

# RECOMMENDER SYSTEMS AND GRAPH PROCESSING

CKME 134 – BIG DATA ANALYTICS TOOLS

RYERSON UNIVERSITY

WINTER 2015

Instructor: Shaohua Zhang

# Course Outline

2

1. Intro to Big Data
2. Distributed Computing and MapReduce
3. Hadoop Ecosystem
4. Programming Hive
5. Advanced Hive
6. Mid-Term Review
7. Programming Pig
8. Advanced Pig
9. Hadoop Use Cases
10. Building Data Product & Next-Gen Hadoop (Spark)
11. **Beyond Hadoop: Graph Analytics and Recommender Systems**

# Exam and Preparation

3

- Final Exam Preparation
  - Course Notes – Session 1 ~ Session 9
  - Hive Intro Lab – Session 4 Lab
  - Pig Intro Lab 1 – Session 7 Lab
- Hangout Session 2
  - Engineering Building (lower level)
  - Saturday, April 11
  - 1 pm ~ 4pm

# Lecture 11 - Outline

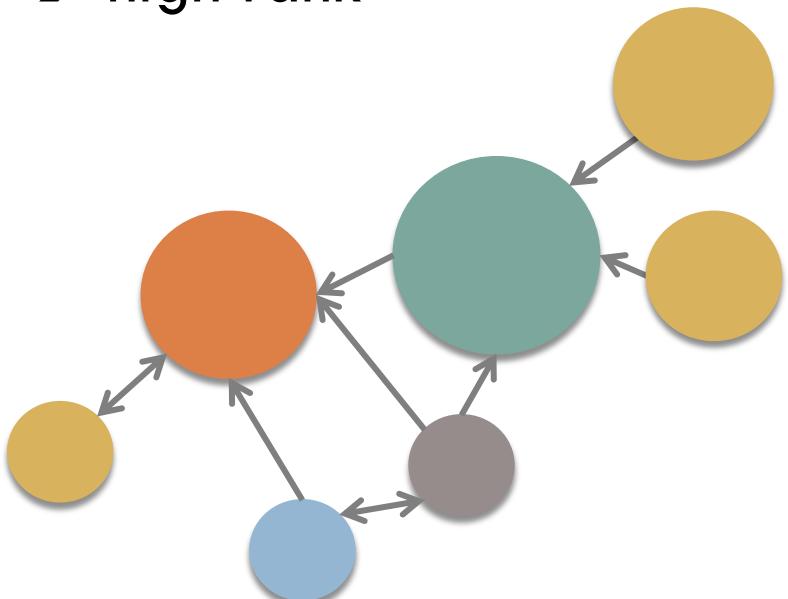
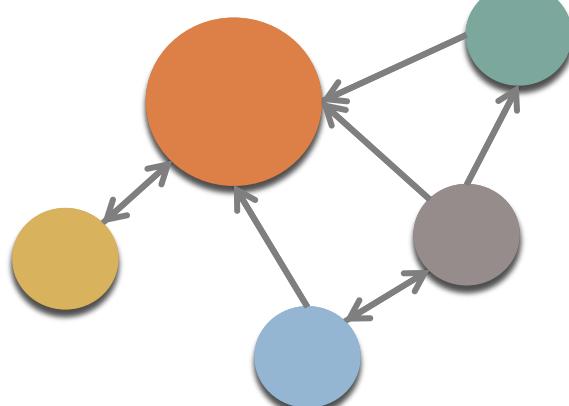
4

- Recommender Systems 101
- Graph Processing
  - Spark GraphX
  - GraphLab
  - Giraph

# PageRank Algorithm

5

- PageRank gives web pages a ranking score on links from other pages
  - Links from many pages → high rank
  - Link from a high-rank page → high rank



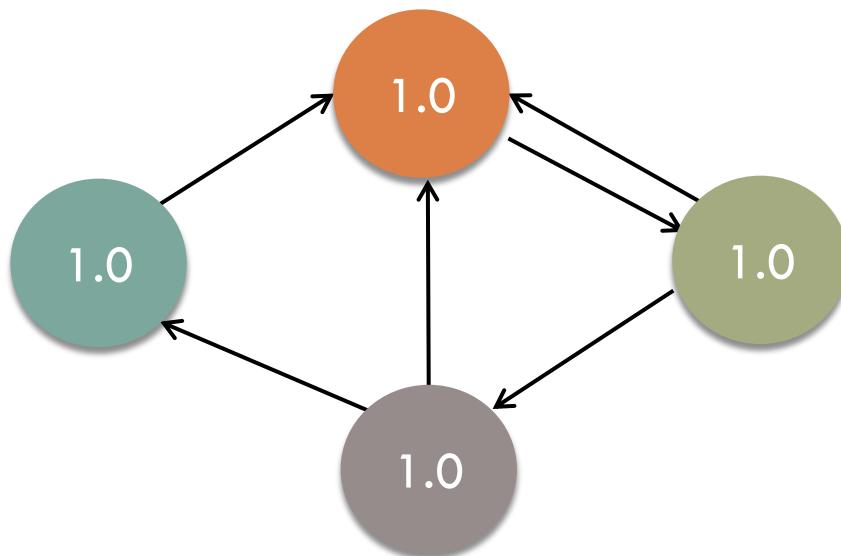
Graph (directed)

# PageRank Explained – Initial State

1. Start each page at a rank of 1

```
# set initial page ranks to 1.0
ranks=links.map(lambda (page,neighbors): (page,1.0))
```

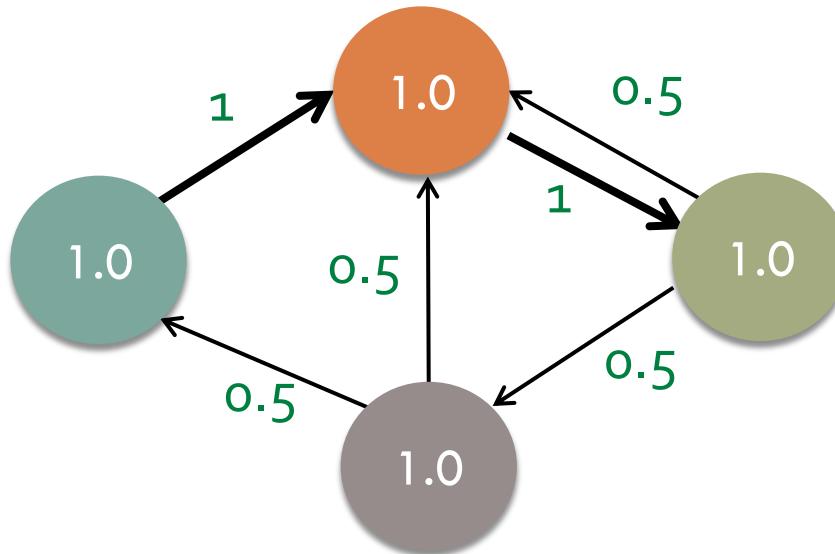
```
links = sc.textFile(linkfile).map(lambda line: line.split())\
    .map(lambda pages: (pages[0],pages[1]))\
    .distinct()\
    .groupByKey()\
    .cache()
```



# PageRank Explained – Iteration 1

1. Start each page at a rank of 1
2. On each iteration, have page  $p$  contribute  
 $\text{contrib}_p = \text{rank}_p / \text{neighbors}_p$  to its neighbors

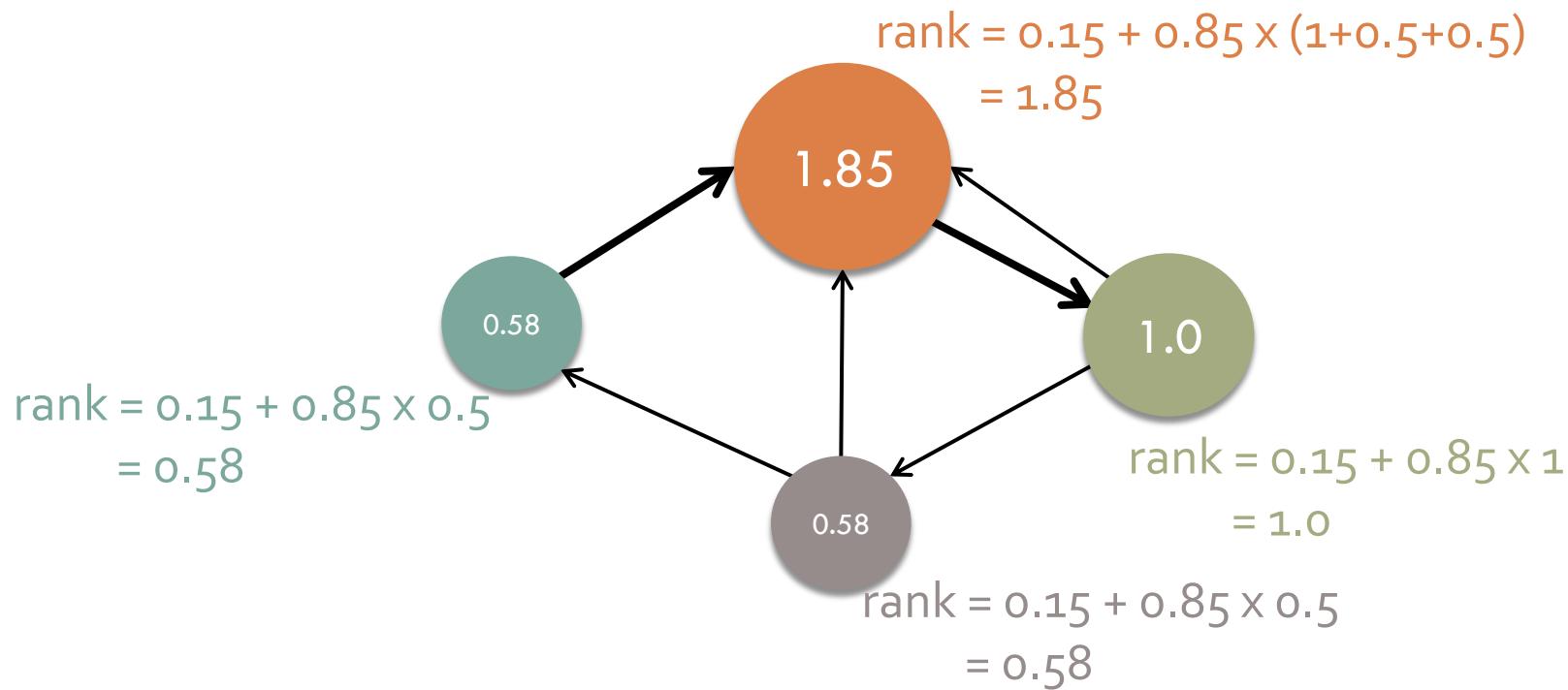
```
def computeContribs(neighbors, rank):  
    for neighbor in neighbors:  
        yield(neighbor, rank/len(neighbors))
```



# PageRank Explained – Iteration 1

1. Start each page at a rank of 1
2. On each iteration, have page p contribute  
 $\text{contrib}_p = \text{rank}_p / \text{neighbors}_p$  to its neighbors
3. Set each page's rank to  $0.15 + 0.85 \times \text{contribs}$

```
for x in xrange(n):
    contribs=links\
        .join(ranks)\.
        flatMap(lambda (page,(neighbors,rank)): \
            computeContribs(neighbors,rank) ) |
    ranks=contribs\
        .reduceByKey(lambda v1,v2: v1+v2)\.
        map(lambda (page,contrib): \
            (page,contrib * 0.85 + 0.15))
    print "Iteration ",x
    for pair in ranks.take(10): print pair
```

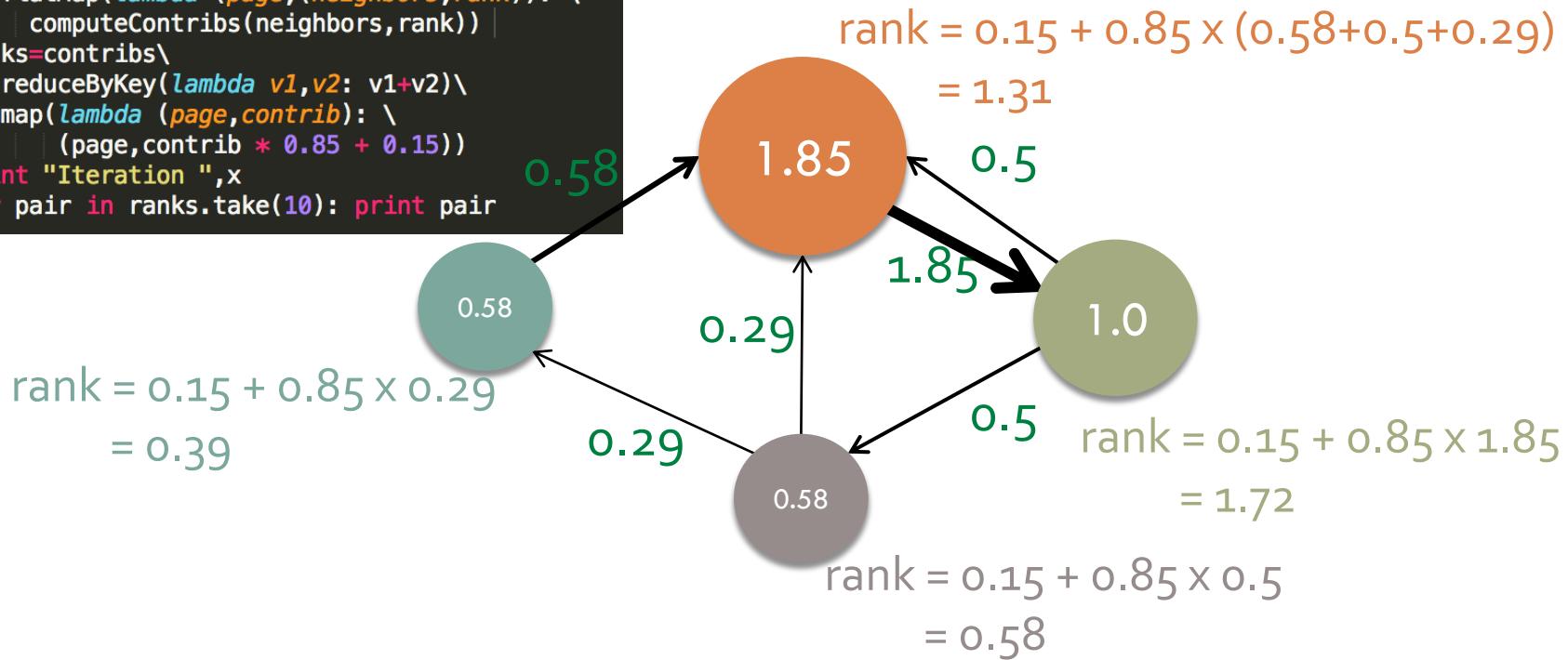


# PageRank Explained – Iteration 2

1. Start each page at a rank of 1
2. On each iteration, have page  $p$  contribute  
 $\text{contrib}_p = \text{rank}_p / \text{neighbors}_p$  to its neighbors
3. Set each page's rank to  $0.15 + 0.85 \times \text{contribs}$

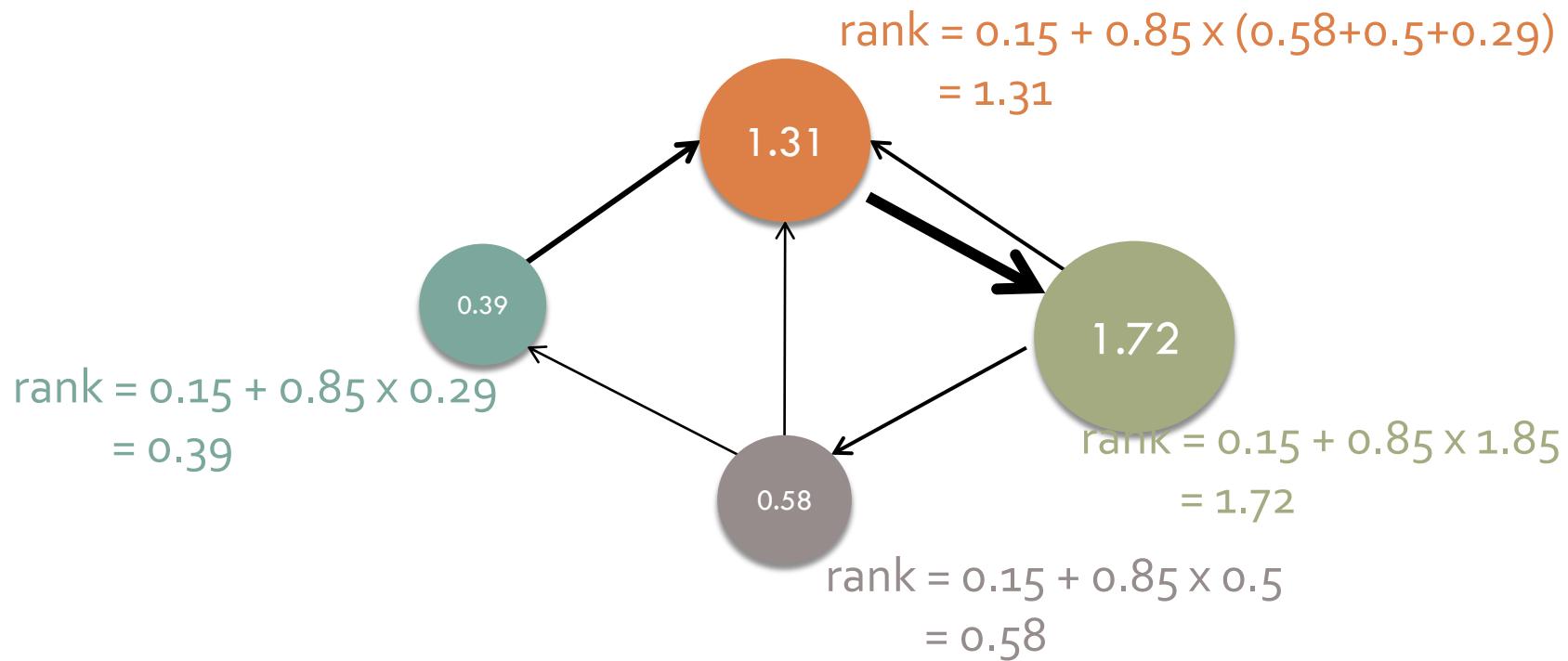
```
def computeContribs(neighbors, rank):  
    for neighbor in neighbors:  
        yield(neighbor, rank/len(neighbors))
```

```
for x in xrange(n):  
    contribs=links\  
    .join(ranks)\  
    .flatMap(lambda (page,(neighbors,rank)): \  
        computeContribs(neighbors,rank)) |  
    ranks=contribs\  
    .reduceByKey(lambda v1,v2: v1+v2)\  
    .map(lambda (page,contrib): \  
        (page,contrib * 0.85 + 0.15))  
    print "Iteration ",x  
    for pair in ranks.take(10): print pair
```



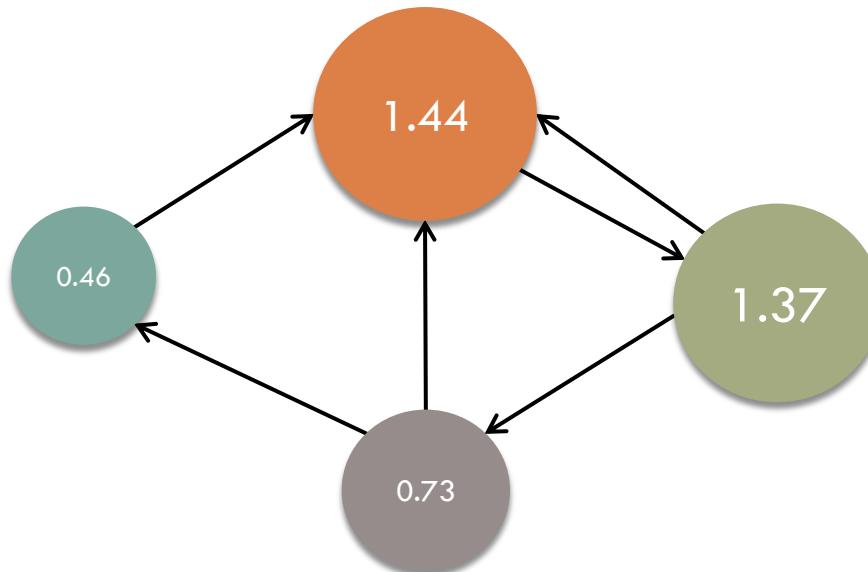
# PageRank Explained – Iteration 2

1. Start each page at a rank of 1
2. On each iteration, have page p contribute  
 $\text{contrib}_p = \text{rank}_p / \text{neighbors}_p$  to its neighbors
3. Set each page's rank to  $0.15 + 0.85 \times \text{contribs}$



# PageRank Explained – Final State

1. Start each page at a rank of 1
2. On each iteration, have page p contribute  
 $\text{contrib}_p = \text{rank}_p / \text{neighbors}_p$  to its neighbors
3. Set each page's rank to  $0.15 + 0.85 \times \text{contribs}$



# Spark Demo

12

## □ PageRank Implementation in Spark

- Download and install Cloudera 5.3 Virtual Machine (Lecture 10 Lab)
- In Virtualbox, create a new VM
- Name "Cloudera 5.3 VM with Spark"
- Type "Linux"
- Version "Other Linux (64-bit)"
- Set memory size to 4096
- Choose "Use an existing virtual hard drive file" and select in the directory "Cloudera-QuickStart-VM-5.3.0-0-virtualbox-disk1.vmdk"
- Create and start the VM
- Once you're in the Cloudera 5.3 VM, download the pagelinks.txt dataset and PageRank.py script into "/home/cloudera/Downloads" folder
- Go to this folder, and run the PageRank.py script
- \$ pyspark PageRank.py
- After several iterations, you should be able to see the pagerank results printed



# Introduction to Recommender Systems

# Recommender Systems

## Examples

14

- Movie
  - Netflix
- Music
  - Spotify/Pandora/Songza
- Social
  - LinkedIn/Facebook/Twitter
- e-Commerce
  - Amazon/eBay/Taobao
- LBS
  - Foursquare/Yelp!
- News
  - Yahoo!/Flipboard/FB News Feeds
- Events
  - Eventbrite
- Fashion
  - Pose (fashion style graph)
- Dating
  - eHarmony/OkCupid

# Recommender Systems

## Why Recommendations?

15

- Consumer Perspective
  - Information Overload
    - Too many products/items
  - Efficiency and relevancy
  - Personalized experience
- Company Perspective
  - Product discoveries → Cross-sell
  - Engagement → Ads
  - Customer satisfaction

# Recommender Systems

## *Types of Recommendations*

16

- Editorial and hand curated
  - Newspaper
- Simple aggregates
  - Top 10, Most Popular
- Personalized – tailored to users
  - Amazon, Netflix
  - Facebook News Feeds

# Recommender Systems

## *Types of Data*

17

- Star Ratings – Explicit
  - Netflix Prize
- Beyond Ratings
  - Implicit Cues
    - Clicks, Likes, Comments
    - Download/Install
    - Time Spent on page
  - User Profiles
    - Demographics → Gender, Age, Marital Status
    - Interests → Sailing, Music, Karate
  - Social Signals
    - Facebook graph
    - Tweets
  - Contextual
    - Location
    - Weather
    - Mode of Transportation
    - etc.

# Recommender Systems

## *Types of Algorithms*

18

- Content-based
- Collaborative Filtering
  - Memory-based
    - User-User
    - Item-Item
  - Model-based
- Hybrid

# Recommender Systems

## *Content-based Recommenders*

19

- Main Idea: Recommend items to customer x similar to previous items rated highly by x
- Examples
  - Movie Recommendation
    - Recommend movies with same directors, actors, genre
  - News Recommendation
    - Recommend news with similar content/topics

# Recommender Systems

## Content-based Recommenders – Item Profiles

20

- For each item, create an item profile
- Profile is a set of features
  - Movies: author, title, actors, director, ...
  - News: Set of important keywords in document

$$TF_{ij} = \frac{f_{ij}}{\max_k f_{kj}}$$

$$w_{ij} = TF_{ij} \times IDF_i$$

$$IDF_i = \log \frac{N}{n_i}$$

	Director		Title, Metadata					Actor					Genre				
	Steven	Woody	Speed	War	Love	Future	Home	Brad Pitt	Jennifer Laurence	Bradley Cooper	Tom Hanks	Jennifer Aniston	Horror	Action	Love	Cartoon	International
Movie 1	1				1		1	1	1			1	1				
Movie 2		1		1					1		1	1		1			
Movie 3	1			1	1			1				1			1		
Movie 4	1		1			1			1		1		1				

# Recommender Systems

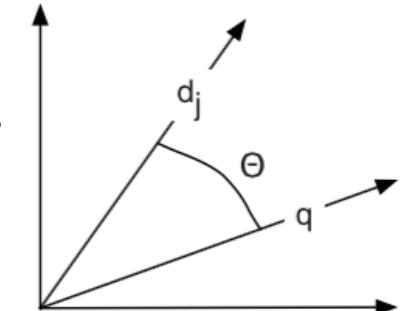
## Content-based Recommenders – User Profiles & Predictions

21

	Director		Title, Metadata					Actor					Genre				
	Steven	Woody	Speed	War	Love	Future	Home	Brad Pitt	Jennifer Laurence	Bradley Cooper	Tom Hanks	Jennifer Aniston	Horror	Action	Love	Cartoon	International
Tom	0.87		0.3	0.5				0.9	0.77		0.34		0.67			0.33	0.54

### User Profiles

- Weighted average of rated item profiles
- Taxonomy/Ontology based
- etc.



$$d_j = \langle w_{1,j}, w_{2,j}, \dots, w_{n,j} \rangle$$

$$q = \langle w_{1,q}, w_{2,q}, \dots, w_{n,q} \rangle$$

w = weight assigned to term

### Prediction

- Cosine similarity between item profiles and user profiles

	Director		Title, Metadata					Actor					Genre				
	Steven	Woody	Speed	War	Love	Future	Home	Brad Pitt	Jennifer Laurence	Bradley Cooper	Tom Hanks	Jennifer Aniston	Horror	Action	Love	Cartoon	International
Movie 1	1				1		1	1	1			1	1		1		
Movie 2		1		1					1		1	1			1		
Movie 3	1				1	1		1				1				1	
Movie 4	1		1				1		1		1		1				
Tom	0.87		0.3	0.5				0.9	0.77		0.34		0.67			0.33	0.54

# Recommender Systems

## *Content-based Recommenders – Pros and Cons*

22

- No cold start or sparsity problems
  - No dependency on other items/users
- Able to recommend to users with unique tastes
- Able to recommend new and unpopular items
- Easy to explain to users

# Recommender Systems

## *Collaborative Filtering – User-based Recommender*

23

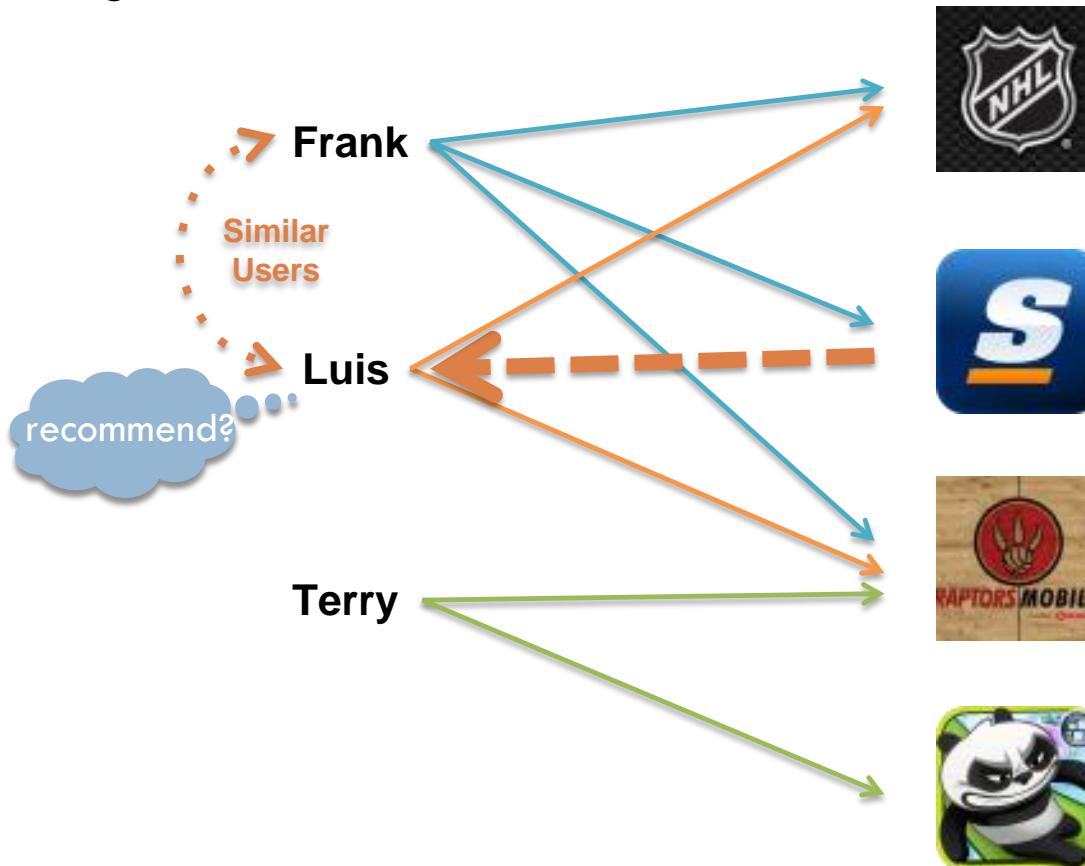
- Consider user  $x$
- Find set of  $k$  other users whose ratings are “similar” to  $x$ ’s ratings
- Estimate  $x$ ’s ratings based on ratings of users in  $k$

# Recommender Systems

## *Collaborative Filtering – User-based Recommender*

24

### □ Finding similar users



# Recommender Systems

## *Collaborative Filtering – User-based Recommender*

25

### □ Similarity Metrics

□ Jaccard

Jaccard similarity

$$sim(\vec{x}_i, \vec{x}_j) = \frac{|\vec{x}_i \cap \vec{x}_j|}{|\vec{x}_i \cup \vec{x}_j|}$$

□ Cosine Similarity

Cosine Similarity

□ Pearson Correlation

$$sim(\vec{x}_i, \vec{x}_j) = \frac{\vec{x}_i \cdot \vec{x}_j}{|\vec{x}_i| * |\vec{x}_j|} = \frac{\sum_u r_{u,x_i} * r_{u,x_j}}{\sqrt{\sum_u r_{u,x_i}^2} * \sqrt{\sum_u r_{u,x_j}^2}}$$

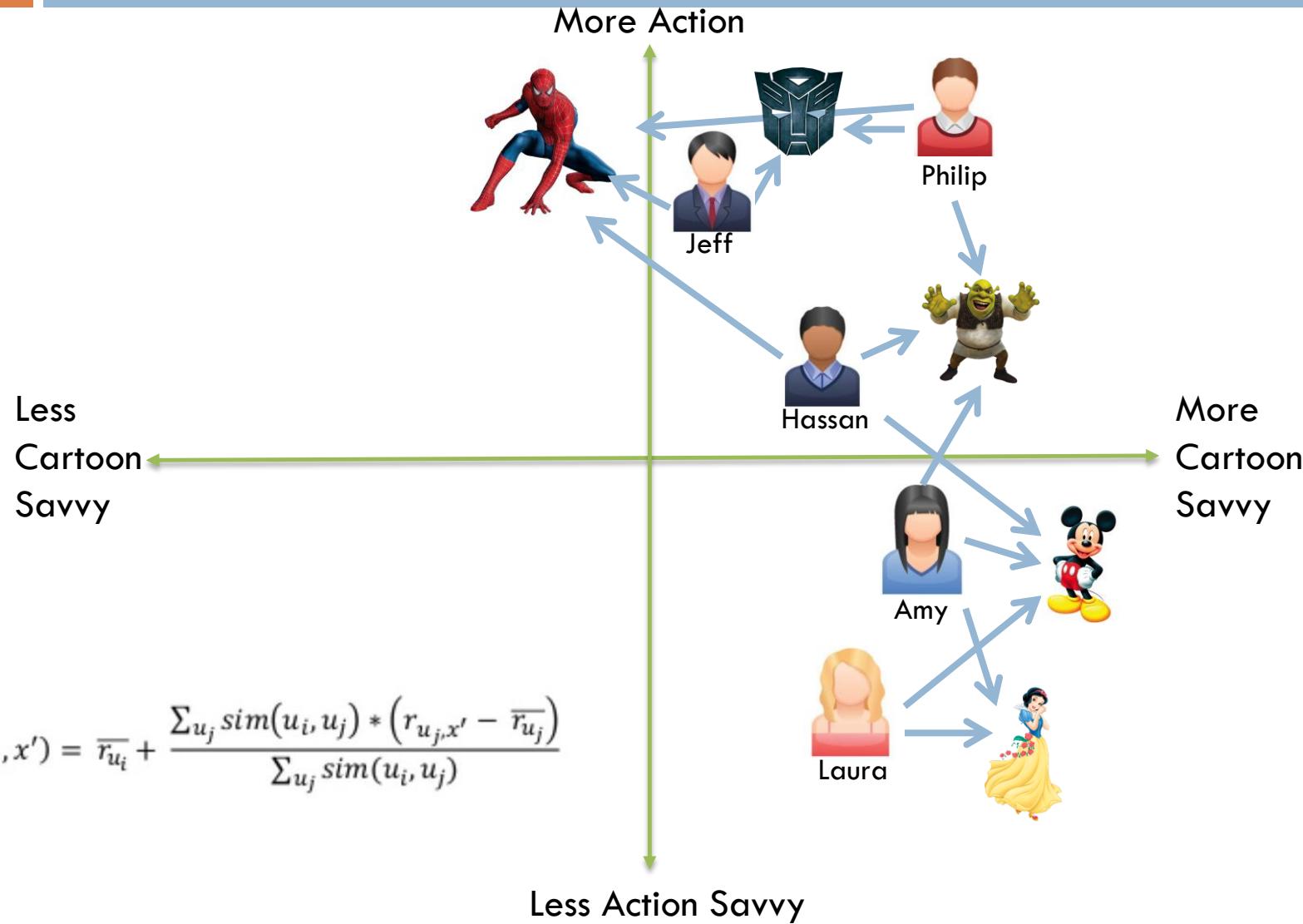
Pearson's correlation coefficient

$$sim(u_i, u_j) = \frac{\sum_{x \in X} (r_{u_i,x} - \bar{r}_{u_i}) * (r_{u_j,x} - \bar{r}_{u_j})}{\sqrt{\sum_{x \in X} (r_{u_i,x} - \bar{r}_{u_i})^2} * \sqrt{\sum_{x \in X} (r_{u_j,x} - \bar{r}_{u_j})^2}}$$

# Recommender Systems

## Collaborative Filtering – User-based Recommender

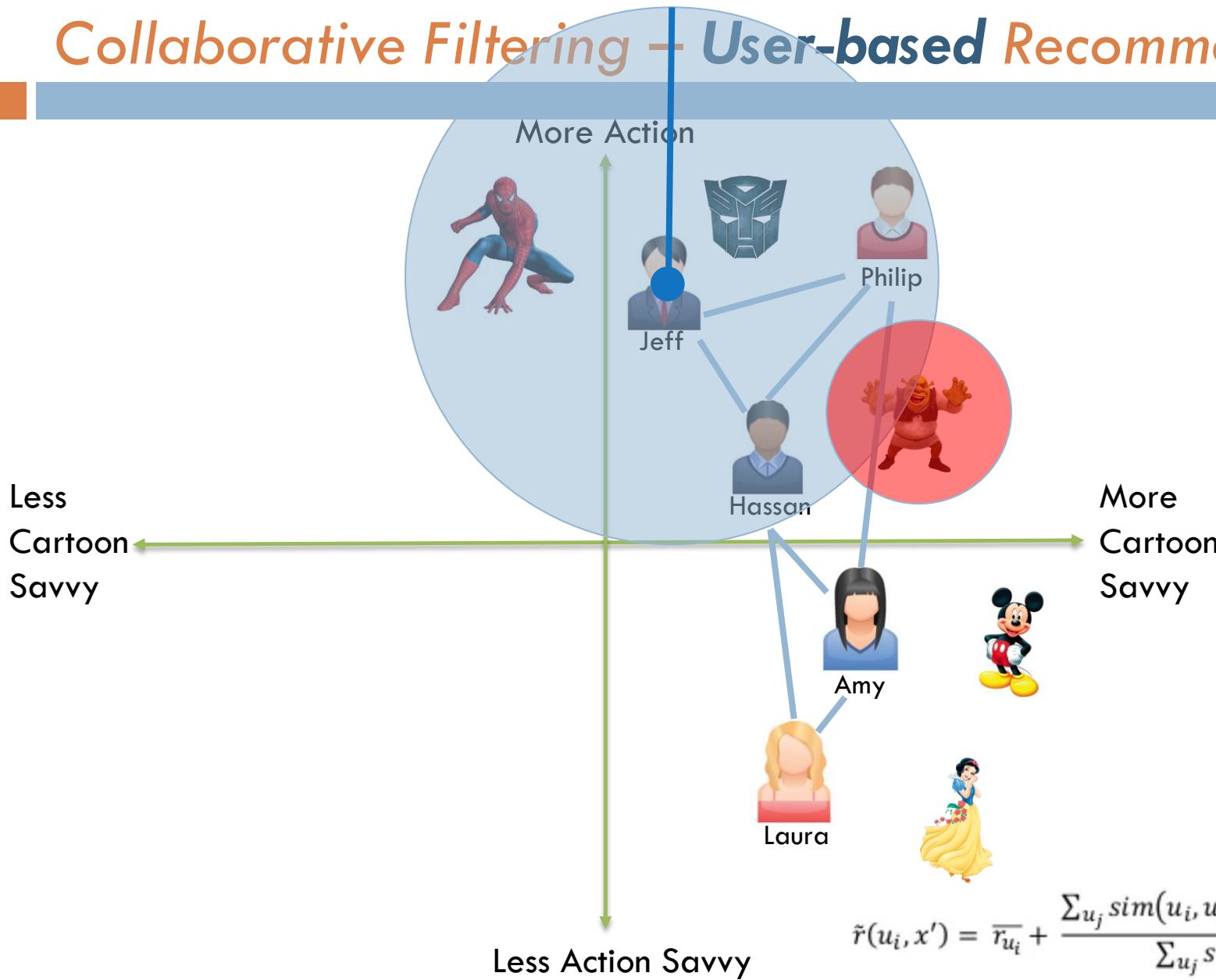
26



# Recommender Systems

## Collaborative Filtering – User-based Recommender

27

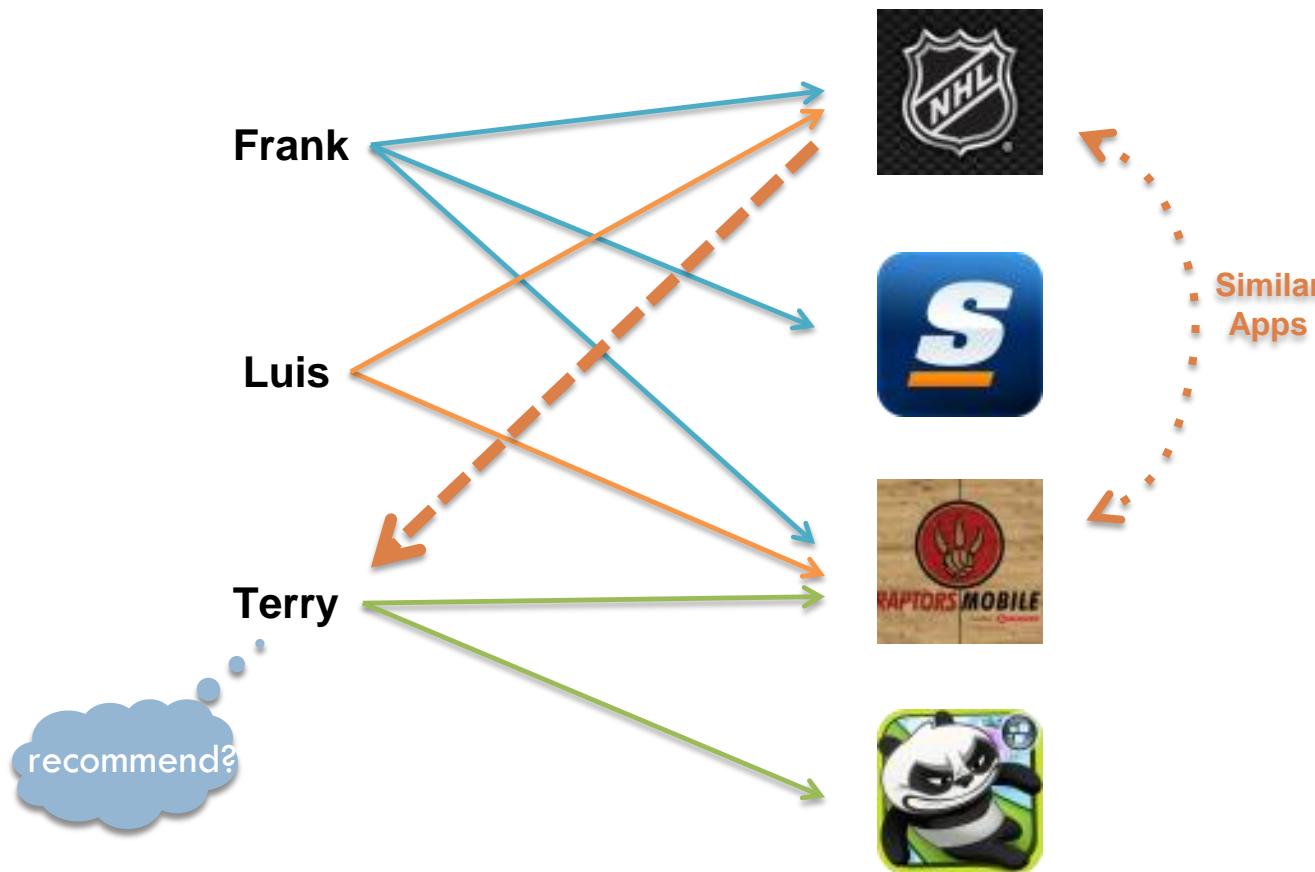


$$\tilde{r}(u_i, x') = \bar{r}_{u_i} + \frac{\sum_{u_j} sim(u_i, u_j) * (r_{u_j, x'} - \bar{r}_{u_j})}{\sum_{u_j} sim(u_i, u_j)}$$

# Recommender Systems

## Collaborative Filtering – Item-based Recommender

28



# Recommender Systems

## Collaborative Filtering – Item-based Recommender

29



Movie Watching Logs	Spiderman	Shrek	Snowwhite	Transformers	Mickey Mouse
Jeff	1			1	
Amy		1	1		1
Laura			1		1
Hasan	1	1			1
Fillip	1	1		1	

Movie Watching Matrix



Co-occurrence Matrix	Spiderman	Shrek	Snowwhite	Transformers	Mickey Mouse
Spiderman	3	2	0	2	1
Shrek	2	3	1	1	2
Snowwhite	0	1	2	0	2
Transformers	2	1	0	2	0
Mickey Mouse	1	2	2	0	3

Co-occurrence Matrix

# Recommender Systems

## Collaborative Filtering – Item-based Recommender

30



Co-occurrence Matrix	Spiderman	Shrek	Snowwhite	Transformers	Mickey Mouse
Spiderman	3	2	0	2	1
Shrek	2	3	1	1	2
Snowwhite	0	1	2	0	2
Transformers	2	1	0	2	0
Mickey Mouse	1	2	2	0	3

Co-occurrence Matrix



Cosine Similarity Matrix	Spiderman	Shrek	Snowwhite	Transformers	Mickey Mouse
Spiderman	1	0.67	0	0.82	0.33
Shrek	0.67	1	0.41	0.41	0.67
Snowwhite	0	0.41	1	0	0.82
Transformers	0.82	0.41	0	1	0
Mickey Mouse	0.33	0.67	0.82	0	1

Cosine-Similarity Matrix



# Recommender Systems

## Collaborative Filtering – Item-based Recommender

31



Movie Watching Logs	Spiderman	Shrek	Snowwhite	Transformers	Mickey Mouse
Jeff	1			1	
Amy		1	1		1
Laura			1		1
Hasan	1	1			1
Fillip	1	1		1	

Movie Watching Matrix



Cosine Similarity Matrix	Spiderman	Shrek	Snowwhite	Transformers	Mickey Mouse
Spiderman	1	0.67	0	0.82	0.33
Shrek	0.67	1	0.41	0.41	0.67
Snowwhite	0	0.41	1	0	0.82
Transformers	0.82	0.41	0	1	0
Mickey Mouse	0.33	0.67	0.82	0	1

Cosine Similarity Matrix



Movie Watching Logs	Spiderman	Shrek	Snowwhite	Transformers	Mickey Mouse
Jeff	1			1	
Amy		1	1		1
Laura			1		1
Hasan	1	1			1
Fillip	1	1		1	

Movie Watching Matrix



Prediction	Spiderman	Shrek	Snowwhite	Transformers	Mickey Mouse
Jeff	1	<b>1.08</b>	0	1	0.33
Amy	1	1	1	0.41	1
Laura	0.33	1.08	1	0	1
Hasan	1	1	1.23	1.23	1
Fillip	1	1	0.41	1	1

Movie Watching Logs	Spiderman	Shrek	Snowwhite	Transformers	Mickey Mouse
Jeff	1			1	
Amy		1	1		1
Laura			1		1
Hasan	1	1			1
Fillip	1	1		1	

Cosine Similarity Matrix	Shrek
Spiderman	0.67
Shrek	1
Snowwhite	0.41
Transformers	0.41
Mickey Mouse	0.67

Movie Watching Logs	Spiderman	Shrek	Snowwhite	Transformers	Mickey Mouse
Jeff	1			1	
Amy		1	1		1
Laura			1		1
Hasan	1	1			1
Fillip	1	1		1	

Movie Watching Logs	Spiderman	Shrek	Snowwhite	Transformers	Mickey Mouse
Jeff	1			1	
Amy		1	1		1
Laura			1		1
Hasan	1	1			1
Fillip	1	1		1	

Prediction	Shrek
Jeff	1.08

Cosine Similarity Matrix	Shrek
Spiderman	0.67
Shrek	1
Snowwhite	0.41
Transformers	0.41
Mickey Mouse	0.67

# Recommender Systems

## *Collaborative Filtering – Pros and Cons*

32

### □ Pros

- Works for any kind of item
  - No feature selection needed
- Predictive when you have lots of data

### □ Cons

- Cold start (not enough users/items)
- Sparsity
  - The user/ratings matrix is sparse
  - Hard to find users that have rated the same items
- Popularity-driven
  - Cannot recommend items to someone with unique tastes
  - Cannot recommend items that are less popular

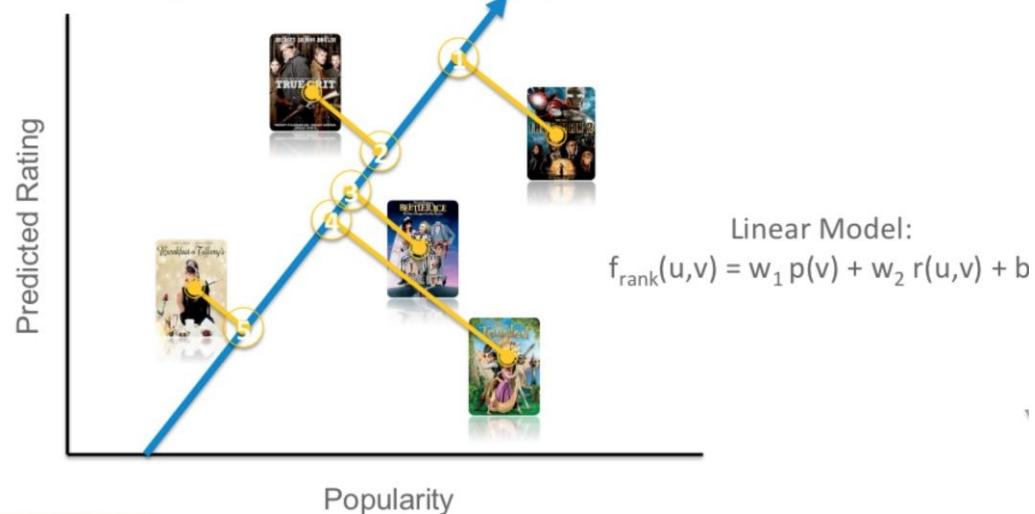
# Recommender Systems

## Hybrid Recommenders

33

### □ Hybrid Approaches

- Implement two or more different recommenders and combine predictions
- Add content-based methods to CF
  - Item profiles for new item problems
  - User profiles for new user problems
- Learning to Rank!



# Recommender Systems

## Challenges

34

- Cold Start
  - New User Problem
  - New Item Problem
- Sparsity & High Dimensionality
  - Matrix Factorization
- Deployment
  - Graph database
  - NoSQL (Cassandra)

# Graph Processing Tools

# Problems that can be solved with graph

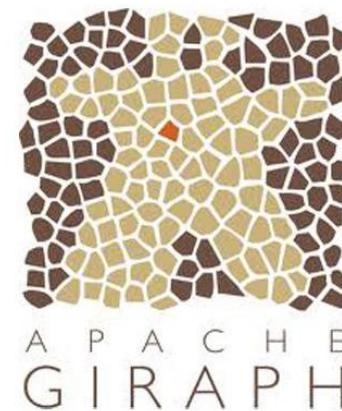
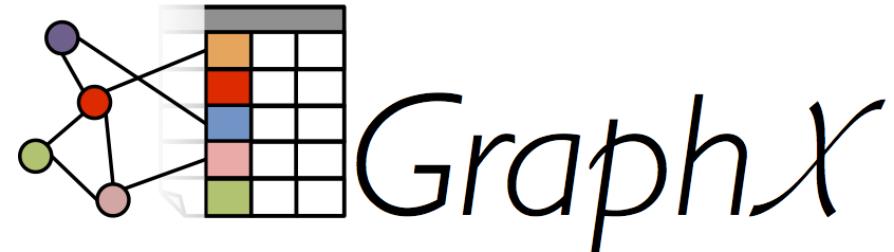
36

- Recommendation Algorithms
- Machine Learning
- Graph Computation
- Topic Modeling
- Label Propagation
- Probabilistic Graphic Modeling
- etc.

# Graph Processing Tools

37

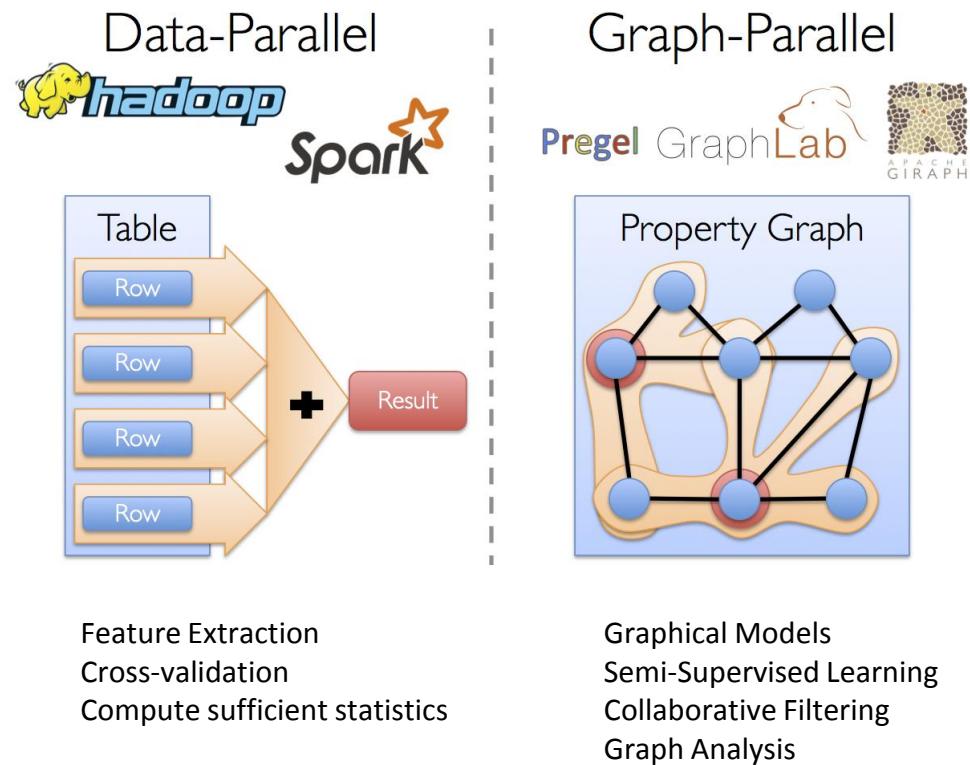
- GraphX
- GraphLab
- Giraph



# Data Parallel vs Graph Parallel

38

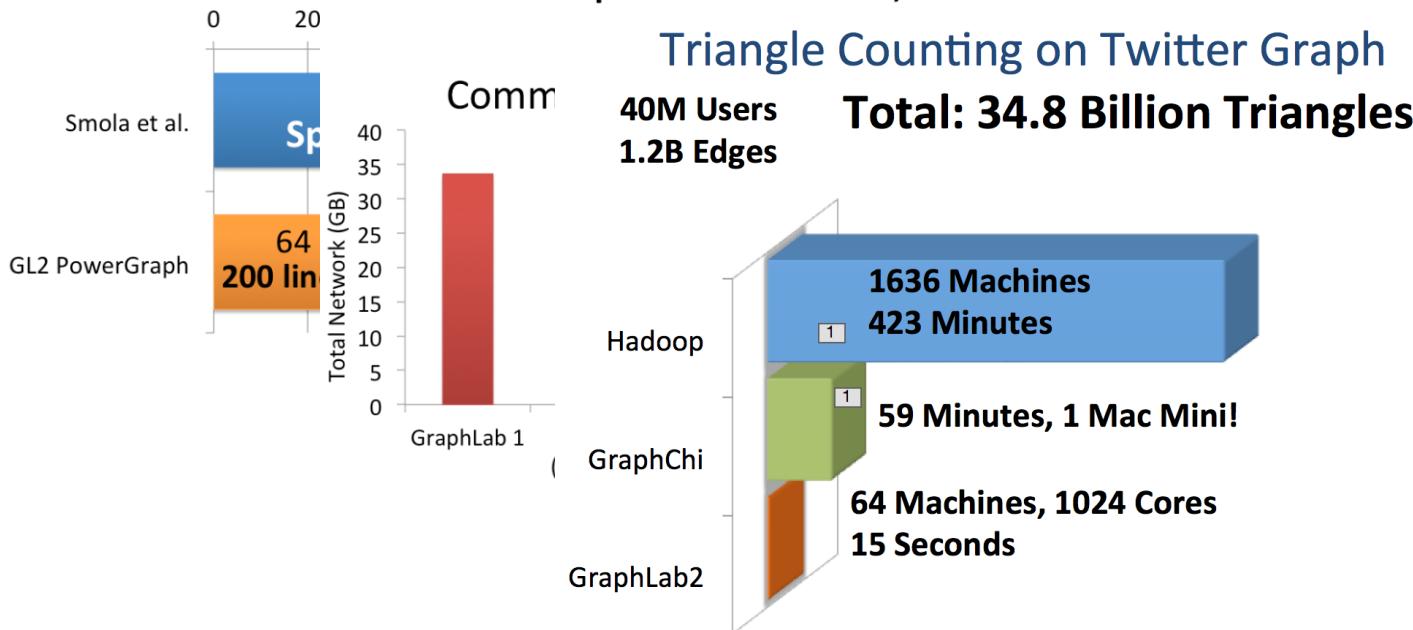
- Challenges in partition and distribute graphs
- Orders of magnitude faster than more general *data-parallel* systems
- Dependency graph



# GraphLab

39

- English language Wikipedia
  - 2.6M Documents, 8.3M Words, 500M Tokens
  - Computationally intensive algorithm



# Additional Lab

40

- Install GraphLab-Create on Cloudera 5.3 VM
  - Follow the instructions posted on blackboard
- Run a basic recommender on movie-lens dataset

# Where to go from here?

# SHARE 8,333 <sup>41</sup> VIDEOS.

# SKYPE USERS CONNECT FOR **23,300** HOURS.

**YELP** USERS  
POST  
**26,380**  
**REVIEWS.**

# APPLE USERS DOWNLOAD **48,000** apps

# PANDORA USERS LISTEN TO **81.14%**

# AMAZON

# INSTAGRAM USERS »

# TWITTER USERS TWEET **277,000**

# 2,460,000 PIECES OF CONTENT.

# TINDER USERS SWIPE 416,667 TIMES.

# WHATSAPP — USERS SHARE — **347,222** **PHOTOS.**

# Where to go from here?

## *Practice, Practice, Practice!*

42

- Taking 50 data science courses don't make you a great data scientist
  - <http://www.datasciencecentral.com/m/blogpost?id=6448529:BlogPost:262632>
  - Practice
  - Read best practices
  - Understanding why!
- Books → read 2 to 3 good ones
  - One Practical predictive modeling
  - One statistical learning
  - One data preparation
  - One big data (hadoop, etc.)

# Where to go from here?

## *Capstone Project*

43

- Find a project that will allow you to practice Hadoop, R/Python and Visualization
  - Find a dataset
  - Define your problem
  - Set up your analytics environment (AWS)
  - Data munging in Hadoop
  - Modeling in R, Python and Spark
    - Try at least 4 different algorithms
  - Build a data pipeline
  - Publish the results to github as notebook
  - Create visualizations (data exploration, final product)

# Where to go from here?

## *Work with Interesting Datasets*

44

- Retail/Marketing
  - KDD Cup 1998 Direct Mail Campaign
  - Alibaba TianChi Competition
- Telco
  - Churn dataset (KDD Cup 2009 Orange)
- Banking
  - German credit data
- Social
  - Twitter graph
  - Foursquare graph
  - Tencent KDD 2012
- Fraud/Intrusion Detection
  - German credit fraud
  - Enron email fraud
  - Intrusion detection kdd99
- Geo
  - Geo-tagged tweets
  - NYC yellow cab
  - Foursquare
- Textual
  - Yelp! challenge
- Recommender
  - MovieLens
  - Netflix
  - Event Recommender (Kaggle)

# Where to go from here?

## Learn from Best Practices

45

- Kaggle Competition
  - Interesting problems
  - Tons of Python/R solutions posted
  - Learn from the best!
  - Join the competition to practice



Completed • \$50,000 • 1,568 teams

### Allstate Purchase Prediction Challenge

Tue 18 Feb 2014 – Mon 19 May 2014 (10 months ago)

Dashboard ▾

Competition Forum

All Forums » Allstate Purchase Prediction Challenge

Search

« Prev Topic

I wrote a paper on the Allstate competition - feedback is welcome!

Next Topic »

Start Watching

11

Hi everyone,

I just finished a class in Data Science, and competed in the Allstate competition as my class project. My [project paper](#), [presentation slides](#), and [R code](#) are posted online. (EDIT: I also just recorded the presentation and put it on [YouTube](#).)

If anyone takes a look through these materials, I would love to get your feedback! I welcome any type of feedback, from the general to the detailed, and from Kaggle masters or novices. This was my first "real" Kaggle competition, and I want to learn as much as possible since I plan to enter many more Kaggle competitions.

And of course, if people have any questions about my techniques or code, I'm happy to answer them.

I'll also summarize my best solution in the [solution sharing thread](#).

Thanks!

Kevin

<http://www.kaggle.com/c/allstate-purchase-prediction-challenge/leaderboard/private>

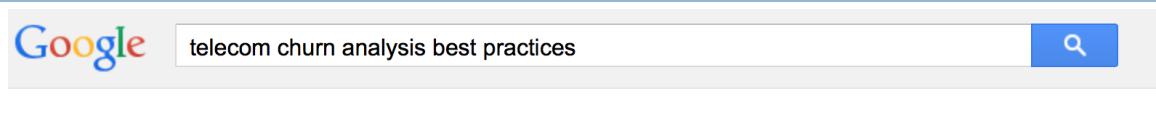
<https://github.com/justmarkham/kaggle-allstate/blob/master/allstate-paper.md>

# Where to go from here?

## Learn from Best Practices

46

### □ Google



#### Customer Churn Analysis - A scientific way to treat churn

Ad [www.preact.com/churn-analysis](http://www.preact.com/churn-analysis) ▾

Data science makes retention simple  
Customer Health · Customer Success

[Request Demo](#)

[Free Trial](#)

[Interactive Tour](#)

#### Telecom Analysis - We Only Get Paid If You Save

Ad [www.abilita.com/Telecom-Analysis](http://www.abilita.com/Telecom-Analysis) ▾ +1 888-836-4968

Free Consultation & No Upfront Fee.  
About The Company · Case Studies · Request a Call · Corporate Contact  
[Telecom Solutions - Schedule Appointment](#) - [Contact Us](#) - [Case Studies](#)

#### Big Data to improve churn analysis in the telecoms industry

[www.reply.eu/.../analytics/big-data-to-improve-churn-analysis-in-the-tele...](http://www.reply.eu/.../analytics/big-data-to-improve-churn-analysis-in-the-tele...) ▾

Best Practice ... The telecoms market provides a good example of why the high acquisition costs and slim profit margins for each customer make churn analysis vital to help companies identify and retain the most profitable among them.

#### Analytics, KPIs for effective Churn & Loyalty management

[www.slideshare.net/EhtishamRao1/churnloyalty-ccm](http://www.slideshare.net/EhtishamRao1/churnloyalty-ccm) ▾

Oct 29, 2011 - Broad agenda A look at Loyalty and Churn in Telcos Aligning and ... of loyalty Top reasons behind voluntary Churn in Telecom What would have ... Share best practice templates Planning & Marketing Analytics Resource ...

#### Breaking the back of customer churn - Bain Brief - Bain ...

[www.bain.com/publications/.../breaking-the-back-of-customer-churn.asp...](http://www.bain.com/publications/.../breaking-the-back-of-customer-churn.asp...) ▾

Feb 5, 2014 - Communications service providers know that churn corrode their business, ... Sentiment analysis, for instance, can mine social media for what people ... To do that, leading companies identify and invest in those few .... Frédéric Debruyne is a Brussels-based partner in the firm's Telecommunications practice.

# Where to go from here?

## Learn from Best Practices

47

### □ GitHub

- Learn from others
- Start building your own repos!

The screenshot shows a GitHub repository page for 'ipython / ipython'. The title is 'A gallery of interesting IPython Notebooks'. It includes a sidebar with navigation links like Home, A gallery of interesting IPython Notebooks, Code blocks and other ideas, Cookbook: Branding the IPython notebook, etc. The main content area contains a note about contributing content, a link to nbviewer, and a note about bookmarklets and extensions. Below this is a 'Table of Contents' section with two main categories: 'Entire books or other large collections of notebooks on a topic' and 'Scientific computing and data analysis with the SciPy Stack'.

A gallery of interesting IPython Notebooks

Fernando Perez edited this page 2 days ago · 229 revisions

This page is a curated collection of IPython notebooks that are notable for some reason. Feel free to add new content here, but please try to only include links to notebooks that include interesting visual or technical content; this should *not* simply be a dump of a Google search on every ipynb file out there.

**Important contribution instructions:** If you add new content, please ensure that for any notebook you link to, the link is to the rendered version using [nbviewer](#), rather than the raw file. Simply paste the notebook URL in the nbviewer box and copy the resulting URL of the rendered version. This will make it much easier for visitors to be able to immediately access the new content.

Note that [Matt Davis](#) has conveniently written a set of [bookmarklets and extensions](#) to make it a one-click affair to load a Notebook URL into your browser of choice, directly opening into nbviewer.

## Table of Contents

1. Entire books or other large collections of notebooks on a topic
  - [Introductory Tutorials](#)
  - [Programming and Computer Science](#)
  - [Statistics, Machine Learning and Data Science](#)
  - [Mathematics, Physics, Chemistry, Biology](#)
  - [Earth Science and Geo-Spatial data](#)
  - [Linguistics and Text Mining](#)
  - [Signal Processing](#)
2. Scientific computing and data analysis with the SciPy Stack

<https://github.com/ipython/ipython/wiki/A-gallery-of-interesting-IPython-Notebooks>

# Where to go from here?

## Learn from Best Practices

48

### □ Read Data Science News/Blogs

- DataTau
- Data Science Weekly
- Data Science Central
- etc. (just Google)

The screenshot shows a website with a chalkboard background featuring mathematical equations like  $x = \log_{10} t$ . The top navigation bar includes links for 'Data Science Weekly', 'Blog', 'Data Scientist Interviews', and 'Data Science Resources'. A breadcrumb navigation 'Home / Data Science Resources' is visible above the main content area. The main title 'Data Science Resources' is displayed prominently. Below it is a list of categories with descriptions:

- **Data Science Books**  
A list of books covering Data Analysis, Data Science, Machine Learning, Data Visualization, Statistics & Associated Programming Languages
- **Data Science Meetups**  
A list of Data Science Meetups from around the world
- **Data Science MOOCs**  
A list of Data Science related MOOCs
- **Data Science Datasets**  
A list of publicly available datasets
- **Data Science Most Read Articles**  
Most read articles from the Data Science Weekly Newsletter by Quarter
- **Data Scientist Talks**  
A list of talks from prominent Data Scientists

<http://www.datascienceweekly.org/data-science-resources>

# Where to go from here?

## Learn from Best Practices

49

### □ Read Industry Whitepapers

- It is important to understand how things are actually applied
- You'll find a lot of goodies in whitepapers

The screenshot shows the SAS website's navigation bar at the top, featuring links for Log In, Worldwide Sites, Contact Us, Chat Now, and a search bar. Below the navigation, a horizontal menu bar includes Products & Solutions, Industries (which is highlighted in orange), Support & Training, Customer Stories, Partners, Community, and About SAS. The main content area is titled 'Industries' and lists various sectors with corresponding links. Below this, there are two examples of how SAS has helped companies: a story about streamlining typhoon recovery efforts in the Philippines and another about improving freight logistics for Followmont Transport.

**Industries**

- › Automotive
- › Banking
- › Capital Markets
- › Casinos
- › Communications
- › Consumer Goods
- › Defense & Security
- › Government
- › Health Care Providers
- › Health Insurance
- › High-Tech Manufacturing
- › Higher Education
- › Hotels
- › Insurance
- › K-12 Education
- › Life Sciences
- › Manufacturing
- › Media
- › Oil & Gas
- › Retail
- › Small & Midsize Business
- › Sports
- › Travel & Transportation
- › Utilities

SAS streamlines response efforts during Philippines typhoon recovery.  
[Read the story](#)

Followmont Transport improves on-the-fly decisions – and optimizes freight logistics.  
[Read the story](#)

Imagine you're presenting to your firm's board about your AML program. What will happen if you don't have answers to all of their questions at your fingertips?  
[http://www.sas.com/en\\_us/insights/articles/risk-fraud- see-and-touch-data.html](http://www.sas.com/en_us/insights/articles/risk-fraud- see-and-touch-data.html)

#### SAS® Visual Analytics customers get a double shot of updates | SAS

New upgrades to SAS Visual Analytics provide more options, features and analytical methods.  
[http://www.sas.com/en\\_us/news/press-releases/2014/march/ sas-visual-analytics-version-8dot4-sgf14.html](http://www.sas.com/en_us/news/press-releases/2014/march/ sas-visual-analytics-version-8dot4-sgf14.html)

#### Bank of North Carolina speeds reporting, insights with SAS® Visual Analytics | SAS

SAS Visual Analytics provides superior report control and validation while remaining fast and easy for users.  
[http://www.sas.com/en\\_us/news/press-releases/2014/june/ bankofnc-visual-analytics.html](http://www.sas.com/en_us/news/press-releases/2014/june/ bankofnc-visual-analytics.html)

# Where to go from here?

## *Sharpen your analytics skills*

50

- Online Courses
  - Coursera – Machine Learning, Intro to Data Science
  - CMU – Machine Learning Summer School (free online)
- Blogs
- Data Science Bootcamps
- Kaggle
- Internship/Consulting
- Get Certified
  - Ryerson Certificate
  - Coursera Certificate
  - SAS Certificate in Predictive Modeling, Programming
  - Cloudera/Hortonworks

# Where to go from here?

## Preparing for Job Interviews

51

- Find an industry of interest
- Learn specific use cases
- Build your skills/portfolio (github)
- Don't get intimidated by the job requirements
- Find a way to talk to the hiring manager directly (LinkedIn)
  - Send her your github page
- Demonstrate your ability to learn new things
  - Most tools are easy to learn such as visualization tools, BI tools, etc.
- Remember that you're also interviewing the hiring manager

### Data Exploration and Visualization

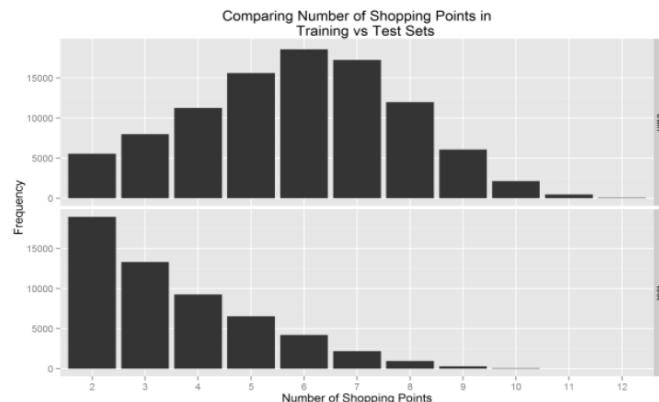
Here are some of my key findings from the exploratory process, and what I concluded from those findings.

#### 1. Missing values:

- risk\_factor was NA in 36.1% of the training set and 38.0% of the test set. As a predictor that I considered potentially useful, I decided to impute the risk\_factor for those customers using a linear regression model based on other customer characteristics.
- C\_previous and duration\_previous were NA for 2.8% of the training set and 4.9% of the test set. I decided that those NA values were probably indicative of new customers, and thus I imputed values of 0 for duration\_previous and "none" (a categorical variable) for C\_previous.
- location was NA for 0.3% of the test set. I decided to impute the location for each customer by copying the location from another customer in the same state.

#### 3. Number of shopping points:

- As seen in the plot below, the training set contained a roughly normal distribution of "shopping points" (the number of quotes a customer reviewed), whereas the test set contained a very different distribution.



<https://github.com/justmarkham/kaggle-allstate/blob/master/allstate-paper.md>

- ✓ Good presentation
- ✓ Documentation skills, knowledge sharing
- ✓ Learning, constantly improving

▼	By Application		
►	Call Centre Service		Today, 1:13 AM
►	Clinical.Bioinformatic		Dec 23, 2014, 4:38 PM
►	Credit Risk		Dec 23, 2014, 5:47 PM
►	Cross_Up_sell		Dec 23, 2014, 5:46 PM
►	Customer Churn & Retention		Dec 23, 2014, 5:50 PM
►	Customer LTV		Dec 23, 2014, 5:43 PM
►	Database Marketing		Dec 23, 2014, 4:52 PM
►	Forecasting		Dec 23, 2014, 4:42 PM
►	Fraud Detection		Dec 23, 2014, 4:48 PM
►	Insurance		Dec 23, 2014, 5:46 PM
►	Real-time		Dec 23, 2014, 4:36 PM
►	Retailing		Dec 23, 2014, 4:42 PM
►	STOCK		Dec 23, 2014, 4:46 PM
►	Survival Analysis		Dec 23, 2014, 5:48 PM
▼	Telecomm		
	Adaptive Customer Value Optimization for Communications Service Providers.pdf		Sep 11, 2007, 10:45 PM
	Advanced Marketing Automation for telecomm.pdf		Sep 11, 2007, 10:45 PM
	managing customer profitability in the telecomm industry.pdf		Sep 11, 2007, 10:45 PM
	MG104_Overcoming%20CRM%20Challenges.pdf		Sep 6, 2008, 9:34 PM
	MG105_How%20Cust%20Intelligence%20Improves%20CRM.pdf		Sep 6, 2008, 9:34 PM
	The_Role_of_Mobile_Phones_in_Sustainable_Rural_Poverty_Reduction_June_2008.pdf		Sep 12, 2008, 4:00 PM
►	Text Mining		
►	Uplift Modeling		
▼	By Techniques		
►	# Data Mining Best Practice		Yesterday, 9:19 PM
►	# In General		Yesterday, 9:19 PM
►	# SAS - Useful Macros		Today, 12:25 PM
►	1 - Data Preparation		Yesterday, 9:15 PM
►	2 - Feature Selection		Dec 23, 2014, 4:35 PM
►	3 - Algorithms		Dec 23, 2014, 4:38 PM
►	4 - Assessment & Ensemble		Yesterday, 6:01 PM
►	5 - Big Data		Yesterday, 6:00 PM
	Books		Jan 4, 2015, 12:54 AM
	References		Feb 22, 2015, 3:36 PM
►	5 - Model Deployment & Validation		Jan 4, 2015, 12:50 AM
			Yesterday, 5:59 PM

5 - Big Data					
Books					
Cascading	✓	Jan 4, 2015, 12:54 AM	--		
Cassandra	✓	Today, 12:51 PM	--		
Cloud	✓	Dec 23, 2014, 6:22 PM	--		
Coding Interview	✓	Jan 4, 2015, 12:46 AM	--		
Data Science	✓	Dec 23, 2014, 4:37 PM	--		
ElasticSearch	✓	Jan 11, 2015, 2:06 PM	--		
Graph Databases	✓	Jan 17, 2015, 11:03 PM	--		
h2oworld	✓	Jan 23, 2015, 1:23 PM	--		
Hadoop	✓	Feb 23, 2015, 4:08 PM	--		
Hadoop KaFaZhe	✓	Jan 4, 2015, 12:46 AM	--		
HBase	✓	Mar 3, 2015, 11:08 PM	--		
Information Retrieval	✓	Dec 23, 2014, 4:53 PM	--		
InternetOfThings	✓	Jan 16, 2015, 11:57 PM	--		
Linux Shell	✓	Jan 11, 2015, 2:06 PM	--		
Mahout	✓	Dec 23, 2014, 6:26 PM	--		
MapReduce Algorithms	✓	Jan 20, 2015, 7:46 PM	--		
NewSQL	✓	Jan 11, 2015, 2:05 PM	--		
NLP	✓	Jan 11, 2015, 2:07 PM	--		
Node.js	✓	Feb 22, 2015, 3:37 PM	--		
Pig/Hive	✓	Jan 18, 2015, 10:26 AM	--		
Programming	✓	Jan 4, 2015, 11:07 AM	--		
Python Data Analysis	✓	Feb 28, 2015, 2:13 PM	--		
R	✓	Feb 19, 2015, 10:37 PM	--		
Redis	✓	Feb 14, 2015, 10:53 AM	--		
REST API	✓	Jan 11, 2015, 2:05 PM	--		
Semantic Web	✓	Jan 2, 2015, 2:21 PM	--		
SocialWeb	✓	Jan 11, 2015, 2:03 PM	--		
Solr	✓	Jan 11, 2015, 2:05 PM	--		
Spark	✓	Feb 14, 2015, 11:42 PM	--		
Storm	✓	Jan 4, 2015, 12:46 AM	--		
Wiley.Data.Smart.Nov.2013.pdf	✓	Jan 4, 2015, 12:46 AM	73.5 MB		
References					
Hadoop	✓	Jan 9, 2015, 2:28 PM			
In-Database	✓	Jan 4, 2015, 12:50 AM			
MapReduce	✓	Dec 23, 2014, 5:46 PM			
Calculating the Jaccard Similarity Coefficient with Map Reduce for Entity Pairs in Wikipedia.pdf	✓	Dec 23, 2014, 4:53 PM			
Implementing best practices for fraud detection on an online advertising platform.pdf	✓	Dec 23, 2014, 5:50 PM			
Map-Reduce for Machine Learning on Multicore.pdf	✓	Jul 6, 2011, 6:46 PM	265 KB		
	✓	Jul 6, 2011, 6:42 PM	1.1 MB		
	✓	Dec 30, 2010, 3:22 PM	477 KB		

# Stay In Touch

54

- My Email [shaohua.zhang@live.com](mailto:shaohua.zhang@live.com)
- My Blog <http://www.dreamcre8or.com>
  - working on my new blog
- LinkedIn <http://ca.linkedin.com/in/shaohuazhang/en>
- LinkedIn Group (Ryerson – Certificate in Big Data)  
<https://www.linkedin.com/grp/home?gid=8258220>
  - Thanks to Ali Kazim ☺