universität wien

Advanced quantitative text analysis (2022W)

,

Day 3

universität wien

Today

9:45-10:40	Input: Dictionaries, Regular Expressions
10:40-10:50	Coffee Break
10:50-11:45	Coding
11:45-12:45	Lunch
12:45-13:40	Input: Supervised Machine Learning
13:40-13:50	Coffee Break
13:50-14:45	Coding

universität wien

Dictionaries, Regular Expressions

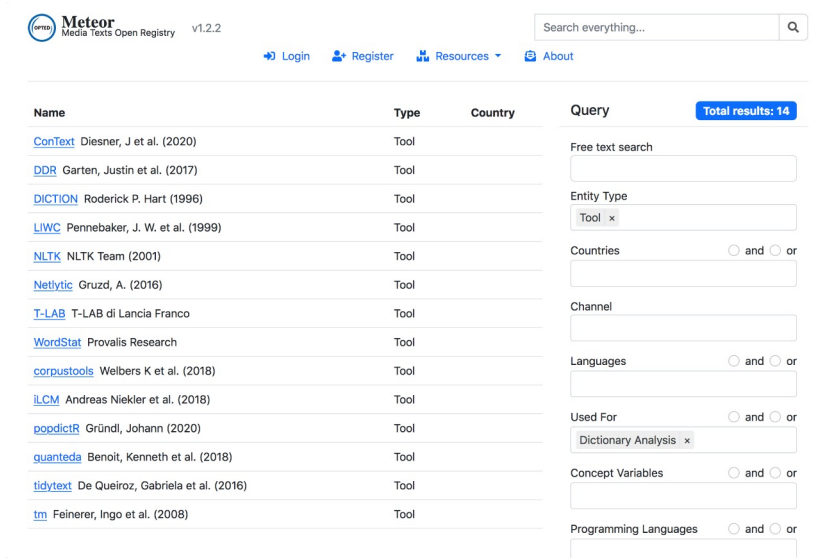
Day 3 Session 1

Dictionary method

Rule-based classification approach when categories are known

- List of words (or phrases) that indicate a category
- Create your own or use/edit existing dictionaries
- Always: validate!

Some existing dictionaries and dictionary application tools



https://meteor.opted.eu/query?dgraph.type=Tool&used_for=0x1ade5

universität wien

Dictionary use cases

- Selecting text data according to some defined category (‘what is relevant data’)

- Classifying documents into known categories

universität wien

Example: Search string for migration news

Table 3 Boolean search strings used for retrieval of migration-related news articles

Country	Language	Search string
Spain	Spanish	asilo* OR inmigra* OR refugiad* OR migrante* OR migratori* OR "sin papeles" OR "campo de desplazados" OR patera* OR emigra* OR "libre circulación" OR "fuga de cerebros"
UK	English	asyl* OR immigrant* OR immigrat* OR migrant* OR migrat* OR refugee* OR foreigner* OR "undocumented worker*" OR "guest worker*" OR "foreign worker*" OR emigrat* OR "freedom of movement" OR "free movement"
Germany	German	asyl* OR immigrant* OR immigriert* OR immigrat* OR migrant* OR migrat* OR flüchtling* OR ausländer* OR zuwander* OR zugewander* OR einwander* OR eingewander* OR gastarbeiter* OR "ausländische arbeitnehmer*" OR emigr* OR auswander* OR ausgewander* OR personenfreizügigkeit* OR arbeitnehmerfreizügigkeit* OR "freier personenverkehr*"

Heidenreich et al., 2020; Lind et al., 2020

universität wien

Example: Classify actor salience

Objective: salience of women and men migrant's in the news across time and outlet

- Dictionary approach to measure mentions of women and men migrants in German news articles

Lind & Meltzer, 2020

Example: Classify actor salience

Keyword list more complex than initially thought

Recall: .67, Precision: .91

```
person_f_migrant_endIn_regex  = c("[Aa]sylantin","[Aa]sylbewerberin", "[Aa]
                                 "[Zz]uwanderin", "[Ee]inwanderin", "[Gg]a
                                 "[Aa]usländische\\w{0,2}\\s[Bb]ürgerin",

f_relation_regex = c("[Mm]eine\\w{0,2}\\s[Ff]rau(en|)(\\s|\\W)", "[Mm]ein\\
                     "[Dd]eine\\w{0,2}\\s[Ff]rau(en|)(\\s|\\W)", "[Dd]ein\\w{
                     "[Ss]eine\\w{0,2}\\s[Ff]rau(en|)(\\s|\\W)", "[Ss]ein\\w{
                     "[Ii]hr\\w{0,2}\\s[Ff]rau(en|)(\\s|\\W)", "[Ii]hr\\w{0,2
                     "[Uu]ns\\w{0,2}\\s[Ff]rau(en|)(\\s|\\W)", "[Uu]ns(er|re)
                     "[Ee]u(er|re)\\w{0,2}\\s[Ff]rau(en|)(\\s|\\W)", "[Ee]u(e

f_relation_regex <- paste(apos_closed, f_relation_regex, "", sep = "") #add
df_f_relation_regex <- data.frame(f_relation_regex) #as. dataframe
```

Table 2. Dictionary subcategories for the measurement of the concepts: migrant women's and migrant men's salience.

			Measured concepts (examples)	
	Subcategory name	Description	Migrant women's salience	Migrant men's salience
1.	Impersonal designation	General person nominations for migrants	(immigrant, asylum seeker, etc.) with female or gender-neutral word endings (-In,/-in, etc.)	(immigrant, asylum seeker, etc.) with generic masculine form or gender-neutral word endings (-In,/-in, etc.)
2.	Origin: Single words	Nominations that refer to the territorial origin of a person [a]	(French, African, Syrian, etc.) with female or gender-neutral word endings (-In, -/in, etc.)	(French, African, Syrian, etc.) with generic masculine form or gender-neutral word endings (-In, -/in, etc.)
3.	Origin: Combinations of different word groups [b]			
a		General gender-related person nominations + from + general territorial denominations [c]	(women, girl, mother, sister) + from + (France, Africa, Syria, etc.)	(man, boy, father, brother) + from + (France, Africa, Syria, etc.)
b		General territorial denominations (adjectives) [c] + General gender-related person nominations	(French, African, Syrian, etc.) + (women, girl, mother, sister)	(French, African, Syrian) + (man, boy, father, brother)
4.	In relation	Relational expressions: possessive pronouns + General gender-related person nominations	(my, your, her, his, our, yours, their) + (women, girl, mother, sister)	(my, your, her, his, our, yours, their) + (man, boy, father, brother)
5.	As phrase	Phrases	e.g., "women and children"	e.g., "men and children"
6.	Other expressions		e.g., "women from abroad"	e.g., "men from abroad"

As general notes, the dictionary includes the singular and plural version of all words, word endings (e.g., for prepositions) consider the different cases used in the German language.
[a]Downloaded from the CLDR (Unicode Common Locale Data Repository) http://cldr.unicode.org/, which holds standard name translations of countries and regions (version v33.1).
[b]Measured at the sentence level.
[c]Manually compiled by a native speaker for all CLDR territorial denominations, which the German language allows (e.g., no separate word for many smaller islands, e.g., Isle of Man, Curaçao). Assisted by the preeminent German language dictionary duden.de.

Lind, F., & Meltzer, C. E. (2020).

universität wien

Validate, validate, validate

universität wien

Data validation

We ended our class yesterday by sharing resources and experiences about data collection.

What possibilities do we have to evaluate the quality of our data?

universität wien

Data validation

Motivation: source and data selection determines results and conclusions

Are the selected data sources and selected data points representative for my target concept or discourse?

- Are they relevant?
- Are the representative?

Stryker et al. 2006

universität wien

Side note: Data source validation

Validation techniques:

- Rely on expert opinion
- Rely on data source selection of similar research

universität wien

Relevance of search string validation

- Sampling based on search strings popular (Stryker et al. 2016) and recommended (Barberá et al., 2021)
- Reviews of search string validation procedures
  - out of 83 content analyses, 39% stated the search terms they used, and only 6% discussed their validity (Stryker et al. 2016)
  - out of 105 content analysis studies, 73.3% stated the search terms they used, only 12.4% reported validity metrics (Mahl et al., 2022)
- Careless application of non-validated search terms may lead to noisy inferences (Mahl et al., 2022)

universität wien

Search string validation guidelines

(1) formalize the validation process—that is, carefully protocol every step and decision taken prior to and during data collection

(2) define the universe of relevant data beforehand, which includes an explicit definition of relevance criteria that have to be met

(3) triangulate inductive and deductive approaches, such as frequency or keyness analyses, to identify the most indicative terms during search term mining, including a consideration of semantic changes of concepts over time as well as varying national or cultural contexts

(4) carefully validate search terms identified in search term mining by using well-established performance metrics such as precision, recall, and F1-score

Mahl et al., 2022, p.18-19

universität wien

Key validation approach

How close is an automated measurement to a more trusted measurement:

Human understanding of text?

universität wien

Dictionary validation with manually created baseline

Steps

- Code a subset manually and compare manual with automated classification decisions (via recall, precision, F1)
- Iterative dictionary improvement
- Ideally: manual coding and dictionary development is performed by different persons

universität wien

Creation of a manual baseline

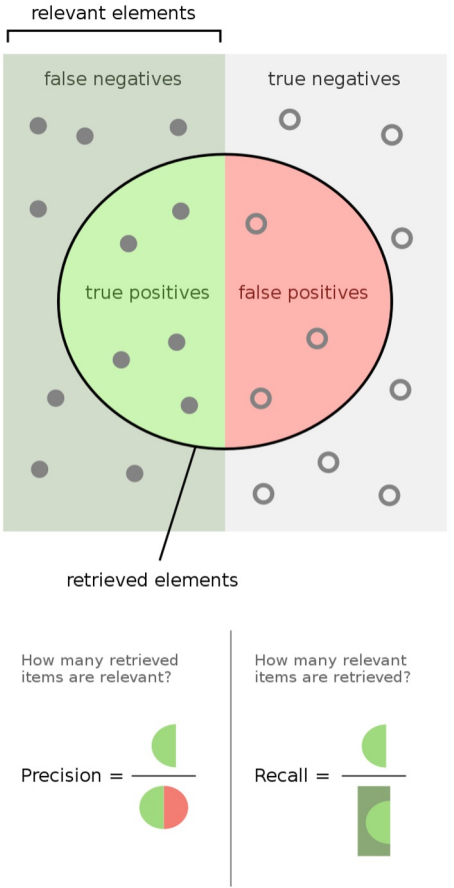
Decisions

- Codebook creation
- Who:
  - Expert coders: Coder recruitment and training sessions
  - Crowdcoders: test questions, majority choice
- Quality assessment: e.g., Inter-coder reliability of involved coders, majority vote
- Documents selected for baseline should be representative for target discourse (e.g., random selection or artificial week)

Recall, precision, F1?

Metrics frequently used to express the validity of a search string & more generally also of automated classification methods

- Precision (P)
- Recall ®(R)
- F1 = 2*(P * R)/R+R)



Scharkow, 2012, 133–36

universität wien

Tools for manual coding

- Google sheets
- AmCAT: https://vu.amcat.nl/accounts/login/?next=/
- AnnoTinder: https://github.com/ccs-amsterdam/CCS_annotator

Example: Baseline creation templates For search strings

Create sampling plan (goal: representative for universe of texts)

Database	Date	Outlet	Number of all articles published that day
APA	Mon 8.10.2018	Standard	136
APA	Tue 9.07.2019	Standard	87
APA	Wed 12.02.2020	Standard	89
APA	Thr 15.04.2021	Standard	98
APA	Fri 27.05.2022	Standard	94
Note: Ideally repeat this procedure for each outlet included; cover the full range of time period investigated			

Collect articles

Code manually

Search with search string

Article id	Date	Text	Manually perceived as relevant (1=yes, 0 = No)	Perceived as relevant by search string (1=yes, 0 = No)
1	Mon 8.10.2018	Kern ist an sich selbst gescheitert. Die SPÖ braucht jetzt mehr Geradlinigkeit und weniger Gockelhaftigkeit …	1	1
2	Mon 8.10.2018	Impressum und Offenlegung: Herausgeber: Oscar Bronner…	0	0
3	Mon 8.10.2018	Einseitiger Vorschlag. Zu viele Waffen in der Hand der Bürger sind gefährlich. Ein Blick in die USA zeigt, warum. Im Kern geht es…	0	1
…	…	…	…	

Calculate recall and precision

universität wien

Dictionary

Pros

- Often needed to select data (search strings)
- High reliability and control
- High transparency and reproducibility

Cons

- Difficulty increases with the latency of the construct

universität wien

Regular Expressions (regex)

universität wien

Regular Expressions (regex)

- How to pronounce?

universität wien

Regular Expressions (regex)

- How to pronounce?
- /gɪf/
- /dʒɪf/

universität wien

Regular Expressions (regex)

- How to pronounce?
- /gɪf/
- /dʒɪf/
- /ɹɛ.gɛks/
- /ɹɛ.dʒɛks/

universität wien

Regular Expressions (regex)

- How to pronounce?
- ***/g□f/***
- /dʒɪf/
- ***/□□.g□ks/***
- /ɹɛ.dʒɛks/

universität wien

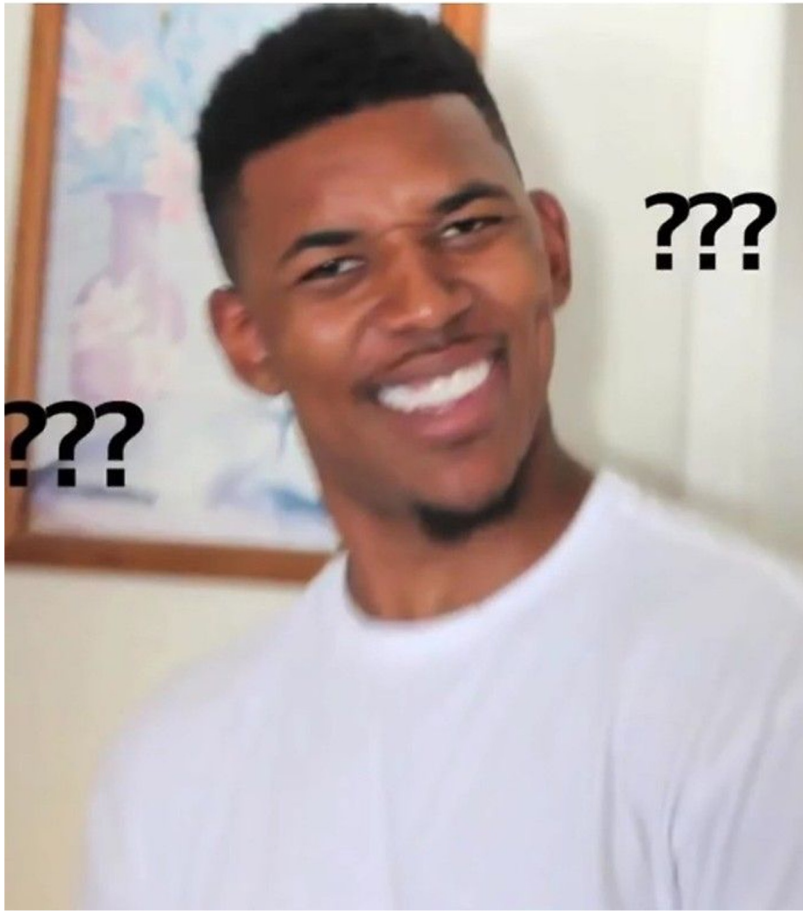
Regular Expressions (regex)

- How to pronounce?
- /gɪf/
- ***/d□□f/***
- /ɹɛ.gɛks/
- ***/□□.d□□ks/***

universität wien

Regular Expressions (regex)

- How to pronounce?
- /gɪf/
- **/d□□f/**
- /ɹɛ.gɛks/
- **/□□.d□□ks/**



universität wien

regex

- Formal language to specify search strings

universität wien

regex

- Formal language to specify search strings
- Insanely difficult

universität wien

regex

- Formal language to specify search strings
- INSANELY difficult

universität wien

regex

- Formal language to specify search strings
- ***INSANELY*** difficult

universität wien

regex

- Formal language to specify search strings
- ***INSANELY*** difficult
- Nobody can remember anything

universität wien

regex

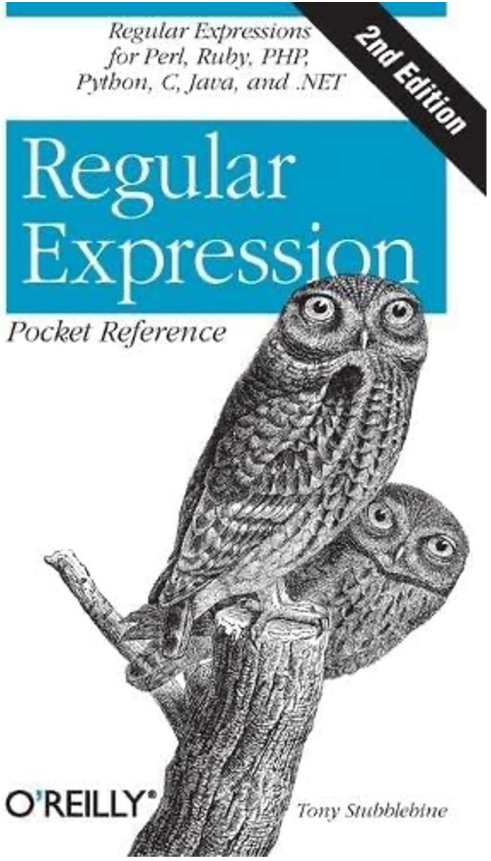
- Formal language to specify search strings
- ***INSANELY*** difficult
- Nobody can remember anything
- Different ***flavours***

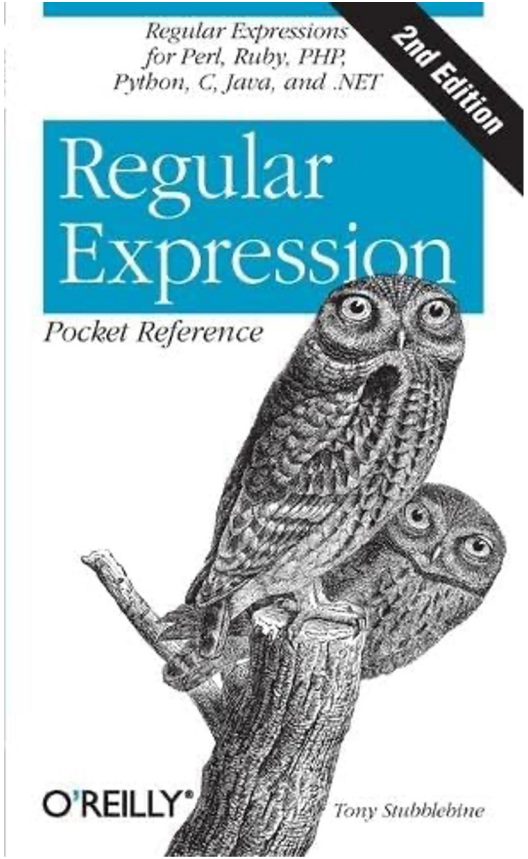
universität wien

regex

- Formal language to specify search strings
- ***INSANELY*** difficult
- Nobody can remember anything
- Different ***flavours***

- “Some people, when confronted with a problem, think “I know, I’ll use regular expressions.” Now they have two problems.” *Jamie Zawinski*


Regular Expressions
for Perl, Ruby, PHP,
Python, C, Java, and .NET
2nd Edition
Regular
Expression
Pocket Reference
O'REILLY
Tony Stubblebine


Regular Expressions
for Perl, Ruby, PHP,
Python, C, Java, and .NET
2nd Edition
Regular
Expression
Pocket Reference
O'REILLY
Tony Stubblebine

128 pages!!!

universität wien

Grimmer / Jurafsky
Cheat-sheet

- Disjunctions

RE	Match	Example Patterns Matched
`[mM]oney`	Money or money	"Money"
`[abc]`	'a', 'b', *or* 'c'	"Investing in Iran"
		"is dangerous business"
`[1234567890]`	any digit	"sitting on $7.5 billion dollars"
		"2005 and 2006, more than "
		"$150 million dollars"
`[\.]`	A period	" 'Run!', he screamed."

- Ranges

RE	Match	Example Patterns Matched
`[A-Z]`	an upper case letter	"Rep. Anthony Weiner
		(D-Brooklyn & Queens)"
`[a-z]`	a lower case letter	"ACORN's"
`[0-9]`	a single digit	"(9th CD) "

universität wien

Grimmer / Jurafsky
Cheat-sheet

- Negations

RE	Match	Example Patterns Matched
`[^A-Z]`	not an upper case letter	"ACORN**'s**"
`[^Ss]`	neither 'S' nor 's'	"**ACORN'**s"
`[^\.]`	not a period	" 'Run!', he screamed."

- Optional Characters: `?`, `*`, `+`

RE	Match	Example Patterns Matched
`colou?r`	Words with `u` 0 or 1 times	"color" or "colour "
`oo*h!`	Words with `o` 0 or more times	"oh!" or "ooh!" or "oooh!"
`o+h!`	Words with `o` 1 or more times	"oh!" or "ooh!" or "oooooh!" or

universität wien

Grimmer / Jurafsky
Cheat-sheet

- Start of the line anchor ^, end of the line anchor **$**

RE	Match	Example Patterns Matched
`^[A-Z]`	Upper case start of line	"Palo Alto" "the town of **P**alo Alto"
`^[^A-Z]`	Not upper case start of line	"the town of Palo Alto" "**P**alo Alto"
`^.`	Start of line	"Palo Alto" "the town of Palo Alto"
`.$`	Identify character that ends a line	"Wait**!**" "This is the end**.**"

- "Or"| statements, Useful short hand

RE	Match	Example Patterns Matched
yours|mine	Matches"yours" or "mine"	"it's either yours or mine"
\ d	Any digit	"1-Mississippi"
\ D	Any non-digit	"1-Mississippi"
\ s	Any whitespace character	"1, 2"
\ S	Any non-whitespace character	"1, 2"
\ w	Any alpha-numeric	"1-Mississippi "
\ W	Any non-alpha numeric	"1-Mississippi"

universität wien

How difficult to regex an email

universität wien

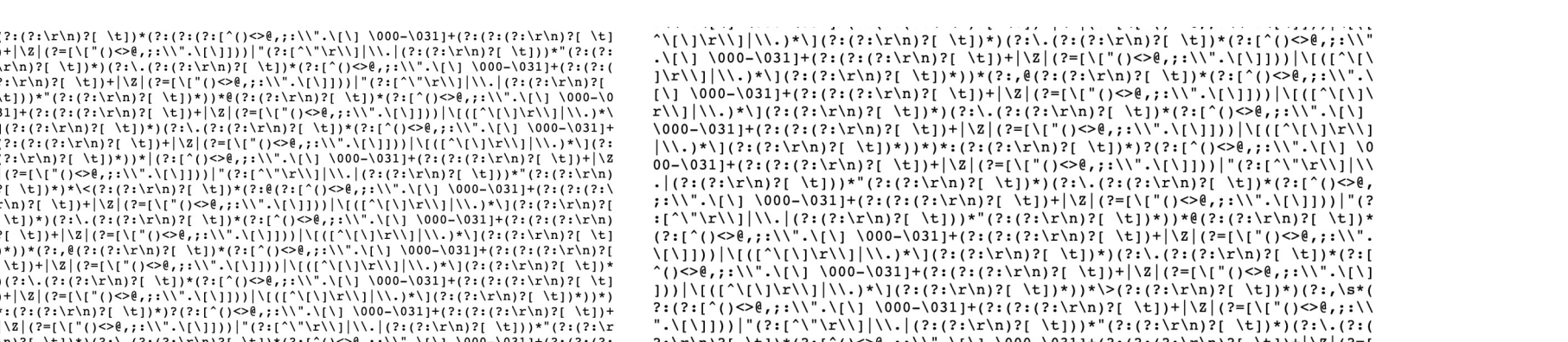
How difficult to regex an email

Rather

universität wien

How difficult to regex an email

```
(?:[a-z0-9!#$%&'*+/=?^_`{|}~-]+(?:\.[a-z0-9!#$%&'*+/=?^_`{|}~-]+)*|"(?:[\x01-
\x08\x0b\x0c\x0e-\x1f\x21\x23-\x5b\x5d-\x7f]|\\[\x01-\x09\x0b\x0c\x0e-\x7f])*")@(?:
(?:[a-z0-9](?:[a-z0-9-]*[a-z0-9])?\.)+[a-z0-9](?:[a-z0-9-]*[a-z0-9])?|\[(?:(?:(2(5[0-
5]|[0-4][0-9])|1[0-9][0-9]|[1-9]?[0-9]))\.){3}(?:(2(5[0-5]|[0-4][0-9])|1[0-9][0-9]|[1-
9]?[0-9])|[a-z0-9-]*[a-z0-9]:(?:[\x01-\x08\x0b\x0c\x0e-\x1f\x21-\x5a\x53-\x7f]|\\
[\x01-\x09\x0b\x0c\x0e-\x7f])+)\])
```



universität wien

Helpers

https://regex101.com

Recommended R tutorial (Wickham & Grolemund, 2017):
https://r4ds.had.co.nz/strings.html

To test regular expressions quickly:
https://spannbaueradam.shinyapps.io/r_regex_tester/

universität wien

Exercise in R

RStudio

- Dictionary creation: Visibility of political actors in news headlines
- Calculate recall, precision, F1, Krippendorff's alpha

universität wien

Supervised classification

Day 3 Session 2

universität wien

Types of machine learning

1. Supervised
    - An outcome variable is defined
    - Focus is on prediction
2. Unsupervised
    - No outcome variable has been defined
    - Focus is on patterns

universität wien

Types of machine learning

1. Supervised
    - An outcome variable is defined
    - Focus is on prediction
2. Unsupervised
    - No outcome variable has been defined
    - Focus is on patterns

universität wien

Supervised machine learning

Objective:

Classification of documents into pre existing categories

universität wien

Steps

1. Create a labeled data set
2. Classify documents with supervised learning algorithm
3. Check performance
4. Using the measures

Grimmer et al., 2022

universität wien

Create a labeled data set

How:

- Human coders annotate parts of the corpus (see slides in session 1 of today)
- Found data (e.g., self-reported profession in users' profile)

Considerations:

- Sampling should be representative for the corpus (e.g., Random, Stratified sample e.g., across time and source)
- Quality of human coding matters (Assess the intercoder reliability)
- Number of documents

Grimmer et al., 2022, Barberá

Create a labeled data set

Number of documents

- the higher the number of categories and the lower the reliability of the coders, the higher the number of documents (Barberá et al., 2021)
- increase the sizes of manually coded validation dataset as large as possible, preferably to more than N = 1,300 (i.e., more than 1% of all data to be examined) assuming acceptable reliability (equal to or higher than .7) (Song et al., 2021)

Table 2. Simulation input parameters.

Factors	Input Parameters
N of human coders	2 (minimum), 5 (intermediate), & 10 (large manual coding)
Intercoder reliability	0.5 (low), 0.7 (acceptable), & 0.9 (high levels of reliability)
N of validation data	600 (0.5%), 1300 (1%), 6500 (5%), & 13000 (10%) of total data
Sampling variability	Random sample vs. nonrandom (biased) subset for validation
Coding per entry	Sole coding vs. duplicated coding for each entry

Song et al., 2021, p. 555

universität wien

Create a labeled data set

Split labeled data in training data and test data

Training data

- The subset that is used to learn the model parameters

Test data

- Another subset used to evaluate the model's predictive quality
- Not used for learning!

universität wien

Classify documents with supervised learning

Classifier learns the mapping between features and the labels in the training set

- We define a model $f(Y)=g(X)$
- And apply a learning algorithm to establish which features in X (features extracted from the training documents) matter to recover Y (i.e, the labels of the training documents)
- We fit the model

Grimmer et al., 2022

universität wien

Classify documents with supervised learning

Considerations:

- Feature representation (Bag of words representation or embeddings)
- Feature selection (remove irrelevant features)
- Classifier selection
  - E.g., Naive Bayes, SVM, KNN, or ensemble methods

Grimmer et al., 2022

universität wien

Check performance

The fitted model (the trained classifier) is applied to a held-out test set (which is a part of the labeled set but was not used for training the model).

Considerations:

- Danger of overfitting (focus on features that work well with training set but do not generalize)
  - Solutions: cross-validation
- Performance metric (i.e., recall, precision)

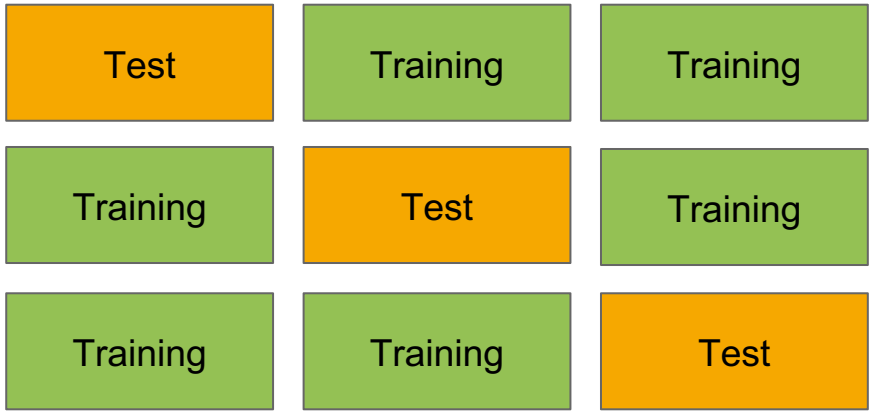
Grimmer et al., 2022

universität wien

Check performance

k-fold cross-validation

- We randomly split the data into *k* sets (“folds”) of roughly equal size
- Each set is hold out once as test set, while training on the remaining sets
- The problem of a lucky split is reduced



universität wien

Check performance

Performance metrics

$$Accuracy = \frac{True\ Negative + True\ Positive}{True\ Negative + True\ Positive + False\ Negative + True\ Positive}$$

$$Precision_{positive} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$Recall_{positive} = \frac{True\ Positive}{True\ Positive + False\ Negatives}$$

Confusion matrix

	Actual label	
Classification (algorithm)	Negative	Positive
Negative	True negative	False negative
Positive	False negative	True positive

Grimmer et al., 2022

universität wien

Check performance

Check convergent validity (Adcock & Collier, 2001)

- Compare the measures with other established measures, e.g.,
  - Use trained model to classify texts (open-ended answers relationship uncertainty as described by participants of an online survey) and compare it with the self-assessment in response to a closed question, see Pilny et al. 2019

universität wien

Using the measures

The classifier is applied to all documents in the corpus

Considerations

- Dealing with the measurement error which occurs through error in the human coding, errors in the classifier

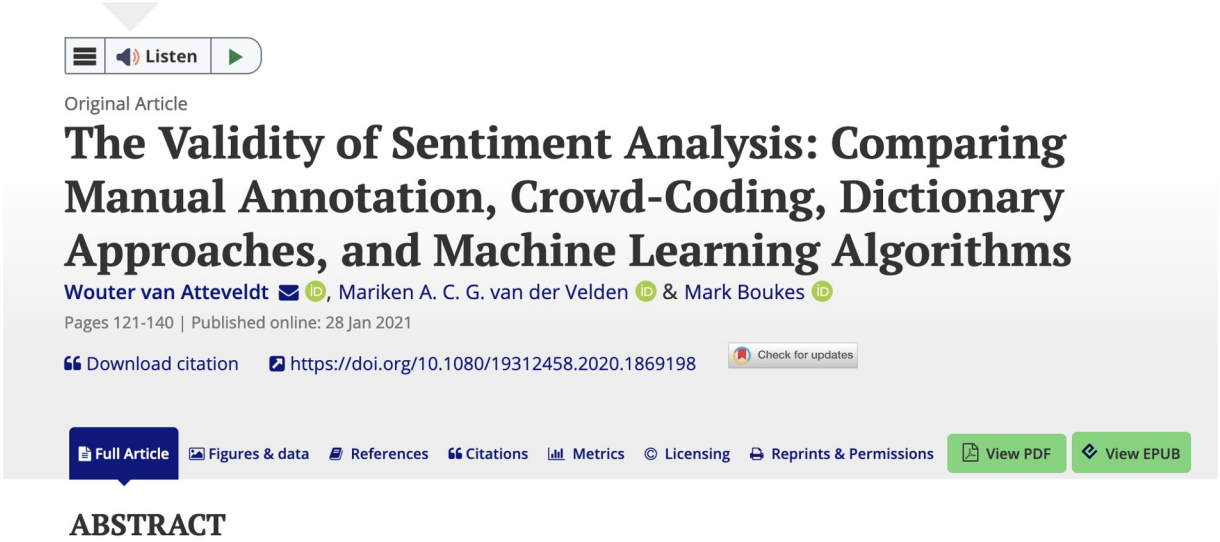
  (assess: if error is systematic or not, if consequential for the analysis)

Grimmer et al., 2022

Dictionary vs. supervised machine learning

Category: sentiment

Result: machine learning significantly outperformed dictionary coding



Van Atteveldt el al. (2021)

universität wien

Dictionary vs. supervised machine learning

- Dictionaries can be applied directly to a new corpus (but validate!)
- Supervised machine learning requires (potentially larger amounts) labeled data
- If the training sample is large enough supervised learning will outperform dictionaries

González-Bailón & Paltoglou (2015)

universität wien

Additional considerations

- Hyperparameter selection
  - Via systematic comparison of different hyperparameters per algorithm
- Random undersampling (Galar et al., 2011)
  - Method to deal with unbalanced classes: use the max. number of positive instances per class and randomly sample the same number of instances of the negative class

universität wien

Naïve Bayes Classifier

- Probabilistic classifier
- Simple
- Fast
- Good Accuracy

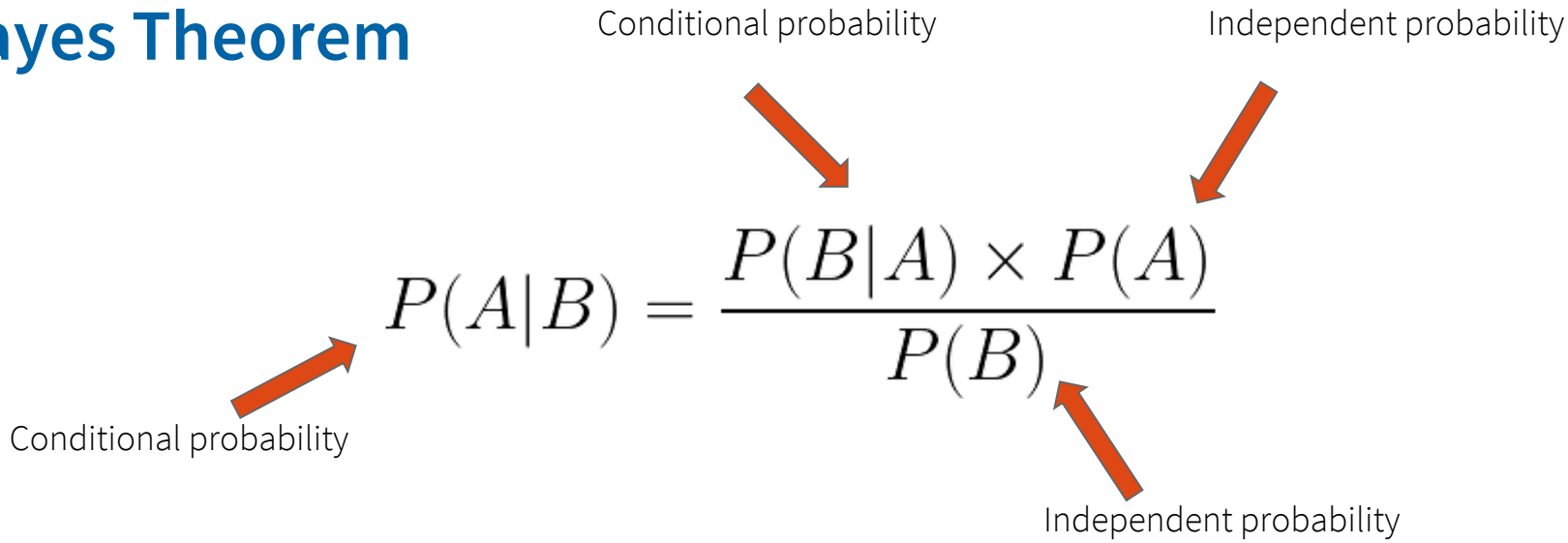
universität wien

Bayes Theorem

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

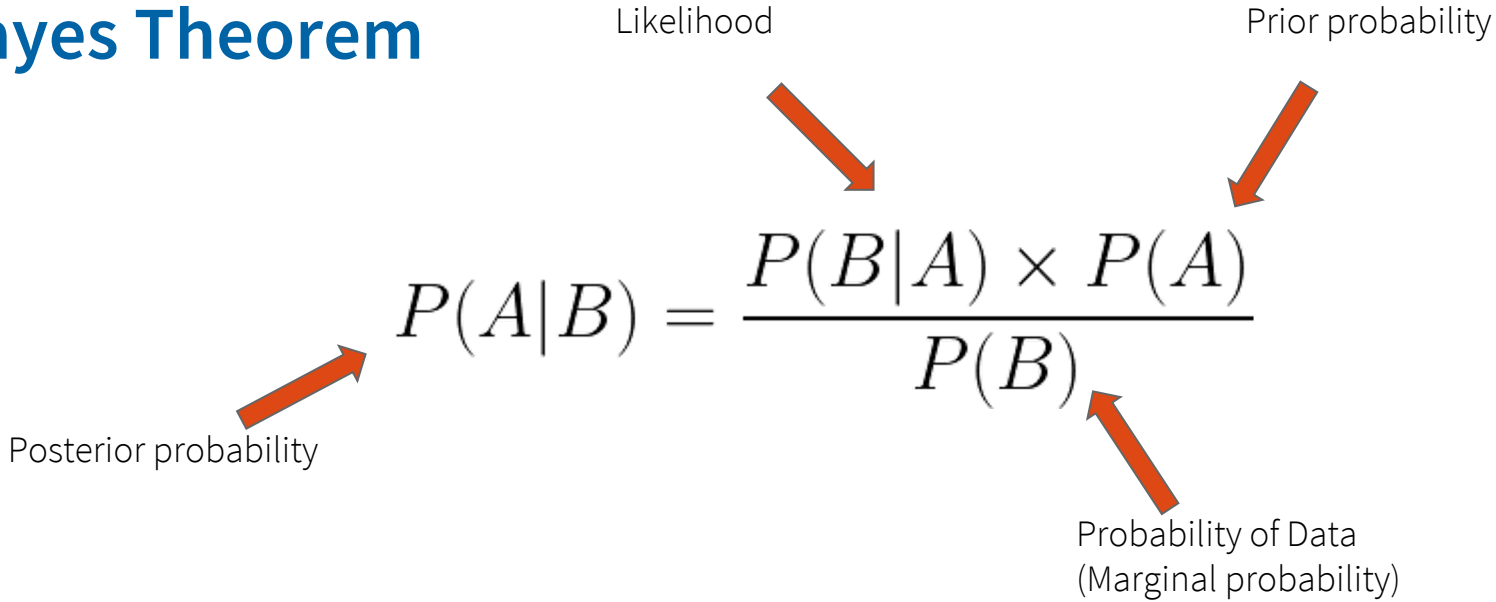
universität wien

Bayes Theorem



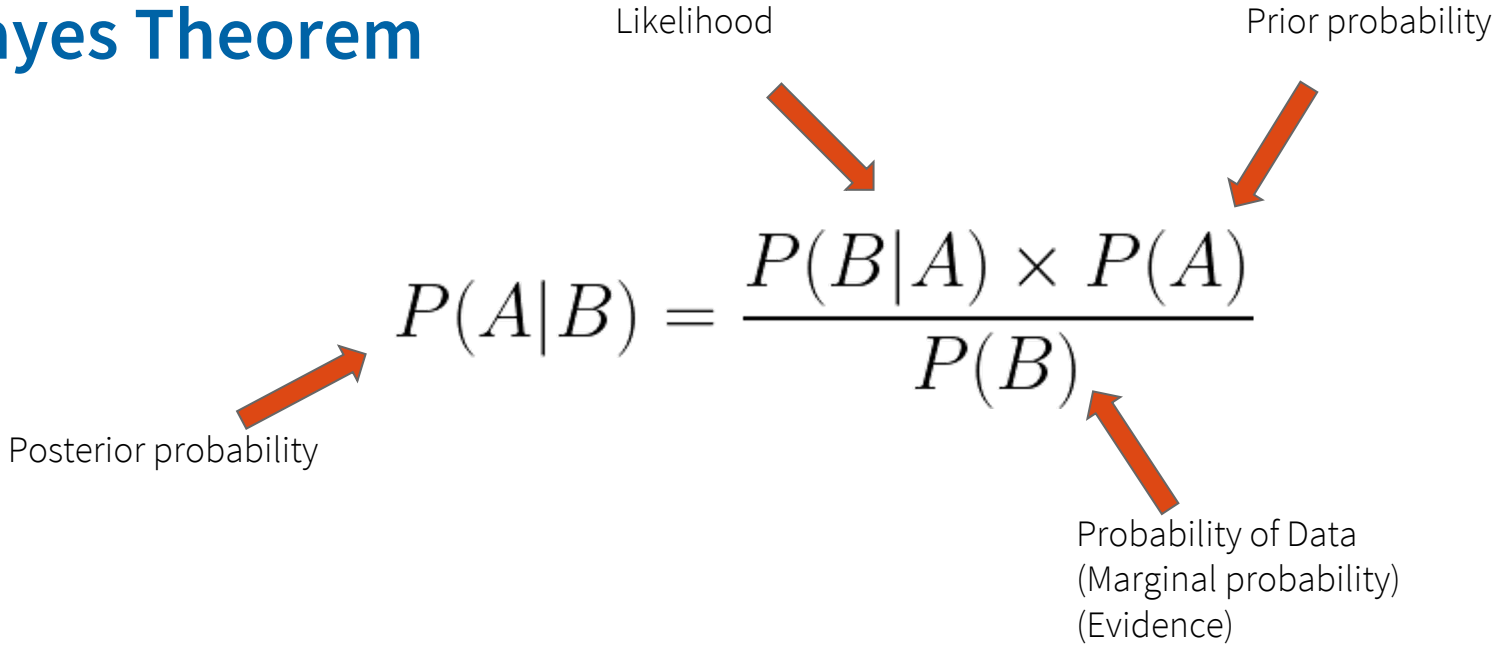
universität wien

Bayes Theorem



universität wien

Bayes Theorem



universität wien

Bayes Theorem

$$P(A|B) \propto P(B|A) \times P(A)$$

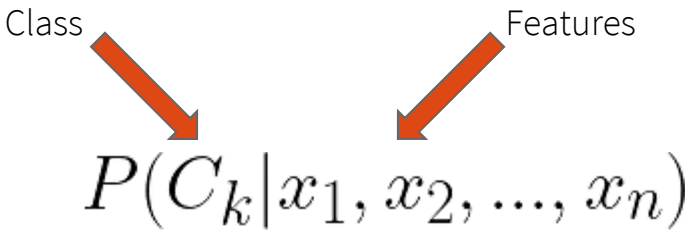
universität wien

Naïve Bayes Classifier

$$P(C_k|x_1, x_2, ..., x_n)$$

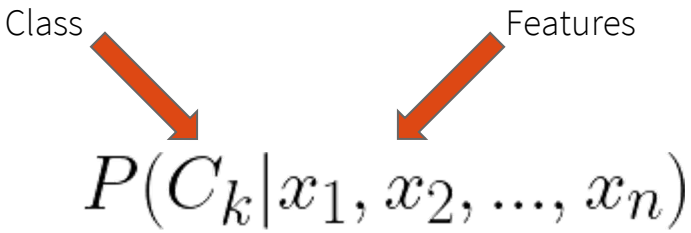
universität wien

Naïve Bayes Classifier



universität wien

Naïve Bayes Classifier



Features are assumed to be independent. Hence, "***Naïve***"

Naïve Bayes Classifier

$$P(C_k|\mathbf{x}) = \frac{P(C_k) \times P(\mathbf{x}|C_k)}{P(\mathbf{x})}$$

universität wien

Naïve Bayes Classifier

$$P(C_k|\mathbf{x}) \propto P(C_k) \times P(\mathbf{x}|C_k)$$

Naïve Bayes Classifier

$$
\begin{aligned}
p(C_k \mid x_1, \ldots, x_n) &\propto p(C_k, x_1, \ldots, x_n) \\
&\propto p(C_k)\, p(x_1 \mid C_k)\, p(x_2 \mid C_k)\, p(x_3 \mid C_k)\, \cdots \\
&\propto p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k)\,,
\end{aligned}
$$

Decision Rule

$$\hat{y} = argmax\ p(C_k) \prod_{i=1}^{n} p(x_i|C_k)$$

universität wien

Implemented in many stats/ML packages

universität wien

universität wien

Support Vector Machine

- Comes from computer science
- Very good
- Rather difficult math

- Considered one of the best of-the-shelf classification algorithms

universität wien

Hyperplane

- n-1 dimensional plane that separates the n-dimensional space

universität wien

Hyperplane

- n-1 dimensional plane that separates the n-dimensional space
- 2-dimensional hyperplane:

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 = 0$$

universität wien

Hyperplane

- n-1 dimensional plane that separates the n-dimensional space
- 2-dimensional hyperplane:
- Line equation

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 = 0$$

Hyperplane

- n-1 dimensional plane that separates the n-dimensional space
- 2-dimensional hyperplane:
- Line equation

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 = 0$$

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p = 0$$

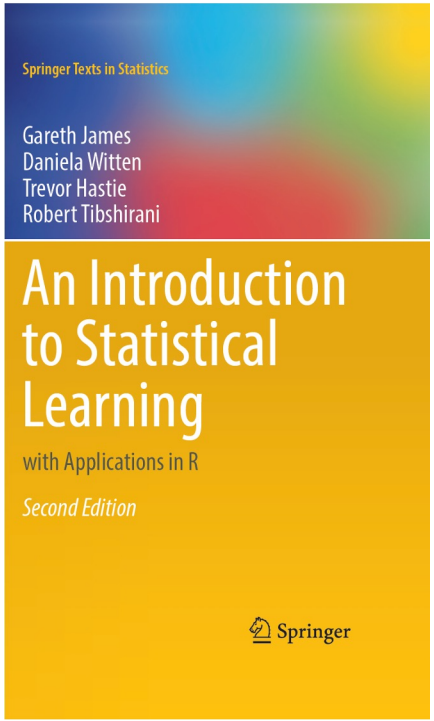
Classification

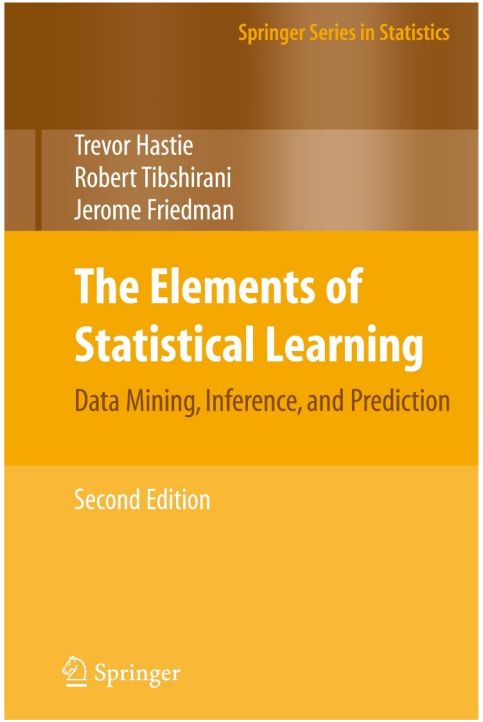
$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p > 0$$

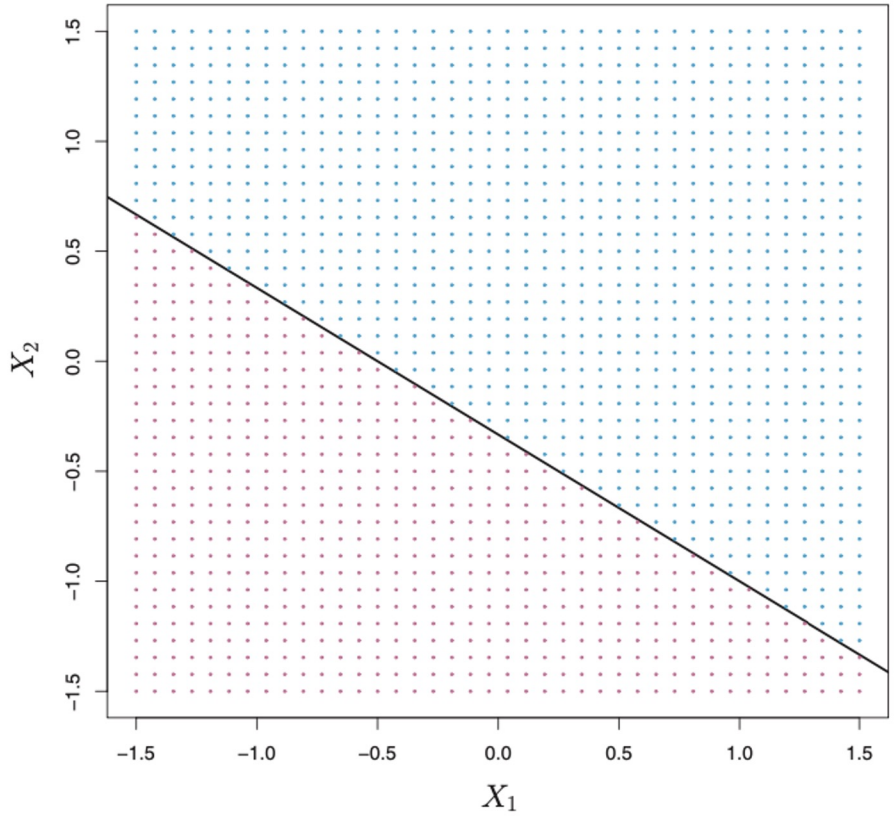
$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p < 0$$

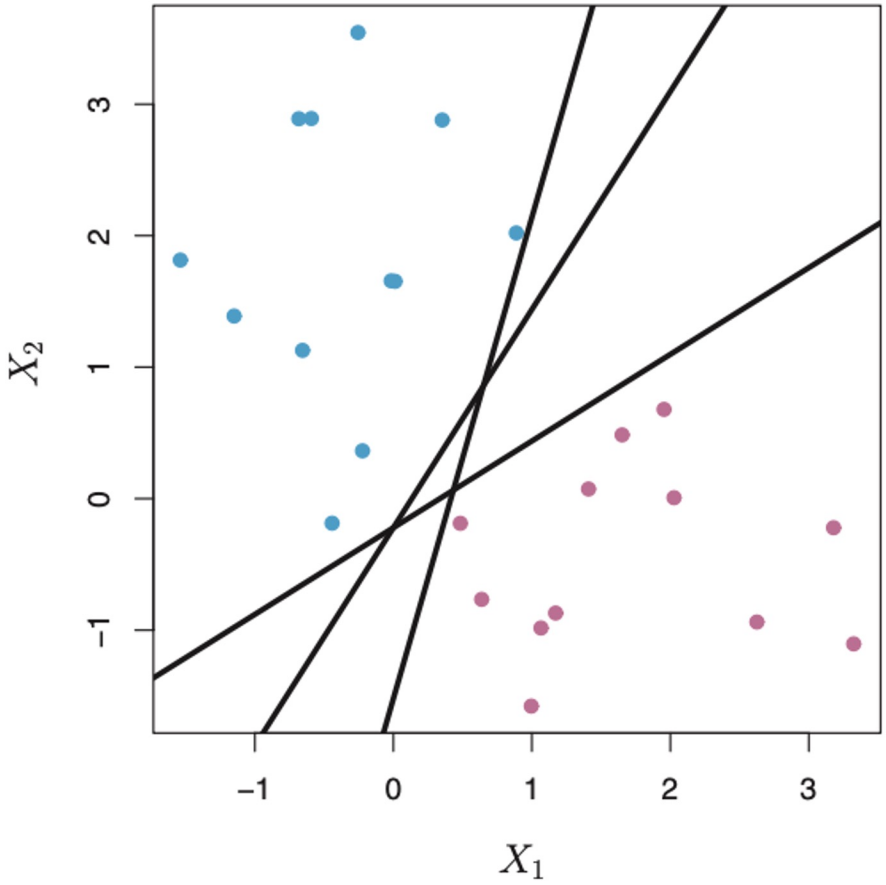
universität wien

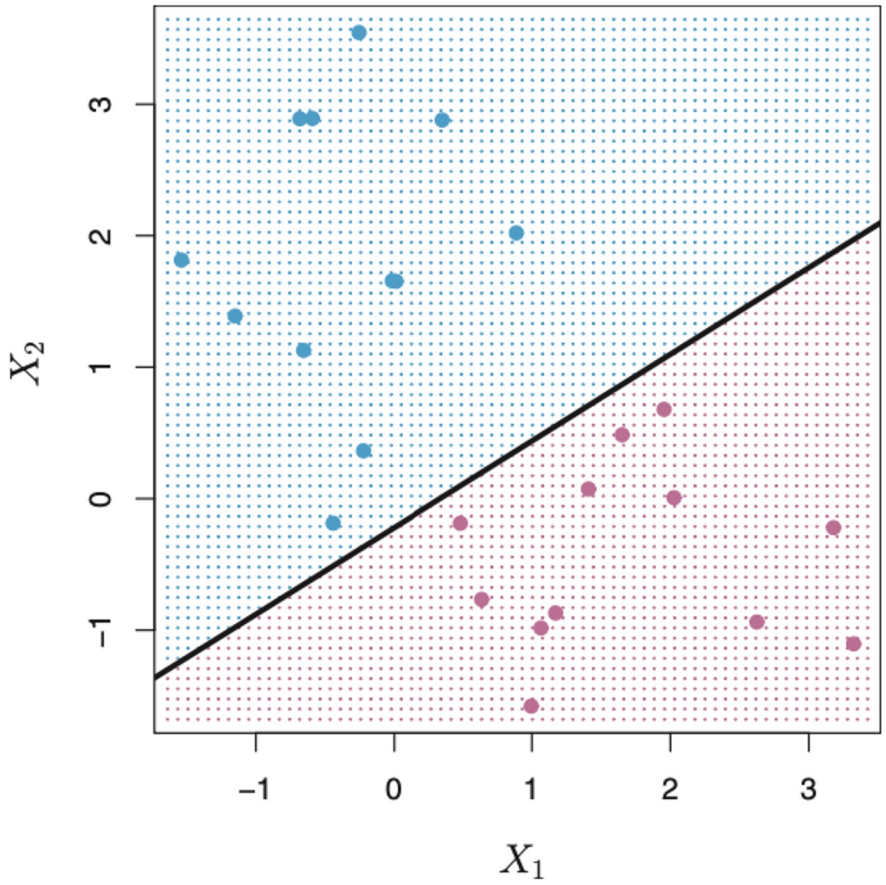
Following images from:





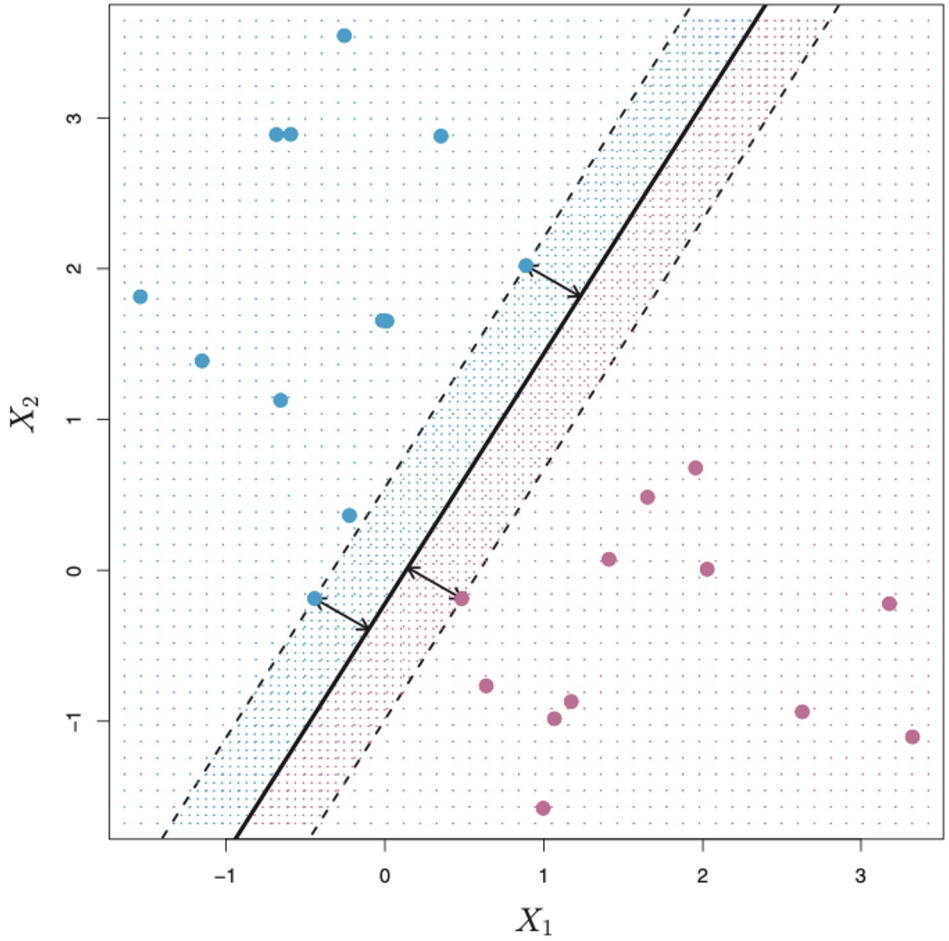

1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
X_2
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
X_1


X_1
X_2
−1
0
1
2
3


X_1
X_2
−1
0
1
2
3

universität wien

SV Classifier



universität wien

Support Vector Machine

- Non-linear version of the Support Vector Classifier
- Extension using ***Kernels***

$$f(x) = \beta_0 + \sum_{i \in \mathcal{S}} \alpha_i \langle x, x_i \rangle$$

$$f(x) = \beta_0 + \sum_{i \in \mathcal{S}} \alpha_i K(x, x_i)$$

$$f(x) = \beta_0 + \sum_{i \in \mathcal{S}} \alpha_i \langle x, x_i \rangle$$

Kernel function

$$f(x) = \beta_0 + \sum_{i \in \mathcal{S}} \alpha_i K(x, x_i)$$

universität wien

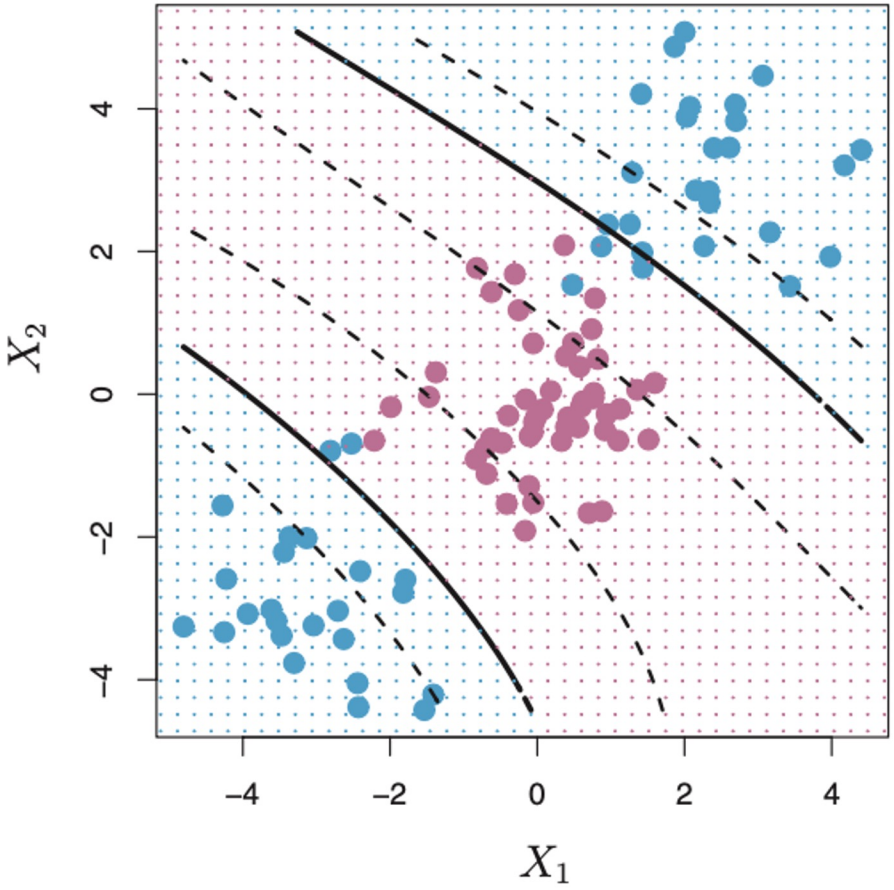
$$f(x) = \beta_0 + \sum_{i \in \mathcal{S}} \alpha_i \langle x, x_i \rangle$$

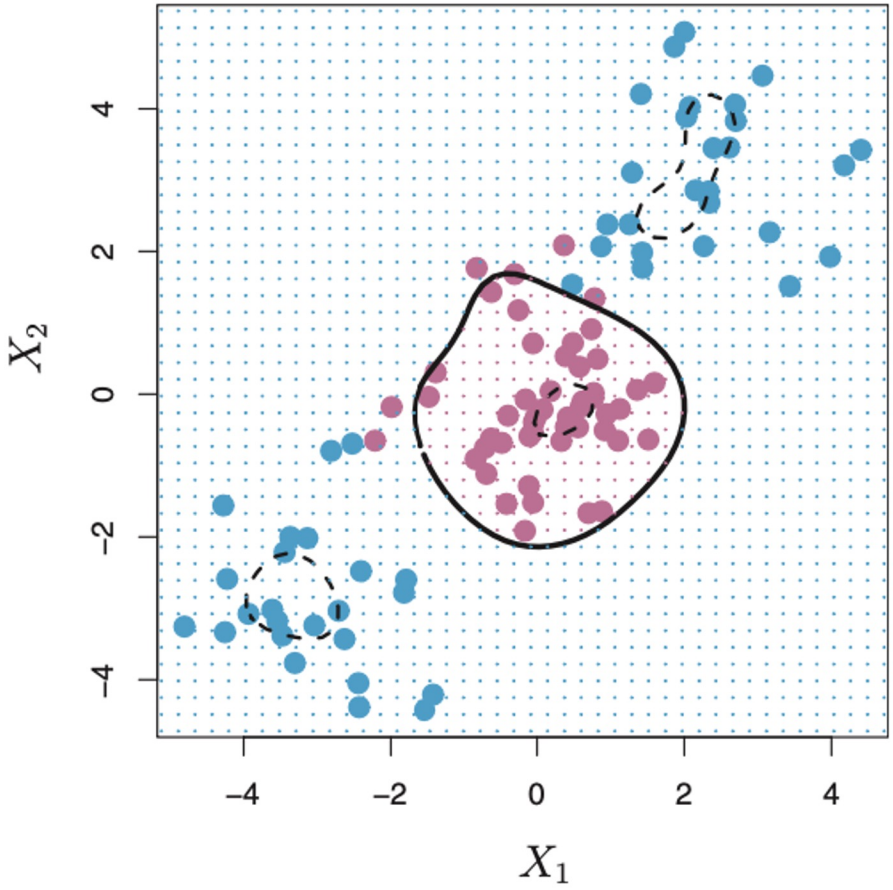
Kernel function

$$f(x) = \beta_0 + \sum_{i \in \mathcal{S}} \alpha_i K(x, x_i)$$

Polynomial Kernel
Non-linear

$$K(x_i, x_{i'}) = (1 + \sum_{j=1}^{p} x_{ij} x_{i'j})^d$$


X_2
4
2
0
−2
−4
−4
−2
0
2
4
X_1


X_2
4
2
0
−2
−4
−4
−2
0
2
4
X_1

universität wien

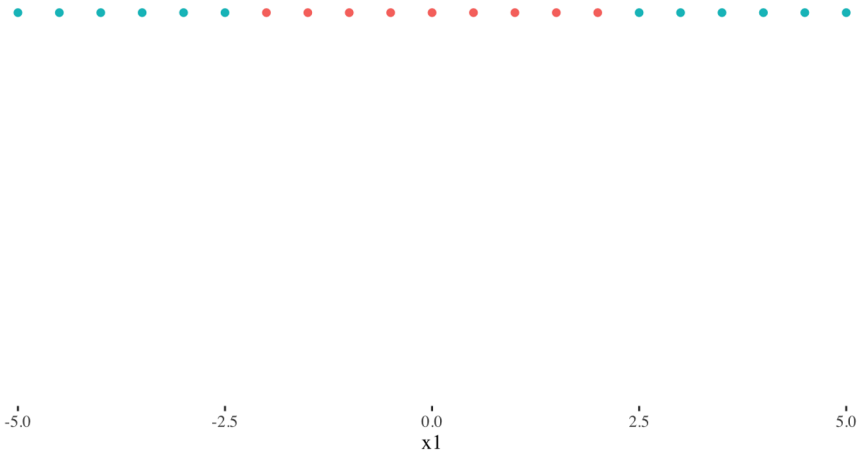
Kernel Trick

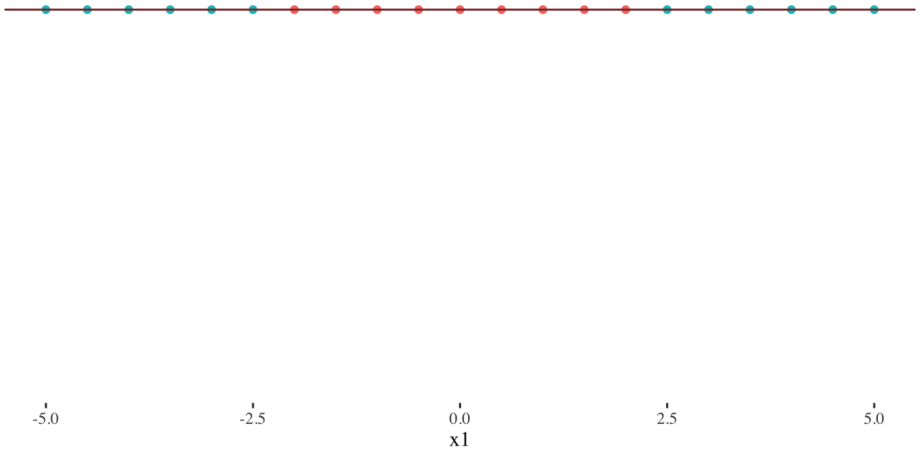
- ***Actual name***

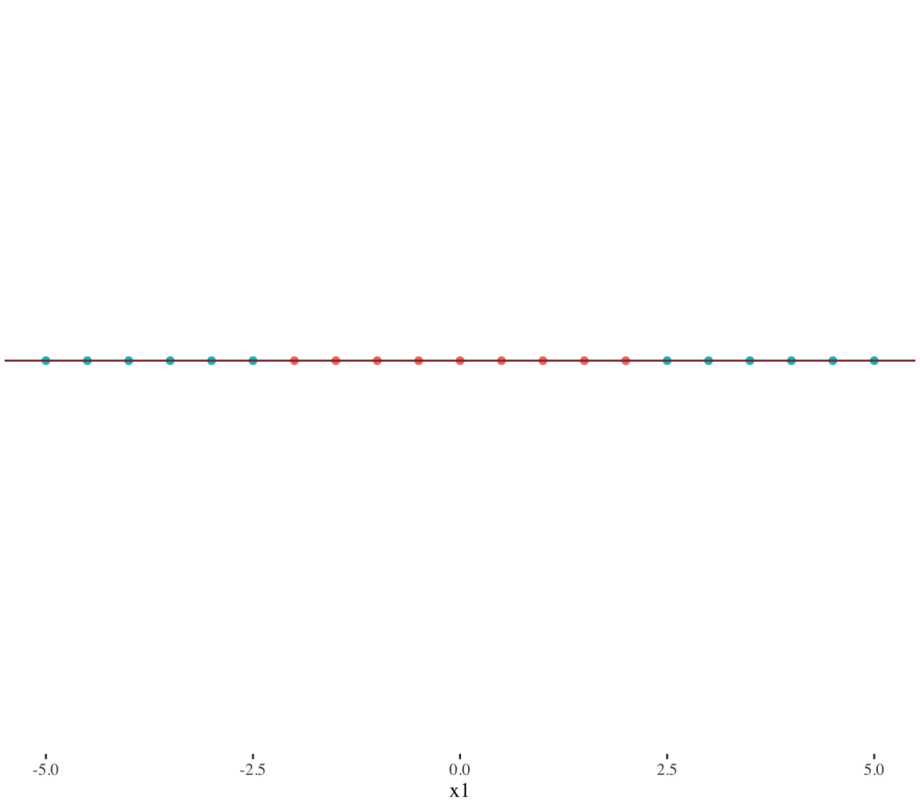
universität wien

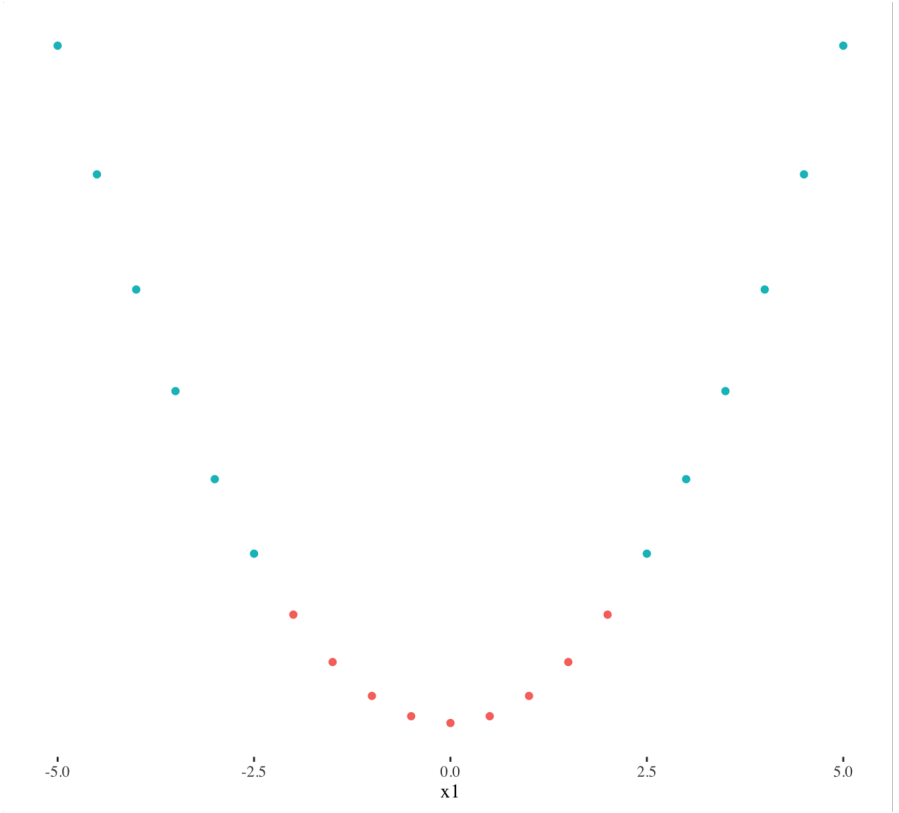
Kernel Trick

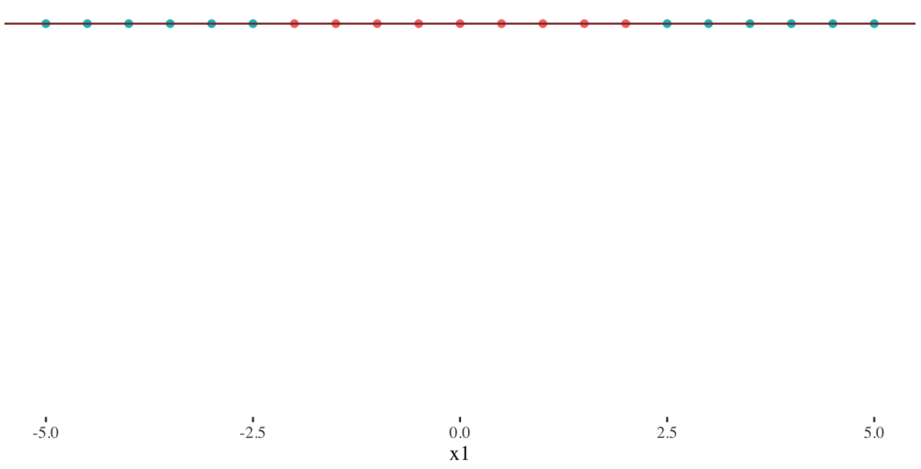
- ***Actual name***
- Attempt to place n-dimensional data into n+1 dimensional space

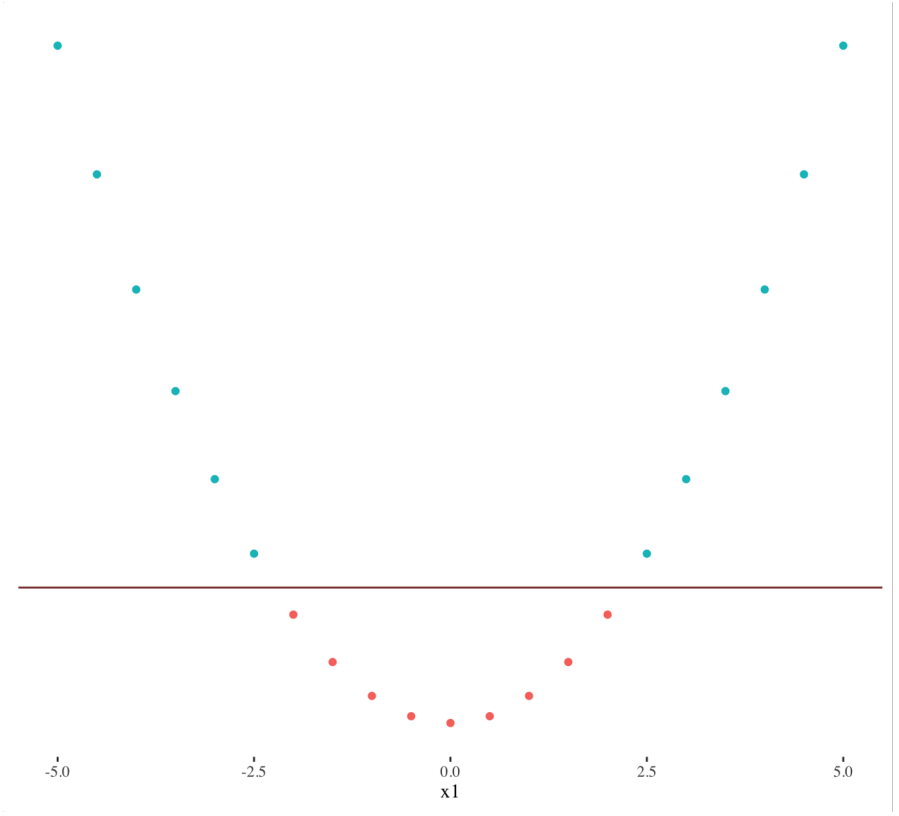

-5.0
-2.5
0.0
2.5
5.0
x1


-5.0
-2.5
0.0
2.5
5.0
x1


-5.0
-2.5
0.0
2.5
5.0
x1


-5.0
-2.5
0.0
2.5
5.0
x1


-5.0
-2.5
0.0
2.5
5.0
x1


-5.0
-2.5
0.0
2.5
5.0
x1