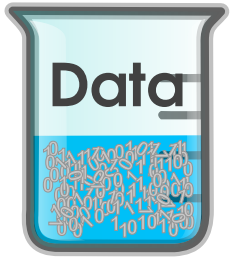
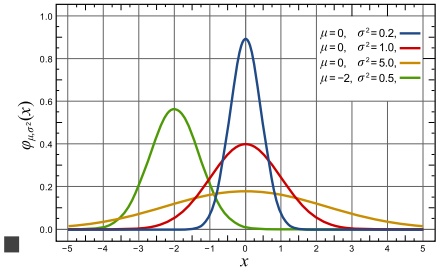
Data Science,
Big Data,

and other
buzz words

1. 

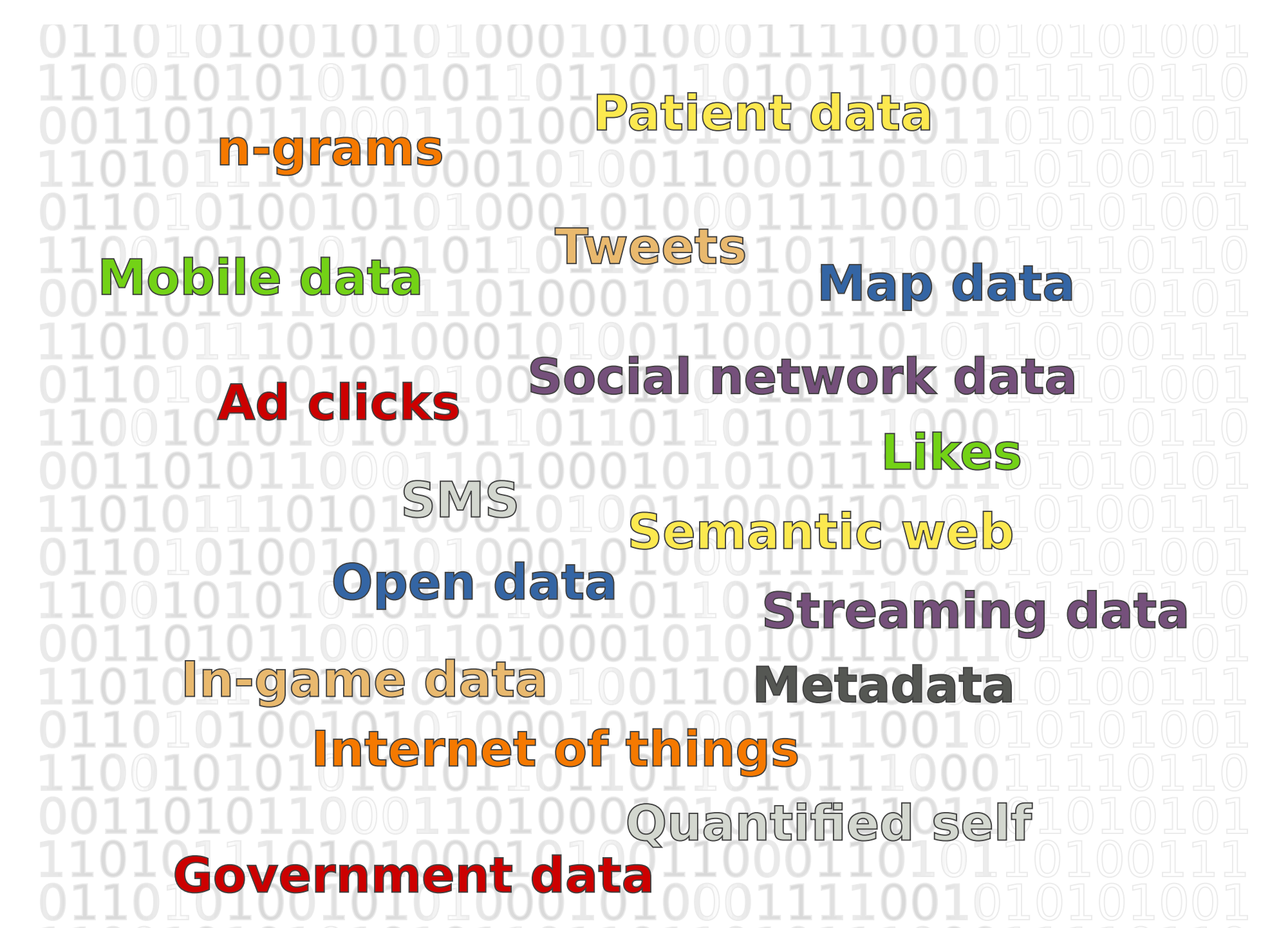
2. 

3. ? ? ?

4. Profit!

Roy Keyes
Zefs Data Science

Steve Koch
UNM Libraries


Patient data
n-grams
Tweets
Mobile data
Map data
Social network data
Ad clicks
Likes
SMS
Semantic web
Open data
Streaming data
In-game data
Metadata
Internet of things
Quantified self
Government data

BIG DATA

- How big is big?
- Bigger than a single machine(?)
- What makes it interesting?

BIG DATA

- How big is big?
- Bigger than a single machine(?)
- What makes it interesting?

BIG DATA

- How big is big?
- Bigger than a single machine(?)
- What makes it interesting?

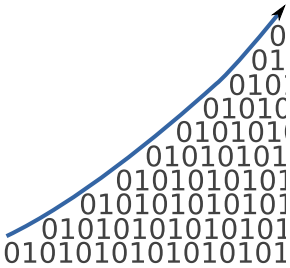
BIG DATA

- How big is big?
- Bigger than a single machine(?)
- What makes it interesting?

How did we get to "Big Data"?

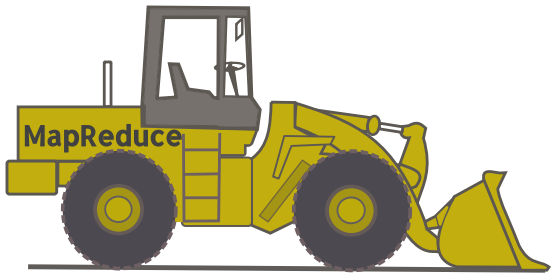
Exponential drop in storage costs.

- 1GB: 2000 = \$10 $\rightarrow$ 2010 = \$0.10



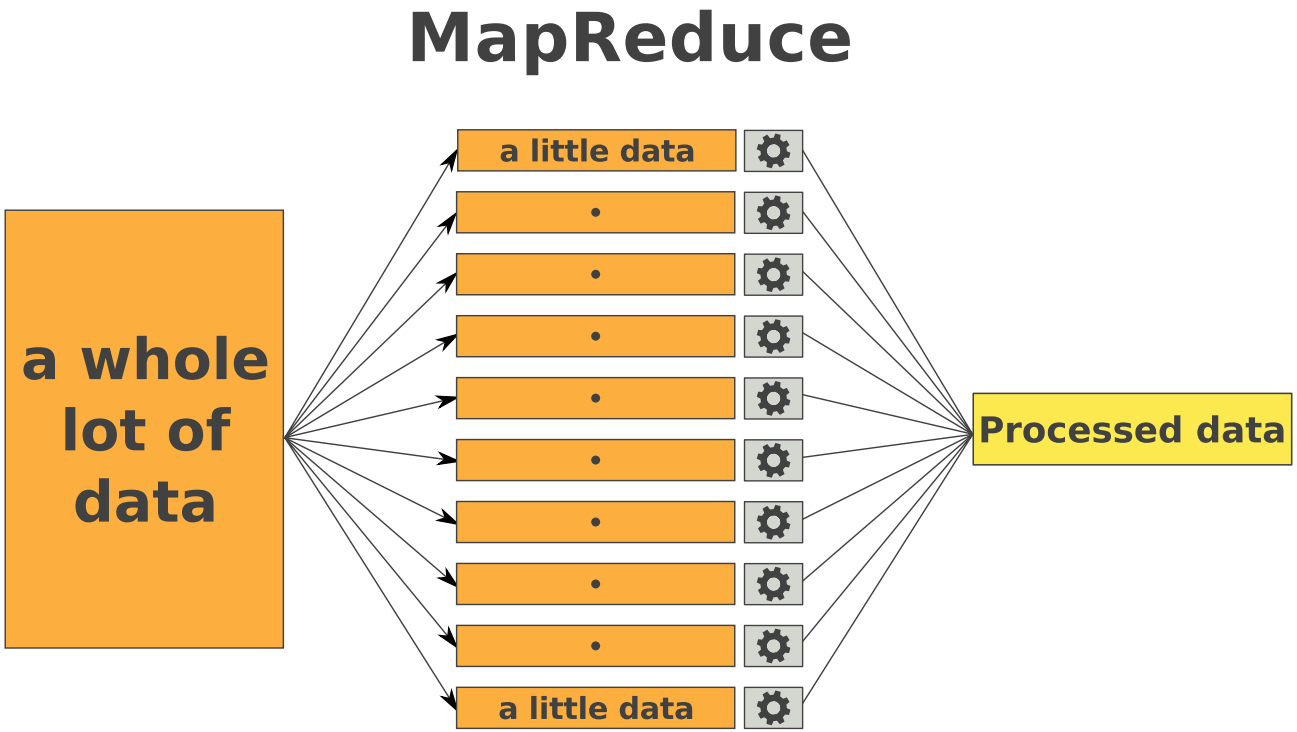
Much more data available.

- Social network and smart phone data
- Global internet access up 500% in past decade



Large scale data processing tools developed.

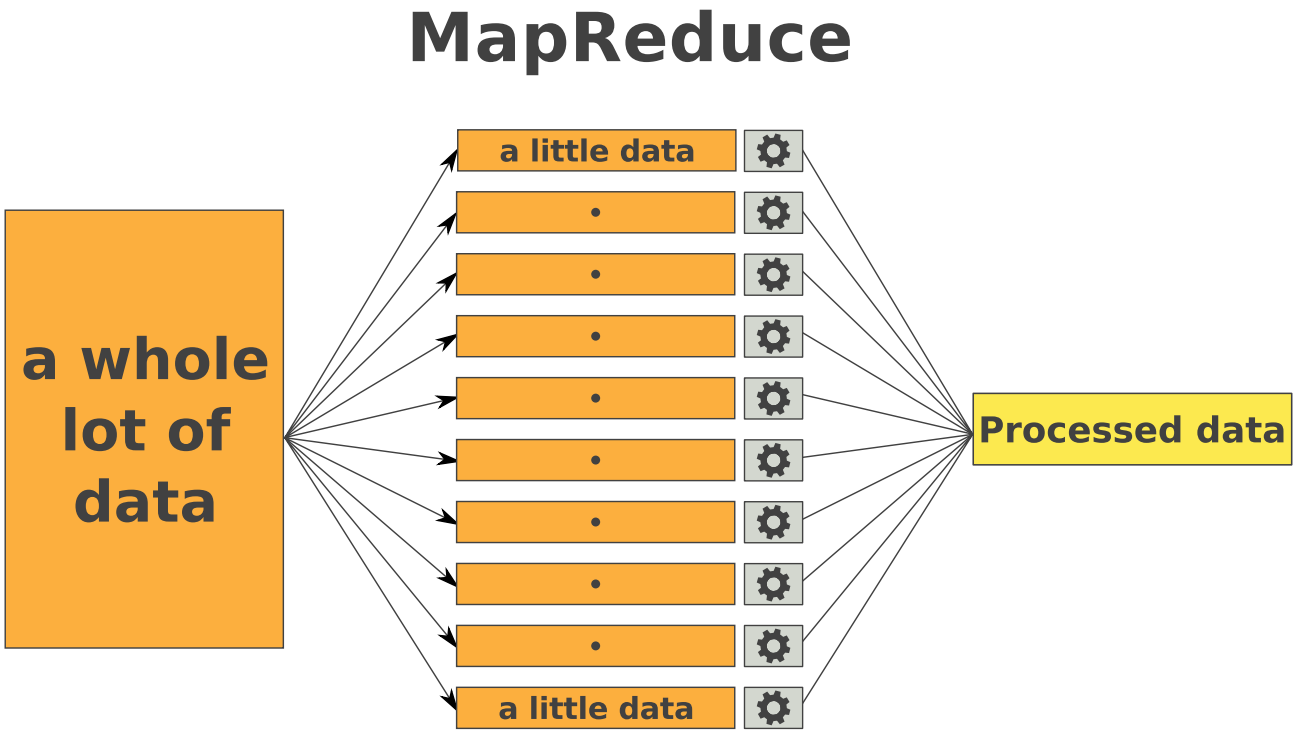
MapReduce, Hadoop, and friends



Hadoop

- Open source implementation of MapReduce from Yahoo (2005).
- Hadoop + commercial cloud = big data processing.
- Basis for an ecosystem of tools.

MapReduce, Hadoop, and friends



Hadoop

- Open source implementation of MapReduce from Yahoo (2005).
- Hadoop + commercial cloud = big data processing.
- Basis for an ecosystem of tools.

(Big) Data in action

Recommendations

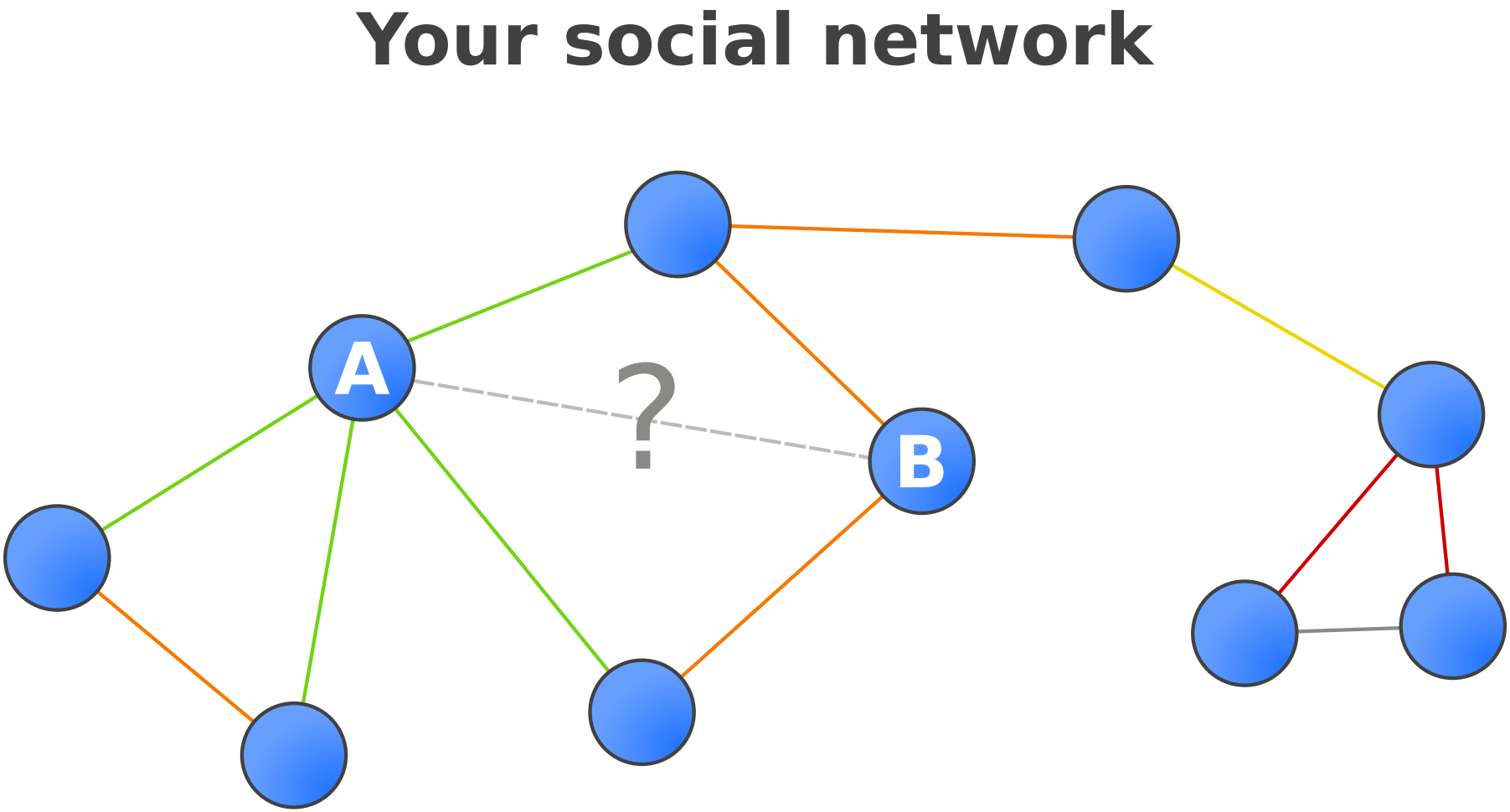
- Recommends "similar" movies.
- Uses customer ratings to infer statistical connections.
- Algorithms scale much better than humans.

Other product recommendation systems

- Retail: Amazon
- Radio: Last.fm, Pandora, Spotify
- Apps: iTunes, Google Play

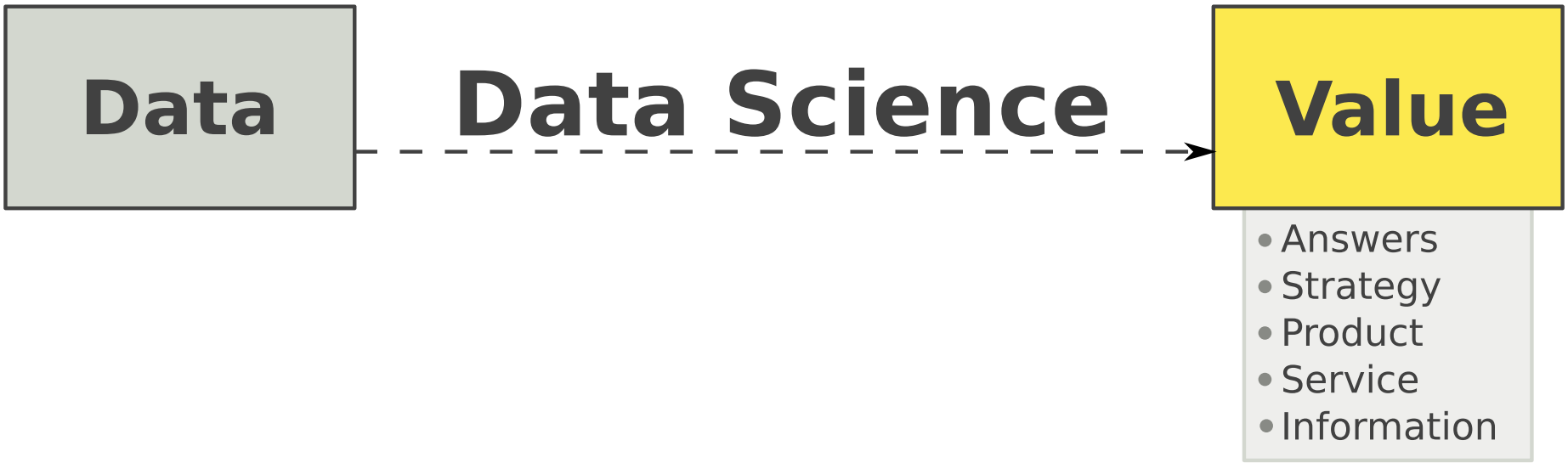
(Big) Data in action

You might know



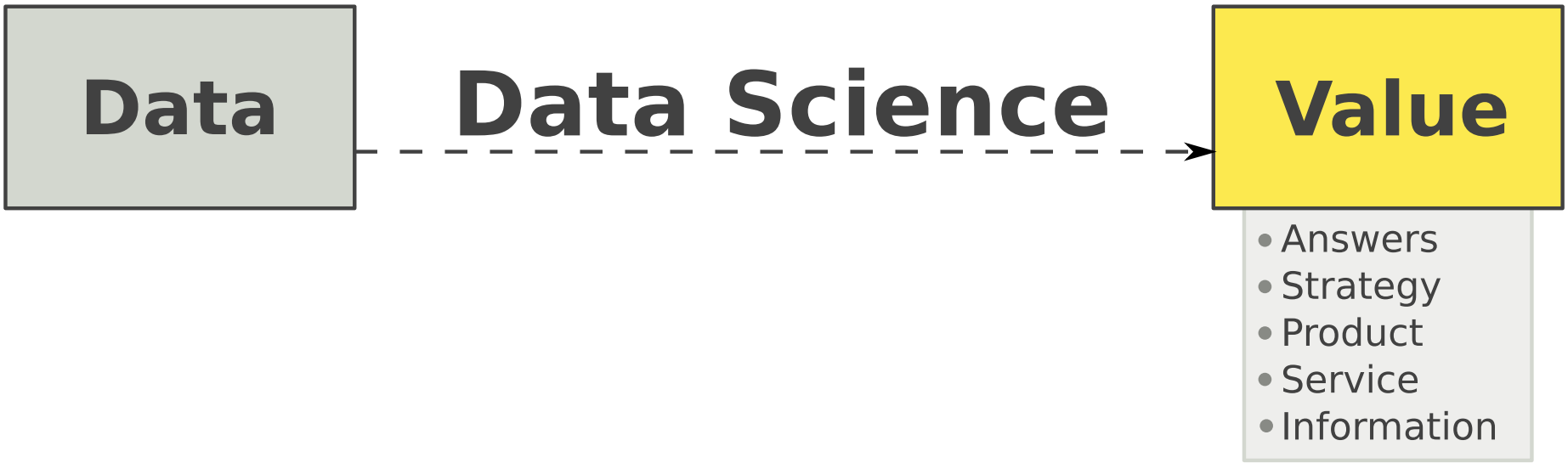
- As seen on LinkedIn, Facebook, Google+, Twitter, etc.

Data Science



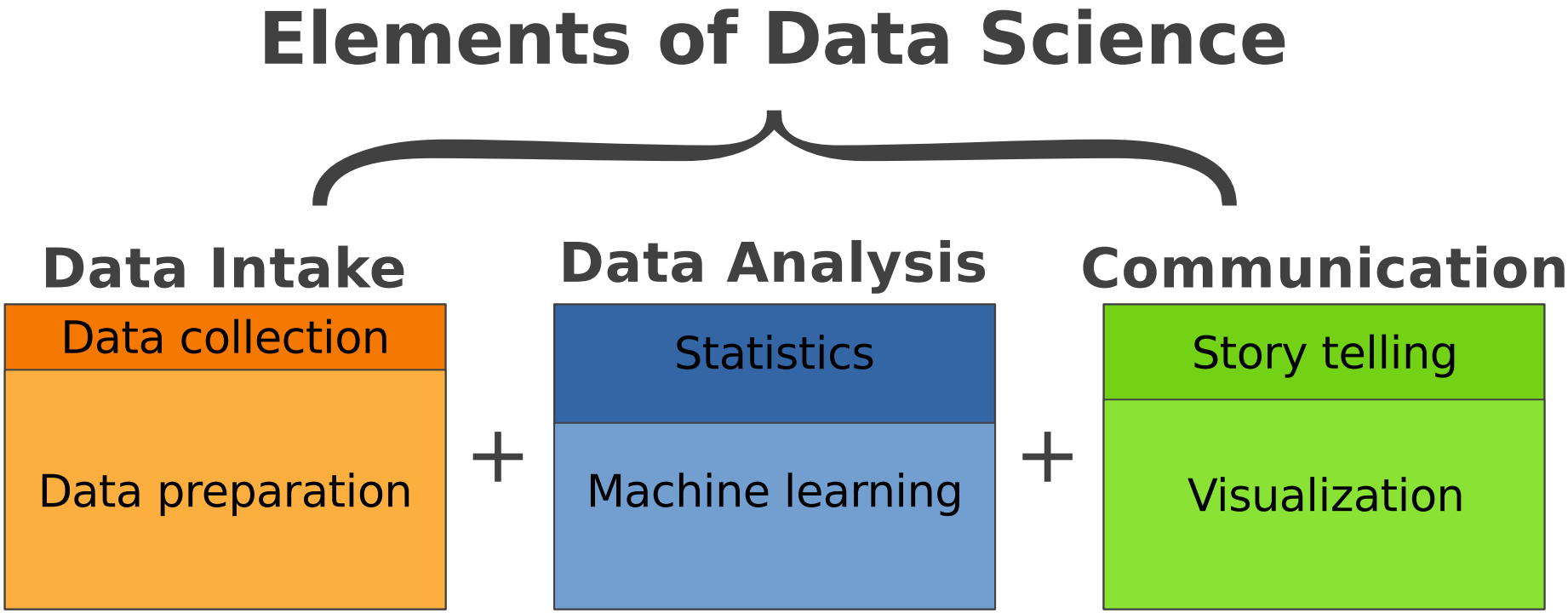
- Data science is the process of extracting value from data.
- Data science is a buzz word.

Data Science

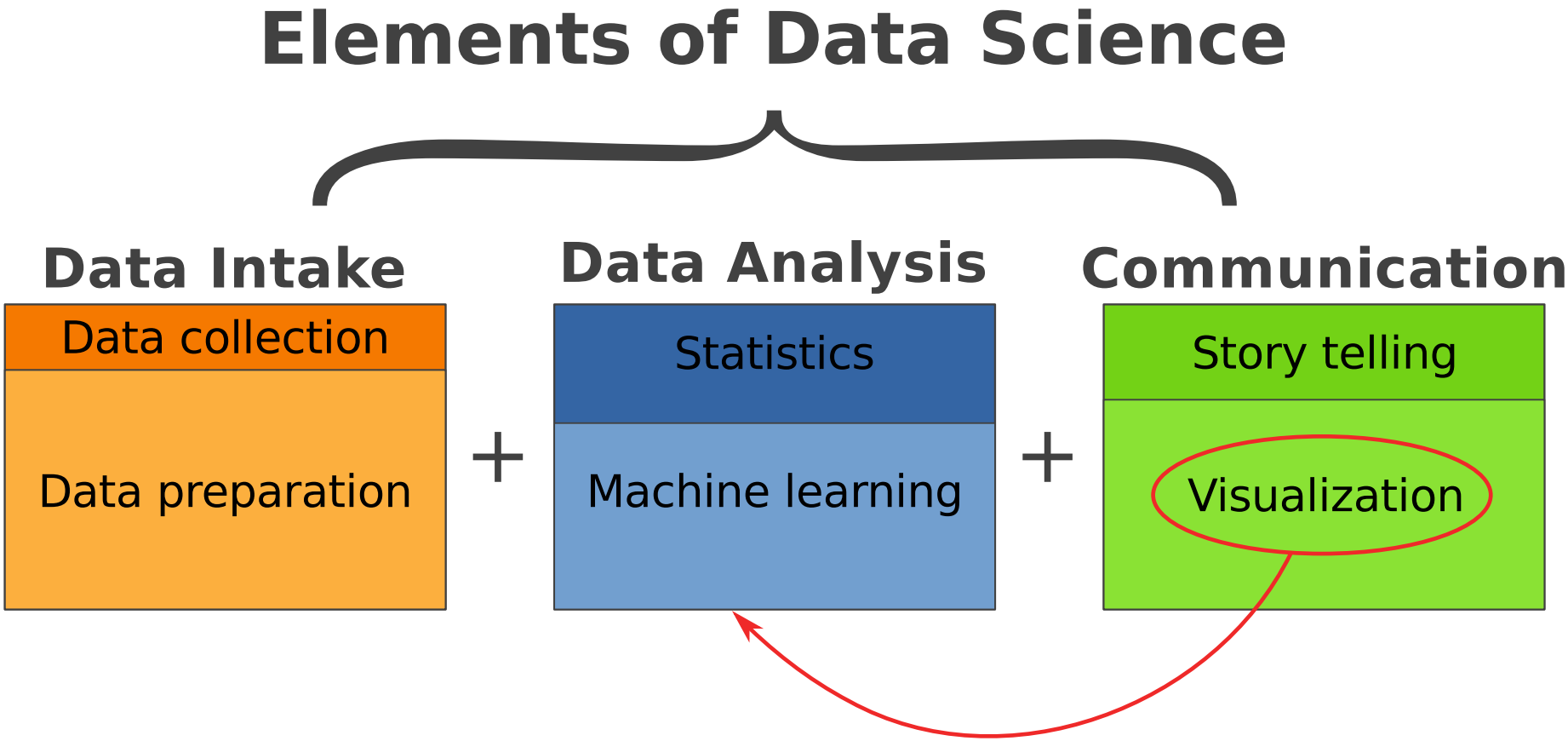


- Data science is the process of extracting value from data.
- Data science is a buzz word.

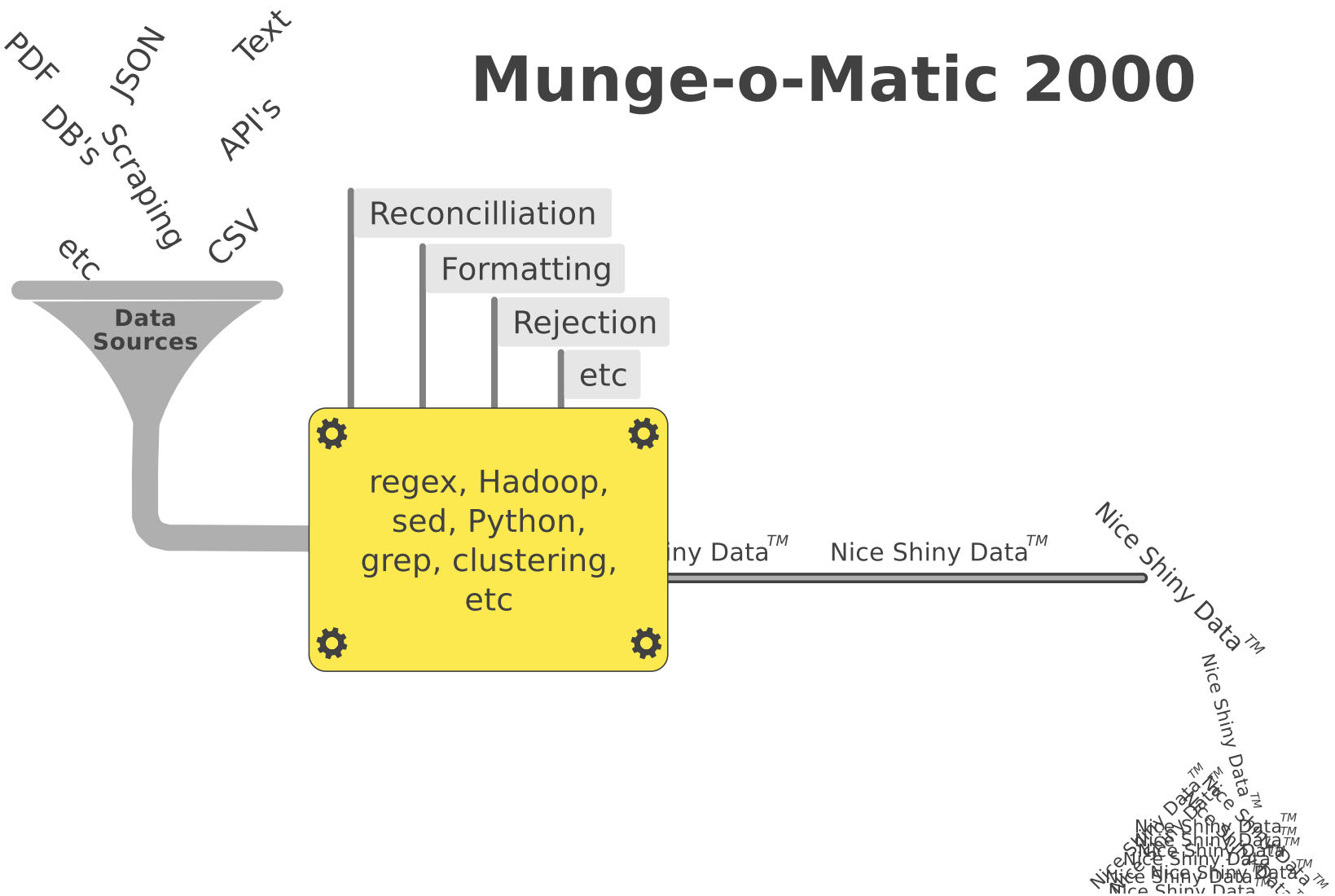
Data Science



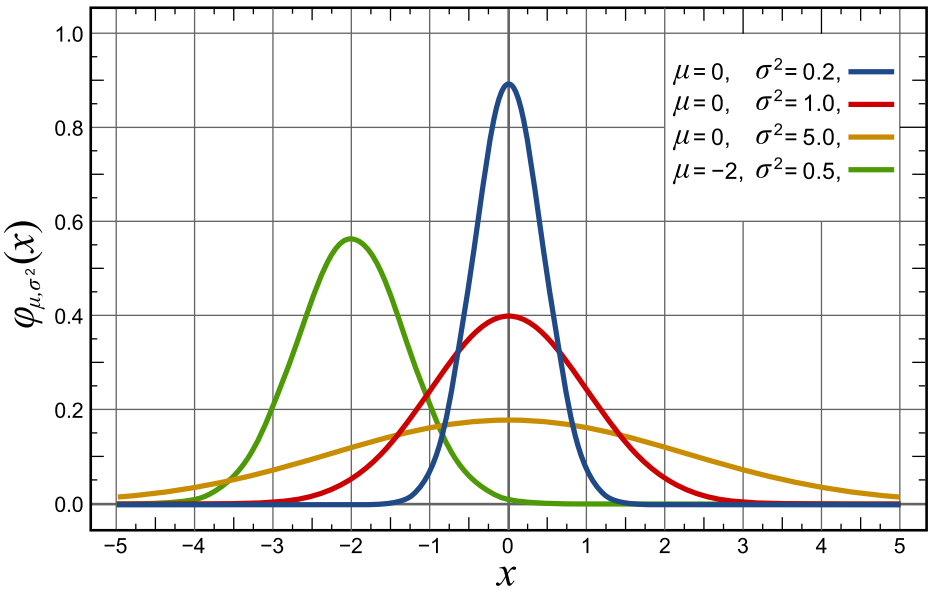
Data Science



Data collection and preparation

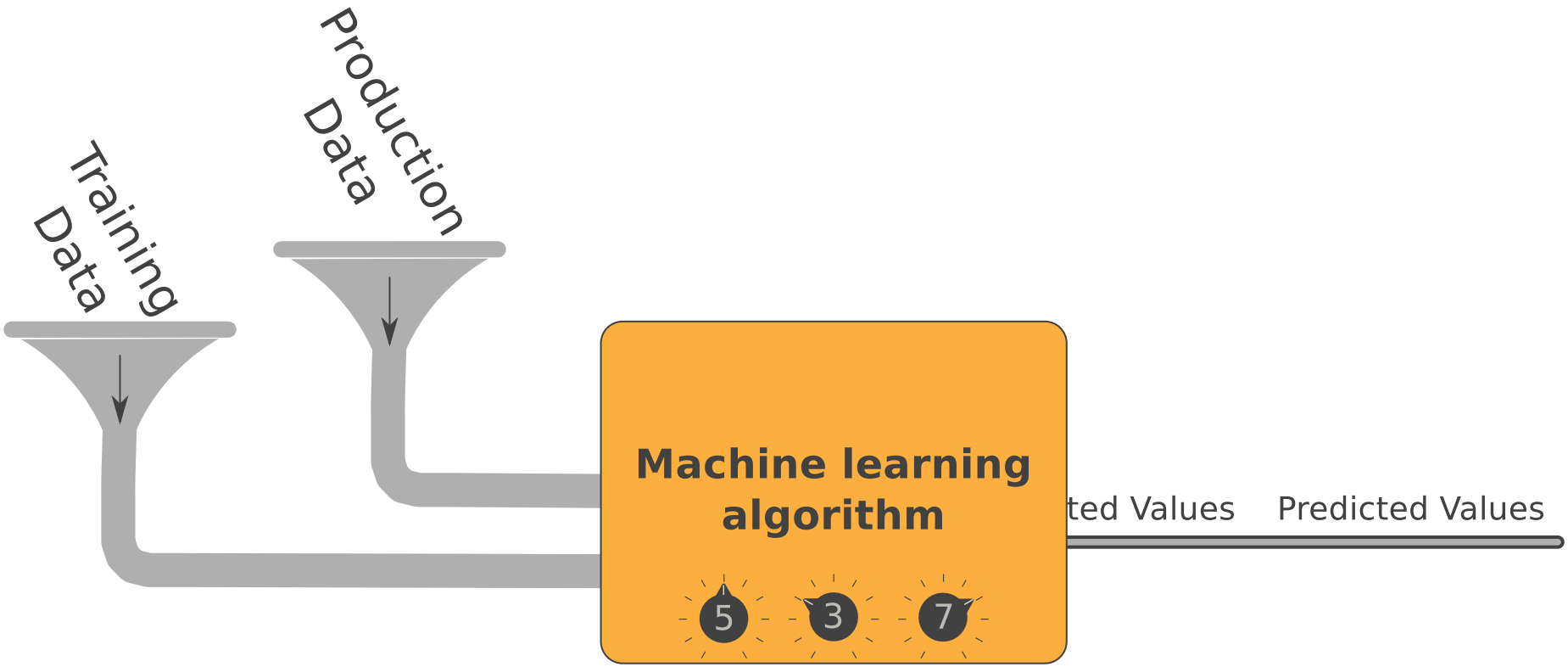


Statistics



- Traditional statistics heavily employed.
- Hypothesis testing.
- Experimental design.
- Regression analysis.
- Statistical modeling.
- Bayesian statistics.

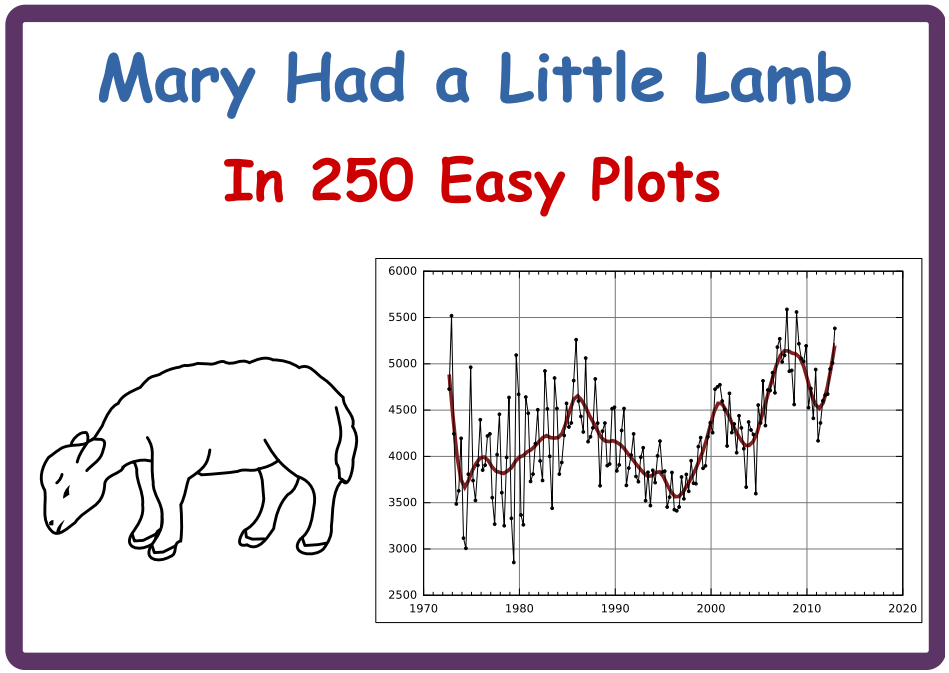
Machine Learning



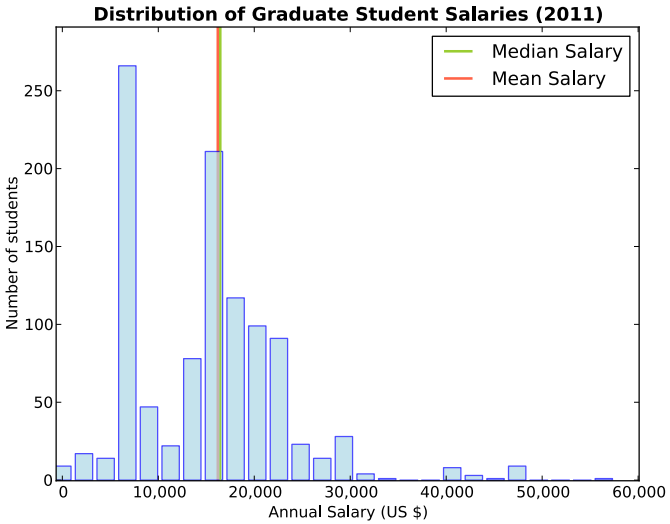
- Value prediction via AI.
  - Regression, categorization, clustering.
- Random forests, support vector machines, neural networks, etc.
- More training data = better.
- Less concerned with explanatory power.

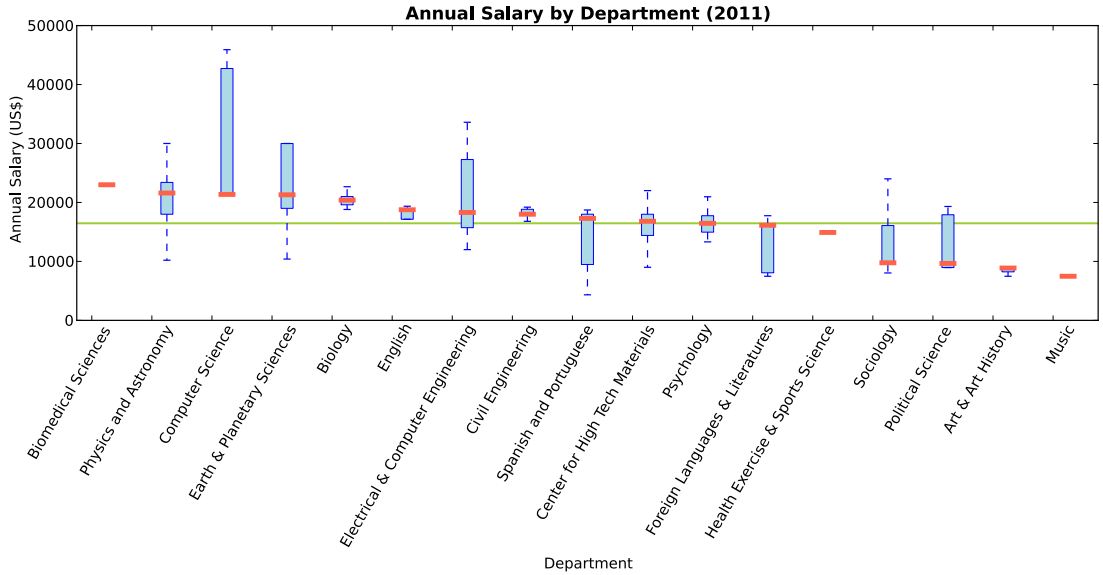
Story Telling

- Presenting results in an audience-appropriate and understandable way.
- Providing results and clear limitations on those results.



Visualization

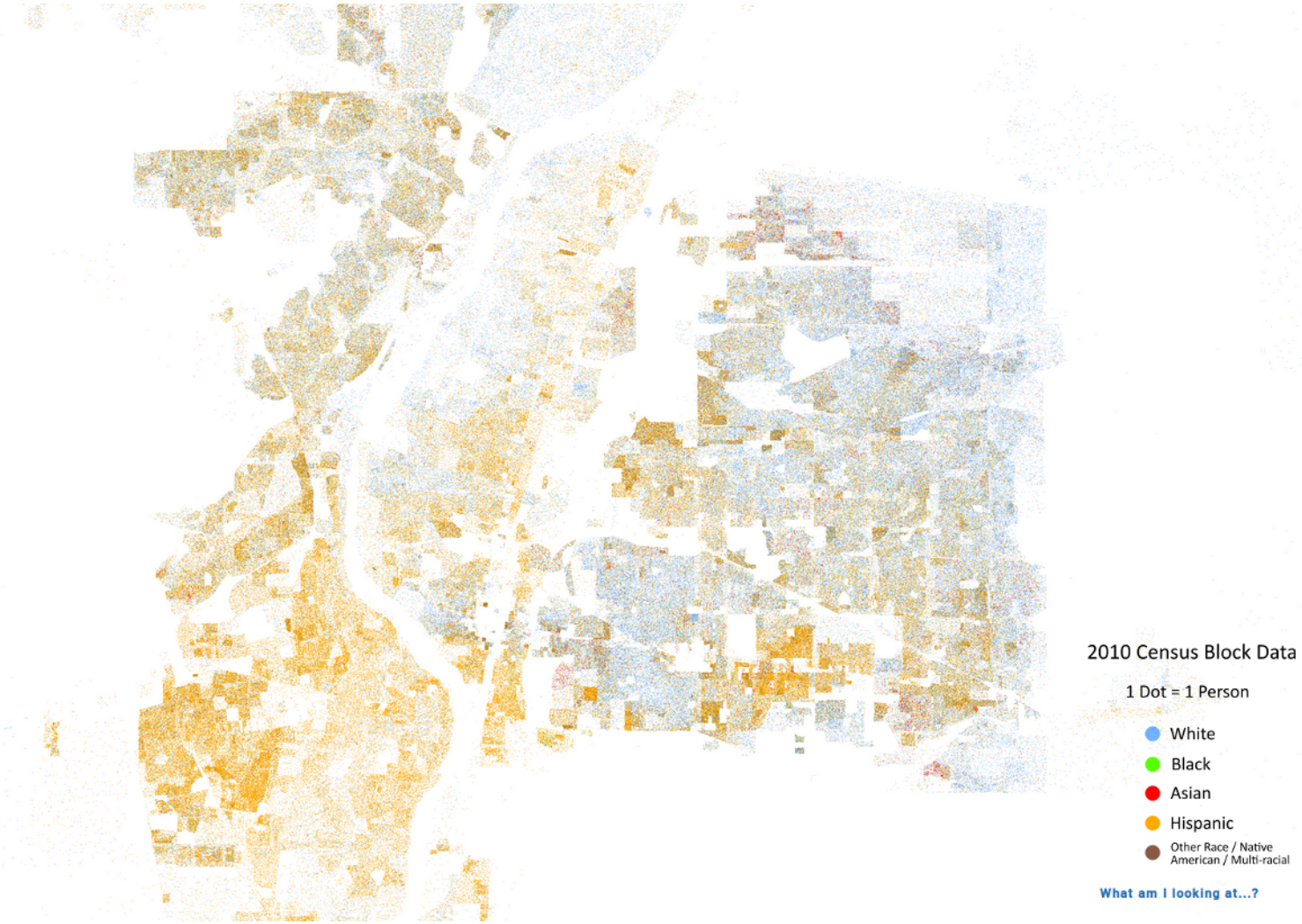




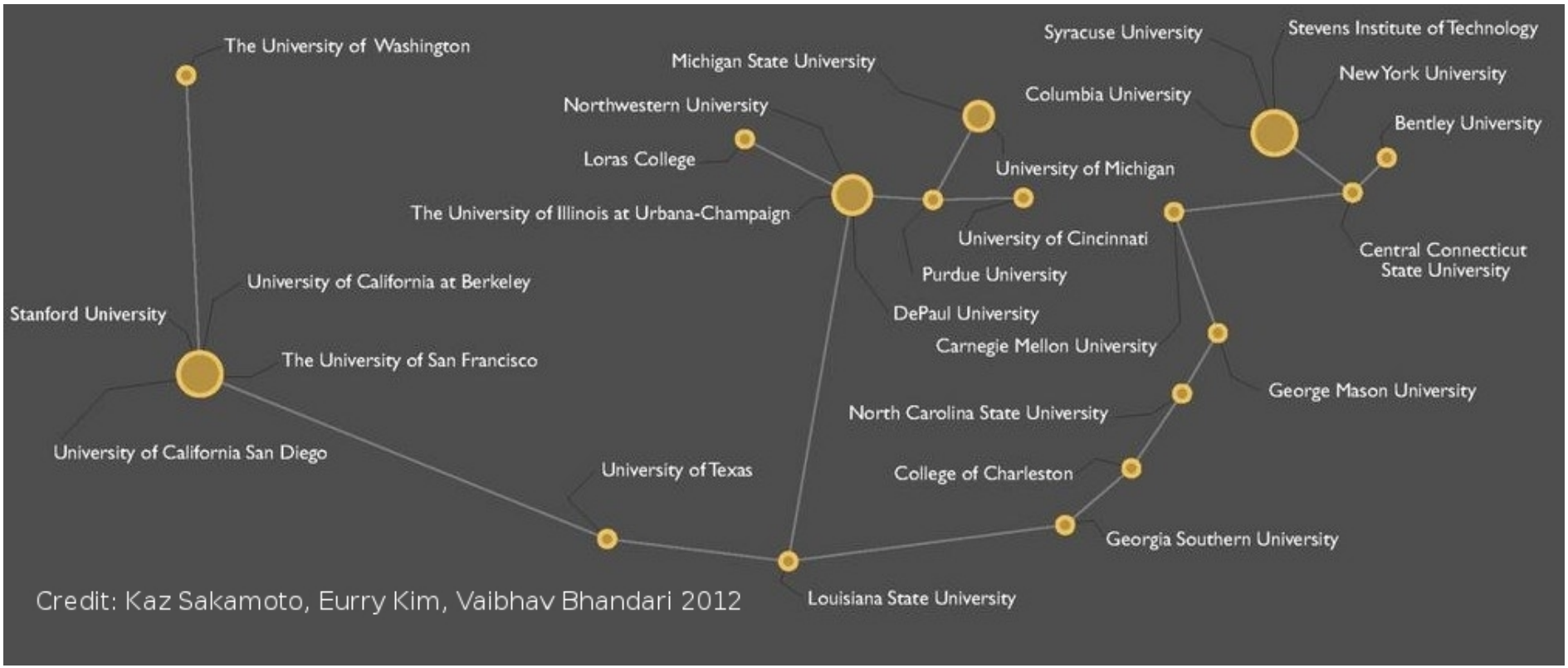
- Used for exploratory data analysis and presenting results.
- Static, dynamic, or interactive as appropriate.
- Many standard and emerging tools available
  - D3.js, Matplotlib, ggplot2, etc.

Visualization

Albuquerque mapped by race/ethnicity (CooperCenter.org)



Data Science as a buzz word is slowing training of workforce



- Academia has been responding to the need
- Goal not clear: Mixture of certificate programs to masters
- Harlan Harris described the problem at DataGotham 2012 `http://goo.gl/zuYyaD`
- Data scientists need major research experience?
- Parallel: Physics Ph.D.s moving to Quant Finance
- A clear need to connect quantitative problem solvers with data problems

Roy Keyes, PhD
Data Scientist
roy@zefsdata.com
@roycoding

Steve Koch, PhD
Research Data Scientist
stevekochscience@gmail.com
@skoch3

Zefs Data Science

UNM

Data Science ABQ

Google+ or @DataSciABQ
bit.ly/ABQDataScience
First meeting: 2nd week of October
Location: TBA