http://poloclub.gatech.edu/cse6242

CSE6242 / CX4242: Data & Visual Analytics

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

Collection

Cleaning

Integration

Analysis

Visualization

Presentation

Dissemination

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?

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

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

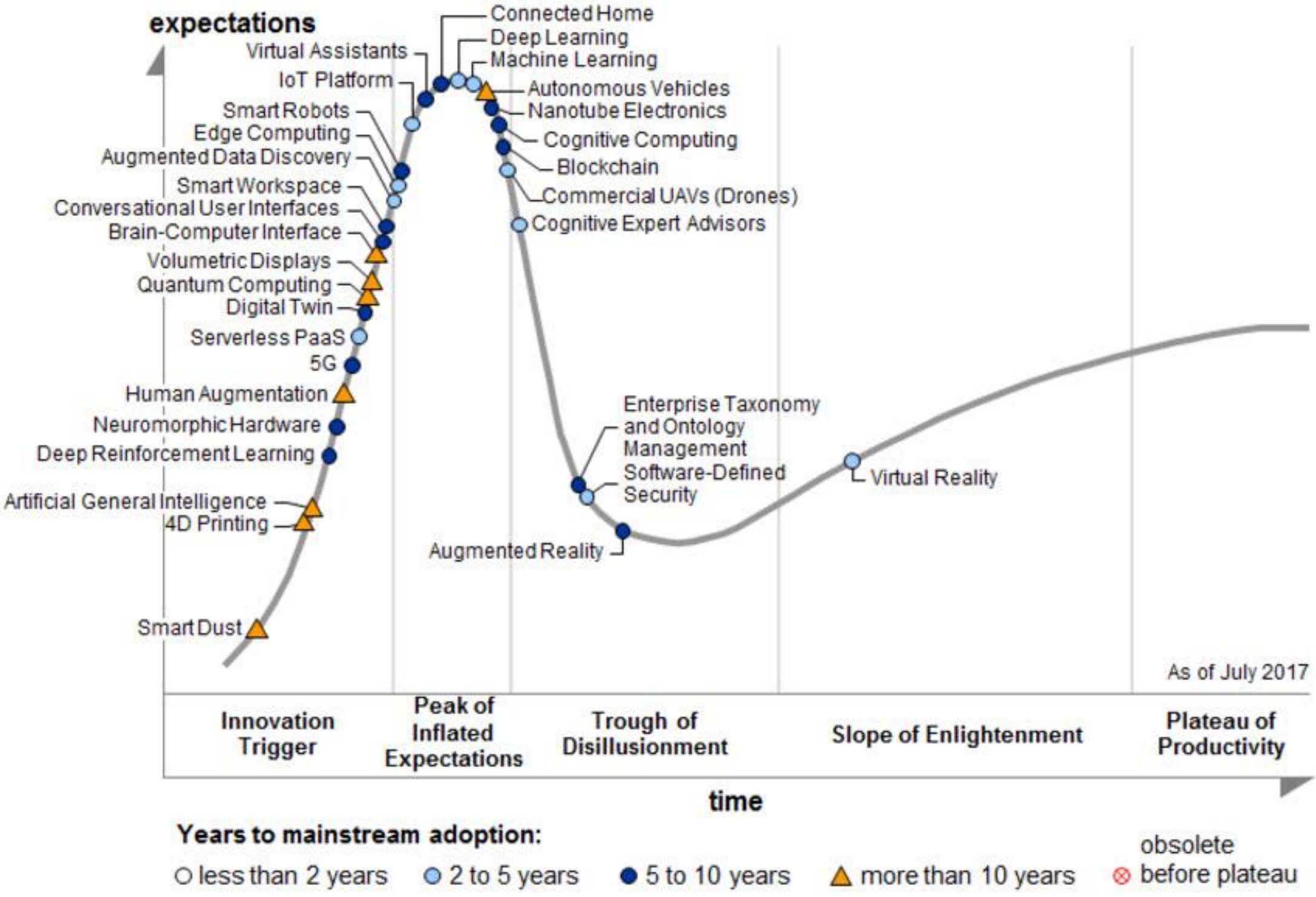
Variability

Visualization

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

Source: Gartner (July 2017)

https://www.forbes.com/sites/louiscolumbus/2017/08/15/gartners-hype-cycle-for-emerging-technologies-2017-adds-5g-and-deep-learning-for-first-time/#3855c7405043
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

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

Posted Mar 15, 2016 by Jon Russell (@jonrussell)

Next Story

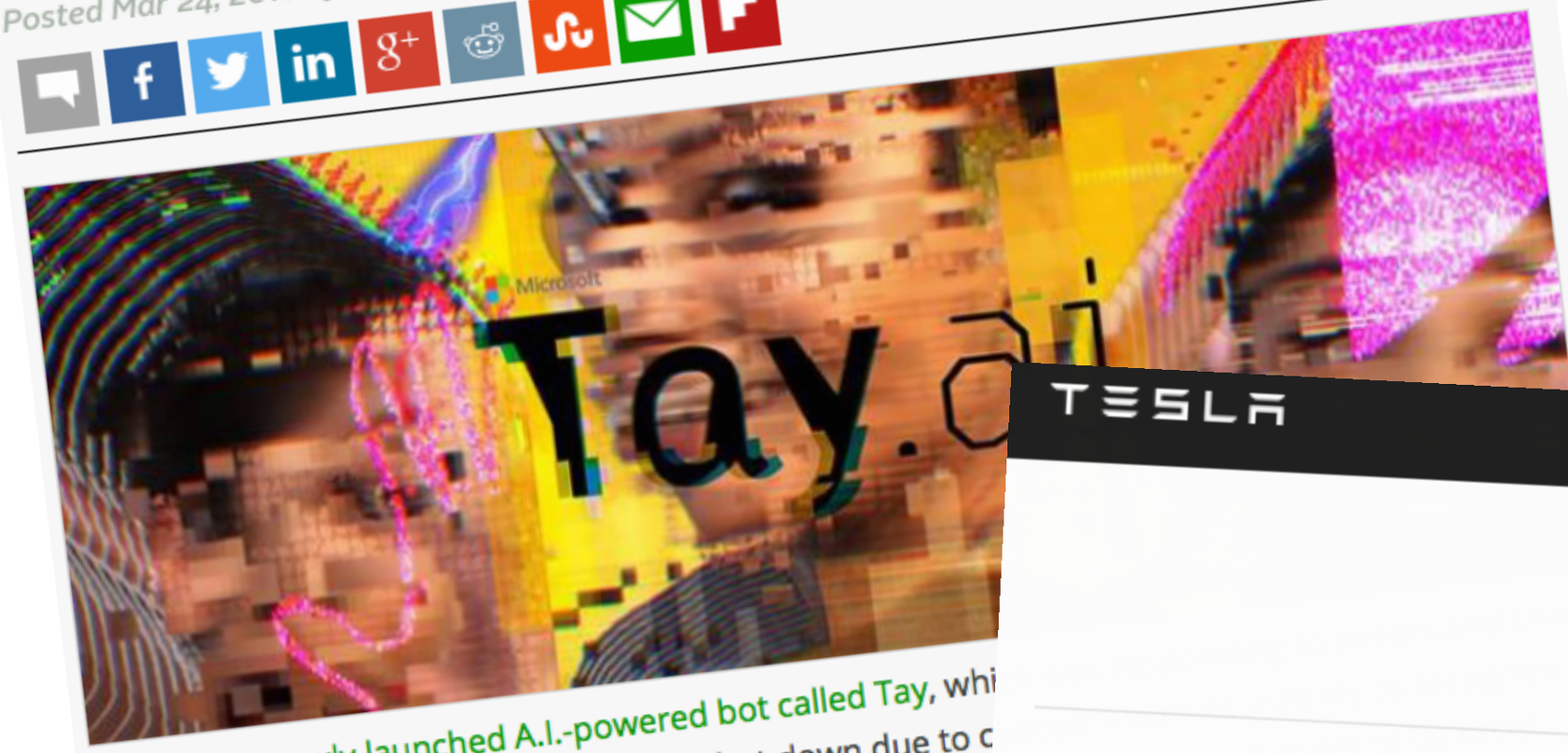
We’re in the 3rd wave of “AI” boom

- Two “AI winters” before https://en.wikipedia.org/wiki/History_of_artificial_intelligence
- 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)

Next Story



Microsoft's newly launched A.I.-powered bot called Tay, whi
on GroupMe and Kik, has already been shut down due to c
when it was making offensive or racist statements. Of cou
but it "learns" from those it interacts with. And naturally, g
first things online users taught Tay was how to be racist,
inflammatory political opinions. [**Update**: Microsoft now

"Neither Autopilot nor the driver noticed the white side of the tractor trailer against a brightly lit sky, so the brake was not applied"

TESLA MODEL S MODEL X MODEL 3 EN

Blog Videos

A Tragic Loss

The Tesla Team • 30 June 2016

We learned yesterday evening that NHTSA is opening a preliminary evaluation into the performance of Autopilot during a recent fatal crash that occurred in a Model S. This is the first known fatality in just over 130 million miles where Autopilot was activated. Among all vehicles in the US, there is a fatality every 94 million miles. Worldwide, there a fatality approximately every 60 million miles. It is important to emphasize that the NHTSA action is simply a preliminary evaluation to determine whether the system worked according to expectations

https://www.tesla.com/en_GB/blog/tragic-loss?redirect=no

Good Read about AI:
White House Report

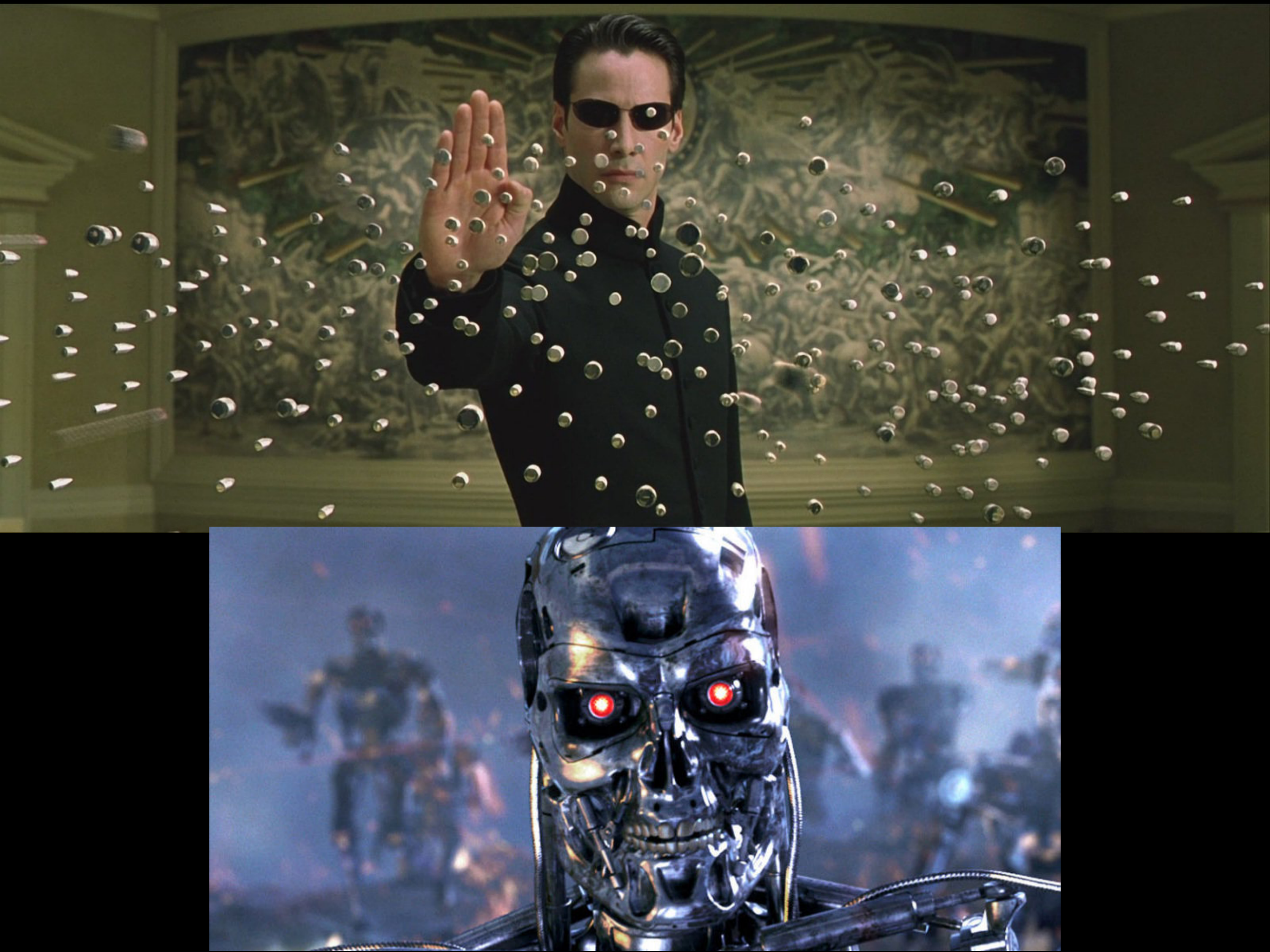
Preparing for The Future of Artificial Intelligence

https://www.whitehouse.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf

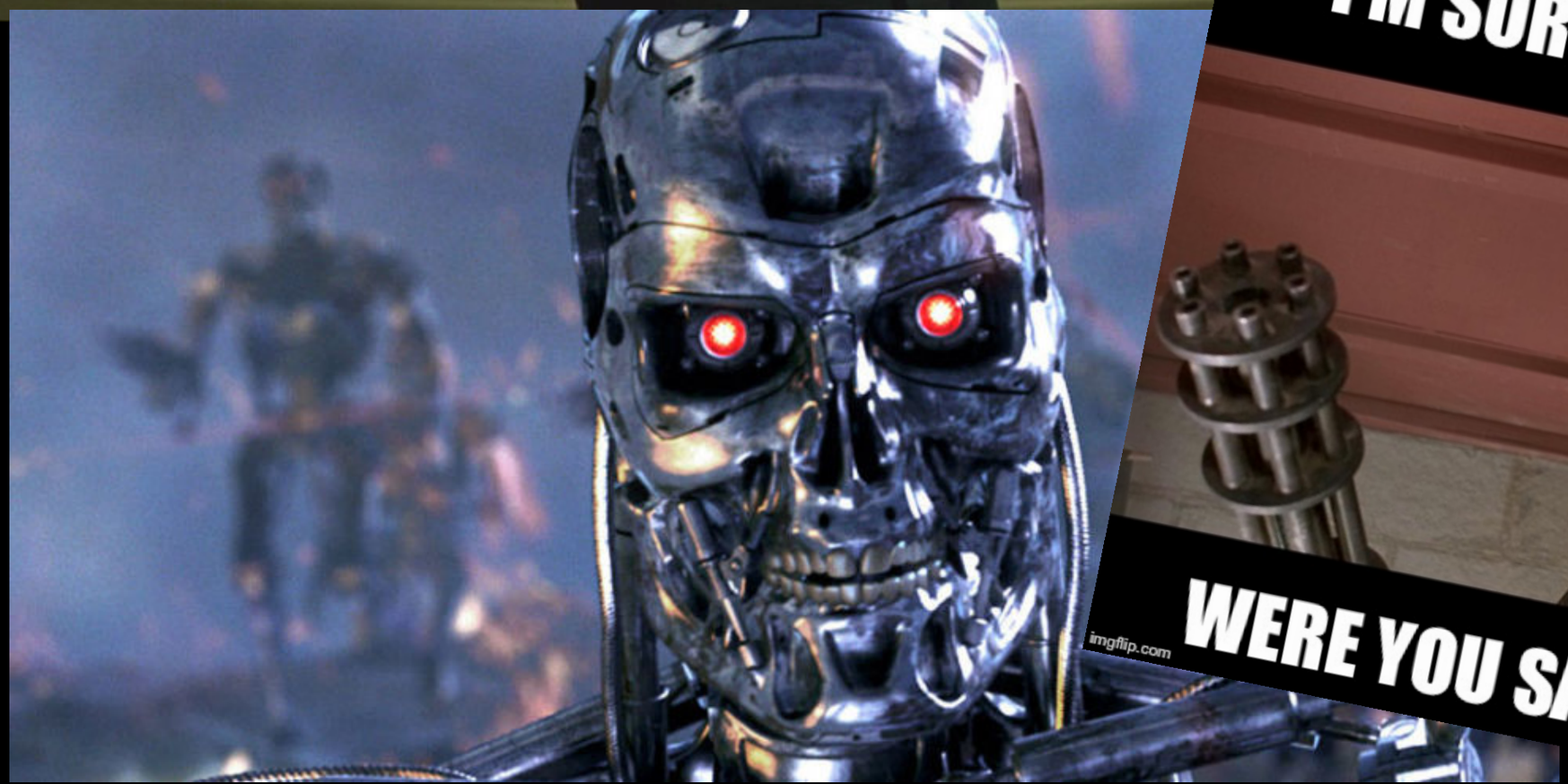
“The Current State of AI

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

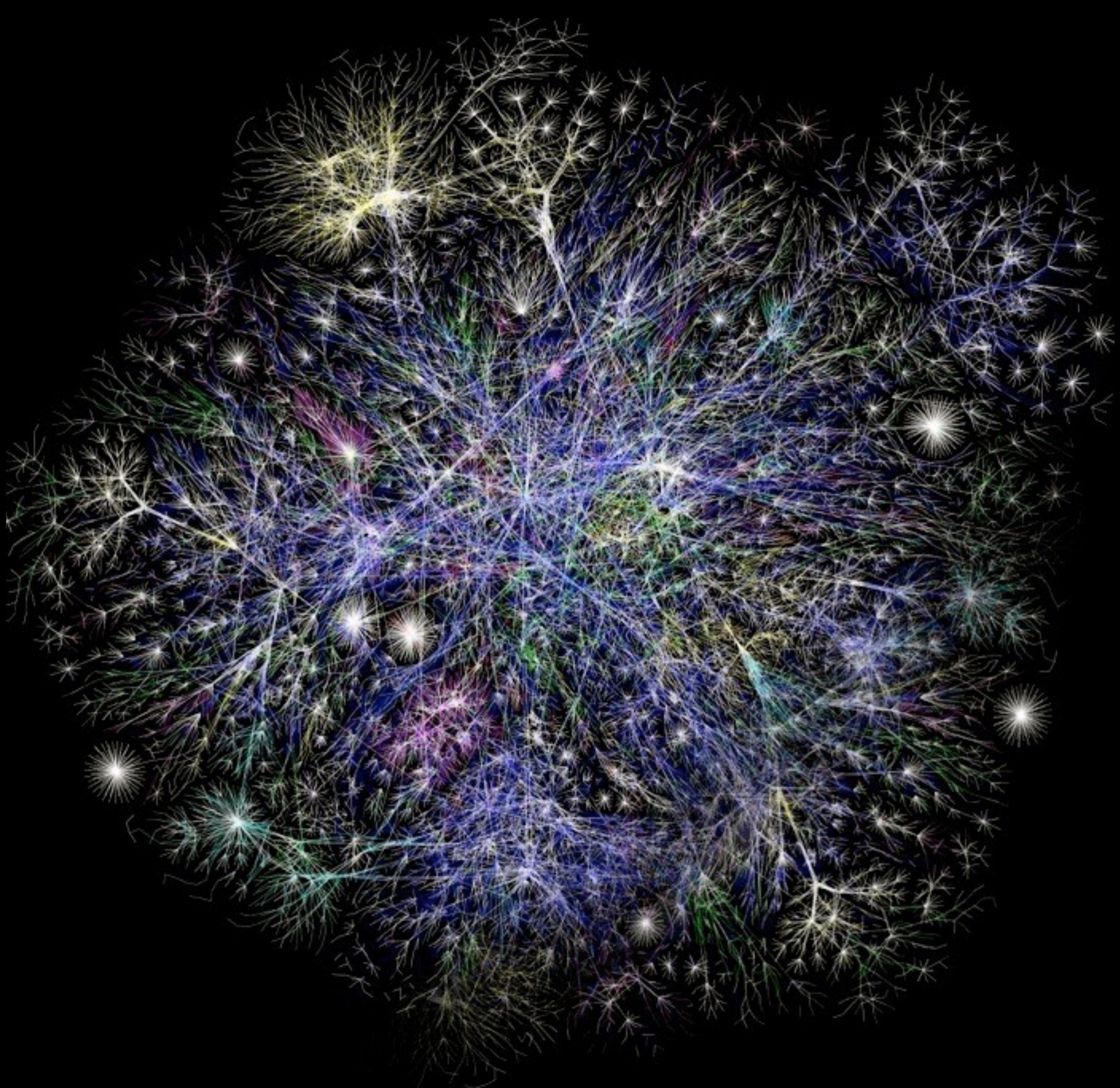
Dissemination

Two Example Projects

from Polo Club

Apolo Graph Exploration:
Machine Learning + Visualization

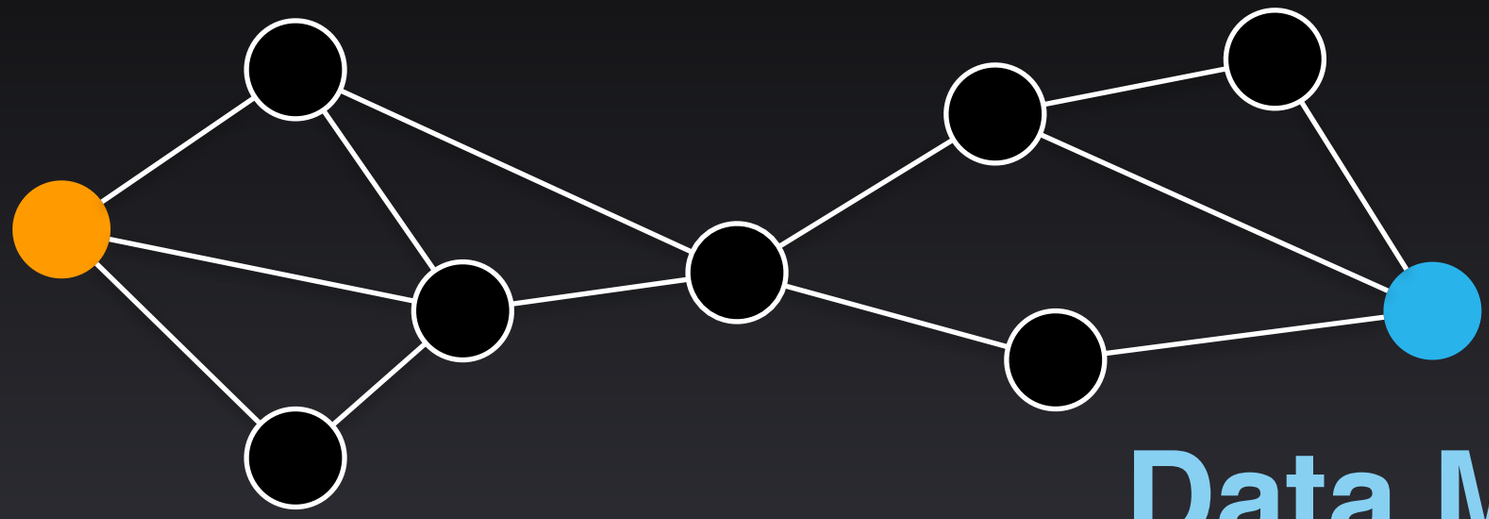
Apolo: Making Sense of Large Network Data by Combining Rich User Interaction and Machine Learning.
Duen Horng (Polo) Chau, Aniket Kittur, Jason I. Hong, Christos Faloutsos. CHI 2011.




BEAUTIFUL HAIRBALL
DEATH STAR
SPAGHETTI

Finding More Relevant Nodes

HCI
Paper

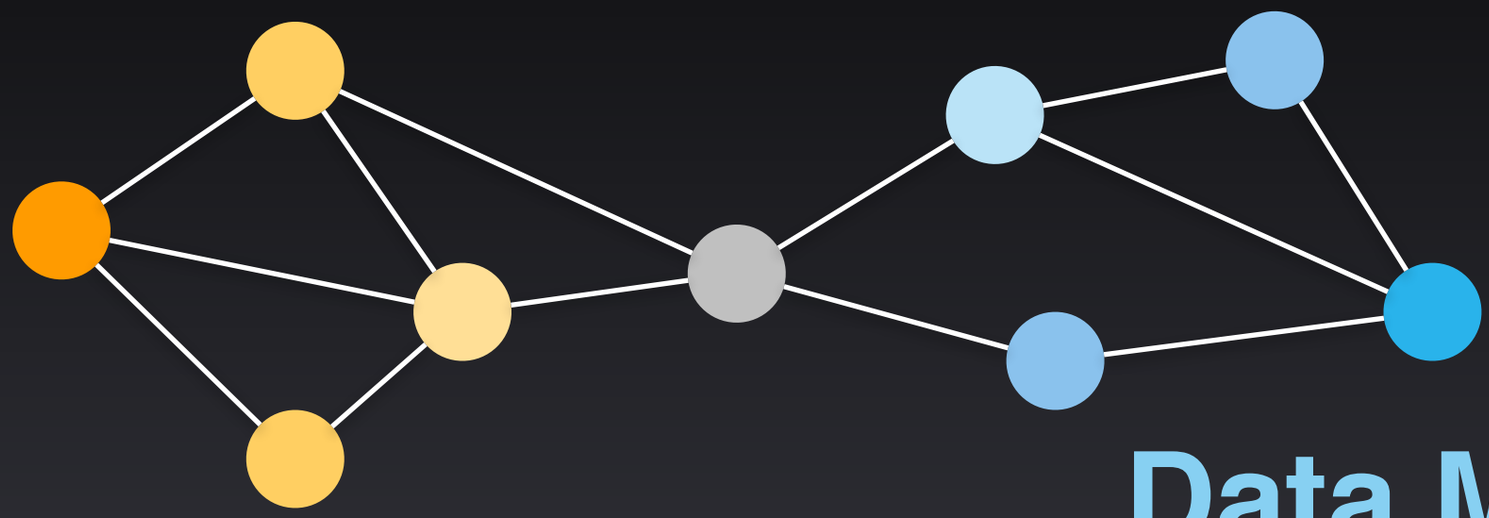


Data Mining
Paper

Citation network

Finding More Relevant Nodes

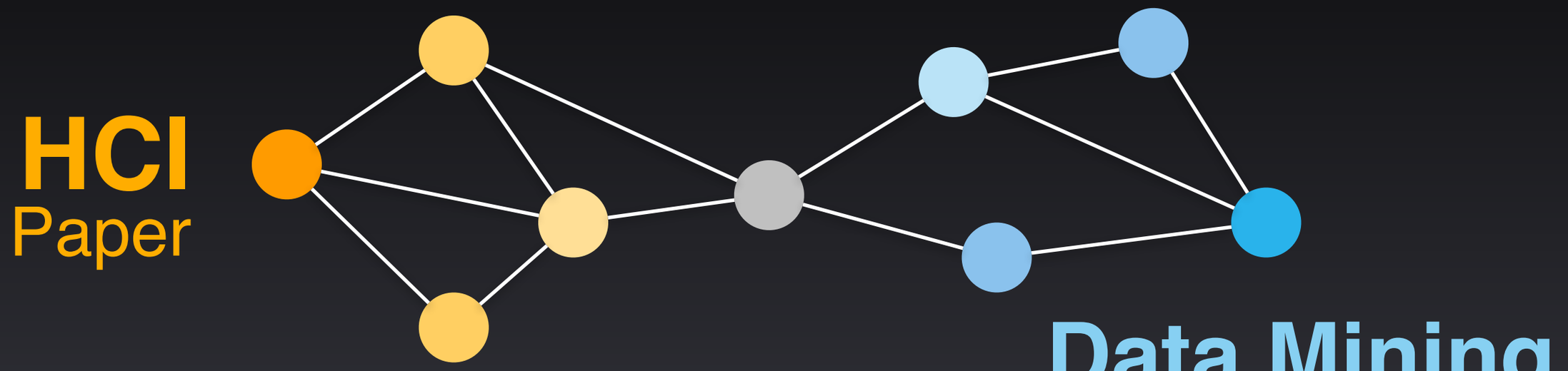
HCI
Paper



Data Mining
Paper

Citation network

Finding More Relevant Nodes

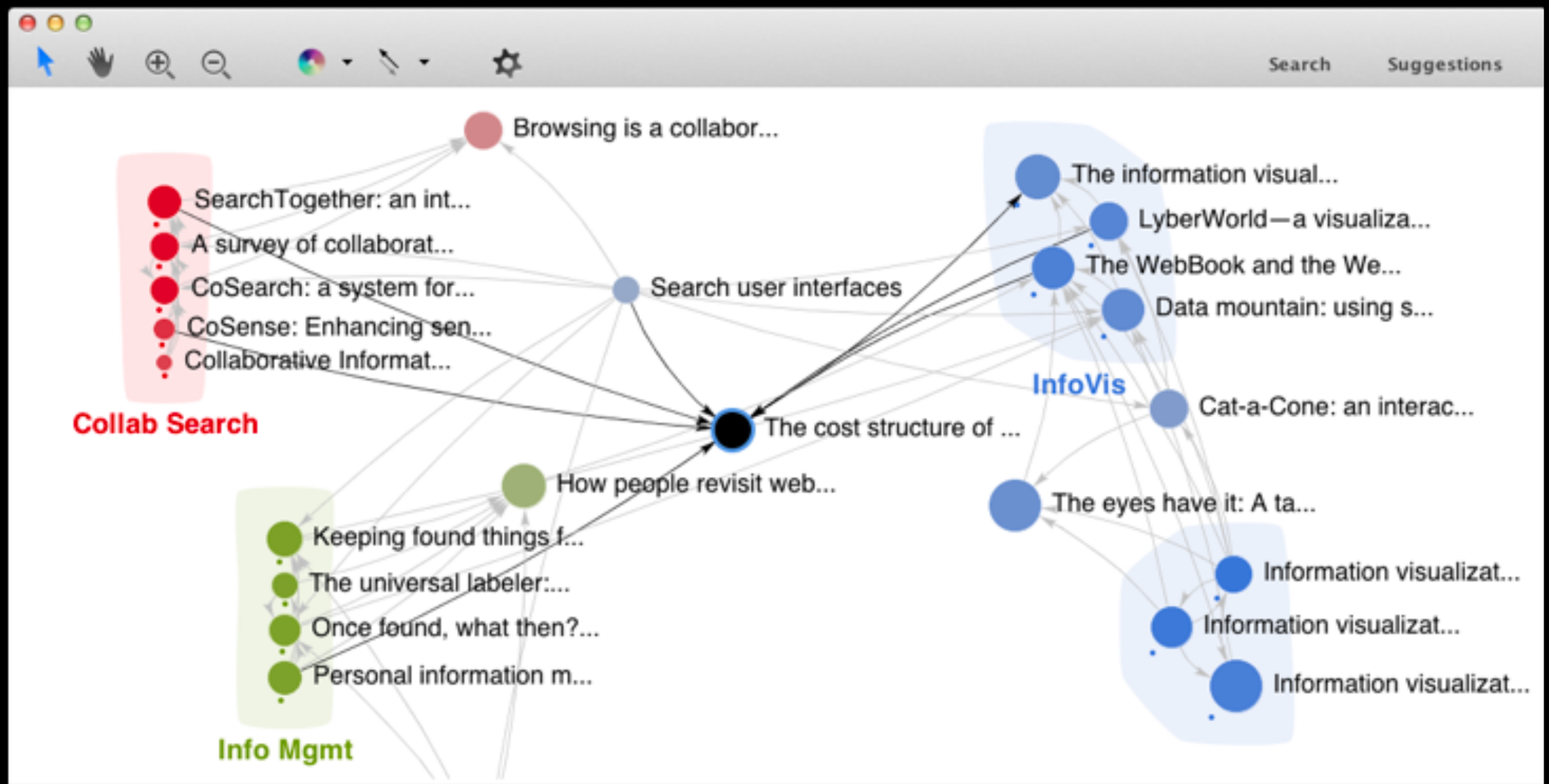


Citation network

Apolo uses **guilt-by-association**
(Belief Propagation)

Demo: Mapping the Sensemaking Literature

Nodes: 80k papers from Google Scholar (node size: #citation)
Edges: 150k citations



Search | Suggestions

For **The cost structure of sensemaking**

The information visualizer, an inf...	1991	
Card, S.K. and Robertson, G.G. and Macki...	532	
The WebBook and the Web Forag...	1996	
Card, S.K. and Robertson, G.G. and York, W.	403	
LyberWorld—a visualization user...	1994	
Hemmje, M. and Kunkel, C. and Willett, A.	223	
The structure of the information...	1997	
Card, S.K. and Mackinlay, J.	198	
Information visualization	2009	
Card, S. and Mackinlay, JD and Shneiderm...	180	
"I"ll get that off the audio": a cas...	1997	
Moran, T.P. and Palen, L. and Harrison, S....	143	
An organic user interface for sear...	1995	
Mackinlay, J.D. and Rao, R. and Card, S.K.	123	
Using a landscape metaphor to re...	1993	
Chalmers, M.	122	
Personal information management	2007	
Jones, W.P. and Teevan, J.	109	
SearchTogether: an interface for c...	2007	
Morris, M.R. and Horvitz, E.	108	
Information foraging theory: Ada...	2007	
Pirolli, P.	107	
Investigating behavioral variabilit...	2007	
White, R.W. and Drucker, S.M.	79	
Jigsaw: Supporting investigative...	2008	
Stasko, J. and Görg, C. and Liu, Z.	71	
The cost-of-knowledge character...	1994	
Card, S.K. and Pirolli, P. and Mackinlay, J.D.	54	
Collaborative conceptual design:...	1996	
Potts, C. and Catledge, L.	45	

The cost structure of sen...

The cost structure of sensemaking PDF 1993

Russell, D.M. and Stefik, M.J. and Pirolli, P. and Card, S.K.

245 citations 8 versions

googlescholar.db

Search | Suggestions

For **The cost structure of sensemaking**

The information visualizer, an inf...	1991	
Card, S.K. and Robertson, G.G. and Macki...	532	
The WebBook and the Web Forag...	1996	
Card, S.K. and Robertson, G.G. and York, W.	403	
LyberWorld—a visualization user...	1994	
Hemmje, M. and Kunkel, C. and Willett, A.	223	
The structure of the information...	1997	
Card, S.K. and Mackinlay, J.	198	
Information visualization	2009	
Card, S. and Mackinlay, JD and Shneiderm...	180	
"I"ll get that off the audio": a cas...	1997	
Moran, T.P. and Palen, L. and Harrison, S....	143	
An organic user interface for sear...	1995	
Mackinlay, J.D. and Rao, R. and Card, S.K.	123	
Using a landscape metaphor to re...	1993	
Chalmers, M.	122	
Personal information management	2007	
Jones, W.P. and Teevan, J.	109	
SearchTogether: an interface for c...	2007	
Morris, M.R. and Horvitz, E.	108	
Information foraging theory: Ada...	2007	
Pirolli, P.	107	
Investigating behavioral variabilit...	2007	
White, R.W. and Drucker, S.M.	79	
Jigsaw: Supporting investigative...	2008	
Stasko, J. and Görg, C. and Liu, Z.	71	
The cost-of-knowledge character...	1994	
Card, S.K. and Pirolli, P. and Mackinlay, J.D.	54	
Collaborative conceptual design:...	1996	
Potts, C. and Catledge, L.	45	

The cost structure of sen...

The cost structure of sensemaking PDF 1993

Russell, D.M. and Stefik, M.J. and Pirolli, P. and Card, S.K.

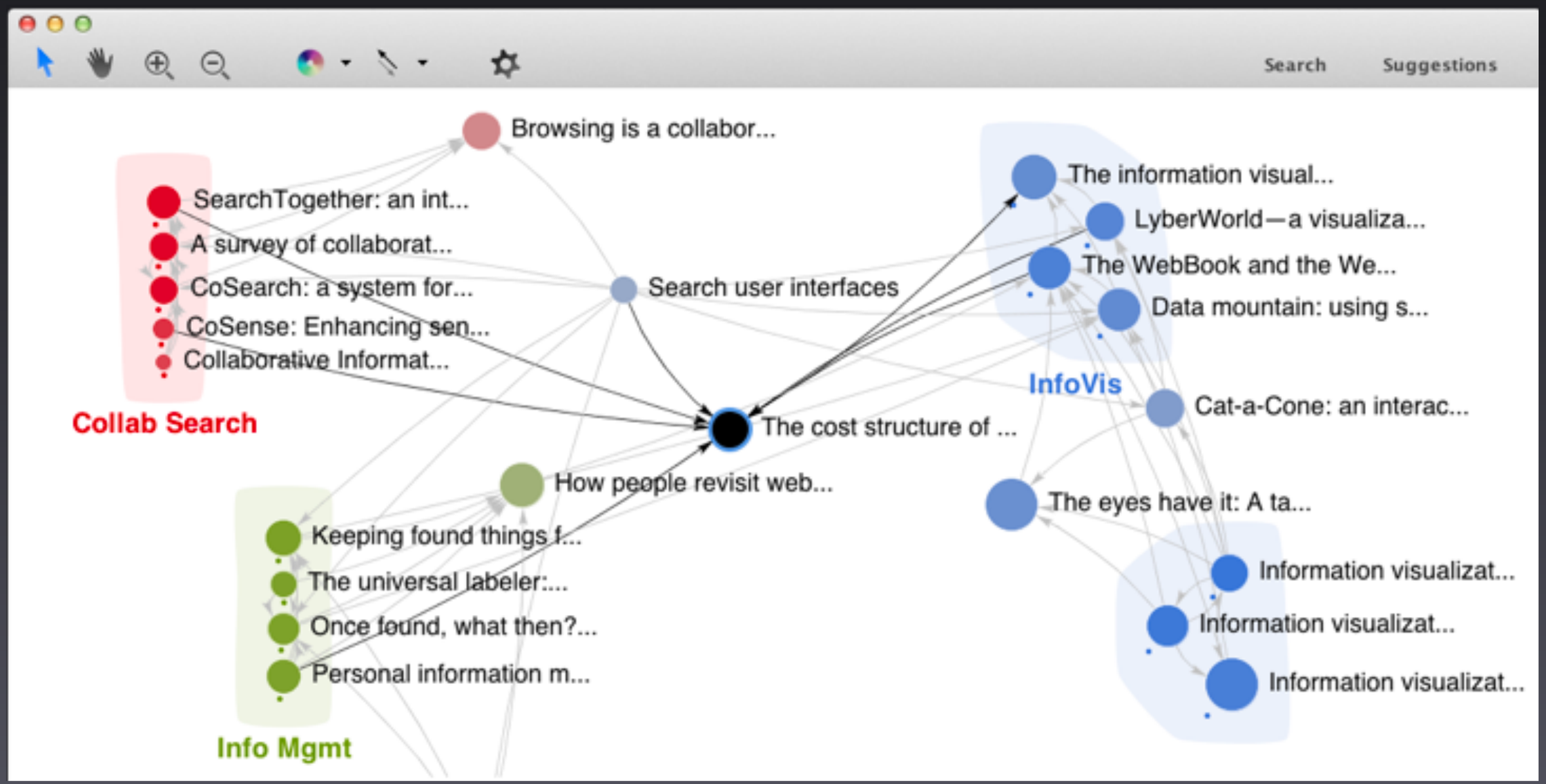
245 citations 8 versions

googlescholar.db

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.

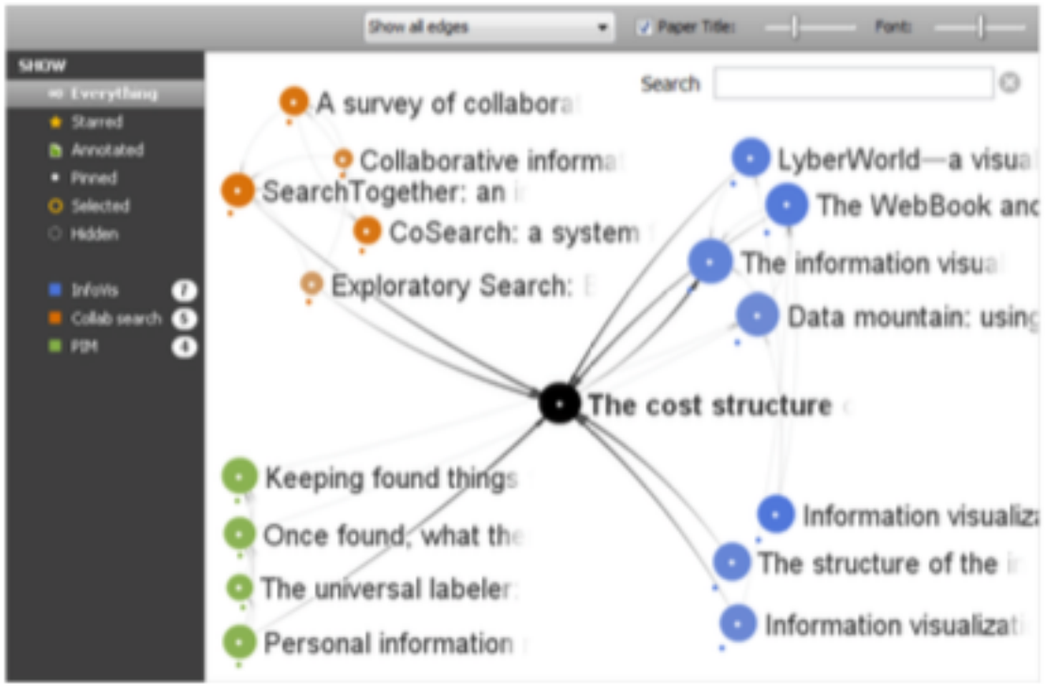


Figure 1. Apolo displaying citation network data around the article *The Cost Structure of Sensemaking*. The user gradually builds up a mental model of the research areas around the article by manually inspecting some neighboring articles in the visualization and specifying them as exemplar articles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

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. *ACM Conference on Human Factors in Computing Systems (CHI) 2011*. May 7-12, 2011.

back; H.5.2 Information Interfaces and Presentation: User

NetProbe:
Fraud Detection in Online Auction

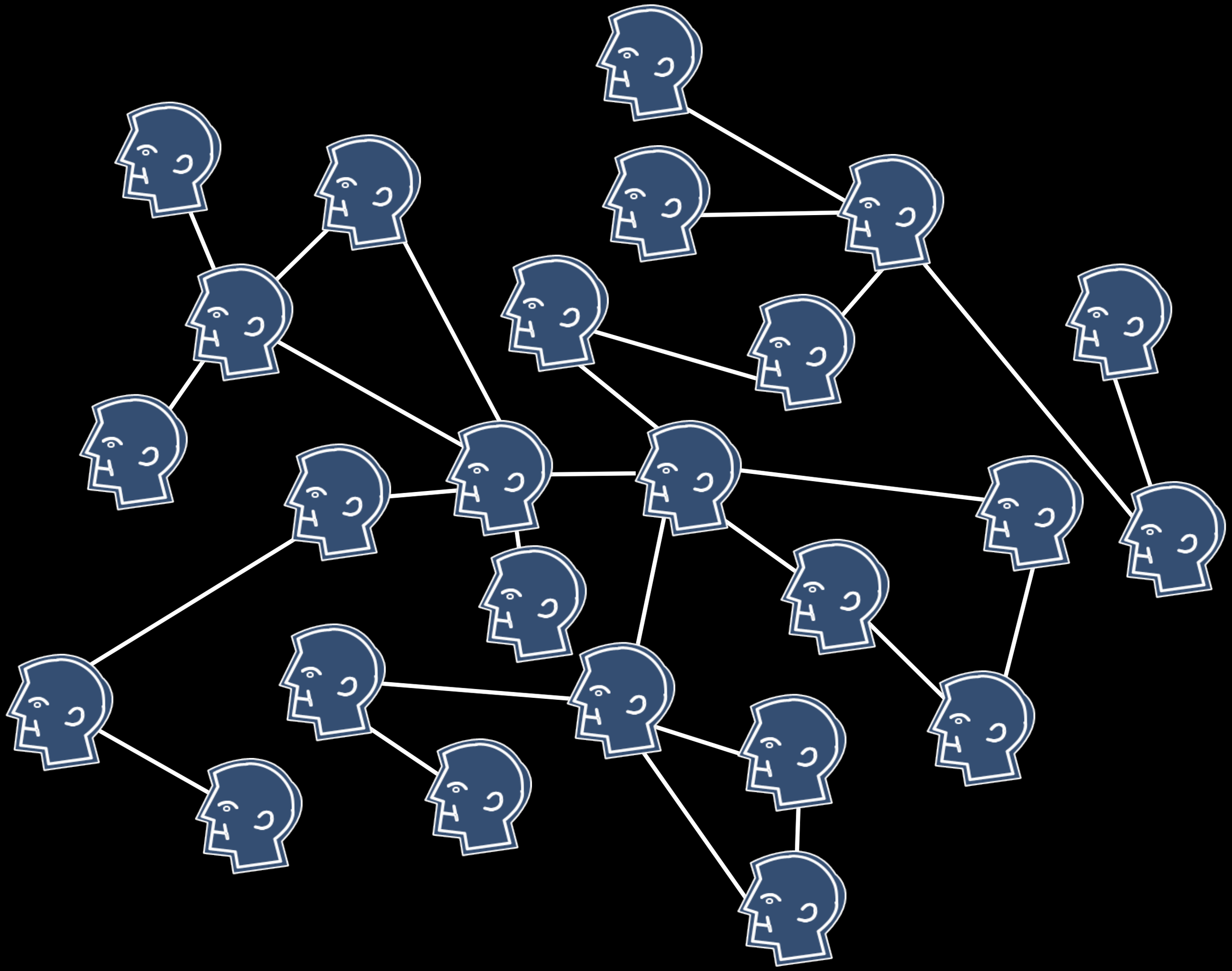
NetProbe: A Fast and Scalable System for Fraud Detection in Online Auction Networks. Shashank Pandit, Duen Horng (Polo) Chau, Samuel Wang, Christos Faloutsos. WWW 2007

NetProbe: The Problem

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

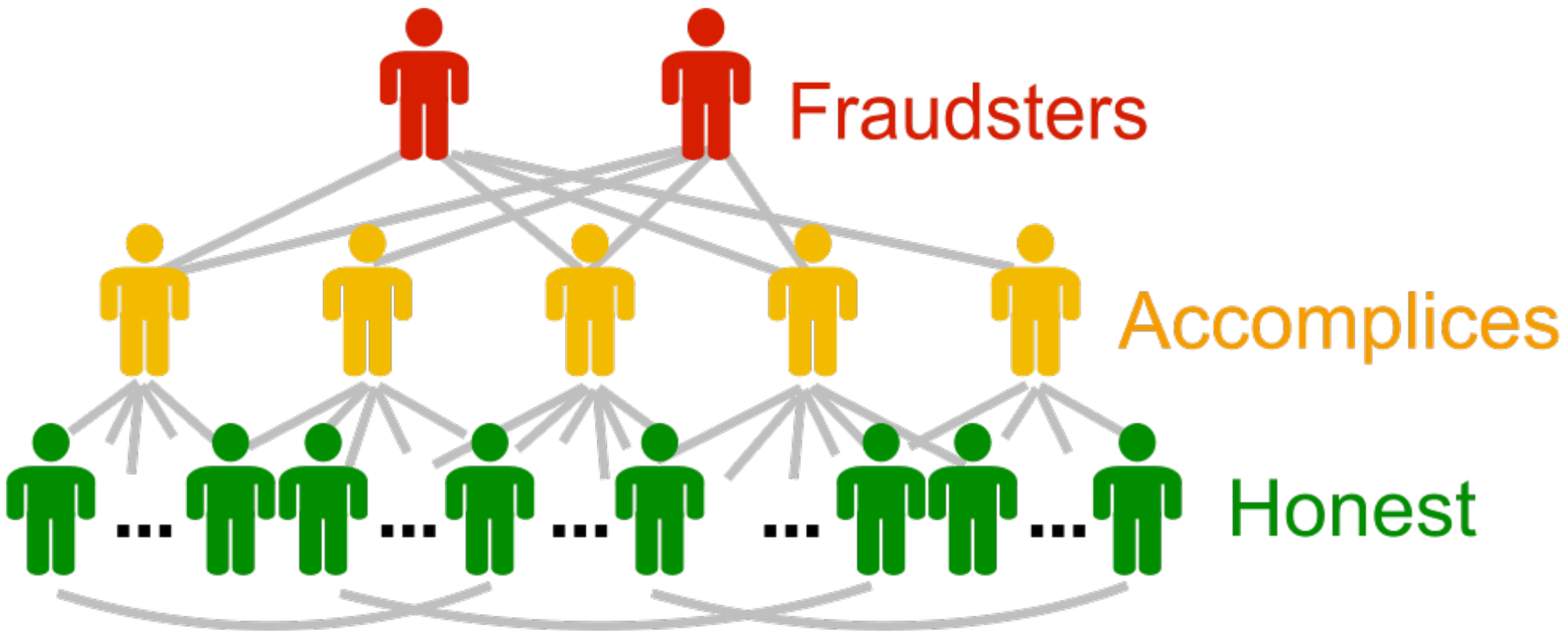
Auction fraud is **#3** online crime in 2010

source: www.ic3.gov



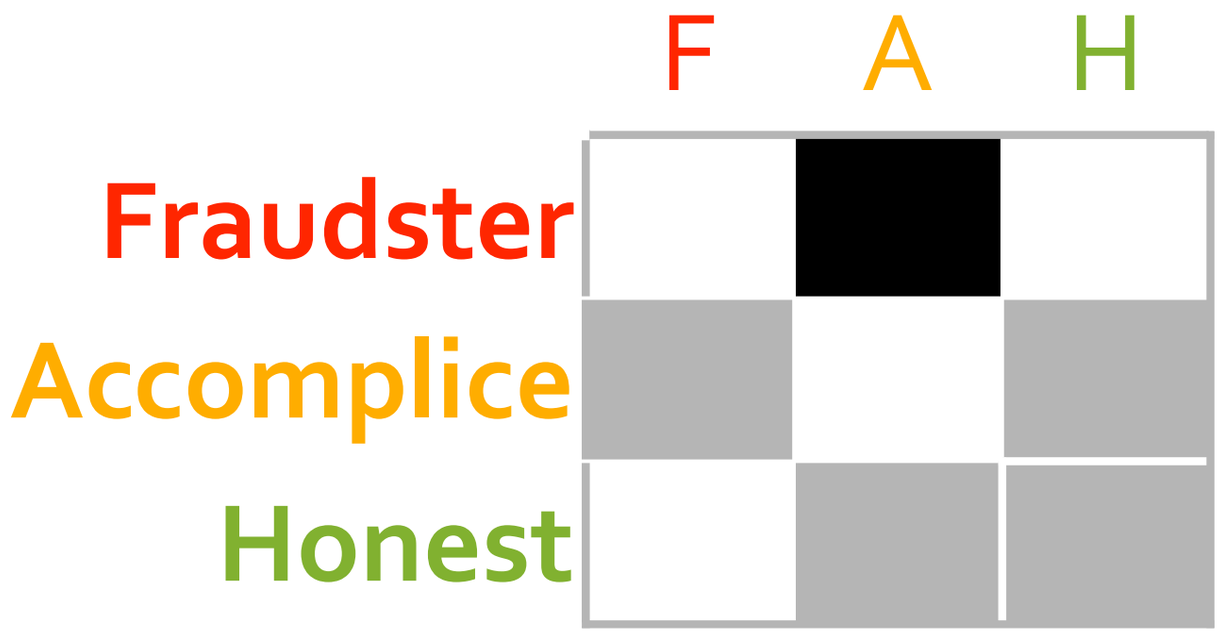
NetProbe: Key Ideas

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

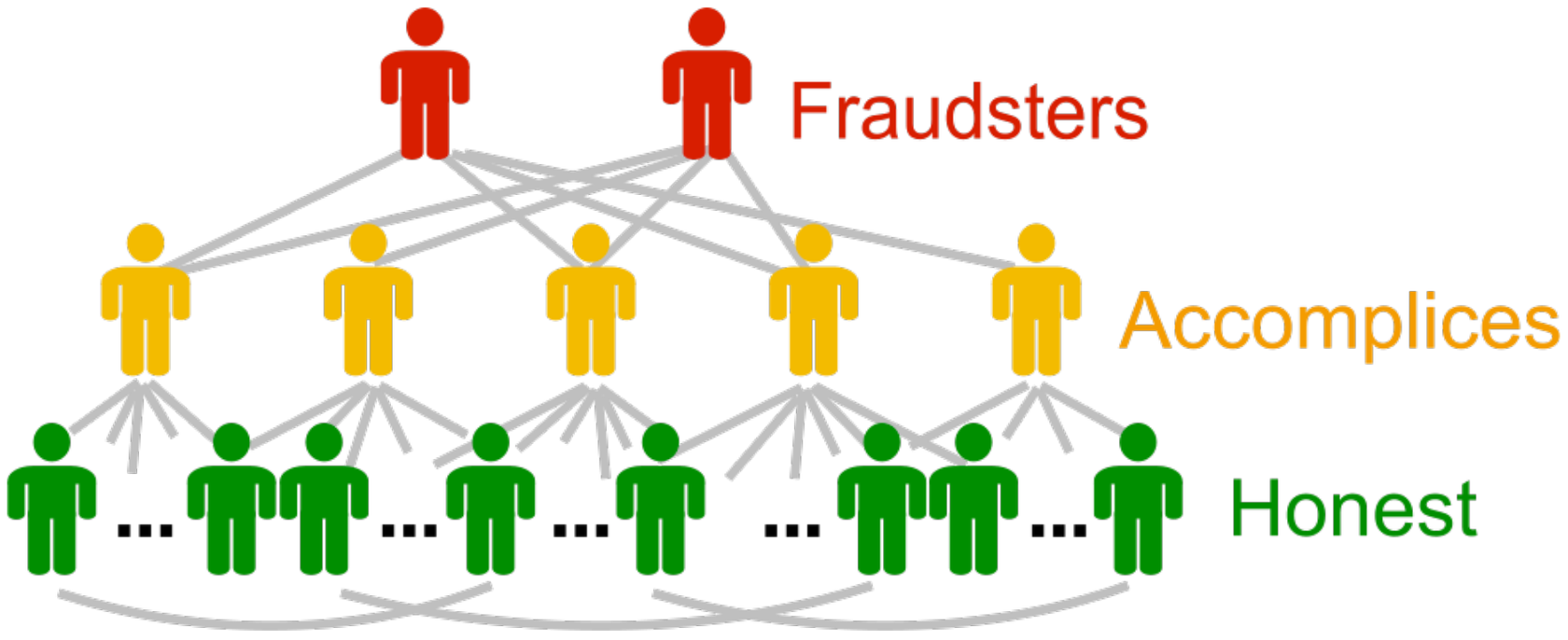


NetProbe: Key Ideas

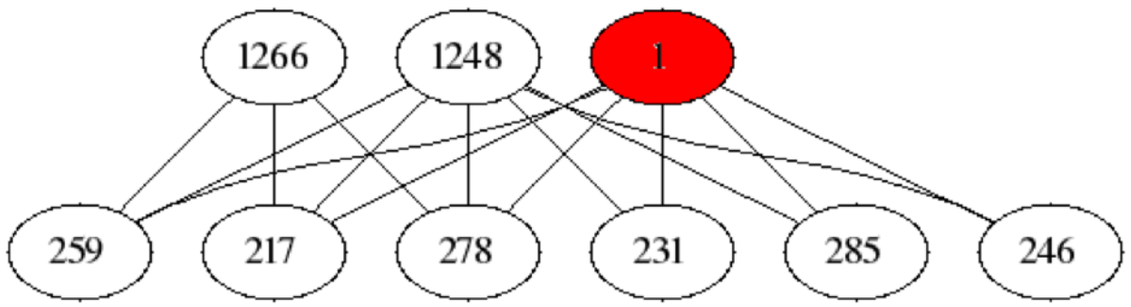
Use **Belief Propagation**

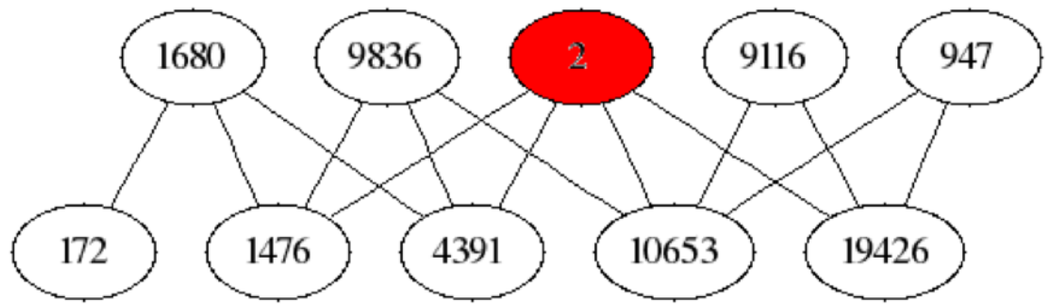


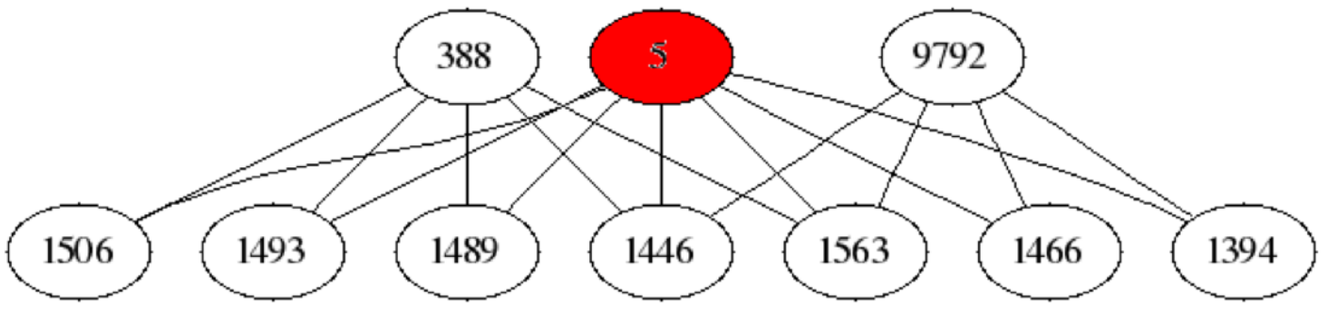
Darker means more likely

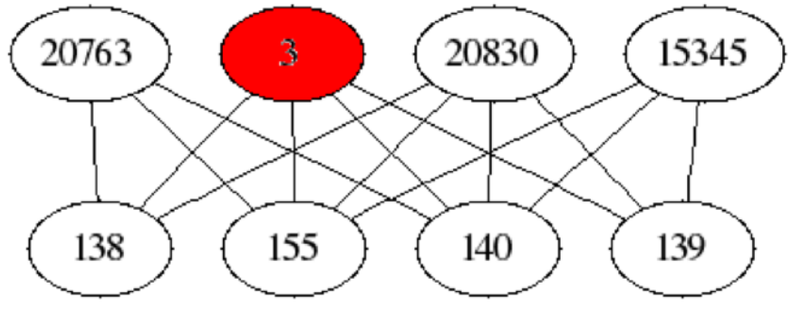


NetProbe: Main Results









eBay

USA TODAY

CNNMoney.com
A Service of CNN, Fortune & Money

THE WALL STREET JOURNAL.

MSNBC

KDKA 2

PITTSBURGH TRIBUNE-REVIEW

Symantec

eBay

USA TODAY

CNNMoney.com
A Service of CNN, Fortune & Money

THE WALL STREET JOURNAL.

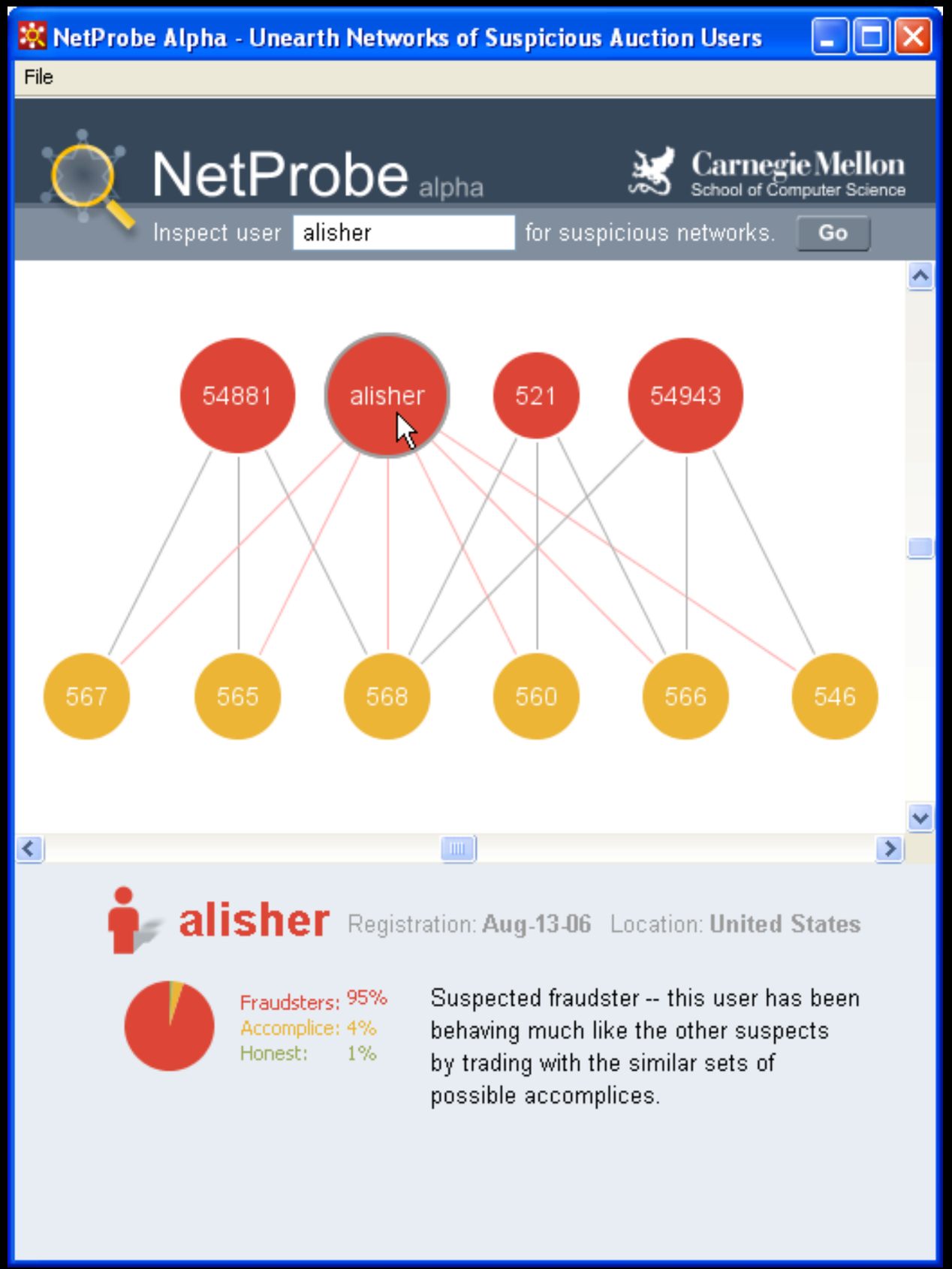
MSNBC

KDKA 2

PITTSBURGH TRIBUNE-REVIEW

Symantec

"Belgian Police"


NetProbe Alpha - Unearth Networks of Suspicious Auction Users
File
NetProbe alpha
Carnegie Mellon
School of Computer Science
Inspect user
alisher
for suspicious networks.
Go
54881
alisher
521
54943
567
565
568
560
566
546
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 did **NetProbe** go through?

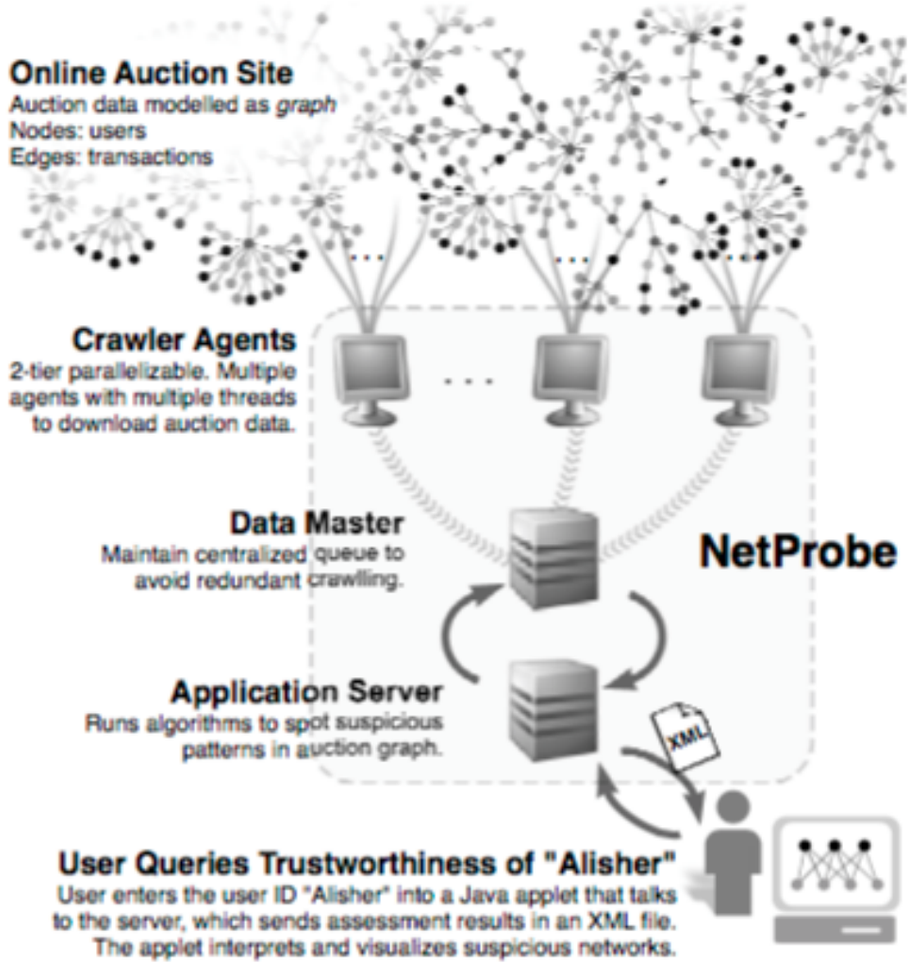
Collection	Scraping (built a “scraper”/“crawler”)
Cleaning	
Integration	
Analysis	Design detection algorithm
Visualization	
Presentation	Paper, talks, lectures
Dissemination	Not released

NetProbe: A Fast and Scalable System for Fraud Detection in Online Auction Networks

Shashank Pandit, Duen Horng 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



NetProbe: A Fast and Scalable System for Fraud Detection in Online Auction Networks. Shashank Pandit, Duen Horng (Polo) Chau, Samuel Wang, Christos Faloutsos. *International Conference on World Wide Web (WWW) 2007*. May 8-12, 2007. Banff, Alberta, Canada. Pages 201-210.

Homework 1 (out next week; tasks subject to change)

Collection

Cleaning

Integration

Analysis

Visualization

Presentation

Dissemination

- 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