CSE6242 / CX4242

Data & Visual Analytics

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

Google “Polo Chau”

Georgia Tech | College of Computing Follow

CV (PDF) | Bio | Publications | Design | Students | Blog | Address

POLO CHAU

Legal name:
Duen Horng Chau

Associate Director, MS in Analytics
Assistant Professor, School of Computational Science & Engineering
Adjunct Assistant Professor, School of Interactive Computing
College of Computing
Georgia Tech

Admin: Carolyn Young Financial Manager: Arlene Washington
polo@gatech.edu www.cc.gatech.edu/~dchau
Office: Klaus 1324 404-385-7682
Google Scholar (h-index: 20) YouTube videos

LinkedIn profile Follow @PoloChau Follow @PoloChau

POSITIONS

May 2014 - Associate Director
MS in Analytics, Georgia Tech

Aug 2012 - Assistant Professor
School of Computational Science & Engineering, Georgia Tech

Dec 2012 - Adjunct Assistant Professor
School of Interactive Computing, Georgia Tech

EDUCATION

Aug 2012 **Ph.D. Machine Learning** Carnegie Mellon University
Thesis: Data Mining Meets HCI: Making Sense of Large Graphs

Polo Club of DATA SCIENCE

NIH MD2K Center of Excellence, Co-PI

MD2K Center of Excellence for Mobile Sensor Data-to-Knowledge

Visualization at Georgia Tech

IDEA workshop at KDD 2013, 2014, 2015

WSDM'16 Publicity Chair

TAs

Gopi Krishnan Nambiar

Nilaksh Das

Pradeep Rajendra

Ajitesh Jain

Vishakha Singh

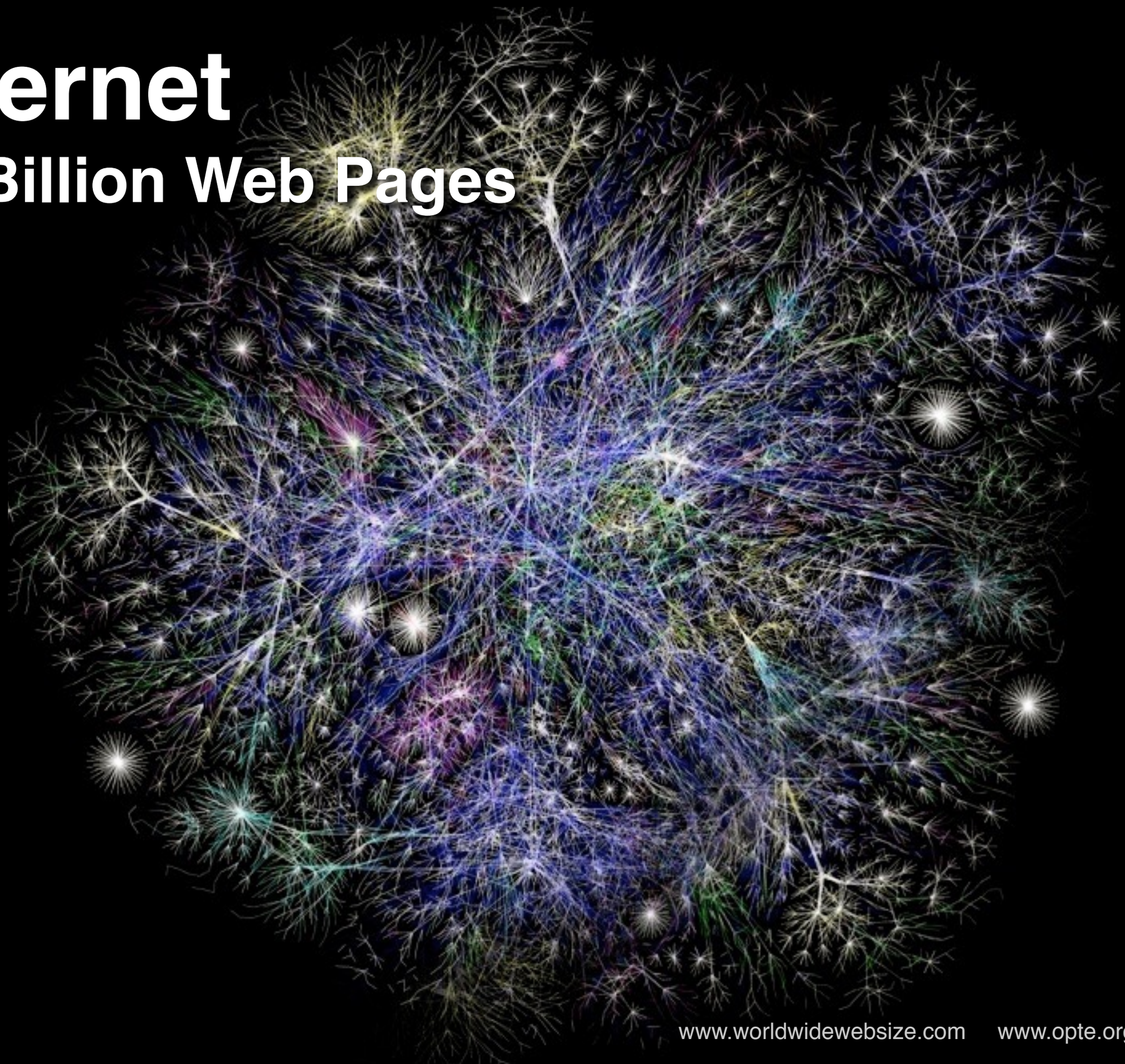
Office hours to be listed on course homepage
poloclub.gatech.edu/cse6242

Polo Club
of
DATA SCIENCE

We work with (really) large data.

Internet

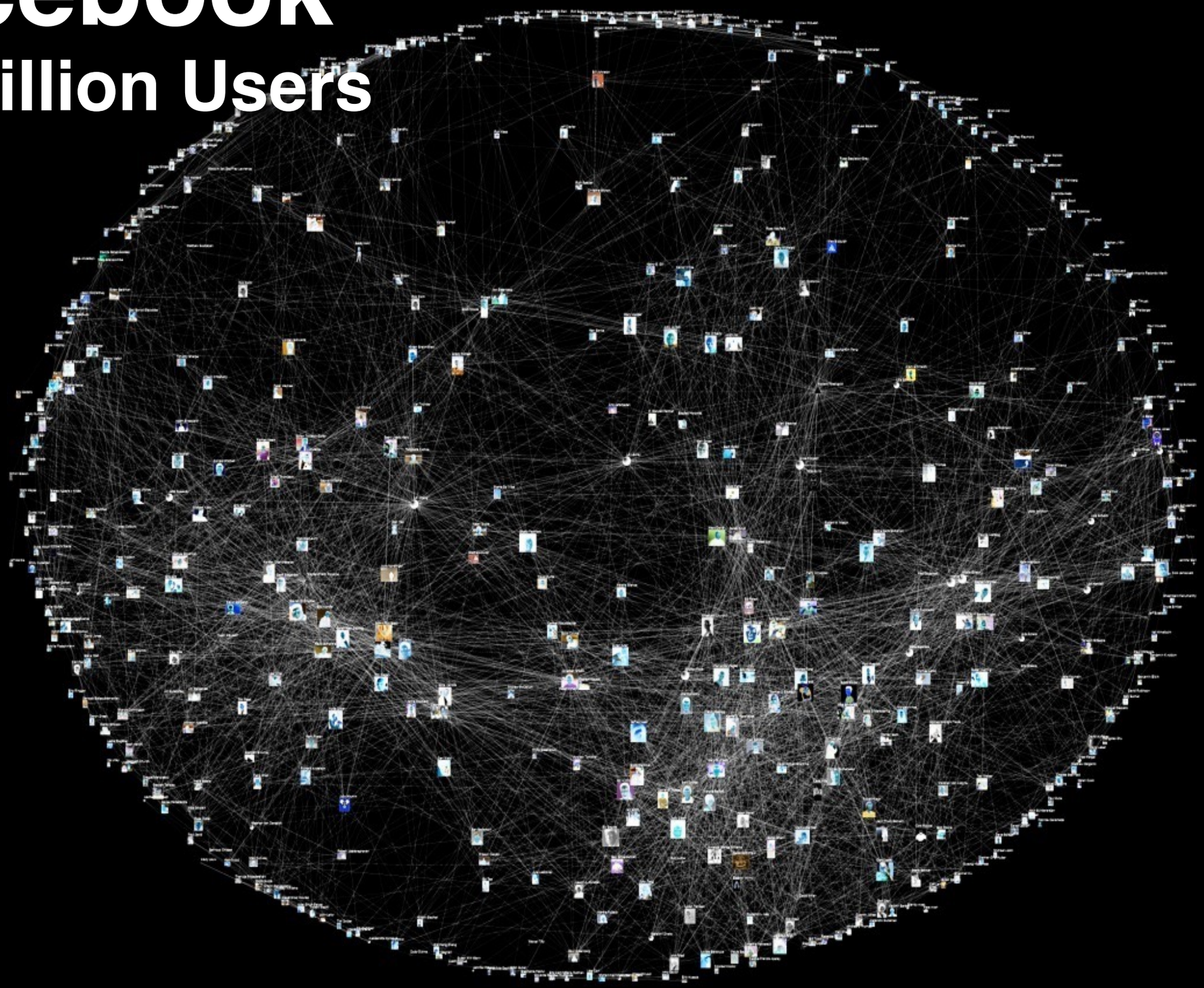
50 Billion Web Pages



www.worldwidewebsize.com www.opte.org

Facebook

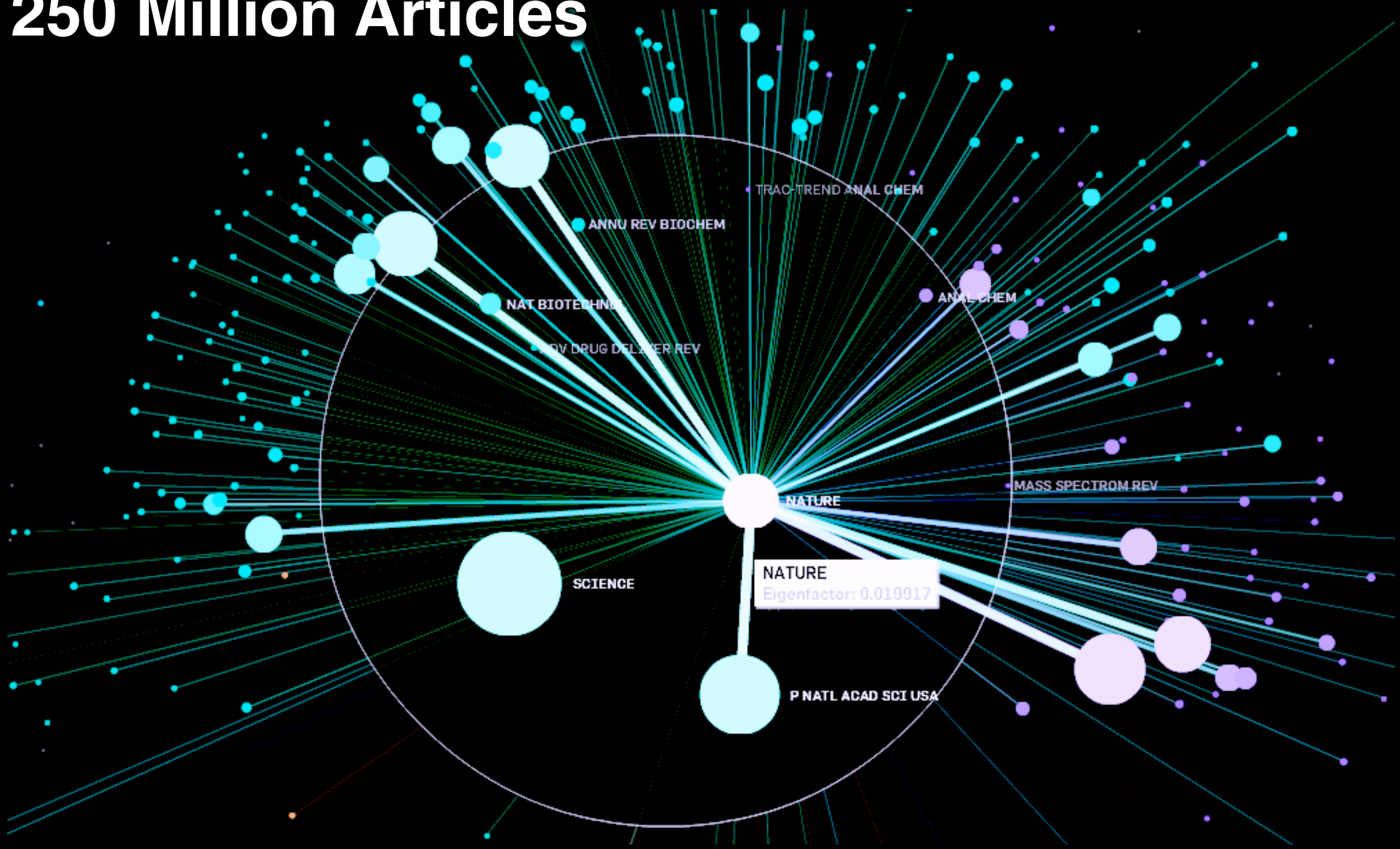
1.2 Billion Users



Modified from Marc_Smith, flickr

Citation Network

250 Million Articles



www.scirus.com/press/html/feb_2006.html#2 Modified from well-formed.eigenfactor.org

Many More

twitter

Who-follows-whom (500 million users)

amazon

Who-buys-what (120 million users)

at&t **cellphone network**

Who-calls-whom (100 million users)

Protein-protein interactions

200 million possible interactions in human genome

Sources: www.selectscience.net www.phonedog.com www.mediabistro.com www.practicalecommerce.com/

Large Networks We Analyzed

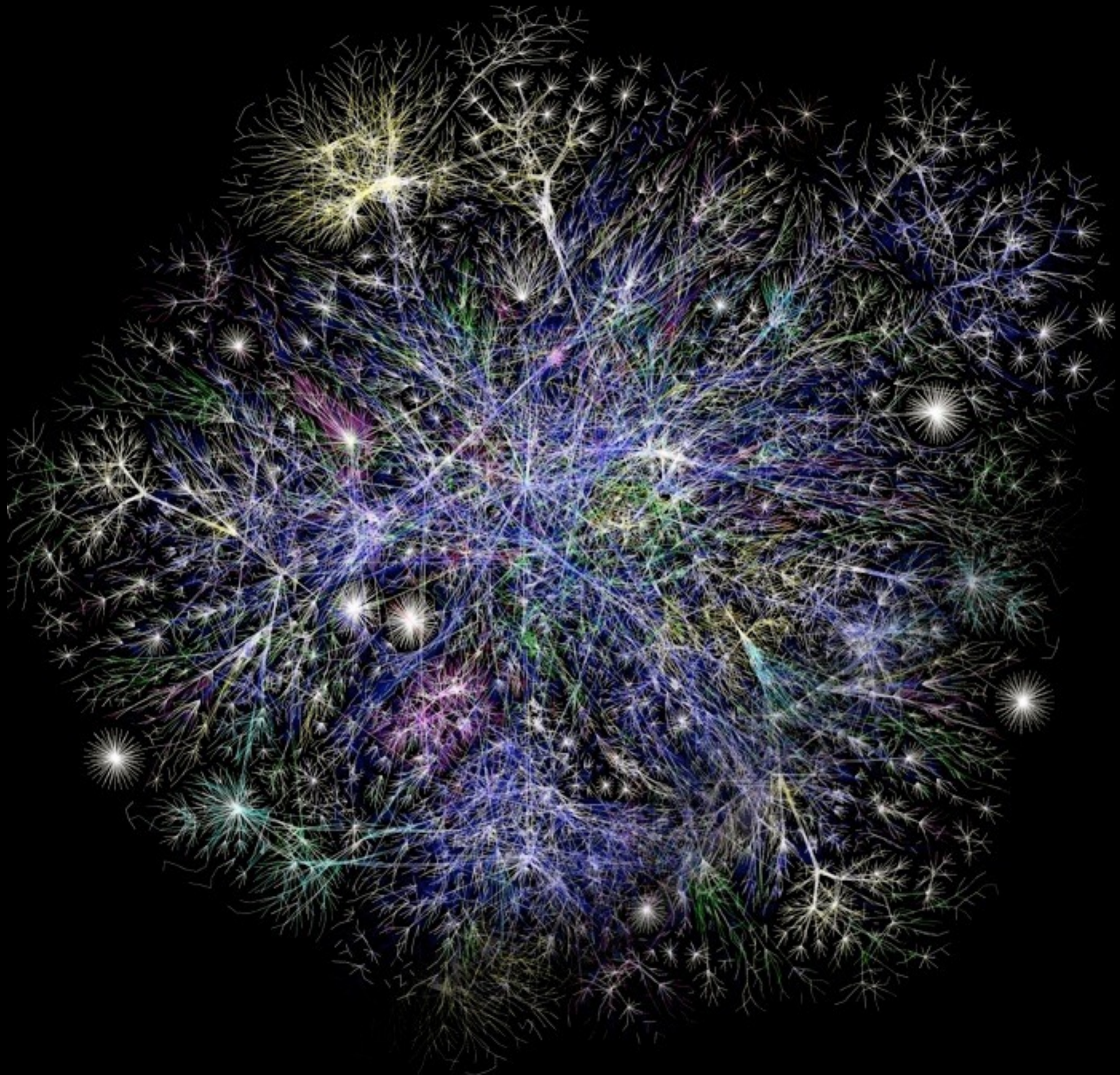
Graph	Nodes	Edges
YahooWeb	1.4 Billion	6 Billion
Symantec Machine-File Graph	1 Billion	**37 Billion**
Twitter	104 Million	3.7 Billion
Phone call network	30 Million	260 Million

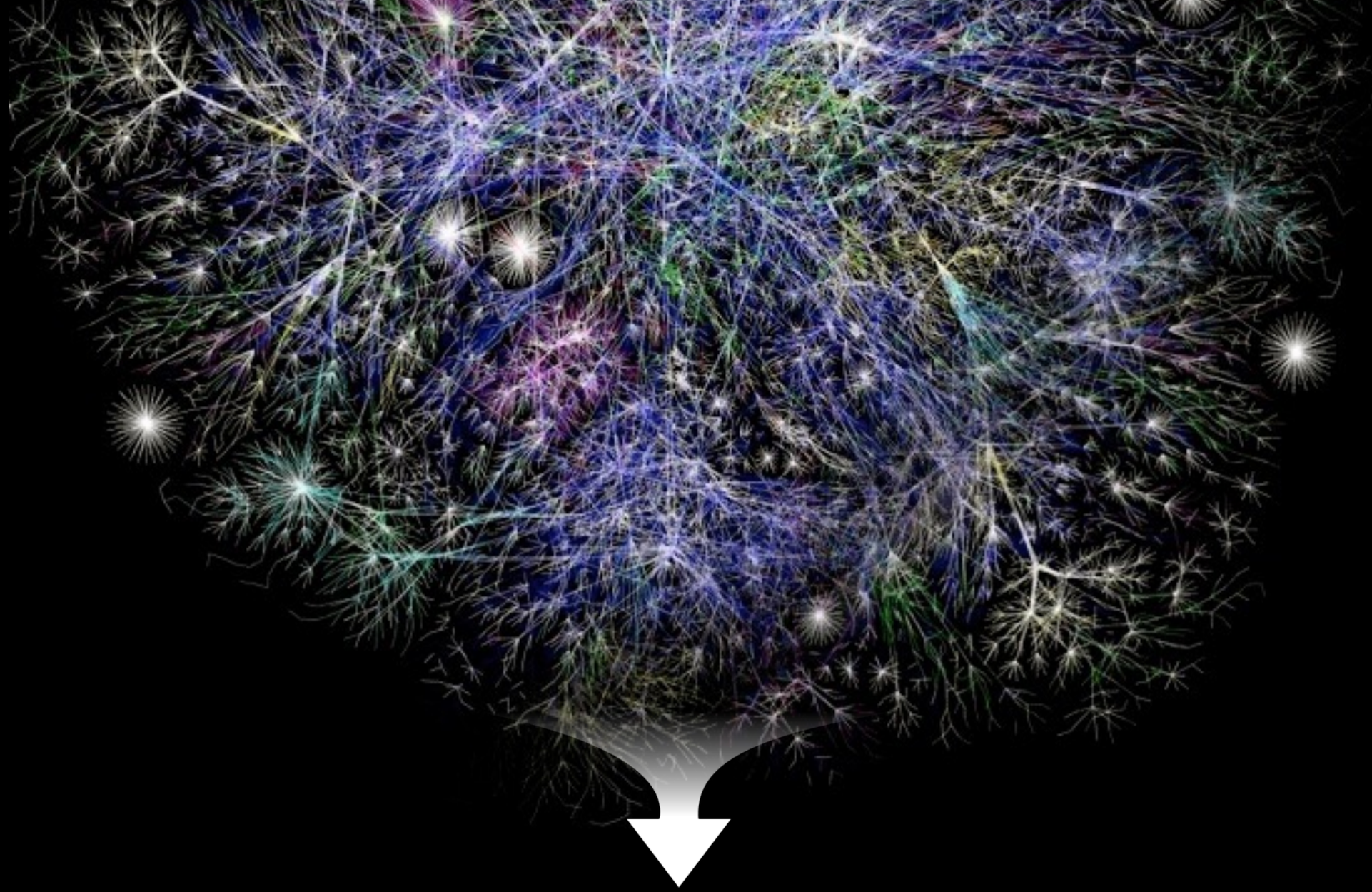
7

7±2

Number of items an average human holds in working memory

George Miller, 1956





7

Data

Insights

How to do that?

COMPUTATION
+
HUMAN INTUITION

How to do that?

COMPUTATION	INTERACTIVE VIS
Automatic	User-driven; iterative
Summarization, clustering, classification	Interaction, visualization
>Millions of nodes	Thousands of nodes

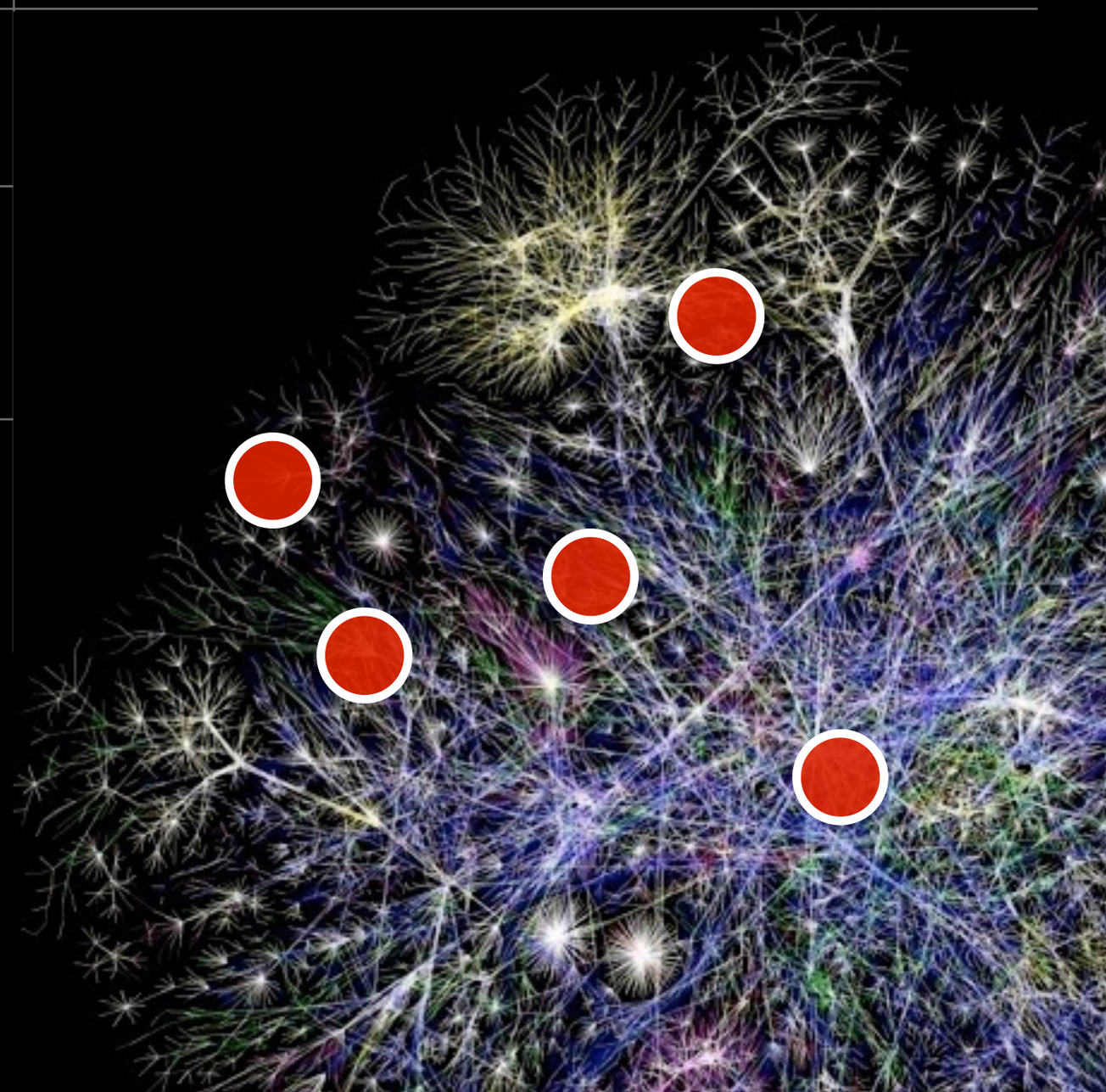
Both develop methods for making sense of network data

How to do that?

COMPUTATION	INTERACTIVE VIS
Automatic	
Summarization, clustering, classification	
>Millions of nodes	

How to do that?

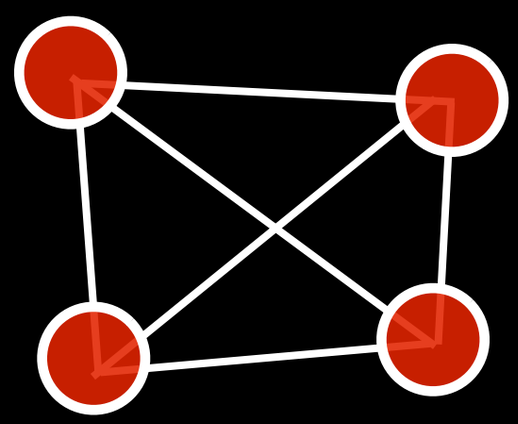
COMPUTATION	INTERACTIVE VIS
Automatic	
Summarization, clustering, classification	
>Millions of nodes	



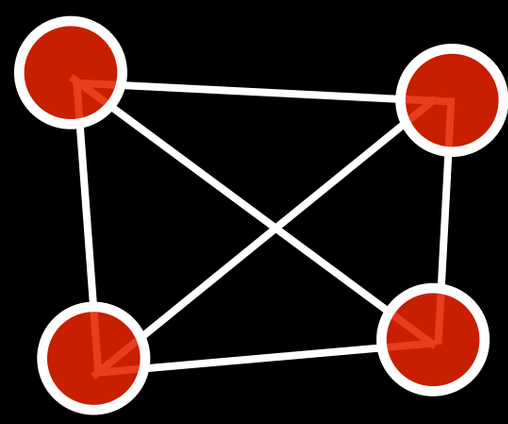
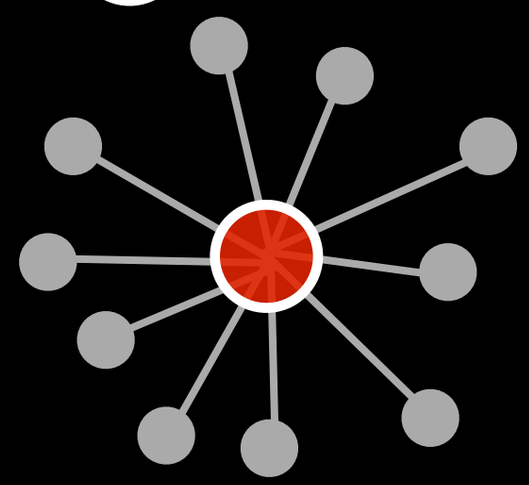
How to do that?

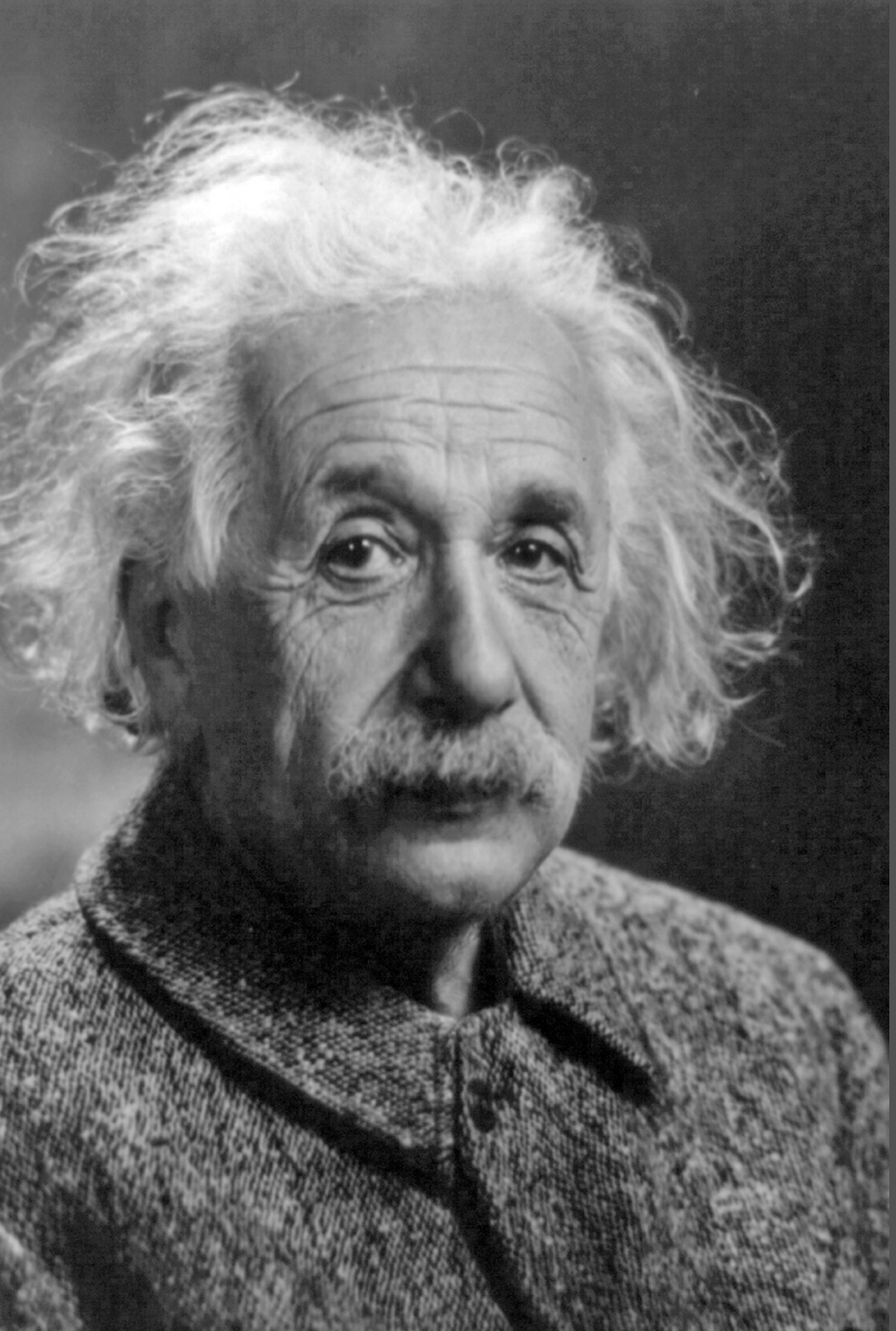
COMPUTATION	INTERACTIVE VIS
⋮	User-driven; iterative
	Interaction, visualization
	Thousands of nodes

How to do that?

COMPUTATION	INTERACTIVE VIS
	User-driven; iterative
	Interaction, visualization
	Thousands of nodes

How to do that?

COMPUTATION	INTERACTIVE VIS
	User-driven; iterative
	Interaction, visualization
	Thousands of nodes



"Computers are incredibly fast, accurate, and stupid.

Human beings are incredibly slow, inaccurate, and brilliant.

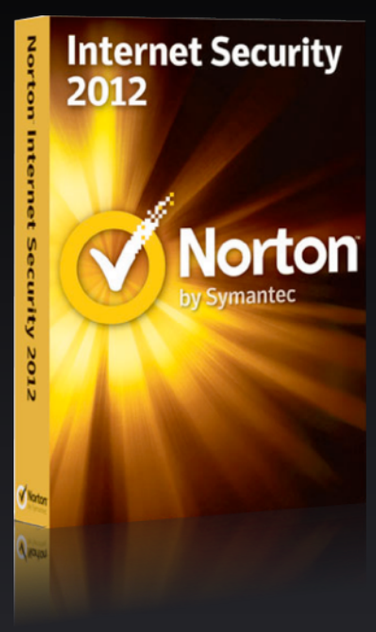
Together they are powerful beyond imagination."

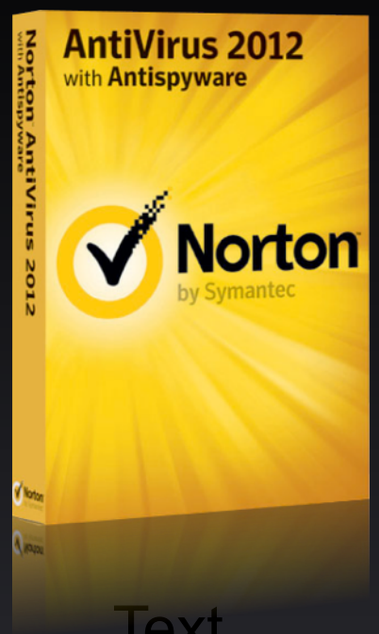
(Einstein might or might not have said this.)

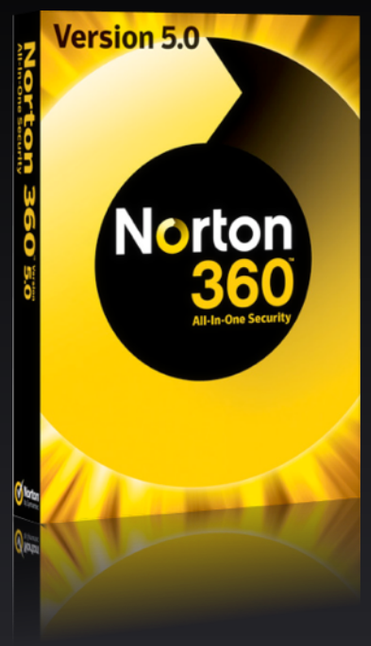
Our Approach for Big Data Analytics

DATA MINING	HCI Human-Computer Interaction
Automatic	User-driven; iterative
Summarization, clustering, classification	Interaction, visualization
>Millions of items	Thousands of items

Our research combines the
Best of Both Worlds







Text

Polonium

Patented with Symantec

Finds malware from **37 billion** file relationships

Serving **120 million** users worldwide

Published at SDM’11

MARCO

Detecting Fake Yelp Reviews

Best papers of SDM 2014

(top data mining conference)

Latent Gesture

CS undergraduate students:
Prem Saravanan, Samuel Clarke

won prestigious
Astronaut Scholarship

GIZMODO

WIRED

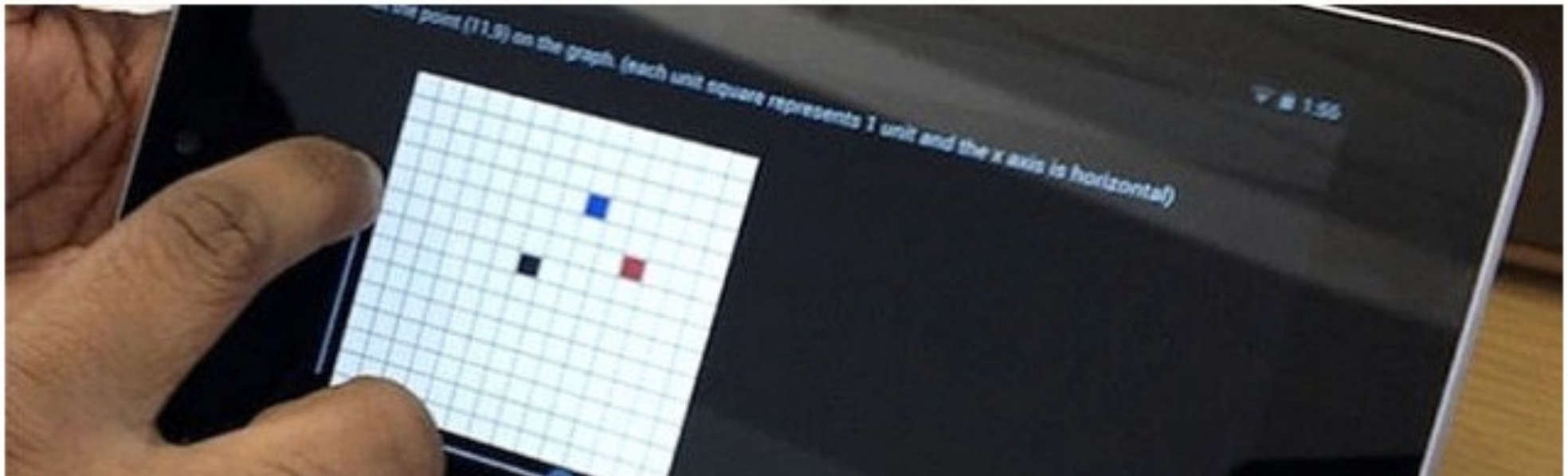
Your Touchscreen Usage Is So Unique It Can Be Used as a Password

engadget

Andrew Liszewski
Filed to: SECURITY 4/17/14 1:20pm

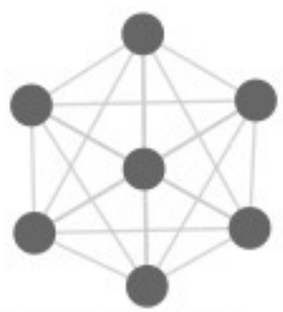
YAHOO!

7,508 2

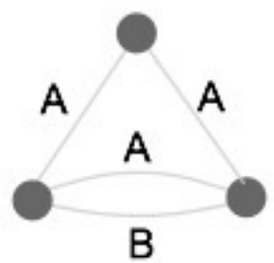


Insider Trading Detection

with Securities and Exchange Commission (SEC)

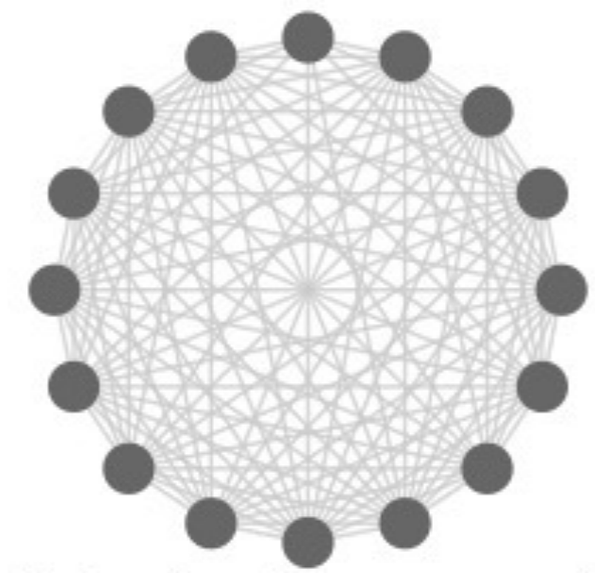


6-node Clique: each edge is an electrical company

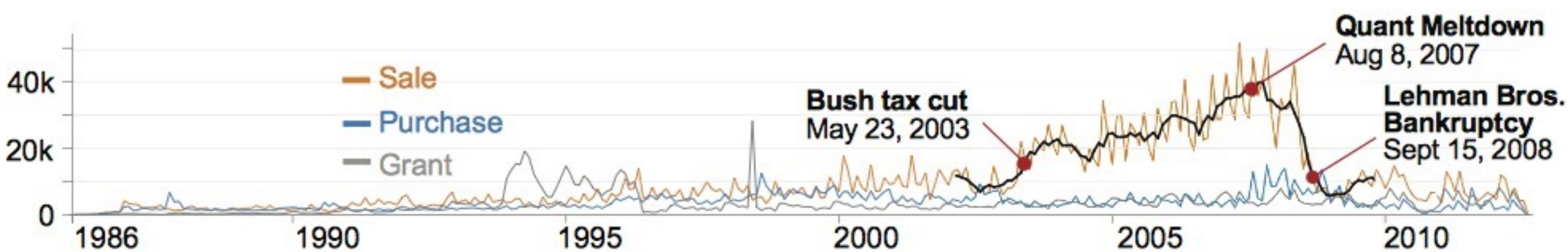


Triangle: company A: biotech; company B: medical supplies

Chain: each edge is an electrical utilities company



Each edge above corresponds to an Electrical Utilities Company



NetProbe

Auction Fraud Detection on eBay

THE WALL STREET JOURNAL.

MSNBC

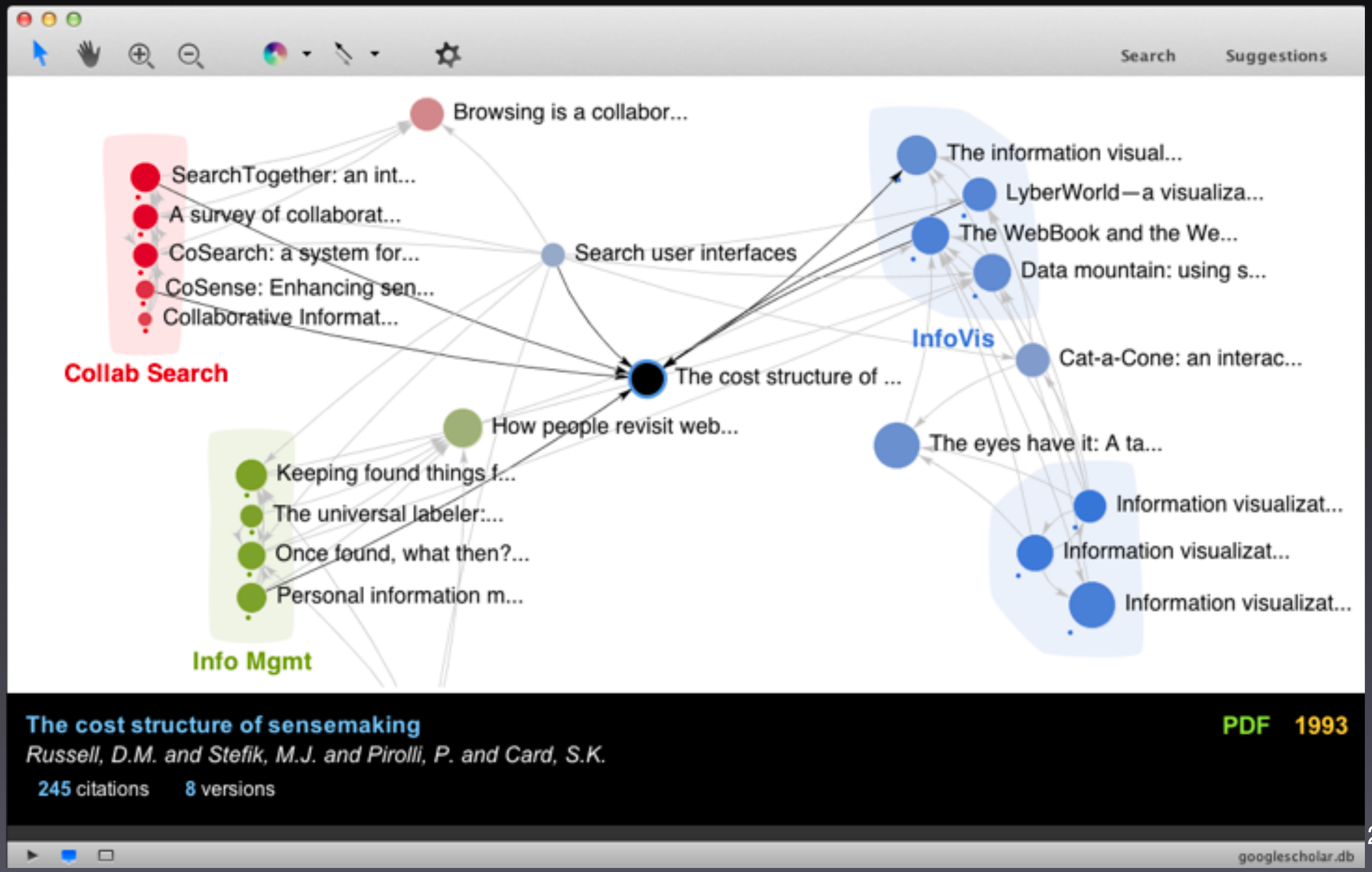
CNNMoney.com
A Service of CNN, Fortune & Money

USA TODAY

KDKA 2

Apolo: Machine Learning + Visualization

Explore million-scale graphs in real time



Logistics

Course homepage	poloclub.gatech.edu/cse6242/
Discussion, Q&A, find teammates	Piazza: goo.gl/9RYh5r or piazza.com/class/ijbuzug4gs43de
Assignment Submission	T-Square (Use Piazza for discussion)

Course Goals

- Learn **scalable** visual and computation techniques and tools, for typical data types
- Learn how to **combine** both kinds of methods (how they complement each other)
- Work on real data & problem
- Learn **practical** know-how (useful for jobs, research)
- Gain **breath** of knowledge

Schedule

See course homepage

poloclub.gatech.edu/cse6242/

Grading

- 4 homework assignments (50%)
  - End-to-end analysis
  - Techniques (computation and vis)
  - “Big data” tools, e.g, Hadoop, Spark, etc.
- Group project (50%) -- 4 to 5 people

From Previous Classes…

- Class projects turned into papers at top conferences (KDD, IUI, etc.)
- Projects as portfolio pieces on CV
- Increased job and internship opportunities
  - Former students sent me “thank you” notes

What Polo expects from you

- Actively participate throughout the course!
- Ask questions during class and on Piazza
- Help out whenever you can, e.g., help answer questions on Piazza
- Polo will reserve last 5-10min of every lecture for Q&A

PASSAGE: A Travel Safety Assistant With Safe Path Recommendations For Pedestrians

Matthew Garvey
College of Computing
Georgia Institute of Technology
Atlanta, GA 30332, USA
mgarvey6@gatech.edu

Meghna Natraj
College of Computing
Georgia Institute of Technology
Atlanta, GA 30332, USA
mnatraj@gatech.edu

Nilaksh Das
College of Computing
Georgia Institute of Technology
Atlanta, GA 30332, USA
nilakshdas@gatech.edu

Bhanu Verma
College of Computing
Georgia Institute of Technology
Atlanta, GA 30332, USA
bhanuverma@gatech.edu

Jiaxing Su
College of Engineering
Georgia Institute of Technology
Atlanta, GA 30332, USA
Jiaxingsu@gatech.edu

IUI'16 Poster paper

Abstract

Atlanta has consistently ranked as one of the most dangerous cities in America with over 2.5 million crime events recorded within the past six years. People who commute by walking are highly susceptible to crime here. To address this probl has developed a mobile application, PASSAG d crime data to find "safe pat Atlanta. user int

Autho Safe P Pulse

ACM H.5.2 Use ory

Int rgia ute of nology Hi ho th

made on social media

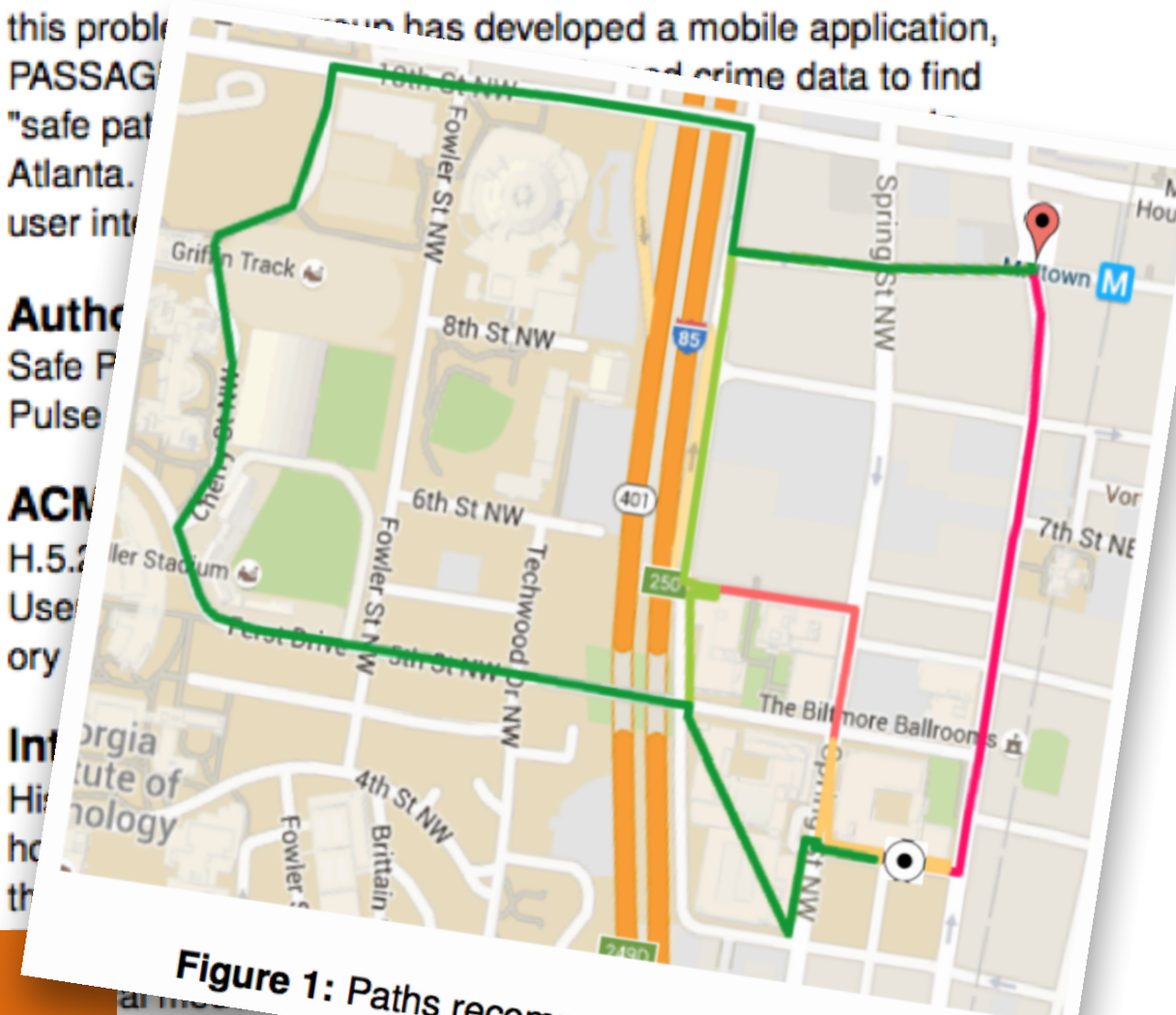


Figure 1: Paths recommended by PASSAGE

Aurigo: An Interactive Tour Planner for Personalized Itineraries

Alexandre Yahi*, Antoine Chassang*, Louis Raynaud*, Hugo Duthil*, Duen Horng (Polo) Chau
Georgia Institute of Technology
{alexandre.yahi, antoine.chassang, l.raynaud, hduthil, polo}@gatech.edu

ABSTRACT

Planning personalized tour itineraries is a complex and challenging task for both humans and computers. Doing it manually is time-consuming; approaching it as an optimization problem is computationally NP hard. We present Aurigo, a tour planning system combining a recommendation algorithm with interactive visualization to create personalized itineraries. This hybrid approach enables Aurigo to take into account both quantitative and qualitative preferences of the user. We conducted a within-subject study with 10 participants, which demonstrated that Aurigo helped them find points of interest quickly. Most participants chose Aurigo over Google Maps as their preferred tools to create personalized itineraries. Aurigo may be integrated into review websites or social networks, to leverage their databases of reviews and ratings and provide better itinerary recommendations.

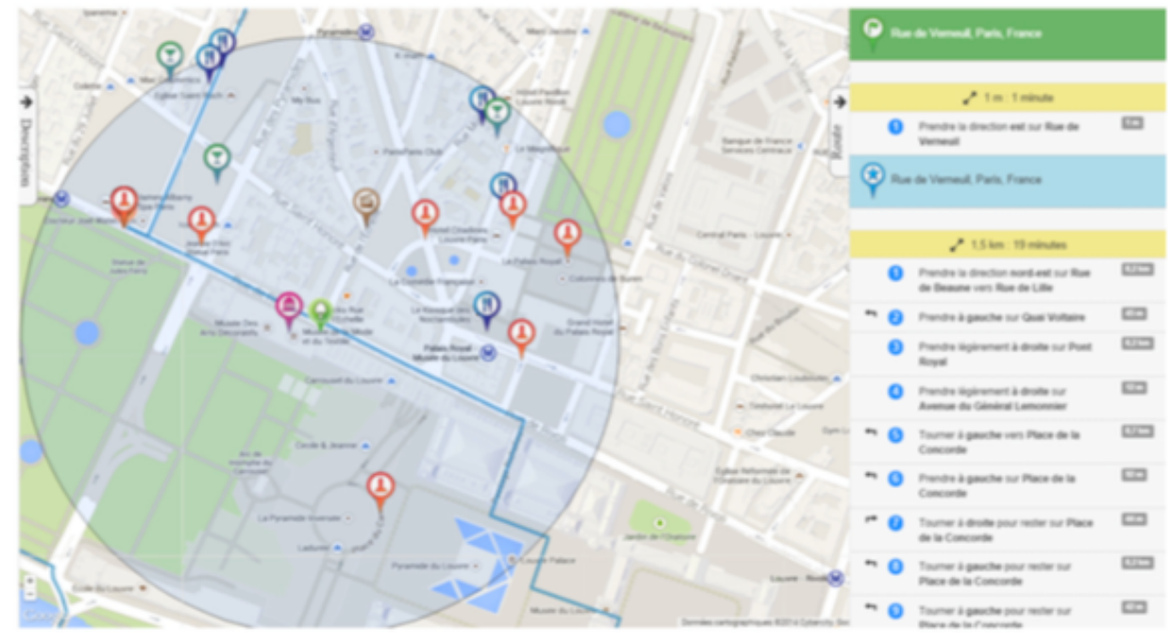
Author Keywords

User Interfaces; Visualization; Recommendation; Tour itinerary planning

ACM Classification Keywords

(e.g. HCI): User Interfaces

IUI'15 Full conference paper



ISPARK: Interactive Visual Analytics for Fire Incidents and Station Placement

Subhajit Das, Andrea McCarter, Joe Minieri, Nandita Damaraju, Sriram Padmanabhan, Duen Horng (Polo) Chau
Georgia Tech
Atlanta, GA, USA
{das, andream, jminieri, nandita, sriramp, polo}@gatech.edu

ABSTRACT

In support of helping to reduce the response time of firefighters, and thus deaths, injuries, and property loss due to fires, we introduce ISPARK. The ISPARK system determines where fire stations should be located, analyzes the primary causes of fires, the existing infrastructure, and response times, by using visualizations which show the GIS mapping of fire stations on a dashboard. Incidents and response times are shown as additional layers, with clustering of fire incidents to determine predicted fire station locations, forecasting of fire incidents using regression, causal, infrastructure, and personnel analysis, creating an interactive, multi-faceted method for locating fire stations. A comparison of urban and rural fire incident response times is another dimension of this study. We demonstrate ISPARK's usage and benefits using a publicly available dataset describing 300,000 fire incidents in the states of Massachusetts and Maine. ISPARK is generalizable to other geographic areas

Figure 1: Screenshot of ISPARK showing actual (pink) and predicted (green) fire station locations n Maine determined by our approach, using coordiates with actual driving distances from fire stations o actual fire incidents. Fire incidents are shown as small yellow dots. ISPARK reduces the average

KDD'15 Workshop paper

Categories and Subject Descriptors