Big Data Analytics
Process & Building Blocks

Duen Horng (Polo) Chau
Georgia Tech

CSE 6242 A / CS 4803 DVA
Jan 10, 2013

Partly based on materials by Professors Guy Lebanon, Jeffrey Heer, John Stasko, Christos Faloutsos
What is Data & Visual Analytics?
What is **Data** & **Visual** Analytics?

No formal definition!
What is **Data & Visual Analytics**?

No formal definition!

**Polo’s definition:**
the *interdisciplinary* science of combining 
*computation techniques* and 
*interactive visualization* 
to transform and model data to aid 
*discovery, decision making, etc.*
What are the “ingredients”? 
What are the “ingredients”?  

Need to worry (a lot) about storage, complex system design, scalability of algorithms, visualization techniques, etc.

Used to be “simpler” before the big data era (why?)
What is **big data**? Why care?

- Many companies’ businesses are based on big data (Google, Facebook, Amazon, Apple, Symantec, LinkedIn, and many more)
- Web search
  - Rank webpages (PageRank algorithm)
  - Predict what you’re going to type
- Advertisement (e.g., on Facebook)
  - Infer users’ interest; show relevant ads
  - Infer what you like, based on what your friends like
- Recommendation systems (e.g., Netflix, Pandora, Amazon)
- Online education
Good news! Many big data jobs

- What jobs are hot?
- “Data scientist”
- Emphasize breadth of knowledge
- This course helps you learn some of the skills
Big data analytics process and building blocks
Process, not “steps”

- Can skip some
- Can go back (two-way street)
- Examples
  - Data types inform visualization design
  - Data informs choice of algorithms
  - Visualization informs data cleaning (dirty data)
  - Visualization informs algorithm design (user finds that results don’t make sense)
How big data affects the process?

- **The 3V of big data**
  - **Volume**: “billions”, “petabytes” are common
  - **Velocity**: think Twitter, fraud detection, etc.
  - **Variety**: text (webpages), video (e.g., youtube), etc.

http://www-01.ibm.com/software/data/bigdata/
Schedule

Wk 2
- Collection
- Cleaning
- Integration
- Analysis
- Visualization

Week 3
- Presentation
- Dissemination

Wk 8 (Hadoop)
- Wk 4: Classification
- Wk 5: Clustering
- Wk 6: Feature selection
- Wk 7: Graphs
- Wk 9: Hadoop
- Wk 12: Text
- Wk 13: Anomaly detection
Two example analytics processes
NetProbe:
Fraud Detection in Online Auction
WWW 2007

NetProbe: The Problem

Find **bad sellers** (fraudsters) on eBay who don’t deliver their items

Auction fraud is **#3** online crime in 2010

source: [www.ic3.gov](http://www.ic3.gov)
NetProbe: Key Ideas

- Fraudsters **fabricate their reputation** by “trading” with their accomplices
- Fake transactions form **near bipartite cores**
- How to detect them?
NetProbe: Key Ideas

Use Belief Propagation

Darker means more likely

Fraudster
Accomplice
Honest

Fraudsters
Accomplices
Honest
NetProbe: Main Results
“Belgian Police”
NetProbe Alpha - Unearth Networks of Suspicious Auction Users

Inspect user alisher for suspicious networks.

NetProbe alpha

Carnegie Mellon
School of Computer Science

alisher

Registration: Aug-13-06
Location: United States

Fraudsters: 95%
Accomplice: 4%
Honest: 1%

Suspected fraudster -- this user has been behaving much like the other suspects by trading with the similar sets of possible accomplices.
What analytics process does NetProbe go through?

- Collection: Scraping
- Cleaning
- Integration
- Analysis: Design detection algorithm
- Visualization
- Presentation: Paper, talks, lectures
- Dissemination: Not released
Discovr movie app
What analytics process would you go through to build the app?

- **Collection**: IMDB, Rotten tomatoes, youtube
- **Cleaning**: May have duplicate trailers
- **Integration**: Determine which movies are related
- **Analysis**: 
- **Visualization**: 
- **Presentation**: 
- **Dissemination**: Mac app, iOS app
Homework 1 (out Jan 17)

- Simple “End-to-end” analysis
- Collect data from Rotten Tomatoes (using API)
  - Movies (Actors, directors, related movies, etc.)
  - Store in SQLite database
- Transform data to movie-movie network
- Analyze, using SQL queries (e.g., create graph’s degree distribution)
- Visualize, using Gephi
- Describe your discoveries