

Big Data Analysis with Apache Spark



This Lecture

Course Objectives and Prerequisites

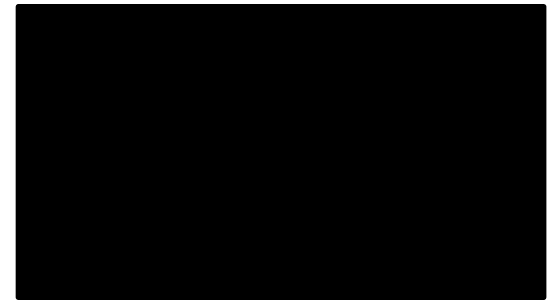
Brief History of Data Analysis

Correlation, Causation, and Confounding Factors

Big Data and Data Science – Why All the Excitement?

So What is Data Science?

Doing Data Science



Course Objectives

Know basic Data Science concepts

- » Extract-Transform-Load operations, data analytics and visualization

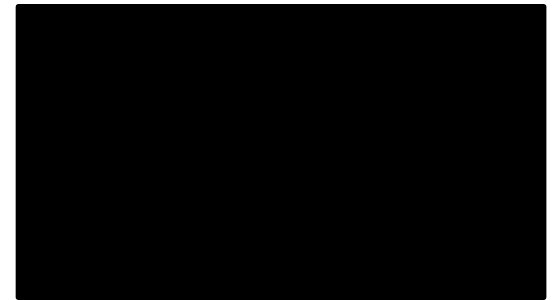
Understand correlation, causation, and confounding factors

Understand the elements of Data Science:

- » Data preparation, Analysis, and Presentation
- » Basic Machine Learning algorithms

Know Apache Spark tools for Data Science

- » DataFrames, RDDs, and ML Pipelines



Course Prerequisites

Basic programming skills and experience

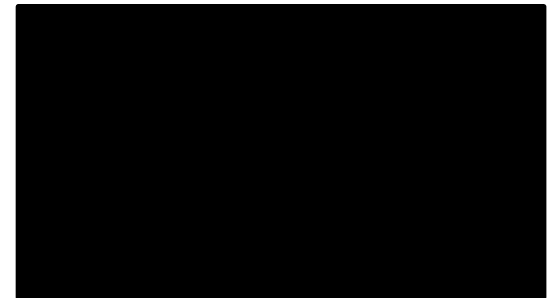
Basic Apache Spark experience

» *CS 105x is required*

» Some experience with [Python 2.7](#)

[Google Chrome web browser](#)

» *Internet Explorer, Edge, Safari are not supported*



What is Data Science?

Drawing useful conclusions from data using computation

- **Exploration**

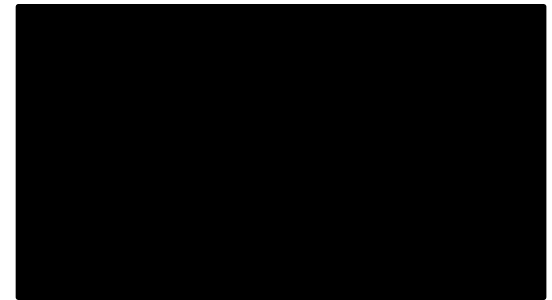
- » Identifying patterns in information
- » Using visualizations

- **Prediction**

- » Making informed guesses
- » Using machine learning and optimization

- **Inference**

- » Quantifying our degree of certainty



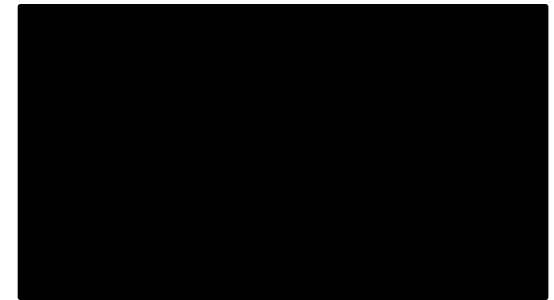
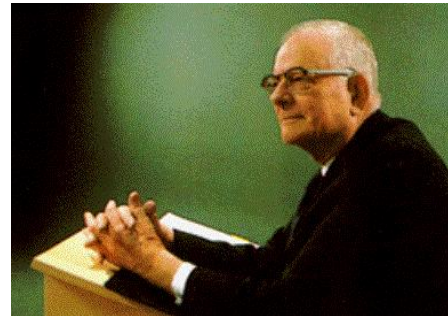
Brief Data Analysis History

- R. A. Fisher
 - » 1935: “The Design of Experiments”

“correlation does not imply causation”



- W. E. Demming
 - » 1939: “Quality Control”

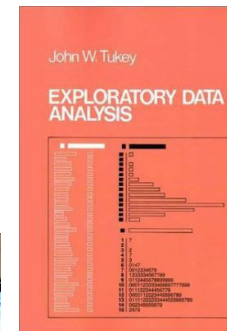


Brief Data Analysis History

- Peter Luhn
 - » 1958: “A Business Intelligence System”



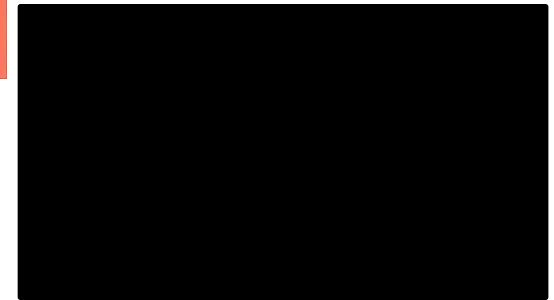
- John W. Tukey
 - » 1977: “Exploratory Data Analysis”



- Howard Dresner
 - » 1989: “Business Intelligence”

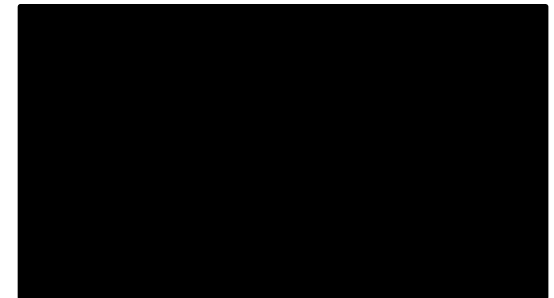
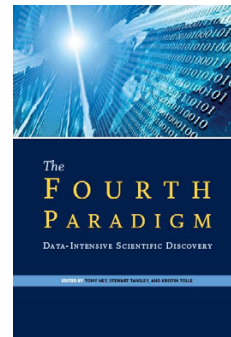


Images: <http://www.businessintelligence.info/definiciones/business-intelligence-system-1958.html>
<http://www.betterworldbooks.com/exploratory-data-analysis-id-0201076160.aspx>
<https://www.flickr.com/photos/42266634@N02/4621418442>



Brief Data Analysis History

- Tom Mitchell
 - » 1997: “Machine Learning book”
- Google
 - » 1996: “Prototype Search Engine”
- Data-Driven Science eBook
 - » 2007: “[The Fourth Paradigm](http://research.microsoft.com/en-us/collaboration/fourthparadigm/)”

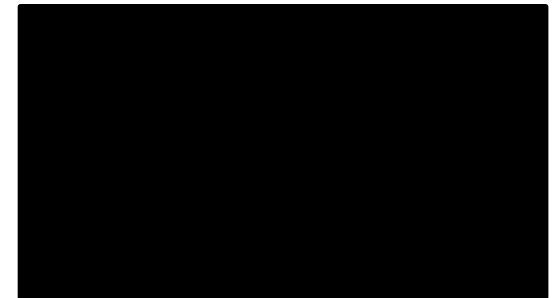


Images: <http://www.amazon.com/Machine-Learning-Tom-M-Mitchell/dp/0070428077>
<http://www.google.com/about/company/history/>
<http://research.microsoft.com/en-us/collaboration/fourthparadigm/>

Brief Data Analysis History

- Peter Norvig
 - » 2009: “The Unreasonable Effectiveness of Data”

- Exponential growth in data volume
 - » 2010: “The Data Deluge”



Why All the Excitement?

elections2012

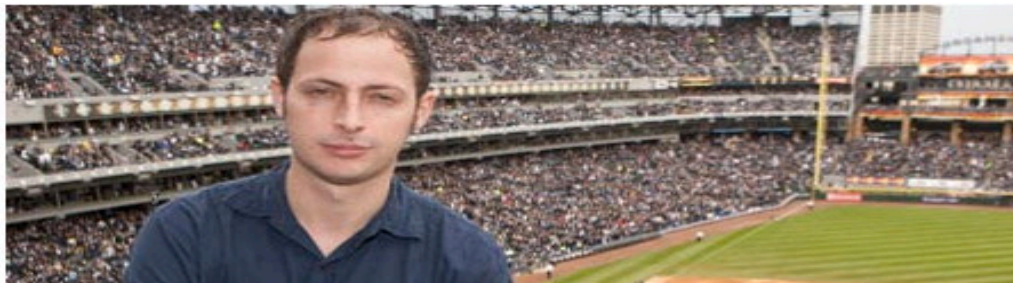
Live results | [President](#) | [Senate](#) | [House](#) | [Governor](#) |

Numbers nerd Nate Silver's forecasts prove all right on election night

FiveThirtyEight blogger predicted the outcome in all 50 states, assuming Barack Obama's Florida victory is confirmed

Luke Harding

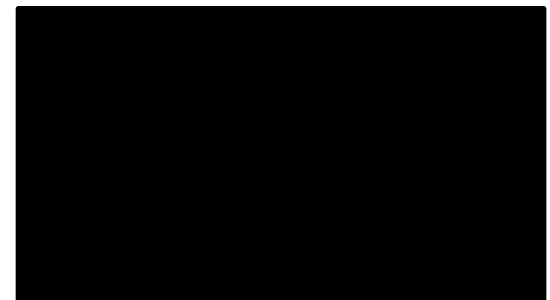
guardian.co.uk, Wednesday 7 November 2012 10.45 EST



<http://www.theguardian.com/world/2012/nov/07/nate-silver-election-forecasts-right>

USA 2012
Presidential
Election

*the signal and the
and the noise and
the noise and the
noise and the no
why most noise a
predictions fail t
but some don't n
and the noise an
the noise and the
nate silver noise
noise and the wa*

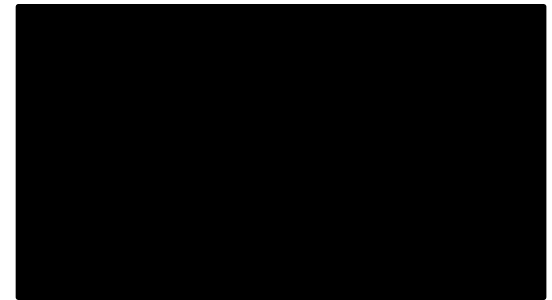


Big Data and USA 2012 Election

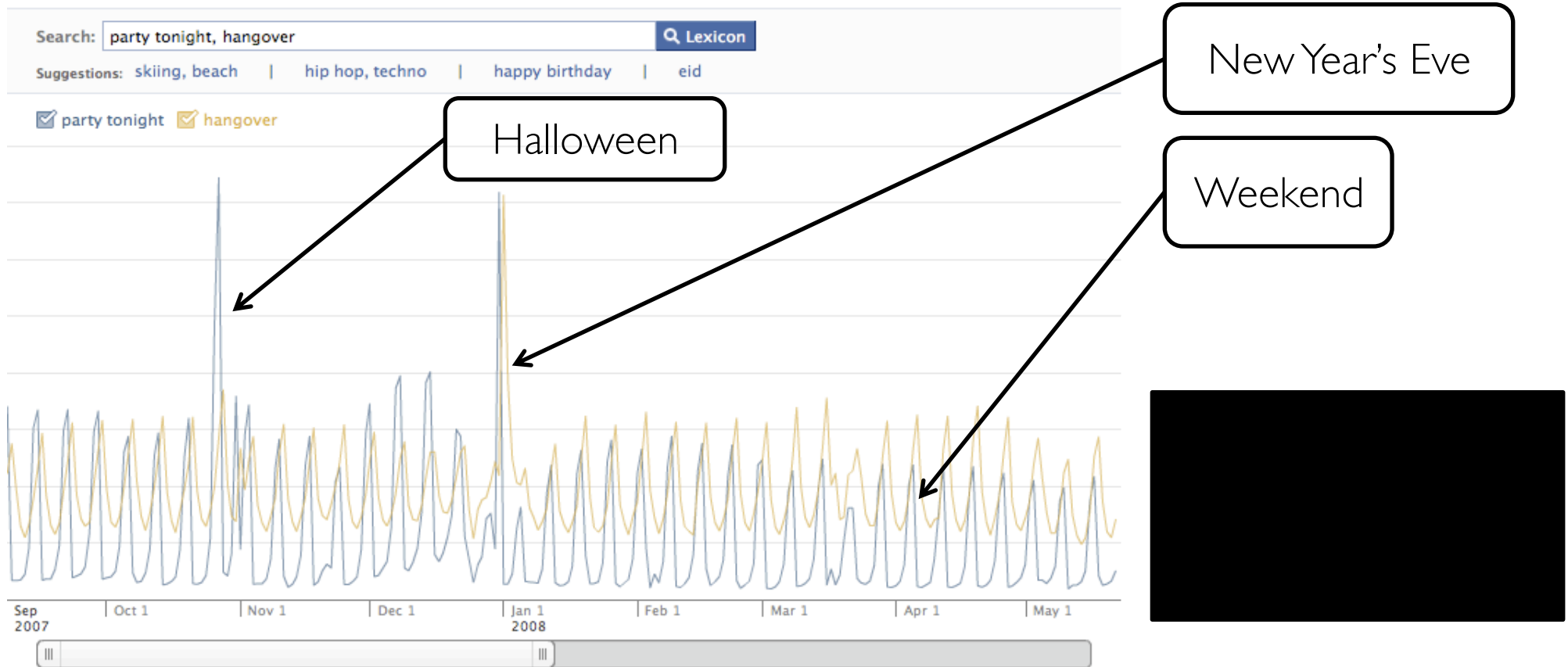
...that was just one of several ways that Mr. Obama's campaign operations, some unnoticed by Mr. Romney's aides in Boston, **helped save the president's candidacy**. In Chicago, the campaign recruited a team of behavioral scientists to build an **extraordinarily sophisticated database**

...that allowed the Obama campaign not only to alter the very nature of the electorate, making it younger and less white, but also to create a portrait of shifting voter allegiances. **The power of this operation stunned Mr. Romney's aides on election night**, as they saw voters they never even knew existed turn out in places like Osceola County, Fla.

[New York Times, Wed Nov 7, 2012](#)



Example: Facebook Lexicon

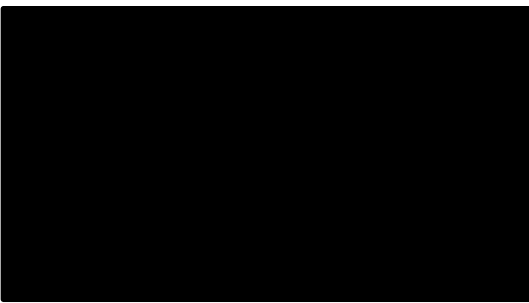


Example: Facebook Lexicon



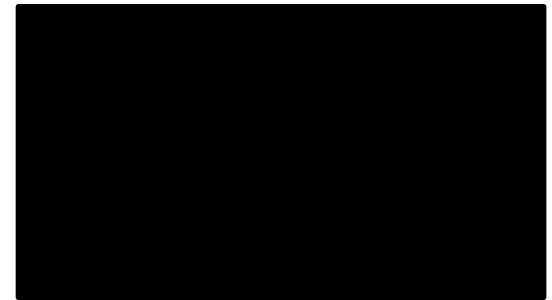
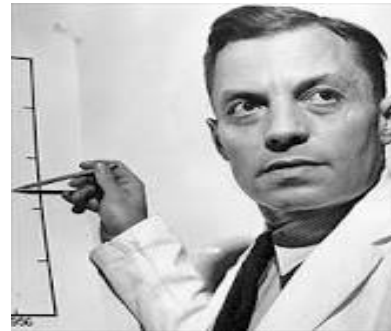
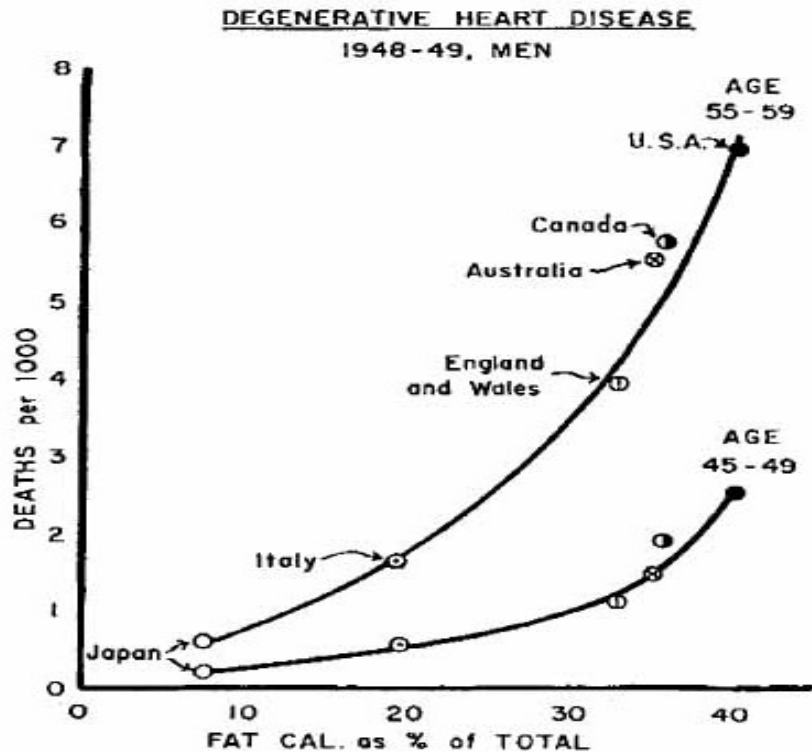
Facebook availability in new countries and languages

Hypothesis: A possible explanation



Data Makes Everything Clearer (part I)?

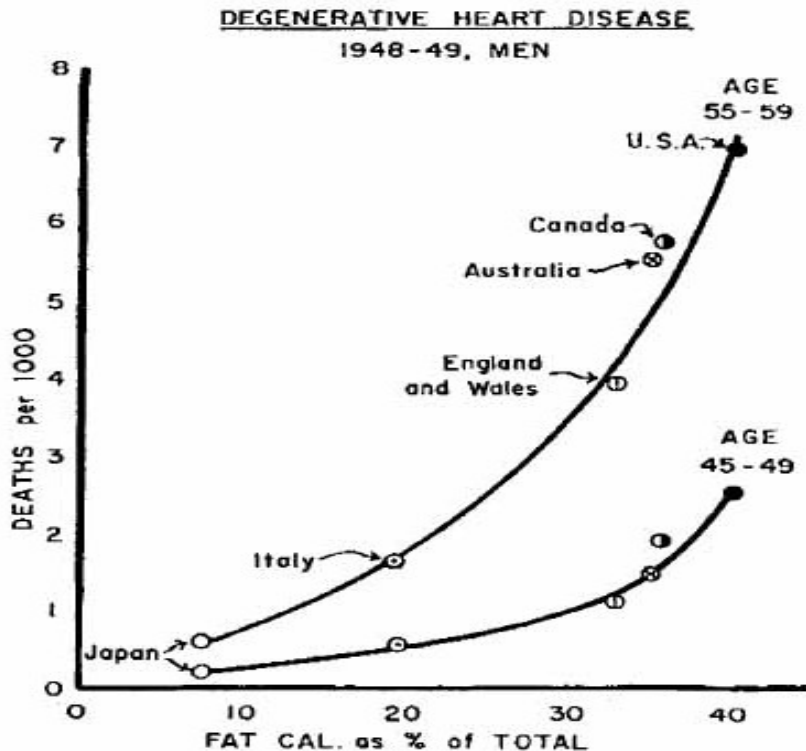
- Seven Countries Study (Ancel Keys)
 - » Started in 1958, followed 13,000 subjects total for 5-40 years



http://en.wikipedia.org/wiki/Seven_Countries_Study

Data Makes Everything Clearer (part I)?

- Seven Countries Study (Ancel Keys)
 - » Started in 1958, followed 13,000 subjects total for 5-40 years

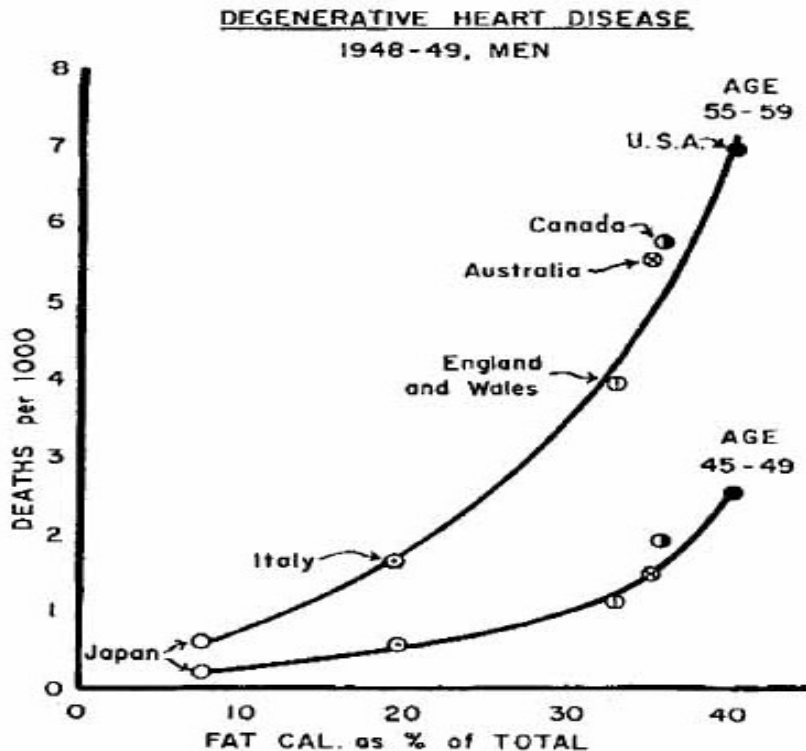


Is there any **relation** between fat consumption and heart disease?

- **Association** \equiv "any relation"

Data Makes Everything Clearer (part I)?

- Seven Countries Study (Ancel Keys)
 - » Started in 1958, followed 13,000 subjects total for 5-40 years



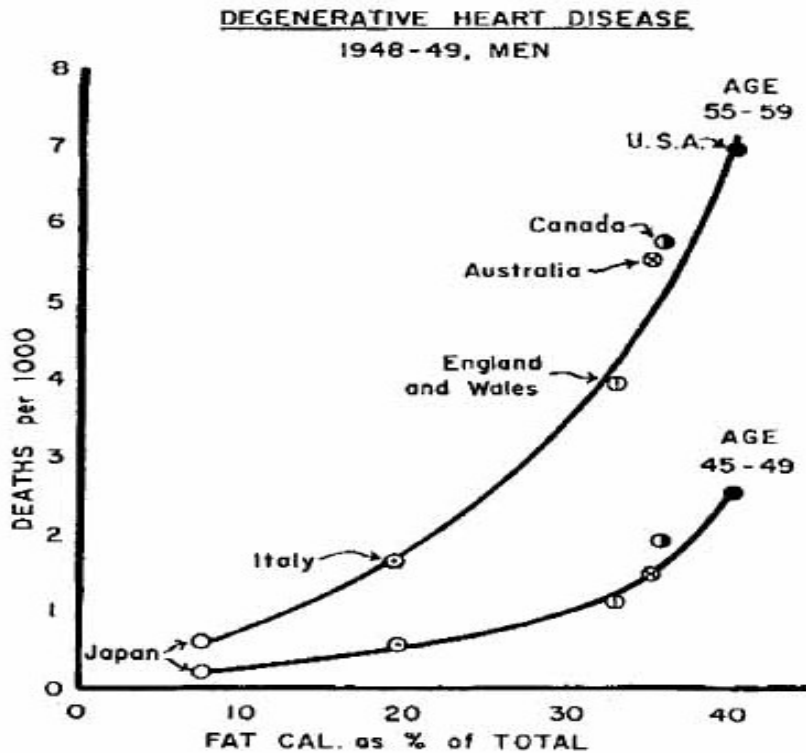
Is there any **relation** between fat consumption and heart disease?

- **Association** \equiv "any relation"

YES – the graph points to an association

Data Makes Everything Clearer (part I)?

- Seven Countries Study (Ancel Keys)
 - » Started in 1958, followed 13,000 subjects total for 5-40 years



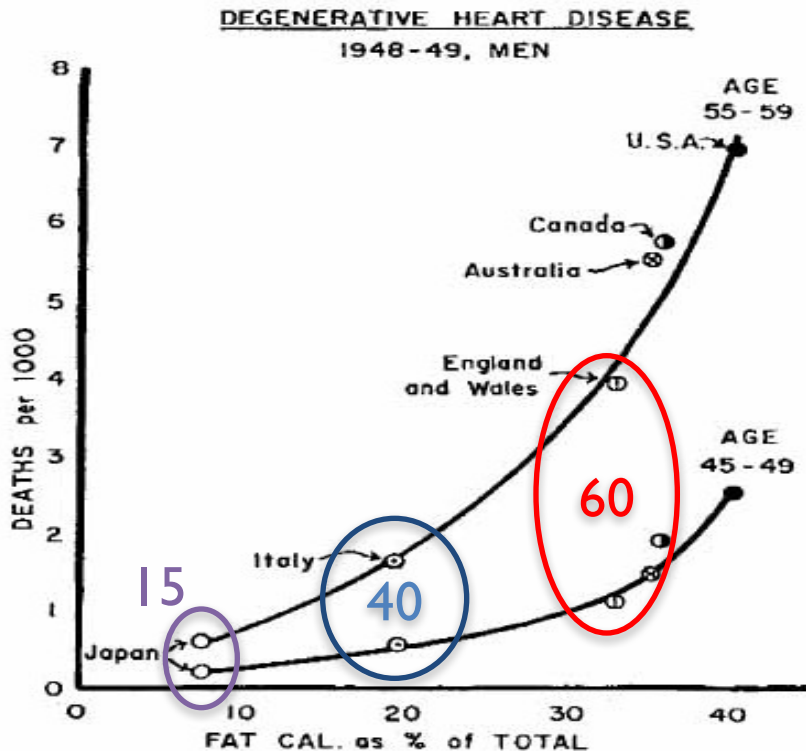
Does fat consumption **increase** heart disease?

- **Causality**

This question is often harder to answer

Data Makes Everything Clearer (part I)?

- Seven Countries Study (Ancel Keys)
 - » Started in 1958, followed 13,000 subjects total for 5-40 years



Significant controversy

- Only studied subset of 21 countries with data
- Failed to consider other factors (e.g., per capita annual sugar consumption in pounds)

“correlation does not imply causation”

Miasmas & Miasmatisers (pre-20th century)

Bad smells given off by waste and rotting matter

» Believed to be the main source of diseases such as Cholera

Suggested remedies:

» “A pocket full o’posies”

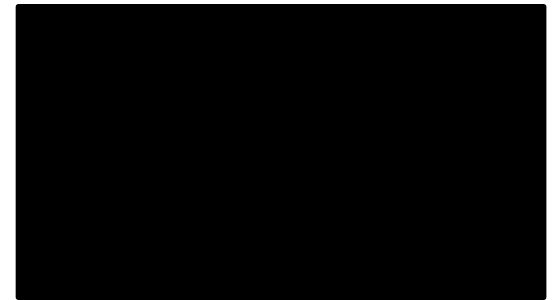
» Fire off barrels of gunpowder

Staunch believers:

» Florence Nightingale

» Edwin Chadwick, Commissioner of the
General Board of Health

https://en.wikipedia.org/wiki/Miasma_theory



John Snow, 1813-1858

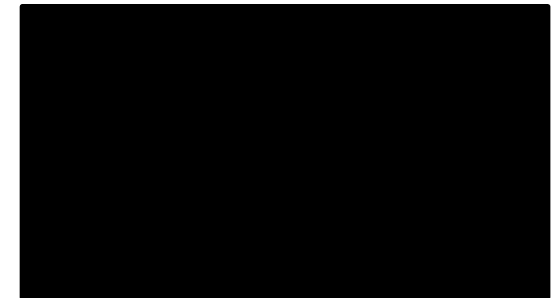
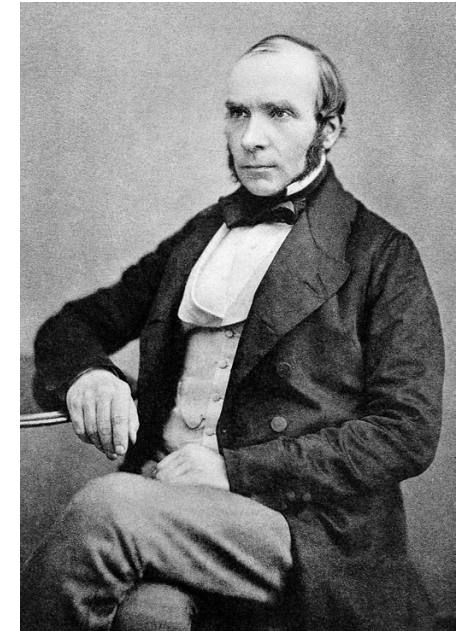
London doctor in the 1850's

Devastating waves of cholera

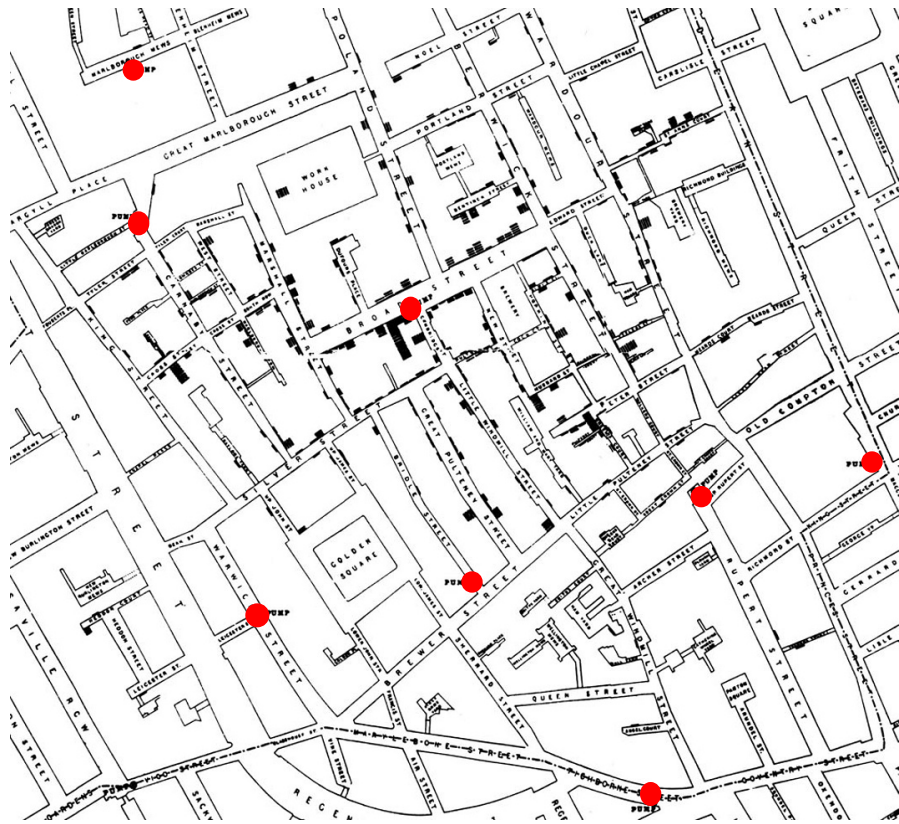
- » Sudden onset
- » People died within a day or two of contracting it
- » Hundreds died in a week
- » Tens of thousands could die in each outbreak

Snow suspected cause was drinking water contaminated by sewage

<https://en.wikipedia.org/wiki/User:Rsabbatini>



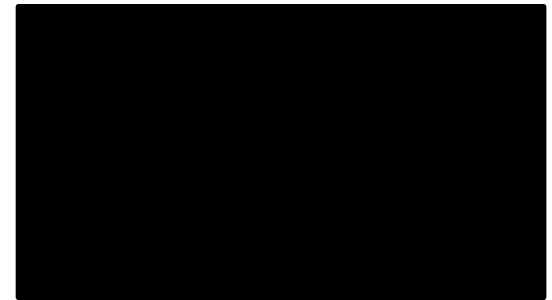
August 1854 London Soho Outbreak



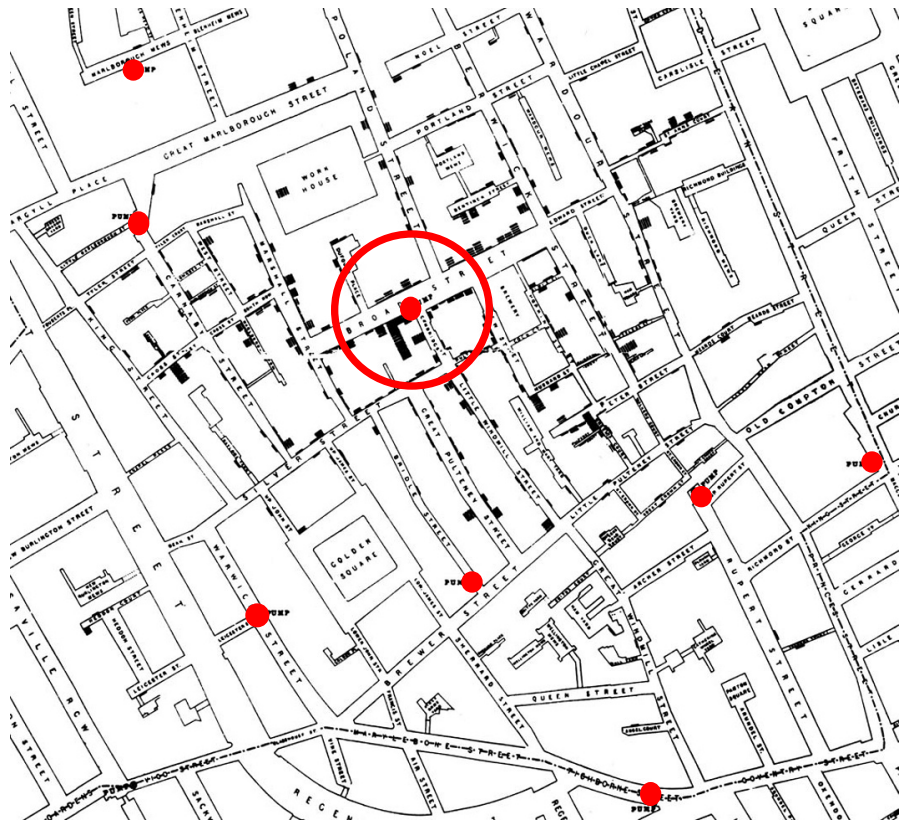
Snow took detailed notes on each death – each bar is a death

Red discs are water pumps

“Spot Map”



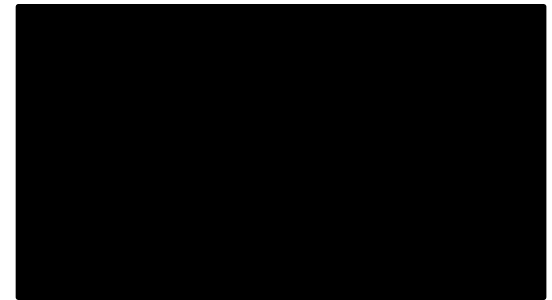
August 1854 London Soho Outbreak



Snow took detailed notes on each death – each bar is a death

Red discs are water pumps

Deaths clustered around Broad Street pump



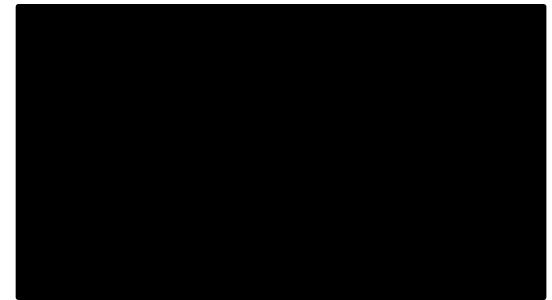
Snow's Analysis

Map has some anomalies, so Snow researched the causes

- » People used pump based on street layout, not distance
- » Brewery workers drank what they brewed and used private well
- » Children from other areas drank pump's water on way to school
- » Two former residents had Broad St water delivered to them

Snow used his map to convince local authorities to close Broad St pump by removing the pump handle

Later a leaking cesspool was found nearby



Snow's Analysis

One of the earliest/most powerful uses of data visualization

Still referred to today!

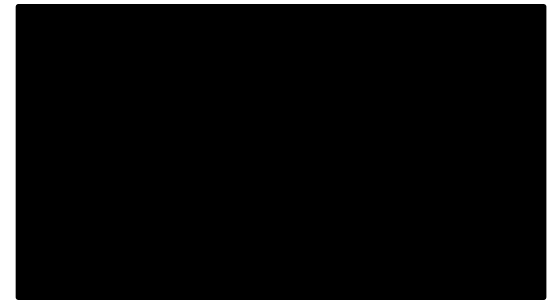
» Scientists at the Centers for Disease Control (CDC) in Atlanta researching outbreaks sometimes ask each other:

“Where is the handle to this pump?”

Is the map a convincing scientific argument?

No! A correlation, not necessarily causation

Hypothesis: A possible explanation



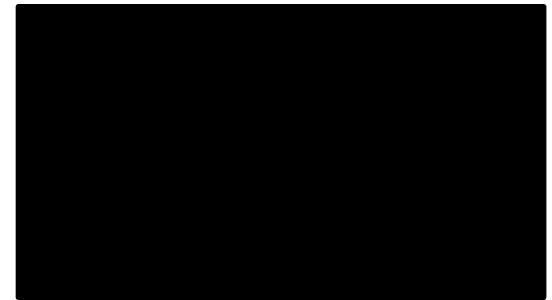
Comparison

Scientists use **comparison** to identify association between a treatment and an outcome

- » Compare outcomes of group of individuals who got treatment (**treatment group**) to outcomes of group who did not (**control group**)

Different results mean evidence of association

- » Determining causation requires even more care



Snow's “Grand Experiment”

Scientific analysis of Cholera deaths based on water source

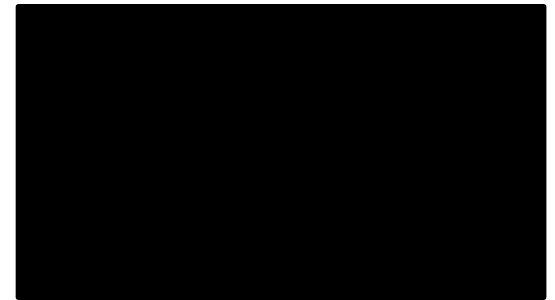


- Water companies used Thames river
- Lambeth drew water from upriver of sewage discharge
 - S&V company from below sewage discharge

Snow's "Grand Experiment"

"... there is no difference whatever in the houses or the people receiving the supply of the two Water Companies, or in any of the physical conditions with which they are surrounded ..."

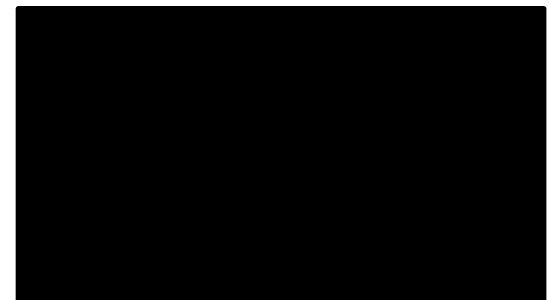
The two groups were similar except for the treatment



Snow's Table

Supply Area	Number of Houses	Cholera Deaths	Deaths per 10,000 Houses
S&V	40,046	1,263	315
Lambeth	26,107	98	37
Rest of London	256,423	1,422	59

S&V death rate was nearly 10x Lambeth-supplied houses



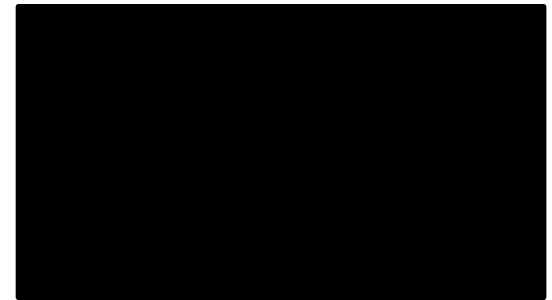
Confounding Factors

If treatment and control groups are **similar apart from the treatment**, then difference in outcomes can be ascribed to the treatment

If treatment and control groups have **systematic differences other than the treatment**, then might be difficult to identify causality

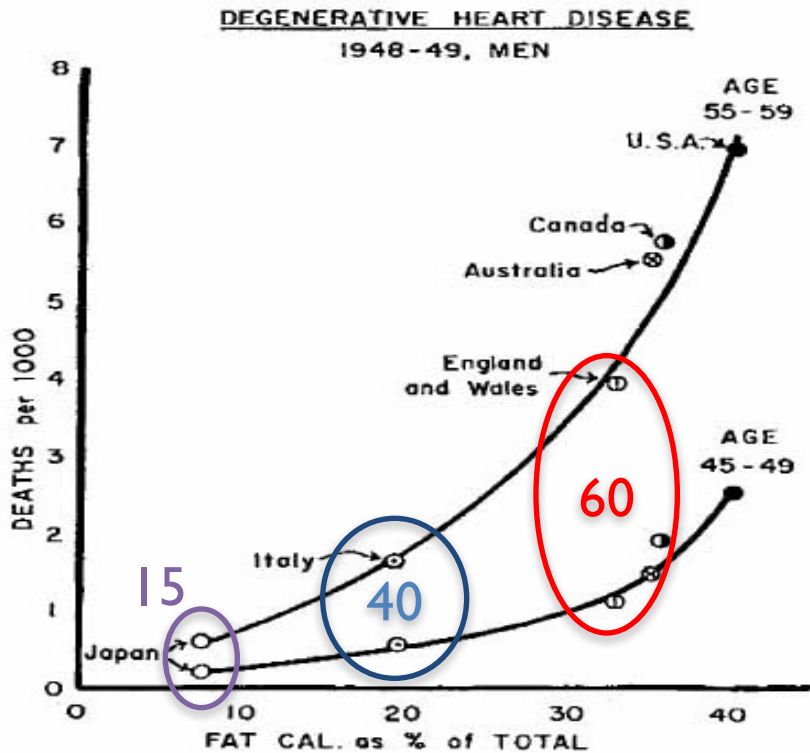
- » Such differences are often present in **observational studies**
(no control over assignment)

They are called **confounding factors** and can lead researchers astray



7 Countries Study Confounding Factors

- Seven Countries Study (Ancel Keys)
 - » Started in 1958, followed 13,000 subjects total for 5-40 years



Confounding Factors:

- Only studied subset of 21 countries with data
- Other factors (e.g., per capita annual sugar consumption in pounds)

“correlation does not imply causation”

Randomize!

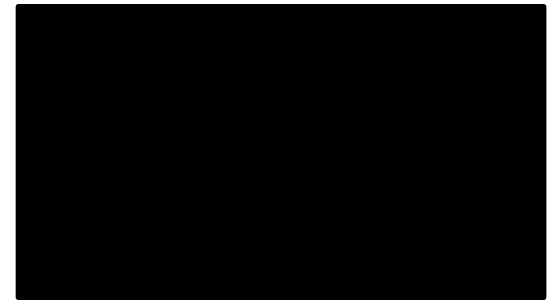
If you assign individuals to treatment and control at **random**, then the two groups will be similar apart from the **treatment**

Can account – *mathematically* – for variability in assignment

Randomized Controlled Experiment

May run **blind** experiment (placebo drug)

Be careful with **observational studies!**



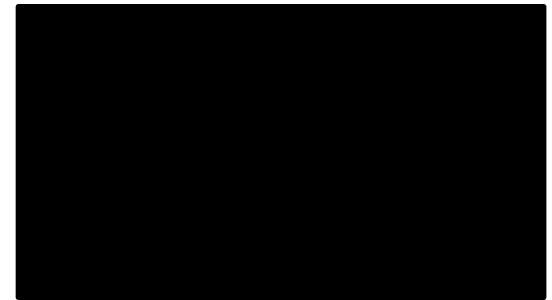
Comparison

Group by some *treatment* and measure some *outcome*

» Simplest setting: a *treatment group* and a *control group*

If the *outcome* differs between these two groups, that's evidence of an *association* (or *relation*)

» E.g., lowest tier of fat consumption had lower rate of heart disease



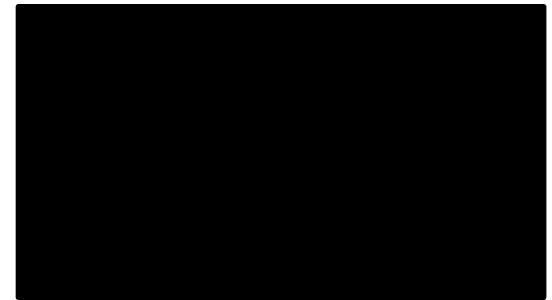
Comparison

Group by some *treatment* and measure some *outcome*

» Simplest setting: a *treatment group* and a *control group*

If the two groups are similar in all ways but the *treatment* itself, a difference in the *outcome* is evidence of *causality*

When a group is divided *randomly*, it's unlikely that there are systematic differences between sub-groups



Data Makes Everything Clearer (part II)?

Epidemiological modeling of online social network dynamics

John Cannarella¹, Joshua A. Spechler^{1,*}

¹ Department of Mechanical and Aerospace Engineering, Princeton University, Princeton, NJ, USA

* E-mail: Corresponding spechler@princeton.edu

Abstract

The last decade has seen the rise of immense online social networks (OSNs) such as MySpace and Facebook. In this paper we use epidemiological models to explain user adoption and abandonment of OSNs, where adoption is analogous to infection and abandonment is analogous to recovery. We modify the traditional SIR model of disease spread by incorporating infectious recovery dynamics such that contact between a recovered and infected member of the population is required for recovery. The proposed infectious recovery SIR model (irSIR model) is validated using publicly available Google search query data for “MySpace” as a case study of an OSN that has exhibited both adoption and abandonment phases. The irSIR model is then applied to search query data for “Facebook,” which is just beginning to show the onset of an abandonment phase. Extrapolating the best fit model into the future predicts a rapid decline in Facebook activity in the next few years.

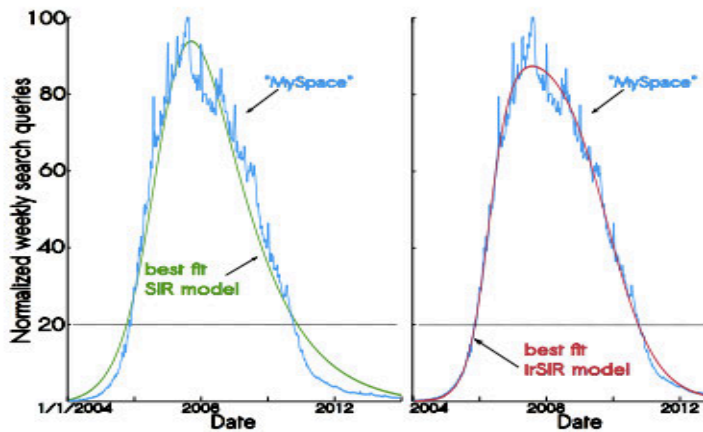
*Beware of
observational studies*

“Extrapolating the best fit model into the future predicts a rapid decline in Facebook activity in the next few years.”

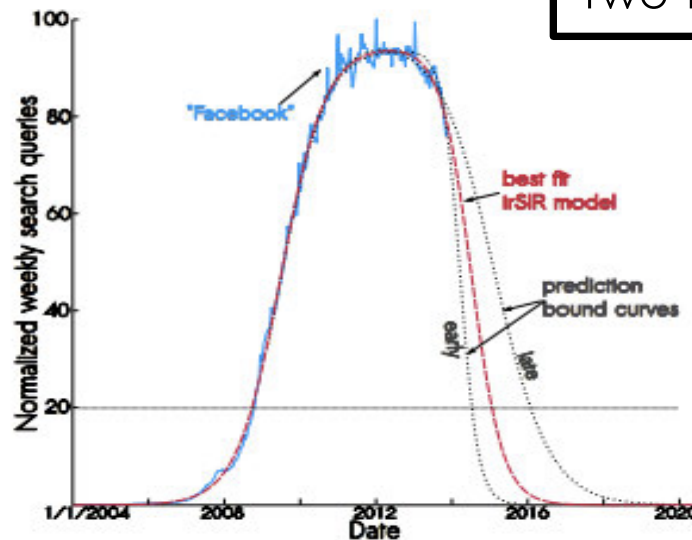
<http://arxiv.org/abs/1401.4208>

Data Makes Everything Clearer (part II)?

Google Trends searches for “*MySpace*”

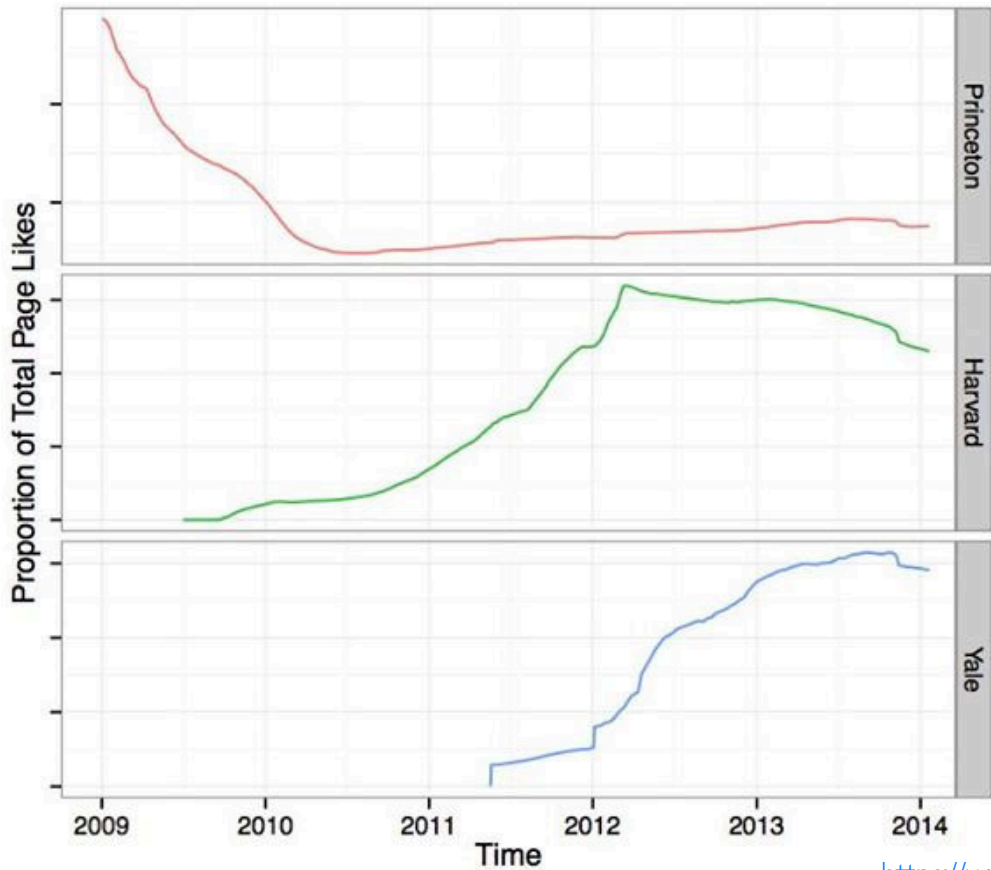


Two Figures from the paper



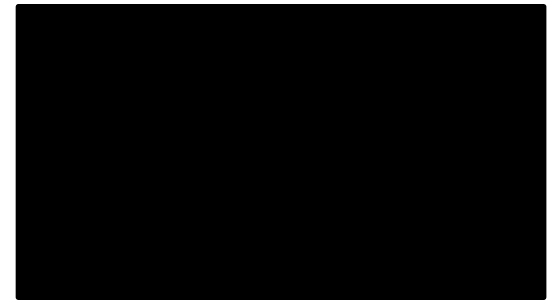
Searches for “*Facebook*”

Data Makes Everything Clearer (part II)?



In keeping with the scientific principle "*correlation equals causation*," our research unequivocally demonstrated that Princeton may be in danger of disappearing entirely.

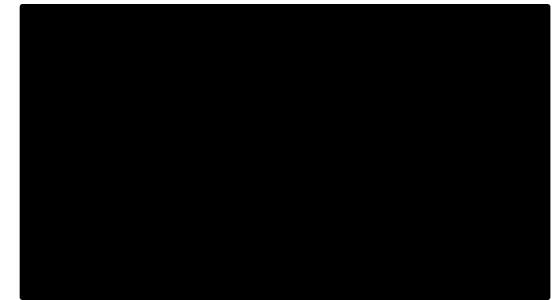
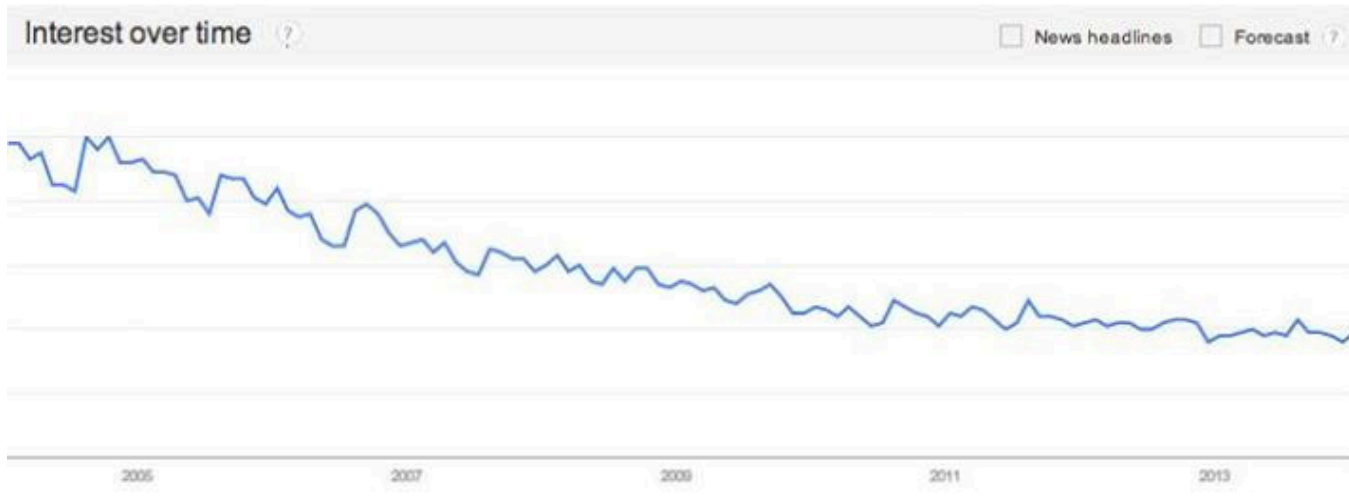
— Princeton
— Harvard
— Yale



Data Makes Everything Clearer (part II)?

... and based on “*Princeton*” search trends:

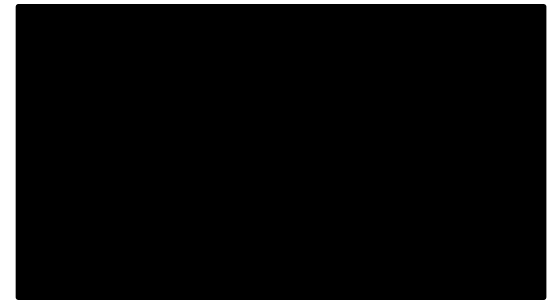
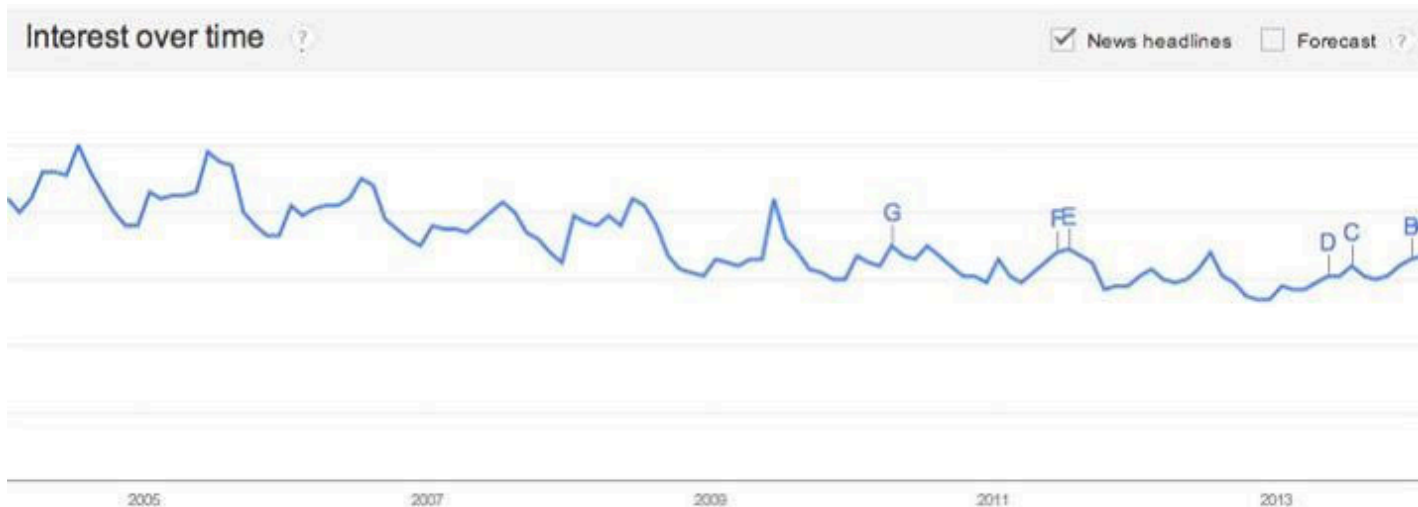
“This trend suggests that Princeton will have only half its current enrollment by 2018, and by 2021 it will have no students at all,…”



<https://www.facebook.com/notes/mike-develin/debunking-princeton/10151947421191849>

Data Makes Everything Clearer (part II)?

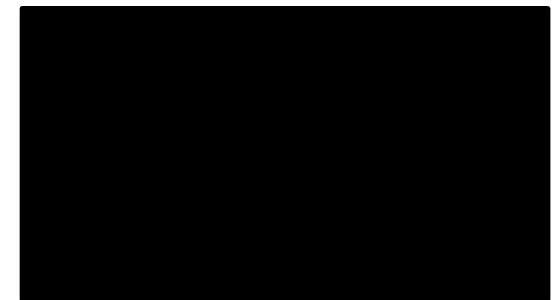
While we are concerned for Princeton University, we are even more concerned about the fate of the planet — Google Trends for “*air*” have also been declining steadily, and our projections show that by the year 2060 there will be no air left:



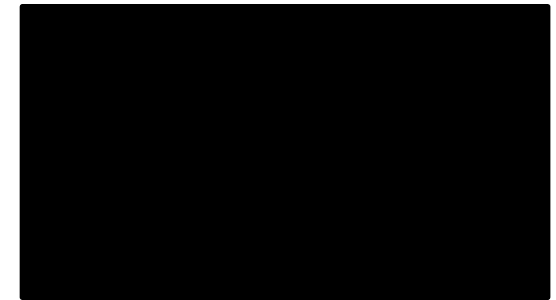
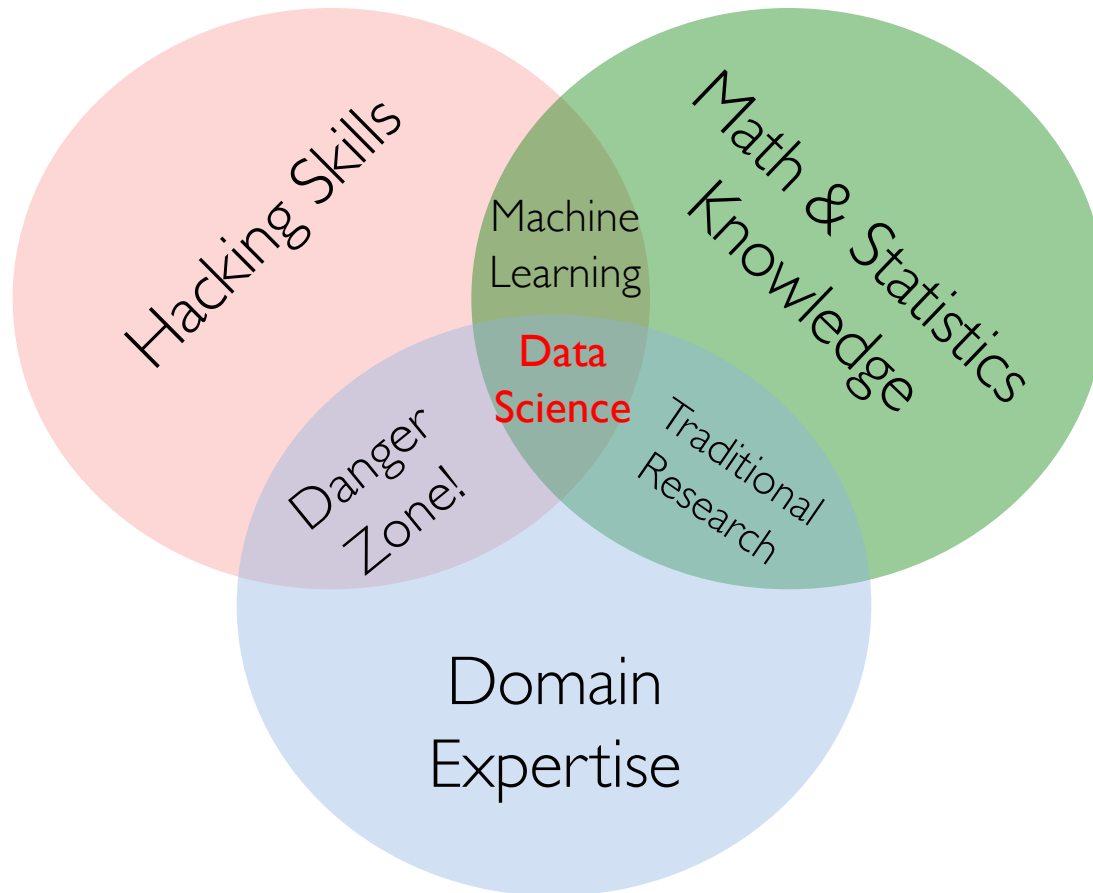
What is Data Science?

Data Science aims to derive knowledge from big data, efficiently and intelligently

Data Science encompasses the set of activities, tools, and methods that enable data-driven activities in science, business, medicine, and government



Data Science – One Definition

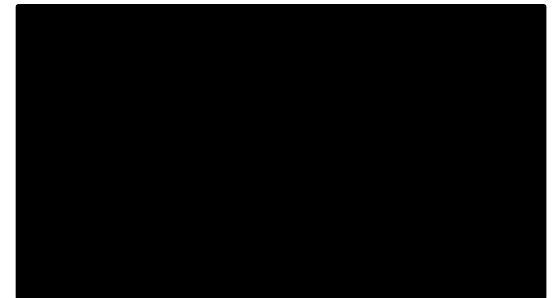


Contrast: Databases

Element	Databases	Data Science
Data Value	“Precious”	“Cheap”
Data Volume	Modest	Massive
Examples	Bank records, Personnel records, Census, Medical records	Online clicks, GPS logs, Tweets, tree sensor readings
Priorities	Consistency, Error recovery, Auditability	Speed, Availability, Query richness
Structured	Strongly (Schema)	Weakly or none (Text)
Properties	Transactions, ACID ⁺	CAP [*] theorem (2/3), eventual consistency
Realizations	Structured Query Language (SQL)	NoSQL : Riak , Memcached , Apache Hbase , Apache River , MongoDB , Apache Cassandra , Apache CouchDB ,...

*CAP = Consistency, Availability, Partition Tolerance

+ACID = Atomicity, Consistency, Isolation and Durability



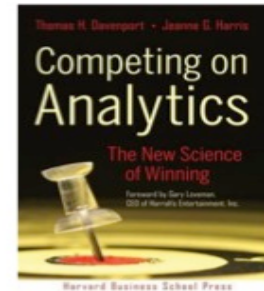
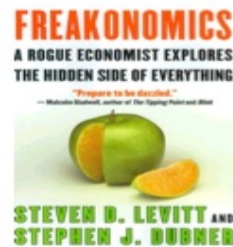
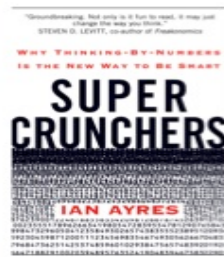
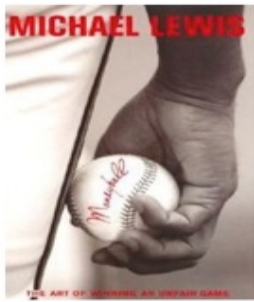
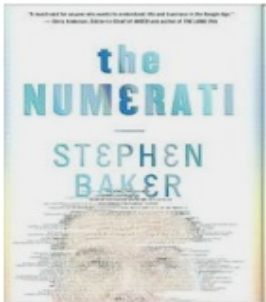
Contrast: Databases

Databases

Data Science

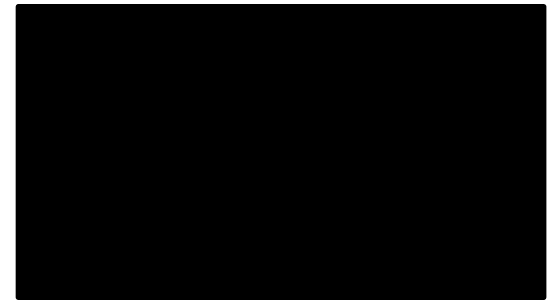
Querying the past

Querying the future

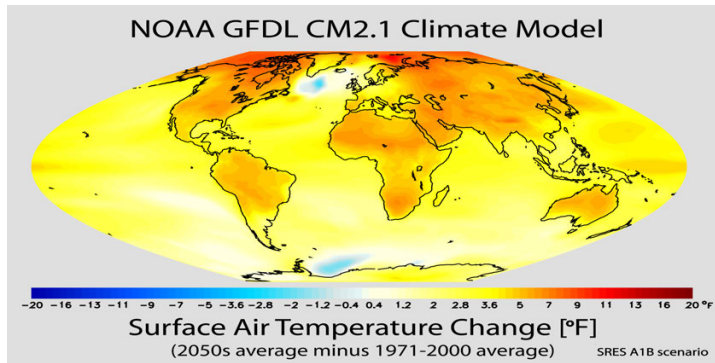


Related – Business Analytics

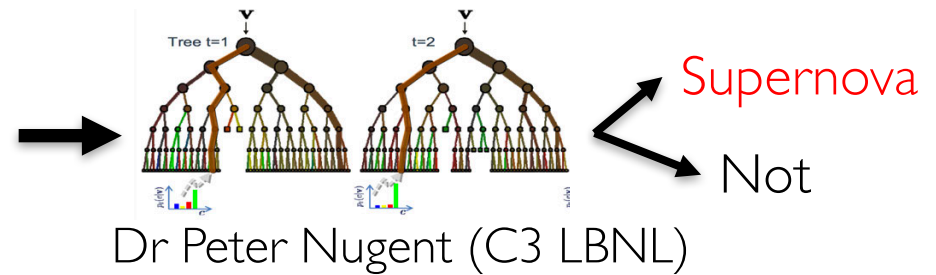
- » Goal: obtain “actionable insight” in complex environments
- » Challenge: vast amounts of disparate, unstructured data and limited time



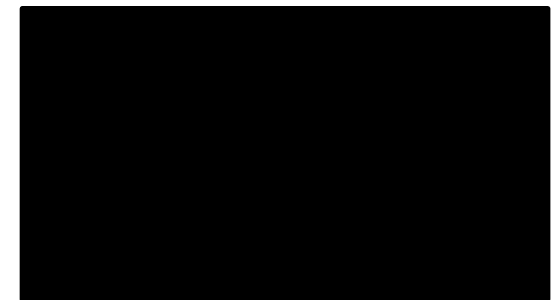
Contrast: Scientific Computing



General purpose ML classifier

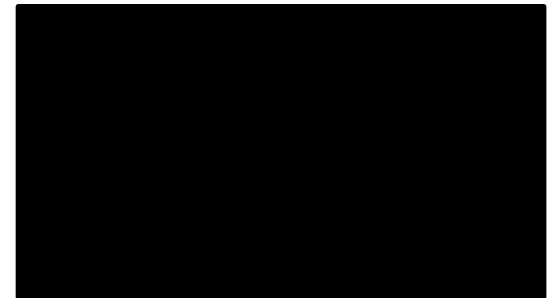


Scientific Modeling	Data-Driven Approach
Physics-based models	General inference engine replaces model
Problem-Structured	Structure not related to problem
Mostly deterministic, precise	Statistical models handle true randomness, and unmodeled complexity
Run on Supercomputer or High-end Computing Cluster	Run on cheaper computer Clusters (EC2)



Contrast: Traditional Machine Learning

Traditional Machine Learning	Data Science
Develop new (individual) models	Explore many models, build and tune hybrids
Prove mathematical properties of models	Understand empirical properties of models
Improve/validate on a few, relatively clean, small datasets	Develop/use tools that can handle massive datasets
Publish a paper	Take action!

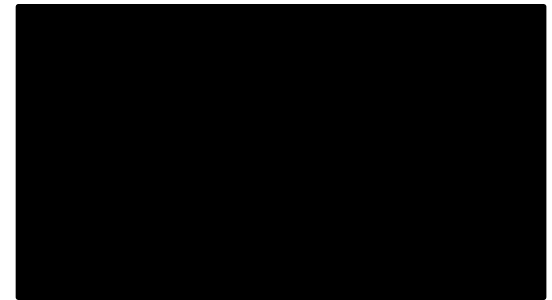


Doing Data Science

The views of three Data Science experts

- » Jim Gray (Turing Award winning database researcher)
- » Ben Fry (Data visualization expert)
- » Jeff Hammerbacher (Former Facebook Chief Scientist, Cloudera co-founder)

Cloud computing: Data Science enabler



Key Data Science Enabler: Cloud Computing

Cloud computing reduces computing operating costs

Cloud computing enables data science on massive numbers of inexpensive computers

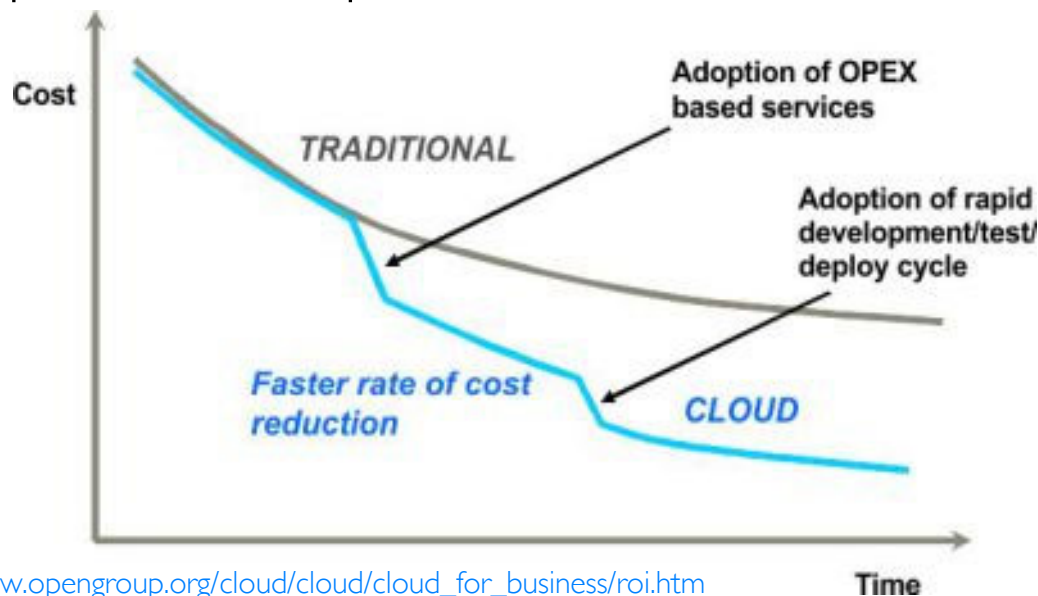
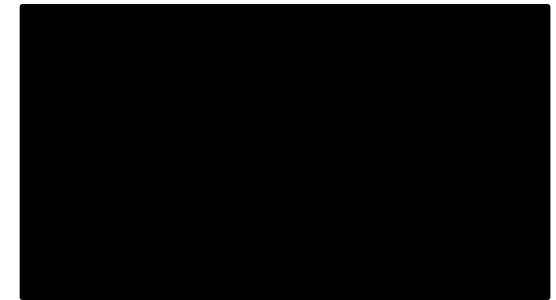
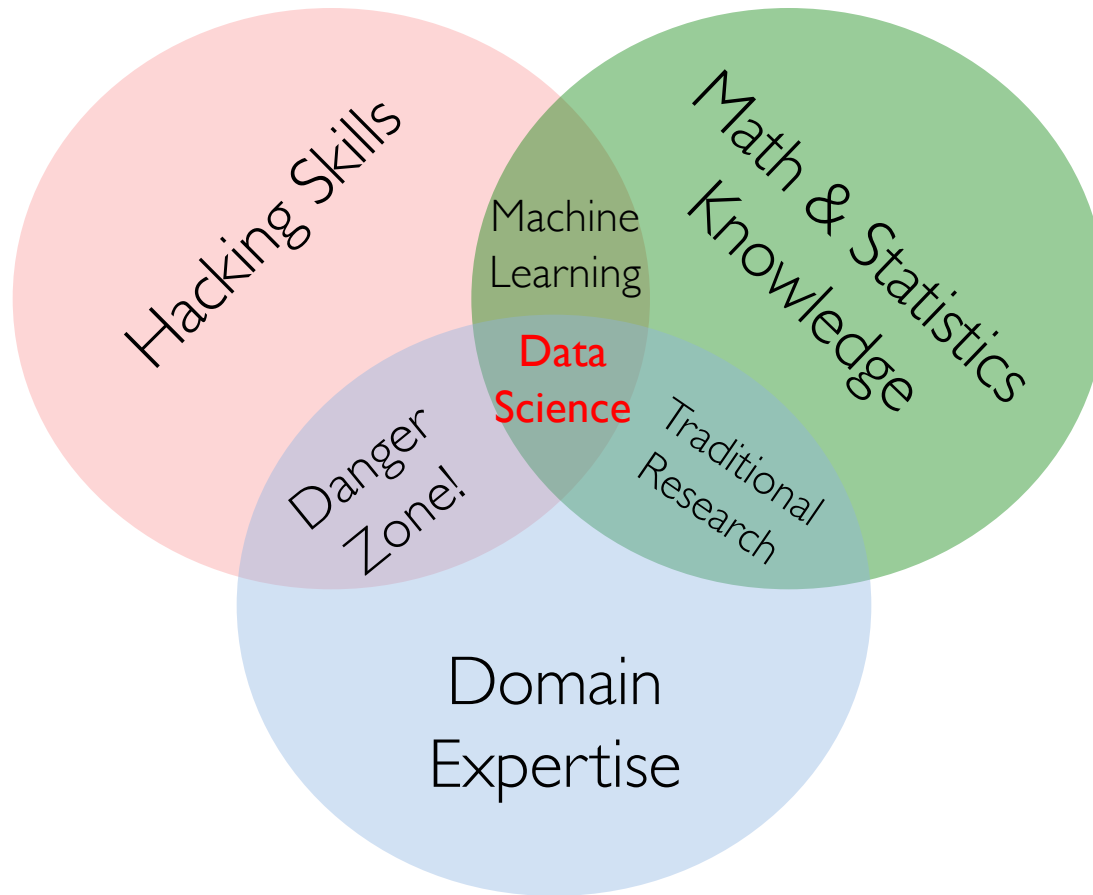


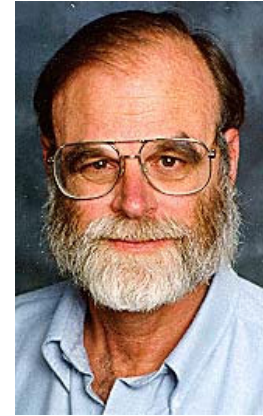
Figure: http://www.opengroup.org/cloud/cloud/cloud_for_business/roi.htm

Data Science – One Definition

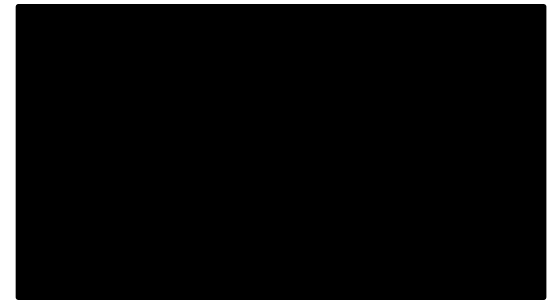


Jim Gray's Model

1. Capture
2. Curate
3. Communicate



Turing award winner

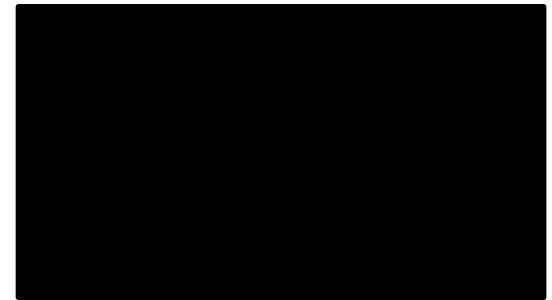


Ben Fry's Model

1. Acquire
2. Parse
3. Filter
4. Mine
5. Represent
6. Refine
7. Interact



Data visualization expert

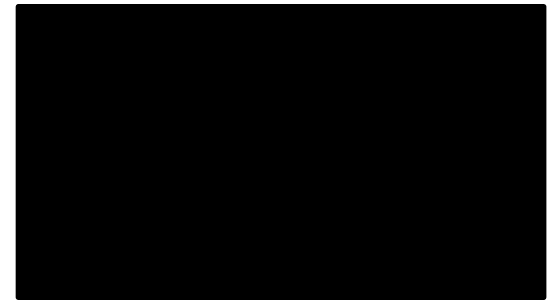


Jeff Hammerbacher's Model

1. Identify problem
2. Instrument data sources
3. Collect data
4. Prepare data (integrate, transform, clean, filter, aggregate)
5. Build model
6. Evaluate model
7. Communicate results

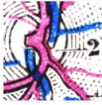








Facebook, Cloudera

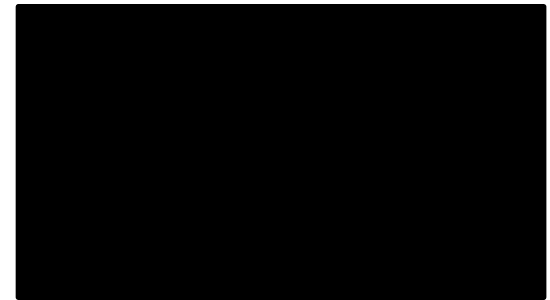


Data Science Competitions

Using Data
Science to find
Data Scientists!

Competition Name	Reward	Teams
 Diabetic Retinopathy Detection Identify signs of diabetic retinopathy in eye images	\$100,000	283
 West Nile Virus Prediction Predict West Nile virus in mosquitos across the city of Chicago	\$40,000	264
 Restaurant Revenue Prediction Predict annual restaurant sales based on objective measurements	\$30,000	2340
 Otto Group Product Classification Challenge Classify products into the correct category	\$10,000	2950
 How Much Did It Rain? Predict probabilistic distribution of hourly rain given polarimetric radar measurements	\$500	282
 ECML/PKDD 15: Taxi Trajectory Prediction (I) Predict the destination of taxi trips based on initial partial trajectories	\$250	72
 ECML/PKDD 15: Taxi Trip Time Prediction (II) Predict the total travel time of taxi trips based on their initial partial trajectories	\$250	35

kaggle



Data Scientist's Practice



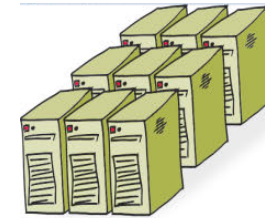
Digging Around
in Data

Clean, prep

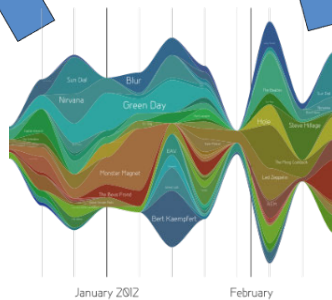
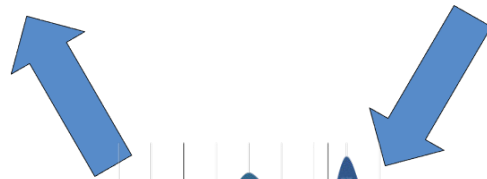


$$\begin{bmatrix} \cos 90^\circ & \sin 90^\circ \\ -\sin 90^\circ & \cos 90^\circ \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

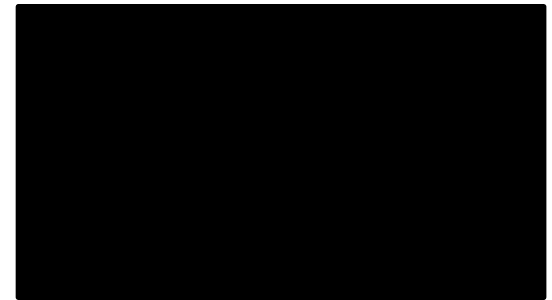
Hypothesize Model



Large Scale
Exploitation



Evaluate
Interpret



Data Science Topics

Data Acquisition

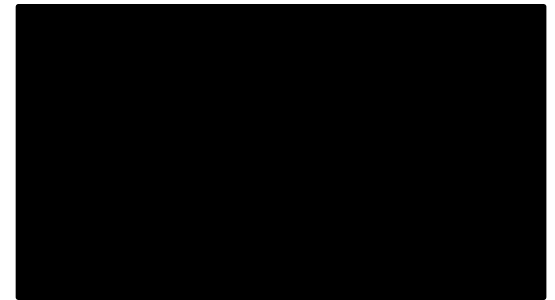
Data Preparation

Analysis

Data Presentation

Data Products

Observation and Experimentation



What's Hard about Data Science?

Overcoming assumptions

Making ad-hoc explanations of data patterns

Not checking enough (validate models, data pipeline integrity, etc.)

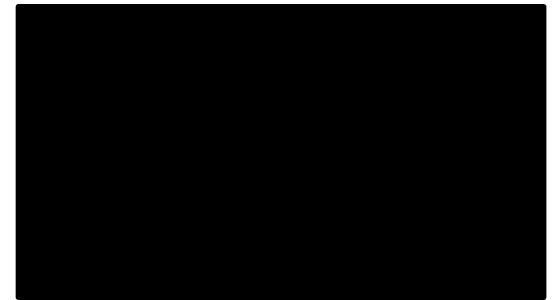
Overgeneralizing

Communication

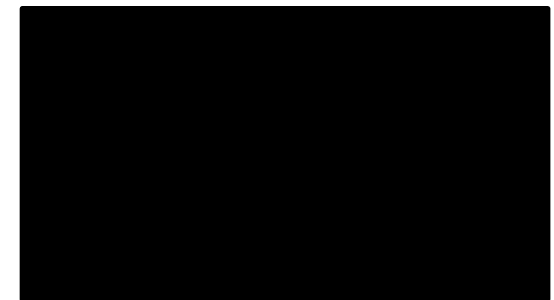
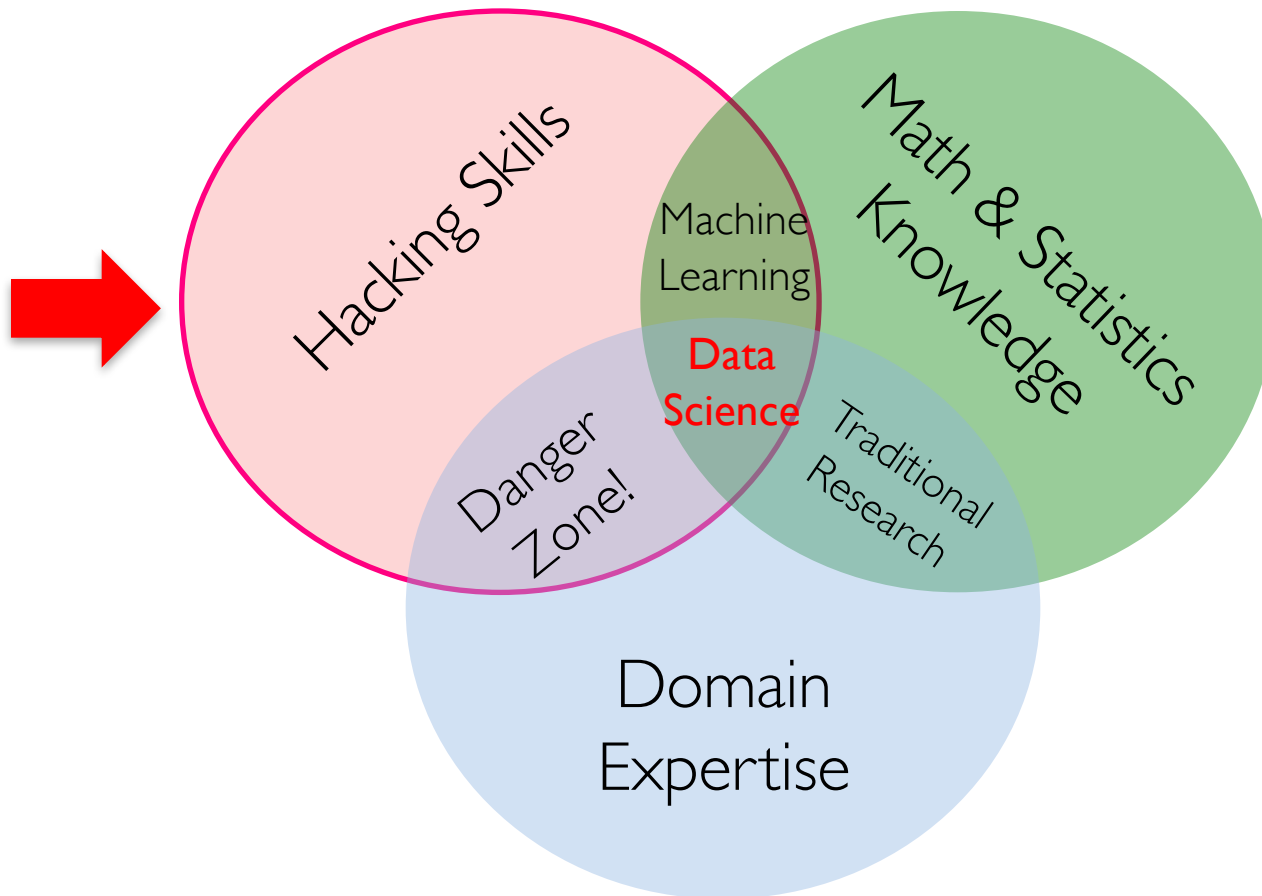
Using statistical tests correctly

Prototype → Production transitions

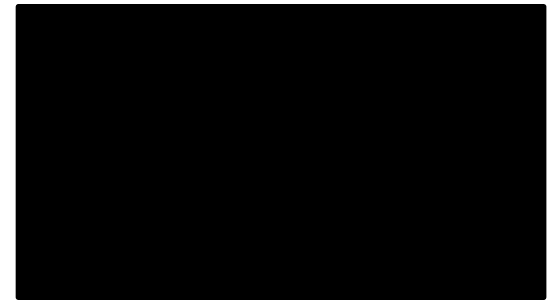
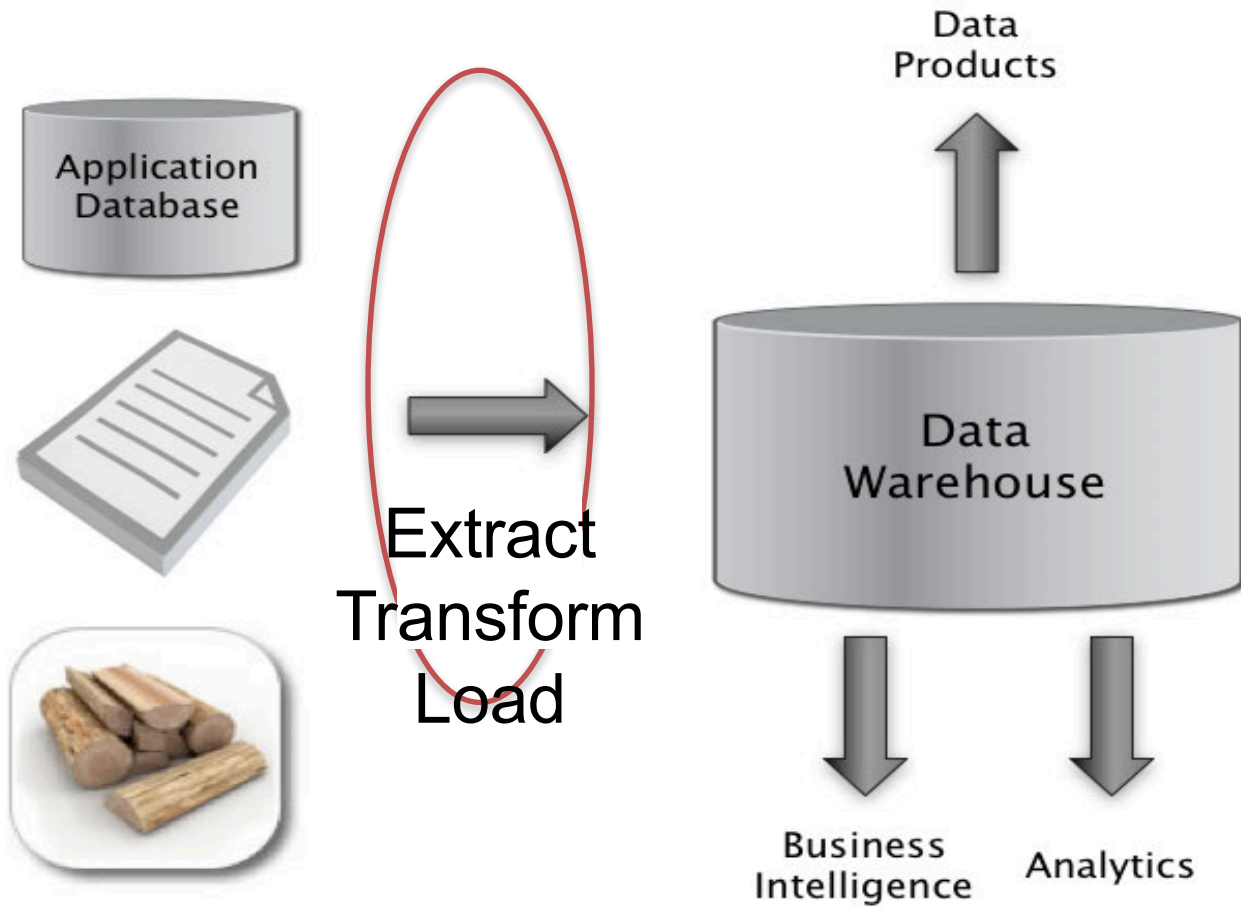
Data pipeline complexity (who do you ask?)



Data Science – One Definition



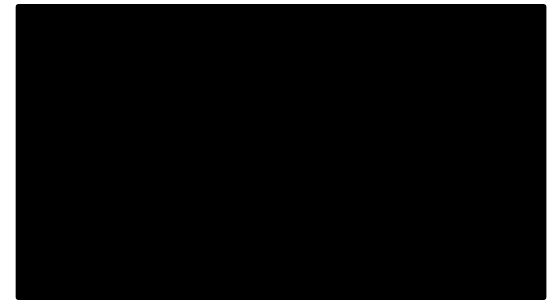
The Big Picture



Data Acquisition (Sources) in Web Companies

Examples from Facebook

- » Application databases
- » Web server logs
- » Event logs
- » Application Programming Interface (API)
server logs
- » Ad and search server logs
- » Advertisement landing page content
- » Wikipedia
- » Images and video

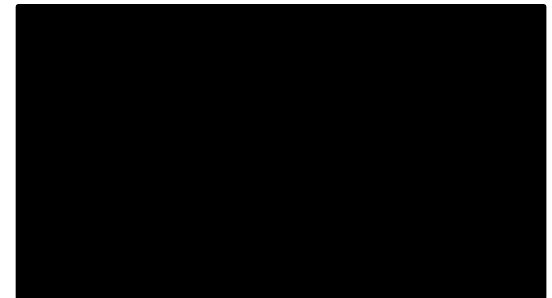


Data Acquisition & Preparation Overview

Extract, Transform, Load (ETL)

- » We need to **extract** data from the **source(s)**
- » We need to **load** data into the **sink**
- » We need to **transform** data at the source, sink, or in a **staging area**

- » Sources: file, database, event log, web site, [Hadoop Distributed FileSystem \(HDFS\)](#), ...
- » Sinks: [Python](#), [R](#), [SQLite](#), [NoSQL store](#), files, [HDFS](#), [Relational DataBase Management System \(RDBMS\)](#), ...



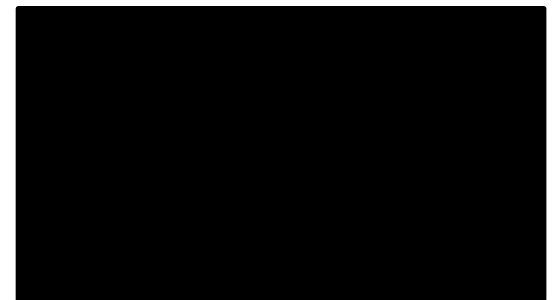
Data Acquisition & Preparation Process Model

The construction of a new data preparation process is done in many phases

- » Data **characterization**
- » Data **cleaning**
- » Data **integration**

We must efficiently move data around in space and time

- » Data **transfer**
- » Data **serialization** and **deserialization** (for files or network)



Data Acquisition & Preparation Workflow

The transformation **pipeline** or **workflow** often consists of many steps

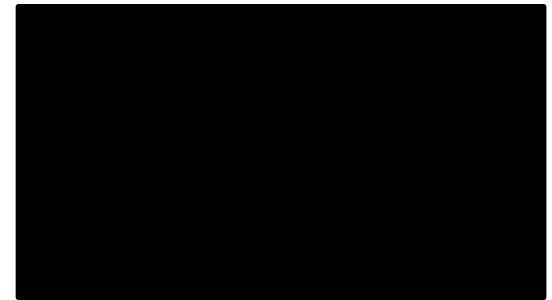
- » For example: Unix pipes and filters
- » `cat data_science.txt | wc | mail -s "word count" myname@some.com`

If a workflow is to be used more than once, it can be **scheduled**

- » Scheduling can be time-based or event-based
- » Use publish-subscribe to register interest (e.g., Twitter feeds)

Recording the execution of a workflow is known as capturing **lineage** or provenance

- » Spark's **DataFrames** do this for you automatically



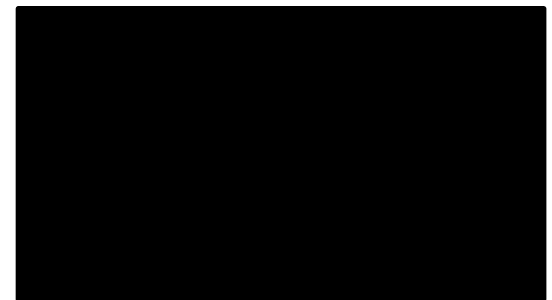
Impediments to Collaboration

The diversity of tools and programming/scripting languages makes it hard to share

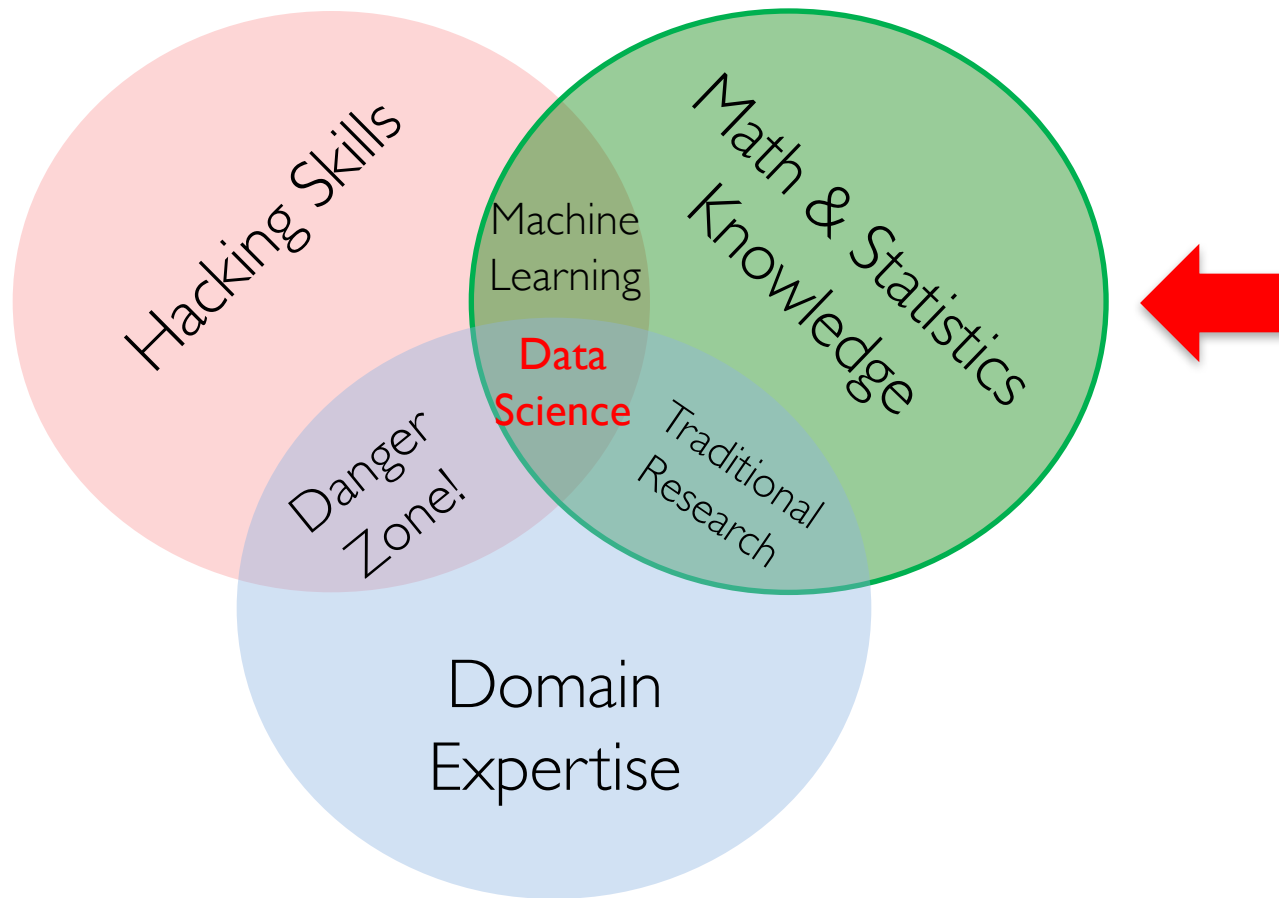
Finding a script or computed result is often harder than just writing the program from scratch!

» Question: How could we fix this?

View that most analysis work is
“throw away”



Data Science – One Definition

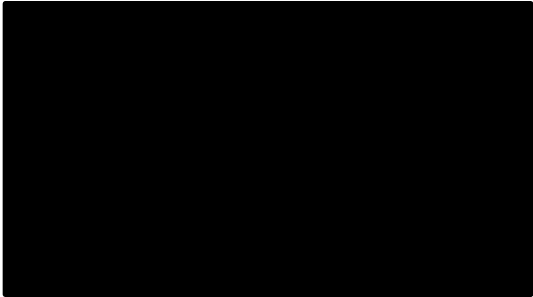


Descriptive vs. Inferential Statistics

Descriptive:

- » E.g., Median – describes data but can't be generalized beyond that
- » We will talk about Exploratory Data Analysis in this lecture

Inferential:

- » E.g., t-test – enables inferences about population beyond our data
 - » Techniques leveraged for Machine Learning and Prediction
 - » Making conclusions based on data in random samples
- 

Examples of Business Questions

Simple (descriptive) Stats

- » “Who are the most profitable customers?”

Hypothesis Testing

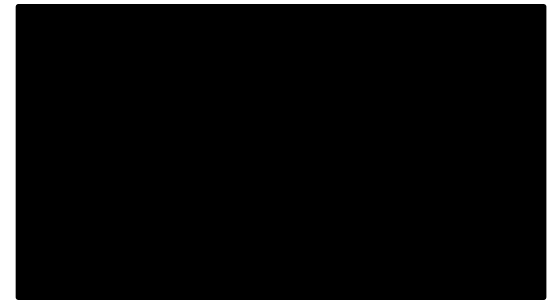
- » “Is there a difference in value to the company of these customers?”

Segmentation/Classification

- » What are the common characteristics of these customers?

Prediction

- » Will this new customer become a profitable customer?
- » If so, how profitable?



Applying Techniques

Most business questions are causal

» What would happen if I show this ad?

Easier to ask correlational questions

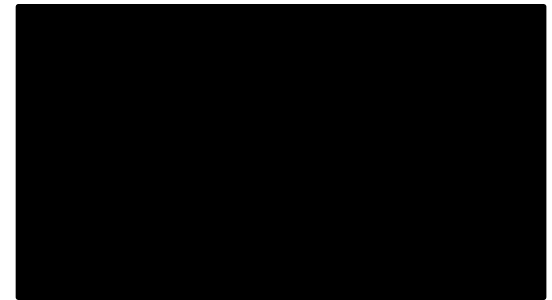
» What happened in this past when I showed this ad?

Supervised Learning: Classification and Regression

Unsupervised Learning: Clustering and Dimension reduction

Note: UL often used inside a larger SL problem

» E.g., auto-encoders for image recognition neural nets



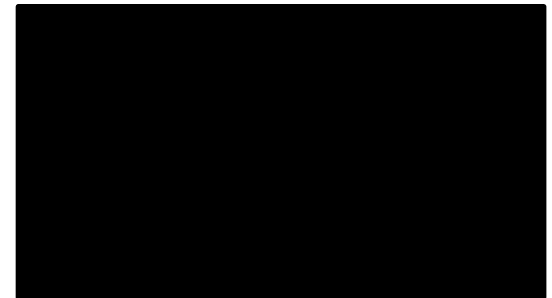
Learning Techniques

Supervised Learning:

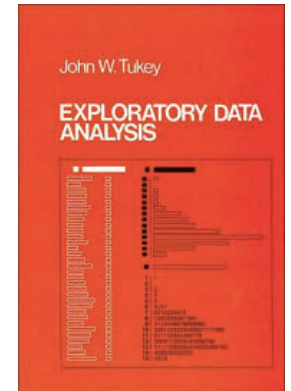
- » kNN (k Nearest Neighbors)
- » Naive Bayes
- » Logistic Regression
- » Support Vector Machines
- » Random Forests

Unsupervised Learning:

- » Clustering
- » Factor Analysis
- » Latent Dirichlet Allocation



Exploratory Data Analysis (1977)



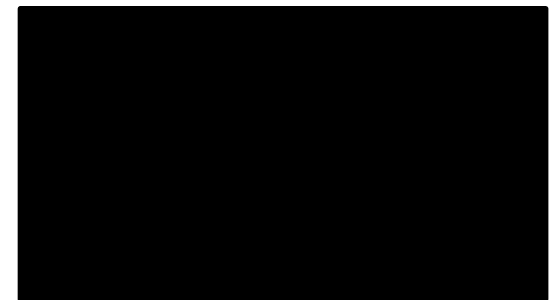
Based on insights developed at Bell Labs in 1960's

Techniques for visualizing and summarizing data

What can the data tell us? (vs “confirmatory” data analysis)

Introduced many basic techniques:

- » 5-number summary, box plots, stem and leaf diagrams,...



The “R” Language

Evolution of the “S” language developed at Bell labs for EDA

Idea: allow interactive exploration and visualization of data

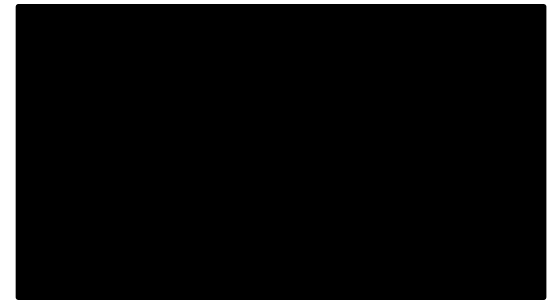
Preferred language for statisticians, used by many data scientists

Features:

- » The most comprehensive collection of statistical models and distributions
- » CRAN: large resource of open source statistical models

Supported by Apache Spark:

- » <http://spark.apache.org/docs/latest/sparkr.html>

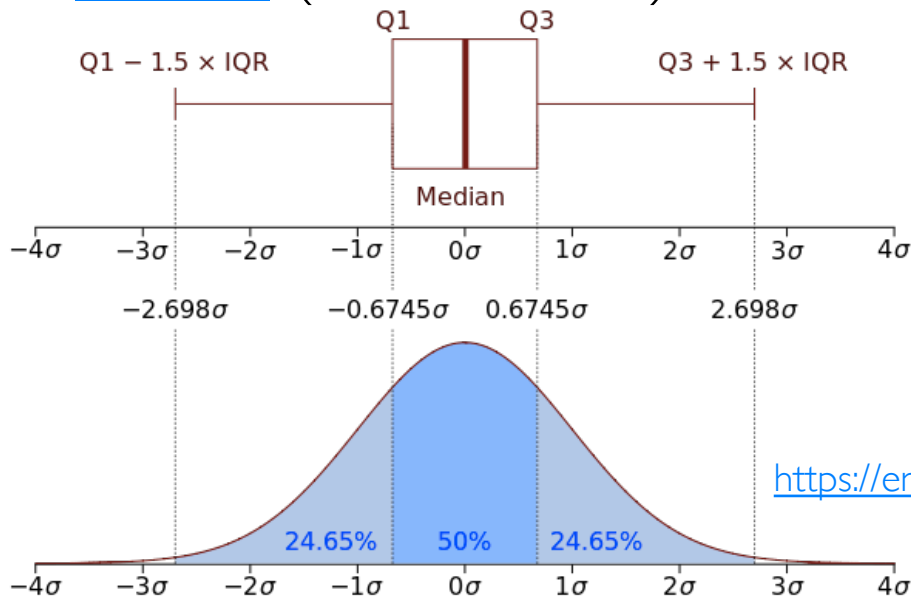


5-Number Summary Statistic

Summary statistic provides:

- » [minimum](#) and [maximum](#) (smallest and largest observations)
- » [lower quartile](#) (Q1) and [upper quartile](#) (Q3)
- » [median](#) (middle value)

More robust to skewed and long-tailed distributions



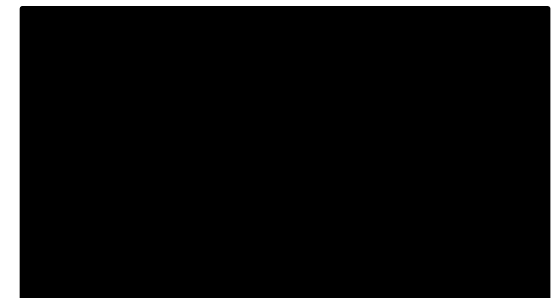
<https://en.wikipedia.org/wiki/User:Jhguch>

https://en.wikipedia.org/wiki/Five-number_summary

The Trouble with Summary Statistics

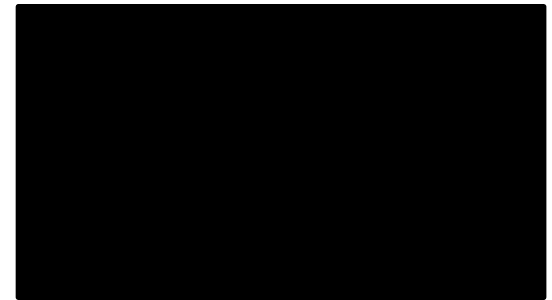
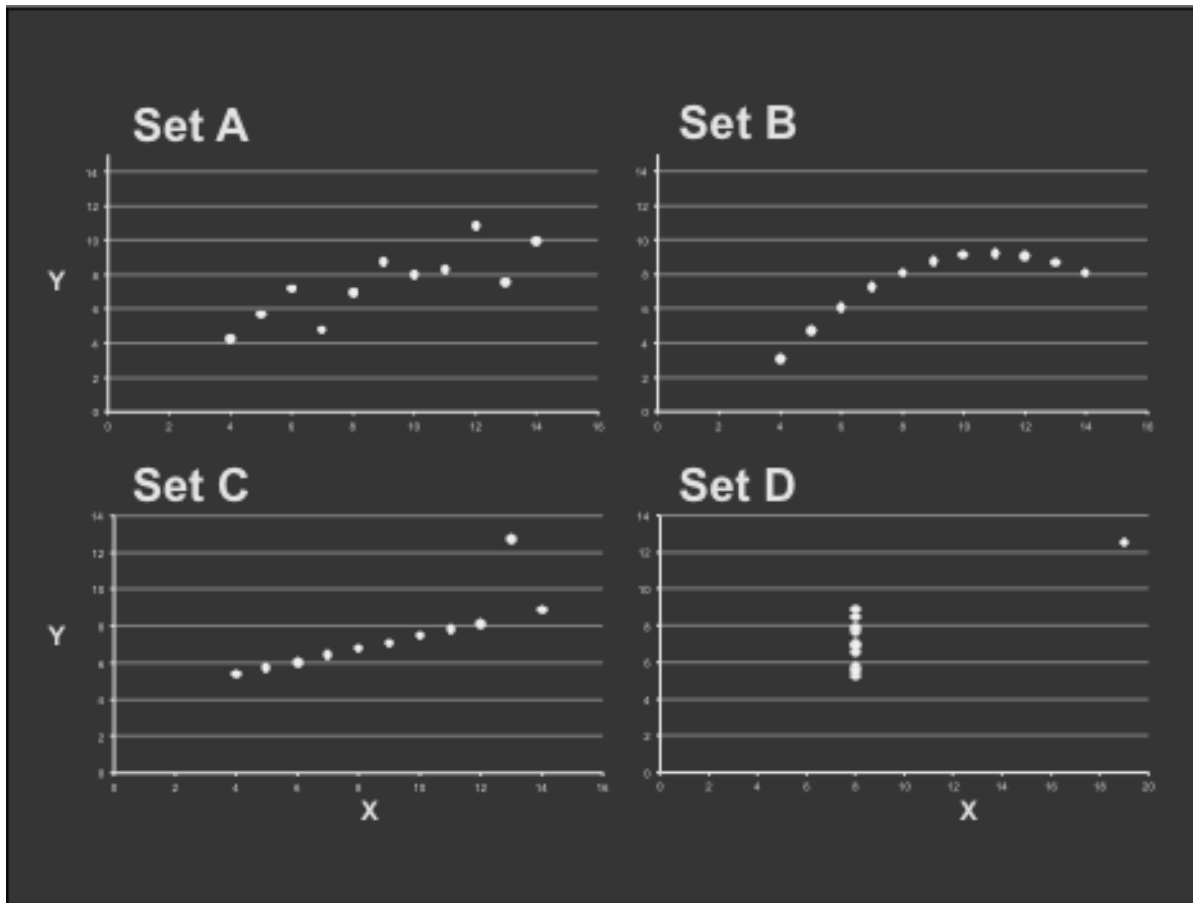
Set A		Set B		Set C		Set D	
X	Y	X	Y	X	Y	X	Y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.11	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

Property in each set	Value
Mean of x	9
Sample variance of x	11
Mean of y	7.50
Sample variance of y	4.122
Linear Regression	$y = 3 + 0.5x$

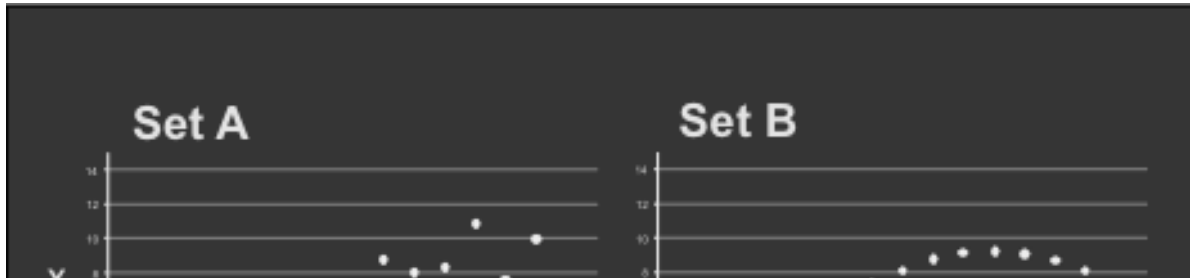


Anscombe's Quartet 1973

Looking at the Data

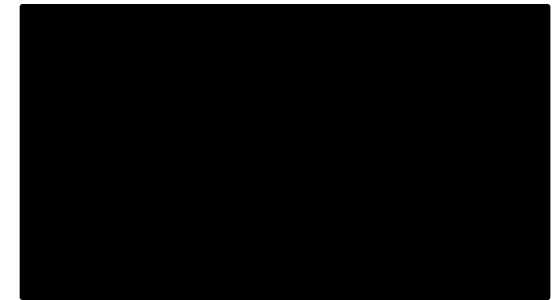
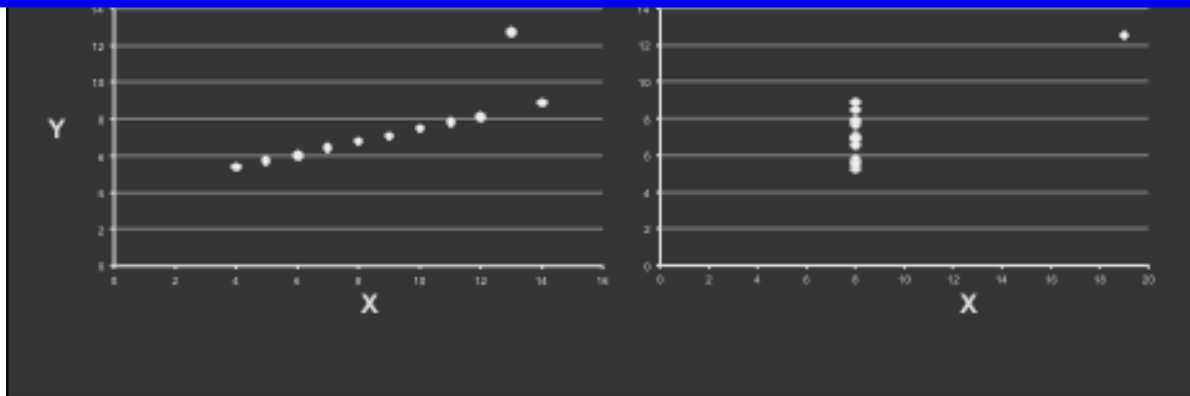


Looking at the Data



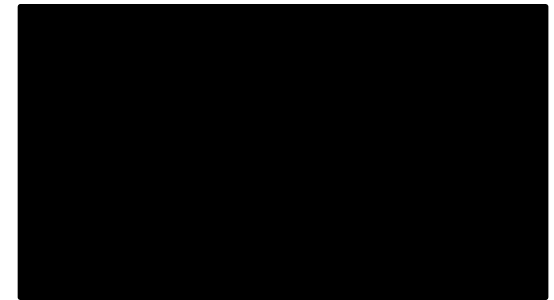
Takeaways:

- Important to look at data graphically before analyzing it
- Basic statistics properties often fail to capture real-world complexities



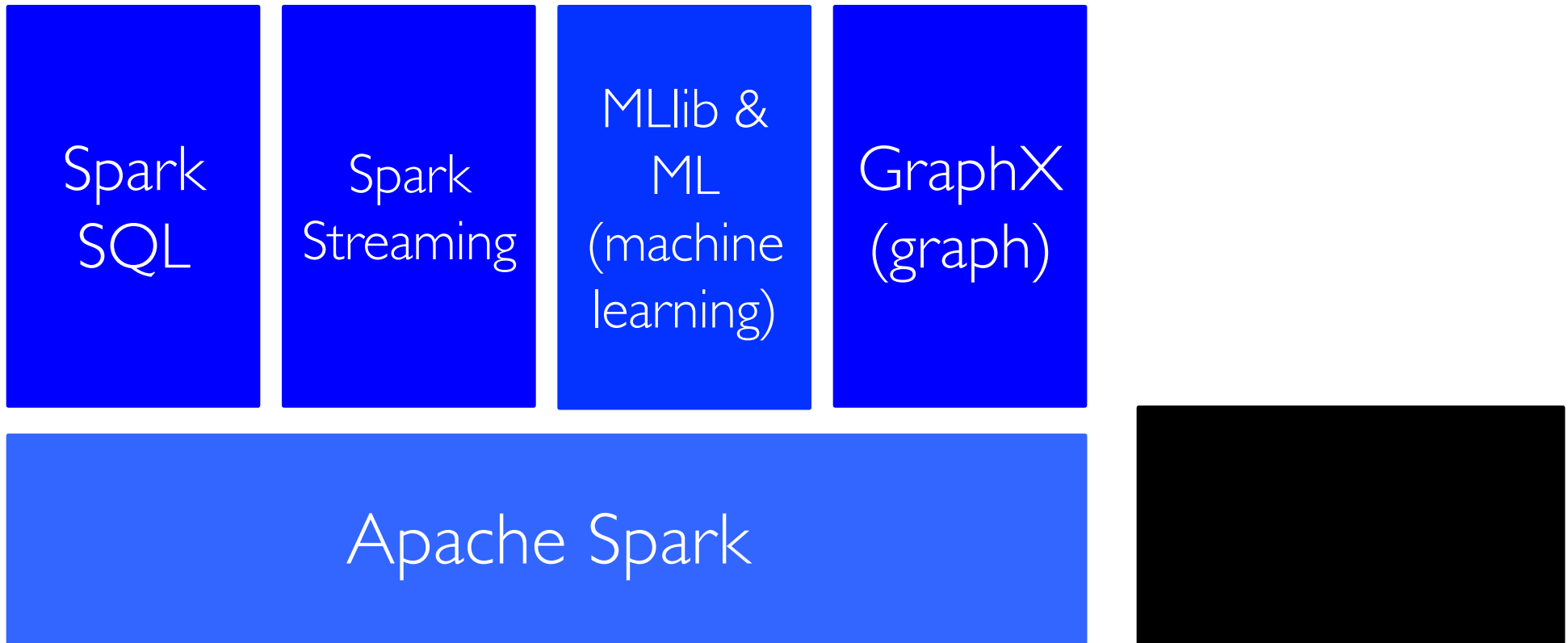
Data Presentation

Data Art – Visualizing Friendships

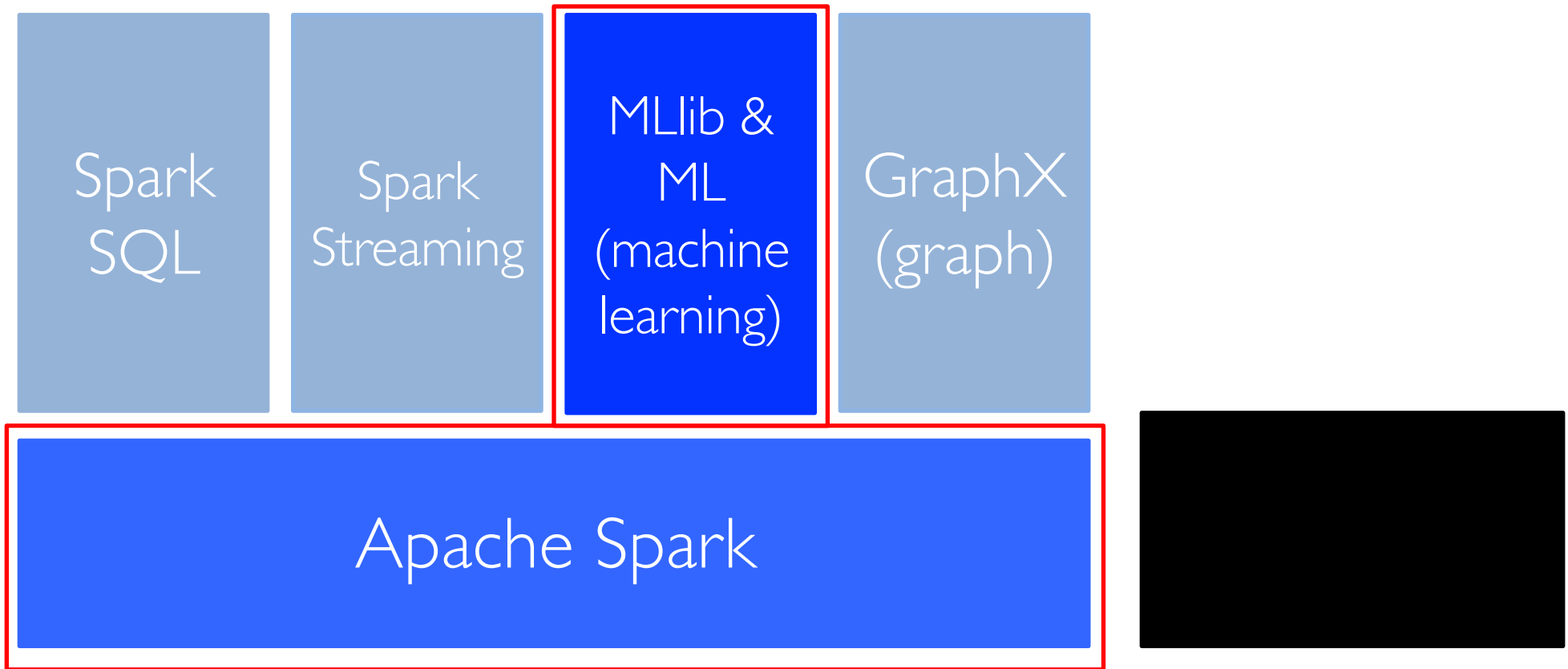


https://www.facebook.com/note.php?note_id=469716398919

Apache Spark Components



Apache Spark Components



Spark's Machine Learning Toolkit

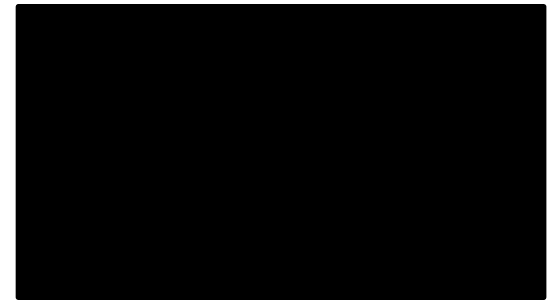
[mllib](#) and [ML Pipelines](#): scalable, distributed ML libraries

- » Scikit-learn like ML toolkit, Interoperates with [NumPy](#)
- » Pipelines: tools for constructing, evaluating, and tuning ML Pipelines
- » Persistence: saving and load algorithms, models, and Pipelines

Classification, regression, clustering, and collaborative filtering

Featurization: feature extraction, selection, transformation, dimensionality reduction

Utilities: linear algebra, statistics, data handling, etc.



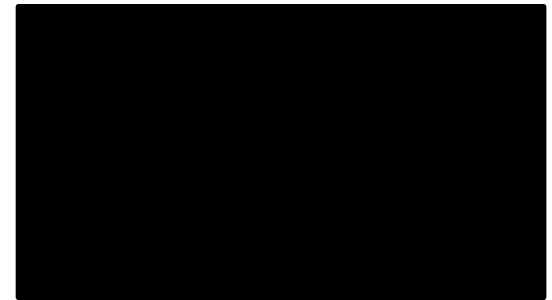
Lab: Regression

Goal: predict gas-fired power plant's power output based on environmental variables

Visualize: Graph data to understand it

Model: Use several ML regression algorithms to explore correlations and prediction

Evaluate: Measure accuracy of models



Lab: Regression

Goal: predict gas-fired power plant's power output based on environmental variables

Visualize: Graph data to understand it

Model: Use several ML regression algorithms to explore correlations and prediction

Evaluate: Measure accuracy of models

