHEMATICAL DECISION MODELS	
System	An assembly of interacting components
Network	Numerical values or functions are assigned to each component and connection of a system allowing calculations to be made about that system
Logic model	A chain (system) of causes and effects leading to an specific outcome
Causal loop diagram	A visualization of the variables and the cause-and-effect connections between any two variables of a logic model
Systems dynamics	A method of modeling the behavior of complex systems over time using stock-and-flow, feedback loops, and statistical models
Index	A composite statistic that aggregates multiple indicators providing a simple way to summarize, compare, and rank different objects and see change over time
Statistical model	A model that applies statistical assumptions or rules to observations
Machine learning	Broad class of decision models that use historical observations or collected data about a smaller group to predict future behavior or behavior of a larger group

MAKING A MODEL LESS WRONG

QUESTION THE MODEL

If indeed all models are wrong, it makes good sense to start by understanding what makes them wrong. One approach is to consult different subject matter experts familiar with similar models early in the design process to check assumptions and simplifications. Diversity in expertise can also help mitigate groupthink. If an organization building a model does not have in-house domain expertise, it is importance to work closely with external subject matter experts both locally and internationally during model planning, development, review, and testing.

Red teaming is another way to question a model. A red team is an independent group that challenges an organization, by assuming an adversarial point of view, to improve an organization's effectiveness. They can help staff to think critically and creatively, to see things from varying perspectives, and to avoid biases and group thinking.[20] Red teaming can help model designers to think about human decision making, to overcome failure of imagination, to re-examine assumptions, and to reconsider a solution's appropriateness in complex development scenarios.

Be wary of overly complicated models, especially assessment models. An assessment model uses inductive, bottom-up logic starting with complex data and synthesizes them to simpler data, perhaps even a single defining idea, leading to a broad understanding of the system. Analysis uses deductive, top-down logic to identify and understand the root cause in an assessment. Assessment and analysis go hand-in-hand. Assessment identifies a problem, and analysis searches for recommendation on solving the problem. Knowing that the Liberal Democracy Index of Switzerland jumped by 23 percent in 1972 is insightful, being able to analyze the assessment model to determine that the main driver was women's suffrage is powerful.

Indexes are useful for quick comparisons and rankings of countries, especially in domains where there may be many complex underlying variables such as human development and gender inequality. But constructing a good index is challenging. Notably, aggregation of the indices sometimes seems a bit like sleight of hand.[21] Statisticians often incorporate factor analysis to find correlations between variables and determine suitable aggregation weights to reduce the variables to a manageable number of dimensions. Common methods of aggregation include the arithmetic mean—e.g., $(0.1+0.7+0.\ 8+1.0) \times 1/4 = 0.5$—and the geometric mean—e.g., $(0.1 \times 0.4 \times 0.5 \times 1.0)^{1/4} \approx 0.37$. An arithmetic mean allows a high-achieving component to linearly compensate for a low-achieving component; while a geometric mean does not nearly allow such a compensation—a low-achieving component penalizes the index more and notably any zero-valued component results in a zero-valued index. The JSR Commitment and Capacity Metrics use an arithmetic mean for aggregation, and the Human Development Index uses a geometric mean. The V-Dem Liberal Democracy Index takes a different approach. It starts with factor analysis to reduce sixty-one indicators down to eight (all between 0 and 1). It then uses a "compromise" average of sums and products to further reduce the eight down to two components: *polyarchy* and *liberal principle*.[22] The polyarchy component is stretched to help set the anchor point 0.5 as the minimum for electoral

democracies. Finally, it finishes with another sum-product average to get the final LDI score.[23] The method is complex—scaling and recombining one index into another. To be sure, distilling sixty-one components down to just one measure in a meaningful and representative way is a challenge. Assessment models should not only help identify problem areas, they should also help identify root causes. And they should provide confidence that the problem is a reflection of the data and not the model itself.

Natural language processing models are increasingly used to automate decision making about people, such as screening job applications. If data used to train these models differ in their representation of men and women, they may influence how these people are perceived and who gets recommended. Therefore, it's important to question both the model and the data. In 2019 a team of researchers built a machine learning model to question underlying gender stereotypes in language, training their model using adjective–gendered-noun bigrams (pairs of words appearing together) from the massive Google Books data set.[24] The set contains over eleven billion words collected from 3.5 million books over the past several centuries. It goes without saying that an adjective like *pregnant* is much more likely to be associated with the noun *woman* than the noun *man*—but what about other adjectives? For example, what noun is most likely to follow the adjective *beautiful*? Using Google Books Ngram Viewer we find that *woman* is the most common noun, then *girl*, *women*, *things*, and *face*.[25] Other female nouns *wife*, *daughter*, *lady*, and *maiden* appear further down the list all before the first male noun *man*. In fact, *beautiful woman* occurs 20 times more often than *beautiful man*. The noun *man* itself is two to eight times more common than *woman*, so a simple bigram comparison is complicated.[26] Instead, we can use a mathematical quantity called the pointwise mutual information, or PMI, to measure the deviation of a bigram.[27] A large, positive PMI indicates that the two words occur frequently together, a negative PMI indicates that the words rarely appear together, and a zero PMI indicates that the words are uncorrelated. The PMI of *pregnant woman* is 6.3, and the PMI of *pregnant man* −0.9. For *beautiful woman* it's 4.7, and for *beautiful man* it's 0.7.

Instead of restricting comparison of the gendered-nouns to typical *man–woman*, *boy–girl*, and *father–mother* pairs, the researchers used a total of 23 gendered-noun pairs including *god-goddess*, *waiter–waitress*, and *wizard–witch*. The researchers exhaustively checked adjective–gendered-noun bigrams and noted the largest-deviation adjectives used to describe male and female nouns. This approach is problematic for several reasons. First, several nouns that the researchers categorized as masculine are often gender neutral or gender ambiguous, like *man*, *god*, *hero*, or *actor*. For example, the noun *gods* in the bigram *false gods* (PMI=5.4) is inherently neither male nor female. Second, some of the chosen gendered-nouns don't have equivalent pairings, such as *wizard* and *witch* or *lord* and *lady*. And many bigrams are idiomatic expressions. Take the phrase *financial wizard* (5.9). Between 1950 and 1980 almost 20 percent of all *wizards* were *financial wizards*.[28] There are several other idiomatic expressions like *feudal lords* (9.1), *dumb waiter* (6.9), *weird sisters* (6.9), and *virgin queen* (6.1). Scores of some bigrams are inflated because their base words are less common as in *gallant knight* (6.1) and *topless waitresses* (9.4). Furthermore, Google Books Ngram relies on an abundance of scientific literature.[29] So bigrams like *infected mothers* (5.0) are over-represented because of global health research concerning HIV-infected women which peaked in the 1990s.[30] Given these shortcomings, it is not surprising that among the top ranked gender-stereotyped adjectives that the researchers found are *false*, *financial*, *feudal*, *dumb*, *virgin*, *weird*, *topless*, *gallant*, *infected*, and *beautiful*.

What to ask: How will domain experts be used to develop or review model design?

EXAMINE MODEL ERROR AND SENSITIVITY

A model is said to be unstable when a very small input can result in very large output. Real-world data has real-world measurement errors and sampling biases, and these errors can be magnified within an unstable model and overshadow the desired output. On the other hand, if large changes in the input only produce small changes in the output, the model is likely not going to be useful for decision makers, because it will be difficult to differentiate between options. Overfitting and underfitting—creating an oversensitive, unstable model and creating an undersensitive model—are real concerns in machine learning models. Because of this, quite a bit of effort is put into model regularization and pre-conditioning, the science of finding a balance between overfit and underfit models.

The IFs model forecasts scenarios decades into the future. Because population growth is exponential, one should expect parts of IFs to be inherently unstable. Because political changes are often nonlinear shocks, one should expect parts of IFs to be overly stable. In the IFs model a country's governance tends to track its level of economic development—rich and autocratic countries like China and Saudi Arabia slowly become more democratic over time while poor and democratic countries like Malawi and Liberia slide gradually to lower levels of democracy. The direction might be right, but those changes are never quite so smooth in reality. By knowing this, a decision maker is able to cautiously and critically assess the outputs of the IFs model especially when using the model far into the future. In fact, it is laughable to blindly trust a forecasting model. *Science*, one of the top academic journals, published an article in 1960 that uses a mathematical model to conclude that the human population would be infinite before the end of 2026.[31] Such an obvious absurdity should cause one to stop and question basic assumptions of the model.[32] Still, the article has been cited 442 times. (Make that 443 times.) In linear models, where the output changes proportionally to the input, the sensitivity can usually be computed directly. In more complicated models, Monte Carlo methods can be applied to test sensitivity, by adjusting one or more of the parameters or inputs and comparing the change in the output.

What to ask: What sensitivity analysis has been performed on your model?

TEST THE MODEL

Models should be verified and validated, ideally independently. Verification checks if the model is implemented correctly, and validation checks if a correctly implemented model answers the problem statement. Without an objective, independent verification and validation process, it is difficult to know the limitations of a model. Academic researchers rely on peer review of other experts in their field. The military uses wargaming and exercises to verify and validate war plans. Computer models can often be tested using other computer models such as Monte Carlo simulation, a method that tests hundreds, thousands, or millions of what-ifs to estimate statistical likelihoods. Data scientists building machine learning models divide the sample data into three categories: one called training data to build the model, another called test data to verify the model, and the final called validation data to validate the model.

In 2016 two researchers announced a machine learning model that they claimed could predict with

90 percent accuracy whether a person was a criminal using only facial features.[33] The researchers trained the model using photos of faces of 730 criminals and 1126 non-criminals—all of them were Chinese, male, clean-shaven, without distinguishing features. The faces of non-criminals were scraped from the internet, presumably from their professional profiles on business websites. The faces of criminals came from police records. Some critical reviews noted that the research was simply automating racism, highlighting the discredited field of physiognomy. One critical review revealed that instead of developing a model detect "criminality," the model learned to simply detect the difference between a scowl (evident in the police photos) and a smile (evident in the professional photos).[34]

While a model can be verified with test data, validating the model is often more challenging. For example, how might one validate an epidemiological forecasting model without an active outbreak. This leaves open questions about how assumptions impact the mode like "does the sample population used to build the model represent the general population?" If a model is unable to be validated, at very minimum, it should be tabletop tested before being fielded to identify potential weakness.[35]

How do you validate a futures scenario model like IFs without waiting ten or twenty years to see if the forecast is accurate? The Pardee Center's approach is to start the model from a date in the past (say 1960) and see how well it agrees with the real world today.[36] But because the parameters of the model are derived from historical observations, the model is likely to perform well in validation. This is a bit like using training data in the place of test data.

What to ask: How will your model be verified and validated?

START FROM FIRST PRINCIPLES

First principles are the fundamental assumptions on which a system is modeled. A common pitfall among data scientists is modeling data without sufficiently understanding the first principles of the system generating the data. In machine learning, artificial neural networks—often simply neural networks—are inspired by biological systems of neural circuits. Data scientists can swap in and out any number of exotic activation functions or hidden layers to construct a custom artificial neural network. That the custom artificial neural network doesn't actually model a real biological neural network is irrelevant. Its purpose is to classify data in a way that is fast, robust, and accurate. Data scientists have a number of tools that they use to tame an otherwise ill-posed problem—take, for example, regularization in regression. An epidemiological model, on the other hand, ought to model real infectious diseases. Data-science quick fixes to make a model work may hide underlying errors and misassumptions in the model.

Occam's razor is a principle of decision making that tells us, when choosing between different explanations of an outcome, to favor the one with the fewest or simplest assumptions. Einstein is credited to have said: "everything should be as simple as possible, but not simpler." The military is more blunt: "keep it simple, stupid." When applied to selecting decision models Occam's razor tells us, all other things being equal, to use the one with the simplest set of assumptions and least number of variables. By using Occam's razor, a model is easier to implement, easier to test, and easier to interpret. Occam's razor applied to machine learning, for example, tells us to use low-dimensional, linear models

over high-dimensional, nonlinear ones. As a consequence we avoid overfitting that often results in an unstable model and a wrong prediction. When constructing an index such as the Self-Reliance Metrics, Occam's razor tells to reduce the number of superfluous variables by using perhaps statistical factor analysis. As a consequence, root cause analysis becomes easier and recommendations clearer.

When down to its bare essentials, a decision model should be explainable. There's an expression that says that you don't really understand something unless you can explain it to your grandmother. Mathematical formalism, the gears of a model's machinery, often obfuscates a model's underlying first principles. While designing a mathematical model may require a technical understanding of say Bayesian statistics and stochastic calculus, describing it should only require plain language. Statistical models, for example, can many times be explained by the functional form. When modeling the influences on voting for Salvador Allende in Chile, researchers Soares and Hamblin repurposed an empirical relationship found in psychophysiology (the study of how the mind and body interact) to model voters' socio-economic alienation.[37] Whether or not such a model was appropriately appropriated is call for re-examination, but at least the model was explainable. The World Bank's Human Capital Index formulation grows out of standard equations in development accounting.[38] And the IFs system-of-systems model can be broken down into subsystems. Each subsystem can then be modeled using the appropriate first principles. For example, the population subsystem is largely birth, death, and migration.

What to ask: Can you explain your model in plain language?

DEFINE PARAMETERS AND VARIABLES PRECISELY

Parameters and variables in a model should be meaningful, unambiguous, and as much as possible in plain language. This is important not only to avoid confusion but also to build trust in the model. It can also simplify dimensional analysis, an easy but often neglected check that units of measurement are consistent within a model. Dimensional analysis is not particularly rocket science, but after a $125 million NASA Mars orbiter crashed due precisely to a mix-up between metric and English systems, maybe it ought to be.[39]

The IFs model has hundreds of different parameters measuring everything from the quantity of amphetamine use to the world forest area to the likelihood of nuclear war. Pardee uses a searchable online wiki to help users know the precise definitions of the variables and how they are used in the model.[40] The V-Dem Project provides a 389-page codebook for the hundreds of variables that it uses.[41] A codebook is an explanation of the structure, methodology, weights. etc. for each variable of a data collection. Without precise definitions decision makers may misinterpret precisely what the model is saying.

What to ask: Will your model be accompanied by a technical lexicon or codebook?

TELL THE TRUTH

It should go without saying that above all a model and the use of that model should be truthful. Don't force model outputs to fit your agenda. Don't let your own cognitive biases lead to self-deception. Don't persuade the decision-maker by overwhelming them with a disregard to the truth. Information journalism and design expert Alberto Cairo has called this deceptive use of data "trumpery."[42] Information scientist and academic Jevin West simply calls it "bullshit."[43]

A 2019 World Economic Forum video highlighted the gender stereotype research discussed on page 7. The video stated that "beautiful, sexy and gorgeous were used most often to describe women," while "men were most frequently called brave, rational and righteous."[44] This statement isn't quite true. *Young*, *old*, *black*, *white*, *good*, and *poor* are the words most often used to describe men or women regardless of gender. Women are also *beautiful* and *pregnant*, and men are also *great* and *dead*. The research instead found adjectives from the most frequent adjective–gendered-noun collocations.[45] But even after excusing these semantics, it's still not true. Of the words *beautiful*, *sexy*, *gorgeous*, *rational*, *brave*, and *righteous*, only *beautiful* and *rational* are among the top 150 adjectives reported in the research article. So why would WEF choose these six words? Because they appear with 38 others in a prominent, colorful table on the article's first page—less than a quarter of which are in the a table of the top 150 several pages into the article . The summary table is clearly misleading. Researchers should work harder to communicate truthfully.

In 1954 Darrell Huff wrote *How to Lie with Statistics*, showing how misuse of statistics and data visualizations could falsely represent reality. Although Huff was not a trained statistician, the book quickly became a standard textbook for many college students and has since become perhaps the most popular book on statistics ever published. In 1964 Huff was hired by the Tobacco Institute to write a sequel *How to Lie with Smoking Statistics*. This book, which continued Huff's successful tongue-in-cheek style to popularizing statistical literacy, was authored in collaboration with an industry congressional and legal defense attorney. Huff now sought to use humor and anecdotes to undermine known health concerns and prove to the public that the Surgeon General's claims about smoking were based on merely poor statistics. The book, however, largely demonstrated Huff's misunderstanding of statistical significance and was never published over concern that it might potentially result in a false advertising lawsuit for misleading readers.[46]

Misunderstanding of statistical significance is so widespread including among many practising scientists that in 2016 the American Statistical Association was prompted to release a statement on the proper use of statistical significance.[47] Indeed a prominent meta-study found that most published research findings are false.[48] A 2016 survey in the journal *Nature* furthermore found that more than 70 percent of researchers were unable to reproduce another scientist's experiments and half of them were unable to reproduce their own.[49] Because it is usually impossible to study an entire population, the standard technique in inferential statistics is to construct statistical models using a much smaller sample of the population. When testing for some behavior in the population, two rival models are compared: a null hypothesis model that lacks the behavior in question and an alternative hypothesis model that has the behavior. Evidence is then collected to determine which of the two competing models better represents the entire population. A *p*-value is the probability that the observations would occur under the null hypothesis model. A *p*-value of 0.05 (five percent) is standard for rejecting the null hypothesis and choosing the alternative hypothesis as a model for the population. While 0.05 is standard, there is nothing magical about using this value as the cutoff. It is really important to note that a *p*-value is the probability that the observa-

tions occur if the null hypothesis were true and not the probability that the null hypothesis is true given the observations.

Suppose we wanted to know which of two political candidates Maria or Peter are favored to win a district election. We conduct a straw poll of say fifty voters—equally men and women–in the district. Our null hypothesis is that everyone is equally likely to vote for either Maria or Peter, and our alternative hypothesis is that either Maria or Peter is a favored candidate. We decide to use a binomial distribution which models likelihood of several trials with two possible outcomes, such as flipping coins.[50] Suppose 20 people we polled said that they will vote for Maria and 30 said that they will vote for Peter. The calculated *p*-value is 0.10 and using 0.05 as a *p*-value cut-off, we would conclude that we have insufficient evidence to decide whether Maria or Peter was the favored candidate. The behavior in the poll may simply be due to randomness of having such a small sample size. Discouraged that our straw poll failed the hypothesis test, suppose we now reexamine our data using voter gender as a discriminating factor. This time we see that 7 of the women polled would vote for Maria and 18 of them would vote for Peter. Now the calculated *p*-value is 0.02, well-below our 0.05 as cut-off. Can we conclude that women would likely vote for Peter over Maria? Perhaps and perhaps not.

Data dredging, also known as p-hacking, is trying multiple different factors until you get a desired result.[51] Put another way, data dredging is looking for patterns in data that can be presented as statistically significant when in fact there is no real underlying effect. This is done by performing many tests on the data and only paying attention to those that come back with significant results, instead of stating a single hypothesis about an underlying effect before the analysis and then conducting a single test for it. Data dredging is easy to do, and often it's done mistakenly. Indeed, it is natural for us to look for and want to find patterns in data.[52] We might wonder whether women are more or less likely to vote for a Maria or Peter? What about men? Young people? People who skip breakfast? People who don't like spinach? Out of twenty factors, we can expect that at least one would result in a *p*-value less than 0.05 just by chance. And we have misled ourselves and others into thinking that there is an underlying factor.

Using an adaptive management approach, we may try several different interventions each with different factors. During implementation, we learn which factors lead to successful interventions and adapt the program with a deeper understanding of those factors. Malawi is one of the world's most tobacco-dependent economies and raw tobacco in Malawi accounts for up to 71 percent of its exports.[53] This is in part an outcome of successful USAID and World Bank programs in the 1990s to economically develop the desperately poor country by supporting tobacco cultivation among smallholder farmers.[54] But, because tobacco is now mostly sold on contract, four out of five farms lose money growing tobacco.[55] And because of declining global tobacco demand and prices, tobacco cultivation is simply not sustainable development.[56] Furthermore, tobacco cultivation uses child labor, can cause tobacco poisoning among farmers, and is a major driver in deforestation.[5758] Now, Feed the Future is helping smallholder farmers in Malawi diversify away from tobacco to other crops such as soybeans, groundnuts, and sweet potato and adopt technologies to increase yield.[59] There are several different factors that affect a successful program. Under adaptive management, implementers would make decisions on limited data throughout the program that represents the larger population of smallholder farms in Malawi.

What to ask: How do you ensure that your model is truthful?

MAKING A MODEL MORE USEFUL

KNOW THE DECISION MAKER AND THE DECISION-MAKING PROCESS

The Principles for Digital Development are nine guidelines for using digital technologies in international development programs.[60] The first and fourth principles "design with the user" and "understand the existing ecosystem" can, when applied to decision support models, be "know the decision maker" and "understand the decision-making process." Each person or organization may have a specific decision-making style. Each situation may have a unique decision-making process such as rational planning, crisis management, adaptive management, or behavior change. The decision maker may be a ministry of health or a front-line health worker or an ordinary citizen. Communicating model outputs could be through a white paper, a weekly staff briefing, tweets, or an interactive dashboard. Start by asking who are the decision makers? What are their unique processes? How can the decision model be integrated into the existing process? Is the model too simple or too complex? Does the model answer the questions of the decision maker?

The District Health Information System 2 (DHIS2) is open-source software used as the national health information system in 67 low and middle-income countries.[61] The system includes aggregated data such as health facility data and population estimates, and event data such as disease outbreaks and patient records. Health care decision makers in these countries are familiar with this system and have integrated it into their decision processes. Developing an epidemiological model's front-end so that it is visually intuitive for someone trained in DHIS2 can make it easier for health care workers to use the model. Going a step further, actually integrating the model into DHIS2 makes it even more useful and more suited for the greater user adoption.

During strategy development, the IFs model can be combined with other futures scenario planning techniques such as workshopping, backcasting, and scenario driver focus groups. Using the IFs model interactively during workshop forecasting and backcasting exercises helps decision makers to anchor on objective evidence and avoid unachievable goals. At the same time these exercises help build technical capacity for workshop participants to be able use the IFs model independently in future strategic planning. Analyses can be further also presented as a white paper along with workshop out-briefs to help senior decision makers better understand and provide context and feedback on what is missing from the model.

What to ask: How does your model integrate into existing decision processes?

START WITH THE DECISIONS BEING MADE

A common mistake in decision making is starting with the data being collected and not the decisions being made. This error has permeated into policy and business management jargon as "data-driven decision making," implying to some that data leads to decisions. Starting with the data often limits the range of decisions being made by curtailing the brainstorming and divergent thinking required for problem solving and the structured thinking required for strategic planning. Data that is easily collected or at hand may lead to anchoring, i.e., relying on initial information to make subsequent decisions. And training machine learning models on available historical data may result in a biased model. For example, arrest records reflect information about criminals who have been arrested but not necessarily all crimes that have occurred.[62] Using predictive policing models off of arrest records may itself perpetuate and reinforce the bias by creating a positive feedback loop.

In the management book *Good to Great: Why Some Companies Make the Leap…and Others Don't* author Jim Collins and his team of researchers analyzed data to identify seven characteristics of companies that went from "good to great." Several years later economist Steven Levitt pointed out that few of the "great" companies had subsequently outperformed the S&P 500 and some had gotten into serious trouble.[63] He concluded that such books are mostly backward-looking and shouldn't guide the future. The error that Collins and his team made is a form of selection bias called survivorship bias, focusing only on data that made it through some selection process and overlooking those that didn't. International development programs often rely on case studies to identify best practices for sustainable development. Because failed pilot programs may not survive long enough to be fully documented, especially in a risk-averse culture, they may not be considered when examining best practices. Ultimately, this may lead to a false belief that certain successful interventions have some special property instead of merely coincidence. Organizations often realize this short-sightedness and embrace the concept of "failing forward," openly learning from failures as well as successes.

Moving from data-driven decision making to decision-driven-data-driven decision making requires being mindful of the role of and use of data throughout the entire decision-making process. Consider a *data-to-action framework*, adapted from the Manchester Centre for Development Informatics' "information value chain" framework.[64] During planning, the data-to-action framework starts by anchoring us on the outcomes and impact that we are trying to achieve. By focusing on the objectives, we are able to more clearly define and frame a problem statement. Under adaptive management these outcomes may change over time. A well-defined problem statement allows us to begin model development and selection. It also helps us identify the right decision makers and the decision-making processes in the organization. We can then confirm that the model is appropriate for the decisions being made. Now, identify the assumptions, constraints, and the critical information requirements. Models simplify the real-world situation to make them solvable. There are real-world constraints such as data availability, staff technical capacity, and project timelines. And there are ethical and legal constraints such as data privacy laws like the European Union General Data Protection Regulation (GDPR) and fairness guidelines. Often only after initial analyses do we uncover information gaps that lead us to reassess the problem statement. One exercise to help identify information gaps is stakeholder mapping that examines all those involved in the decision process to help identify how data is collected and shared. Finally, we consider the data requirements and assess the information gaps. We think about what happens if we don't have all the data, such as making changes to the model or using proxy variables.

DATA-TO-ACTION PLANNING	
1. ACTION	Define the objectives and identify the problem statement.
2. DECISION	Identify the decision makers and the decision-making processes.
3. INFORMATION	Identify the assumptions, constraints, and the critical information requirements.
4. DATA	Determine the data requirements. Assess the information gaps.

Execution starts with data collection and data cleaning, identifying and validating the data. Real data gives us the opportunity to finally test the model. We turn data into information and insight through analysis and visualization. Information should reflect the model outputs in context with model assumptions and limitations. It should also be contextualized to the real world. Finally, we provide recommended courses of action. While a mathematical model may be rational, the human decision maker often is not. Think critically about decision-making biases. Also, think about the best ways to communicate with the decision maker: a dashboard, a report, a tweet? A decision maker chooses a course of action. These actions have effects—intended and unintended. Think about how feedback loops help learning and adaptive management.

DATA-TO-ACTION EXECUTION	
1. DATA	Collect and clean data.
2. INFORMATION	Turn data into information through analysis and visualization. Provide recommended courses of action.
3. DECISION	Think critically about biases. Choose a course of action.
4. ACTION	Think about how feedback loops help learning and adaptive management.

What to ask: What questions does your model help answer?

PRESENT ACTIONABLE RECOMMENDATIONS

"What's the so what?" Mathematical models by design are abstract, and the outputs of them are removed from real-world context. And because mathematical models are simplifications of the real world, it is important to be transparent about assumptions and constraints in any recommendations. Without the assistance of experts who understand both the model and the real-world context, decision makers may misinterpret model recommendations.

The Liberal Democracy Index (LDI) combines 61 dif-

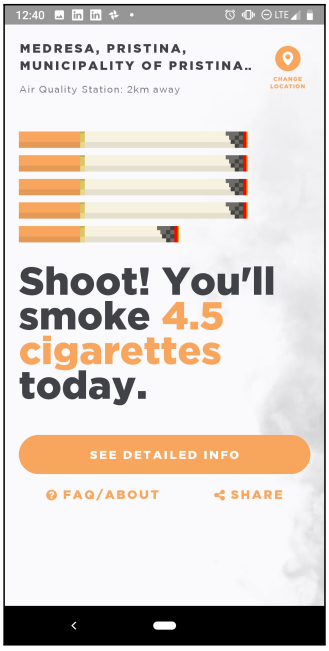
ferent indicators to provide a single decimal number between zero and one. By itself the number is hardly actionable. So, the V-Dem Institute does not report an index by itself. Instead, they present them together as perhaps a time series to tell a story about the progress of a country with coverage that may span 228 years. By comparing the change in the LDI over a 10-year period, V-Dem is able to tell the story of "advancers" and "backsliders," countries where citizens most experienced advances or declines in access to democracy, political rights, and civil liberties. By examining a side-by-side comparison of the 178 states, the V-Dem Institute is able to tell a story that autocratization is gaining momentum and that political exclusion due to socio-economic status is making the rich even more powerful.[65] And all of this is presented in context of the components like suffrage, clean elections, and rule of law. What becomes strikingly clear in V-Dem's presentation is that the LDI less a number and more a framework.

What to ask: How will your model help us make better decisions?

USE INFORMATION TO TELL A STORY OR ELICIT CURIOSITY

"Tell me the facts and I'll learn. Tell me the truth and I'll believe. But tell me a story and it will live in my heart forever."[66] Not all decision support models need to lead to explicit recommendations. Some present information in a compelling way to raise the awareness of a decision maker or to influence behavior.

Air pollution is a major health concern in many developing countries, causing 4.2 million early deaths every year and affecting 91 percent of the world's population.[67] Because people often don't understand the dangers of air pollution, fighting air pollution often starts by informing civil society before they can take action. Governments typically report particulate pollution matter using an Air Quality Index (AQI). This information can be hard for a citizen to use for making decisions. Analogously, if you grew up in the United States or the handful of countries in the world that solely use Fahrenheit to gauge temperature, you may have a hard time knowing whether or not 15°C is sweater weather—it is. The smart phone app "Sh**t! I smoke" shows air pollution visually as the equivalent number of cigarettes that you smoke a day.[68] Knowing that you are getting the equivalent pollution of ten cigarettes or more a day is often more meaningful to a citizen than reporting an AQI of 280. Even in countries where cigarette smoking is common, a billboard image of a baby smoking would evoke an emotional response.



*"Sh**t! I Smoke" smart phone app.*[69]

Advances in digital visual and user design have made data storytelling increasingly immersive. Smart infographics are the norm and software for producing them is common and affordable. The New York Times frequently includes interactive data visualization with their online articles to better draw the reader into the story. Data-driven documents or D3, the technology behind the New York Times visuals, is available as the open-source JavaScript library D3.js. And there is a growing community of users and developers who realize the power of dynamic data exploration. For example, Keshif, a data exploration platform often used in development projects, is built on top of D3.js. Data visualization platforms such as Flourish Studio, Datawrapper, Google Data Studio, and Tableau are making it easier for people to tell their data stories using innovative charts, maps, and interactive web apps. The V-Dem Institute uses Highcharts, yet another immersive data visualization platform, to draw decision-makers and analysts into interactive radar charts, time-series plots, and forecasting tools.

Millenium Foundation Kosovo (MFK) is an implementing entity of the Threshold Program agreed between the Government of Kosovo and MCC. One of MFK's projects, the DigData challenge, is an open data challenge to improve analytical use of labor force data by civil society, business, and the Government.[70] The challenge included a communications track that specifically focused on creating viral messages about labor data. Finalists of the challenge produced series of videos, interactive exploration dashboards, and infographics that could easily be shared on social media, liked and retweeted. Small nudges can be effective in influencing behavior and improving decision making.

What to ask: What about your approach makes people listen?

MAKE THE MODEL SUSTAINABLE

The best model is worthless if it is never used. The model may be too complex and require specific domain expertise. Data may be too costly to collect or too slow to analyze. Software may not be updated, or it may be proprietary and require licenses for additional users. Decision makers may not have the training to take action on the results. Or the model may simply not fit into current organizational decision processes. And as sankey diagrams, bump graphs, treemaps, and sparklines replace overused, outmoded, and often misused pie charts and bar graphs, decision makers may need to be trained to interpret them. To help improve model adoption incorporate user training and model familiarization into the decision-making process.

One step toward making a model sustainable is by providing user documentation. The IFs model has an in-depth Wiki that explains virtually all aspects of the model's subcomponents. The V-Dem Institute provides hundreds of pages of documentation to help users better understand the Liberal Democracy Index.

Design with the user as well as the decision maker. The IFs model has a richly developed data access layer but an outmoded presentation layer that overlooks many human-centered design principles. Frequently used menu items are several layers deep, exporting data is difficult, and variable names are unintuitive. User design of the platform makes adoption difficult. In the 1960s the Lockheed Skunk Works Engineering division operated under the "keep-it-simple-stupid" principle of design, also known as the KISS principle. It called for designing aircraft components so "simple-stupid" that they

could be fixed by the average mechanic in combat conditions with minimal tools. Incorporate the KISS principle into model design for a development program. Otherwise, in all likelihood, the user will find a way to break it. While many advanced econometric models exists to forecast food prices, such models require trained economists and statisticians. Instead FEWS NET staff and partners come from a wide variety of backgrounds and rely on tools that can be easily learned, adapted, and implemented often using a simple spreadsheet.[71]

What to ask: How will your model be maintained in the future?

BE TRANSPARENT AND OPEN

Every decision maker should be wary of any black-box model.[72] To build trust with the decision maker acknowledge the weakness of the model. Be transparent on the model limitations. Specify the simplifying assumptions and the effect of those assumptions on recommendations. Communicate error margins and error sources. Capture and communicate constructive lessons learned during model development and potential next steps to strengthen a model and its sustainability.The lessons not only help others on future projects, but they also highlight that the design team thought critically about the model.

One of the Principles for Digital Development advocates using open standards and existing open platforms.[73] Jupyter notebooks are open-source web applications for creating and sharing documents that contain text, mathematical equations, executable code, and visualizations. Jupyter—a portmanteau of its three core programming languages Julia, Python and R—is frequently used in numerical simulation, mathematical and statistical modeling, and machine learning.[74] Colaboratory and CoCalc are two online computing environments that use Jupyter and support collaboration by synchronizing changes in real-time. Former World Bank Chief Economist, Nobel laureate, and Jupyter user Paul Romer has noted "The more I learn about the open source community, the more I trust its members. The more I learn about proprietary software, the more I worry that objective truth might perish from the earth."[75]

When researchers published their findings on language gender stereotypes discussed on page 7, not only did they provide a mathematical description of their model, but they provided the Python code that they used to implement the model as Jupyter notebook.[76] This allows others to check the results and to easily adapt the model to for other projects. And the creators of the "Sh**t! I Smoke" smart phone app provide their code on GitHub, a software repository commonly used to host open-source projects, allowing others to learn from and build on their work.[77]

What to ask: How will others observe, critique, or build on your model?

FIRST, DO NO HARM

Few people have ever heard of the Gaussian Copula model let alone know of its impact on the world. The Gaussian Copula is a statistical decision model that helps financial managers assess risk between multiple complex, interacting components by flattening 7,750 correlations into just one.[78] Shortly after it was developed in 2000 the model experienced exploding popularity, even though many that used it didn't understand its shortcomings. The Gaussian Copula has been called *"beautiful," "simple,"* and *"ingenious;"* it has also been called *"the formula that killed Wall Street"* for its role in the 2008 global financial crisis.[79] The financial meltdown and the ensuing recession has been estimated to have wiped $10 trillion out of the global economy (one-sixth of the world GDP) and caused millions to lose their homes and livelihoods.[80] A maxim of health and humanitarian aid practitioners is to "first, do no harm." As data and artificial intelligence become evermore ubiquitous in our lives and have greater power to do harm, the "first, do no harm" maxim is being frequently applied to mathematical decision models. Big data can undermine our privacy, black box models can forfeit our liberties, and invisible algorithms can change our lives in profound ways. Mathematician and data scientist Cathy O'Neil calls such algorithms WMDs for widespread, mysterious, and destructive.[81] When algorithms make decisions about people, they may replicate and even amplify human biases. These models create vicious, positive feedback loops that reinforce discrimination and other forms of bias. But, in the age of the algorithm, who is to blame?

When an epidemiological model is wrong, scarce resources may be misallocated. When a Self-Reliance Metric is wrong, a country's strengths and weaknesses may be misrepresented. When the IFs model is wrong, strategic planners may underestimate the long-term effects of the climate crisis or of a youth bulge. When a FEWS NET crop model is wrong, relief agencies may be inadequately plan for a humanitarian crisis. We may wonder "What is the impact of using a model?" We may just as well wonder "What is the impact of not using a model?"

What to ask: What happens when your model is wrong?

PRZEMYSŁAW TROJAN | PIXABAY

FINAL THOUGHTS

Even when we make the best attempts to build and implement mathematical decision models with these principles in mind, we will encounter the realities of the real world. There may be delays procuring needed training data. Changeover in key personnel may make maintaining project champions difficult. In addition there are ethical and legal constraints that rightly limit how personal data can be collected and shared.[82] So, perhaps a parting principle should be "Be prepared for the realities of the real world."

While this guide focused on mathematical decision models, the considerations and critical questions for being *less wrong* and *more useful* apply to any number of mechanism and tools uses to solve development challenges. These include innovation challenge proposals and emerging digital technologies. The word "model" could just as well be replaced by "concept," "approach," or "solution." By thinking critically and strategically we can continue to find better solutions that save lives, reduce poverty, strengthen democratic governance, and help people emerge from humanitarian crises and progress beyond assistance.

UNKNOWN | C. 1910

KEY QUESTIONS TO ASK

QUESTION THE MODEL
How will domain experts be used to develop or review model design?

TEST THE MODEL
How will your model be verified and validated?

EXAMINE MODE ERROR AND SENSITIVITY
What sensitivity analysis has been performed on your model?

DEFINE PARAMETERS AND VARIABLES PRECISELY
Will your model be accompanied by a technical lexicon?

START FROM FIRST PRINCIPLES
Can you explain your model in plain language?

TELL THE TRUTH
How do you ensure that your model is truthful?

KNOW THE DECISION MAKER AND THE DECISION-MAKING PROCESS
How does your model integrate into existing decision processes?

START WITH THE DECISION BEING MADE
What questions does your model answer?

PRESENT ACTIONABLE RECOMMENDATIONS
How will your model help us make better decisions?

USE INFORMATION TO TELL A STORY OR ELICIT CURIOSITY
What about your approach makes people listen?

MAKE THE MODEL SUSTAINABLE
How will your model be maintained in the future?

BE TRANSPARENT AND OPEN
How will others observe, critique, or build on your model?

FIRST, DO NO HARM
What happens when your model is wrong?

RESOURCES MENTIONED IN THIS GUIDE

The following list summarizes tools and resources mentioned in this guide. The list is wholly incomplete and serves only as a representative sample. Inclusion does not equal endorsement.

DECISION MODELS

International Futures A dynamical system-of-systems model used to explore long-term global trends and support strategic planning

Liberal Democracy Index A score for the strength of democratic institutions aggregated across 61 dimensions (suffrage rights, equality before the law, freedom of expression, etc.)

Self-Reliance Metrics Seventeen capacity and commitment indexes that comprise USAID's Journey to Self-Reliance Country Roadmaps, one of which is the Liberal Democracy Index

Famine Early Warning Systems Network, or FEWS NET, An integrated system of models, methods, and analysts to help decision makers plan for food emergencies

Bayesian networks Probabilistic models with applications to adaptive management used to determine the likelihood of factors contributing to observed outcome

Monte Carlo method A technique for implementing a model by randomly trying thousands or millions of possible scenarios to statistically determine the likelihood of an outcome

DATA VISUALIZATION TOOLS

Tableau A data visualization and dashboard application focused on business analytics

Google Data Studio A web-based platform for interactive dashboard and data visualization

CARTO A cloud-computing platform that provides geographic and spatial data science tools

Flourish A data visualization and storytelling platform targeted towards digital journalism

D3.js A JavaScript library for dynamic and interactive data visualization

Keshif A web-based data exploration platform designed for use with minimal training

COLLABORATION PLATFORMS

Github A software repository commonly used to host open-source projects

Jupyter An open-source environment consisting of input/output cells that can contain code (e.g., Julia, Python, or R), text, and data visualizations

Colaboratory Also known as Google Colab, a collaborative, cloud-computing Jupyter notebook environment that supports Python

CoCalc A collaborative, cloud-computing platform that supports computational math and Jupyter notebooks

District Health Information System 2 (DHIS2) Open-source software used as the national health information system in many low and middle-income countries

PUBLICATIONS

Principles for Digital Development Nine guidelines to help development practitioners integrate best practices into technology-enabled programs

Reflecting the Past, Shaping the Future: Making AI Work for International Development A guide for understanding and accessing the appropriateness of machine learning and artificial intelligence in USAID programming

Considerations for Using Data Responsibly at USAID A guide that provides a framework for balance tensions between privacy and security, transparency and accountability, and data use

ENDNOTES

[1]Katharine Kripke et al., "Voluntary Medical Male Circumcision for HIV Prevention in Swaziland: Modeling the Impact of Age Targeting," *PLoS One* 11, no. 7 (2016): e0156776

[2]Victor Bulmer-Thomas, *The New Economic Model in Latin America and Its Impact on Income Distribution and Poverty* (Springer, 1996)

[3]Famine Early Warning Systems Network (FEWS NET) Website: http://fews.net/

[4]John L Newman, "A stochastic dynamic model of fertility," *Research in Population Economics* 6 (1988): 41–68

[5]Engineering National Academies of Sciences, Medicine, et al., *Quantum Computing: Progress and Prospects* (National Academies Press, 2019)

[6]Threshold Program Agreement between United States and the Republic of Kosovo: https://www.mcc.gov/resources/doc/tpaa-kosovo, September 2017

[7]USAID Learning Lab, "How-To Note: Developing a Project Logic Model (and Its Associated Theory of Change)"

[8]USAID, *ADS Chapter 201: Program Cycle Operational Policy*

[9]The mathematical field that studies CLDs and similar structures is called graph theory.

[10]USAID Learning Lab, "How-To Note: Developing a Project Logic Model (and Its Associated Theory of Change)"

[11]USAID Learning Lab, *Technical Note: The Logical Framework*, December 2012

[12]Wesley Langat, *Finding Fakes: Mobile Phones Help Detect Counterfeit Seeds in Kenya*, November 2018

[13]USAID, *ADS Chapter 201: Program Cycle Operational Policy*

[14]USAID Learning Lab, "How-To Note: Developing a Project Logic Model (and Its Associated Theory of Change)"

[15]International Futures Wiki: https://pardee.du.edu/wiki

[16]Frederick S. Pardee Center for International Futures, *International Futures Model*

[17]*USAID Self-Reliance Metrics FY 2019 Methodology Guide*

[18]Michael Coppedge et al., *V-Dem Methodology v8*, 2018

[19]Amy Paul, Craig Jolley, and Aubra Anthony, *Reflecting the Past, Shaping the Future: Making AI Work for International Development* (USAID, 2018)

[20]Joint Chiefs of Staff, *5-0 Doctrine for Joint Operations*, 2017

[21]Gerardo L Munck and Jay Verkuilen, "Conceptualizing and Measuring Democracy: Evaluating Alternative Indices," *Comparative political studies* 35, no. 1 (2002): 5–34

[22]Coppedge et al., *V-Dem Methodology v8*

[23]LDI $= \frac{1}{4}(P^x + Q) + \frac{1}{2}P^xQ$ where P is the polyarchy component, Q is the liberal principle component, and $x = \log_2 3$.

[24]Maria Hornbek, *This Machine Read 3.5 Million Books Then Told Us What It Thought About Men And Women*, September 2019

[25]Google Ngram Viewer: beautiful *_NOUN

[26]Google Ngram Viewer: man/woman

[27]The pointwise mutual information of two events x and y is $\text{pmi}(x, y) = \log p(x, y)/p(x)p(y)$ where $p(x)$ is frequency of occurrence.

[28]Google Ngram Viewer: financial wizard/_ADJ_ wizard

[29]Sarah Zhang, *The Pitfalls of Using Google Ngram to Study Language*, June 2017

[30]Google Ngram Viewer: infected mothers,infected women

[31]Heinz Von Foerster, Patricia M Mora, and Lawrence W Amiot, "Doomsday: Friday, 13 November, AD 2026," *Science* 132, no. 3436 (1960): 1291–1295

[32]The authors of the article were electrical engineers. That electrical engineers are developing models of human population growth underscores the need of having domain experts question the models.

[33]Xiaolin Wu and Xi Zhang, "Automated Inference on Criminality using Face Images," *arXiv preprint:1611.04135*, 2016,

[34]Carl T. Bergston and Jevin West, "Calling Bullshit Case Study on Criminal Machine Learning," *University of Washington Information School INFO 270*

[35]A tabletop exercise is a discussion-based session that simulates a situation to help identify possible gaps or weaknesses in a plan. Tabletop exercises originated in emergency response preparation.

[36]Barry B Hughes, "Assessing the Credibility of Forecasts Using International Futures (IFs): Verification and Validation," Unpublished IFs working paper on the IFs website, 2006,

[37]Adam Przeworski and Glaucio AD Soares, "Theories in Search of a Curve: a Contextual Interpretation of Left Vote," *American Political Science Review* 65, no. 1 (1971): 51–68

[38]Aart Kraay, *Methodology for a World Bank Human Capital*

Index (The World Bank, 2018)

[39]Robert Lee Hotz, "Mars Probe Lost Due to Simple Math Error," *Los Angeles Times*, October 1999,

[40]International Futures Wiki: https://pardee.du.edu/wiki

[41]Michael Coppedge et al., *V-Dem Codebook v8*, 2018

[42]Alberto Cairo, *Visual Trumpery: How to Fight against Fake Data and Visualizations—from the Left and from the Right*, January 2018

[43]Carl T. Bergston and Jevin West, *Calling Bullshit*

[44]Hornbek, *This Machine Read 3.5 Million Books Then Told Us What It Thought About Men And Women*

[45]Alexander Miserlis Hoyle et al., "Unsupervised Discovery of Gendered Language through Latent-Variable Modeling," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Florence, Italy: Association for Computational Linguistics, July 2019), 1706–1716

[46]Alex Reinhart, "Huff and Puff," *Significance* 11, no. 4 (2014)

[47]Ronald L Wasserstein, Nicole A Lazar, et al., "The ASA's Statement on *p*-Values: Context, Process, and Purpose," *The American Statistician* 70, no. 2 (2016): 129–133

[48]John PA Ioannidis, "Why Most Published Research Findings are False," *PLoS medicine* 2, no. 8 (2005): e124

[49]Monya Baker, "1,500 Scientists Lift the Lid on Reproducibility," *Nature* 533, no. 7604 (2016): 452

[50]A binomial distribution is perhaps not the best model to use for voting, because there are many factors that may prevent identical and independent trials.

[51]Christie Aschwanden, "Science isn't Broken. It's Just a Hell of a Lot Harder than We Give It Credit for," *FiveThirtyEight*, August 2015,

[52]Christie Aschwanden, "Your Brain is Primed to Reach False Conclusions," *FiveThirtyEight*, February 2015,

[53]Alexander Simoes, *Observatory of Economic Complexity*

[54]Richard J Tobin and Walter I Knausenberger, "Dilemmas of Development: Burley Tobacco, the Environment and Economic Growth in Malawi," *Journal of Southern African Studies* 24, no. 2 (1998): 405–424

[55]Julia Smith, "It's Time for Malawi to Quit Tobacco," *The Conversation*, May 2018,

[56]Duc-Quang Nguyen and Simon Bradley, *A Graphic Look at Global Smoking Trends*, May 2018

[57]Sarah Boseley, "Child Labour Rampant in Tobacco Industry," *The Guardian*, June 2018,

[58]John Vidal, "Malawi's Forests Going up in Smoke as Tobacco Industry Takes Heavy Toll," *The Guardian*, July 2015,

[59]"Partnership for Alternative Crops in Malawi," Accessed: July 24, 2019

[60]Principles for Digital Development: https://digitalprinciples.org

[61]District Health Information System 2: https://github.com/dhis2

[62]Paul, Jolley, and Anthony, *Reflecting the Past, Shaping the Future: Making AI Work for International Development*

[63]Steven D Levitt, "From Good to great. . . to Below Average," *Freakonomics*., 2008,

[64]Richard Heeks, "A Structural Model and Manifesto for Data Justice for International Development," *Development Informatics Working Paper Series*, no. 69 (2017)

[65]Anna Lührmann et al., "Democracy for All?," 2018,

[66]A Native American proverb

[67]The World Health Organization's Air Pollution Website: https://www.who.int/airpollution/en/

[68]Marcel Coelho and Amaury Martiny, Sh**t! I Smoke Website: https://github.com/amaurymartiny/shoot-i-smoke

[69]ibid.

[70]DigData Kosovo: https://millenniumkosovo.org/digdata

[71]FEWS NET. 2018. *Developing Price Projections for Food Security Early Warning*. Washington, DC: FEWS NET.

[72]Bruce Y Lee and Sarah M Bartsch, "How to Determine if a Model is Right for Neglected Tropical Disease Decision Making," *PLoS neglected tropical diseases* 11, no. 4 (2017): e0005457

[73]Principles for Digital Development: https://digitalprinciples.org

[74]James Somers, "The Scientific Paper is Obsolete," *The Atlantic*, 2018,

[75]Paul Romer, *Jupyter, Mathematica, and the Future of the Research Paper.*

[76]Hornbek, *This Machine Read 3.5 Million Books Then Told Us What It Thought About Men And Women*

[77]Coelho and Martiny,

[78]David X Li, "On Default Correlation: A Copula Function Approach," *The Journal of Fixed Income* 9, no. 4 (2000): 43–54

[79]G Celeux et al., "Recipe for Disaster: The Formula That Killed Wall Street," *Journal of the American Statistical Association* 109, no. 507 (2014): 1325–1337

[80]Damiano Brigo, Andrea Pallavicini, and Roberto Torresetti, "Credit Models and the Crisis, or: How I Learned to Stop Worrying and Love the CDOs," *arXiv preprint arXiv:0912.5427*, 2009,

[81]Cathy O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (Broadway Books, 2016)

[82]USAID Development Informatics, *Considerations for Using Data Responsibly at USAID* (2019)