Analytics Building Blocks

Duen Horng (Polo) Chau
Assistant Professor
Associate Director, MS Analytics
Georgia Tech

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, interaction techniques, statistical tests, etc.

Wasn’t this complex before this big data era. Why?
THE WORLD OF DATA

NUMBER OF EMAILS SENT EVERY SECOND
2.9 MILLION

DATA CONSUMED BY HOUSEHOLDS EACH DAY
375 MEGABYTES

VIDEO UPLOADED TO YOUTUBE EVERY MINUTE
20 HOURS

DATA PER DAY PROCESSED BY GOOGLE
24 PETABYTES

TWEETS PER DAY
50 MILLION

TOTAL MINUTES SPENT ON FACEBOOK EACH MONTH
700 BILLION

DATA SENT AND RECEIVED BY MOBILE INTERNET USERS
1.3 EXABYTES

PRODUCTS ORDERED ON AMAZON PER SECOND
72.9 ITEMS

SOURCES: Cisco, comScore, MapReduce, Radiant Group, Twitter, YouTube

IN THE 21ST CENTURY, we live a large part of our lives online. Almost everything we do is reduced to bits and sent through cables around the world at light speed. But just how much data are we generating? This is a look at just some of the massive amounts of information that human beings create every single day.

http://spanning.com/blog/choosing-between-storage-based-and-unlimited-storage-for-cloud-data-backup/
What is big data? Why care?

("big data" is buzz word, so is “IoT” - Internet of Things)

- **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
- Health IT: patient records (EMR)
- Bio and Chemical modeling:
  - Finance
  - Cybersecurity
- Internet of Things (IoT)
Good news! Many jobs!

Most companies are looking for “data scientists”

The data scientist role is critical for organizations looking to extract insight from information assets for ‘big data’ initiatives and requires a broad combination of skills that may be fulfilled better as a team
- Gartner (http://www.gartner.com/it-glossary/data-scientist)

Breadth of knowledge is important.

This course helps you learn some important skills.
Analytics Building Blocks
Collection
Cleaning
Integration
Analysis
Visualization
Presentation
Dissemination
Building blocks, 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 Vs of big data (3Vs, 4Vs, now 7Vs)

**Volume**: “billions”, “petabytes” are common

**Velocity**: think Twitter, fraud detection, etc.

**Variety**: text (webpages), video (youtube)…

**Veracity**: uncertainty of data

**Variability**

**Visualization**

**Presentation**

**Dissemination**

http://www.ibmbigdatahub.com/infographic/four-vs-big-data

http://dataconomy.com/seven-vs-big-data/
Gartner's 2016 Hype Cycle

http://www.gartner.com/newsroom/id/3412017
https://en.wikipedia.org/wiki/Hype_cycle
“Artificial Intelligence”

Self-Driving Taxis Hit the Streets of Singapore

by Kirsten Korosec  @kirstenkorosec  AUGUST 25, 2016, 4:09 AM EDT
We’re in the 3rd wave of “AI” boom

• Two “AI winters” before

• We should be cautiously optimistic
  (Polo’s motto)
Microsoft silences its new A.I. bot Tay, after Twitter users teach it racism [Updated]

Posted Mar 24, 2016 by Sarah Perez (@sarahintampa)

Microsoft’s newly launched A.I.-powered bot called Tay, which was responding to tweets and chats on GroupMe and Kik, has already been shut down due to concerns with its inability to recognize when it was making offensive or racist statements. Of course, the bot wasn’t *coded* to be racist, but it “learns” from those it interacts with. And naturally, given that this is the Internet, one of the first things online users taught Tay was how to be racist, and how to spout back ill-informed or inflammatory political opinions. [Update: Microsoft now says it’s “making adjustments” to Tay in...
We (along with researchers from Berkeley and Stanford) are co-authors on today’s paper led by Google Brain researchers, Concrete Problems in AI Safety. The paper explores many research problems around ensuring that modern machine learning systems operate as intended. (The problems are very practical, and we’ve already seen some being integrated into OpenAI Gym.)

Advancing AI requires making AI systems smarter, but it also requires preventing accidents — that is, ensuring that AI systems do what people actually want them to do. There’s been an increasing focus on safety research from the machine learning community, such as a recent paper from DeepMind and FHI. Still, many machine learning researchers have wondered just how much safety research can be...
Schedule

Collection
Cleaning
Integration
Analysis
Visualization
Presentation
Dissemination
Two Example Projects
from Polo Club
Apolo Graph Exploration: Machine Learning + Visualization

Duen Horng (Polo) Chau, Aniket Kittur, Jason I. Hong, Christos Faloutsos. CHI 2011.
Finding More Relevant Nodes

Citation network

HCI Paper

Data Mining Paper
Finding More Relevant Nodes

Citation network
Finding More Relevant Nodes

Apolo uses guilt-by-association (Belief Propagation)

Citation network
**Demo: Mapping the Sensemaking Literature**

**Nodes:** 80k papers from Google Scholar (node size: `#citation`)

**Edges:** 150k citations
The cost structure of sensemaking


245 citations 8 versions
The cost structure of sensemaking

245 citations 8 versions
Key Ideas (Recap)

Specify **exemplars**

Find **other** relevant nodes (BP)
What did **Apolo** go through?

- **Collection**
  - Scrape Google Scholar. No API :(

- **Cleaning**

- **Integration**

- **Analysis**
  - Design inference algorithm
    (Which nodes to show next?)

- **Visualization**
  - Interactive visualization you just saw

- **Presentation**
  - Paper, talks, lectures

- **Dissemination**
Apolo: Making Sense of Large Network Data by Combining Rich User Interaction and Machine Learning

Duen Horng (Polo) Chau, Aniket Kittur, Jason I. Hong, Christos Faloutsos
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213, USA
{dchau, nkittur, jasonh, christos}@cs.cmu.edu

ABSTRACT
Extracting useful knowledge from large network datasets has become a fundamental challenge in many domains, from scientific literature to social networks and the web. We introduce Apolo, a system that uses a mixed-initiative approach—combining visualization, rich user interaction and machine learning—to guide the user to incrementally and interactively explore large network data and make sense of it. Apolo engages the user in bottom-up sensemaking to gradually build up an understanding over time by starting small, rather than starting big and drilling down. Apolo also helps users find relevant information by specifying exemplars, and then using a machine learning method called Belief Propagation to infer which other nodes may be of interest. We evaluated Apolo with twelve participants in a between-subjects study, with the task being to find relevant new papers to update an existing survey paper. Using expert judges, participants using Apolo found significantly more relevant papers. Subjective feedback of Apolo was also very positive.

Figure 1. Apolo displaying citation network data around the article The Cost Structure of Sensemaking. The user gradually builds up a mental model of the research areas around the article by manually inspecting some neighboring articles in the visualization and specifying them as exemplar articles (with colored dots underneath) for some ad hoc groups, and instructs Apolo to find more articles relevant to them.

NetProbe: A Fast and Scalable System for Fraud Detection in Online Auction Networks. Shashank Pandit, Duen Horng (Polo) Chau, Samuel Wang, Christos Faloutsos. 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
NetProbe: Key Ideas

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

[Diagram showing a network with Fraudsters, Accomplices, and Honest individuals]
NetProbe: Key Ideas

Use Belief Propagation

Fraudster
Accomplice
Honest

Darker means more likely

Fraudsters
Accomplices
Honest
NetProbe: Main Results
THE WALL STREET JOURNAL.

“Belgian Police”
NetProbe Alpha - Unearth Networks of Suspicious Auction Users

Inspect user alisher for suspicious networks.

alisher

54881

521

54943

567

565

568

560

566

546

alisher

Registration: Aug. 13, 2006
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 did **NetProbe** go through?

- **Collection**: Scraping (built a “scraper”/“crawler”)
- **Cleaning**
- **Integration**
- **Analysis**: Design detection algorithm
- **Visualization**
- **Presentation**: Paper, talks, lectures
- **Dissemination**: Not released
ABSTRACT

Given a large online network of online auction users and their histories of transactions, how can we spot anomalies and auction fraud? This paper describes the design and implementation of NetProbe, a system that we propose for solving this problem. NetProbe models auction users and transactions as a Markov Random Field tuned to detect the suspicious patterns that fraudsters create, and employs a Belief Propagation mechanism to detect likely fraudsters. Our experiments show that NetProbe is both efficient and effective for fraud detection. We report experiments on synthetic graphs with as many as 7,000 nodes and 30,000 edges, where NetProbe was able to spot fraudulent nodes with over 90% precision and recall, within a matter of seconds. We also report experiments on a real dataset crawled from eBay, with nearly 700,000 transactions between more than 66,000 users, where NetProbe was highly effective at unearthing hidden networks of fraudsters, within a realistic response time of about 6 minutes. For scenarios where the underlying data is dynamic in nature, we propose Incremental NetProbe, which is an approximate, but fast, variant of NetProbe. Our experiments prove that Incremental NetProbe
Homework 1 (out next week; tasks subject to change)

- Simple “End-to-end” analysis
- Collect data 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