E+ ADSEE

JobMarket Signalling
1. Business Problem Decomposition

ADSEE

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

Design of course

Audience: Non technical researchers who are interested in data science, who are time limited.

Aim: By the end of the course the author (**Dr. Alan Berg**) expects that the participant can read R code.

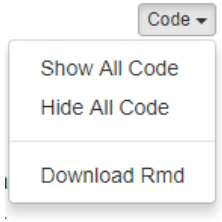
Method: Hands on approach to coding where the researcher gets used to reading code. No emphasis is placed on writing code, though code and useful links are provided. A Follow on course can later zoom into the details.

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 wish to learn R for the first time then try the following:
  - https://rafalab.github.io/dsbook/r-basics.html
  - https://bookdown.org/dli/rguide/

Course Prerequisites

- Web browser to read **.nb.html** files
- Rstudio installed
- Optional, drag and drop tooling
  - Orange3
  - Butter
- Patience, read code one line at a time



NOW WE DEMONSTRATE: Reviewing an HTML file with results from a web browser

Let’s Start

https://wallpapersafari.com/w/4mdrRF

Overview

It is important to analysis the problem that you want to solve for the business. However, there is a potential contradiction, a Data Science workflow is iterative in its nature. Therefore, you should feel free to review your business problem as new information is discovered.

In this module we will review Job Market Signalling and discuss gender issues.

The **R code** associated with this module deals with exploring the data and it's weaknesses as early in the cycle as possible. **The motivation is to discover issues early, where it is cheapest to solve.**

Data Science &Text Mining

Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data

Text Mining is the Process of gaining actionable insights from a large amount of unstructured text.

- **Job Descriptions** → Large Amounts of unstructured text
- **Signals** → Knowledge we wish to extract
- **Domain Expert** → has an understanding of which signals we want to extract and their context.

Text Input

1. More and more Big data sets
2. Can transcribe video and audio to text
3. Other dimensions such as demographic and location data when combined provides more context
4. Even simple methods such as counting the frequency of words can distinguish between categories
5. In general, more complex and newer methods generally provide better accuracy but cost most computational effort and time
6. No amount of Machine Learning can replace a domain expert in setting the right questions

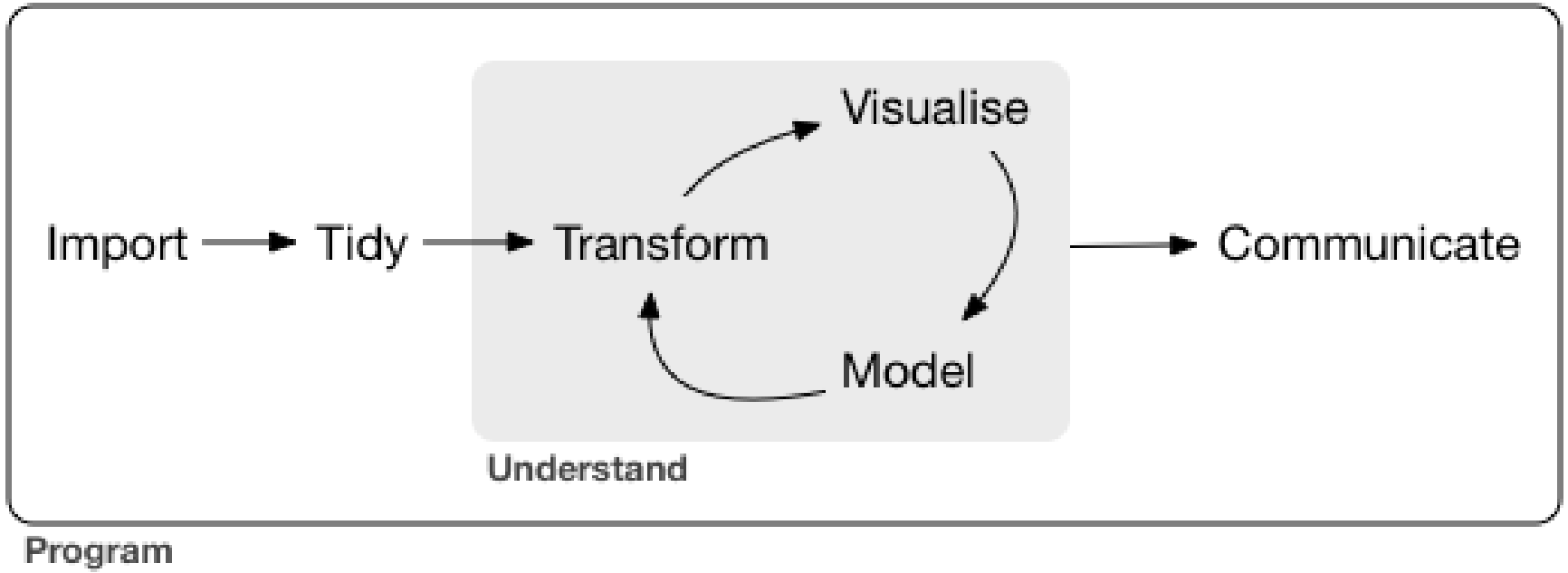
Text Mining Methods

- Rule based
  - Sentiment Analysis
  - Custom dictionaries
  - Stemming
- Supervised Learning
  - Characterization
  - Spam Not Spam
- Unsupervised Learning
  - Topic models
  - Clustering of similar documents

Boundaries are fuzzy
Many possible techniques
New techniques emerging
There is a significant role for the domain expert

Looks complex, but there are a lot of good packages available that allow you to perform techniques within only a few lines of code

R for DataScience



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

Why these methods?

Because we want to:
Discover the main properties of the data and understand what we can and cannot use.

LESSON: Signalling

In this lesson we will detail a number of different types of actionable information that can be derived from ***Big volumes*** of Job advertisements.

Big Data

Is a field that treats ways to analyse, systematically extract information from, or otherwise deal with data sets that are too large or complex to be dealt with by traditional data-processing application software.

Though there are cheats to cram analysis onto a Laptop computer if you are prepared to focus on simple methods and computers running while you sleep.

Burning Glass – US Job Market

Burning Glass spiders over 40,000 online job related websites and generated 300,000,000,000 words worth of Job descriptions for 60 million jobs for the two years 2019, 2020

CEDFOP

CEDFOP Skills Panorama brings insights on jobs and Skill requested in online job advertisements. More than 100 million of online job ads were collected and analysed, covering period of July 2018 till September 2020.

Benefits

(Turrell et al. 2019)

- One of the major benefits of using individual online job postings is that they are a direct measure of the economic activity associated with trying to hire workers.
- Another is the sheer volume they offer — These large numbers allow for very granular analysis.

Theory

A definitive work (Shannon 1948)
When transmitting information noise is introduced.

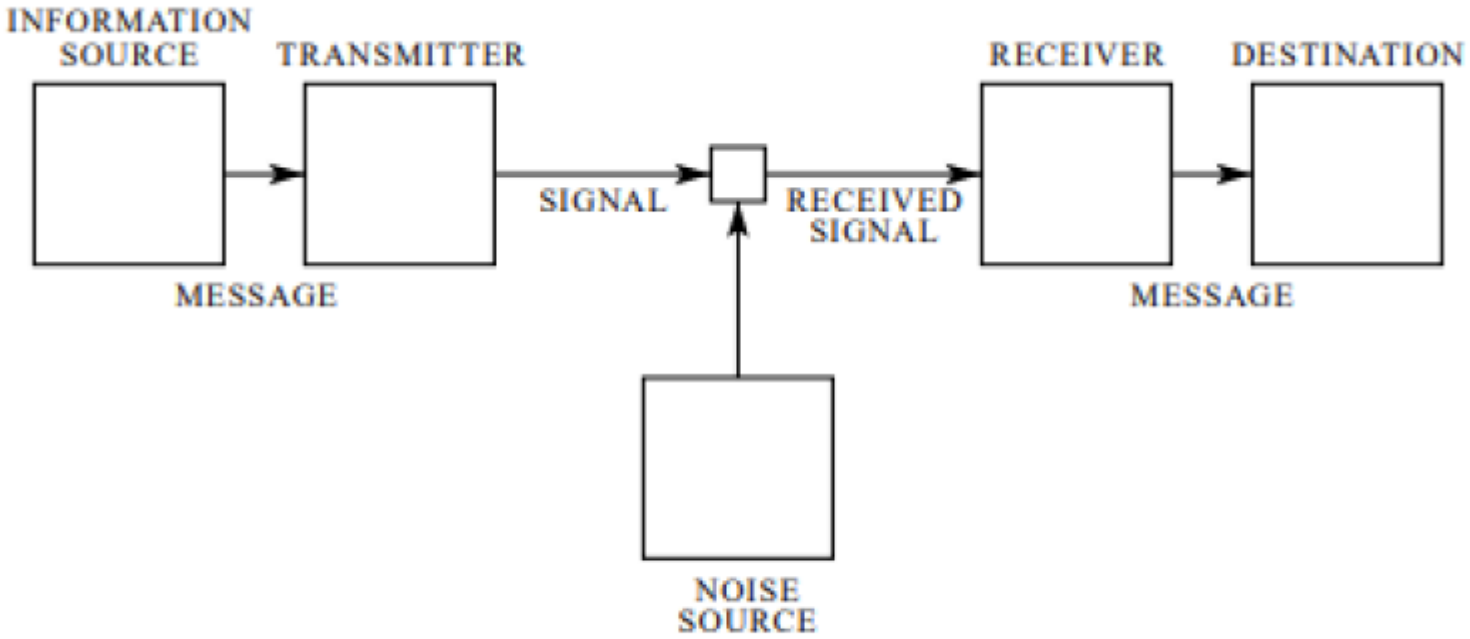
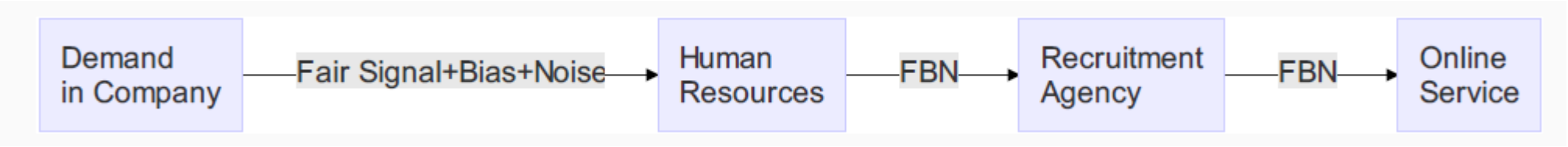


Fig. 1—Schematic diagram of a general communication system.

Noise & Bias

There is no one source of ***noise*** as we have more than one transmitter and receiver in the logistic chain for vacancy creation, formatting and transmission of the message.

In the end we are all human with ***our own*** cultural biases set within an ***organizations*** culture that are transmitted.



NOTE: The diagram is an over simplification and dependent on the aggregation of a vast number of workflows over the job market.

Noise

- Lots of noise and Bias
- Data is complex as many extra pieces of information are also included
- Data is missing or incorrect
- We do not see the whole of the job market
- Adverts have multiple Languages
- Duplication of adverts is possible
- Need for simplification and automated approaches

The majority of the work around a data science project is getting your data into a state that you can use it. Exploratory Data Analysis (EDA) enables you to understand the structural problems in your datasets

Bias by example

- Gender
- Race
- illegal activity
- Physical looks
- Barriers such as having a driving license when they are not needed.
- Salary
- Skills, e.g. unnecessary skills for the function
- Personality
- Education level
- Ageism

Signals: Job volumes, skill extraction.

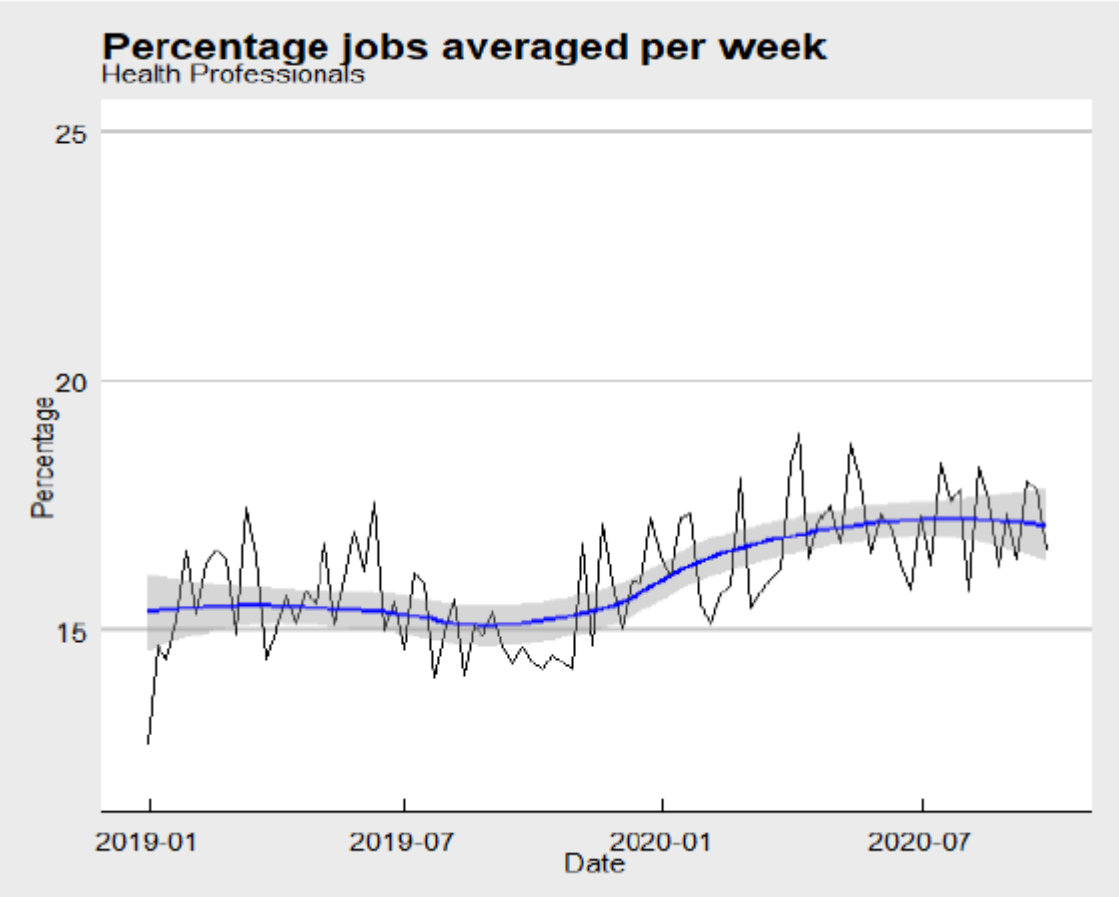
In this section we look at the research around signals available in job adverts

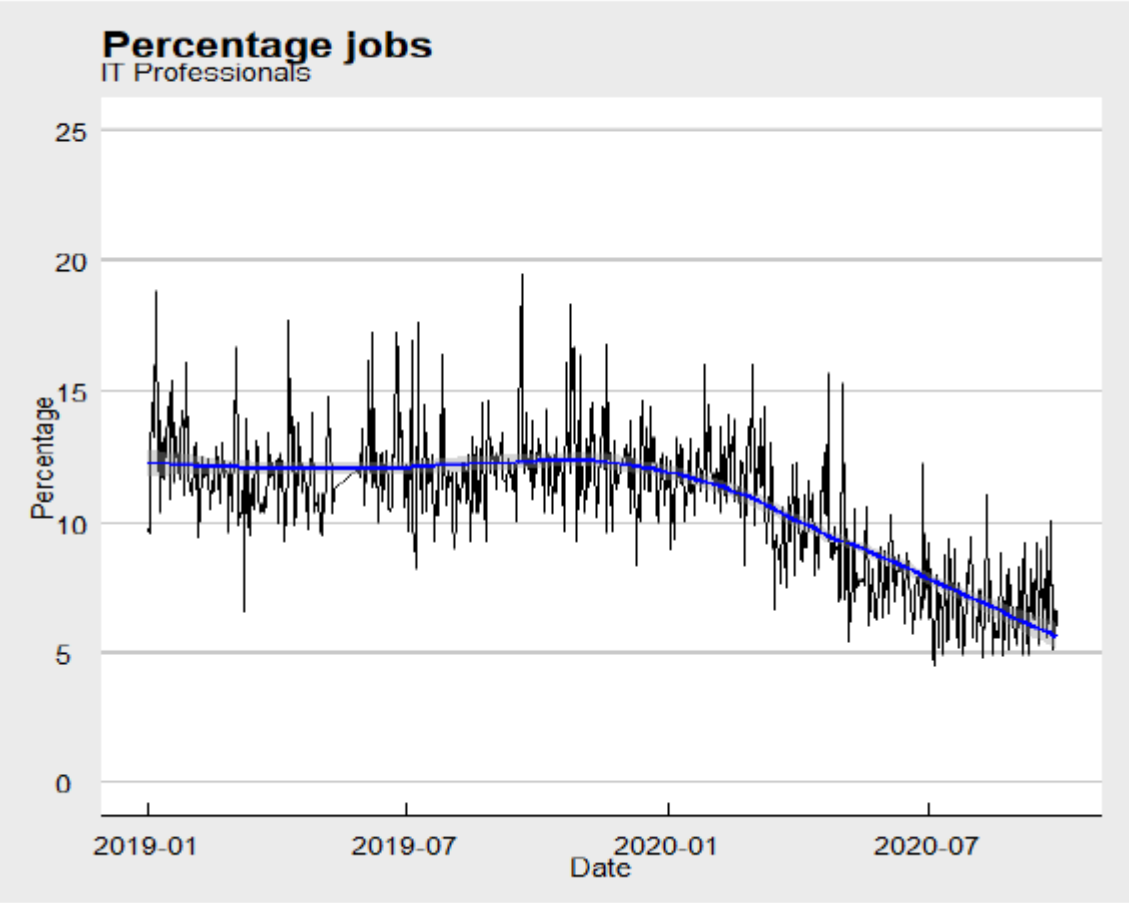
(Turrell et al. 2019)

Using a dataset of ***15 million UK job adverts*** from a recruitment website, we construct new economic statistics measuring labour market demand.

- Skill extraction
- Taxonomy of skills
- Volumes of job adverts
- Modelling
- Per Region

Volumes of Vacancies





Be careful of confounders

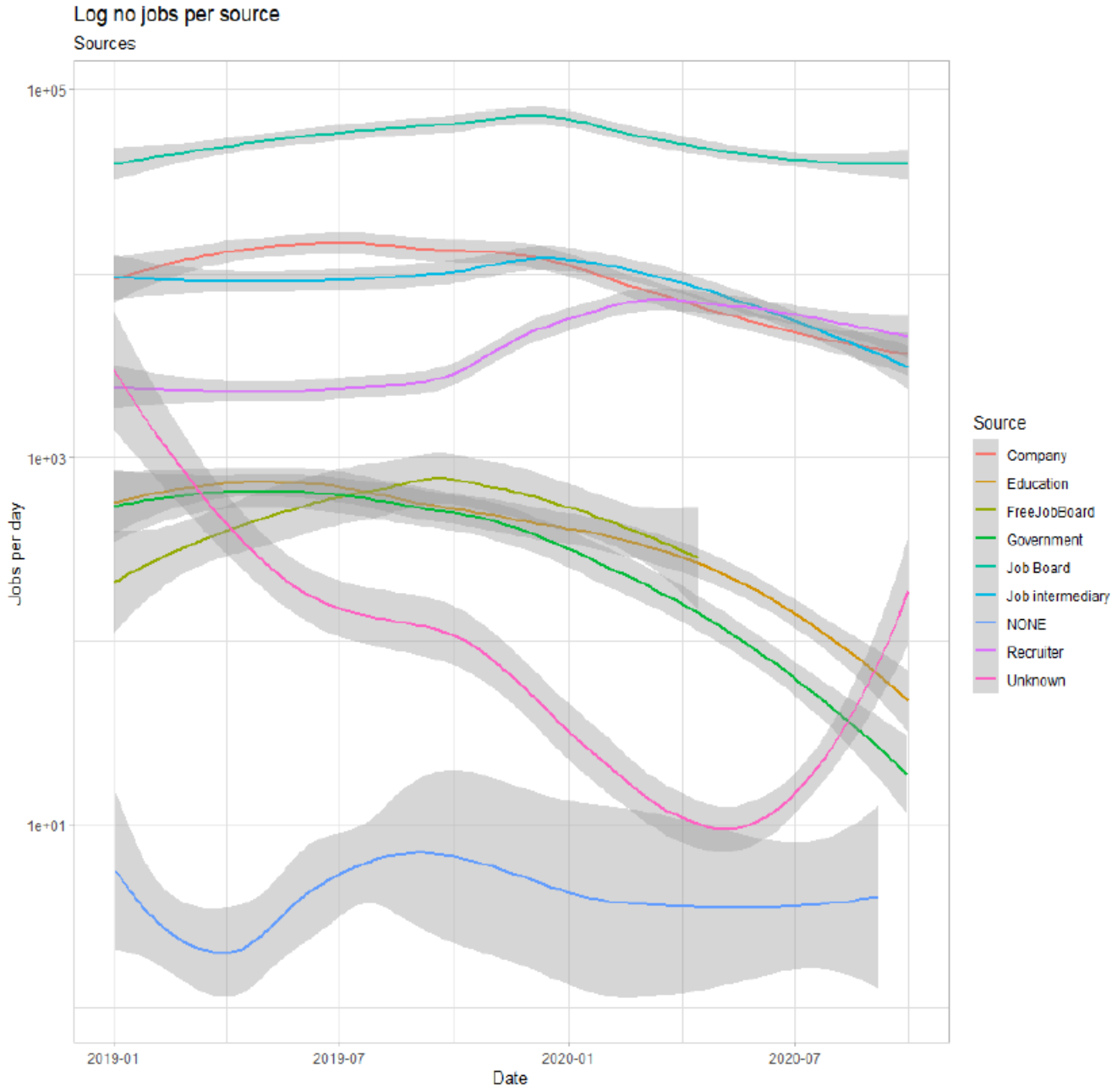


Figure 6: *Sources of Job advertisements*

Signals: Dictionary

Measurable by a **dictionary approach**
They has been evidence of bias in the use of gendered words in Job advertisements. (Gaucher, Friesen, and Kay 2011)

The Gaucher paper is an excellent source as:

- It provides a description of methodology to create dictionaries
- It provides a dictionary of biased wording

Although in a more recent study ("AN EXPLORATION OF GENDER BIAS IN INFORMATION TECHNOLOGY JOB ADVERTISEMENTS" 2020) suggests that they can find little evidence of the previously discovered bias in wording.

Here we find that experts and datasets do not always agree.

Signals: Job Titles

(Hodel et al. 2017)

Results show that gender-fair job titles were more frequent in more egalitarian countries with higher levels of socioeconomic gender equality our findings suggest that the language use in job advertisements indeed corresponds with linguistic, cultural, and socioeconomic aspects and may contribute to the transmission of gender (in)equalities and gender stereotypes.

Signals: Personality Traits

Measurable by a dictionary approach

Creative, Innovative, and Collaborative Librarians

1. Words such as ***"dynamic"*** or ***"passionate"*** may indicate a personality type. If the personality type is not the same as the candidate then they may not apply.
2. Majority of job advertisements chose to use personality traits to attract job Candidates
3. Although the Big Five dimension of conscientiousness has been linked to job performance, only 16% of job advertisements used personality traits in thatdimension

Signals: Soft Skills

Measurable by a dictionary approach

(Calanca et al. 2019)

• Our analysis is based on a dataset containing 245,000 job advertisements (ads) from the United Kingdom (UK)

• We find that soft skills are a crucial component of job ads, especially of low-paid jobs and jobs in female-dominated professions

• The three most distinctive skills for Teaching are ***enthusiastic, dedicated, professional***, whereas for Accounting & Finance they are ***accurate, responsible, analytical*** abilities.

Signals: Crime

(Volodko, Cockbain, and Kleinberg 2019)

Human trafficking offenders may use the Internet to recruit their victims.

While there may be value in screening job advertisements to identify potential labour trafficking and exploitation, **additional information** is needed to ascertain actual labour trafficking.

Enrichment

Signals: Ageism

(Burn et al. 2021)

This research focuses on measuring employer behavior specifically, whether there is less hiring of qualified older workers. We find that language classified by the machine learning algorithm as closely related to ageist stereotypes is perceived as ageist by experimental subjects.

Key points

- Burning Glass and CEDFOP are examples of suppliers of Big volumes of vacancies
- Data tends to be dirty and costs time and effort to clean up.
- Exploratory Data Analysis can help you understand the state of your data.
- Domain experts have managed to extract signals out of vacancies.
- Methods included:
  - A dictionary approach
  - Review job titles
  - Entity extraction
  - Machine Learning for characterisation
  - Modelling of regional differences
  - Looking at volumes over time
- Experts do not always agree
- Data can be enriched to help find signals

References

"AN EXPLORATION OF GENDER BIAS IN INFORMATION TECHNOLOGY JOB ADVERTISEMENTS." 2020. *Issues In Information Systems*. https://doi.org/10.48009/3_iis_2020_189-199.

Burn, Ian, Daniel Firoozi, Daniel Ladd, and David Neumark. 2021. "Machine Learning and Perceived Age Stereotypes in Job Ads: Evidence from an Experiment." https://doi.org/10.3386/w28328.

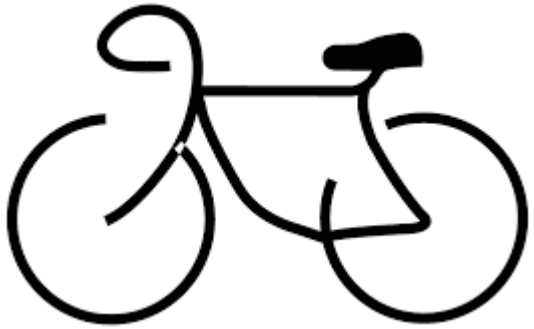
Calanca, Federica, Luiza Sayfullina, Lara Minkus, Claudia Wagner, and Eric Malmi. 2019. "Responsible Team Players Wanted: An Analysis of Soft Skill Requirements in Job Advertisements." *EPJ Data Science* 8 (1). https://doi.org/10.1140/epjds/s13688-019-0190-z.

Gaucher, Danielle, Justin Friesen, and Aaron C. Kay. 2011. "Evidence That Gendered Wording in Job Advertisements Exists and Sustains Gender Inequality." *Journal of Personality and Social Psychology* 101 (1): 109–28. https://doi.org/10.1037/a0022530.

Hodel, Lea, Magdalena Formanowicz, Sabine Sczesny, Jana Valdrová, and Lisa von Stockhausen. 2017. "Gender-Fair Language in Job Advertisements." *Journal of Cross-Cultural Psychology* 48 (3): 384–401. https://doi.org/10.1177/0022022116688085.

Shannon, C. E. 1948. "A Mathematical Theory of Communication." *Bell System*

Exercise



Which labour market themes are mentioned on the CEDEFOP and Burning Glass websites?

How are the themes relevant within your context?

Codebook

```
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/

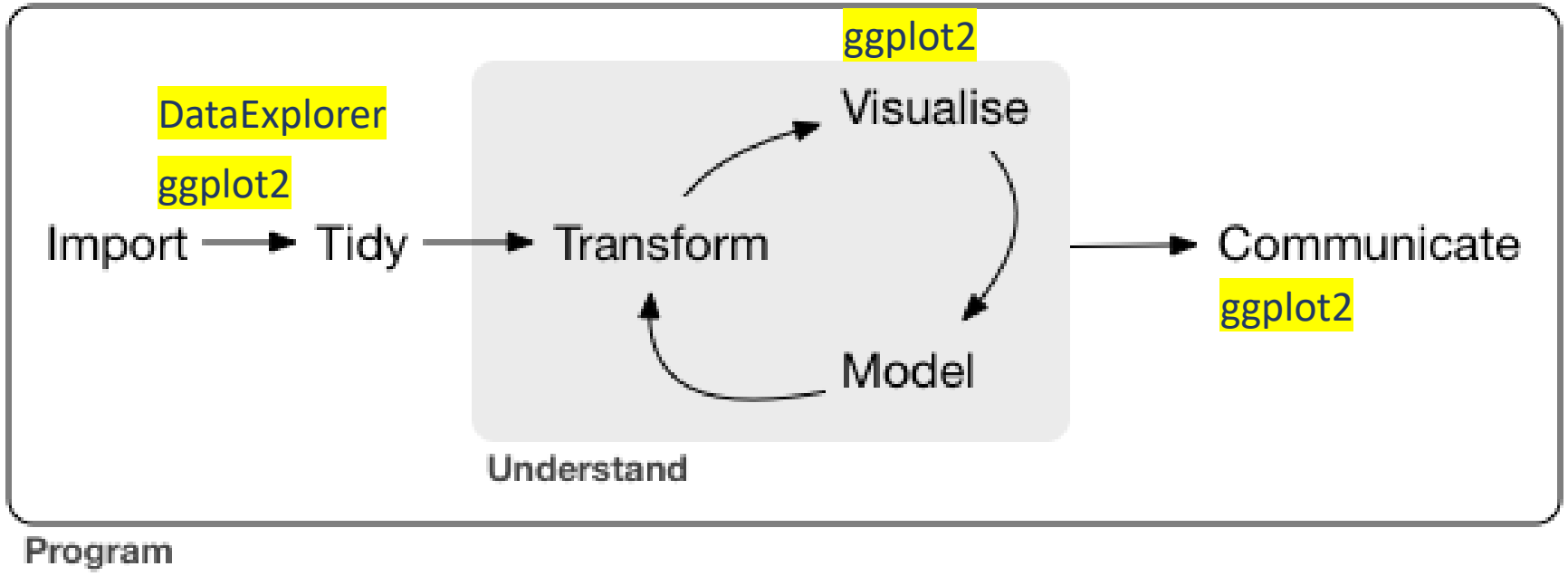
Exploratory Data Analysis Slides and code

EDA is an iterative cycle. You:

1. Generate questions about your data.
2. Search for answers by visualising, transforming, and modelling your data.
3. Use what you learn to refine your questions and/or generate new questions.

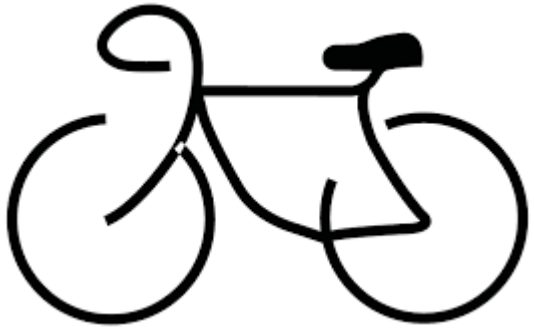
EDA is not a formal process with a strict set of rules. More than anything, EDA is a state of mind. During the initial phases of EDA you should feel free to investigate every idea that occurs to you. Some of these ideas will pan out, and some will be dead ends. As your exploration continues, you will home in on a few particularly productive areas that you'll eventually write up and communicate to others.

R for DataScience



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

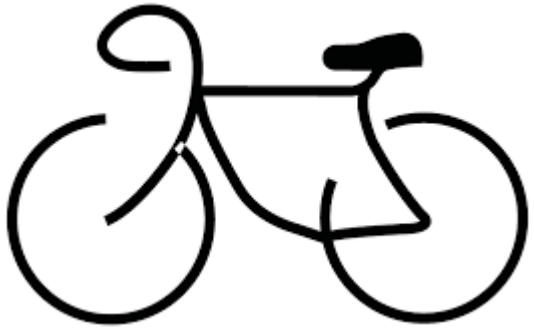
Exercise



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

Exercise



1. Review the EDA output from the notebook.
2. Write in words the main opportunities and issues are related to which fields starting with the **JobBody** and **postcode** fields. Feel free to copy images into your report.
3. Note the top six fields that are most useful for your field of study and why
4. Conclude with a short summary of what would be needed to improve the data

Expected length is approximately 1 page of A4 of written word.
The graphics do not count as part of the final length.

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