Data Science

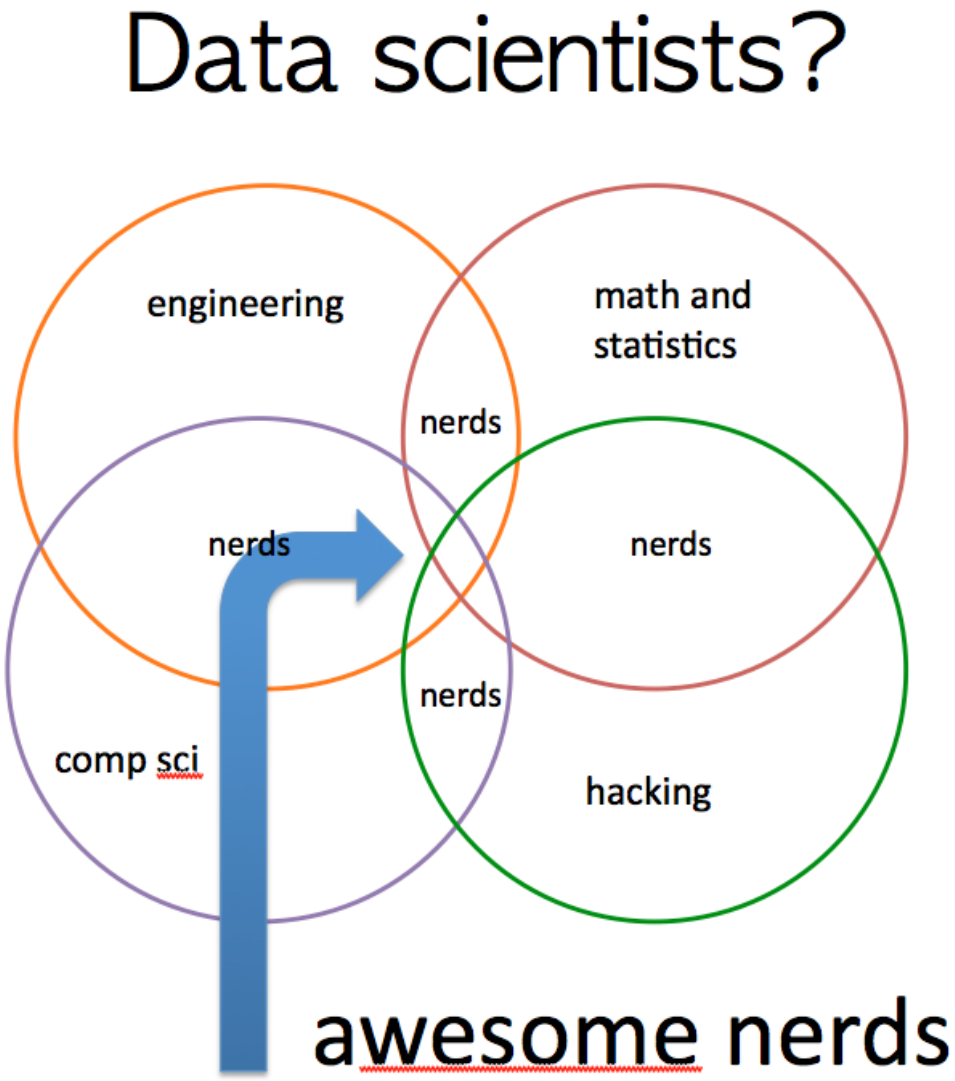
Yi Fang, PhD

Department of Computer Engineering
Santa Clara University

yfang@scu.edu

http://www.cse.scu.edu/~yfang/

What is a Data Scientist?

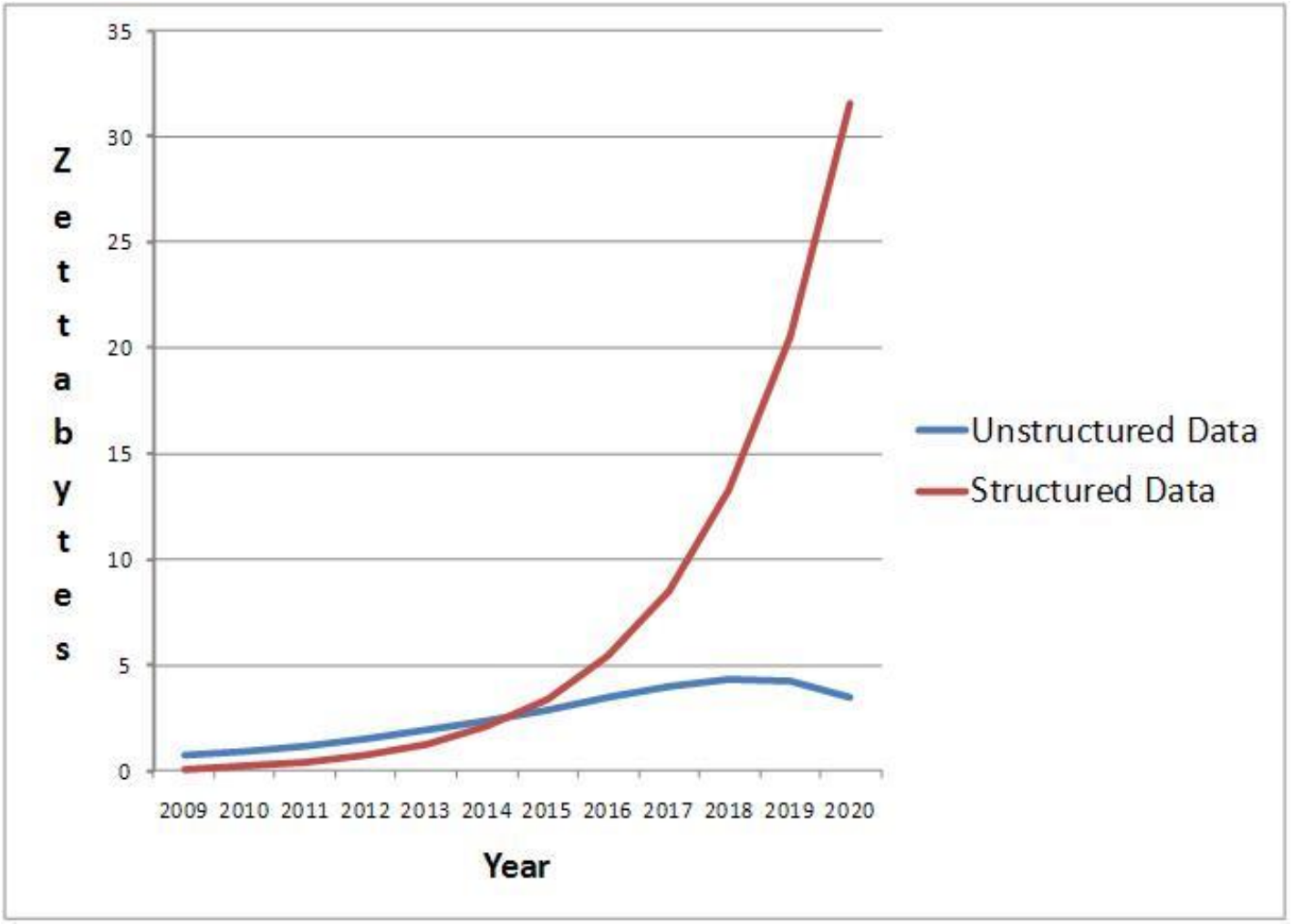


Why is data science different from other fields?

Unstructured Data

- Documents
- Webpages
- Images
- Audio
- Video
- More…

Growth



http://www.emc.com/leadership/programs/digital-universe.htm

Big Data

Any dataset where the size or speed of incoming data causes difficulties in processing

- Volume
- Velocity
- Variety

Law of Data

18 Months

the amount of time for digital data to double

Why do you care?

"Every single industry will be totally revolutionized by big data"

- Joe Tucci, EMC

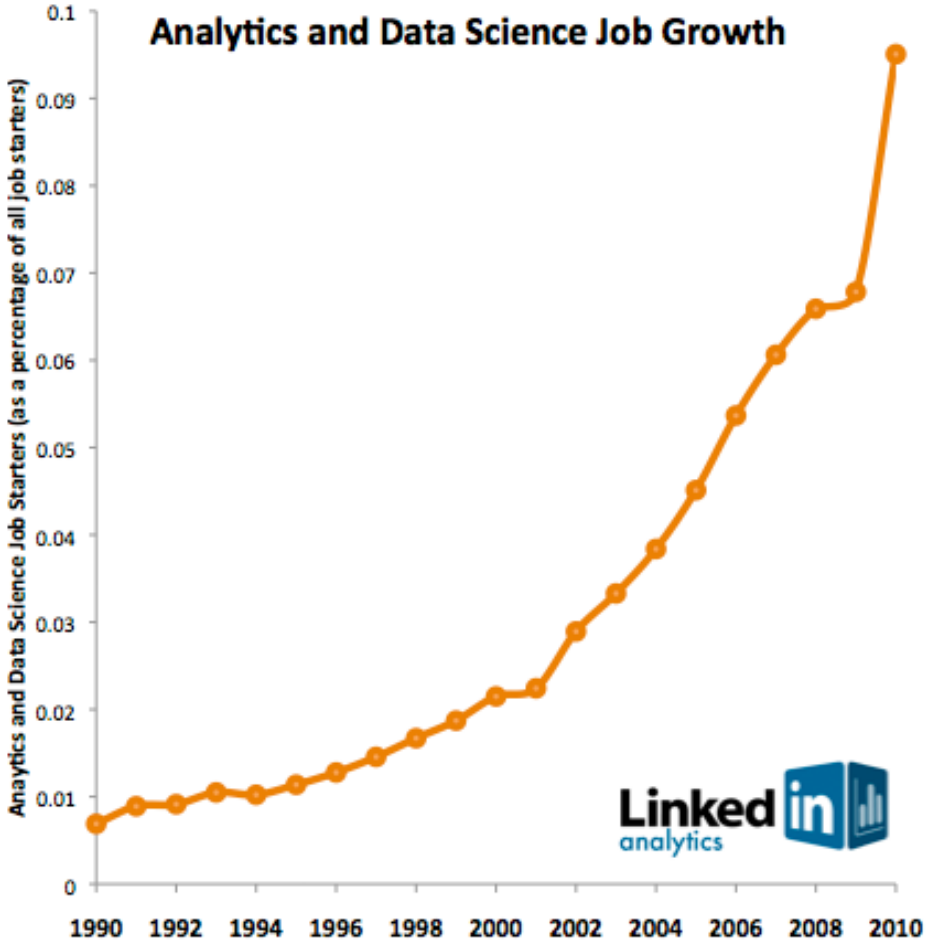
Big Data Examples

- Google: > 100 PB; > 1T indexed URLs
- Facebook: 1 billion users; 40 billion photos
- YouTube: > 750 PB
- Twitter: > 55 billion tweets/year;
  > 150 million/day; 1700/second
- Text messages: 6.1 T/year; 876/person/year
- US cell calls: 2.2 T minutes/year;
  19 minutes/person/day
  ~ size of a YouTube

Driving Forces

Sensors and The Internet of Things

Data Science Job Listing



Data Scientist:

The Sexiest Job of the 21st Century

Meet the people who can coax treasure out of messy, unstructured data.

by Thomas H. Davenport and D.J. Patil

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—and you probably leave early."

The World's 7 Most Powerful Data Scientists

"The success of companies like Google, Facebook, Amazon, and Netflix, not to mention Wall Street firms and industries from manufacturing to retail and healthcare, is increasingly driven by better tools for extracting meaning from very large quantities of data."

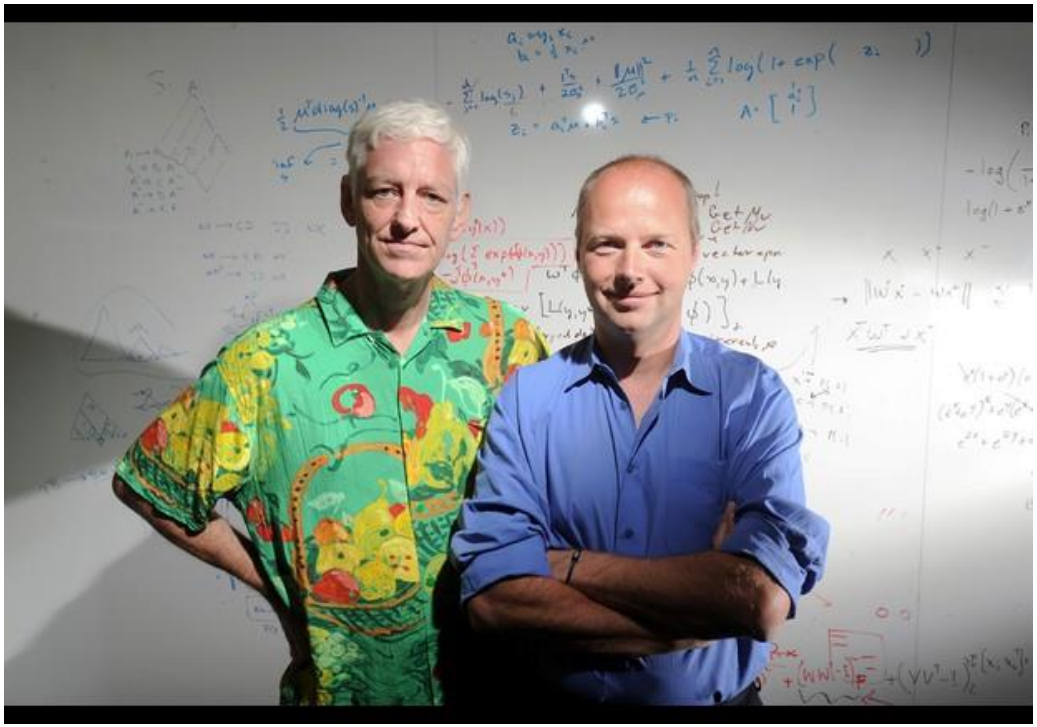
- Tim O'Reilly

#1 Larry Page, founder, Google

#2

- Jeff Hammerbacher, Chief Scientist, Cloudera
- DJ Patil, U.S. Chief Data Scientist, White House

#3



- Peter Norvig, Director of Research, Google
- Sebastian Thrun, Professor, Stanford University

My Own List

Michael Jordan

Andrew Ng

Hilary Mason

Amit Singhal

Recommendation Systems

Information Overload

Book Recommendation systems

- **Amazon.com** recommends books based on your purchase history (and others')

Movie Recommendation systems

- **Netflix** predicts other "Movies You'll Love"

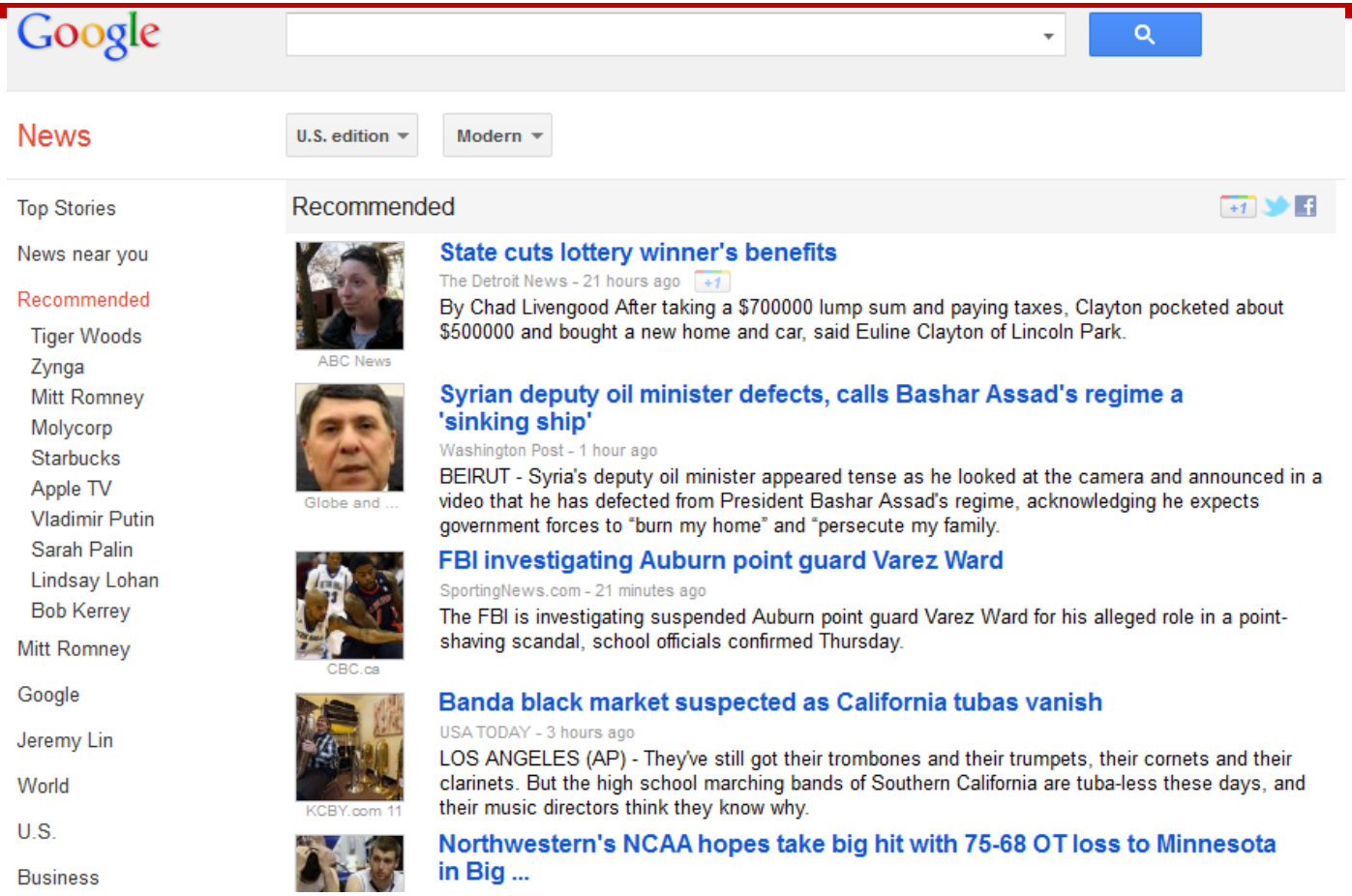
➢ Recommendations drives more than 60% Netflix's DVD rentals [Thompson, 2011]

Recommendation algorithms

- **Netflix Prize:**
  Beat Netflix's own recommender system with 10% margin,
  Win $1 million

- **Testbed:**
  480,000 users
  18,000 movies

News Recommendation



➢ **Google News** recommends news articles based on clicks and browse history

Personalized Job Recommendation

User Click Prediction in Personalized Job Recommendation

Miao Jiang, Yi Fang
Department of Computer Engineering
Santa Clara University
Santa Clara, California, USA
yfang@scu.edu

Huangming Xie, Jike Chong, Meng Meng
Simply Hired, Inc.
Sunnyvale, California, USA
{jike, huangming}@simplyhired.com

ABSTRACT

Major job search engines aggregate tens of millions of job postings online to enable job seekers to find valuable employment opportunities. Predicting the probability that a given user clicks on jobs is crucial to job search engines as the prediction can be used to provide personalized job recommendations for job seekers. This paper presents a real-world job recommender system in which job seekers subscribe to email alert to receive new job postings that match their specific interests. The architecture of the system is introduced with the focus on the recommendation and ranking component. Based on observations of click behaviors of a large number of users in a major job search engine, we develop a set of features that reflect the click behavior of individual job seekers. Furthermore, we observe that patterns of missing features may indicate various types of job seekers. We

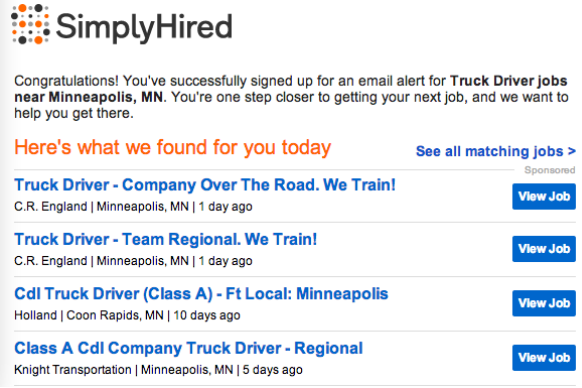
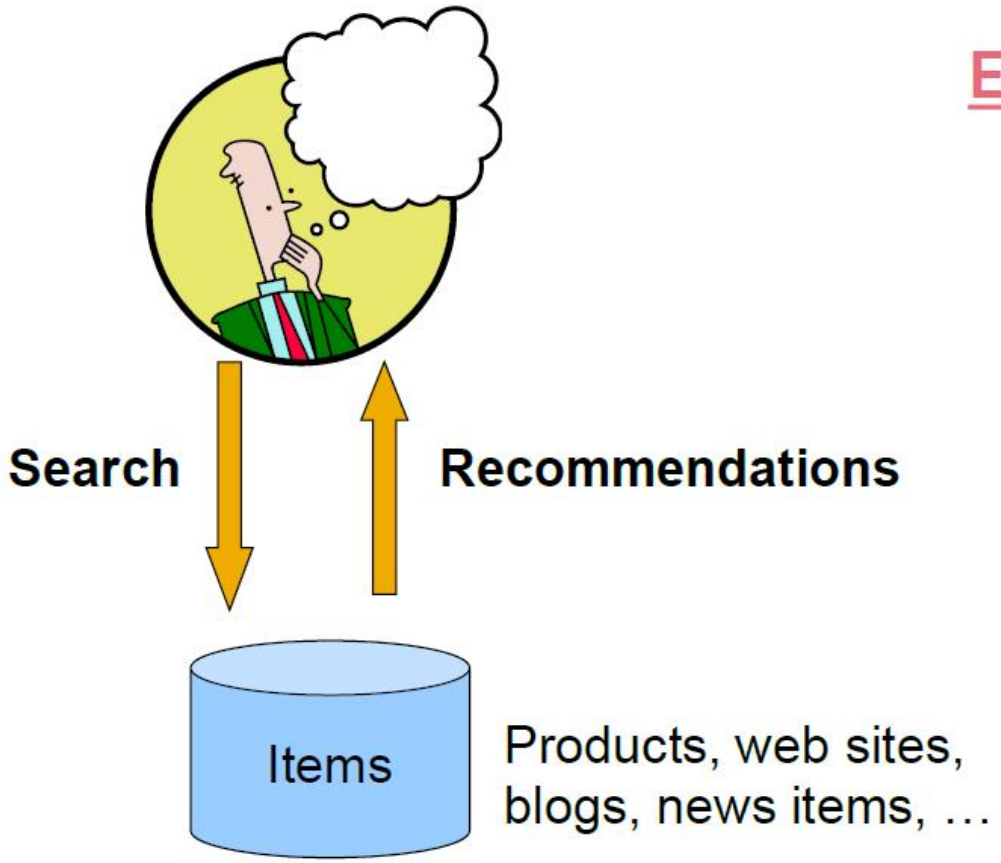


Figure 1: An example of *Simply Hired*'s email alert service for job recommendation.

Point-of-Interest Recommendation

- Foursquare check-in
- Google Place API

Recommendation vs Search



Examples:

amazon.com

PANDORA

StumbleUpon

del.icio.us

NETFLIX

movielens
helping you find the *right* movies

last·fm
the social music revolution

Google News

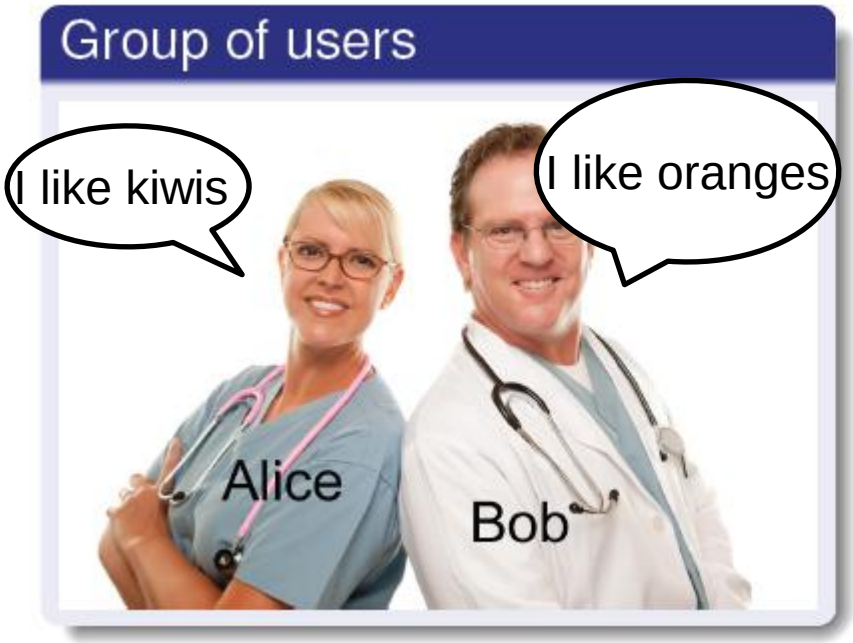
YouTube

XBOX LIVE

Types of Recommendations

- Editorial
- Simple aggregates:
  - Top 10, Most Popular, Recent Uploads
- Tailored to individual users
  - Amazon, Netflix, …

What is Collaborative Filtering?



- Observe some user-item preferences
- Predict new preferences

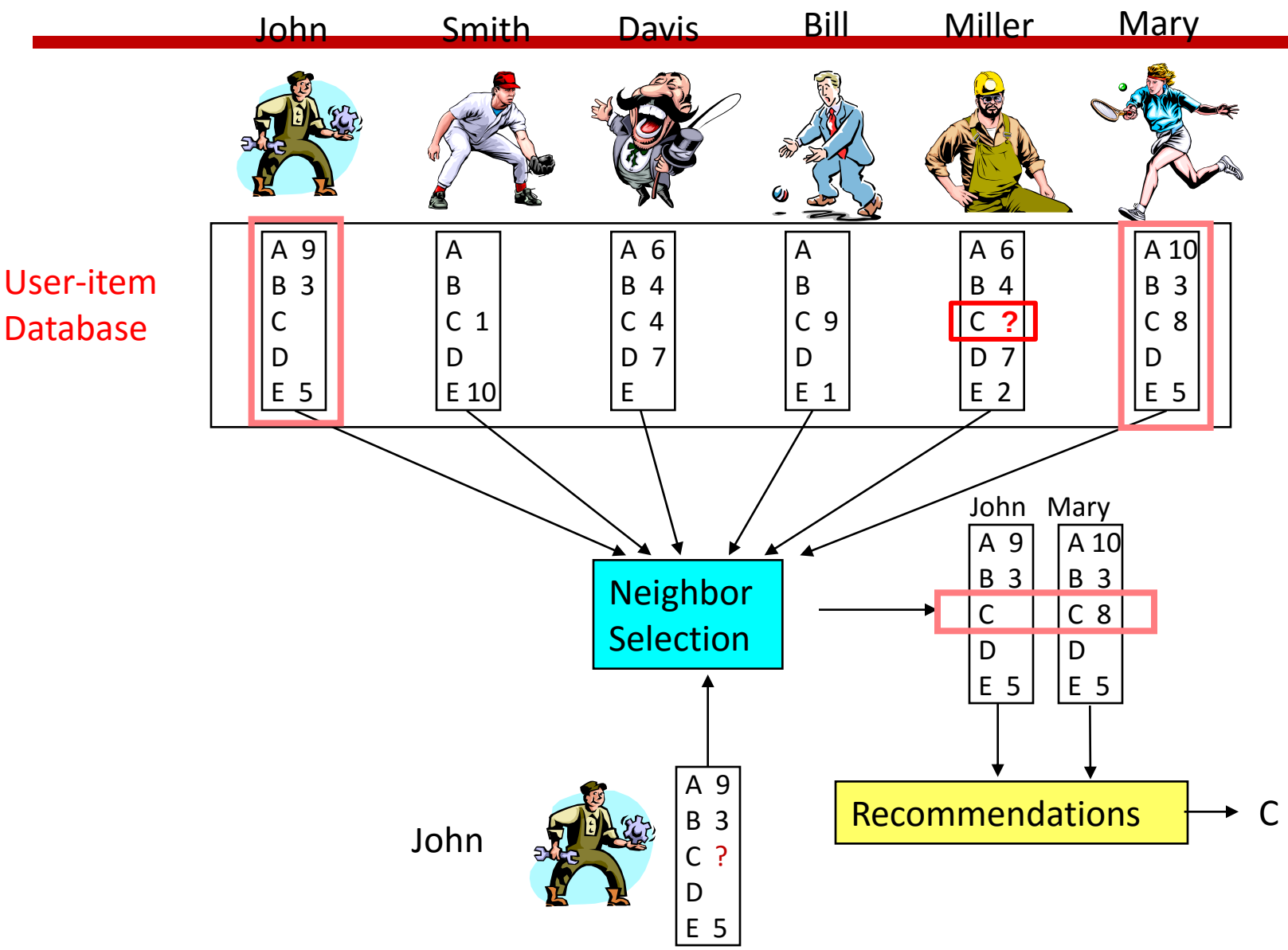
Does Bob like strawberries???

Data and Task

- Set U={u_1, …, u_m} of m users
- Set I={i_1, …, i_n} of n items (e.g. Movies, books)
- Set R={$r_{u,i}$ } of ratings/preference
  (e.g., 1-5, 1-10, binary)

- Task:
  - Recommend new items for an active user a
  - Usually formulated as a rating prediction problem

User-based Collaborative Filtering



User-based Collaborative Filtering

"Similar users rate similarly!"

User-based Collaborative Filtering

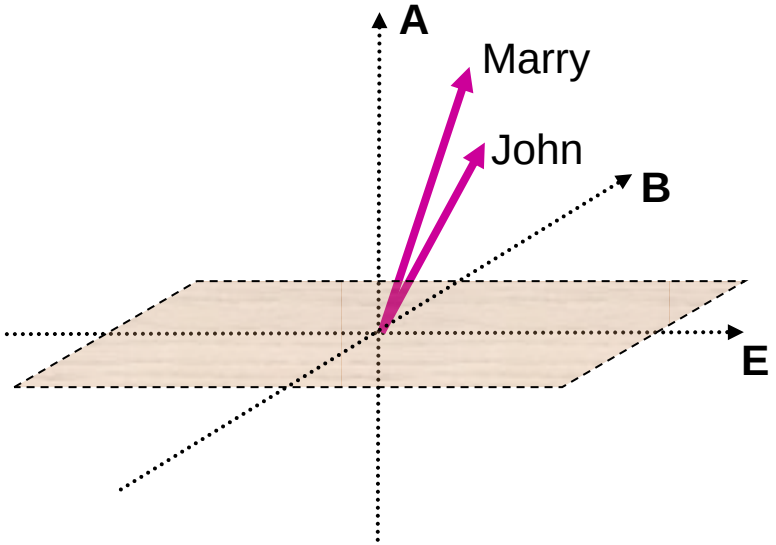
- Consider the active user *a*
- Find *k* other users whose ratings are "similar" to *a*'s ratings
- Estimate *a*'s ratings based on ratings of the *k* similar users
- Called *k*-nearest neighborhood method

Neighbor Selection

- How similar are the users?

John		Mary	
A	9	A	10
B	3	B	3
C		C	8
D		D	
E	5	E	5

- Cosine Vector Similarity



Rating Prediction

- For a given active user, select the most similar *k* users, based on their similarity

- Take the average of the k similar users' ratings on the target item

Exercise

- Predict User D's rating on Item 4

	$Item1$	$Item2$	$Item3$	$Item4$	$Item5$
User A	4	4	1	4	3
User B	2	1	4	2	5
User C	3	1	3	2	1
User D	5	4	2		3

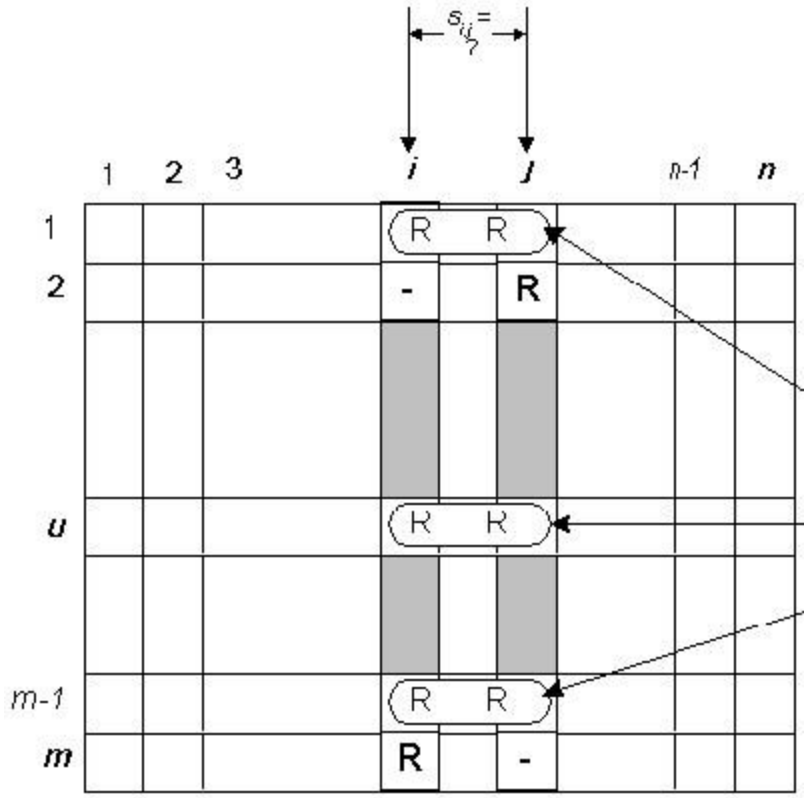
Item-based Collaborative Filtering

"Similar items are rated similarly!"

Item-based Collaborative Filtering

- Rather than matching the active user to similar customers, finding items that get similar ratings

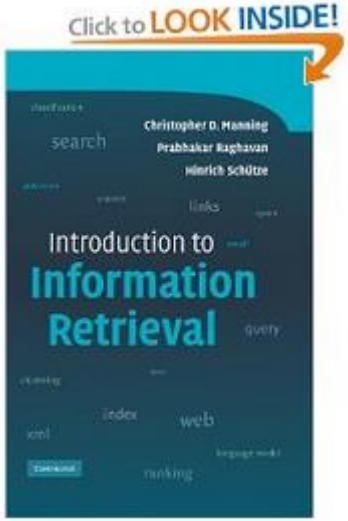
Finding Similar Items



Computed by looking into co-rated items only. These co-rated pairs are obtained from different users.

Amazon's book recommendation

"Users who bought this book, also bought that book"



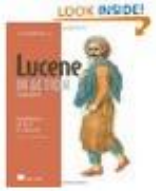
Customers Who Bought This Item Also Bought

Speech and Language Processing (2nd Edition)
Daniel Jurafsky
★★★★☆ (32)
Hardcover
$112.47

Modern Information Retrieval: The Concepts ...
› Ricardo Baeza-Yates
★★★★★ (1)
Paperback
$55.68

Foundations of Statistical Natural Language ...
› Christopher D. Manning
★★★★½ (14)
Hardcover
$56.84

Natural Language Processing with Python
› Steven Bird
★★★★☆ (16)
Paperback
$37.59



Lucene in Action, Second Edition: Covers Apache ...
› Michael McCandless
★★★★½ (30)
Paperback
$31.36