

# Course Introduction and Overview

Internet Analytics (COM-308)

Prof. Matthias Grossglauser  
School of Computer and Communication  
Sciences

**EPFL**

# Logistics

- Lectures on Mon 15:15-17:00h
- Moodle module “Internet Analytics”: slides, lab handouts & deliverables ([moodle.epfl.ch](https://moodle.epfl.ch))
- Labs: Thu afternoon 16:15-18:00h
  - CM1111/1112 (BYOD ok)
- Live Q&A homeworks & lectures
  - Thu after lab 18:15-19:00h, back in BC01
- Forum on lectures & labs ([Ed](#))
- Action items:
  - **By end of next week**: Groups of 4 - self-organize & register in moodle
  - **By Tuesday 23:59h**: Spark account setup: please register for the class in IS-academia

# Team

- Instructor:
  - Matthias Grossglauser
- Assistants:
  - Edouard Dufour
- Team of student-assistants:
  - Nika Horvat
  - Timael Andrié

Edouard



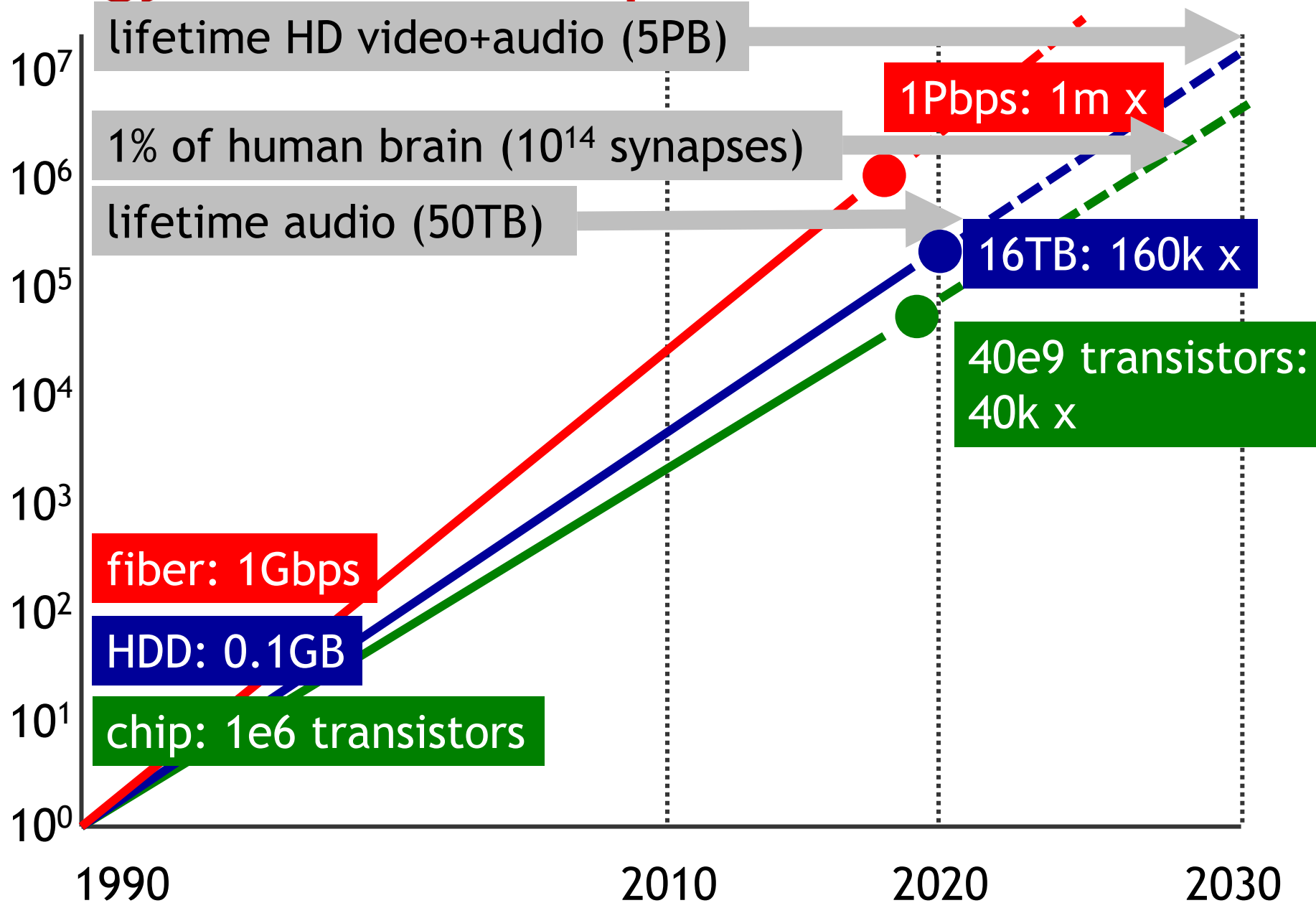
Potential mystery TA



# Grading

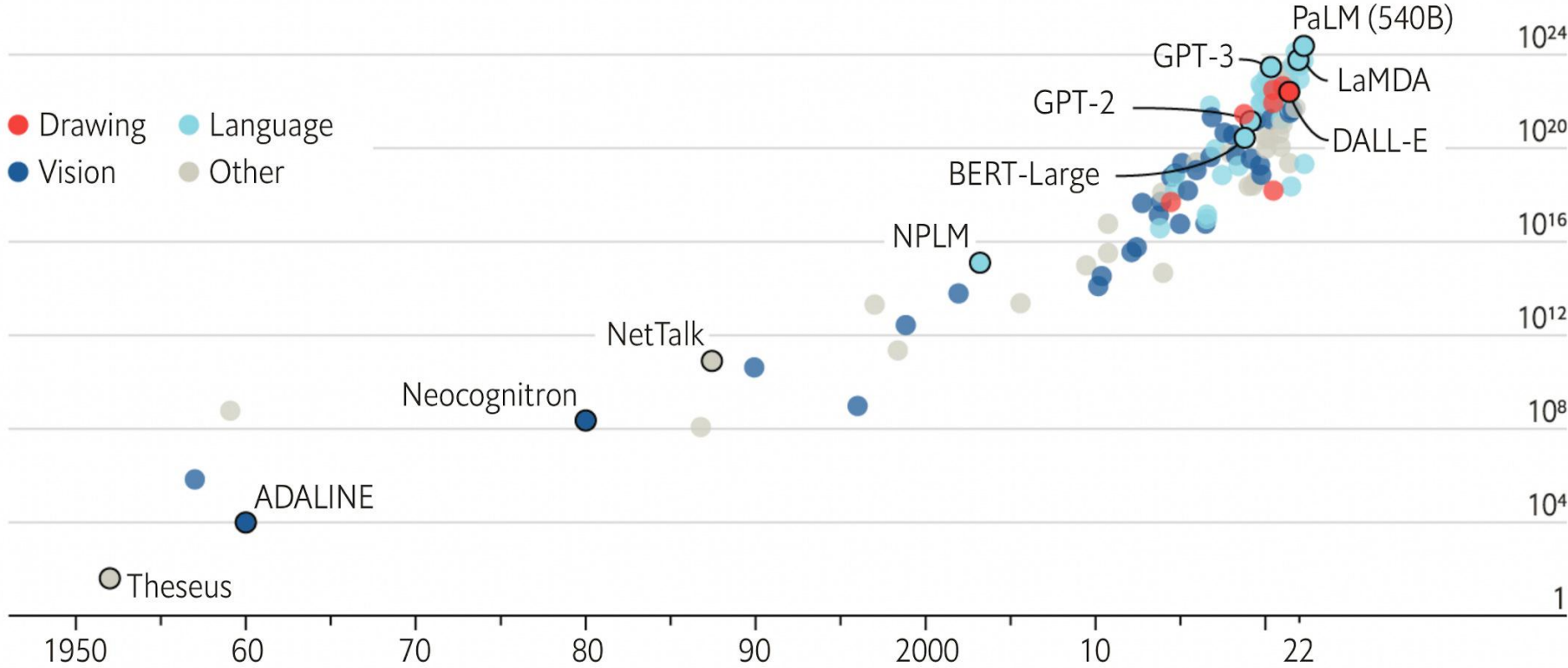
- Final exam:  $x\%$
- Labs:  $(100-x)\%$ 
  - Lab 0: not graded
  - Hand-ins (code, plots, interpretation, ...)
  - Deadline for lab  $i$  usually at the start of lab  $(i+1)$
  - Selected mini-interviews with whole team to go over your deliverables and to check your understanding
- Homeworks: not graded
  - But strongly encouraged to do them regularly

# Technology evolution: exponential trends



# Training cost for largest ML models

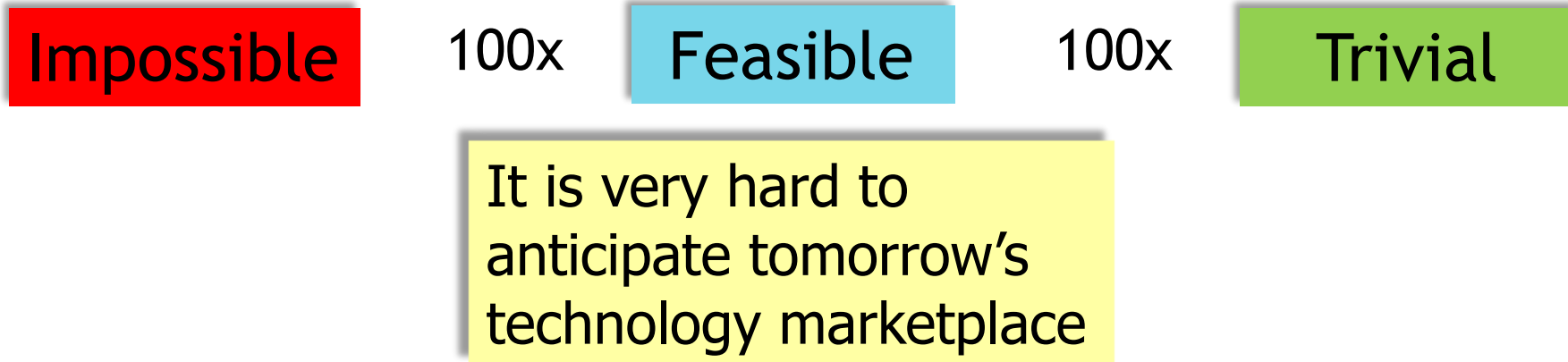
Floating-point operations, selected systems, by type, log scale



Sources: "Compute trends across three eras of machine learning", by J. Sevilla et al., arXiv, 2022; Our World in Data

- Note:  $10^{24}$  ns = 31 million years!

# Exponential technology evolution



- Technology fundamentals
  - Several surprising exponential scaling laws
    - Stable and predictable over decades
  - What applications in 10-20 years: no idea
  - Clear trend: measuring, storing, analyzing everything!

# Limited prediction horizons

"This 'telephone' has too many shortcomings to be seriously considered as a means of communication." – Western Union

"Computers in the future may have only 1,000 vacuum tubes and perhaps only weigh 1 1/2 tons." – Popular Mechanics

"There is no reason for any individual to have a computer in their home." – Ken Olson, President DEC

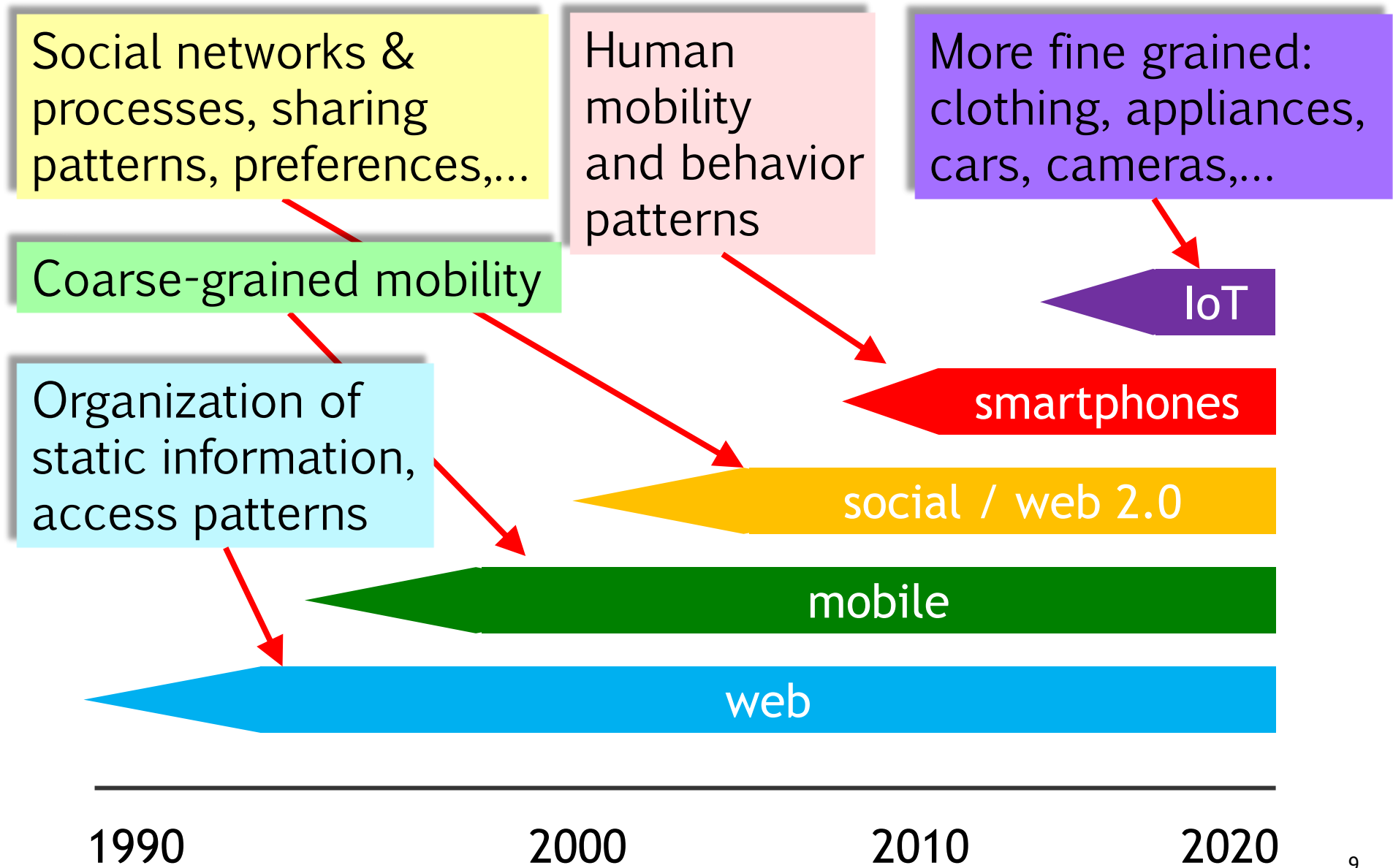
1850

1900

1950

2012

# Evolution of interfaces & user data



# Example IoT: tracking via Apple AirTags

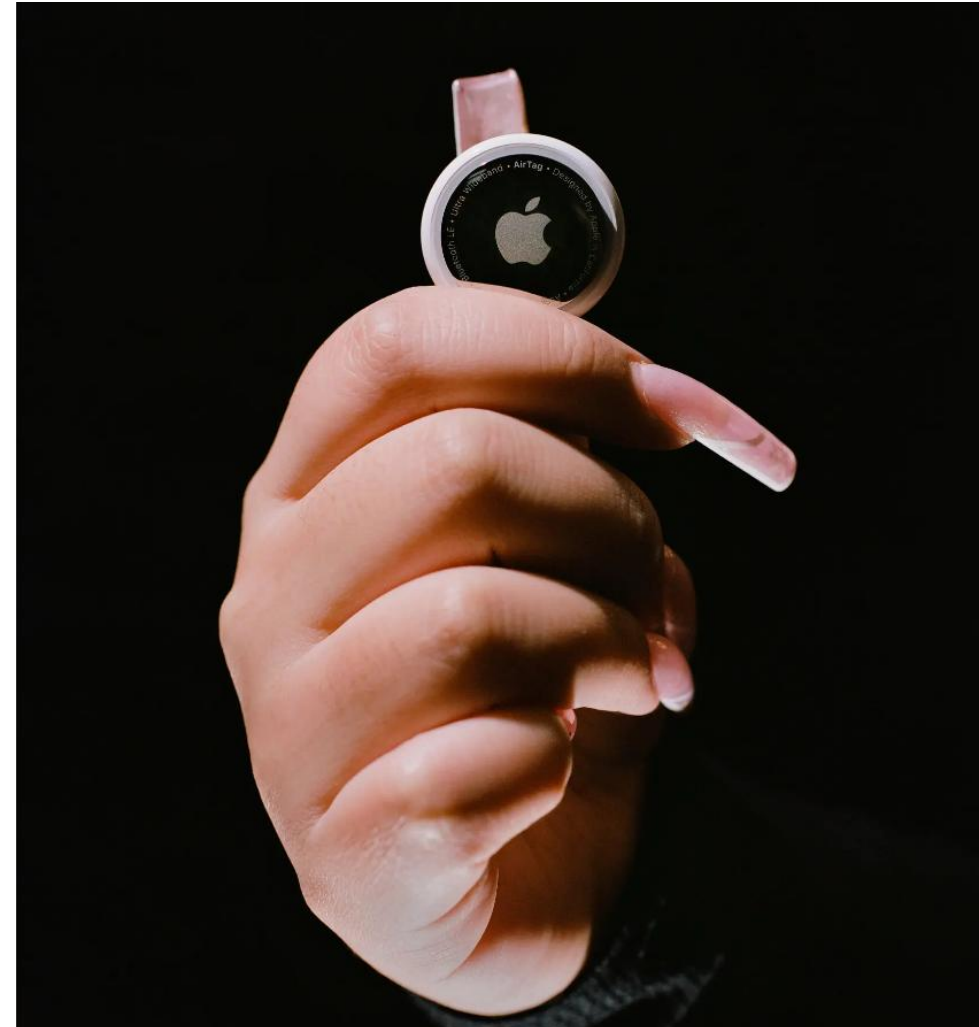


The New York Times

Account ▾

## *Are Apple AirTags Being Used to Track People and Steal Cars?*

Privacy groups sounded alarms about the coin-sized location-tracking devices when they were introduced. Now people are concerned those fears are being realized.



# Example IoT: Clearview AI

Le Temps  
1209 Genève  
022 575 80 50  
<https://www.letemps.ch/>

Medienart: Print  
Medientyp: Tages- und Wochenpresse  
Auflage: 35'370  
Erscheinungsweise: 6x wöchentlich

Seite: 9  
Fläche: 49'647 mm<sup>2</sup>

Auftrag: 3007101  
Themen-Nr.: 999.222

Referenz: 83456893  
Ausschnitt Seite: 1/2

## Aspirer 100 milliards de visages, le projet délirant de Clearview AI



Clearview AI affirme être à bout touchant pour posséder dans ses bases de données 100 milliards de visages d'ici un an.

(ARCHIVE/ LESZEK LATA)  
ANOUCHE SEYDTAGHIA  
@Anouch

**DONNÉES** L'entreprise new-yorkaise, pourtant visée par de nombreuses plaintes, poursuit le développement d'outils de reconnaissance faciale. Elle veut désormais atteindre les 100 milliards

de nouveaux services. Les risques de dérive sont déjà avérés. Le nom de Clearview AI avait émergé il y a deux ans, lorsqu'une enquête du *New York Times* avait révélé la technologie développée

Dans les jours qui avaient suivi, face à la colère suscitée par ces pratiques, la plupart des géants de la tech – hormis Amazon – avaient annoncé arrêter leurs projets de reconnaissance faciale. Mais en

# What is this class about?

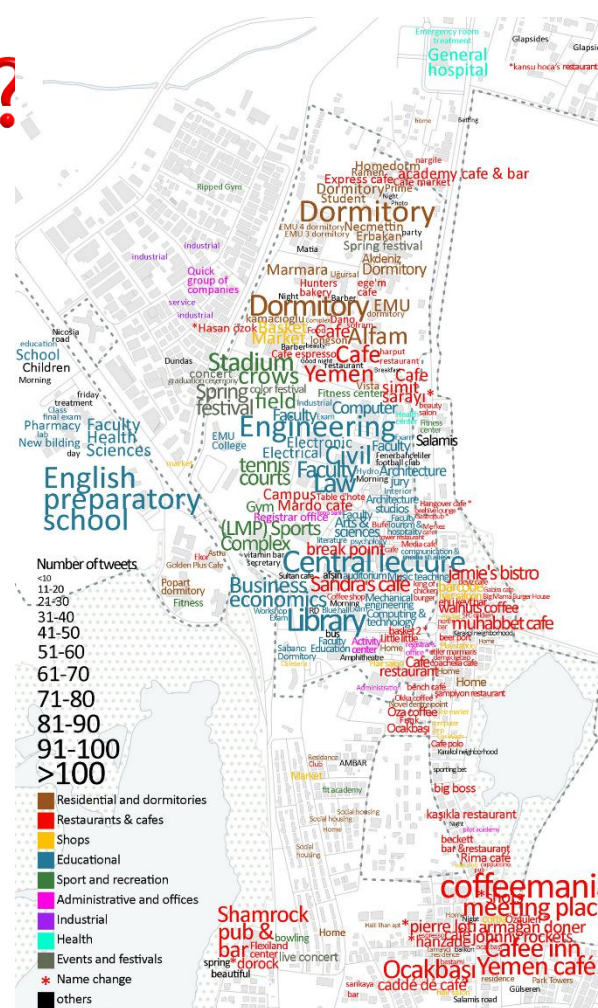
- Data growth: social networking and social media; messaging; location & navigation; smart things;...
- Pre-cloud engineering challenges:
  - **Networking**: connectivity, routing, traffic engineering, intrusion detection,...
  - **Data-center design**: databases, server farms, traffic monitoring, energy, hosting,...
- Post-cloud engineering challenges:
  - **Cloud**: outsourcing of many “lower-layer functions”: connectivity, storage, computing, security,...
  - **Data science**: drives user functionality & user experience; monetization (ads, etc.); management (capacity, etc.),...

# What is this class about?

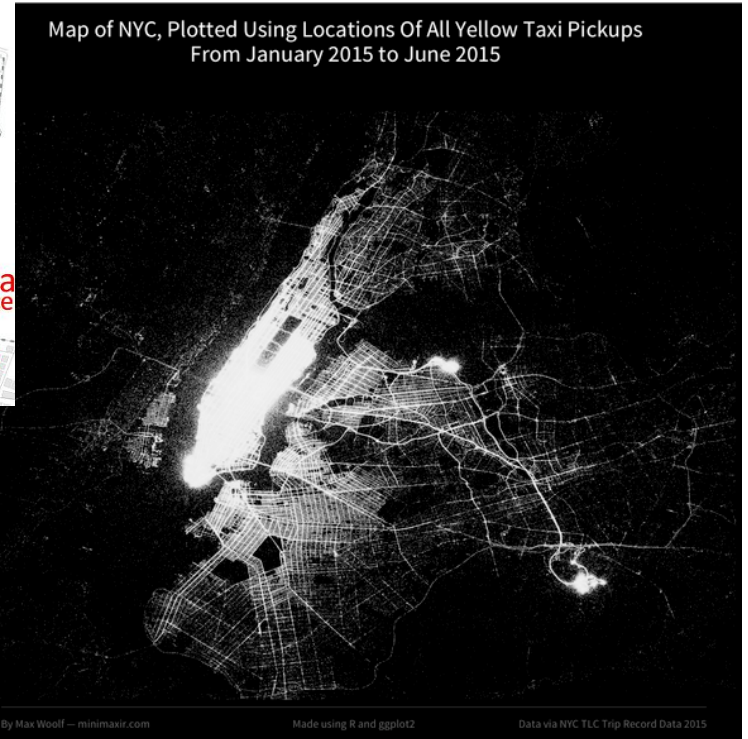
- 
- Real data:
    - Degree of realism
    - Interesting in its own right
    - Real-world challenges: gaps, errors, lack of documentation,...

# What is this class about?

- Many sources of data → room for creativity in how to answer a question!
- Examples:
  - Traces of geo-tagged + timestamped tweets → characterize the neighborhoods in a city (residential, financial district, bars&clubbing, restaurants, artsy,...)
  - Taxi GPS traces in NYC → estimates of financial performance of retail stores



[Iranmanesh, Cömert, Hoskara: Reading urban land use through spatio-temporal and content analysis of geotagged Twitter data, Springer Nature, 2021]



[reddit]

# What this class is *not* about?

- Little on business models, policy & legal issues
- Little coverage of systems issues
  - Cloud architectures, energy,...
- Not an exhaustive ML class
  - E.g., no self-contained treatment of “classical” topics like regression, SVM, deep learning - but introduced as needed
  - No LLMs (but word embeddings, key ingredient)
- Criterion in choice of topics: real-world relevance of...
  - Data & models
  - Functions & applications

# Matrix of data and functions

	Networks	Ratings	Document	Corpus	Streams
Characterize Model	Small worlds, scale-free			Topic models	Counts, moments
Predict Infer	Link prediction				
Rank	PageRank	Collaborative filtering	Content-based recommend		
Filter			Spam filtering		
Search Retrieve				Keyword search, similar docs	
Associate Summarize	Community detection			Clustering	Random projections

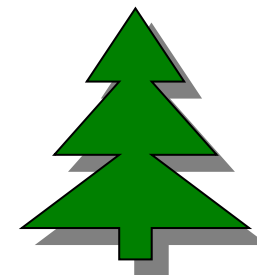
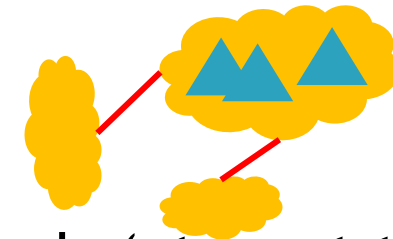
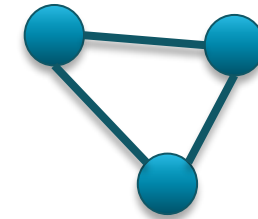
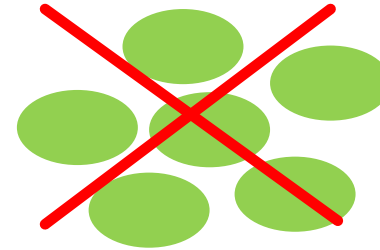
# Overview: social and info networks

- Social and information networks
  - How are people connected?
  - How is information connected?



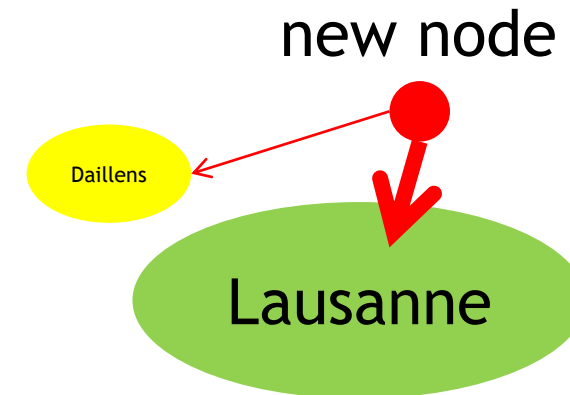
# Networks: structure

- Structure:
  - Single snapshot in time
  - Generic properties?
- 1: Giant Component
  - Almost all pairs of nodes are reachable
- 2: Clustering
  - Many triangles
- 3: Strong and weak ties
  - Interconnected sub-communities
  - Your acquaintances are more important than your friends (at least to find a job ;-)
- 4: Short distances
  - Six degrees of separation



# Networks: evolution

- Evolution:
  - How does the network change over time?
- Herding behavior:
  - We tend to copy behavior of those around us
  - Benefitting from each others' decisions
- 1: Information cascades
  - Sequences of wrong decisions
- 2: Preferential attachment
  - Skewed degree distribution
  - The rich get richer
- 3: Friendship paradox
  - Your friends have more friends than you

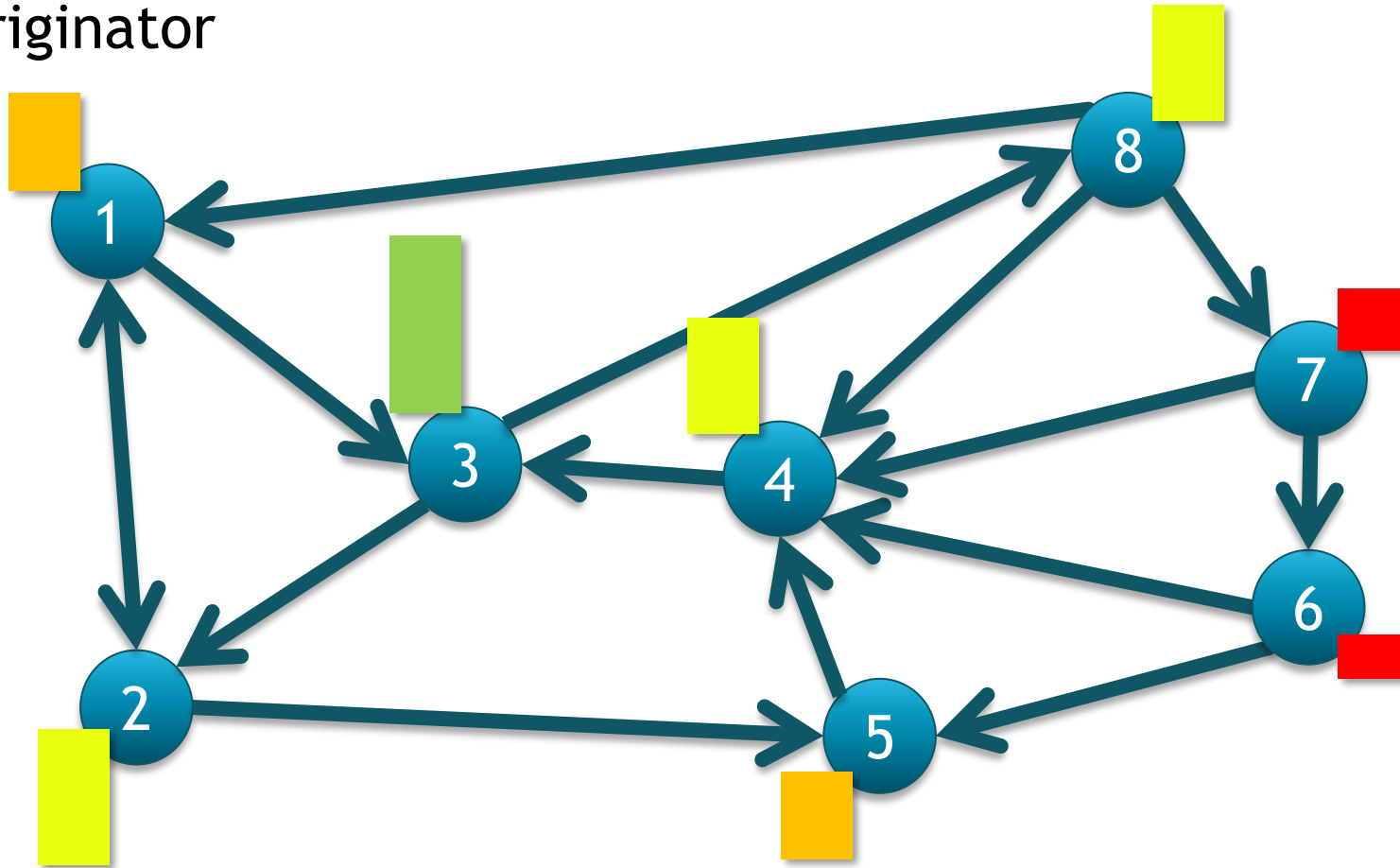


# Networks: processes

- Processes:
  - Nodes have state and influence each other
  - How does net structure influence processes?
- 1: Epidemics
  - How does a rumor spread through a social network?
  - How does a disease progress, and who should be vaccinated?
- 2: Sampling
  - Very large network: how to estimate properties without visiting all the nodes?

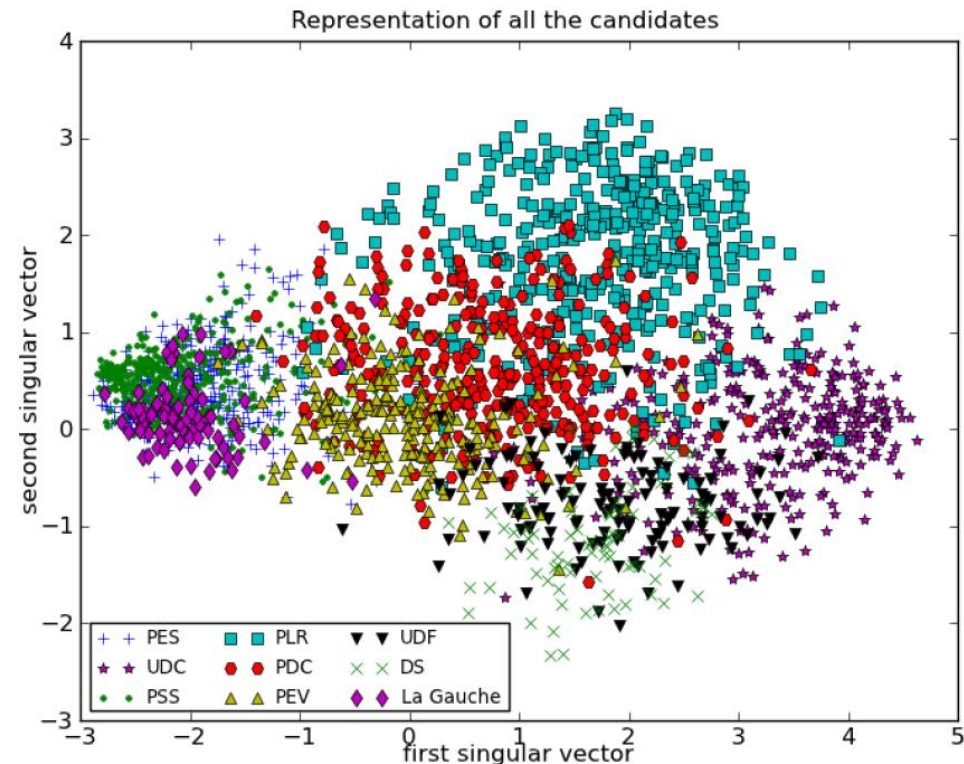
# Overview: relevance and filtering

- PageRank:
  - A hyperlink “endorses” the target
  - An endorsement depends on the “relevance” of the originator



# Overview: dimensionality reduction

- Raw data:
  - Often high-dimensional
  - But has “structure” = low-dimensional signal + noise
- Challenge:
  - How to find low-dimensional structure?
- Applications:
  - Visualizing
  - Explaining
  - Modeling
- Example:
  - SmartVote dataset on political candidates



# Overview: recommender systems

**amazon** [Join Prime](#) [matthias's Amazon.com](#) | [Today's Deals](#) | [Gift Cards](#) | [Help](#)

Shop by **Department** ▾ Search

[Your Amazon.com](#) | [Your Browsing History](#) | [Recommended For You](#) | [Amazon Betterizer](#) | [Improve Your Recommendations](#) | [Your Profile](#) | [Learn More](#)

[Your Amazon.com](#) > **Recommended for You**  
(If you're not matthias grossklauser, click here.)

These recommendations are based on [items you own](#) and more.

## Just For Today

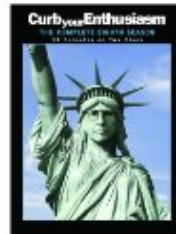
[Browse Recommended](#)

## Recommendations

- [Amazon Instant Video](#)
- [Amazon MP3 Store](#)
- [Appliances](#)
- [Appstore for Android](#)
- [Arts, Crafts & Sewing](#)
- [Automotive](#)
- [Baby](#)
- [Beauty](#)
- [Books](#)
- [Books on Kindle](#)
- [Camera & Photo](#)
- [Cell Phones & Accessories](#)
- [Clothing & Accessories](#)
- [Computers](#)
- [Electronics](#)
- [Grocery & Gourmet Food](#)
- [Health & Personal Care](#)
- [Home & Kitchen](#)
- [Home Improvement](#)

view: [All](#) | [New Releases](#) | [Coming Soon](#)

1.



### [Curb Your Enthusiasm: The Complete Eighth Season](#)

DVD ~ Larry David (June 5, 2012)  
Average Customer Review: [★★★★☆](#) (88)  
In Stock

List Price: ~~\$39.98~~

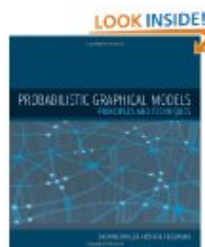
Price: **\$25.77**

[16 used & new from \\$25.77](#)

I own it  Not interested  [★★★★☆](#) Rate this item

Recommended because you purchased [Curb Your Enthusiasm: The Complete Sixth Season](#) and more ([Fix this](#))

2.



### [Probabilistic Graphical Models: Principles and Techniques \(Adaptive\)](#)

by Daphne Koller (July 31, 2009)  
Average Customer Review: [★★★★☆](#) (17)  
In Stock

List Price: ~~\$99.00~~

Price: **\$95.04**

[81 used & new from \\$64.98](#)

I own it  Not interested  [★★★★☆](#) Rate this item

Recommended because you added [Introduction to Information Retrieval](#) to your Shopping Cart and more ([Fix this](#))

# Overview: recommender systems

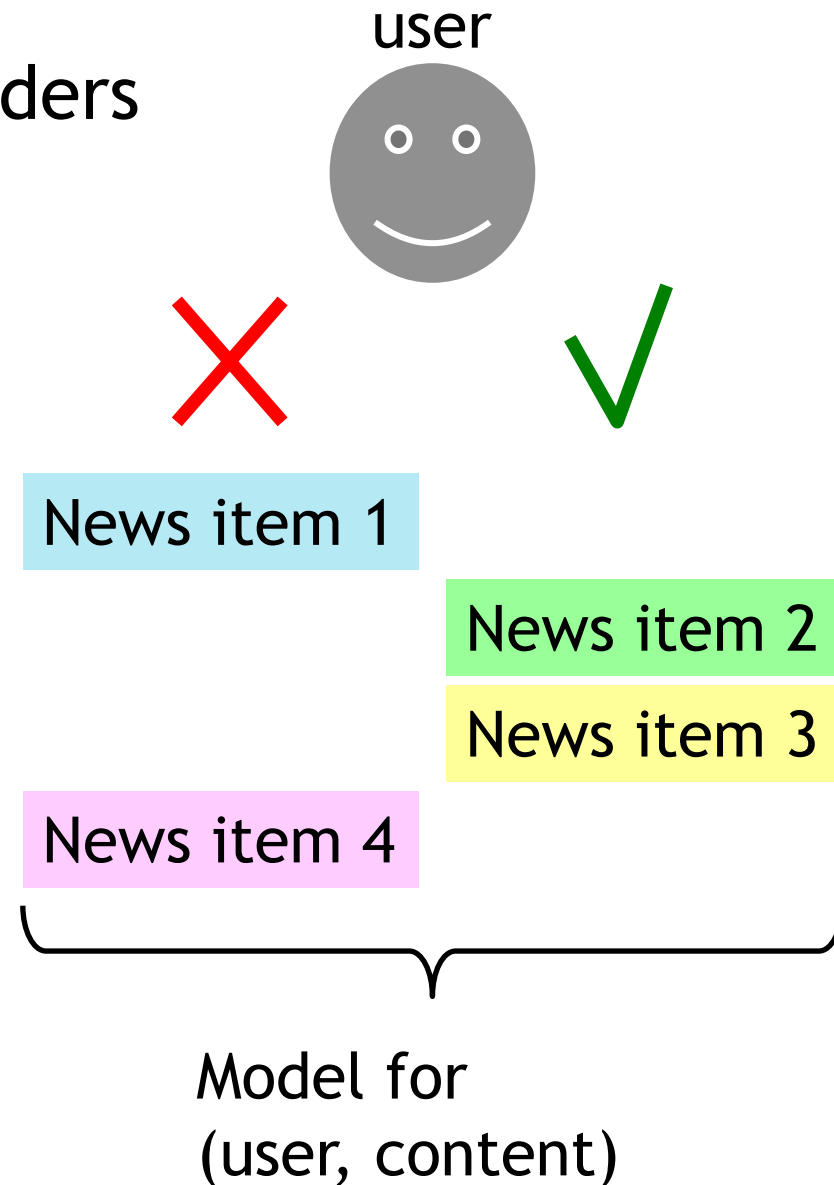
- Content-based recommenders

item 1:  
“Plane hijacked...”

item 2:  
“soccer game...”

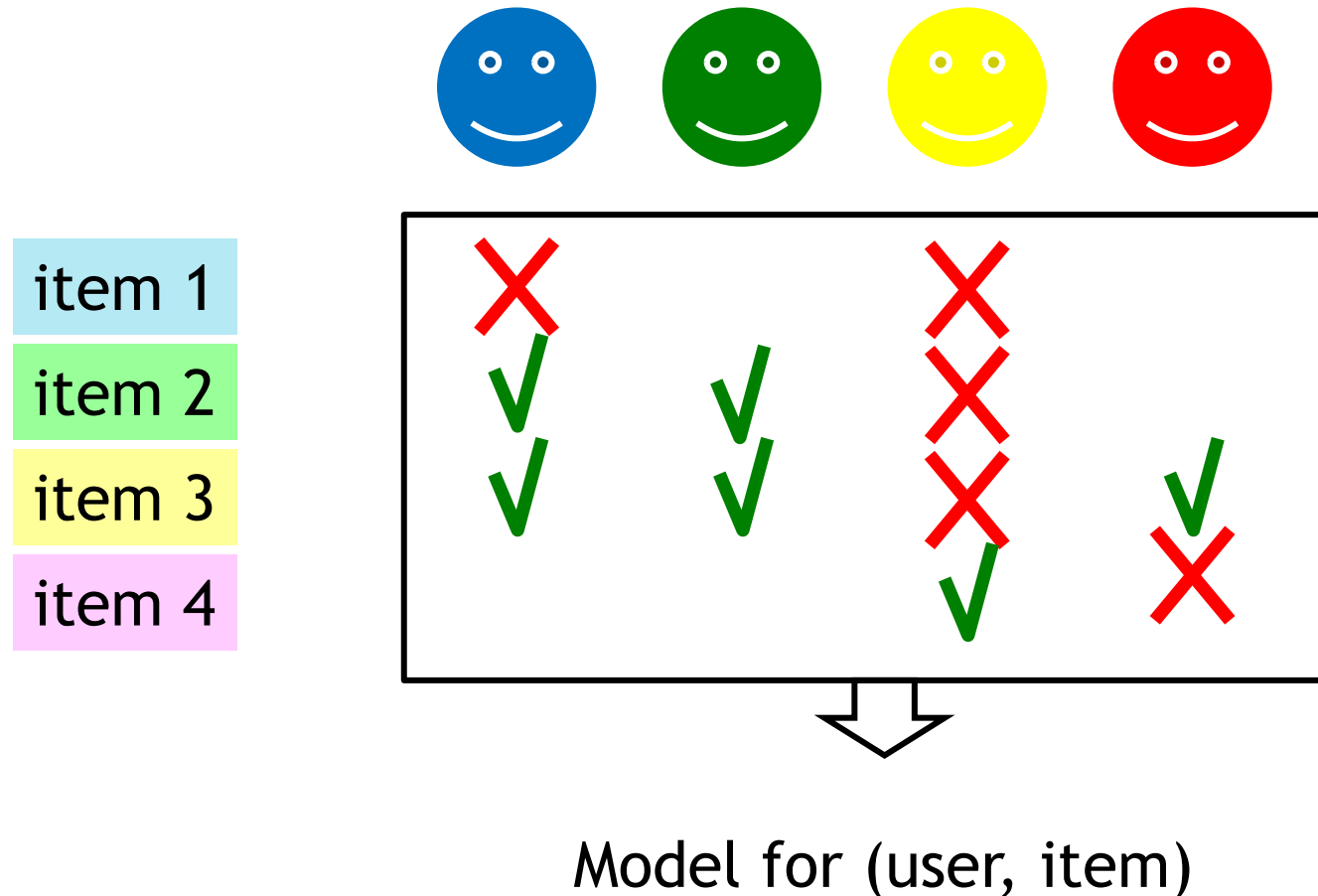
item 3:  
“swiss skiers win...”

item 4:  
“50.3% vote yes...”



# Overview: recommender systems

- Collaborative filtering-based recommenders

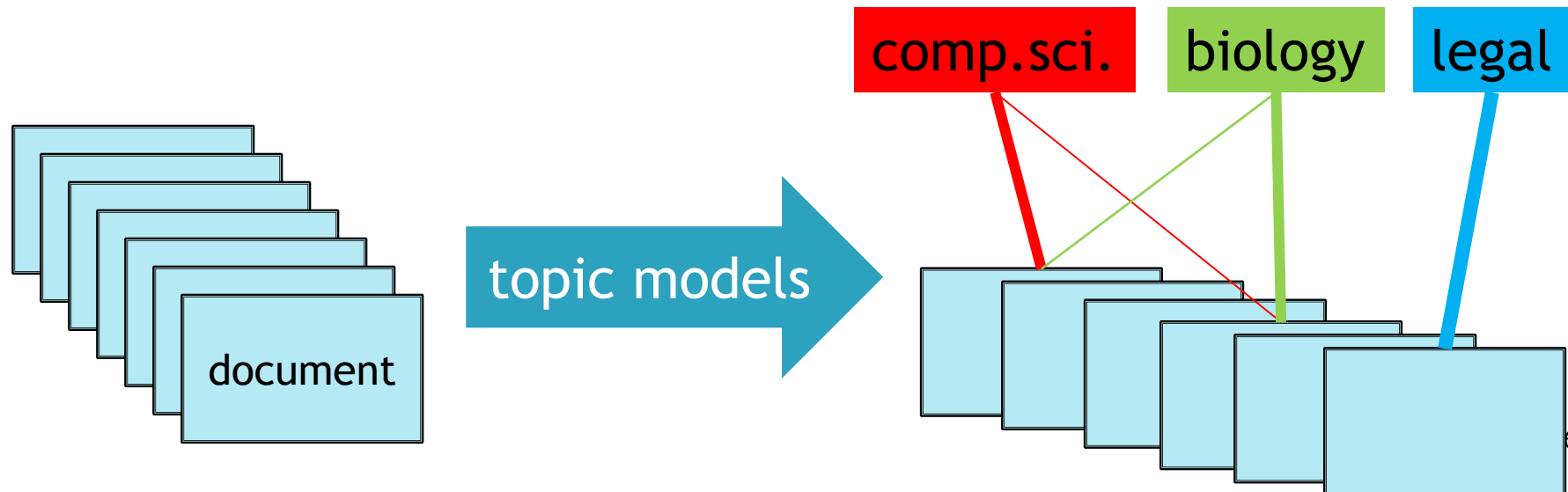


# Overview: search and retrieval

- Given a query, how to find best matches?

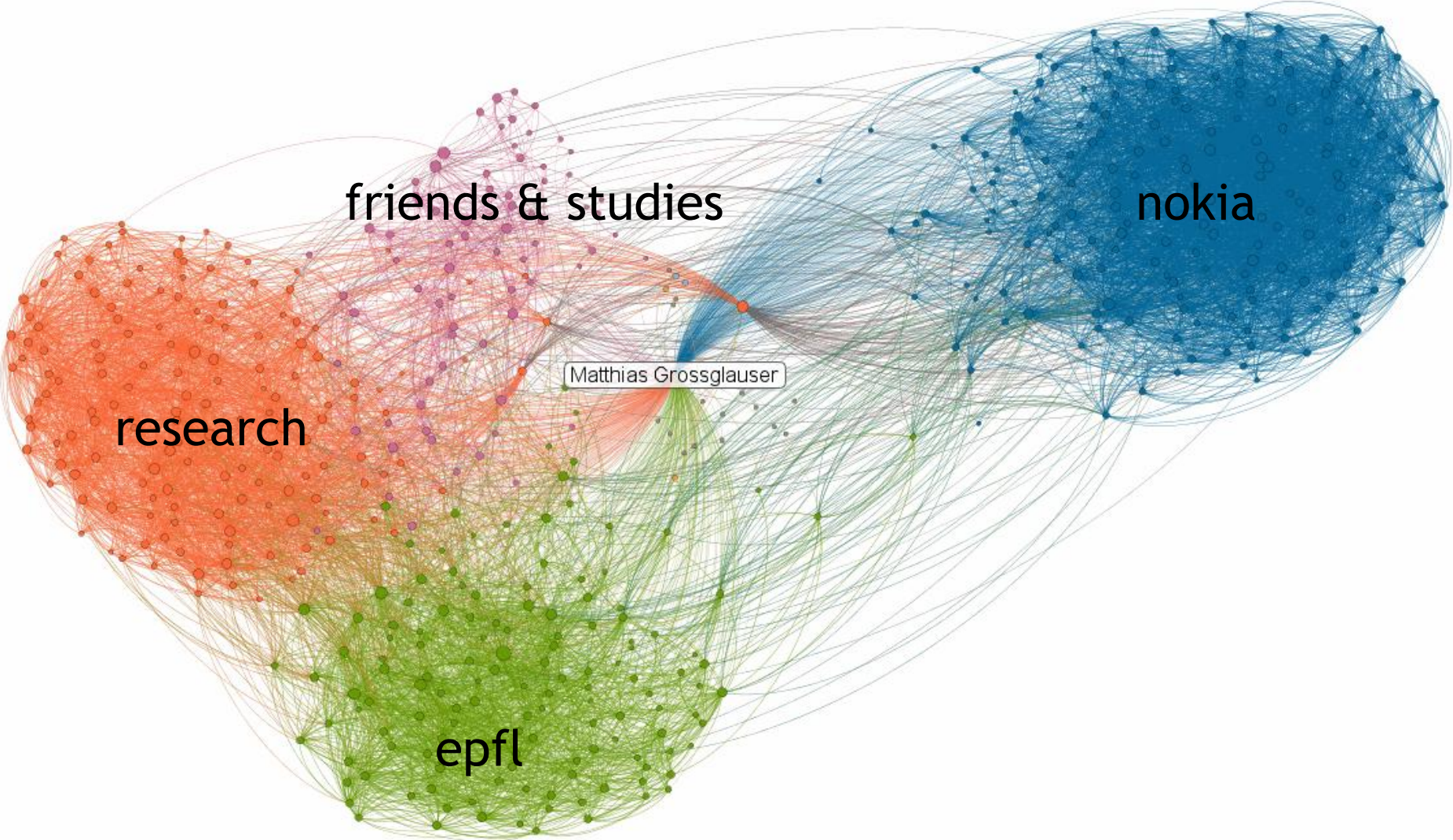


- Without a query, how to describe a corpus?



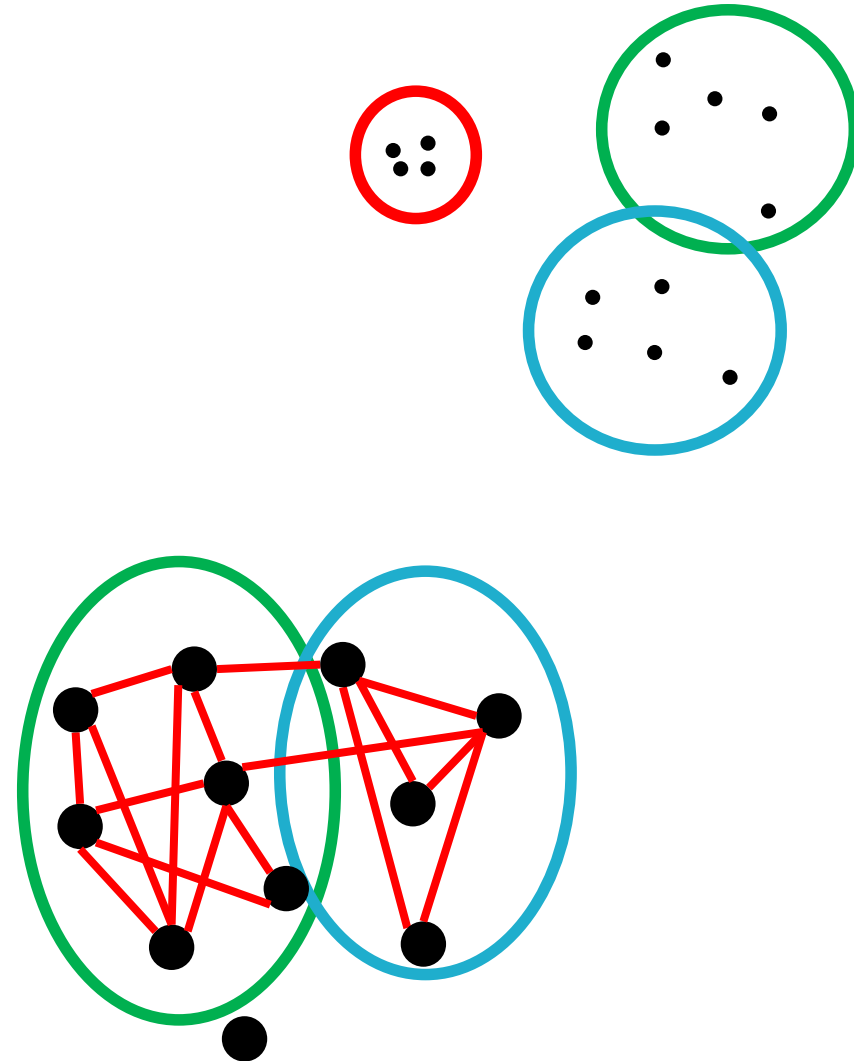
# Overview: clusters, groups, communities

LinkedIn Maps Matthias Grossglauer's Professional Network  
as of February 16, 2013



# Overview: clusters, groups, communities

- Another type of structure
- Cluster:
  - Set of points close to each other, and far from other points
- Community:
  - A set of nodes with more links among them than to other parts of network



# Overview: streams

- Internet backbone router
- Order of magnitude:
  - 100s of interfaces at 10s of Gbps
  - = several billion pkts/sec!
- Traffic analysis app to detect DDoS attack:
  - How many \*different\* (unique) source IP addresses in a minute?
  - If too large -> suspicious (fake addresses)!



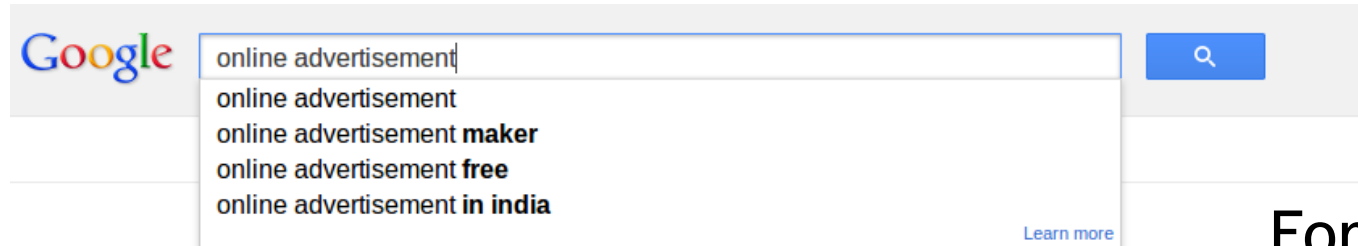
# Overview: streams

- Computing statistics with sub-linear memory
- Example:
  - $n$  numbers: how many unique values  $k$ ?



- How to solve with  $\theta(n)$  memory?
  - Keep every value in some efficient data structure; compare & count
- How to solve with  $o(n)$  memory?
  - Cannot solve exactly
- Streaming algorithms:  $\infty$  data, finite memory
  - Approximation
  - (Pseudo-)randomization

# Overview: ad auctions



Ads related to **online advertisement** ⓘ

**Online Email Advertising - VerticalResponse.com**  
[www.verticalresponse.com/FreeTrial](http://www.verticalresponse.com/FreeTrial)  
Trusted by 500k+ for Reliable Email Marketing Since 2001. Try it Free!  
238 people +1'd or follow VerticalResponse  
Flexible Pricing Options - Features - Pay As You Go - Start Free Trial Now

**Online Advertising - Create your online marketing strat.**  
[www.agencevirtuelle.com/OnlineMarketing](http://www.agencevirtuelle.com/OnlineMarketing)  
AdWords, SEO, Mobile, Social Media.

For each search, this table of “sponsored search results” is the result of an online auction

[Online advertising - Wikipedia, the free encyclopedia](https://en.wikipedia.org/wiki/Online_advertising)

[en.wikipedia.org/wiki/Online\\_advertising](https://en.wikipedia.org/wiki/Online_advertising)

**Online advertising**, also known as **online advertisement**, internet marketing, online marketing or e-marketing, is the marketing and promotion of products or ...

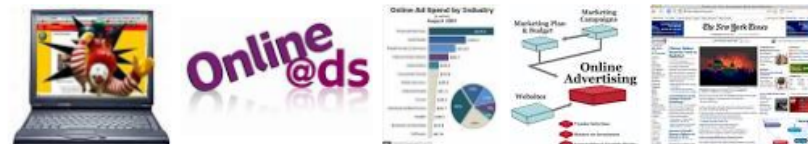
[History of online advertising - Competitive advantage over ... - Online advertisement](#)

[Online Advertising: How to Do It Right | Small Business Trends](http://smallbiztrends.com/2010/11/online-advertising-how-to-do.html)

[smallbiztrends.com/2010/11/online-advertising-how-to-do.html](http://smallbiztrends.com/2010/11/online-advertising-how-to-do.html)

Nov 4, 2010 – Helpful tips on using **online advertising** for small businesses.

[Images for online advertisement - Report images](#)



[Internet advertising: The ultimate marketing machine | The Economist](http://www.economist.com/node/7138905)

[www.economist.com/node/7138905](http://www.economist.com/node/7138905)

Jul 6, 2006 – IN TERMS of efficiency, if not size, the **advertising** industry is only now

# Overview: ad auctions

- Online advertisement:
  - Key business model for the (consumer) internet & mobile
- Keyword auctions:
  - Each request to google -> auction
  - Advertiser bid for (keyword, bid)
- Generalized Second Price (GSP) auction

# Class Schedule (1)

1	Introduction and Overview	<ul style="list-style-type: none"><li>- class overview</li><li>- logistics</li></ul>
2	Social and Information Networks 1: Structure	<ul style="list-style-type: none"><li>- intro: social networks, web, social web</li><li>- social networks, key properties, effects &amp; metrics</li><li>- giant component, relation to tree percolation</li><li>- strong/weak ties, clustering</li><li>- "everything is close": distances; random graphs</li></ul>
3	Social and Information Networks 2: Evolution	<ul style="list-style-type: none"><li>- "the rich get richer": power laws, cumulative advantage, pref attachment</li><li>- my friends have more friends than i-phenomenon</li><li>- link prediction</li></ul>
4	Social and Information Networks 3: Processes	<ul style="list-style-type: none"><li>- processes: epidemics, cascades</li><li>- sampling a network: undirected vs directed</li><li>- random walks on networks, mixing times, spectral properties</li></ul>

# Class Schedule (2)

5	Ranking	<ul style="list-style-type: none"><li>- Intro web structure</li><li>- PageRank algorithm</li><li>- Large-scale computation</li><li>- HITS variant: hubs and authorities</li></ul>
6	Dimensionality Reduction	<ul style="list-style-type: none"><li>- intro: "finding hidden structure", visualization</li><li>- PCA and derivatives</li><li>- singular-value decomposition</li><li>- manifold unwrapping</li></ul>
7	Recommender Systems 1	<ul style="list-style-type: none"><li>- Collaborative filtering</li><li>- graph-based, item-item vs user-user</li><li>- spectral/matrix completion</li><li>- case study: netflix</li></ul>
8	Recommender Systems 2	<ul style="list-style-type: none"><li>- Intro: "the long tail"</li><li>- Applications, models</li><li>- TF-IDF</li><li>- kNN classifier</li><li>- Naïve Bayes classifier</li></ul>

# Class Schedule (3)

9	Clusters and Communities	<ul style="list-style-type: none"><li>- Gaussian Mixture Model (GMM)</li><li>- EM algorithm</li><li>- Communities and modularity measures</li><li>- Louvains clustering algorithm</li></ul>
10	Text Search and Retrieval 1	<ul style="list-style-type: none"><li>- Latent Semantic Indexing (LSI)</li><li>- Intro: Bayesian networks</li><li>- Gibbs sampling</li></ul>
11	Text Search and Retrieval 2	<ul style="list-style-type: none"><li>- Probabilistic LSI (pLSI)</li><li>- Latent Dirichlet Allocation (LDA)</li></ul>
12	Streams	<ul style="list-style-type: none"><li>- Intro</li><li>- motivating apps</li><li>- streaming model</li></ul>
13	Internet Ad Economy & Online Auctions	<ul style="list-style-type: none"><li>- intro: sponsored search, keyword auctions</li><li>- VCG and Generalized Second Price (GSP)</li><li>- incentive compatibility</li></ul>

# Caveats

- Breadth
  - Intersection of data mining, machine learning, network science, statistics, algorithms
- No textbook
  - Combination of several textbooks & other sources
  - The slides + your notes are the course support
- Focus on real applications & data
  - Details often unknown: trade secrets
  - E.g., google practice vs original PageRank
- Lecture/lab overlap
  - Lectures+homeworks: models, theory, background
  - Labs: try it out on real data
  - Overlap is deliberately partial: you learn something new in labs

# Summary

- Problems & data from the real world...
- But enough theory, models to understand the foundations
- Required background:
  - Linear algebra
  - Probability & statistics
  - Algorithms
  - Python
- Next:
  - Thu 16:15h: Tutorial on Spark and using infrastructure
  - No HW session this week yet
- Reminder: sign up for class in ISA by tomorrow!