

---

# Data Science

**Yi Fang, PhD**

Department of Computer Engineering  
Santa Clara University

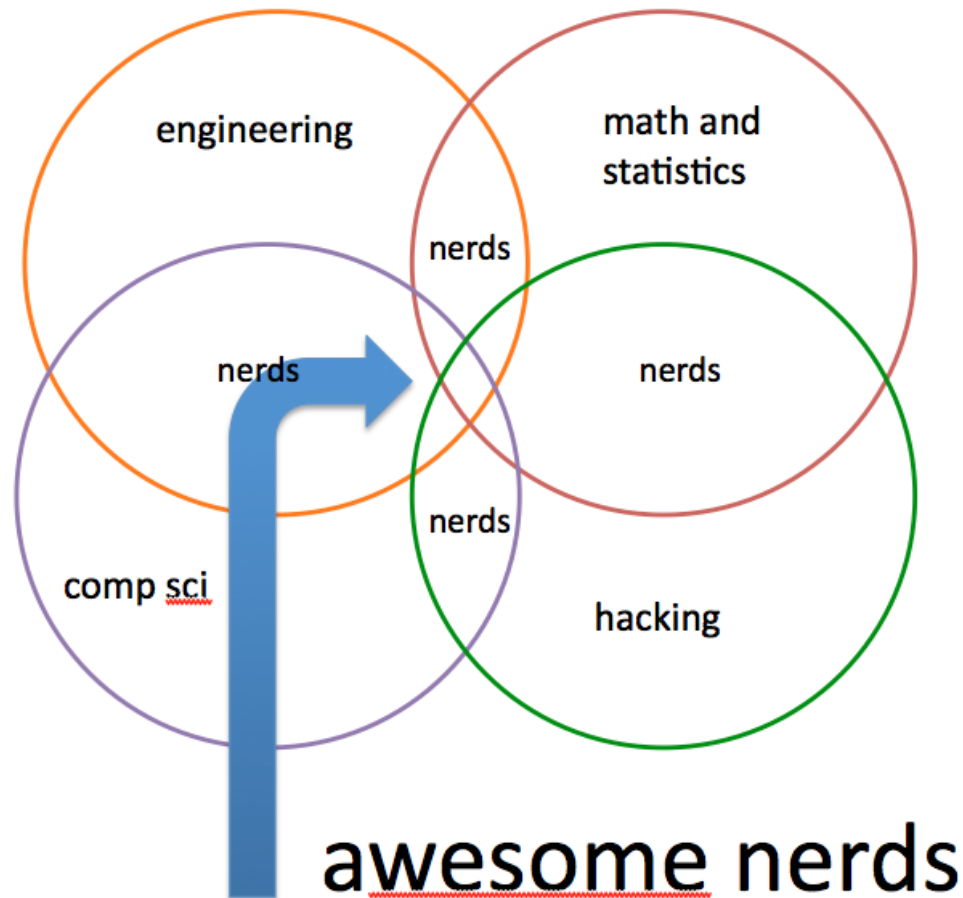
yfang@scu.edu

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

# What is a Data Scientist?

---

Data scientists?



---

Why is data science different  
from other fields?

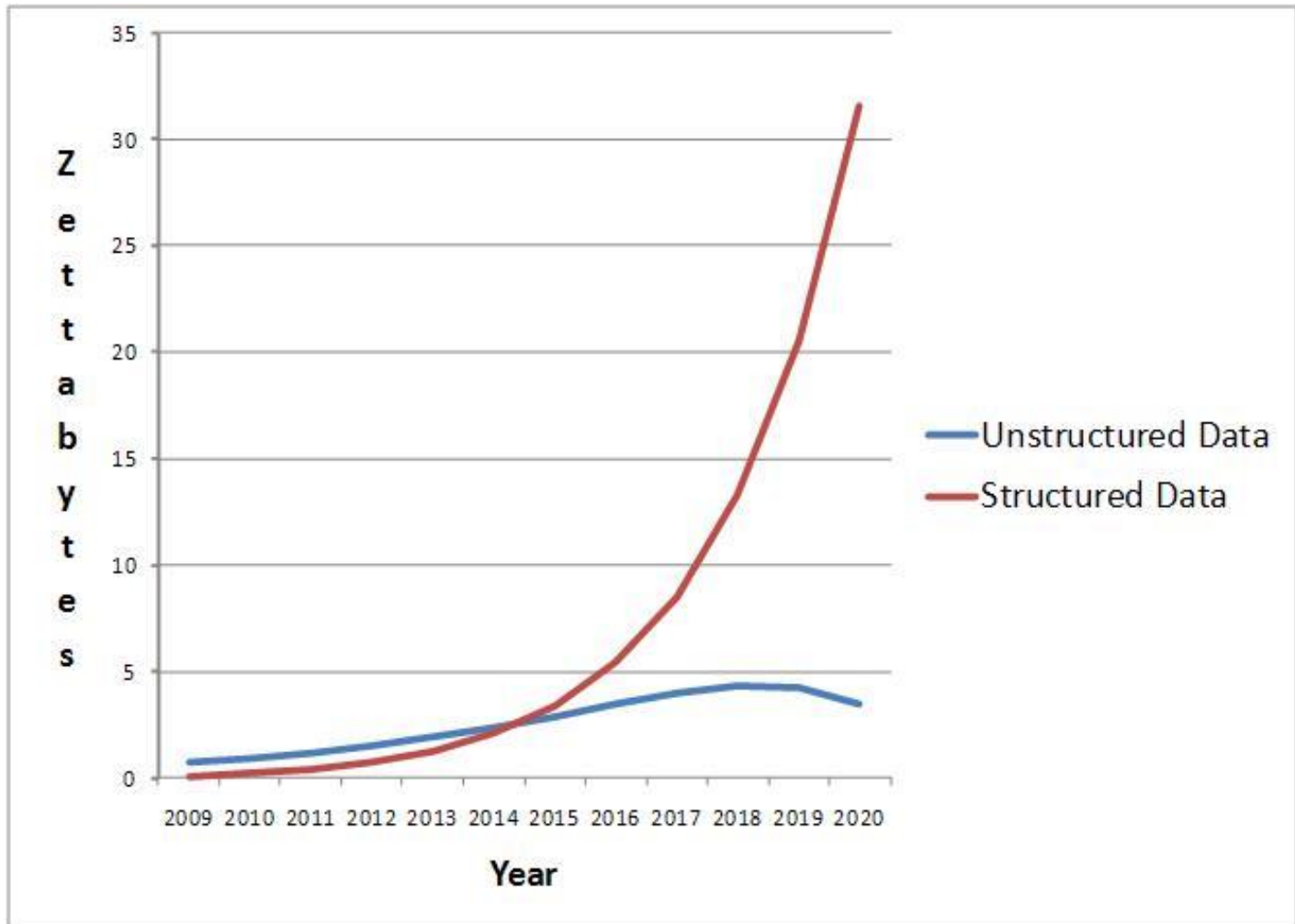
# Unstructured Data

---

- Documents
- Webpages
- Images
- Audio
- Video
- More...

# Growth

---



# 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

---



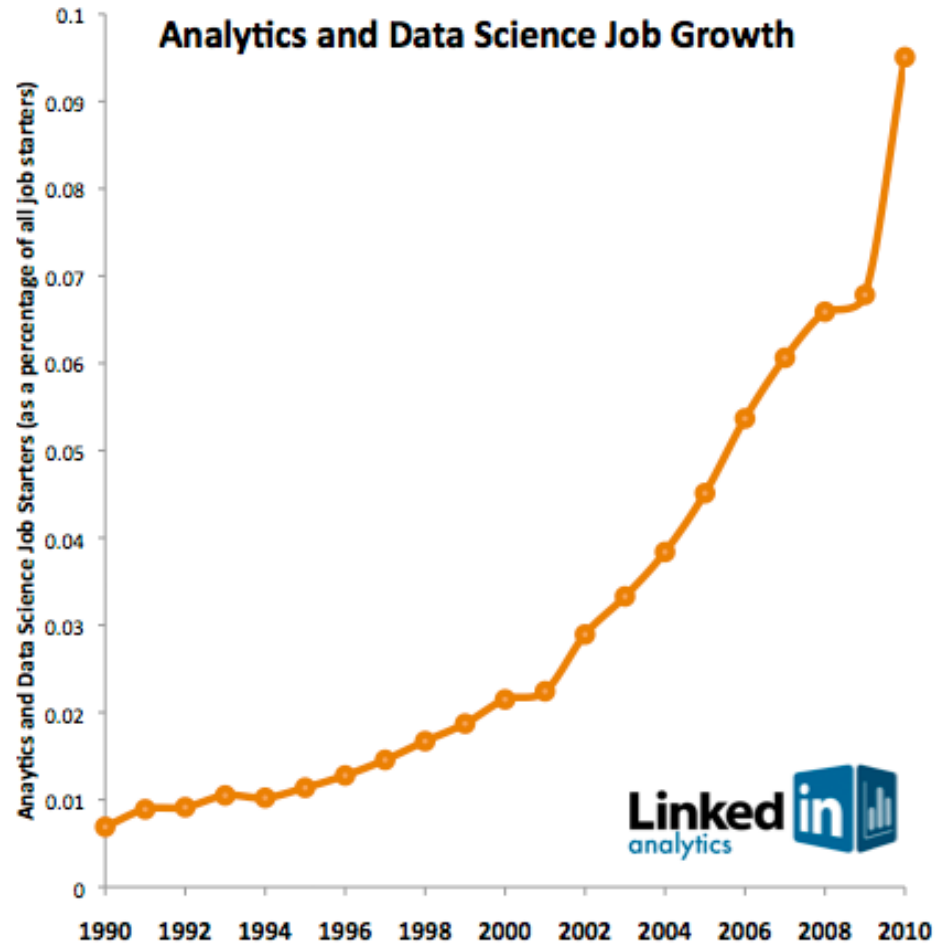
# Sensors and The Internet of Things

---



# 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*

**W**hen 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

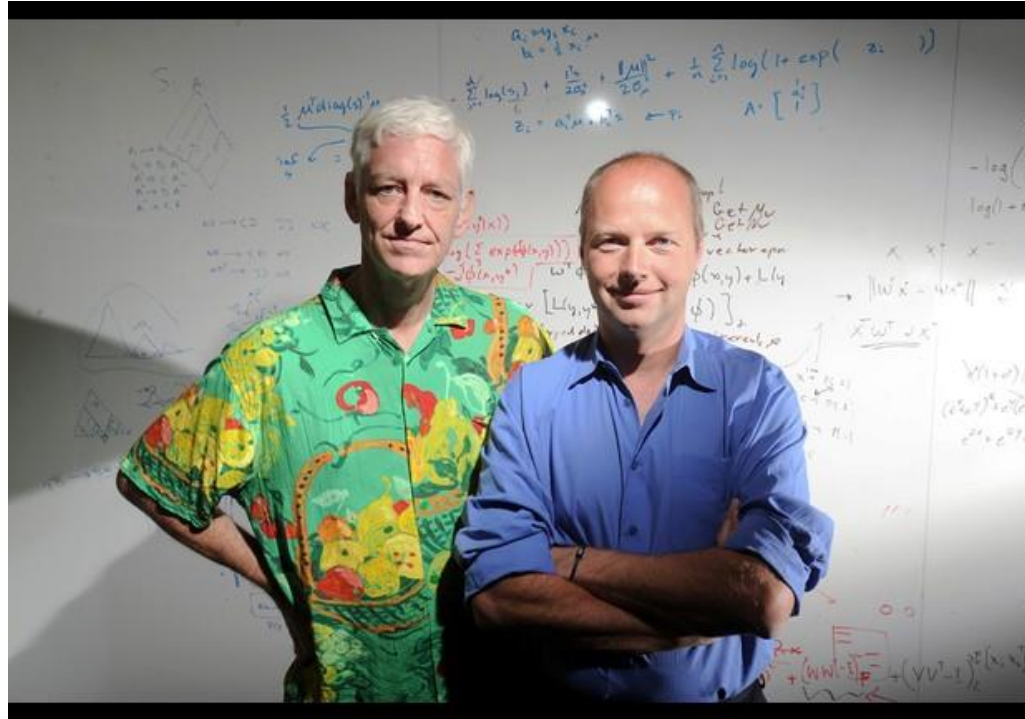
---



- 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')

Yi, Welcome to Your Amazon.com (if you're not Yi Fang, click here.)

**Today's Recommendations For You**

Here's a daily sample of items recommended for you. Click here to [see all recommendations](#). Page 2 of 35 (Start over)

 <p>LOOK INSIDE!</p>	 <p>LOOK INSIDE!</p>	 <p>LOOK INSIDE!</p>	 <p>LOOK INSIDE!</p>	
<p><a href="#">Data Analysis with Open Source Tools</a> (Paperback) by Philipp K. Janert</p> <p>★★★★☆ (23) \$25.30</p> <p><a href="#">Fix this recommendation</a></p>	<p><a href="#">Causality: Models, Reasoning and Inference</a> (Hardcover) by Judea Pearl</p> <p>★★★★☆ (5) \$42.79</p> <p><a href="#">Fix this recommendation</a></p>	<p><a href="#">Design Patterns: Elements of Reusable Object-Oriented Software</a> (Hardcover) by Erich Gamma</p> <p>★★★★☆ (297) \$32.99</p> <p><a href="#">Fix this recommendation</a></p>	<p><a href="#">Code Complete: A Practical Handbook of Software Construction, 2nd Edition</a> (Paperback) by Steve McConnell</p> <p>★★★★☆ (150) \$29.54</p> <p><a href="#">Fix this recommendation</a></p>	<p><a href="#">Fundamentals of Embedded Software: Where C and Assembly Meet</a> by Daniel W. Lewis</p> <p>★★★★☆ (11) \$82.96</p> <p><a href="#">Fix this recommendation</a></p>

# Movie Recommendation systems

- **Netflix** predicts other “Movies You’ll Love”



The screenshot shows the Netflix website interface. At the top, the Netflix logo is on the left, and navigation buttons for "Browse DVDs", "Browse Instant", "Your Queue", "Movies You'll ❤️", "Friends & Community", and "DVD Sale \$5.99" are in the center. A search bar on the right contains the text "Movies, actors, directors, genres" and a "Search" button. Below the navigation bar, there are links for "Home", "Genres", "New Releases", "Netflix Top 100", "Critics' Picks", and "Award Winners".

The main content area is divided into three sections:

- Because you enjoyed:** This section lists movies like [Chinatown](#), [Vertigo](#), and [Dr. Strangelove](#). Below this, it says "We think you'll enjoy:" followed by [The Last Laugh](#) and a red "Add" button.
- Movie Card:** A central card for the movie "THE LAST LAUGH" by F.W. Murnau. It features a star rating of five stars and a "Not Interested" button.
- YOUR RECENT ACTIVITY:** This section shows a list of recent activities: "04/14 We shipped" (with a blurred title), "04/14 We received" (with a blurred title), and "03/26 We received" (with a blurred title).
- SUGGESTIONS FOR YOU:** This section states "You have [new suggestions](#) in Movies You'll ❤️".

➤ Recommendations drives more than 60% Netflix’s DVD rentals [Thompson, 2011]



# Recommendation algorithms

**NETFLIX**

## Netflix Prize

Home Rules Leaderboard Register Update Submit Download

**NETFLIX**

Browse Recommendations Friends Queue Buy DVDs

Home Genres New Releases Previews Netflix Top 100 Crit

### Movies For You

Mandy, the following movies were chosen based on your interest in [The Untouchables](#), [The Untouchables Season 1](#), and [The Untouchables Part 1](#).

**The Big One**

★ ★ ★ ★ ☆

It's a subversive comedy about a man who...

**You really liked it...**

Now only for just \$5.99

Shop as low as \$5.99

**Welcome!**

The Netflix Prize seeks to substantially improve the accuracy of predictions about how much someone is going to love a movie based on their movie preferences. Improve it enough and you win one (or more) Prizes. Winning the Netflix Prize improves our ability to connect people to the movies they love.

Read the [Rules](#) to see what is required to win the Prizes. If you are interested in joining the quest, you should [register a team](#).

You should also read the [frequently asked questions](#) about the Prize. And check out how various teams are doing on the [Leaderboard](#).

Good luck and thanks for helping!

FAQ | Forum | Netflix Home

© 1997-2006 Netflix, Inc. All rights reserved.

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

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

# News Recommendation

The image shows a screenshot of the Google News homepage. At the top is the Google logo and a search bar. Below the search bar are two dropdown menus: 'U.S. edition' and 'Modern'. The main content area is divided into two columns. The left column, titled 'Recommended', lists various topics and names: Tiger Woods, Zynga, Mitt Romney, Molycorp, Starbucks, Apple TV, Vladimir Putin, Sarah Palin, Lindsay Lohan, Bob Kerrey, Mitt Romney, Google, Jeremy Lin, World, U.S., and Business. The right column, titled 'Recommended', features four news articles, each with a thumbnail image, a headline, and a brief description. The first article is 'State cuts lottery winner's benefits' from The Detroit News, dated 21 hours ago. The second is 'Syrian deputy oil minister defects, calls Bashar Assad's regime a 'sinking ship'' from the Washington Post, dated 1 hour ago. The third is 'FBI investigating Auburn point guard Varez Ward' from SportingNews.com, dated 21 minutes ago. The fourth is 'Banda black market suspected as California tubas vanish' from USA TODAY, dated 3 hours ago. The fifth article is 'Northwestern's NCAA hopes take big hit with 75-68 OT loss to Minnesota in Big ...'.

➤ **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

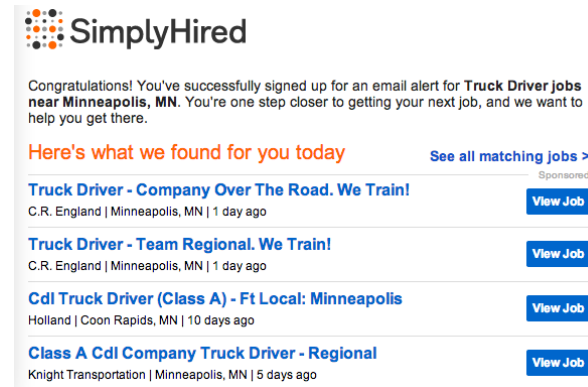


Figure 1: An example of *Simply Hired's* email alert service for job recommendation.

# Point-of-Interest Recommendation

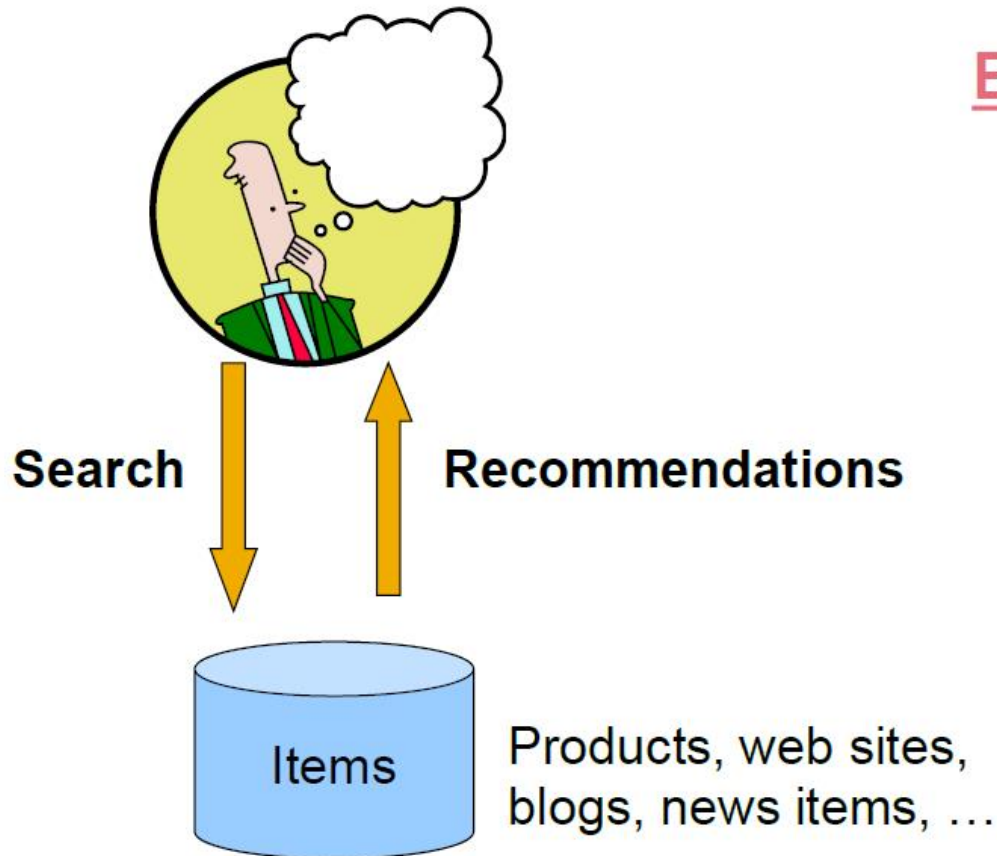
---

- Foursquare check-in
- Google Place API



# Recommendation vs Search

---



## Examples:

amazon.com.



StumbleUpon



del.icio.us



movielens  
helping you find the *right* movies

last.fm™  
the social music revolution

Google™  
News

You Tube

XBOX  
LIVE

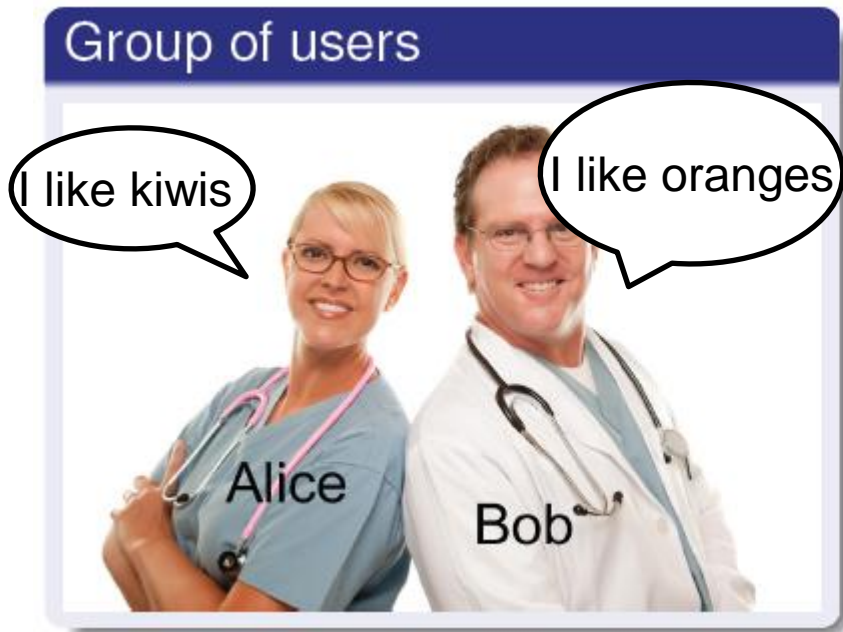
# Types of Recommendations

---

- 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  $U = \{u_1, \dots, u_m\}$  of  $m$  users
- Set  $I = \{i_1, \dots, i_n\}$  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

John

Smith

Davis

Bill

Miller

Mary



User-item Database

A 9	A	A 6	A	A 6	A 10
B 3	B	B 4	B	B 4	B 3
C	C 1	C 4	C 9	C ?	C 8
D	D	D 7	D	D 7	D
E 5	E 10	E	E 1	E 2	E 5

Neighbor Selection

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

Recommendations → C

John



A 9
B 3
C ?
D
E 5

# 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

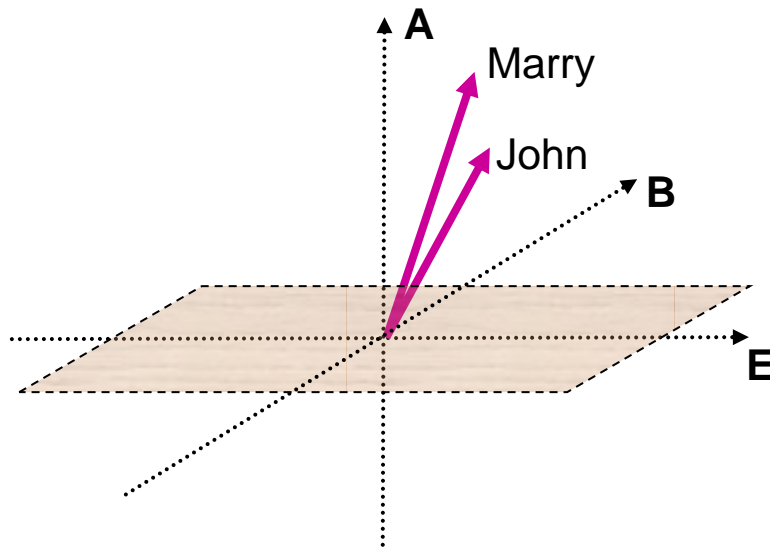
# Neighbor Selection

---

- 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

	<i>Item1</i>	<i>Item2</i>	<i>Item3</i>	<i>Item4</i>	<i>Item5</i>
User <i>A</i>	4	4	1	4	3
User <i>B</i>	2	1	4	2	5
User <i>C</i>	3	1	3	2	1
User <i>D</i>	5	4	2		3

# Item-based Collaborative Filtering

---

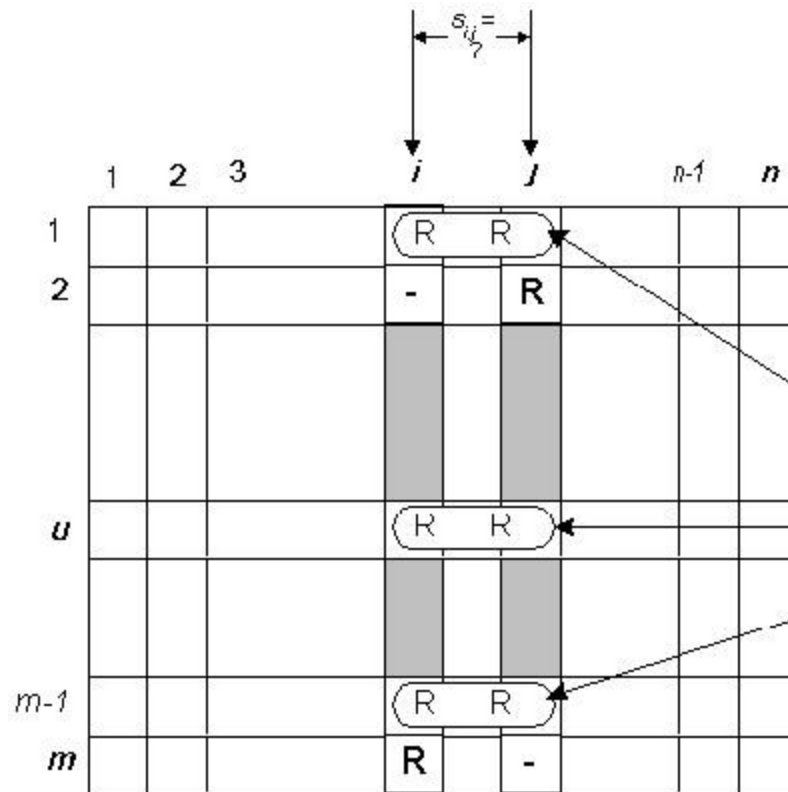
“Similar items are rated similarly!”

# Item-based Collaborative Filtering

---

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

# Finding Similar Items



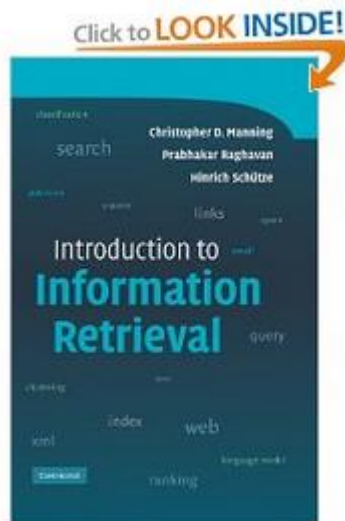
Computed by looking into co-rated items only. These co-rated pairs are obtained from different users.

# Amazon's book recommendation

---

“Users who bought this book, also bought that book”





## Customers Who Bought This Item Also Bought



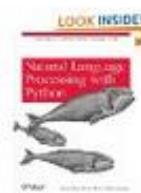
Speech and Language Processing (2nd Edition)  
 Daniel Jurafsky  
 ★★★★★☆ (32)  
 Hardcover  
 \$112.47



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



Foundations of Statistical Natural Language ...  
 > Christopher D. Manning  
 ★★★★★☆ (14)  
 Hardcover  
 \$56.84



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



Lucene in Action, Second Edition: Covers Apache ...  
 > Michael McCandless  
 ★★★★★☆ (30)  
 Paperback  
 \$31.36