the shortcut × moz://a

DATA SCIENCE

TRAINING PROGRAM

Week 1: What is Data Science

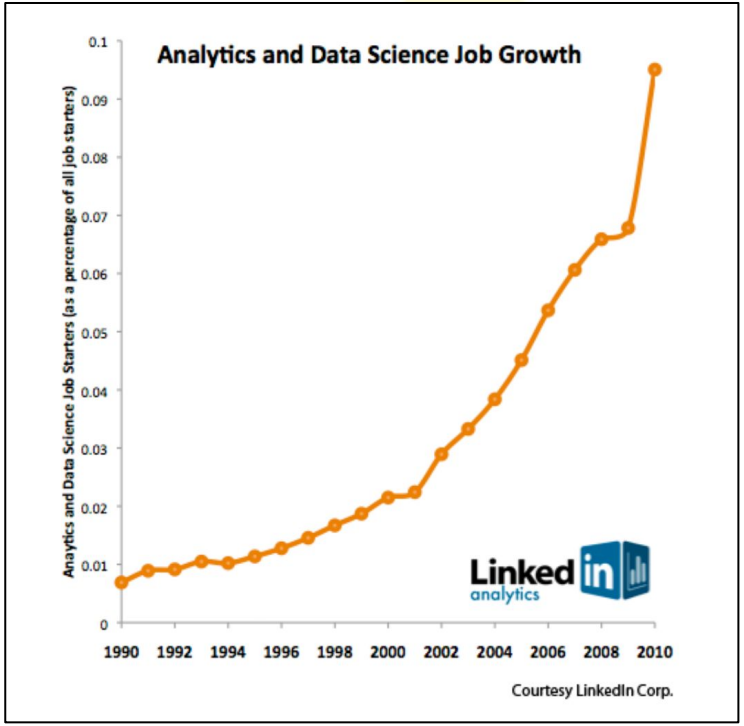
Jesse McCrosky, Mozilla

2019-10-7

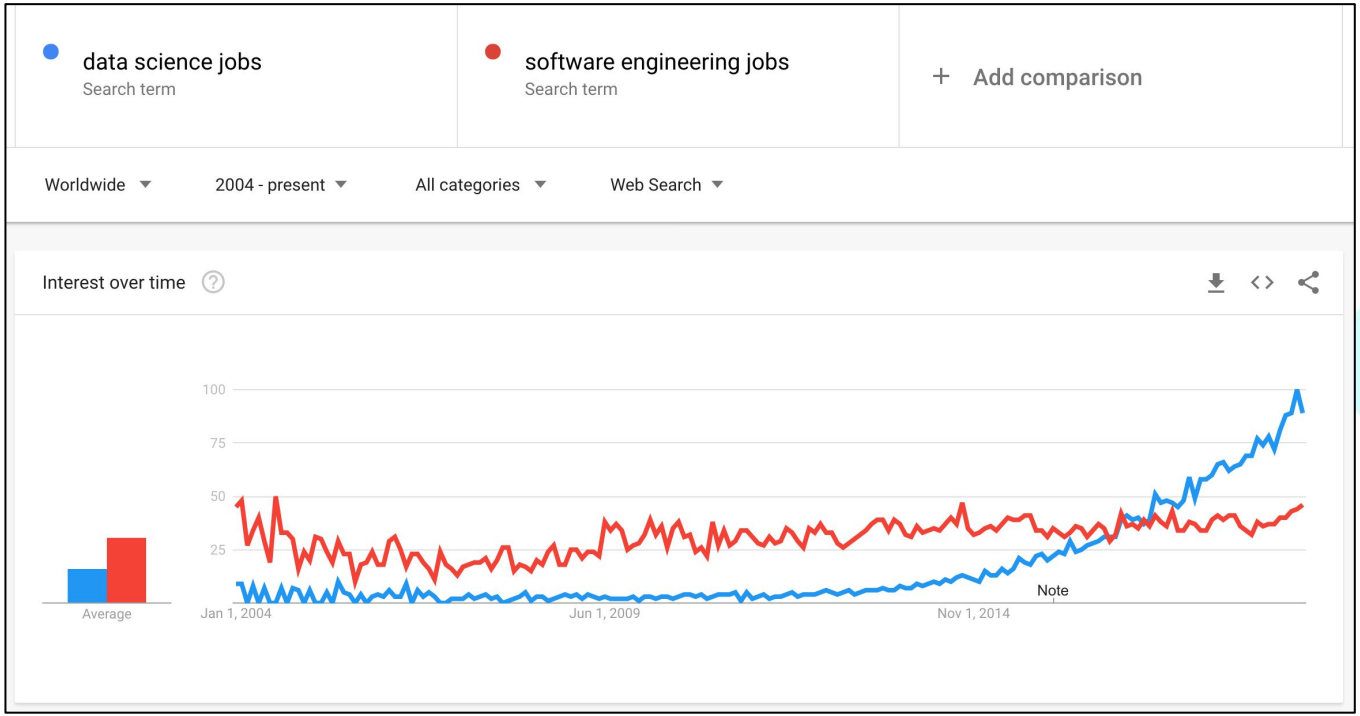
kodit.io selko SILO.AI

What is Data Science?

- Making Good Decisions Using Data
  - (or building products that make good decisions using data)
- How to meaningfully measure how well a product is doing
- How to decide whether a new feature or change should be launched or not
- Generating insights into what features customers want



Realizing the Benefits of Being Data-driven



“Data scientists are kind of like the new Renaissance folks, because data science is inherently multidisciplinary.” - John Foreman, MailChimp

History / Evolution of Data Science

- “Data Science” as a term was first used in 1960 as a substitute for “Computer Science”
- In a 1997 lecture, “Data Science” was used to describe the work of Statisticians
- In a 2001 article an article proposed “Data Science” as an independent field
- In 2012, Harvard Business Review published "Data Scientist: The Sexiest Job of the 21st Century"

From https://en.wikipedia.org/wiki/Data_science

“Junior Data Scientist”

- Mismatch in industry:
  - Many employers can’t find Data Scientists
  - Many job-seekers can’t find Data Science jobs
- Why?
  - Skills mismatch.

Data Science Background

Typical Background

- Most typical: Ph.D. in Statistics
- 88% have at least a Master's degree and 46% have PhDs (reference)
- Other common backgrounds are computer science, engineering, economics (especially econometrics)
- Also see other quantitative disciplines such as biostatistics, physics, etc.
- But some successful Data Scientists have no academic background

My Background

- B.Sc. and M.Math in Computing Science
- Yoga Teacher :)
- Most of Ph.D. in Community Health and Epidemiology
- Analyst with Statistics Canada
- Independent Statistical Consultant
- Data Scientist with Google
- Data Scientist with Mozilla

Broadly Speaking

- Data Science is an umbrella term covering many roles
- Not all Data Science work requires a Ph.D.
  - Many companies want to hire someone that can do it all
  - But “Junior Data Scientist” roles exist and should become more common
  - In this course we aim for an “MVP” - teach you enough to be useful in a data science role
- Some Data Science work does require a lot of education or experience
  - Ideally a Junior Data Scientist works in a company that already has a more experienced Data Scientist that can provide mentoring and support on the job
  - Junior Data Scientist can continue learning “on the job”

Types of Data Scientists and Related Roles

- Statistician
- Quantitative Analyst / Business Intelligence Analyst
- Data Engineer
- Machine Learning Engineer / Researcher
- Data Scientist

"Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician."
- John Wills, Slack

Statistician

More research oriented:

- Developing new statistical methods for complex problems
- Applying relatively new methods that require a deeper understanding of theory to properly apply and interpret
- Developing measurement and metrics

Less requirement for computing skills, but R is important as newly developed methods are usually implemented there first

But some roles, especially in government, are called “Statistician” when they’re really more of an “Analyst” role - extracting, manipulating, and presenting simple data

Quantitative Analyst / Business Intelligence Analyst

Sometimes synonymous with “Data Scientist”

Sometimes refers to a different role: extracting and summarizing data, simple manipulations, fitting basic models, preparing reports and visualizations

Business Intelligence may require more business strategy ability to develop strategies from data

Distinction between Analyst and Data Scientist:

- “The stereotype is than an analyst will answer your question quickly with a number, but a data scientist will interrogate you to uncover what you *really* want to know, and work out how to answer that question instead.” - Felix Lawrence, Data Scientist, Mozilla
- “The job of the data scientist is to ask the right questions. If I ask a question like ‘how many clicks did this link get?’ which is something we look at all the time, that’s not a data science question. It’s an analytics question. If I ask a question like, ‘based on the previous history of links on this publisher’s site, can I predict how many people from France will read this in the next three hours?’ that’s more of a data science question.” - Hilary Mason, Founder, Fast Forward Labs

Data Engineer

Should understand some statistics

- Understand data quality issues
- Foresee data needs for analysis

Focuses on engineering necessary to support the work of Data Scientists

- Managing data storage, automated processing (ETL)
- Building tools for data access, analysis, and visualization
- Productionizing models designed by Data Scientists

Machine Learning Engineer / Researcher

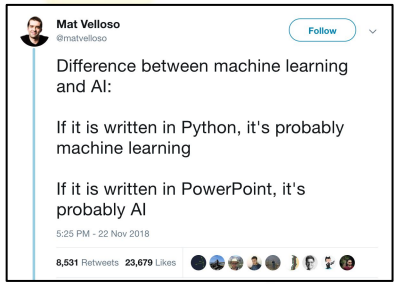
Machine Learning is an often-misunderstood term - most statistical models qualify as “Machine Learning”

Thus, “Machine Learning” is an important tool for Data Science

ML Specialists work specifically on building Machine Learning models or products

- ML Researchers come up with the mathematical model (and often a prototype)
- ML Engineers deal with the implementation and productionization issues

From Wikipedia: “Machine learning (ML) is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead. It is seen as a subset of artificial intelligence.”



“You have to get up pretty early in the morning to beat logistic regression.” - Philip Apps, Pinterest

Data Scientist

Lots of variety depending on company, team, region, etc. Can be similar to any of the related roles.

The set of skills needed to make data meaningful or useful:

- **Statistics** - select appropriate statistical methods and interpret them well
- **Computing** - work with complex and large-scale data and prototype models and tools
- **Analysis** - translating a real-world problem to a statistical model; understanding measurement, careful interpretation to translate from statistical finding to real-world insight and action
- **Communication** - clear and careful communication of findings and ability to work with many other roles (translate from “math-speak” to “PM-speak”)
- **Product** - understanding the product and how it’s used to understand how to approach business questions and interpret product metrics in a holistic “product” manner
- **User** - understand the user; focus on the user; be able to use data to develop insights into what the user is thinking or feeling

"Data science is the process by which data becomes understanding, knowledge, and insight" - Hadley Wickham, Rstudio

What Data Scientists Do

What Data Scientists Do

(The problems we solve and methods we apply)

- Evaluation
- Predicting User Behaviour
- Forecasting
- Measurement / Metrics Development
- Targeting
- Optimization
- Data Products
- DataOps / Data Democratization

Evaluation

- Determining the impact of a proposed product launch or change
- “Change” might be a user interface change, a new feature, an infrastructure change that’s not expected to be visible to users, or anything
- Work includes:
  - Collaborating with the engineering team to ensure necessary data collection is in place
  - Designing and launching an appropriate experiment (usually)
  - Evaluating data to ensure experiment and data collection are running correctly
  - Analysis of results to develop a coherent picture of the impact of the change
    - Often making a go/no-go recommendation

Measuring Impact

- KPI (Key Performance Indicator) Metrics
  - The types of things that track product success: e.g. DAU (Daily Active Users), Conversion Rate (how many users make purchases or sign up), or even just revenue
  - Are things we really care about, but are often hard to change enough to detect in a single experiment
- Performance Metrics
  - How well is the product performing: e.g. how fast, how often does it crash, what proportion of the time does the user click on a search link, etc.
  - Easier to detect changes, but complex to interpret - example of Quantum performance and Google Search clicks
- Other Behavior Metrics
  - Metrics that measure how the user is using the product, e.g. clicks on particular UI elements, etc.
  - Very easy to move but much more difficult to interpret if a movement is good or bad
  - Often requires great skill to put together collection of metric movements to develop coherent picture of what's happening

Predicting User Behaviour

- Logistics: forecasting utilization to effectively allocate resources
    - Example of search volume and rate of adaptation to daylight savings
- Predict interest in new features to prioritize development
- Predict interest or desire: recommender systems
- Modeling determinants of product success - understanding what's important to users

"Predictive analytics (PA)—Technology that learns from experience (data) to predict the future behavior of individuals in order to drive better decisions."

— Eric Siegel, Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die

Predictive vs. Explanatory Modeling

- Predictive Modeling
  - all we care about is predicting the outcome (will the user use the product or make the purchase, will the user like this movie, is the user male or female, etc.)
  - Most “ML” Methods like neural networks are purely predictive, but there is work in “model explainability” that offers some explanatory power as well
- Explanatory Modeling
  - Want to understand what the determinants of the outcome are
    - If the outcome is “will the user use the product” we want to know what kinds of users are more likely to use the product under what circumstances
    - Example: “women with high education are more likely to use it” or “people are more likely to use it if they have a very fast internet connection”

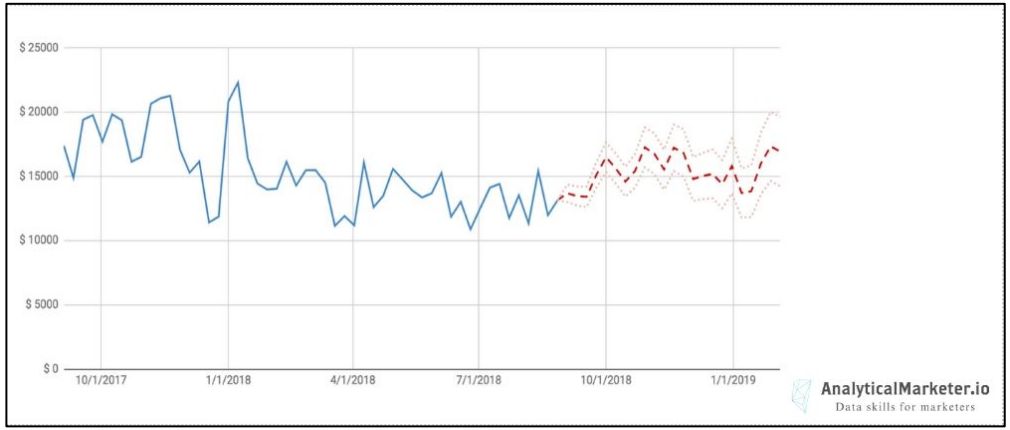
Forecasting

- Typically means modeling KPI-type metrics like revenue, DAU, conversion rate, etc.
- Make predictions of how metrics will move in the future to help set realistic goals and make good plans
- Also useful for understanding impact of events

Forecasting Components - A metric normally moves in seemingly unpredictable ways, but this can be decomposed to individual sources of variation

- Day of week
- Seasonal
- Trend
- Event impact

Controlling for other sources of variation can make it easier to see the impact of an event like a media mention or new release



Measurement / Metrics Development

- Gap between what you want to measure and the data that's actually available
- Data is typically noisy, complex, and does not behave as expected
- Measurement development means understanding what your data is actually telling you
- Metric development means developing some function of the available data that tells you something useful
- Examples:
    - How to measure daily active users? Is it enough that a user opens the browser to count as a user? Or should the user need to open some web site to count? How many?
    - If a user performs more searches on Google does it mean that he or she is more satisfied with the product? Or less?

"In our lust for measurement, we frequently measure that which we can rather than that which we wish to measure... and forget that there is a difference." - George Udny Yule

Targeting

- “Micro-targeting” or “personalization”?
- Biggest example is online advertising - the ad shown to a user can depend on huge amounts of data about that person
- Feature discovery: a product might provide instructions for how to use a feature that the user has never used but might be interested in
- A game might adapt its level of difficulty to the user in order to encourage him or her to keep playing longer

Optimization

- Making processes more efficient
- Industry knowledge tells you how to configure the heating system in a building to be efficient, but with data-driven optimization it may be possible to find an even more efficient solution for a particular building and set of needs
- Optimization is quite general - comparing a set of different onboarding experiences for a product is an optimization problem but also can be considered evaluation
- Other logistics problem - finding more efficient ways to run a business

Data Products

- Some products are fundamentally Data Science Products
- Often in the “Machine Learning” space:
  - Image recognition, natural language processing, etc.
  - “People You May Know” - LinkedIn
  - Fraud Detection
- Often relates to other aspects of Data Science work:
  - “Personalization”: an app to recommend places a person should visit while traveling based on his/her interests
  - “Optimization”: Uber’s system to dispatch cars to riders

DataOps / Data Democratization

- Data “democratization” or “accessibility”
    - Creating tools and processes that allow all stakeholders to work with data
- “Data Literacy”: if a company has good tools and processes such that everyone is empowered to explore and understand the data, it’s much easier to do Data Science
- Self-service: Data Scientists can save time if stakeholders can do some work themselves
    - Danger of “enough rope to hang yourself” - it’s **extremely** easy to come to a wrong conclusion working with data analysis

“If you torture the data long enough, it will confess to anything.” - Ronald H. Coase, Essays on Economics and Economists

Data Science Skills

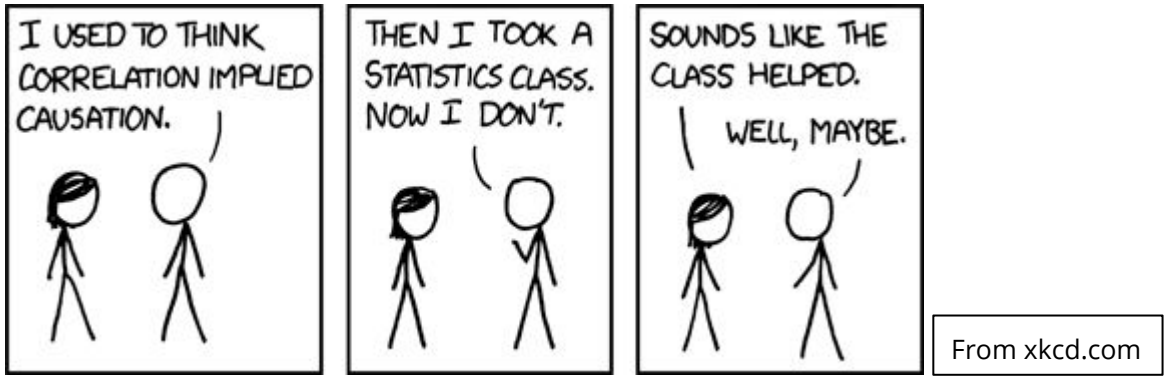
Data Science Skills

(How to "think data")

- Rigor (Skepticism)
- User Focus
- Statistics Skills
- Computational Skills
- "Data Intuition"
- Communication Skills

Rigor (Skepticism)

- Really just scientific thinking - but the nature of statistics makes it very easy to make mistakes
- Basic Idea:
    - I observe some data about a system and my analysis tells me that the data suggests that the system works one way (my hypothesis). I need to think very carefully about other reasons that the data might be that way even if my hypothesis isn't true.
    - Need to dig deeper and consider all possible explanations for the data we observe.
- Metaphor is like reading lips - you only have some information related to what you really want to understand



From xkcd.com

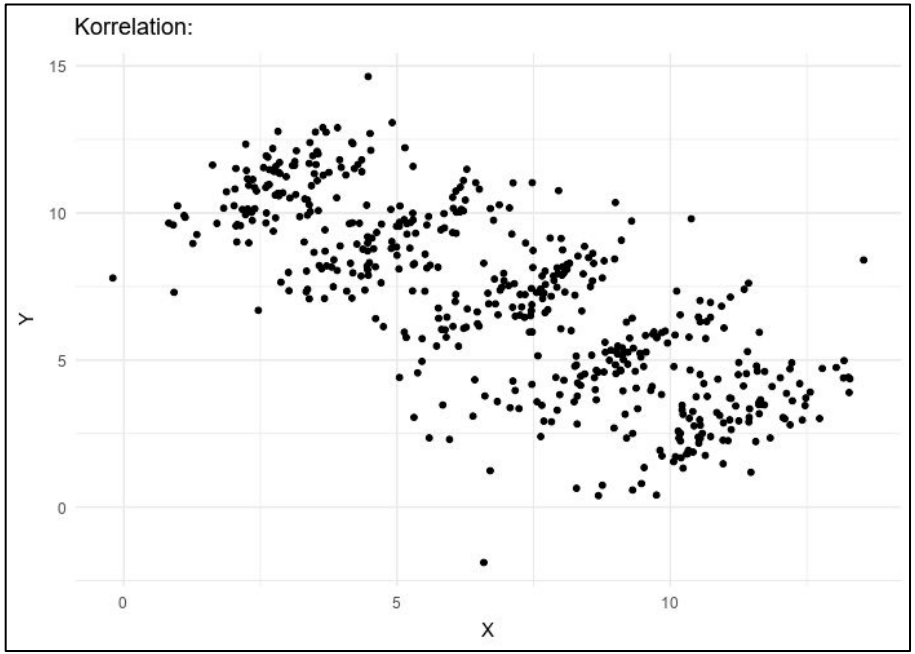
Rigor (Skepticism)

- Understanding if a method is appropriate for a problem
  - Are the assumptions of the method met?
  - If not (as is often the case), is the method sufficiently robust to still give meaningful results?
  - Is there a simpler method that would work
- Being very careful about how the data is interpreted and assumptions that are made about its behaviour
- Careful documentation of analysis process

Example: Simpson's Paradox

	Men		Women	
	Applicants	Admitted	Applicants	Admitted
Total	8442	**44%**	4321	35%

Department	Men		Women	
	Applicants	Admitted	Applicants	Admitted
A	***825***	62%	108	**82%**
B	***560***	63%	25	**68%**
C	325	**37%**	***593***	34%
D	417	33%	375	**35%**
E	191	**28%**	***393***	24%
F	373	6%	341	**7%**



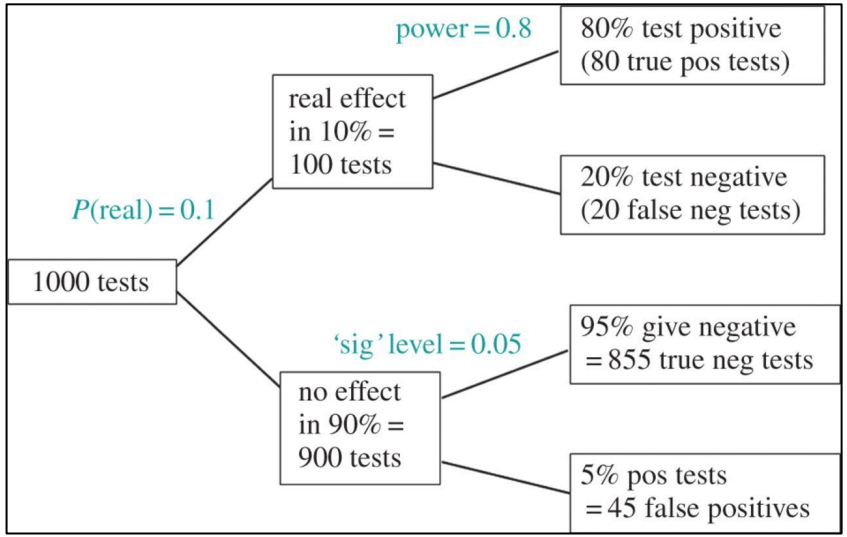
Tables and graph from https://en.wikipedia.org/wiki/Simpson%27s_paradox

"Why Most Published Research Findings Are False"

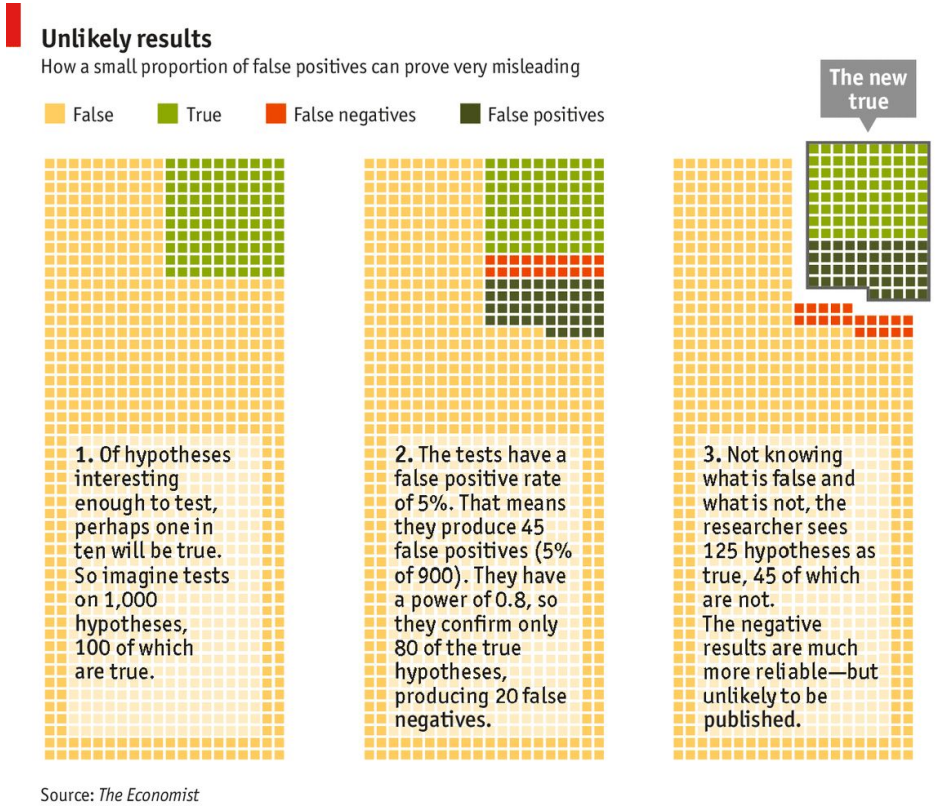
An oversimplification:

- Sometimes data will appear to show something interesting just by chance
- Null hypothesis & Alternative Hypothesis
- Standard practice: p-value <= 5%
- So 1 in 20 statistical findings is false?
- But it gets worse...

	Predicted = TRUE	Predicted = FALSE
Actual =TRUE	TP (True Positive)	FN (False Negative)
Actual =FALSE	FP (False Positive)	TN (True Negative)



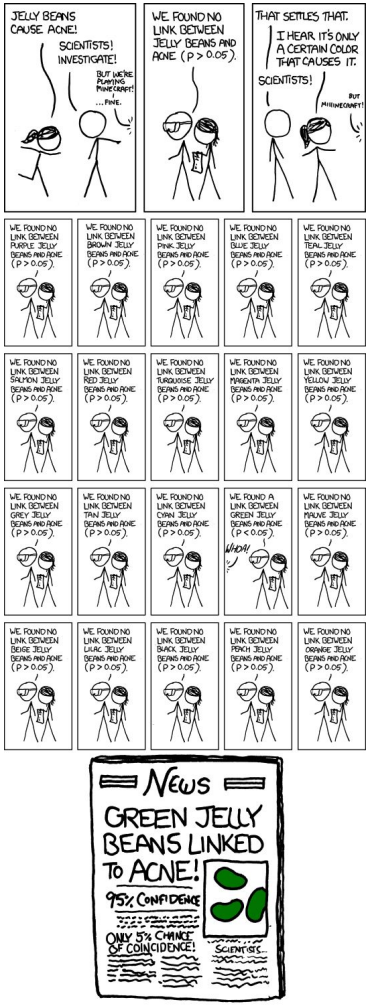
"Why Most Published Research Findings Are False"



Source: *The Economist*

Example: Multiple Comparisons

- If we expect a “false positive” for 1 out of every 20 comparisons, making lots of comparisons is dangerous.
- Example of PM looking at metrics for his launch sliced by country…
- Can be subtle: my example of determinants of trust of tech companies



User Focus

- A measurement issue: measuring how satisfied a user is with a product is very difficult - so we have to use imperfect metrics for evaluation - it's important to think very carefully about whether a change to a product might make the metric seem better but actually make things worse for the user - need to try to understand how data relates to the actual experiences of the user
    - More queries is better?
- Also need to remember the user of the analysis you produce! Statistical results are easy to misinterpret, so consider the interests and statistical literacy of your audience and take responsibility to avoid misinterpretation.

Statistics Skills

- Ability to work with noisy, complex data
- Build appropriate models to understand data and solve problems
- Understand uncertainty!
- Understand methods for:
    - Descriptive analysis and Visualization
    - Inference (hypothesis testing)
    - Exploratory modeling (looking at correlations and meaningful transformations of data)
    - Explanatory modeling (understand what factors influence something of interest)
    - Predictive modeling
- **Most of the work can be done with very simple statistical methods**

Computational Skills

- Need to be able to use tools necessary to solve problems:
  - “Big Data” tools: SQL/Spark/etc.
    - Obtaining data
    - Basic exploration and manipulation/aggregation of data
    - MapReduce paradigm (as in Spark/Hadoop/etc.) is key for handling large data
  - “Analysis” tools: R/Python/etc.
    - More flexible exploration and data manipulation
    - Fitting models, performing inference, creating visualizations, etc.
- Lots of other tools that are used and these categories are fuzzy

Data Intuition

- Perhaps the most important skill
    - Other skills are easier to learn over time - and tools and methods are always changing, so the ability to learn is more important than knowing a laundry-list of tools and statistical models.
- The ability to translate between real-world problems and statistical methods:
    - Given a problem to solve, understand how data can be used to solve it
    - Given statistical results, understand what they really tell you about the real world
- Also, being able to look at data and develop insight:
    - Recognizing patterns and important features
    - Recognizing signs of data quality issues

Communication Skills

- Also the most important skill ;)
- Largely about translation - being able to translate statistical results and concepts to a language that your audience will understand and interpret correctly
    - Generally the audience varies widely: statisticians reviewing your work, engineers that are implementing a feature you're analyzing, PMs on those projects, executives making launch decisions, etc.
- **Storytelling**: knowing how to use data to tell a coherent story of what's happening
- Data Science is fundamentally very collaborative and interdisciplinary. Working well with others is essential.
- "For most data scientists, your value is expressed ultimately as the communication of the result or a finding to a stakeholder. To some extent this is truer for data science than other technical fields; an entry-level data scientist might generate an experiment report that will travel to the C-suite, whereas entry-level programmers direct work outcomes don't necessarily travel beyond committing code for a shipped feature." - Rebecca Weiss (my director)

Data Science Tools

Data Science Tools
(Software and Data Models)

- R / Python
- SQL
- MapReduce (Spark, etc.)
- Visualization
- Many others...

R / Python

- Primary working language of most Data Scientist is either R or Python
    - R is probably more common, especially among Data Scientists with a statistics background
    - There is support for a greater range of statistical methods in R (especially newer methods)
    - But Python is better ;)
- A Data Scientist should really be able to “read” both, but only needs expertise in one of the two
- Important: “expertise” means use as a programming tool!
    - I’ve interviewed many candidates that could use R or Python “interactively” (typing a command, getting results, typing another command, etc.) but could not write a simple algorithm. You don’t need a Computer Science degree but you should be able to solve basic algorithm-writing challenges (I like Project Euler) in your preferred language.

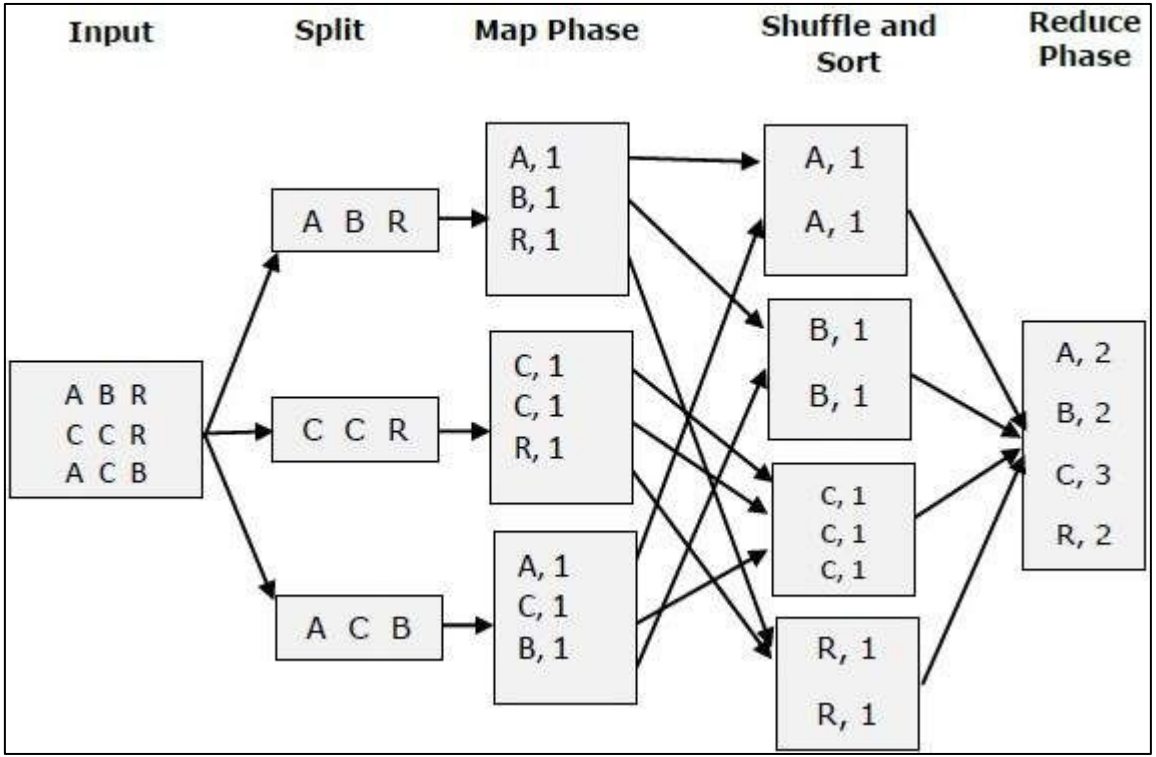
SQL

- Often the first step is obtaining data uses SQL
  - Maybe data is too big to load into memory (needed for R or Python) so we use SQL to select just the subset (or aggregation) that we need
  - We can also do data manipulation and descriptive analysis in SQL
  - Some people do really crazy things too… I searched for how to do linear regression in SQL:

```sql
-- test data (GroupIDs 1, 2 normal regressions, 3, 4 = no variance)
WITH some_table(GroupID, x, y) AS
(       SELECT 1,  1,  1    UNION SELECT 1,  2,  2    UNION SELECT 1,  3,  1.3
  UNION SELECT 1,  4,  3.75 UNION SELECT 1,  5,  2.25 UNION SELECT 2, 95, 85
  UNION SELECT 2, 85, 95    UNION SELECT 2, 80, 70    UNION SELECT 2, 70, 65
  UNION SELECT 2, 60, 70    UNION SELECT 3,  1,  2    UNION SELECT 3,  1, 3
  UNION SELECT 4,  1,  2    UNION SELECT 4,  2,  2),
-- linear regression query
/*WITH*/ mean_estimates AS
(   SELECT GroupID
          ,AVG(x * 1.)                                          AS xmean
          ,AVG(y * 1.)                                          AS ymean
    FROM some_table
    GROUP BY GroupID
),
stdev_estimates AS
(   SELECT pd.GroupID
          -- T-SQL STDEV() implementation is not numerically stable
          ,CASE      SUM(SQUARE(x - xmean)) WHEN 0 THEN 1
           ELSE SQRT(SUM(SQUARE(x - xmean)) / (COUNT(*) - 1)) END AS xstdev
          ,      SQRT(SUM(SQUARE(y - ymean)) / (COUNT(*) - 1))     AS ystdev
    FROM some_table pd
    INNER JOIN mean_estimates  pm ON pm.GroupID = pd.GroupID
    GROUP BY pd.GroupID, pm.xmean, pm.ymean
),
standardized_data AS                    -- increases numerical stability
(   SELECT pd.GroupID
          ,(x - xmean) / xstdev                                 AS xstd
          ,CASE ystdev WHEN 0 THEN 0 ELSE (y - ymean) / ystdev END AS ystd
    FROM some_table pd
    INNER JOIN stdev_estimates ps ON ps.GroupID = pd.GroupID
    INNER JOIN mean_estimates  pm ON pm.GroupID = pd.GroupID
),
standardized_beta_estimates AS
(   SELECT GroupID
          ,CASE WHEN SUM(xstd * xstd) = 0 THEN 0
                ELSE SUM(xstd * ystd) / (COUNT(*) - 1) END      AS betastd
    FROM standardized_data pd
    GROUP BY GroupID
)
SELECT pb.GroupID
      ,ymean - xmean * betastd * ystdev / xstdev                AS Alpha
      ,betastd * ystdev / xstdev                                AS Beta
FROM standardized_beta_estimates pb
INNER JOIN stdev_estimates ps ON ps.GroupID = pb.GroupID
INNER JOIN mean_estimates  pm ON pm.GroupID = pb.GroupID
```

MapReduce

- Was “invented” by Google, but is a pretty simple concept in distributed computing
- When data is too big for a computer, find a way to split the computation among many computers
- Not just any computation can be split like this - but MapReduce specifies a nice wide class of computations that can be parallelized
- Newer tools like Spark are slowly replacing this approach

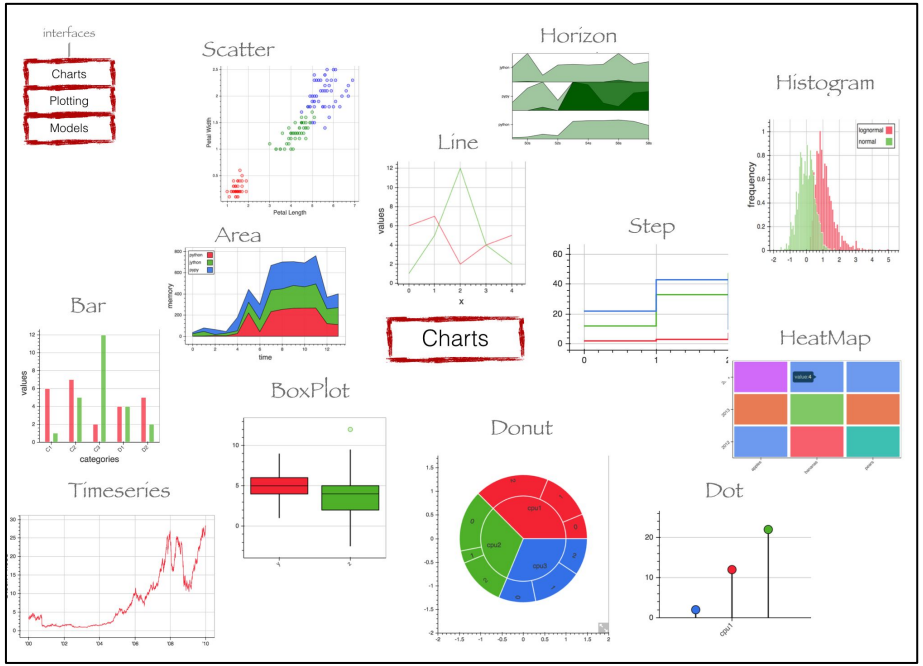


Spark and Similar Tools

- Allows more general manipulation of large data
- May use techniques like MapReduce “under the hood” but hides some of the complexity
- Allows you to pretend that your data is a simple data frame or similar construct

Visualization

- Can be critical in some Data Science roles and unimportant in others
- Part of effective communication
  - Visualization adds a whole new layer on how misleading statistical results can be
- Dashboards are a thing!
  - Can be a way of presenting results on an ongoing experiment or study
  - Tracking performance/growth/satisfaction/whatever
  - Can be tools to make data more accessible to stakeholders without skills/time needed to access raw data



BREAK

Elements of Data Science

- Obtaining and processing data
- Descriptive Statistics
- Visualization
- Inference
- Modeling (predictive & explanatory)
- Forecasting
- Metric Development
- Business Problem Translation
- Experiment Design (counterfactuals)
- Data Operations / Data Democratization
- User Journey / Retention / Segmentation

Data

- Many types: tabular, relational, image/sound/video, spectrum, etc…
  - Will mostly focus on tabular
- Aside: “Senior Data Science” skill - deeply understand the difference between data and the process that generated the data
  - Is the data truly capturing what you think it is?
  - Are there data quality problems that need to be considered?
  - Are there factors that impact the data other than what you’re interested in?

Tabular Data

- Rows & Columns
  - Rows represent some “unit of observation - could be users, events, days, products, etc.
  - Columns are “variables” that provide some information about each “unit of observation”
- Generally a column should always have the same “type” of data:
  - Number, text, date
  - Advanced note: could be array or data structure too!

	A	B	C	D
1	Region	Salesman	Date	Revenue
2	North	Rob	10/21/2017	$2,059
3	South	Joe	10/9/2017	$1,908
4	North	Rikki	9/27/2017	$1,429
5	East	Chris	9/15/2017	$2,588
6	North	Rikki	9/3/2017	$2,085
7	East	Chris	8/22/2017	$1,996
8	South	Joe	8/10/2017	$1,718
9	North	Rikki	7/29/2017	$2,851
10	North	Rikki	7/17/2017	$2,735
11	East	Chris	7/5/2017	$2,864
12				

Data Types (KNOW THIS WELL FOR NEXT WEEK)

- **Dichotomous/Binary**: is either “yes or no”, “true or false”. May be represented as words (“true” and “false”) in data or numbers (1 and 0).
- **Categorical**: takes on one of a set of possible values with no order between them. May be represented and text or integers.
- **Ordinal**: like categorical, but values have some order. Most common example is likert scale: (“strongly disagree”, “disagree”, …, “strongly agree”).
- **Interval/Ratio**: numbers in which you can measure the difference between any two values. Is “Ratio” if it is meaningful to say that one value is, for example, twice as big as another.

Descriptive Statistics

- Provide insight into basic characteristics of data.
- A “statistic” is some number (or set of numbers) that summarizes some section of your data table
    - For example, the mean might summarize one complete column. The correlation coefficient might summarize the relationship between two columns.
- Common tools are: mean, median, mode, min, max, variance, correlation, table, cross table.
- Will learn much more next week!

Visualization

- A tool to communicate data or statistics. Can often be very effective, but needs to be used carefully.
    - Too much “generating a plot just for the sake of generating a plot”
- Deeply related to “data storytelling”, which come later in the course.
- Will talk some about visualization next week!

Statistical Modeling

- Key idea: There is some process that generates data. We want to build a model of that process.
  - Could be to better understand the process (explanatory modeling) or so we can predict some characteristic of the data it generates (predictive modeling)
  - Examples include linear/logistic regression, and pretty much all machine learning techniques
- Basic terminology
  - Dependent variable, “y variable”, “outcome” - the variable that we want to predict or understand how other variables influence it
  - Independent variables, “x variables” “predictors” - the variables from which we make the prediction or that we consider for influence

Statistical Modeling (examples)

- Taking into account sex, how does weight impact the risk of heart attack
  - Logistic regression (explanatory or predictive modelling)
  - Dependent variable is “had heart attack”
  - Independent variables are sex and weight
- Given products viewed on an online shopping site, how much is a person likely to spend that day?
  - Linear regression (or something more complicated) (predictive modeling)
  - Dependent variable is “amount user will spend”
  - Independent variables are “products viewed”

Machine Learning is Statistical Modeling

- Independent variables are the state of the sensors, dependent variable is steering/braking/acceleration actions to take
- Independent variables are the pixels in a photo, dependent variable is the name of the person in the photo...

Inference

- Can be very simple - “t test”?
- Can be very subtle - generalizing from a sample to a population
  - Given that I’ve measured the heights of one of you, what do I know about the height of the rest of you?
  - What if I measure half of you?
- Confidence intervals and hypothesis tests
- Frequentist vs. Bayesian

BREAK

How to Approach Data Science Problems

- Curious, Skeptical, Systematic
- Other words?

Get To The Bottom of the Problem

- Especially if request comes from another (but even if it comes from yourself) take time to deeply understand what's really being asked.
- What does the requestor want to end up learning?
- What actions might the requestor take based on what you provide.
- "Over interpretation" is a major problem
- Telling the Data Scientist "what to do" instead of "what I want" is a major problem

Don’t Listen to “The Call Of The Data”

- Often the data itself will suggest certain ways to approach it
- Be careful of this and make sure everything you do is informed my the problem you want to solve

Rigor and Skepticism

- When you have a finding (especially one you're happy about), think about alternative reasons you might have found that result
  - Data doesn't represent what you think it does
  - Overfitting, multiple comparisons, bad inference
  - Confounding
- What can you do to rule out these alternatives?
- Keep strengthening your analysis

Let's Learn!

- From here, the course will go into much more detail on these themes and many others
- We want to learn too! This is the first time many of us are doing this. Please provide feedback, both formally and informally.
- We want this to be fun and useful - work with us to make it succeed!

Course Structure

- Lectures on Mondays and labs on Wednesdays
- Ask questions of staff after lecture on Mondays and during labs on Wednesdays
- Ask questions of other students on Mighty Networks and in person!
- Work in groups of one to three for labs
    - Learning to collaborate is important! But making sure you learn the material is important too. Take ownership of your own education.

Labs

- Get setup for labs:
  - You will need a Kaggle account to download some data sets: https://www.kaggle.com/
  - You can use any python notebook environment you want but we will not offer any support in case of problems. We recommend Google Colab (https://colab.research.google.com/) as it is fairly “problem-proof”.
- This week’s lab:
  - Now you know the elements of Data Science. Please find a blog post, news article, etc. with some data science elements and share on Mighty Networks for discussion on Wednesday. Try to think about what elements of data science you can identify in the article.