Introduction

1. DM vs. KDD

Data mining (knowledge discovery from data)

Knowledge discovery (mining) in databases (KDD)

2. Steps of KDD; iterative in nature; results need to be validated.

A KDD process includes data cleaning, data integration, data

selection, transformation, data mining, pattern evaluation, and

knowledge presentation

Learning the application domain

relevant prior knowledge and goals of application

Creating a target data set: data selection

Data cleaning and preprocessing: (may take 60% of effort!)

Data reduction and transformation

Find useful features, dimensionality/variable reduction, invariant

representation.

Choosing functions of data mining

summarization, classification, regression, association, clustering.

Choosing the mining algorithm(s)

Data mining: search for patterns of interest

Pattern evaluation and knowledge presentation

visualization, transformation, removing redundant patterns, etc.

Use of discovered knowledge

3. Database (efficiency) vs. Machine learning (effectiveness) vs. Statistics (validity):

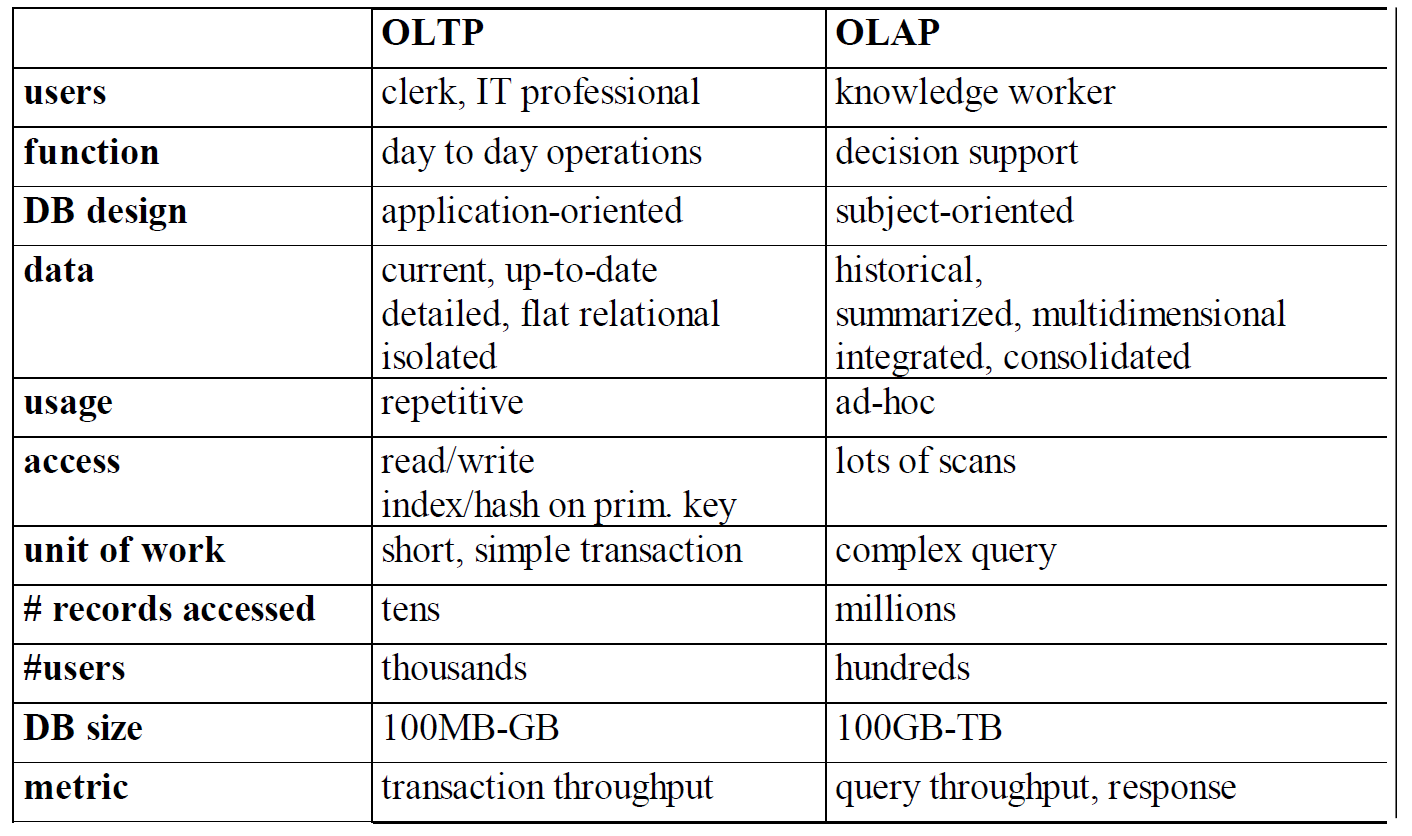
4. Able to cast a real problem into a data mining problem.

Data Warehousing and OLAP

I Understand the four characteristics of DW (DW vs. Data Mart)

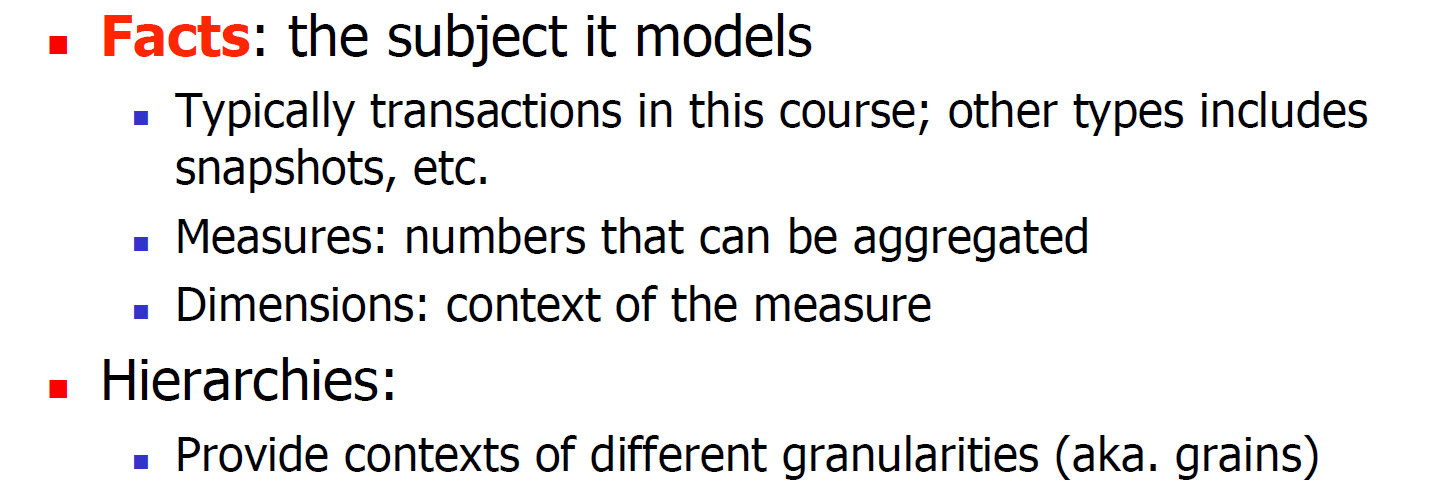
subject-oriented, integrated, time-variant, and nonvolatile collection of data

I Differences between OLTP and OLAP



I Multidimensional data model; data cube;

I fact, dimension, measure, hierarchies



I cuboid, cube lattice

I three types of schemas

Star(A fact table in the middle connected to a set of dimension tables), snow（A refinement of star schema where some dimensional hierarchy is normalized into a

set of smaller dimension tables）, face constellation.（Multiple fact tables share dimension tables）

I four typical OLAP operations

**Roll-up**

**Drill-down**

Slice and Dice

Pivoting

I ROLAP/MOLAP/HOLAP

Using relational DB technology: ROLAP

Using multidimensional technology: MOLAP

Molap:

(Sparse) array-based multidimensional storage engine

I Query processing methods for OLAP servers, including the BUC cubing algorithm.

Data Preprocessing

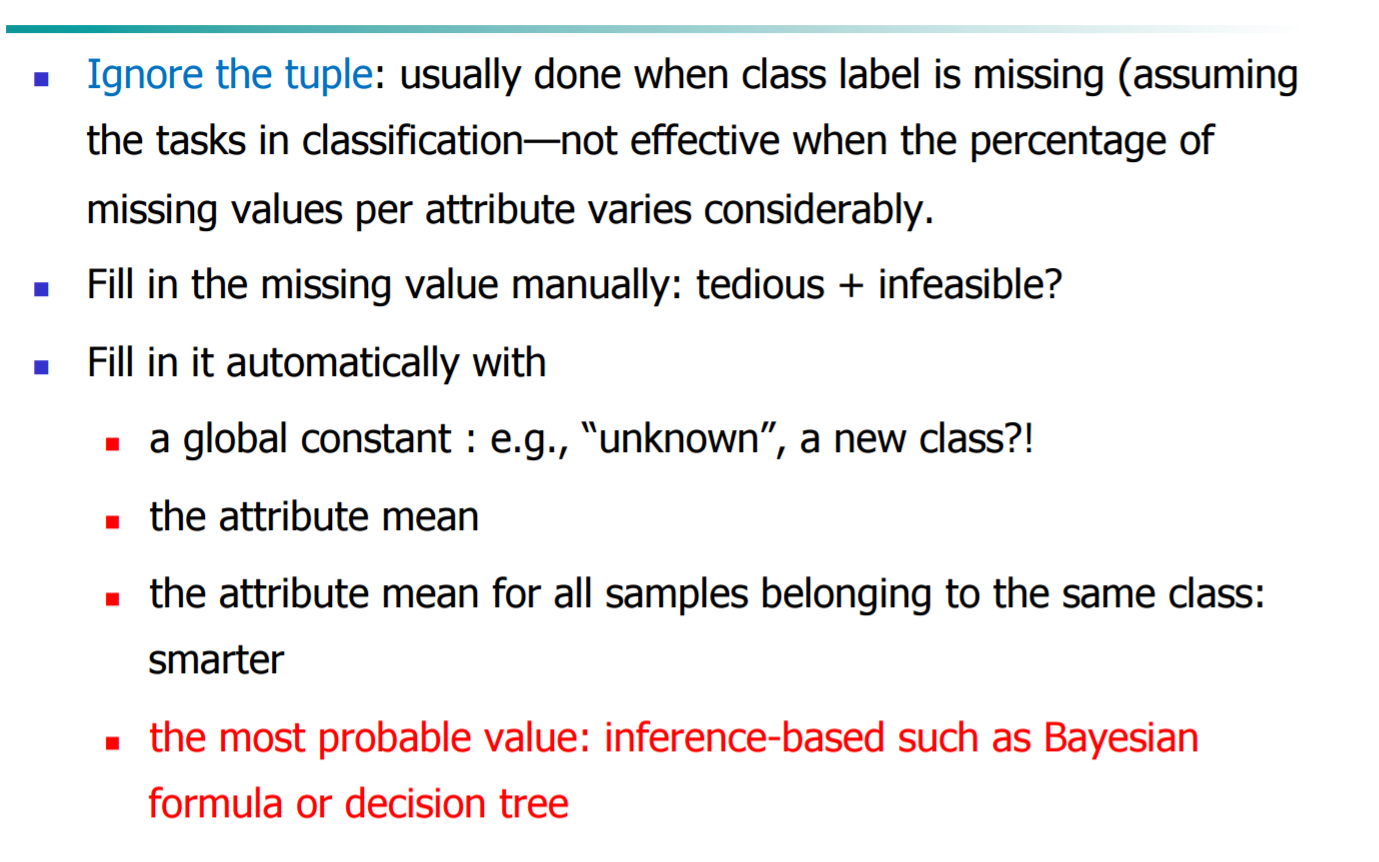
I Understand that real data is “dirty” (incomplete, noisy, inconsistent)

incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data

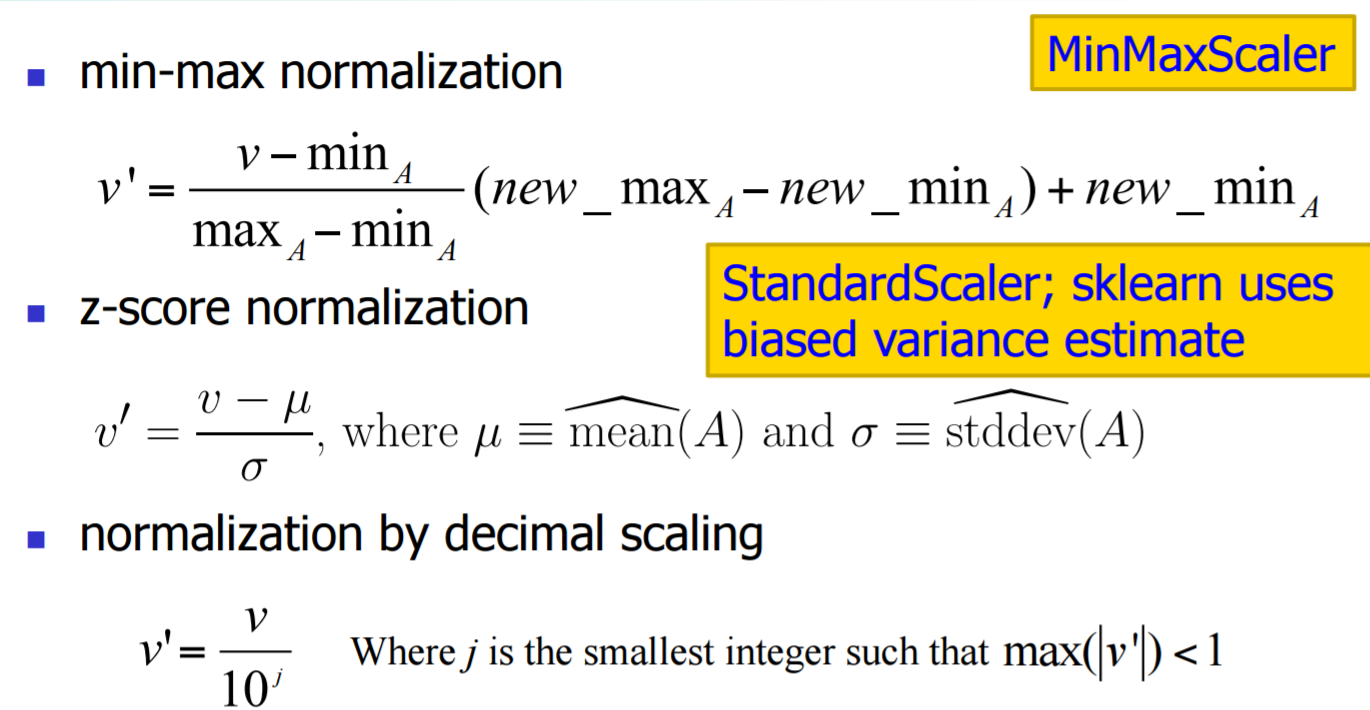
noisy: containing errors or outliers

inconsistent: containing discrepancies in codes or names

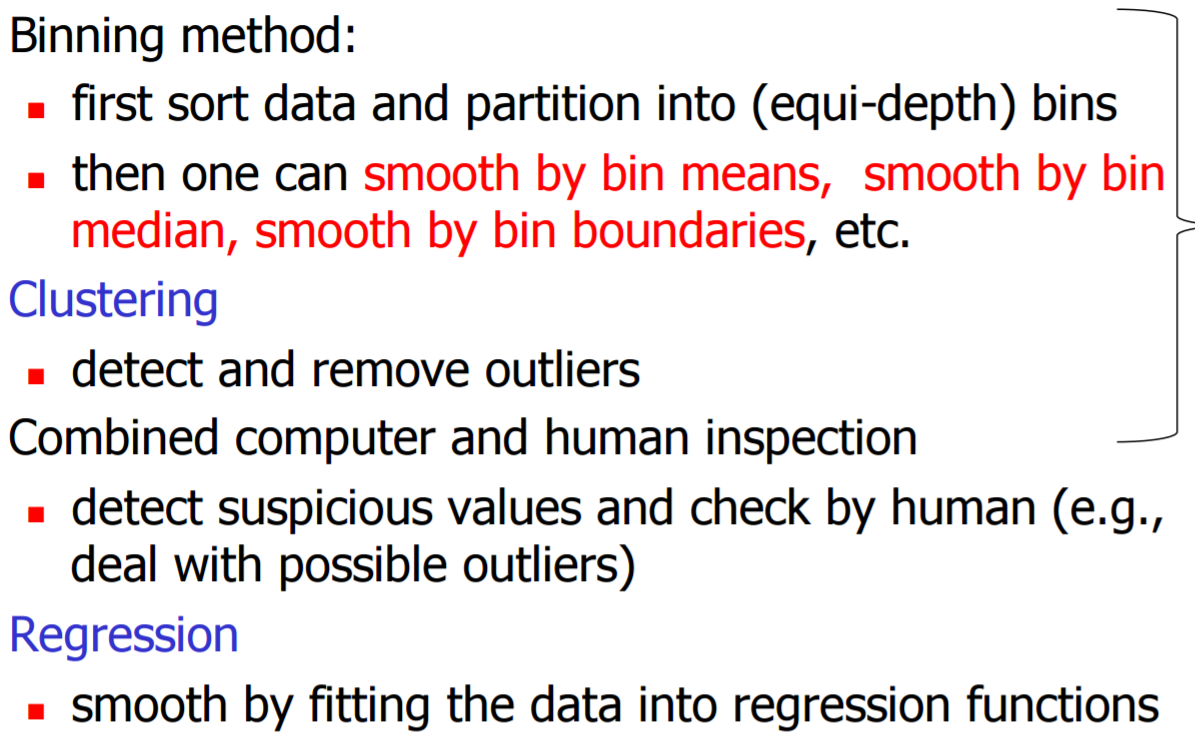
I How to handle missing data?

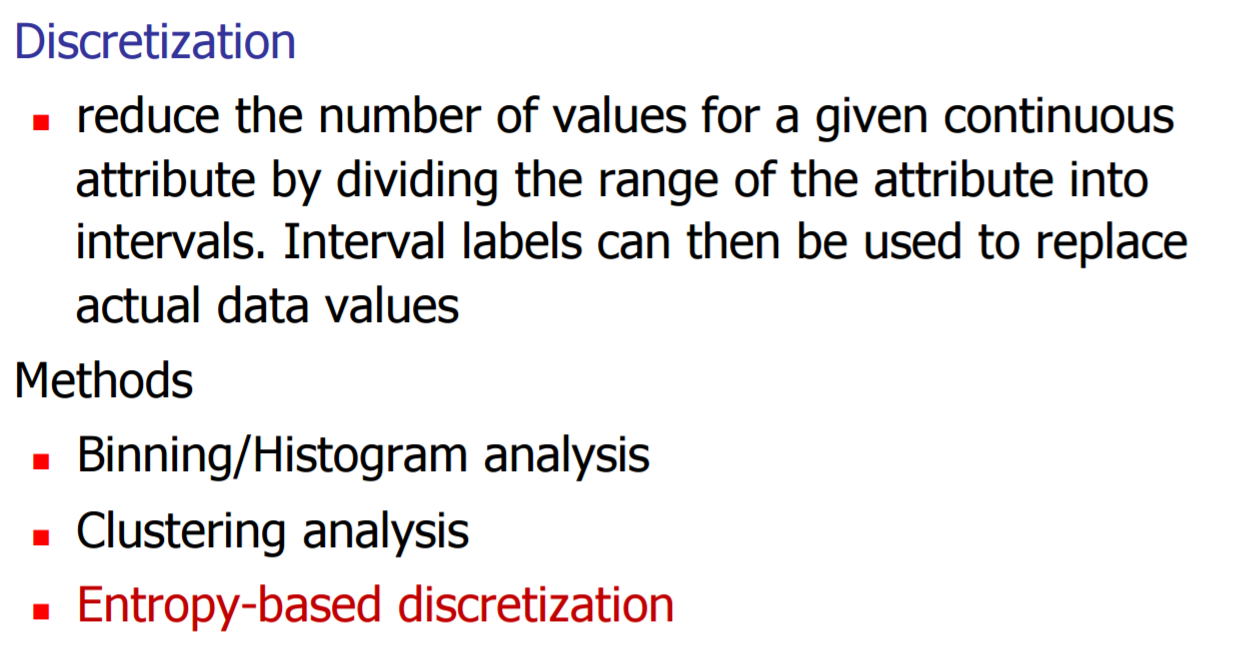


I How to normalize the data?



I How to handle noisy data? different binning/histogram method (including V-optimal and MaxDiff)

 I How to discretize data?



Classification and Prediction

Classification basics:

overfitting/underfitting;

cross-validation

Classification vs prediction vs clustering (unsupervised learning);

eager learning vs. lazy learning (instance-based learning)

Decision tree:

The ID3 algorithm

Decision tree pruning

Derive rules from the decision tree

The CART algorithm (with gin index)

Naive Bayes classifier

Smoothing

Two ways to apply NB on text data

Logistic regression/MaxEnt classifier;

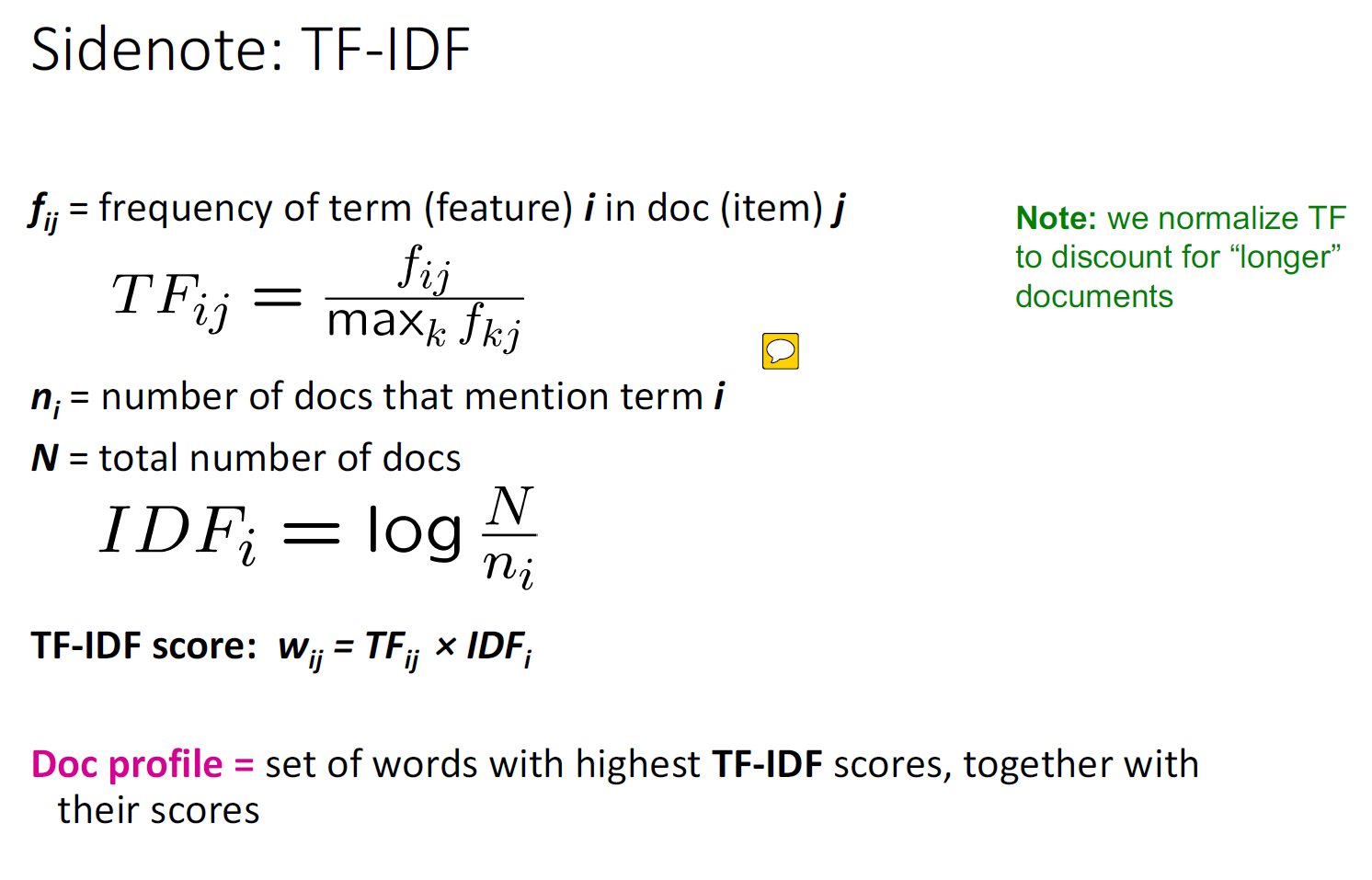
Maximum likelihood estimation of the model parameters + regularization; Gradient ascend.

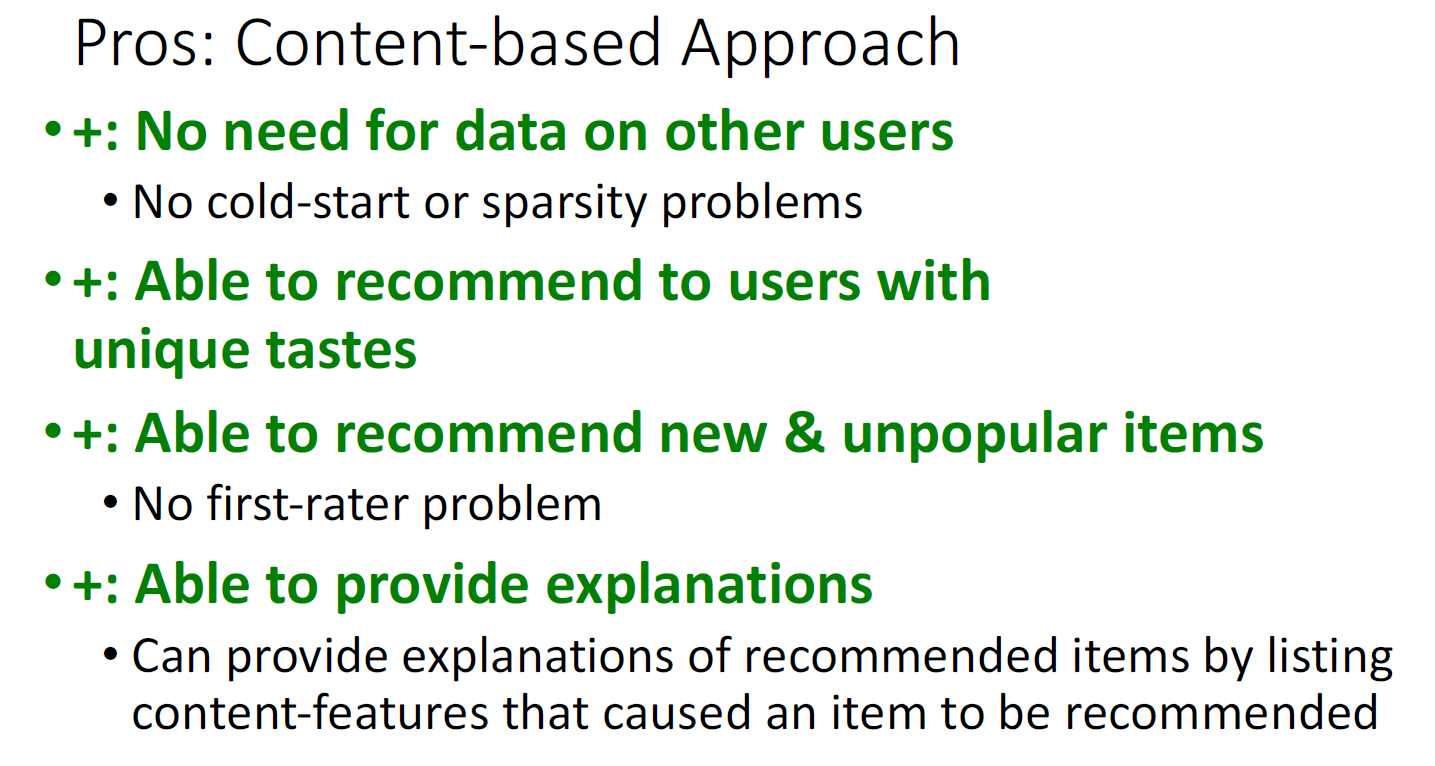
Recommendation System

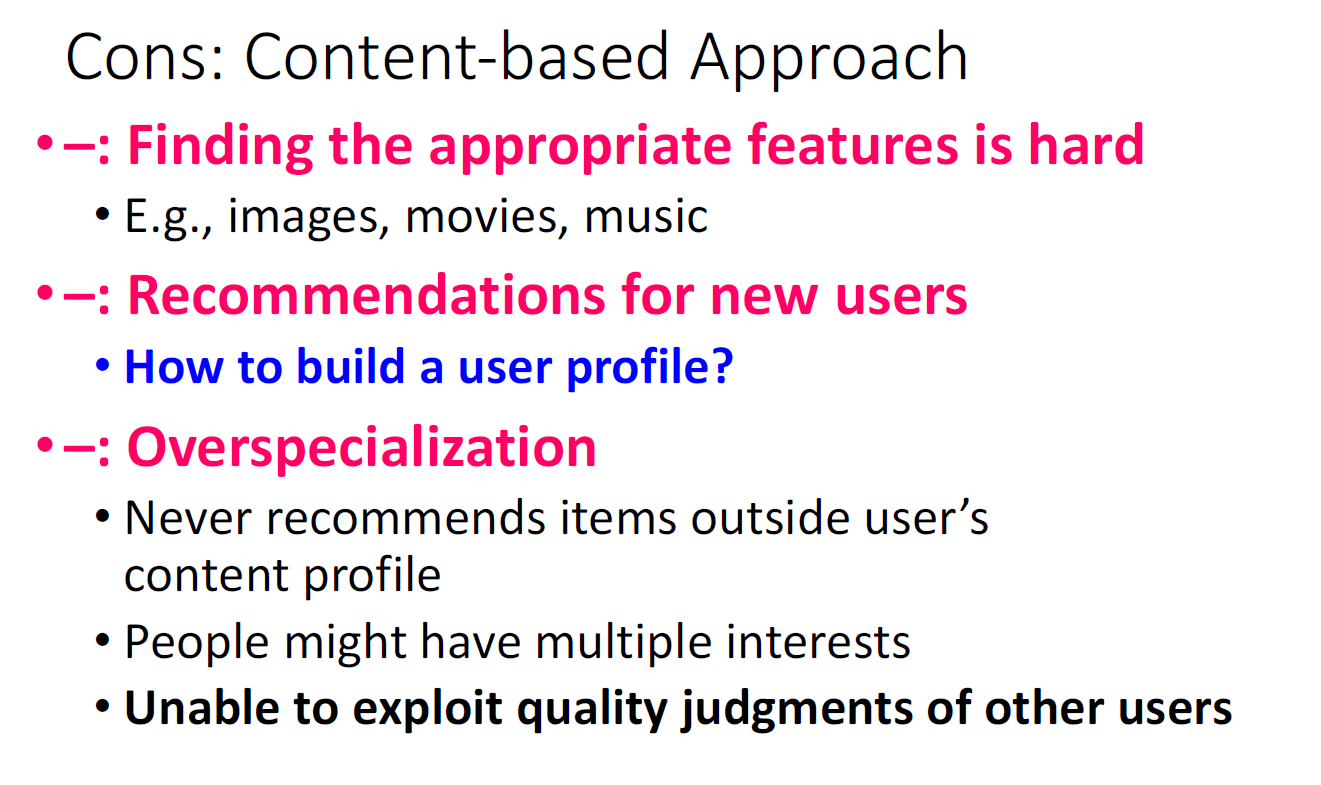
Problem definition and its input.

Three major approaches, including their pros and cons.

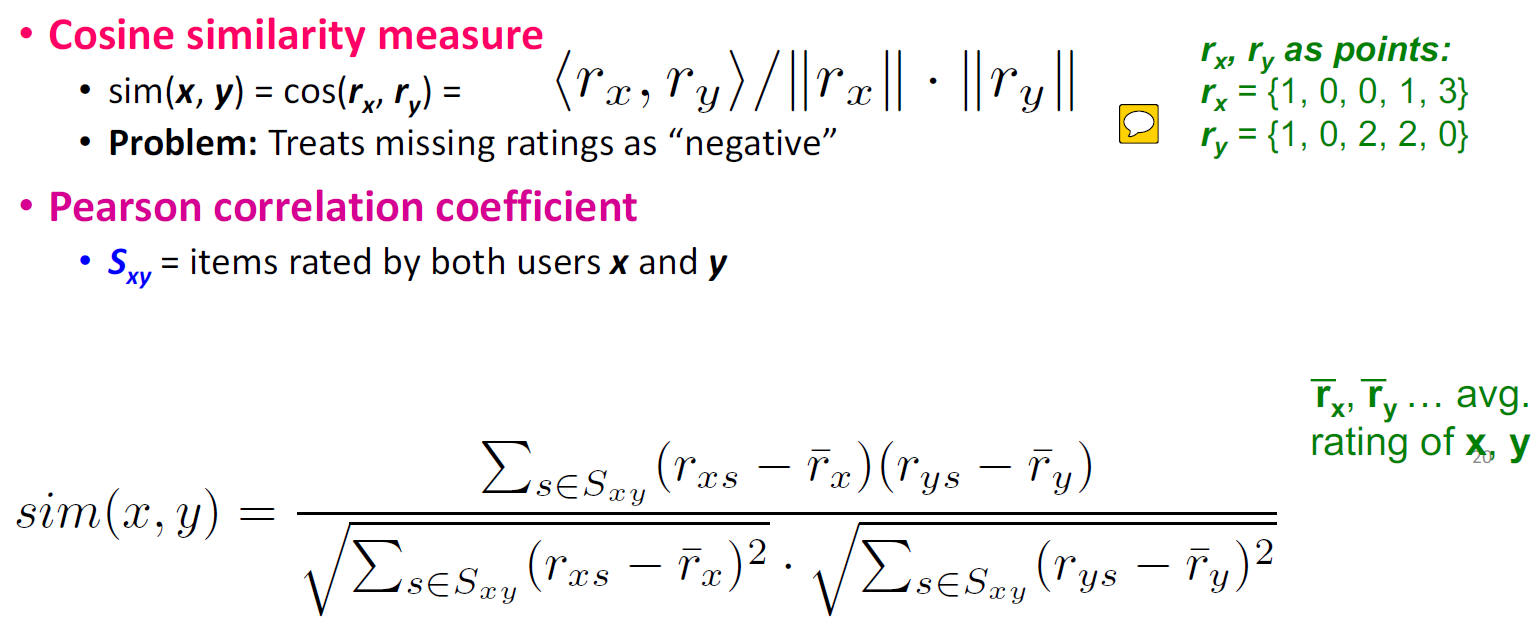
content based:

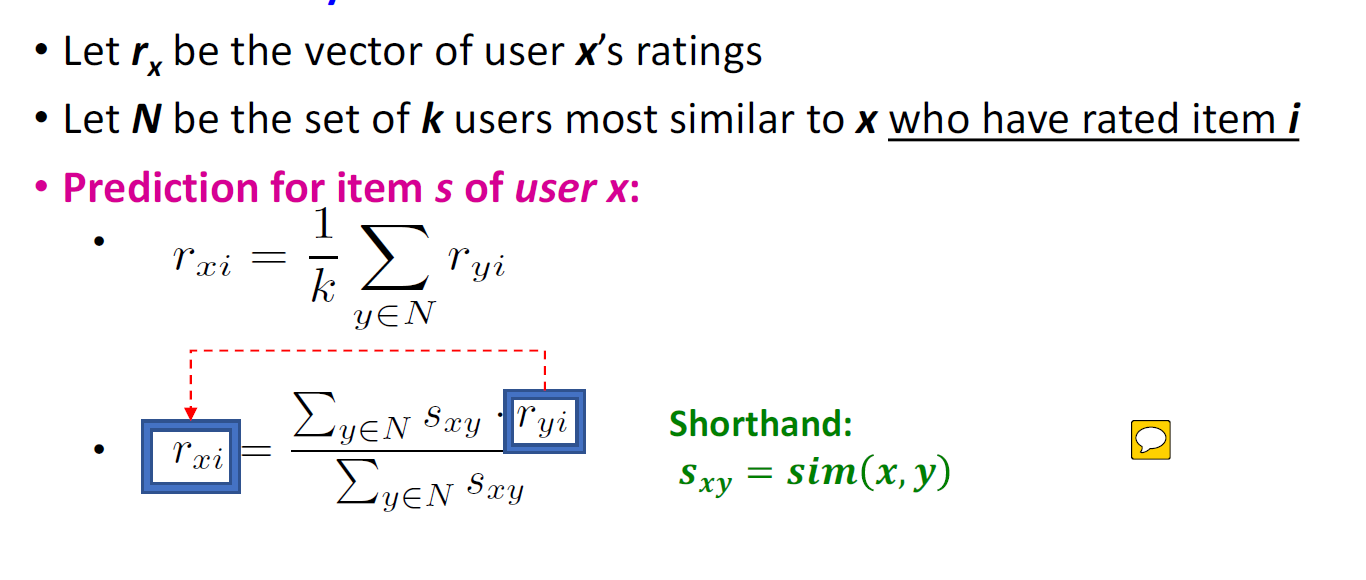


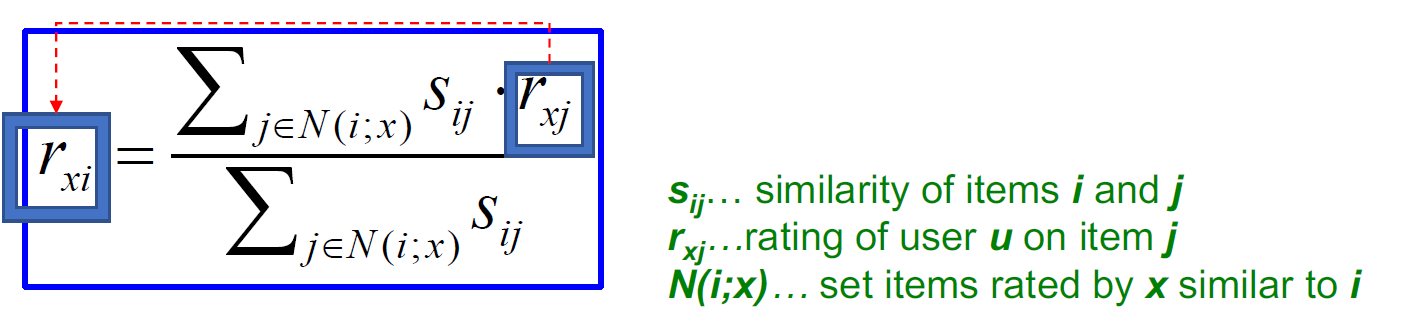


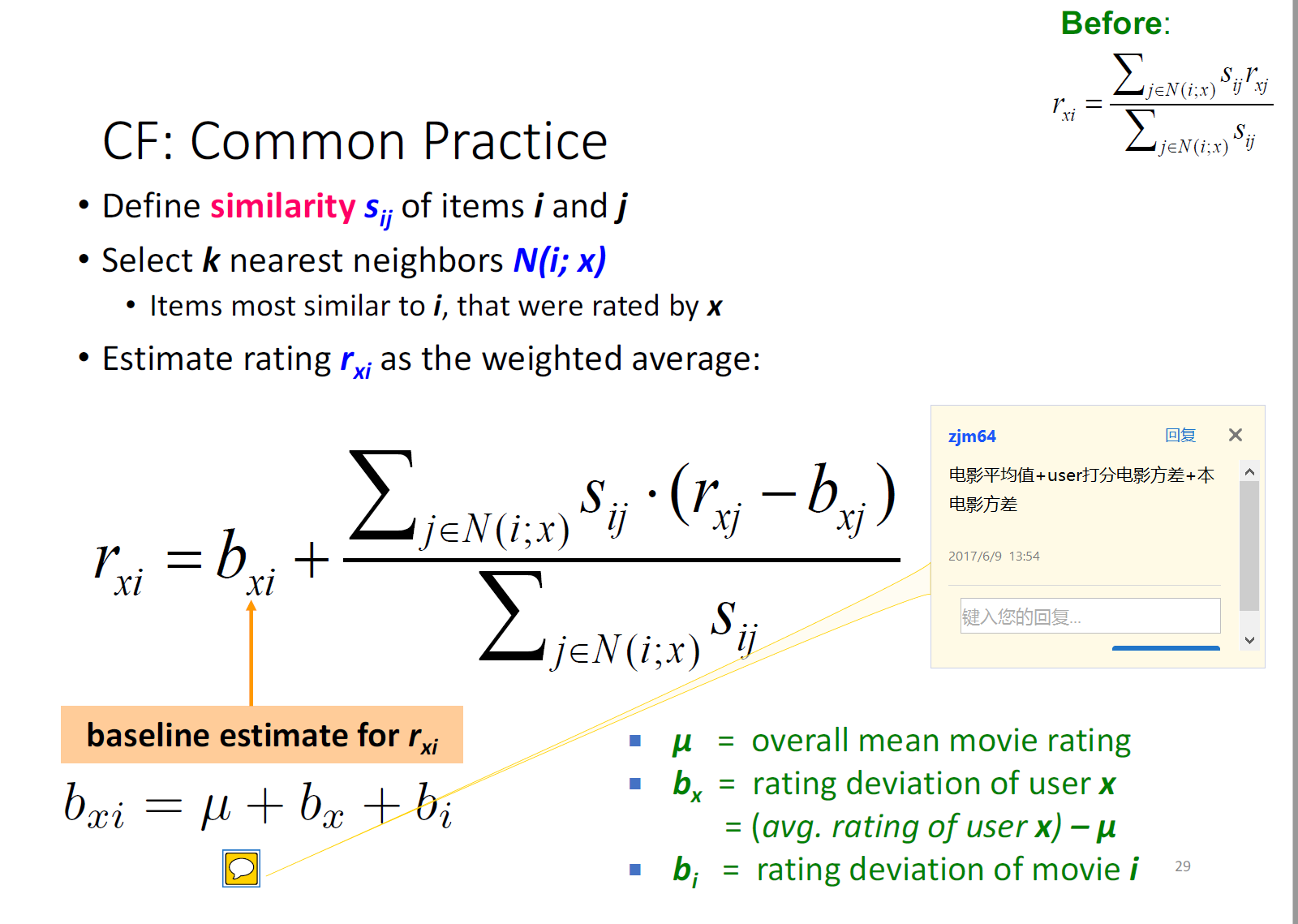


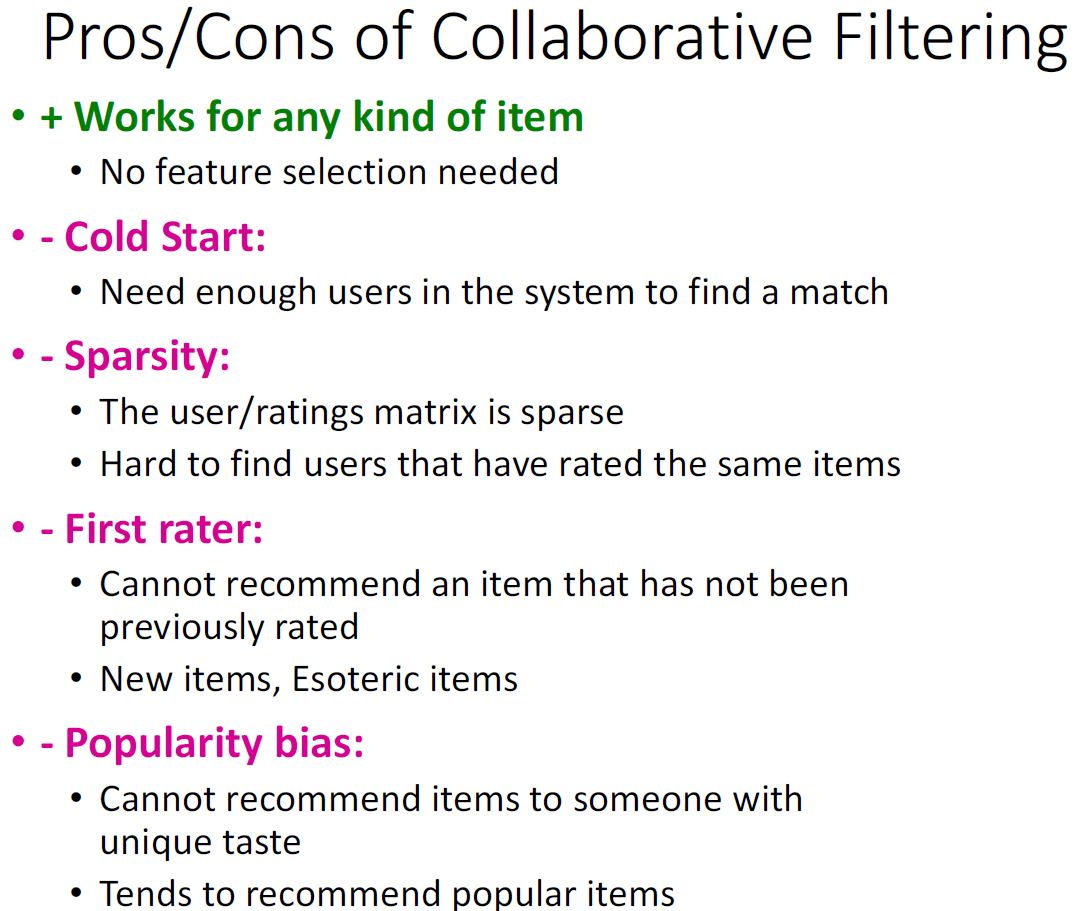
Collaborative filtering; pearson correlation











Matrix factorization-based; stochastic gradient descend.

Variants of Matrix factorization.