About Me

- B.Tech – M.Tech – IIT Kanpur (Electrical Engineering)
- Thesis on Automatic Image Annotation of Large Image Databases
- Opera Solutions – Analytics Specialist
- Info Edge – Senior Data Scientist / Senior Manager
- Tickled Media – VP of Data Science
- HTMedia – Head of Analytics
- Lendingkart - VP of Analytics
- Mudracircle – Cofounder and CTO
- Analytics Vidhya – Head of Engineering

Executive Summary

After this workshop you should have a decent understanding of what goes in developing Recommendation Systems

- **Module 1 - Introduction to Recommendation Systems**
  - Examples of Real Life Recommendation Systems
  - What is Scalability?
  - What is an awesome recommendation system?
  - How to start designing recommendation systems?
  - The Machine Learning Pipeline
- **Module 2 - Association rule mining and market basket analysis**
  - Case Study: Optimizing product placement in Retail to increase sales
  - Association Rule Mining and Market Basket Analysis
  - The Apriori Algorithm
  - Hands on – Market Basket Analysis in Practice
  - Assignment – Identify best product bundles

Executive Summary

After this workshop you should have a decent understanding of what goes in developing Recommendation Systems

- **Module 3 - Collaborative Filtering**
  - Types of Recommendation Systems
  - Collaborative filtering
    - User-user
    - Item-item
  - Hands on: Building a movie rating recommendation system
  - Assignment: Building a Song Recommendation System
  - Cold Start Problem

- **Module 4 - Visualizing and improving Recommendation Engines**
  - Content based Filtering
  - Hands on: Creating a Simple Text Matching System
  - Understanding Singular Value Decomposition (SVD)
  - Hands on: Building a Semantic Search system
  - Assignment : Movie rating recommendation using content based filtering

Executive Summary

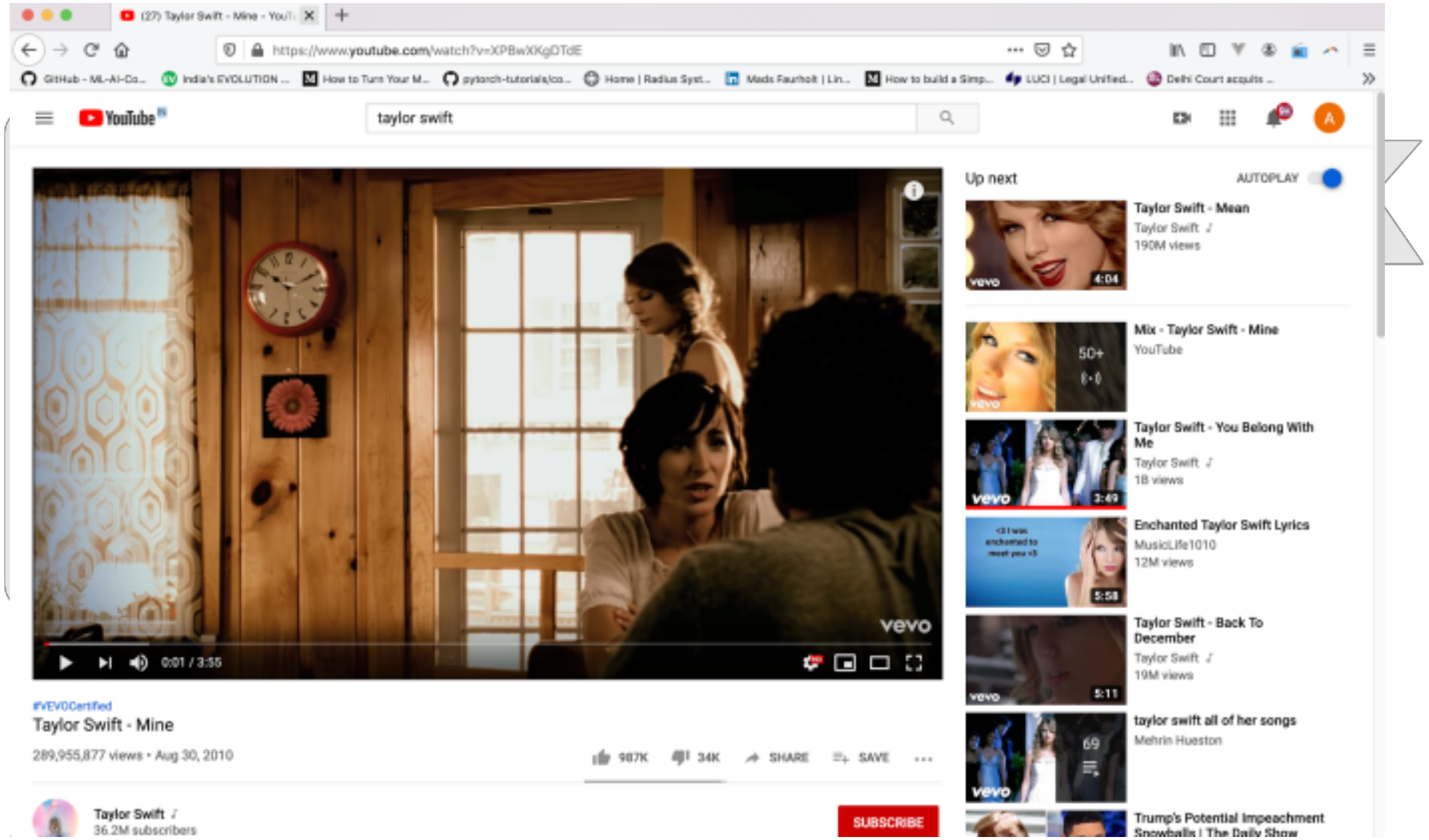
After this workshop you should have a decent understanding of what goes in developing Recommendation Systems

- **Module 5- End to end job recommendation system**
  - Case Study: Building a job recommendation system for a jobs portal
  - When to use Collaborative Filtering vs Content based Filtering
  - Evaluation Metrics:
    - Theoretical: Accuracy, Precision, Recall, ROC
    - Practical Evaluation Metrics – Business Metrics, A/B Testing, Google Analytics
  - Assignment: Video recommender system
  - Assignment: Food discovery similar to Uber Eats
- **Module 6- Deploying your own Recommendation system**
  - Building APIs and scaling them up
  - Deployment of these systems

Introduction to Recommendation Systems

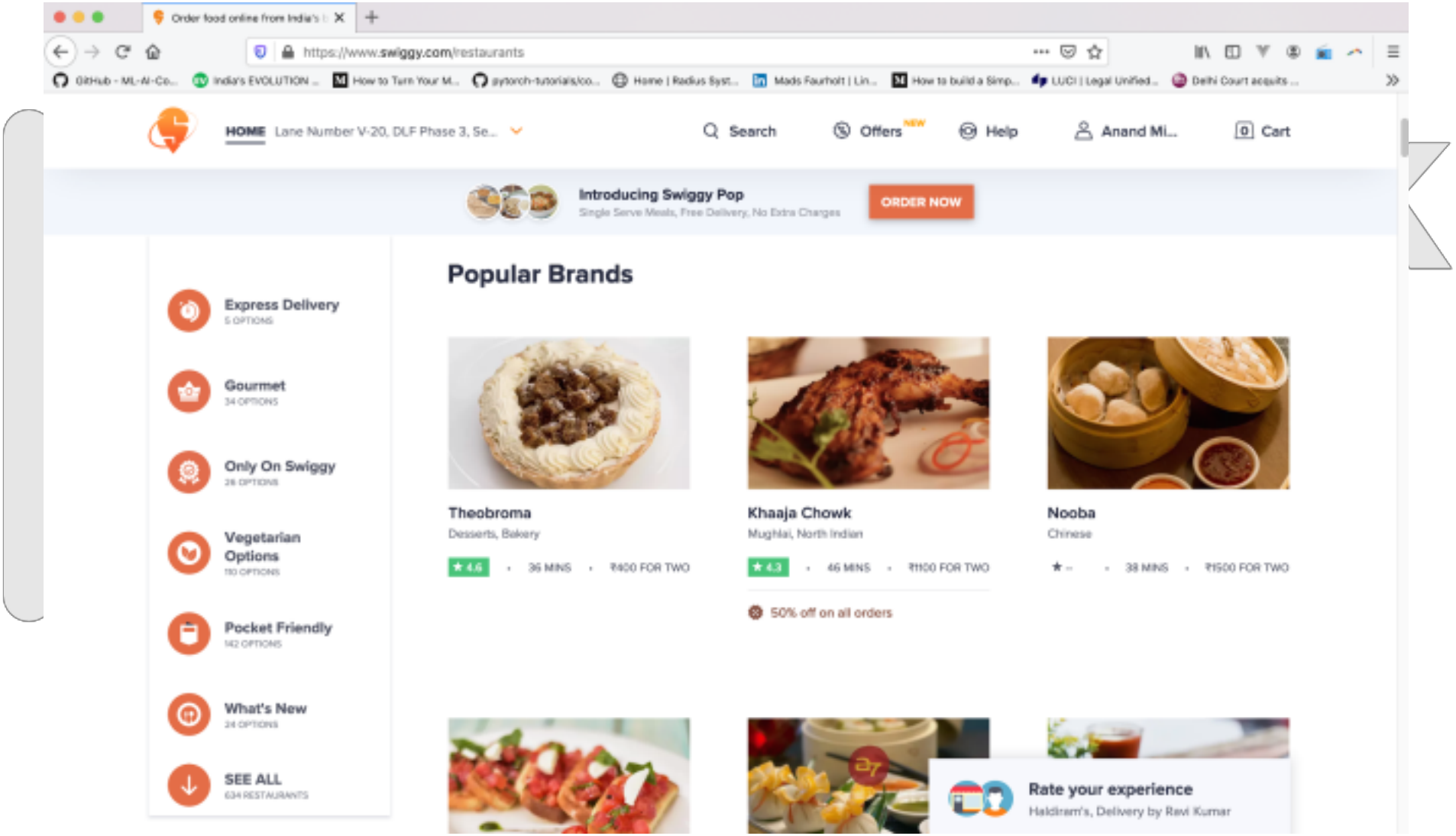
Examples of Real Life Recommendation Systems

Finding Related videos on Youtube



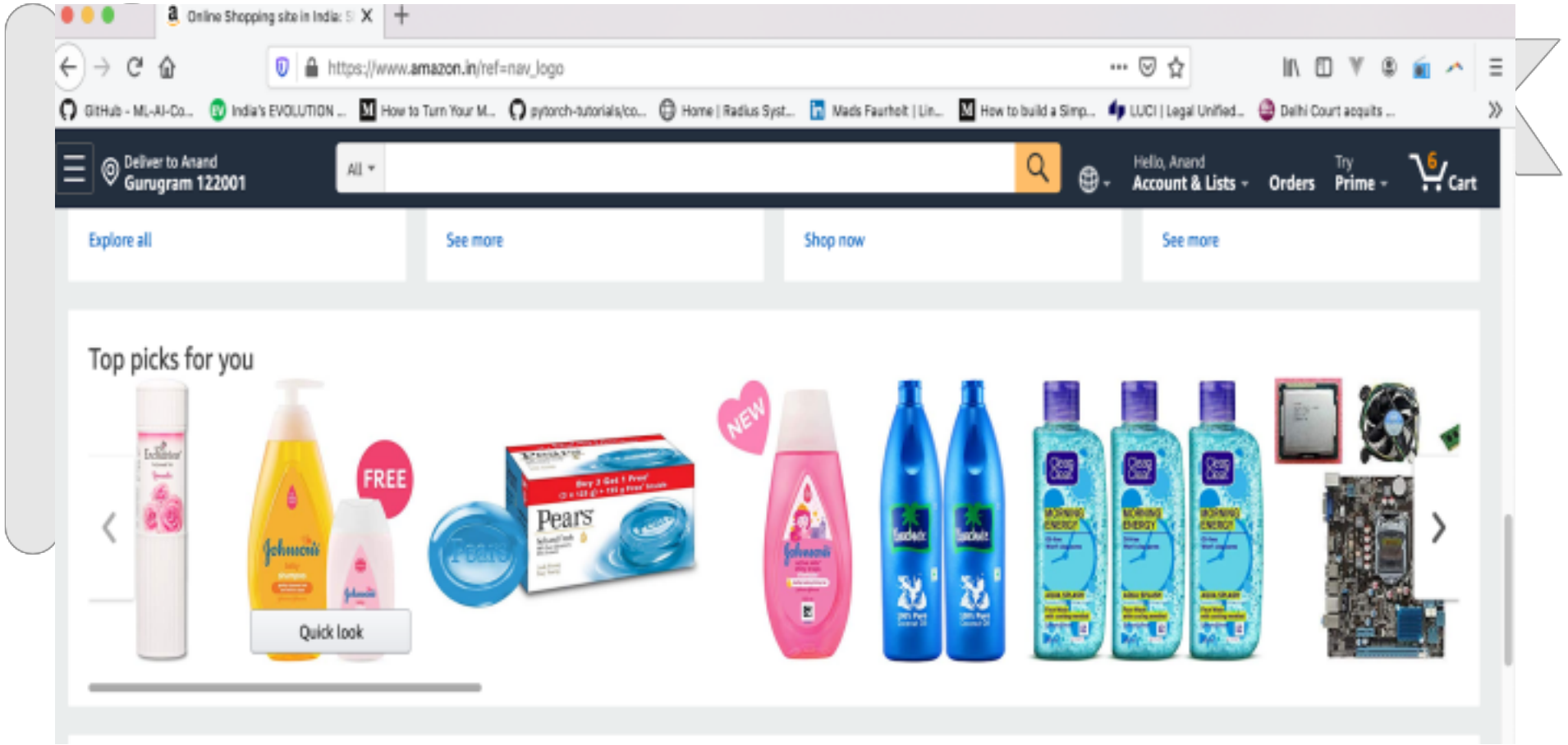
Examples of Real Life Recommendation Systems

Get your favorite dish on Swiggy



Examples of Real Life Recommendation Systems

Personalized product recommendations on Amazon



Examples of Real Life Recommendation Systems

Finding your partner on Tinder

Analytics Vidhya
Learn everything about analytics

To Create Awesome Recommendation Systems

What is an Awesome Recommender System?

Analytics Vidhya
Learn everything about analytics

Features of an awesome recommender system

- Show the relevant results
  - Show programming titles to a software engineer and baby toys to a new mother
- Don't recommend items user already knows or would find anyway
- Expand user's taste without offending or annoying him/her

Challenges in creating an awesome recommender system

- Huge amounts of data, tens of millions of customers and millions of distinct catalog items
- Results may be required to be returned in real time
- Have limited information about new customers
- Old customers can have a glut of useless information

What is Scalability?

Analytics Vidhya
Learn everything about analytics

Scalability

What is Scalability?

- 100 visitors -> 1M visitors -> 100M visitors -> Billion visitors
- Ability to handle increasing loads just by adding resources
- Vertical Scaling
- Horizontal Scaling
- Real time or other constraints should be fulfilled

Analytics Vidhya
Learn everything about analytics

How do we start designing such Recommendation Systems?

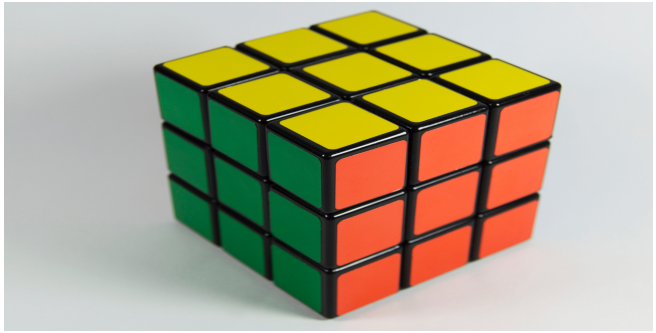
Starting designing Recommendation Systems

Analytics Vidhya
Learn everything about analytics

Know your Goal First

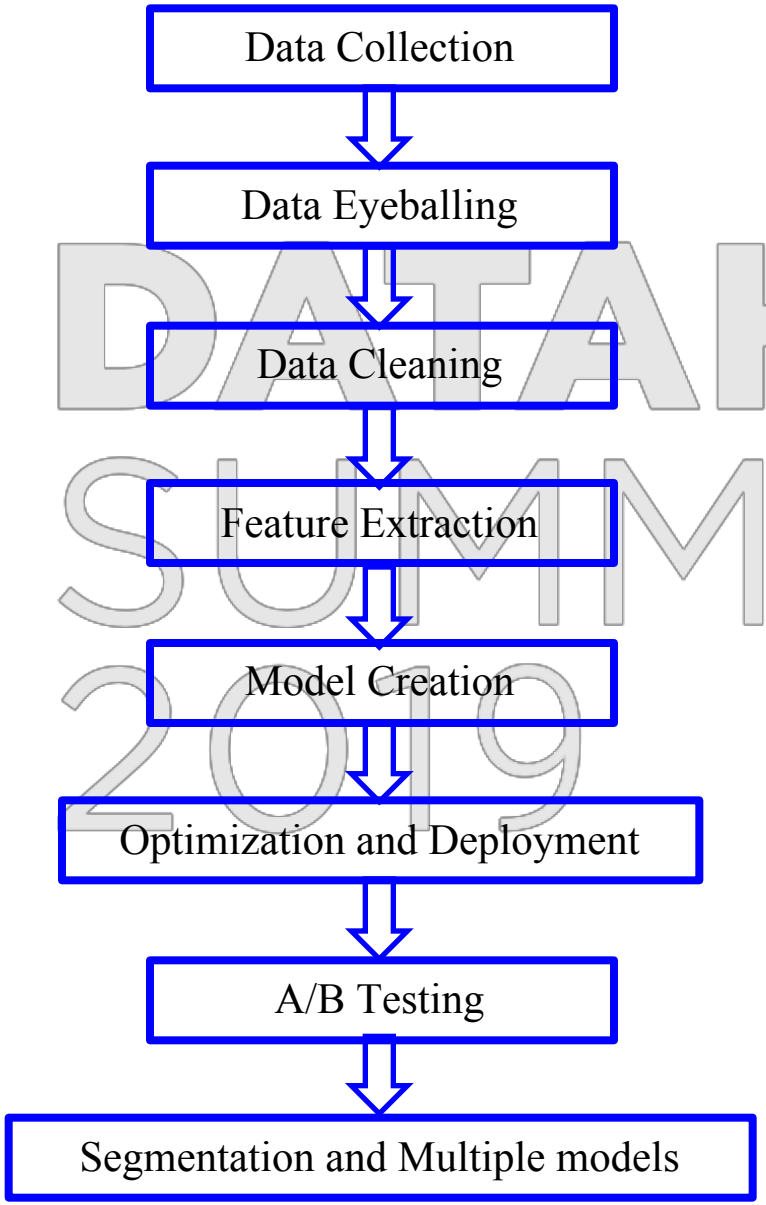
Baselining the Problem

Has this problem being solved before?



Talk to the Stakeholders

The Machine Learning Pipeline



CK

Association Rule Mining and Market Basket Analysis

Case Study: Optimizing Product Placement in Retail

Analytics Vidhya
Learn everything about analytics

Optimizing Product Placement in Retail

- How to place products to increase sales?
- Bundling
- Cross Sell

Market Basket Analysis - Concept

Analytics Vidhya
Learn everything about analytics

Optimizing Product Placement in Retail

- Analyze baskets of users
- See patterns of two items being bought together

Analytics Vidhya
Learn everything about analytics

How to analyze the basket data to get insights about product placement and cross sell?

Transactional Data of our store

ID	Items Bought
1	Bread, Cheese, Egg, Juice
2	Bread, Cheese, Juice
3	Bread, Milk, Yogurt
4	Bread, Juice, Milk
5	Cheese, Juice, Milk

The Apriori Algorithm

Transactional Data of our store

ID	Items Bought
1	Bread, Cheese, Egg, Juice
2	Bread, Cheese, Juice
3	Bread, Milk, Yogurt
4	Bread, Juice, Milk
5	Cheese, Juice, Milk

Frequent Item Sets (of 1 item)

Items	Frequency
Bread	4
Cheese	3
Egg	1
Juice	4
Milk	3
Yogurt	1

The Apriori Algorithm

- Need to decrease the number of combinations
- Idea: Remove less frequent items
- Support = num_items/ num_transactions
- Min Support = 50%

Remove these as the support is < 50%

1 – item Candidate Sets

Items	Freq	Support
Bread	4	4/5 = 80%
Cheese	3	3/5 = 60%
Egg	1	1/5 = 20%
Juice	4	4/5 = 80%
Milk	3	3/5 = 60%
Yogurt	1	1/5 = 20%

The Apriori Algorithm

Candidate sets selected in last iteration

Bread, Cheese, Juice Milk

Transactional Data of our store

ID	Items Bought
1	Bread, Cheese, Egg, Juice
2	Bread, Cheese, Juice
3	Bread, Milk, Yogurt
4	Bread, Juice, Milk
5	Cheese, Juice, Milk

2 Item Candidate Set (Pairs)

Item Pairs	Frequency	Support
(Bread, Cheese)	2	2/5 = 40%
(Bread, Juice)	3	3/5 = 60%
(Bread, Milk)	2	2/5 = 40%
(Cheese, Juice)	3	3/5 = 60%
(Cheese, Milk)	1	1/5 = 20%
(Juice, Milk)	2	2/5 = 40%

The Apriori Algorithm

- Selected 2 – item candidate sets
- (Bread, Juice)
- (Cheese, Juice)

Remove these as the support is < 50%

2 – item Candidate Sets

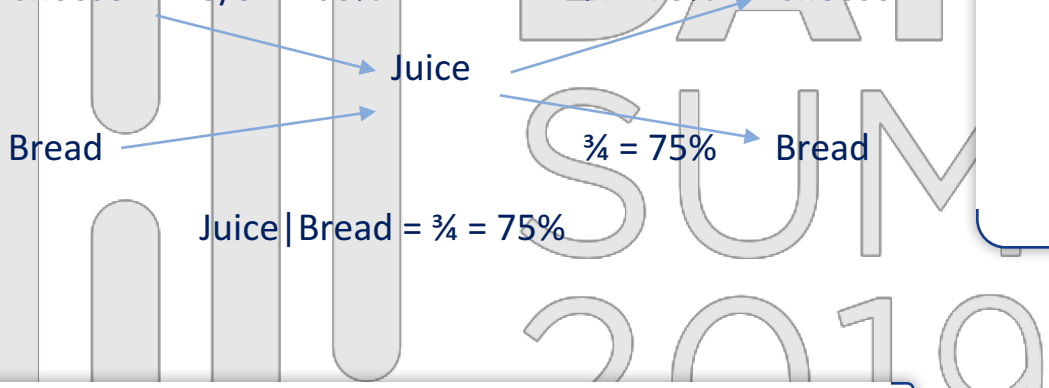
Item Pairs	Frequency	Support
(Bread, Cheese)	2	2/5 = 40%
(Bread, Juice)	3	3/5 = 60%
(Bread, Milk)	2	2/5 = 40%
(Cheese, Juice)	3	3/5 = 60%
(Cheese, Milk)	1	1/5 = 20%
(Juice, Milk)	2	2/5 = 40%

Rules from Apriori

Rules

(Bread, Juice), (Cheese, Juice)

Can be Bread -> Juice or Juice -> Bread



Juice|Bread = ¾ = 75%

Confidence (A->B) = Support(AUB)/Support(A)

Minimum Confidence = 75%

Final Rules

All 4 Rules are good here

Cheese or Bread -> Juice

Juice -> Cheese and Bread

2 – item Candidate Sets

Item Pairs	Frequency	Support
(Bread, Juice)	3	3/5 = 60%
(Cheese, Juice)	3	3/5 = 60%

Frequent Item Sets (of 1 item)

Items	Frequency
Bread	4
Cheese	3
Juice	4
Milk	3

Hands on – Market Basket Analysis in practice

Assignment – Identify best product bundles

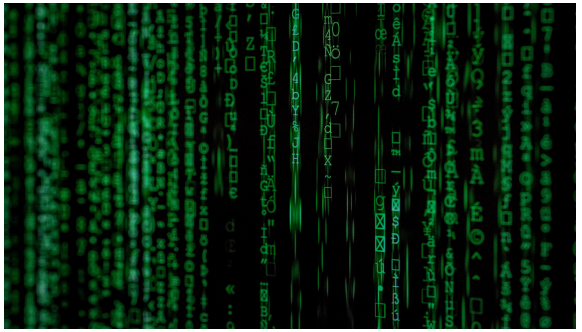
Meal

Collaborative Filtering

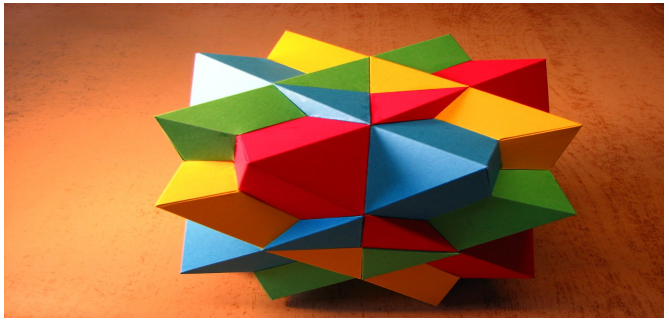
Types of Recommendation Systems

Collaborative Filtering based

Content Based



Hybrid



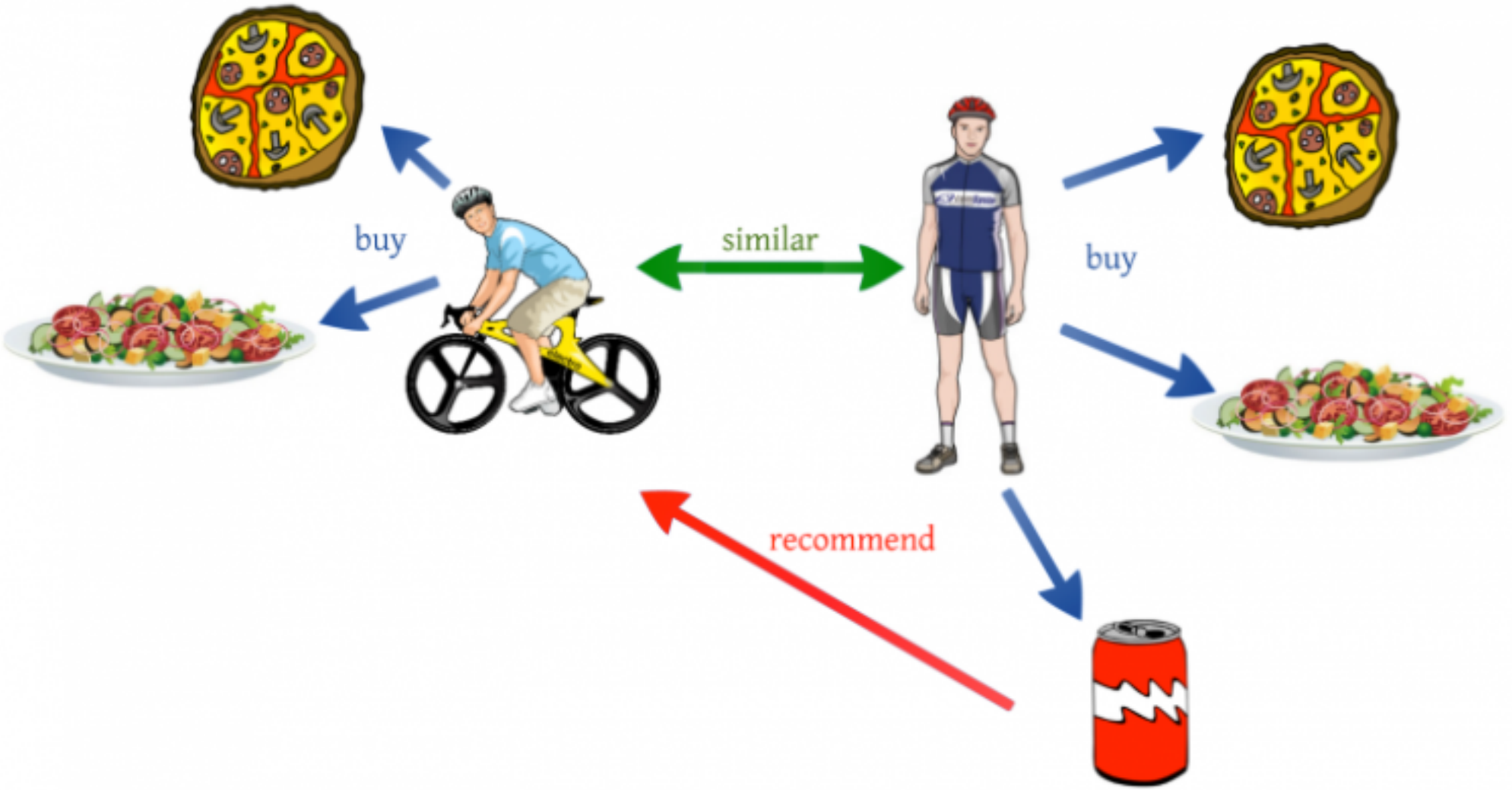
Types of Collaborative Filtering

Analytics Vidhya
Learn everything about analytics

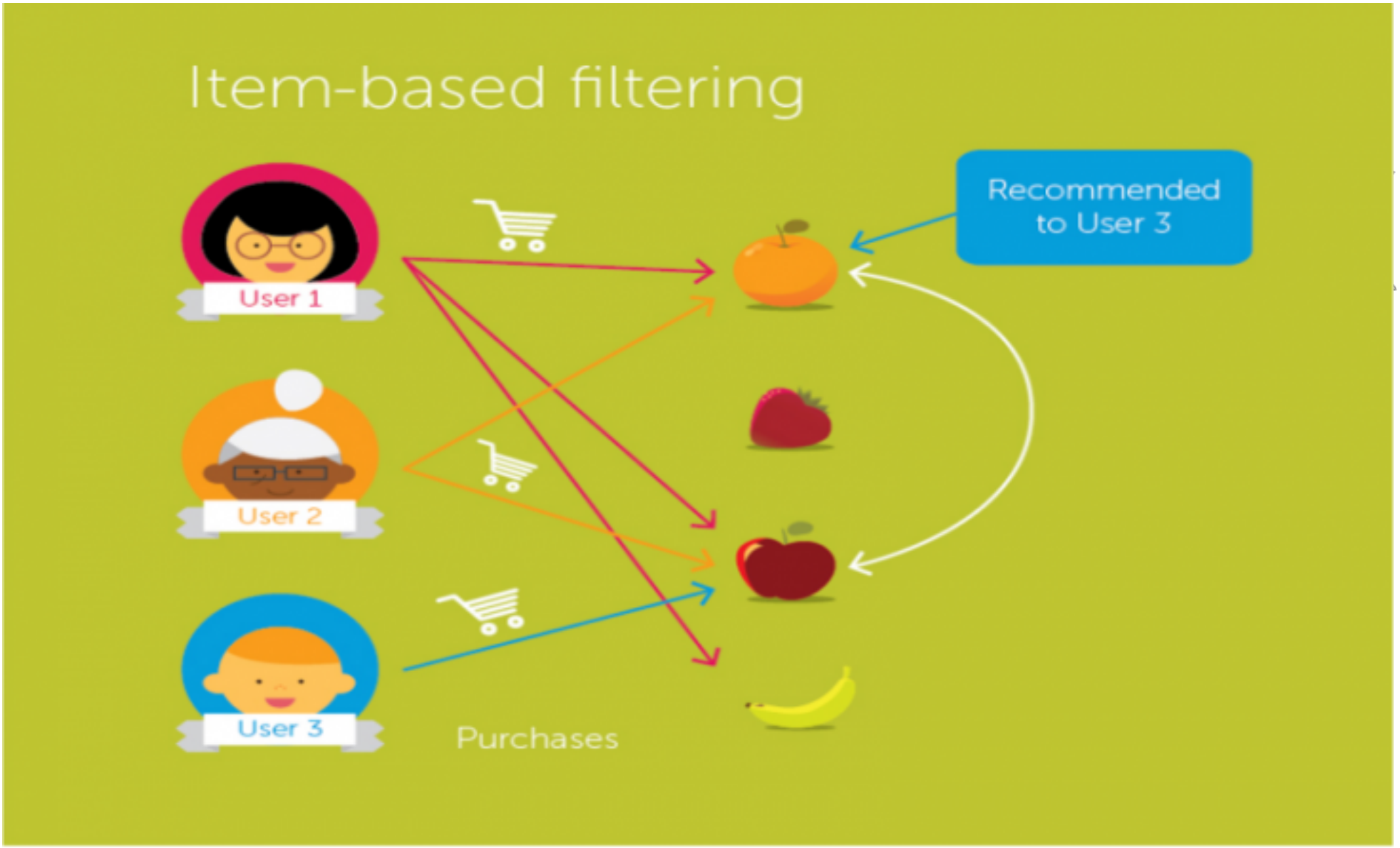
User - User

Item - Item

User – User Collaborative Filtering



Item - Item Collaborative Filtering



Identify the users and items for the following cases:

Cases	Users	Items
LinkedIn		
Facebook		
Netflix		
Amazon		

Identify the users and items for the following cases:

Cases	Users	Items
LinkedIn	Members	Jobs, Members
Facebook	Members	Members, Groups
Netflix	Members	Movies, TV shows
Amazon	Members	Products

Hands on – Building a movie rating recommendation system

Assignment – Building a Song Recommendation System

The Cold Start Problem

Analytics Vidhya
Learn everything about analytics

Problems with Collaborative Filtering

Cold Start with users

Cold Start with items

What to do now?

- Content based approaches

ACK

T

Visualizing and Improving Recommendation Systems

Short Overview: Content based Recommendation Systems

Analytics Vidhya
Learn everything about analytics

Uses the content of the problem

- What do we mean by content?
- Attributes of the user and items

Example – Movie Recommendation System

User
- Gender
- Age
- Hobbies
- Occupation

Movie
- Genre
- Director
- Cast
- Rating
- Duration

How do we blend all these factors together?

- Create a model?
- We will look into how to create these models later

Short Overview - Hybrid approaches

- Mixture of Collaborative Filtering and Content based approaches
- Best of both worlds
  - Having enough behavioral data - use Collaborative Filtering
  - Else - use Content based
  - Mix both the approaches - detail - later - Case Study on Job Recommendation System

Hands on – Simple Text Matching System

Problem

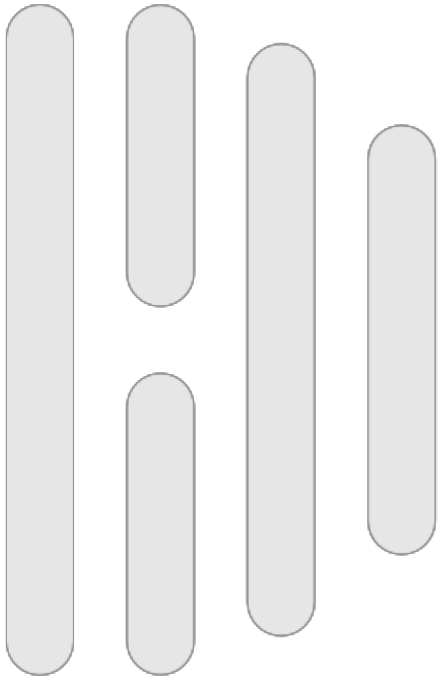
Text1 = ”Data Science is great”

Text2 = “I love data science”

Problem: Find Similarity between these two texts
Example of a Content based system

Approaches?

- Number of matching words
- Issue: Scaling
- Next: Geometrical Interpretation of this problem – Very Important



Geometrical Interpretation

- Board

Really Important part

- (Vector?, Matrix?, Rotation?)
- Matrix 2 ways of viewing geometrically
- Seeing collaborative filtering
  - Geometrically
- Demo - Simple Text Matching

Now

Ok if you don't understand next 15-20 mins

Improving Simple Text Matching System: TFIDF

TFIDF decrease weightage of very common words in data and highlights important words

- Some words will be appearing everywhere - I, am , the
- Useless words for matching
- Remove them - how?
  - NTLK English stopwords
  - Dont want to update the vocab again and again
  - Depending on context this is not what I need. Ex - Amit Bacchan
- What to do now?
  - Suppress words with high freqency in the entire dataset
  - Highlight words appearing in only specific documents
  - tf * idf
  - Many ways - Wikipedia
  - Generally idf - log - to reduce the scale

Issues with Cosine based Similarity

- Lied that dimensions are independent
- In fact, in a real setting dimensions are very much dependent
- So the axis are not perpendicular
- But anyways the angle wont change even if the axis are slanted
  - So why can't I use cosine?
  - The angle wont change, but the formula to calculate cosine in slanted coordinate system certainly will
  - Equation of cosine in slanted coordinate system - complicated - uses Tensors
- Why don't I use this equation?

Why not use Slanted Axis Cosine?

- Why don't we use the slanted axis cosine?
  - Equation different for different pairs
  - Difficult to understand and implement
  - Doesn't give a good picture of what's happening
- Is there a better way?
  - Can I change the axis/ basis to some other basis which are orthogonal (perpendicular)?
- But let's first see how much are these variable correlated to each other?

Correlation Matrix and Diagonalization

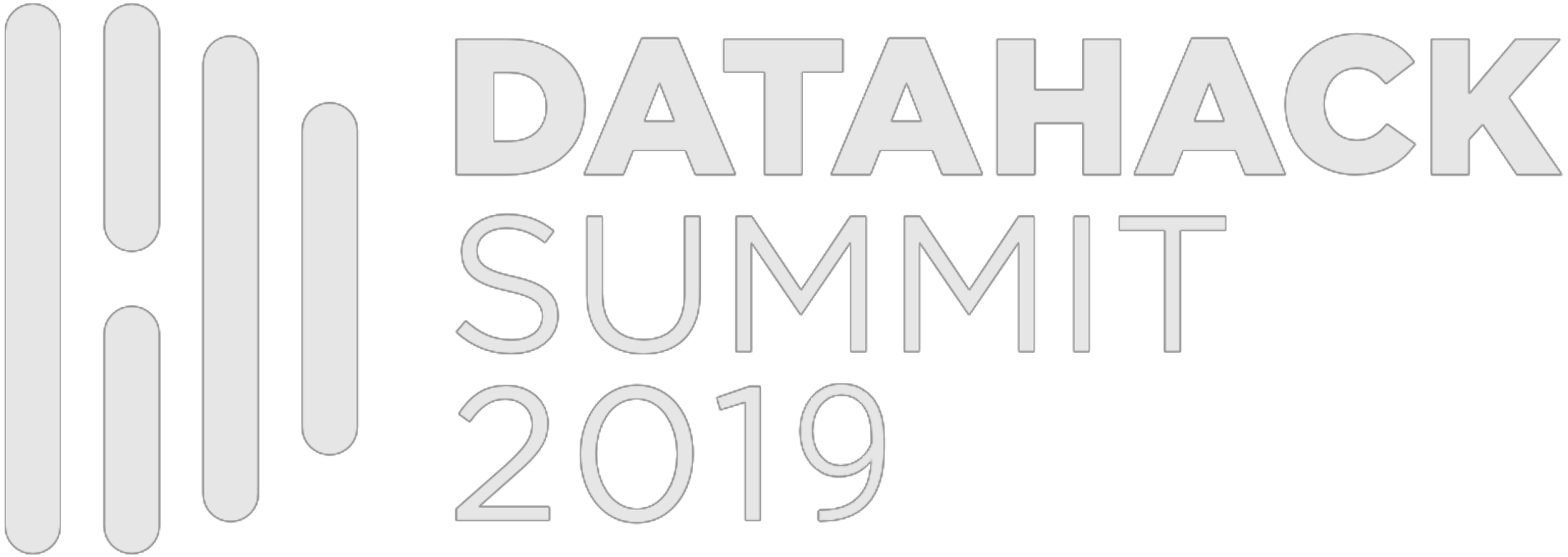
- Lets just create a correlation matrix for it
- A^TA (A is normalized so we get correlations)Now the off diagonal elements represent the correlations of different variables
  - Ideally I want them to be zero
  - But they are non zero
- Is there a way to make these non diagonal elements zero?
- In other words, is there a way to diagonalize this matrix

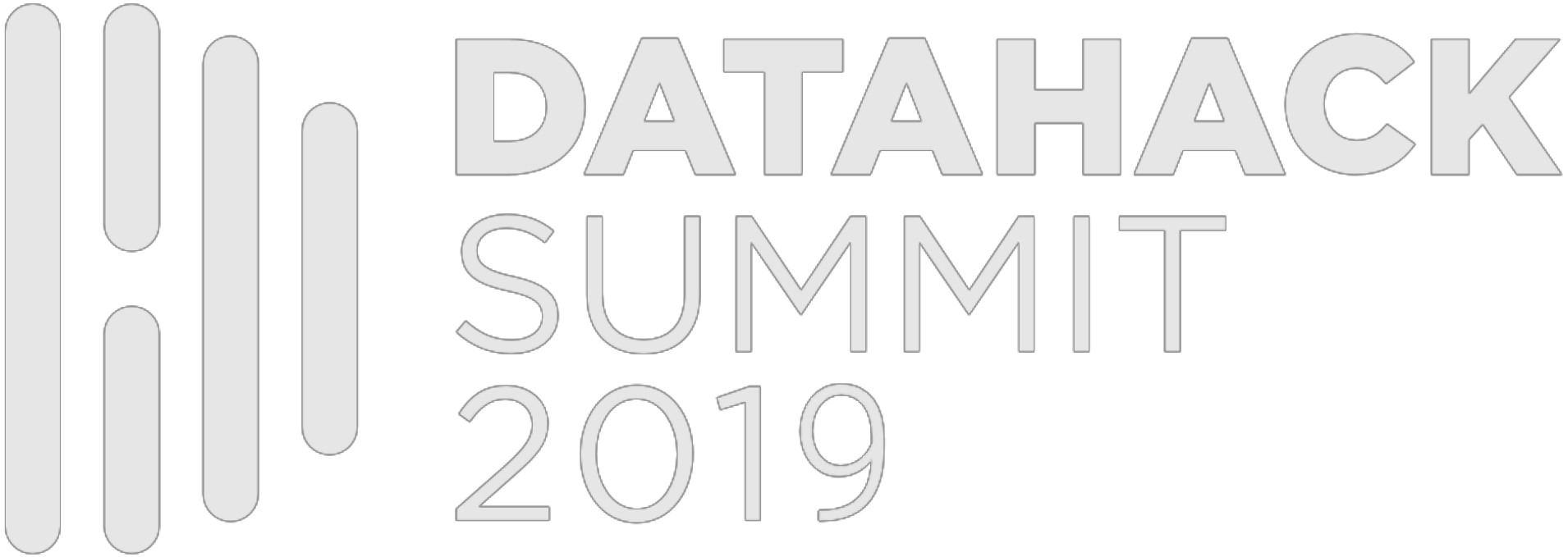
Matrix Diagonalization

- We now turn to math to know the techniques for matrix diagonalization
- Many techniques available
  - SVD
  - PCA
  - NNMF
  - And more
- Most used in the industry is SVD / PCA
- Little difference between the two

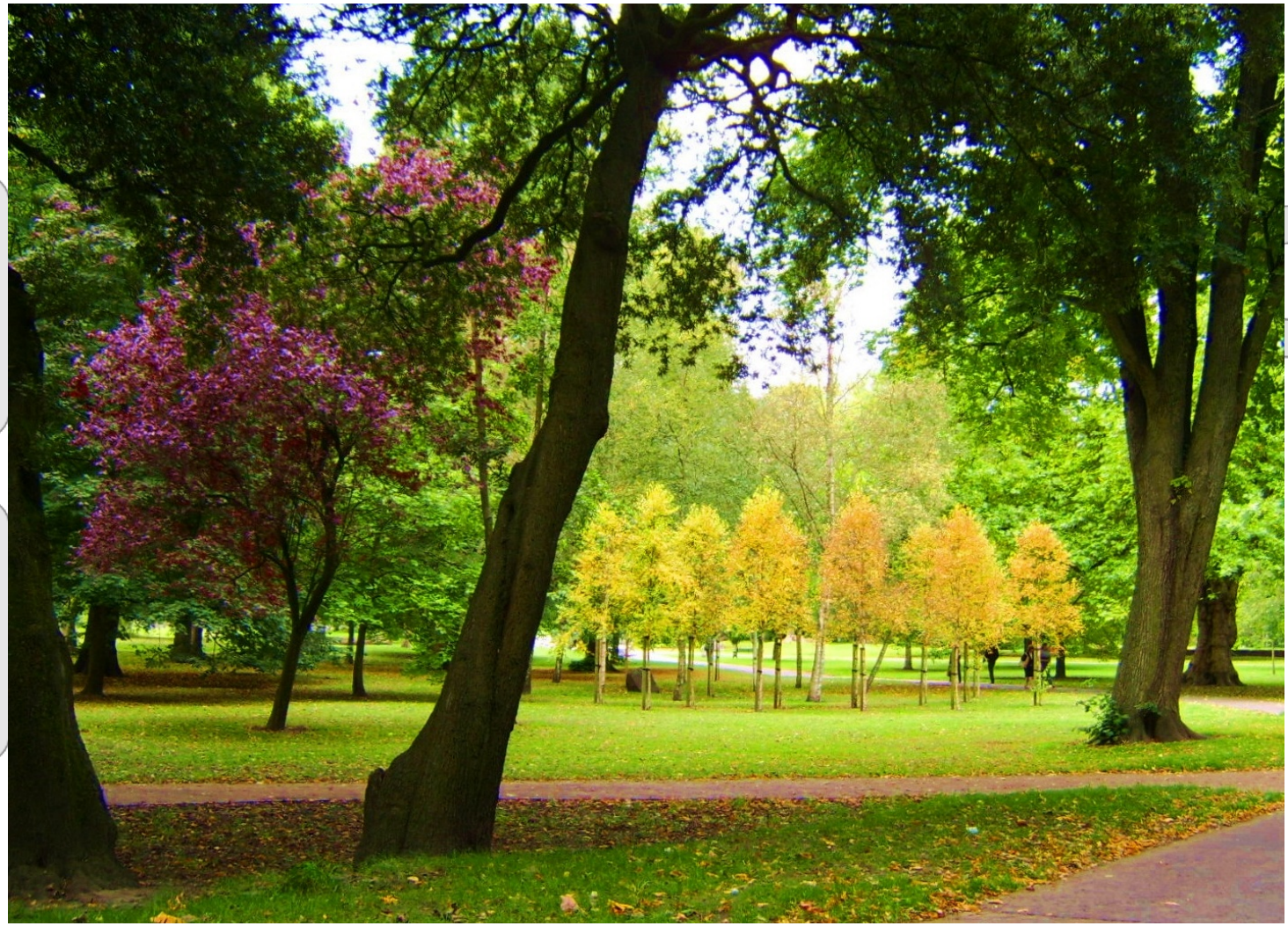
How to diagonalize a Matrix?

- Just an intuition
- Sloppy notation
- Eigenvalues and Eigenvectors – Meaning
- SVD
- Take the principal components


DATAHACK
SUMMIT
2019


DATAHACK
SUMMIT
2019

Break



End to End Job Recommendation System

Analytics Vidhya
Learn everything about analytics

Job Recommendation System: Problem Definition

- We are running a Job Portal
- There are users on the portal who apply to jobs (Add a photo)
- As a data scientist you need to figure out how can you improve the job recommendation system

Recommend the best matching jobs to the user

- **On email (every second day, 10 jobs matching your profile)**
- **On website (real time)**

Job Recommendation: Data Dictionary

Analytics Vidhya
Learn everything about analytics

Users

- Ctc
- Experience
- Skills
- Name
- Resume

Jobs

- Title
- Skills
- JD
- CTC Range
- Experience Range

Applies

- User
- Job
- Timestamp

Job Recommendation: Deciding the Goal

Analytics Vidhya
Learn everything about analytics

Goal of our Recommendation System

Business Metrics

- Probability of Placement
- Placement
  - Apply
  - Shortlist
  - Interview
  - Offer & Negotiation
- **Applies**
- Unique Applicants?

Analytics Vidhya
Learn everything about analytics

Goal: Increase Applies

Job Recommendation: Which Features to consider?

- What is the Goal?
  - Increase the number of applies
- Which things/ features to consider which recommending jobs to a user?
  - Talk to the Stakeholders
  - Go into the shoes of the customer

- But how does the user apply in the first place?
  - Online real time case – simple
  - In email – We send and email, user opens…..a bit complicated
- Lets see it in detail

The User Journey – Email Recommendations

- Mail Sent -> Open -> link click on email -> apply

Mails Sent	Open	Link Click in mail	Apply
10000	2000	200	10

Ratios

- open/mail sent = 20% - subject
- click/open = 10% - Recommendation
- apply/click = 5% - Product, Job,..

apply/mail sent = product of these 3 ratios

Content based Recommendation Systems

- Goal: Increase the number of applies
- Basic Structure
  - Feature -> Model -> Recommender System
  - But how to decide the features?
- And how to decide the model?
- How to really think about it?

- If you think about it, why would someone apply to a job
  - Ctc
  - Exp
  - Profile matching
  - Good company

Content based Recommendation Systems

- Next step is to convert these concepts to actual features
- How do we capture city?
- Goal : Applies
- One idea: based on proximity
  - May not work in practice
  - People in metros generally apply to jobs of metros
- City City movement propensity

Content based Recommendation Systems

- Similarly with Designation?
- Issues in designation
  - Not standardized
  - Abbreviations SSE (can be Senior Software Dev as well as Senior Sales Executive)

Content based Recommendation Systems

- How do we match User profile and Job?
  - Issue of unclean data
    - English stopwords, many useless words
    - Don't want them to match
  - How to fix this problem?
    - Remove stopwords
    - Is there a better way?
      - TFIDF?
  - How should we do the text matching?
    - User (Designation, skills, Resume) – Job (title, skills, JD)
    - Weighted matching
    - Then maybe Dimensionality reduction to improve matching
    - TFIDF + SVD = LSI

Merging many scores

- Multiple such scores
  - lsicosine
  - cityScore
  - designationScore
  - ……..
- Come up with a single score
- Sort and send the best matching jobs to the user
- How to merge these scores to a single score?
- Models?

Modeling

- Create a model to know how to blend all features to create final matching score?
- What is the target variable?
  - What am I trying to optimize
  - Applies?
- Which algorithm to use?
- Scaling up the training process (Caching the model)

Evaluation Metrics

- Practical metrics
- Business Metrics
  - Applies
  - Unique applicants
- A/B Testing
- Google Analytics
- Between theoretical and practical - Simulations

Evaluation Metrics

- How do I know if my model is performing well or not?
- Theoretical metrics
  - RMSE
  - Accuracy
  - Precision
  - Recall
- Can we get a more close to real life metric?

Case Study: Video recommender system similar to YouTube

Problem

- You are the head of data science of a company that serves videos to users, similar to Youtube.
- Create a recommendation system so that once a user views a video, he/ she will be shown similar videos

Assignment - Food Discovery System

- Problem Statement
- You are the Head of Data Science of a leading company in Food E-Commerce Business. Something like UberEats / Zomato
- How do you come up with a personalized recommendation engine for users so that you can increase sales on your platform
- Think about what approach will you use
- What features would you use
- Team A - odd rows
- Team B - even rows

BREAK

Deploying your Recommendation Systems

Making things Real time?

- Want to make things real time?
- Why?
- The challenges in making things real time from batch processing?
- The BIG IDEA: Caching, don't compute same thing again and again
- Some Tech concepts needed
  - How web servers work
  - APIs
  - Types of web requests
  - Caching

Search as a Recommendation System

- The Problem - Create a semantic search engine on news articles
- But how does search tie up to recommendation systems
- To understand this we need to understand how search works
  - Reverse Index
  - Board
- Speed and Relevance tradeoff

Building APIs and Scaling them up

Analytics Vidhya
Learn everything about analytics

- APIs - What are they?
- Why do we need APIs
- Making our recommendation system live on the web
- But how does internet works?
- Client Server Architecture (SQL Slide add)
- APIs - small micro services

- Cloud Providers - AWS, Google
- Scaling - Horizontal, Vertical
- Load Testing - Code
- Distributed Computing - multiple servers
- Paas vs Iaas
- DNS and Routing
- Final Picture

Thank you slide