http://poloclub.gatech.edu/cse6242

CSE6242

Data & Visual Analytics

Duen Horng (Polo) Chau

Associate Professor, College of Computing

Associate Director, MS Analytics

Machine Learning Area Leader, College of Computing

Georgia Tech

Google “Polo Chau” (only one in the world)

Bio CV Students Papers Teaching Funding Press Design

POLO CHAU

Legal name:
Duen Horng Chau

Associate Director, MS in Analytics
Associate Professor, School of Computational Science & Engineering
Machine Learning Area Leader, College of Computing
Georgia Tech

Linkedin Twitter Google Scholar YouTube

Admin: Carolyn Young
polo@gatech.edu
Office: CODA 1321

Financial Manager: Arlene Washington
www.cc.gatech.edu/~dchau
404-385-7682

POSITIONS

May 2014 -	Associate Director MS in Analytics, Georgia Tech
Aug 2018 -	Associate Professor School of Computational Science & Engineering, Georgia Tech
Aug 2012 - Aug 2018	Assistant Professor School of Computational Science & Engineering, Georgia Tech
Dec 2012 - Dec 2015	Adjunct Assistant Professor School of Interactive Computing, Georgia Tech

My research group website:

Polo Club of DATA SCIENCE

Scalable. Interactive. Interpretable.

Students (see all)
Austin Wright, ML PhD
Zijie (Jay) Wang, ML PhD
Haekyu Park, CS PhD
Scott Freitas, ML PhD
Nilaksh Das, CSE PhD
Fred Hohman, CSE PhD
Shang-Tse Chen, CS PhD
Minsuk (Brian) Kahng, CS PhD
Siwei (Bob) Li, CS UG
Ángel (Alex) Cabrera, CS UG
Joon Kim, CS UG
Sudeep Agarwal, CS UG
Kristina Marotta, CS OMS
Matthew Keezer, MS CS

Recent Alumni (see all)

How to address Polo?

Grammatically correct

Prof. Chau

Dr. Chau

Grammatically incorrect, but popular

Prof. Polo

Dr. Polo

Course Registration

This class room seats 300. If you are on the waitlist, please wait for seats to released (some students typically “drop” after today). I’ll also increase the cap to close to 300.

CSE 6242 A

236/250 seats filled

102/250 waitlist slots taken

CSE 6242 Q (distance-learning): 6 students

Course TAs **Be very very nice to them!**

Priyank Madria

Anmol Chhabria

Aastha Agrrawal

Haekyu Park

Hanna Kim

Sharmila Baskaran

Office hours and locations (TBD) on course homepage
poloclub.gatech.edu/cse6242

poloclub.gatech.edu

Polo Club | Human-Centered AI | Graph Vis & Mining | Cybersecurity | Social Good | Members | CSE6242 | IDEA | Sponsors | YouTube | Twitter | GitHub

Polo Club of DATA SCIENCE

Scalable. Interactive. Interpretable.

At Georgia Tech, we innovate **scalable, interactive, and interpretable** tools that amplify human's ability to understand and interact with billion-scale data and machine learning models. Our current research thrusts: human-centered AI (interpretable, fair, safe AI; adversarial ML); large graph visualization and mining; cybersecurity; and social good (health, energy).

Minsuk	Shang	Fred	Nilaksh	Haekyu	Scott	Jay	Austin	Anmol	Bob	Jonathan	Will	Omar
CS PhD	CS PhD	CSE PhD	CSE PhD	CS PhD	ML PhD	ML PhD	ML PhD	MS CSE	CS Undergrad	CS Undergrad	CS Undergrad	CS Undergrad

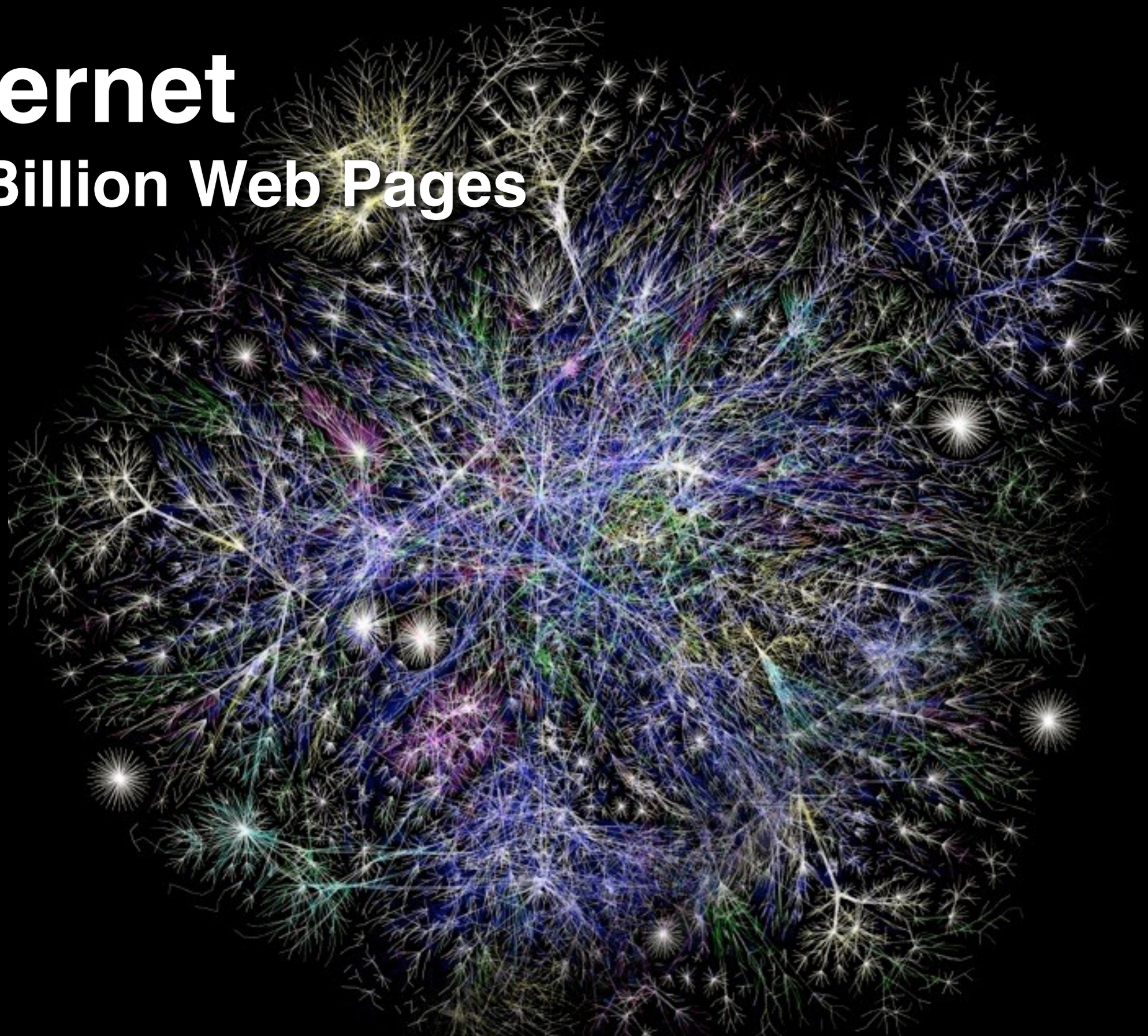
Polo
Associate Prof

Polo Club
of
DATA SCIENCE

We work with (really) large data.

Internet

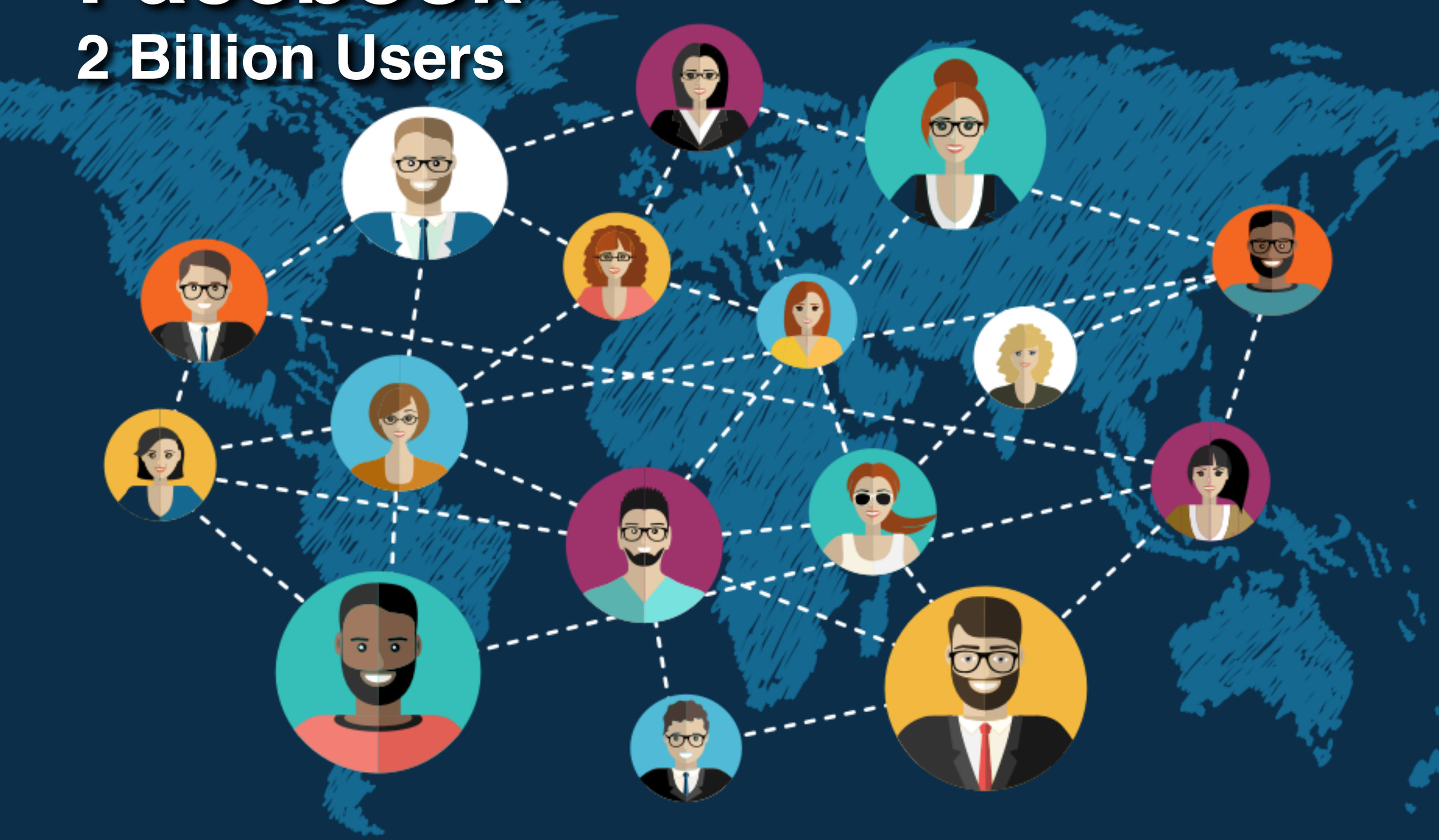
50 Billion Web Pages



www.worldwidewebsize.com www.opte.org

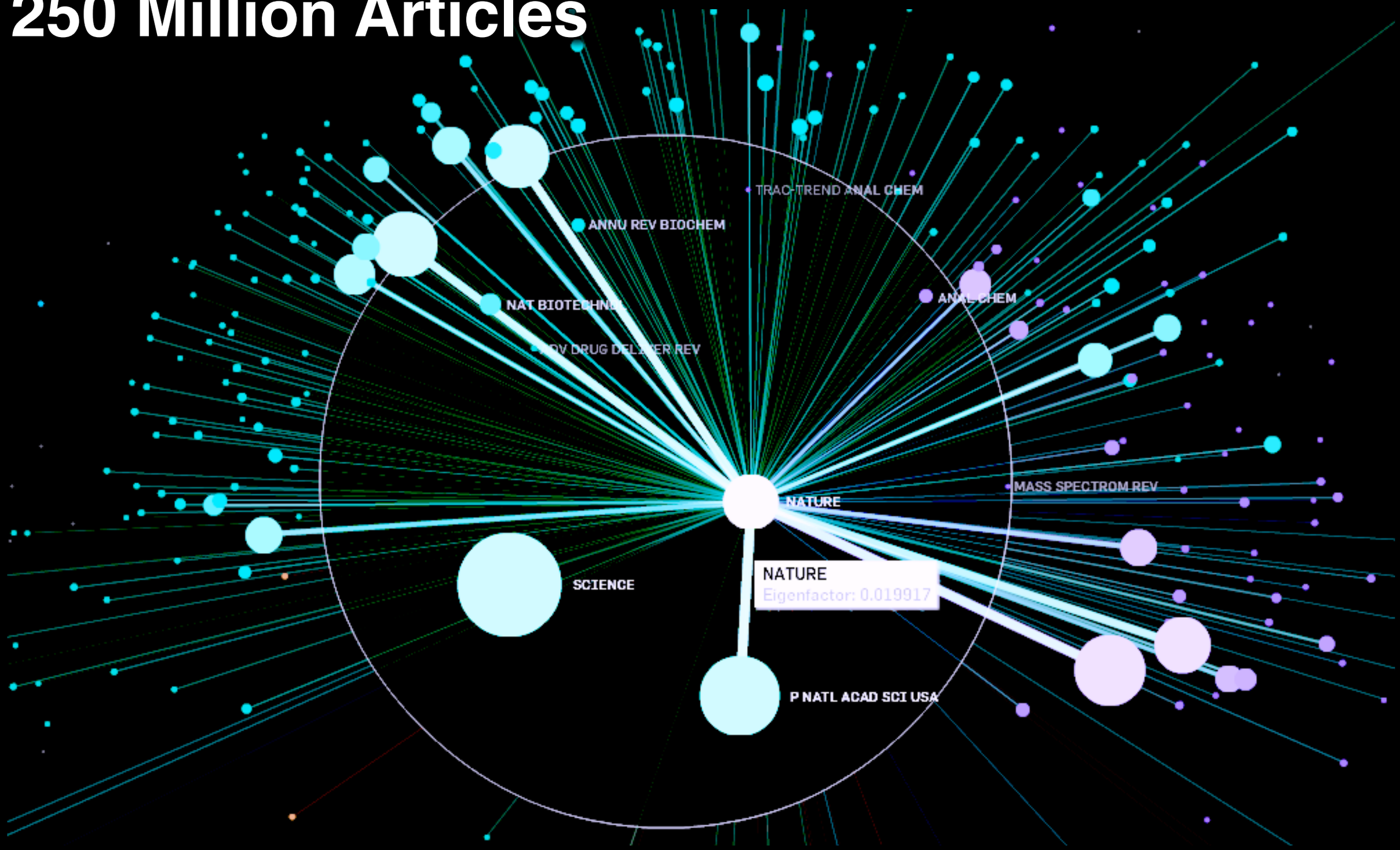
Facebook

2 Billion Users



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/

“Big Data” 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

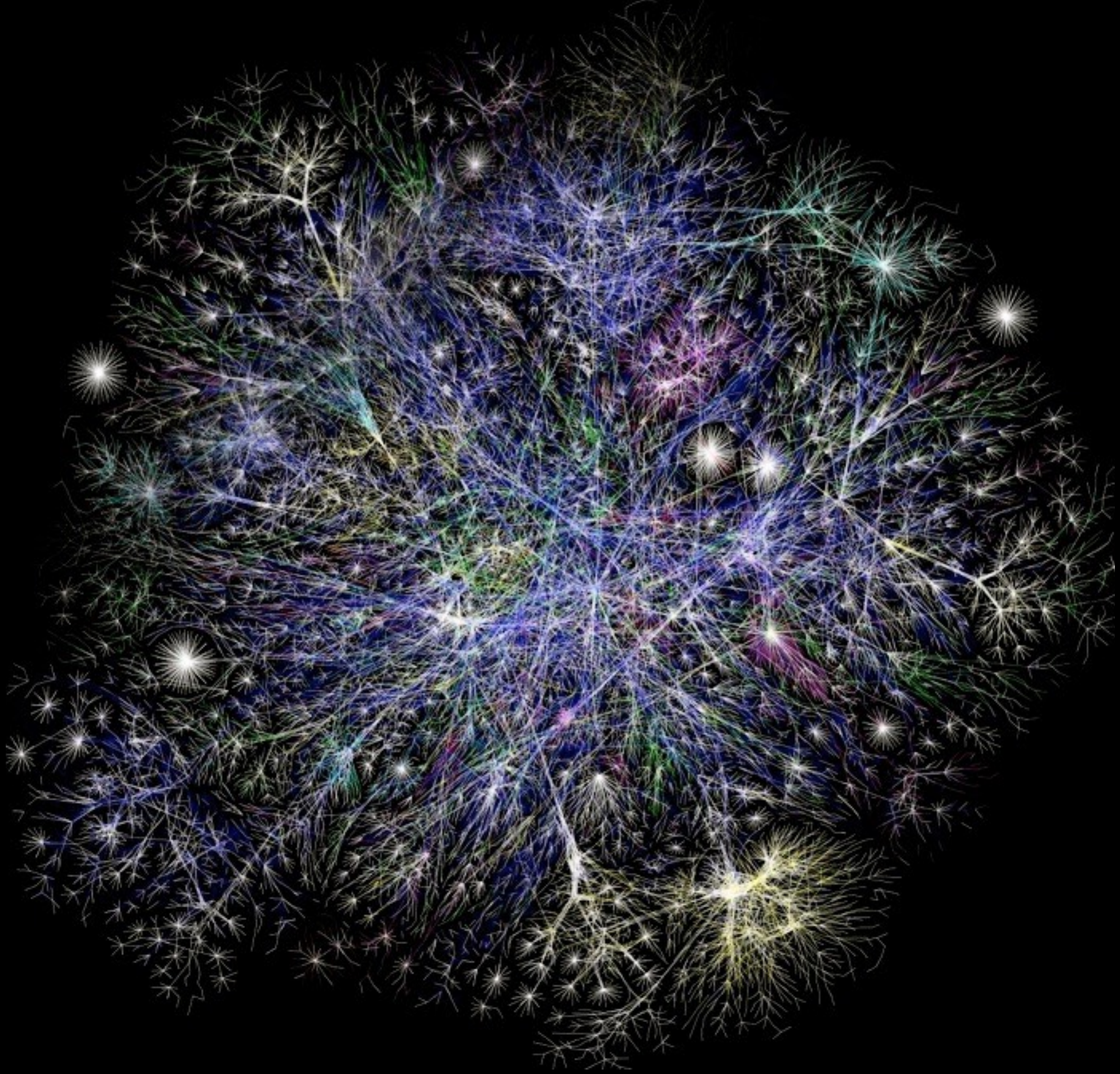
We also work with small data.
Small data also needs love.

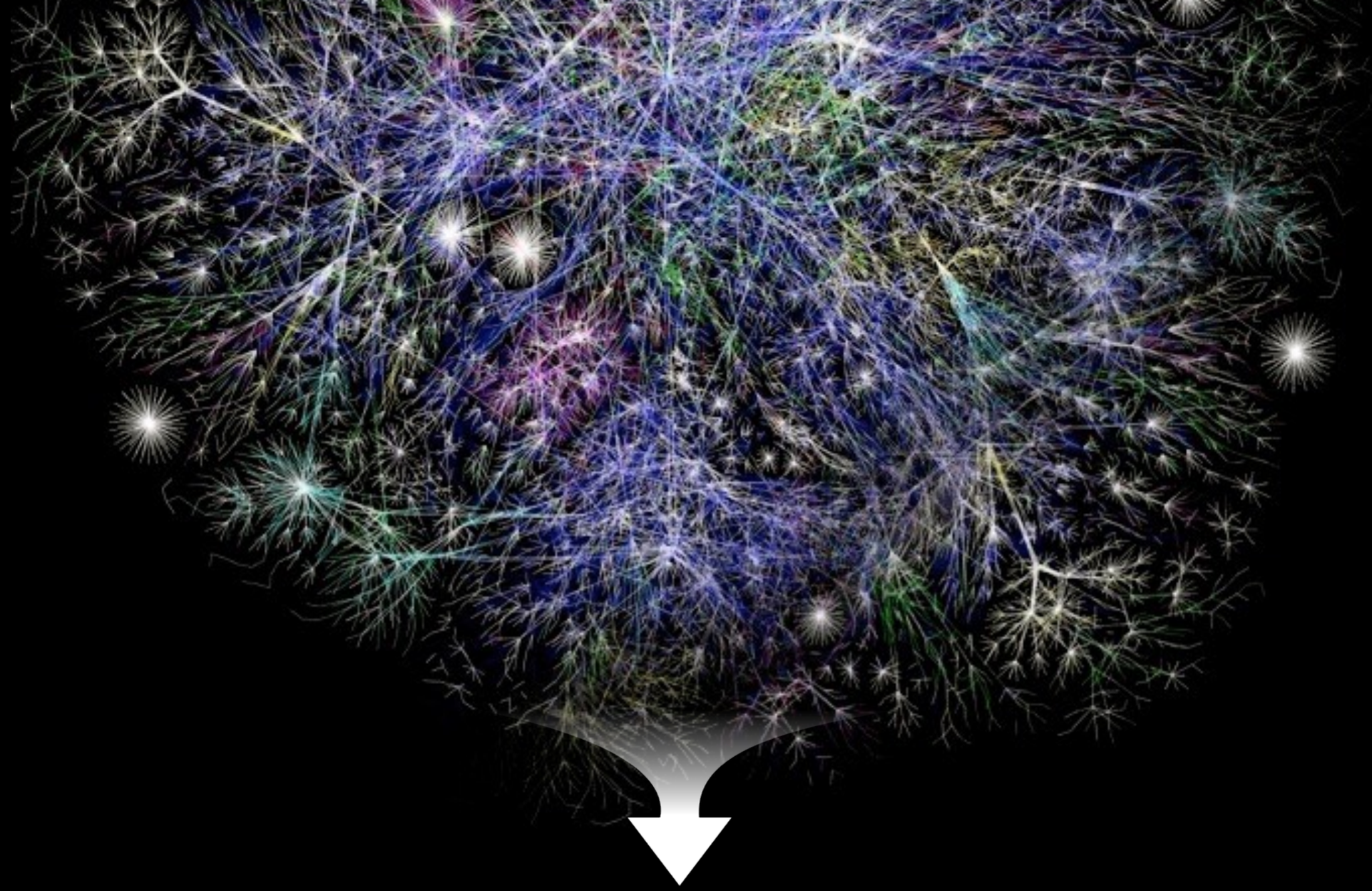
7

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

Or, to ride the AI wave…

ARTIFICIAL INTELLIGENCE

+

HUMAN INTELLIGENCE

How to do that?

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

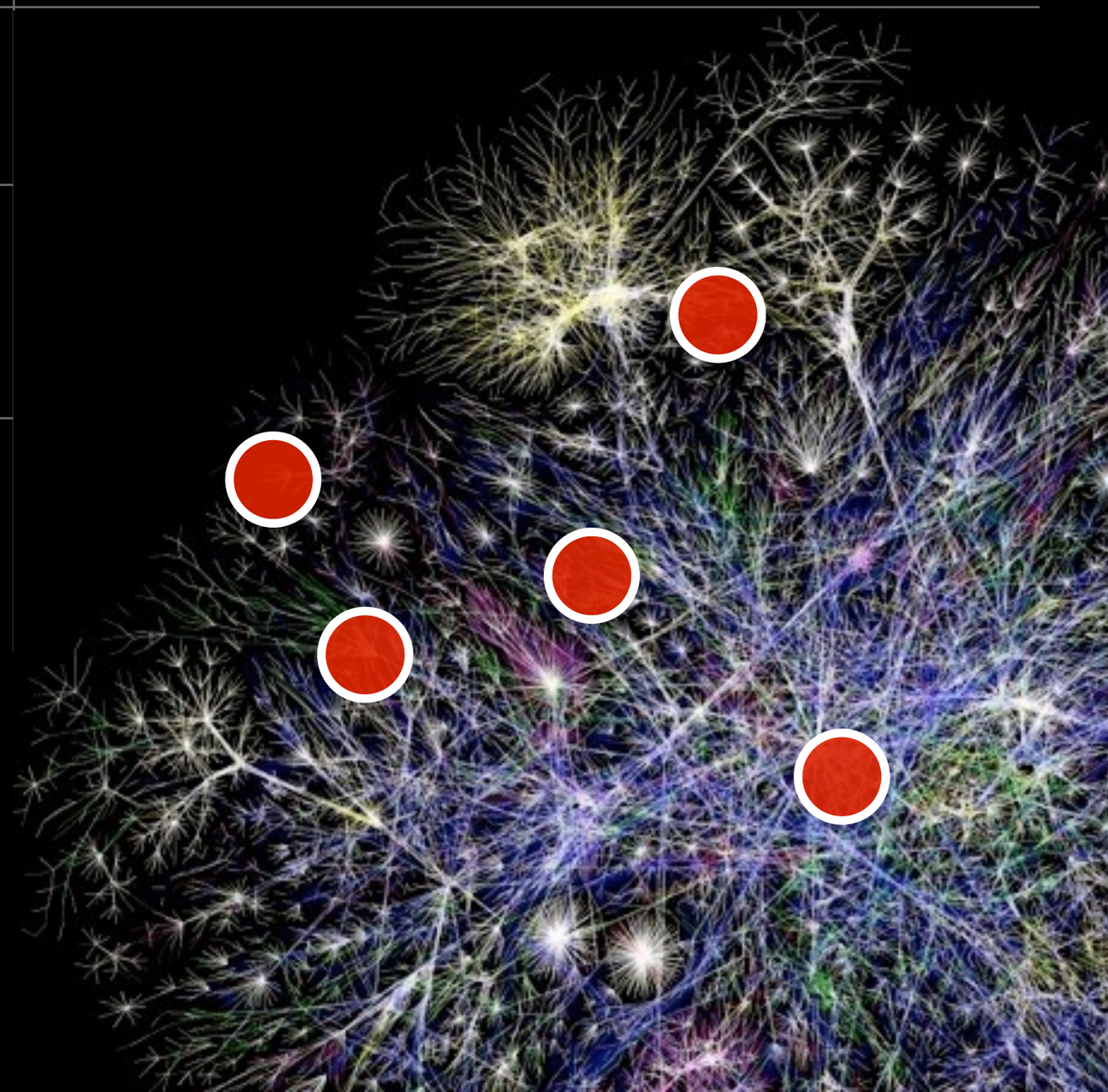
Both develop methods for making sense of network data

How to do that?

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

How to do that?

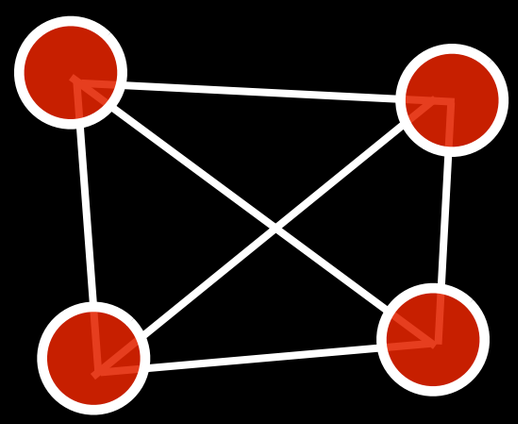
COMPUTATION	INTERACTIVE VIS
Automatic	
Summarization, clustering, classification	
>Millions of nodes	



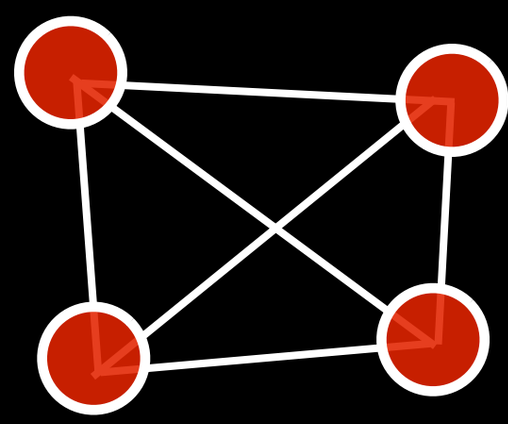
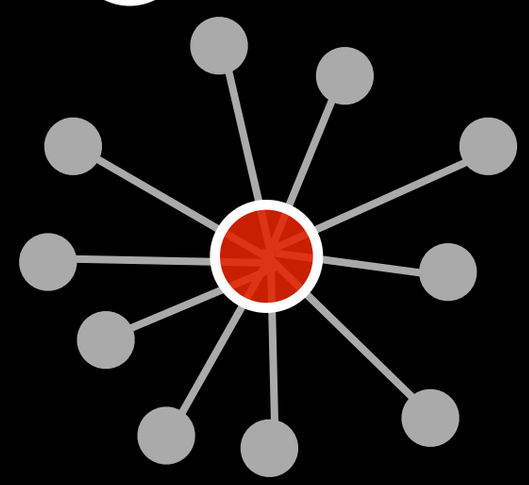
How to do that?

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

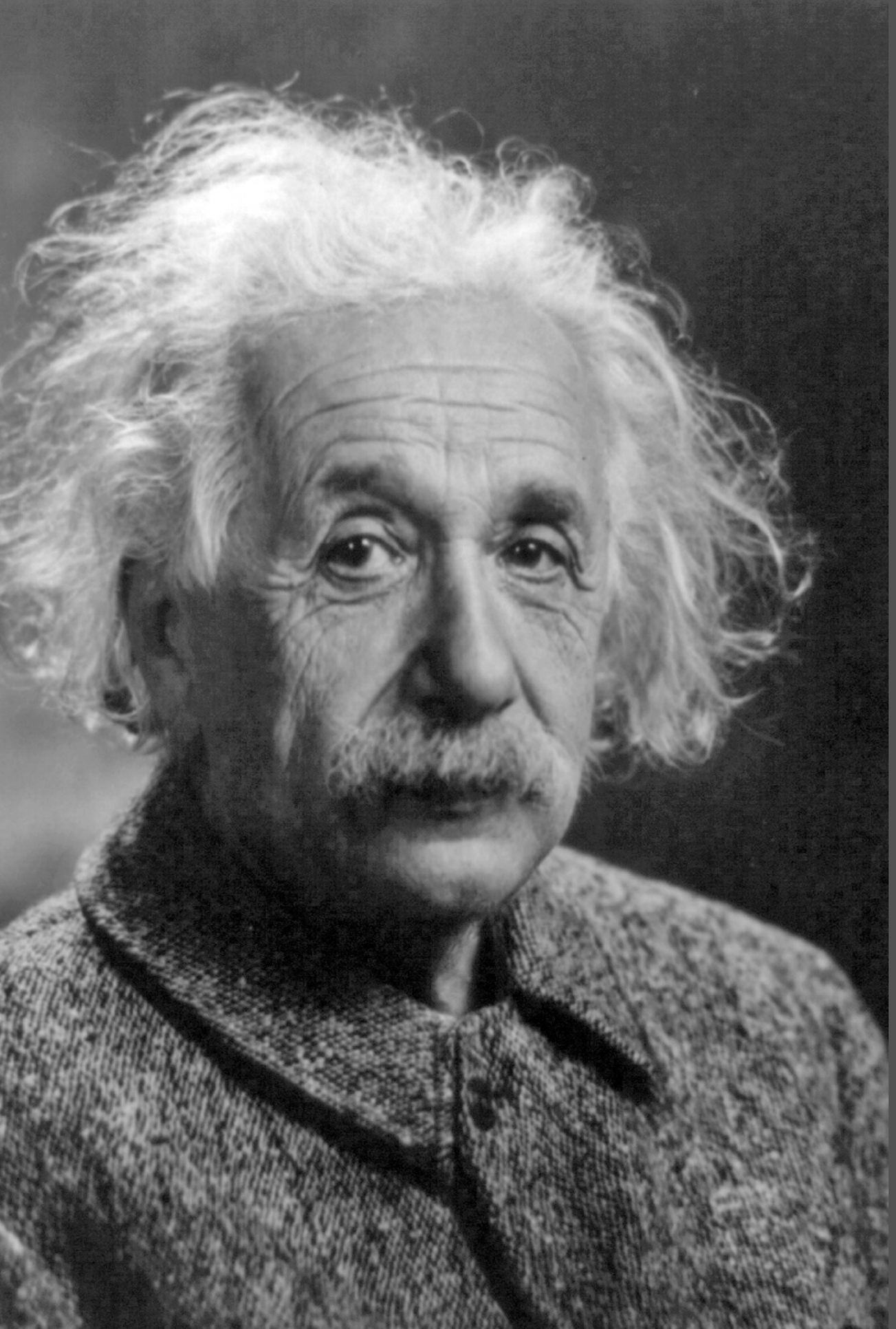
Our Approach for Big Data Analytics

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

Our mission & vision:

Scalable, interactive, usable tools for big data analytics



“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.)

Logistics

Course website
(policies, syllabus, schedule, etc.)

https://poloclub.github.io/cse6242-2019fall-campus/
(link also available on Canvas)

Discussion, Q&A, find teammates

Piazza
(link/tab available on Canvas)

Make sure you're in the right Piazza! (CSE-6242-O01, CSE-6242-OAN have their Piazza forums too)

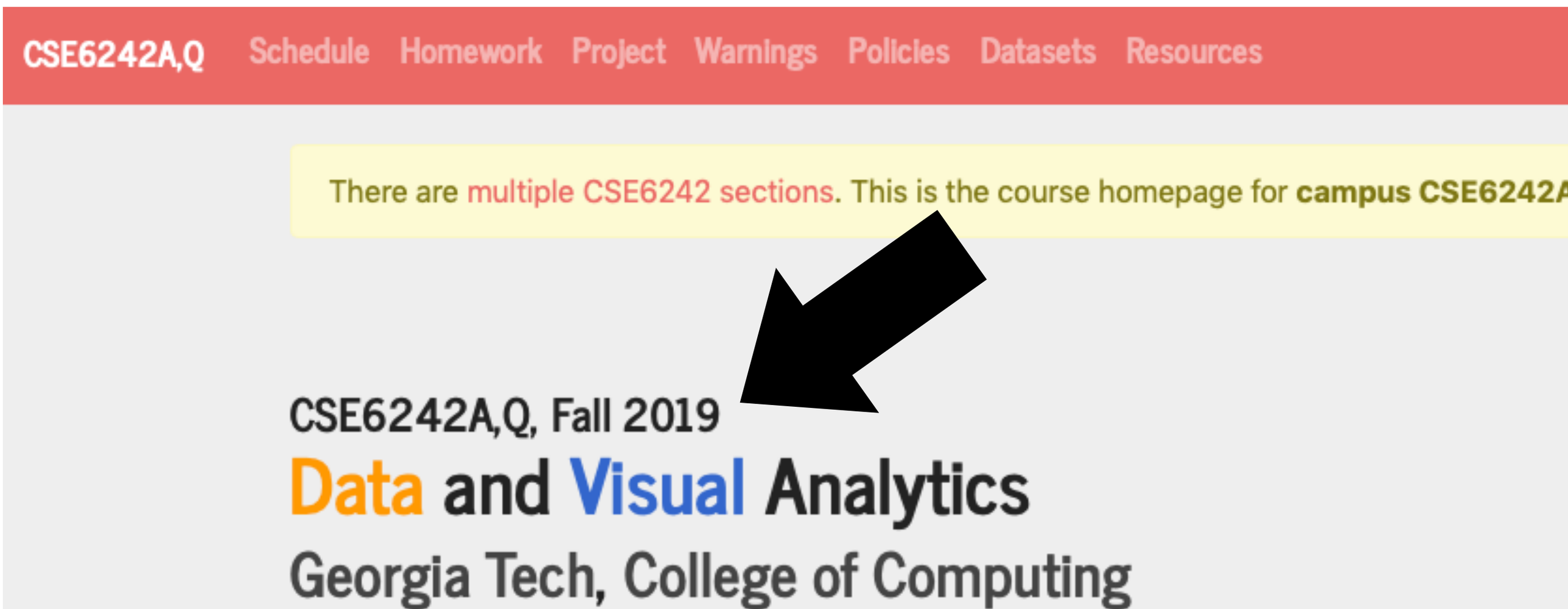
Assignment Submission

Canvas

Course Homepage

For syllabus, schedule, projects, datasets, etc.

If you Google “cse6242”, you will see many matches.
Make sure you click the correct site!



Join Piazza ASAP

(via canvas.gatech.edu)

Announcements and Discussion

We use Piazza for all announcements and discussion. Everyone must join this class's Piazza (link available on Canvas). Double check that you are joining the correct Piazza! There are multiple concurrent course sections with the same name and course number taking place, e.g., online for OMSA and OMSCS, and campus for Atlanta-based students.

The fastest way to get help with homework assignments is to post your questions on Piazza. That way, only our TAs and instructor can help, your peers can too.

If you prefer that your question addresses to only our TAs and the instructor, you can use the private post feature (i.e., check the "Individual Students(s) / Instructors(s)" radio box).

While we welcome everyone to share their experiences in tackling issues and helping each other out, but please do not post your answers, as that may affect the learning experience of your fellow classmates.

Important to join Piazza because…

- Polo will announce events related to this class and data science in general
  - Distinguished lectures
  - Seminars
  - Hackathons (**free food**, prizes)
  - Company recruitment events (**free food**, swag)

Course Goals

What is Data & Visual Analytics?

What is Data & Visual Analytics?

No formal definition!

What is Data & Visual Analytics?

No formal definition!

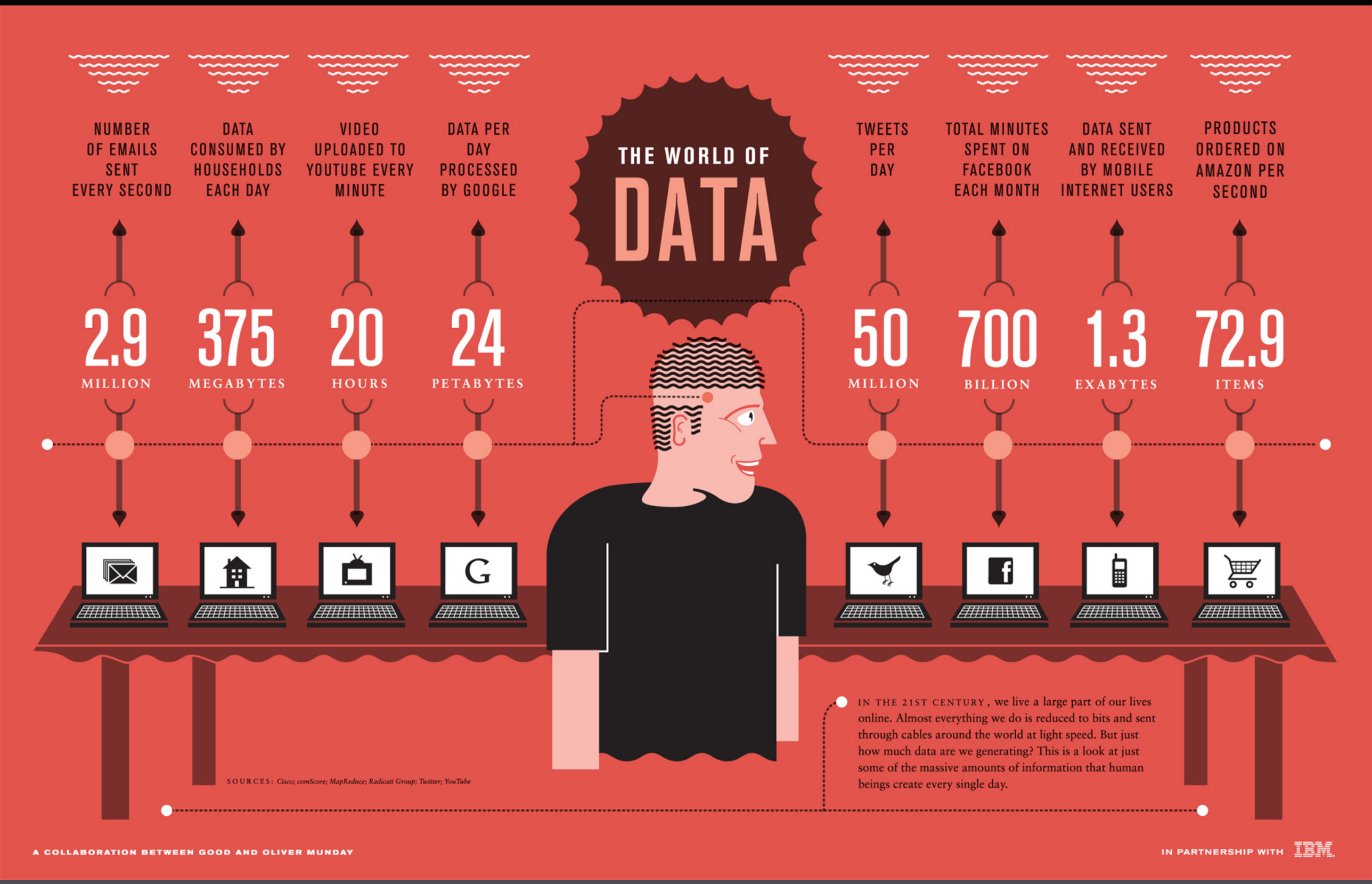
Polo's definition:
the *interdisciplinary* science of combining
computation techniques and
interactive visualization
to transform and model data to aid
discovery, decision making, etc.

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?



http://spanning.com/blog/choosing-between-storage-based-and-unlimited-storage-for-cloud-data-backup/

What is **big data**? Why care?

Many businesses are based on big data.

Search engines: rank webpages, predict what you're going to type

Advertisement: infer what you like, based on what your friends like; show relevant ads

E-commerce: recommends movies/products (e.g., Netflix, Amazon)

Health IT: patient records (EMR)

Finance

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.

Course Schedule

(Analytics Building Blocks)

Collection

Cleaning

Integration

Analysis

Visualization

Presentation

Dissemination

Building blocks. Not Rigid “Steps”.

Collection

Cleaning

Integration

Analysis

Visualization

Presentation

Dissemination

Can skip some

Can go back (two-way street)

- **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

Course Goals

- Learn **visual** and **computation** techniques and use them in **complementary** ways
- Gain a **breadth** of knowledge
- Learn **practical** know-how by working on **real data & problems**

Grading

- [50%] 4 homework assignments
  - End-to-end analysis
  - Techniques (computation and vis)
  - “Big data” tools, e.g., Hadoop, Spark, etc.
- [50%] Group project -- 4 to 6 people
- [Tentative **bonus points**] In-class pop quizzes
  - Each quiz is worth **1% course grade**
- **No exams**

Policies

On website; we go through them now

Grading, plagiarism, collaboration, late submission, and the **“warnings”** about the difficulty this course

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

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.

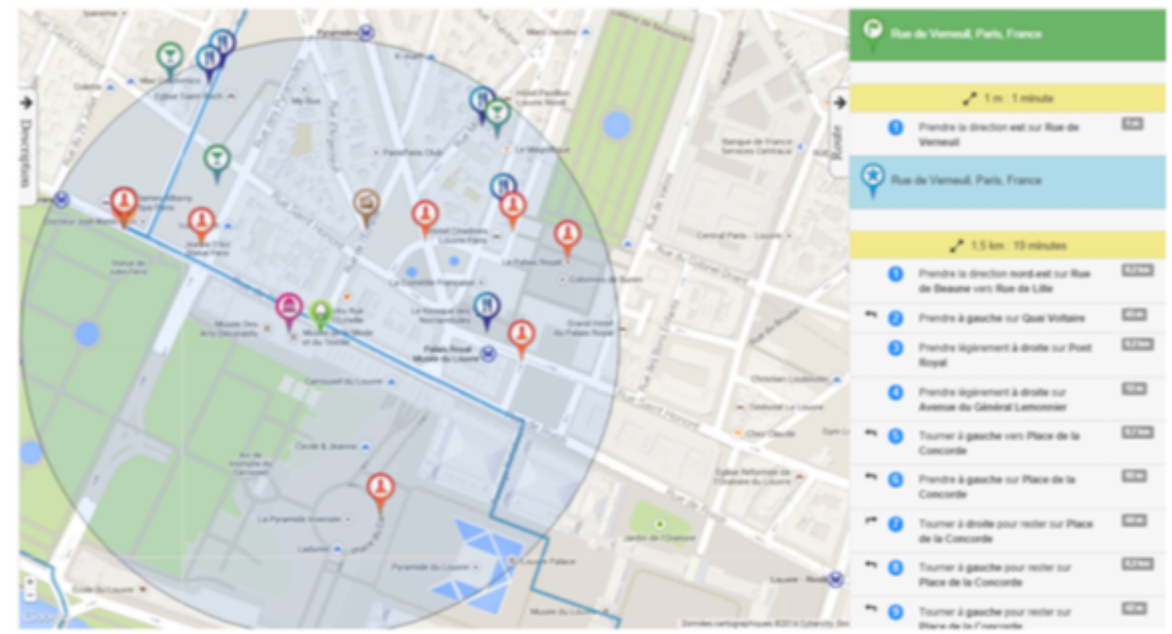
Author Keywords

User Interfaces; Visualization; Recommendation; Tour itinerary planning

ACM Classification Keywords

(e.g. HCI): User interfaces

IUI 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

KDD Workshop paper

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

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 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 crime data to find "safe pat Atlanta. user int

Autho Safe P Pulse

ACM H.5.2 Use ory

Int Hi ho th

made on social media pl

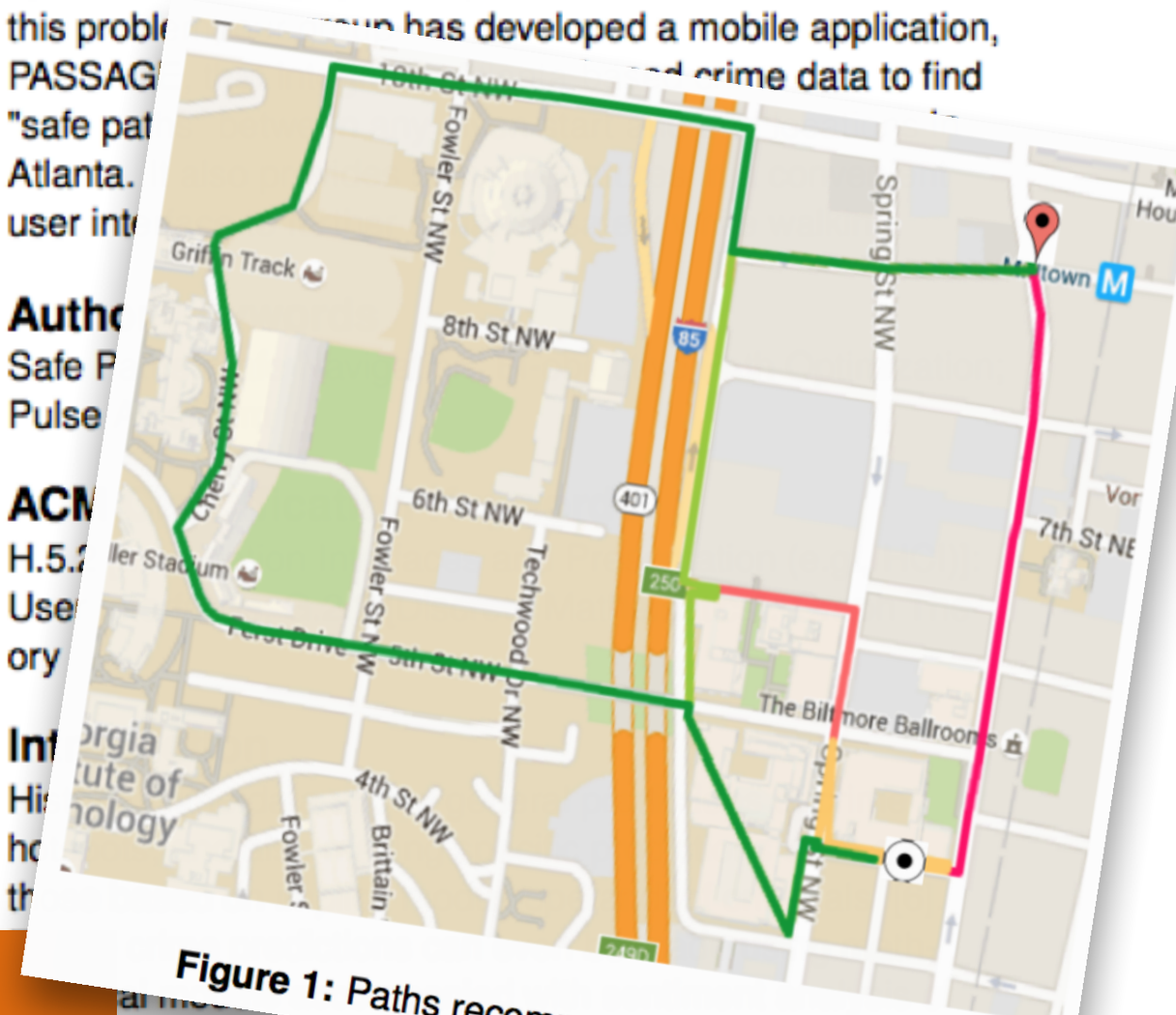


Figure 1: Paths recommended by PASSAGE

"I feel like the concepts from your class are like a ***rite of passage for an aspiring data scientist****. Assignments lead to a feelings of accomplishment and truly progressing in my area of passion."*

"I really get more intuition about how to ***deal with data with some powerful tools in HW3*** *[uses AWS]. That feeling is beyond description for me."*

"I would like to say thank you for your class! Thanks to the skills I got from the class and the project, ***I got the offer****."*

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 reserves last few minutes of every class for Q&A

FREE After-class Coffee 

- After (some) classes, Polo randomly selects 5 students (+2 volunteers) for **FREE** after-class coffee
- Polo's treat. You can order coffee, tea, pastries — whatever you want
- Very casual — you can ask me **ANYTHING**
- Will try doing this starting next week!