

Big Data

What it is and why you should care

David Levitan

Data Scientist, Microsoft Customer Data and Analytics

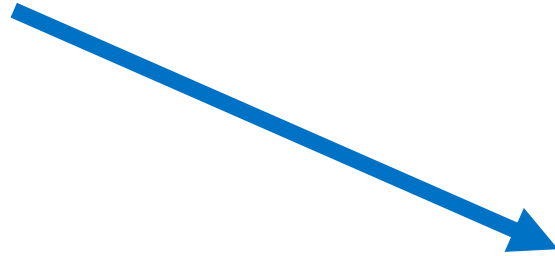
Big Data

What it is and why you should care
but why it's not the right thing to worry about

David Levitan

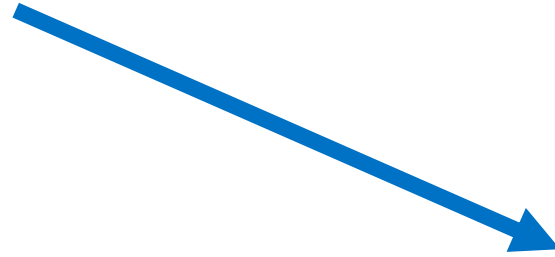
Data Scientist, Microsoft Customer Data and Analytics

Big Data



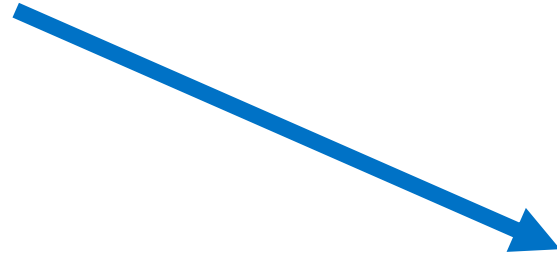
Big Impact

Big Data



Big Impact

Big Data



Big Impact

Big Data

Why?



Big Impact

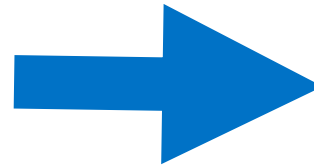
What is big data?

1995



Wikimedia

540MB Hard Drive
\$550 in today's
dollars



2015



WD Red 6TB NAS Hard Drive:
by Western Digital

\$249.00 ~~\$299.99~~ Prime

Get it by **Wednesday, Oct 7**

More Buying Choices

\$234.44 new (96 offers)

\$200.00 used (4 offers)

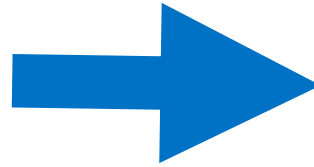
Amazon.com

20x the capacity
for half the cost

1996



ASCI Red Supercomputer
~1 TFLOPS
\$67 million dollars



2015



Xbox One
~1.3 TFLOPS
\$350

What is big data?

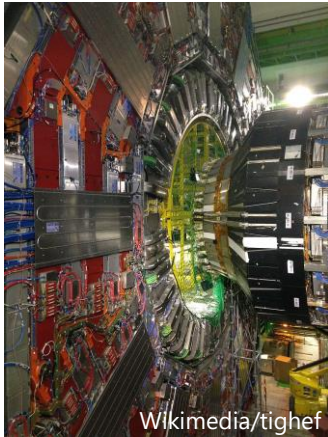
One definition

Enough data that you need to build a custom solution to store/manage/work with it

Currently, probably multiple petabytes

Who has that much data? Tech companies and scientists

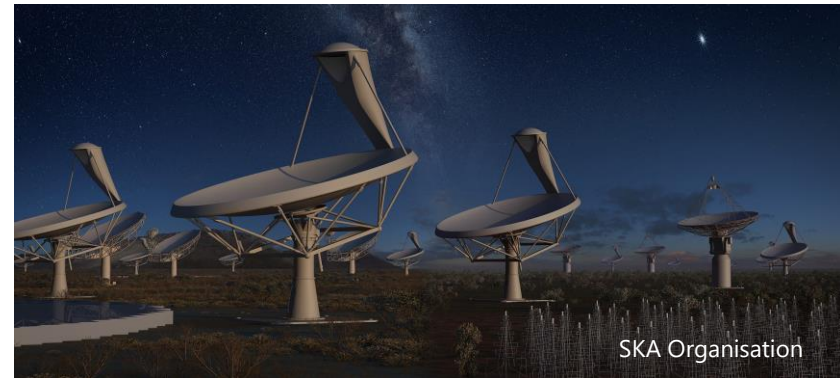
What is big data?



Large Hadron Collider

2010

~30TB/day



Square Kilometer Array

Mid-2020's

~30TB/(4 minutes)



Dan Ariely ✓

January 6, 2013 · 🌐

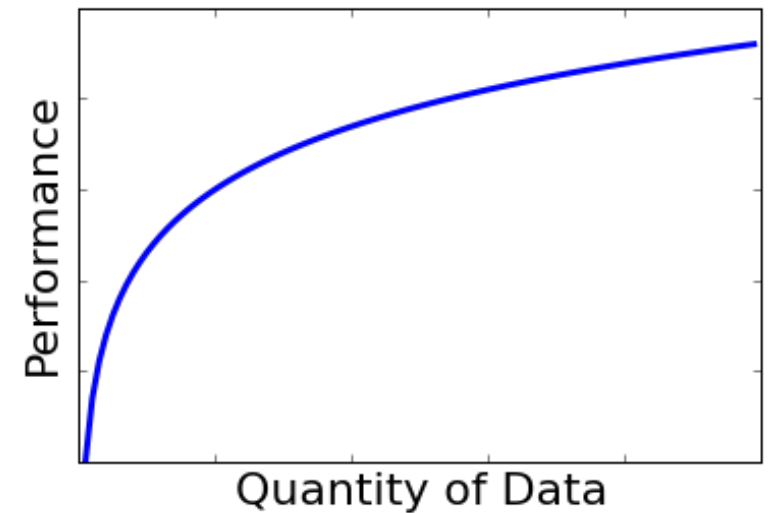
Big data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it...

2.2k Likes 104 Comments 1k Shares

Data size is not very important

Smaller data:

- is easier and cheaper to store
- is easier and faster to work with
- can often produce more value (particularly as return on investment) than larger data sets



Understand
Problem



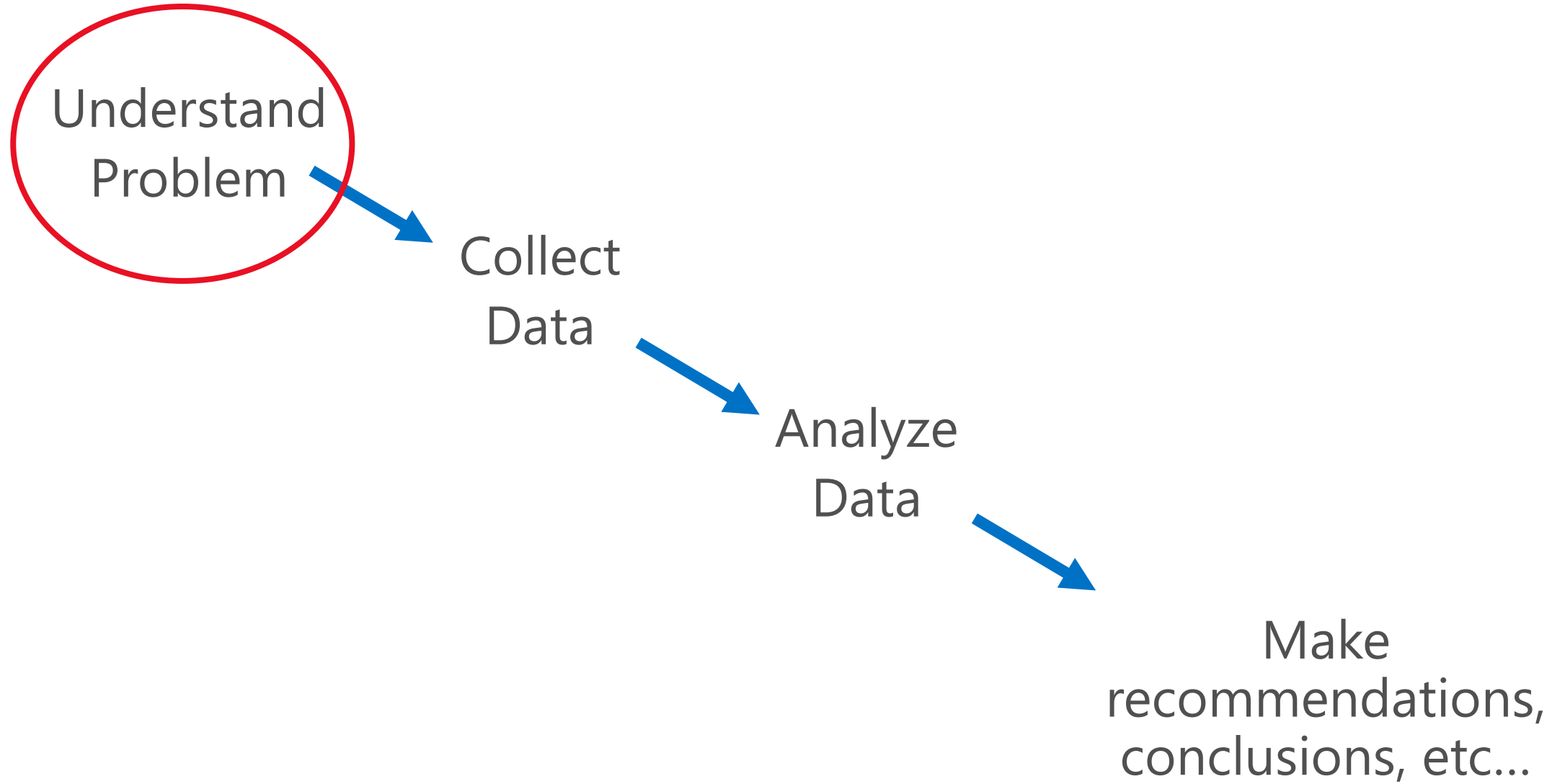
Collect
Data

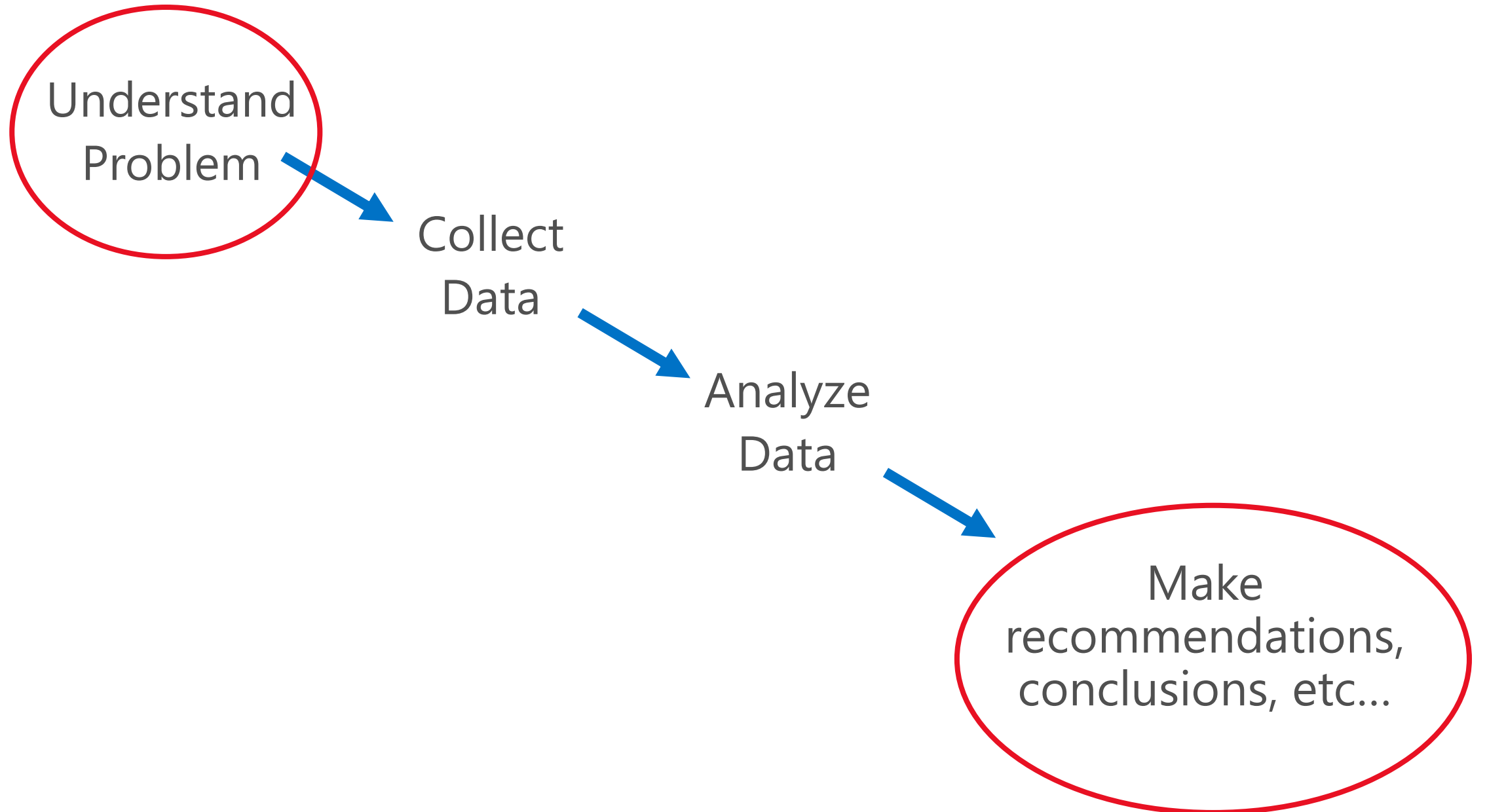


Analyze
Data



Make
recommendations,
conclusions, etc...





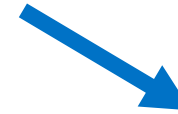
Understand
Problem



Collect
Data



Analyze
Data



Make
recommendations,
conclusions, etc...

Data analysis and collection is somewhat straightforward. Identifying and understanding the problem is much harder.

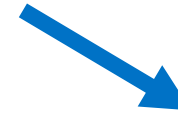
Understand
Problem



Collect
Data



Analyze
Data



Make
recommendations,
conclusions, etc...

Data analysis and collection is somewhat straightforward. Identifying and understanding the problem is much harder.

Technology is less important than your problem.

Understand your data



Google trends data



Two similar pieces of data may have vastly differently behavior. Look at your data to understand what is going on.

- Seasonality expected?
- Are there outliers? Why?
- Do you want to keep outliers?

Don't make assumptions

Oftentimes certain operations are done during data acquisition and/or initial processing. Understand what was done before you work with your data.

Don't make assumptions

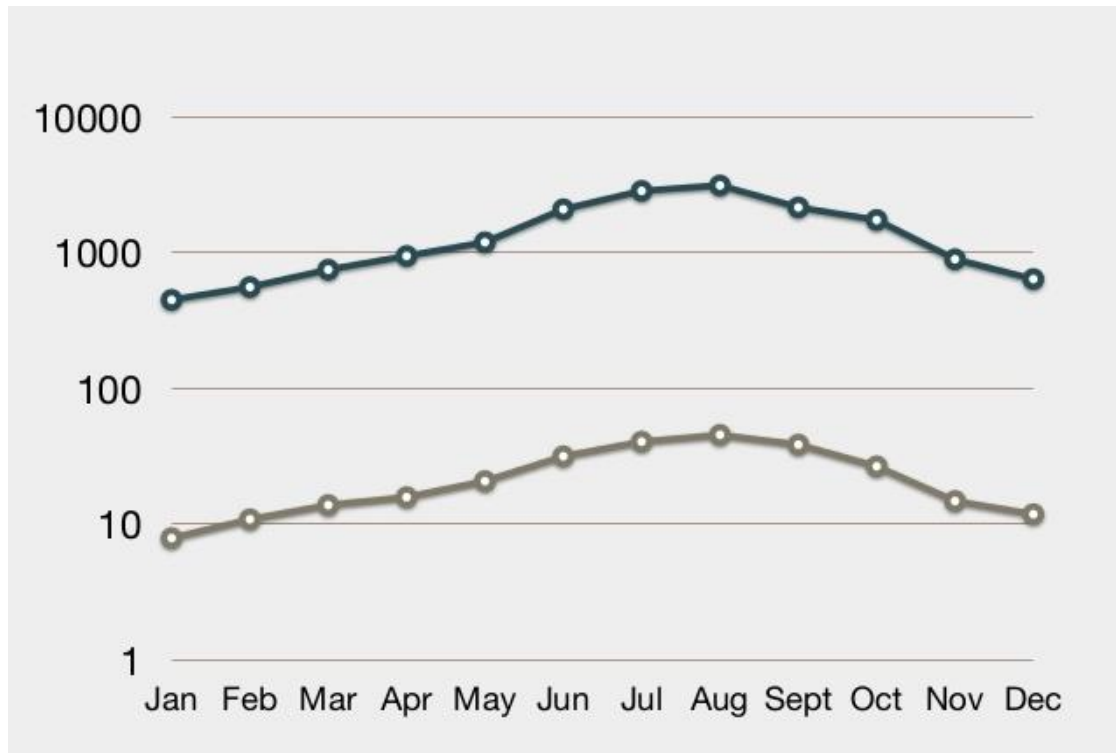
Oftentimes certain operations are done during data acquisition and/or initial processing. Understand what was done before you work with your data.



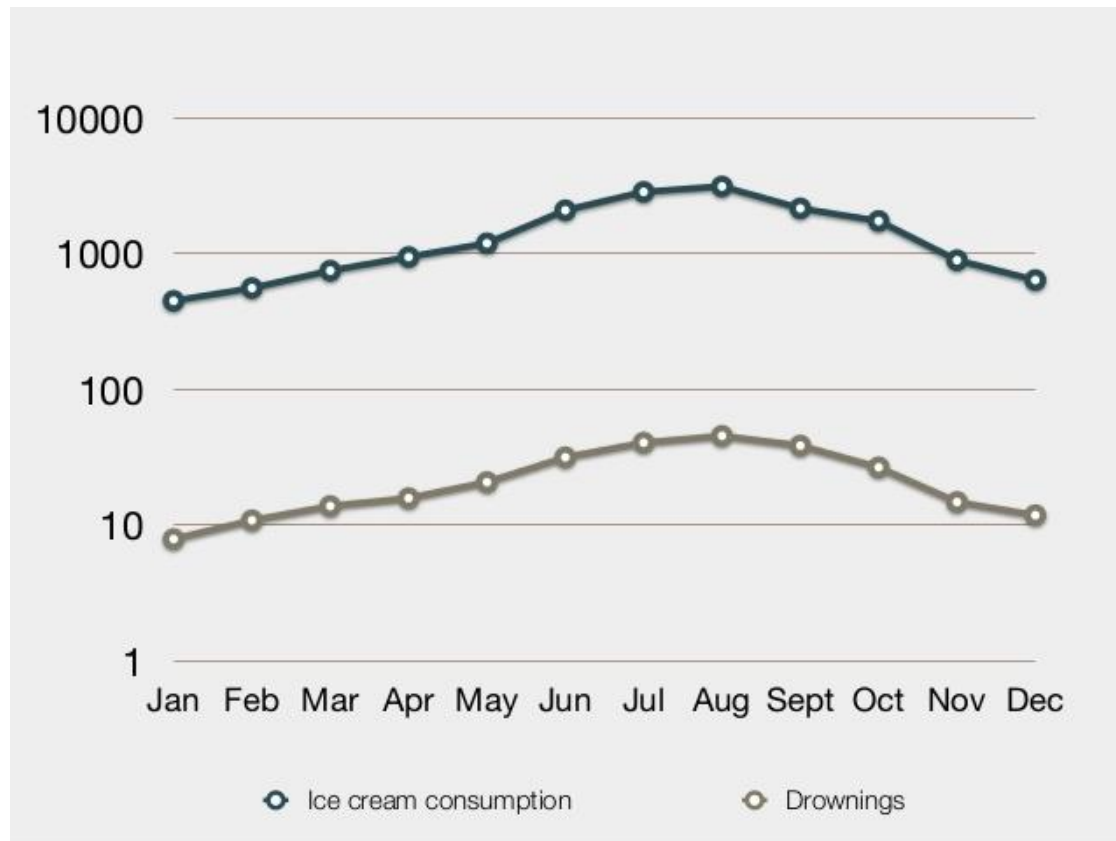
On July 23rd, 1983, an Air Canada flight ran out of fuel half way. Why?

Maintenance crew and pilots had assumed the plane was calibrated for imperial units as was typical for most planes of this type. Instead, it had been calibrated for metric.

Correlation does not imply causation

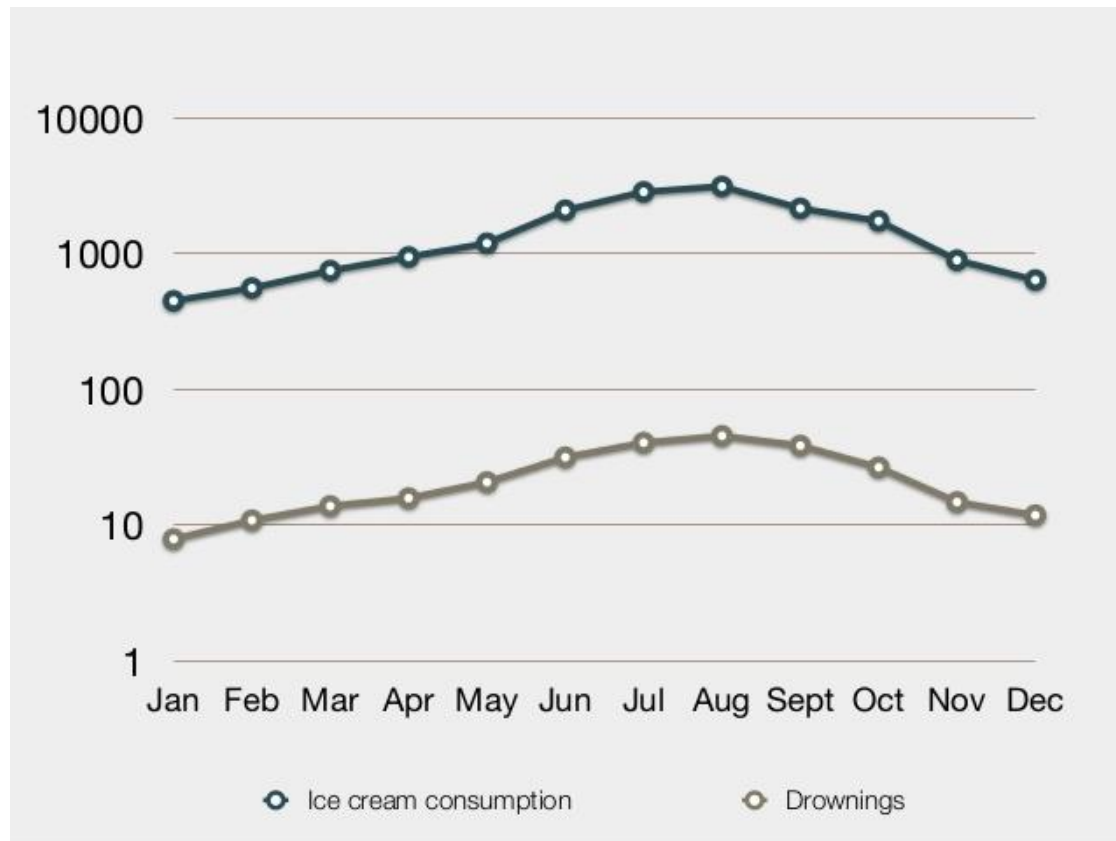


Correlation does not imply causation



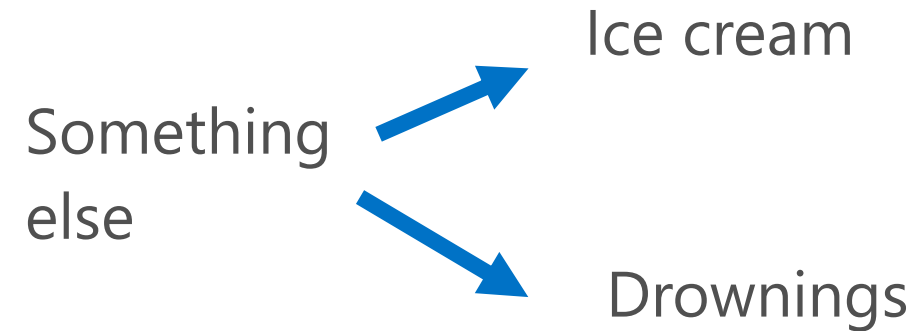
Ice cream  Drownings

Correlation does not imply causation



Ice cream → Drownings

OR



How to have a happy marriage...

In Sept. 2014, a paper titled “The Relationship between Wedding Expenses and Marriage Duration” by Andrew M. Francis and Hugo M. Mialon found that happiness in marriage was correlated with a big, cheap wedding.

How to have a happy marriage...

In Sept. 2014, a paper titled “The Relationship between Wedding Expenses and Marriage Duration” by Andrew M. Francis and Hugo M. Mialon found that happiness in marriage was correlated with a big, cheap wedding.



CNN News Video TV Opinions More... Search CNN

U.S. World Politics Tech Health Entertainment Living Travel **Money** Sports

ANORO[®] ELLIPTA[®] (lumecidinium 62.5 mcg and vilanterol 25 mcg inhalation powder)
Get your free 30-day trial*
Supported by GSK. ANORO ELLIPTA was developed in collaboration with THERAVANCE.
*Subject to eligibility. Restrictions apply. ANORO.com

Important Safety Information
■ ANORO ELLIPTA is only approved for use in COPD. ANORO is NOT approved for use in asthma.
■ People with asthma who take long-acting beta₂-adrenergic
Prescribing info & Boxed Warning, Medication Guide

Want a happy marriage? Have a big, cheap wedding

By Brandon Griggs, CNN
Updated 7:20 PM ET, Mon October 13, 2014

✉️ 📘 🐦 🗨️

More from CNN

Bobbi Kristina Brown undergoes a tracheostomy

Blogger's brutal death for speaking his mind

Unfortunately not quite that easy

Big, cheap wedding



Happy Marriage

The screenshot shows the top navigation bar of the CNN website. The main menu includes 'News', 'Video', 'TV', 'Opinions', and 'More...'. Below this, there are sub-menus for 'U.S.', 'World', 'Politics', 'Tech', 'Health', 'Entertainment', 'Living', 'Travel', 'Money', and 'Sports'. A search bar is located on the right. Below the navigation bar is a large advertisement for ANORO ELLIPTA, featuring a red and white inhaler device. The ad includes the text 'Get your free 30-day trial' and 'Important Safety Information'.

Want a happy marriage? Have a big, cheap wedding

By Brandon Griggs, CNN
Updated 7:20 PM ET, Mon October 13, 2014



A section titled 'More from CNN' featuring four article thumbnails. The first thumbnail shows two women, with the text 'Bobbi Kristina Brown undergoes a tracheostomy'. The second thumbnail shows a person in a dark hooded garment, with the text 'Blogger's brutal death for speaking his mind'. The other two thumbnails are partially visible but their text is not legible.

Unfortunately not quite that easy

Big, cheap wedding → Happy Marriage

Other factors

Big family
Few financial problems

Big, cheap wedding → Happy Marriage



Real Time Economics

Economic insight and analysis from The Wall Street Journal.



- MINIMUM WAGE
- EMPLOYMENT
- INFLATION

HOT TOPICS: WSJ ECONOMIST SURVEY GRAND CENTRAL NEWSLETTER CENTRAL BANK WATCH

1:22 pm ET
Oct 24, 2014 CONSUMPTION

Will a Cheap Wedding Help Your Marriage? A Lesson in Causation

- ARTICLE
- COMMENTS (19)

CB RESEARCH CENSUS BUREAU CONSUMER SPENDING CONSUMPTION DIVORCE

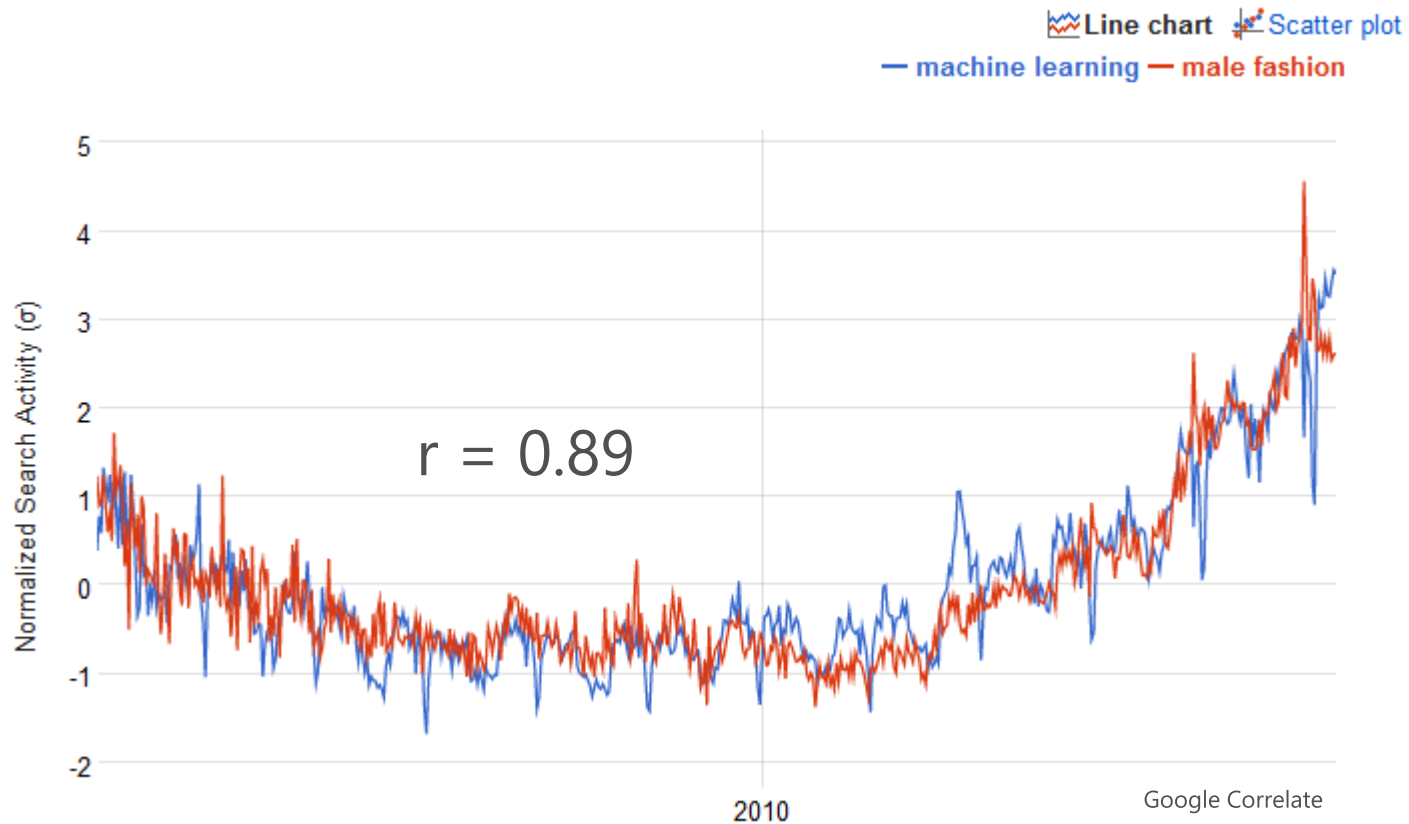
- Email
- Print
- Facebook 2.6k
- Twitter 184
- Google+
- LinkedIn

By SARAH PORTLOCK CONNECT



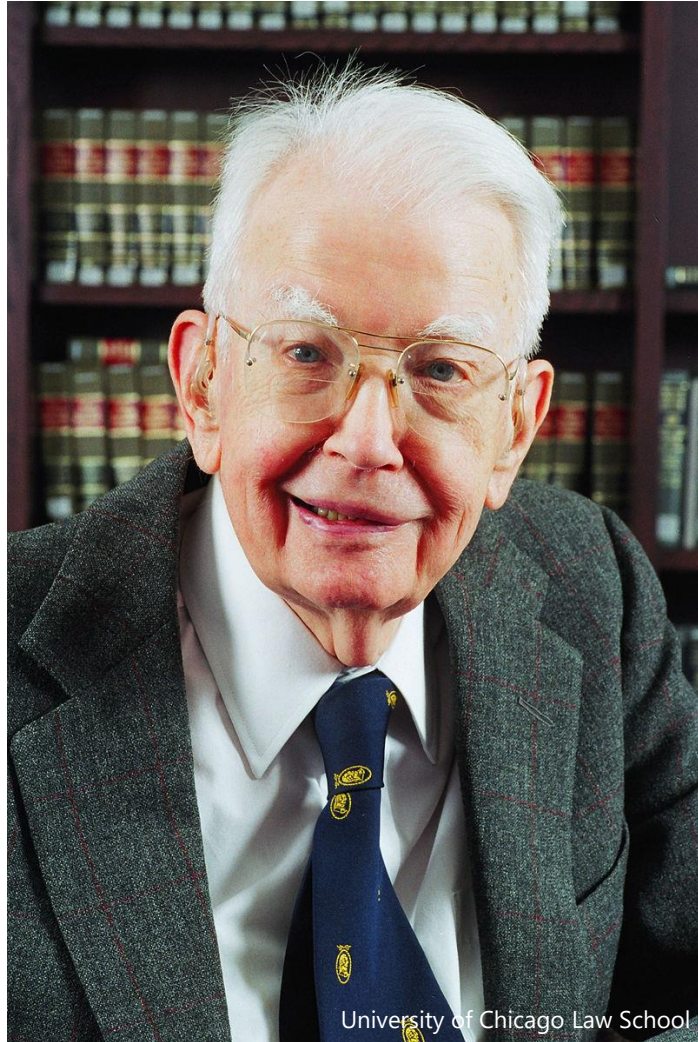
Random correlations are everywhere

United States Web Search activity for **machine learning** and **male fashion** ($r=0.8903$)



Apparently, the more searches for male fashion, the more searches for machine learning.

Good to know all data scientists dress well!



"If you torture the data long enough, it will confess."

- Ronald Coase

Be cognizant of privacy

Google's Wi-Fi Spying: What Were They Thinking?

By [Jeff Bertolucci](#), PCWorld

"Don't be evil" has gone all [1984](#) on us. Or so it seems after [Google](#) revealed Friday that its [Street View cars](#), in addition to snapping photos of the world's roadways, have also been [collecting sensitive personal information](#) from unencrypted wireless networks.

It was no secret that Google's cars had already [been collecting publicly broadcast SSID information](#) (Wi-Fi network names) (unique numbers for devices like Wi-Fi routers). But this technology, used for location-based services such as Google Maps, didn't include [collecting personal information sent over the network](#).

Even if you can gather more data, it's not always the best idea.

Newly Obtained Records Reveal Extensive Monitoring of E-ZPass Tags Throughout New York



By [Mariko Hirose](#), Staff Attorney, NYCLU

APRIL 24, 2015 | 1:00 PM



New documents obtained by the New York Civil Liberties Union reveal that wireless E-ZPass tollbooth transponders are being read routinely throughout New York City to systematically collect location data about drivers.



Privacy also extends to your output

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did



Kashmir Hill
FORBES STAFF

Welcome to *The Not-So Private Parts* where technology & privacy collide

FOLLOW ON FORBES (2081)



FULL BIO >

Opinions expressed by Forbes Contributors are their own.



TARGET

Target has got you in its aim

Every time you go shopping, you share intimate details about your consumption patterns with retailers. And many of those retailers are studying those details to figure out what you like, what you need, and which coupons are most likely to make you happy. **Target**, for example, has figured out how to data-mine its way into your womb, to figure out whether you have a baby on the way long before you need to start buying diapers.

What's Even Creepier Than Target Guessing That Pregnant?



It can be spooky to contemplate living in a world where Google and Facebook and

Why have I not focused on tools or technology?

Predictive Analytics

Microsoft Officially Launches Azure Machine Learning Platform

Posted Feb 18, 2015 by [Ron Miller \(@ron_miller\)](#)

4,069
SHARES



THE WALL STREET JOURNAL.

Home World U.S. Politics Economy Business **Tech** Markets Opinion Arts Life Re



Meet DJI's Osmo, a Robotic Selfie Stick



Review: Dell XPS 15 Fixes the Worst Thing About Windows Laptops



Facebook's Dislike Button Is Here--In the Form of Emoji Reactions



Sam Persona Data No LoopPa

TECH

Amazon Web Services to Add Analytics

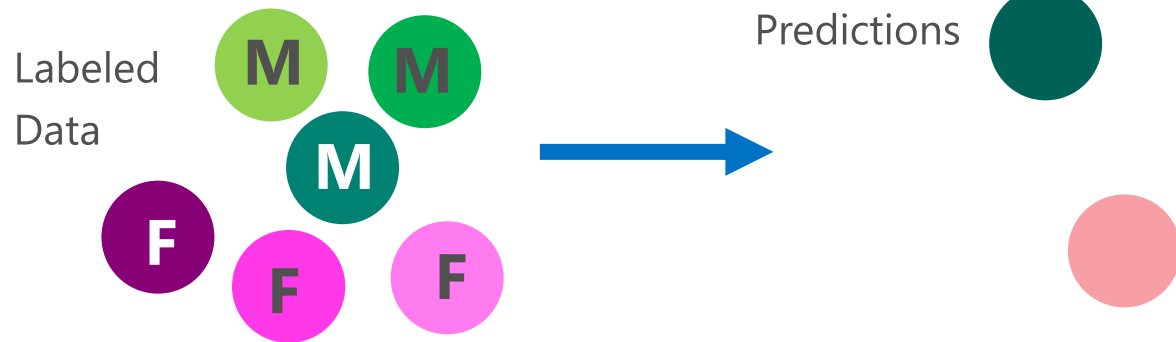
Cloud-computing division enters field designed to make better use of collected data



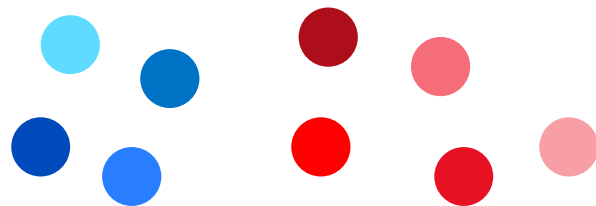
Data science is more than machine learning

Machine Learning

Predictions



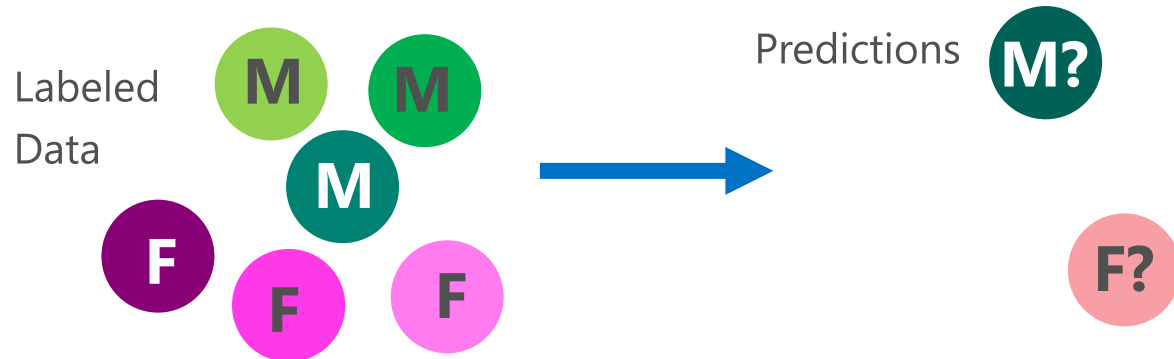
Clustering



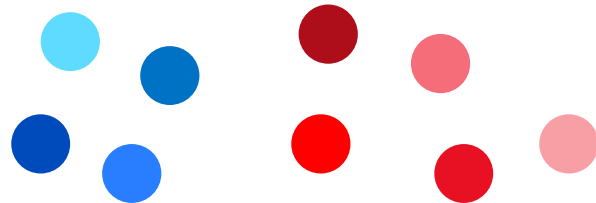
Data science is more than machine learning

Machine Learning

Predictions



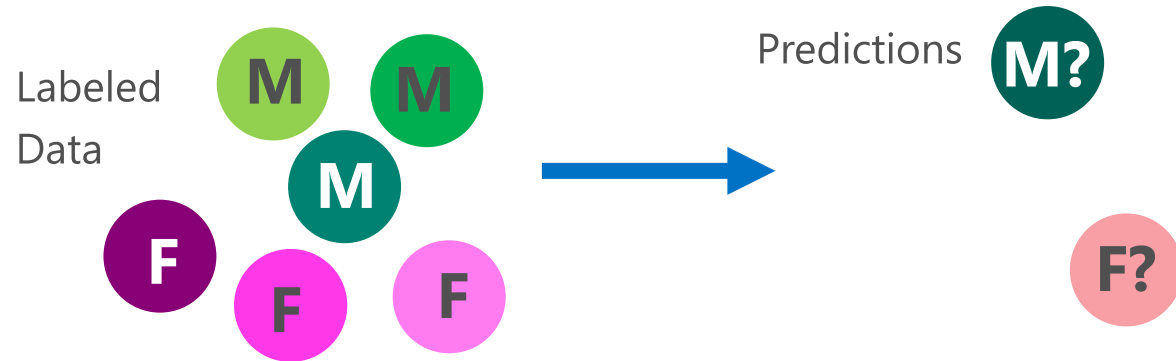
Clustering



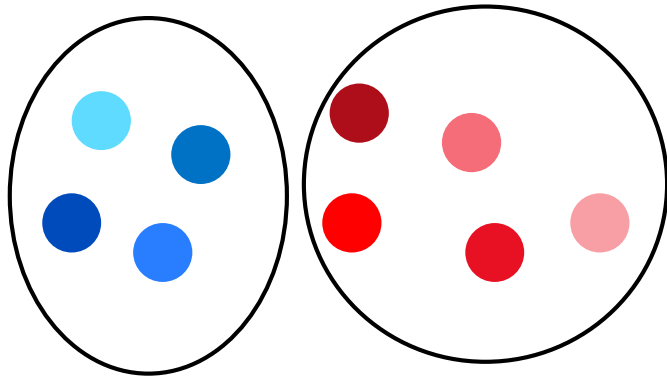
Data science is more than machine learning

Machine Learning

Predictions



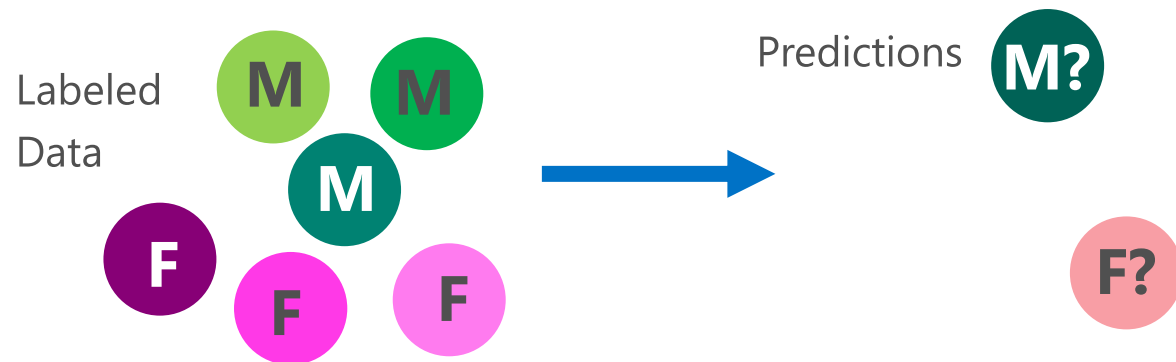
Clustering



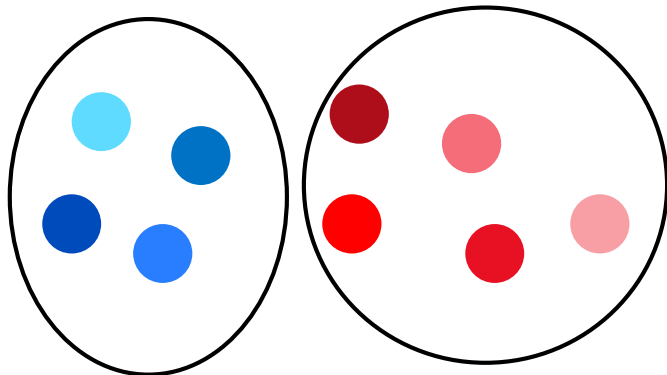
Data science is more than machine learning

Machine Learning

Predictions



Clustering



Data science/"big data"

- Predictions
- Clustering
- Estimations
- Measurements
- Explanations
- Visualizations

Machine learning: easier than you think

MNIST dataset

60k handwritten digits

Raw pixel values provided

Goal is to train a classifier that can recognize handwritten digits



Azure ML Example

Microsoft Azure Machine Learning | Home Studio Gallery PREVIEW David Levitan-Free-Works... ? +

Experiment created on 10/6/2015 Finished running ✓

Search experiment items

- ▶ Saved Datasets
- ▶ Data Format Conversions
- ▶ Data Input and Output
- ▶ Data Transformation
- ▶ Feature Selection
- ▶ **Machine Learning**
 - ▶ Evaluate
 - Cross Validate Model
 - Evaluate Model
 - Evaluate Recommender
 - ▶ Initialize Model
 - ▶ Score
 - Apply Transformation
 - Assign Data to Clusters
 - Score Matchbox Recom...
 - Score Model
 - ▶ Train
 - OpenCV Library Modules
 - Python Language Modules
 - R Language Modules
 - Statistical Functions
 - Text Analytics
 - Web Service
 - Deprecated

```
graph TD; A[MNIST Train 60k 28x28 dense] --> B[Split]; B --> C[Multiclass Neural Network]; B --> D[Train Model]; D --> E[Score Model]; E --> F[Evaluate Model];
```

Properties

- ▶ **Experiment Properties**
 - START TIME 10/6/2015...
 - END TIME 10/6/2015...
 - STATUS CODE Finished
 - STATUS DETAILS None
- ▶ **Summary**

Enter a few sentences describing your experiment (up to 140 characters).
- ▶ **Description**

Enter the detailed description for your experiment.
- ▶ **Quick Help**

NEW RUN HISTORY SAVE DISCARD CHANGES RUN SET UP WEB SERVICE PUBLISH TO GALLERY

Azure ML Example Results

- A few minutes to construct experiment
- 5 minutes for it to train
- 97.8% accuracy with default parameters
- Record is 99.8%

But it's easy to make mistakes:
overfitting, improper tuning, etc...

So read up on the basics.

Want to try it yourself?

<https://studio.azureml.net/>

Experiment created on 10/6/2015 > Evaluate Model > Evaluation results

Metrics

Overall accuracy	0.977583
Average accuracy	0.995517
Micro-averaged precision	0.977583
Macro-averaged precision	0.977397
Micro-averaged recall	0.977583
Macro-averaged recall	0.977364

Confusion Matrix

	Predicted Class								
	0	1	2	3	4	5	6	7	8
0	99.2%		0.1%	0.2%		0.1%			0.3%
1		98.7%	0.3%	0.1%	0.1%		0.1%	0.2%	0.2%
2	0.3%		97.8%	0.3%	0.3%	0.1%	0.1%	0.4%	0.6%
3	0.1%		1.2%	97.0%		0.7%		0.3%	0.3%

APR 15, 2011 1:14PM ET

Quote: The Ad Generation

ELI ROSENBERG



FLICKR/MARI SMITH

"The best minds of my generation are thinking about how to make people click ads."

--Jeff Hammerbacher, a 28-year-old Silicon Valley tech whiz who went from being an early employee at Facebook to co-founding the data analysis start-up Cloudera, in Ashlee Vance's *BusinessWeek* [story](#) about the advertising and social-media driven bubble in Silicon Valley.

Questions?