



DATA SCIENCE: A GUIDE FOR SOCIETY

The Data Science Revolution: Why it's changing every aspect of our lives

AI and satellites could
predict volcanic eruptions

Machine-learning
techniques used to
accurately predict
battery life

Data science to support cyclone relief efforts

University stakes claim
on first AI-powered
travel guide

AI can evaluate
treatment
response for brain
tumours

The data science revolution is transforming aviation

AI systems explain cause of religious violence

**Big data predicts terrorist
attacks with more than
90% accuracy**

Can a computer write a story?
Machine learning goes to work

SHOULD WE HAVE LEFT THE BREXIT DECISION TO AI?

AI and CSI: How big
data is reshaping the
field of forensics

Could data science turn the tide
in the fight against cybercrime?

**DATA ALGORITHM PREDICTS
PREMIER LEAGUE FINISH
Where Liverpool, Manchester
United and Arsenal Finish**

How AI and satellites could
help predict volcanic eruptions

Facebook is using AI to map population density around the world

You want AI to
predict McDonald's
sales? Machine learning
can help

How machine learning is key to solving Alzheimer's

IBM hopes machine learning
is the key to solving Alzheimer's

AI can evaluate treatment
response for brain tumours

Model can more naturally
depression in conversational
AI

Machine learning and AI have massive potential

THIS AI CALCULATES

Hospital develops
AI to predict
patient
me

Meanings

3

The Stages of Data Science

4

..... The 3 questions to ask:

Q1. Where does it come from?

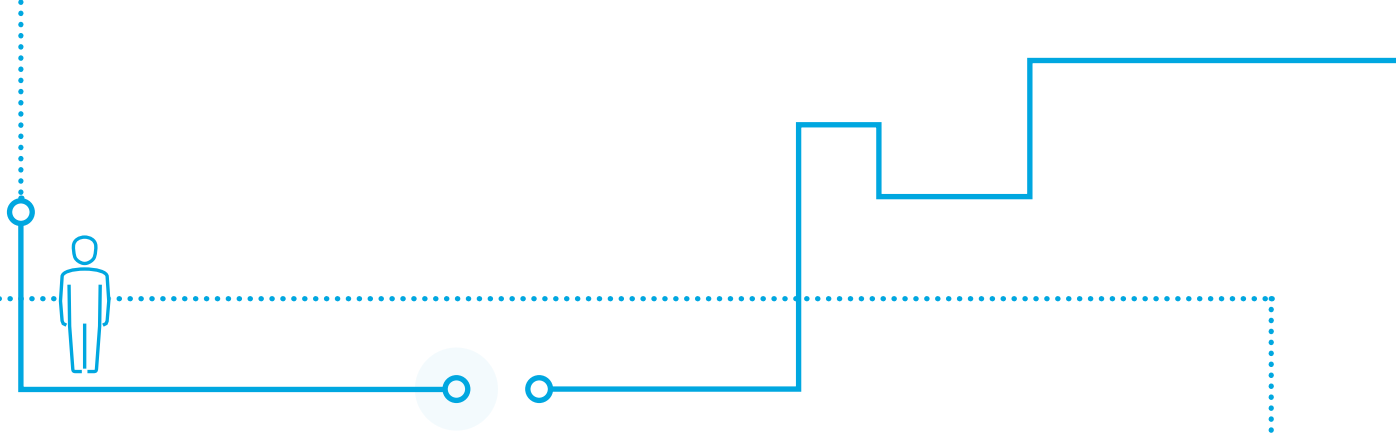
5-9

Q2. What assumptions are being made?

10-13

Q3. Can it bear the weight being put on it?

14-18



According to the headlines, we're in the middle of a "data revolution" — allowing us to make predictions on anything from the football results to who is likely to commit a crime. Governments have "big data strategies" to address ever growing expectations about how data could help answer some of the toughest questions in research and in policy. The public is promised benefits from fairer credit assessments to better predictions of drug side effects. Data science is clearly a powerful tool. If used properly, people will be able to make better decisions far more efficiently.

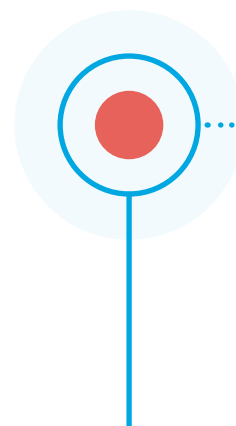
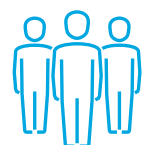
But what if our journalists, politicians and other decision makers don't know how to ask the right questions to check the quality of these claims? What if they're left wondering what terms like "big data" or "AI" even mean? Or they're passing on flawed information or making bad decisions, because they don't know how to scrutinize data science analysis? Equally, we can't always rely on researchers or data scientists to know exactly how their work actually applies to the real world.

The public's ability to question the quality of evidence can make a huge difference. We know this from debates about patient information, climate science, misleading product claims and military interventions. Society would reject a public health program rolling out on the basis of a study with a 10-person sample size. So we must

not let data science become a "black box", where these questions about evidence quality drop away.

A compelling reason to act now is the increasing weight being placed on the results of data science, such as making custodial decisions about offenders and allocating health and social resources. The quality of this evidence really matters.

This guide isn't about becoming a data science expert. It explains the language to help talk about it and highlights the key questions to ask those people using data science as evidence in decision making. If we ask more of these questions and turn the discussion to the quality of the data analysis, we can reject hype, recognize what's useful and ensure data-led decisions are transparent and accountable.



MEANINGS



The term “data science” has only been widely used for the past decade. It has spread fast across continents and sectors, earning titles like “sexiest job of the 21st century”. With such recent beginnings and fast advances, it’s no wonder that data scientists themselves debate some of these definitions. So first off, this isn’t a textbook glossary on academic data science, more a guide to help you enter into conversations:

Data science

A field that tries to extract meaningful information from data, to help make real-life decisions. Usually this involves collecting data, organizing it, analyzing it (often using machine-learning algorithms) and then presenting findings for a decision.

Big data

This is just a type of dataset that needs special methods of analysis to cope with the fact that it is very large or diverse in format. For example, Instagram data from everyone in the UK would be both: billions of data entries in a variety of formats (images, videos, numbers, text and so on), so we call this big data. Data science isn’t just for big data, it can be used on standard data too.

Machine learning

The writing of computer programs, called algorithms, which tell a computer how to learn from data by looking for patterns. Using these patterns, predictions can be made on new data. We interact with this everyday – such as when Amazon or Netflix suggest something to watch based on people’s past habits and then adjust the algorithm if lots of us don’t respond. Machine learning is the data science element of artificial intelligence (AI). AI refers to the many different ways of getting computers to simulate intelligent behavior, such as playing chess or facial recognition.

Variable

A factor or characteristic that might be relevant to answering a question. These could be numbers like age, temperature, test score or number of films downloaded. Or they might fall into categories, such as pass or fail, religion or eye color.

Algorithm

A set of instructions – mathematical instructions – to find or calculate something. Algorithms have many roles in data science, including figuring out the relationships between things (variables).

Models

A representation of how conclusions can be made from new data. For example, new data on a person’s age and driving record could be used in a model for an insurer to make conclusions about that person’s driving risk. These are usually “supervised learning” models, where algorithms learn from data that has labels, in order to predict how to label the new data. There are two different types of supervised learning models:

1. Classification models predict the categories the data belongs to, such as taking emails and predicting whether they should have the label of “spam” or “not spam”.
2. Regression models make numerical predictions, such as estimating how many people will die of flu by looking at how it has spread in recent months.

These are the methods this guide is going to focus on as they are the most commonly used methods in practical decision-making. But many of the points we raise also apply to what is known as “unsupervised learning techniques”. This is essentially when we don’t know the categories or groups within the data so the algorithms have to figure it out.

THE STAGES OF DATA SCIENCE

Let's take a question and see how data science is being used to answer it.

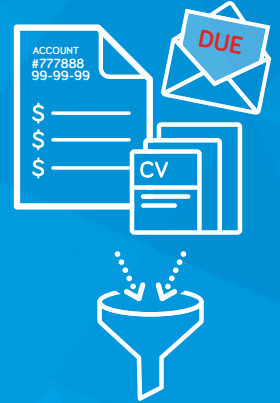
Some local governments might seek to use data science to see if people at risk of becoming homeless could be identified and offered help before things get to that point. To do this, they ask a data scientist to create a way of categorizing citizens into two categories: at risk, or not at risk.



Data collection

Data scientists gather existing data from a variety of sources that they think are relevant to the question they are asking.

They gather data on citizens from last year, variables such as income, age, debt level, employment history or use of government assistance. This data would be collected for all citizens being studied, along with whether they became homeless in that year or not.



Data preparation and cleaning

This involves getting the data ready for analysis and arranged into the right format.

Errors in the format of address would be corrected, missing data dealt with where possible, and all the data converted into one type of format.

Street Address	Zip Code
94 True Avenue	95814
1 Real Street 85001	
6 Main Road	80202
113 Other Side	22313



Using algorithms on the data to create a model

Algorithms are used to estimate the relationships between the variables by learning from the data. This information is used to create a model.

An algorithm would work out how the different variables affected someone's risk of becoming homeless. For example, if lots of homeless citizens had an income below \$25,000 before becoming homeless, the algorithm would flag this as a risk factor. All this information is used to create a model, which is an equation that links these variables together to predict if someone is at risk or not.

$$\begin{aligned} &\text{Income} < \$15,000 \\ &+ \\ &> 6 \text{ months unemployment} \\ &+ \\ &> 6 \text{ months defaulted debt} \\ &= \\ &\text{At risk of homelessness} \end{aligned}$$



Testing the model

Some of the data collected at the start is held back and not used to create the model. Instead, this "test dataset" is used to test how good the model is at predicting things.

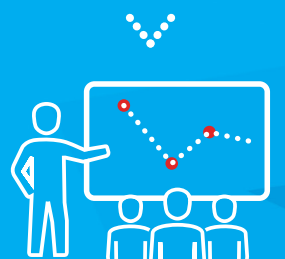
Using this test data, we can see how good the model is at predicting which citizens were at risk of becoming homeless.



Deploying the model

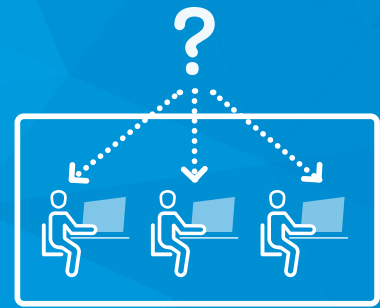
Data scientists will communicate the findings and performance of the model to decision makers, often using visualization techniques, such as graphs or charts. The decision makers may then choose to use the model in real life.

A local official may decide, after hearing from the data scientist, that the model was good enough at categorizing people to trial an intervention program to offer help to citizens flagged as "at risk."



WHERE DOES IT COME FROM?

Q1



The first thing we need to know is where the data has come from and whether it has been collected for the current purpose.

The essential aspects of this conversation are:

- A The actual question the data is answering.**
- B Where the data came from,** and whether it was obtained under controlled conditions to answer that question (experimental data); or from looking at data that was gathered for another purpose, such as Facebook data (observational).
- C If the data is representative of the population or thing we are interested in.** If it comes from observation, representativeness is harder to control for.
- D The type of conclusion being drawn from the data.** If someone claims their analysis shows that A causes B, they should usually have experimental evidence to back it up. They might also have a good theory for *how* A causes B. Analysis from observational data is most likely to tell us how variables “correlate” or fluctuate together. This is not the same as “causation.” People used to think bad smells caused diseases because where the smells were worst, disease was too. They correlated.
- E Whether the relationships seen in the data are even real.** When data scientists look for correlations in large datasets, they are bound to find some that are just due to random chance – lots of ear infections among non-drinkers for example.

In more detail...

- A If decision makers are using the data to answer a question, they should explain why they think it does.**

We could have the most rigorously produced piece of data science but if it wasn't designed to answer the question it's now being used for, it may not be useful in decision making.

The UK longitudinal education outcomes project (LEO) links long-term data on graduates' earnings to their student record. Recently, policymakers have tried to use the results of this analysis to make funding decisions for universities. But LEO offers no information on occupation, industry or other employer characteristics. This means, for example, that a university that provides nurse and teacher training will inevitably appear to perform less well than one focused on financial services and corporate law. They may then be given less funding based on this data science research.

B**Where the data came from: observation or experiment?**

Let's say we want to find out whether Hormone Replacement Therapy (HRT), used to alleviate symptoms of menopause, is linked to women having heart problems.

This could be studied experimentally...

Recruit a large enough group of post-menopausal women to be able to calculate their representativeness of the population and randomly assign them to a program of HRT treatment or a program of treatment with a placebo. Study the subsequent heart health of both groups.

Such experiments involve the researchers varying one factor (the medication) to see whether this produces an effect. This is usually necessary to determine cause-and-effect relationships.

If results come from an experiment, the collection of the data should have been specifically designed to answer the research question. The markers of quality and reliability to look for are:

- Large enough sample sizes (researchers should be able to explain how well they can extrapolate from the sample size).
- A "control group" for comparison against – with the same characteristics as the group being studied, except for the variable being tested.
- Random allocation of subjects to these groups where possible.
- Reasonable generalization to real life, with discussion about the limits of the experiment for this.
- Clear estimates of the scope for error in the results.

Or studied observationally...

Use existing health records of women who have and haven't taken HRT and analyze their rates of heart problems.

Observational data analysis involves looking at data that already exists and seeing whether any relationships can be found. It allows researchers to observe many more variables than in experiments. For example, researchers can study people's lifestyle using datasets from social media activity, health app data and phone GPS. There are advantages to this, such as being able to discover patterns that hadn't been thought of or couldn't have been measured in an experiment. However, there are also many possible limitations with the data samples. They may be self-selecting (people using health apps may be more healthy in general) or missing relevant information such as whether people had a pre-existing heart problem.

Markers of quality for observational data analysis include:

- A clearly stated data source that is reliable and impartial where possible.
- Open discussion about any limitations of the sample and possible bias.
- Justification given to why variables were included or excluded.
- Explanation of how the conclusions allow for missing data and other errors.

C**The data should be representative of the population being studied.**

Whatever data type – observational, experimental, big or small – we must ask whether the data that is used to create the model is representative of the population for which claims are being made about.

From observational data, researchers noticed that women taking HRT appeared to have fewer heart problems. However, after HRT had been prescribed for decades, a large randomized controlled experiment* found that in fact, HRT was actually slightly worse for women's hearts. The observational data had not accounted for the fact that wealthier women were more likely to be prescribed HRT. It was being wealthier that lowered their incidence of heart problems, not the HRT. The randomized trial overcame this over-representation of wealthier women by comparing like with like.

There can be whole groups whose characteristics make them essentially invisible in datasets. This is common when using mobile phone or social media data:

City officials in Boston MA in the USA tried to fix their pothole problem by allocating resources around the city to fix them based on the results of a mobile phone app. The app passively recorded information on pothole locations when owners drove over the bumps in the road. However, older residents and low income groups were less likely to own a smartphone. This meant that areas where older or lower in-come people lived were less likely to get resources to fix their potholes.

One way this problem of selective data can be overcome is by combining the results with other methods, such as focus groups or interviews. However, ideally a data collection method should be interrogated to ensure it does not omit a significant group.

D**The type of relationship should be explained: correlation vs causation.**

Very often in data science, observation is used to find variables that are “correlated”. This is where two variables have a mutual relationship, for example if one goes up so does the other.

“Causation” – where one thing actually causes another – is harder to discover. It needs experimental evidence and if possible a theory of how it causes it. If experiments aren't possible for ethical, financial or practical reasons, there are some statistical methods that can be used to try to show causality and data scientists should be able to explain this.

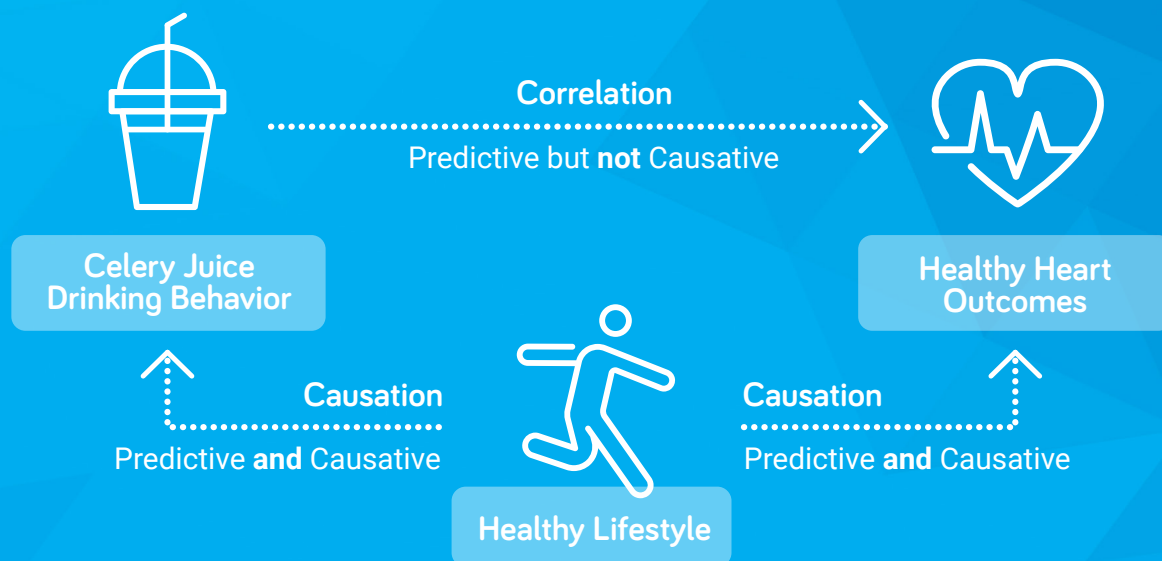
Even where two variables are strongly correlated, and may be used as predictors in data science models, it does not mean one thing causes the other. Ice cream sales have been found to correlate with increased murder rates**. This may just be chance, but it seems persistent so it could actually be that both are linked to a hidden third variable, such as the hot weather.

* Rossouw, JE. et al., (2002) Risks and Benefits of Estrogen Plus Progestin in Healthy Postmenopausal Women Principal Results From the Women's Health Initiative Randomized Controlled Trial. Available at: <https://www.ncbi.nlm.nih.gov/pubmed/12117397>

** Scheidegger, K. (2012) Rebutting the Myths About Race and the Death Penalty. Available at: <https://ssrn.com/abstract=2178223>

As we go to press, celery-juicing mania has taken off across social media sites where “influencers” are promoting the benefits of drinking celery juice, such as for improved heart health. Often claims like this come from analyzing observational data and discovering that juicing celery is a predictor of better health (ie they are correlated).

Aha! So let’s all start juicing celery to improve our hearts? Well, the data doesn’t actually tell us anything about celery being **the cause**. More likely, celery juicing is an indicator of other variables — people who follow these nutrition trends are more likely to have an overall healthier lifestyle, diet and higher income (the biggest predictors of heart health).



E

The researchers should ensure the relationships they’ve found are real.

Traditionally, researchers have set out with a specific question or theory in mind and collected and analyzed the variables relevant to this, leading them to a result that either disproves or strengthens the theory. With the rise of big data though, many researchers celebrated the fact that huge datasets can be explored to *reveal* connections between variables wherever they exist, without a specific question.

However, large datasets always contain false relationships — variables that look connected but are just due to chance. Searching large datasets for any possible relationships is called “data dredging.” To illustrate this potential for finding false leads, US researchers have shown that Americans’ cheese consumption in the 2000s correlated very closely to the number of people dying by becoming tangled in their bedsheets*.

Even though researchers know data dredging is riddled with false leads, it sometimes happens because of the pressure on researchers for publishable results, because they get drawn in by how strong a false relationship seems, or because people without statistical training are using datasets to promote their ideas or products.

*Vigen, T. Spurious correlations. Available at: <http://www.tylervigen.com/spurious-correlations>

While this issue is not new, it is far more common and problematic when handling big data due to the vast numbers of variables that can be compared.

Google Flu Trends is often held up as an exemplary use of big data. It uses search engine data to predict flu outbreaks. Data scientists found 45 search terms which people in the areas in which the flu had spread seemed to be Googling more often. These 45 search terms were then declared predictors of flu outbreak.

However, in 2013, the model's prediction vastly overestimated the cases of flu by almost double. So why wasn't this working? One reason could be that the relationships it found between search terms and flu outbreak were false, having been dredged from a huge dataset of over 50 million search terms*.

It was impossible to tell because Google refused to release the 45 search terms.

The chance of the relationships being false would be lower if the data scientists:

- Started out with a specific research question.
- Only studied variables which they can justify as relevant.
- Use statistics to correct for the fact they've searched for so many relationships – this is a very important indicator of quality and reliability. Put simply, if you search 1000 combinations of variables and find two that strongly correlate, you must compare this to the likelihood that such a pattern in the data exists by chance among so many variables.
- Make sure the model has been rigorously tested, which we'll discuss in section 3, but in short, if the model makes accurate predictions on a separate test dataset, then the model has good "predictive power" and it is more likely that the relationships used in the model are true.

* Butler, D. (2013) When Google got flu wrong. Available at: <https://www.nature.com/news/when-google-got-flu-wrong-1.12413>

Q1

WHAT ASSUMPTIONS ARE BEING MADE?

Q2



Data science tackles tricky questions about things like human behavior, effects of interventions or forecasting rare events. It has enabled researchers to advance knowledge in areas that were previously too complex to handle. To reduce these complex issues down to numbers and equations requires lots of different assumptions about what the data represents, no matter how good the model.

How true these assumptions are can make a massive impact on how true the results of the analysis turn out to be. The important assumptions to discuss are:

- A** **That the right thing has been measured**, especially if a “proxy” or “surrogate” measure has been used for something that is difficult to measure, such as using SAT scores as an estimation for ability, or body mass index (BMI) for body fat. We need to know how good an estimate these proxies really are.
- B** **That variables missing from the model are irrelevant.** Are there any that could actually have an impact?
- C** **That the results can be generalized to other times, places or groups.** Have the data scientists considered things the model would not have picked up in historical data (such as a recent political change) or whether a model trained on one country’s data can be applied in another?
- D** **That using algorithms eliminates human prejudice from decision making.** The output of data analysis can only be as good as the input data and the ability of data scientists to correct for prejudice within it.

[In more detail...](#)

A Measuring the right things.

The rings on a tree can tell us not just about its age, but about the weather conditions of the past. Researchers measure trees rings because they (obviously) can’t directly measure past weather. The ring widths are a good proxy.

Using proxies as estimations of things that are hard to measure is really useful, but it must be done with caution. Markers of quality include:

- An explanation of how good an estimation the proxy is for the thing we *really* care about.
- A direct relationship between the proxy and variable we care about, without too many interfering factors.

With tree rings, we know trees grow faster in warmer temperatures and slower in the cold, so the link between their width and temperature is well established. There aren't many other factors that can change the tree rings, and any there are (the age of the tree) are predictable and can be factored in to the analysis.

On the other hand, while it is well-observed that life expectancy in a country is correlated to the number of televisions per 100 people, it would be unwise to use televisions as a proxy measure. There are so many interfering factors that could affect one thing and not the other, such as a sudden surge in the cost of televisions.

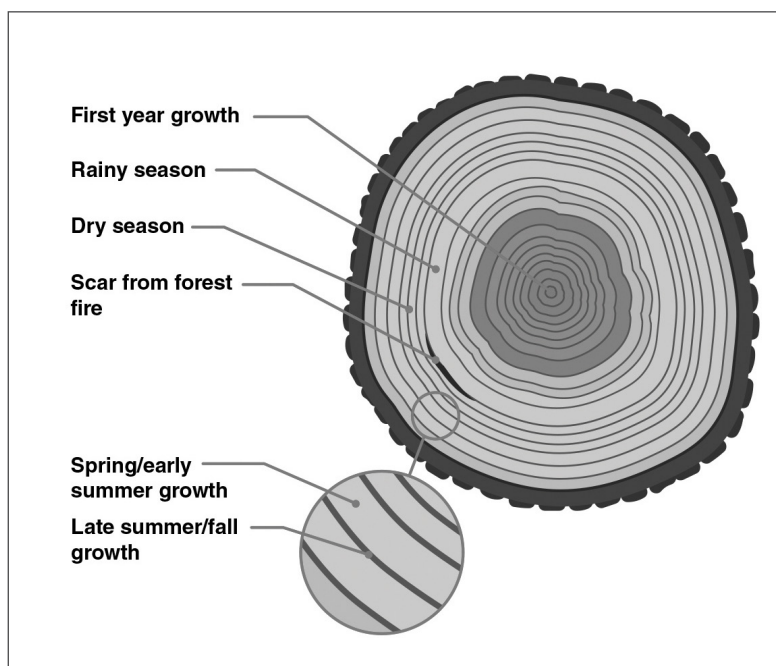


Image source: Stoller-Conrad, J. (2017) Tree rings provide snapshots of Earth's past climate. Available at: <https://climate.nasa.gov/news/2540/tree-rings-provide-snapshots-of-earths-past-climate/>

B

Missing variables that could affect the results.

With today's computing power and data science methods, models can be highly sophisticated, but they still require a decision about which relevant variables to construct them from. This is harder in some areas of research. For example, economic models have historically not included psychological factors as these factors are not easy to measure in a dataset.

Economists failed to predict the 2008 financial crisis and subsequent crash in the US housing market. Their models assumed that everyone took financial decisions in the same way. When it became clear that finance executives were trading heavily in mortgage markets that were riskier than they had realized, panic set in.

High-quality data science requires discussion about the possibility that an important variable is missing.

C**Generalizing to other times, places or groups.**

Predictive models, by necessity, are created using the data from what has happened in the past. This means we have to assume the future is going to be similar. Of course in some areas it may not be.

You should ask data scientists whether they've considered changes that could affect the validity of the model and how the model takes account of this. This is particularly important when trying to predict something far away from the range the training data is in. For example, if using 2018 data on the impact of a new public health initiative to predict the impact in 2030, when for example, age distribution in the population will have changed.

Similarly, applying the model to a different location may introduce many variables not present when the model was created: models of pollution effects in the Atlantic may not translate to the Indian Ocean for example. The same goes for generalizing to other groups, such as models of human behavior on people from different cultures.

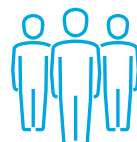
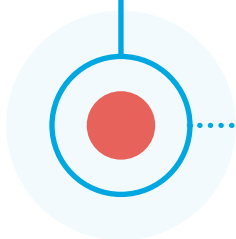
D**Reducing the impact of prejudice.**

Often people want to use data models to make decisions to remove the human biases that prejudice a decision-making process. However, machine-learning techniques learn by analyzing the world the way it is, not as it ought to be. So using historical data creates the possibility of encoding those prejudices. Recent examples of this have hit headlines, initiating debate from high-profile figures like US politician Alexandria Ocasio-Cortez.

***“THEY’RE JUST AUTOMATED ASSUMPTIONS.
AND IF YOU DON’T FIX THE BIAS, THEN YOU
ARE JUST AUTOMATING THE BIAS.”***

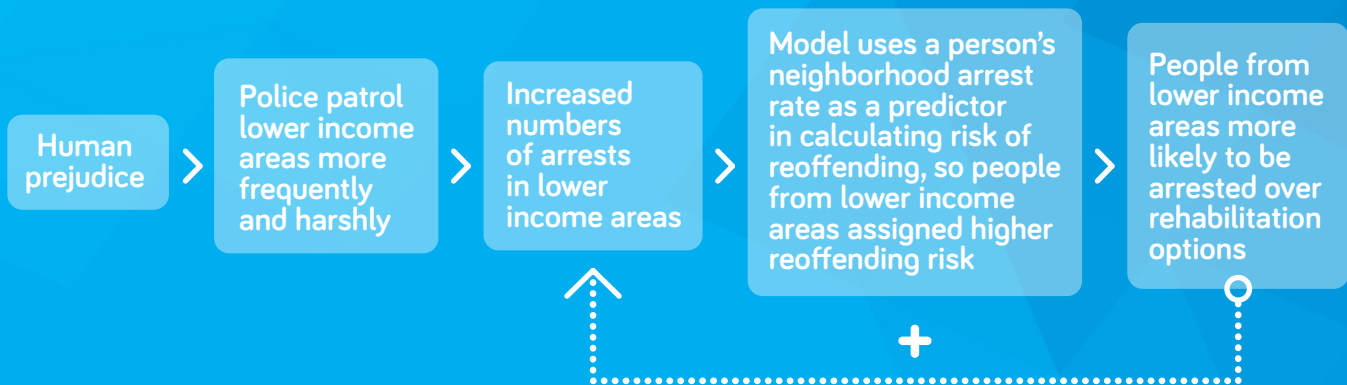
— ALEXANDRIA OCASIO-CORTEZ

Data scientists at Amazon created a model to automate their recruitment process by searching application forms and predicting the best candidates for roles. However, it was discovered that the model was heavily favoring men over women when recruiting for technology-based roles due to the data reflecting the previous male dominance of the field. Looking at this past data, the model “learned” that men outperformed women in previous selection decisions and so was disregarding resumes that indicated they were from a woman.

**Q2**

Not only can models be influenced by human prejudices, they can compound the effects of it:

In the UK, Durham police force's Harm Assessment Risk Tool (HART) was designed to categorize the risk of offenders reoffending, in order to decide who should be referred to a rehabilitation program instead of being dealt with in the courts. The model uses postcodes (zipcodes) as a predictor. This has been criticized because it is likely that there is some prejudice in the decision to arrest and charge someone from less-affluent areas, meaning HART would amplify this prejudice via a "runaway feedback loop."



Even when data is processed to be representative, this in itself can embed prior assumptions. For example, processing political polls to make them reflect the sex split of the population embeds the assumption that people vote differently dependent on their sex. There is no easy solution for these issues. Data scientists are discussing them all the time and in high-quality studies you can expect to find a description of how these issues have been taken into account.

Once Amazon data scientists realized what was going on they coded the model to ignore explicitly gendered words or whether the applicant came from a women-only college, but the researchers also had the problem of implicitly gendered words. These are words that are more likely to be used by men or by women, which can still give the model an indication of the gender of the applicant. For example, they found words like "executed" and "captured" were more likely to be used by men.

Markers of quality to watch out for include:

- Description of how variables that may encode past prejudices have been dealt with.
- The use of algorithms which help to detect and mitigate biases either in the training data or model.

CAN IT BEAR THE WEIGHT BEING PUT ON IT?

Q3



Decisions made on the basis of data science research can have far-reaching consequences, sometimes for years to come. Therefore you don't just ask whether the data and analysis are of good quality, but whether they're good *enough* for a particular use. What this means in data science is asking about the performance of the model that's being used and how robust it is.

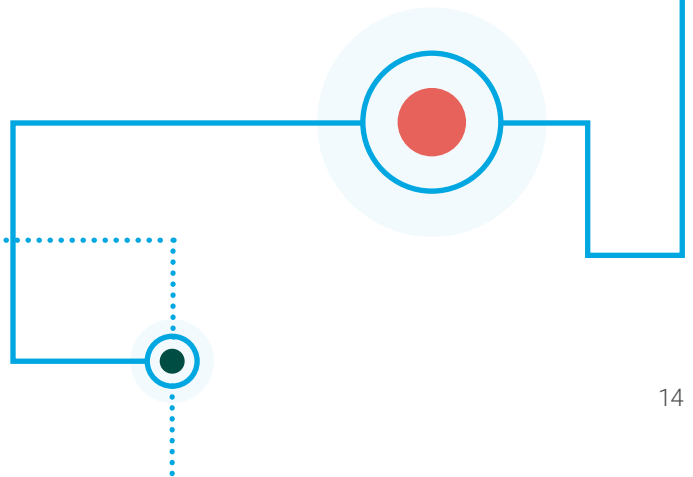
There are four elements to the conversation that you'd expect to have about this. These are:

- A** **How well the model performs.** There are many ways to measure this. For categorical models you would expect people promoting the model to describe measures of its accuracy, sensitivity or specificity. For models making numerical predictions, people should have calculated how much "error" the model contains.
- B** **Whether the model has been tested correctly, with a separate test dataset.** This is to make sure the model actually has the power to make predictions on new data and isn't just "overfitted" to the training data used to build it.
- C** **Whether it would make valuable real-world recommendations, and if it's really better than what we've already got.** This involves further testing of the model against a baseline, such as the current best method in place.
- D** **How precise these predictions are.** Can we really know it will be *exactly* 25°C two weeks on Thursday? Models often produce their own estimates of how precise their predictions are, such as giving "prediction intervals" for a numerical prediction (eg it will be 25°C, give or take 3°C).

In more detail...

A How well the model performs.

Data scientists should give you some performance measures for the model. These basically tell you how good the model is at predicting things.



For categorical models, the most common measure is accuracy – how often the model predicts the category correctly. We should never rely purely on accuracy as a measure though, as there are ways a model can get a high accuracy score, whilst being a terrible model:

“New airport terrorist screening system shown to be over 99% accurate!”

- 78 million travelers through airport per year.
- Model labels all travelers as not terrorists.
- If there were 10 terrorists flying through each year, then the model is only wrong 10 times out of 78 million.
- This means it is over 99% accurate, but doesn't identify a single terrorist, which makes it useless!

To make sure a model isn't cheating, you should look past advertized percentages and ask how well it is answering the most important question, and at what cost. For a model that categorizes data this will often involve looking at the trade off between:



When trying to identify terrorists in the airport, a false negative is much worse than a false positive: so the sensitivity measure is most important. A sensitive test may waste extra resources on security checks with passengers who aren't actually a threat, but is much less likely to miss individuals who do mean harm.

Different measures are used for numerical prediction models (such as RMSE*) which look at how much difference there is between model predictions and real values when the model is tested.

No matter what measures are used, data scientists should be able to comment on these measures and explain what they mean about the model's predictive ability.

*RMSE stands for root mean square error.

B**Testing properly with a separate test dataset.**

It's crucial that the performance measures are calculated on a completely separate test dataset, at the end of the analysis. This gives you an estimate of how well the model would perform on unseen new data.

Sometimes, models that work really well on the data used to make them are actually terrible at making predictions from new data, perhaps because the model has not been very successful at weeding out the relevant predictors from the irrelevant ones.

Let's take a non-data science example to illustrate this. This recruiter has created a "model" of what a good replacement for John would look like, but they haven't done a good job of picking out what the relevant variables about John are. Someone could match the description well, fitting points 2-6. But they would make a terrible accountant.

**"Oh no, John from accounts is leaving the company!
He was great, we'll have to find a good replacement!"**

Job ad — wanted:

- 5+ years experience in accountancy
- A masters degree in photography
- Be an avid cyclist
- Must be 5ft 8
- A tendency to tell bad "dad jokes"
- Takes their tea with two sugars

This is known as "overfitting," and it's a common problem in machine learning. It essentially means incorporating too much irrelevant noise from the training data into how the model works, meaning the model is not generalizable.

The more complex the model, the higher the chance of overfitting, so you would want more evidence of complex models' predictive power on new data.

Overfitting has been suggested as a contributor to the energy accident at the Fukushima nuclear plant following the Tōhoku earthquake*. The engineers at the plant made an overly complicated model to predict how often they should expect large earthquakes in the area. This model predicted these only once every 13,000 years — in other words, near impossible, so safety precautions for earthquakes of this size were not taken. If a simpler, less overfitted model had been used, it would have predicted an earthquake that large one once every 300 years — often enough that adequate safety measures would have likely been put in place.

*Stacey, B. (2015) Fukushima: The failure of predictive models. Available at: https://mpra.ub.uni-muenchen.de/69383/1/MPRA_paper_69383.pdf

C**Proving it's worth using for real world decisions.****Testing against the baseline**

A really important step in deciding if a model is worth using in decision-making is whether it performs any better than the existing method. This is known as "testing against the baseline." For example, if we were testing a model for a cancer screening test, this may mean comparing to an earlier, less advanced screening test or even comparing to no test.

External validation

Even better than just testing against a baseline using statistics would be if the model has also been tested out in the real world. This testing should be subject to usual ethical considerations, where it involves decisions that affect people's lives. This has three major advantages:

- Often this is done by a group of people independent to the model creators. This group is able to question the methods used by creators in more detail.
- It can highlight potential problems with implementing the model in a real world scenario: a model that processes image data to calculate traffic hotspots is not much good if half the cameras are never repaired.
- Real life testing can highlight any major biases that exist in the dataset used to create the model. These biases could very easily have also existed in the separate test dataset discussed in section 2C, so this external validation is a useful step.

D**How precise can the model really be?**

When models use data to predict a number, like profits or temperatures, they usually give a measure of how confident they can be. A prediction interval tells us the range within which the true answer is likely to be.

For example, if a bank uses a model to predict its profits for the following year, it may predict \$5.5bn with a prediction interval of \pm \$1bn - ie profits should be between \$4.5bn-\$6.5bn.

Often when results from data analysis like this are communicated, large prediction intervals are not reported despite their importance for how we interpret the results. Amid excitement about big data and advanced computer models, there is a tendency to overstate the precision of things like hurricane path predictions or political polls, even when the researchers themselves don't claim the predictions are sure things.



In the 2016 US presidential election, polling organizations were criticized for predicting that Hillary Clinton would win. However, it was largely the way the poll results were reported by the media, rather than the analysis itself, that was misleading. The website FiveThirtyEight, for example, predicted that Clinton would have a 71.4% chance of winning. This percentage was picked up and widely reported by other news outlets. However, what few of them reported was that FiveThirtyEight also illustrated the prediction intervals around the number of electoral votes Clinton and Trump were predicted to receive — which showed there was actually a lot of overlap:

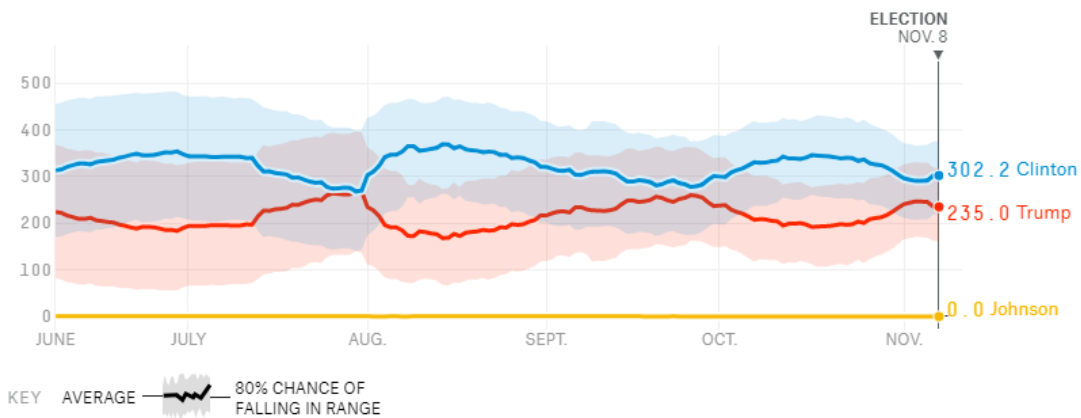


Image source: FiveThirtyEight (2016) Who will win the presidency?
Available at: <https://projects.fivethirtyeight.com/2016-election-forecast/>

ABOUT US

Sense about Science is an independent charity that works to ensure the public interest in sound science and evidence is recognized in public life and policy making. A small team working with thousands of supporters, from world leading researchers to community groups, we focus on socially and scientifically difficult issues where evidence is neglected, politicised or misleading.

This guide is about the biggest challenges of data science for policy and society. So we are delighted to have developed and produced it in partnership with **Elsevier**, one of the biggest providers of curated data content, and the **International Network of Government Science Advice**, the global network of those interested in the science-policy interface.

We owe a debt to Philip Dawid, Michiel Kolman and Yuko Harayama who challenged us, and to the many people — Voice of Young Science members, data scientists and members of the public — who have contributed to the workshops, arguments and questions in these pages. Responsibility and any error lies fully with Sense about Science, but thanks are owed to them.

This guide was compiled by Errin Riley.


Published in 2019 by



Sense about Science

2 Stephen Street
Bloomsbury
London W1T 1AN

Registered Charity No. 1146170
Company No: 6771027

This document is licensed under Creative Commons Attribution-Noncommercial-No Derivative Works 2.0 UK: England & Wales License. 

For an electronic version of this guide and further resources, please visit:

askforevidence.org