CSE 590
Data Science Fundamentals

Introduction

Klaus Mueller

Computer Science Department
Stony Brook University and SUNY Korea

What Is data Science?

What Does It Enable?

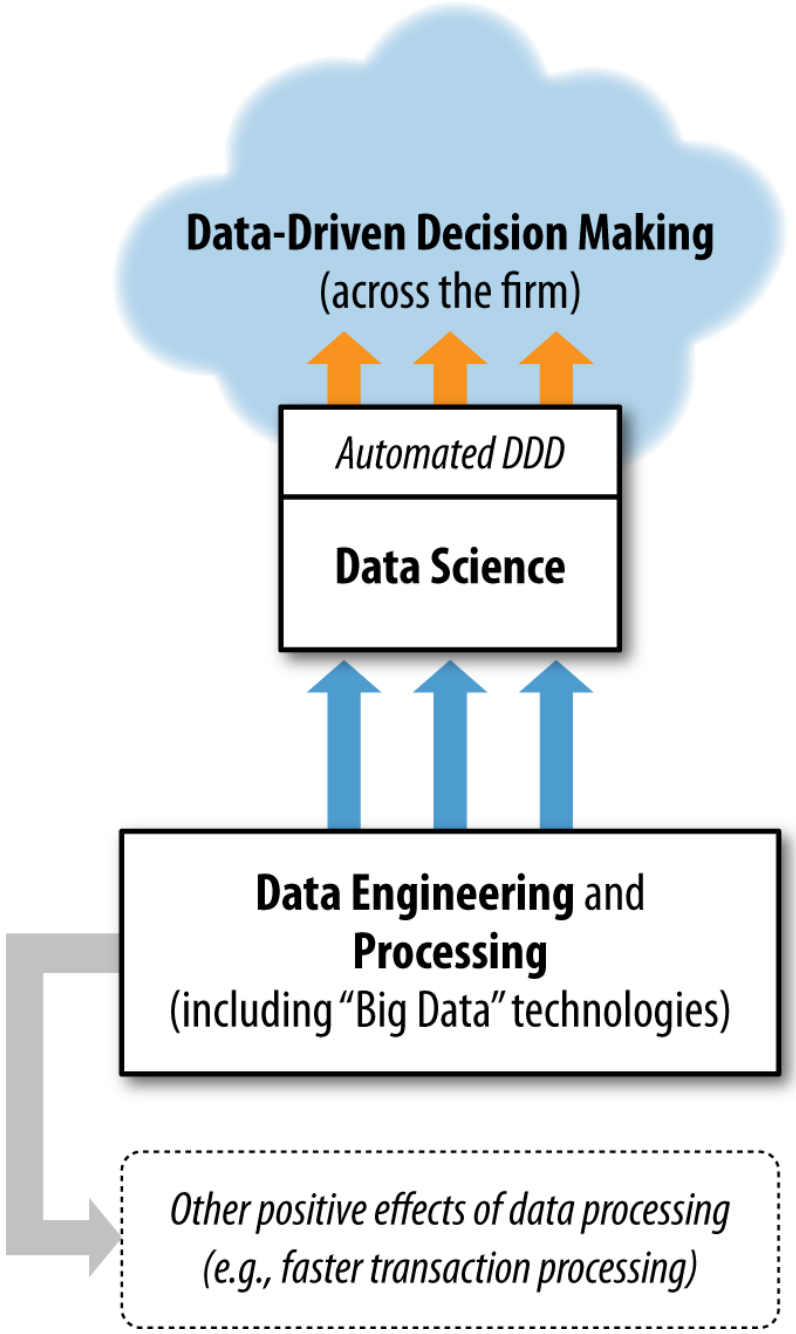
What Does It require?

Isn't It Just data Mining?

DDD

Make decisions based on data

- not purely on intuition and long business experience
- use a combination of these

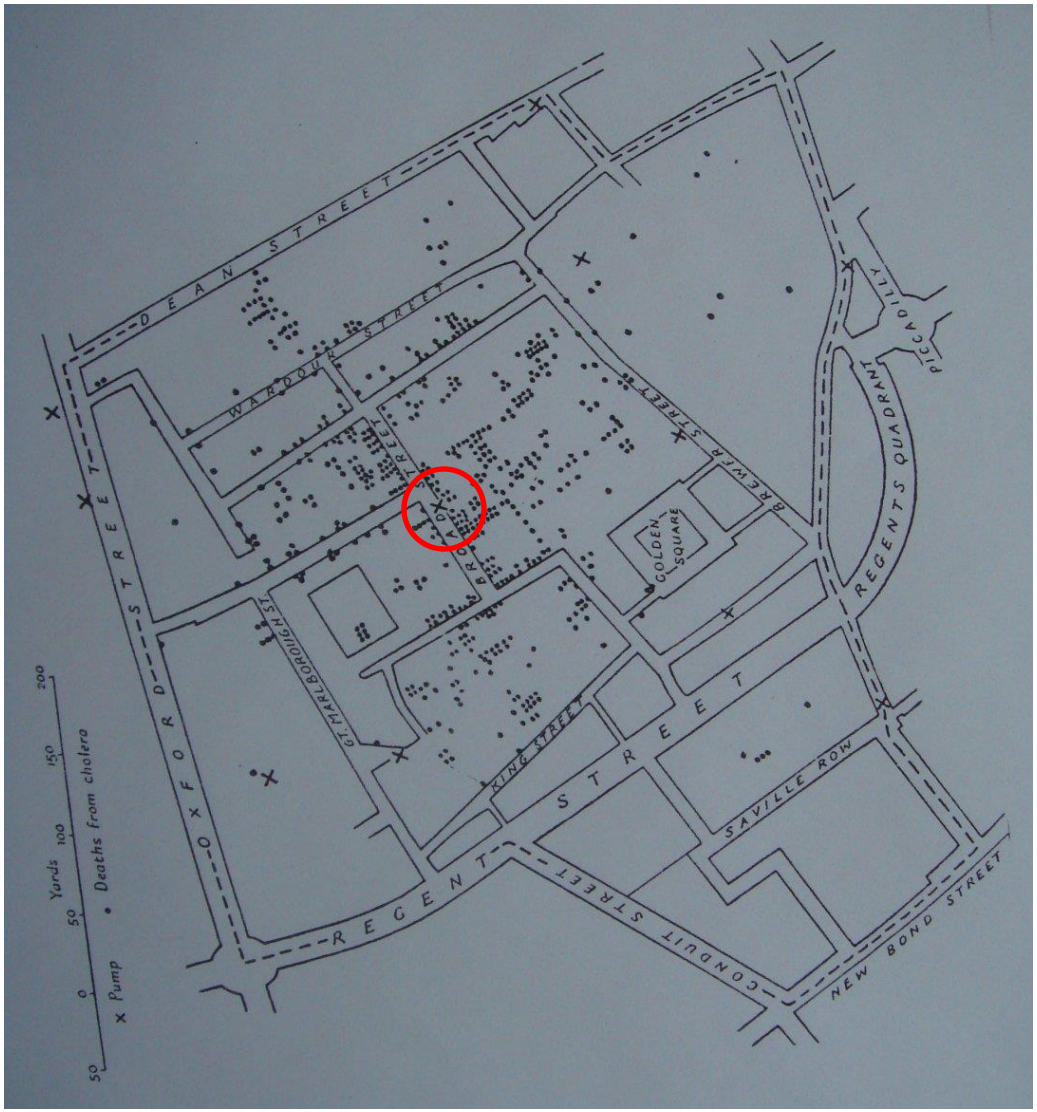


Early Application of Data Science

Dr. John Snow's London Cholera Map (1854)

- data collection
- data assimilation
- statistical testing
- visualization
- computational analysis (brain)
- domain knowledge

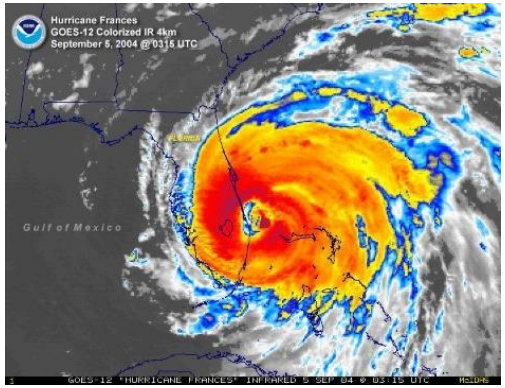
Very early example of data science



Hurricane Francis (2004)

How Walmart made a huge profit using principles of data science

- Linda Dillman, Walmart CIO



What are customers likely to buy when a hurricane is approaching?

- look at past local hurricanes (like Hurricane Charley)
- flashlights, water, sure….
- everyone will stock these
- but also beer and strawberry Pop-Tarts
- look for outliers and unusual patterns….
- it will give your business the edge

The race for the Shopping Moms (2012)

How Walmart competitor Target won that race

Key observations

- once moms buy diapers they will buy everything else in the store too
- everyone will have good deals on diapers to attract the moms
- key is to hook moms before the baby is even out – and so beat the competition
- use predictive analytics from past data to detect revealing changes in shopping behavior (diet, wardrobe, vitamins, …)
- then use targeted marketing – this will give you the edge
- unfortunately this backfired as a bit too creepy …

Data Science Has Two Ingredients

The Data

+

The capability to extract useful knowledge from data (The Data Scientists)

The Amazing Rise of Signet Bank (1990's)

Back in those days all credit cards had uniform pricing

- profit was limited
- now there is pricing, credit limits, cash back, loyalty points, etc.

Richard Fairbank and Nigel Morris changed that using principles from data science

- gave customers random terms on their credit cards – called *scientific tests*
- this incurred losses, but collected valuable data (=business assets)
- profiling enabled accurate modeling of profitability per customer
- after a few years this scheme became profitable
- gave rise to a highly successful credit card

Capital One

Other Early Examples

Data is Power

amazon

facebook

But Beware..

Not all data attributes are useful

- which one is?

If you look too hard at a set of data, you will find something

- but overfitting might result
- the finding might not generalize

Not all findings are useful knowledge

- how can we tell what is useful?

MODERN DATA SCIENTIST

21th century, requires a mixture of
, computer science, communicatio
who a data scientist is, is equally h
the modern data scientist really i

MATH & STATISTICS

- Machine learning
- Statistical modeling
- Experiment design
- Bayesian inference
- Supervised learning: decision trees, random forests, logistic regression
-
-

PROGRAMMING & DATABASE

- Computer science fundamentals
- Scripting language e.g. Python
- Statistical computing packages, e.g., R
- Databases: SQL and NoSQL
- Relational algebra
-

DOMAIN KNOWLEDGE & SOFT SKILLS

- Passionate about the business
- Curious about data
- Influence without authority
- Hacker mindset
- Problem solver
- Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- Able to engage with senior management
- Story telling skills
- Translate data-driven insights into decisions and actions
- Visual art design
- R packages like ggplot or lattice
- Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

Data Scientists

The U.S. will need 140,000-190,000 predictive analysts and 1.5 million managers/analysts by 2018

- McKinsey Global Institute's June 2011

Why do we need many more knowledgeable managers?

- because data scientists may work for more than one group

Google Flu Trends

Predict emerging flu from search terms in specific regions

Could predict regional outbreaks of flu up to 10 days before reported by the CDC

Nate Silver's Election Predictions

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

Takes a big-picture approach

- use multiple sources of unique data
- combine with historical data
- apply principles of sound statistical analysis

Opportunities galore

IBM

Government achieves significant cost savings and ability to react to potential threats quickly

Government cuts acoustic analysis from hours to **70 Milliseconds**

Utility provider improves prediction of power outages

Utility avoids power failures by analyzing **10 PB** of data in minutes

Hospital detects and intervenes in potentially life-threatening conditions

Hospital analyzes streaming vitals to intervene **24 hours earlier**

Retailer optimizes inventory levels and product mix

Retailer reduces time to run queries by **80%**

Stock exchange reduces time to insights to achieve optimal buying / selling strategies

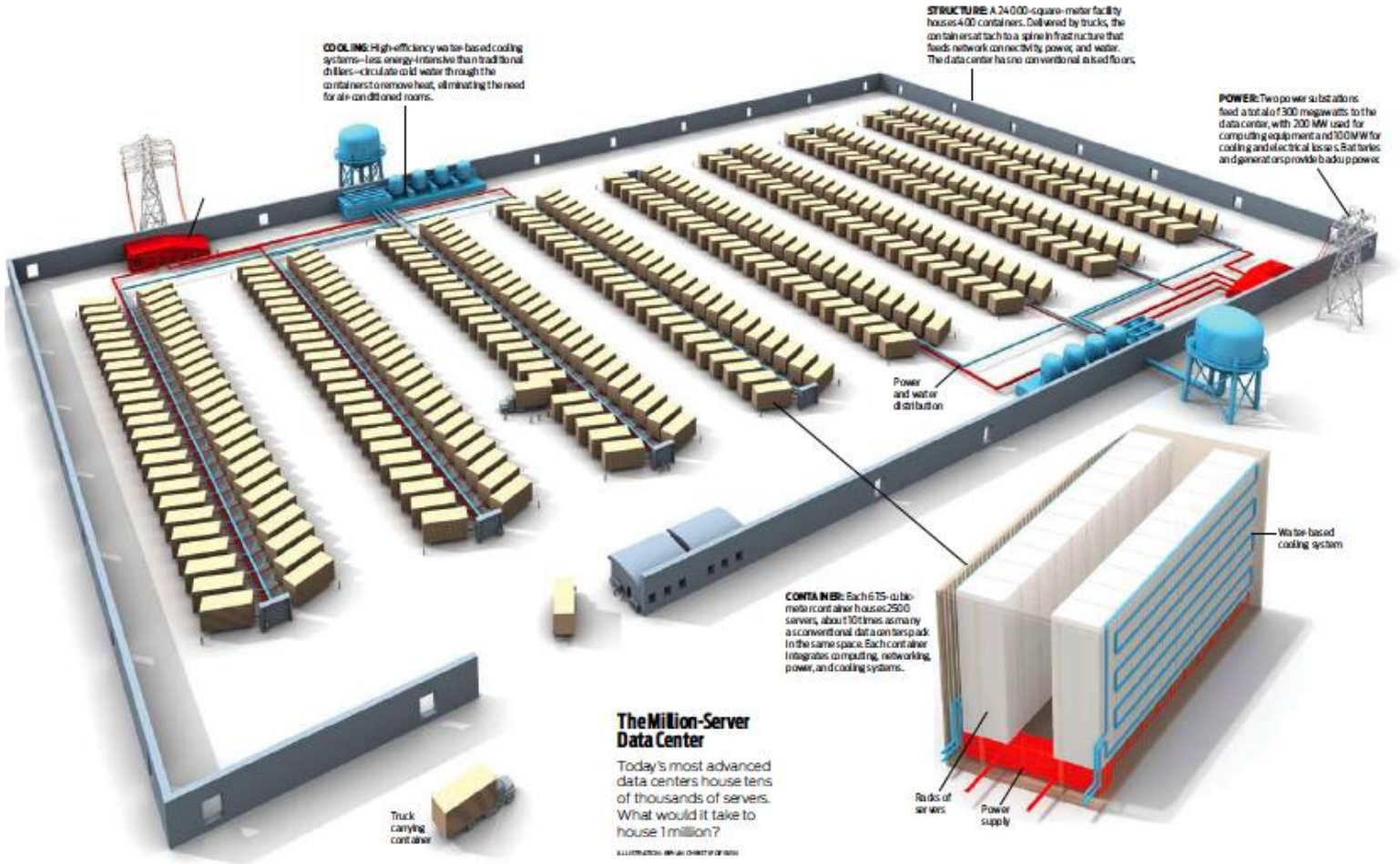
Stock Exchange cuts queries from 26 hours to **2 minutes** on **2 PB**

Telco provider improves ability to quickly address network issues / opportunities

Telco analyses streaming network data to reduce hardware costs by **90%**

Million Server Data Center

Not always needed



Databases vs. data Science

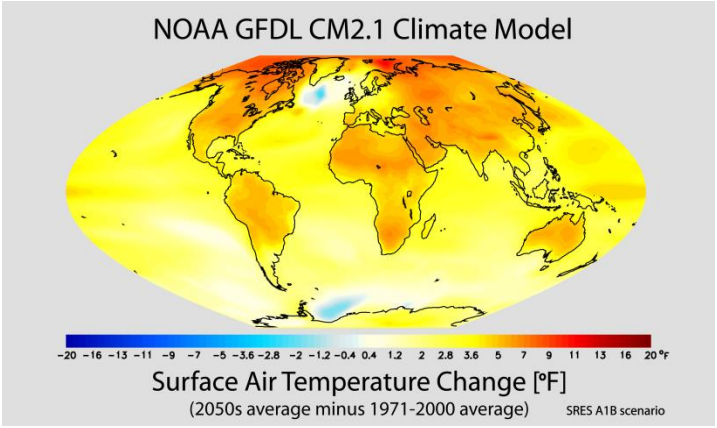
	Databases	Data Science
Data Value	“Precious”	“Cheap”
Data Volume	Modest	Massive
Examples	Bank records, Personnel records, Census, Medical records	Online clicks, GPS logs, Tweets, Building sensor readings
Priorities	Consistency, Error recovery, Auditability	Speed, Availability, Query richness
Structured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

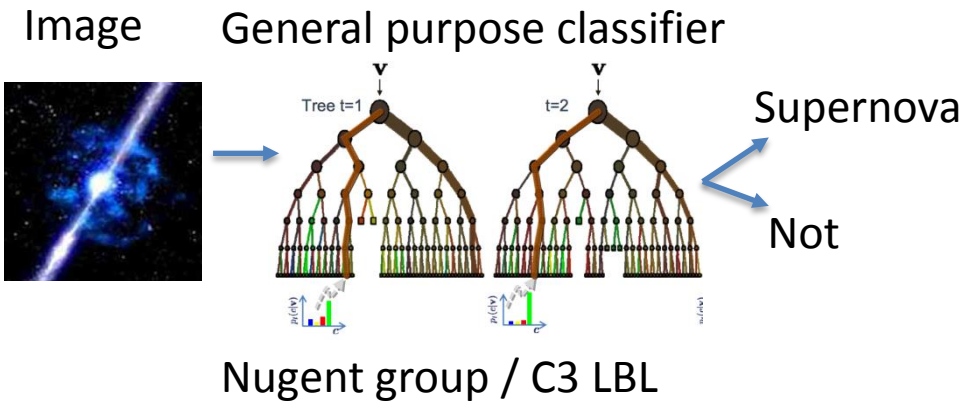
CAP = Consistency, Availability, Partition Tolerance

ACID = Atomicity, Consistency, Isolation and Durability

John Canny, Berkeley

Scientific vs. data-Driven Modeling





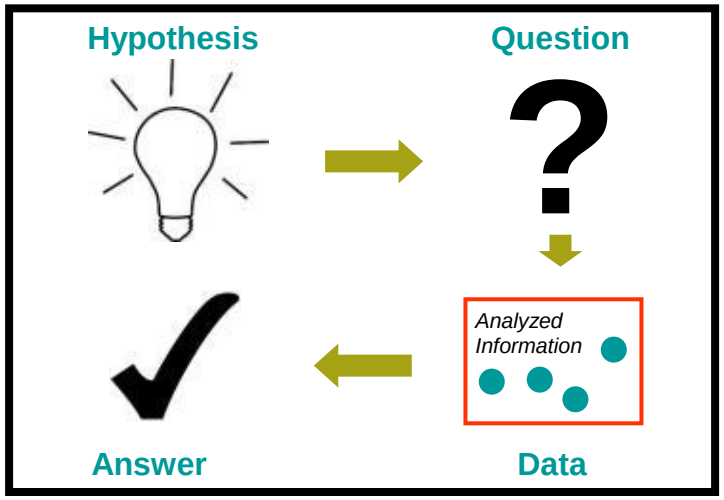
Scientific Modeling	Data-Driven Approach
Physics-based models	General inference engine replaces model
Problem-Structured	Structure not related to problem
Mostly deterministic, precise	Statistical models handle true randomness, and unmodeled complexity
Run on Supercomputer or High-end Computing Cluster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

John Canny, Berkeley

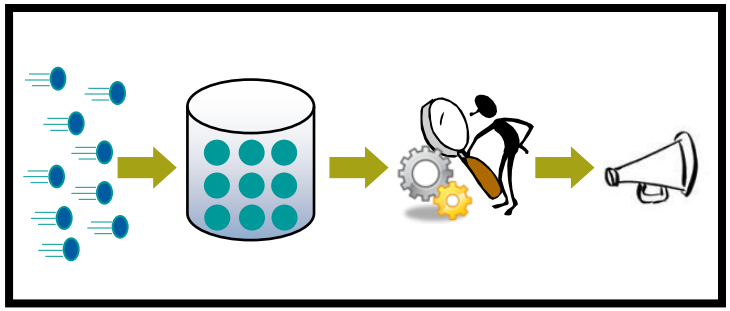
Big Data Approach to Science

Traditional Analytics

Structured & Repeatable
Structure built to store data



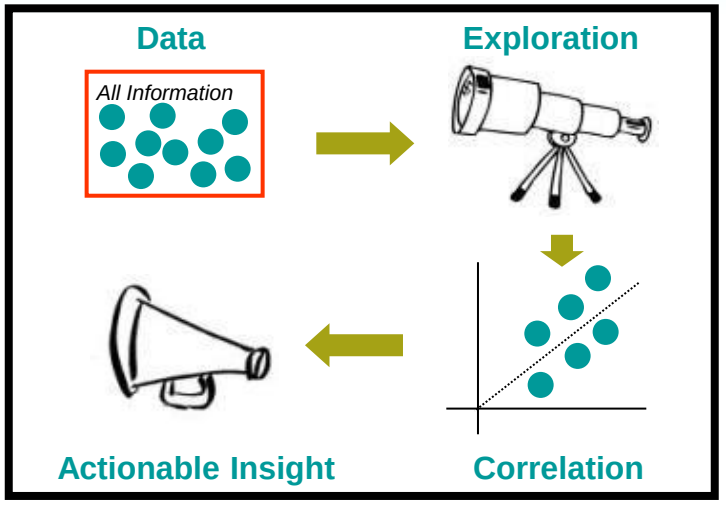
Start with hypothesis
Test against selected data



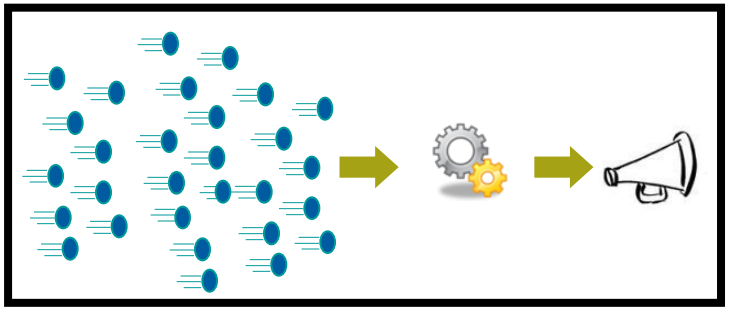
Analyze after landing…

Big Data Analytics

Iterative & Exploratory
Data is the structure



Data leads the way
Explore *all* data, identify correlations



Analyze in motion…

Characteristics of Big Data

Volume

Data at scale

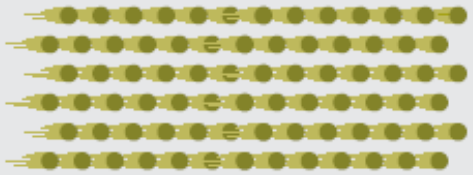
Terabytes to petabytes of data

Variety

Data in many forms

Structured, unstructured, text, multimedia

Velocity



Data in motion

Analysis of streaming data to enable decisions within fractions of a second

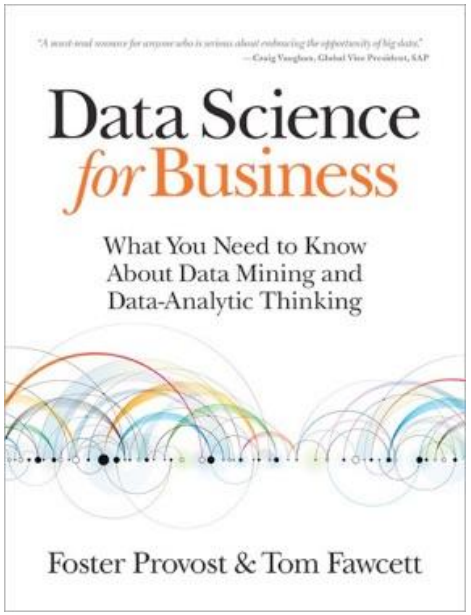
Veracity

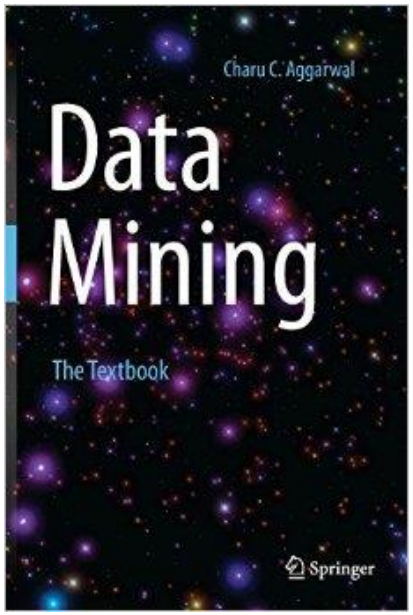
Data uncertainty

Managing the reliability and predictability of inherently imprecise data types

IBM

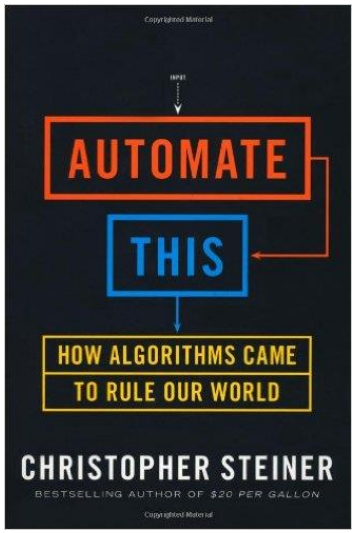
Text Books

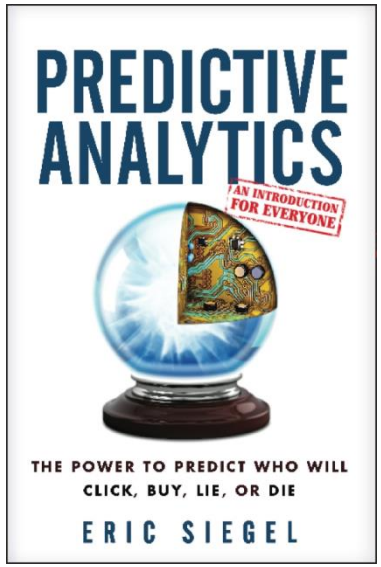


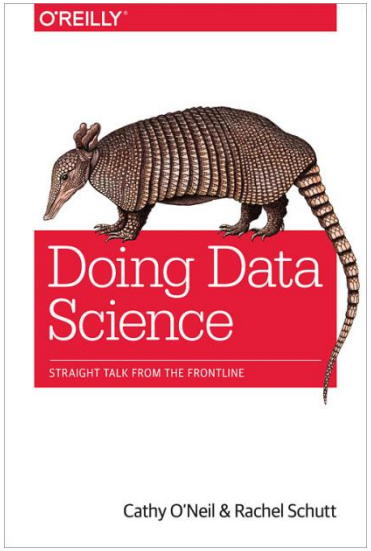


Required

Optional







Tentative Schedule

Lecture	Topic	Projects
1	Intro, schedule, and logistics	
2	Data Science components and tasks	
3	Data types	Project #1 out
4	Introduction to R, statistics foundations	
5	Introduction to D3, visual analytics	
6	Data preparation and reduction	
7	Data preparation and reduction	Project #1 due
8	Similarity and distances	Project #2 out
9	Similarity and distances	
10	Cluster analysis	
11	Cluster analysis	
12	Pattern miming	Project #2 due
13	Pattern mining	
14	Outlier analysis	
15	Outlier analysis	Final Project proposal due
16	Classifiers	
17	Midterm	
18	Classifiers	
19	Optimization and model fitting	
20	Optimization and model fitting	
21	Causal modeling	
22	Streaming data	Final Project preliminary report due
23	Text data	
24	Time series data	
25	Graph data	
26	Scalability and data engineering	
27	Data journalism	
	Final project presentation	Final Project slides and final report due

Grading

Projects (2): 15% each

Midterm: 30%

Final Project: 40%

- proposal: 10%
- prelim report: 10%
- final report: 10%
- presentation: 10%

Participation

- not graded, but I hope you will attend regularly and participate actively

For late submission policy see website