Quanlify with ease: Combining quantitative and qualitative corpus analysis

Andreas Blaette
SSHOC Webinar | April 16, 2020

Corpus analysis: An emerging field.

What are the challenges?

Corpus analysis: An emerging field.

What are the challenges?

There is no lack ...

- of algorithms
- of ideas and projects on research data management (implementing FAIR data principles)

But ...

Corpus analysis: An emerging field.

What are the challenges?

There is no lack ...

- of algorithms
- of ideas and projects on research data management (implementing FAIR data principles)

But ...

- Acquisition of NLP techniques in the social sciences & humanities still piecemeal

Corpus analysis: An emerging field.

What are the challenges?

There is no lack ...

- of algorithms
- of ideas and projects on research data management (implementing FAIR data principles)

But ...

- Acquisition of NLP techniques in the social sciences & humanities still piecemeal
- Tools for processing text that scale well are just emerging

Corpus analysis: An emerging field.

What are the challenges?

There is no lack ...

- of algorithms
- of ideas and projects on research data management (implementing FAIR data principles)

But ...

- Acquisition of NLP techniques in the social sciences & humanities still piecemeal
- Tools for processing text that scale well are just emerging
- Availability of data for replication is missing

Corpus analysis: An emerging field.

What are the challenges?

There is no lack ...

- of algorithms
- of ideas and projects on research data management (implementing FAIR data principles)

But ...

- Acquisition of NLP techniques in the social sciences & humanities still piecemeal
- Tools for processing text that scale well are just emerging
- Availability of data for replication is missing
- Reproducibility of data (getting FAIRER) is wanting

Corpus analysis: An emerging field.

What are the challenges?

There is no lack ...

- of algorithms
- of ideas and projects on research data management (implementing FAIR data principles)

But ...

- Acquisition of NLP techniques in the social sciences & humanities still piecemeal
- Tools for processing text that scale well are just emerging
- Availability of data for replication is missing
- Reproducibility of data (getting FAIRER) is wanting
- Integration of quantitative and qualitative approaches to text is technically difficulat

The PolMine Project (www.polmine.de)

Research - Data - Code - Tutorials - Centre

Research - Data - Code - Tutorials - Centre

Corpora

- **GermaParl**: Corpus of the German Bundestag, regional parliaments DOI 10.5281/zenodo.3742113
- **UNGA** Verbatim Records of the United Nations General Assembly DOI 10.5281/zenodo.3748858
- **MigParl**: Debates on migration and integration in Germany's Regional Parliaments
- **MigPress**: Corpus of reports on migration and integration in Süddeutsche Zeitung und Frankfurter Allgemeine Zeitung (2000-2019)

The PolMine Project (www.polmine.de)

Research - Data - Code - Tutorials - Centre

Corpora

- **GermaParl**: Corpus of the German Bundestag, regional parliaments DOI 10.5281/zenodo.3742113
- **UNGA** Verbatim Records of the United Nations General Assembly DOI 10.5281/zenodo.3748858
- **MigParl**: Debates on migration and integration in Germany's Regional Parliaments
- **MigPress**: Corpus of reports on migration and integration in Süddeutsche Zeitung und Frankfurter Allgemeine Zeitung (2000-2019)

*(Toolchain for corpus preparation: **frapp**, **bignlp**, **biglda**, **trickypdf**)*

The PolMine Project (www.polmine.de)

Research - Data - Code - Tutorials - Centre

Corpora

- **GermaParl**: Corpus of the German Bundestag, regional parliaments DOI 10.5281/zenodo.3742113
- **UNGA** Verbatim Records of the United Nations General Assembly DOI 10.5281/zenodo.3748858
- **MigParl**: Debates on migration and integration in Germany's Regional Parliaments
- **MigPress**: Corpus of reports on migration and integration in Süddeutsche Zeitung und Frankfurter Allgemeine Zeitung (2000-2019)

*(Toolchain for corpus preparation: **frapp**, **bignlp**, **biglda**, **trickypdf**)*

Packages for corpus analysis

- **polmineR**: Elementary vocabulary for corpus analysis CRAN 0.8.0
- **RcppCWB**: R Wrapper for the C Code of the Corpus Workbench (using C++/Rcpp) CRAN 0.2.8
- **cwbtools**: Tools to create and manage CWB indexed corpora CRAN 0.2.0

Things are evolving.

But where do we stand?

Things are evolving.

But where do we stand?

- Acquisition of NLP techniques in the social sciences & humanities:
  Working with large-scale, linguistically annotated corpora

Things are evolving.

But where do we stand?

- Acquisition of NLP techniques in the social sciences & humanities:
  Working with large-scale, linguistically annotated corpora
- Tools for processing text that scale well are just emerging:
  Packages such as bignlp, biglda

Things are evolving.

But where do we stand?

- Acquisition of NLP techniques in the social sciences & humanities:
  Working with large-scale, linguistically annotated corpora
- Tools for processing text that scale well are just emerging:
  Packages such as bignlp, biglda
- Availability of data for replication is missing:
  Depositing data with Zenodo (or other open science repositories)

Things are evolving.

But where do we stand?

- Acquisition of NLP techniques in the social sciences & humanities:
  Working with large-scale, linguistically annotated corpora
- Tools for processing text that scale well are just emerging:
  Packages such as bignlp, biglda
- Availability of data for replication is missing:
  Depositing data with Zenodo (or other open science repositories)
- Reproducibility of data (getting FAIRER):
  100% reproducibility of data

Things are evolving.

But where do we stand?

- Acquisition of NLP techniques in the social sciences & humanities:
  Working with large-scale, linguistically annotated corpora
- Tools for processing text that scale well are just emerging:
  Packages such as bignlp, biglda
- Availability of data for replication is missing:
  Depositing data with Zenodo (or other open science repositories)
- Reproducibility of data (getting FAIRER):
  100% reproducibility of data
- Integration of quantitative and qualitative approaches to text is an unfulfilled promise

Assumption, Problem, Vision and Plan

Assumption: The validity of our research with large-scale corpora depends on our ability to combine the quantitative and the qualitative analysis of textual data.

Assumption, Problem, Vision and Plan

Assumption: The validity of our research with large-scale corpora depends on our ability to combine the quantitative and the qualitative analysis of textual data.

Problem: It takes a software engineer to implement an integrated environment for the quantitative and qualitative analysis of text. You need funding to entertain the cooperation with a software engineer. Funding ends. Solutions are not sustainable.

Assumption, Problem, Vision and Plan

Assumption: The validity of our research with large-scale corpora depends on our ability to combine the quantitative and the qualitative analysis of textual data.

Problem: It takes a software engineer to implement an integrated environment for the quantitative and qualitative analysis of text. You need funding to entertain the cooperation with a software engineer. Funding ends. Solutions are not sustainable.

Vision: Wouldn't it be great to have an open source, modular toolset that can flexibly be used by ordinary computational social scientists to implement "quanlitative" workflows?

Assumption, Problem, Vision and Plan

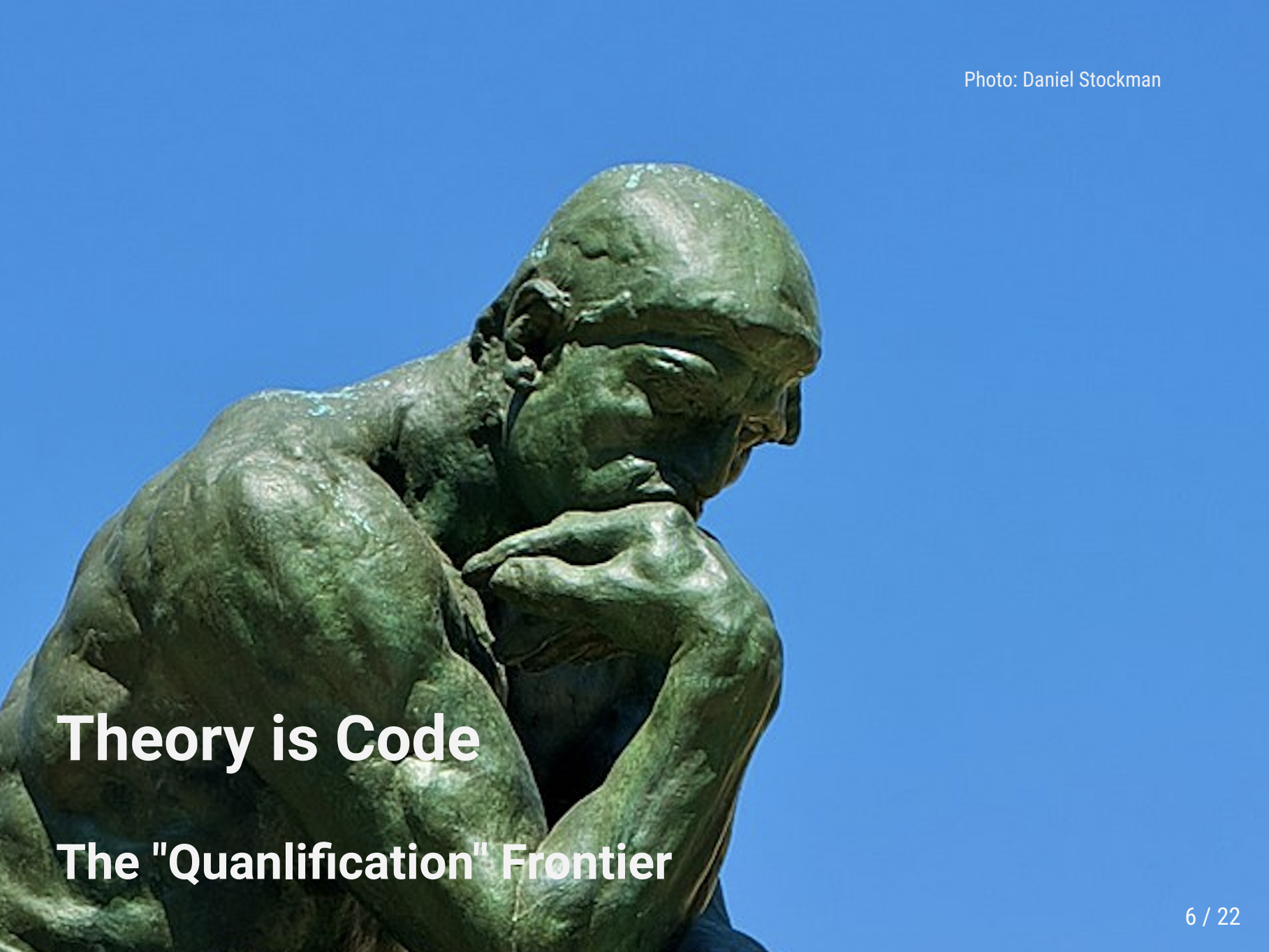
Assumption: The validity of our research with large-scale corpora depends on our ability to combine the quantitative and the qualitative analysis of textual data.

Problem: It takes a software engineer to implement an integrated environment for the quantitative and qualitative analysis of text. You need funding to entertain the cooperation with a software engineer. Funding ends. Solutions are not sustainable.

Vision: Wouldn't it be great to have an open source, modular toolset that can flexibly be used by ordinary computational social scientists to implement "quanlitative" workflows?

Plan of this Talk

1. Theory is Code - The "Quanlification" Frontier
2. Quanlification as a Matter of Design
3. Implementing Quanlifictation
4. Work Ahead - Getting Things Done
5. Discussion



Theory is Code

The "Quanlification" Frontier

The idea of "distant reading"

- "[…] the trouble with close reading […] is that it necessarily depends on an extremely small canon. […] what we really need is a little pact with the devil: we know how to read texts, so now let‘s learn how not to read them. Distant reading, where distance, let me repeat is, is a condition of knowledge. It allows you to focus on units that are much smaller or much larger than the text: devices, themes, types – or genres and systems. And if, between the very small and the very large, the text itself disappears, well, this is one of the cases where one can justifiably say, Less is more.“ (Moretti [2000] 2013: 49)

From text to numbers

The idea of "distant reading"

- "[…] the trouble with close reading […] is that it necessarily depends on an extremely small canon. […] what we really need is a little pact with the devil: we know how to read texts, so now let‘s learn how not to read them. Distant reading, where distance, let me repeat is, is a condition of knowledge. It allows you to focus on units that are much smaller or much larger than the text: devices, themes, types – or genres and systems. And if, between the very small and the very large, the text itself disappears, well, this is one of the cases where one can justifiably say, Less is more.“ (Moretti [2000] 2013: 49)

The "text as data" movement:

- scaling party positions (wordscore and wordfish) as a driver -"[…] while our method is designed to analyse the content of a text, it is not necessary for an analyst using the technique to understand, or even read, the texts to which the technique is applied. " (Laver, Benoit & Garry 2003)

Text + Numbers = Quanlification

An obsolete methodological divide?

Text + Numbers = Quanlification

An obsolete methodological divide?

Quantity

- Text as Data
- Natural Language Processing (NLP)
- Data / Text Mining
- Machine Learning (ML)
- "Validate, validate, validate" (Grimmer et al. 2013)

Text + Numbers = Quanlification

An obsolete methodological divide?

Quantity

- Text as Data
- Natural Language Processing (NLP)
- Data / Text Mining
- Machine Learning (ML)
- "Validate, validate, validate" (Grimmer et al. 2013)

Quality

- eHumanities / Digital Humanities
- Corpus Linguistics
- Computational Linguistics
- "blended reading" (Stulpe, Lemke 2015), "scalable reading" (Weitin 2017) & related concepts

Text + Numbers = Quanlification

An obsolete methodological divide?

Quantity

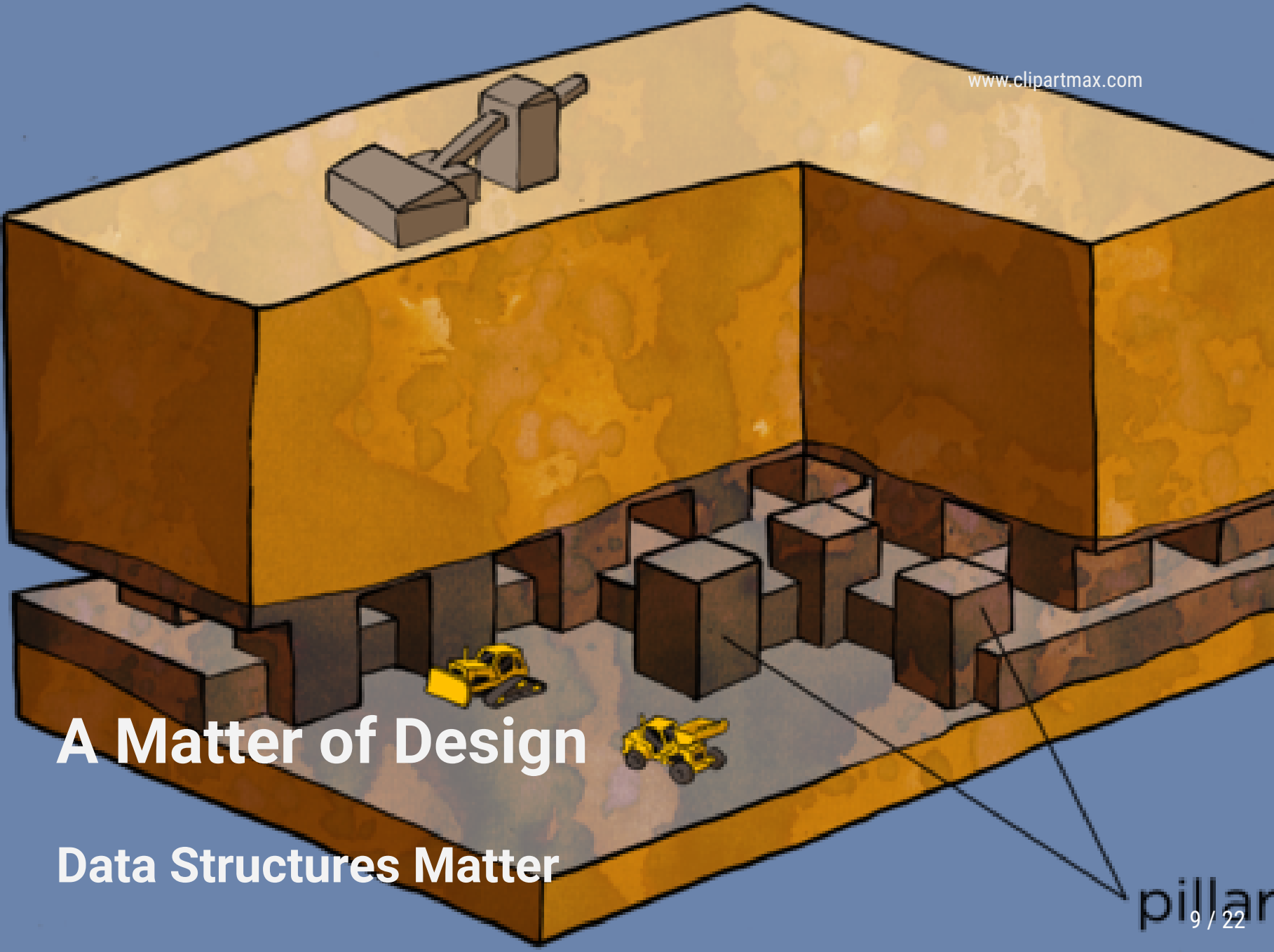
- Text as Data
- Natural Language Processing (NLP)
- Data / Text Mining
- Machine Learning (ML)
- "Validate, validate, validate" (Grimmer et al. 2013)

Quality

- eHumanities / Digital Humanities
- Corpus Linguistics
- Computational Linguistics
- "blended reading" (Stulpe, Lemke 2015), "scalable reading" (Weitin 2017) & related concepts

Quanlification

- Epistemological disputes notwithstanding: The necessity to combine qualitative and quantitative approaches to text is conceptually undisputed.
- Software inhibits combining quantity and quality: Tools are there, but setting up a quanlitative project is expensive: Difficult without a dedicated software engineer



A Matter of Design

Data Structures Matter

Three-Tier Architecture: C & R & More

polmineR: A basic vocabulary for quanlification

- **Corpora and subcorpora**
  - corpus objects: *corpus()*
  - subsetting corpora: *partition()* / *subset()*

Verbs and nouns for corpus analysis

polmineR: A basic vocabulary for quanlification

- **Corpora and subcorpora**
  - corpus objects: *corpus()*
  - subsetting corpora: *partition()* / *subset()*
- **Quantification**
  - counting: *hits()*, *count()*, *dispersion()* (and *size()*)
  - cooccurrences: *cooccuurrences()*, *Cooccurrences()*
  - feature extraction: *features()*
  - term-document-matrices: *as.sparseMatrix()*, *as.TermDocumentMatrix()*

polmineR: A basic vocabulary for quanlification

- **Corpora and subcorpora**
  - corpus objects: *corpus()*
  - subsetting corpora: *partition()* / *subset()*
- **Quantification**
  - counting: *hits()*, *count()*, *dispersion()* (and *size()*)
  - cooccurrences: *cooccuurrences()*, *Cooccurrences()*
  - feature extraction: *features()*
  - term-document-matrices: *as.sparseMatrix()*, *as.TermDocumentMatrix()*
- **Qualitative analysis**
  - Keywords-in-context / concordances: *kwic()*
  - full text (of a subcorpus): *get_token_stream()*, *as.markdown()*, *as.html()*, *read()*

polmineR - the People's Corpus Miner

polmineR - the People's Corpus Miner

- Prerequisites:
  - Any kind of computer that still has keys (Windows, Linux, macOS)
  - Installation of R/RStudio

polmineR - the People's Corpus Miner

- Prerequisites:
  - Any kind of computer that still has keys (Windows, Linux, macOS)
  - Installation of R/RStudio
- Three lines of code will get you polmineR and GermaParl (or any other corpus)

```
install.packages("polmineR")
install.packages("GermaParl") # includes small sample dataset
GermaParl::germaparl_download_corpus() # get the full corpus (1 GB)
```

polmineR - the People's Corpus Miner

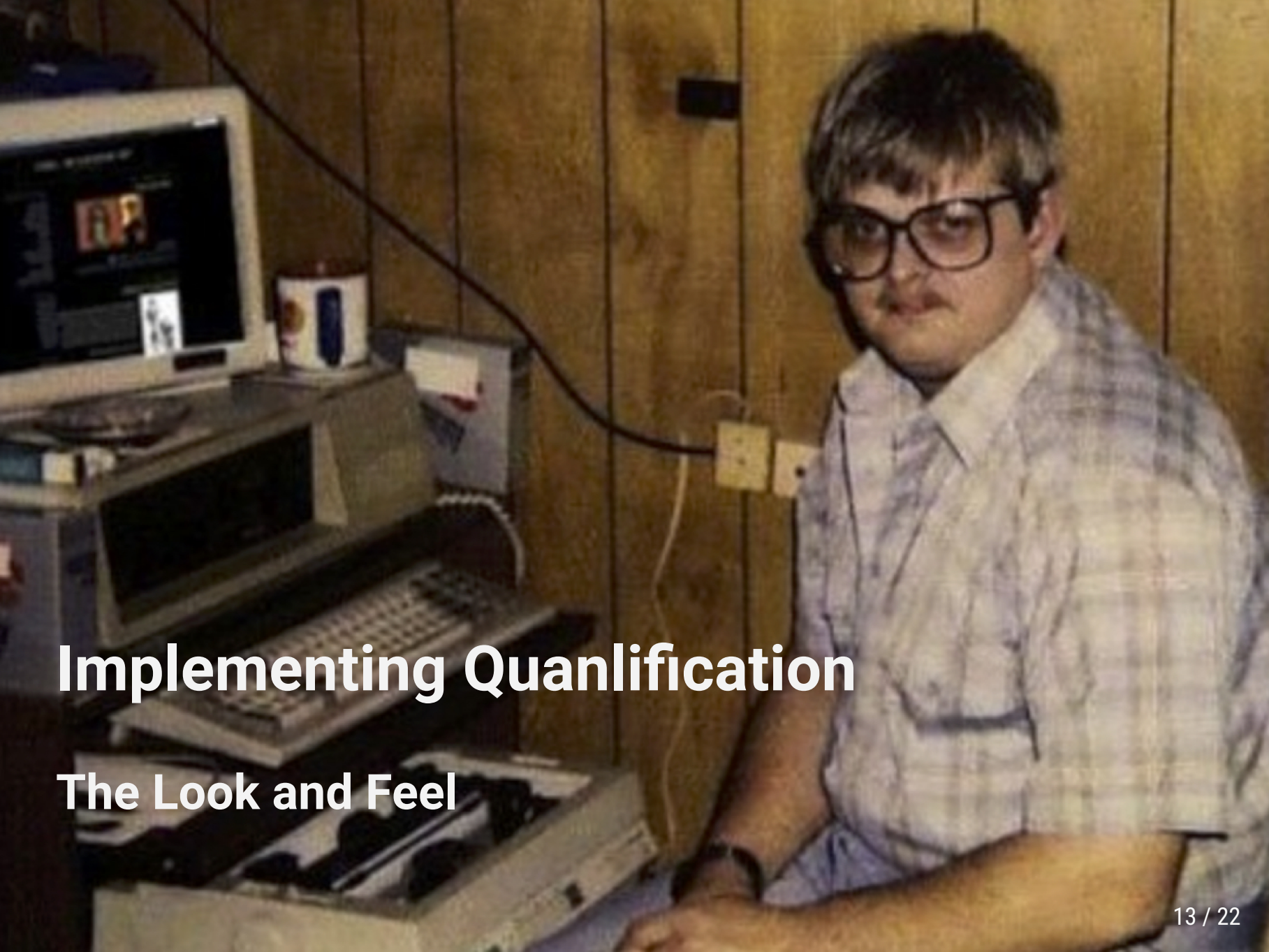
- Prerequisites:
  - Any kind of computer that still has keys (Windows, Linux, macOS)
  - Installation of R/RStudio
- Three lines of code will get you polmineR and GermaParl (or any other corpus)

```
install.packages("polmineR")
install.packages("GermaParl") # includes small sample dataset
GermaParl::germaparl_download_corpus() # get the full corpus (1 GB)
```

- Start your inquiry

```
library(polmineR)
count("GERMAPARL", query = "Europa")
kwic("GERMAPARL", query = "Europa")
```

- Installation options: Local install, or R, RStudio, OpenCPU on server (remote corpus access)



Implementing Quanlification

The Look and Feel

Reading Anywhere: 'fulltext' htmlwidget

Problem Statement

- **Read Anything:** Cooccurrences, concordances, subcorpora, topic models - you want to contextualize all of it
- **Read Anywhere:** Include fulltext output into (html) documents and slides, and in GUIs

Implementation

- polmineR: Implementation of a `read()`-method
- GUI: Package 'fulltext' that renders input data into an "htmlwidget" (a truly flexible device)

Demo

- HTML documents with scrollable fulltext -> example
- Slides with `fulltext` htmlwidget -> example
- polmineR shiny App -> example

Highlighting and Tooltipping

Problem Statement

- **Highlighting with multiple colors:** The statistical analysis of text yields variable dictionaries with word weights - visualising multiple dictionaries at the same time will help to gather the semantic sense of numeric analyses
- **Tooltipping:** Colors alone may be misleading. Show numeric information on demand

Implementation

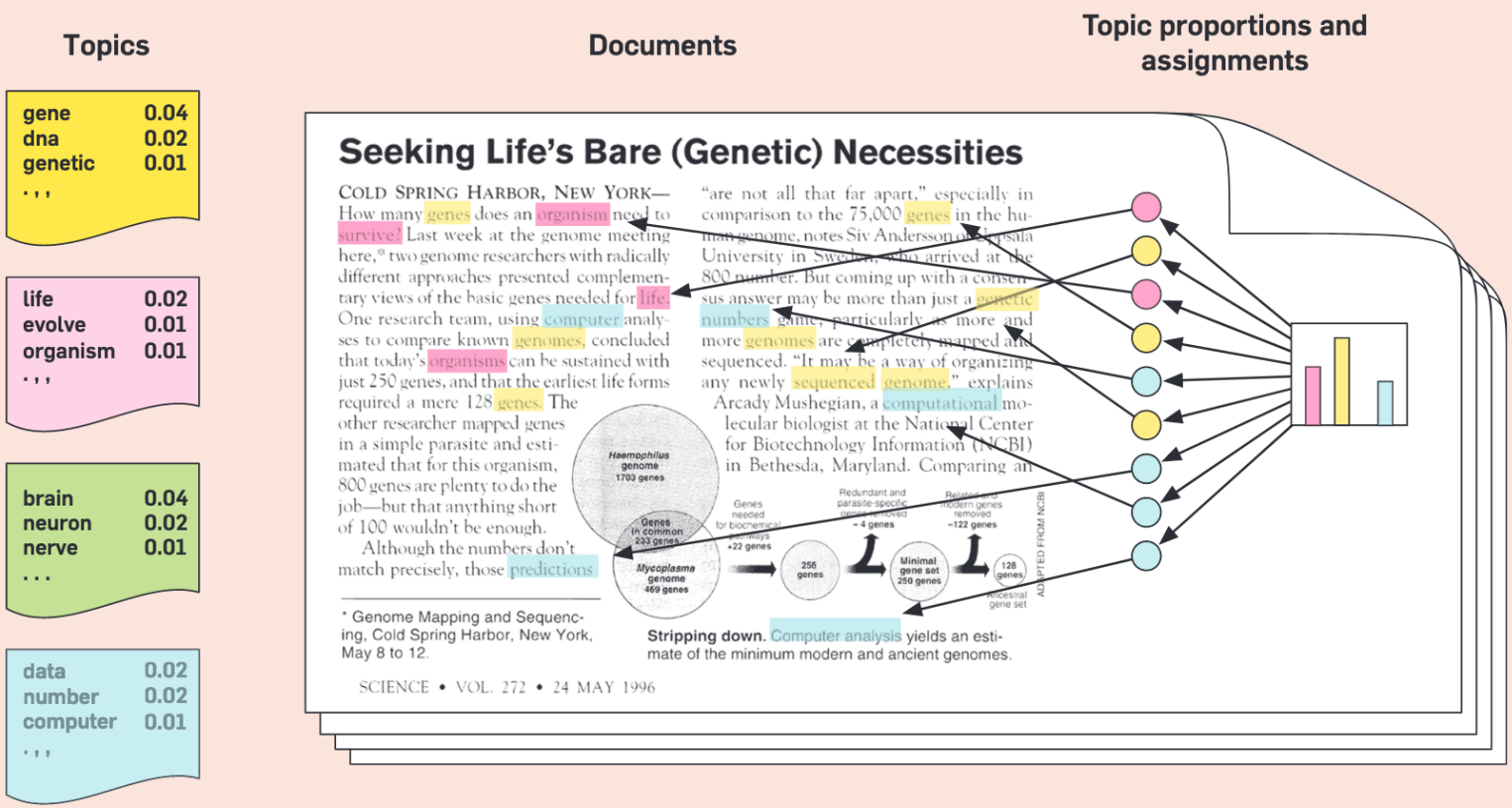
- polmineR package: Implementation of methods `highlight()` and `tooltips()`
- GUI: Enrich input data for htmlwidgets and amend CSS

Demo

- Validating sentiment analyses with interactive KWIC tables -> example
- Evaluate topic models with flexdashboard -> example
- Understanding the data behind cooccurrence graphs -> example

Blei 2012: Intuition behind LDA

Figure 1. The intuitions behind latent Dirichlet allocation. We assume that some number of "topics," which are distributions over words, exist for the whole collection (far left). Each document is assumed to be generated as follows. First choose a distribution over the topics (the histogram at right); then, for each word, choose a topic assignment (the colored coins) and choose the word from the corresponding topic. The topics and topic assignments in this figure are illustrative—they are not fit from real data. See Figure 2 for topics fit from data.



Annotation

Problem Statement

- **Annotation and intersubjectivity**: In the qualitative research tradition, annotations are a means to communicate evaluative decisions to other researchers
- **Annotation and machine learning** Annotations (generating labelled data) are a precondition for machine learning

Implementation

- polmineR level: Implementation of `edit()`-methods
- Htmlwidget with annotation functionality (different from `fulltext` htmlwidget, as it returns values)

Demo

- Annotate any class inheriting from the `textstat` class (i.e. tables - kwic, cooccurrences, features) -> example
- Simple text annotation with the `annolite` package
- Annotate cooccurrence graphs -> example

Annotation

- Code Example: Annotate a cooccurrences object

```
library(polmineR)
s <- cooccurrences("UNGA", "sustainability") # could also be kwic etc.
annotations(s) <- list(name = "annotation", what = "")
edit(s)
```

- Code Example: Annotate fulltext

```
library(polmineR)
library(annolite) # at github.com/PolMine/annolite, dev-branch

data <- corpus("GERMAPARLMINI") %>%
  subset(speaker == "Volker Kauder") %>%
  subset(date == "2009-11-10") %>%
  as("partition") %>%
  as.fulltextdata(headline = "Volker Kauder (CDU)")
anno <- annotate(data)
```

Work Ahead

Getting Things Done

The Nature of the Beast

A modular toolset (not a framework)

- turn whatever is already there into tools for quanlification
- set of htmlwidgets (crosstalk enabled) as elements
- shiny modules, shiny apps, and shiny gadgets
- flexdashboards
- Rmarkdown templates (for documents, slides, flexdashboards)

The Nature of the Beast

A modular toolset (not a framework)

- turn whatever is already there into tools for quanlification
- set of htmlwidgets (crosstalk enabled) as elements
- shiny modules, shiny apps, and shiny gadgets
- flexdashboards
- Rmarkdown templates (for documents, slides, flexdashboards)

Skills required

- Know when to use what (presenting research results is different from research)
- It will take sound documentation, tutorials, recipes to illuminate the toolset!

Conscious Uncoupling and Modularization

- *fulltext*: Toolset to generate fulltext display from corpus data

- *gradget*: Graph annotation widget

- *annolite*: Leightweight Fulltext Display and Annotation Tools

- *topicanalysis*: Auxiliary functions for topicmodelling.

- *quanlify*:Toolset for the qualitative validation of quantitative text analysis

Beware! All of this is experimental!

Discussion. Frontier.

Vision

- Offer a very flexible set of tools to implement all kinds of workflows that entail distant and close reading with minimal cost
- A **l**eightweight **i**nfrastructure for the **qu**anlitat**i**ve analysis of text **d**ata ("liquid")
- Towards a people's framework for quanlification

Discussion. Frontier.

Vision

- Offer a very flexible set of tools to implement all kinds of workflows that entail distant and close reading with minimal cost
- A **l**eightweight **i**nfrastructure for the **qu**anlitat**i**ve analysis of text **d**ata ("liquid")
- Towards a people's framework for quanlification

Discussion Points

- Alternative approaches, relevant previous work
- Balance between GUI and console?
- Will there be users?
- How to build a community?
- Role for EOSC/SSHOC?