Data & Visual Analytics

Duen Horng (Polo) Chau
Georgia Tech

CSE 6242 A / CS 4803 DVA

Jan 8, 2013

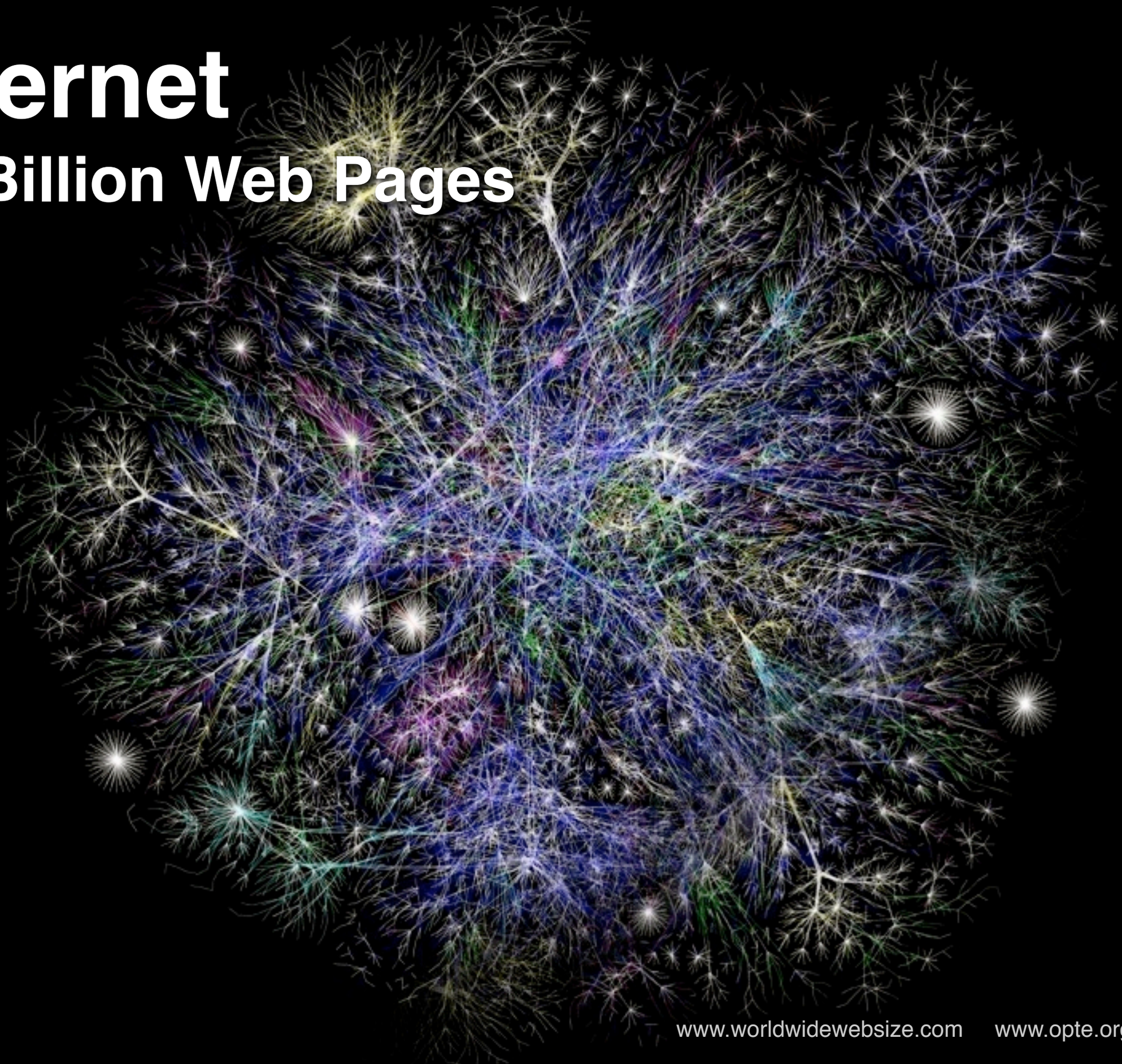
I Work with Large Graphs

I Work with Large Graphs

= Large Network Data

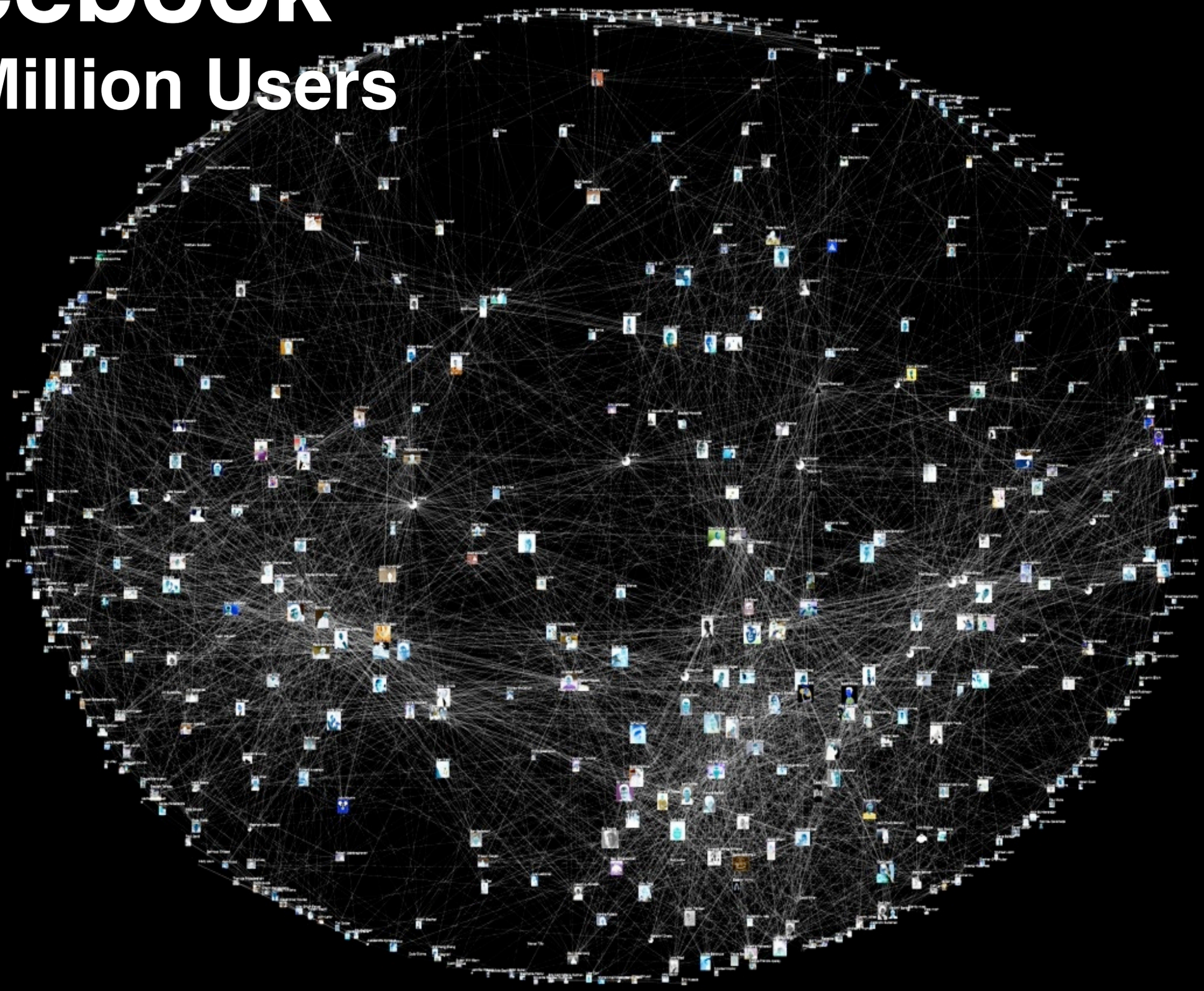
Internet

50 Billion Web Pages



Facebook

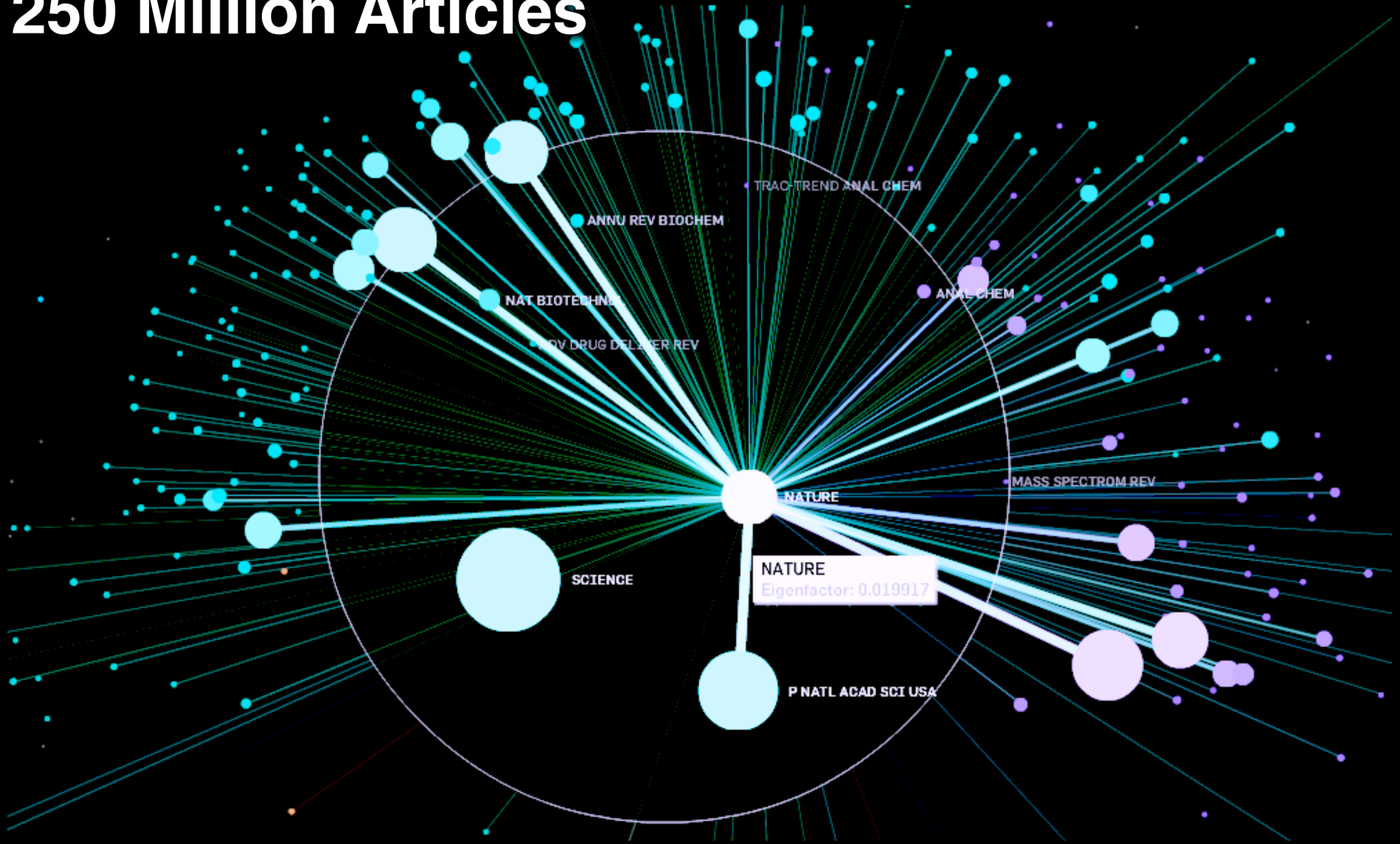
800 Million Users



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 Graphs I Analyzed

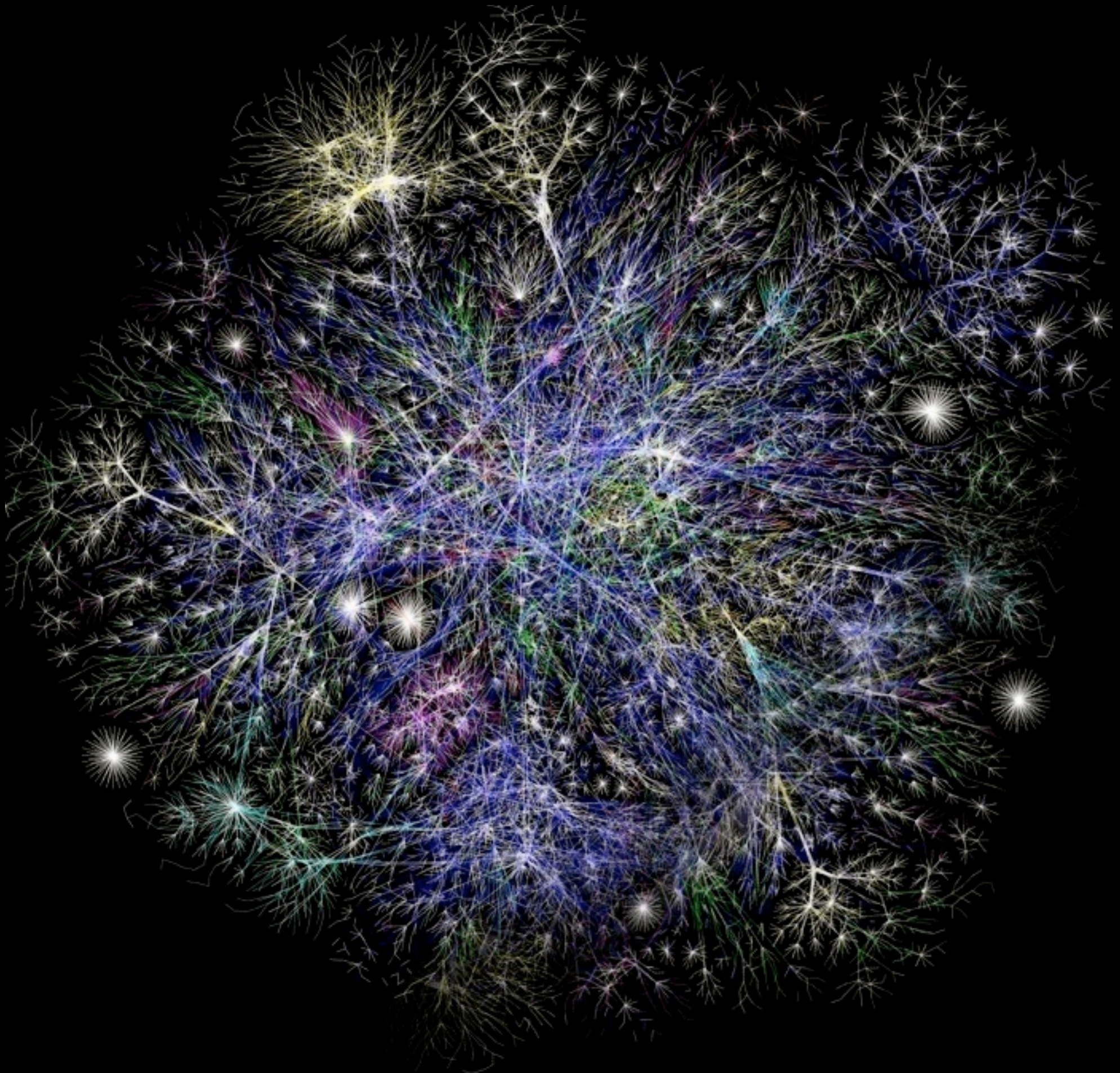
Graph	Nodes	Edges
YahooWeb	1.4 Billion	6 Billion
Symantec Machine-File Graph	1 Billion	**37 Billion**
Twitter	104 Million	3.7 Billion
Phone call network	30 Million	260 Million

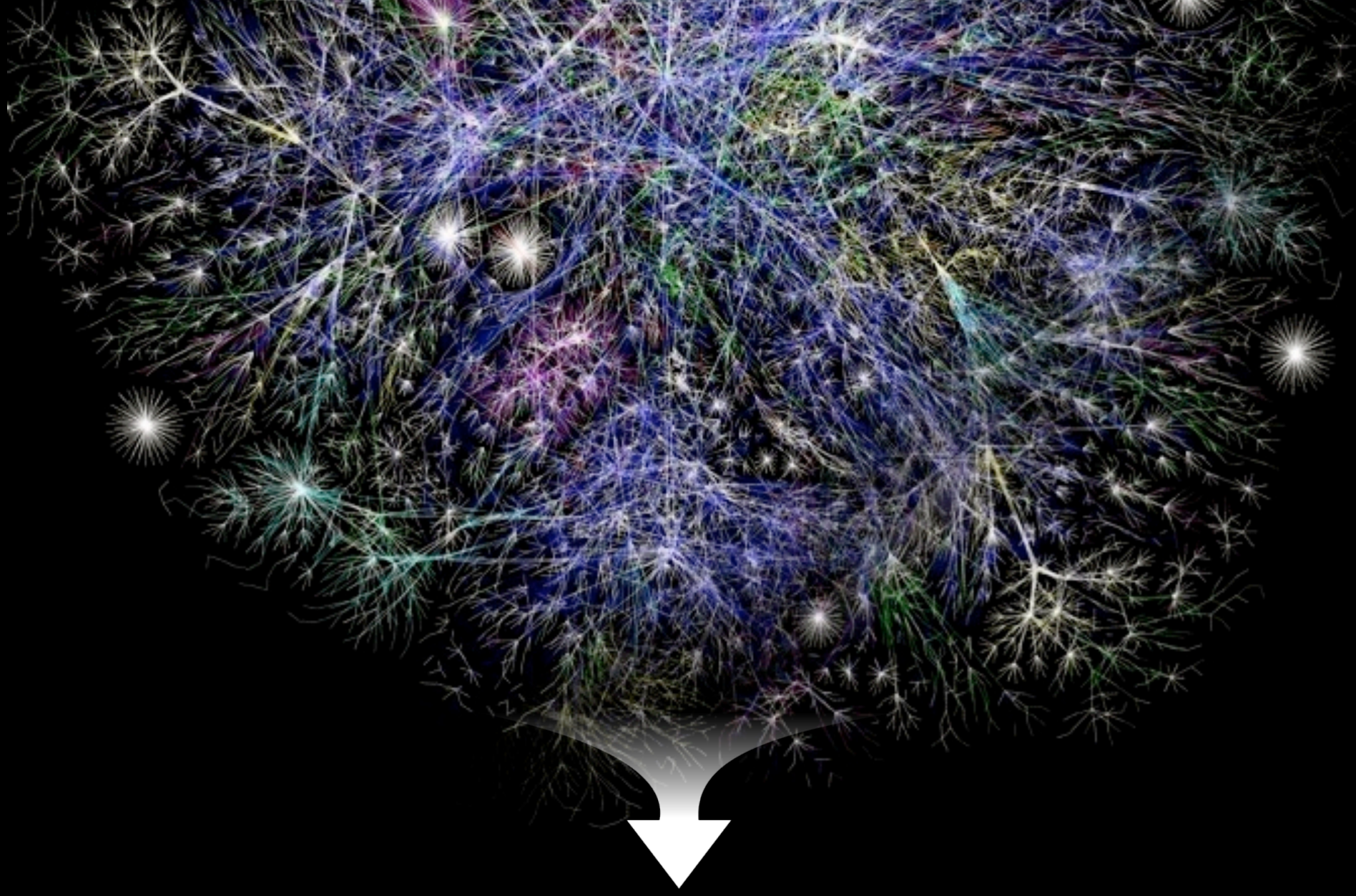
7

7±2

Number of items an average human holds in working memory

George Miller, 1956





7

Data

Insights

How to do that?

COMPUTATION + VISUALIZATION

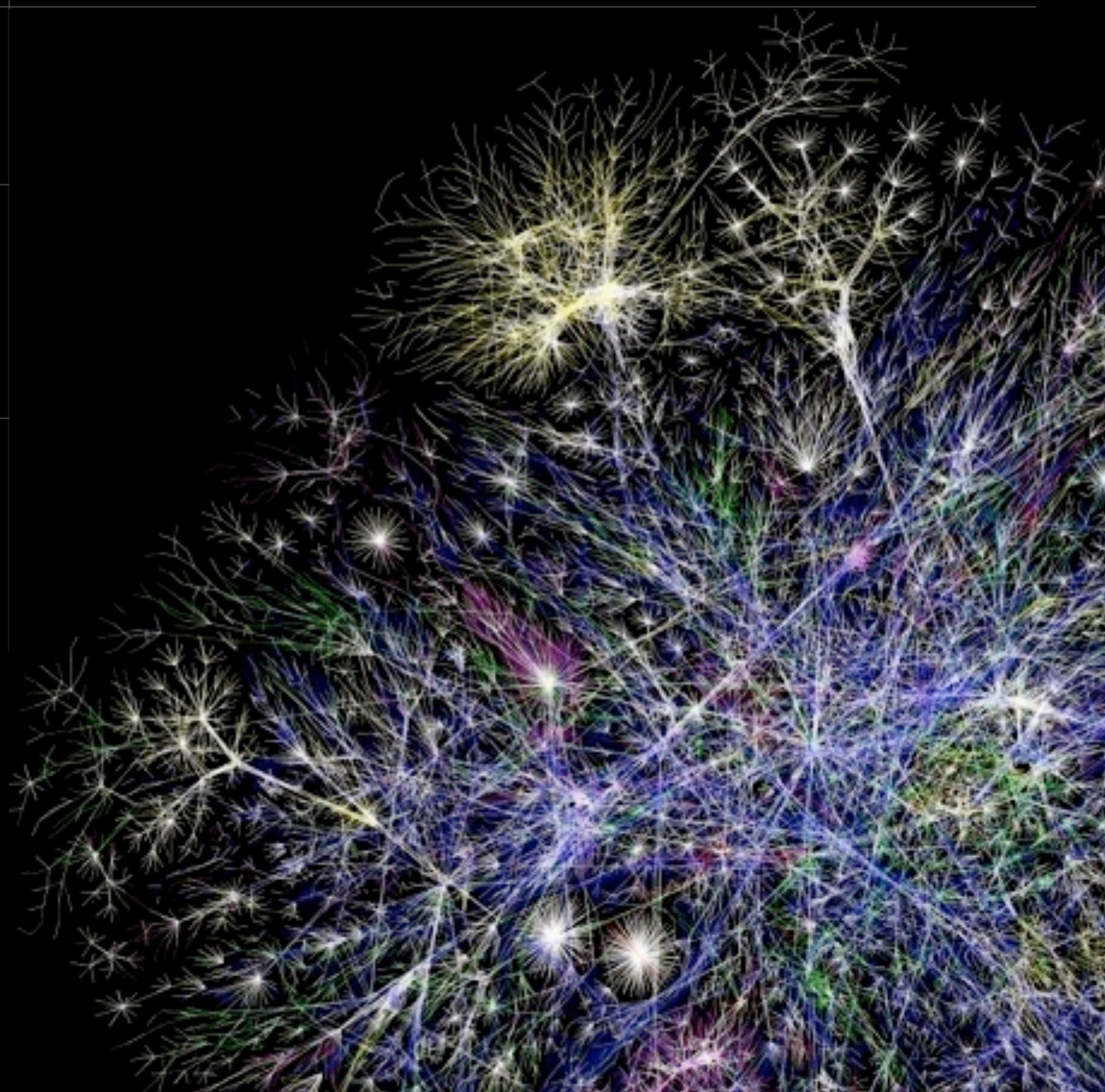
How to do that?

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

Both develop methods for making sense of network data

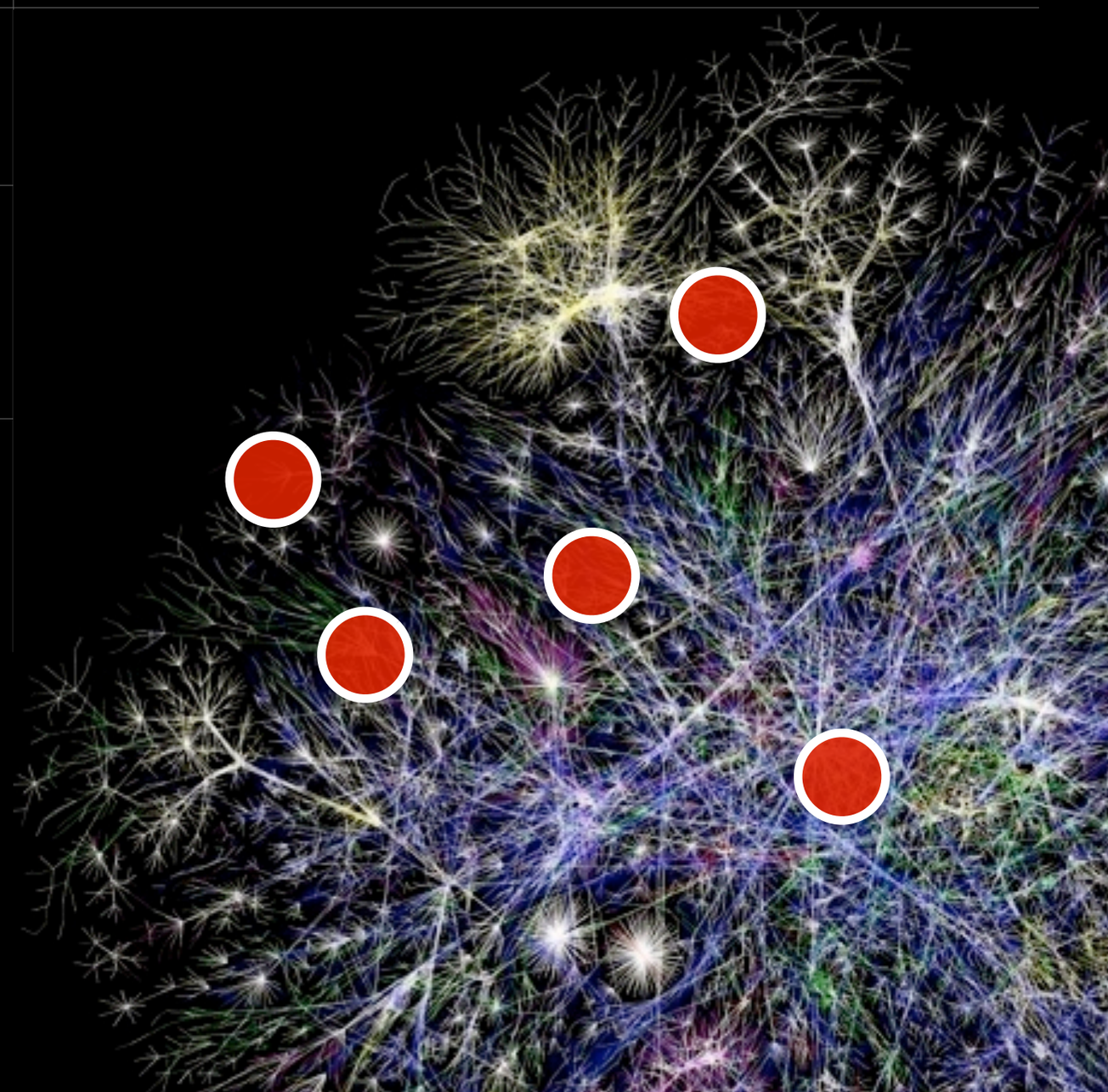
How to do that?

COMPUTATION	VISUALIZATION
Automatic	
Summarization, clustering, classification	
>Millions of nodes	



How to do that?

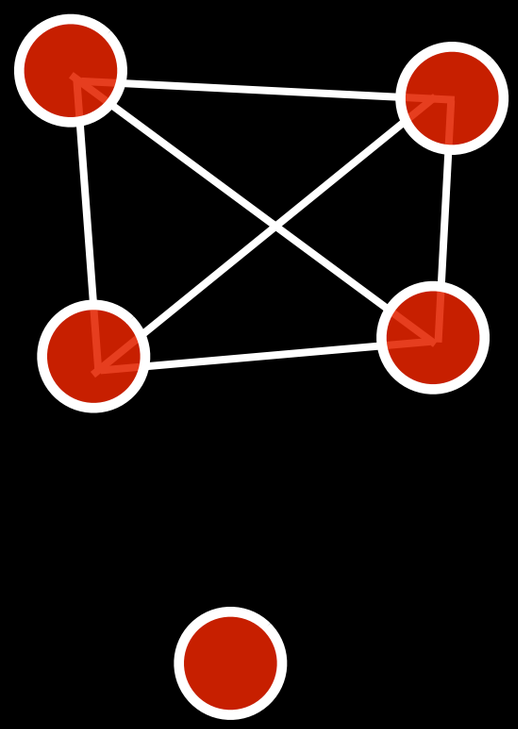
COMPUTATION	VISUALIZATION
Automatic	
Summarization, clustering, classification	
>Millions of nodes	



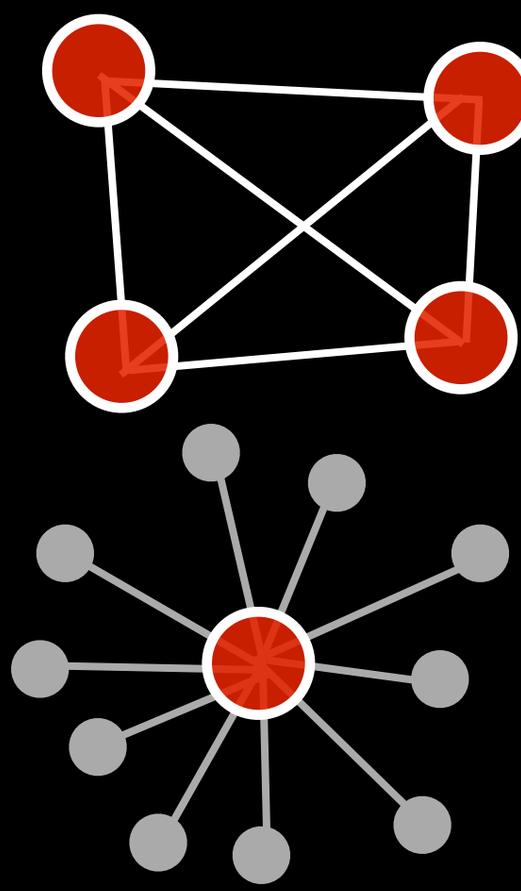
How to do that?

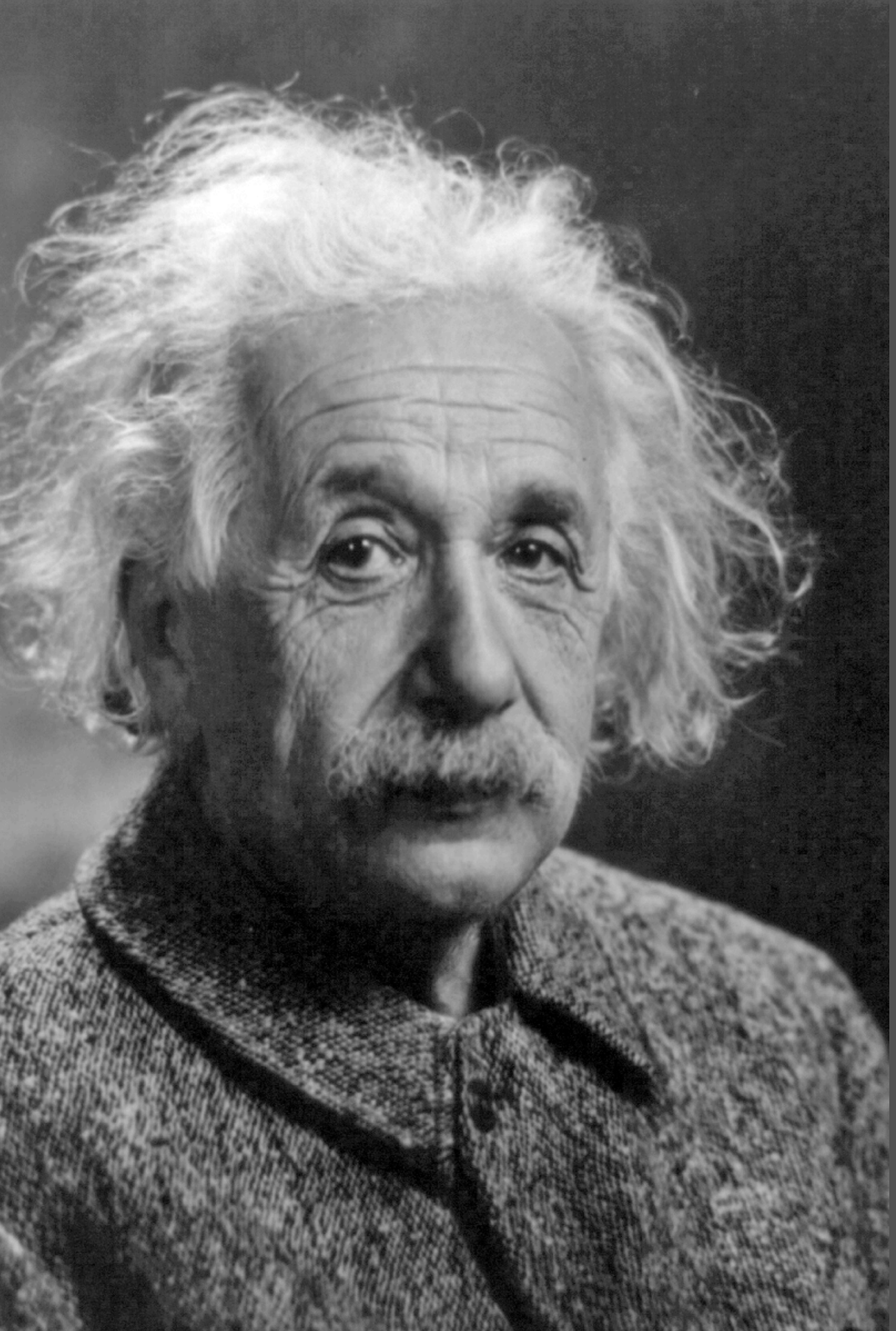
COMPUTATION	VISUALIZATION
	User-driven; iterative
	Interaction, visualization
	Thousands of nodes

How to do that?

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

How to do that?

COMPUTATION	VISUALIZATION
	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."

Logistics

Course homepage	poloclub.gatech.edu/cse6242/
Discussion, Q&A, find teamates	Piazza (link on homepage)
Submission	T-Square

Course Goals

- Learn a broad class of **scalable** visual and computation techniques and tools, for typical data types
- Learn how to **combine** both kinds of methods (how they complement each other)
- Gain **practical** know-how
- Gain **breath** of knowledge

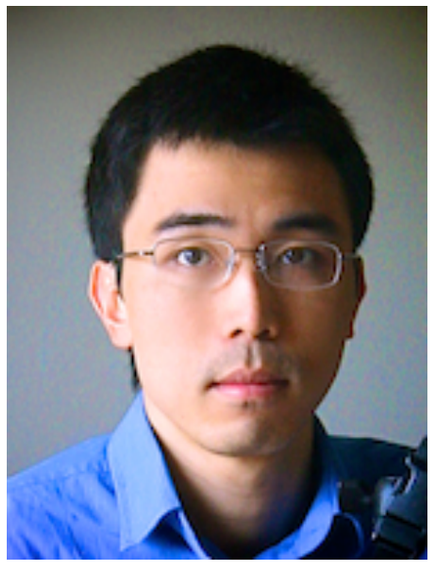
Course Expectation

- Overview of scalable visual and computation techniques and tools
- Gain knowledge & experience (useful for jobs, research)
- Experience with designing and developing an interactive analysis tool

Schedule

See course homepage

Course Staff



Instructor
Duen Horng (Polo) Chau
Assistant Professor, CSE
Thu 3-4pm, Klaus 1324

TA
Parikshit Ram
PhD student, CSE
Mon 4-5pm, Klaus 1315

Grading

- 3 homework assignments (40%)
  - End-to-end analysis
  - Techniques (computation and vis)
  - Hadoop (+ other "big data" tools)
- Group project (50%) -- 2 to 3 people
- Participation (10%) -- in class, on Piazza