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
Building blocks, not “steps”

- Collection
- Cleaning
- Integration
- Analysis
- Visualization
- Presentation
- Dissemination

Can skip some

Can go back (two-way street)

Examples

- **Data types** inform **visualization** design
- **Data size** informs choice of **algorithms**
- **Visualization** motivates more **data cleaning**
- **Visualization** challenges **algorithm** assumptions e.g., user finds that results don’t make sense
How big data affects the process?

The Vs of big data (used to be 3Vs, now 7Vs)

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

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

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

**Veracity**: uncertainty of data

**Variability**

**Value**

http://www.ibmbigdatahub.com/infographic/four-vs-big-data
http://dataconomy.com/seven-vs-big-data/
Gartner's 2017 Hype Cycle (debatable)

Note: PaaS = platform as a service; UAVs = unmanned aerial vehicles

https://en.wikipedia.org/wiki/Hype_cycle
“Artificial Intelligence”

Self-Driving Taxis Hit the Streets of Singapore

by Kirsten Korosec  @kirstenkrososec  AUGUST 25, 2016, 4:09 AM EDT

Google AI beats Go world champion again to complete historic 4-1 series victory

Posted Mar 15, 2016 by Jon Russell (@jonrussell)
We’re in the 3rd wave of “AI” boom

- Two “AI winters” before 

- We should be **cautiously optimistic**
  (Polo’s motto)
Neither Autopilot nor the driver noticed the white side of the tractor trailer against a brightly lit sky, so the brake was not applied.
Good Read about AI: White House Report

Preparing for The Future of Artificial Intelligence

https://www.whitehouse.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf
“The Current State of AI

Remarkable progress has been made on what is known as **Narrow AI**, which addresses specific application areas such as playing strategic games, language translation, self-driving vehicles, and image recognition. Narrow AI underpins many commercial services such as trip planning, shopper recommendation systems, and ad targeting, and is finding important applications in medical diagnosis, education, and scientific research. These have all had significant societal benefits and have contributed to the economic vitality of the Nation.
General AI (sometimes called Artificial General Intelligence, or AGI) refers to a notional future AI system that exhibits apparently intelligent behavior at least as advanced as a person across the full range of cognitive tasks. A broad chasm seems to separate today’s Narrow AI from the much more difficult challenge of General AI. Attempts to reach General AI by expanding Narrow AI solutions have made little headway over many decades of research. The current consensus of the private-sector expert community, with which the NSTC Committee on Technology concurs, is that General AI will not be achieved for at least decades.”
Likely no Matrix or SkyNet in Your Life Time
Schedule

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

BEAUTIFUL HAIRBALL
DEATH STAR
SPAGHETTI
Finding More Relevant Nodes

HCI Paper

Data Mining Paper

Citation network
Finding More Relevant Nodes

Citation network

HCI Paper

Data Mining Paper
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

PDF 1993
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**: You *may* use a new Apolo prototype (called Argo)
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.
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
Find **bad sellers** *(fraudsters)* on eBay who don’t deliver their items

**NetProbe: The Problem**

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?
Use Belief Propagation

Fraudster

Accomplice

Honest

Darker means more likely

Fraudsters

Accomplices

Honest
NetProbe: Main Results
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
NetProbe: A Fast and Scalable System for Fraud Detection in Online Auction Networks
Shashank Pandit, Duen Horng (Polo) Chau, Samuel Wang, Christos Faloutsos
Carnegie Mellon University
Pittsburgh, PA 15213, USA
{shashank, dchau, samuelwang, christos}@cs.cmu.edu

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 Twitter API
- Store in SQLite database
- Great graph from data
- Analyze, using SQL queries (e.g., create graph’s degree distribution)
- Visualize graph using Gephi (and maybe Argo)
- Describe your discoveries