Course Introduction and Overview

Internet Analytics (COM-308)

Prof. Matthias Grossglauser
School of Computer and Communication Sciences

EPFL

Logistics

- Offline lectures + offline and online fora
- Moodle module “Internet Analytics”: slides, lab handouts & deliverables (moodle.epfl.ch)
- Labs: Wed morning 08:15-10:00h
  - BC07-08 (or BYOD)
- Live Q&A homeworks & lectures
  - Wednesday after lab 10:15-11:00
- Offine forum on lectures & labs (oknoname.com)
- Action items:
  - By end of next week: Groups of 3 (tbc) – 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:
  - Arnout Devos
  - Aswin Suresh
- Team of student-assistants:
  - Lucas Trognon: homework sessions
  - Antoine Magron, Giovanni Monea, Cyrille Pittet: labs

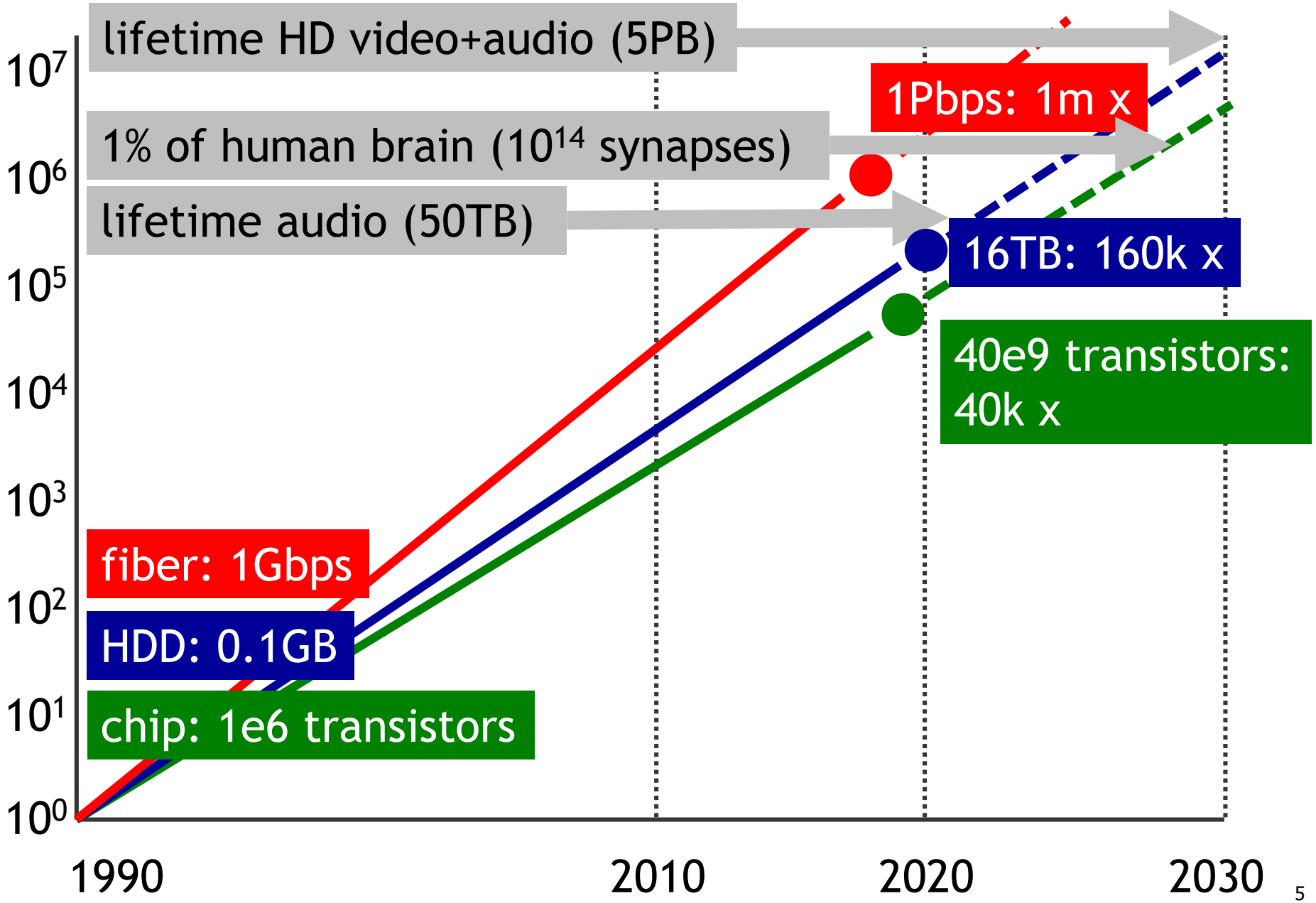
Arnout

Aswin

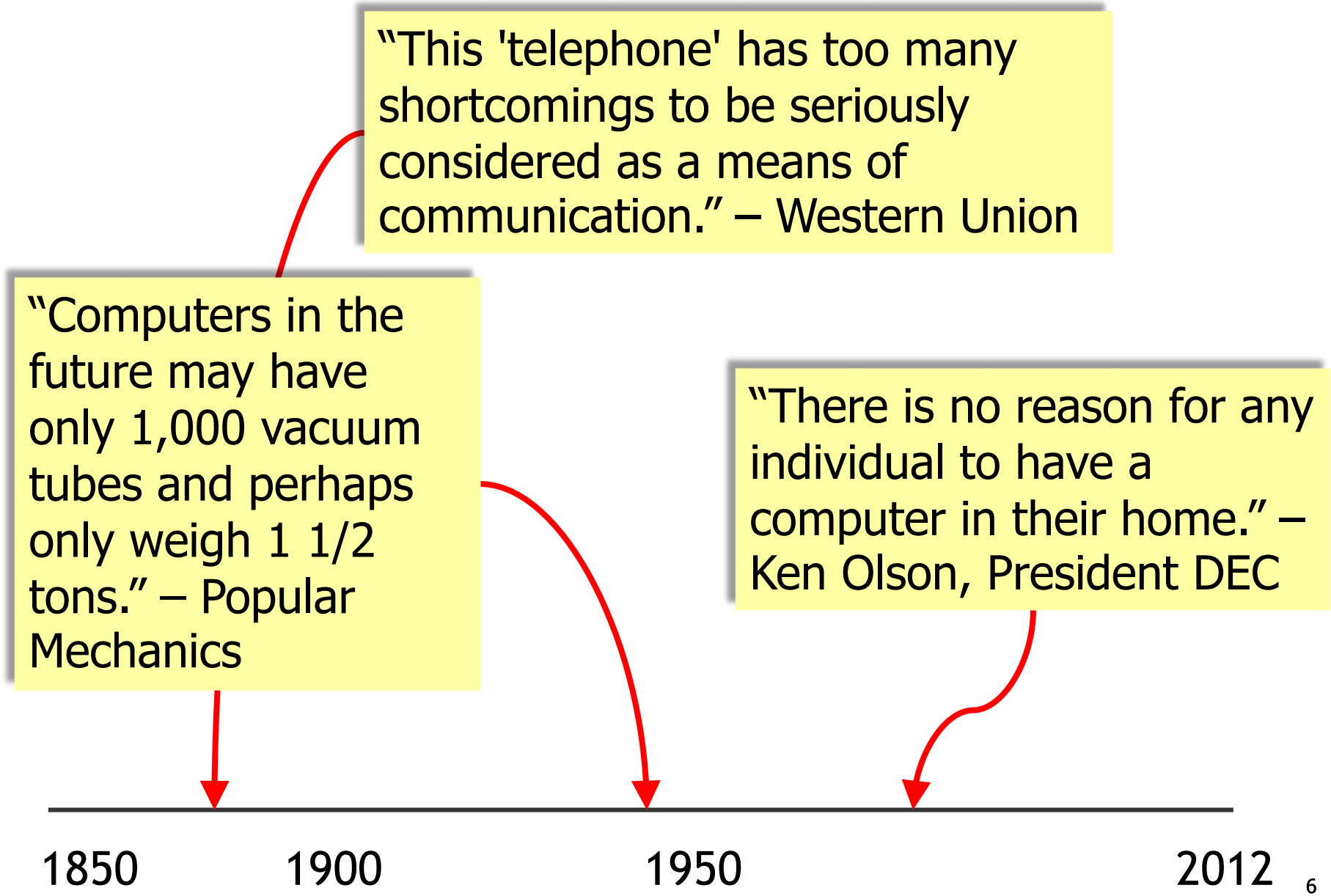
Grading

- Midterm exam: 30%
  - You will get practice midterms and finals from previous years
- Final exam: 40%
  - Last week of semester
- Labs: 30%
  - 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



Limited prediction horizons



Exponential technology evolution

Impossible 100x Feasible 100x Trivial

It is very hard to anticipate tomorrow's technology marketplace

- 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!

Example: Clearview AI

Datum: 21.02.2022

LE TEMPS

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 mm²

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)*

ANOUCH SEYDTAGHIA
@Anouch

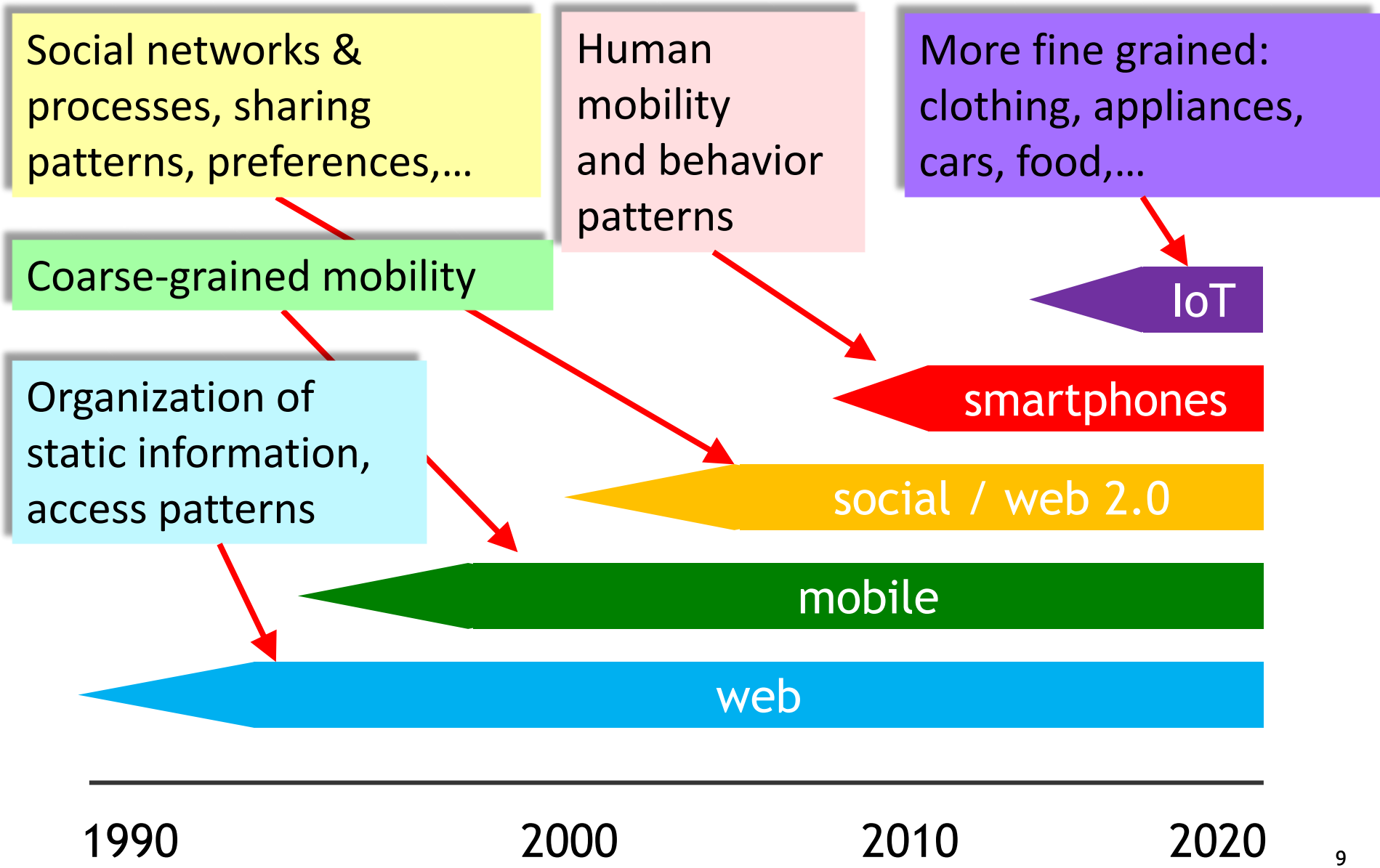
DONNÉES L'entreprise new-yorkaise, pourtant visée par de nombreuses plaintes, poursuit le développement d'outils de reconnaissance faciale. Elle veut désormais atteindre les 100 milliards

cer 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

Evolution of interfaces & user data



Example: 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.

What is this class about?

- Web and mobile services
  - Explosive growth: social networking and social media; messaging; blogs; location & navigation; …
- 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?

- Social web, online social networks, mobile, IoT:
  - Explosion of user data
  - Increasing correlation of user data: more than sum of its parts
    - Example: many geo-tagged tweets → neighborhood characteristics; taxi GPS traces → retail performance
  - Demand for data scientists
- Data: a huge variety
  - What are the main user data types?
  - What underlying models describe them?
- Function/application-oriented
  - How is data turned into decisions, actions, insights?
- Working with real data

What is this class about?

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

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
- Criterion in choice of topics: real-world relevance of…
  - Data & models
  - Functions & applications

Matrix of data and functions

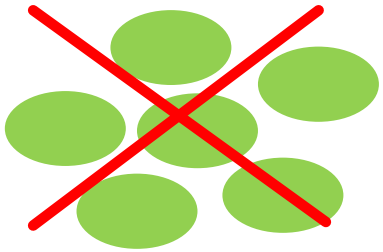
	Networks	Ratings	Document	Corpus	Streams
Characterize Model	Small worlds, scale-free			Topic models	Counts, moments
Predict Infer	Link prediction				
Rank	PageRank	Collaborative filtering	Content-based recommend		
Filter			Spam filtering		
Search Retrieve				Keyword search, similar docs	
Associate Summarize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

Overview: social and info networks

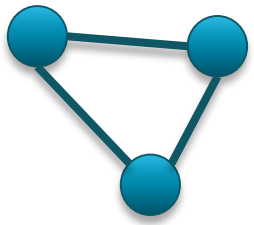
- Social and information networks
  - How are people connected?
  - How is information connected?

Networks: structure

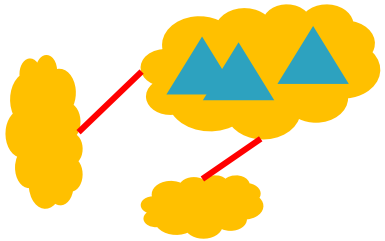
- Structure:
  - Single snapshot in time
  - Generic properties?



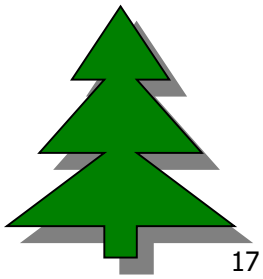
- 1: Giant Component
  - Almost all pairs of nodes are reachable



- 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 ;-)

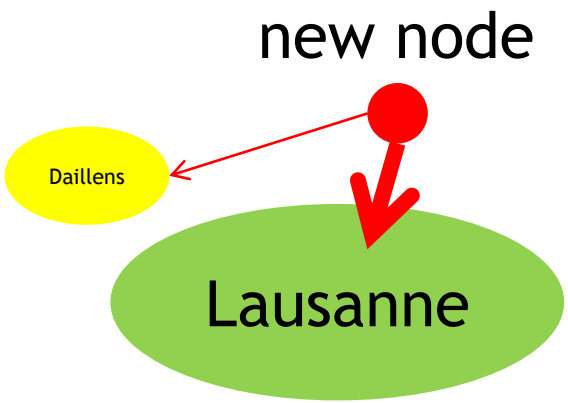


- 4: Short distances
  - Six degrees of separation



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

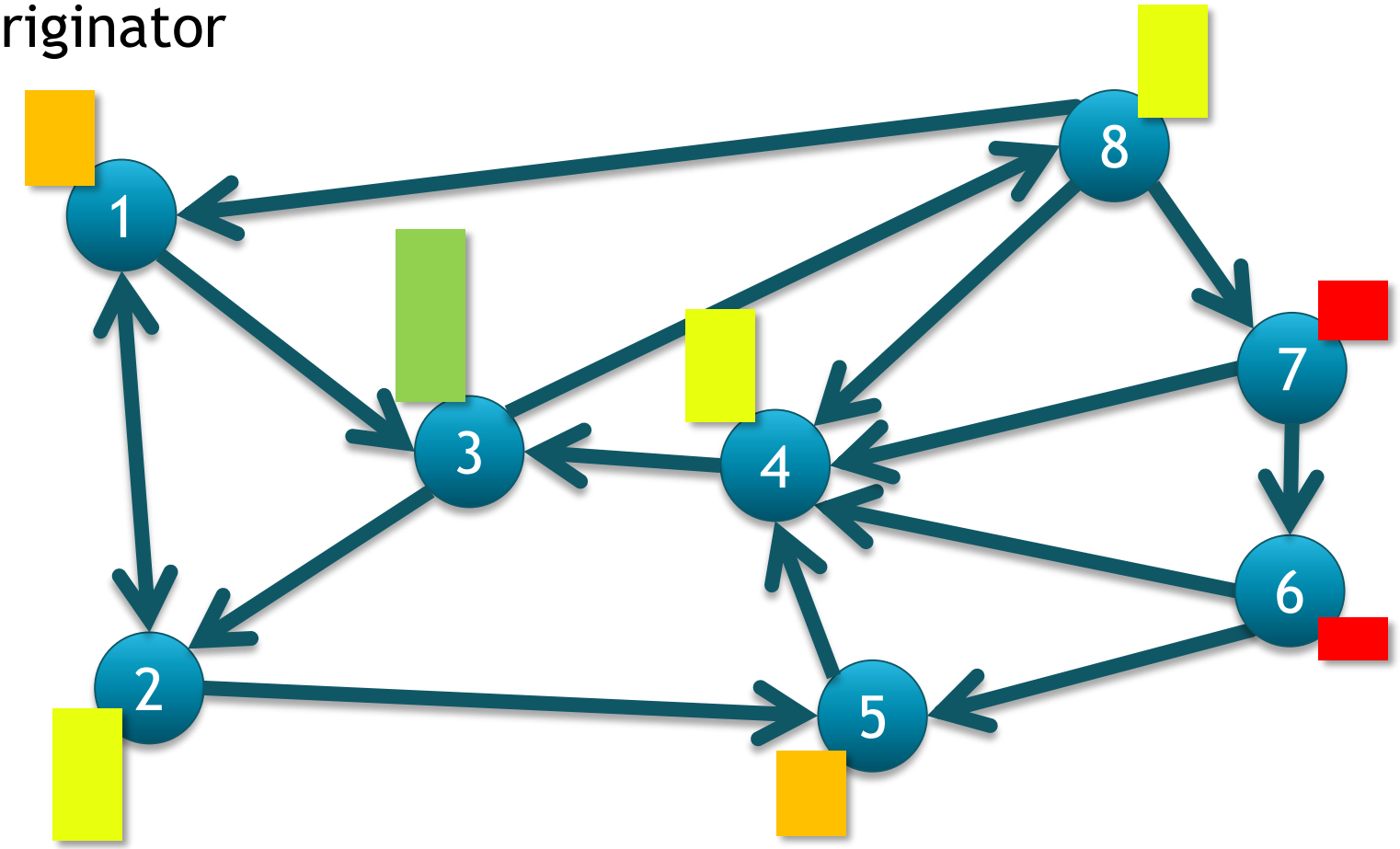


Networks: processes

- 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?

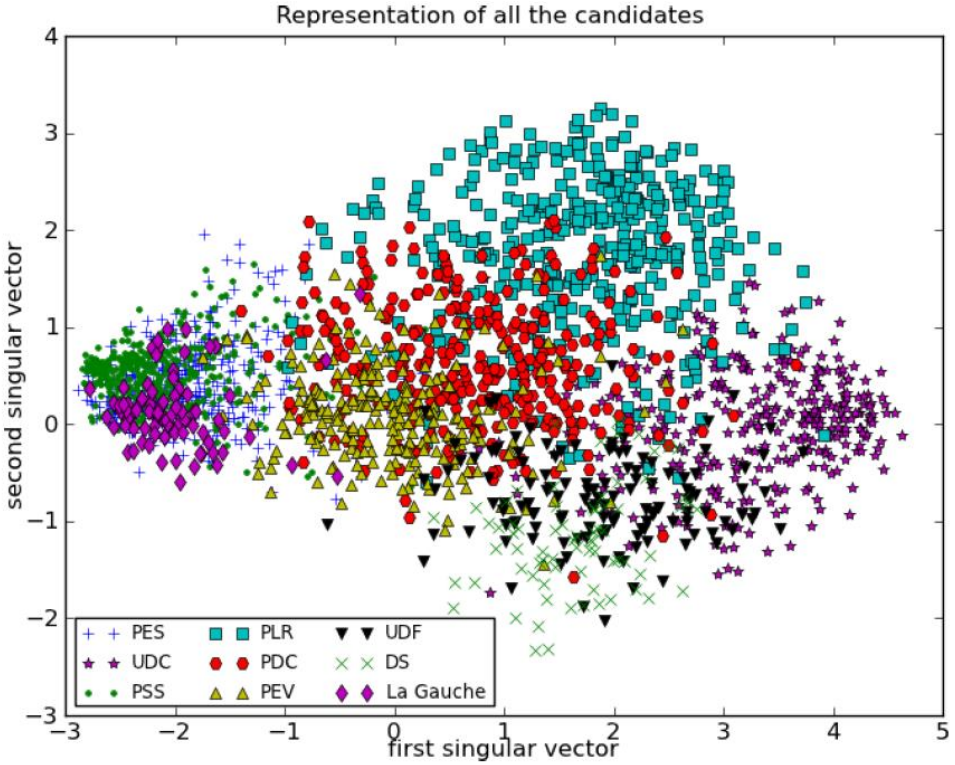
Overview: relevance and filtering

- 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 | All

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 grossglauser, 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.

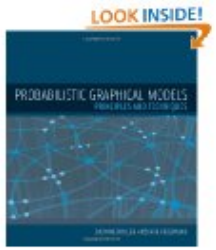
Curb Your Enthusiasm: The Complete Eighth Season
DVD ~ Larry David (June 5, 2012)
Average Customer Review: ★★★★½ (88)
In Stock

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 C
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

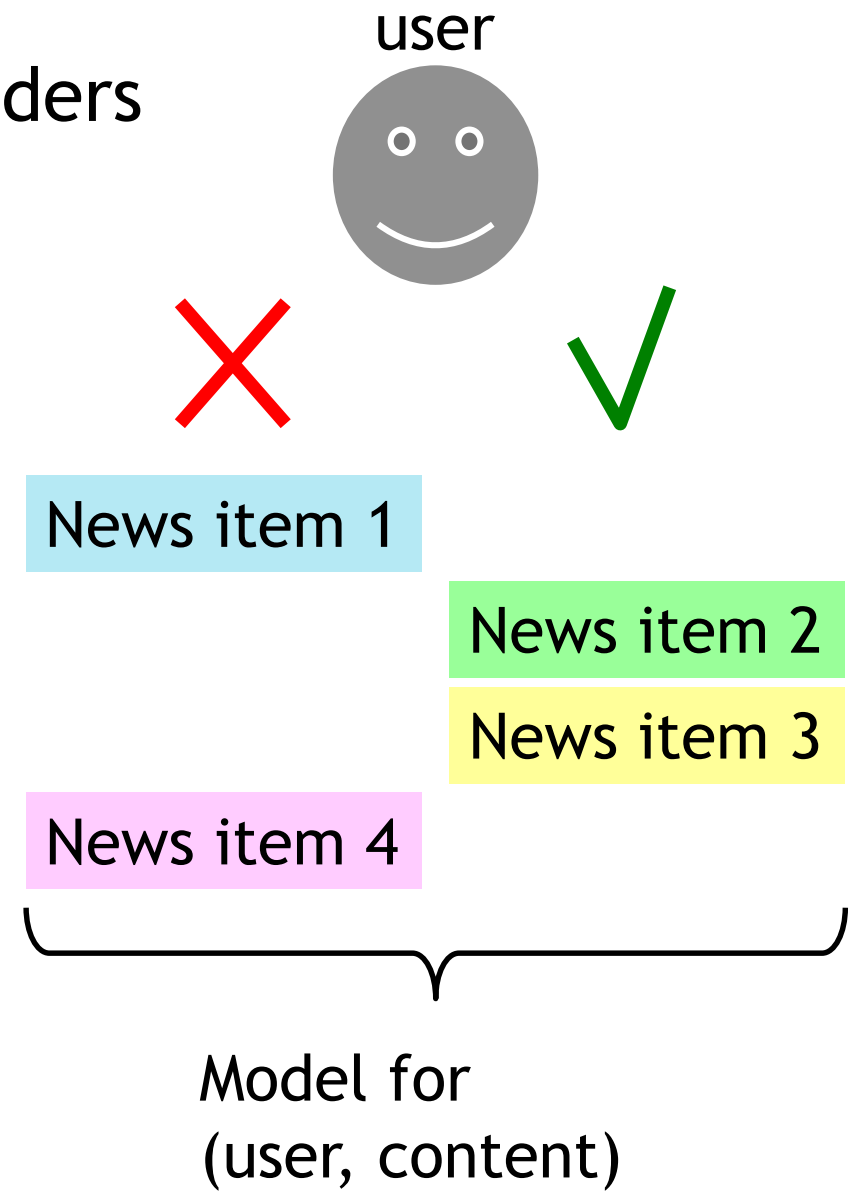
- Content-based recommenders

item 1:
“Plane hijacked…”

item 2:
“soccer game…”

item 3:
“swiss skiers win…”

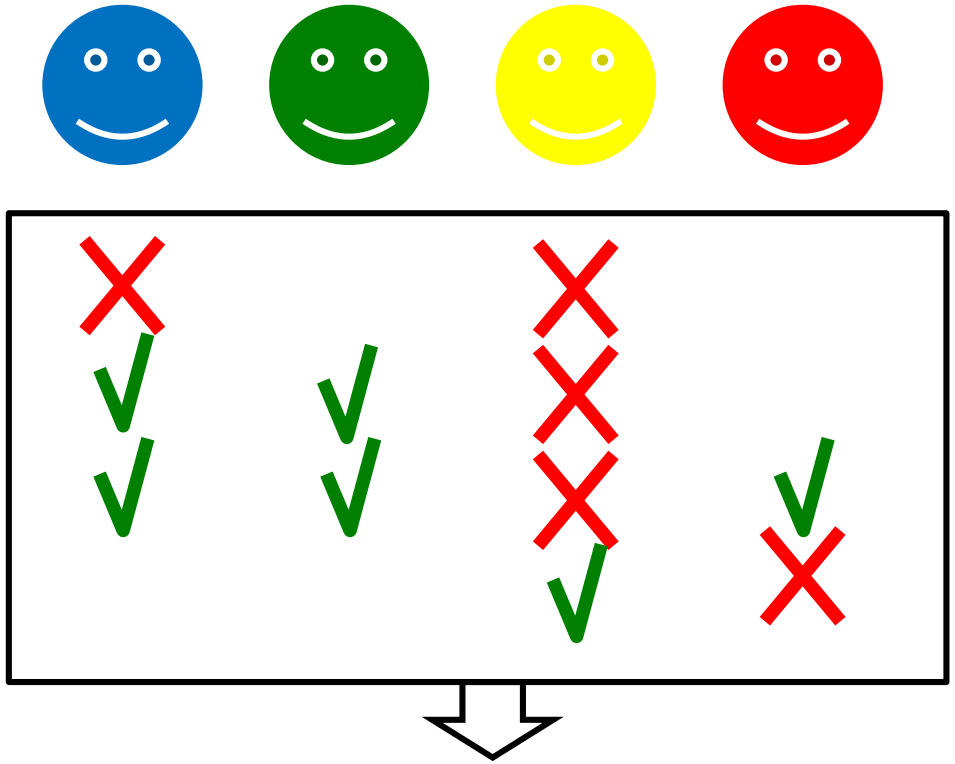
item 4:
“50.3% vote yes…”



Overview: recommender systems

- Collaborative filtering-based recommenders

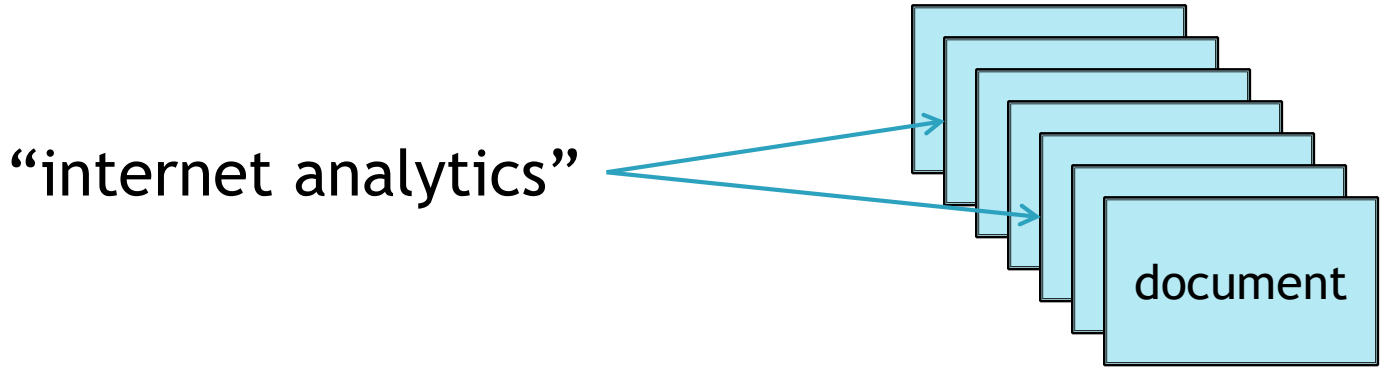
item 1
item 2
item 3
item 4



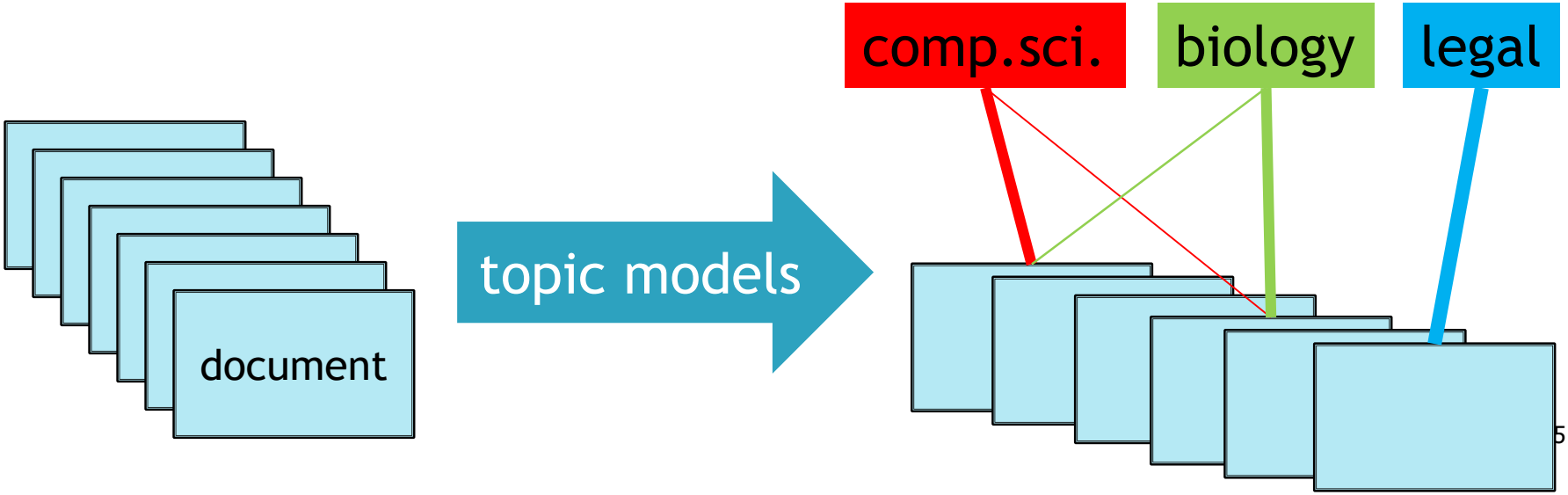
Model for (user, item)

Overview: search and retrieval

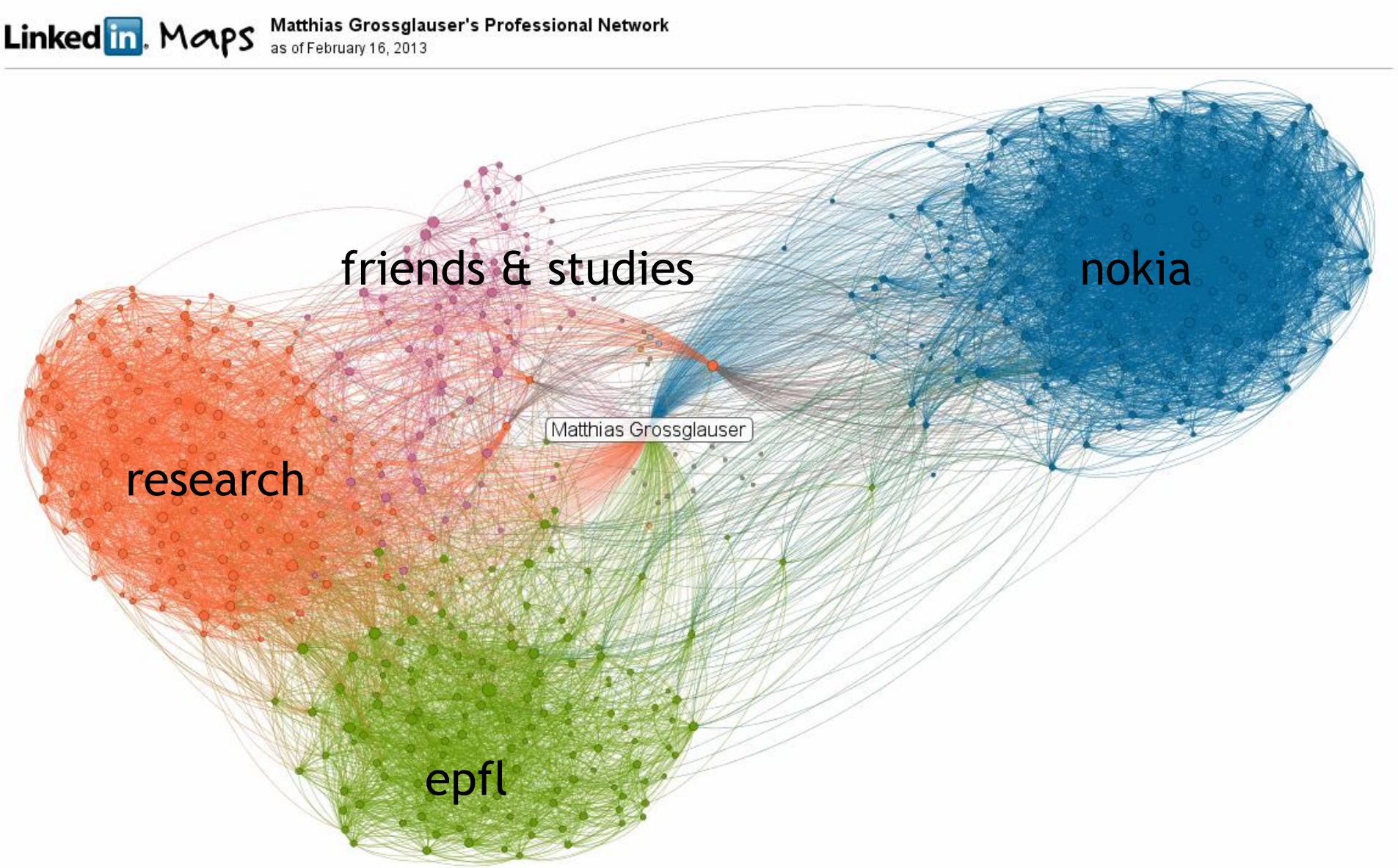
- Given a query, how to find best matches?



- Without a query, how to describe a corpus?

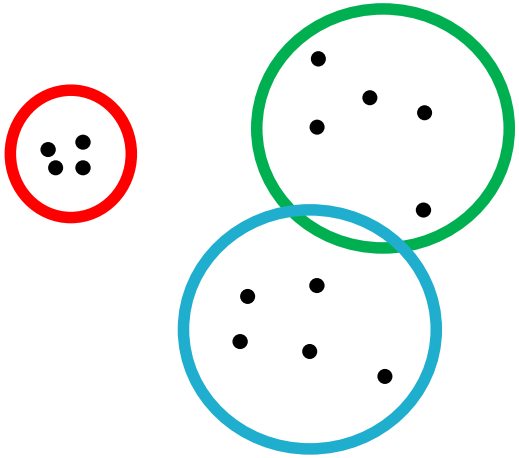


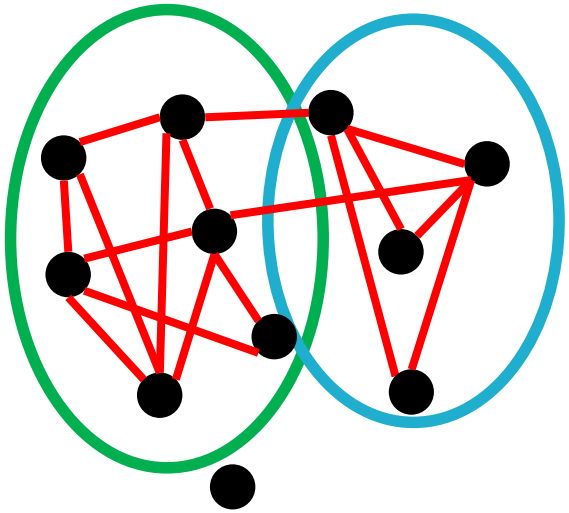
Overview: clusters, groups, communities



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

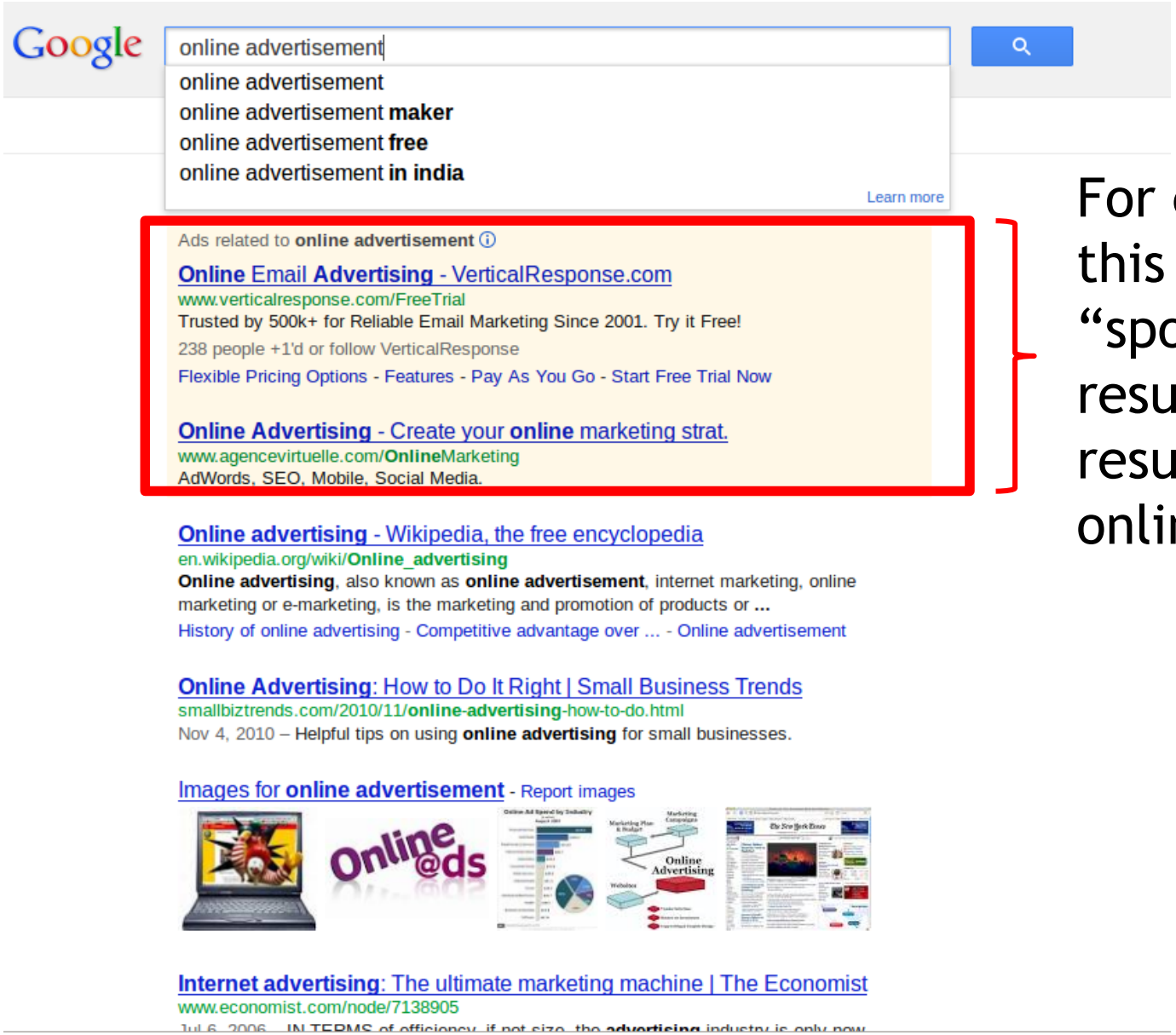
- Internet backbone router
- Order of magnitude:
  - 100s of interfaces at 10s of Gbps
  - = several billion pkts/sec!
- Traffic analysis app to detect DDoS attack:
  - How many *different* (unique) source IP addresses in a minute?
  - If too large -> suspicious (fake addresses)!

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: ∞ data, finite memory
  - Approximation
  - (Pseudo-)randomization

Overview: ad auctions



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

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	- class overview - logistics
2	Social and Information Networks 1: Structure	- intro: social networks, web, social web - social networks, key properties, effects & metrics - giant component, relation to tree percolation - strong/weak ties, clustering - "everything is close": distances; random graphs
3	Social and Information Networks 2: Evolution	- "the rich get richer": power laws, cumulative advantage, pref attachment - my friends have more friends than i-phenomenon - link prediction
4	Social and Information Networks 3: Processes	- processes: epidemics, cascades - sampling a network: undirected vs directed - random walks on networks, mixing times, spectral properties

Class Schedule (2)

5	Ranking	- Intro web structure - PageRank algorithm - Large-scale computation - HITS variant: hubs and authorities
6	Dimensionality Reduction	- intro: "finding hidden structure“, visualization - PCA and derivatives - singular-value decomposition - manifold unwrapping
7	Recommender Systems 1	- Collaborative filtering - graph-based, item-item vs user-user - spectral/matrix completion - case study: netflix
8	Recommender Systems 2	- Intro: "the long tail" - Applications, models - TF-IDF - kNN classifier - Naïve Bayes classifier

Class Schedule (3)

9	Clusters and Communities	- Gaussian Mixture Model (GMM) - EM algorithm - Communities and modularity measures - Louvains clustering algorithm
10	Text Search and Retrieval 1	- Latent Semantic Indexing (LSI) - Intro: Bayesian networks - Gibbs sampling
11	Text Search and Retrieval 2	- Probabilistic LSI (pLSI) - Latent Dirichlet Allocation (LDA)
12	Streams	- Intro - motivating apps - streaming model
13	Internet Ad Economy & Online Auctions	- intro: sponsored search, keyword auctions - VCG and Generalized Second Price (GSP) - incentive compatibility

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:
  - Wed 08:15h: Tutorial on Spark and using infrastructure
  - No HW session this week yet
- Reminder: sign up for class in ISA by tomorrow!