

Quantitative Text Analysis

Meeting 5

Petro Tolochko

Dictionaries and Concepts

Dictionaries

Dictionaries

- Rule-based method
- List of words (or phrases) that indicate a category
- Create your own or use/edit existing dictionaries

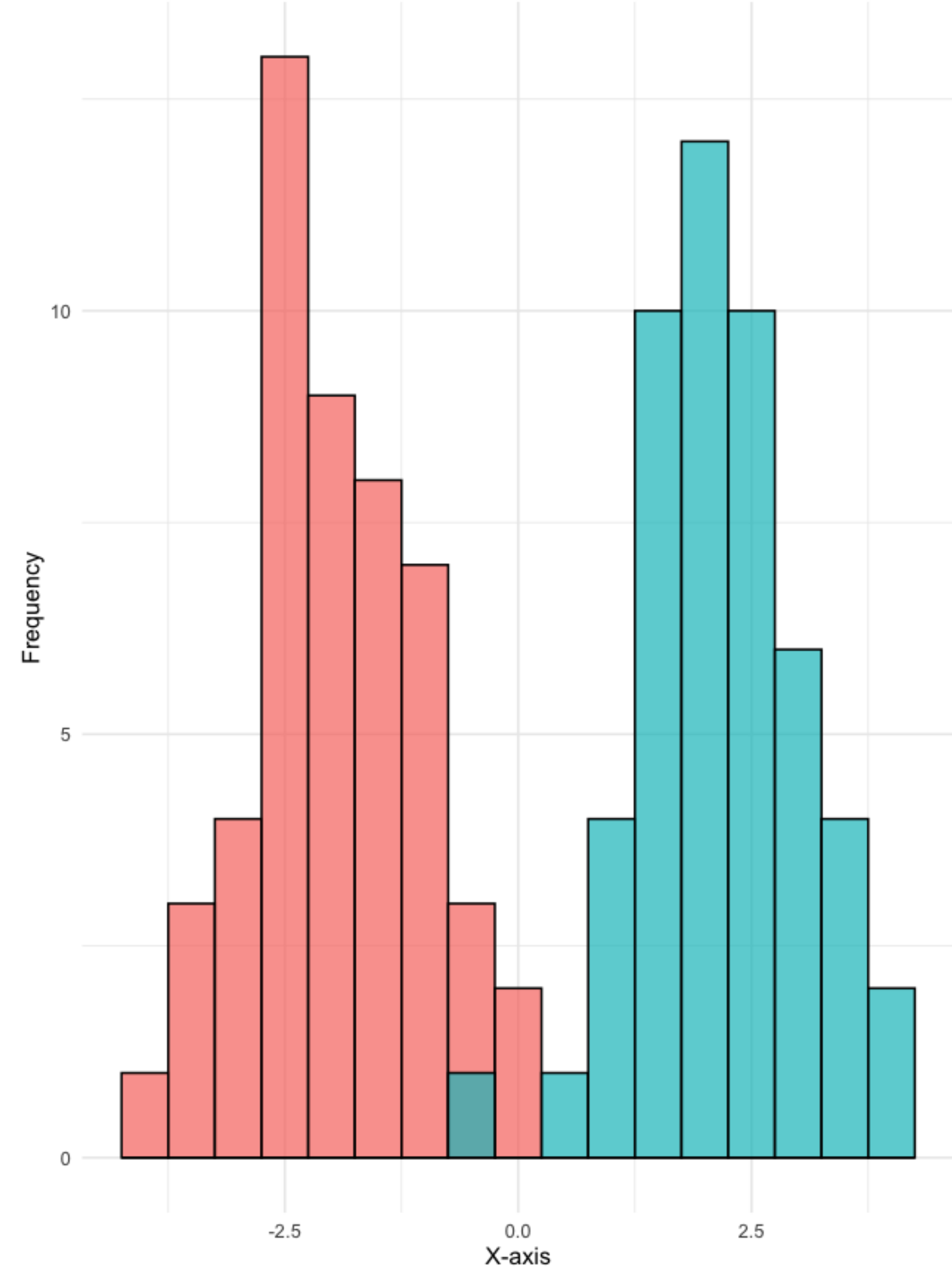
Use cases for dictionaries

Use cases for dictionaries

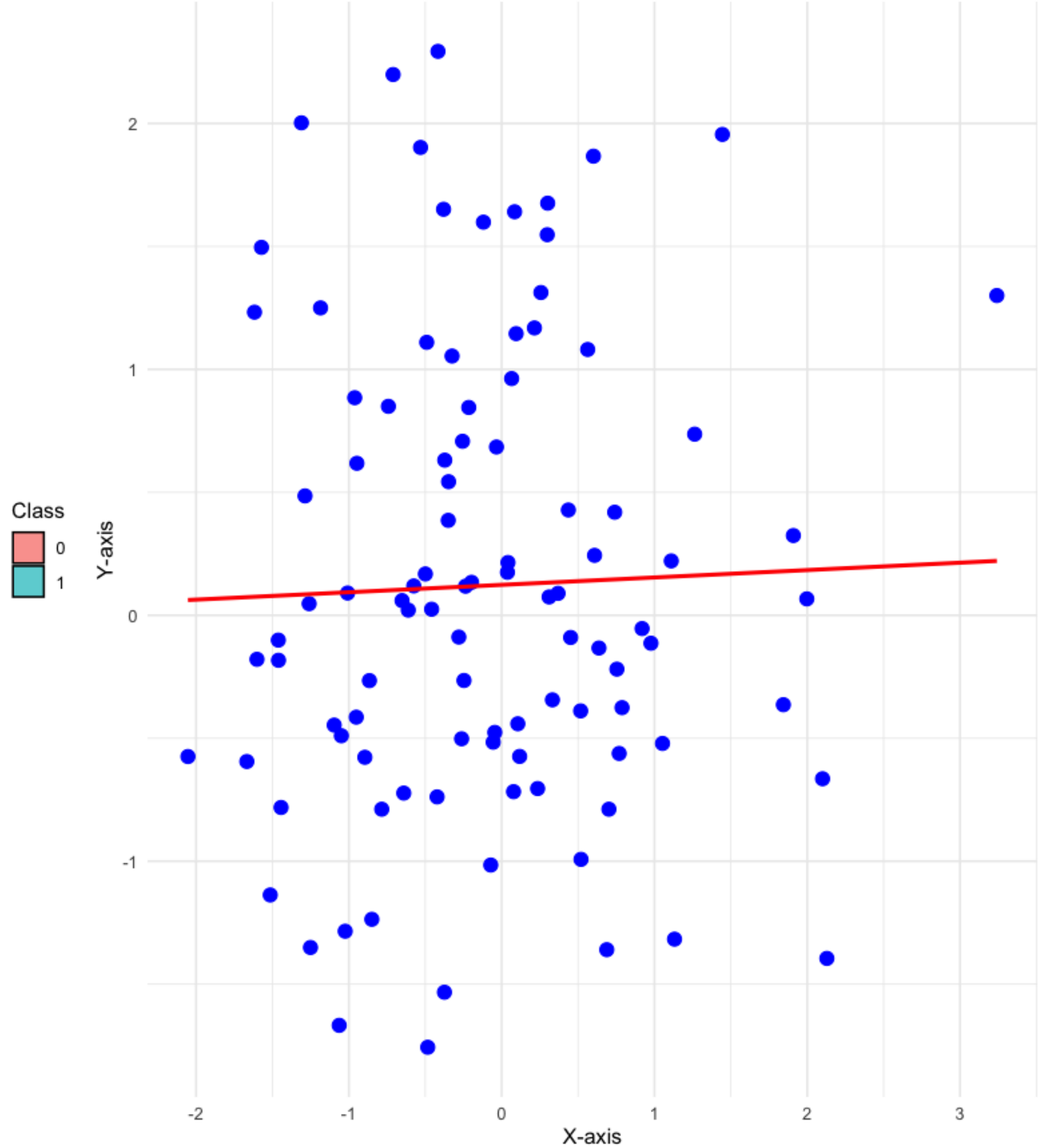
- Classification
- Regression
- Search string (form of classification)

Classification vs. Regression

Classification



Regression



Classification Examples

- Classify texts into:
 - positive/negative
 - populist/non-populist
 - related/unrelated to a certain category

Regression Examples

- “Score” texts on some dimension:
 - Positive/negative
 - Emotional content (anger, sadness, etc.)

Existing vs. Own Dictionaries

Existing vs. Own Dictionaries

- Many existing and validated dictionaries:

Name	Type	Country	Query	Total results: 14
ConText Diesner, J et al. (2020)	Tool		Free text search	
DDR Garten, Justin et al. (2017)	Tool			
DICTION Roderick P. Hart (1996)	Tool		Entity Type	
LIWC Pennebaker, J. W. et al. (1999)	Tool		Tool ×	
NLTK NLTK Team (2001)	Tool		Countries <input type="radio"/> and <input type="radio"/> or	
Netlytic Gruzd, A. (2016)	Tool			
T-LAB T-LAB di Lancia Franco	Tool		Channel	
WordStat Provalis Research	Tool			
corpustools Welbers K et al. (2018)	Tool		Languages <input type="radio"/> and <input type="radio"/> or	
iLCM Andreas Niekler et al. (2018)	Tool			
popdictR Gründl, Johann (2020)	Tool		Used For <input type="radio"/> and <input type="radio"/> or	
quanteda Benoit, Kenneth et al. (2018)	Tool		Dictionary Analysis ×	
tidytext De Queiroz, Gabriela et al. (2016)	Tool		Concept Variables <input type="radio"/> and <input type="radio"/> or	
tm Feinerer, Ingo et al. (2008)	Tool			
			Programming Languages <input type="radio"/> and <input type="radio"/> or	

Existing vs. Own Dictionaries

- Many existing and validated dictionaries:
- Many instances of creating ad-hoc dictionaries:

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"

Validation in Dictionary Analysis

Validation in Dictionary Analysis

- Source and data selection determines results and conclusions
- Are the selected data sources and selected data points representative for your target concept or discourse?
 - Relevant?
 - Representative?

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)

Key validation approach

- How close is an automated measurement to a more trusted measurement:
 - Human understanding of text

Dictionary validation with manually created baseline

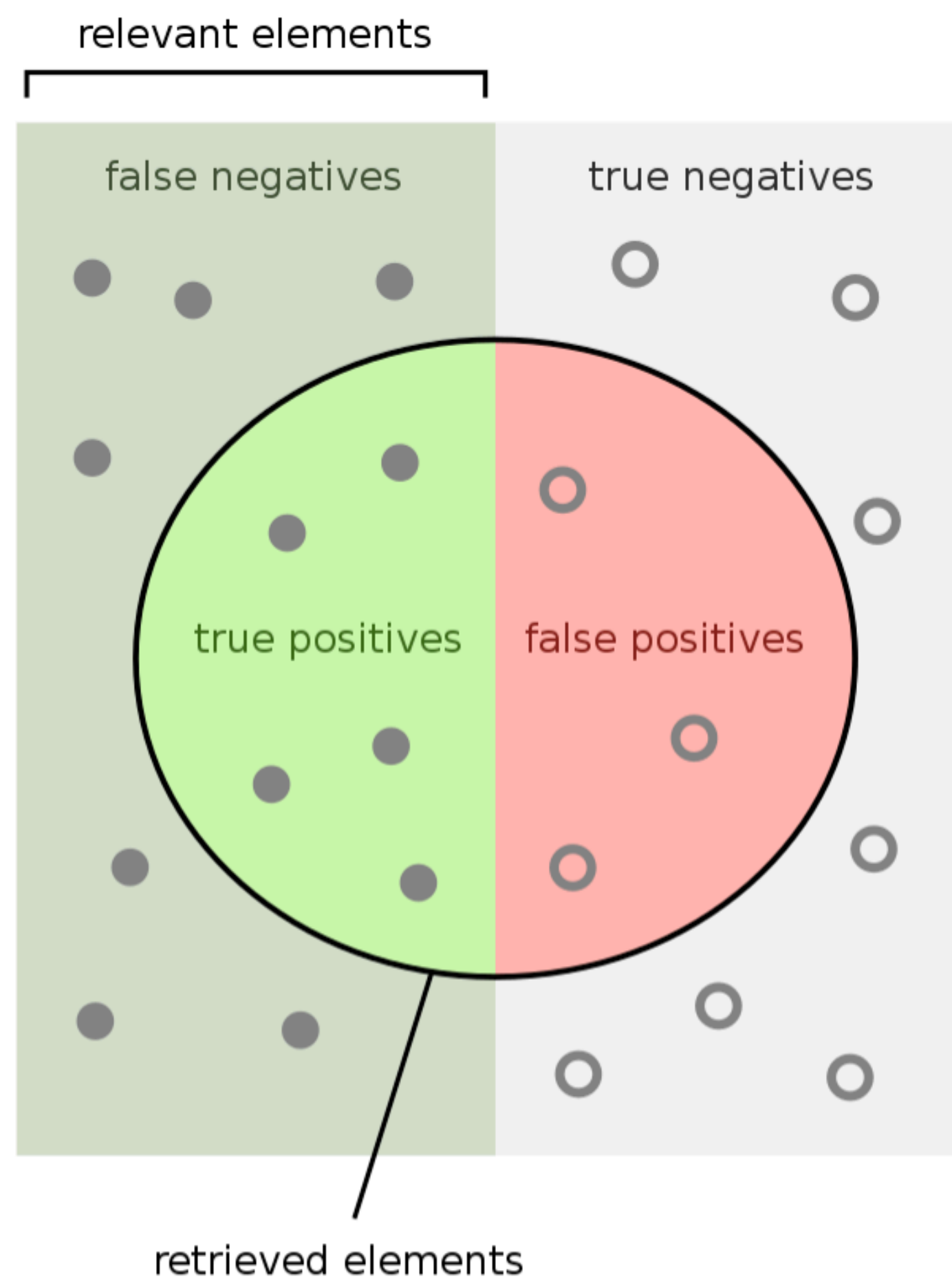
- Code a subset manually (consider intercoder reliability)
- Compare manual decisions with automated classification decisions (via recall, precision, F1)
- Iterative dictionary improvement
- Ideally: manual coding and dictionary development is performed by different persons

Creation of a manual baseline

- Codebook creation
- Who codes manually?
 - Expert coders: Coder recruitment and training sessions
 - Crowdcoders: test questions, majority choice
- Quality assessment: e.g., Inter-coder reliability of involved coders, majority vote
 - How reliable? Consider valid disagreement (Baden et al., 2023)
- 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)
- $F1 = 2 * (P * R) / (P + R)$



How many retrieved
items are relevant?

Precision = $\frac{\text{true positives}}{\text{true positives} + \text{false positives}}$

How many relevant
items are retrieved?

Recall = $\frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$

Dictionary pros vs cons

- 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
 - Language nuances

Questions?

Regular Expressions (regex)

Regular Expressions (regex)

- How to pronounce?
- /gɪf/
- /dʒɪf/

Regular Expressions (regex)

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

Regular Expressions (regex)

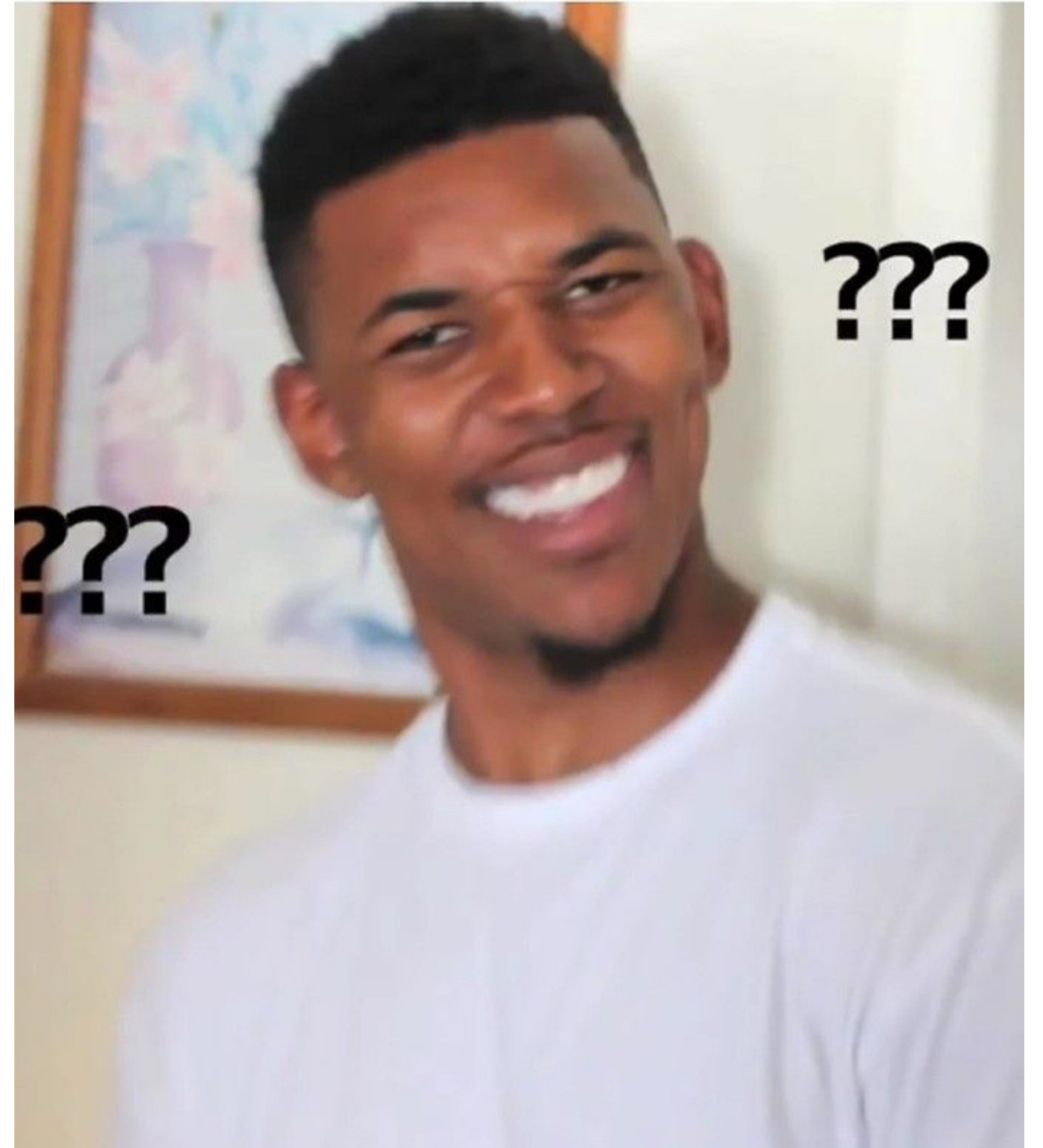
- How to pronounce?
- **/gɪf/**
- /dʒɪf/
- **/ɹɛ.gɛks/**
- /ɹɛ.dʒɛks/

Regular Expressions (regex)

- How to pronounce?
- /gɪf/
- **/dʒɪf/**
- /ɹɛ.ɡɛks/
- **/ɹɛ.dʒɛks/**

Regular Expressions (regex)

- How to pronounce?
- /gɪf/
- /dʒɪf/
- /ʁε.gεks/
- /ʁε.dʒεks/



regex

- Formal language to specify search strings

regex

- Formal language to specify search strings
- Insanely difficult

regex

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

regex

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

regex

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

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

- Disjunctions

RE	Match	Example Patterns Matched
[mM]oney	Money or money	" <u>Money</u> "
[abc]	'a', 'b', <i>or</i> 'c'	"Investing in Ir <u>a</u> n" "is d <u>a</u> ngerous <u>b</u> usiness" "sitting on \$ <u>7</u> . <u>5</u> billion dollars" " <u>2005</u> and <u>2006</u> , more than " "\$ <u>150</u> million dollars" " 'Run!', he screamed. <u>.</u> "
[\.]	A period	

- Ranges

RE	Match	Example Patterns Matched
[A-Z]	an upper case letter	"Rep. <u>A</u> nthony <u>W</u> einer (<u>D</u> - <u>B</u> rooklyn & <u>Q</u> ueens)"
[a-z]	a lower case letter	"ACORN' <u>s</u> "
[0-9]	a single digit	"(<u>9</u> th CD) "

- Negations

RE	Match	Example Patterns Matched
[^A-Z]	not an upper case letter	“ACORN <u>s</u> ”
[^Ss]	neither ‘S’ nor ‘s’	“ <u>ACORN</u> ’s”
[^\.]	not a period	“ ‘Run!’ <u>, he screamed.</u> ”

- Optional Characters: ?, *, +

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

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	" <u>P</u> alo Alto" "the town of P alo Alto"
^[^A-Z]	Not upper case start of line	" <u>t</u> he town of Palo Alto" " P alo Alto"
^.	Start of line	" <u>P</u> alo Alto" " <u>t</u> he town of Palo Alto"
.\$	Identify character that ends a line	"Wait <u>!</u> " "This is the end <u>.</u> "

- "Or" | statements, Useful short hand

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

How difficult to regex an email?

How difficult to regex an email?

- Rather...

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])?|\b(?:2(5[0-
5]|[0-4][0-9])|1[0-9][0-9]|1[0-9]?[0-9]))\b){3}(?:2(5[0-5]|[0-4][0-9])|1[0-9][0-9]|1[0-
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]))+))\b)
```


[illegible][illegible]

(?:?:\r\n)?[\t])*(?:?:?:[^(())<>@,;:\\".\\[\] \000-\031]+(?:?:?:\r\n)?[\t])
)+|\Z|(?=[\["()<>@,;:\\".\\[\]])|"(?:?:[\^"\r\\]|\\.|(?:?:\r\n)?[\t]))*"(?:?:?:
\r\n)?[\t])*(?:?:\.(?:?:\r\n)?[\t])*(?:?:[^(())<>@,;:\\".\\[\] \000-\031]+(?:?:?:(
?:\r\n)?[\t])+|\Z|(?=[\["()<>@,;:\\".\\[\]])|"(?:?:[\^"\r\\]|\\.|(?:?:\r\n)?[\t]))
)(?:?:\r\n)?[\t])*(?:?:\.(?:?:\r\n)?[\t])*(?:?:[^(())<>@,;:\\".\\[\] \000-\0
31]+(?:?:?:\r\n)?[\t])+|\Z|(?=[\["()<>@,;:\\".\\[\]])|\\([([^\[\]\r\\]|\\.)*\
](?:?:\r\n)?[\t])*(?:?:\.(?:?:\r\n)?[\t])*(?:?:[^(())<>@,;:\\".\\[\] \000-\031]+
(?:?:\r\n)?[\t])+|\Z|(?=[\["()<>@,;:\\".\\[\]])|\\([([^\[\]\r\\]|\\.)*\](?:?:
(?:\r\n)?[\t])*)*)|(??:[^(())<>@,;:\\".\\[\] \000-\031]+(?:?:\r\n)?[\t])+|\Z
|(?=[\["()<>@,;:\\".\\[\]])|"(?:?:[\^"\r\\]|\\.|(?:?:\r\n)?[\t]))*"(?:?:\r\n)
?:[\t])*)*\<(?:?:\r\n)?[\t])*(?:?:@(?:?:[^(())<>@,;:\\".\\[\] \000-\031]+(?:?:\r\n)
?:[\t])+|\Z|(?=[\["()<>@,;:\\".\\[\]])|\\([([^\[\]\r\\]|\\.)*\](?:?:\r\n)?[\t]
?:[\t])*(?:?:\.(?:?:\r\n)?[\t])*(?:?:[^(())<>@,;:\\".\\[\] \000-\031]+(?:?:\r\n)
?:[\t])+|\Z|(?=[\["()<>@,;:\\".\\[\]])|\\([([^\[\]\r\\]|\\.)*\](?:?:\r\n)?[\t]
?:[\t])*)*(?:?:,@(?:?:\r\n)?[\t])*(?:?:[^(())<>@,;:\\".\\[\] \000-\031]+(?:?:\r\n)?[\t]
?:[\t])+|\Z|(?=[\["()<>@,;:\\".\\[\]])|\\([([^\[\]\r\\]|\\.)*\](?:?:\r\n)?[\t])*)
(?:?:\.(?:?:\r\n)?[\t])*(?:?:[^(())<>@,;:\\".\\[\] \000-\031]+(?:?:\r\n)?[\t]
?:[\t])+|\Z|(?=[\["()<>@,;:\\".\\[\]])|\\([([^\[\]\r\\]|\\.)*\](?:?:\r\n)?[\t])*)*)
: (?:?:\r\n)?[\t]))?(?:?:[^(())<>@,;:\\".\\[\] \000-\031]+(?:?:\r\n)?[\t])+
|\Z|(?=[\["()<>@,;:\\".\\[\]])|"(?:?:[\^"\r\\]|\\.|(?:?:\r\n)?[\t]))*"(?:?:\r
\n)?[\t])*(?:?:\.(?:?:\r\n)?[\t])*(?:?:[^(())<>@,;:\\".\\[\] \000-\031]+(?:?:\r\n)
?:[\t])+|\Z|(?=[\["()<>@,;:\\".\\[\]])|"(?:?:[\^"\r\\]|\\.|(?:?:\r\n)?[\t]
?:[\t]))*"(?:?:\r\n)?[\t])*(?:?:\.(?:?:\r\n)?[\t])*(?:?:[^(())<>@,;:\\".\\[\] \000-\031
]+(?:?:\r\n)?[\t])+|\Z|(?=[\["()<>@,;:\\".\\[\]])|\\([([^\[\]\r\\]|\\.)*\](?:?:
(?:\r\n)?[\t])*)*(?:?:\.(?:?:\r\n)?[\t])*(?:?:[^(())<>@,;:\\".\\[\] \000-\031]+(?:?:
?:\r\n)?[\t])+|\Z|(?=[\["()<>@,;:\\".\\[\]])|\\([([^\[\]\r\\]|\\.)*\](?:?:
?:\r\n)?[\t])*)*)*\>(?:?:\r\n)?[\t])*(?:?:[^(())<>@,;:\\".\\[\] \000-\031]+(?:?:
?:\r\n)?[\t])+|\Z|(?=[\["()<>@,;:\\".\\[\]])|"(?:?:[\^"\r\\]|\\.|(?:?:\r\n)?[\t]
?:[\t]))*"(?:?:\r\n)?[\t])*(?:?:\r\n)?[\t])*(?:?:[^(())<>@,;:\\".\\[\] \000-\031]+
(?:?:\r\n)?[\t])+|\Z|(?=[\["()<>@,;:\\".\\[\]])|"(?:?:[\^"\r\\]|\\.|(?:?:\r\n)?[\t]
?:[\t]))*"(?:?:\r\n)?[\t])*(?:?:\.(?:?:\r\n)?[\t])*(?:?:[^(())<>@,;:\\".\\[\] \000-\031]+
(?:?:\r\n)?[\t])+|\Z|(?=[\["()<>@,;:\\".\\[\]])|\\([([^\[\]\r\\]|\\.)*\](?:?:\r\n)?[\t])
(?:?:[^(())<>@,;:\\".\\[\] \000-\031]+(?:?:\r\n)?[\t])+|\Z|(?=[\["()<>@,;:\\".\\[\]
:\\".\\[\]])|\\([([^\[\]\r\\]|\\.)*\](?:?:\r\n)?[\t])*(?:?:\.(?:?:\r\n)?[\t])*(?:?:
?:[^(())<>@,;:\\".\\[\] \000-\031]+(?:?:\r\n)?[\t])+|\Z|(?=[\["()<>@,;:\\".\\[\]
:\\".\\[\]])|\\([([^\[\]\r\\]|\\.)*\](?:?:\r\n)?[\t])*)*)|(??:[^(())<>@,;:\\".\\[\] \000-\
031]+(?:?:\r\n)?[\t])+|\Z|(?=[\["()<>@,;:\\".\\[\]])|"(?:?:[\^"\r\\]|\\.|(
?:\r\n)?[\t]))*"(?:?:\r\n)?[\t])*(?:?:\r\n)?[\t])*(?:?:@(?:?:[^(())<>@,;
:\\".\\[\] \000-\031]+(?:?:\r\n)?[\t])+|\Z|(?=[\["()<>@,;:\\".\\[\]])|\\([([
^\[\]\r\\]|\\.)*\](?:?:\r\n)?[\t])*(?:?:\.(?:?:\r\n)?[\t])*(?:?:[^(())<>@,;:\\".
\\[\] \000-\031]+(?:?:\r\n)?[\t])+|\Z|(?=[\["()<>@,;:\\".\\[\]])|\\([([^\[\]
]\r\\]|\\.)*\](?:?:\r\n)?[\t])*)*(?:?:,@(?:?:\r\n)?[\t])*(?:?:[^(())<>@,;:\\".\\
[\] \000-\031]+(?:?:\r\n)?[\t])+|\Z|(?=[\["()<>@,;:\\".\\[\]])|\\([([^\[\]
]\r\\]|\\.)*\](?:?:\r\n)?[\t])*)*(?:?:\.(?:?:\r\n)?[\t])*(?:?:[^(())<>@,;:\\".\\[\]
 \000-\031]+(?:?:\r\n)?[\t])+|\Z|(?=[\["()<>@,;:\\".\\[\]])|\\([([^\[\]
]\r\\]|\\.)*\](?:?:\r\n)?[\t])*)*)*(?:?:\r\n)?[\t])*(?:?:[^(())<>@,;:\\".\\[\] \0
00-\031]+(?:?:\r\n)?[\t])+|\Z|(?=[\["()<>@,;:\\".\\[\]])|"(?:?:[\^"\r\\]|\\

Concepts and Data

Concepts and Data

- What do we actually want to measure?
- And where do we find it?

From the research questions/hypotheses to key concepts

- Example research question: Comparing the Twitter accounts of the alright influencers Paula Winterfeldt, Lauren Southern and Lana Lokteff, which images of women are conveyed most frequently?
- Example hypothesis: On Twitch, male streamers are rated more positively by viewers than female streamers?
 - Which concepts must be recorded?

Discovery

- Idea from qualitative research
- A “method” to discover new concepts through descriptive analysis
- Grimmer et al., 2022:
 - I. Context Relevance
 - II. No Ground Truth
 - III. Concept vs. Method
 - IV. Data Separation

Principles of Discovery

- Principle 1: ***Context relevance.***
- Text as data models complement theory and substantive knowledge.
Contextual knowledge amplifies our ability to make computational discoveries

Principles of Discovery

- Principle 2: *No ground truth.*
- There is no ground truth conceptualization; only after a concept is fixed can we talk meaningfully about it being right or wrong

Principles of Discovery

- Principle 3: ***Judge the concept, not the method.***
- The method you used to arrive at a conceptualization does not matter for assessing the concept's value – its utility does

Principles of Discovery

- Principle 4: ***Separate data is best.***
- Ideally after data is used for discovery it should be discarded in favor of new data for confirming/testing discoveries.

Codebook

- The codebook is the tool you use to code your content
- It is a kind of questionnaire that you use to inquiry the examined texts/photos/videos
- The codebook should be detailed enough so that
 - you can apply it again in the same way after some time (intracoder reliability)
 - other people (with a little training) can also use it in the same way as you (intercoder reliability)

Codebook

- In automated text analysis:
 - Validation
 - Documentation

Codebook Development

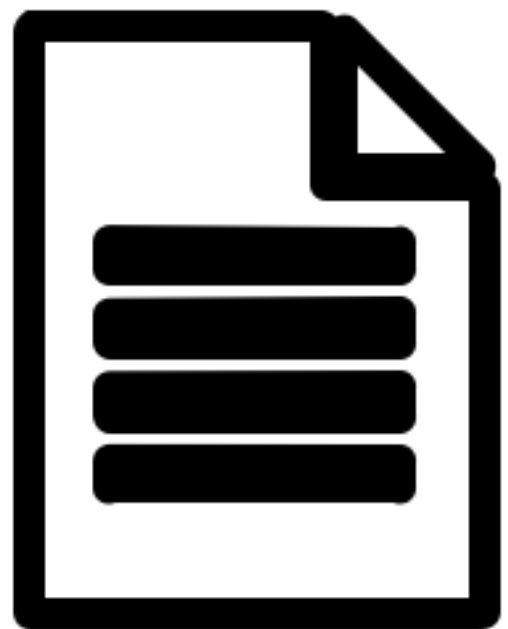
- Iterative process:
 - First draft based on the variables identified in the research question(s)/ hypothesis(s).
 - Tip: take other studies as a model
 - Code material examples using the draft codebook
 - Then refine/edit the codebook

Codebook Function

- Used for dimension reduction (e.g., Egami et al., 2022)

Codebook Function

- Used for dimension reduction (e.g., Egami et al., 2022)



$g(\cdot)$

