EXPLORING DATA SCIENCE WITH »R«

Created: 2021-05-19 Mi 19:47

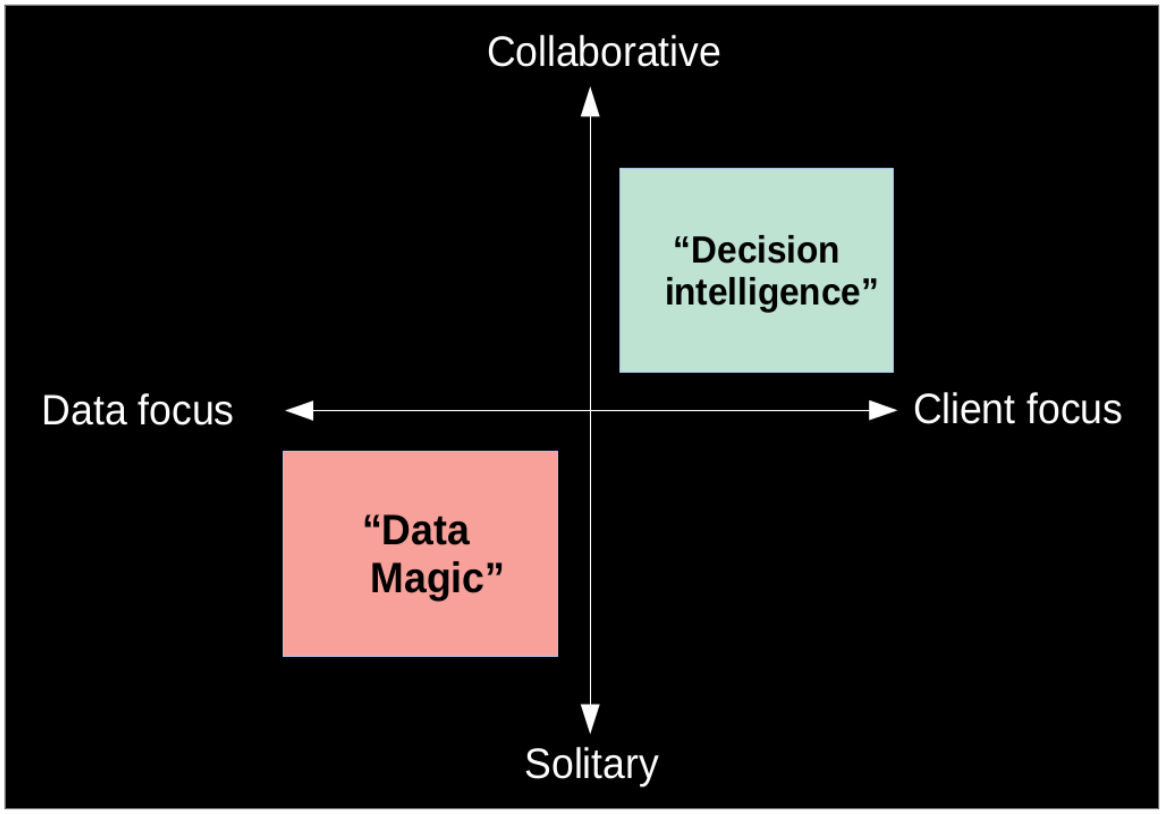
WHAT WILL YOU LEARN?

- Two ways of looking at data
- Data science $\leftarrow$ modeling $\cup$ coding
- Data science examples using R
- R vs. Python for data science

WHAT CAN YOU DO?

- Answer polls, like "**do you know R?**"
- Leave comments/questions in the chat
- Download the slides (PDF) **here**
- Learn R, e.g. **here**

TWO WAYS OF LOOKING AT DATA



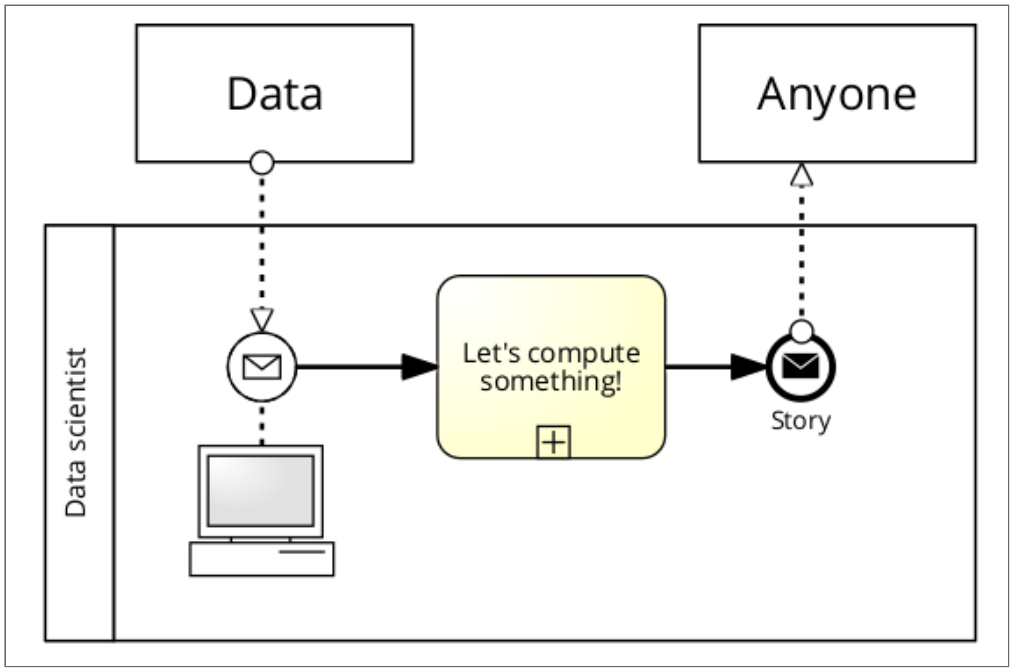
SCENARIO 1: "DATA MAGIC"

- We've got data!
- We've got computers!
- Let's compute something!

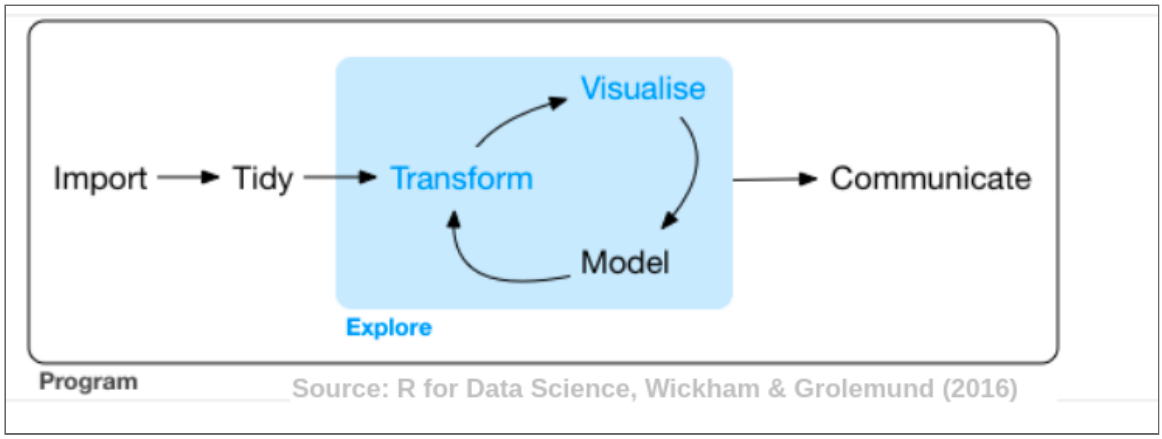
GOOD QUESTION, BAD SETUP

What's the story behind the data?

"DATA MAGIC" WORKFLOW



EXPLORATORY DATA ANALYSIS



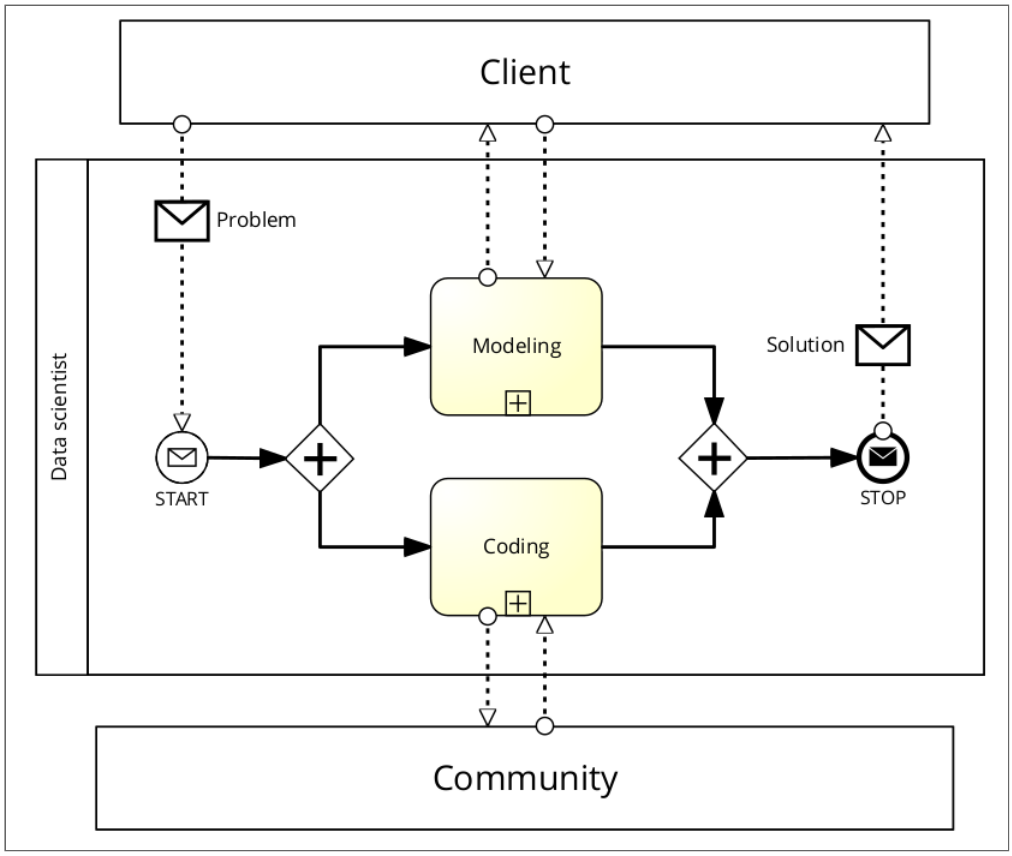
SCENARIO 2: "DECISION INTELLIGENCE"

- A client got a problem!
- A client needs a solution!

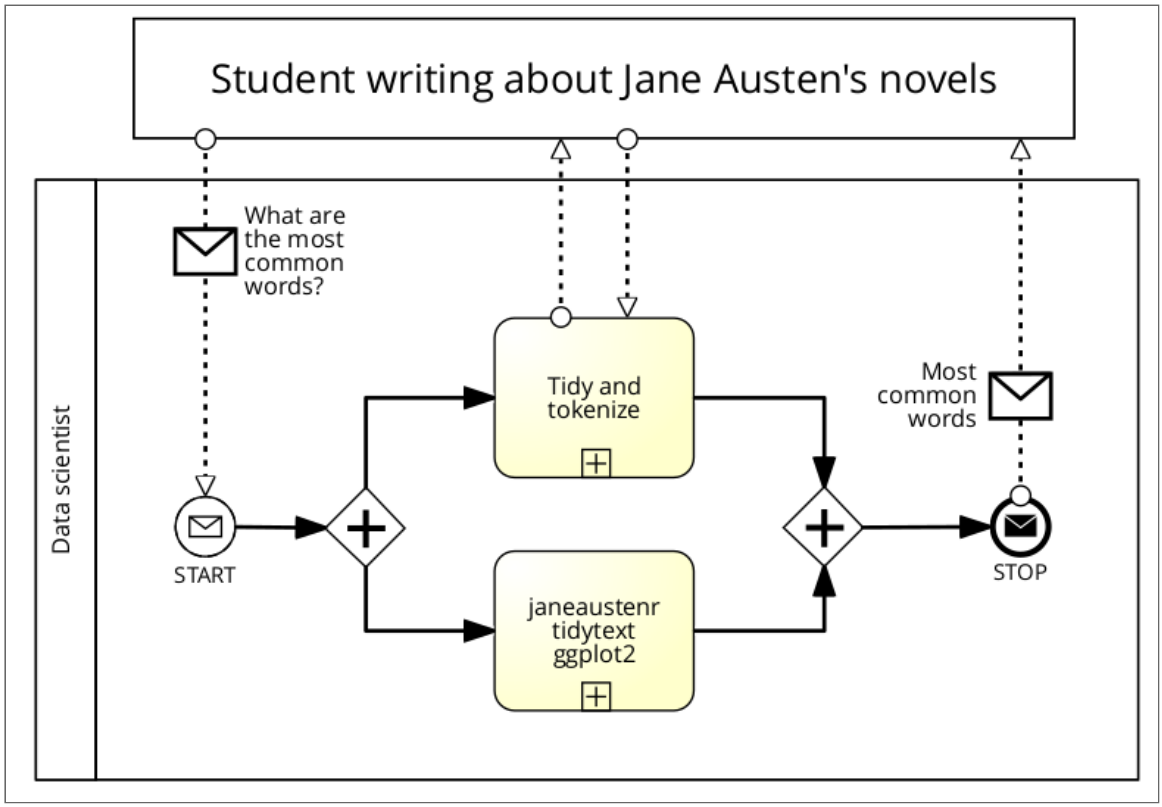
BETTER QUESTION, GOOD SETUP

What are your options?

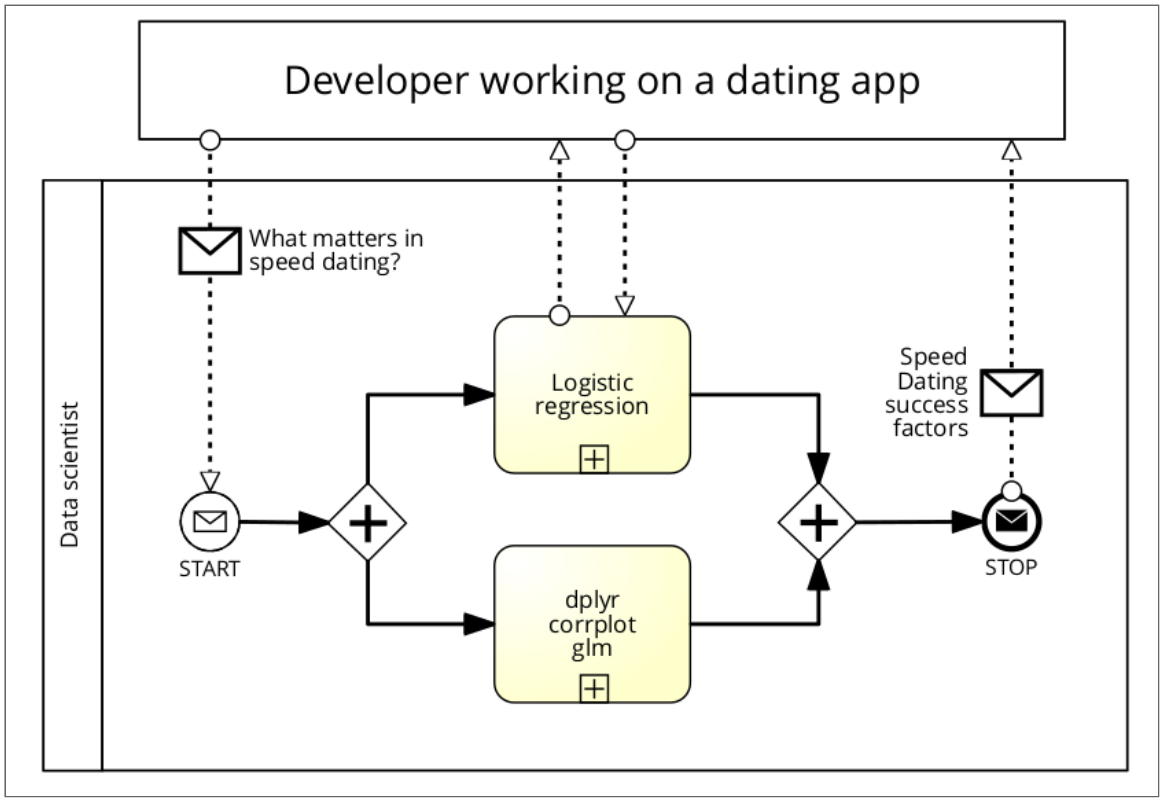
PROBLEM-BASED COLLABORATIVE EXPLORATORY DATA ANALYSIS



EXAMPLE: TEXT MINING



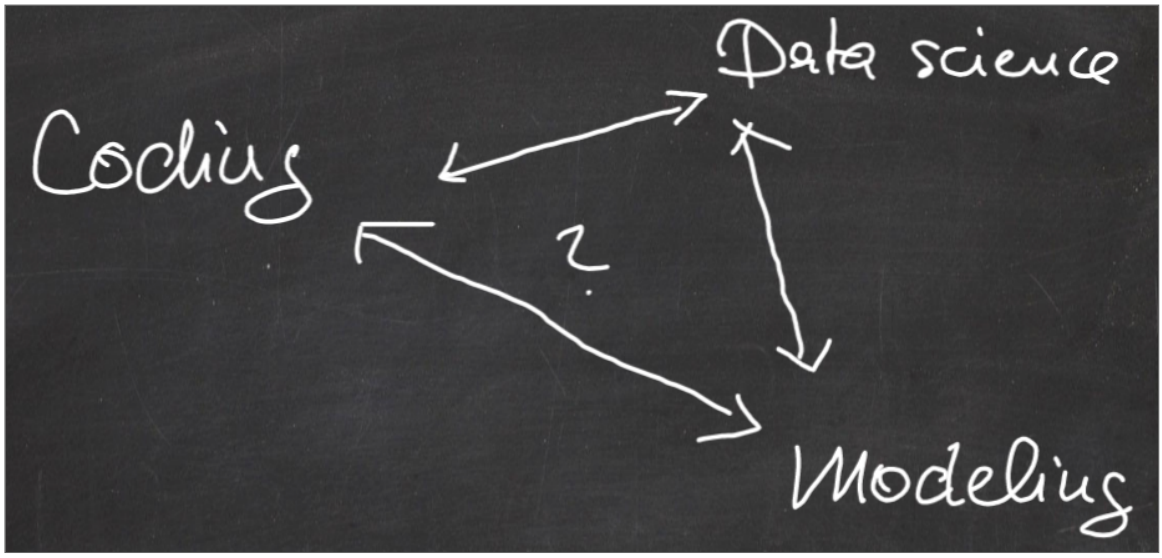
EXAMPLE: SPEED DATING



SUMMARY

- "Data magic" vs. "decision intelligence"
- Data/solitary vs. Client/collaborative
- Text mining and recommender systems
- Modeling/coding as collaborative activities

DS $\leftarrow$ MODELING $\cup$ CODING



CODING SKILLS

- Algorithmic thinking
- Fixed at run-time
- Procedure oriented
- Depends on data structures
- Object-oriented programming (OOP)
- Functional Programming (FP)

MODELING SKILLS

- Heuristic thinking
- Adaptive at run-time
- Pattern oriented
- Cognitive bias
- Analogy, induction
- Petri nets, BPMN, UML

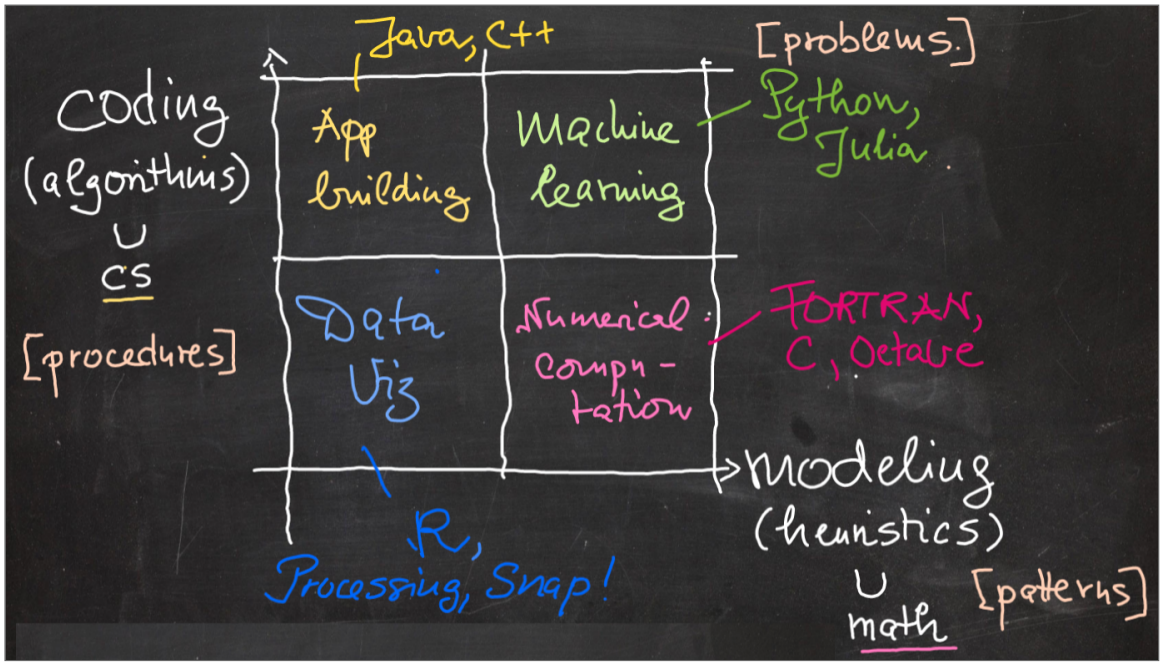
FOUR SCENARIOS

- Data Visualization
- Application Building
- Numerical Computation
- Machine Learning

PROGRAMMING LANGUAGES

- Apps: Java, C++
- Numerics: FORTRAN, C, Octave
- DataViz: R / Python, Snap!, Processing
- Machine Learning: Python / R, Julia

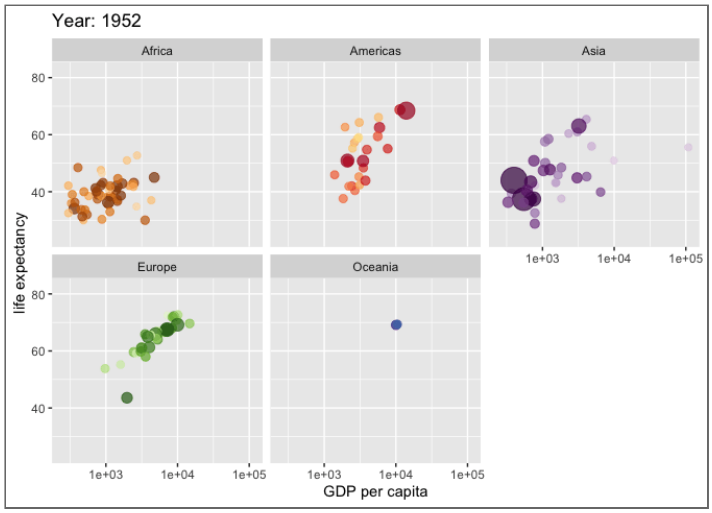
DATA SCIENCE FRAMEWORK



SUMMARY

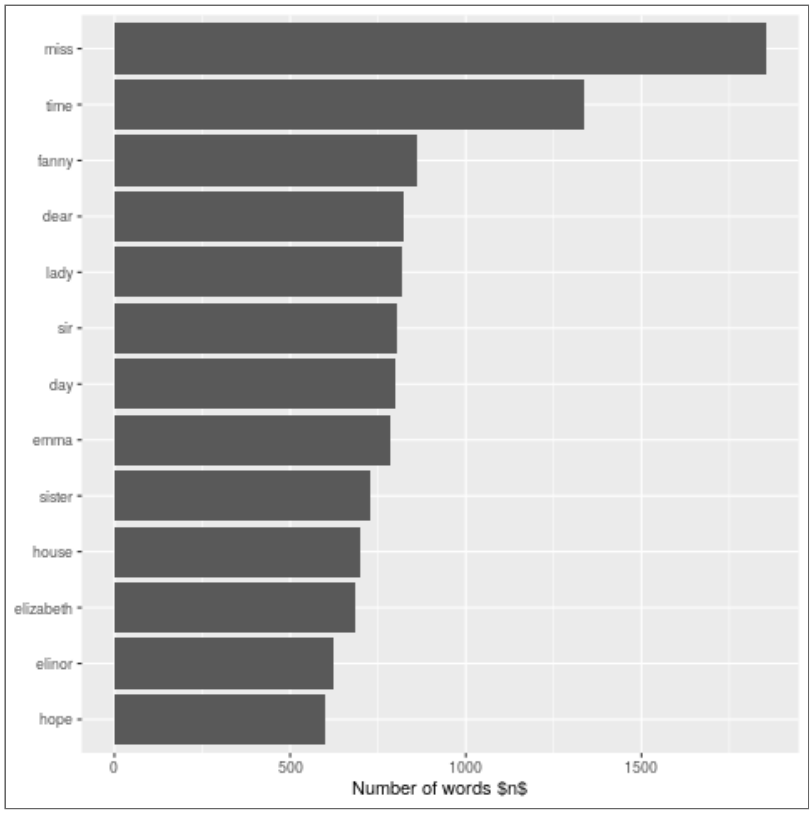
- Coding is algorithmic, modeling is heuristic
- Paradigms: OOP/FP, BPMN/UML
- Scenarios: DataViz, Apps, Numerics, ML
- Languages: R, Python, and many more
- Two pathways: CS/coding vs. DS/modeling

DATA SCIENCE EXAMPLES USING R



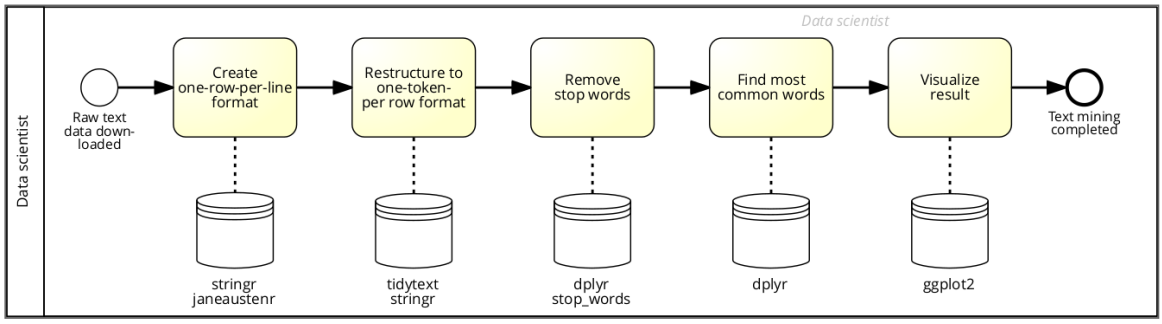
- Text mining Jane Austen
- What matters in speed dating
- Current student projects

TEXT MINING JANE AUSTEN



The most common words in Jane Austen's novels

TEXT MINING JANE AUSTEN: STEPS



BPMN diagram showing required R libraries

TEXT MINING JANE AUSTEN: CODE

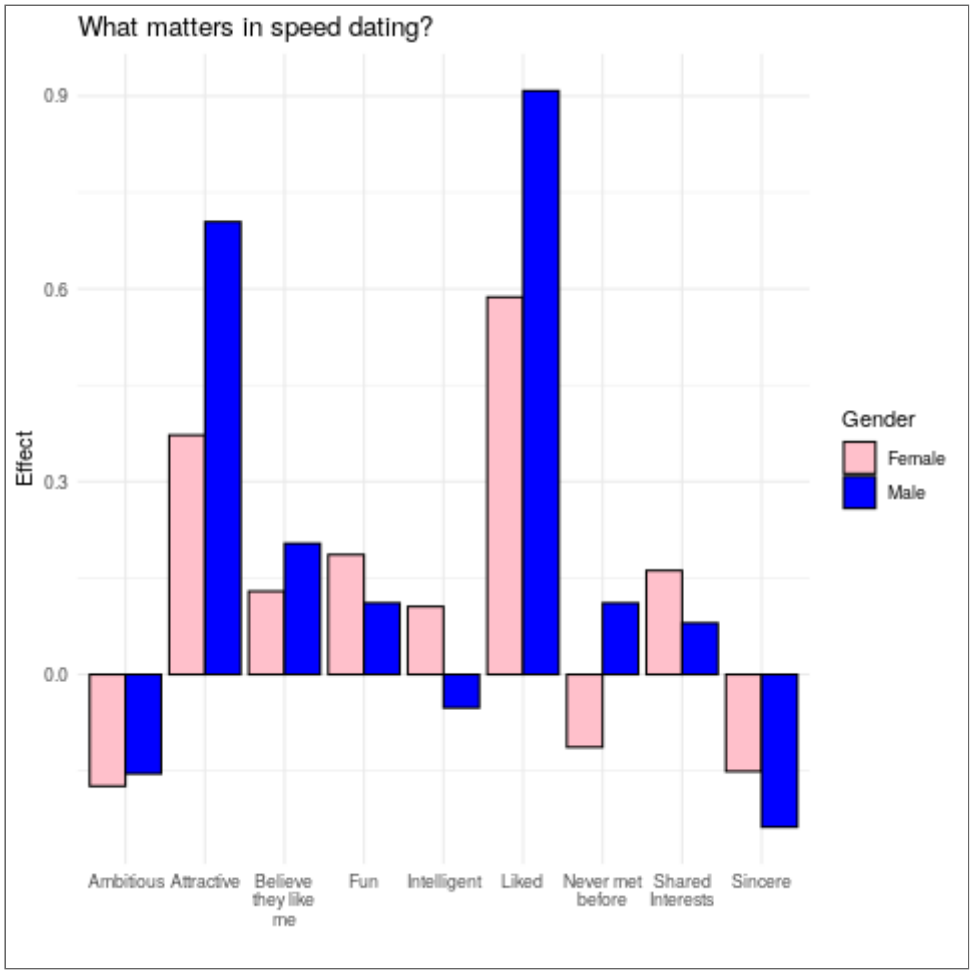
```
library(janeaustenr)
library(dplyr)
library(stringr)
library(tidytext)
library(ggplot2)
data(stop_words)

original_books <- austen_books() %>%
    group_by(book) %>%
    mutate(linenumber = row_number(),
           chapter = cumsum(str_detect(text,
                                       regex("^chapter [\\divxlc]",
                                             ignore_case = TRUE)))) %>%
    ungroup()

tidy_books <- original_books %>%
    unnest_tokens(word, text)

tidy_books <- tidy_books %>%
    anti_join(stop_words)

tidy_books  %>%
    count(word, sort = TRUE)

tidy_books %>%
    count(word, sort = TRUE) %>%
    filter(n > 600) %>%
    mutate(word = reorder(word, n)) %>%
    ggplot(aes(n, word)) +
    geom_col() +
    labs(x="Most common words in Jane Austen's novels",y = NULL)
```

Source: Silge & Robinson (2017)

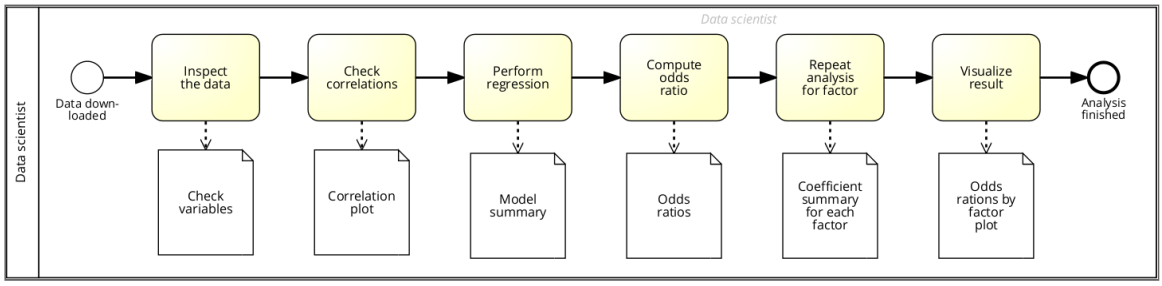
REAL WORLD APPLICATION

Arxiv Sanity Preserver (recommender)

WHAT MATTERS IN SPEED DATING



WHAT MATTERS IN SPEED DATING: STEPS



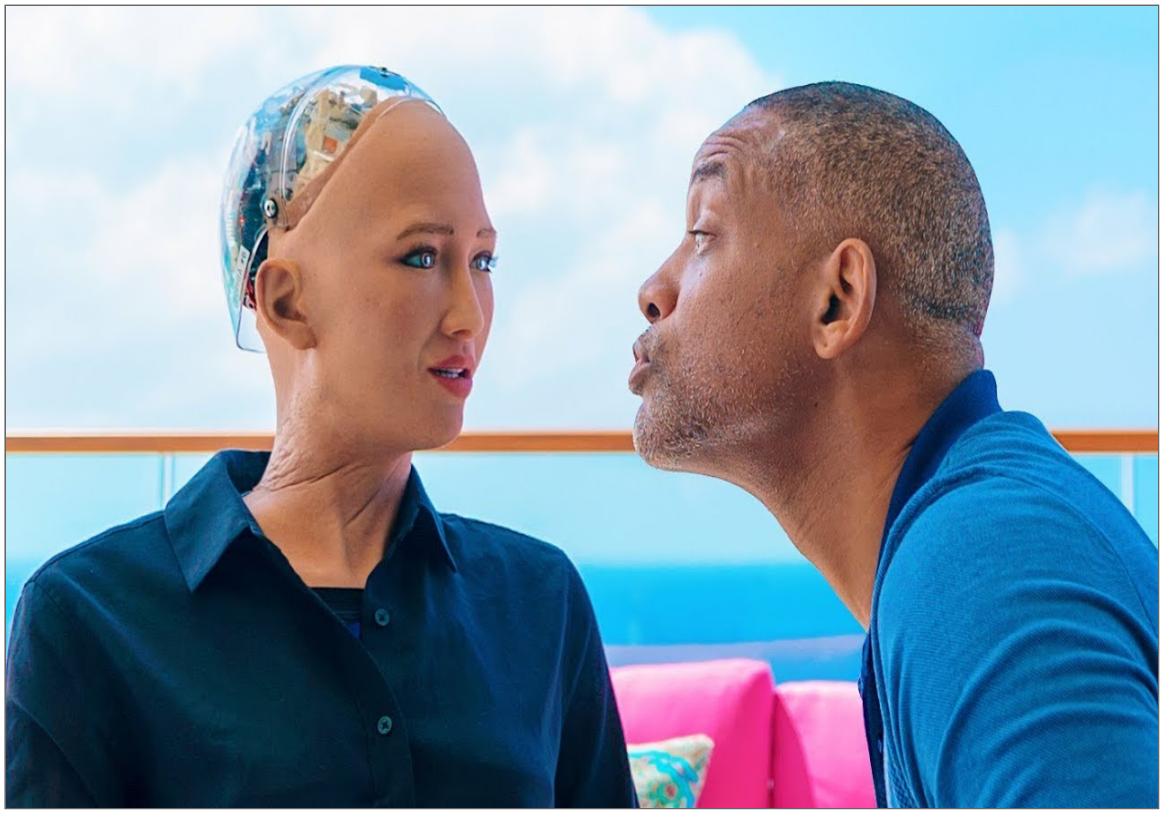
BPMN diagram showing process & output

WHAT MATTERS IN SPEED DATING: CODE

```
library(tidyverse)
library(corrplot)

download.file(
    "https://raw.githubusercontent.com/keithmcnulty/speed_dating/master/speed_data_data.RDS",
    "speed_dating_data.RDS")
data <- readRDS("speed_dating_data.RDS")

head(as.data.frame(data,3))

corr_matrix <- data %>%
    dplyr::select(attr, sinc, intel, fun, amb, shar, like, prob, met) %>%
    as.matrix()

M <- cor(corr_matrix, use="complete.obs")
corrplot::corrplot(M)

model_m <- glm(dec ~ attr+sinc+intel+fun+amb+shar+like+prob+met,
                data = data %>% dplyr::filter(gender == 1),
                family = "binomial")
ctable_m <- coef(summary(model_m))
odds_ratio_m <- exp(coef(summary(model_m))[ , c("Estimate")])
coef_summary_m <- cbind(ctable_m,
                        as.data.frame(odds_ratio_m,
                                      nrow = nrow(ctable_m),
                                      ncol = 1))
chart_data <- coef_summary_f %>%
    tibble::rownames_to_column() %>%
    dplyr::left_join(coef_summary_m %>%
                     dplyr::add_rownames(), by = "rowname") %>%
    dplyr::select(rowname, odds_ratio_f, odds_ratio_m) %>%
    tidyr::pivot_longer(cols = c("odds_ratio_f", "odds_ratio_m"),
                        names_to = "odds_ratio") %>%
    dplyr::mutate(Effect = value - 1,
                  Gender = ifelse(odds_ratio == "odds_ratio_f", "Female", "Male"),
                  Factor = dplyr::recode(rowname,
                                         amb = "Ambitious", attr = "Attractive",
                                         fun = "Fun", intel = "Intelligent",
                                         like = "Liked", met = "Never met\nbefore",
                                         prob = "Believe\nthey like\nme",
                                         shar = "Shared\nInterests",
                                         sinc = "Sincere"))

ggplot(data = chart_data %>% dplyr::filter(rowname != "(Intercept)"),
       aes(x=Factor, y=Effect, fill=Gender)) +
    geom_bar(stat="identity", color="black", position=position_dodge()) +
    theme_minimal() +
    labs(x="", title = "What matters in speed dating?") +
    scale_fill_manual(values=c('#FFC0CB', '#0000FF'))
```

Source: McNulty (2020)

REAL WORLD APPLICATION



"Will Smith Tries Online Dating"

CURRENT STUDENT PROJECTS

- Avocado sales in different US cities
- Rental prices in German cities
- What is my Netflix consumption like?
- Popularity of different US bills
- Lifestyle habits and weight issues
- Musical consumption during a pandemic
- Influence of Elon Musk tweets on bitcoin
- Influence of Queens Gambit on Chess.com

WHICH LANGUAGE?

Data visualization	Numerical computation	Application building	Machine learning
R	FORTRAN	Java	R
Snap!	GNU Octave	C++	Python
Processing	C	Kotlin	Julia

R VS. PYTHON

- Product
- Popularity
- Prediction
- Proposal

PRODUCT - DIFFERENCES

	R [4.1.0]	Python [3.9.5]
Release:	1991 (1976)	2008 (1989)
Written in:	FORTRAN, C, R	C
Purpose:	Math & Statistics	Productivity
License:	GPL	**PSF**
Libraries:	17,634 (**CRAN**)	137,000 (**PyPI**)
Vectors:	Start at `1`	Start at `0`
Typedness:	Weak	Strong

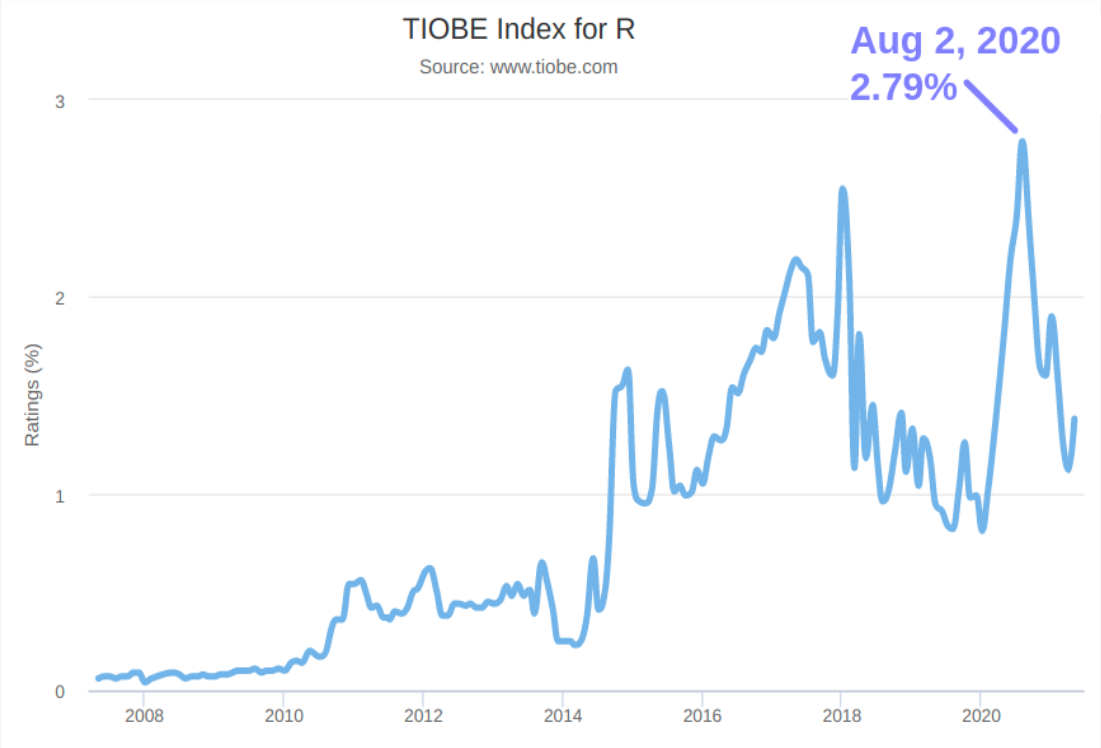
PRODUCT - SIMILARITIES

Access:	Free	Easy to learn
Paradigms:	OOP	Functional
Memory:	Dynamical[14]	Garbage[15]
Translation:	Interpreter	REPL/shell
Analytics:	Visualization	Machine Learning
Creativity:	Packages	Portability
Interfaces:	Shiny/Django	SQL/SQLite

POPULARITY - SEARCH

Highest Position (since 2007): #8 in Aug 2020

Lowest Position (since 2007): #73 in Dec 2008



POPULARITY - NO CONTEST

	R	Python
TIOBE Index	1.38%	11.74%
Loved (Dev)	44.5%	66.7%
Dreaded (Dev)	55.5%	33.3%
Wanted (Dev)	5.1%	30.0%
Salary (US)	$109k	$120k
Salary (World)	$59k	$57k

PREDICTION - THE WAR IS OVER!

PROPOSAL - WHAT YOU SHOULD DO

- R
- R + Python
- R + Python + Java
- R + Python + Java + C++
- R + Python + Java + C++ + ^[\w]+$

SUMMARY

Data science:	Decision intelligence
Core skills:	Modeling and coding
Applications:	ML / EDA / NLP
R vs. Python:	R + Python

TOOLS

- BPMN: **Signavio** Process Manager
- **R**: Emacs + **ESS** + Org-Mode (+ **Python**)
- Slides: GNU **Emacs** + **Org-Mode** + **reveal.js**
- Computer: Dell 7300 i7 1.9GHz (2019)
- OS: Ubuntu 18.04 LTS Linux (2018)
- Wacom Intuos Art Tablet

THANK YOU! QUESTIONS? COMMENTS?

Contact: birkenkrahe@outlook.com

REFERENCES

1. Cox. Translating Statistics to Make Decisions. Apress 2017.
2. Huang, Evans & Chattopadhyay. Deep Learning Without Neural Networks: Fractal-nets for Rare Event Modeling. Preprint. **10.21203/rs.3.rs-86045/v1**
3. Landin. The next 700 programming languages. In: Communications of the ACM 9(3) (1966). **10.1145/365230.365257**
4. McNulty. What Matters in Speed Dating? Online: **towardsdatascience.com** (02/14/2020)
5. Pearl & Mackenzie. The Book of Why: The New Science of Cause and Effect. **New York: Basic Books (2018)**.
6. Page. The Model Thinker. Basic Books (2018).
7. Polya. How to solve it. Doubleday (1957).
8. Programming vs Coding - A Short Comparison Between Both. **Online: GeeksforGeeks** (09/08/2020)
9. R: A language and environment fo statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL **https://www.r-project.org/**.
10. Silge & Robinson. Text Mining with R. O'Reilly (2017). **Online: tidytextmining.com**
11. Wickham & Grolemund. R for Data Science: Visualize, Model, Transform, Tidy, And Import

Data. O'Reilly (2016). Online:**r4ds.had.co.nz** (2016)]]

12. Wing. The Data Life Cycle. In: Harvard Data Science Review 1(1) (2019). **doi:10.1162/99608f92.e26845b4**