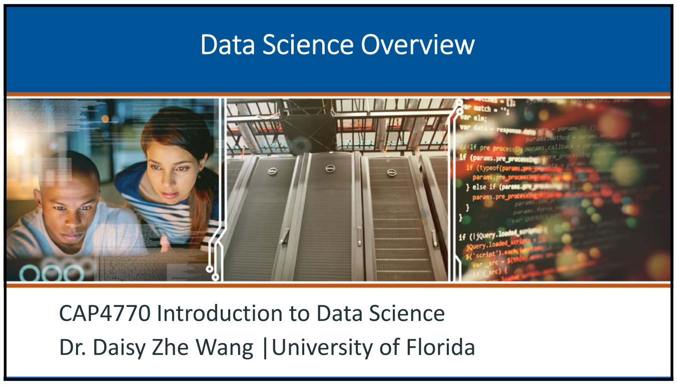
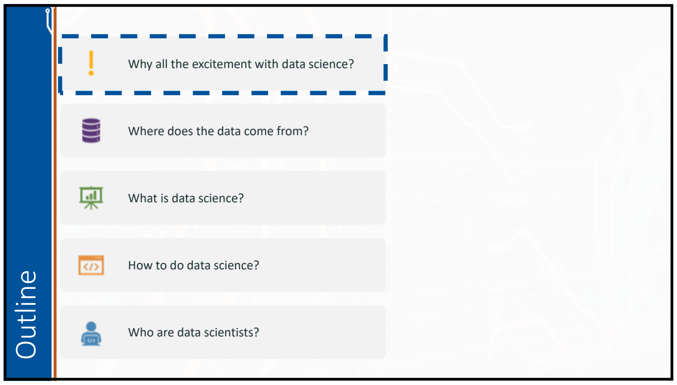


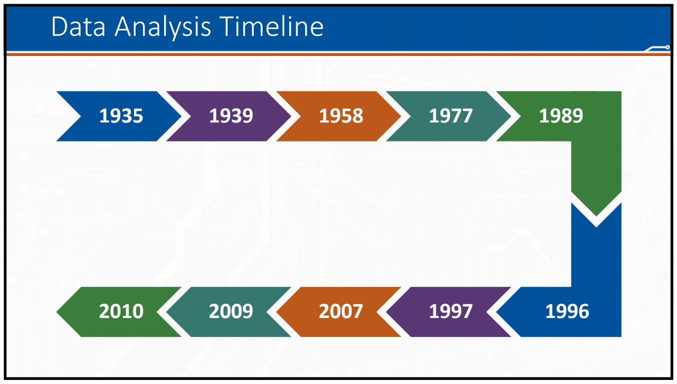
Data Science Overview

CAP4770 Introduction to Data Science
Dr. Daisy Zhe Wang |University of Florida



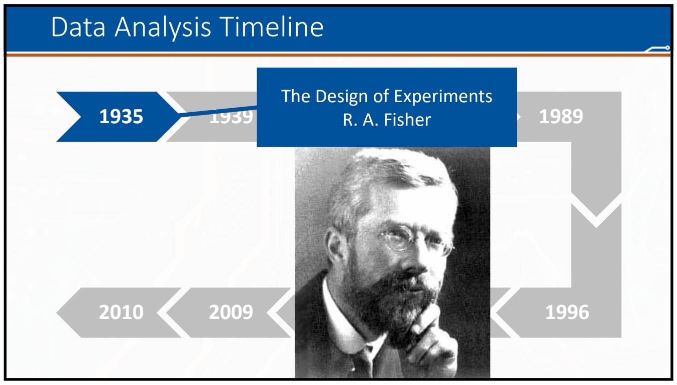
Outline

- ! Why all the excitement with data science?
- Where does the data come from?
- What is data science?
- How to do data science?
- Who are data scientists?



Data Analysis Timeline

1935 → 1939 → 1958 → 1977 → 1989 → 1996 → 1997 → 2007 → 2009 → 2010



Data Analysis Timeline

1935

1939

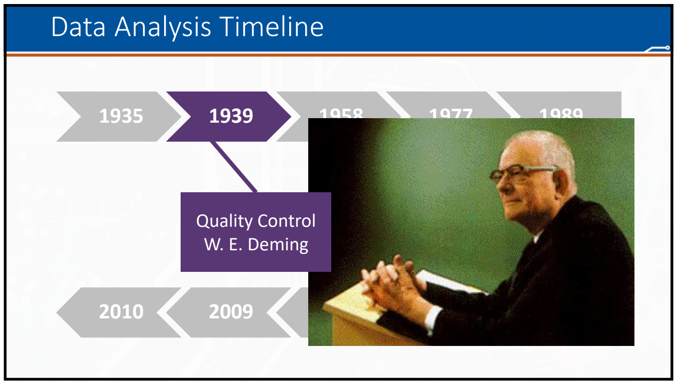
1989

1996

2010

2009

The Design of Experiments
R. A. Fisher



Data Analysis Timeline

1935

1939

1958

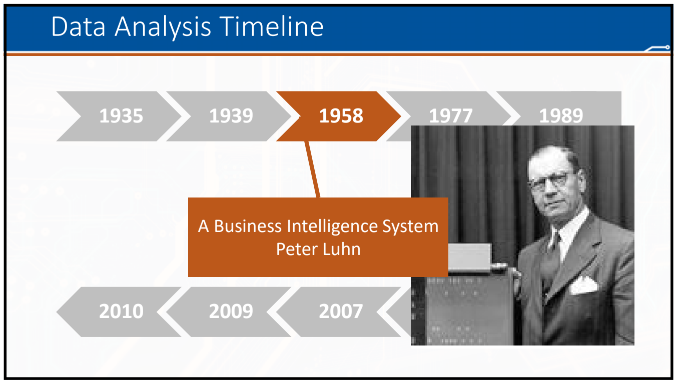
1977

1989

2010

2009

Quality Control
W. E. Deming



Data Analysis Timeline

1935

1939

1958

1977

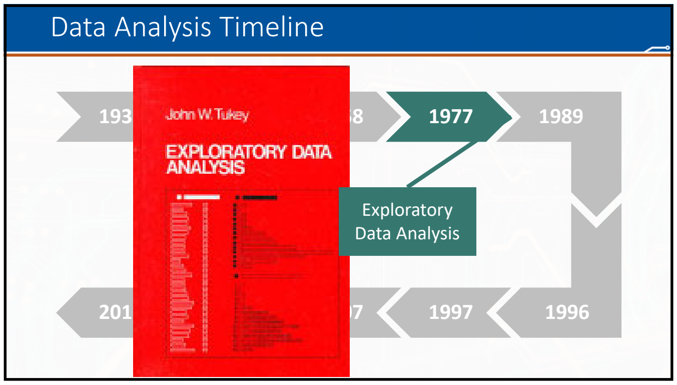
1989

2010

2009

2007

A Business Intelligence System
Peter Luhn



Data Analysis Timeline

193

John W. Tukey

EXPLORATORY DATA ANALYSIS

8

1977

1989

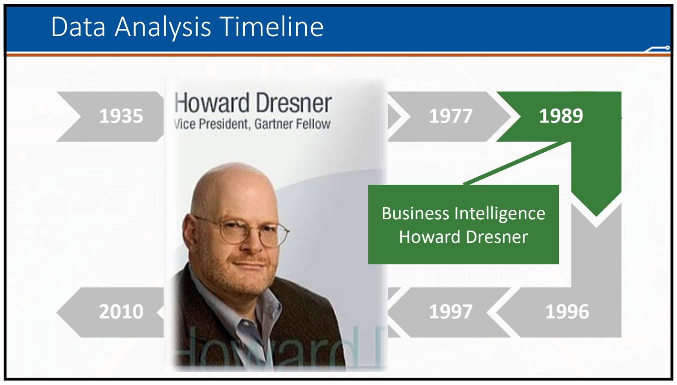
Exploratory Data Analysis

201

7

1997

1996



Data Analysis Timeline

1935

Howard Dresner

Vice President, Gartner Fellow

1977

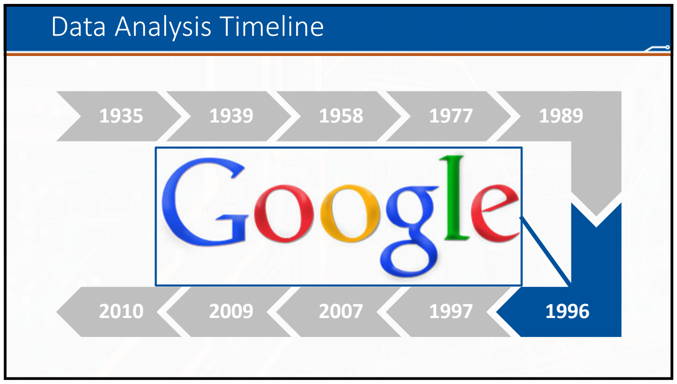
1989

Business Intelligence
Howard Dresner

2010

1997

1996

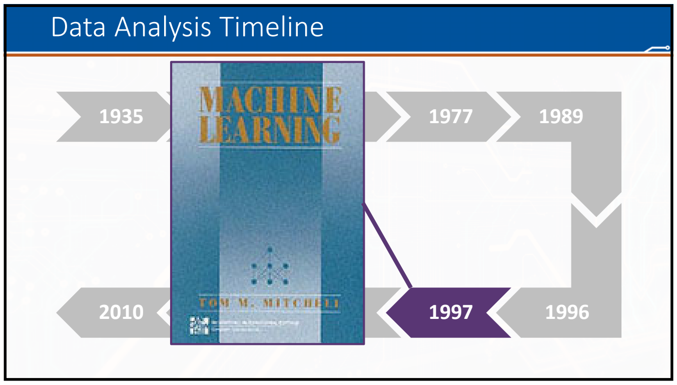


Data Analysis Timeline

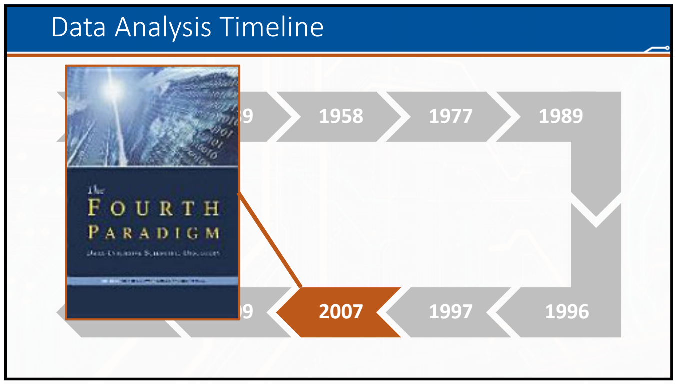
1935 1939 1958 1977 1989

Google

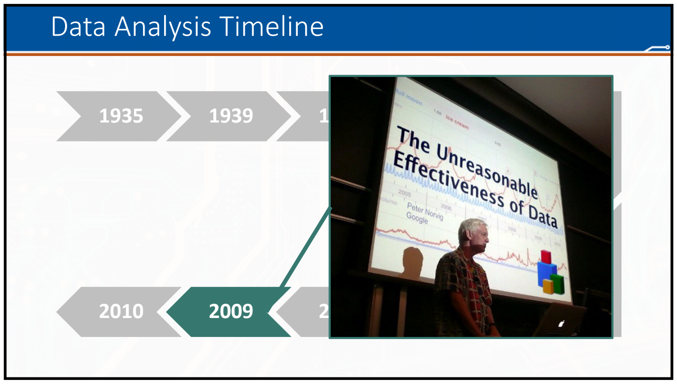
2010 2009 2007 1997 1996



Data Analysis Timeline

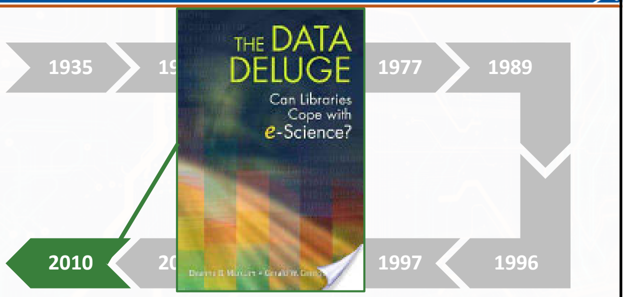


Data Analysis Timeline



Data Analysis Timeline

Data Analysis Timeline



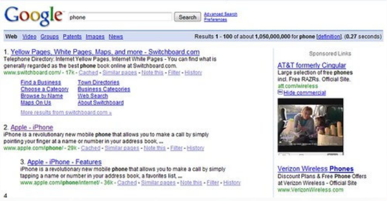
Sponsored Search

Google revenue around $50 bn/year from **marketing/ sponsored search**, 97% of the companies revenue.

Google Adwords and Adsense:
There are around 30 billion search requests a month. Perhaps a trillion events of history between search providers.

Sponsored search uses an **auction** – a pure competition for marketers /advertiser trying to win access to consumers.

In other words, a competition for models of consumers – their likelihood of responding to the ad – and of determining the right bid for the item.



Results **1 - 100** of about **1,050,000,000** for **phone** [definition]. (**0.27** seconds)

1. Yellow Pages, White Pages, Maps, and more - Switchboard.com
Telephone Directory: Internet Yellow Pages, Internet White Pages - You can find what is generally regarded as the best **phone** book online at Switchboard.com.
www.switchboard.com/ - 17k - Cached - Similar pages - Note this - Filter - History
Find a Business · Town Directories · Choose a Category · Business Categories · Browse by Name · Web Search · Maps On Us · About Switchboard
More results from switchboard.com »

2. Apple - iPhone
iPhone is a revolutionary new mobile **phone** that allows you to make a call by simply pointing your finger at a name or number in your address book, ...
www.apple.com/**iphone**/ - 29k - Cached - Similar pages - Note this - Filter - History

3. Apple - iPhone - Features
iPhone is a revolutionary new mobile **phone** that allows you to make a call by simply tapping a name or number in your address book, a favorites list, ...
www.apple.com/**iphone**/internet/ - 36k - Cached - Similar pages - Note this - Filter - History

Sponsored Links

AT&T formerly Cingular
Large selection of free **phones** incl. Free RAZRs. Official Site.
att.com/wireless
Hide commercial

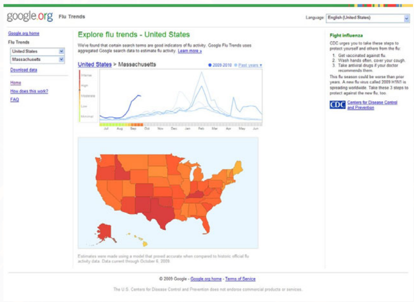
Verizon Wireless **Phones**
Discount Plans & Free **Phone** Offers at Verizon Wireless - Official Site
www.VerizonWireless.com

4

Why all the Excitement with Data Science?

Exciting new effective applications of data analytics e.g., Google Flu Trends:

- Detecting outbreaks two weeks ahead of CDC data
- New models are estimating which cities are most at risk for spread of the Ebola virus.
- Prediction model is built on various data sources, types and analysis.



Explore flu trends - United States

We've found that certain search terms are good indicators of flu activity. Google Flu Trends uses aggregated Google search data to estimate flu activity. Learn more »

Fight influenza

CDC urges you to take these steps to protect yourself and others from the flu:

1. Get vaccinated against flu.
2. Wash hands often, cover your cough.
3. Take antiviral drugs if your doctor recommends them.

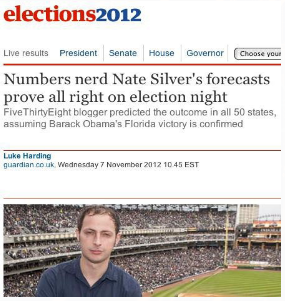
This flu season could be worse than prior years. A new flu virus called 2009 H1N1 is spreading worldwide. Take these 3 steps to protect against the new flu, too.

Estimates were made using a model that proved accurate when compared to historic official flu activity data. Data current through October 6, 2009.

© 2009 Google - Google.org home - Terms of Service

The U.S. Centers for Disease Control and Prevention does not endorse commercial products or services.

Why all the Excitement with Data Science?



elections2012

Live results | President | Senate | House | Governor | Choose your

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

Predicting political champagne and election

Outcome:

The signal and the noise and the noise and the noise **why most predictions fail but some don't** noise and the noise and the noise **nate silver** noise and the noise.

Data and Election 2012

…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

The White House Names Dr. DJ Patil as the **First U.S. Chief Data Scientist**, Feb. 18th 2015

Splunking

Grab data from many machines

Index it

Check for unusual events:

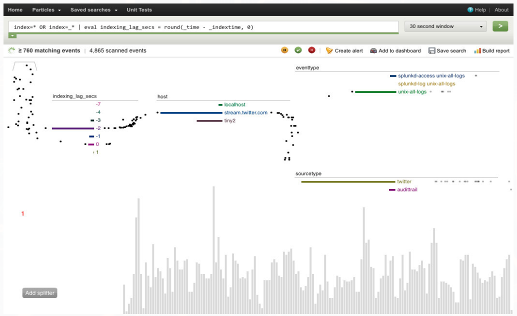
- Disk problems
- Network congestion
- Security attacks

Monitor resources:

- Network
- Memory usage
- Disk use, latency
- Threads

Dashboard for cloud administration.

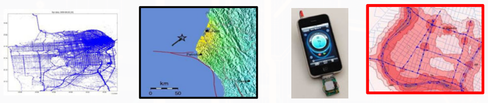
Splunking



```
index=* OR index=_* | eval indexing_lag_secs = round(_time - _indextime, 0)
```

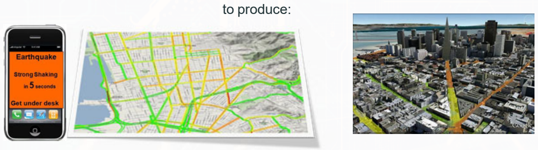
≥ 760 matching events | 4,865 scanned events

Traffic Prediction and Earthquake Warning



Crowdsourcing + physical modeling + sensing + data assimilation

to produce:



From Alex Bayen, UCB, Director, Institute for Transportation Studies

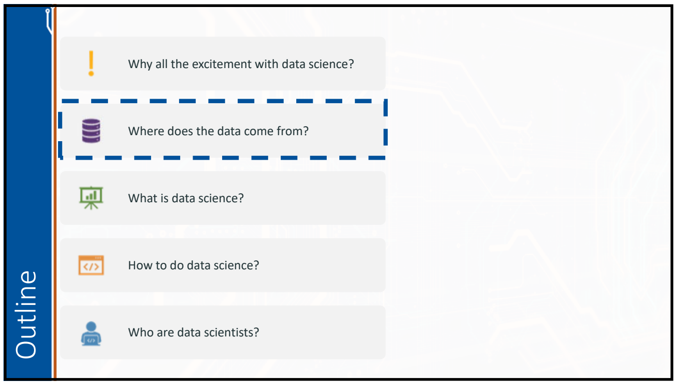
Other Data Science Applications

Transaction Databases → Recommender systems (Netflix), Fraud Detection (Security and Privacy)

Wireless Sensor Data → Smart Home, Real-time Monitoring, Internet of Things Applications

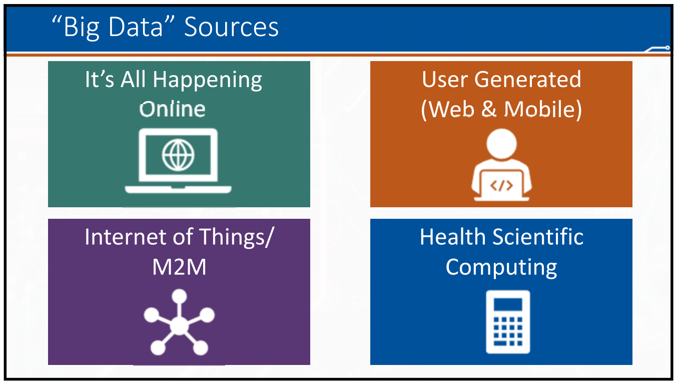
Text Data, Social Media Data → Product Review and Consumer Satisfaction (Facebook, Twitter, LinkedIn), E-discovery

Genotype and Phenotype Data → Epic, 23andme, Patient-Centered Care, Personalized Medicine



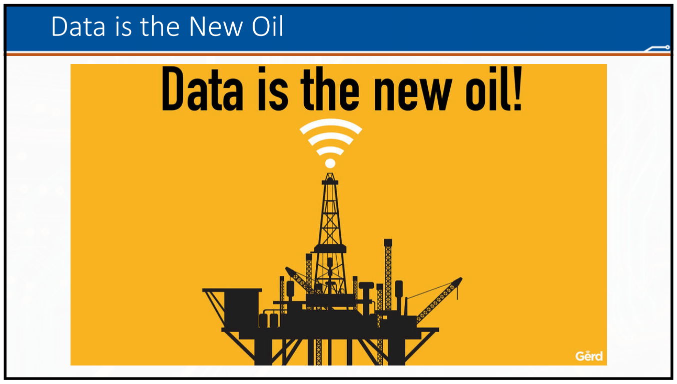
Outline

- Why all the excitement with data science?
- Where does the data come from?
- What is data science?
- How to do data science?
- Who are data scientists?

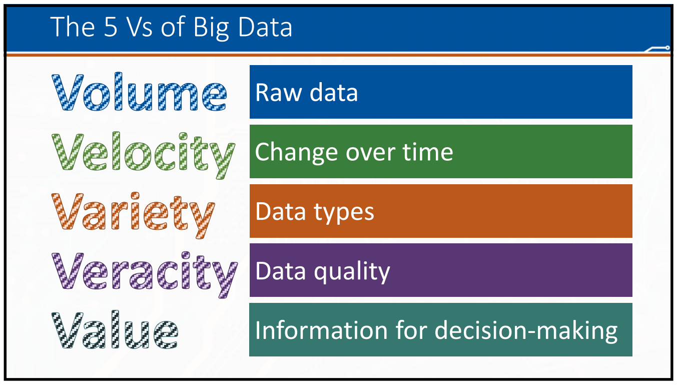


"Big Data" Sources

- It's All Happening Online
- User Generated (Web & Mobile)
- Internet of Things/ M2M
- Health Scientific Computing

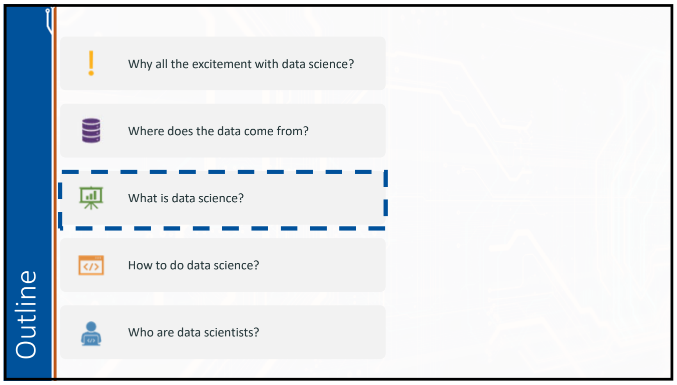


Data is the New Oil



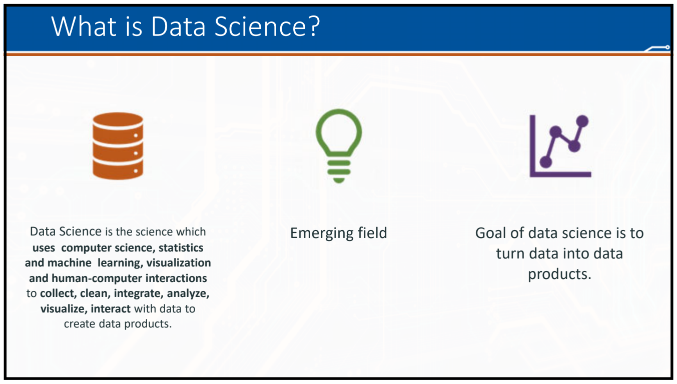
The 5 Vs of Big Data

Volume	Raw data
Velocity	Change over time
Variety	Data types
Veracity	Data quality
Value	Information for decision-making



Outline

- Why all the excitement with data science?
- Where does the data come from?
- What is data science?
- How to do data science?
- Who are data scientists?

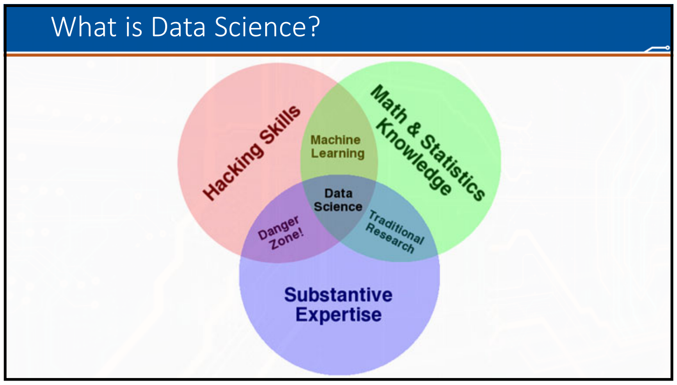


What is Data Science?

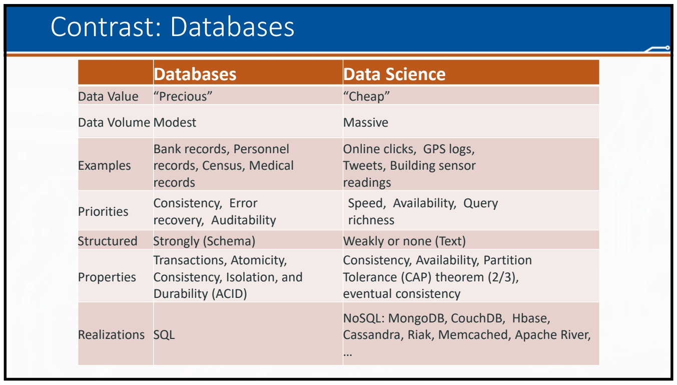
Data Science is the science which **uses computer science, statistics and machine learning, visualization and human-computer interactions** to **collect, clean, integrate, analyze, visualize, interact** with data to create data products.

Emerging field

Goal of data science is to turn data into data products.

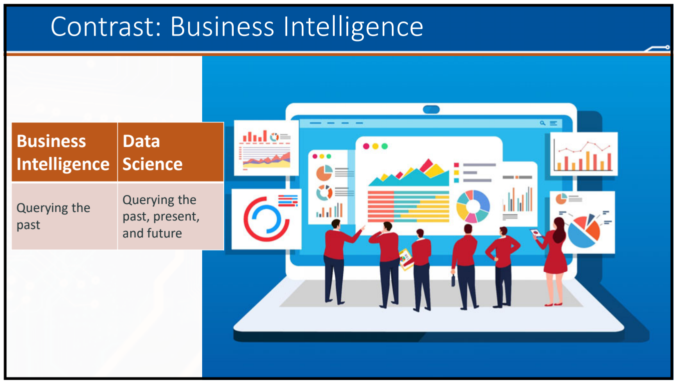


What is Data Science?



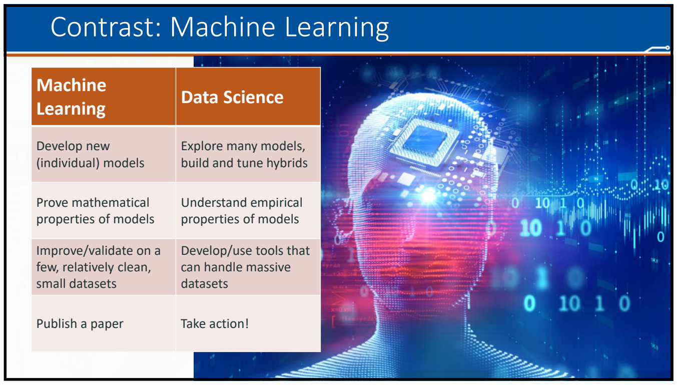
Contrast: Databases

	Databases	Data Science
Data Value	"Precious"	"Cheap"
Data Volume	Modest	Massive
Examples	Bank records, Personnel records, Census, Medical records	Online clicks, GPS logs, Tweets, Building sensor readings
Priorities	Consistency, Error recovery, Auditability	Speed, Availability, Query richness
Structured	Strongly (Schema)	Weakly or none (Text)
Properties	Transactions, Atomicity, Consistency, Isolation, and Durability (ACID)	Consistency, Availability, Partition Tolerance (CAP) theorem (2/3), eventual consistency
Realizations	SQL	NoSQL: MongoDB, CouchDB, Hbase, Cassandra, Riak, Memcached, Apache River, …



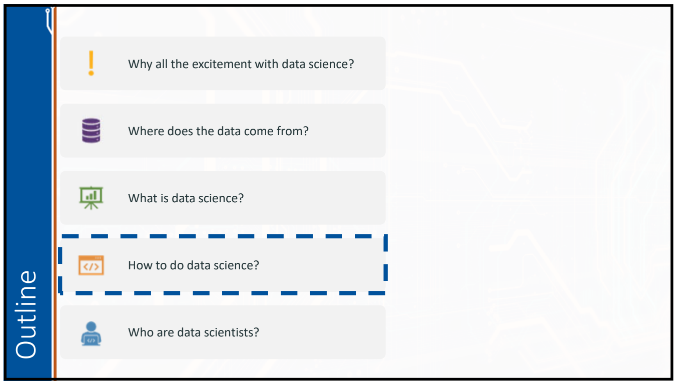
Contrast: Business Intelligence

Business Intelligence	Data Science
Querying the past	Querying the past, present, and future



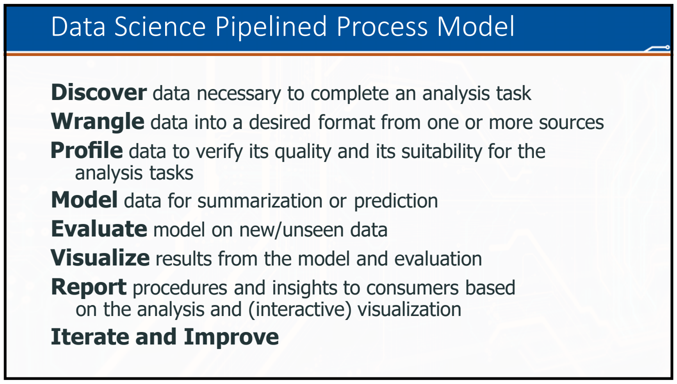
Contrast: Machine Learning

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!



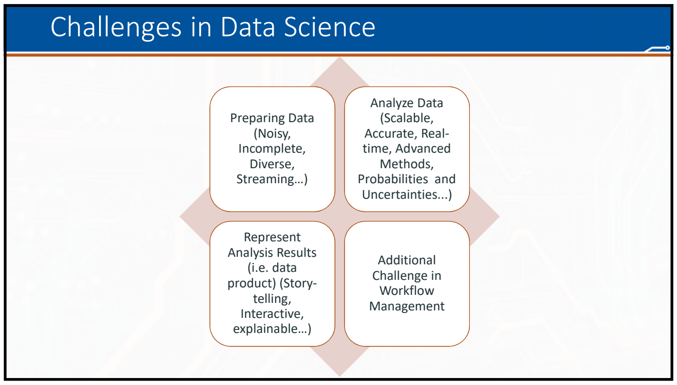
Outline

- ! Why all the excitement with data science?
- Where does the data come from?
- What is data science?
- How to do data science?
- Who are data scientists?



Data Science Pipelined Process Model

Discover data necessary to complete an analysis task
Wrangle data into a desired format from one or more sources
Profile data to verify its quality and its suitability for the analysis tasks
Model data for summarization or prediction
Evaluate model on new/unseen data
Visualize results from the model and evaluation
Report procedures and insights to consumers based on the analysis and (interactive) visualization
Iterate and Improve



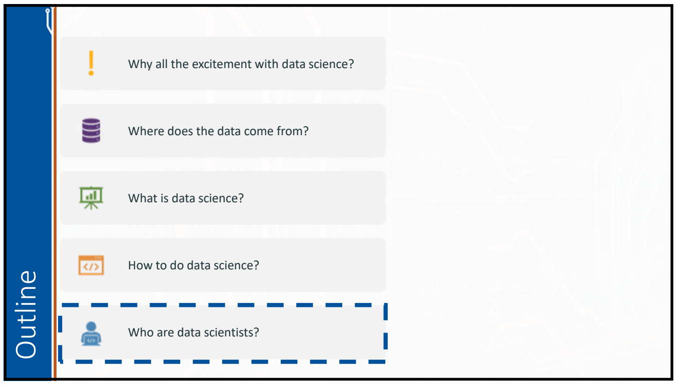
Challenges in Data Science

Preparing Data (Noisy, Incomplete, Diverse, Streaming…)

Analyze Data (Scalable, Accurate, Real-time, Advanced Methods, Probabilities and Uncertainties…)

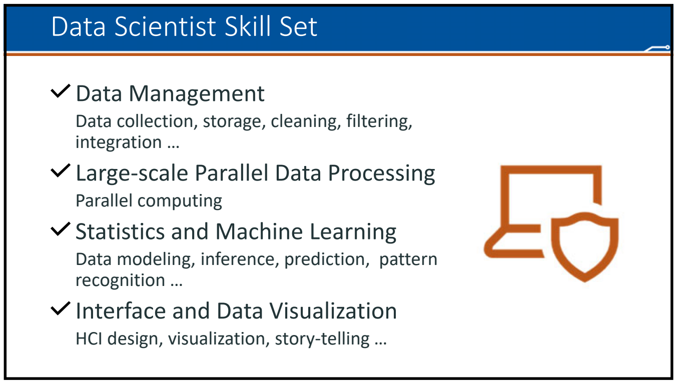
Represent Analysis Results (i.e. data product) (Story-telling, Interactive, explainable…)

Additional Challenge in Workflow Management



Outline

- ! Why all the excitement with data science?
- Where does the data come from?
- What is data science?
- How to do data science?
- Who are data scientists?



Data Scientist Skill Set

- ✓ Data Management
  Data collection, storage, cleaning, filtering, integration …
- ✓ Large-scale Parallel Data Processing
  Parallel computing
- ✓ Statistics and Machine Learning
  Data modeling, inference, prediction, pattern recognition …
- ✓ Interface and Data Visualization
  HCI design, visualization, story-telling …

Analyzing the Analysts

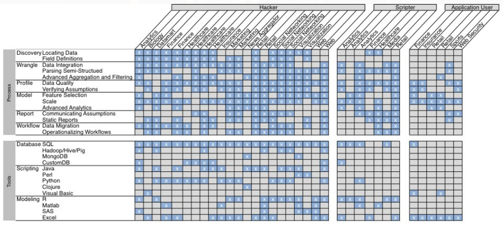
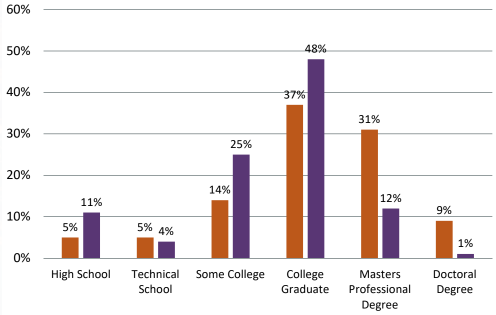

Hacker
Scripter
Application User
Analytics
Biology
Datamart
Finance
Finance
Healthcare
Healthcare
Healthcare
Healthcare
Insurance
Marketing
Marketing
News Aggregator
Retail
Retail
Social Networking
Social Networking
Social Networking
Visualization
Visualization
Web
Web
Analytics
Analytics
Analytics
Finance
Healthcare
Media
Retail
Finance
Insurance
Retail
Retail
Sports
Web Security
Process
Discovery Locating Data
Field Definitions
Wrangle Data Integration
Parsing Semi-Structued
Advanced Aggregation and Filtering
Profile Data Quality
Verifying Assumptions
Model Feature Selection
Scale
Advanced Analytics
Report Communicating Assumptions
Static Reports
Workflow Data Migration
Operationalizing Workflows
Tools
Database SQL
Hadoop/Hive/Pig
MongoDB
CustomDB
Scripting Java
Perl
Python
Clojure
Visual Basic
Modeling R
Matlab
SAS
Excel
X

Fig. 1. Respondents, Challenges and Tools. The matrix displays interviewees (grouped by archetype and sector) and their corresponding challenges and tools. *Hackers* faced the most diverse set of challenges, corresponding to the diversity of their workflows and toolset. *Application users* and *scripters* typically relied on the IT team to perform certain tasks and therefore did not perceive them as challenges.

From Kandel, Paepcke, Hellerstein and Heer, "Enterprise Data Analysts and Visualization: An Interview Study", IEEE VAST 2012

Data Science Requires Higher Education



40% of data science professionals have an advanced degree and nearly one in ten have a doctorate. In contrast, less than 1% of BI professionals have a PhD. (EMC, 2011)

Big Data Science in UF and COE

UF Pre-Eminence Hiring

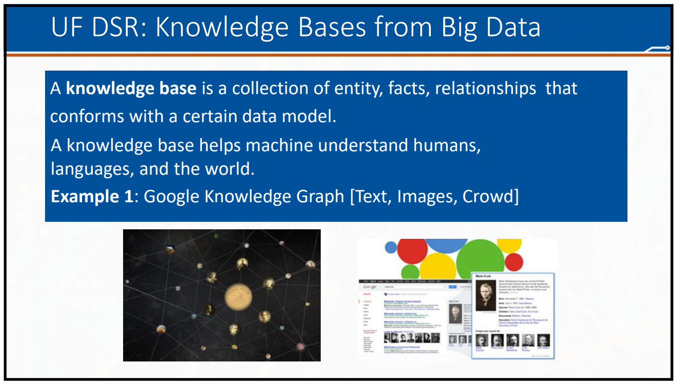
UF Informatics Institute

UF HyPerGator Systems

UF Data Science Research Lab

Student Organization: Data Science and Informatics (DSI)

NSF Center for Big Learning at UF

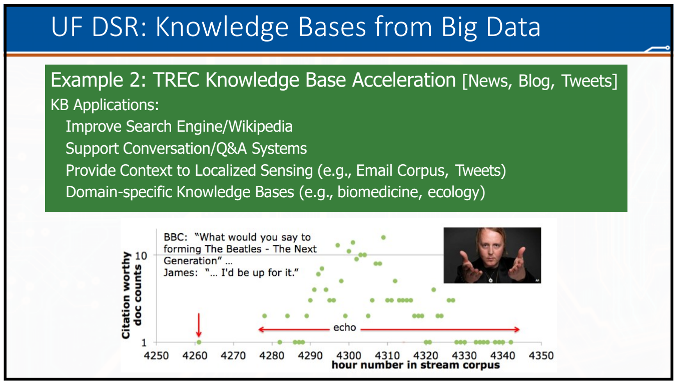


UF DSR: Knowledge Bases from Big Data

A **knowledge base** is a collection of entity, facts, relationships that conforms with a certain data model.

A knowledge base helps machine understand humans, languages, and the world.

Example 1: Google Knowledge Graph [Text, Images, Crowd]



UF DSR: Knowledge Bases from Big Data

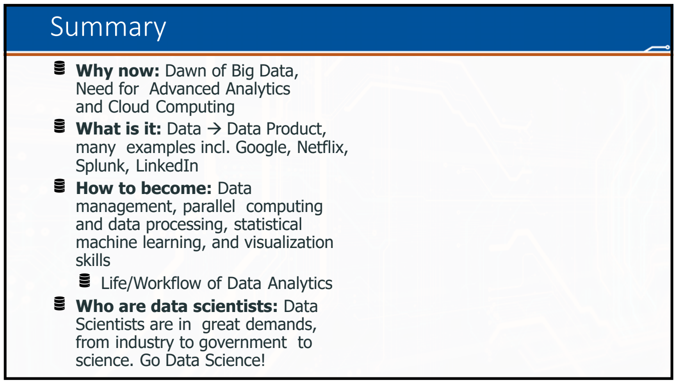
Example 2: TREC Knowledge Base Acceleration [News, Blog, Tweets]

KB Applications:

- Improve Search Engine/Wikipedia
- Support Conversation/Q&A Systems
- Provide Context to Localized Sensing (e.g., Email Corpus, Tweets)
- Domain-specific Knowledge Bases (e.g., biomedicine, ecology)

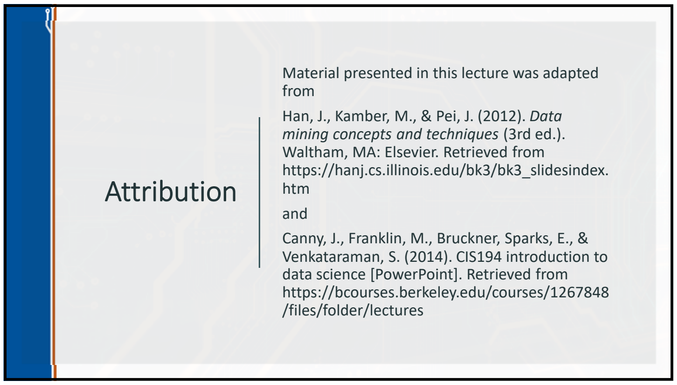
BBC: "What would you say to forming The Beatles - The Next Generation" …
James: "… I'd be up for it."

Citation worthy doc counts
10
1
echo
4250
4260
4270
4280
4290
4300
4310
4320
4330
4340
4350
hour number in stream corpus



Summary

- **Why now:** Dawn of Big Data, Need for Advanced Analytics and Cloud Computing
- **What is it:** Data → Data Product, many examples incl. Google, Netflix, Splunk, LinkedIn
- **How to become:** Data management, parallel computing and data processing, statistical machine learning, and visualization skills
  - Life/Workflow of Data Analytics
- **Who are data scientists:** Data Scientists are in great demands, from industry to government to science. Go Data Science!



Attribution

Material presented in this lecture was adapted from

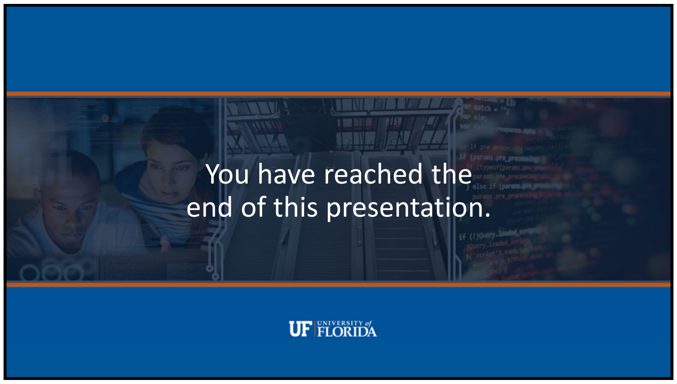
Han, J., Kamber, M., & Pei, J. (2012). *Data mining concepts and techniques* (3rd ed.). Waltham, MA: Elsevier. Retrieved from https://hanj.cs.illinois.edu/bk3/bk3_slidesindex.htm

and

Canny, J., Franklin, M., Bruckner, Sparks, E., & Venkataraman, S. (2014). CIS194 introduction to data science [PowerPoint]. Retrieved from https://bcourses.berkeley.edu/courses/1267848/files/folder/lectures

Image Credits

https://live.staticflickr.com/3094/2363451168_4386113c29.jpg
https://makingscience.royalsociety.org/s/rs/people/fst00034451
https://upload.wikimedia.org/wikipedia/commons/7/73/W._Edwards_Deming.jpg
https://mikeurbonas.files.wordpress.com/2012/08/hans-peter-luhn01.png
https://upload.wikimedia.org/wikipedia/commons/2/20/Google-Logo.svg
https://books.google.com/books/about/Exploratory_Data_Analysis.html?id=UT9dAAAAIAAJ
https://live.staticflickr.com/5504/31254136671_66e440182a_b.jpg
https://static1.squarespace.com/static/5150aec6e4b0e340ec52710a/t/51525c33e4b0b3e0d10f77ab/1364352052403/Data_Science_VD.png?format=750w
https://books.google.com/books/about/The_Fourth_Paradigm.html?id=oGs_AQAAIAAJ
https://books.google.com/books?id=0wjB8H2cUXcC&dq=data+deluge&source=gbs_navlinks_s
https://live.staticflickr.com/5202/5255522197_4a2fb3ca1e.jpg
http://trec-kba.org/trec-kba-streaming-slot-filling-example.png
https://live.staticflickr.com/2604/3993294706_bae6d14b4f_b.jpg
https://commons.wikimedia.org/wiki/File:Business_Intelligenece,_Govard_Drezner.jpg



You have reached the end of this presentation.