

poloclub.github.io/#cse6242

CSE6242 / CX4242

Data & Visual Analytics



Duen Horng (Polo) Chau

Professor, College of Computing

Associate Director, MS Analytics

Director of Industry Relations, The Institute for Data Engineering and Science

Associate Director of Corporate Relations, The Center for Machine Learning

Georgia Tech

Course Registration

Classroom has capacity for 305 students. **We will raise the number of seats to 305.**

If you have decided not to take this course, **please free up your seat ASAP**, so other students can get in.

If you are on the waitlist, please wait for seats to open up. **Enrollment changes a lot during first week of class.**

CSE 6242 A

233/235 seats filled

102 waitlist slots taken

CX 4242 A

54/55 seats filled

13 waitlist slots taken

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

Feyzi Can Eser

Yiwei Kuang

Ishan A Desai

Arya Mohan

Huayi Peng

Mahek Mishra

Zilu Zhu

Aniruddha Prashant Deshpande

Google "Polo Chau" (easy to find; should be first result)



Polo Chau | Legal name: Duen Horng Chau

Professor, [School of Computational Science & Engineering](#)
Associate Director, [MS in Analytics](#)
Director of Industry Relations, [The Institute for Data Engineering and Science](#)
Associate Director of Corporate Relations, [The Center for Machine Learning](#)
Georgia Tech

[LinkedIn](#) [Twitter](#) [Google Scholar](#) [YouTube](#)

Admin: [Kevelyn Cormier](#) Financial Managers: [Holly Rush](#)
[http://www.cc.gatech.edu/~dchau](#)

If you are interested in joining my group.

Welcome to connect on LinkedIn!

POSITIONS

My research group website:



Polo Club
of
DATA SCIENCE

Scalable. Interactive.
Interpretable.

Students (see all)

- [Zijie \(Jay\) Wang](#), ML PhD
- [Austin Wright](#), ML PhD
- [Seongmin Lee](#), CS PhD
- [Anthony Peng](#), CS PhD

How to address Polo?

Grammatically correct

Prof. Chau

Dr. Chau

Grammatically incorrect, but popular

Prof. Polo

Dr. Polo



**The course focuses on
working with large datasets.**

(Also the focus of Polo's research group)

Polo Club of Data Science

 poloclub.github.io

AI

ARTIFICIAL
INTELLIGENCE

+

HI

HUMAN
INTELLIGENCE

Scalable, interactive, interpretable tools to make sense of complex large-scale datasets and models



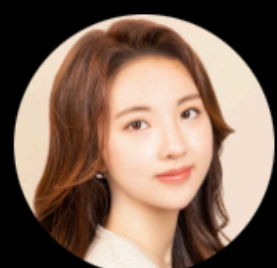
Haekyu
CS PhD



Jay
ML PhD



Austin
ML PhD



Seongmin
CS PhD



Ben
ML PhD



Anthony
CS PhD



Matthew
ML PhD



Alec
ML PhD



Mansi
CS Master



Harsha
CS Masters



Pratham
CS Undergrad



David
CS Undergrad



Justin
CS Masters



Sri
CS Undergrad



Aishwarya
CS Undergrad




Polo
Prof



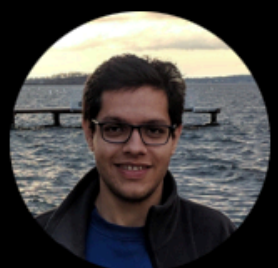
Fred
 Research Scientist, Apple




Nilaksh
 Applied Scientist, Amazon AWS




Scott
 Senior Applied Scientist,




Rahul
 Applied Scientist, AWS AI




Sivapriya
 AI Research, JPMorgan



Chakri





Omar
 CS PhD, Stanford



Megan



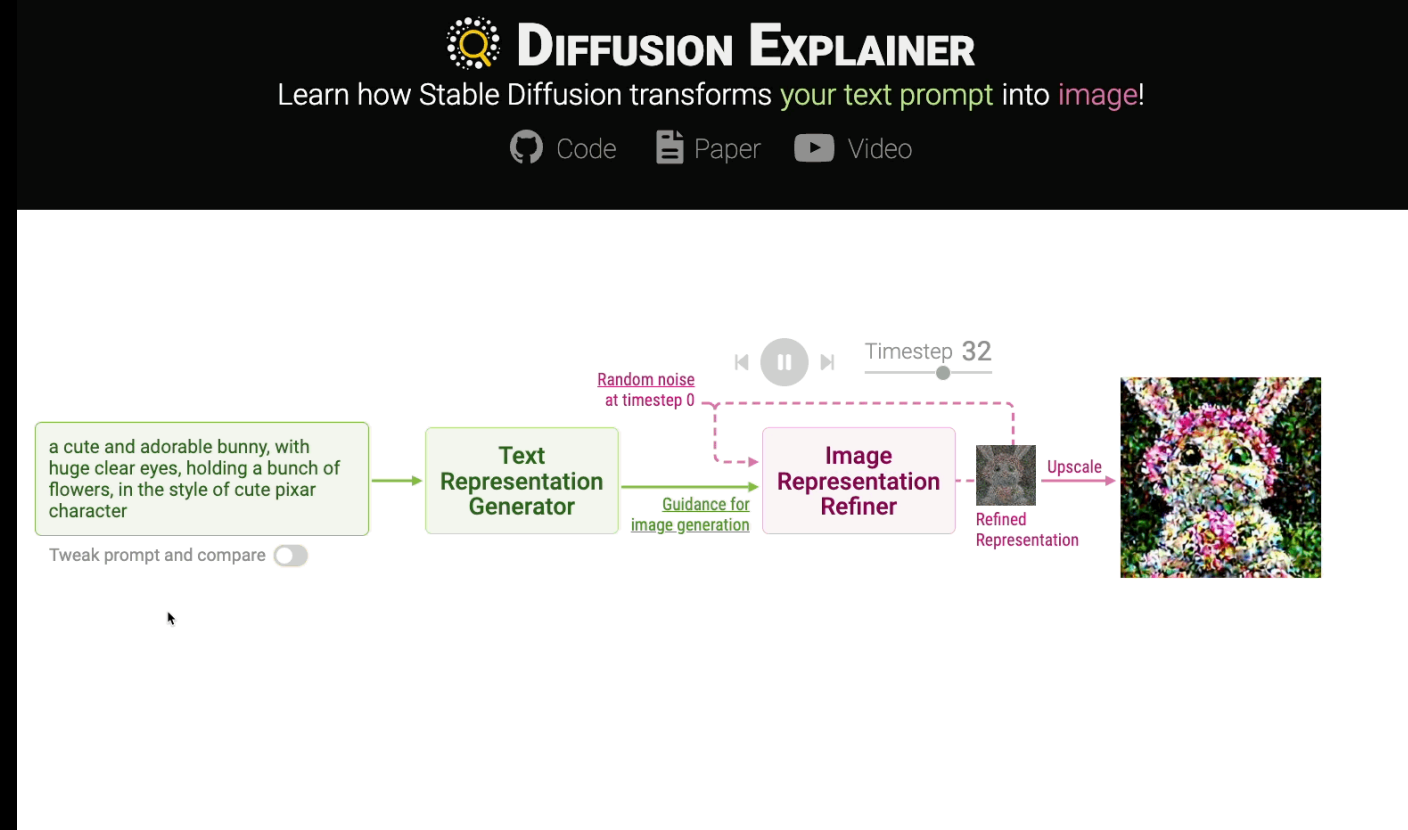

Alex
 Software Engineer, Goldman



Kevin
 ML PhD, CMU

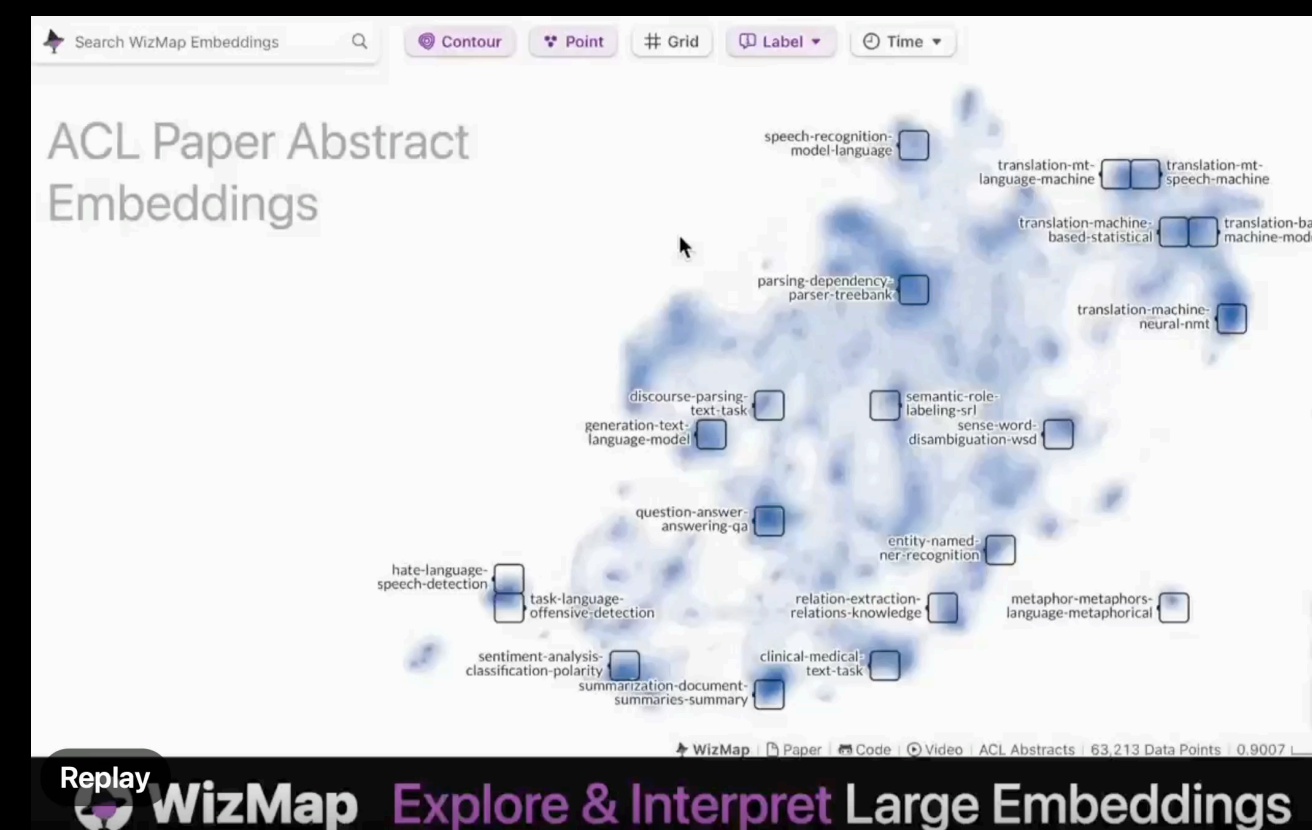
HUMAN CENTERED AI FOR EVERYONE

 poloclub.github.io



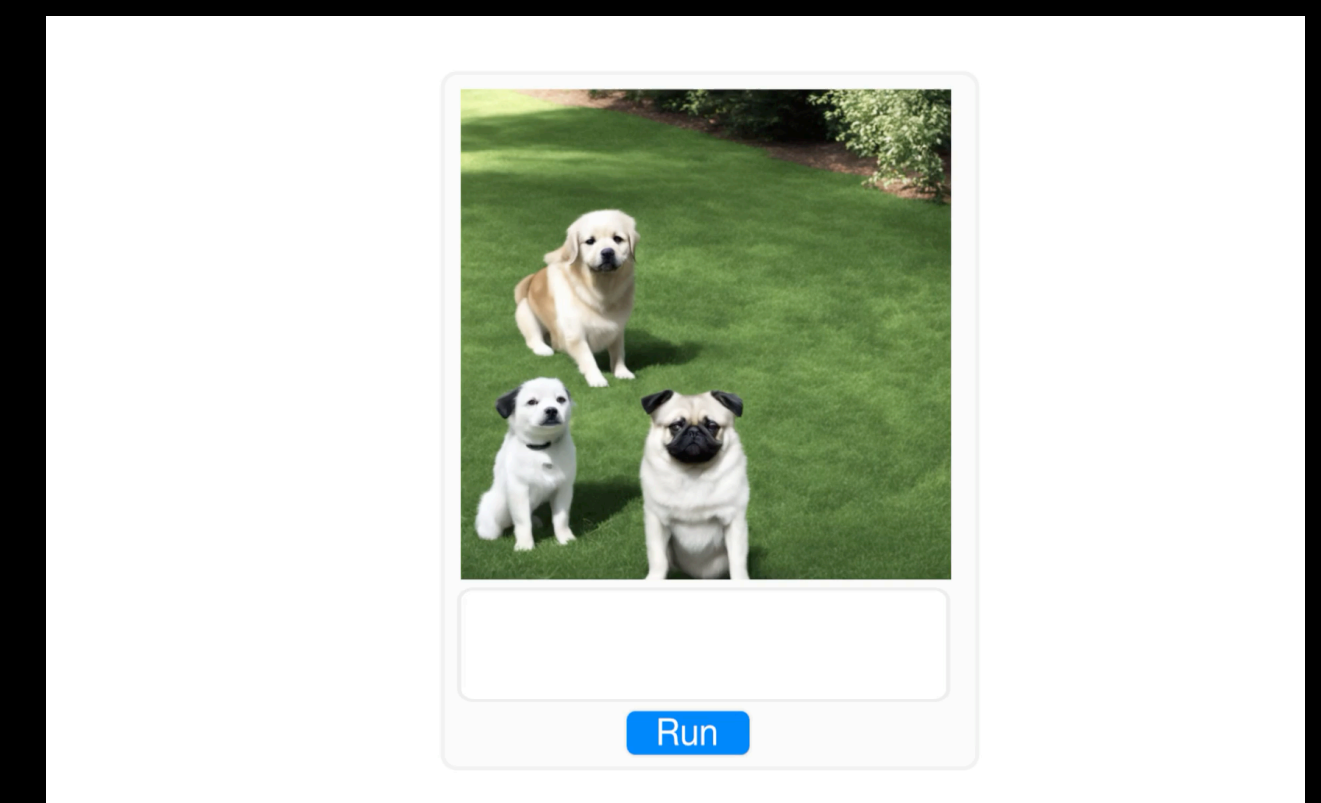
Learn & Explain

- 👉 Diffusion Explainer
- Transformer Explainer



Interpret & Attribute

- 👉 WizMap
- LLM Attributor
- MeMemo

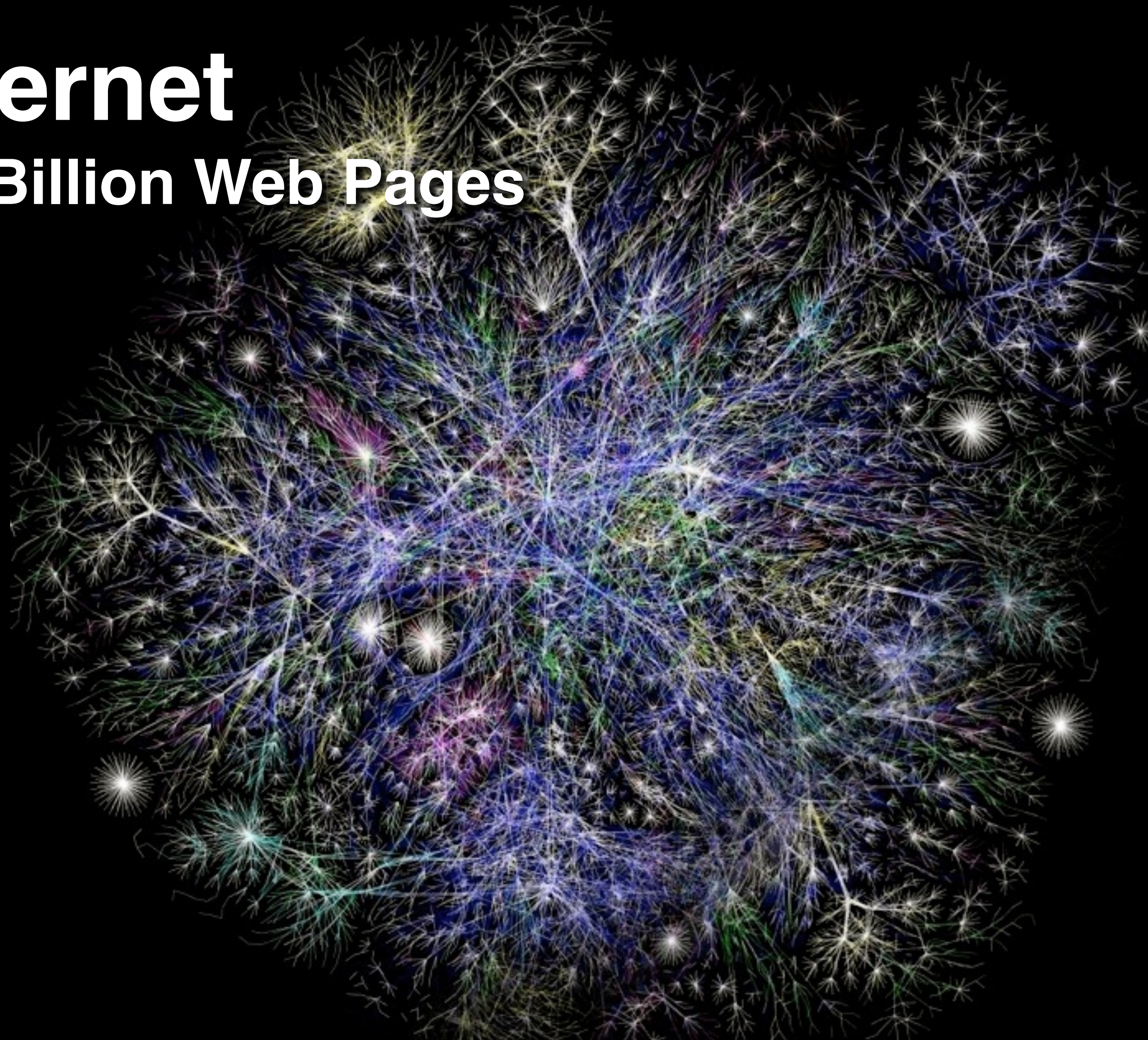


Guide & Safeguard

- 👉 LLM Self Defense
- Click Diffusion

Internet

50 Billion Web Pages



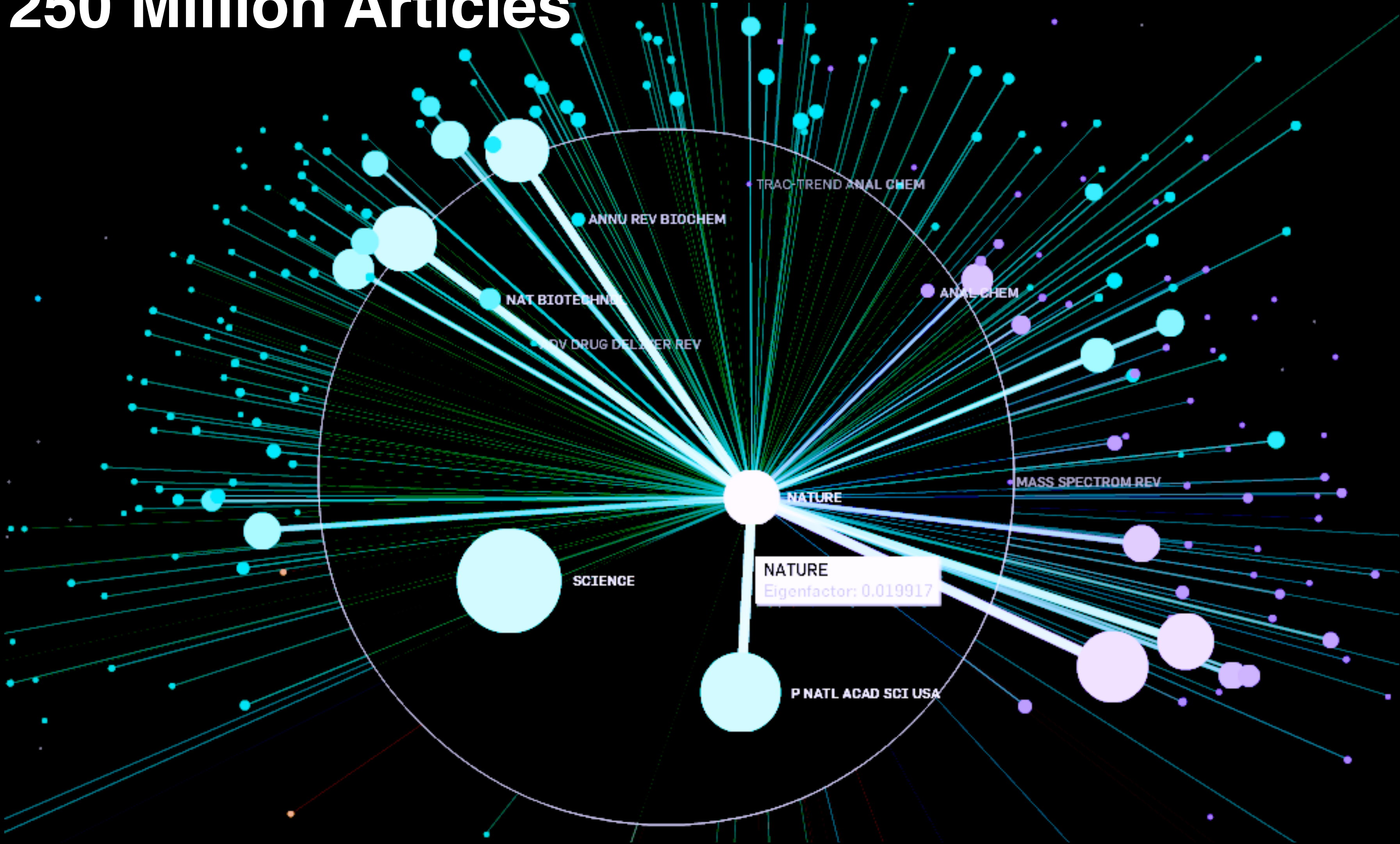
Facebook

2 Billion Users



Citation Network

250 Million Articles



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

“Big Data” Analyzed

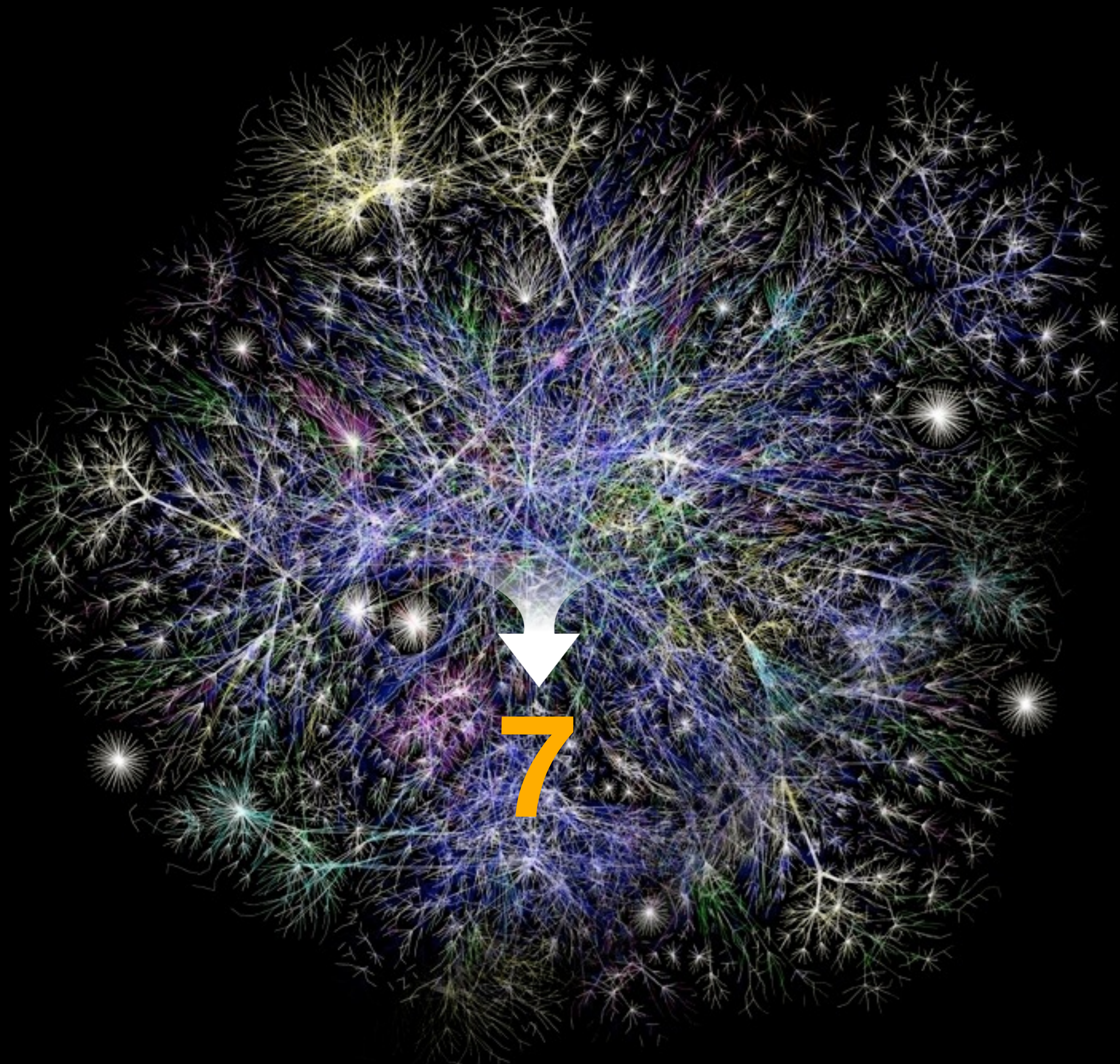
Graph	Nodes	Edges
YahooWeb	1.4 Billion	6 Billion
Symantec Machine-File Graph	1 Billion	37 Billion
Twitter	104 Million	3.7 Billion
Phone call network	30 Million	260 Million

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

7 ± 2

Number of **items** an average human
holds in **working memory**

George Miller, 1956



Data



Insights

How to do that?

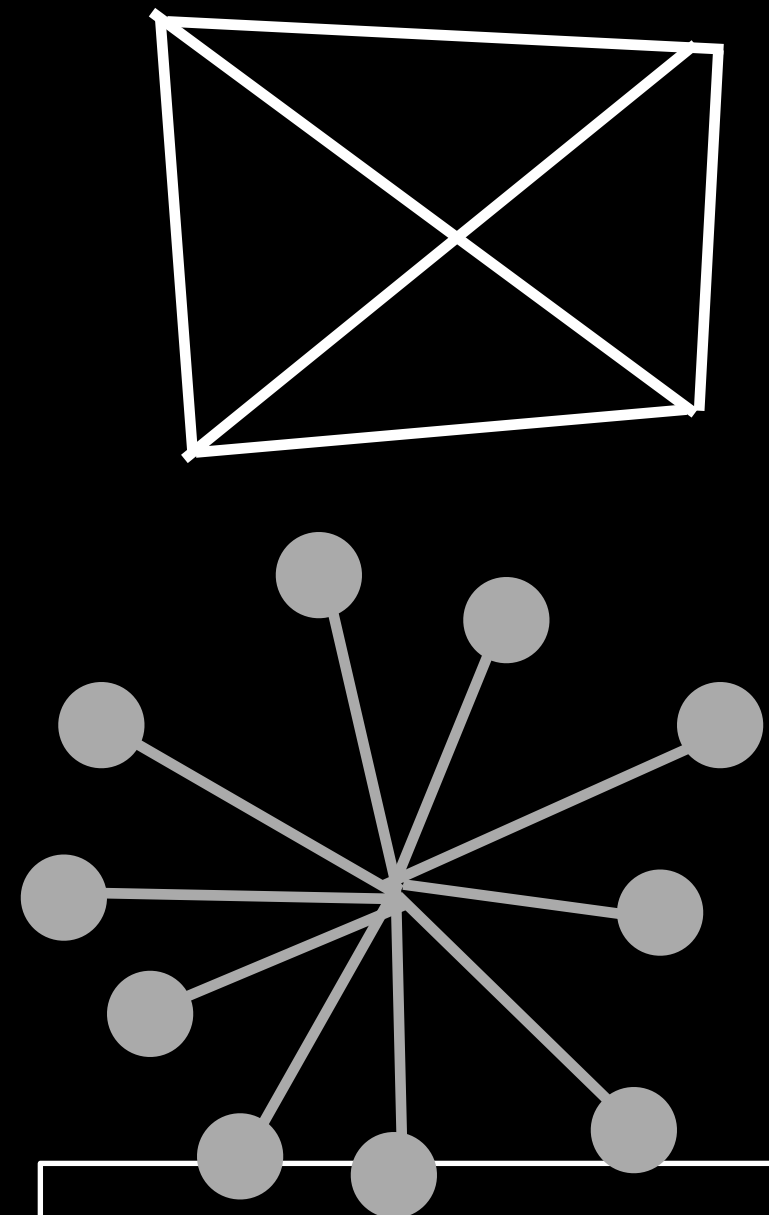
COMPUTATION
+
HUMAN INTUITION

Or, to ride the AI wave...

ARTIFICIAL INTELLIGENCE
+
HUMAN INTELLIGENCE

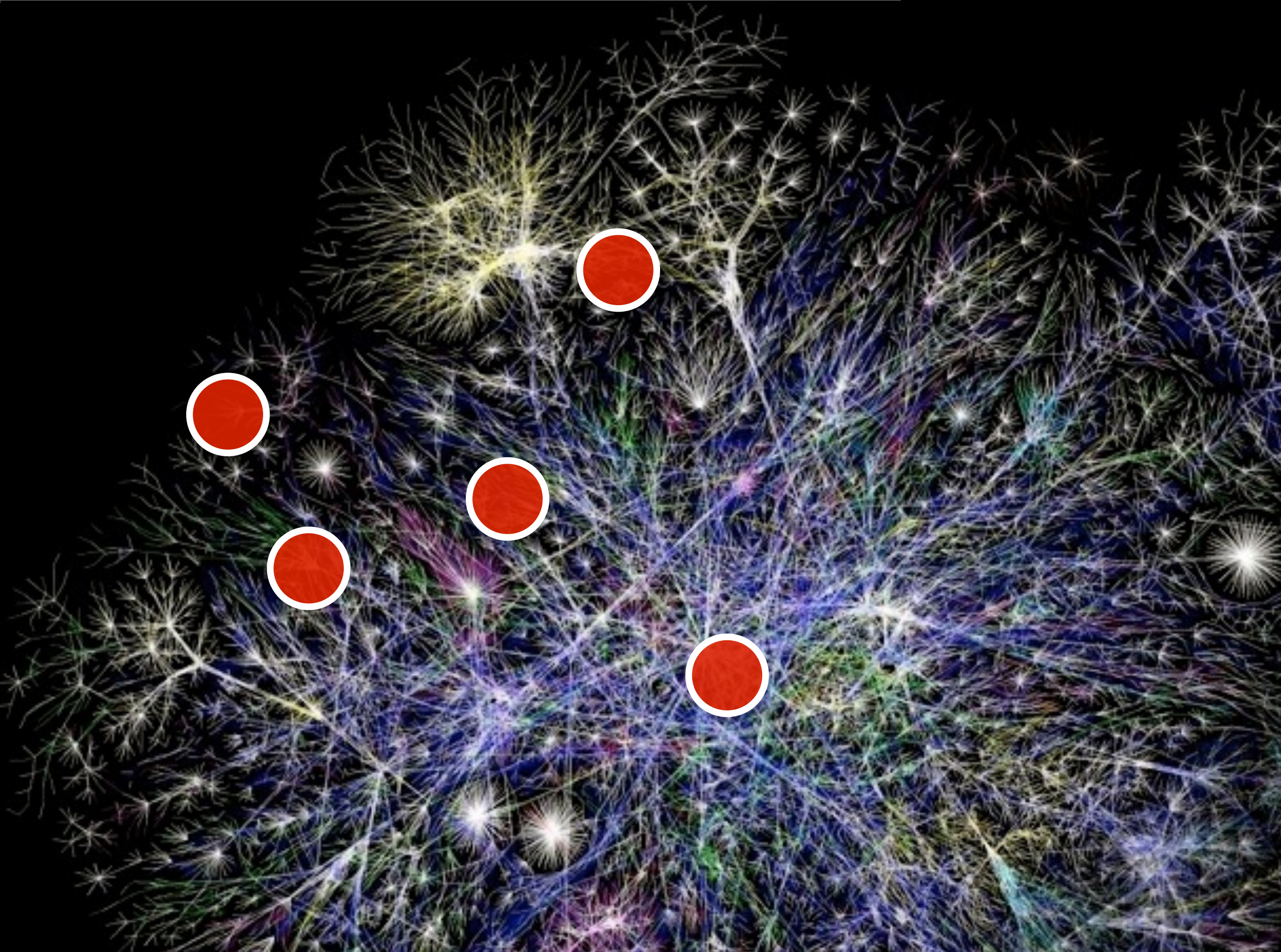
How to do that?

COMPUTATION



Both develop met
sense of ne

INTERACTIVE VIS



Our Approach for Big Data Analytics

MACHINE LEARNING

HCI

Human-Computer
Interaction

Automatic

User-driven; iterative

Summarization,
clustering, classification

Interaction, visualization

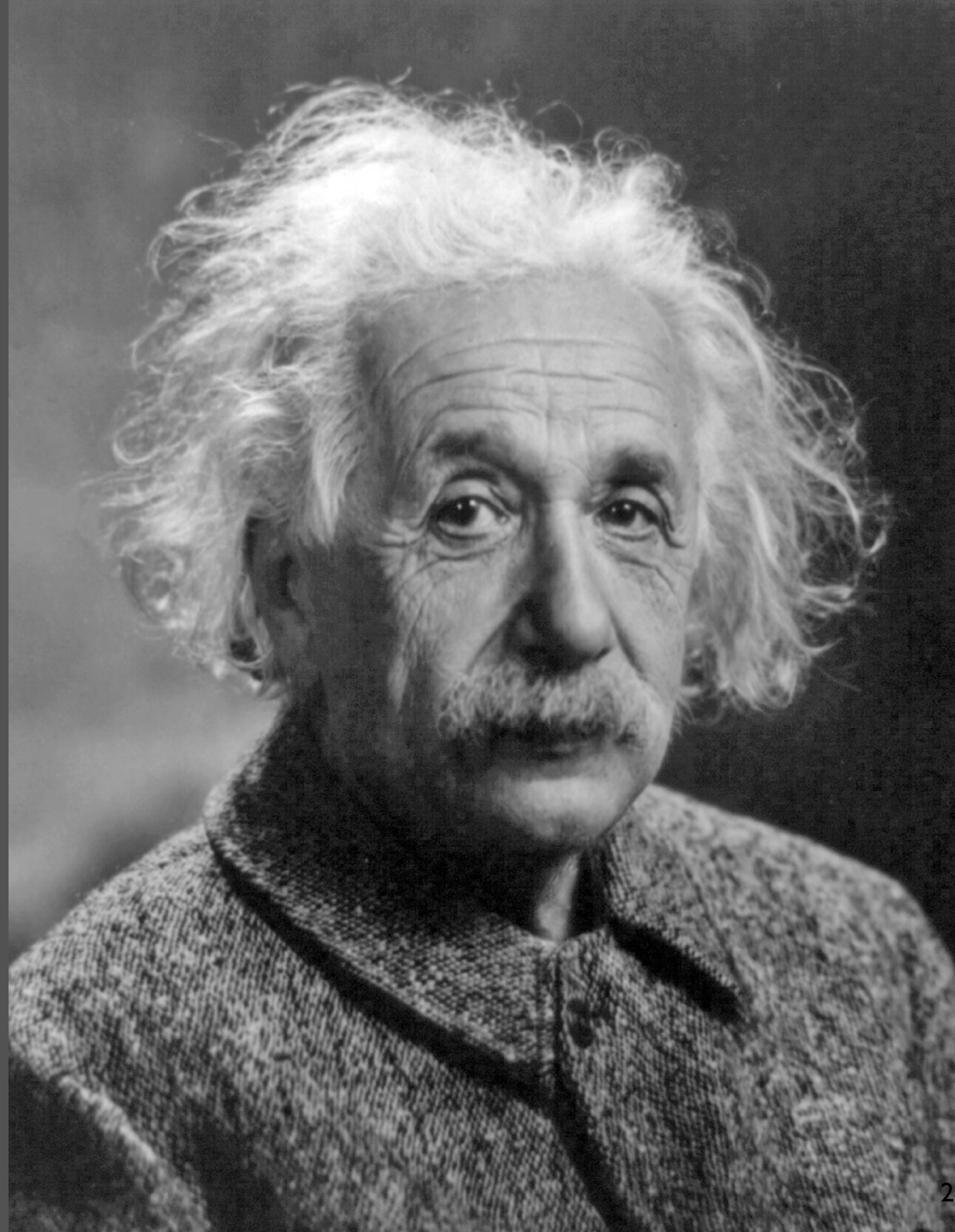
>Millions of items

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/cse6242-2024fall-campus/>
(link also available on Canvas)

**Discussion, Q&A,
find teammates**

Ed Discussion
(link available on Canvas)

**Assignment
Submission**

Canvas/Gradescope

Course Homepage

For syllabus, schedule, projects, datasets, etc.

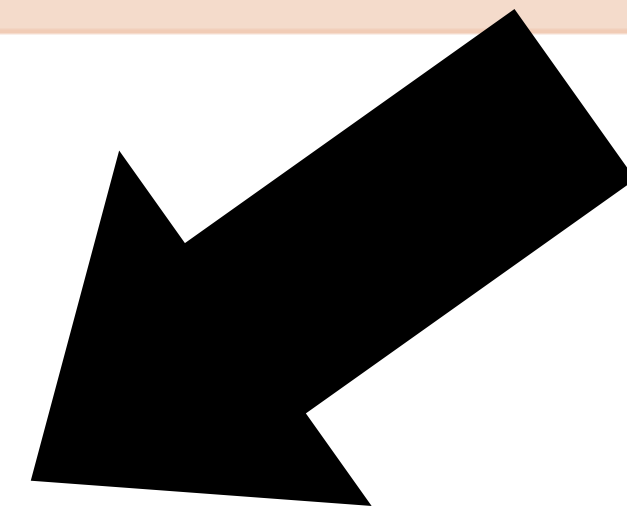
If you Google “cse6242”, you will see many matches.
Make sure you click the correct site!

[CSE6242A/CX4242A](#) [Schedule](#) [Homework](#) [Project](#) [Warnings](#) [Policies](#) [Datasets](#) [Resources](#)

There are [multiple CSE6242 sections](#). This is the course homepage for **campus CSE6242A/CX4242A**.

[CSE6242A/CX4242A Fall 2024](#)

Data and Visual Analytics



Join Ed Discussion Right Away

via canvas.gatech.edu

Announcements and Discussion

Home

Announcements

Modules

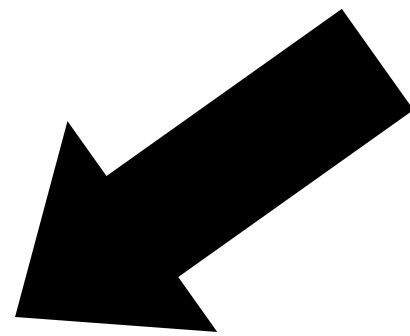
Ed Discussion

Assignments

Gradescope

Quizzes

People



We use Edstem for all announcements and discussion. Everyone must join this class's Ed Discussion through Canvas. Double check that you are joining the correct Edstem!

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. Students must always use **Ed Discussion** to communicate with course staff or for any class-related questions. Ed Discussion will be used for general posts, including private and public posts, threads, mega threads, Q&A, and announcements.

If course staff needs to communicate with specific students (i.e. members of a project team), the **Ed Chat** feature of Ed Discussion will be used. Students can benefit from this feature to communicate with other students. e.g., to discuss forming a project.

IMPORTANT: Everyone must ensure that the notification setting is on for both Ed Discussion and its Ed Chat feature to stay up to date with the class requirements and prevent losing points because of missing updates and announcements on Ed Discussion.

Important to join Ed Discussion because...

- We will announce events related to this class and data science in general
- Distinguished lectures, seminars
- Hackathons
- Company recruitment events (with free food, swags!)

Add your photo to help us and your classmates recognize you!

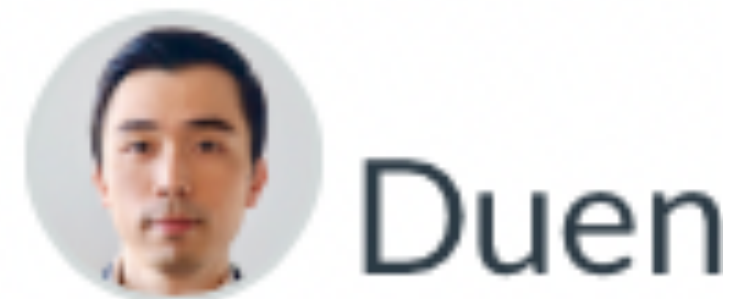
Canvas

☰ Duen Horng Chau's settings

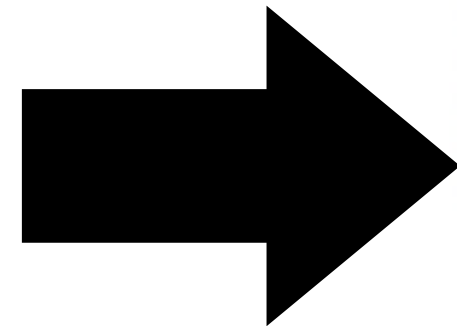
Notifications

Files

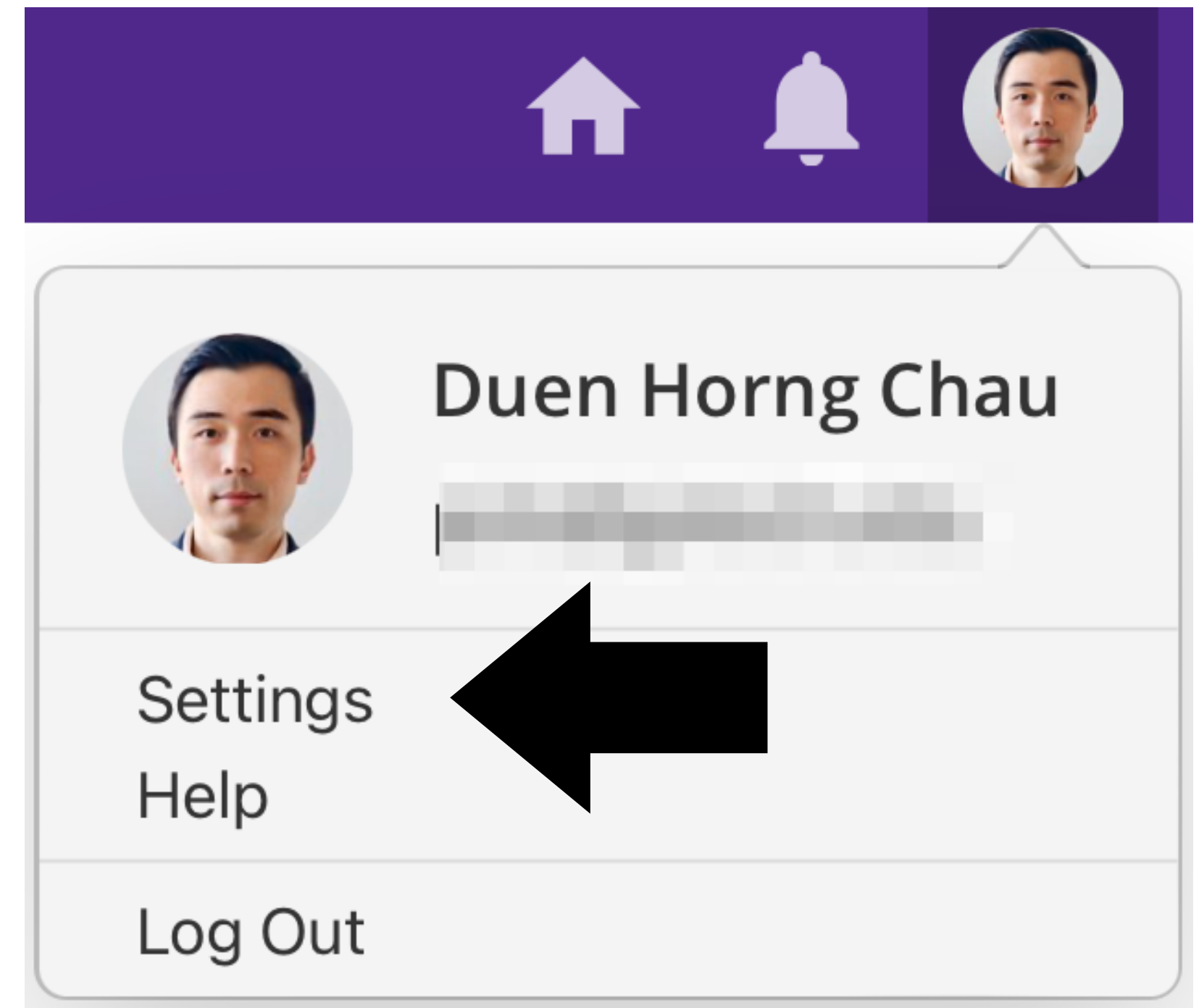
Settings



Full Name: Duen



Ed Discussion



If you need help cropping headshot photo into square shape, use **Magic Crop** (<https://poloclub.github.io/magic-crop/>)

Course Goals

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”?

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



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.

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

Can skip some

Cleaning

Can go back (two-way street)

Integration

- **Data types** inform **visualization** design

Analysis

- **Data size** informs choice of **algorithms**

Visualization

- **Visualization** motivates more **data cleaning**

Presentation

- **Visualization** challenges algorithm assumptions

Dissemination

e.g., user finds that results don't make sense

Course Goals

- Learn **visual** and **computational** 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
- **[Bonus points]** Quizzes
 - 4 online quizzes in total; ~10min each
 - **1% course grade point** each; lowest score dropped
- **No Exams** 🎉🎉🎉

Policies. Very Important!

(on course website)

Attendance, COVID-19, grading, plagiarism, collaboration, late submission, and the **“warnings”** about the difficulty this course

From Previous Classes...

- Class projects turned into papers at top conferences
- 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 Yahia*, Antoine Chassang*, Louis Raynaud*, Hugo Duthil*, Duen Horng (Polo) Chau
Georgia Institute of Technology
{alexandre.yahia, antoine.chassang, l.raynaud, hduthil, polo}@gatech.edu

ABSTRACT

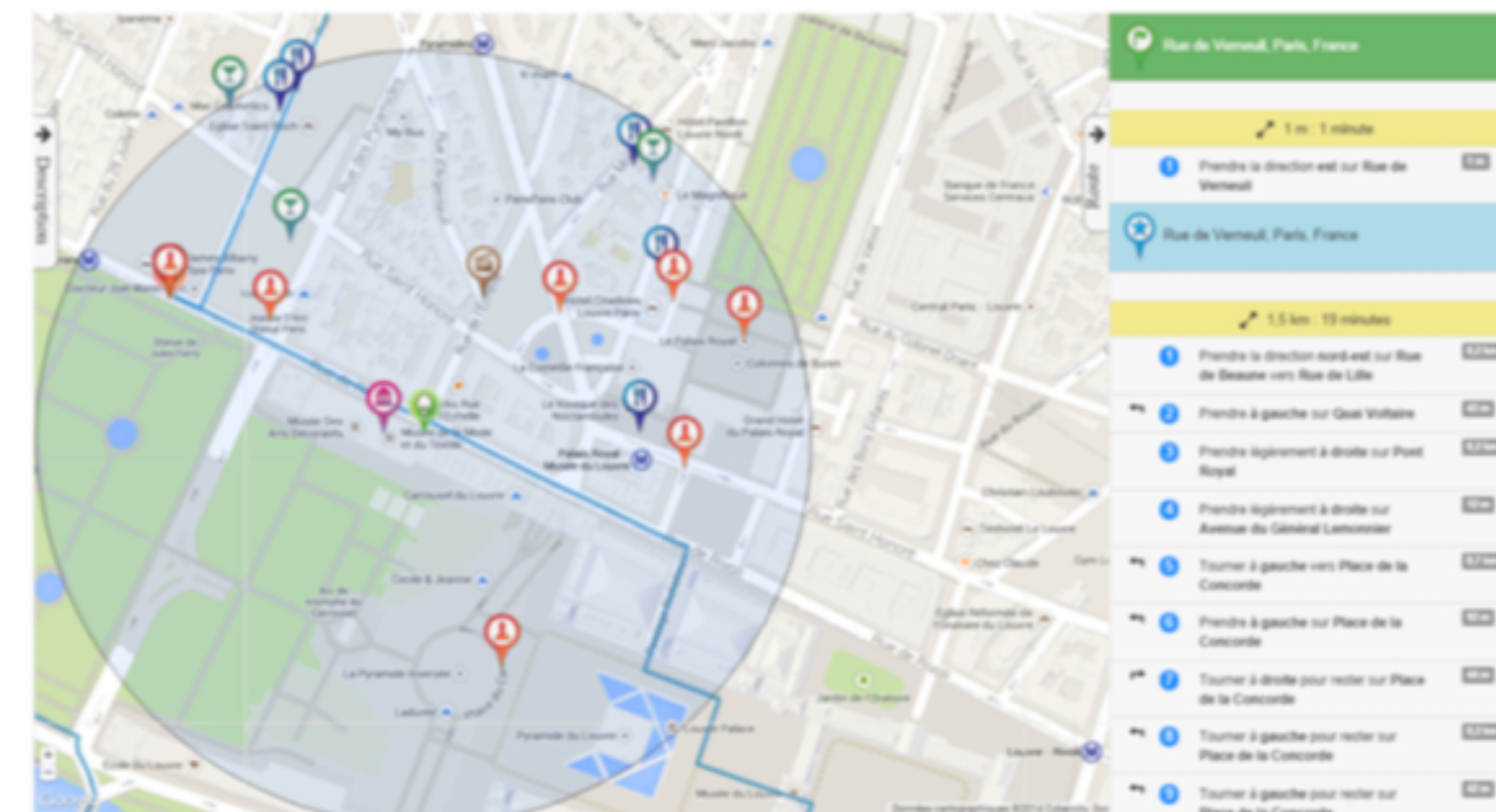
Planning personalized tour itineraries is a complex and challenging task for both humans and computers. Doing it manually is time-consuming; approaching it as an optimization problem is computationally NP hard. We present Aurigo, a tour planning system combining a recommendation algorithm with interactive 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



Full conference paper

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

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

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 problem, our group has developed a mobile application, PASSAGE, that uses real-time crime data to find "safe paths" for pedestrians in Atlanta. The application uses a user interface to allow users to input their starting and ending points.

Author

Safe Path
Pulse

ACM

H.5.2
User-Centered Computing
User Interfaces

Int

Georgia
Institute of
Technology

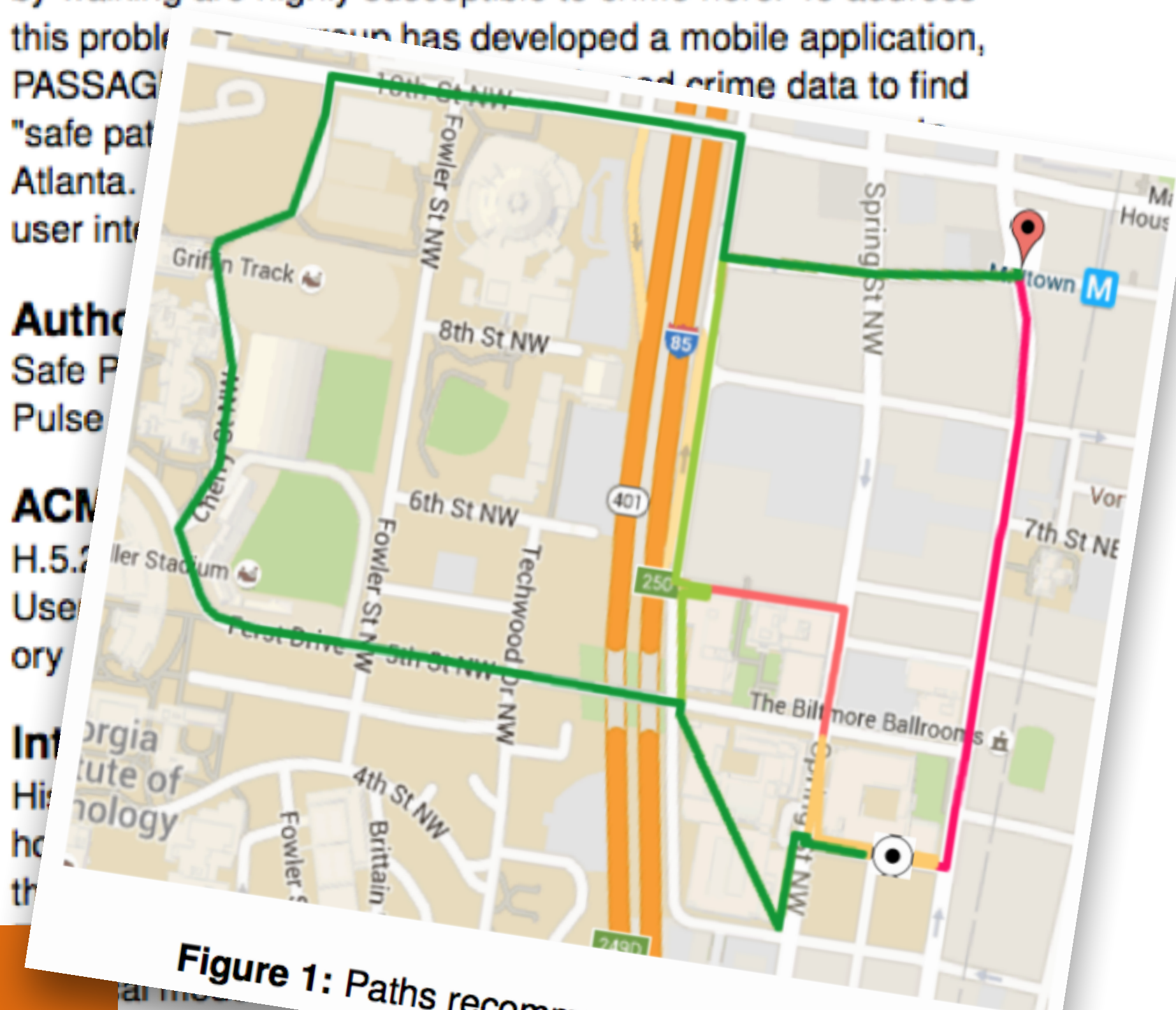


Figure 1: Paths recommended by PASSAGE

“As someone with 25 years work experience, I find my self **directly applying what I am learning within days**. The skill set of rapid learning that you are teaching is the main thing I interview for.”

“...thank you for the materials taught in DVA. As it was **perfectly aligned** with the what employers are looking out for. It made less challenging for me to secure this new job [Business Intelligence engineer at Amazon] in this competitive job market.”

“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**.”

“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.”

What we expects from you

- **Actively participate** throughout the course!
- If you need help, **let us know early** — the earlier you let us know, the more help we can offer
- **Help your fellow classmates**, e.g., help answer questions on Ed Discussion
- **Share your ideas!** Ideas for improving learning experiences, let us know

FREE After-class Coffee

- After (some) classes, we'll have 5-7 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 two weeks from now!