

Data Science Basics

Rebecca Bilbro and Pri Oberoi

3/14/2016



Dr. Rebecca Bilbro (rbilbro@doc.gov)
Data Scientist, Commerce Data Service
Board Member, Data Community DC
Faculty, Georgetown School of Continuing Studies
and District Data Labs

Pri Oberoi (poberoi@doc.gov)
Data Scientist, Commerce Data Service
Chair of Mentors, Women in Bio



Goals

Our goals for the class

- Provide a functional definition for data science
- Explain where the field came from
- Describe the key skills and tools
- Demonstrate what data products are
- Walk through the data science pipeline

Goals

Your goals for the class

- Be able to spot data science in the wild
- Understand the data science pipeline
- Think about your data strengths and growth areas
- Consider what role data science could play in your work
- Brainstorm potential data science projects

What is data science?

What is data science?

Thoughts?

Examples?

Is it just rebranding?

New methods, old questions?

Old methods, new questions?

Something new?

A NETFLIX ORIGINAL SERIES

HOUSE 
of CARDS





“Data science is the practice of **transforming raw data into insights, products, and applications to empower data-driven decision making**. It combines proven, time-tested methods from fields including statistics, natural sciences, computer science, operations research, and design in ways that are particularly well-suited to the data age. These methods, which range from data mining and visualization to predictive modeling, can scale from small to large datasets and can handle structured data as well as unstructured data like text and images.”

Jeff Chen, Chief Data Scientist
U.S. Department of Commerce

What does “data science” have
to do with “big data”?

The Economist

FEBRUARY 27th - MARCH 5th 2010

Economist.com

Obama the warrior

Misgoverning Argentina

The economic shift from West to East

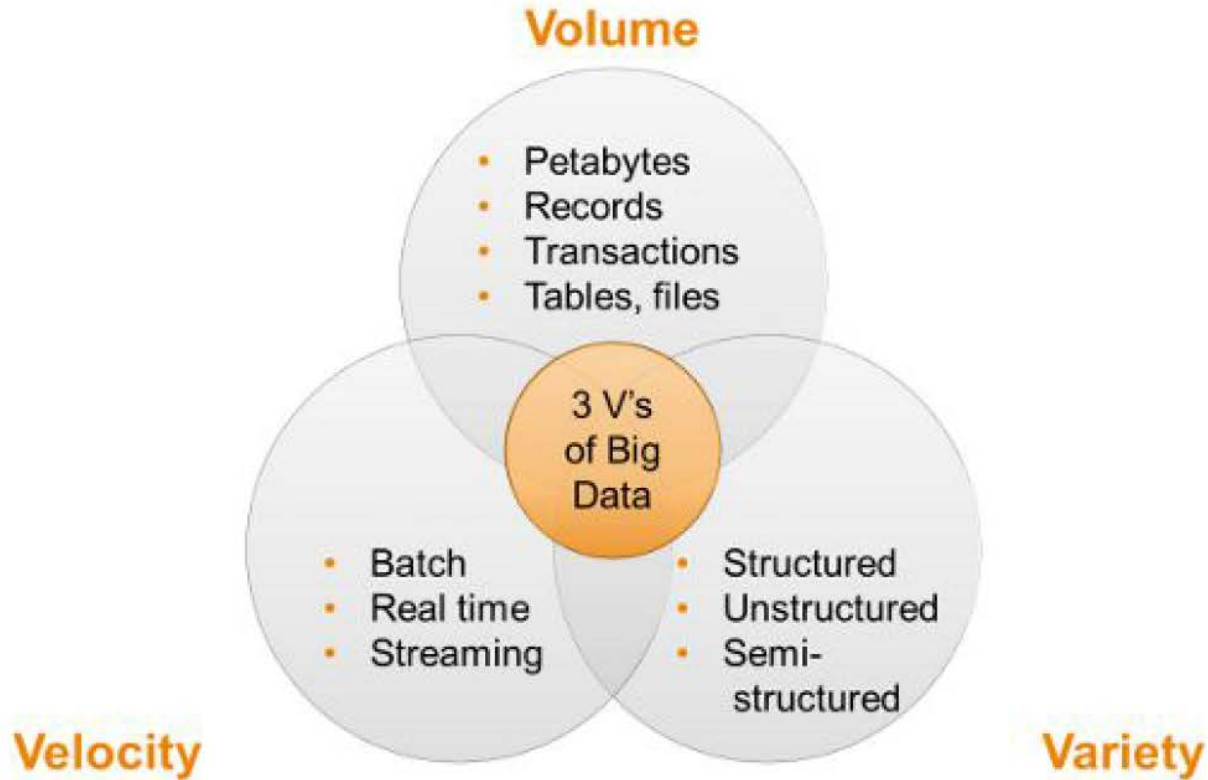
Genetically modified crops blossom

The right to eat cats and dogs

The data deluge

AND HOW TO HANDLE IT: A 14-PAGE SPECIAL REPORT





Where does data science come from?

Hal Varian (2009)

“The sexy job in the next ten years will be statisticians... The ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it—that’s going to be a hugely important skill.”

[Hal Varian on how the Web challenges managers
McKinsey Quarterly](#)



Steve Lohr, NYT (2009)

“The new breed of statisticians... use powerful computers and sophisticated mathematical models to hunt for meaningful patterns and insights in vast troves of data.”

[For Today's Graduate, Just One Word: Statistics](#)



Mike Driscoll (2009)

“I believe that the folks to whom Hal Varian is referring are not statisticians in the narrow sense, but rather people who possess skills in three key, yet independent areas:

statistics, data munging, and data visualization.”

[The Three Sexy Skills of Data Geeks](#)
[dataspora](#)

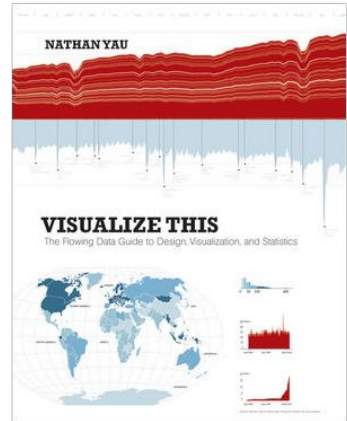


Nathan Yau (2009)

“We're seeing *data scientists* - people who can do it all - emerge from the rest of the pack.”

“Statisticians should know APIs, databases, and how to scrape data; designers should learn to do things programmatically; and computer scientists should know how to analyze and find meaning in data.”

[Rise of the Data Scientist](#)
[FlowingData](#)



Ben Fry (2004)

Computational Information Design



[PhD Thesis](#)

[MIT Media Arts & Sciences](#)



Hilary Mason & Chris Wiggins (2010)

1. Obtain: pointing and clicking does not scale.
2. Scrub: the world is a messy place
3. Explore: You can see a lot by looking
4. Models: always bad, sometimes ugly
5. Interpret: “The purpose of computing is insight, not numbers.” (Hamming)

"Data science is clearly a blend of the hackers' arts; statistics & machine learning; expertise in mathematics & the domain of the data for the analysis to be interpretable. It requires creative decisions & open-mindedness in a scientific context."

[A Taxonomy of Data Science](#)

[“Dataists”](#)

Mike Loukides (2010)

"Data science enables the creation of data products."

"Whether... data is search terms, voice samples, or product reviews,... users are in a feedback loop in which they contribute to the products they use. That's the beginning of data science."

[What is data science?](#)

[O'Reilly Radar](#)

What is
Data Science?

The future belongs to the companies
and people that turn data into products



A Melting Pot?

John D. Cook (2011)

"Calling someone a jack of all trades could be a way of saying that you don't have a mental category to hold what they do."

"Take an expert programmer back in time 100 years. What are his skills? Maybe he's pretty good at math. He has good general problem solving skills, especially logic. He has dabbled a little in linguistics, physics, psychology, business, and art. He has an interesting assortment of knowledge, but he's not a master of any recognized trade."

[Jack of all trades?](#)

[The Endeavour](#)

Venkatesh Rao (2011)

“I find myself feeling strangely uncomfortable when people call me a generalist and imagine that to be a compliment... I just look like a generalist because my path happens to cross many boundaries that are meaningful to others, but not to me.”

“[T]he primary real value of an extrinsically defined discipline... is *predictable boundedness*. Mathematicians can trust that they won’t have to suddenly start dancing halfway through their career to progress further.”

“You might wake up one fine day and realize that your life... actually adds up to expertise in some domain you’d never identified with at all.”

[The Calculus of Grit](#)
[ribbonfarm](#)

Drew Conway on Academia (2011)

"With respect to how academics have been impacted by data science, I think the impact has mostly flowed in the other direction. One major component of data science is the ability to extract insight from data using tools from math, statistics and computer science. Most of this is informed by the work of academics, and not the other way around."

"As so much more data gets pushed into the open, I believe basic data hacking skills — scraping, cleaning, and visualization — will be prerequisites to any academic research project."

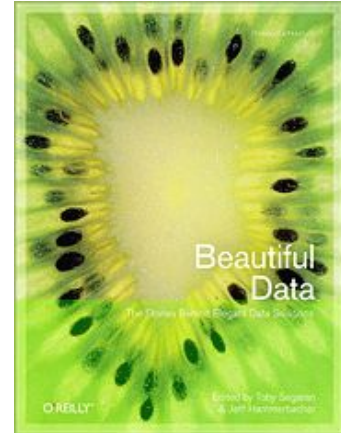
[Data science is a pipeline between academic disciplines](#)

[O'Reilly Radar](#)

Jeff Hammerbacher (2009)

"... on any given day, a team member could author a multistage processing pipeline in Python, design a hypothesis test, perform a regression analysis over data samples with R, design and implement an algorithm for some data-intensive product or service in Hadoop, or communicate the results of our analyses to other members of the organization."

[Beautiful Data, O'Reilly](#)



A Practicality?

DJ Patil (2015)

“Back in 2008, Jeff [Hammerbacher] and I got together to talk about our experiences building data teams at Facebook and LinkedIn. We basically came up with the term ‘data scientist’ because HR was being a pain.”

"Data Science" = cooler "Analytics"?

ASA President Nancy Geller (2011)

"If the "S" word falls into disfavor and disuse, I fear our discipline will lose its identity and, instead of a single discipline, Statistics will become subservient to data analysis, data mining, bioinformatics, business analytics, etc."

"We need to tell people that Statisticians are the ones who make sense of the data deluge occurring in science, engineering, and medicine; that Statistics provides methods for data analysis in all fields, from art history to zoology; that it is exciting to be a Statistician in the 21st century because of the many challenges brought about by the data explosion in all of these fields."

[Don't Shun the 'S' Word](#)
[Amstat News](#)

INFORMS' Michael Gorman on "Analytics"

"...[T]here is not a big difference between analytics and OR/MS, but a difference in their relative emphases. Both areas discuss the use or application of advanced techniques by organizations. However, OR clearly emphasizes the tools and techniques; analytics emphasizes more the analytical process, the tool application and integration, and their impact on organizational competitiveness and efficiency."

Jeff Drazen - NEJM (2015)

“[we worry] that a new class of research person will emerge— people who had nothing to do with the design and execution of the study but use another group’s data for their own ends, possibly stealing from the research productivity planned by the data gatherers, or even use the data to try to disprove what the original investigators had posited. There is concern among some front-line researchers that the system will be taken over by what some researchers have characterized as **‘research parasites’** ”



DJ Patil ✓

@DJ44



Following

[#IAmAResearchParasite](#). The best science is done as in collaboration not in silos. Data is a team sport.

Atul Butte @atulbutte

Wow! Editor-in-chief of @ScienceMagazine writes article titled #IAmAResearchParasite! buff.ly/21P6ktd

RETWEETS

60

LIKES

78



4:37 AM - 5 Mar 2016



Kirk Borne (2016)

“Fake data scientists are often experts in one particular discipline and insist that their discipline is the one and only true data science. That belief misses the point that data science refers to the application of the full arsenal of scientific tools and techniques (mathematical, computational, visual, analytic, statistical, experimental, problem definition, model-building and validation, etc.) to derive discoveries, insights, and value from data collections.”

[20 Questions to Detect Fake Data Scientists \(KDNuggets\).](#)

So, if there are “real” and “fake”
data scientists...
what *are* the key skills?

Thoughts?



Josh Wills

@josh_wills



Follow

Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician.

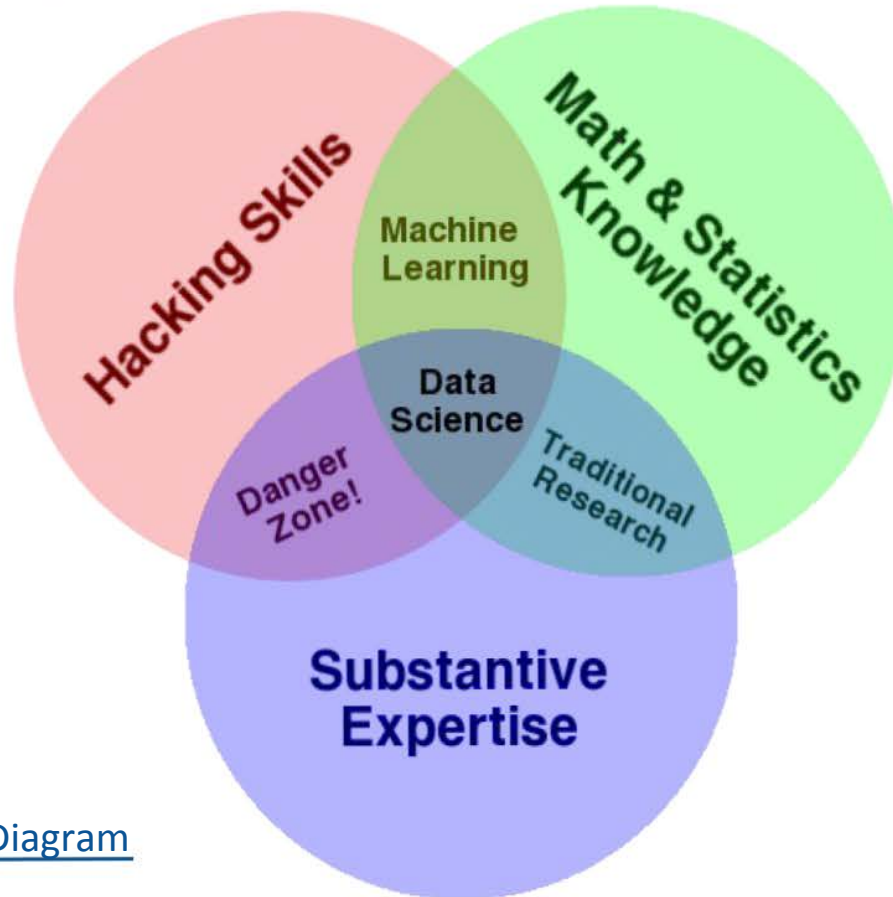
 Reply  Retweet  Favorited  More

RETWEETS
891

FAVORITES
406



12:55 PM - 3 May 2012



“Science” vs. “Scientist”

The state or fact of knowing; knowledge or cognizance of something specified or implied.

vs.

A person with expert knowledge of a science; a person using scientific methods.

Survey Time!

Data Scientist Survey



What kind of Data Scientist are you?

Data Scientist is a hot new term for people who apply advanced statistical, analytical, and machine learning tools to organizational data, and particularly Big Data. But if there's one thing we've learned, it's that not all Data Scientists are alike. We come from different backgrounds, we attack problems from a variety of angles, and we think of our own career paths taking different routes. Several DC-area Data Scientists conducted a survey in 2012, and found out a lot about the variation in people who could arguably be identified by the term. Now, **you** can take advantage of their hard work and find out what sort of highly-in-demand, brilliant, dare-we-say "sexy" Data Scientist you are!

Just take a few minutes to rank your skills and tell us how you view yourself. In exchange, we'll tell you more and describe how you fit in! Advice provided is for entertainment value only!

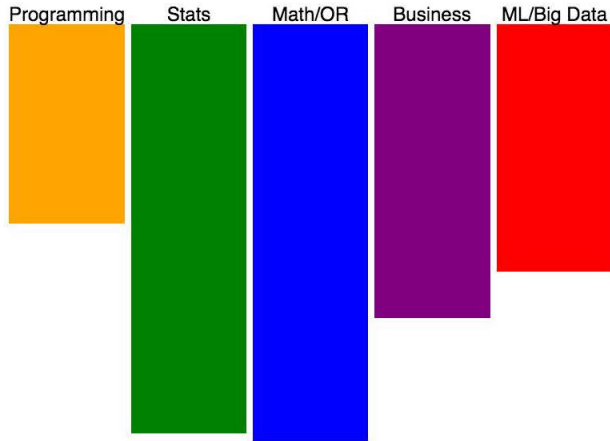
Just in case you were wondering, we will ****NEVER**** publish or provide to any third party unaggregated responses or identifying data.

[Get Started](#)

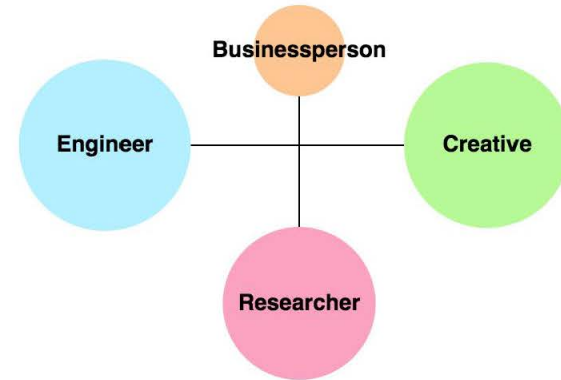
<http://survey.datacommunitydc.org/>

You're a **Data Engineer** with top skills in **Math/OR!**

Skills T-Chart



Self ID Chart



(Me)

The Variety of Data Scientists

Data Businesspeople:

Businessperson, leader, entrepreneur

Data Creatives:

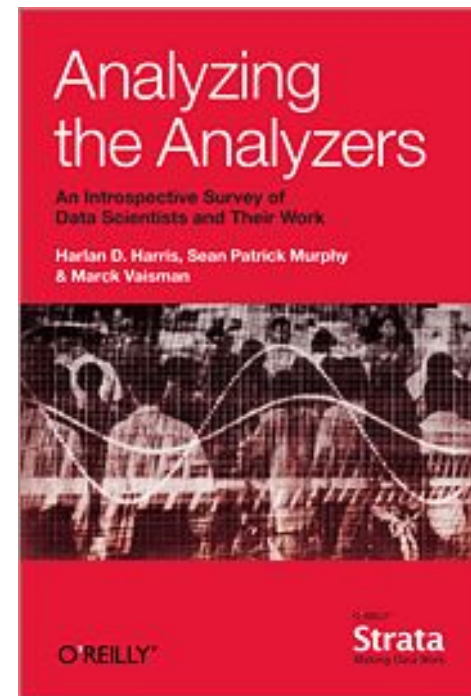
Artist, Jack of all Trades, Hacker

Data Developers:

Engineer, Programmer

Data Researchers:

Scientist, Researcher, Statistician



What tools do data scientists use?

What tools do data scientists use?
Suggestions?

Business Logic and Spreadsheet Computation



Price: \$139.99



Price: \$139.99

Google docs

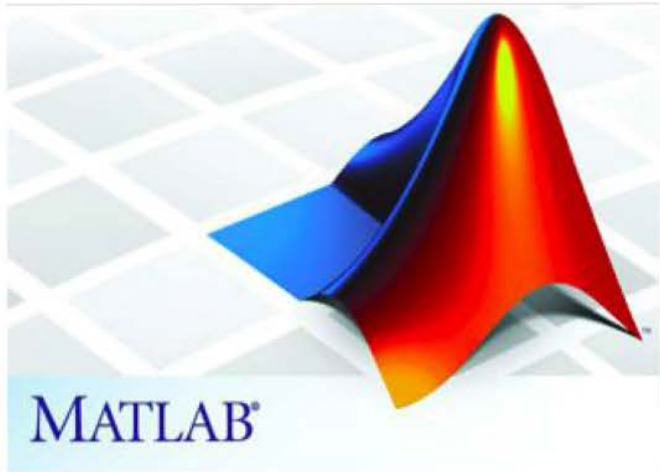


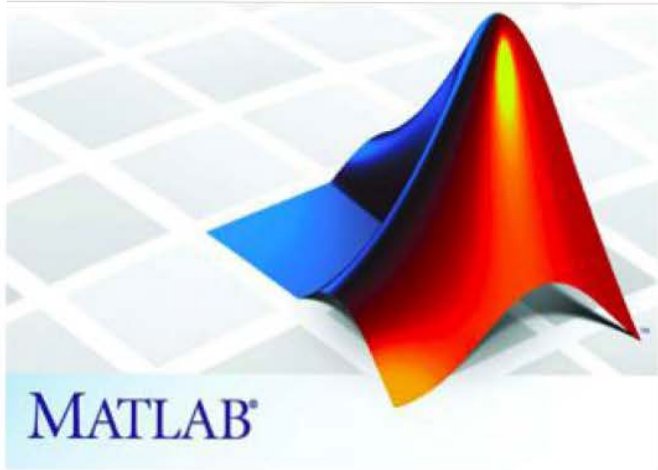


Price: \$139.99



Mathematical and Scientific Computation

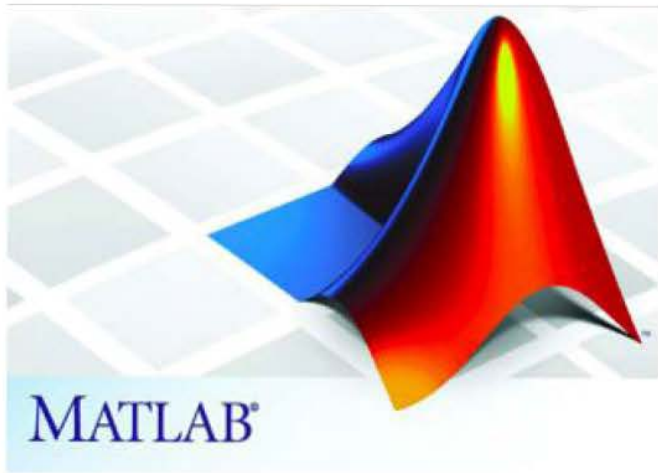




Price: \$\$\$
Beware of "Login Required"
to learn about individual
license pricing



```
> julia
julia | A fresh approach to technical computing
      | Version 0.8.0-1335217393.r9135
      | Commit 9135075fe2 (2012-04-23 17:43:13)
julia> str = "Hello, world!\n"
"Hello, world!\n"
julia> printf(str)
Hello, world!
julia> |
```



Price: \$\$\$
Beware of "Login Required"
to learn about individual
license pricing



python™



SciPy



NumPy



```
> julia
julia: a fresh approach to technical computing
Version 0.8.0+1335217383.+9135
git 9135075fe2 (2012-04-23 17:43:13)

julia> str = "Hello, world!\n"
"Hello, world!\n"

julia> printf(str)
Hello, world!

julia> 
```

Statistical Modeling and Analysis



THE POWER TO KNOW.

Price: Don't even ask



Price: A price quote is required



Price: \$700 and up



THE POWER TO KNOW[®]

Price: Don't even ask



Price: A price quote is required



Price: \$700 and up





THE POWER TO KNOW.

Price: Don't even ask

STATA Data Analysis and Statistical Software

Price: A price quote is required



Price: \$700 and up



Information and Knowledge Sharing





HAPPY CODING.



django



Databases

ORACLE®

IBM

ORACLE®

IBM



MySQL®

PostgreSQL



ORACLE®

IBM



Big Data and Distributed Computation







Infrastructure and Computing Resources

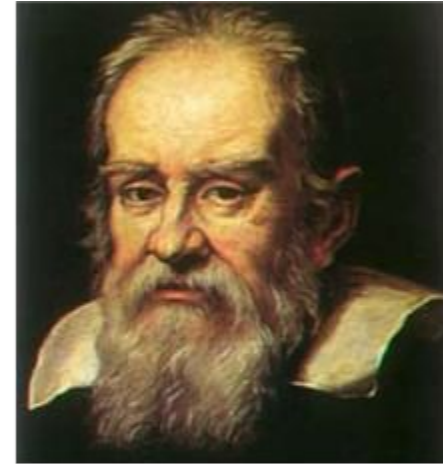






Google Compute Engine

But even with these tools,
you still need brains!



Hypothesis Driven Development

Practicing Hypothesis-Driven Development is thinking about the development of new ideas, products and services – even organizational change – as a series of experiments to determine whether an expected outcome will be achieved. The process is iterated upon until a desirable outcome is obtained or the idea is determined to be not viable.

We need to change our mindset to view our proposed solution to a problem statement as a hypothesis, especially in new product or service development – the market we are targeting, how a business model will work, how code will execute and even how the customer will use it. **We do not do projects anymore, only experiments.**

Hypothesis Driven Development **ThoughtWorks®**

We Believe That *<this capability>*

Will Result In *<this outcome>*

We Will Know We Have Succeeded When

<we see a measurable signal>

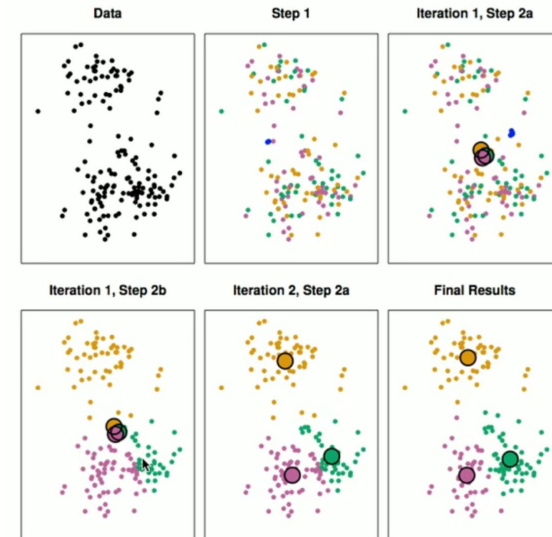
Making data science work in an organization

1. **Product Ownership:** A committed client-side domain expert.
2. **Theory of Change:** An idea of how a data science approach will improve outcomes at a clearly defined decision point.
3. **Delivery Strategy:** Could be a recommendation engine, dashboard, new SOPs...
4. **Domain Growth Potential:** Highest in qualitative and social service fields.
5. **Data Availability/Accessibility:** Data must exist!
6. **Data Alignment:** Data must be appropriate to the hypothesis.
7. **Signal Strength:** Data must contain sufficient signal for accurate prediction.
8. **Appeal:** The innovation factor.

Brainstorm:
Where could data science
be used in your field?

What news topics get published by NIST?

- K-Means Clustering on news published on [NIST's newsfeed](#) in 2014
- Their website doesn't indicate which subject area the article is about, so our data is unlabeled
- We know NIST publishes news on 15 subject areas, so we know $k=15$
- The goal is to find homogeneous clusters in your data, where we try to minimize the amount of variation within the cluster (Euclidean distance)
- Each iteration slightly improves the clustering



Clustering NIST headlines and description

Introduction:

In this workshop we show you an example of a workflow in data science from initial data ingestion, cleaning, modeling, and ultimately clustering. In this example we scrape the news feed of of [NIST](#). For those not in the know, NIST is the National Institute of Standards and Technology. It is comprised of multiple research centers which include:

- Center for Nanoscale Science and Technology (CNST)
- Engineering Laboratory (EL)
- Information Technology Laboratory (ITL)
- NIST Center for Neutron Research (NCNR)
- Material Measurement Laboratory (MML)
- Physical Measurement Laboratory (PML) This makes it an easy target in topic modeling.

You can use also this guide to scrape other data from a webpage: <http://docs.python-guide.org/en/latest/scenarios/scrape/>

Import the necessary modules for the workshop.

- [lxml](#) is a package for processing XML and HTML
 - If trouble installing on OSX, try running 'xcode-select --install'
- [requests](#) is a package for processing HTTP requests
- [future](#) to make a print function
- [scikit-learn](#) is a package with broad tool sets for machine learning
 - [TfidfVectorizer](#) to vectorize raw documents into a TF-IDF matrix
 - [KMeans](#)

What is a data product?

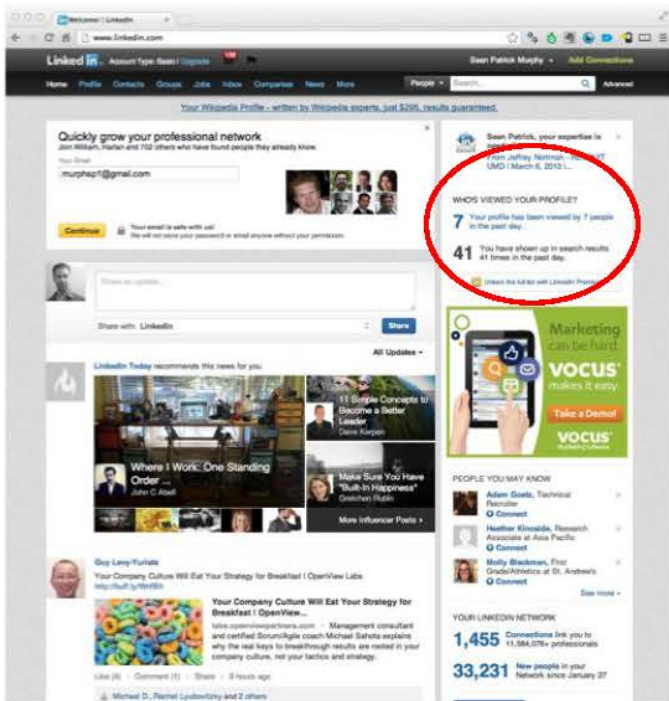
Ideas?

A data product is a product that is based on the combination of data and algorithms.

A data application acquires its value from the data itself, and creates more data as a result. It's not just an application with data; it's a data product.

Data products are self-adapting, broadly applicable economic engines that derive their value from data and generate more data by influencing human behavior or by making inferences or predictions upon new data.

What are some examples?



Quickly grow your professional network

WHO'S VIEWED YOUR PROFILE?

7 Your profile has been viewed by 7 people in the past day.

41 You have shown up in search results 41 times in the past day.



Carrier 3:19 PM

flights Search Results

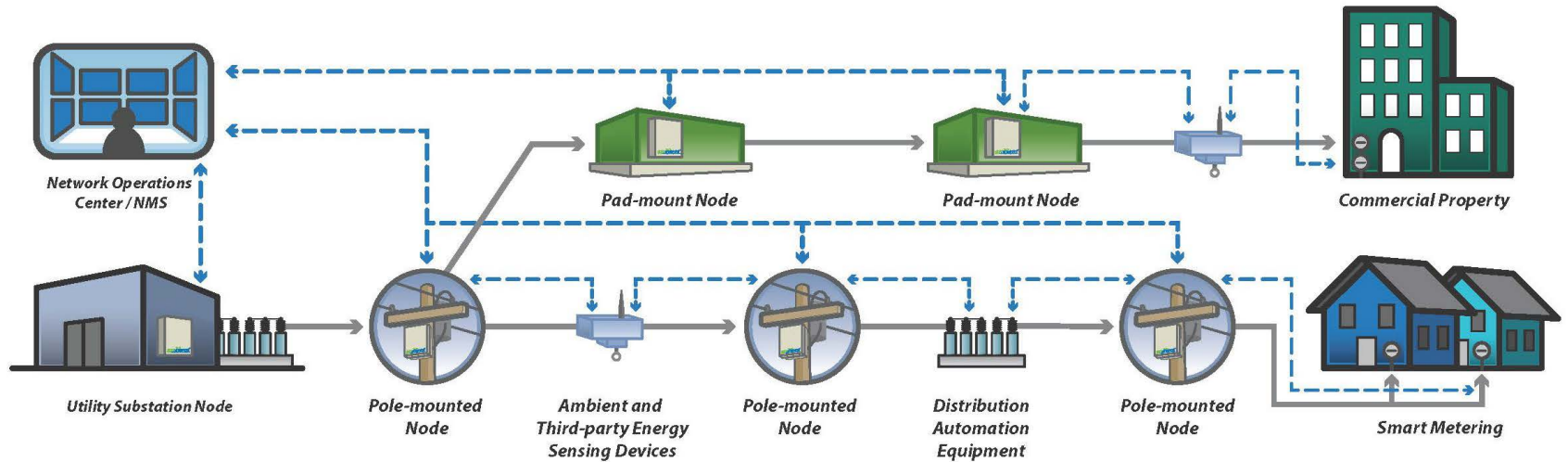
San Francisco → Hong Kong, 3/11/13

Depart:	12PM	12AM	12PM	12AM
Arrive:	3AM	3PM	3AM	3PM
\$1,017		Singapore		
\$1,048		United		
\$1,030		Air Canada		
\$1,067		China Air		
\$1,148		Cathay*		
\$1,055		Korean		
\$1,068		ANA*		
\$1,068		United		
\$1,068		ANA*		
\$1,079		Air China		

Agony Price Depart Length









How Data and a Good Algorithm Can Help Predict Where Fires Will Start

The New York City Fire Department is using a tool called FireCast to predict which buildings are most likely to have fires

FireCast 2.0 targets the most fire-prone buildings, many of which haven't been inspected in years. (© Paul A. Souders/CORBIS)

Data science for precision policy

Netflix Challenge

- Improve accuracy of predictions about how much someone is going to enjoy a movie based on their movie preferences.
- Netflix provided a large data set on how nearly half a million people have rated about 18,000 movies.
- Based on these ratings, you are asked to predict the ratings of these users for movies in the set that they have not rated.
- The first team to beat the accuracy of Netflix's proprietary algorithm by a certain margin wins a prize of \$1 million!

Anand Rajaraman, Datawocky

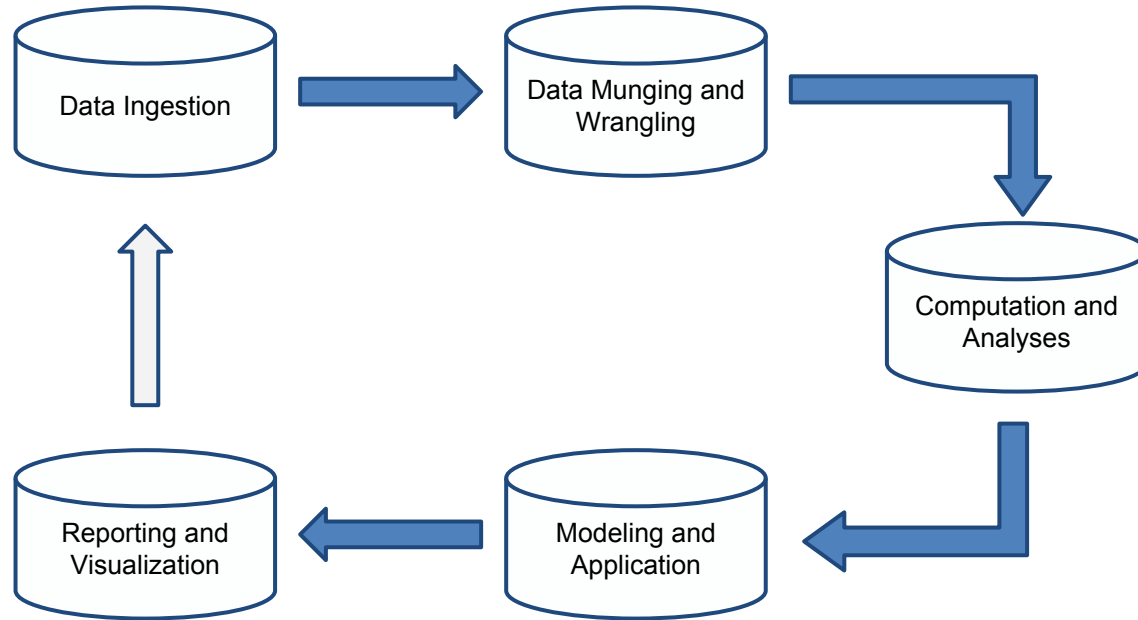
“Team A came up with a very sophisticated algorithm using the Netflix data. Team B used a very simple algorithm, but they added in additional data beyond the Netflix set: information about movie genres from the Internet Movie Database (IMDB). Guess which team did better?”

Anand Rajaraman, Datawocky

“Team A came up with a very sophisticated algorithm using the Netflix data. Team B used a very simple algorithm, but they added in additional data beyond the Netflix set: information about movie genres from the Internet Movie Database (IMDB). Guess which team did better?”

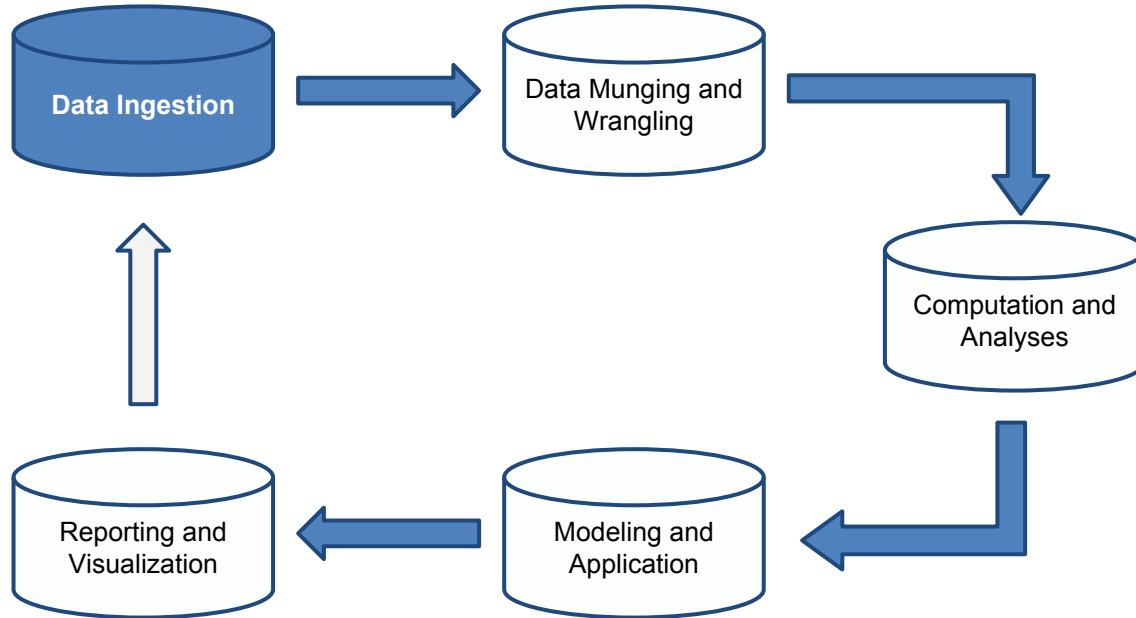
=> More data usually beats better algorithms

What is the data science pipeline?

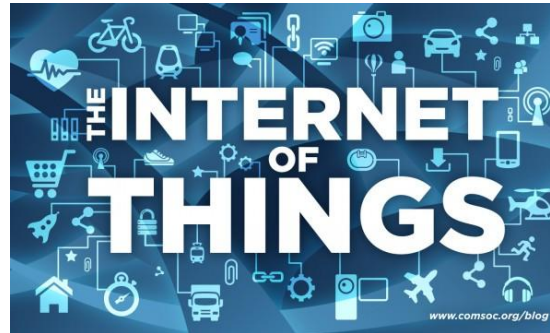
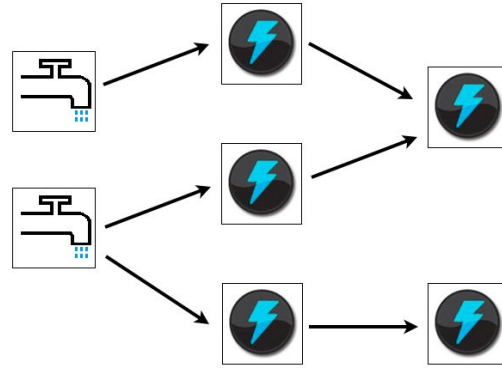


Data Ingestion

Means
Source
Question
Size
Velocity



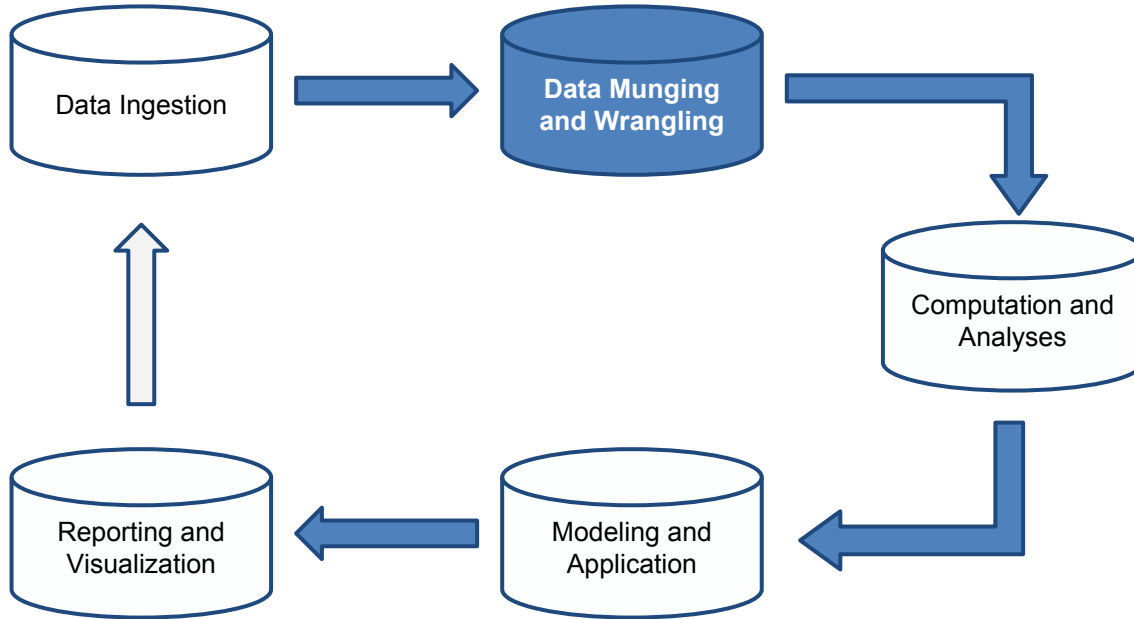
- There is a world of data out there-how to get it? Web crawlers, APIs, Sensors? Python and other web scripting languages are custom made for this task.
- The real question is how can we deal with such a giant volume and velocity of data?
- Big Data and Data Science often require ingestion specialists!



RESTful API
GET PUT POST DELETE



Data Munging and Wrangling



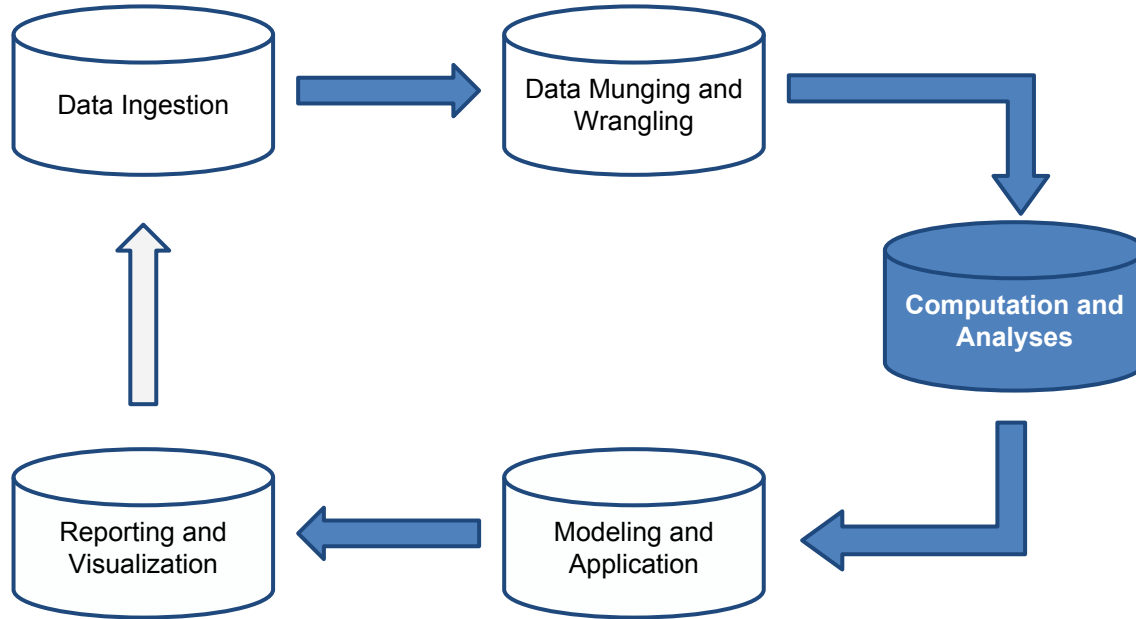
Warehouse
Extract
Transform
Filter
Aggregation
Training

- Warehousing the data means storing the data in as raw a form as possible.
- Extract, transform, and load operations move data to operational storage locations.
- Filtering, aggregation, normalization and denormalization all ensure data is in a form it can be computed on.
- Annotated training sets must be created for ML tasks.

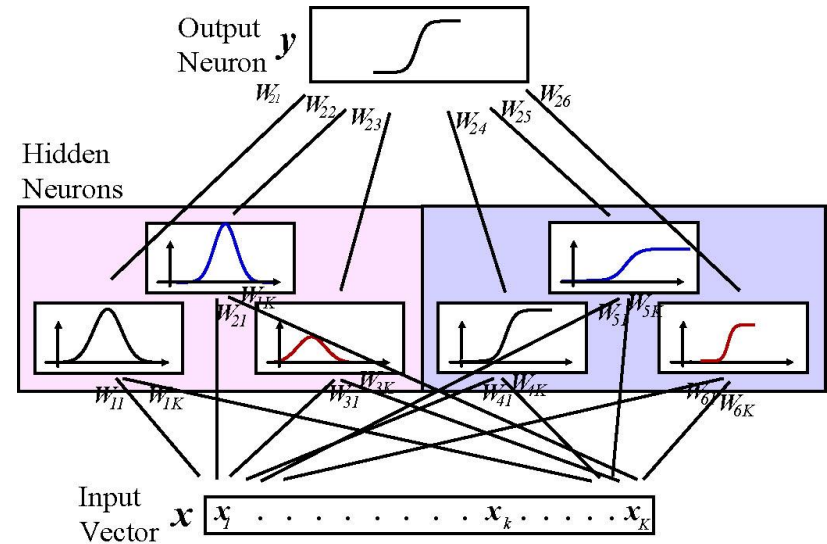


Computation and Analyses

Hypothesis
Design
Method
Time

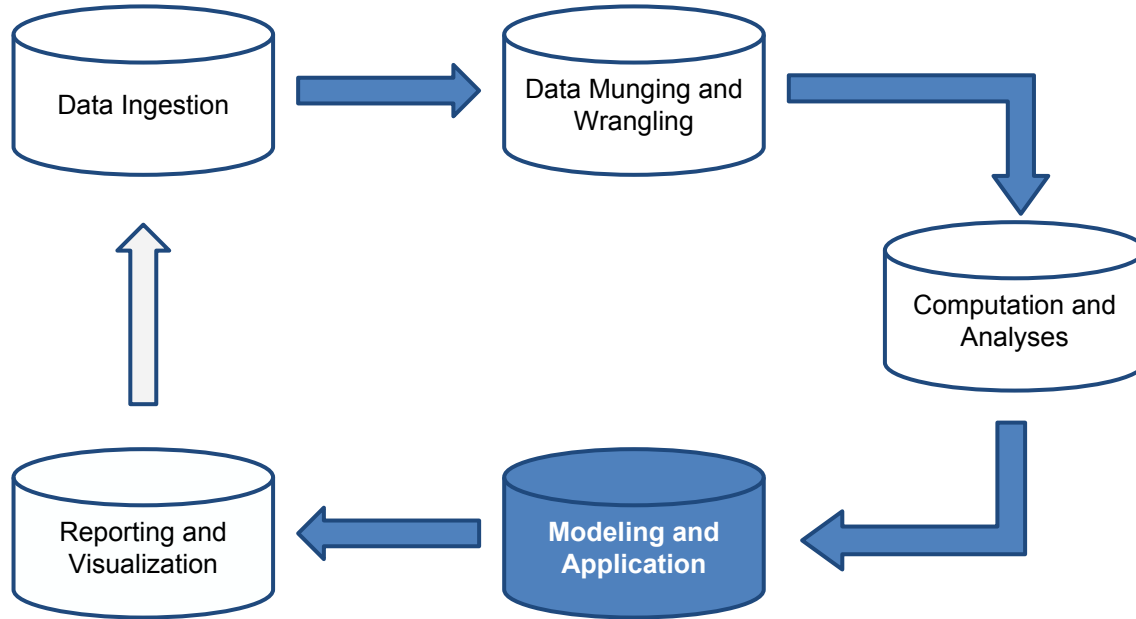


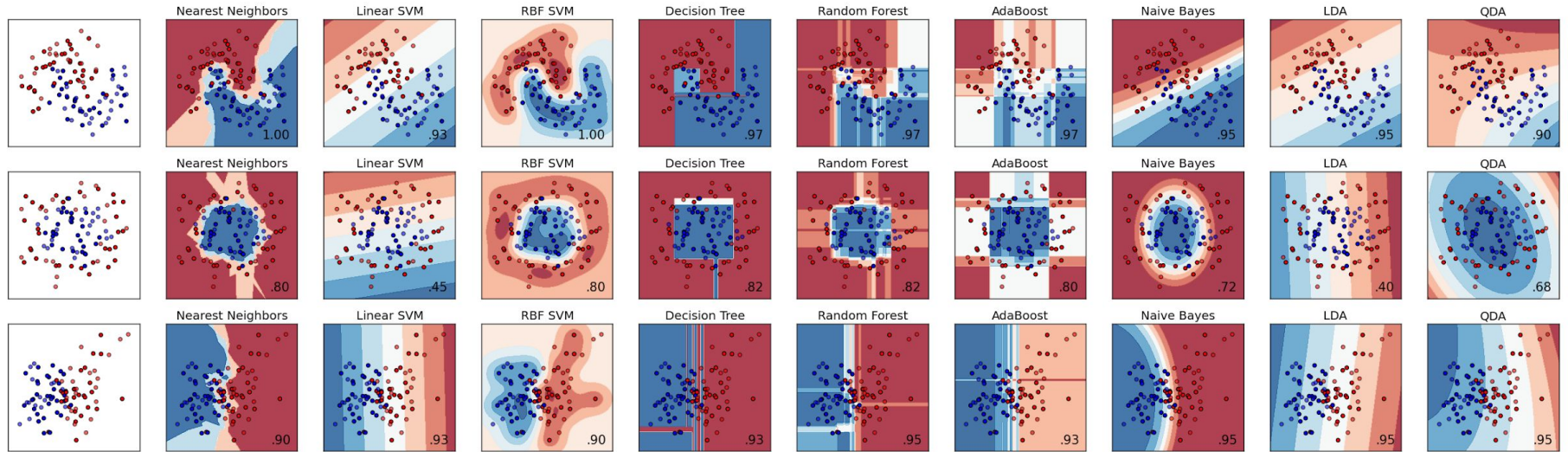
- Hypothesis driven computation includes design and development of predictive models.
- Many models have to be trained or constrained into a computational form like a Graph database, and this is time consuming.
- Other data products like indices, relations, classifications, and clusters may be computed.



Modeling and Application

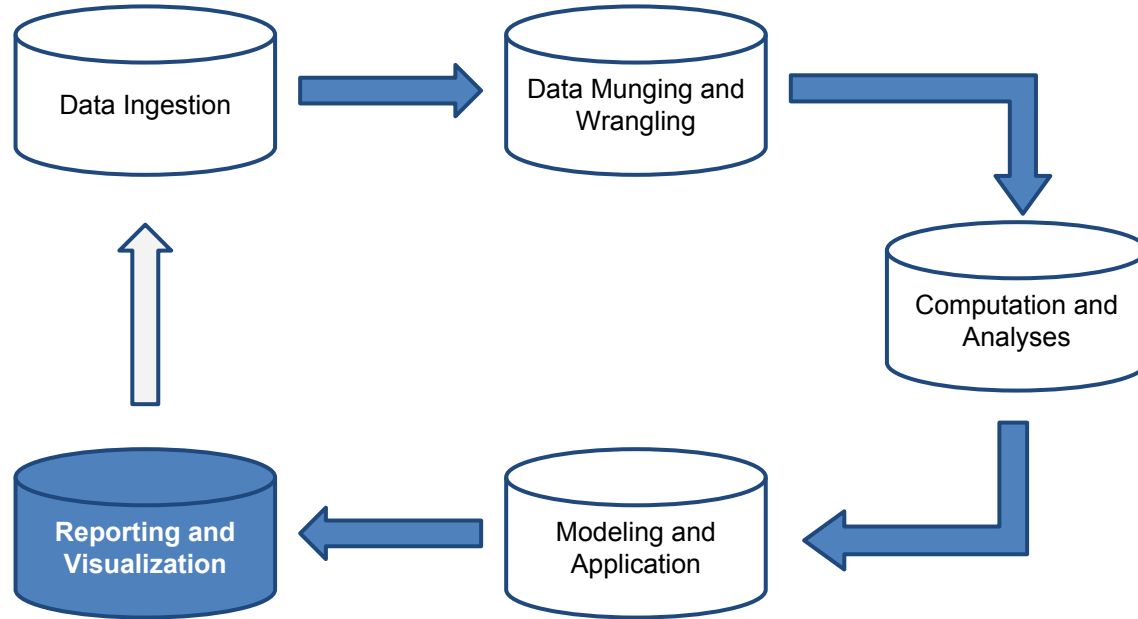
Supervised
Unsupervised
Regression
Classification
Clustering
Etc...





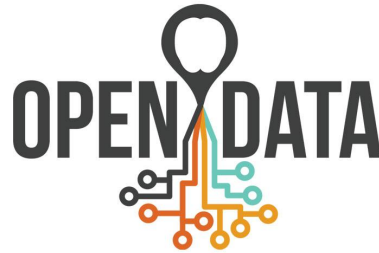
- Supervised vs. unsupervised
- Regression vs. classification
- Clustering
- Bayes, Logistic Regression, Decision Trees, KNN, etc

Reporting and Visualization



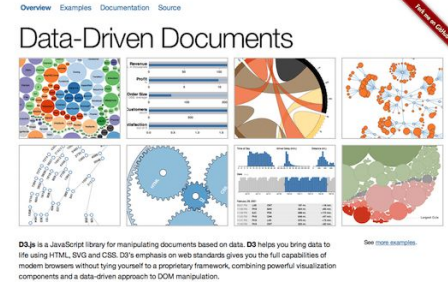
Crucial
Active Learning
Error Detection
Mashups
Value

- Often overlooked, this part is crucial, even if we have data products.
- Humans recognize patterns better than machines. Human feedback is crucial in Active Learning and remodeling (error detection).
- Mashups and collaborations generate more data- and therefore more value!



SUNLIGHT
FOUNDATION

django
REST
framework



Where to go from here?



COMMERCE.GOV



Search Engage Share

COMMERCE DATA USABILITY PROJECT

With tens of thousands of datasets ranging from satellite imagery to material standards to demographic surveys, the U.S. Department of Commerce has long been in the business of Open Data. Through the Commerce Data Usability Project, go on a series of guided tours through the Commerce data lake and learn how you can leverage this free and open data to unlock the possible.

Check out more of our open work at:

<http://www.commerce.gov/datausability>

and

<https://github.com/CommerceDataService>



Special thanks to my teacher:

Benjamin Bengfort

PhD Candidate at the University of Maryland; Data Scientist at District Data Labs.

Twitter: twitter.com/bbengfort

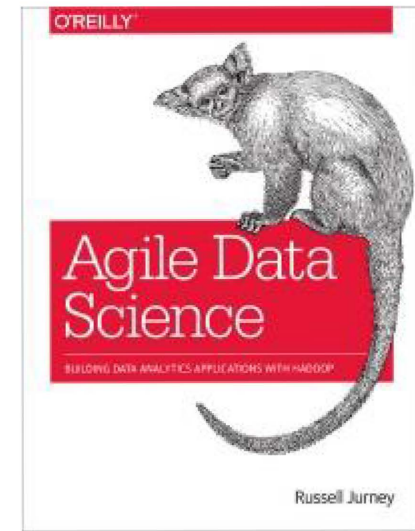
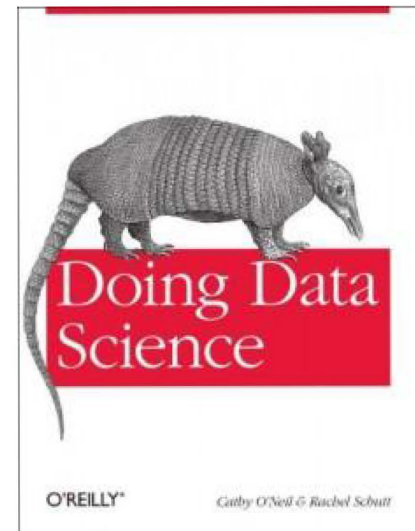
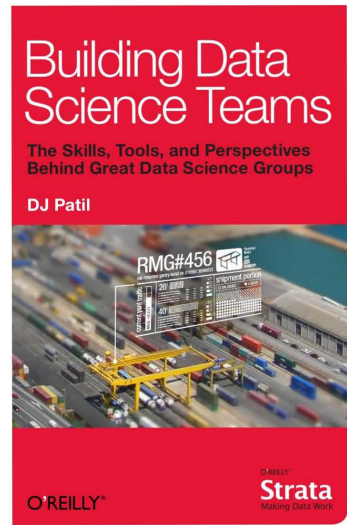
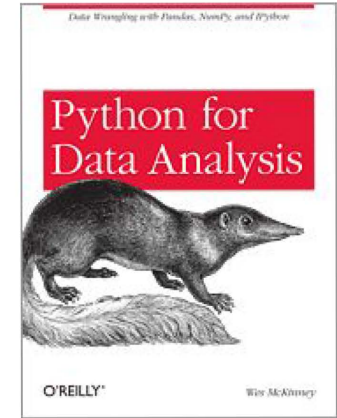
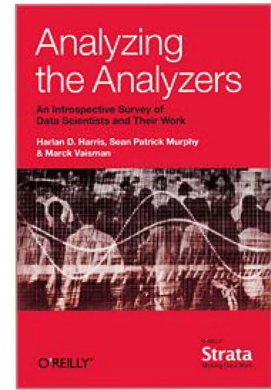
LinkedIn: linkedin.com/in/bbengfort

Github: github.com/bbengfort

Email: bb830@georgetown.edu

(These are mostly his slides!)







Find us at:

rbilbro@doc.gov and poberoi@doc.gov