

Introduction to Data Science: Tools and Techniques Week 1

Dr Muhammad Atif Tahir

Including notes from Michael Franklin

Dan Bruckner, Evan Sparks,

Shivaram Venkataraman, John Canny, Alexander Lex,

Braxton Osting

Outline

- Data Science – Why all the excitement?
 - examples
- Where does data come from
- So what is Data Science
- Doing Data Science
- About the course
 - what we'll cover
 - requirements, workload etc.

Data Analysis Has Been Around for a While

1935: "The Design of Experiments"

R.A. Fisher



W.E.
Demming

1939: "Quality Control"



1958: "A Business Intelligence System"

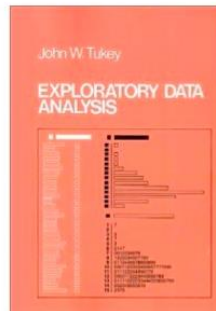


Peter Luhn

1997: "Machine Learning"



1977: "Exploratory Data Analysis"



1989: "Business Intelligence"

Howard
Dresner

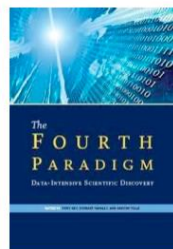


2010: "The Data Deluge"

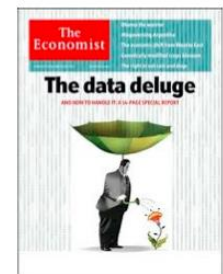
1996: Google



2007: "The Fourth Paradigm"

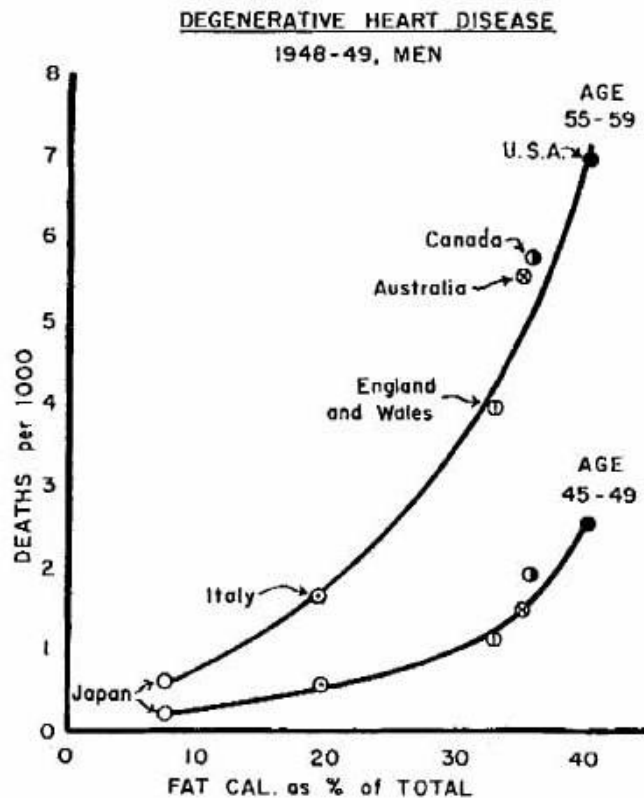


2009: "The Unreasonable Effectiveness of Data"



Data makes everything clearer

- Seven Countries Study (Ancel Keys, UCB 1925,28)
- 13,000 subjects total, 5-40 years follow-up.



Data Science: Why all the Excitement?



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.

Data Makes Everything Clearer?

Epidemiological modeling of online social network dynamics

John Cannarella¹, Joshua A. Spechler^{1,*}

¹ Department of Mechanical and Aerospace Engineering, Princeton University, Princeton, NJ, USA

* E-mail: Corresponding spechler@princeton.edu

Abstract

The last decade has seen the rise of immense online social networks (OSNs) such as MySpace and Facebook. In this paper we use epidemiological models to explain user adoption and abandonment of OSNs, where adoption is analogous to infection and abandonment is analogous to recovery. We modify the traditional SIR model of disease spread by incorporating infectious recovery dynamics such that contact between a recovered and infected member of the population is required for recovery. The proposed infectious recovery SIR model (irSIR model) is validated using publicly available Google search query data for “MySpace” as a case study of an OSN that has exhibited both adoption and abandonment phases. The irSIR model is then applied to search query data for “Facebook,” which is just beginning to show the onset of an abandonment phase. Extrapolating the best fit model into the future predicts a rapid decline in Facebook activity in the next few years.

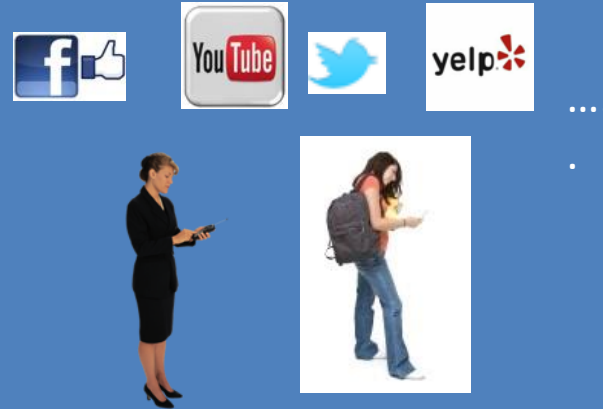
“Big Data” Sources

It's All Happening On-line



Every:
Click
Ad impression
Billing event
Fast Forward, pause,...
Server request
Transaction
Network message
Fault
...

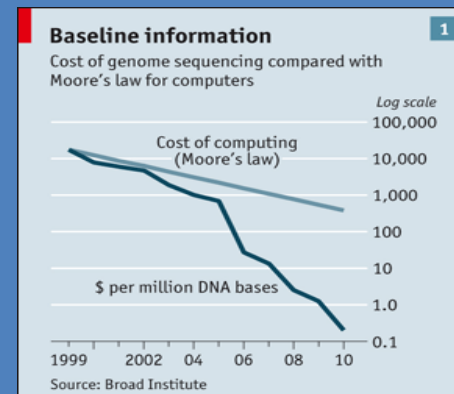
User Generated (Web & Mobile)



Internet of Things / M2M



Health/Scientific Computing

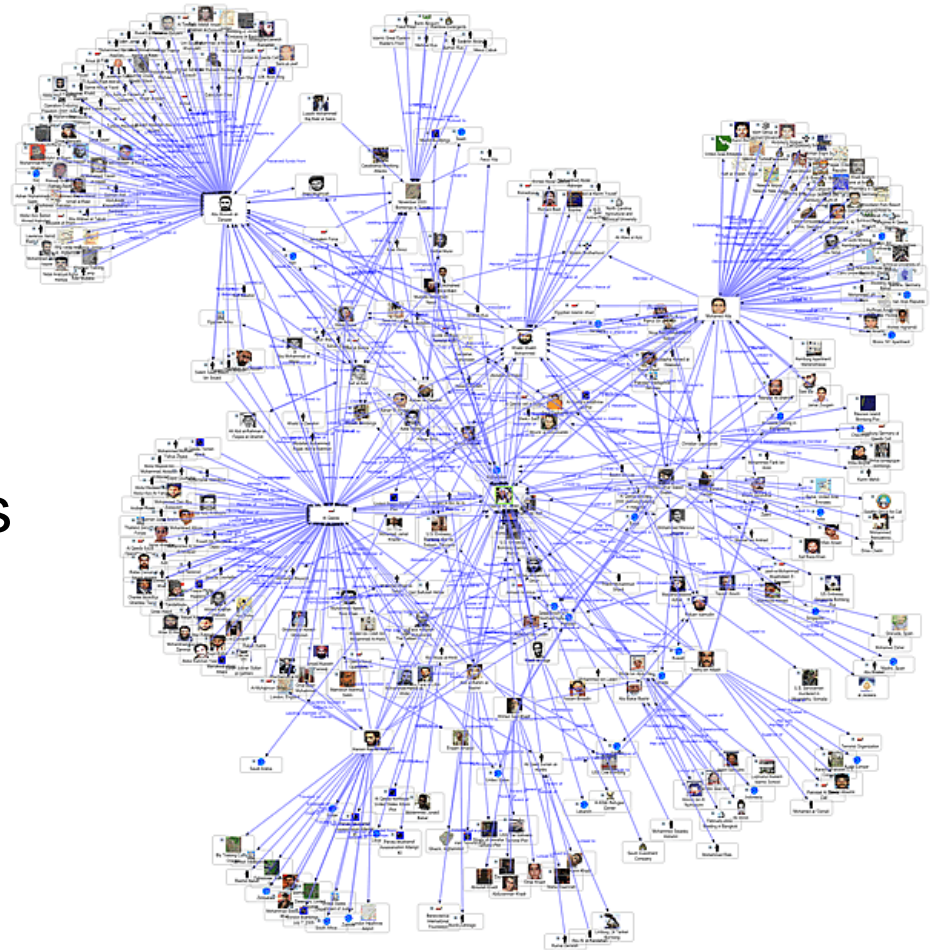


Graph Data

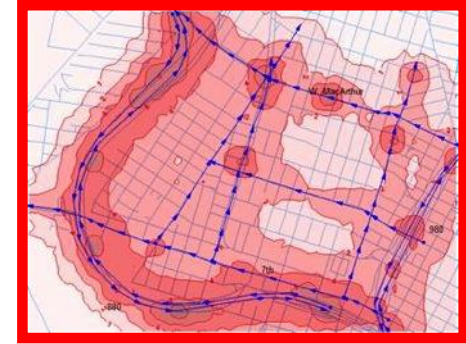
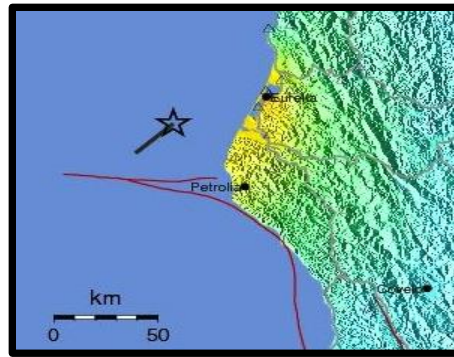
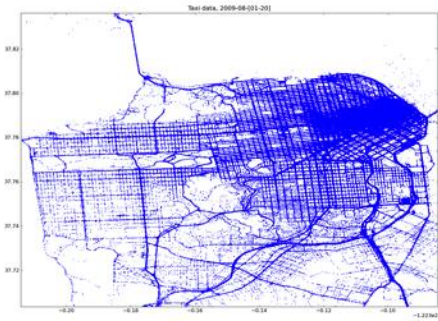
Lots of interesting data has a graph structure:

- Social networks
- Communication networks
- Computer Networks
- Road networks
- Citations
- Collaborations/Relationships
- ...

Some of these graphs can get quite large (e.g., Facebook* user graph)



What can you do with the data?



Crowdsourcing + physical modeling + sensing + data assimilation

to produce:



From Alex Bayen, UCB

DATA SCIENCE – WHAT IS IT?

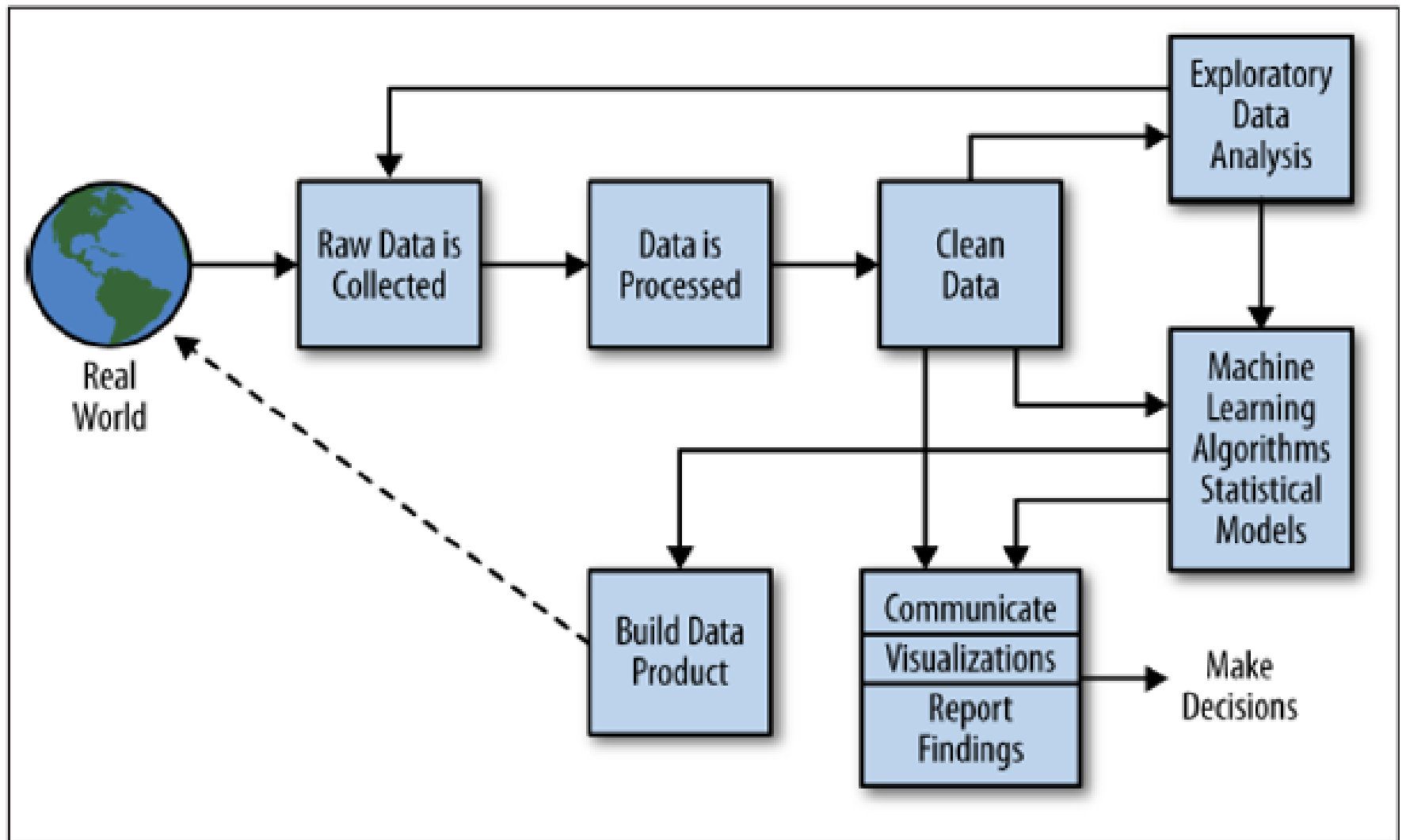
“Data Science” an Emerging Field















O'Reilly Radar report

What is Data Science?

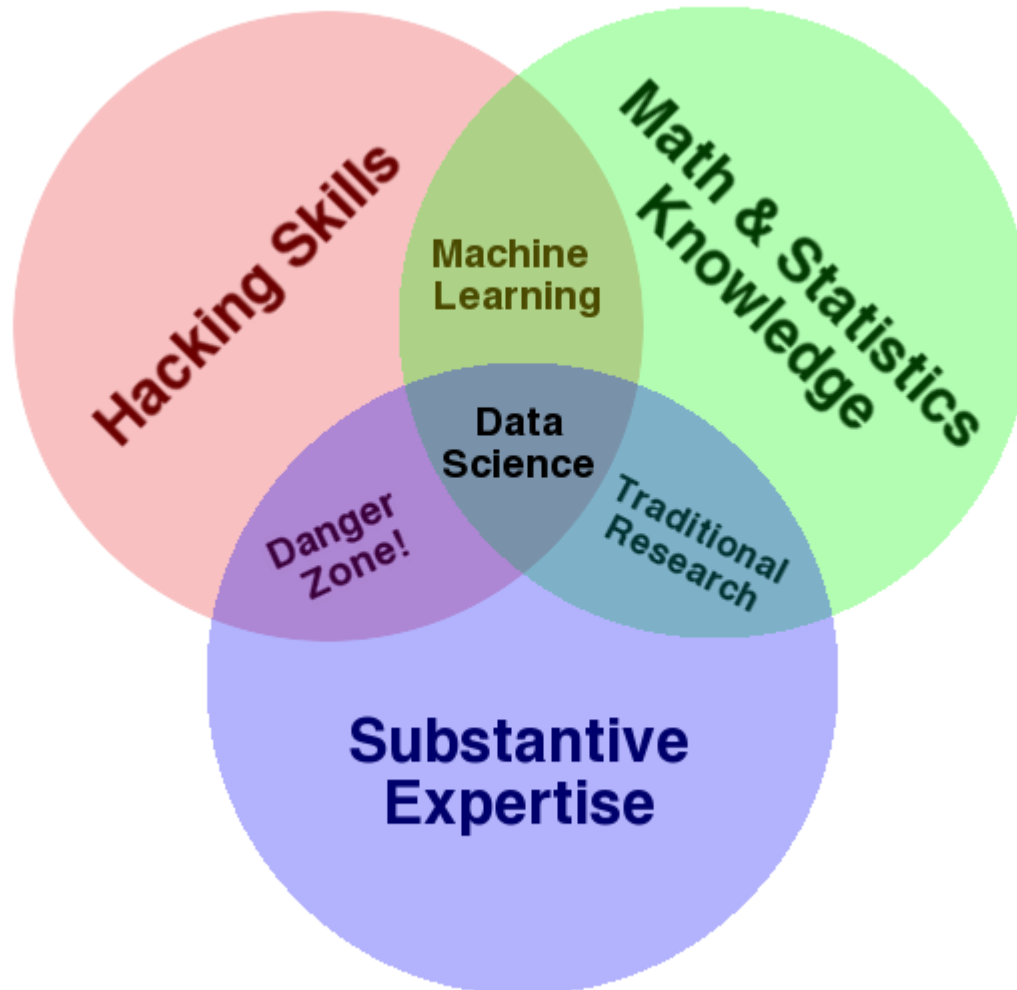
- Data science is an interdisciplinary field about processes and systems to extract knowledge or insights from data in various forms. (Wikipedia)
- Data Science closes the circle from collecting real-world data, to processing and analyzing it, to influence the real world again



Some recent ML Competitions

Active Competitions			
		Flight Quest 2: Flight Optimization Final Phase of Flight Quest 2	33 days Coming soon \$220,000
		Packing Santa's Sleigh He's making a list, checking it twice; to fill up his sleigh, he needs your advice	5.8 days 338 teams \$10,000
		Flu Forecasting  Predict when, where and how strong the flu will be	41 days 37 teams
		Galaxy Zoo - The Galaxy Challenge Classify the morphologies of distant galaxies in our Universe	2 months 160 teams \$16,000
		Loan Default Prediction - Imperial College Lon... Constructing an optimal portfolio of loans	52 days 82 teams \$10,000
		Dogs vs. Cats Create an algorithm to distinguish dogs from cats	11 days 166 teams Swag

Data Science – One Definition



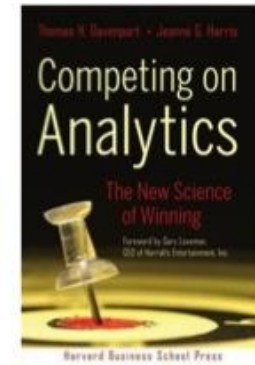
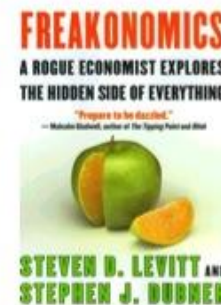
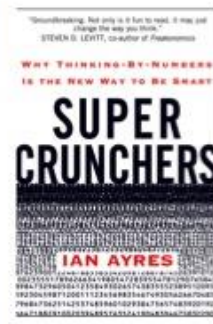
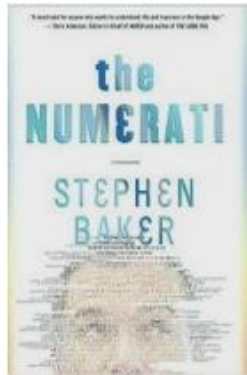
Contrast: Databases

	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, ACID*	CAP* theorem (2/3), eventual consistency
Realizations	SQL	NoSQL: Riak, Memcached, Apache River, MongoDB, CouchDB, Hbase, Cassandra,...

ACID = Atomicity, Consistency, Isolation and Durability CAP = Consistency, Availability, Partition Tolerance

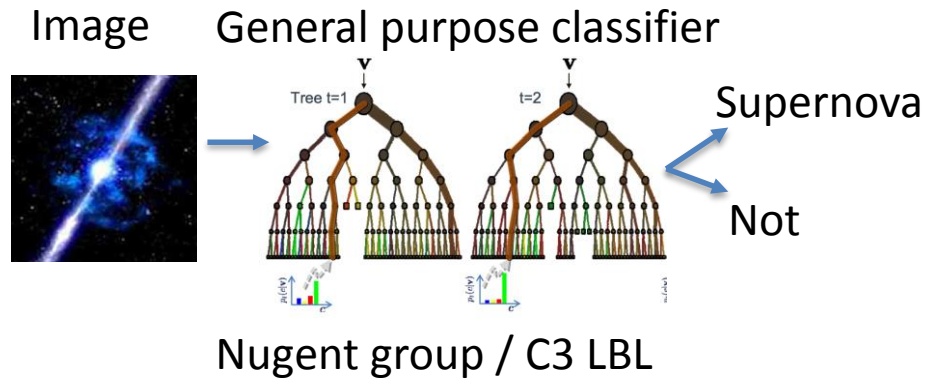
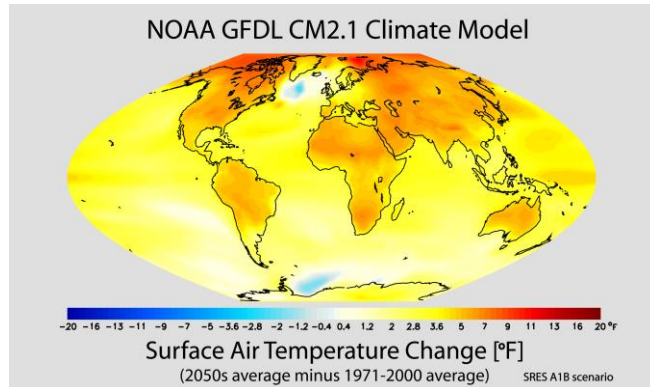
Contrast: Databases

Databases	Data Science
Querying the past	Querying the future



Business intelligence (BI) is the transformation of raw data into meaningful and useful information for [business analysis](#) purposes. BI can handle enormous amounts of unstructured data to help identify, develop and otherwise create new strategic business opportunities - Wikipedia

Contrast: Scientific Computing



Scientific Modeling

Physics-based models

Problem-Structured

Mostly deterministic, precise

Run on Supercomputer or
High-end Computing Cluster

Data-Driven Approach

General inference engine replaces model

Structure not related to problem

Statistical models handle true randomness,
and **unmodeled complexity**.

Run on cheaper computer Clusters (EC2)

Contrast: Machine Learning

Machine Learning	Data Science
Develop new (individual) models	Explore many models, build and tune hybrids
Prove mathematical properties of models	
Improve/validate on a few, relatively clean, small datasets	Understand empirical properties of models
Publish a paper	Develop/use tools that can handle massive datasets
	Take action!

Analyzing the Analysts

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

US faces shortage of 140,000 to 190,000 people “with deep analytical skills, as well as 1.5 million managers and analysts with the know-how to use the analysis of big data to make effective decisions.”

10/18/12 Bill Howe, UW 44
--Mckinsey Global Institute

Demand for workers with specialist data skills like data scientists and data engineers has more than tripled over five years (+231%), according to a labour market analysis commissioned for Dynamics of data science skills, a new Royal Society report published today. Demand for all types of workers grew by 36% over the same period

DOING DATA SCIENCE

Ben Fry's Model

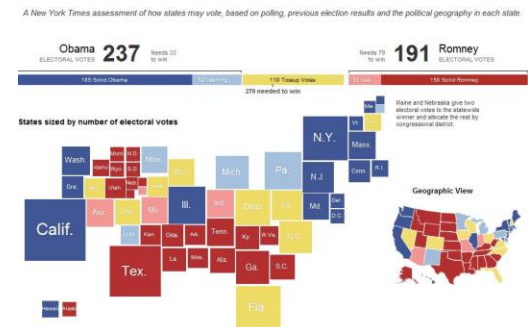
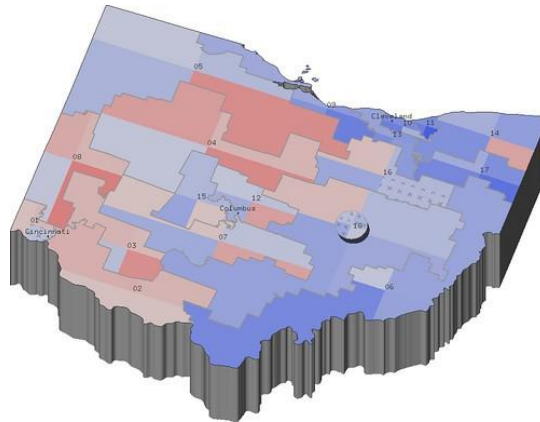
1. Acquire
2. Parse
3. Filter
4. Mine
5. Represent
6. Refine
7. Interact

Jeff Hammerbacher's Model

1. Identify problem
2. Instrument data sources
3. Collect data
4. Prepare data (integrate, transform, clean, filter, aggregate)



5. Build model
6. Evaluate model
7. Communicate results



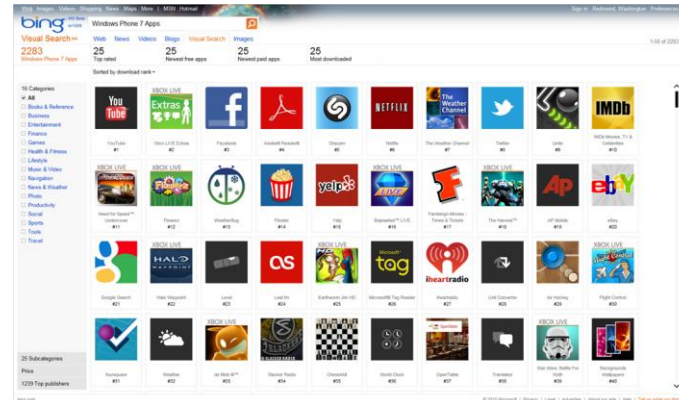
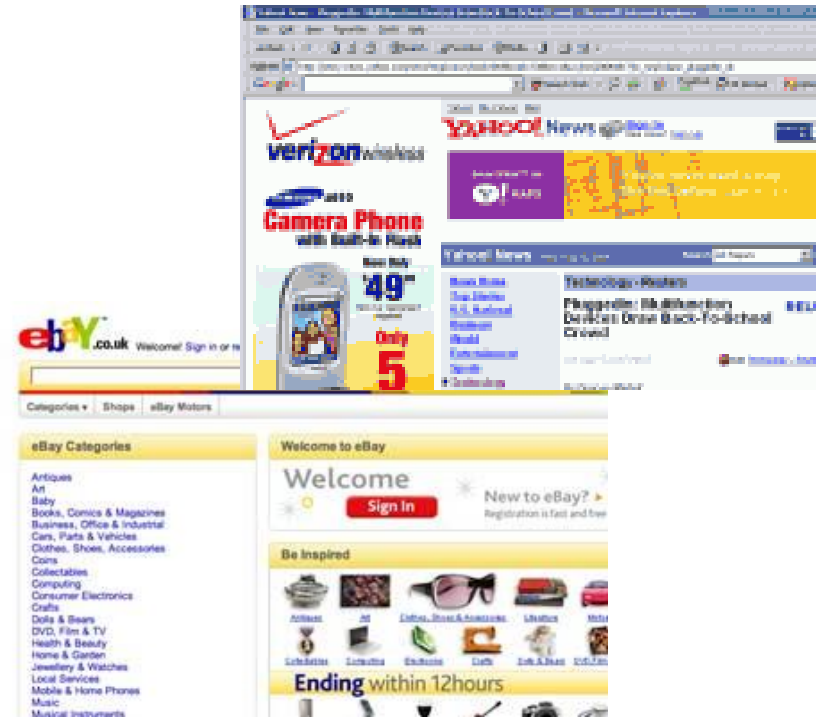
From the Trenches

Yahoo [KDD 2009, best app. paper]

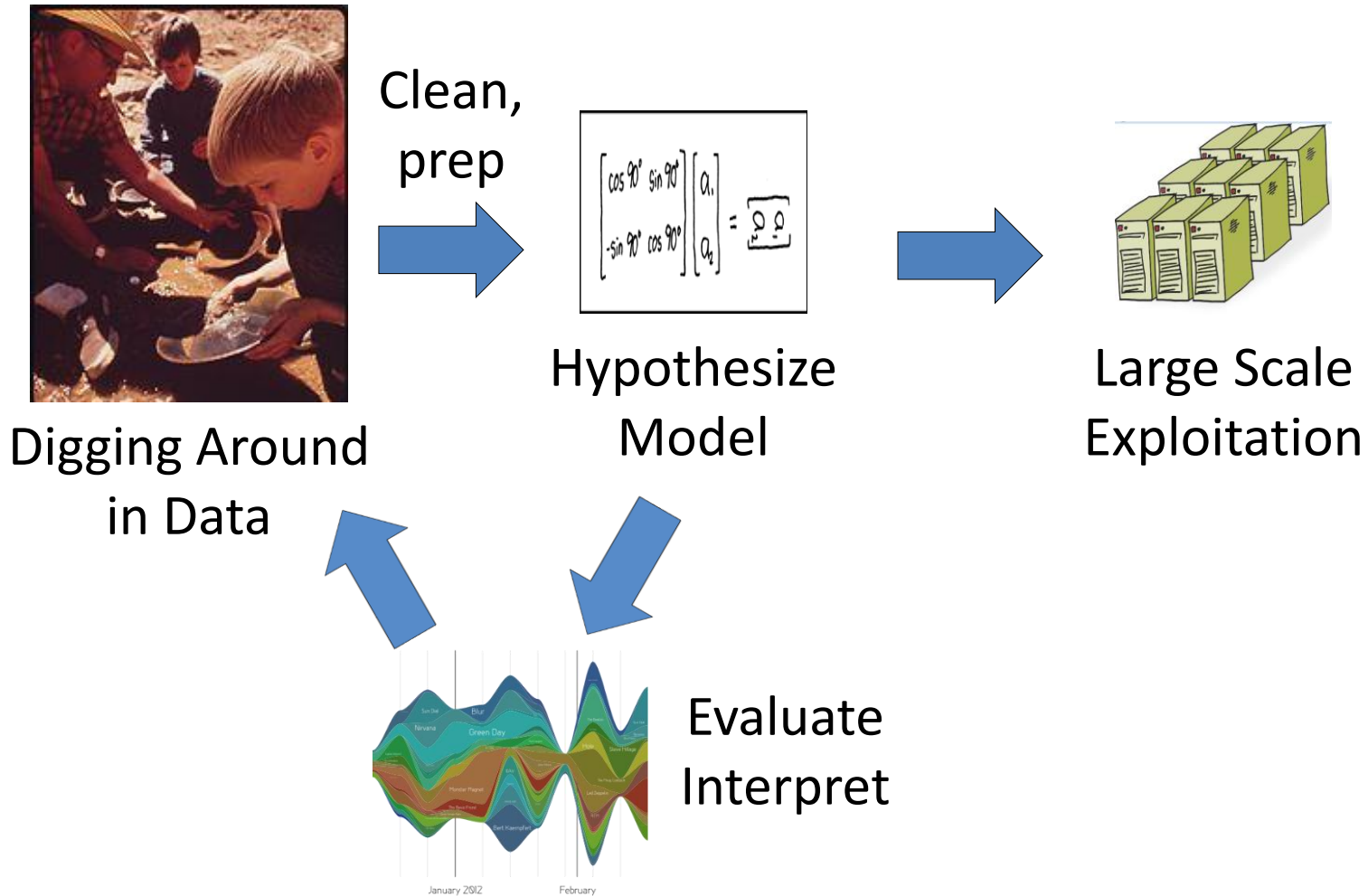
Ebay [SIGIR 2011, hon. mention]

Quantcast [2012]

Microsoft [CIKM 2014]



Data Scientist's Practice



What's Hard about Data Science

- Overcoming assumptions
- Making ad-hoc explanations of data patterns
- Overgeneralizing
- Communication
- Not checking enough (validate models, data pipeline integrity, etc.)
- Using statistical tests correctly
- Prototype → Production transitions
- Data pipeline complexity (who do you ask?)

About the Course

Grading

- Homeworks and in-class labs: 15%
- Midterm 1: 12.5%, Midterm 2: 12.5%
- Project (in groups): 10%
- Final Exam: 50%

Projects

Project teams should form by week 3.

Project proposals will be due by week 6.

You need:

- A clear problem statement
- An accessible dataset
- Modeling plan + appropriate tools

About the Course

Staff Contact:

Instructor: Dr Muhammad Atif Tahir

atif.tahir@nu.edu.pk