Data Science

Deriving Knowledge from Data at Scale

Jun 23, 2016

Wee Hyong Tok, PhD
tokwh@uw.edu

Course Logistics

Deriving Knowledge from Data at Scale

- Course Website:
- Instructor: Wee Hyong Tok
- Teaching Assistant: Debmalya Chandra
- Course Times: Thurs 6pm – 9pm
- Office Hours: on demand
- email: tokwh@uw.edu (weekend reply)
- Please use the Class forum

Open Discussion

- Introduce yourself
- What is your background?
- What experience/expertise do you have relevant to data science?
- What experience/expertise do you feel that you need to acquire?

UW Data Science

- Thoughts, so far?
- What have been the highlights (most enjoyed)?
- What have been the lowlights (least enjoyed, could be improved?)
- What are you hoping to get out of this course in the program?
- Open questions you have, topics you wish to cover?

Big Data and Data Science the case for Advanced Analytics in DW

"Big Data: The Next Frontier for Innovation, Competition and Productivity." – McKinsey Report

- Need 140,000 to 190,000 more people with "deep analytical" skills, typically experts in statistical methods and data-analysis technologies
- Need 1.5 million more data-literate managers, whether retrained or hired
- $300 billion – potential annual value to US health care – more than double the total annual health care spending in Spain
- €250 billion potential annual value to Europe's public sector administration – more than the GDP of Greece
- $600 billion potential annual consumer surplus from using personal location data globally
- 60% potential increase in retailers' operating margins possible with big data

Lecture Outline

- Course Content
- Course Logistics
- Course Motivations
- Doing Data Science
- Open Discussion
- Homework

Course Content

What We **Will** Cover

- Data Preparation
- Practical Techniques of Data Science
- Machine Learning over Data
- Experimentation
- Literature, what are people writing and saying about data science
- Guest Lecturers, practicing data scientists (2)

Course Content

Data Science Vocabulary List, *the language of data science (quick poll)…*

- Machine learning
- Supervised learning
- Unsupervised learning
- Training set or training sample
- Test set or test sample
- K-nearest neighbors
- Regression
- Residual analysis
- Least squares estimators
- Confusion matrix
- Classification
- Prediction
- Forecasting, ARIMA, Holt-Winter
- Over fitting
- K Fold Cross-validation, why use it?
- Loss functions
- Labels
- Euclidean distance
- Bias, variance, bias variance trade-off
- ROC Curve

Course Content

What We **Will Not** Cover

- Probability and Statistics, *covered in second course of the program…*
- NoSQL, *covered in the first course of the program…*
- Machine Learning Theory, *though we will discuss ML techniques…*
- Data Mining*, again we will discuss techniques from data mining…*
- Artificial Intelligence, *techniques often covered AI literature…*

Course Logistics

Rough Structure of Each Lecture

- Review Discussion, *this is important…* 20 minutes
- Data Science Practices 60 minutes
- Break 10 minutes
- Machine Learning Over Data 60 minutes
- Hands On, *again important…* 30 minutes

- During the Week
  - Reading, almost weekly [2 hours], expect you to discuss highlights
  - Homework, ~5 data science exercises.

See Course Outline

Lecture 1

Techniques

- Overview
- The Data Science Workflow
- Best Practices

Data Science Practice

- Introduction to Data Science
- Causal Analysis in Display Advertising

Homework

- Reading

Lecture 2

Techniques

- Machine Learning Primer

Data Science Practice

- Elements of a Time Series
- Time Series Forecasting 1/2

Homework

- Reading

Lecture 3

Techniques

- Decision Trees
- Random Forests

Data Science Practice

- Introduction to Weka

Homework

- Time Series Forecasting
- Install Weka

Class Input

Profile Yourself, upload to Class Folder for Lecture 1 in PDF or Word

Profile yourself, on a **relative scale**, 1…6, rather than absolute scale, with respect to your skill levels in the following domains (1 low … 6 expert):

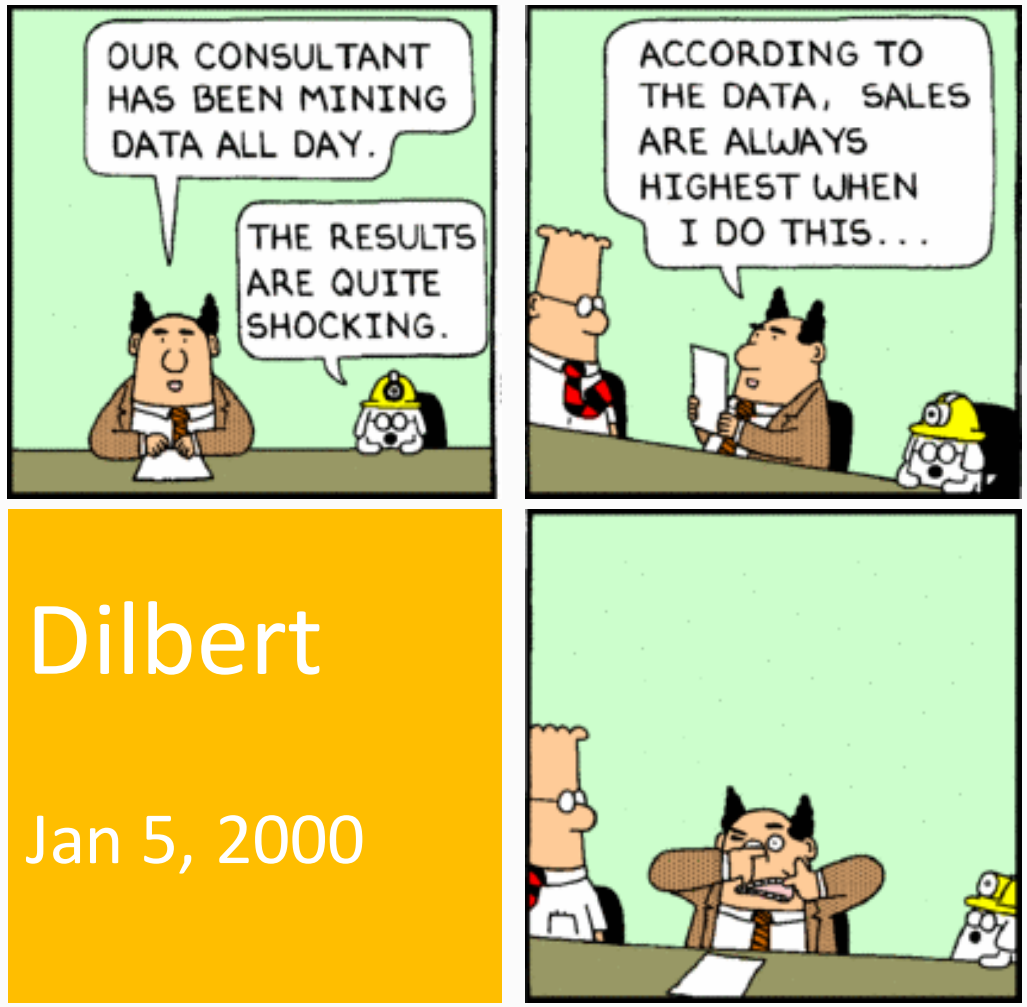
- Software engineering;
- Math;
- Statistics;
- Machine Learning;
- Domain Expertise;
- Communication and Presentation skills, and
- Data Visualization…

Open Discussion

Define Data Science

What kind of things does a data scientist do?...

Define “Data Scientist”

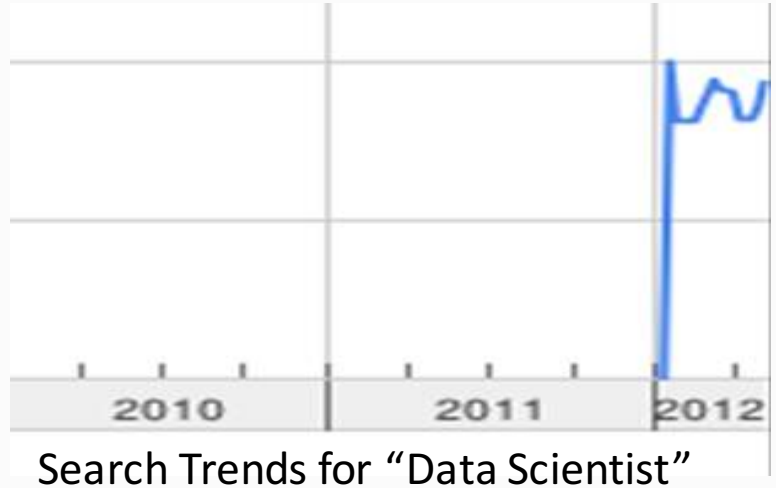


Dilbert

Jan 5, 2000

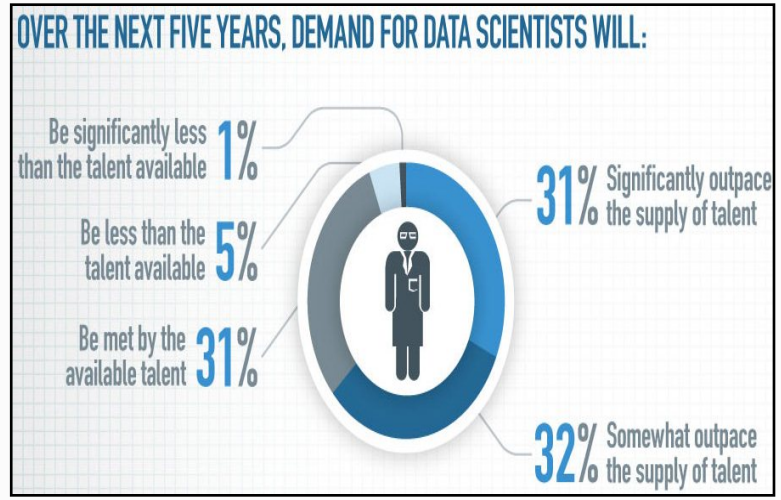
Data Scientist is a HOT new career path

> By definition all scientists are data scientists. In my opinion, they are half hacker, half analyst, they use data to build products and find insights. It's Columbus meets Columbo – starry eyed explorers and skeptical detectives.
>
> Monica Rogati (LinkedIn)



Search Trends for "Data Scientist"

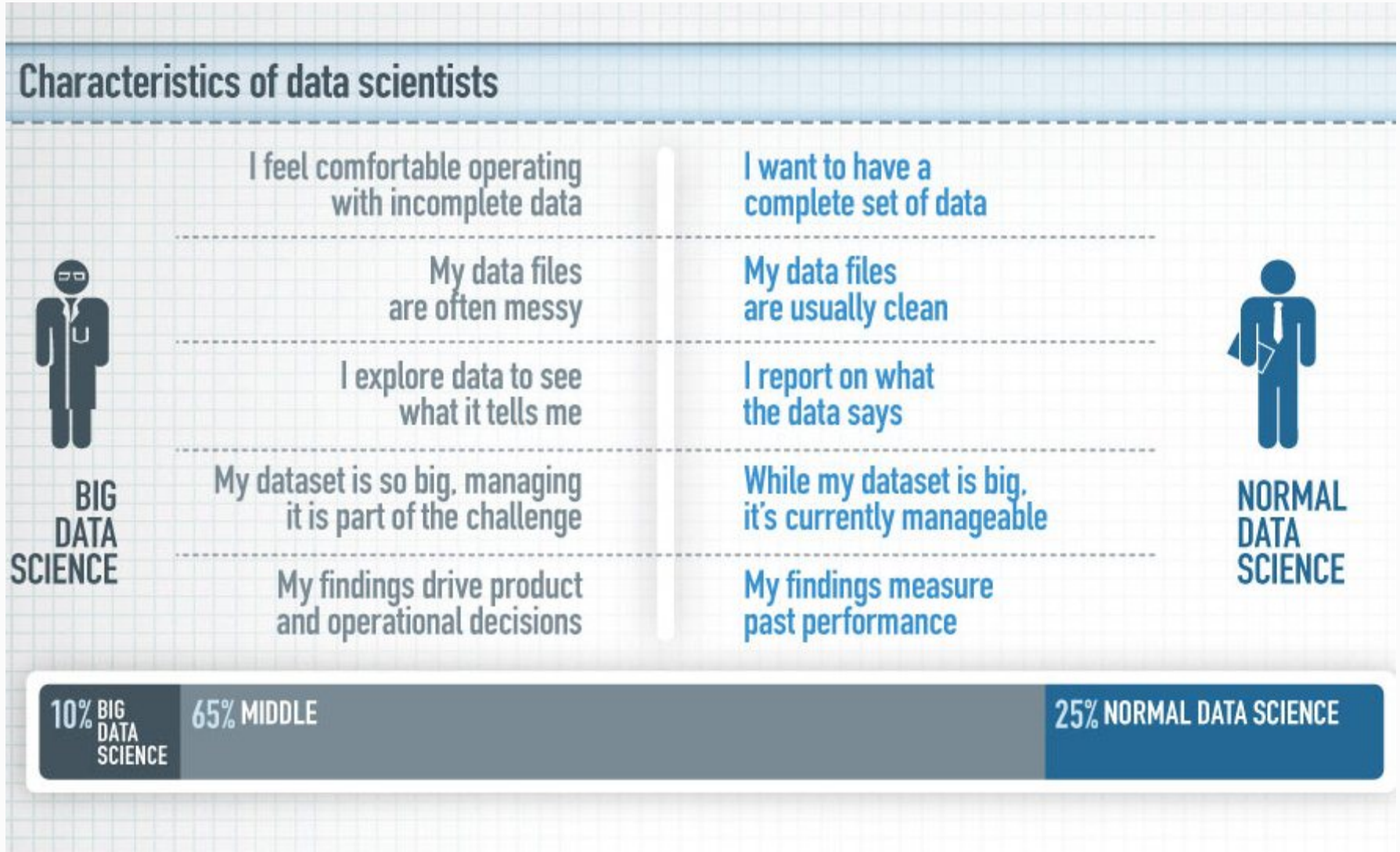
> A data scientist is someone who can obtain, scrub, explore, model and interpret data, blending hacking, statistics and machine learning. Data scientists not only are adept at working with data, but appreciate data itself as a first-class product.
>
> Hilary Mason (Bit.ly)



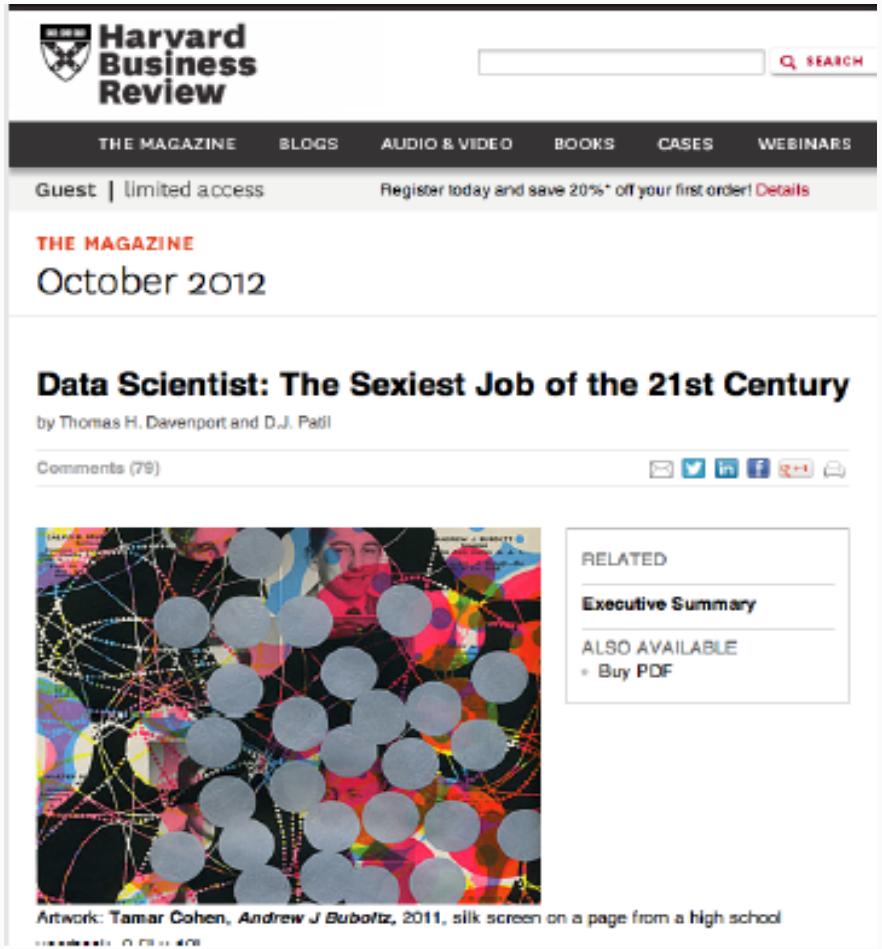
Computer Science

Machine Learning

Statistics



Doing Data Science



65% of enterprises feel they have a **strategic shortage of data scientists**, a role many did not even know existed 12 months ago…

Over 2/3 believe demand for talent will outpace the supply of data scientists

OVER THE NEXT FIVE YEARS, DEMAND FOR DATA SCIENTISTS WILL:

Be significantly less than the talent available **1%**

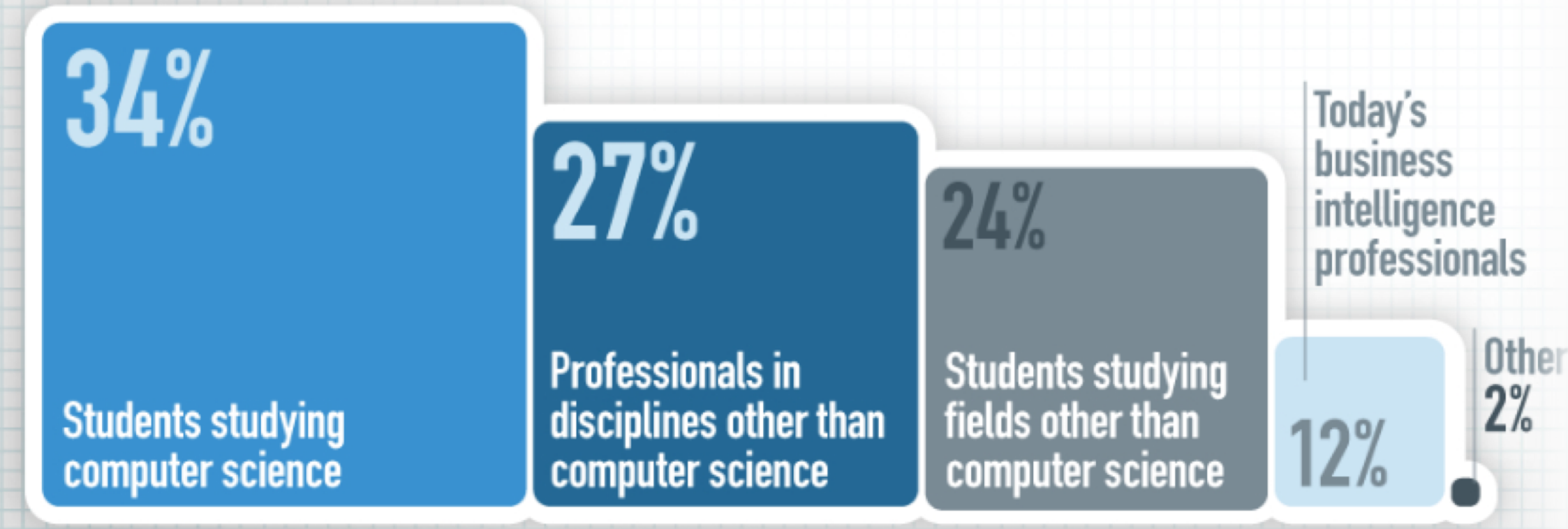
Be less than the talent available **5%**

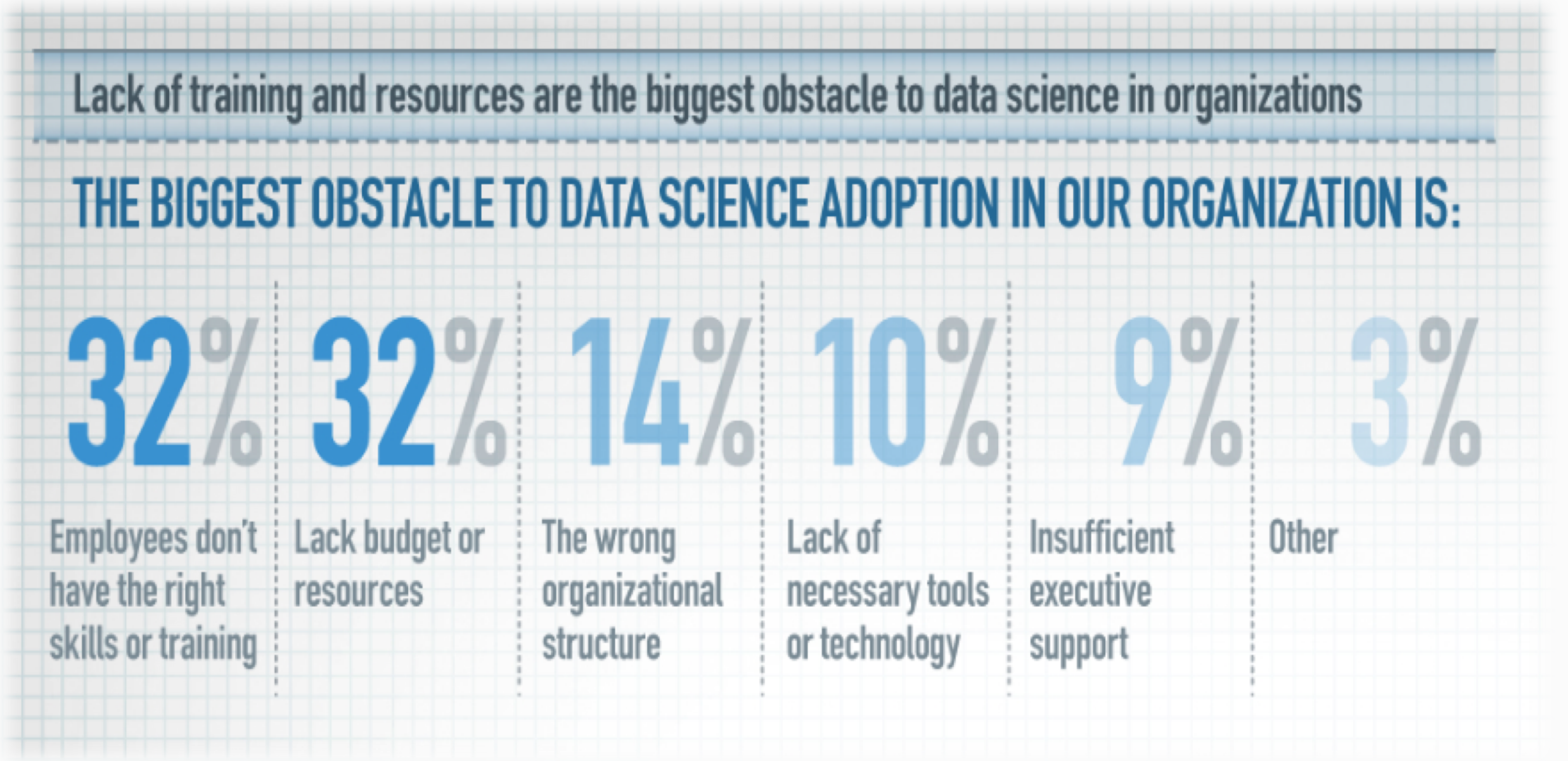
Be met by the available talent **31%**

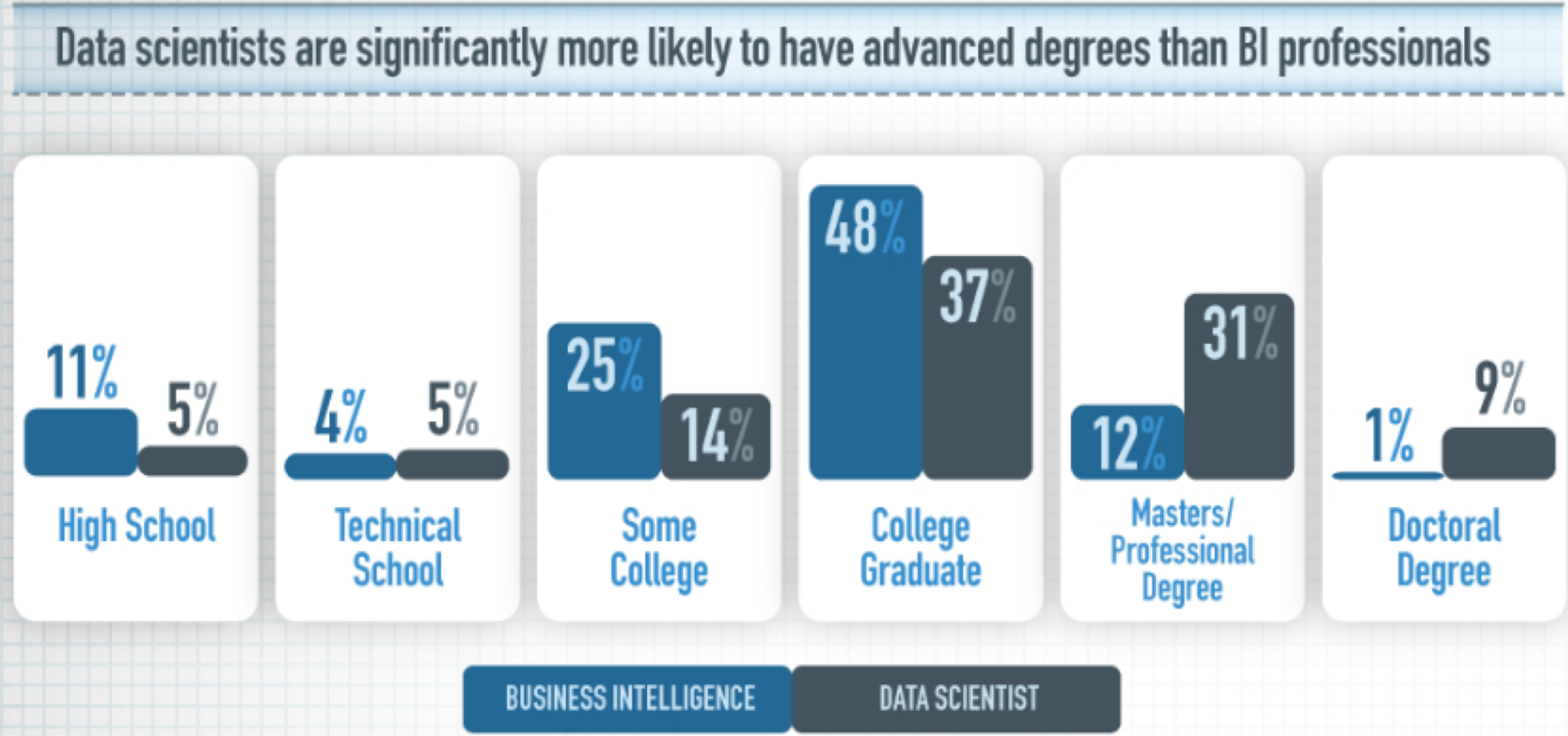
31% Significantly outpace the supply of talent

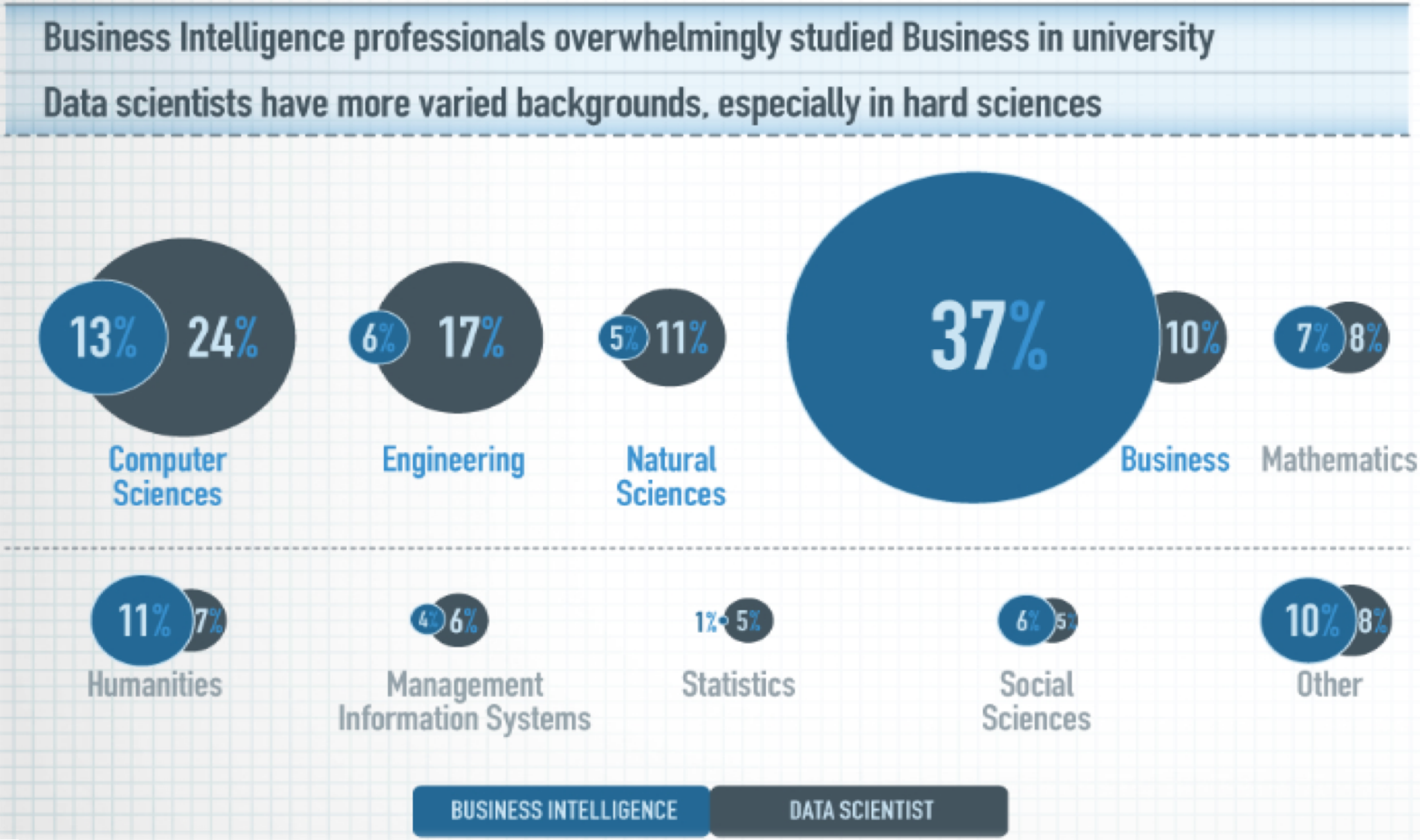
32% Somewhat outpace the supply of talent

Only 12% see today's BI professional as the best source for new data scientists




Lack of training and resources are the biggest obstacle to data science in organizations
THE BIGGEST OBSTACLE TO DATA SCIENCE ADOPTION IN OUR ORGANIZATION IS:
32%
Employees don't have the right skills or training
32%
Lack budget or resources
14%
The wrong organizational structure
10%
Lack of necessary tools or technology
9%
Insufficient executive support
3%
Other


Data scientists are significantly more likely to have advanced degrees than BI professionals
11%
5%
High School
4%
5%
Technical School
25%
14%
Some College
48%
37%
College Graduate
12%
31%
Masters/ Professional Degree
1%
9%
Doctoral Degree
BUSINESS INTELLIGENCE
DATA SCIENTIST


Business Intelligence professionals overwhelmingly studied Business in university
Data scientists have more varied backgrounds, especially in hard sciences
13%
24%
Computer Sciences
6%
17%
Engineering
5%
11%
Natural Sciences
37%
10%
Business
7%
8%
Mathematics
11%
7%
Humanities
4%
6%
Management Information Systems
1%
5%
Statistics
6%
5%
Social Sciences
10%
8%
Other
BUSINESS INTELLIGENCE
DATA SCIENTIST

Data scientists believe that new technology will create a demand for more data scientists

Data scientists decrease the number of data scientists required by automating much of the analytical work **17%**

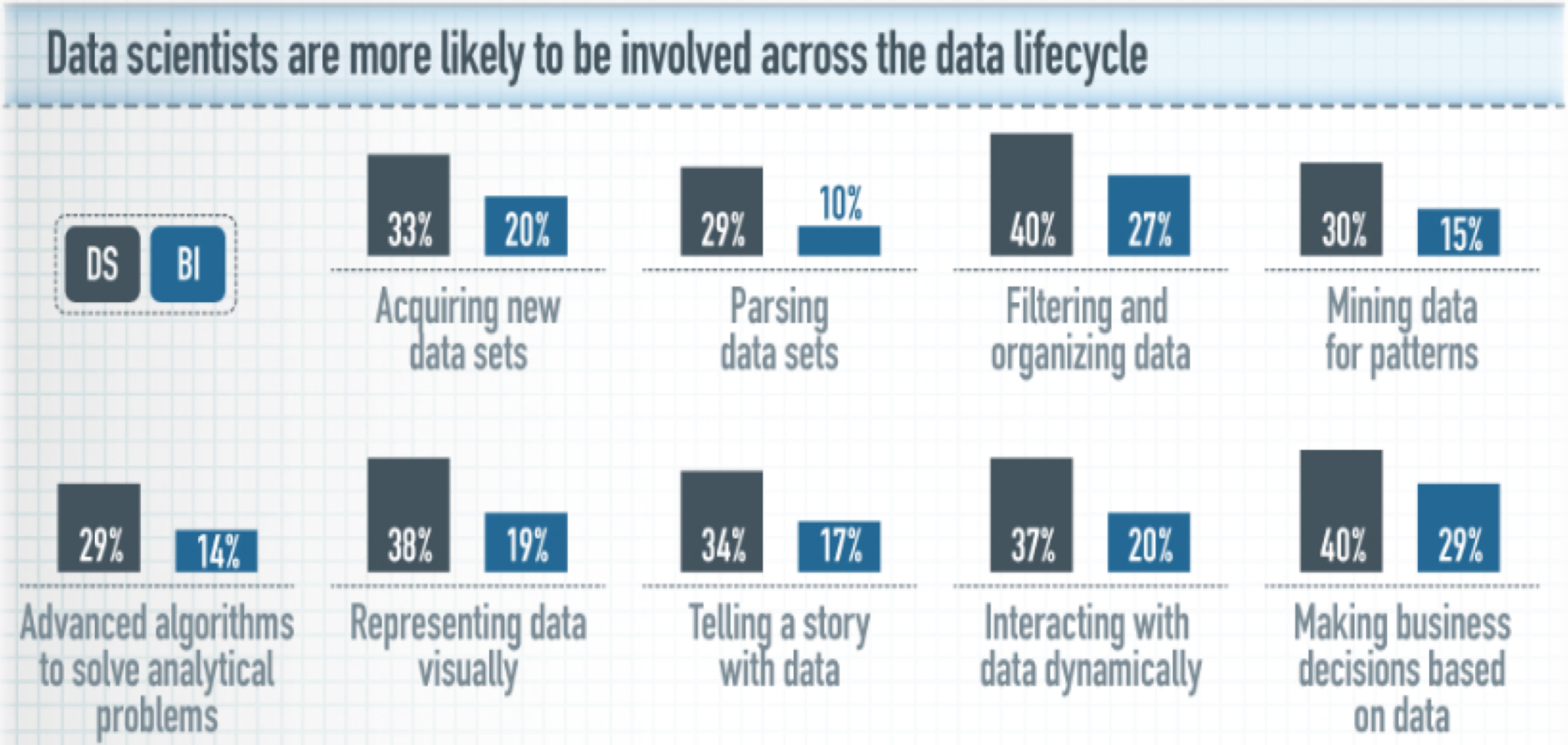
83% Data scientists increase the number of data scientists required by opening up new possibilities

Characteristics of data scientists

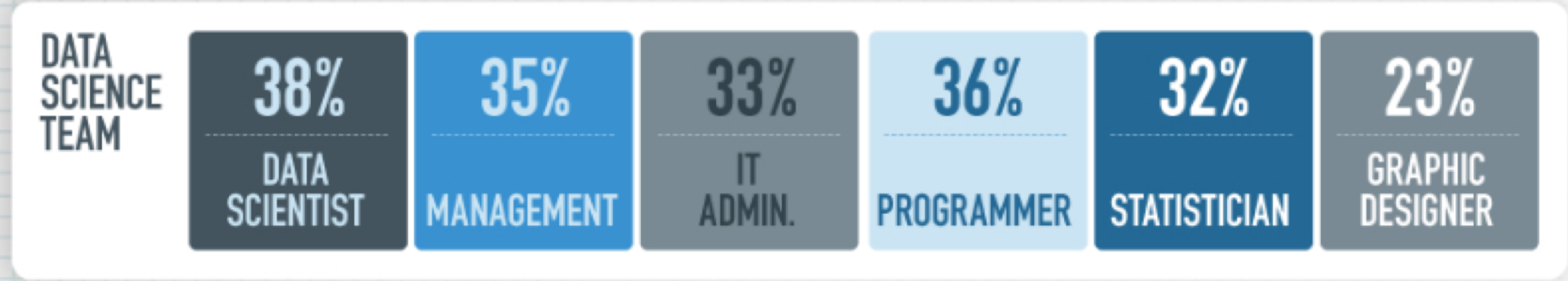
BIG DATA SCIENCE

I feel comfortable operating with incomplete data	I want to have a complete set of data
My data files are often messy	My data files are usually clean
I explore data to see what it tells me	I report on what the data says
My dataset is so big, managing it is part of the challenge	While my dataset is big, it's currently manageable
My findings drive product and operational decisions	My findings measure past performance

NORMAL DATA SCIENCE


Data scientists are more likely to be involved across the data lifecycle
DS
BI
33%
20%
Acquiring new data sets
29%
10%
Parsing data sets
40%
27%
Filtering and organizing data
30%
15%
Mining data for patterns
29%
14%
Advanced algorithms to solve analytical problems
38%
19%
Representing data visually
34%
17%
Telling a story with data
37%
20%
Interacting with data dynamically
40%
29%
Making business decisions based on data

Who does a data scientist work with?



Data Science

10 Important Ideas

10 Important Ideas in data science shaping the field and that differentiate data science from business intelligence or related topics. **Each will be a topic of at least one lecture** in the data science track of our course. Today we will briefly introduce each…

Data Science

10 Important Ideas: #1

Interdisciplinary Data Science teams.

My experience, along with DJ Patil's piece on Building Data Science teams, highlights the growing importance of interdisciplinary teams. The students who showed up to take this class are from across various disciplines.

I want you to build upon your individual strengths, as well as find ways to effectively collaborate with those who have complementary strengths, as this will be a *critical component of your success going forward*.

Data Science

10 Important Ideas: #2

Democratization of Machine and Statistical Learning Algorithms

Machine Learning algorithms used to be primarily used in Computer Science and Statistics departments. Now with the proliferation of new types of data sets, these algorithms are starting to get used across academic disciplines and within companies across sectors. With this democratization, it becomes imperative that those ***using the algorithms understand their meaning and potential impact***.

Data Science

10 Important Ideas: #3

Build a solid foundation of good coding practices

Data scientists need a solid foundation in writing code, and coding practices such as paired programming, code reviews, debugging, and version control.

Data Science

10 Important Ideas: #4

Data Strategy

For data scientists taking leadership positions in start-ups and industry, ***thinking in terms of a data strategy is a useful paradigm***. Data strategy involves figuring out what data to collect or log, how to store it, legal and space constraints, the pipelines that are built on top of it, how data will be used as part of the company's core business and how decisions will be made from data.

Data Science

10 Important Ideas: #5

Little Data

In addition to working with massive data sets and the engineering and infrastructure that's been built to analyze and process that, we also still work with Little Data. Oftentimes we sample from Big Data, which creates a Little Data set which can be used to explore and prototype.

Data Science

10 Important Ideas: #6

The Space between the Data Set and the Algorithm

Many people go straight from a data set to applying an algorithm. But there's a huge space in between of important stuff. It's easy to run a piece of code that predicts or classifies. That's not the hard part. The hard part is doing it well.

Data Science

10 Important Ideas: #7

Being Human

We now have tons of data on user (human) behavior. The data scientist brings with her not just a set of machine learning tools, but her humanity to interpret and find meaning in data, and make ethical data-driven decisions.

Data Science

10 Important Ideas: #8

Causation or Causality, Correlation and Experiments

Related to the fact that Little Data is still important, so are the classical statistical concepts of causation, correlation and experiments. Experiments and causal inference (e.g. propensity score modeling) are important parts of an engineer's and statistician's tools. We'll be exploring these more…

Data Science

10 Important Ideas: #9

Feedback Loop

The data generated by user behavior becomes the building blocks of data products which simultaneously are used by users and influence user behavior. We see this in recommendation systems, ranking algorithms, friend suggestions, etc. And we will see it increasingly across sectors including education, finance, retail and health.

Data Science

10 Important Ideas: #10

Causing the Future

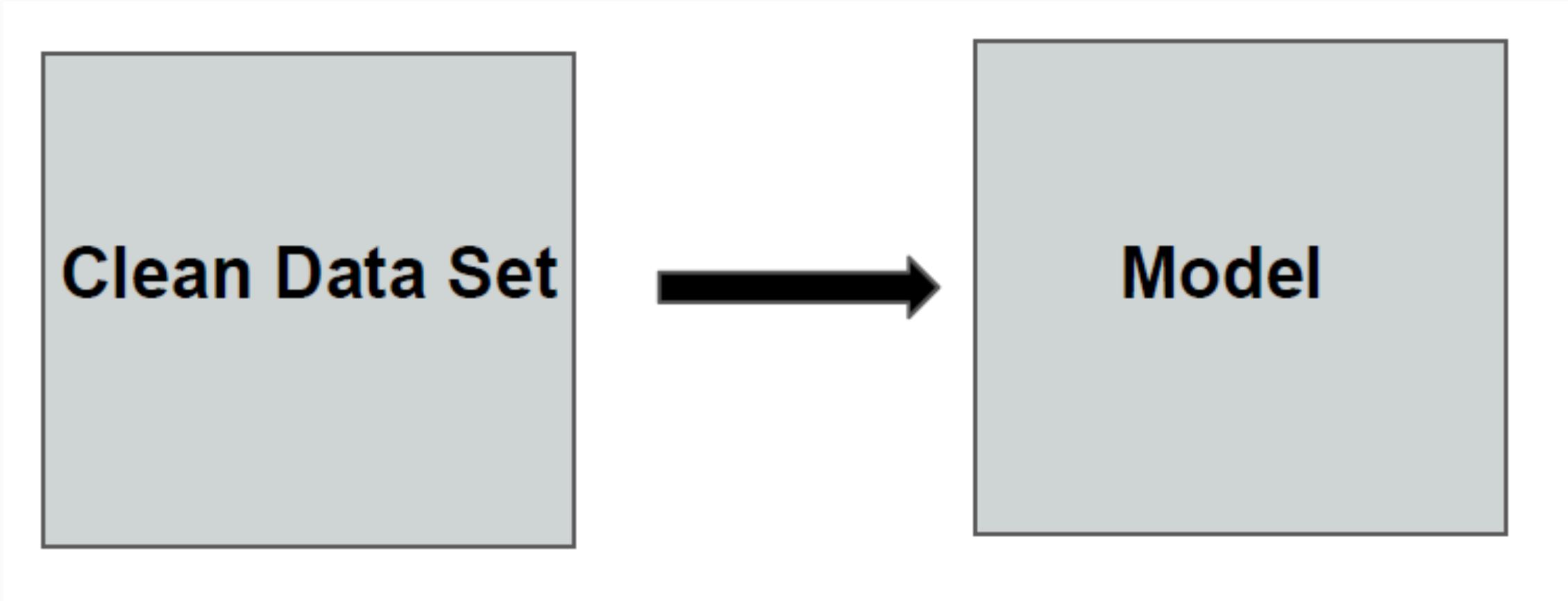
Prediction and **Causation** are two important themes in statistics, machine learning and data science. Much is made about Predicting the Future (see Nate Silver), Predicting the Present (see Hal Varian), and exploring Causal relationships from observed data (the Past) (see Sinan Aral). The next logical concept then is: models and algorithms ***not only capable of Predicting the Future, but also of Causing the Future***.

Doing Data Science

What do data scientists do?....

Doing Data Science

The way it's usually taught



Doing Data Science

Ted Johnson

- Assemble an accurate and relevant data set
- Choose the appropriate algorithm

Doing Data Science

Jim Gray

- Capture
- Curate
- Communicate

Doing Data Science

Colin Mallows

- Identify data to collect and its relevance to your problem
- Statistical specification of the problem
- Method selection
- Analysis of method
- Interpret results for non-statisticians

Doing Data Science

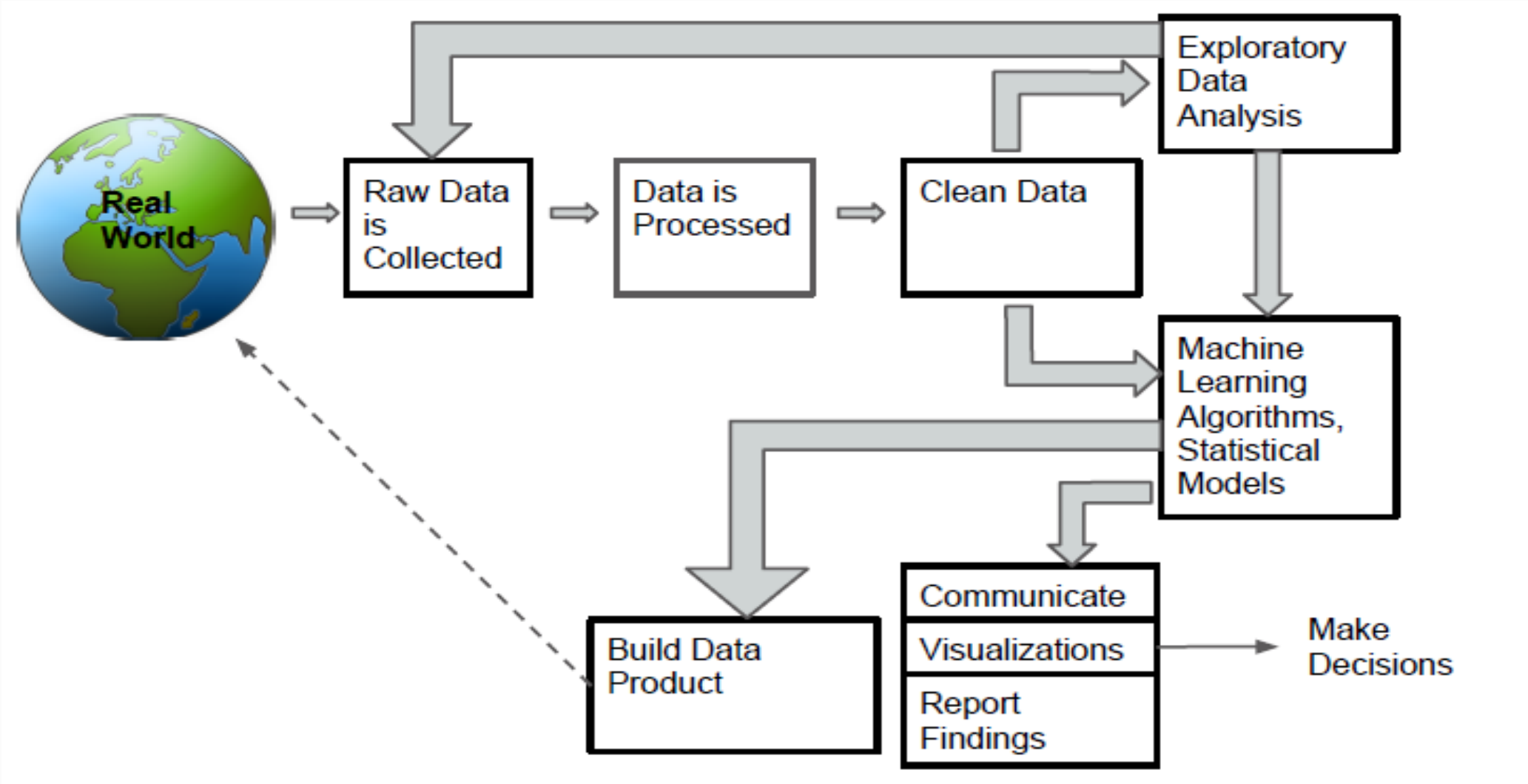
Ben Fry

- Acquire
- Parse
- Filter
- Mine
- Represent
- Refine
- Interact

Ben Fry's dissertation on Computational Information Design (2004)

Doing Data Science

My perspective…



Doing Data Science

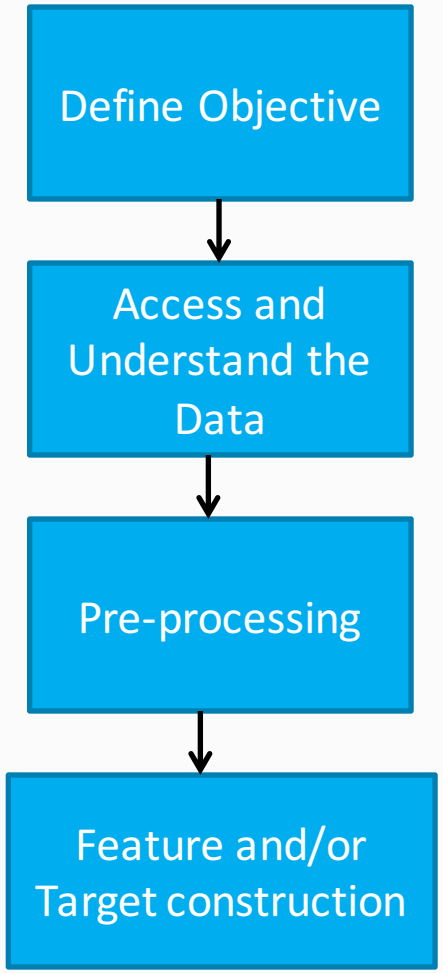
My Process Model

- Identify problem and data proxy that can be measured
- Instrument data sources
- Collect data
- Prepare data (integrate, transform, clean, impute, filter, aggregate)
- Build model
- Evaluate model
- Rinse, lather, repeat

See http://www.dataists.com/

Doing Data Science

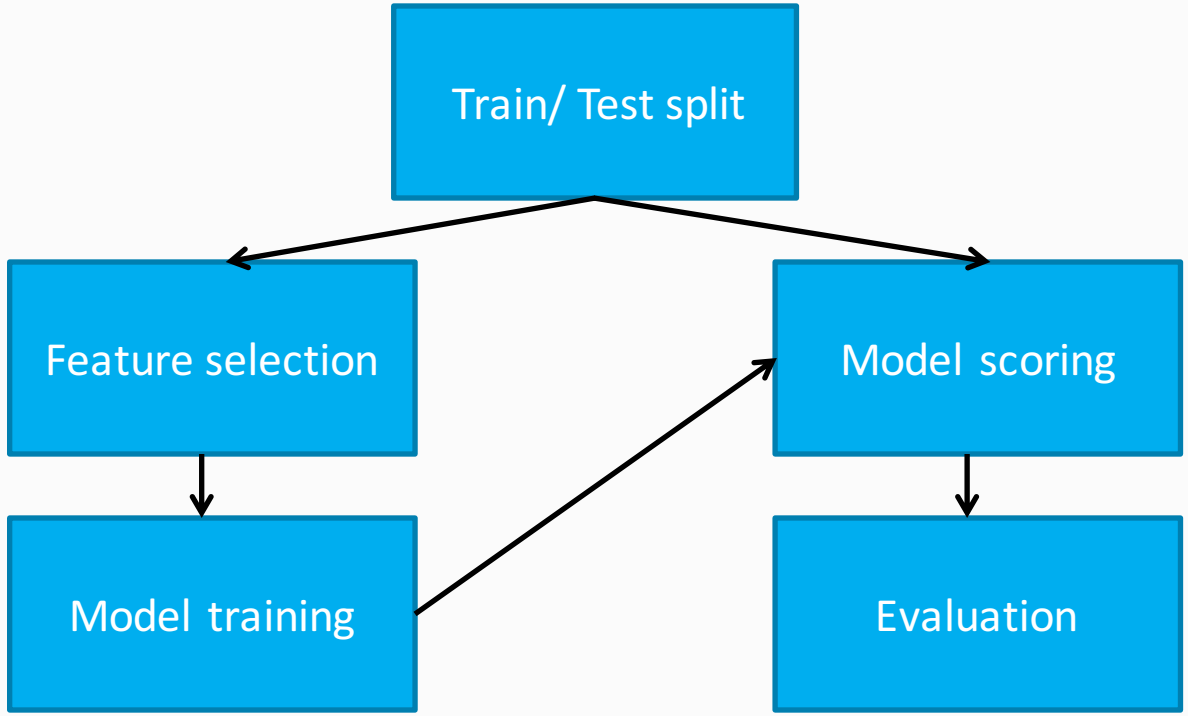
My Process Model



1. Define the objective and quantify it with a metric – optionally with constraints, if any. This typically requires domain knowledge.
2. Collect and understand the data, deal with the vagaries and biases in the data acquisition (missing data, outliers due to errors in the data collection process, more sophisticated biases due to the data collection procedure etc
3. Frame the problem in terms of a machine learning problem – classification, regression, ranking, clustering, forecasting, outlier detection etc. – some combination of domain knowledge and ML knowledge is useful.
4. Transform the raw data into a "modeling dataset", with features, weights, targets etc., which can be used for modeling. Feature construction can often be improved with domain knowledge. Target must be identical (or a very good proxy) of the quantitative metric identified step 1.

Doing Data Science

My Process Model

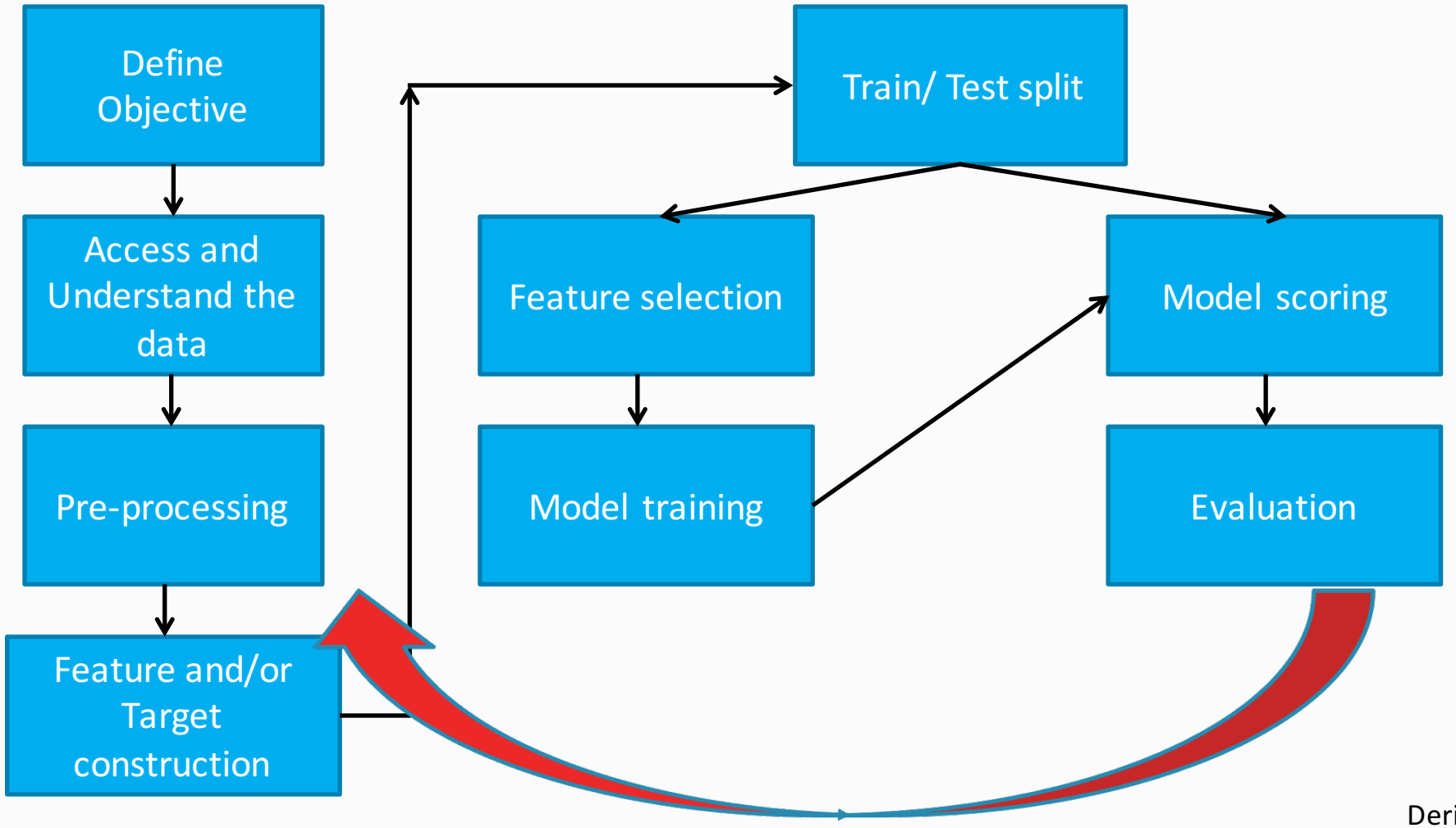


5. Train, test and evaluate, taking care to control bias/variance and ensure the metrics are reported with the right confidence intervals (cross-validation helps here), be vigilant against target leaks (which typically leads to unbelievably good test metrics) – this is the ML heavy step.

Doing Data Science

My Process Model

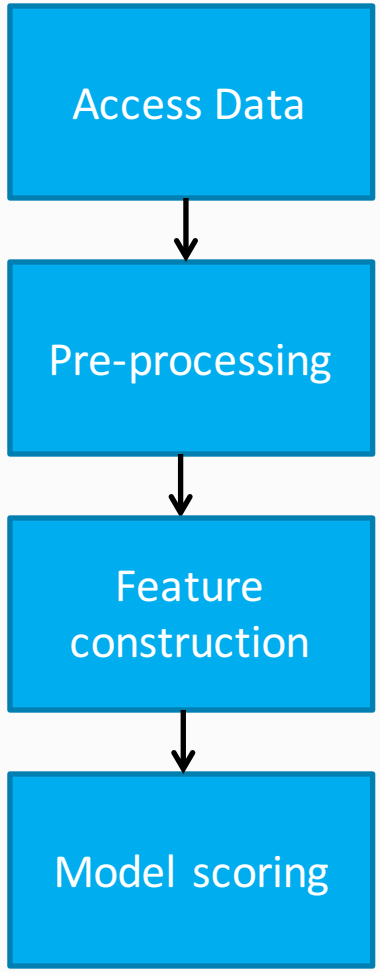
6. Iterate steps (2) – (5) until the test metrics are satisfactory



Doing Data Science

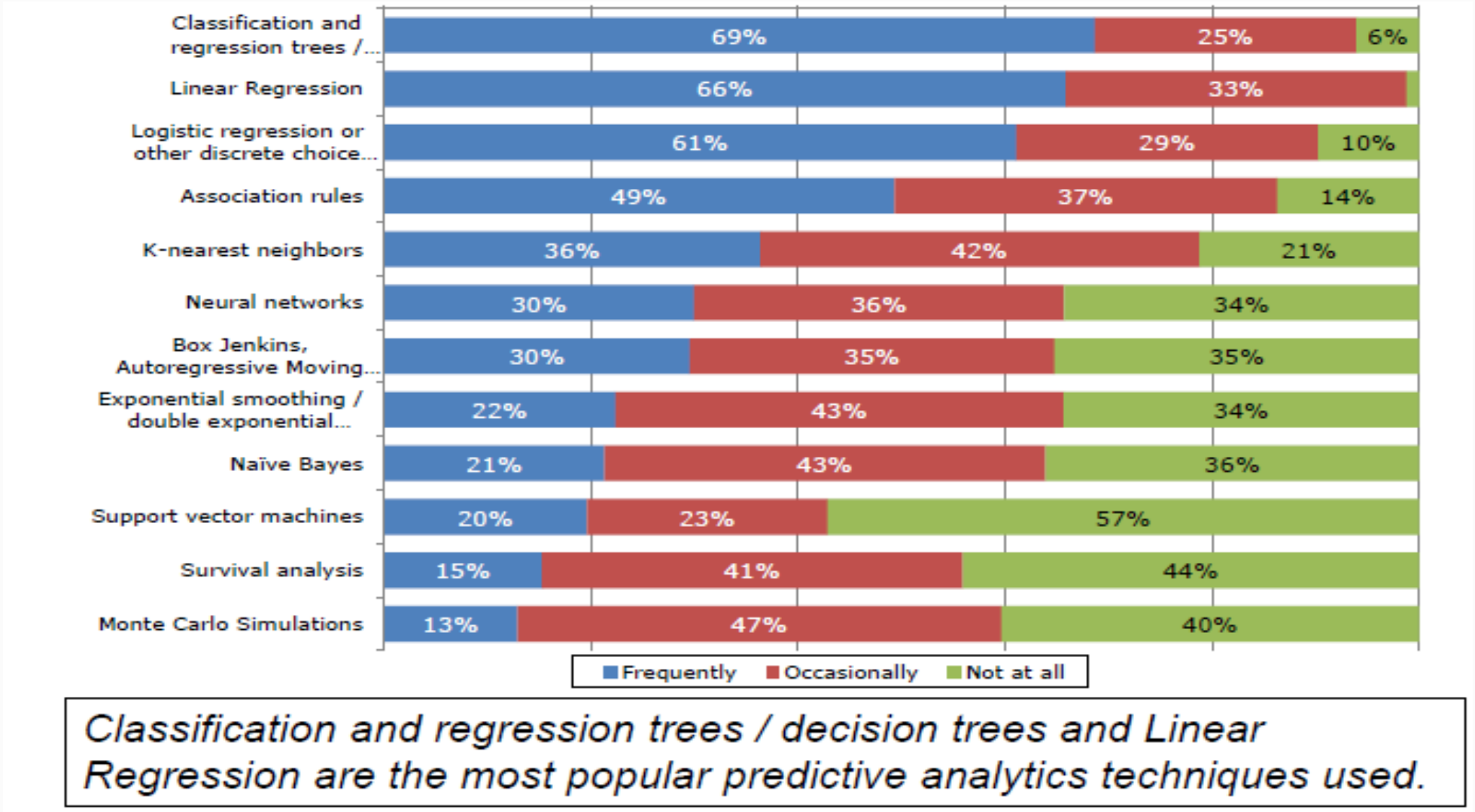
My Process Model

7. Deploy the trained model, ensure that the model is reproduced faithfully in the operationalization environment – monitor the data distributions (both model inputs and outputs), alert if there are large deviations from those seen in training, re-train if necessary – can be made easy with tooling



Organizations Employ a Variety of Predictive Analytics Algorithms

Machine Learning Lectures on Top Techniques

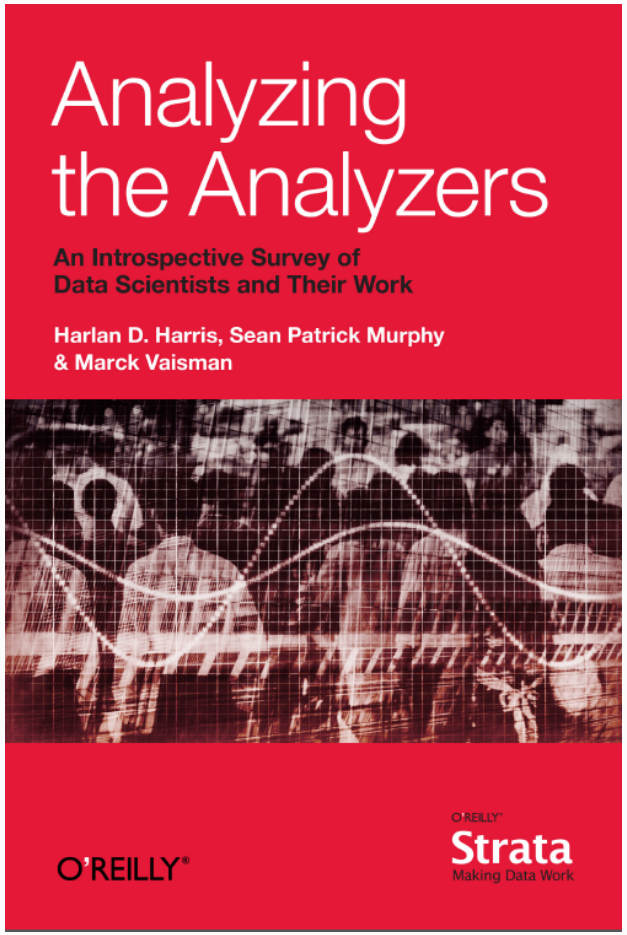


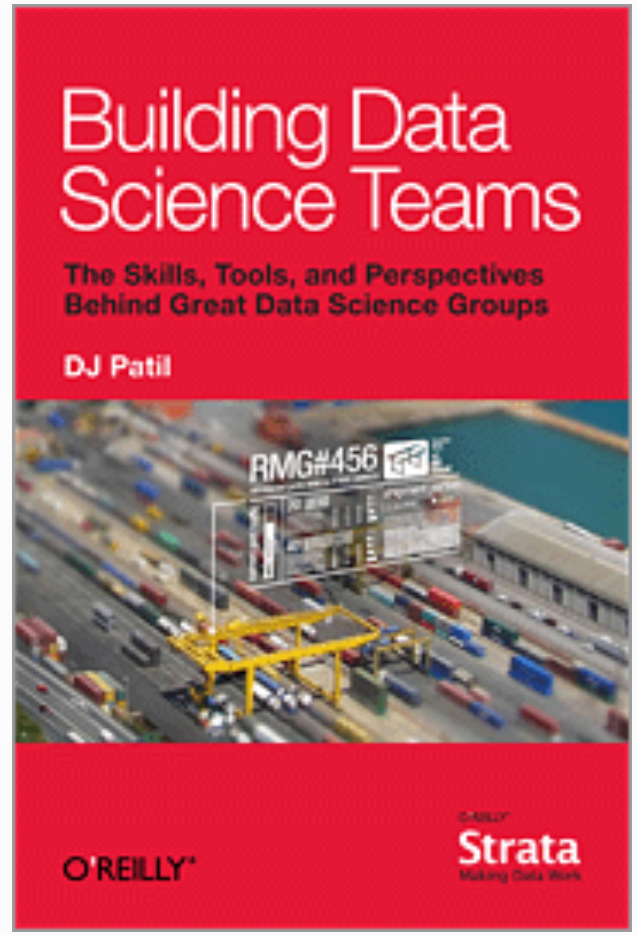
Classification and regression trees / decision trees and Linear Regression are the most popular predictive analytics techniques used.

Break,10 minutes…

Out of Class Reading

Week One





How to be Successful

1. Create value

2. Capture some for yourself

How to create value (as a data scientist)

- Extract insights from data for decision support

Productive Use of Time

1. Have a bias *against* writing learning algorithms
   - Bias *in favor of* leveraging 3rd party implementations

Productive Use of Time

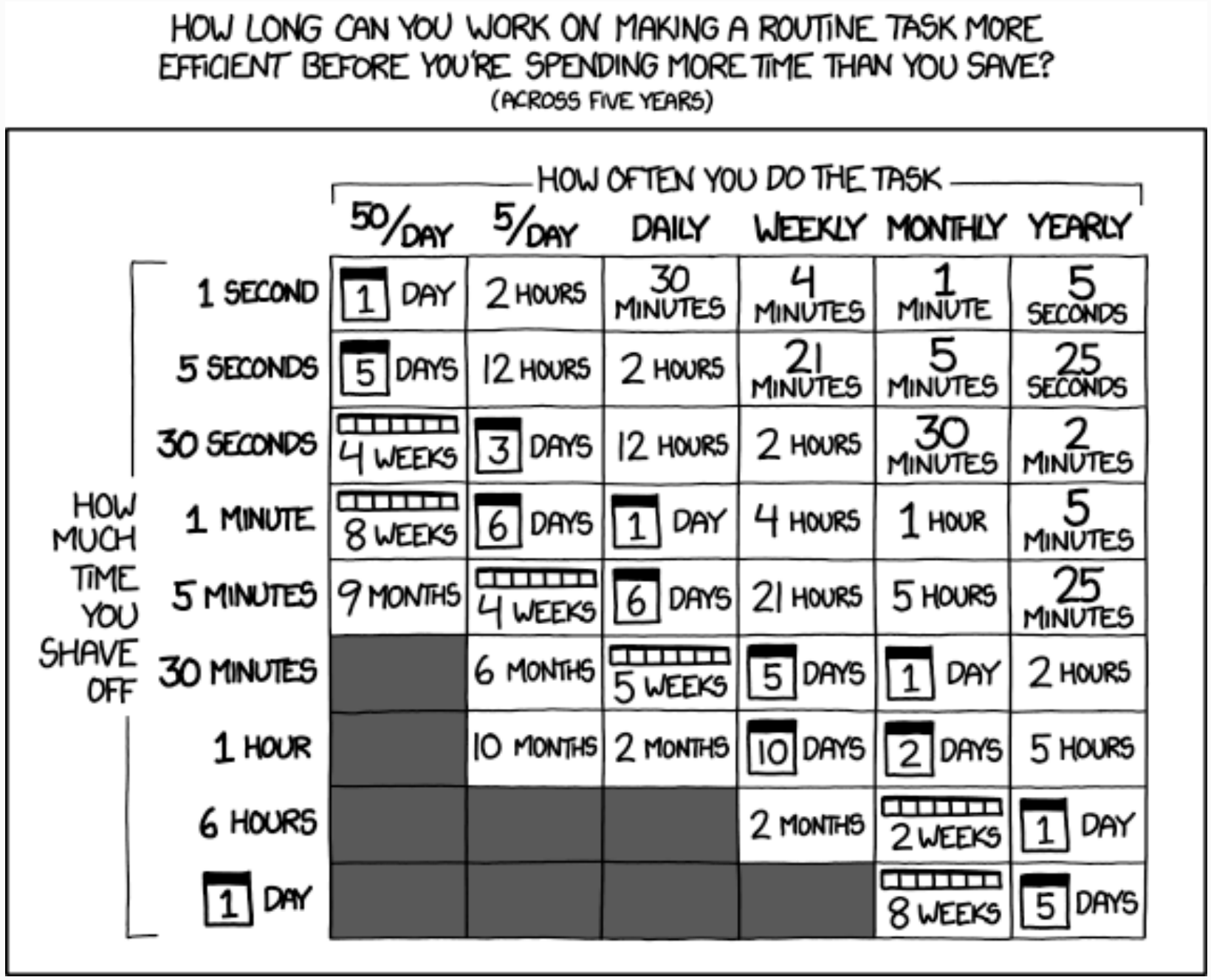
1. Have a bias *against* writing learning algorithms
   - Bias *in favor of* leveraging 3rd party implementations
   - Add data: *more information beats better algorithms*

Productive Use of Time

1. Have a bias *against* writing learning algorithms
   - Bias *in favor of* leveraging 3rd party implementations
   - Add data: *more information beats better algorithms*
2. You *will* write data manipulation algorithms

Productive Use of Time

1. Have a bias *against* writing learning algorithms
   - Bias *in favor of* leveraging 3rd party implementations
   - Add data: *more information beats better algorithms*
2. You *will* write data manipulation algorithms
   - Data is surprising enough, need algorithm certainty
   - Defect count is proportional to line count
   - *Use as high level a language as possible*


HOW LONG CAN YOU WORK ON MAKING A ROUTINE TASK MORE EFFICIENT BEFORE YOU'RE SPENDING MORE TIME THAN YOU SAVE?
(ACROSS FIVE YEARS)
HOW OFTEN YOU DO THE TASK
50/DAY
5/DAY
DAILY
WEEKLY
MONTHLY
YEARLY
HOW MUCH TIME YOU SHAVE OFF
1 SECOND
1 DAY
2 HOURS
30 MINUTES
4 MINUTES
1 MINUTE
5 SECONDS
5 SECONDS
5 DAYS
12 HOURS
2 HOURS
21 MINUTES
5 MINUTES
25 SECONDS
30 SECONDS
4 WEEKS
3 DAYS
12 HOURS
2 HOURS
30 MINUTES
2 MINUTES
1 MINUTE
8 WEEKS
6 DAYS
1 DAY
4 HOURS
1 HOUR
5 MINUTES
5 MINUTES
9 MONTHS
4 WEEKS
6 DAYS
21 HOURS
5 HOURS
25 MINUTES
30 MINUTES
6 MONTHS
5 WEEKS
5 DAYS
1 DAY
2 HOURS
1 HOUR
10 MONTHS
2 MONTHS
10 DAYS
2 DAYS
5 HOURS
6 HOURS
2 MONTHS
2 WEEKS
1 DAY
1 DAY
8 WEEKS
5 DAYS

Analysis and Diminishing Returns

1. First few models tend to capture most of the value

Analysis and Diminishing Returns

1. First few models tend to capture most of the value
2. Distinguish between:
   - Marginal improvements important (e.g., search, WalMart);
   - Marginal improvements unimportant (typical).

Analysis and Diminishing Returns

1. First few models tend to capture most of the value
2. Distinguish between:
   - Marginal improvements important (e.g., search, WalMart);
   - Marginal improvements unimportant (typical).
3. Latter case: get first 80% and move on to new problem

The Importance of Starting Small

The Importance of Starting Small

1. When you first encounter a data set, you know *nothing*.
   - Ergo: first piece of data is *very informative.*
   - Think of data set utility as roughly logarithmic in size.

The Importance of Starting Small

1. When you first encounter a data set, you know *nothing*.
   - Ergo: first piece of data is *very informative.*
   - Think of data set utility as roughly logarithmic in size.

2. Don’t require a large data set before starting analysis.

The Importance of Starting Small

1. When you first encounter a data set, you know *nothing*.
   - Ergo: first piece of data is *very informative.*
   - Think of data set utility as roughly logarithmic in size.

2. Don't require a large data set before starting analysis.

3. Always try things out on small portions of data first.

Timescales and Failing Fast

1. Immediate zone: less than 60 seconds
   - 100s per day
2. Bathroom break zone: less than 5 minutes
   - 10s per day
3. Lunch zone: less than an hour
   - 5 per day
4. Overnight zone: less than 12 hours
   - 1 per day

Timescales and Failing Fast

1. Immediate zone: less than 60 seconds
   - 100s per day
2. Bathroom break zone: less than 5 minutes
   - 10s per day
3. Lunch zone: less than an hour
   - 5 per day
4. Overnight zone: less than 12 hours
   - 1 per day

Timescales and Failing Slow

1. Immediate zone: less than 60 seconds
   - 100s per day
2. Bathroom break zone: less than 5 minutes
   - 10s per day
3. Lunch zone: less than an hour
   - 5 per day
4. Overnight zone: less than 12 hours
   - 1 per day

Sublinear Debugging

Print intermediate information sufficient to terminate bad computations early.

Other Speed Tricks

1. Move code to data, not the converse!
2. Do feature engineering with a fast learning algorithm (e.g., linear), then switch to a slower algorithm for the final product (e.g., GBDT, NN).
3. Subsample your data intelligently.
    1. Less examples (rows), e.g., imbalanced classification.
    2. Less features (columns), e.g., random projections, SVD.

Failing Fast: Summary

Productivity demands debugging as fast as possible.

Stay in the immediate zone.

Proxy Metrics

Proxy Metric: Something you can measure and optimize

- Revenue per impression
- Clickthrough rate
- Reciprocal communication rate
- Polling results
- Gene expression levels
- Value at risk

~~Proxy Metrics~~ Reality

Reality: Something you actually care about

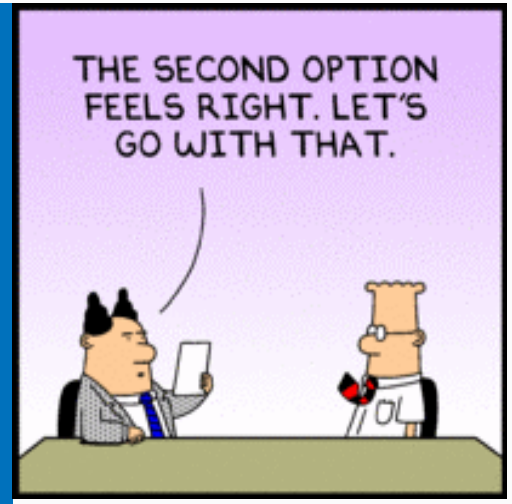
- ~~Revenue per impression~~ → Economic Value Created
- ~~Clickthrough rate~~ → User Experience Quality
- ~~Reciprocal communication~~ rate → Match Quality
- ~~Polling results~~ → Election Outcome
- ~~Gene expression levels~~ → Drug Efficacy in Vivo
- ~~Value at risk~~ → Portfolio Quality

Proxy Metrics vs. Reality

Key Takeaways

- Think about your data, not about your software.
- Productivity is about not waiting for answers.
- Mind the gap (between proxy metrics and reality).
  - Best defense: close collaboration with a business expert.

Causal Analysis in Online Display Advertising



Dilbert

The Life of a Browser Process.

1. Observe people taking actions and visiting content

2. Use observed data to build list of prospects

3. Subsequently observe same browser surfing the web the next day

4. Browser visits a site where a display ad spot exists and bid requests are made

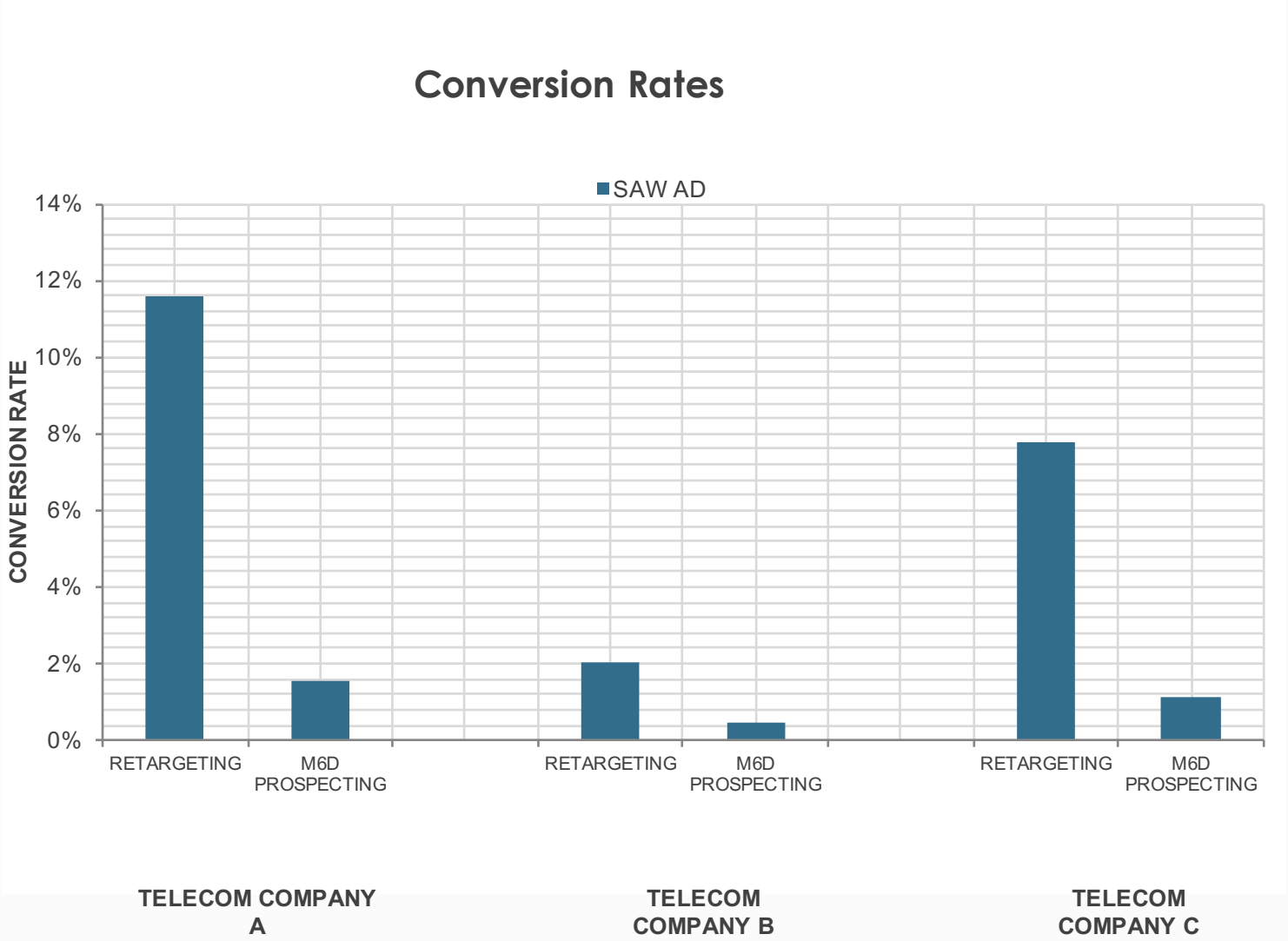
5. Auction is held for display spot

6. If auction is won display the ad

7. Observe browsers actions after displaying the ad

What Do Advertisers Want?

Conversions?



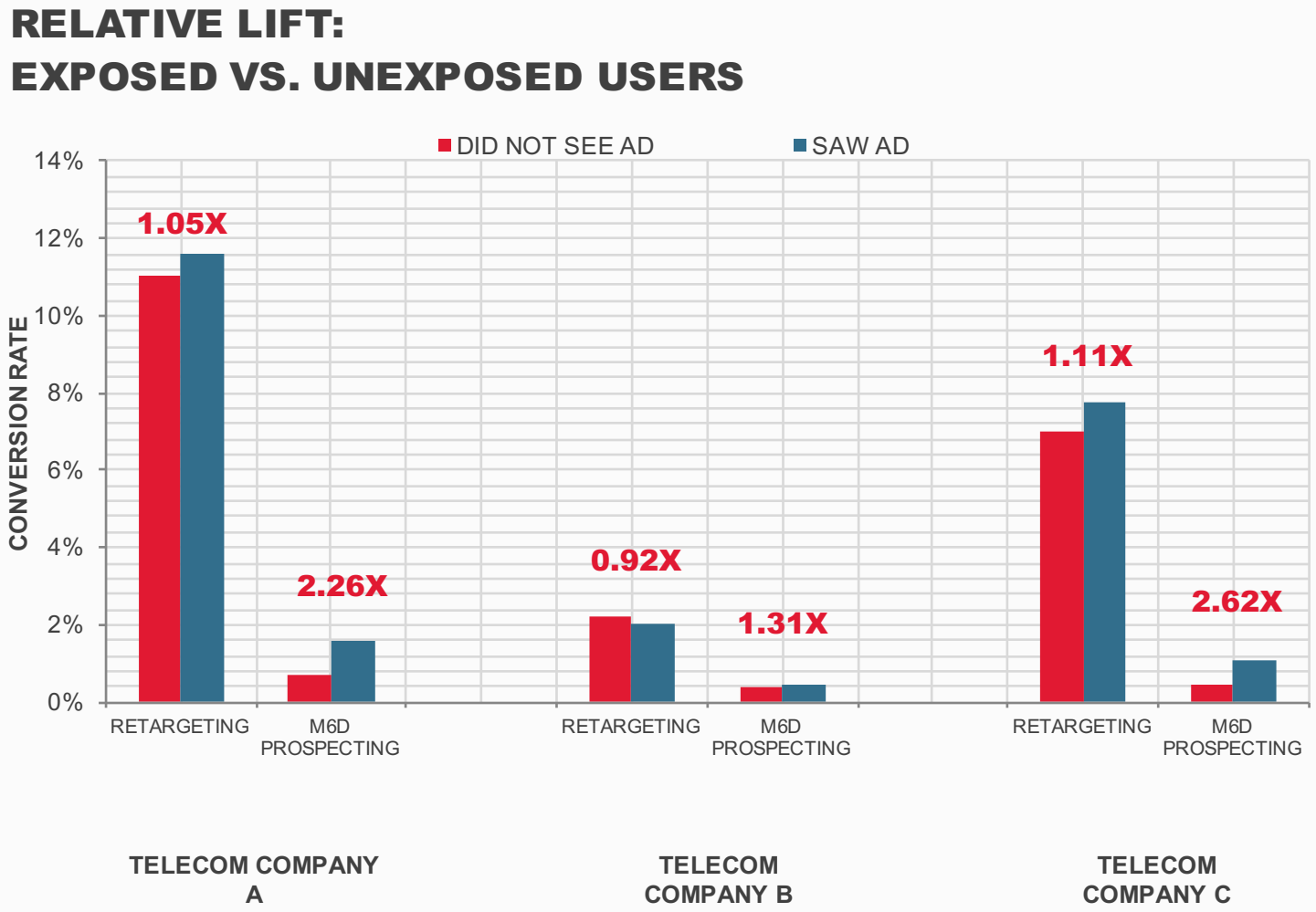
Three different telecoms;

Raw conversion deceiving, connecting data to business value);

What is the effectiveness of the add?

What Do Advertisers Want?

Conversions?



Question of Interest.

What is the **causal** effect of
display advertising
on customer conversion?

display advertising
Showing/Not showing a browser a display ad.

customer conversion
Visiting the advertisers website in the next 5 days.

General Approach.

$?$ 1. Ask the right question

P 2. Understand/express the causal process

$\Psi(P)$ 3. Translate question into a formal quantity

$\Psi(P_n)$ 4. Try to estimate it

1. state question.

What is the effect of display advertising on customer conversion?

display advertising

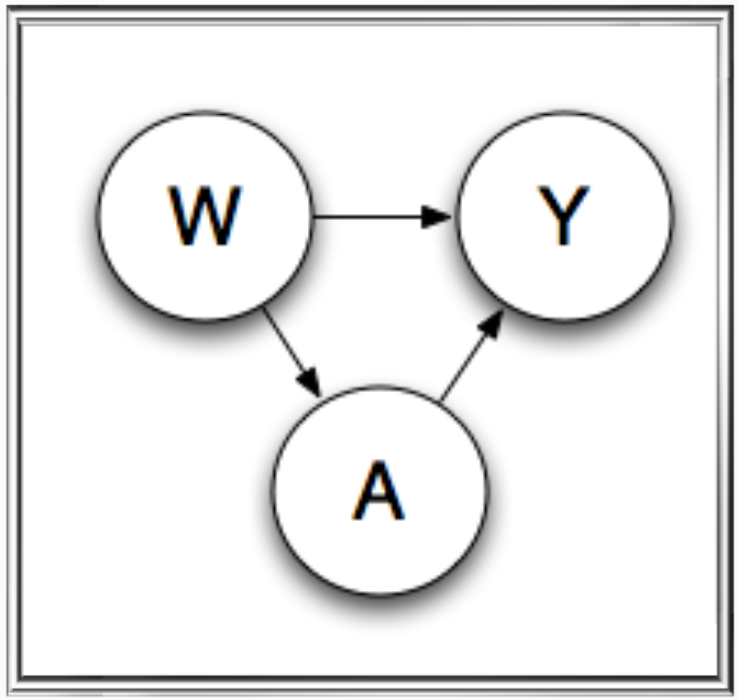
Showing/Not showing a browser a display ad.

customer conversion

Visiting the advertisers website in the next 5 days.

P

2. express causal process.



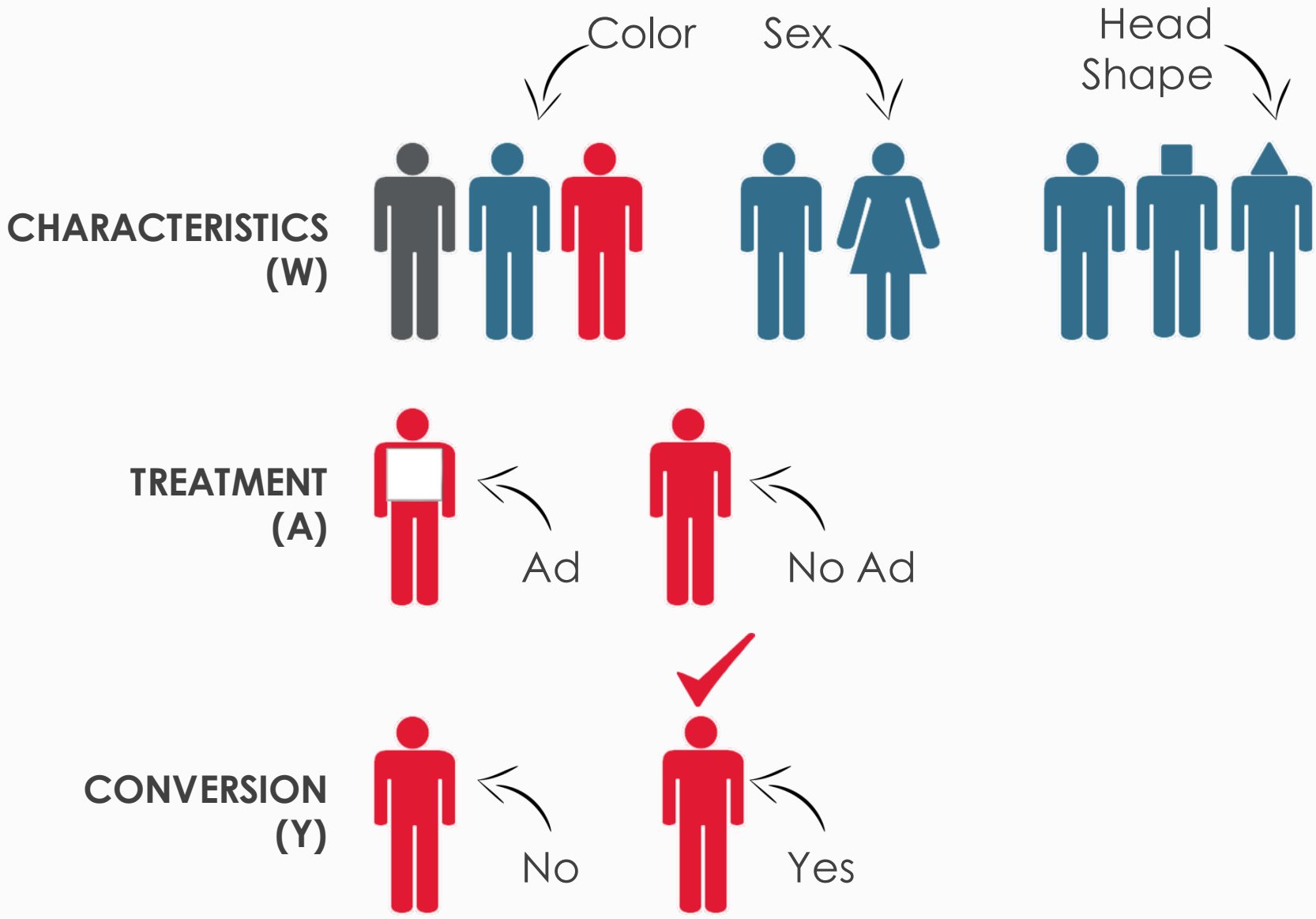
O = (W,A,Y) ~ P_0

W – Baseline Variables

A – Binary Treatment (Ad)

Y – Binary Outcome (Purchase)

Data Structure: Our Viewers.



$\Psi(P)$

3. define quantity.

Additive Impact

$$E[Y_{A=ad}] - E[Y_{A=no\ ad}]$$

Relative Impact

$$E[Y_{A=ad}]/E[Y_{A=no\ ad}]$$

$\Psi(P_n)$ 4. estimate quantity.

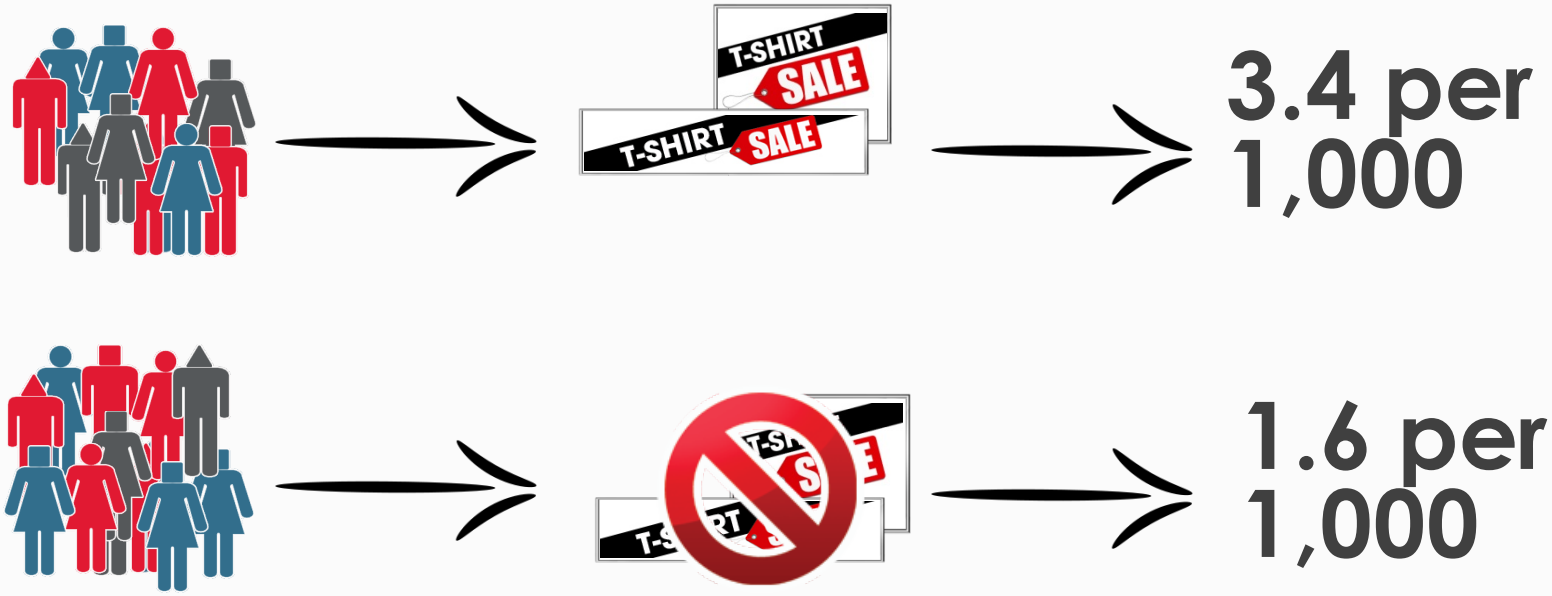
1. A/B testing

2. Modeling Observational Data

Common Approach: A/B Testing

Hard to get right…

Since we can not both treat and not treat the SAME individuals. Randomization is used to create "EQUIVALENT" groups to treat and not treat.



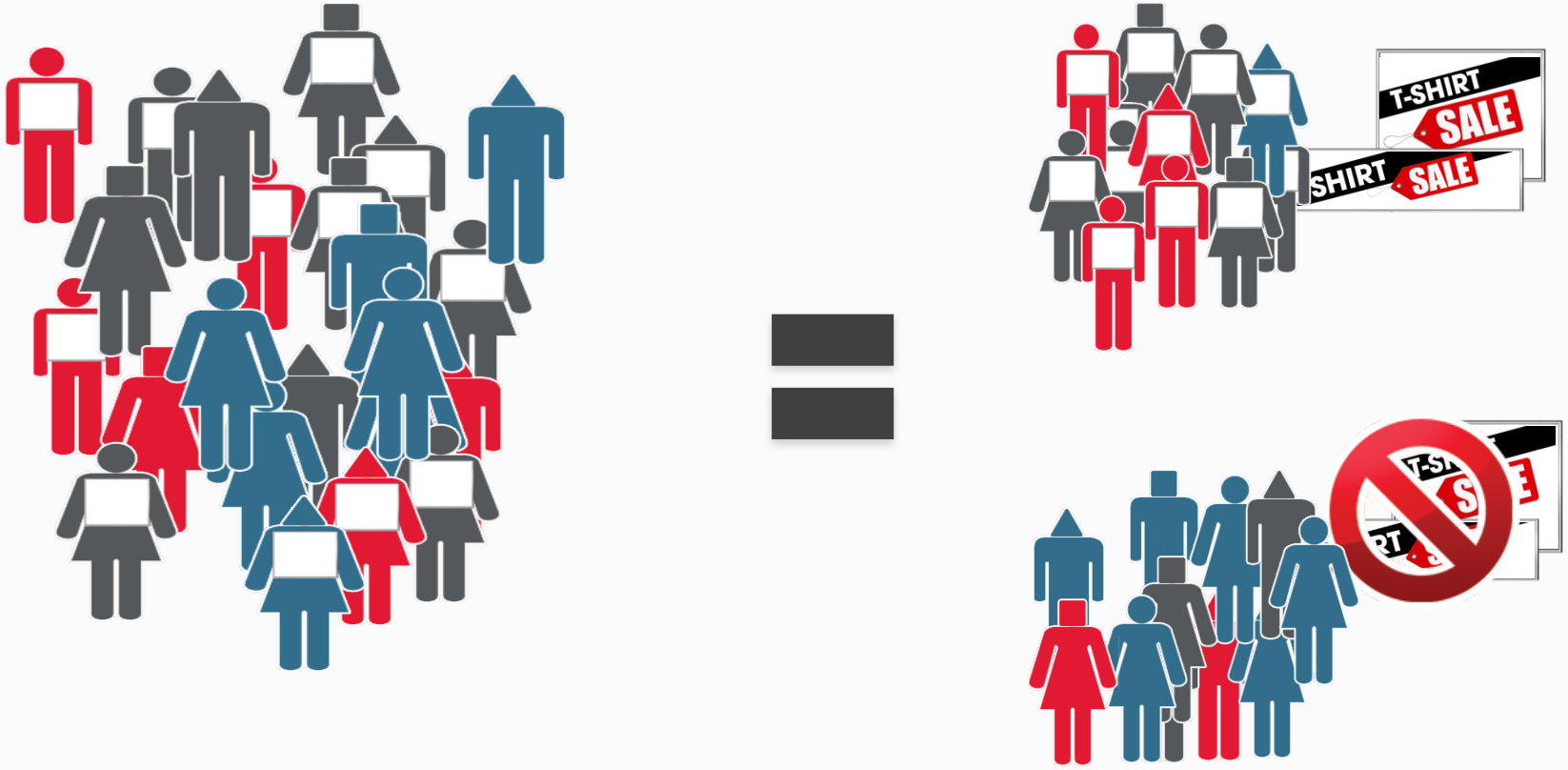
Practical Concerns.

Associated with Doing A/B Testing

1. Cost of displaying PSAs to the control (untreated group).
2. Overhead cost of implementing A/B test and ensuring that it is done **CORRECTLY**.
3. **Wait time** necessary to evaluate the results.
4. No way to analyze past or completed campaigns.

Non Invasive Causal Estimation (NICE).

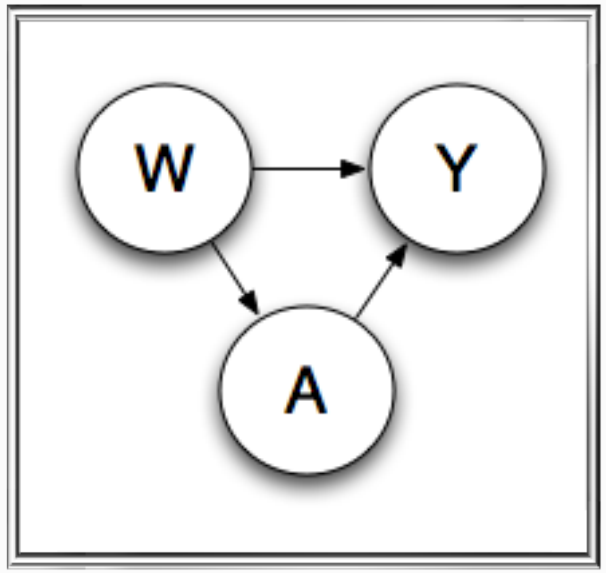
Estimate The Effects in the Natural Environment (Observed Data)



Use the results of a normal campaign. Red people don't convert so unlikely to see ad.
Blue and Grey with round heads are good converters so more likely to see advertisements.
So we have a bias in the presentation and hence the results

“what if”

causal analysis adjusting for confounding



Need to adjust for the fact that the group that saw the advertisement and the group that didn’t may be very different.

Estimation – A Primer.

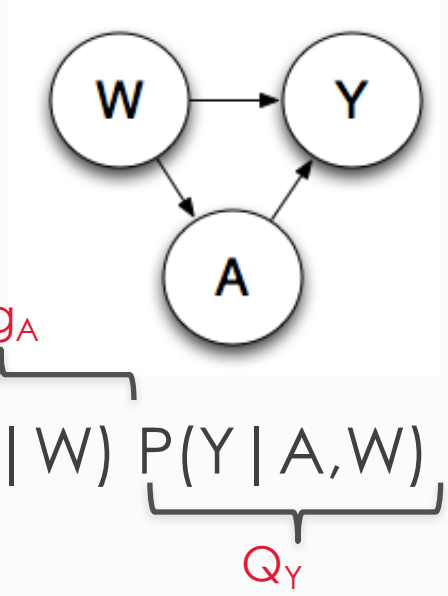
1. When can we estimate it? Necessary conditions:
   - no unmeasured confounding (need to account for all)
   - experimental variability/positivity (present to all groups)

2. Be VERY careful with data collection
   - Define cohorts and follow them over time

3. Estimation techniques
   - Unadjusted
   - Adjust through g_A
   - MLE (max likelihood estimation) estimate of Q_Y
   - Double robust combining g_A and Q_Y
   - TMLE (targeted maximum likelihood estimation)

   Two are conditional probabilities…

4. Many tools exist for estimating **binary conditional distributions**
   - Logistic regression, SVM, GAM, Regression Trees, etc.



Summary

Techniques

- Overview
- Data Science Workflow
- Best Practices

Data Science Practice

- Introduction to Data Science
- Causal analysis in display advertising

Homework

- Reading