Introduction to Data Science

2017 CSUN DataJam
CSU Northridge
Friday, October 6, 2017

Wayne Smith, Ph.D.
Department of Management
CSU Northridge
ws@csun.edu

This presentation is available at:

smithw.org/dsintro.pptx or smithw.org/dsintro.pdf

Overview

1. LA's "Silicon Beach"/"Silicon Valley South
2. High-level introduction to Data Science/Big Data
3. High-level look at Analytics in a big LA firm
4. Examples of Data Science tasks

- Most of the following material is mine, but some came from…
  - Levon Karayan (Disney), and
  - Hovig Tchalian (CGU/Drucker).

LA's "Silicon Beach"/Silicon Valley South

Los Angeles Times
Business

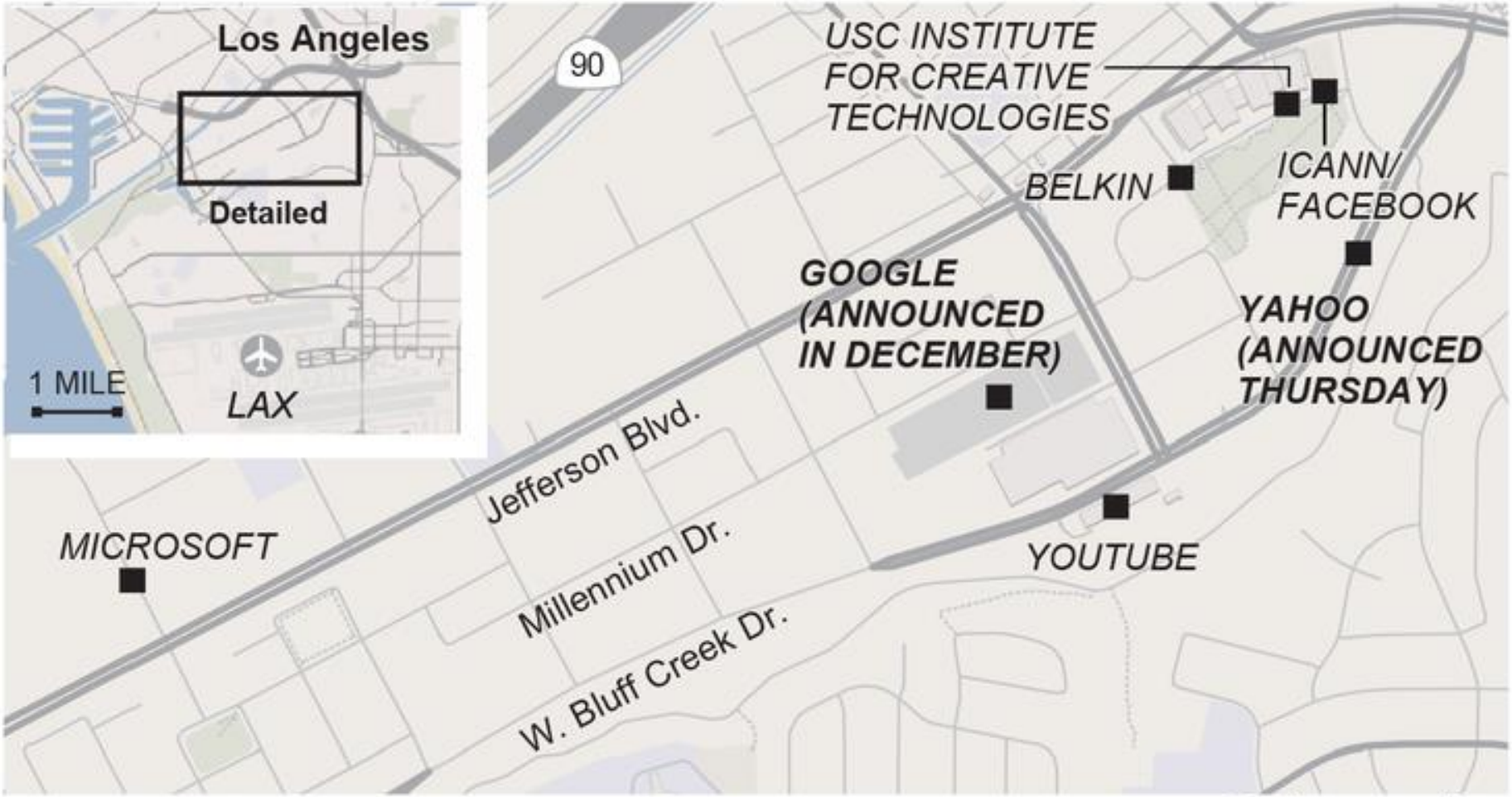
Playa Vista turning into Silicon Valley South as tech firms move in

This article is related to: Business, Eric Garcetti, , Marissa Mayer, LAX

Yahoo has signed a long-term lease for about 130,000 square feet at the new Collective campus, which is still under construction in Playa Vista. (Marcus Yam / Los Angeles Times)

By ANDREA CHANG AND PETER JAMISON
contact the reporters

http://www.latimes.com/business/la-fi-silicon-valley-south-20150118-story.html#page=1



Sources: Mapbox, OpenStreetMap

@latimesgraphics

Gil Press
Contributor

I write about technology, entrepreneurs and innovation.

TECH 12/01/2012 @ 5:32PM | 15,988 views

Big Data News of the Week: Beautiful $300,000 Minds

Jeff Hawkins at eTech 2007 (Photo credit: Wikipedia)

While many saw big data as the winner of the recent elections, I voted for Big Intuition, citing Bill Clinton's insight and advice as an example of how decisions and data science—in political campaigns or any other endeavor—cannot be automated and must rely on human judgment an domain expertise.

This week, Matthew Jones, a historian at Columbia who is working on the history of data mining, came to a similar conclusion after auditing Rachel Schutt's introduction to data science class: "Data science depends utterly on algorithms but does not reduce to those algorithms. The use of those algorithms rests fundamentally on what sociologists of science cal 'tacit knowledge'—practical knowledge not easily reducable to articulated rules—or perhaps impossible to reduce to rules."

This irreducible knowledge is of two kinds: Expertise and experience in a specific domain (as in "Clinton knows how to run political campaigns"); and—specifically for data scientists—experience with and understanding of the tools they apply. Says Jones: "The hubris one might have when using an algorithm must be tempered through a profound familiarity with that algorithm and its particular instantiation."

Skills, knowledge, and abilities are employers looking for in entry-level employees

FIGURE 38 ATTRIBUTES EMPLOYERS SEEK ON A CANDIDATE'S RESUME

ATTRIBUTE	% OF RESPONDENTS
Ability to work in a team	78.0%
Problem-solving skills	77.3%
Communication skills (written)	75.0%
Strong work ethic	72.0%
Communication skills (verbal)	70.5%
Leadership	68.9%
Initiative	65.9%
Analytical/quantitative skills	64.4%
Flexibility/adaptability	63.6%
Detail-oriented	62.1%
Interpersonal skills (relates well to others)	58.3%
Technical skills	56.8%
Computer skills	49.2%
Organizational ability	47.7%

Source: Job Outlook 2017, National Association of College and Employers

What is Data Science?

Curriculum Guidelines for Undergraduate Programs in Data Science (September, 2016)

- Data Science is "the science of planning for, acquisition, management, analysis of, and inference from data."

- Students would demonstrate mastery of skills and concepts, including many traditionally associated with the fields of Statistics, Computer Science and Mathematics. *Data Science blends much of the pedagogical content from all three disciplines, but it is* ***neither the simple intersection, nor the superset of the three****.*

- There is a fourth area of demonstrated mastery too: subject-matter expertise.
- Building upon experimentation, modeling, and computation, there are some that believe that data science *is, in fact, a new, type of scientific discovery*.

- Case-based, hands-on, and interdisciplinary
- Additionally, some existing courses in statistics, math, and computer science, should be partly re-designed for use in a data science curriculum.
- https://www.stat.berkeley.edu/~nolan/Papers/Data.Science.Guidelines.16.9.25.pdf

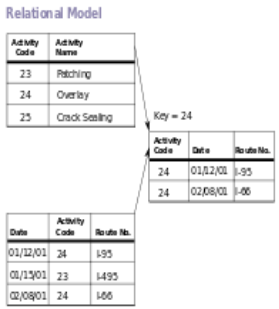
Information Dynamics

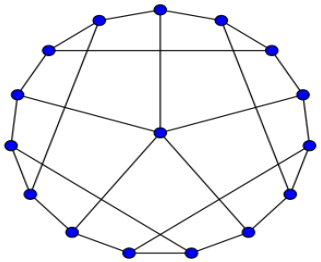
- *Wisdom*
  - Extraordinary Insight (Explanation) for Foresight (Prediction)
  - Restaurant: How should our menu change in the future to best optimize nightly sales?
- *Knowledge*
  - Combination of Explicit Information and Tacit Information
  - Restaurant: What action led to the change in last night's sales?
- *Information*
  - Meaningful Data
  - Restaurant: How does last night's sales compare to that night the previous year? How does last night's sales compare to our goals?
- *Data*
  - Raw, atomic, basic
  - Restaurant: What were the total sales for last night?

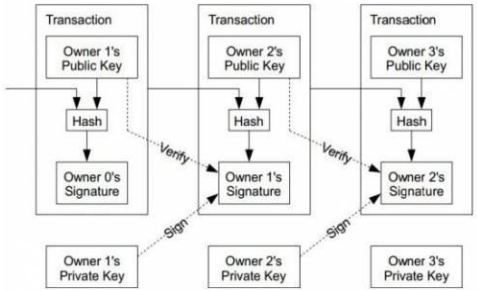
Analytics for Decision-making (e.g., in Management/HR)

- *Prescriptive* Analytics
  - What should we do?
  - HR Department: What should we (the HR Department) do to meet or exceed the organization's hiring and retention goals for next year? What data/information/knowledge/wisdom should we provide to our hiring and technical managers to help? What are we missing?
- *Predictive* Analytics
  - What is likely to happen?
  - HR Department: How many new employees will our organization need next year? How will the mix change? What is our competition likely to do?
- *Diagnostic* Analytics
  - Why did it happen?
  - HR Department: Did our emphasis on recruiting from campus A (over campus B, etc.) matter? What do the managers of these entry-level employees think?
- *Descriptive* Analytics
  - What happened?
  - HR Department: How many entry-level professionals did we hire last year? How many of them are still with us now?

Data Representation







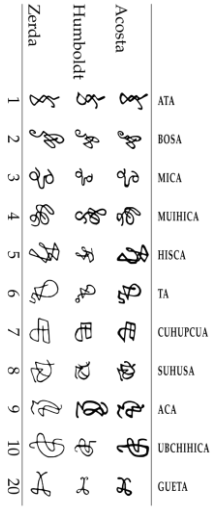
Unstructured

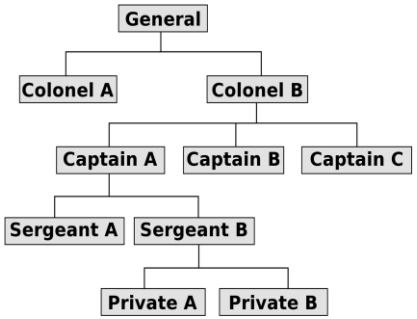
Relational

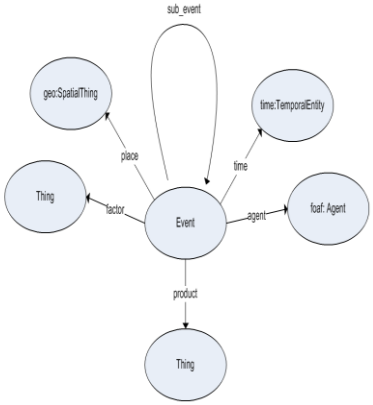
Graph

Blockchain

Progress







Tabular

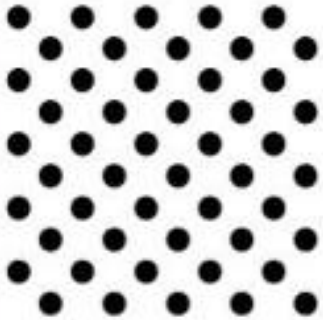
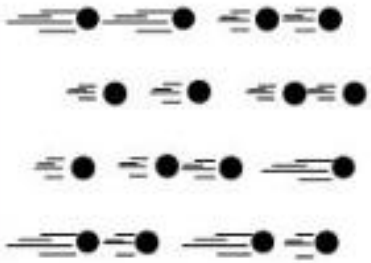
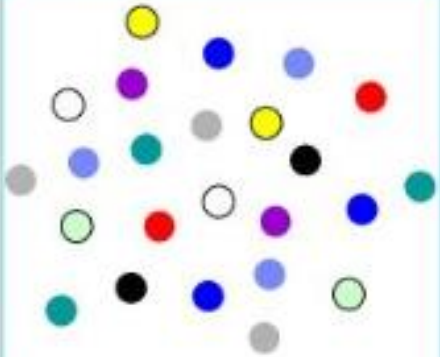
Hierarchical

Ontology

What is Big Data?

Big Data (or, Data Analytics): A Fuller Definition

So, what is "big data"?

Volume	Velocity	Variety	Veracity*
			
Data at Rest	**Data in Motion**	**Data in Many Forms**	**Data in Doubt**
Terabytes to exabytes of existing data to process	Streaming data, milliseconds to seconds to respond	Structured, unstructured, text, multimedia	Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

Big Data (or, Data Analytics): A Fuller Definition

How much data is “big” data?

Common Data Storage Measurements

UNIT	VALUE
bit	1 bit
byte	8 bits
kilobyte	1,024 bytes
megabyte	1,024 kilobytes
gigabyte	1,024 megabytes
terabyte	1,024 gigabytes
petabyte	1,024 terabytes

Big Data (or, Data Analytics): A Fuller Definition

Where does all this data come from, exactly?

By 2016, annual Internet traffic will reach 1.3 Zettabytes

Google processes > 24 Petabytes of data in a single day

Facebook processes 500+ Terabytes of data daily

Twitter processes 12 Terabytes of data daily

150 Exabytes global size of "Big Data" in Healthcare, growing between 1.2 and 2.4 EX / year

AT&T transfers about 30 Petabytes of data through its network daily

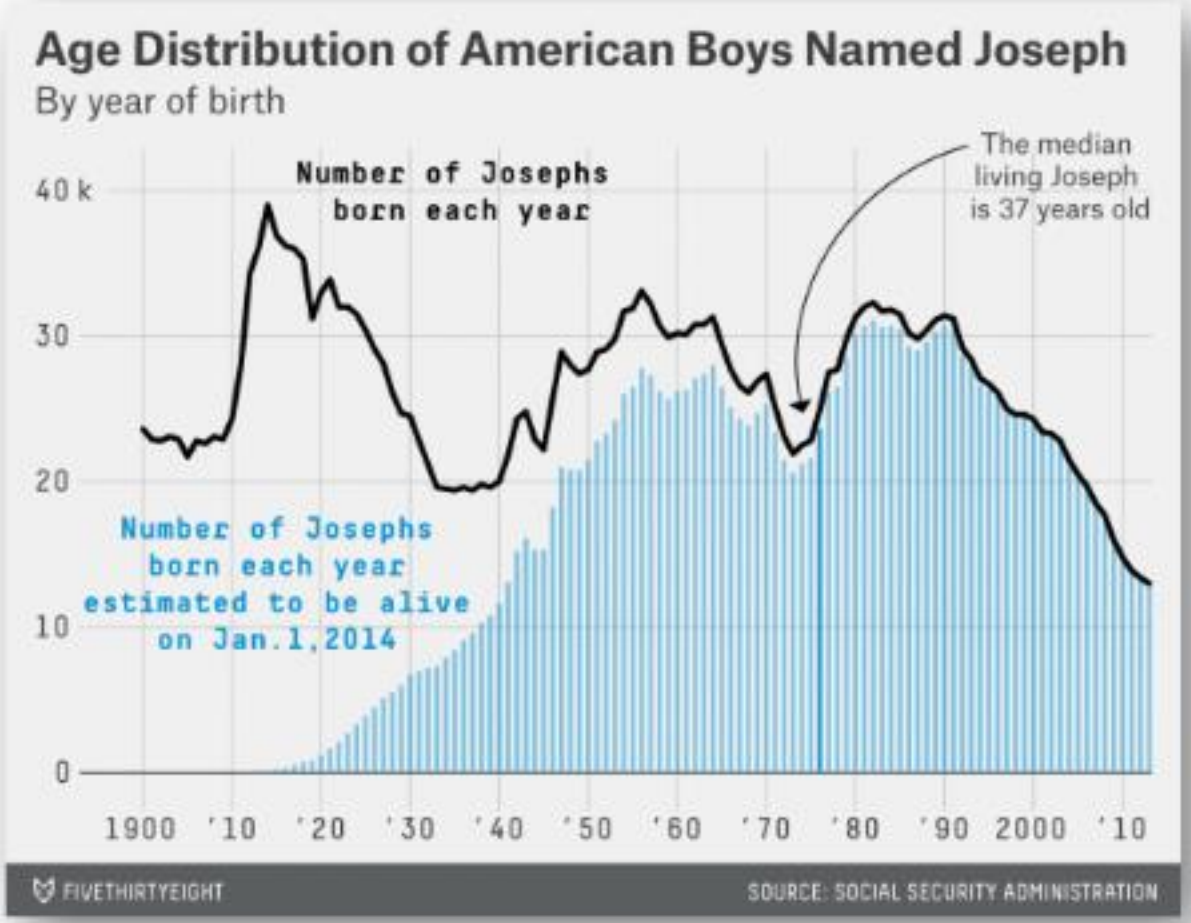
Hadron Collider at CERN generates 40 Terabytes of usable data / day

For every session, NY Stock Exchange captures 1 Terabyte of trade information

IBM

Big Data (or, Data Analytics): A Rough Definition

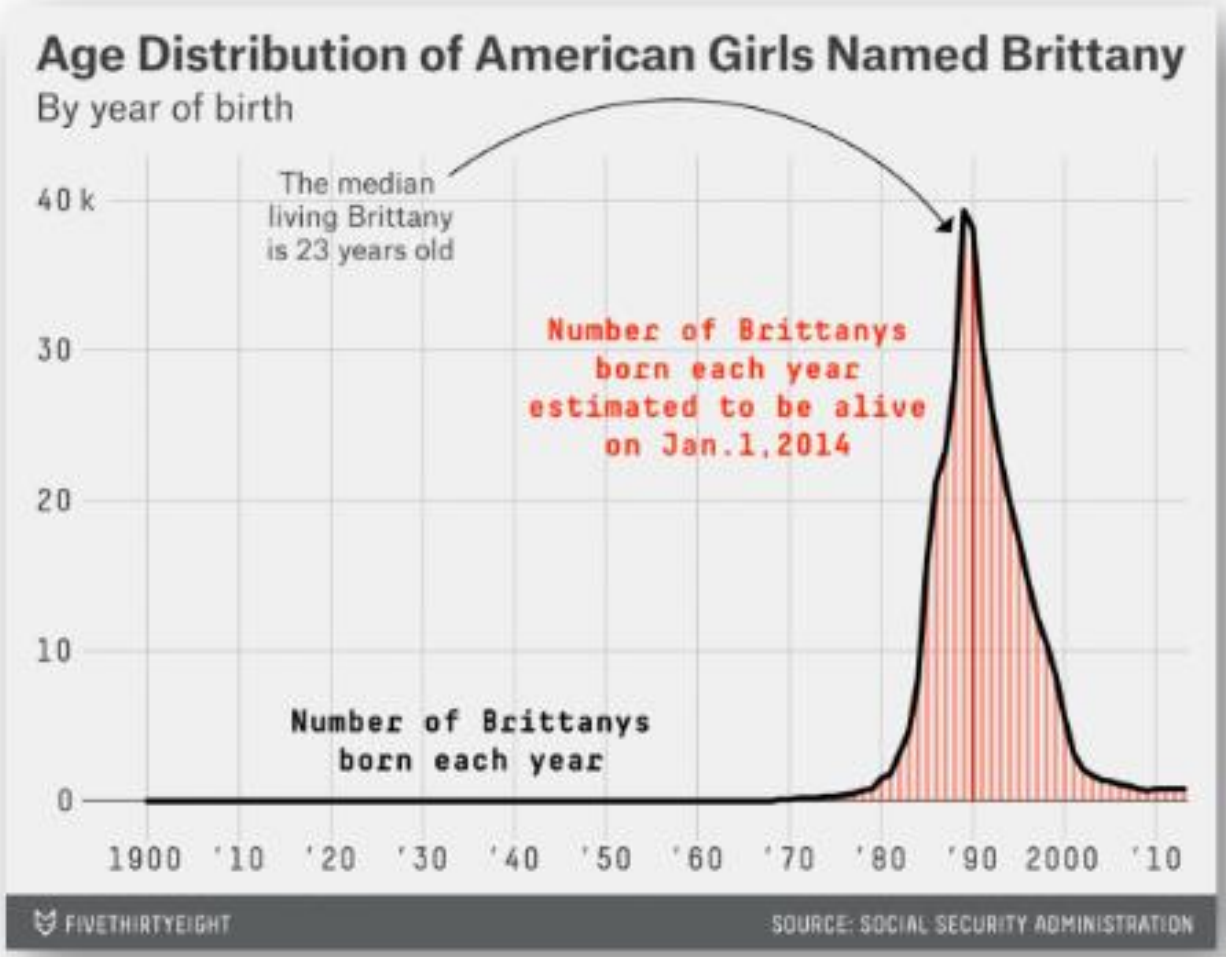
"Joseph has been one of the most enduring American names"



From description...

Big Data (or, Data Analytics): A Rough Definition

Popularity of the name "Brittany" peaked around the year 1990

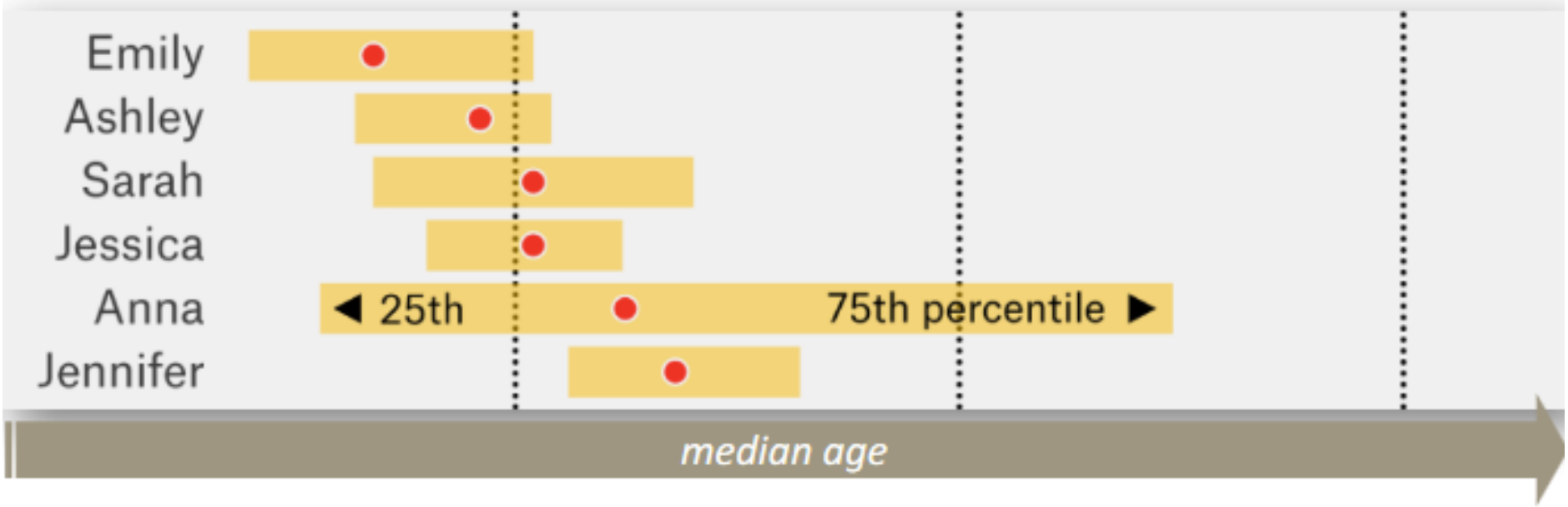


From description...

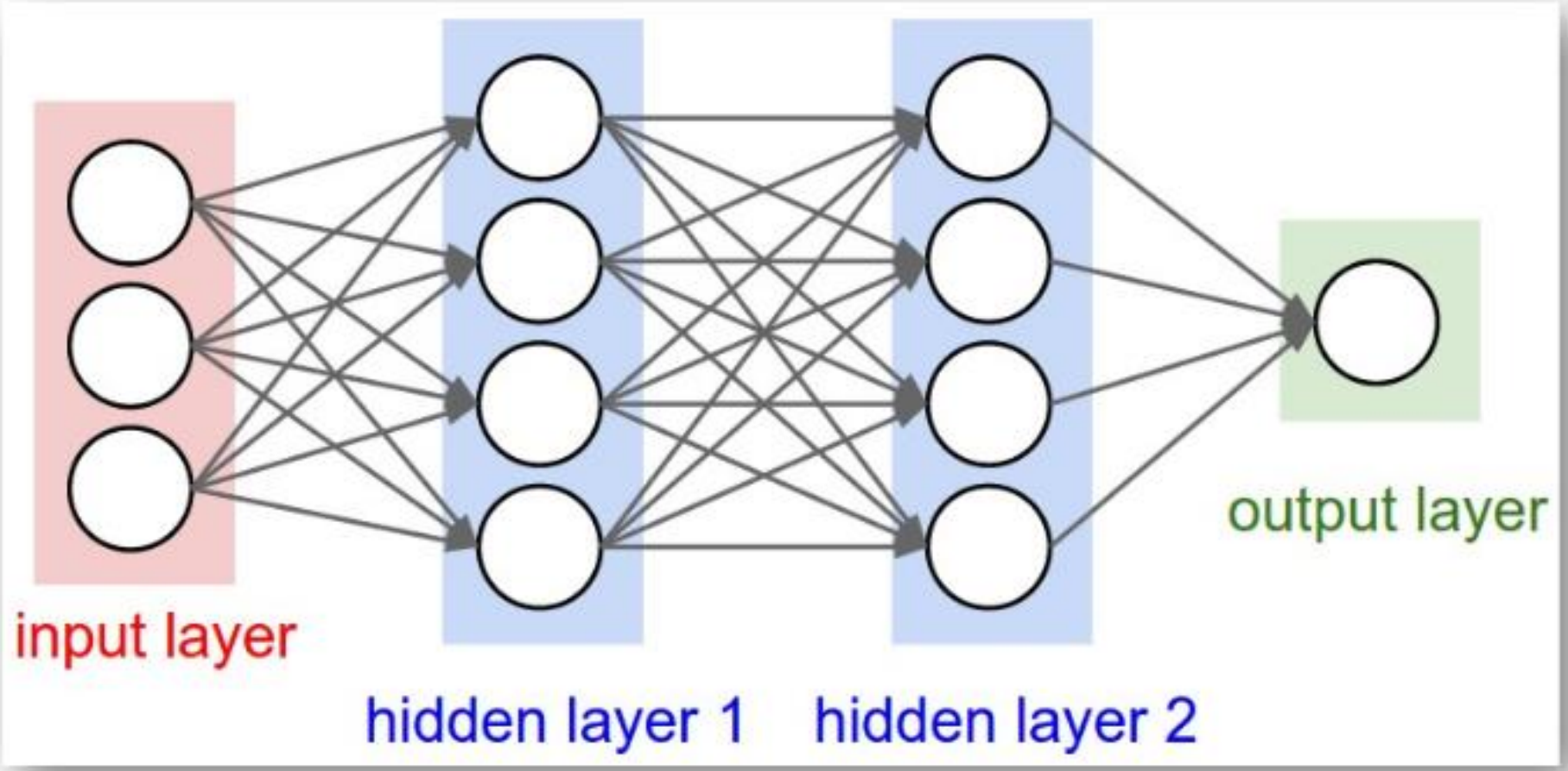
Big Data (or, Data Analytics): A Rough Definition

"How to Tell Someone's Age When All You Know Is Her Name"

To prediction



Practical Applications III: Deep (Machine) Learning



Practical Applications III: Deep (Machine) Learning

Mt. Sinai Hospital (NY) 2015 Research Program: "Deep Patient"

1. Tested on 700,000 patient records

Able to predict disease far better than traditional methods

2. Better than humans at predicting onset of schizophrenia

Not even physicians can accurately predict that psychiatric disorder

3. Algorithm was able to detect a pattern never before discovered

Not only is pattern latent, so is its detection method ("black-box")

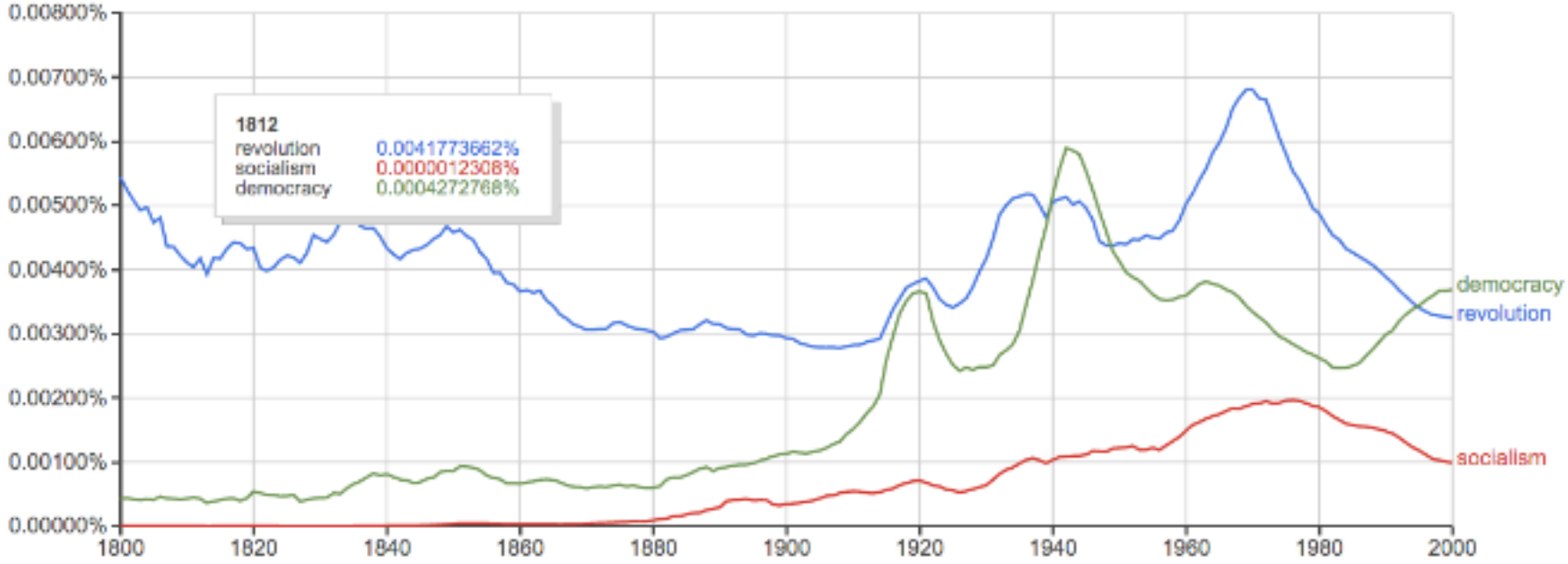
Text Mining & Linguistic Analysis

A simple search provides a great example of language change:

Google Books Ngram Viewer

Graph these comma-separated phrases: revolution,socialism,democracy case-insensitive

between 1800 and 2000 from the corpus English with smoothing of 3 Search lots of books



How does Disney do it?

Studio
Technology

Goals

- Each ingest/job/workflow creates a foundation that can be built on.
- Provide self service tools to prevent spreadmart vs. silo vs. data warehouse flip-flopping.
- Increasing quality of data

Personas: Business

- Information Worker
  - Excel, Powerpoint
  - Prepared BI reports
  - Light Statistics
- Business Analyst
  - Excel, Powerpoint
  - COTS Reporting tool
  - Light Statistics
- Data Analyst
  - Excel, Powerpoint
  - COTS Reporting tool
  - SQL

Personas: Data Scientist

Objective

To use the “right” data analysis techniques to deliver business value.

Skills

- Required: Python & SQL; Nice to have: Java, Scala
- Machine Learning, Statistics, Deep Learning
- Data wrangling skills
- Distributed systems & algorithms
- Data Sampling, approximate aggregations, extrapolation
- Scientific Method - Notebooks
- Data communication, visualization
- Cloud services, Linux CLI
- Bonus: NLP, image recognition

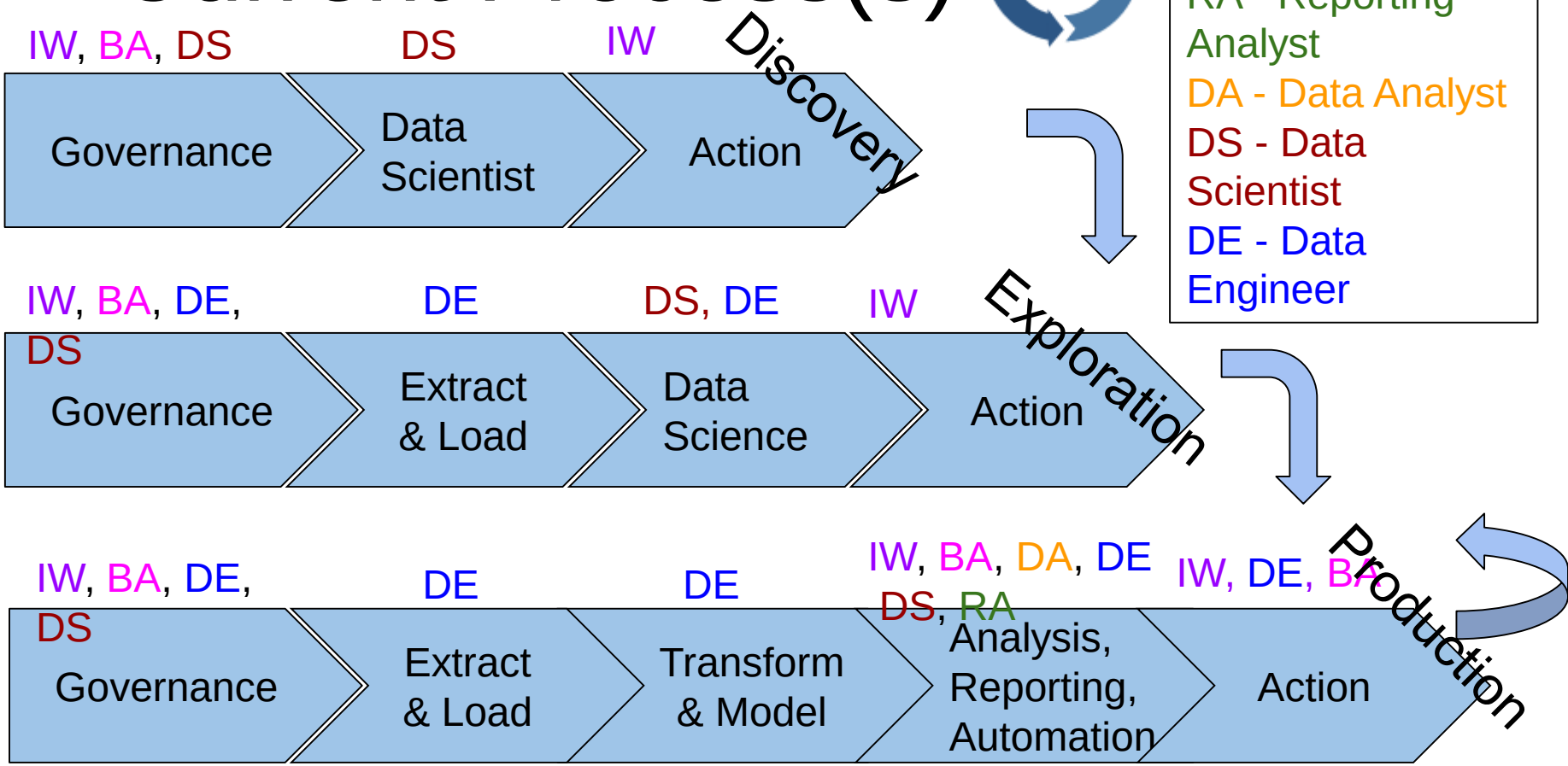
Personas: Technical

- Data Engineer
  - Distributed Systems, Stream Processing
  - Tools, Infrastructure, Frameworks, Services
  - Java, Scala, SQL, Python, R, Bash/Zsh
  - Linux, Git, DevOps, Cloud
  - Medium Stats
  - Medium ML/DL
  - Hadoop, Yarn, HDFS, ElasticSearch
- Reporting Programmer Analyst

  COTS Reporting tool

  SQL
- Technical Analyst
  - Excel, Powerpoint
  - COTS Reporting tool
  - SQL
  - Medium-Strong Statistics
  - Light ML
  - Zeppelin
- System Engineer
  - Infrastructure
  - Automation
  - Continuous Delivery
  - Linux Optimization
  - Monitoring

Current Process(s)



Legend

IW - Info Worker

BA - Business Analyst

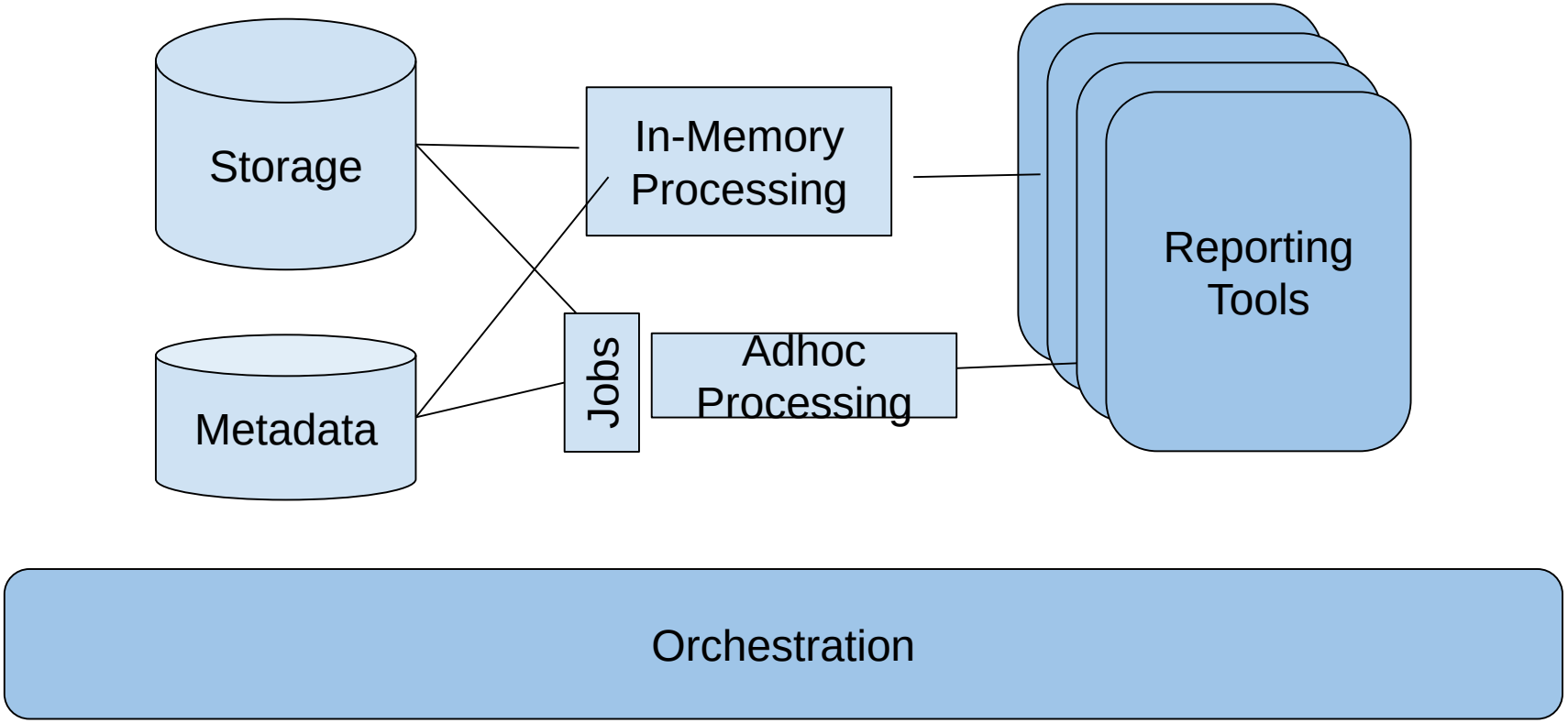
RA - Reporting Analyst

DA - Data Analyst

DS - Data Scientist

DE - Data Engineer

Data Lake


Storage
Metadata
In-Memory Processing
Jobs
Adhoc Processing
Reporting Tools
Orchestration

Business Request	Examples (labels)	Algorithm
Product-to-product	Similar products… Products like this...	Content Filtering TF-IDF Similarity Boolean Model for Approval Cosine similarity for Scoring Tag Affinity
Personalized v1	You might like… Recommended for you… Category for you...	Collaborative Filtering DIMSUM for all-pairs similarity Cold Start = Popular
Personalized v2	You might like… Recommended for you… Category for you...	Hybrid Recommender ALS, SVD, Matrix Factorization Cold Start = Popular / Content Filter
Popular	Most popular of all time... Most popular this month… Popular in Category...	Custom
Because You Purchased	Because you watched [Title]...	Topic Models
People Who Purchased Also Purchased	People who purchased also purchased… Recommended for you...	Latent Factor Models / Matrix Factorization
Trending	Trending this week… Trending this month… Trending in category...	Deviate from predictions

Cultural Cautions

- ⚠ Lack of Strategic thinking
- ⚠ Missing automation
- ⚠ “Tell us something we didn’t know”
- ⚠ I found a case where you’re wrong
- ⚠ Data hoarding
- ⚠ Political blunders
- ⚠ Chasing “Shiny”, no baseline value
- ⚠ Accidental bias
- ⚠ Query is Slow! OMG, Missing Data!

What do real world data science needs look like?

Data Science needs: Current Examples from the City of LA

- Elected Officials
  - Affordable Housing Risk Scoring and Covent Risk Scoring
  - Downtown Transportation Analysis
  - Street Pavement Prioritization and Early Warning System
  - Property Values and Affordable and Low Income Housing
  - LAPD Recruitment Performance Dashboard
  - CAP tracking enhancements, dashboards, and integration with other Personnel systems
  - Attrition prediction tool
  - Homelessness Services Matcher

Data Science needs: Current Examples from the City of LA

- Information Technology Agency
  - ServiceNow Analysis and Dashboard
- Office of Finance
  - Call center operational improvements
  - Bill Collections
  - Revenue Forecasting
- Department of Transportation
  - Projecting Parking Demand
- Department of Cultural Affairs
  - Cultural Events Analytics, Neighborhood Arts Profile, and Cultural Desert Discovery

City of LA: Specific Examples

- **Downtown Transportation Analysis**
  - Analysis of bicyclists and pedestrian use on Spring and Main both before and after Spring and Main Forward project. This will build on existing work from CSULA and LADOT. The Downtown configuration analysis for Project Downtown streets could show an ideal mix of various improvements and interventions. We hope to be able to project throughput for various streets downtown in different configurations.

City of LA: Specific Examples

- **LAPD Recruitment Performance Dashboard**
  - LAPD Personnel recruiters' main metric for success for recruiting candidates is the number of tests administered. However, there is limited visibility into which recruiters are testing the highest proportion of successful candidates, what strategies are most viable, or which geographic areas and events yield the best results. In preparation for anticipated surges in retirement, smarter recruiting is essential. A paradigm shift that is outcome-oriented will lead to greater accountability and flexibility as LAPD and Personnel strive to meet hiring goals. For this project, we wish to provide recruiters with new metrics of success.