E+ ADSEE

JobMarket Signalling
2. Approach Analysis

ADSEE

Dr. Alan Berg

Co-funded by the
Erasmus+ Programme
of the European Union

Overview

In this module we look at a number of basic approaches to analysing Job Market Data centred around text mining. During this course we will exercise a number of the approaches.

The coding section is about looking at the text, seeing specific words in their native context. In this case Job vacancies. This is a well tested method which enables problem domain experts to create dictionaries of words that represent different signals.

In the results decomposition module the code section will detail one method to measure the reliability of the expert opinion.

Why these methods?

Because we want to:
Look at the use of the wording of job adverts (vacancies) in the context that the words are used.

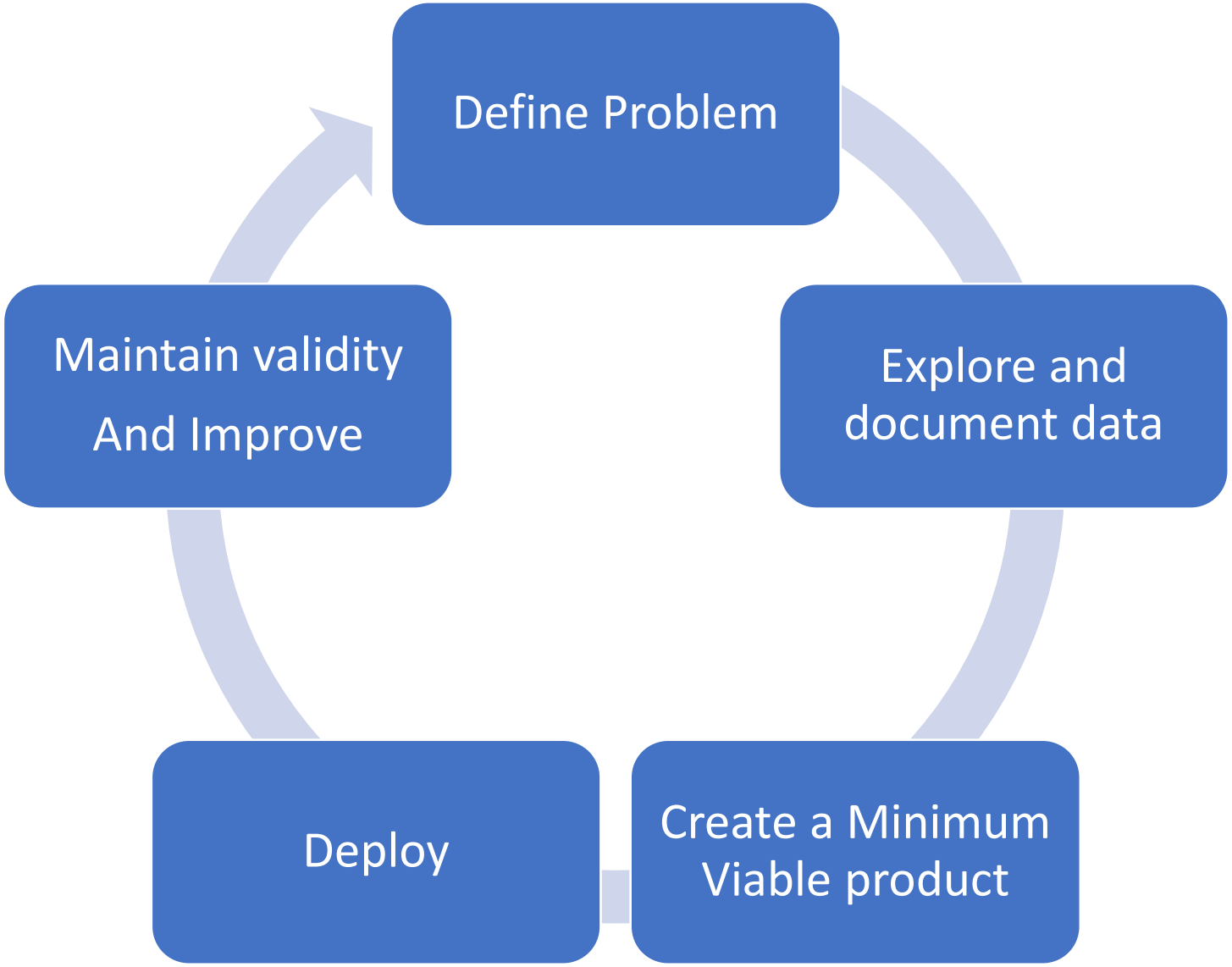
Approaches

Two main lines of attack:

1. You start with a research question and define a hypothesis to test
2. You iteratively explore the data.

Data Science Lifecycle for Business

- Many possible workflows
- Many possible interactions within the workflow
- We are not looking at the engineering aspects
- We are providing you a sense of what can be achieved with little

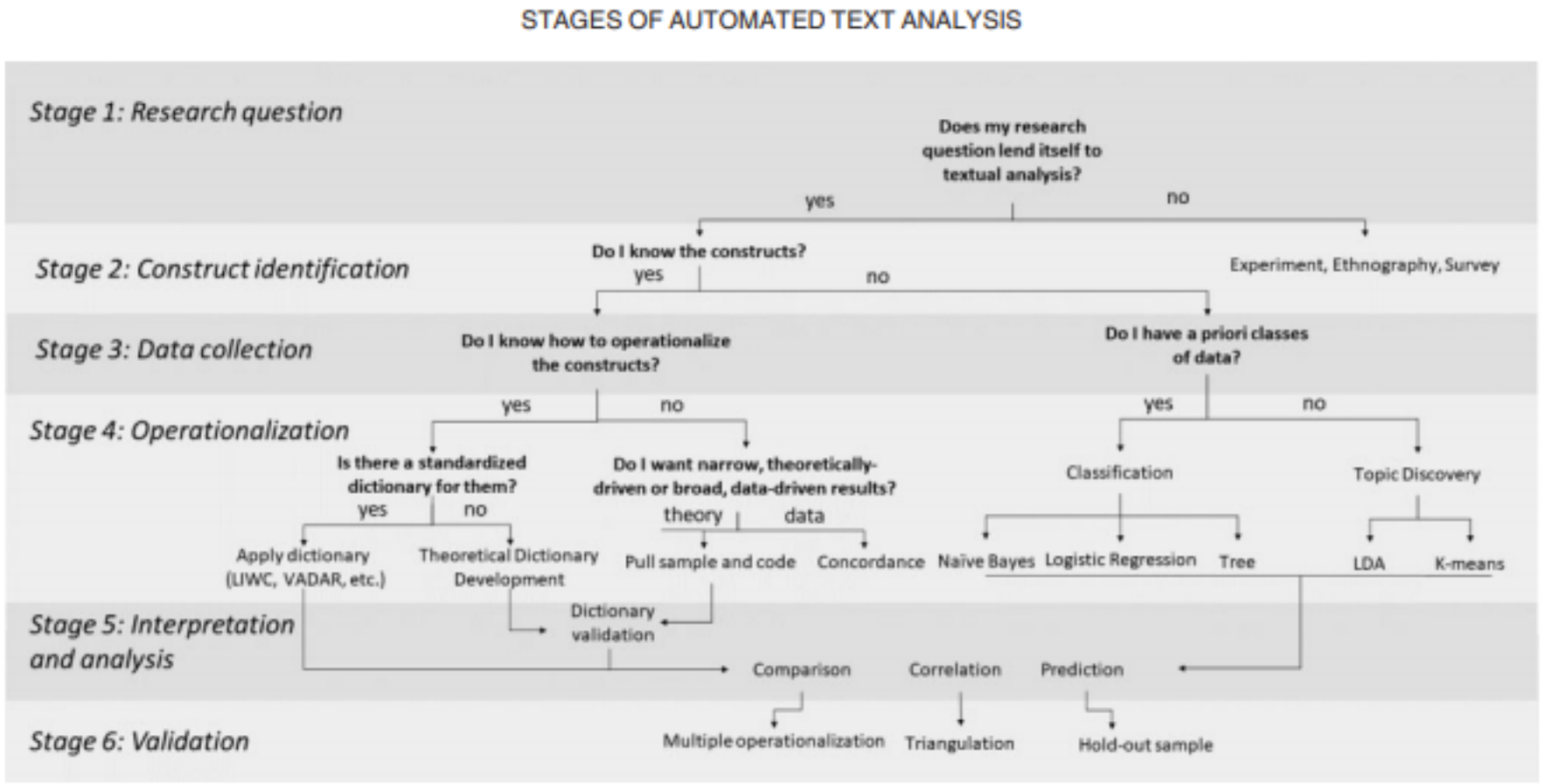


CRISP

https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining

https://github.com/FavioVazquez/ds-cheatsheets/blob/master/Business_Science/img/Business_Science_Problem_Framework.png

Workflows

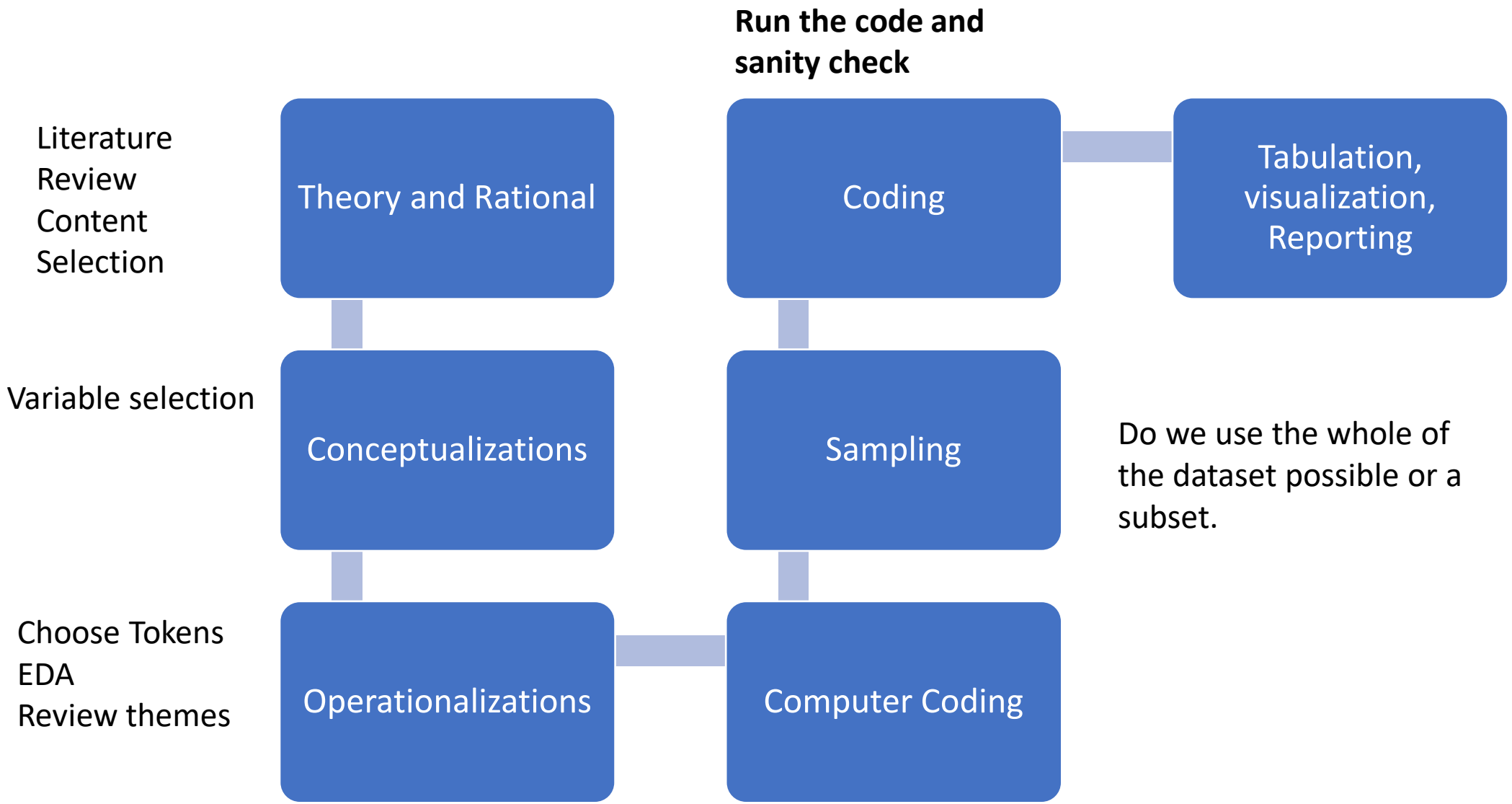


https://www.researchgate.net/publication/324495873_Automated_Text_Analysis_for_Consumer_Research

Example Research questions

Research question	Text	Linguistic aspect	Source
Dictionary-based—Comparison			
How does temporal and spatial distance affect emotions after a tragic event?	Twitter	Semantic	Doré et al. 2015
How do power and affiliation vary by political ideology?	Transcripts (chatrooms, State of the Union), news websites	Semantic	Fetterman et al. 2015
What explains representational gender bias in the media?	Newspapers	Phatic	Shor et al. 2015
How does personal pronoun use in firm-customer interactions impact customer attitude?	Transcripts	Pragmatic	Packard, Moore, and McFerran 2016
Why don't major crises like oil spills provoke broad changes in public discourse concerning the systemic risks inherent to a carbon-dependent economy?	Newspaper articles	Semantic	Humphreys and Thompson 2014
Do people modify warmth to appear competent (and vice versa) when doing impression management?	Emails	Semantic	Holoien and Fiske 2013
Does social hierarchy affect language use? In what ways?	Emails	Pragmatic	Kacewicz et al. 2014
Do Christians and atheists vary in their language use?	Twitter	Semantic	Ritter et al. 2013
How does someone's communication style change based on private versus public communication?	Facebook wall posts and private messages	Semantic, pragmatic	Bazarova 2012
How do letters to shareholders differ in a period of economic growth versus recession?	Letters to shareholders	Semantic	Pollach 2012
Are people with the same linguistic style more likely to form a romantic relationship?	Transcripts, instant messages	Stylistic, pragmatic	Ireland et al. 2011
How does happiness change throughout the lifecycle?	Personal blogs	Semantic	Mogilner et al. 2011
Dictionary-based—Correlation			
Do depressed patients use more self-focused language?	Written essays	Semantic	Brockmeyer et al. 2015

Phatic
relating to, or being speech used for social or emotive purposes rather than for communicating information


Run the code and sanity check
Literature Review Content Selection
Theory and Rational
Coding
Tabulation, visualization, Reporting
Variable selection
Conceptualizations
Sampling
Do we use the whole of the dataset possible or a subset.
Choose Tokens EDA Review themes
Operationalizations
Computer Coding
Look at word frequencies, topics, etc and create dictionaries. Check that the experts agree based on a coding book

What is the tidyverse?

The tidyverse is an opinionated collection of R packages designed for data science. All packages share an underlying design philosophy, grammar, and data structures.

https://www.tidyverse.org/

What is tidy data?

1. Every column is variable.
2. Every row is an observation.
3. Every cell is a single value.

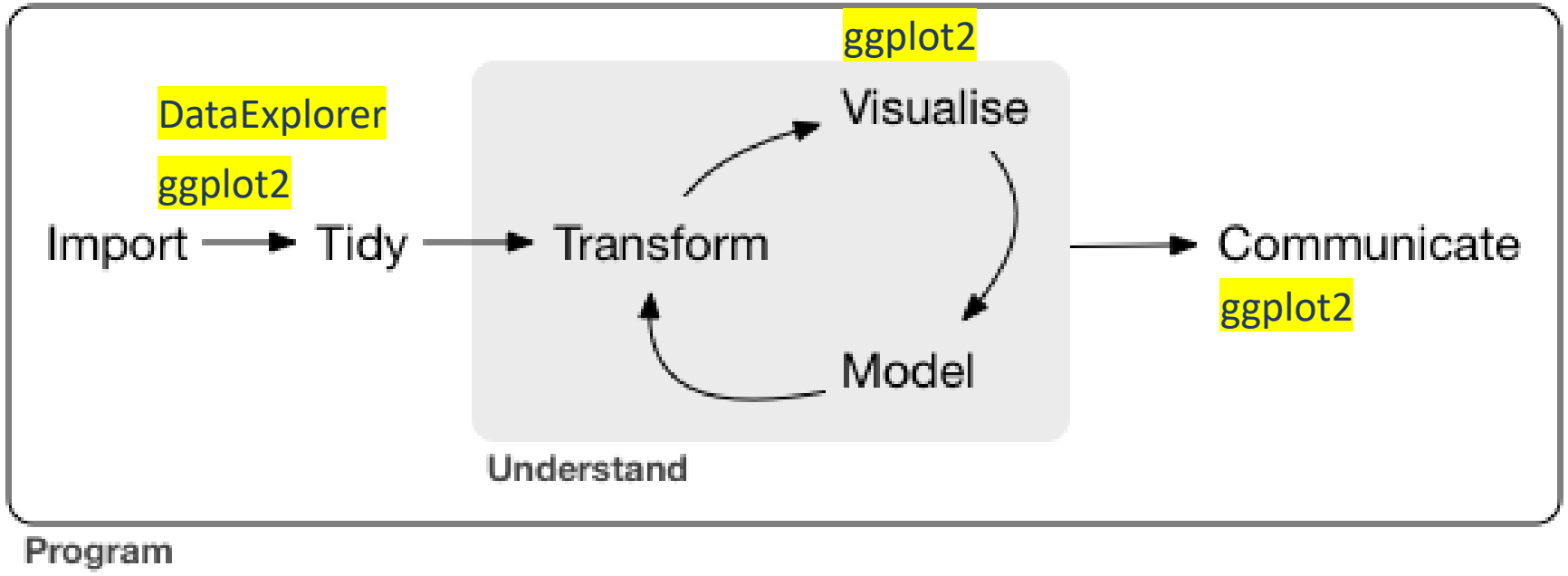
Person	Age	Sex	Height	Education
asdFg1	32	M	174.34	Degree

Data.Frame

Person	Age	Sex	Height	Education
asdFg1	32	M	174.34	Degree
Character	integer	Factor	double	Ordered factor

- Each column is a specific type (known as a class)
- Common class for data manipulation
- Other options such as data.table
- Different packages sometimes expect different classes
- Is a significant source of issues during programming.
- Is not a tibble but can be converted via as.tibble
- Tibbles can hold lists which are a linked set of different classes

R for DataScience



https://r4ds.had.co.nz/introduction.html

What is tidytext?

Using tidy data principles can make many text mining tasks easier, more effective, and consistent with tools already in wide use. Much of the infrastructure needed for text mining with tidy data frames already exists in packages like dplyr, broom, tidyr and ggplot2. In this package, we provide functions and supporting data sets to allow conversion of text to and from tidy formats, and to switch seamlessly between tidy tools and existing text mining packages.

https://cran.r-project.org/web/packages/tidytext/vignettes/tidytext.html

Approaches: Comparison, classification, prediction

- Context, context, context
- Aggregation levels
- Visualize, visualize, visualize
- Volumes
- Word Frequencies
- Word association
- Topic analysis via for example TD-IDF or more complex mechanisms
- NLP and Keyword extraction

Differences with a survey only approach

- Faster reaction time
- Opportunistic resources
- Quantitative
- Incremental (can add to the current methods)
- Requires data science expertise

Risks

- Quality of the constructs
- Biases in the dataset
- Missing data
- Systematic error
- Over representation of specific groups such as location or Education level
- Pre processing from the data provider

Context, context, context

Detecting Disruptive Talk in Student Chat-Based Discussion within Collaborative Game-Based Learning Environments: https://dl.acm.org/doi/10.1145/3448139.3448178

Table 1: Example utterances of disruptive and non-disruptive talk.

Class	Definition	Example
Disruptive Talk	Talk that generates frustration, annoyance, harming communication, or contributes to an increasingly bad mood among the group members	"I want to be right I'm gonna correct you I am right" "Um yea. yep, you can't work"
Non-Disruptive Talk	Normal talk that does not create bad moods among the group members.	"if [the fish] don't come to the top to breath then they are not going to breath at all" "It could also be in the water quality section"

Context, context, context

Detecting Disruptive Talk in Student Chat-Based Discussion within Collaborative Game-Based Learning Environments: https://dl.acm.org/doi/10.1145/3448139.3448178

Table 3: Top words from disruptive and non-disruptive talk.

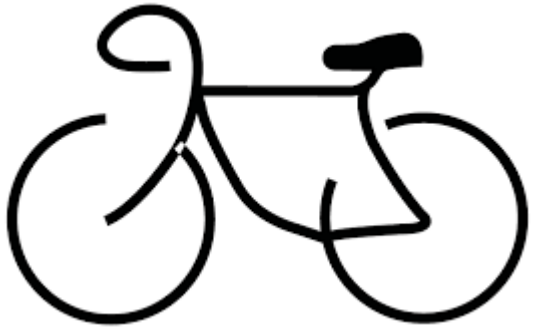
Rank	Disruptive Talk	Non-Disruptive Talk
1	"chat"	"note"
2	"enter"	"think"
3	"name"	"water"
4	"like"	"yes"
5	"boo"	"need"
6	"brother"	"column"
7	"nopy"	"food"
8	"yes"	"okay"
9	"nope"	"space"
10	"do"	"air"
11	Nickname1	"one"
12	Nickname2	"let"
13	"exit"	"fish"
14	"beep"	"say"
15	"exit"	"know"

Bigrams

Bigrams are two words that are next to each other. The following is a Table of the top 10 most frequent Bigrams for 2019-01-04 in the US job market as captured by Burning Glass. **Which of the bigrams suggest huddles?**

word	n	rank
job description	84219	1
education level	57514	2
level experience	56369	3
company information	56201	4
week salary	56097	5
shift shift	55994	6
experience license	55992	7
company phone	55977	8
information company	55966	9
company direct	55960	10

Exercise



- Look up on the Internet the following terms in relationship to text mining
  - Token
  - Feature
  - N gram
  - Lexical complexity
  - Dispersion
  - Attraction

Context

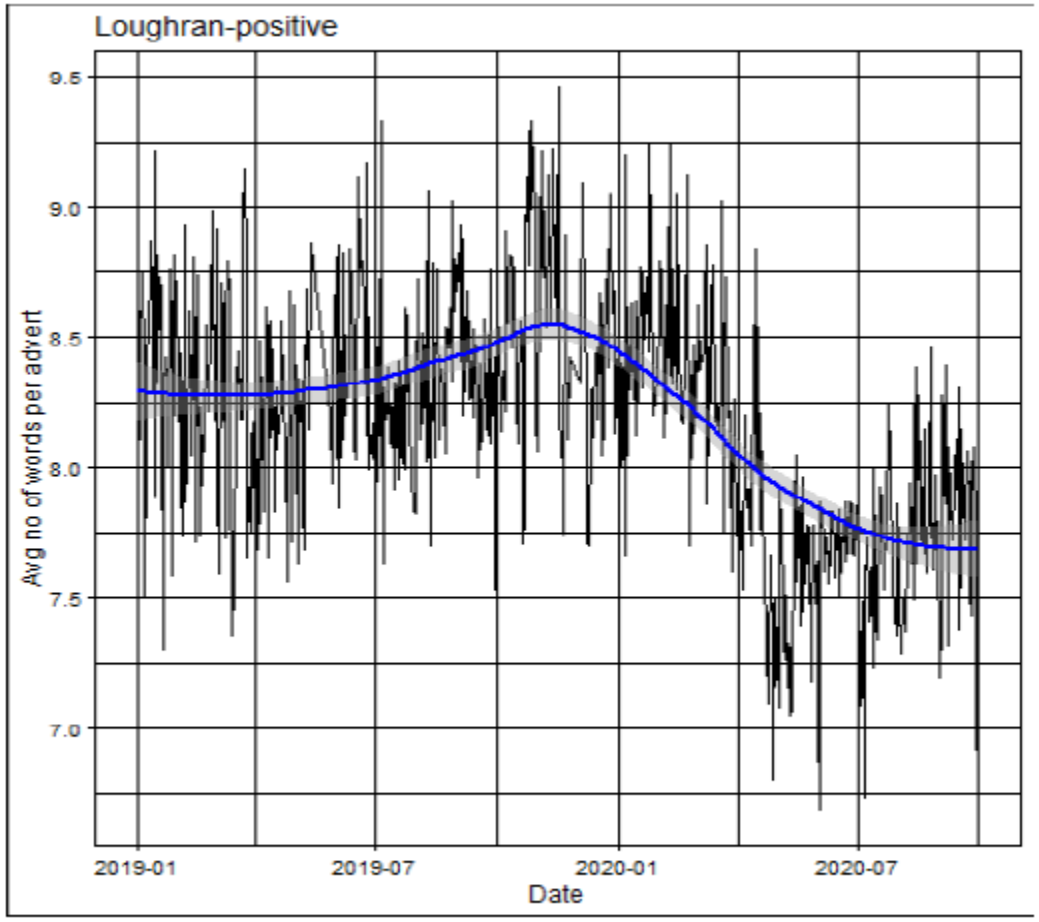


Figure 3: Positive sentiment in the US job market.

KWIC

religion , national origin , sex , sexual orientation ,	gender	identity or expression , pregnancy , age (40 and
upon race , religion , color , national origin ,	gender	(including pregnancy , childbirth , or related medical conditions
childbirth , or related medical conditions), sexual orientation ,	gender	identity , gender expression , hair style , age ,
related medical conditions), sexual orientation , gender identity ,	gender	expression , hair style , age , status as a

Confounders

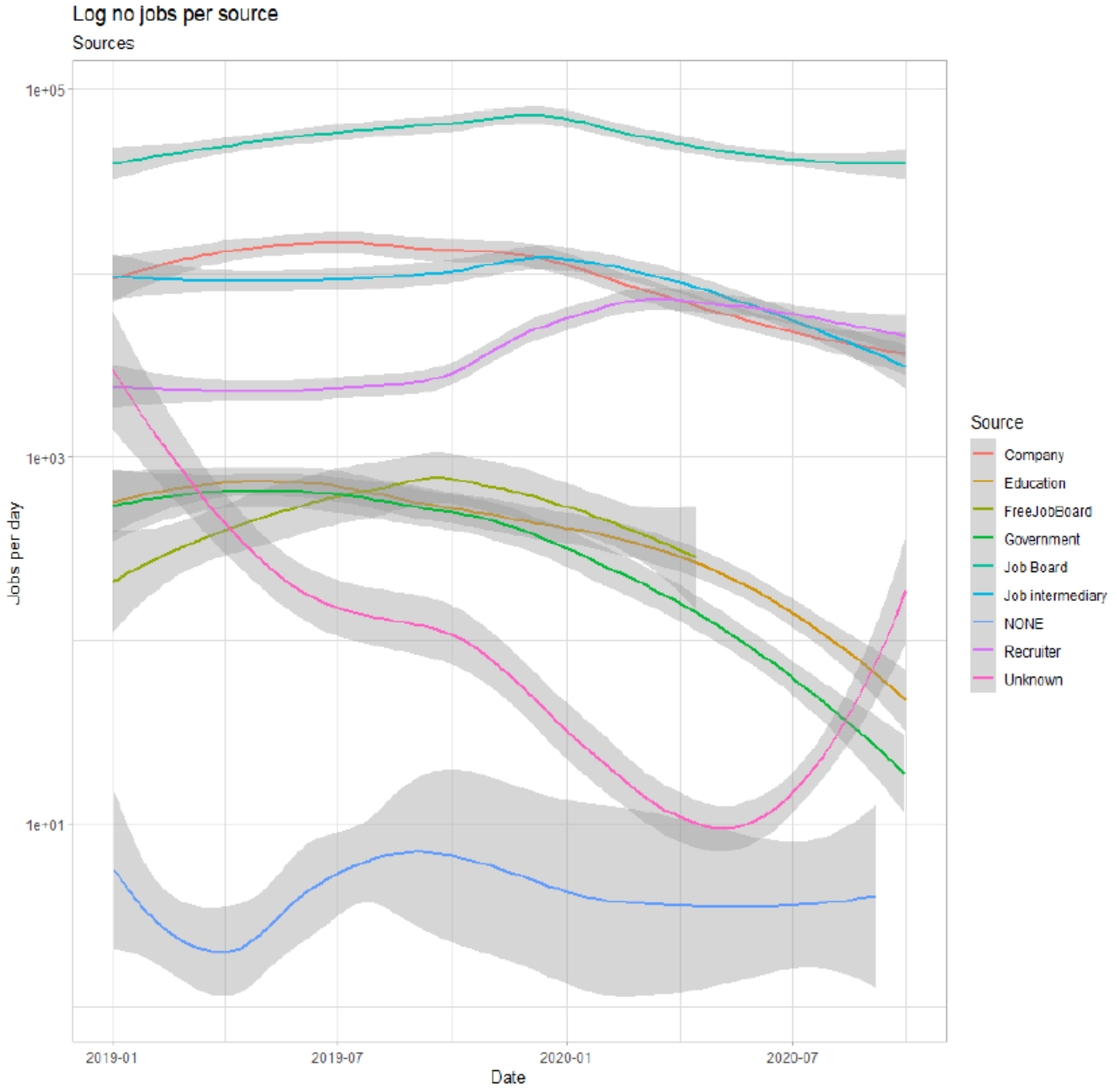
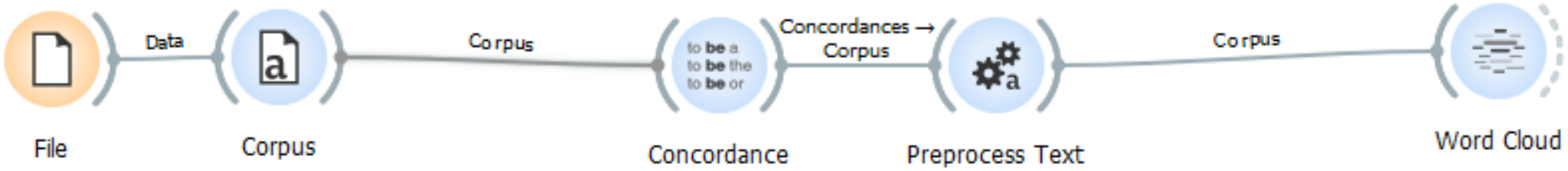


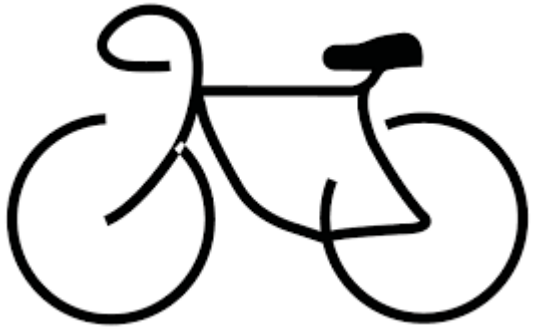
Figure 6: *Sources of Job advertisements*

Orange3

KWIC is popular so there are free tools such as Orange3 which allow you to import text files through a data pipeline that you can make by dragging and dropping. As an example, the following data pipeline imports a file, allows you to create KWIC and visualize. As you get used to the tool more complex pipelines are possible.



Exercise



Review the Orange3 website:
https://orangedatamining.com/workflows/

1. What is the functionality of the tool
2. What is the current version
3. What is a workflow
4. How can this tool help in analyzing text
5. Install Orange3
6. Follow a workflow such as the story arc

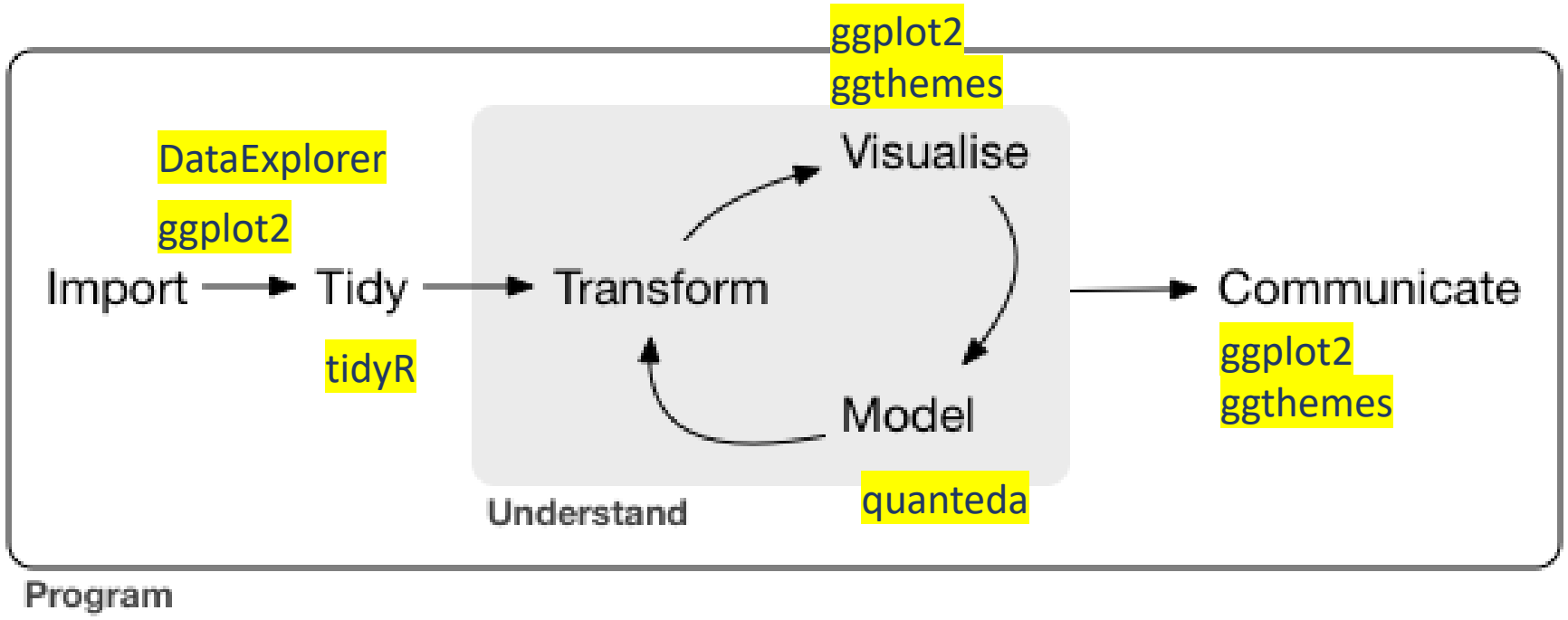
Review Code Book

```
sp(library(tidyverse))

# Load data
file.jobs <- "../../DATA/MonsterBoard-2013-n=20000.Rdata"
load(file.jobs)

# Custom dictionary, replace with your own
# In the topic notebook you have a recipe to divid into topics as well.
word <- c("young","age","race","disability","sexual","discriminate","ethnicity","families", "family","faith","abr
oad","barrier","creed","carer","home", "marital","she", "female","her","mother","minority","hate","care","support
ive","nurture","carer","local","helpful","social","parent","flexible","friendly")

custom.dic <-data.frame(word=word)
#custom.dic

# Place the job adverts into a tibble for easy manipulation
my.job <- tibble(Row = seq_along(sample$JobBody) , text = sample$JobBody)

# Load in stop words
data("stop_words")
stop_words <- rbind(stop_words,c("nbsp","Custom"))
stop_words <- rbind(stop_words,c(" ","Custom"))

# clean words
my.words <- my.job %>%  unnest_tokens(word, text) %>%  anti_join(stop_words)
my.words$EDU <- sample$EducationLevelID[my.words$Row]
my.words$EDU[my.words$EDU=="NULL"]<- 10
table(my.words$EDU)
```

Mentality

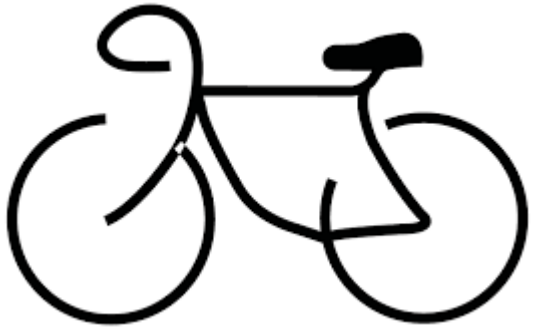
- **Be kind to yourself.** You do not need to understand every detail.
- Be patient.
- Learning does cost energy
- Read small chunks of code at a time
- Keep practicing reading code. Make notes on each function
- Small recipes can do a lot, so try and find those recipes
- The Internet is your friend
- If you want to learn R for the first time then try the following:
  - https://rafalab.github.io/dsbook/r-basics.html
  - https://bookdown.org/dli/rguide/

R for DataScience



https://r4ds.had.co.nz/introduction.html

Exercise



Review the output from the notebook .nb.html file and write in your notes the following.

1. List the packages used and describe their purpose
2. List the new functions used and what they do
3. Describe any short recipes in your own words
4. Search the Internet for at least two links for similar examples

Reading List

Review the links in the reading list for this module.
Write brief notes, answering the questions:

1. Do the links still work
2. Which links are relevant for you
3. Which links are not relevant for you
4. What further information would you of wished