Distributed recipe mining to combine historic and social media knowledge

Sigurd Sippel

Betreuung: Kai von Luck

Motivation

Social Media

Historic Books

develops during the time

redundant

adaptions

Manhattan

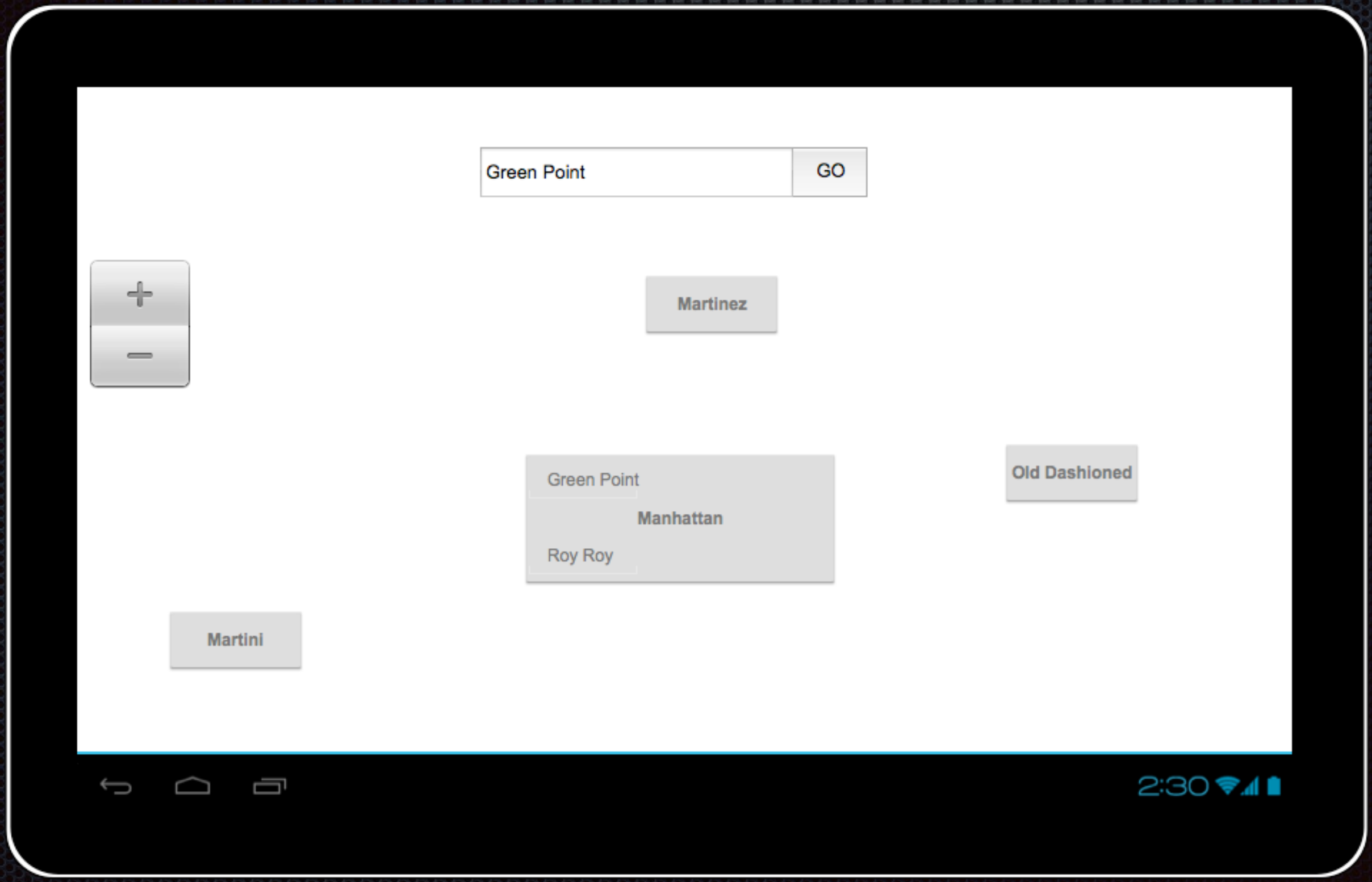
6 cl Rye

3 cl roter Wermut

stir, strain in cocktail glass

What is the main recipe?

Search on cluster map



Inner Cluster

Back to Overview

Recipes, Ingerediants or Tools | GO

Main Recipe

Manhattan

6 cl Rye

3 cl roter Wermut

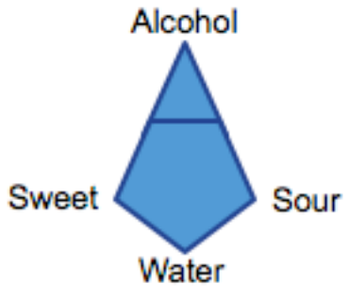
rühren, im Cocktailspitz

Adaptions

- Main Recipe
- Green Point
- Rob Roy

Statistics

How many findings?	122
How old is it?	140 year



Timeline

- 1878 Jerry Thomas S. 122
- 1922 Savoy S. 43
- ...

2:30

KDD process [FPSS96]

Knowledge discovery in database and data mining

Osama Fayyad

SIGKDD

KDD Conference

Gregory Piatetsky-Shapiro

Defining goal

Data selection

Data cleaning

Data reduction

Mining function

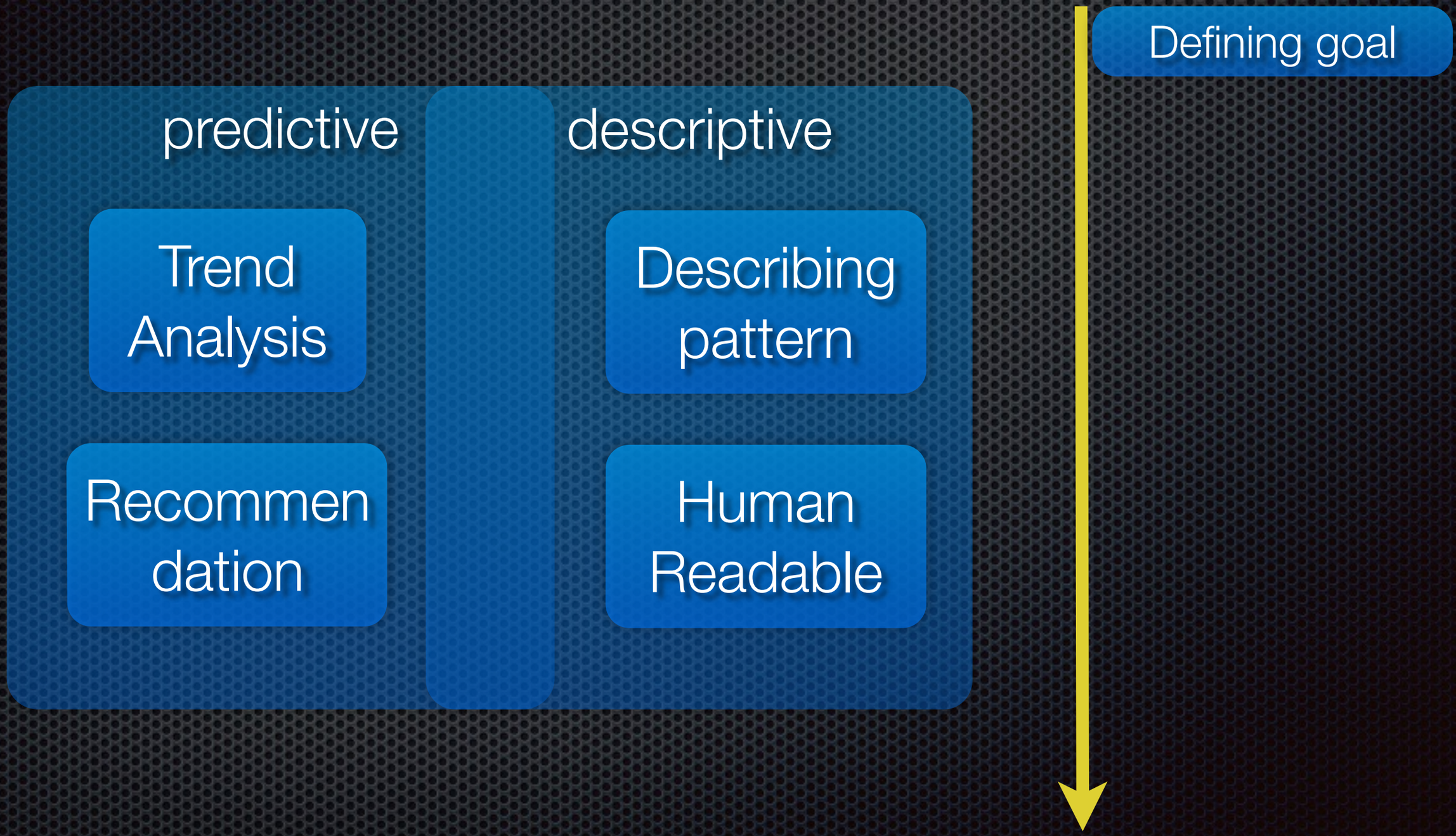
Mining algorithm

Data mining

Interpretation

Performance

Defining goal
domain and feature selection



Feature selection

date

author

description

title

Manhattan

6 cl Rye

3 cl roter Wermut

stir, strain in cocktail glass

ingredients

units

tools

preperation

glassware

Defining goal

Data selection

Data cleaning

Defining goal

Data selection

Data cleaning

Plausibility check

remove noise

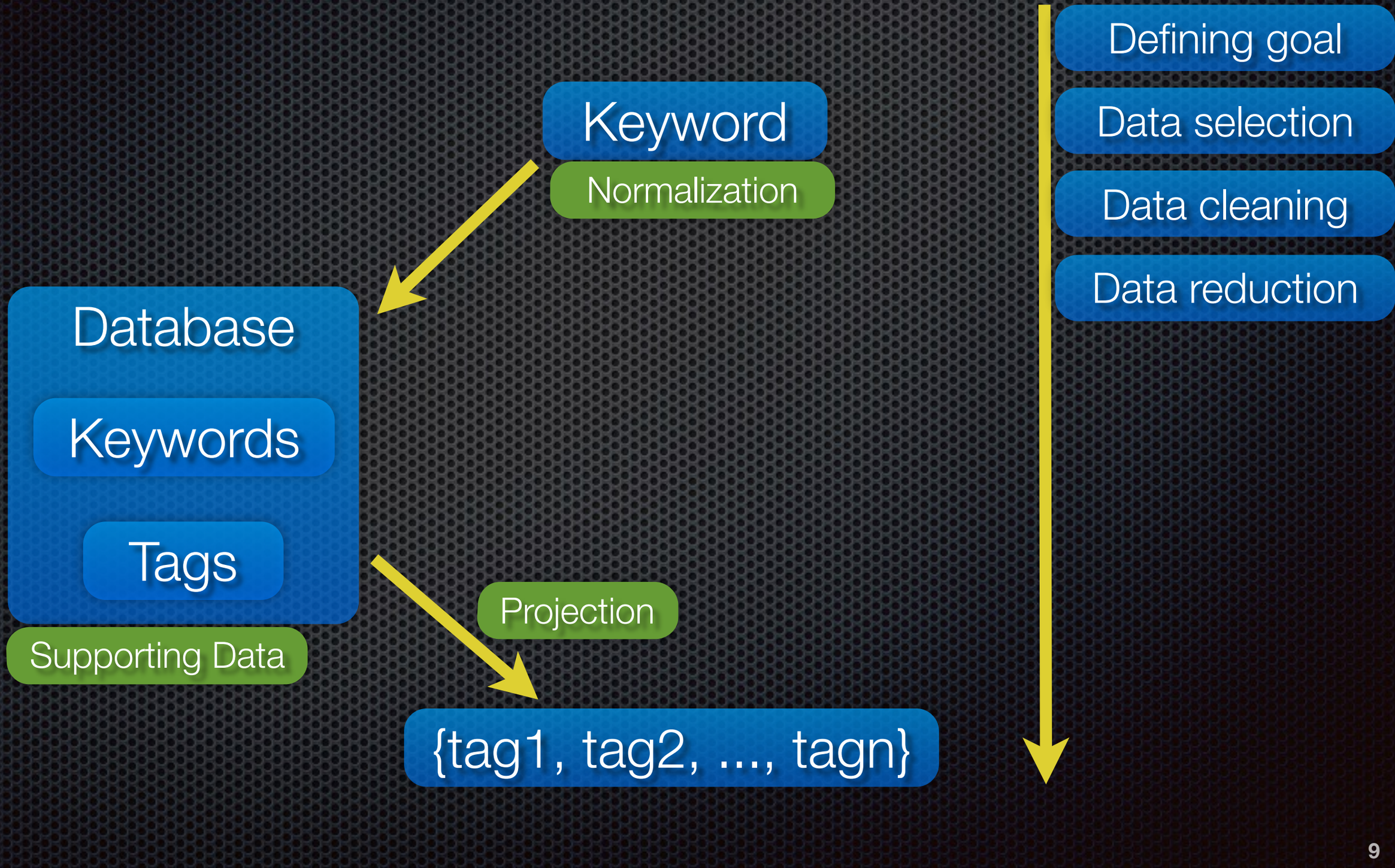
Handling missing data

Implicit information

Fuzzy Logic

Data reduction & projection

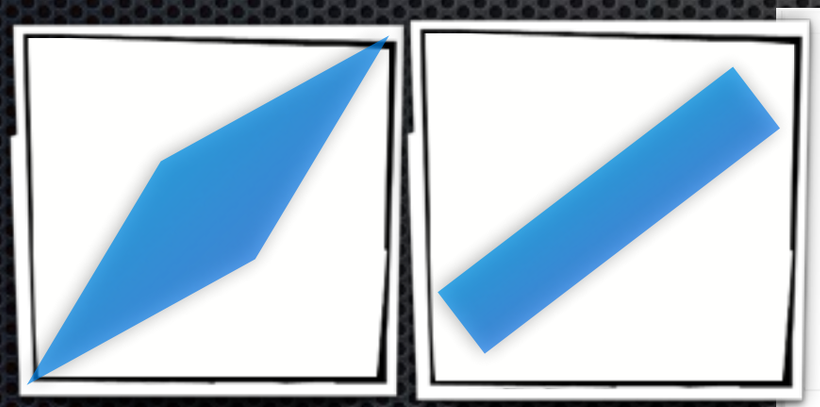
Feature extraction



Distance function

Euklid distance [JMF99]

$$d_2(\mathbf{x}_i, \mathbf{x}_j) = \left(\sum_{k=1}^{d} (x_{i,k} - x_{j,k})^2\right)^{1/2}$$



$$\operatorname{lev}_{a,b}(i,j) = \begin{cases} \max(i,j) & \text{if } \min(i,j) = 0, \\ \min \begin{cases} \operatorname{lev}_{a,b}(i-1,j) + 1 \\ \operatorname{lev}_{a,b}(i,j-1) + 1 \\ \operatorname{lev}_{a,b}(i-1,j-1) + 1_{(a_i \neq b_j)} \end{cases} & \text{otherwise.} \end{cases}$$

Dynamic-Time-Warping

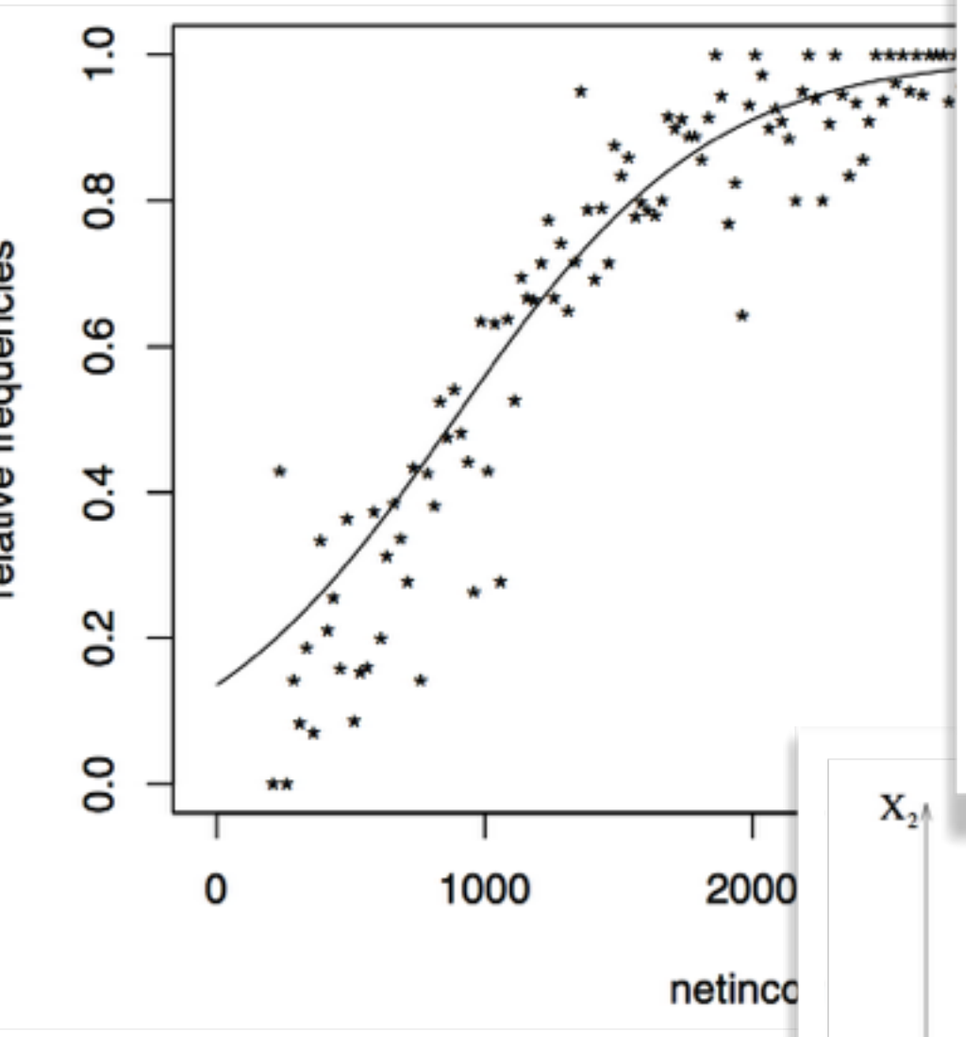
[HLWG08] Speech Recogniction

Levenshtein-Distance

String recognition

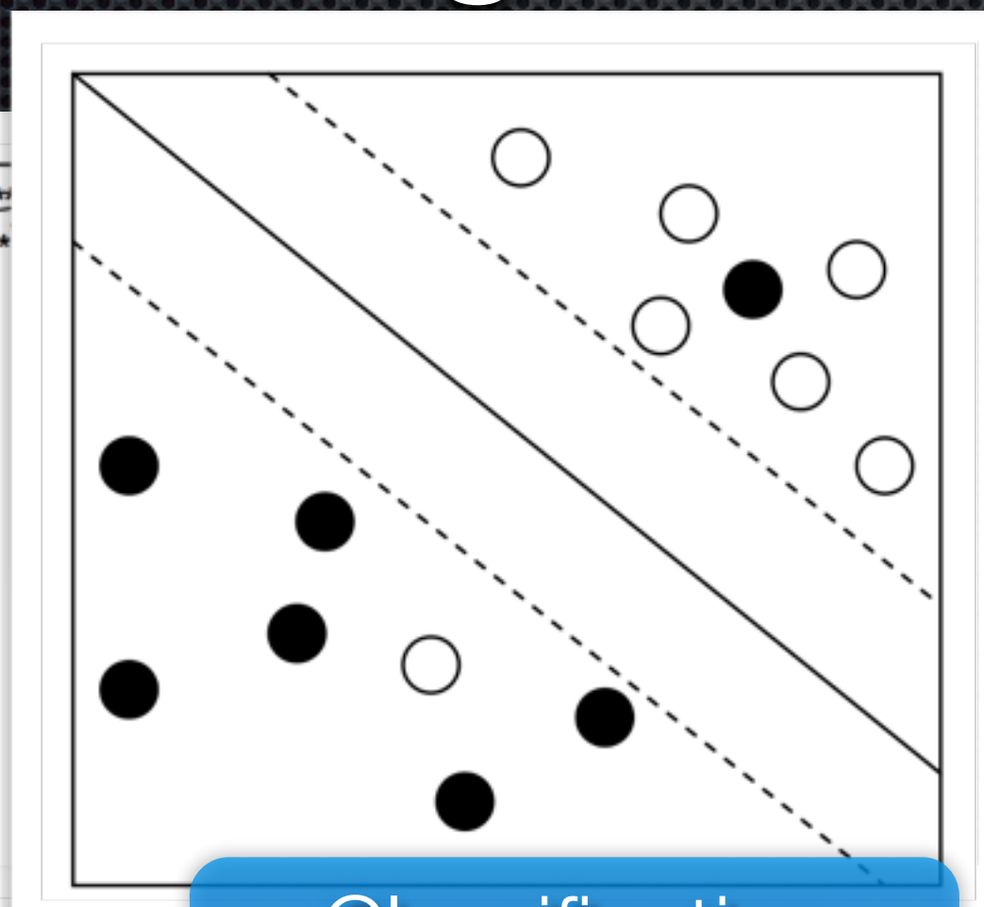
Choosing mining function

[FKPT07, S. 512]

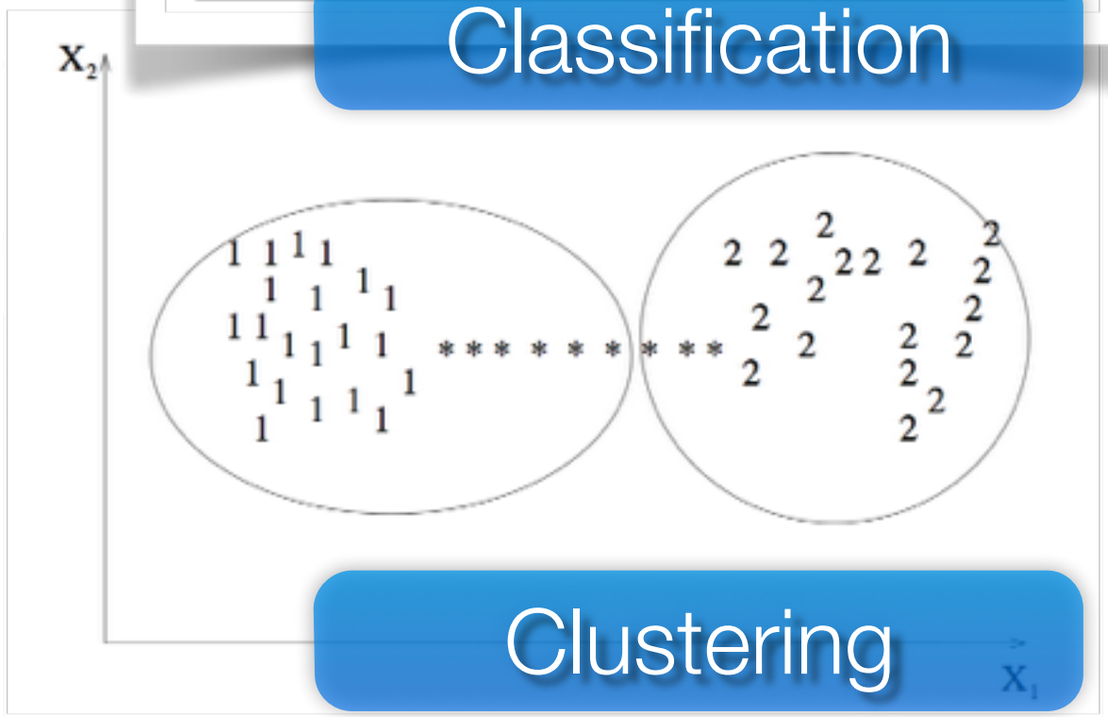


Regression

[LW06, S. 6]



Classification

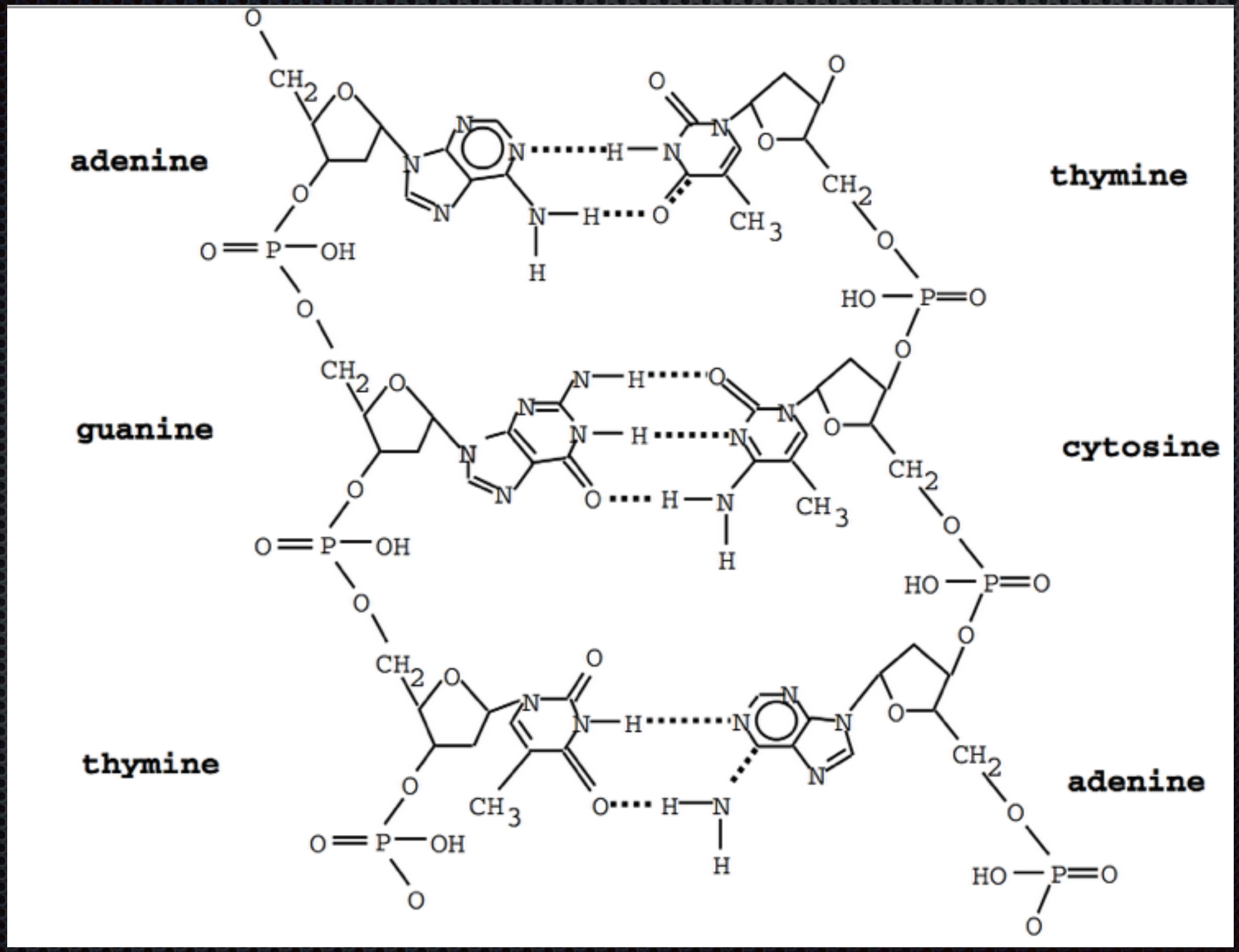


Clustering

- Defining goal
- Data selection
- Data cleaning
- Data reduction
- Mining function

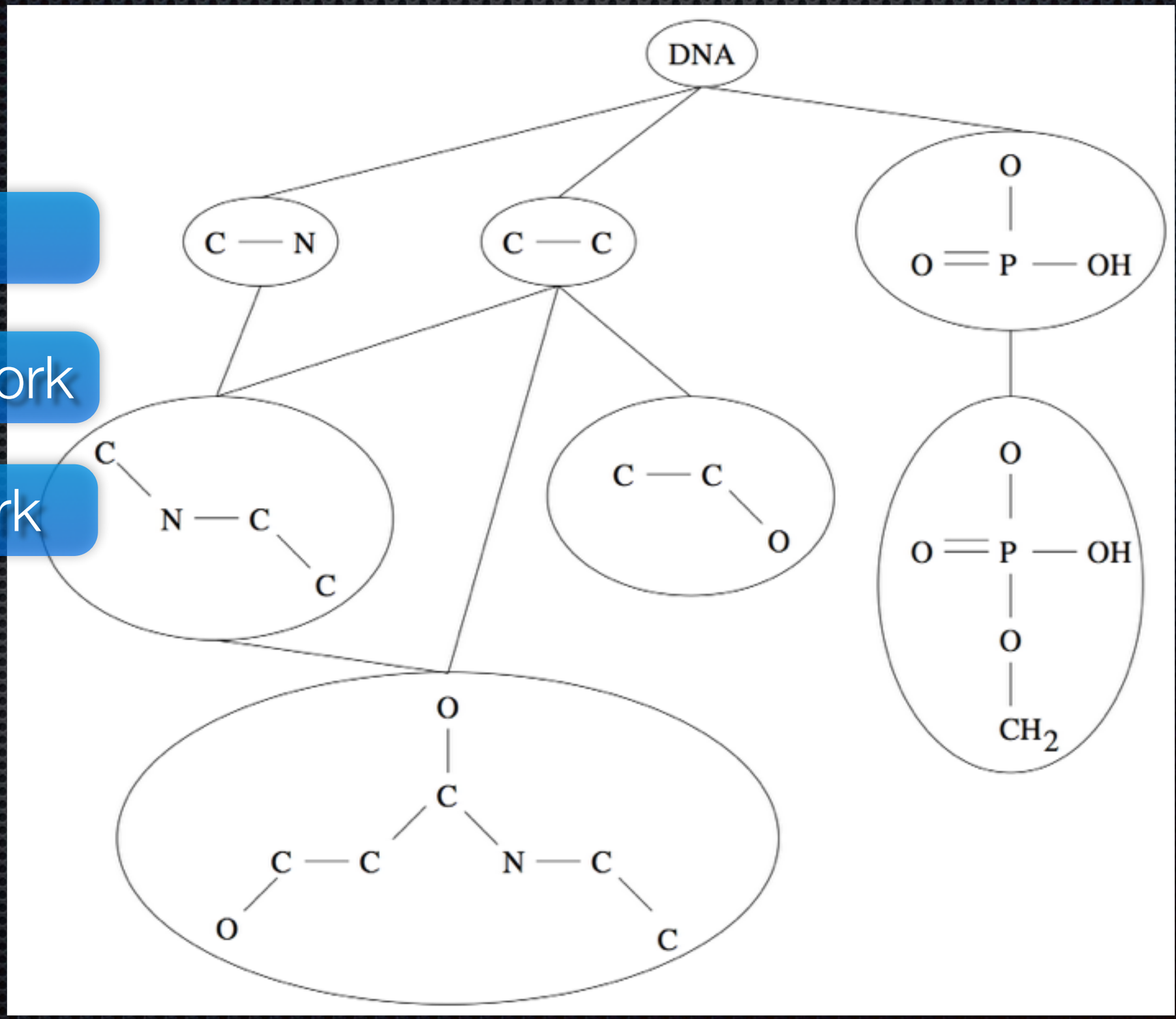
[JMF99, S. 277]

Chemical Model [MHC06, S. 80]

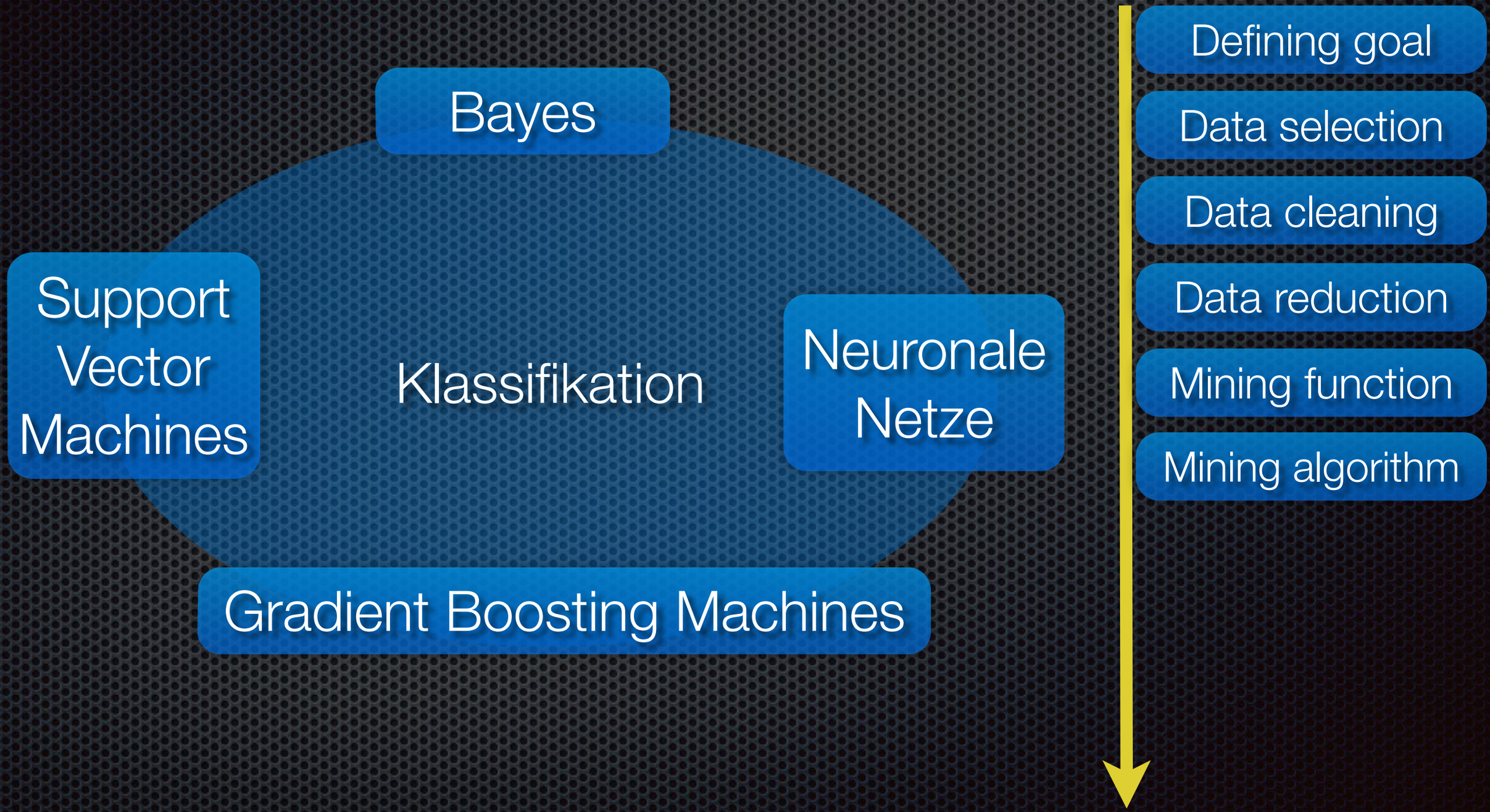


(Graph) Representation [MHC06, S. 80]

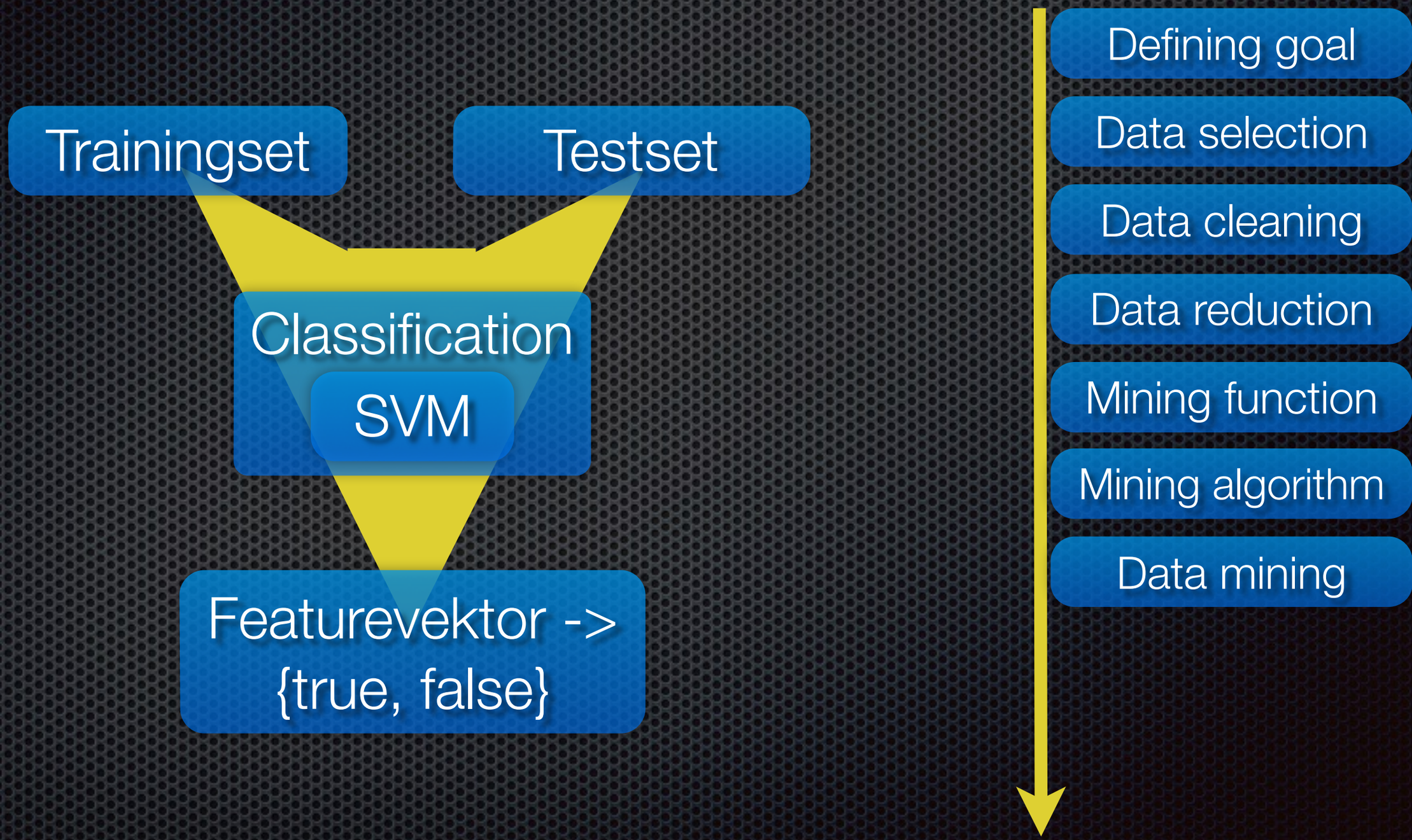
- Trees
- Bayesian network
- Neural network



Choosing mining algorithm



Data mining: Classification



Interpretation

Remove redudant patterns

Translating for users

Visualization

Defining goal

Data selection

Data cleaning

Data reduction

Mining function

Mining algorithm

Data mining

Interpretation

Optimizing

Training set

Gradient descent technique

Hill climbing

Feature Complexity

Defining goal

Data selection

Data cleaning

Data reduction

Mining function

Mining algorithm

Data mining

Interpretation

Performance

Large scale Twitter mining
of drug related adverse events [BTY12]

Socal Media

Facebook

Twitter

Mining

is drug related?

Its hot weather, Tamoxifen is a nightmare

(hot weather, Tamoxifen) => Nightmare

Classification Training

User Training set

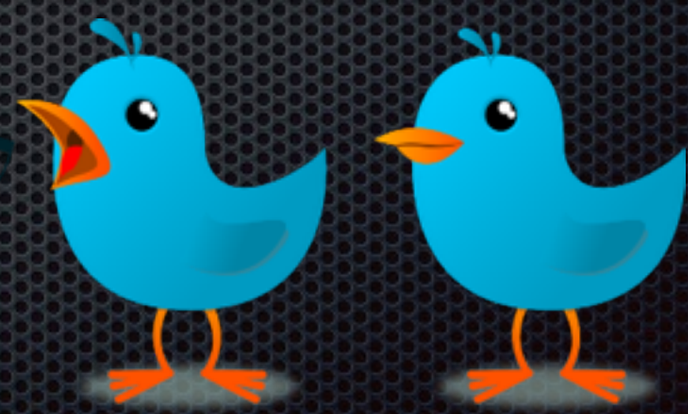
Potential drug users

Corresponding Keywords in Timeline

Tweet Training set

Medicine Ontology

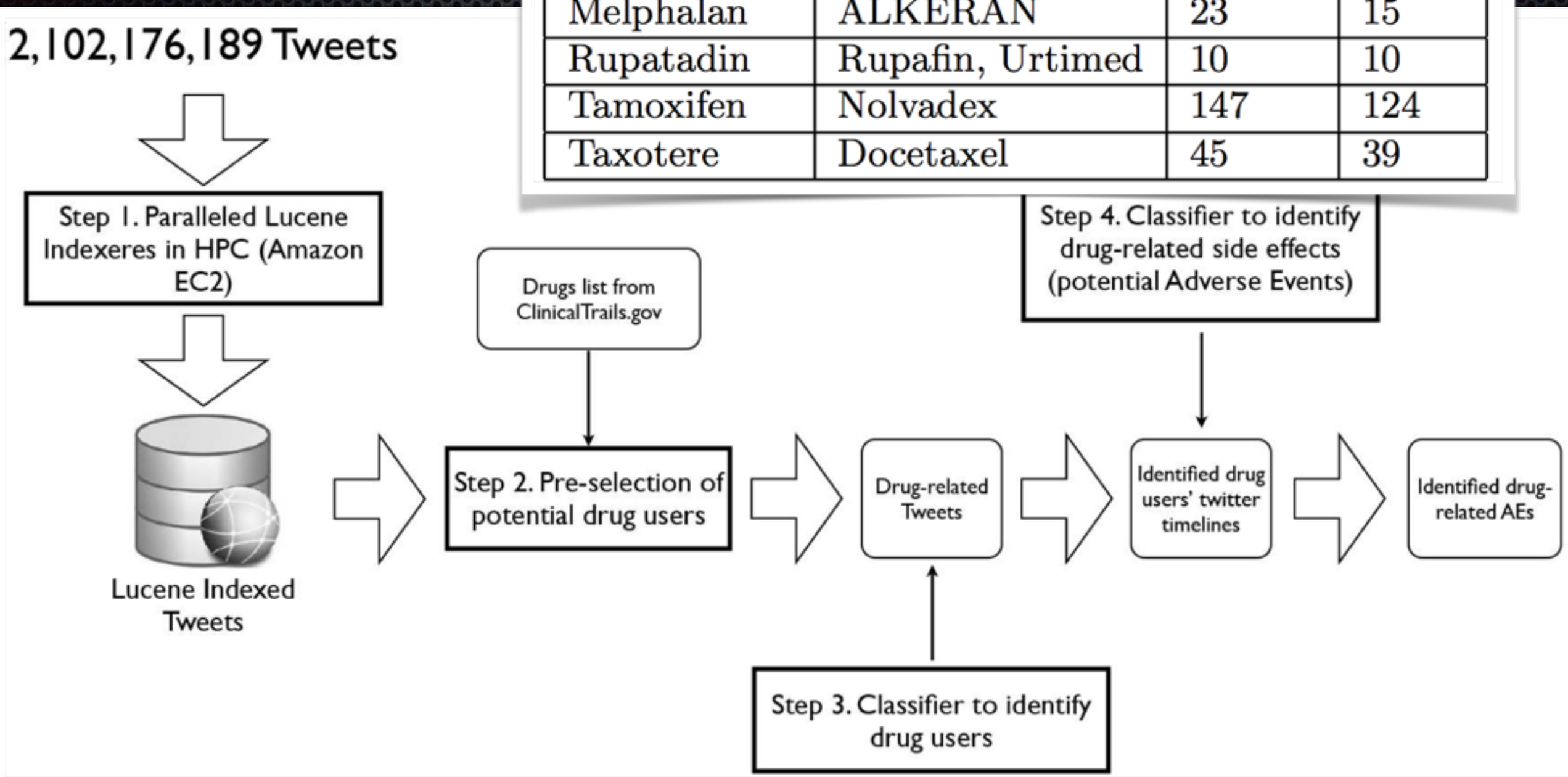
Unified Medicine Language system



Mining Process to find user

[BTY12, S. 27]

Drug name	Synonym(s)	# of tweets	# of users
Avastin	Bevacizumab	264	236
Melphalan	ALKERAN	23	15
Rupatadin	Rupafin, Urtimed	10	10
Tamoxifen	Nolvadex	147	124
Taxotere	Docetaxel	45	39



Feature Extraction of Tweets

Bag of words

Action

State of Drag Using

Count

Hash Tags

Reply Tags

URLs

Pronouns

Drug names

Relevance

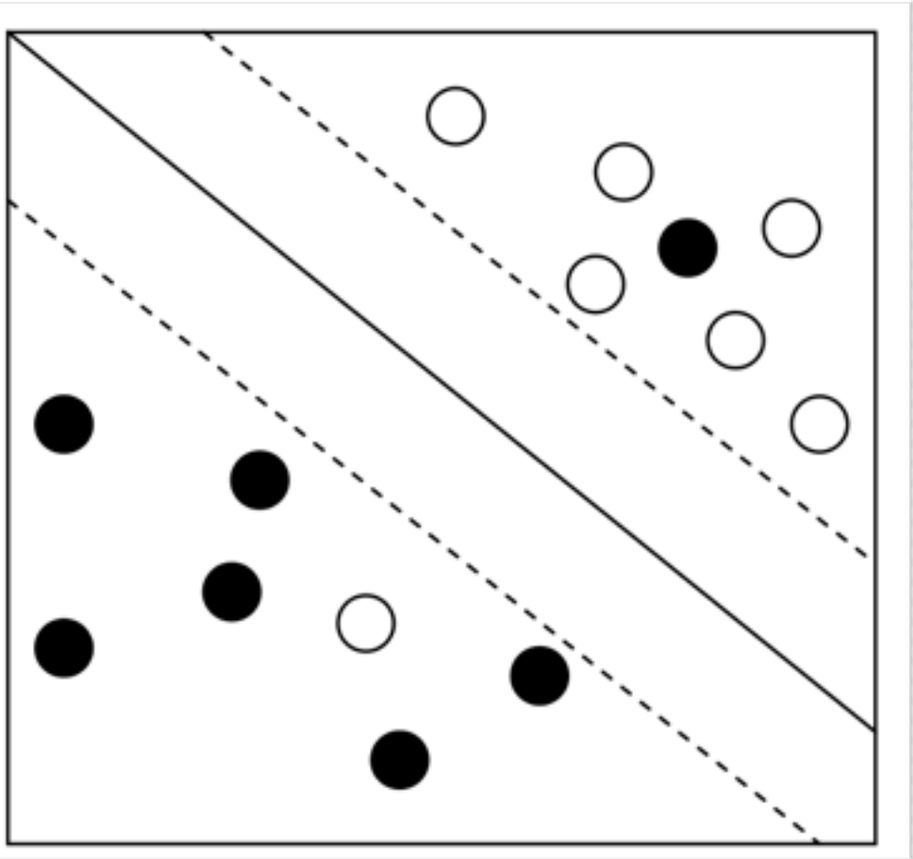
Count

Semantic Types

Semantic Groups

Relevance in Unified Medicine Language system

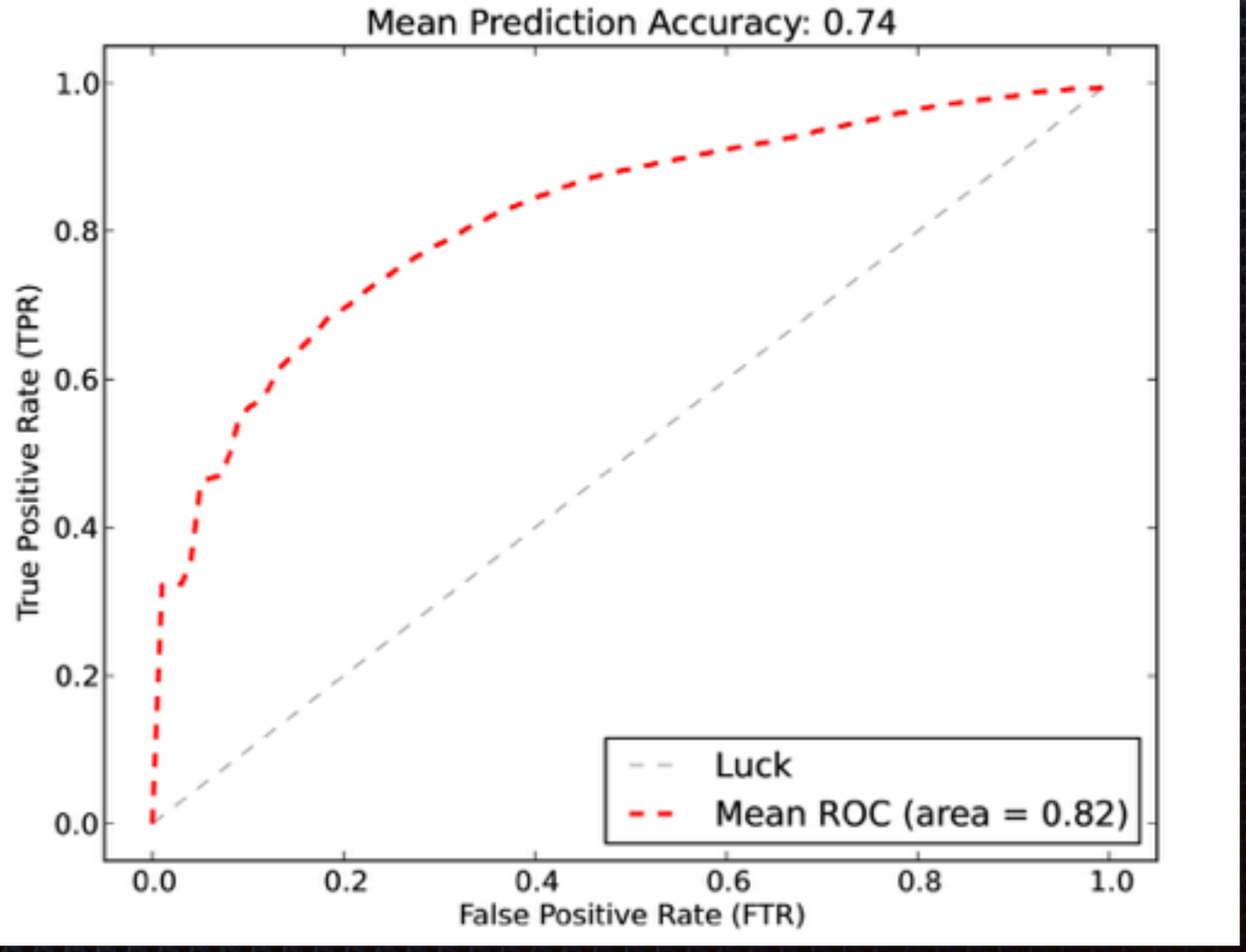
Support Vector Machine



[LW06, S. 6]

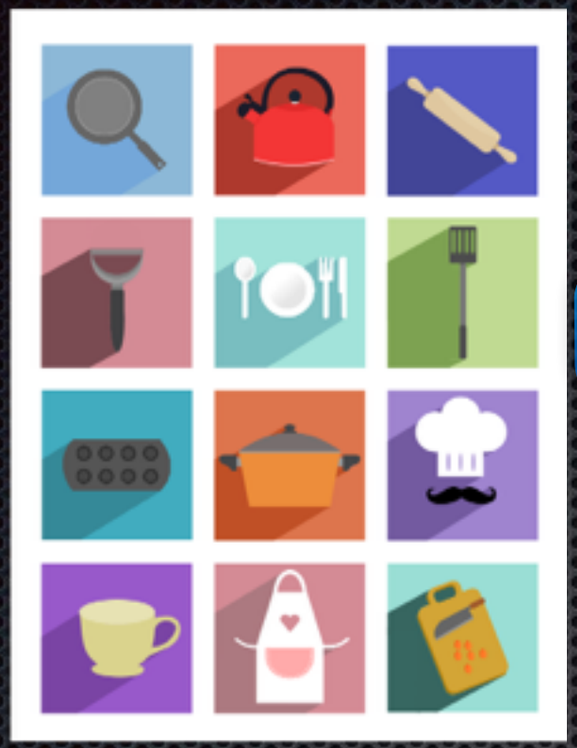
decision function g(x) = sign(f(x))

hyper plane <w,x> + b = 0



[BTY12, S. 30]

Recipe recommendation [TLA12]



Preparation

Dishes

removed quantities

removed temperature

Ingredient combination

key ingredients

modification options

nutritions

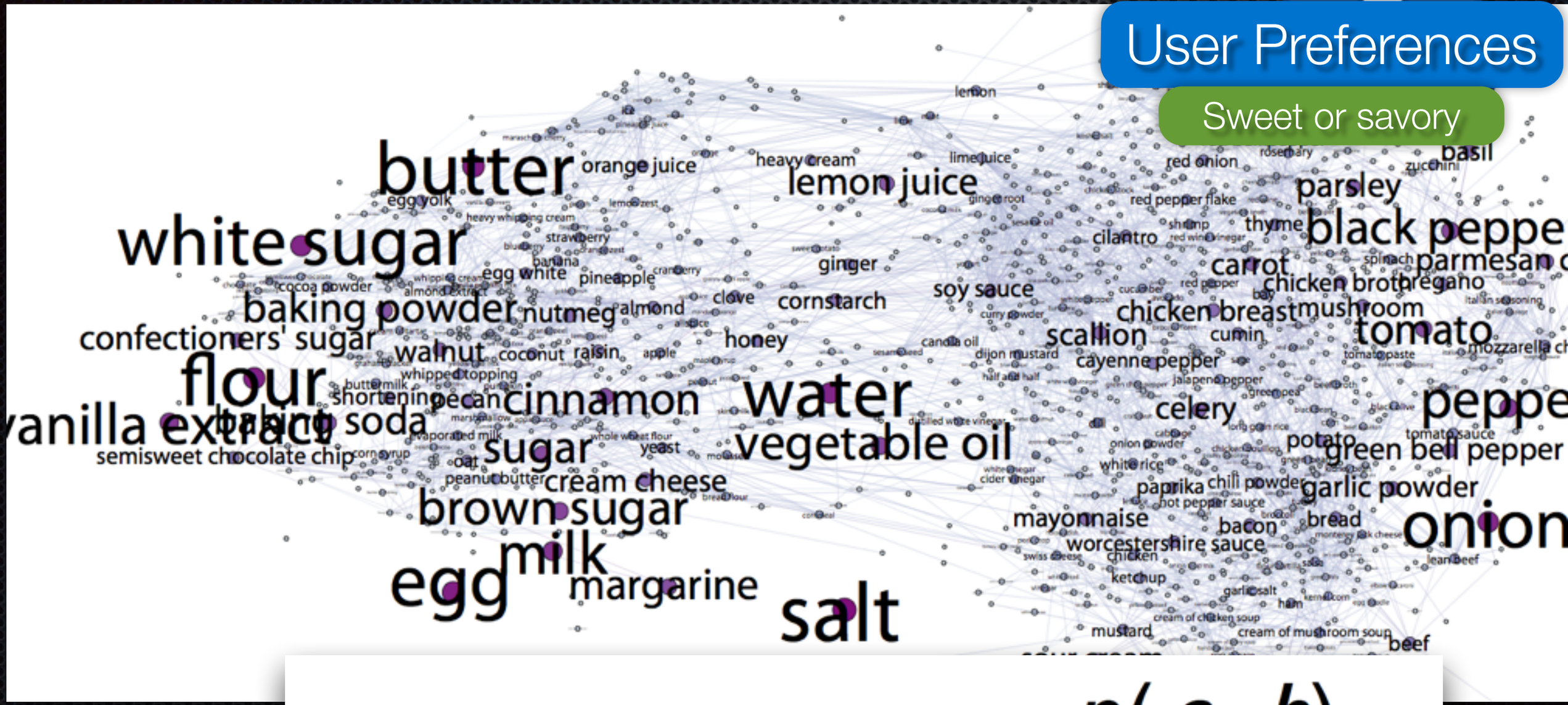
Gradient Boosting Machine

Classification [Fri01]

Co-occurrence network

User Preferences

Sweet or savory



[TLA12, S. 301]

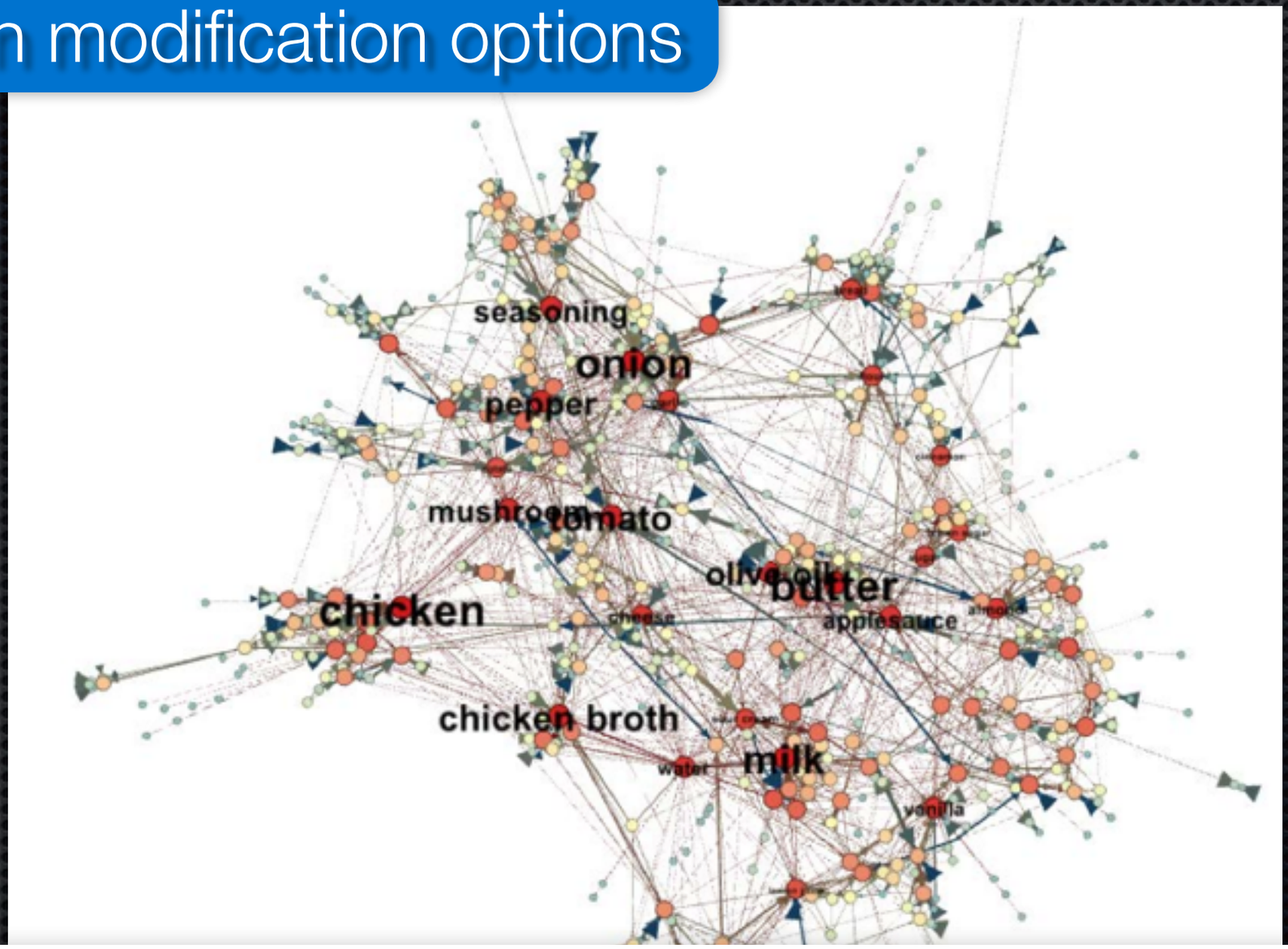
$$PMI(a,b) = log \frac{p(a,b)}{p(a)p(b)},$$

pointwise mutual information

Substitution network

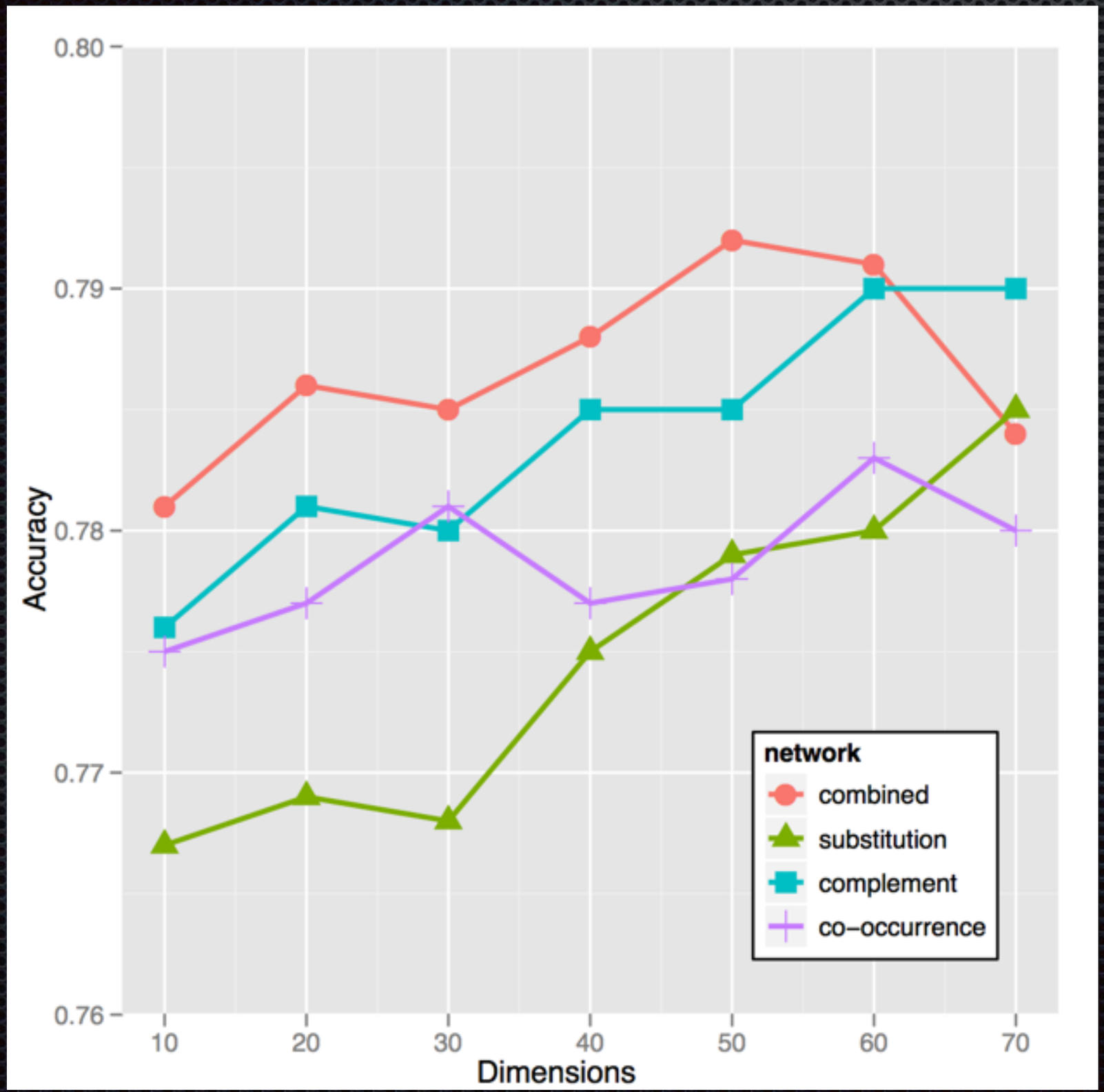
[TLA12, S. 302]

based on modification options



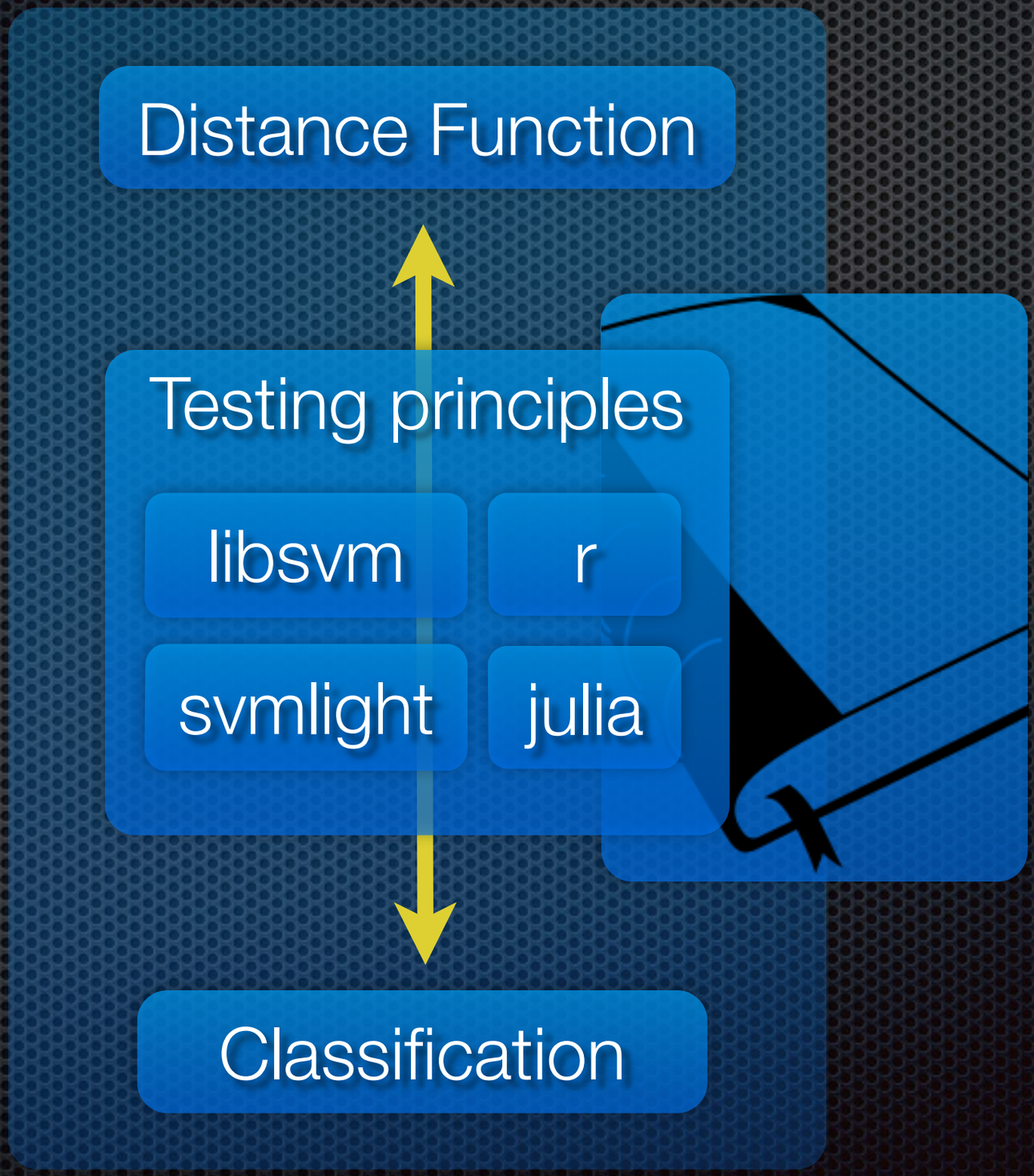
$$\mathrm{PMI}(a \to b) = \log \frac{p(a \to b)}{p(a)\,p(b)},$$

Prediction performance



[TLA12, S. 307]

Roadmap Project 1



Sources

[BTY12] Bian, Jiang ; Topaloglu, Umit ; Yu, Fan: Towards Large-scale Twitter Mining for Drug- related Adverse Events. In: Proceedings of the 2012 International Workshop on Smart Health and Wellbeing. New York, NY, USA : ACM, 2012 (SHB '12). – ISBN 978–1–4503– 1712–2, 25–32

[FKPT07] Fahrmeir, Ludwig ; Künstler, Rita ; Pigeot, Iris ; Tutz, Gerhard: Statistik. Springer- Verlag Berlin Heidelberg, 2007 (Springer-Lehrbuch). http://books.google.de/books? id=ZinjP103iRcC. – ISBN 9783540697398

[FPSS96] Fayyad, Usama ; Piatetsky-Shapiro, Gregory ; Smyth, Padhraic: The KDD Pro- cess for Extracting Useful Knowledge from Volumes of Data. In: Commun. ACM 39 (1996), November, Nr. 11, 27–34. http://dx.doi.org/10.1145/240455.240464. – DOI 10.1145/240455.240464. – ISSN 0001–0782

[Fri01] Friedman, Jerome H.: Greedy function approximation: a gradient boosting machine. In: Annals of Statistics (2001), S. 1189–1232

[JMF99] Jain, A. K. ; Murty, M. N. ; Flynn, P. J.: Data Clustering: A Review. In: ACM Com- put. Surv. 31 (1999), September, Nr. 3, 264–323. http://dx.doi.org/10.1145/331499. 331504. – DOI 10.1145/331499.331504. – ISSN 0360–0300

[LW06] Lovell, Brian C. ; Walder, Christian J.: Support vector machines for business applica- tions. (2006)

[MHC06] Maulik, U. ; Holder, L.B. ; Cook, D.J.: Advanced Methods for Knowledge Discovery from Complex Data. Springer, 2006 (Advanced Information and Knowledge Processing). http://books.google.de/books?id=OOOSx1X2-_sC. – ISBN 9781846282843

[TLA12] Teng, Chun-Yuen ; Lin, Yu-Ru ; Adamic, Lada A.: Recipe Recommendation Using Ingredient Networks. In: Proceedings of the 3rd Annual ACM Web Science Conference. New York, NY, USA : ACM, 2012 (WebSci '12). – ISBN 978–1–4503–1228–8, 298–307