Introduction to Big Data with Apache Spark

BerkeleyX

This Lecture

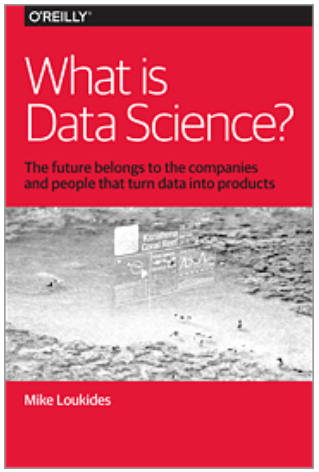
So What is Data Science?

Doing Data Science

Data Preparation

Roles

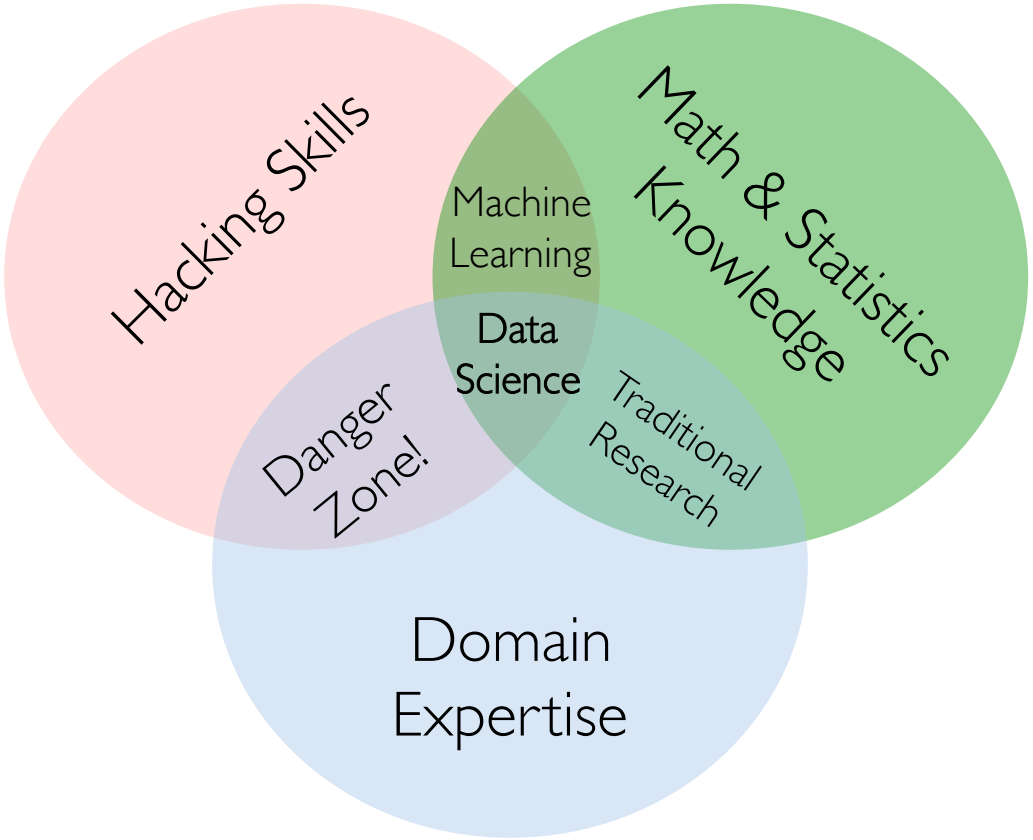
What is Data Science?



- *Data Science* aims to derive knowledge from big data, efficiently and intelligently
- *Data Science* encompasses the set of activities, tools, and methods that enable data-driven activities in science, business, medicine, and government

http://www.oreilly.com/data/free/what-is-data-science.csp

Data Science – One Definition



http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

Contrast: Databases

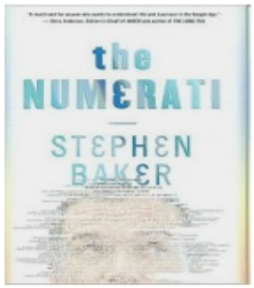
Element	Databases	Data Science
Data Value	"Precious"	"Cheap"
Data Volume	Modest	Massive
Examples	Bank records, Personnel records, Census, Medical records	Online clicks, GPS logs, Tweets, tree sensor readings
Priorities	Consistency, Error recovery, Auditability	Speed, Availability, Query richness
Structured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

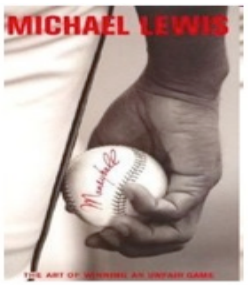
*CAP = Consistency, Availability, Partition Tolerance

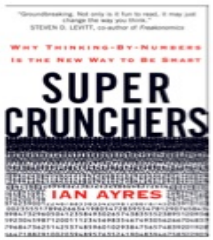
+ACID = Atomicity, Consistency, Isolation and Durability

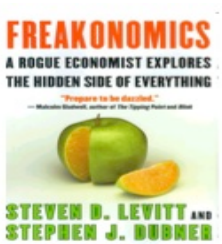
Contrast: Databases

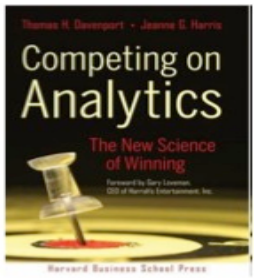
Databases	Data Science
Querying the past	Querying the future





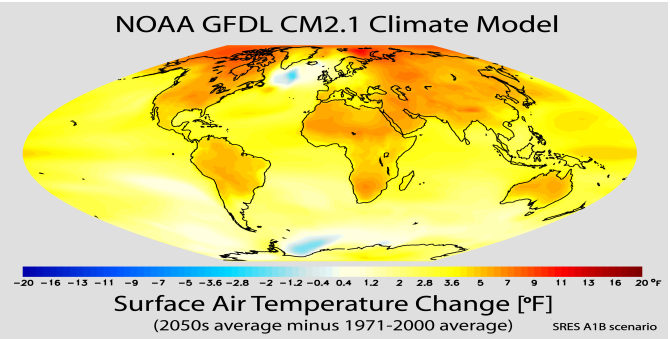






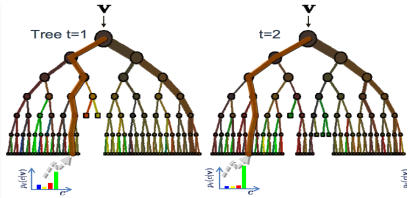
- Related – Business Analytics
  - » Goal: obtain "actionable insight" in complex environments
  - » Challenge: vast amounts of disparate, unstructured data and limited time

Contrast: Scientific Computing



Image

General purpose ML classifier



Supernova

Not

Dr Peter Nugent (C3 LBNL)

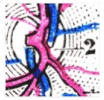
Scientific Modeling	Data-Driven Approach
Physics-based models	General inference engine replaces model
Problem-Structured	Structure not related to problem
Mostly deterministic, precise	Statistical models handle true randomness, and unmodeled complexity
Run on Supercomputer or High-end Computing Cluster	Run on cheaper computer Clusters (EC2)

Contrast: Traditional Machine Learning

Traditional 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!

Recent Data Science Competitions

Using Data Science to find Data Scientists!

Competition Name	Reward	Teams
 **Diabetic Retinopathy Detection** Identify signs of diabetic retinopathy in eye images	$100,000	283
 **West Nile Virus Prediction** Predict West Nile virus in mosquitos across the city of Chicago	$40,000	264
 **Restaurant Revenue Prediction** Predict annual restaurant sales based on objective measurements	$30,000	2340
 **Otto Group Product Classification Challenge** Classify products into the correct category	$10,000	2950
 **How Much Did It Rain?** Predict probabilistic distribution of hourly rain given polarimetric radar measurements	$500	282
 **ECML/PKDD 15: Taxi Trajectory Prediction (I)** Predict the destination of taxi trips based on initial partial trajectories	$250	72
 **ECML/PKDD 15: Taxi Trip Time Prediction (II)** Predict the total travel time of taxi trips based on their initial partial trajectories	$250	35

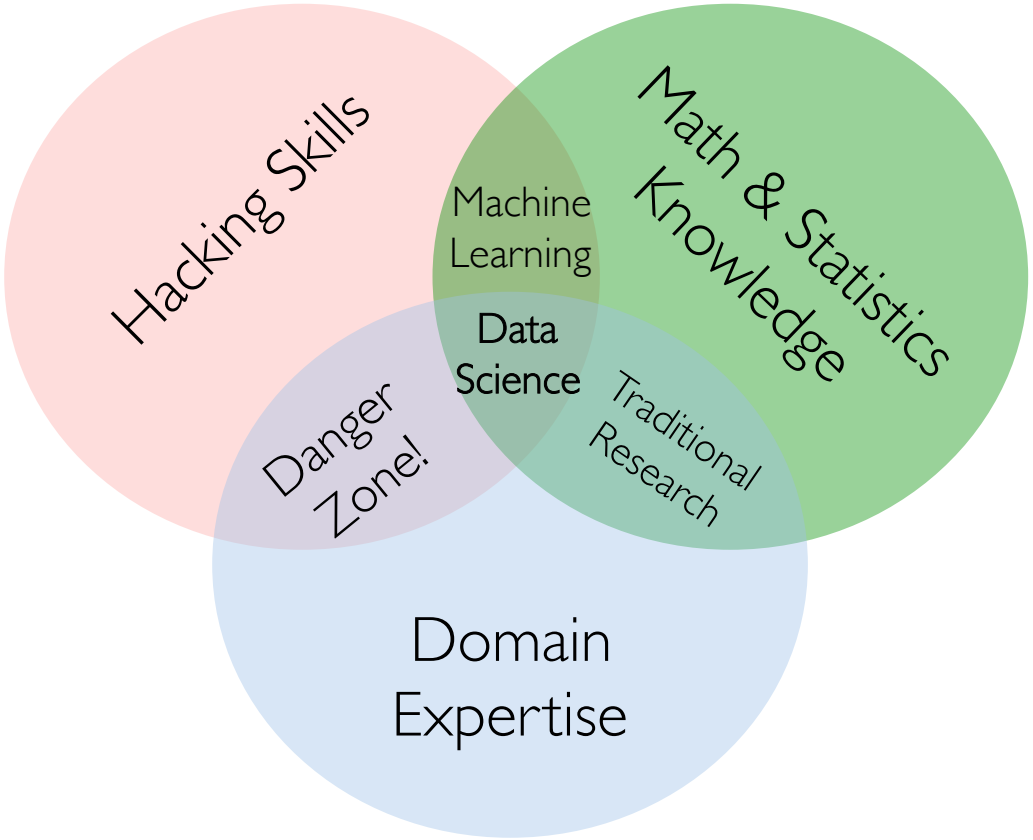
kaggle

Doing Data Science

- The views of three Data Science experts
  - » Jim Gray (Turing Award winning database researcher)
  - » Ben Fry (Data visualization expert)
  - » Jeff Hammerbacher (Former Facebook Chief Scientist, Cloudera co-founder)

- Cloud computing: Data Science enabler

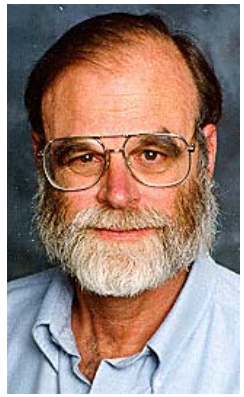
Data Science – One Definition



http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

Jim Gray's Model

1. Capture
2. Curate
3. Communicate



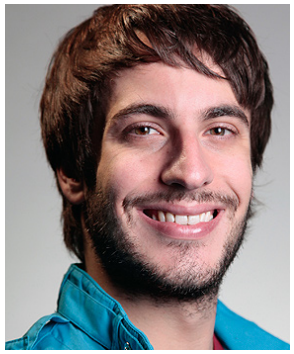
Turing award winner

Ben Fry's Model

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

Data visualization expert

Jeff Hammerbacher's Model



Facebook, Cloudera

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

Key Data Science Enabler: Cloud Computing

- Cloud computing reduces computing operating costs
- Cloud computing enables data science on massive numbers of inexpensive computers

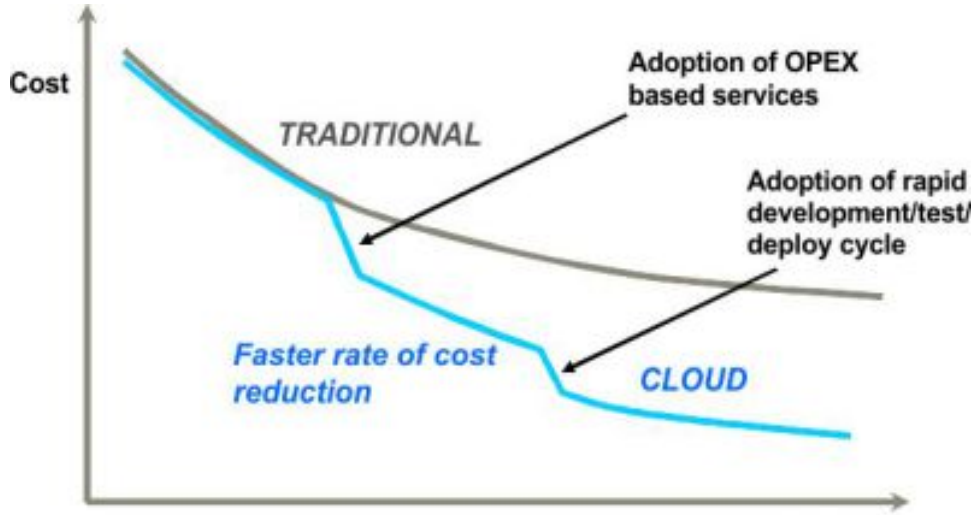
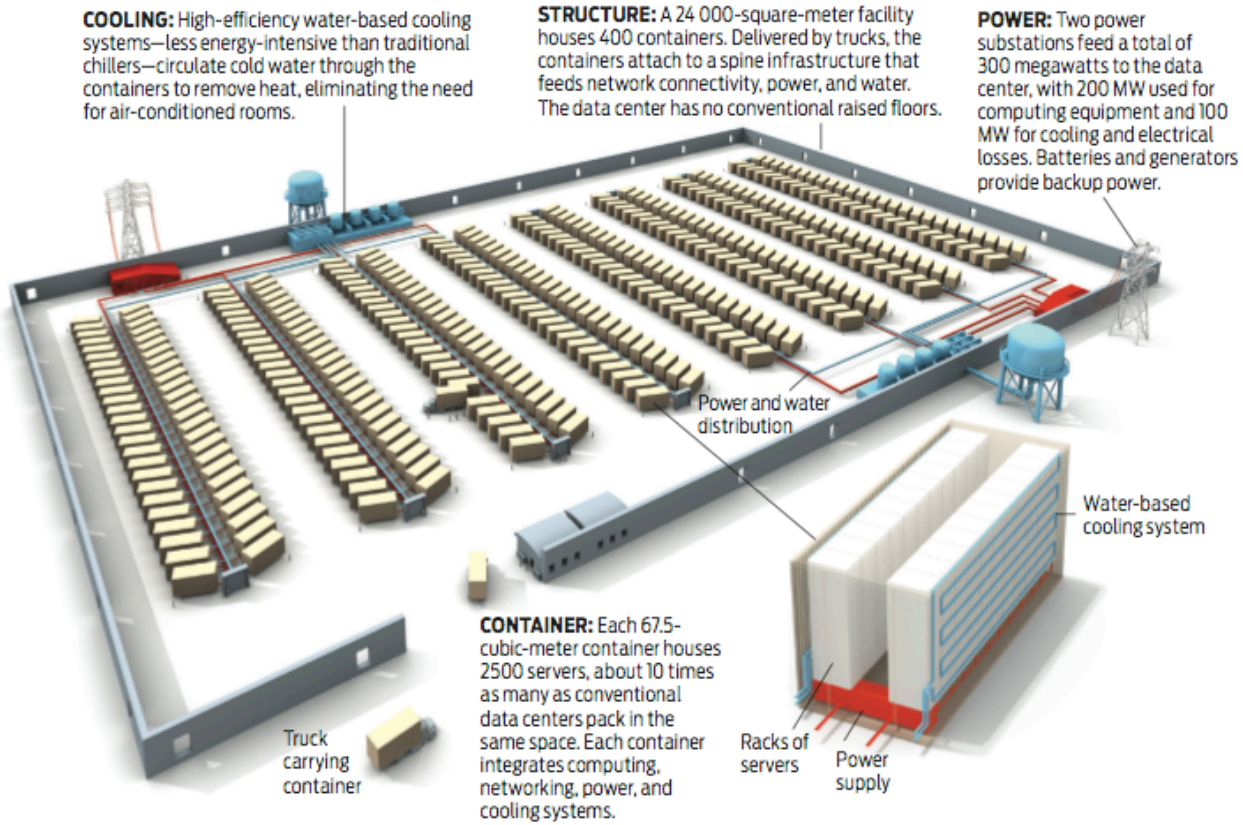


Figure: http://www.opengroup.org/cloud/cloud/cloud_for_business/roi.htm

The Million-Server Data Center

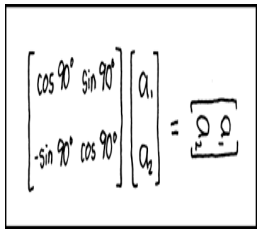


http://spectrum.ieee.org/tech-talk/semiconductors/devices/what-will-the-data-center-of-the-future-look-like

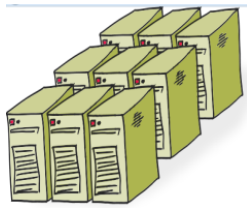
Data Scientist's Practice

Digging Around in Data

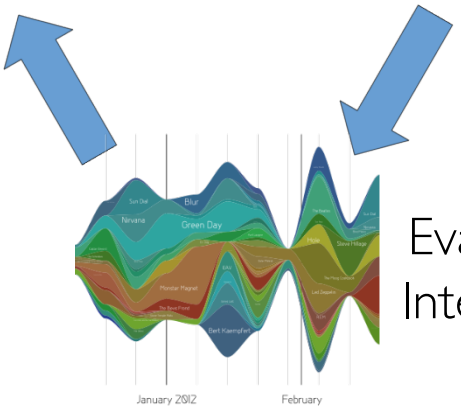
Clean, prep



Hypothesize Model



Large Scale Exploitation



Evaluate Interpret

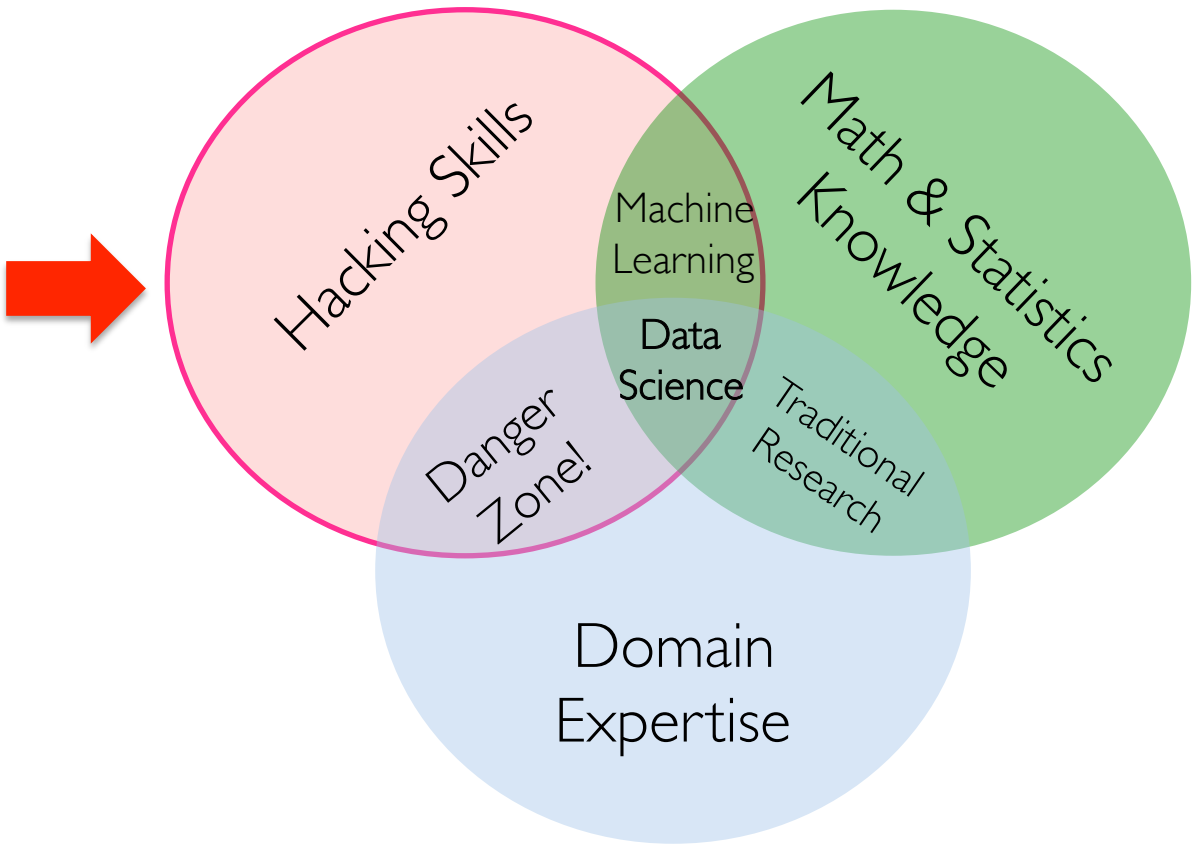
Data Science Topics

- Data Acquisition
- Data Preparation
- Analysis
- Data Presentation
- Data Products
- Observation and Experimentation

What's Hard about Data Science?

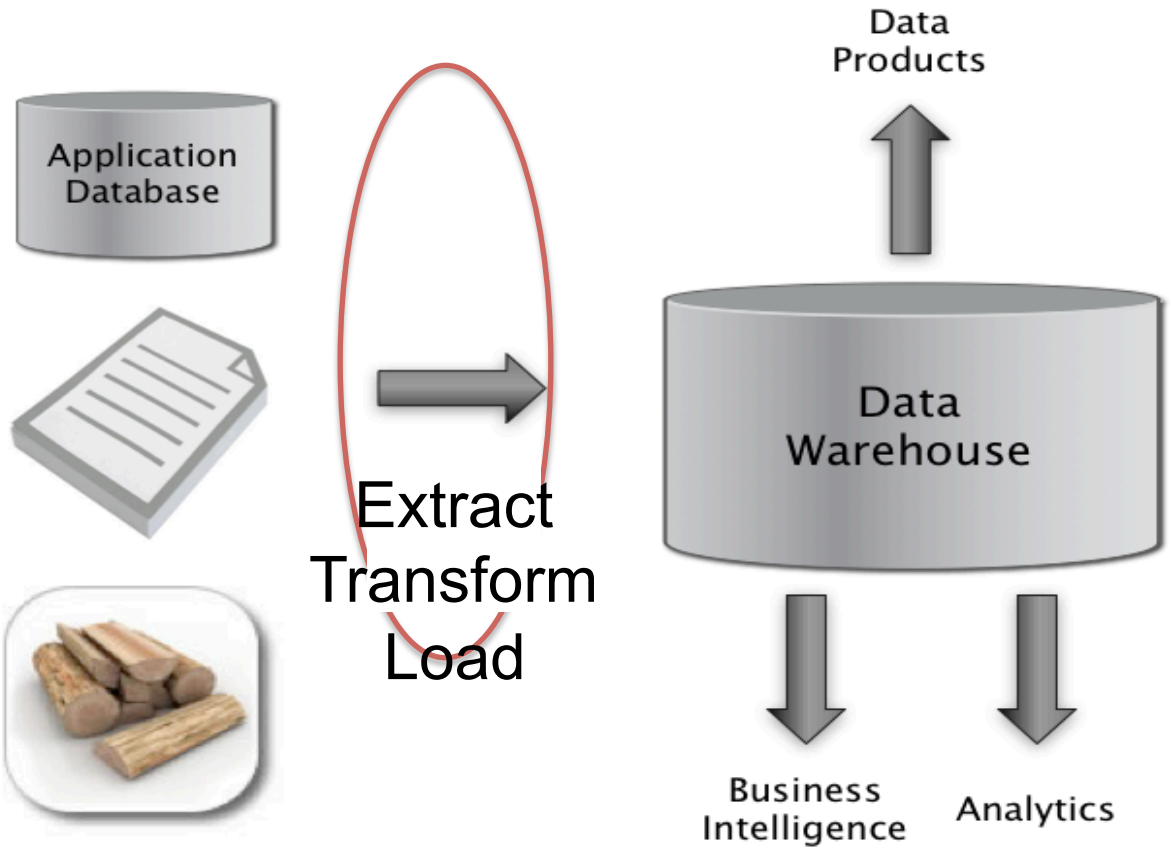
- Overcoming assumptions
- Making ad-hoc explanations of data patterns
- Not checking enough (validate models, data pipeline integrity, etc.)
- Overgeneralizing
- Communication
- Using statistical tests correctly
- Prototype → Production transitions
- Data pipeline complexity (who do you ask?)

Data Science – One Definition



http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

The Big Picture



Data Acquisition (Sources) in Web Companies

- Examples from Facebook
  - » Application databases
  - » Web server logs
  - » Event logs
  - » Application Programming Interface (API) server logs
  - » Ad and search server logs
  - » Advertisement landing page content
  - » Wikipedia
  - » Images and video

Data Acquisition & Preparation Overview

- Extract, Transform, Load (ETL)
  - » We need to *extract* data from the source(s)
  - » We need to *load* data into the sink
  - » We need to *transform* data at the source, sink, or in a staging area

  - » Sources: file, database, event log, web site, Hadoop Distributed FileSystem (HDFS), …
  - » Sinks: Python, R, SQLite, NoSQL store, files, HDFS, Relational DataBase Management System (RDBMS), …

Data Acquisition & Preparation Process Model

- The construction of a new data preparation process is done in many phases
  - » Data characterization
  - » Data cleaning
  - » Data integration

- We must efficiently move data around in space and time
  - » Data transfer
  - » Data serialization and deserialization (for files or network)

Data Acquisition & Preparation Workflow

- The transformation pipeline or workflow often consists of many steps
  - » For example: Unix pipes and filters
  - » `cat data_science.txt | wc | mail -s "word count" myname@some.com`

- If a workflow is to be used more than once, it can be scheduled
  - » Scheduling can be time-based or event-based
  - » Use publish-subscribe to register interest (e.g., Twitter feeds)

- Recording the execution of a workflow is known as capturing lineage or provenance
  - » Spark's Resilient Distributed Datasets do this for you automatically

Impediments to Collaboration

- The diversity of tools and programming/scripting languages makes it hard to share
- Finding a script or computed result is often harder than just writing the program from scratch!
  - » Question: How could we fix this?
- View that most analysis work is "throw away"

Data Science Roles

- Businessperson
- Programmer
- Enterprise
- Web Company

The Businessperson

- Data Sources
  - » Web pages
  - » Excel
- ETL
  - » Copy and paste
- Data Warehouse
  - » Excel
- Business Intelligence and Analytics
  - » Excel functions
  - » Excel charts
  - » Visual Basic

Image: http://www.fastcharacters.com/character-design/cartoon-business-man/

The Programmer

- Data Sources
  - » Web scraping, web services API
  - » Excel spreadsheet exported as Comma Separated Values
  - » Database queries
- ETL
  - » wget, curl, Beautiful Soup, lxml
- Data Warehouse
  - » Flat files
- Business Intelligence and Analytics
  - » Numpy, Matplotlib, R, Matlab, Octave

Image: http://doctormo.deviantart.com/art/Computer-Programmer-Ink-346207753

The Enterprise

- Data Sources
  - » Application databases
  - » Intranet files
  - » Application server log files
- ETL
  - » Informatica, IBM DataStage, Ab Initio, Talend
- Data Warehouse
  - » Teradata, Oracle, IBM DB2, Microsoft SQL Server
- Business Intelligence and Analytics
  - » SAP Business Objects, IBM Cognos, Microstrategy, SAS, SPSS, R

Image: http://www.publicdomainpictures.net/view-image.php?image=74743

The Web Company

- Data Sources
  - » Application databases
  - » Logs from the services tier
  - » Web crawl data
- ETL
  - » Apache Flume, Apache Sqoop, Apache Pig, Apache Oozie, Apache Crunch
- Data Warehouse
  - » Apache Hadoop/Apache Hive, Apache Spark/Spark SQL
- Business Intelligence and Analytics
  - » Custom dashboards: Oracle Argus, Razorflow
  - » R

Image: http://www.future-web-net.com/2011/04/un-incendio-blocca-il-server-di-aruba.html