Introduction to Big Data

CS454: Big Data Analysis and Visualization

John J. Tran
California State University Los Angeles

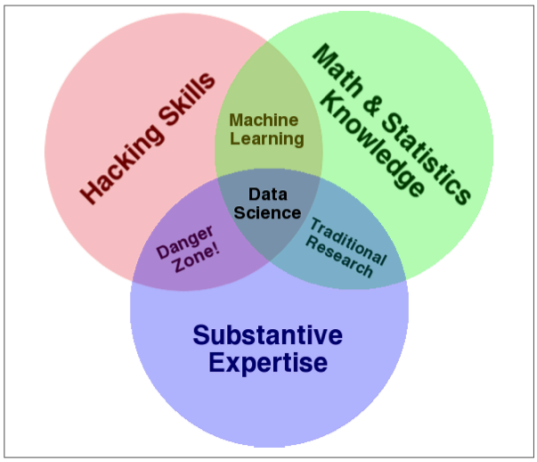
Course Roadmap

- Introduction
- Data generation and collection
- Building infrastructure to do data science
  - Map reduce, statistical analysis, machine learning, and visualization
- Mashing data and information visualization
  - Writing algorithms to coalesce data collected into visual displays
- Automating the process: end-to-end product

What is Big Data?

- Big data is data that
  - Exceeds processing capacity of conventional database
  - Exceeds storage capacity of conventional database
  - Does not follow strict structure
- Big data "hides" data (meaning)
  - We must extract (derive) meaning
    - Manipulation: process, aka slice and dice
    - Visualize: look at data differently

Drew Conway's Venn diagram of Data Science

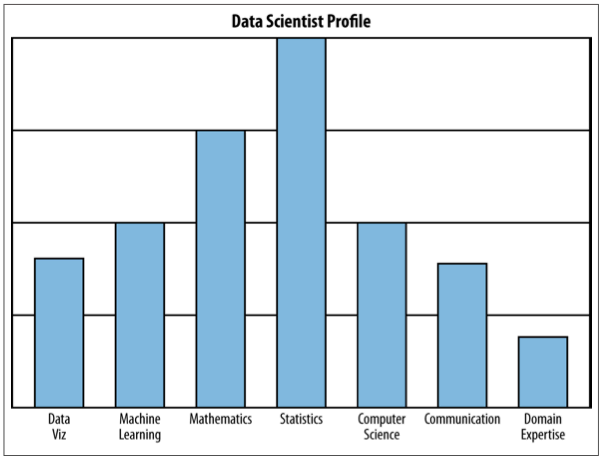


Source: Doing Data Science

Disciplines with Data Science

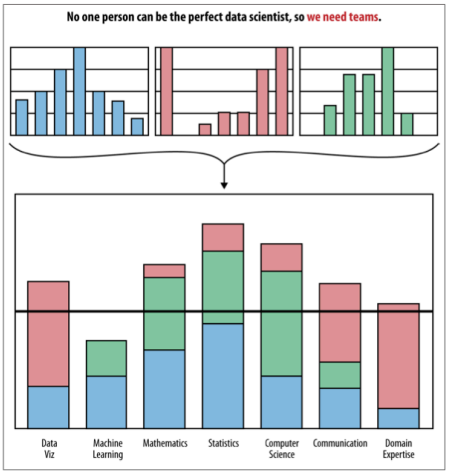
- According to Nathan Yau:
  - Statistics
    - Numbers, stats, theoretical
  - Data "munging"
    - Parsing, scraping and formatting
  - Visualization
    - Graphs and plots

Data Science Profile



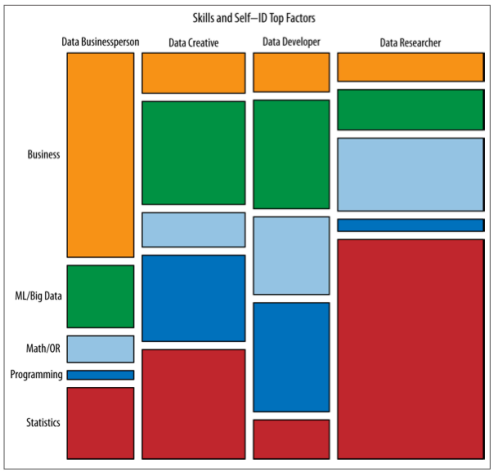
Source: Doing Data Science

Data science is ultimately Teamwork



Source: Doing Data Science

Clustering and Visualization



Source: Doing Data Science

Data Products

- Loukides suggests
  - Use of data and metadata are becoming overt
    - Ubiquitous (iTunes, Facebook)
    - In your face
  - Recent trend: look at data in the background
    - Results-centric
    - Impact society
      - Efficient routing, road condition > energy saving
      - Self-driving car (do we need data?)

Data Deluge

- According Loukides, challenges with growth data dependency
  - Humans cannot "manually" handle large amount of data
  - Re-examine needs:
    - Doctors don't need data, they need to heal patients with good data
    - Hotel companies use data to efficiently manage occupancy, discounts, and promotions
- Trending: ambient data
  - How capitalize on data noise (explicit or implicit)

Combining Data

- Evolve from “single database” to joining of multiple databases to extract key meaning
- Thought experiment: How does facebook know how to identify pictures (this is a very difficult problem):
  - It knows your friends and your network
  - Search space is a whole lot smaller

Key point: combining picture database with social graph

Data Discovery

- 3rd order knowledge (Holy Grail of big data research)
  - Predictive aspect of big data analysis
    - Recommendations
- Examples
  - Google suggestion
  - Apple Genius
  - Facebook friend recommendation

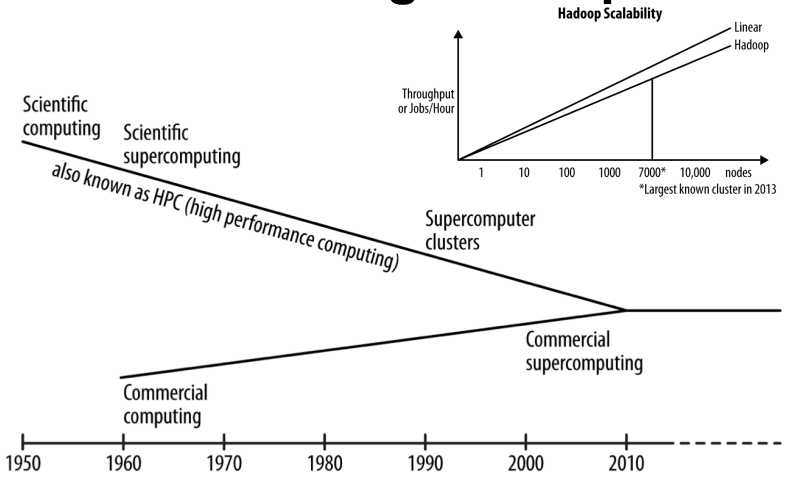
(User) Interfaces

- User interface is extremely important
  - Apple recognize this
- Data is product must include designers
- Case studies:
  - GPS
  - Apple iTunes
- Ultimately: data to users **in a meaningful way**

Evolution of Big Data

- Evolve from early days of High Performance Computing
  - HPC serves scientific computing community (e.g. weather prediction)
  - Big Data (or data science) serves business community (e.g. behavior prediction)
- The Yahoo phenomena
  - Need to index large data (for advertisement purpose)

Evolution of "large" computation

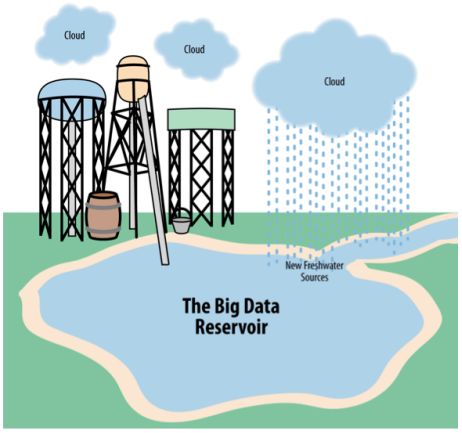


Source: Disruptive Possibilities

Birth of the Cloud

- Open source technologies:
  - Hadoop
  - Linux
  - OpenStack
- Economy of scale
  - Joe "the plumber" can do cloud computing
- Purchase resource on demand
  - Amazon's elastic cloud
  - Microsoft's Azure cloud technology

Clouds & Big Data



Cloud provides computation and storage

Big Data is accessible to/from all cloud platforms

Source: Disruptive Possibilities

Making Sense of Data

- Once infrastructure has been built, what do we do next?
- In short Data Science is the following cycle
  - Collect data from sensors
  - Analyze data collected
    - Derive 1st order meaning
    - Categorize, binning, classification
  - Make sense of data collection
    - Derive 2nd order meaning
    - What is the data telling us?
  - Anticipate trends
    - Derive 3rd order meaning
    - What is the future?

First Order Meaning

- Classification of data collected
  - Requires some domain expertise
    - Temperature and weather data requires climate scientists
    - GPS data requires cartographer
    - Images data requires imagery experts
- Impose some kind of organization on the data
  - Put data into collection
    - Images can be classified as “indoor” vs. “outdoor”
    - Temperature data vs. precipitation data
  - Slice and dice along various dimensions
    - Temporal, spatial, etc.
- Two immediate payoffs:
  - Physical storage of data – organizational value
  - Hierarchical cognitive organization

Second Order Meaning

- Comprehending temporal and spatial location of the data collected
- Asking key pertinent questions about the data:
  - What is it?
  - Where is it?
  - When is it?

  } First order meaning

  - What is the person doing? → Second order meaning
- For example: with a collection of pictures with GPS data we can ask where is the picture taken?

Third Order Meaning

- Much more difficult to derive
- Predictive and anticipatory in nature
- Asking trending questions
- For example, given a set of data
  - What is the tendency of the actors?
  - What is the intention of the image?
  - Where do teenage girls shop when it's raining?

Data Relationship

- Social network – no data is isolated
- Visualize using graph network tools
- Applying centrality theorem
  - Some mathematics to explain relationship between entities and physics behind our data model

Ultimately to tie everything back

- Derive 1st, 2nd, and 3rd order meanings from the social network graph models

Wrapping things up

- This course is about data science
  - Capture data
  - Analyze data
  - Visualize data
- Ultimately to
  - Derive meaning
  - Predict behavior
  - Visualize patterns

Framework for Course Software Development

- We will use python as the primary development tool
- Class will work together as a team requiring some expertise with CM
  - Will use either SVN or Git
- Code will have to be properly documented
- Will use third-party open source software
- Expectations:
  - Our code will be open source
  - We will write a “paper” for our findings and discovery

Next Step

- **Learn python** (this is non-negotiable)
  - Write simple trivial examples:
    - Hello World
    - Parse command line arguments
- Learn the python eco system
  - Rome was not built in one day, but Rome was built; so you will need to develop python code within the eco system
- **Learn LaTeX** (this is not negotiable)
  - Required for our paper
- Adopt third-party python libraries
  - No one builds a car from scratch anymore – let's go shopping!