CSE6242 / CX4242

Data & Visual Analytics

Max Mahdi Roozbahani
Senior Lecturer at SCI
Joint faculty at IC, CSE, OMSCS, OMSA
SCI ML Committee
Founder of Filio

Course Registration

Classroom has capacity to accommodate a bit more (CSE6242 + CX4242). **We will raise the number of seats if it is possible.**

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.

Course TAs Be very nice to them!

Yiwei Kuang (Co-Head TA)

Ian Shu-Hei Wong (Co-Head TA)

Xiaoai Zhu

Neha Rajesh Lakhani

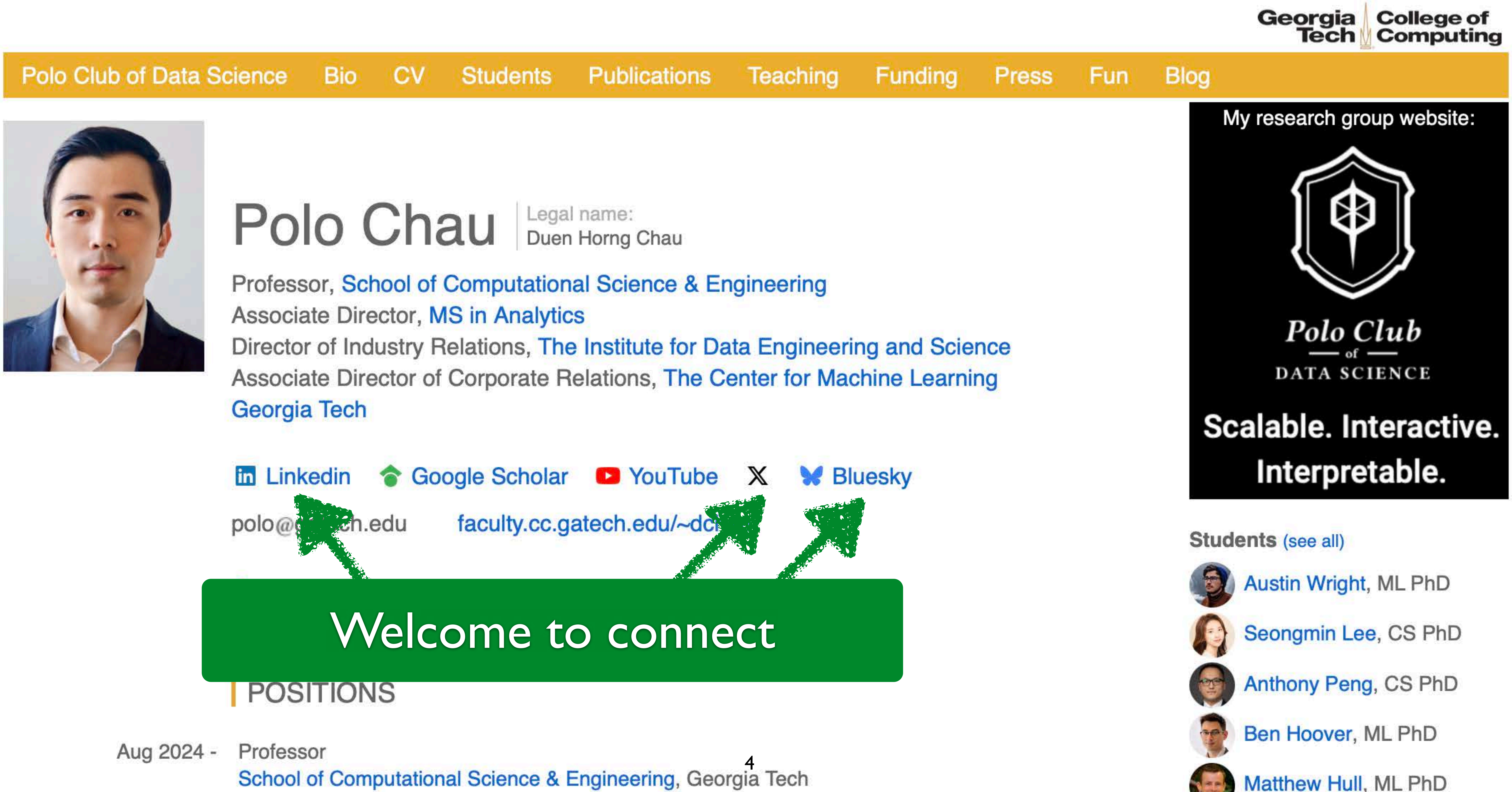
Yiting Zhang

Sai Shivanand Gokhale

Madhur Milind Rajadhyaksha

Aiswarya Jayachandran

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



Georgia Tech | College of Computing

Polo Club of Data Science | Bio | CV | Students | Publications | Teaching | Funding | Press | Fun | Blog

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 | Google Scholar | YouTube | X | Bluesky

polo@gatech.edu faculty.cc.gatech.edu/~dchau

Welcome to connect

POSITIONS

Aug 2024 - Professor
School of Computational Science & Engineering, Georgia Tech

My research group website:

Polo Club of DATA SCIENCE

Scalable. Interactive. Interpretable.

Students (see all)

- Austin Wright, ML PhD
- Seongmin Lee, CS PhD
- Anthony Peng, CS PhD
- Ben Hoover, ML PhD
- Matthew Hull, ML PhD

The course focuses on working with large datasets.

(Also the focus of Polo’s research group)

🌐 poloclub.github.io

Polo Club of Data Science

AI + HI

ARTIFICIAL INTELLIGENCE

HUMAN INTELLIGENCE

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

Austin
ML PhD

Seongmin
CS PhD

Ben
ML PhD

Anthony
CS PhD

Matthew
ML PhD

Alec
ML PhD

Mansi
CS PhD

Aeree
ML PhD

Alex
CS PhD

Pratham
CS MS

Sri
CS
Undergrad

Aishwarya
CS
Undergrad

Rahul
CS
Undergrad

Kunal
CS
Undergrad

Polo
Prof

Haekyu
🎓 Machine Learning Engineer, Stripe

Jay
🎓 Member of Technical Staff, OpenAI

Sivapriya
🎓 AI Research, JPMorgan Chase

Harsha
🎓 Deep Learning Engineer, Toyon Research

Omar
🎓 CS PhD, Stanford

Megan
🎓 CS MS, Stanford

Alexander
🎓 CS PhD

Kevin
🎓 ML PhD, CMU

David
🎓 Machine Learning Engineer, Apple

Justin
🎓 Software Engineer, Meta

HUMAN CENTERED AI FOR EVERYONE

poloclub.github.io

Austin

Seongmin

Ben

Anthony

Matthew

Alec

Mansi

Aeree

Alex

Pratham

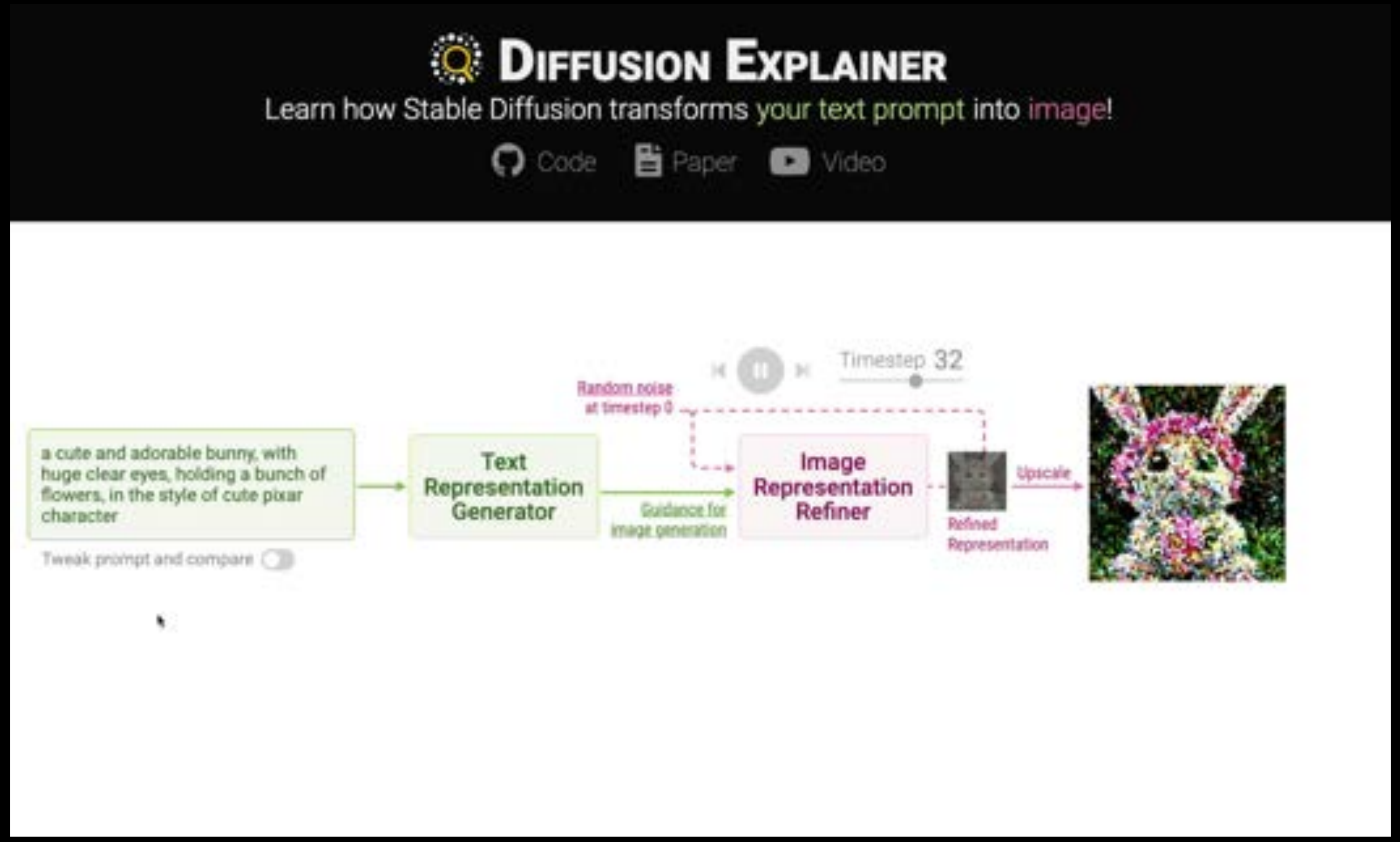
Sri

Aishwarya

Rahul

Kunal

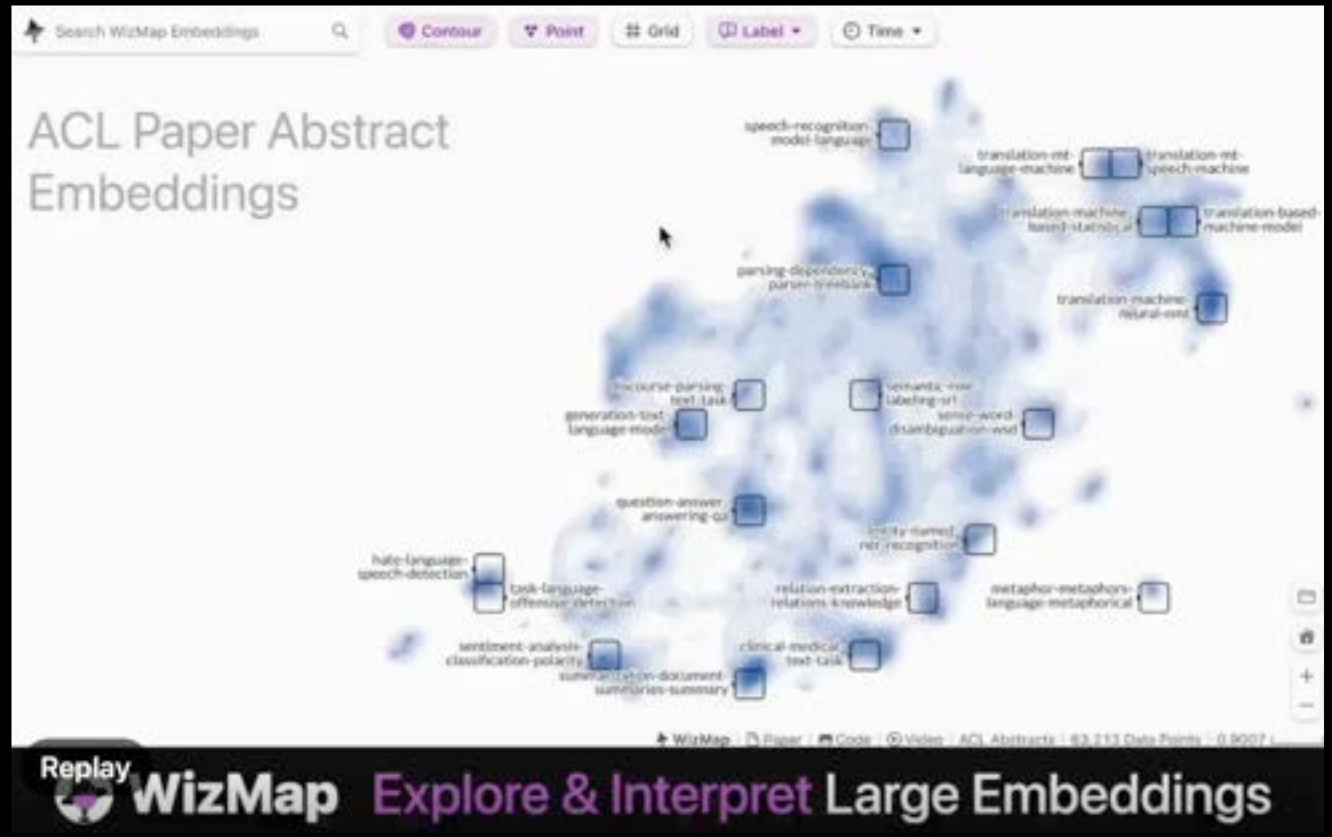
Polo

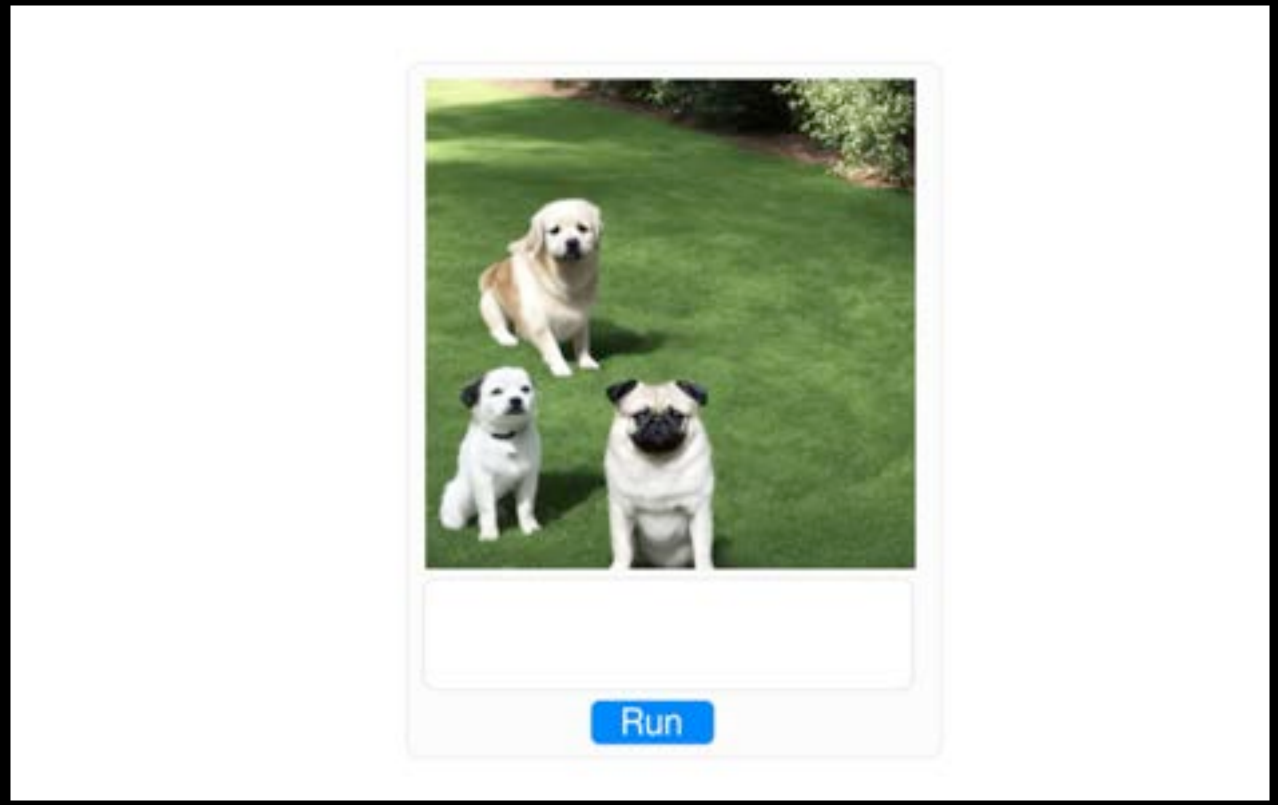

DIFFUSION EXPLAINER

Learn how Stable Diffusion transforms your text prompt into image!

a cute and adorable bunny, with huge clear eyes, holding a bunch of flowers, in the style of cute pixar character

Code
Paper
Video
Timestep 32
Random noise at timestep 0
Text Representation Generator
Guidance for image generation
Image Representation Refiner
Refined Representation
Upscale
Tweak prompt and compare


ACL Paper Abstract Embeddings



Learn & Explain

☝ Diffusion Explainer
Transformer Explainer

Interpret & Attribute

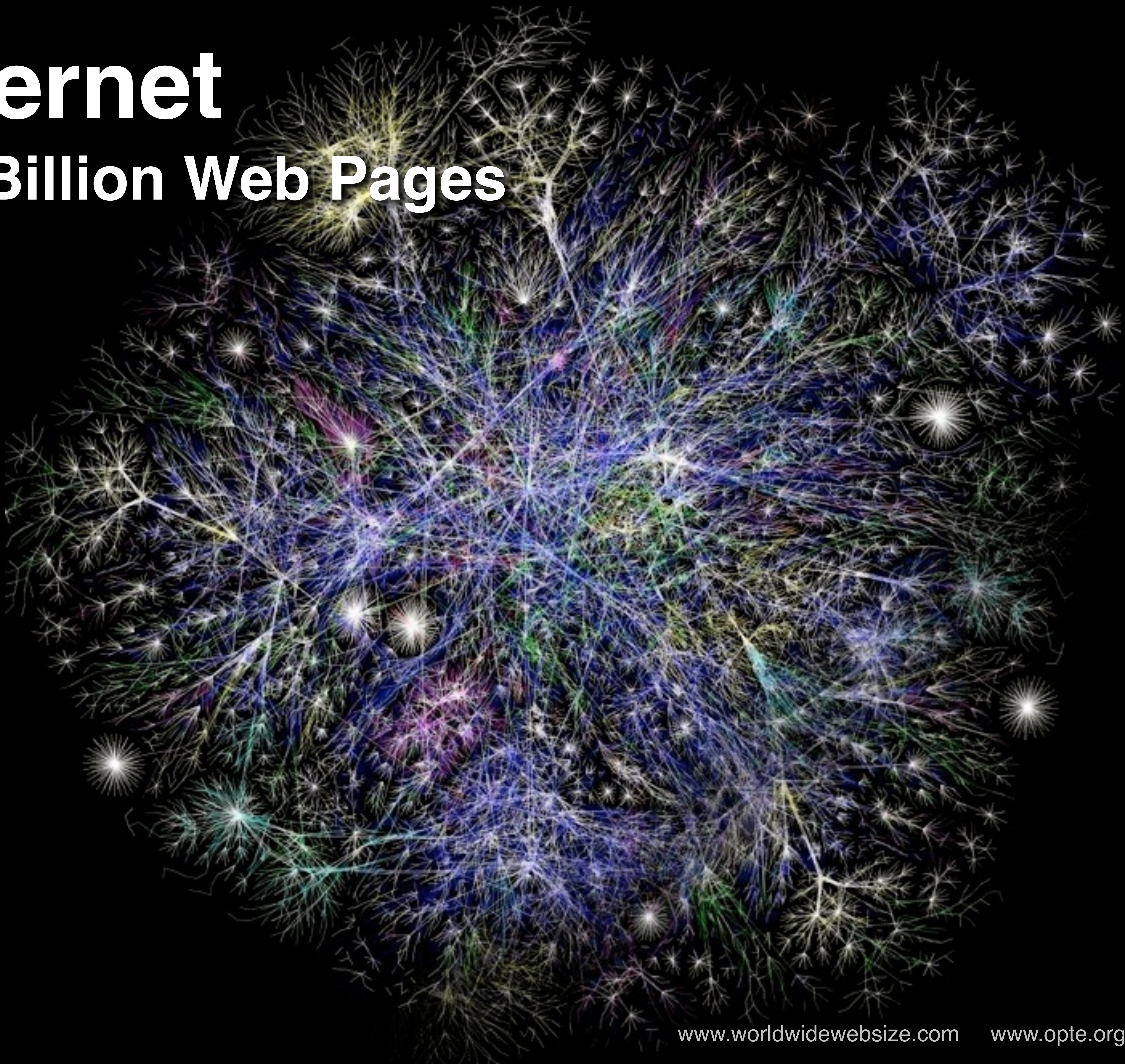
☝ WizMap
LLM Attributor
MeMemo

Guide & Safeguard

☝ LLM Self Defense
Click Diffusion

Internet

50 Billion Web Pages



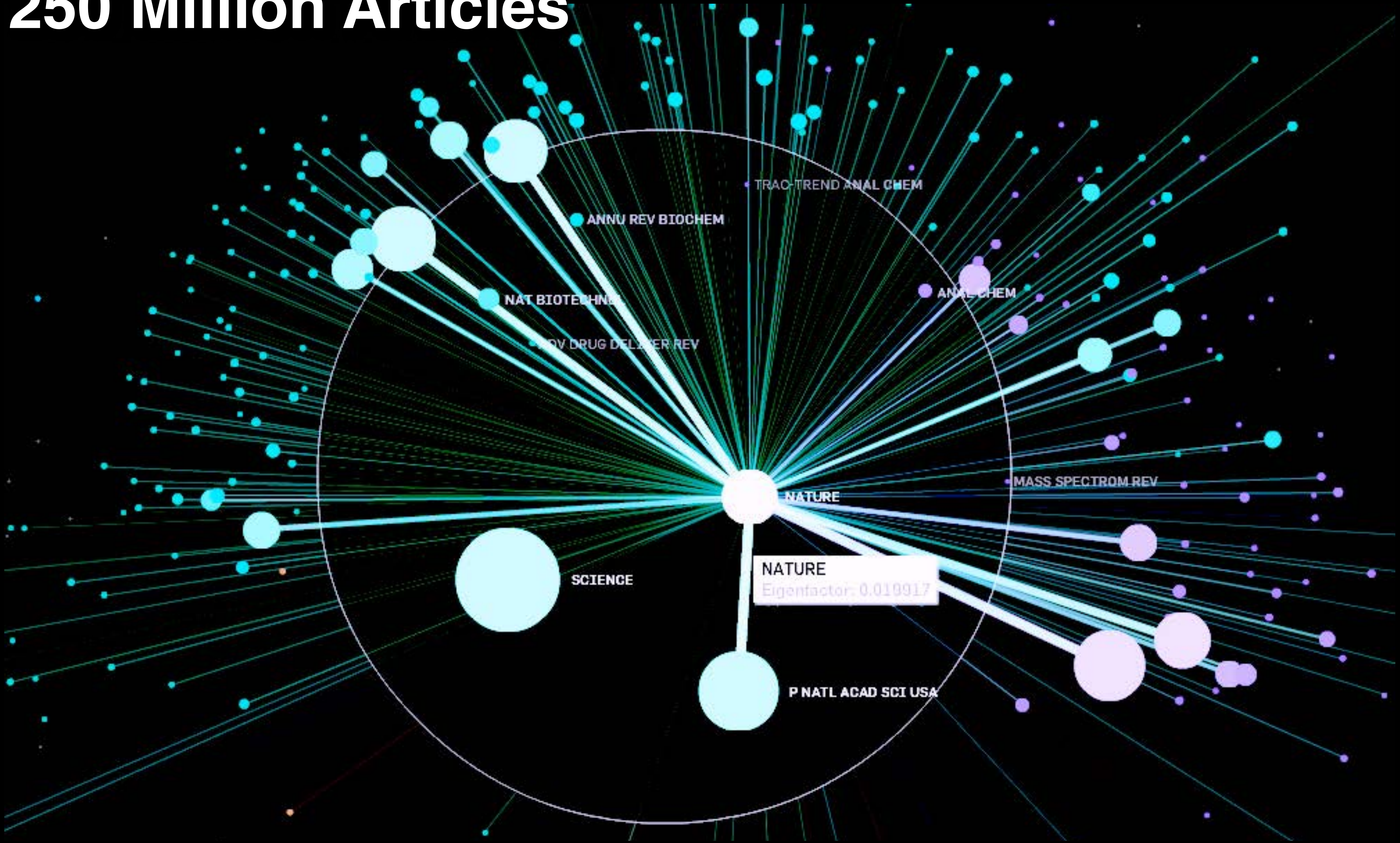
www.worldwidewebsize.com www.opte.org

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

X 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

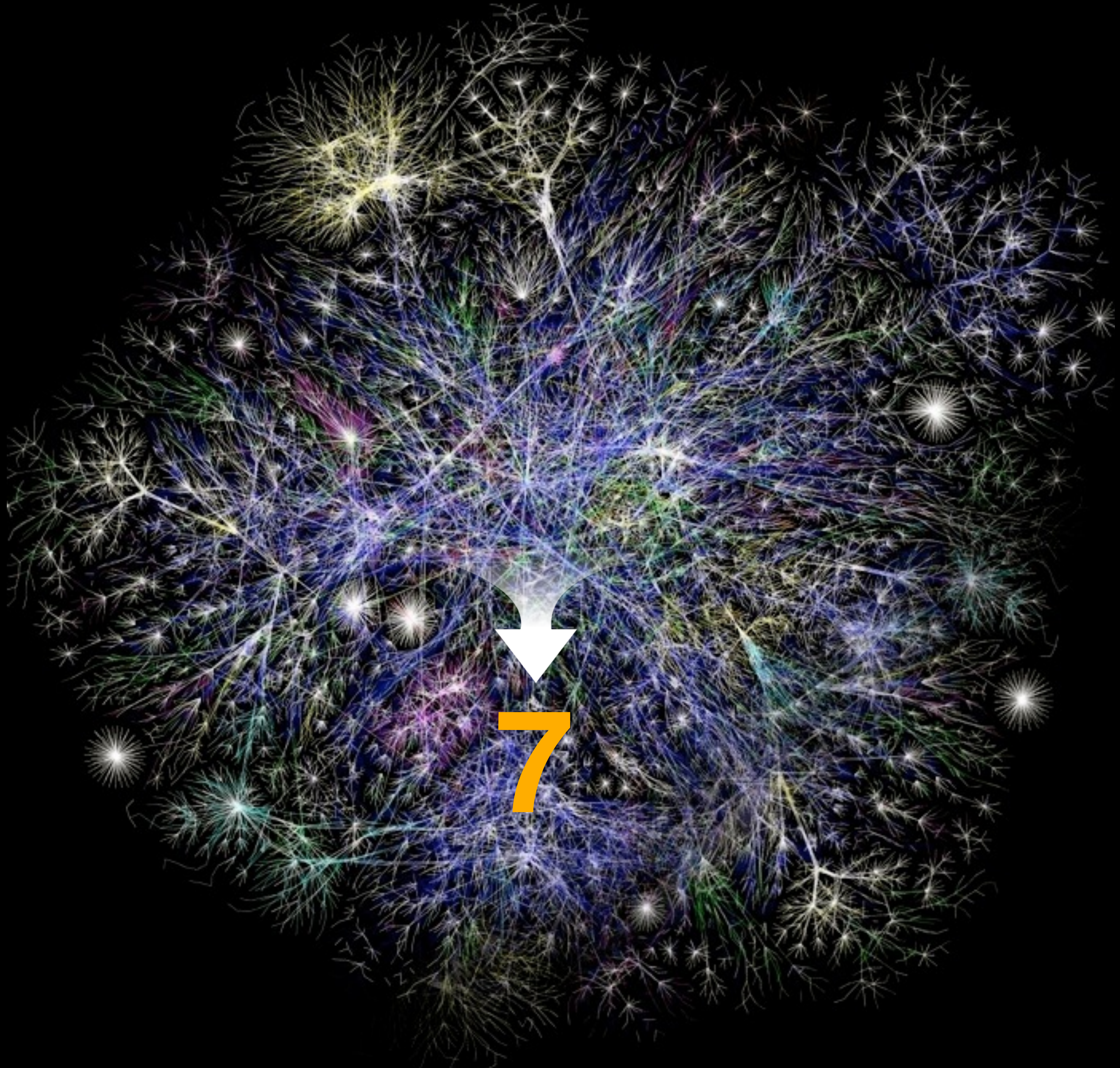
Graph	Nodes	Edges
YahooWeb	1.4 Billion	6 Billion
Symantec Machine-File Graph	1 Billion	**37 Billion**
X (Twitter)	104 Million	3.7 Billion
Phone call network	30 Million	260 Million

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



7

Data

Insights

How to do that?

COMPUTATION

+

HUMAN INTUITION

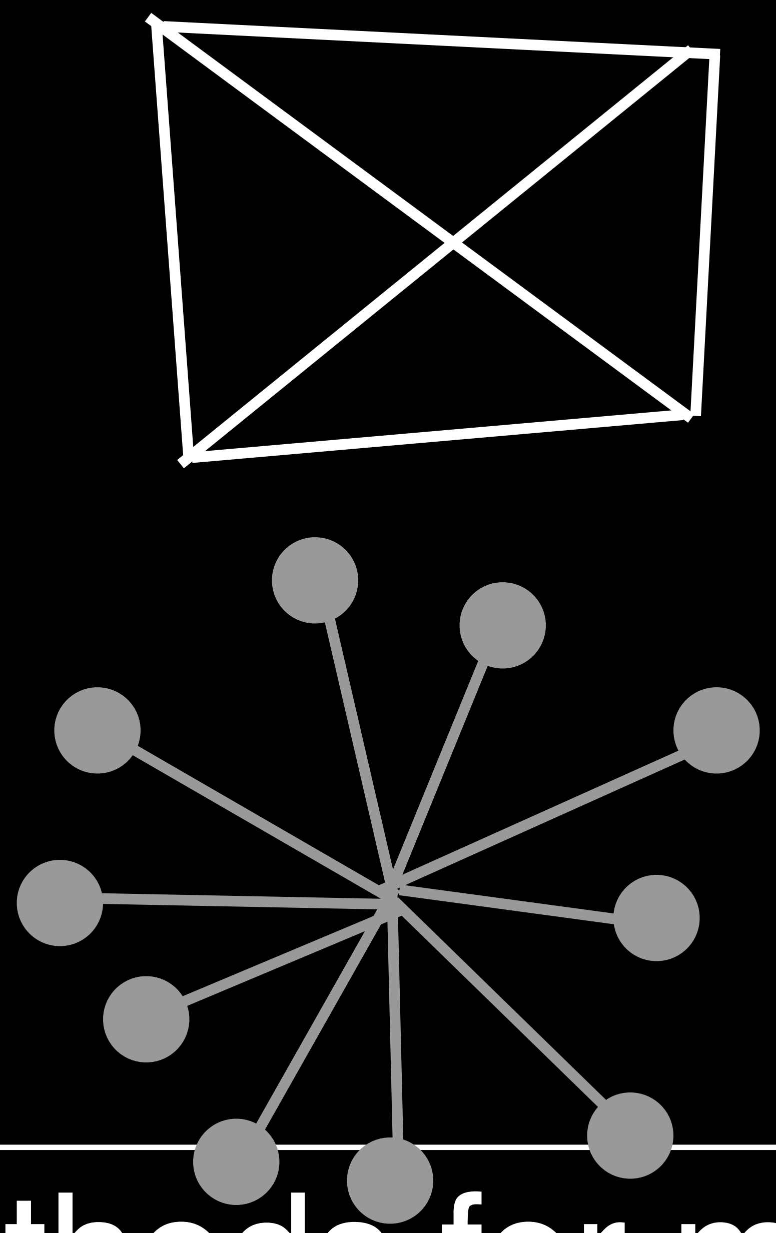
Or, to ride the AI wave…

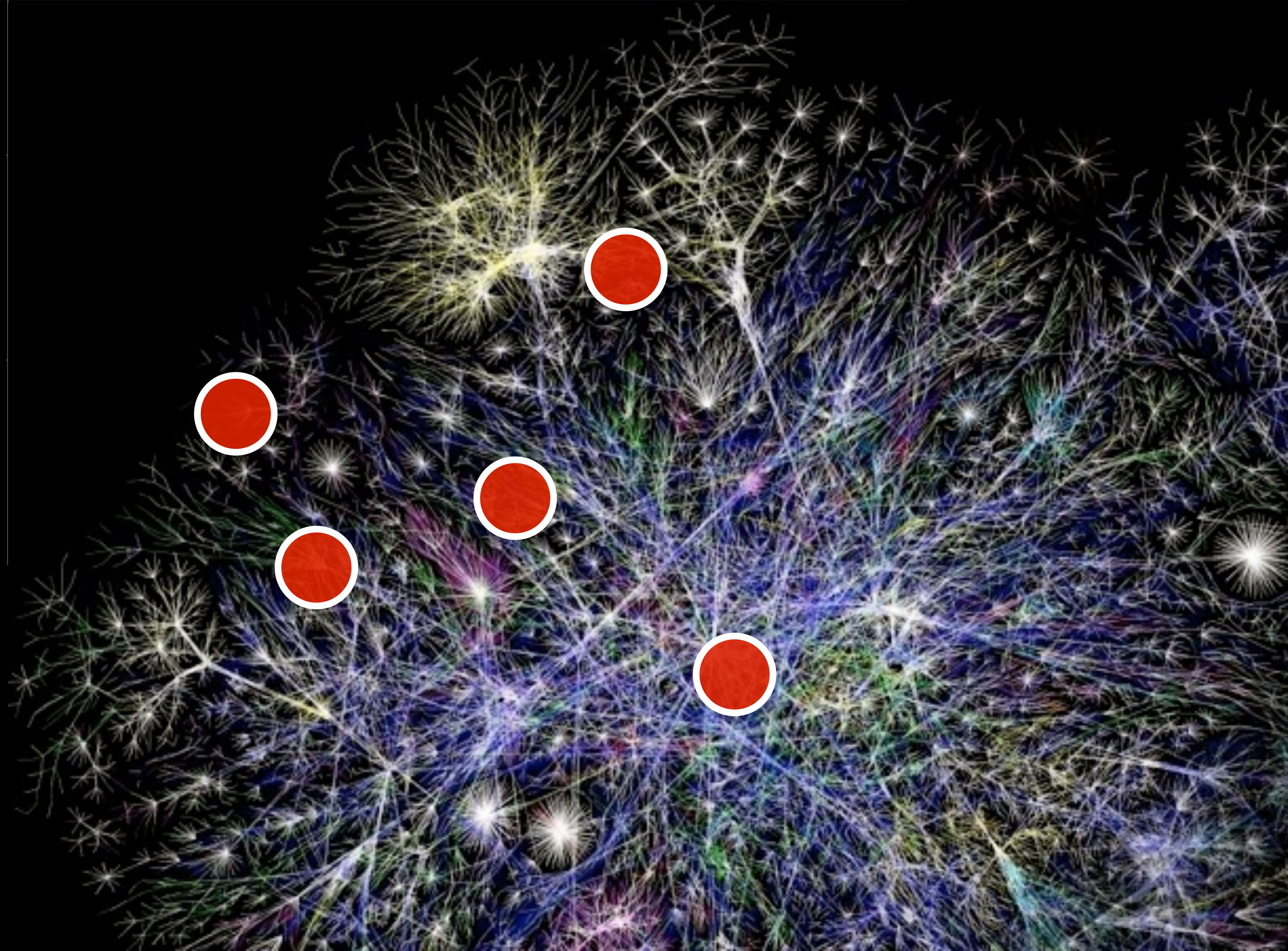
ARTIFICIAL INTELLIGENCE
+
HUMAN INTELLIGENCE

How to do that?

COMPUTATION

INTERACTIVE VIS





Both develop methods for making sense of network data

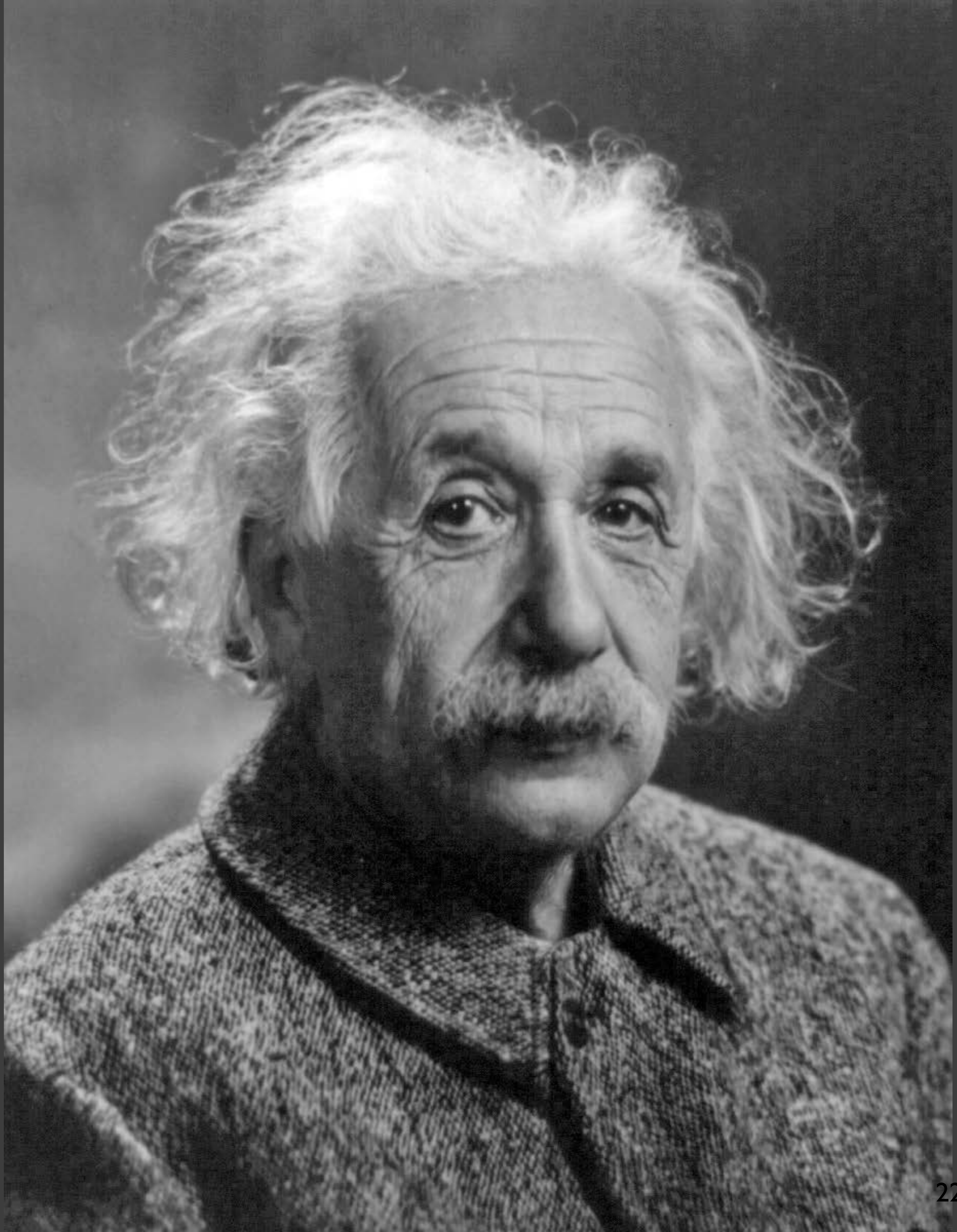
Our Approach for Big Data Analytics

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.)

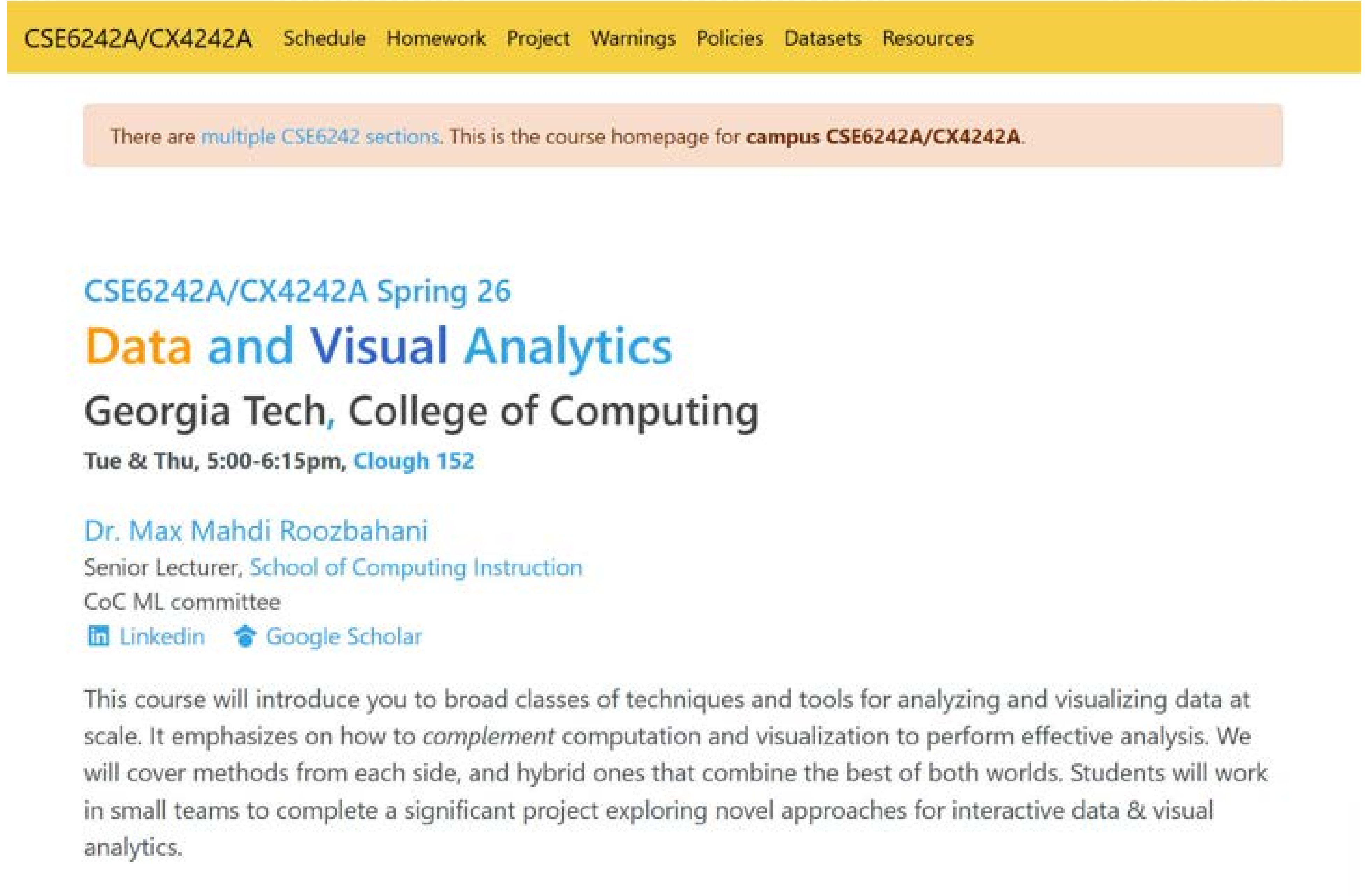
Logistics

Course website Policies, syllabus, schedule, etc.	https://poloclub.github.io/cse6242-2026spring-campus/ (link also available on Canvas)
Discussion, Q&A, find teammates	**Ed Discussion** (access via 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!



There are multiple CSE6242 sections. This is the course homepage for **campus CSE6242A/CX4242A**.

CSE6242A/CX4242A Spring 26

Data and Visual Analytics

Georgia Tech, College of Computing

Tue & Thu, 5:00-6:15pm, Clough 152

Dr. Max Mahdi Roozbahani

Senior Lecturer, School of Computing Instruction

CoC ML committee

Linkedin Google Scholar

This course will introduce you to broad classes of techniques and tools for analyzing and visualizing data at scale. It emphasizes on how to *complement* computation and visualization to perform effective analysis. We will cover methods from each side, and hybrid ones that combine the best of both worlds. Students will work in small teams to complete a significant project exploring novel approaches for interactive data & visual analytics.

Join Ed Discussion Right Away

via canvas.gatech.edu

Home
Announcements
Modules
Ed Discussion
Assignments
Gradescope
Quizzes
People

Announcements and Discussion

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.

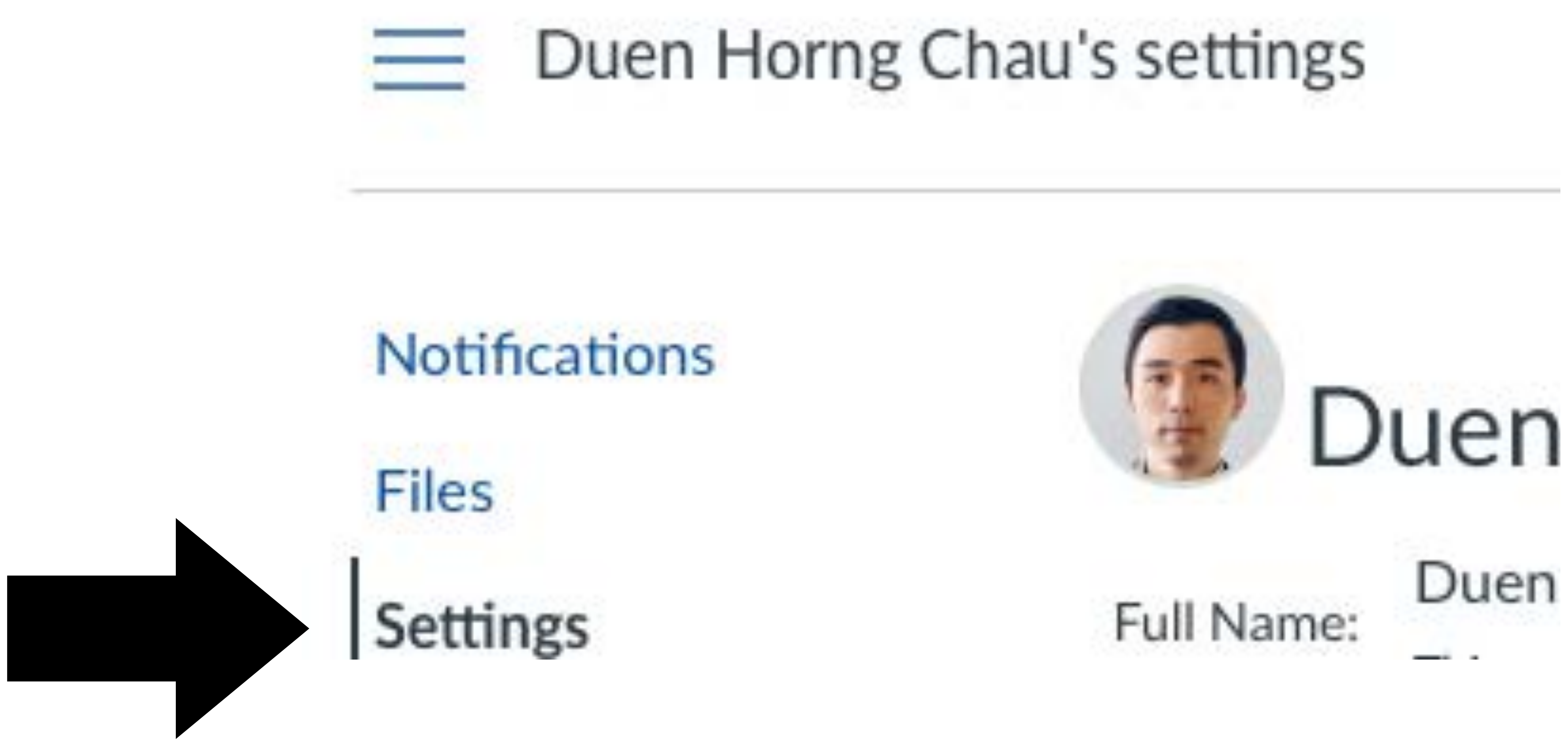
IMPORTANT: Everyone must ensure that the notification setting is on for both Ed Discussion and its Ed Chat feature to stay up to date with the class requirements and prevent losing points because of missing updates and announcements on Ed Discussion.

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!)

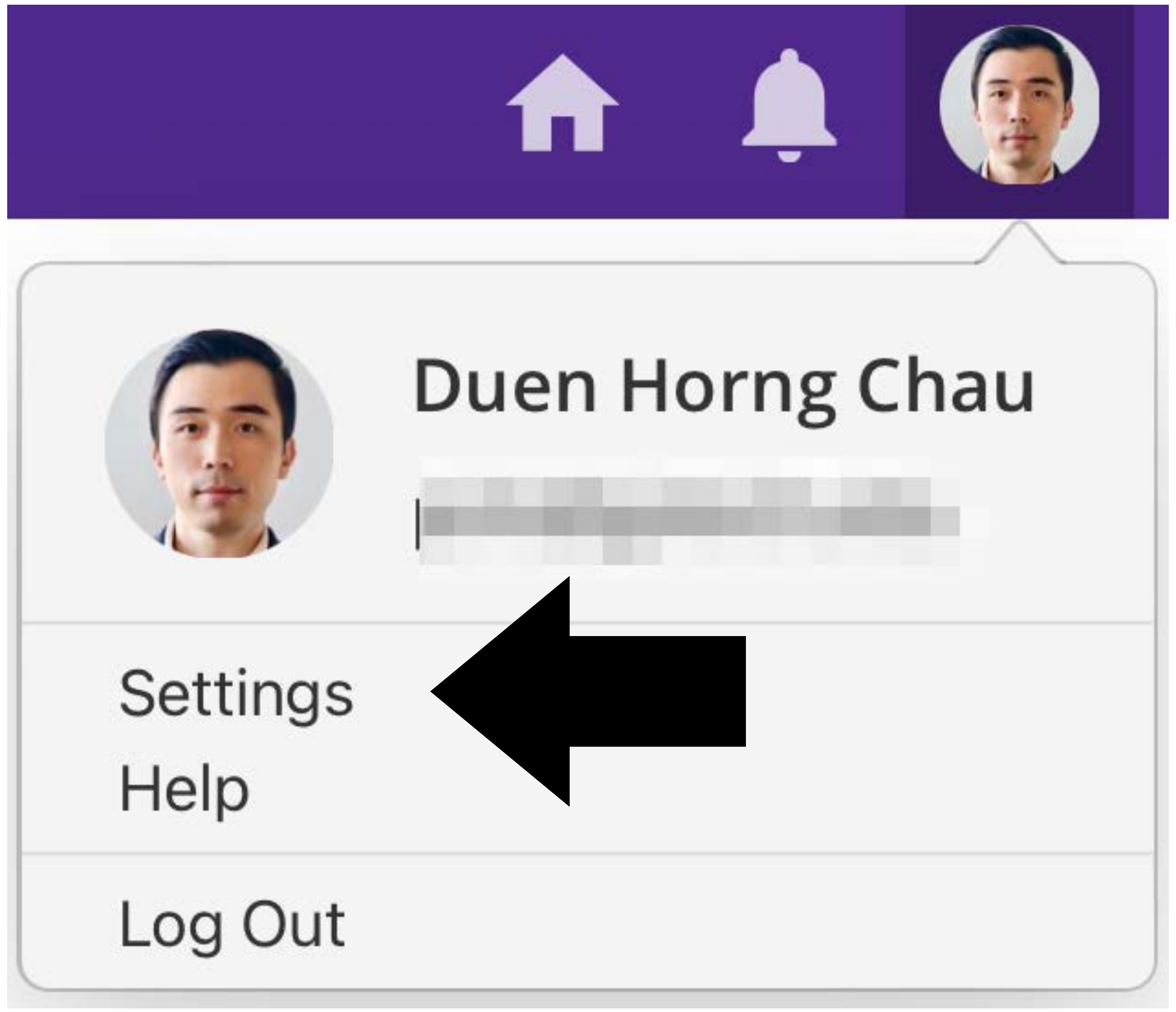
Add your photo to help us and your classmates recognize you!

Canvas


Duen Horng Chau's settings

Full Name: Duen

Ed Discussion


Duen Horng Chau

[redacted]

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?

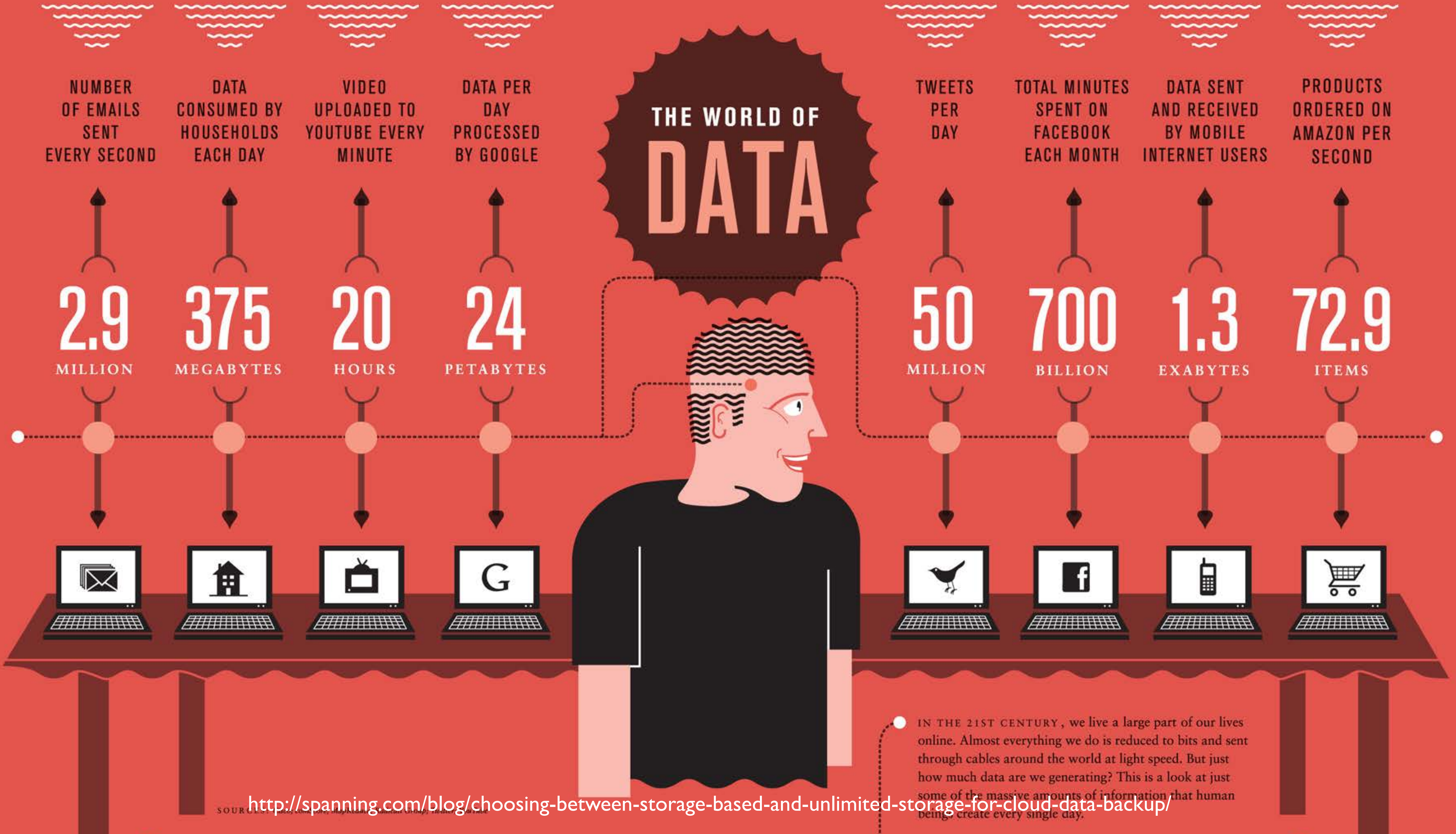
No formal definition!

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.

SOURCES: [illegible]

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 **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
- **5.67% bonus points**
  - 1.67% for HW2
  - 3% for bonus quizzes; 4 online quizzes (~10min each); lowest-scoring quiz dropped
  - 1% for CIOS
- **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…

- Projects as portfolio pieces on CV
- Increased job and internship opportunities
  - Former students sent me “thank you” notes
- Class projects turned into publications

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.

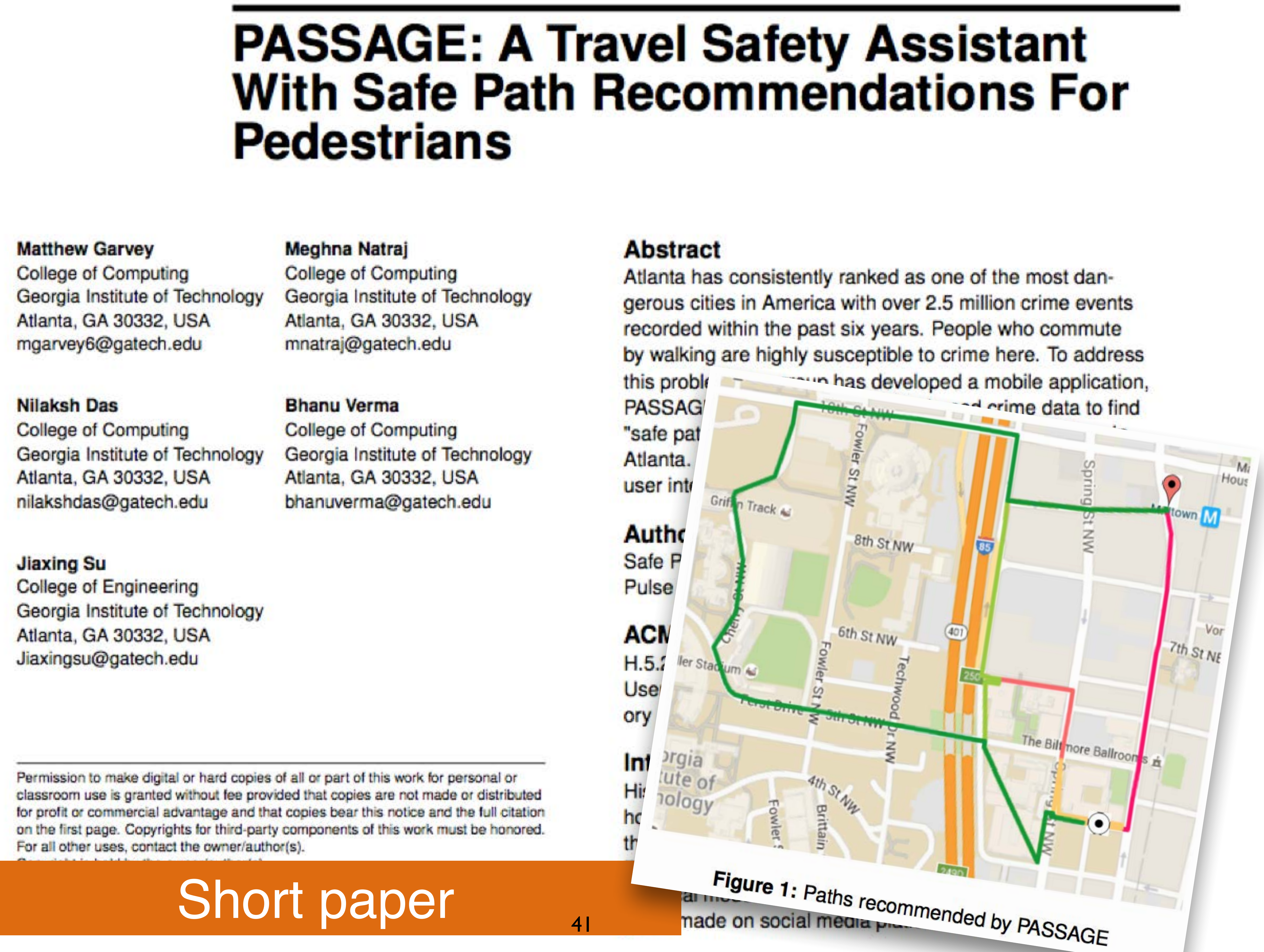
Author Keywords

User Interfaces; Visualization; Recommendation; Tour itinerary planning

ACM Classification Keywords

(e.g. HCI): User interfaces

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

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
hc
th

made on social media p

Figure 1: Paths recommended by PASSAGE

Short paper

“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