An End-to-End Data Science Project

Deena Gergis
Lead Data Scientist @ Bayer

#ACCELERATE

Workshop’s goal

The workshop will guide you through the process of completing an **end-to-end Data Science project**.

We will start with a **problem statement** and end with a **deployed product** that our client will be able to use.

We will utilize and connect various technologies, packages and programming paradigms to produce a functional product for our (fictional) client.

What to expect

Not this

- Course about the different technologies
- Deep development of any of the steps
- Information about specific markets or industries

But that

- Simplified end-to-end life cycle of an AI solution development
- Connecting all the different tech and analytics pieces together
- Reflections on real commercial operations and projects & the associated best practices

Workshop overview:

Session 1
Prep & Analytics
12.09.2021

Start with the business problem, set the foundation up, find data source, preprocess Start the descriptive analytics pipeline)

Session 2
Machine learning
19.09.2021

Implemented analytics pipeline, Build and evaluate prediction model(s), use Mlflow to keep track of the various experiments

Session 3
Deployment in Prod
26.09.2021

Create prediction functions and production class, develop an API, create a dashboard that the user will access and call the API

What you will do:

- **During the sessions:** You will get tasks to be done
- **After the sessions:**
  - You will complete the whole covered phases
  - Dig deeper into the various technologies discussed

i.e.: No Spoon-feeding :-)

#ACCELERATE

& let's get started

Problem statement

Our *(fictional)* client is an IT educational institute. They have reached out to us has reach out with the following:

“IT jobs and technologies keep evolving quickly. This makes our field to be one of the most interesting out there. But on the other hand, such fast development confuses our students. They do not know which skills they need to learn for which job.

“Do I need to learn C++ to be a Data Scientist?” “Do DevOps and System admins use the same technologies?” “I really like JavaScript; can I use it in Data Analytics?” Those are some of the questions that our students ask.

Could you please develop a data-driven solution for our students to answer such questions? They mostly want to understand the relationships between the jobs and the technologies.

Data Science Workflow

https://www.researchgate.net/figure/Guos-Data-Science-Workflow_fig2_319317441

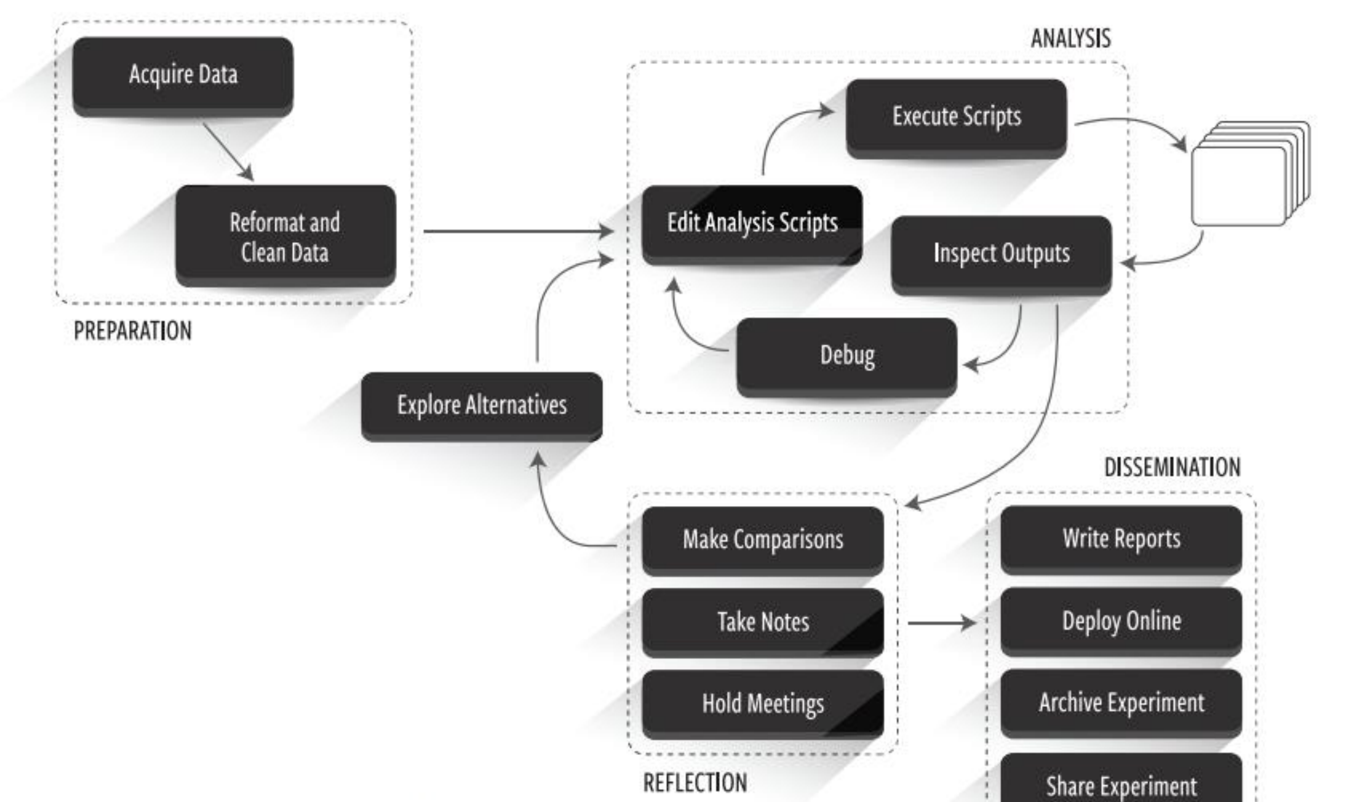


Figure 4: Guo's Data Science Workflow

#ACCELERATE

1. Business Problem

It's your turn:
What is your Business case?

You are asking a commercial business to invest in a new project. You need to prove that your work will have a positive financial impact.

How will you prove this? What are the KPIs that you will positively impact?

Business case

You are asking a commercial business to invest in a new project. You need to prove that your work will have a positive financial impact.
How will you prove this? What are the KPIs that you will positively impact?

1. **Higher enrollment rate due to the higher certainty**

2. **Decrease in drop-out rate**

3. **Time saved for the academic advisors**

Learn more: https://www.youtube.com/watch?v=zQ5WqAz3myo

2. Data

#ACCELERATE

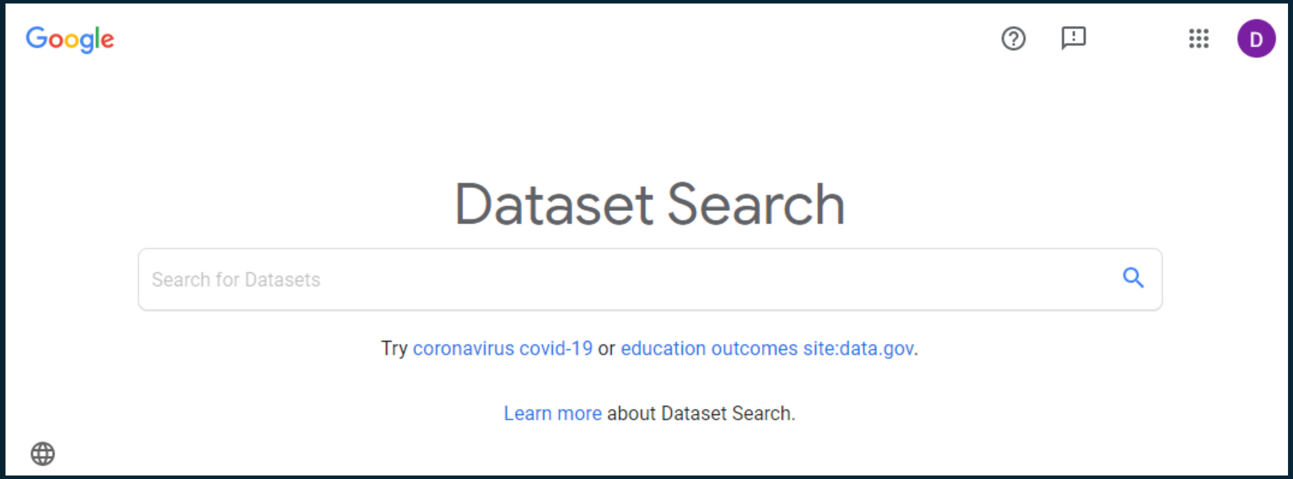
It's your turn:
What is your Data Source?

Our client doesn't have any internal data sources that could be used for this project. **Find the data source that you will use to build the solution**

Data source

Where to start?

https://datasetsearch.research.google.com/



Google

Dataset Search

Search for Datasets

Try coronavirus covid-19 or education outcomes site:data.gov.

Learn more about Dataset Search.

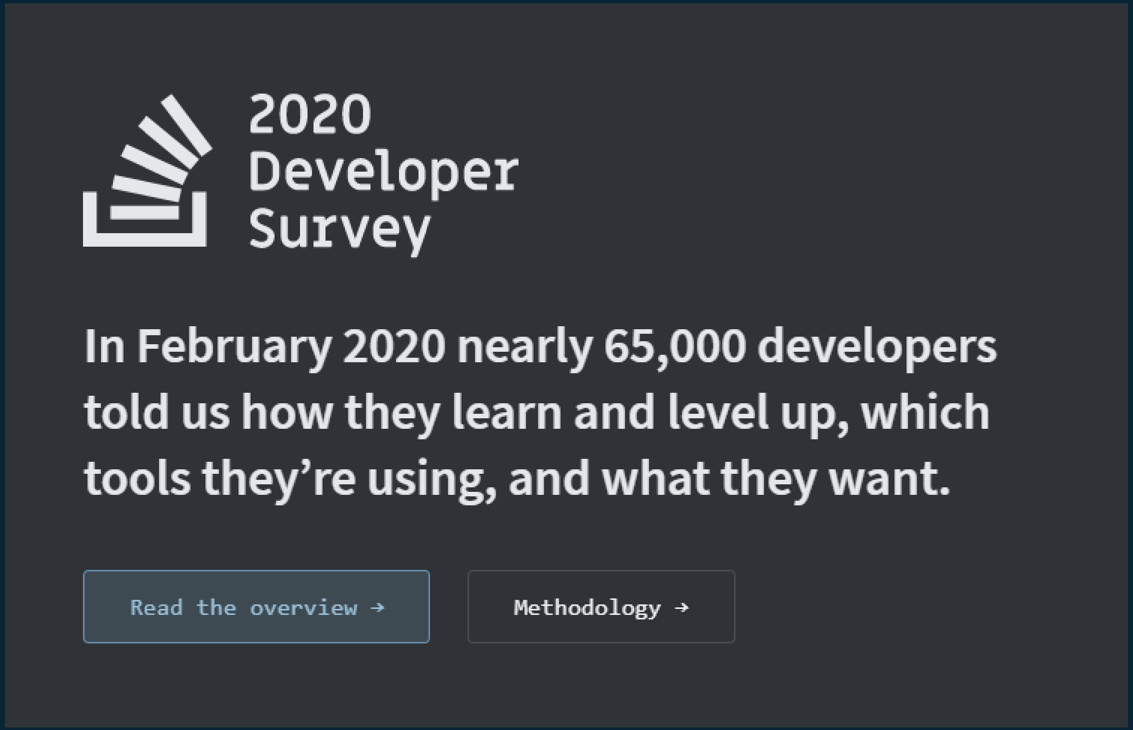
Be careful:

- Be thorough with the quality checks
- Make sure that your data will be updated on a regular base

Data source

Chosen: Stack Overflow developers survey

https://insights.stackoverflow.com/survey/2020



2020 Developer Survey

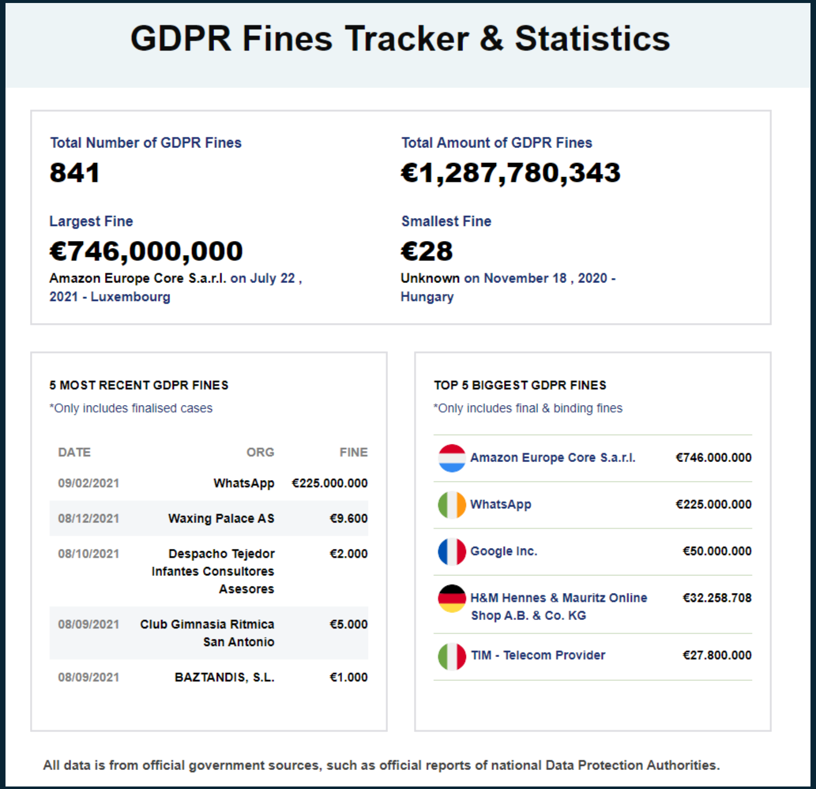
In February 2020 nearly 65,000 developers told us how they learn and level up, which tools they’re using, and what they want.

3. Foundations

1. Legal and data privacy check

Global:

https://www.privacyaffairs.com/gdpr-fines/



GDPR Fines Tracker & Statistics

Total Number of GDPR Fines
841

Total Amount of GDPR Fines
€1,287,780,343

Largest Fine
€746,000,000
Amazon Europe Core S.a.r.l. on July 22 , 2021 - Luxembourg

Smallest Fine
€28
Unknown on November 18 , 2020 - Hungary

5 MOST RECENT GDPR FINES
*Only includes finalised cases

DATE	ORG	FINE
09/02/2021	WhatsApp	€225.000.000
08/12/2021	Waxing Palace AS	€9.600
08/10/2021	Despacho Tejedor Infantes Consultores Asesores	€2.000
08/09/2021	Club Gimnasia Ritmica San Antonio	€5.000
08/09/2021	BAZTANDIS, S.L.	€1.000

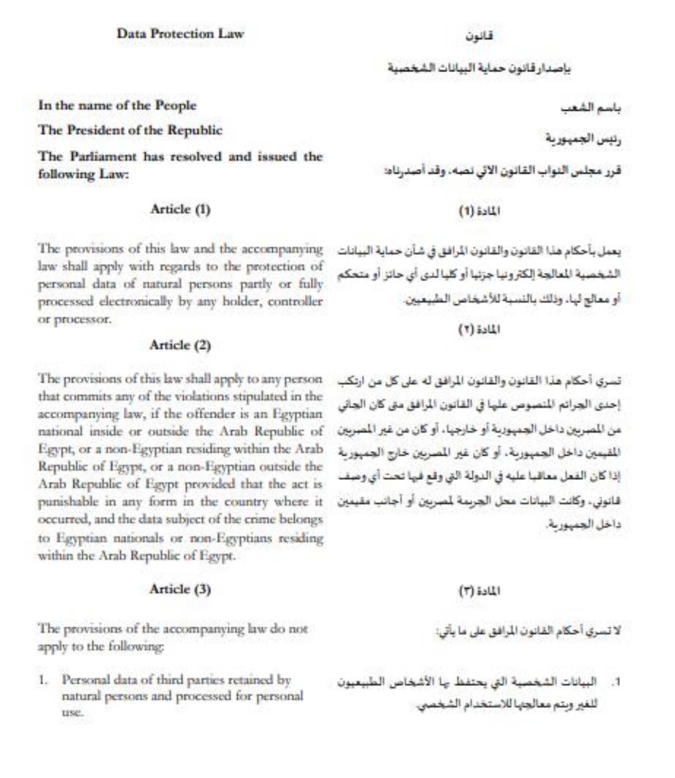
TOP 5 BIGGEST GDPR FINES
*Only includes final & binding fines

Amazon Europe Core S.a.r.l.	€746.000.000
WhatsApp	€225.000.000
Google Inc.	€50.000.000
H&M Hennes & Mauritz Online Shop A.B. & Co. KG	€32.258.708
TIM - Telecom Provider	€27.800.000

All data is from official government sources, such as official reports of national Data Protection Authorities.

Local:

https://www.privacylaws.com/media/3263/egypt-data-protection-law-151-of-2020.pdf



Data Protection Law

In the name of the People

The President of the Republic

The Parliament has resolved and issued the following Law:

Article (1)

The provisions of this law and the accompanying law shall apply with regards to the protection of personal data of natural persons partly or fully processed electronically by any holder, controller or processor.

Article (2)

The provisions of this law shall apply to any person that commits any of the violations stipulated in the accompanying law, if the offender is an Egyptian national inside or outside the Arab Republic of Egypt, or a non-Egyptian residing within the Arab Republic of Egypt, or a non-Egyptian outside the Arab Republic of Egypt provided that the act is punishable in any form in the country where it occurred, and the data subject of the crime belongs to Egyptian nationals or non-Egyptians residing within the Arab Republic of Egypt.

Article (3)

The provisions of the accompanying law do not apply to the following:

1. Personal data of third parties retained by natural persons and processed for personal use.

قانون

بإصدار قانون حماية البيانات الشخصية

باسم الشعب

رئيس الجمهورية

قرر مجلس النواب القانون الآتي نصه، وقد أصدرناه:

المادة (١)

يعمل بأحكام هذا القانون والقانون المرافق في شأن حماية البيانات الشخصية المعالجة إلكترونيا جزئيا أو كليا لدى أي حائز أو متحكم أو معالج لها، وذلك بالنسبة للأشخاص الطبيعيين.

المادة (٢)

تسري أحكام هذا القانون والقانون المرافق له على كل من ارتكب إحدى الجرائم المنصوص عليها في القانون المرافق متى كان الجاني من المصريين داخل الجمهورية أو خارجها، أو كان من غير المصريين المقيمين داخل الجمهورية، أو كان غير المصريين خارج الجمهورية إذا كان الفعل معاقبا عليه في الدولة التي وقع فيها تحت أي وصف قانوني، وكانت البيانات محل الجريمة لمصريين أو أجانب مقيمين داخل الجمهورية.

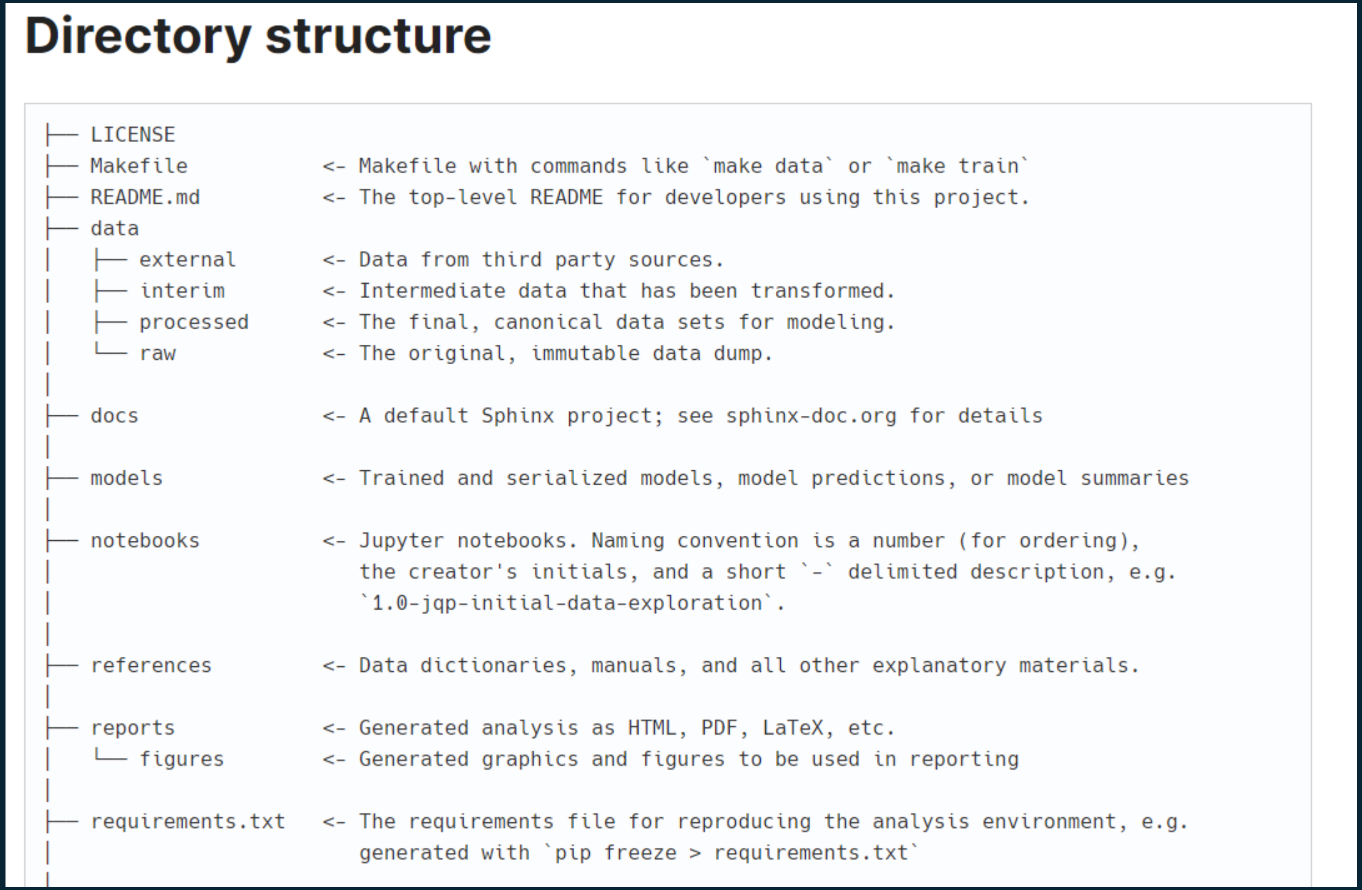
المادة (٣)

لا تسري أحكام القانون المرافق على ما يأتي:

1. البيانات الشخصية التي يحتفظ بها الأشخاص الطبيعيون للغير ويتم معالجتها للاستخدام الشخصي.

2. How to structure your project

https://drivendata.github.io/cookiecutter-data-science/

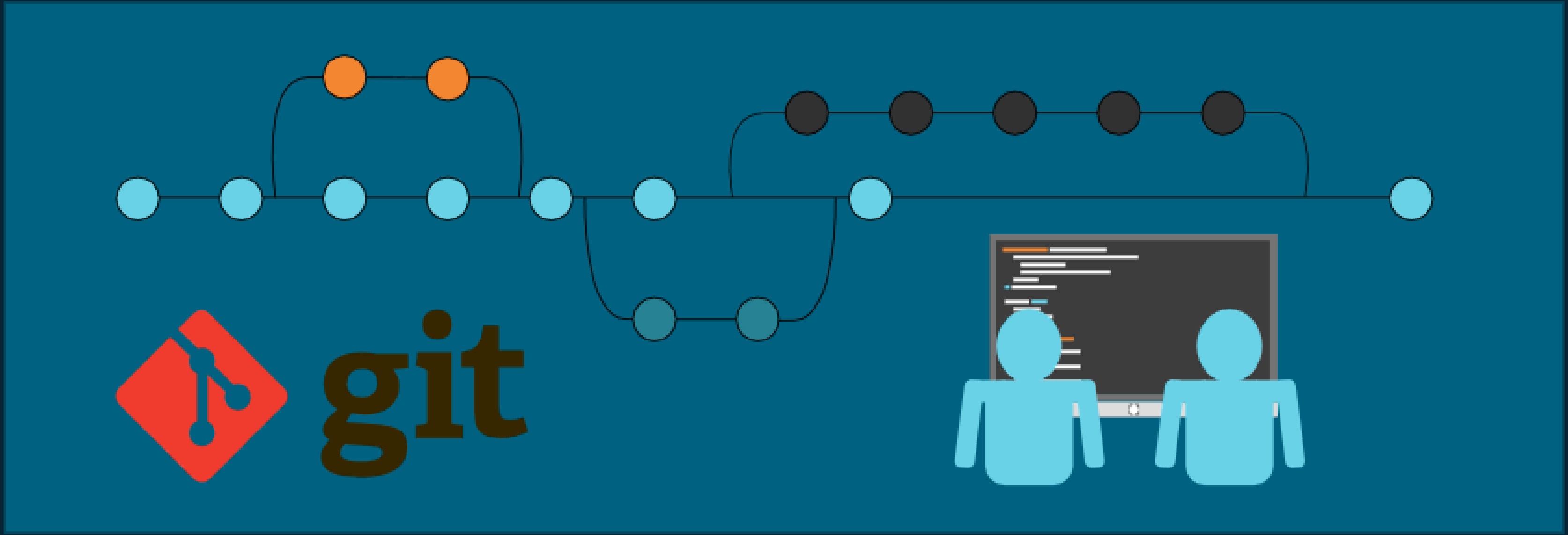


Directory structure

```
├── LICENSE
├── Makefile           <- Makefile with commands like `make data` or `make train`
├── README.md          <- The top-level README for developers using this project.
├── data
│   ├── external       <- Data from third party sources.
│   ├── interim        <- Intermediate data that has been transformed.
│   ├── processed      <- The final, canonical data sets for modeling.
│   └── raw            <- The original, immutable data dump.
│
├── docs               <- A default Sphinx project; see sphinx-doc.org for details
│
├── models             <- Trained and serialized models, model predictions, or model summaries
│
├── notebooks          <- Jupyter notebooks. Naming convention is a number (for ordering),
│                         the creator's initials, and a short `-` delimited description, e.g.
│                         `1.0-jqp-initial-data-exploration`.
│
├── references         <- Data dictionaries, manuals, and all other explanatory materials.
│
├── reports            <- Generated analysis as HTML, PDF, LaTeX, etc.
│   └── figures        <- Generated graphics and figures to be used in reporting
│
├── requirements.txt   <- The requirements file for reproducing the analysis environment, e.g.
│                         generated with `pip freeze > requirements.txt`
│
```

3. Your Git repo

https://developerhowto.com/2018/10/12/git-for-beginners/



It's your turn:

- Create your project directory using the cookie cutter
- Track your project in a new GitHub repo
- Download & save your data

4. Preprocessing

#ACCELERATE

Preprocessing at first glance

…

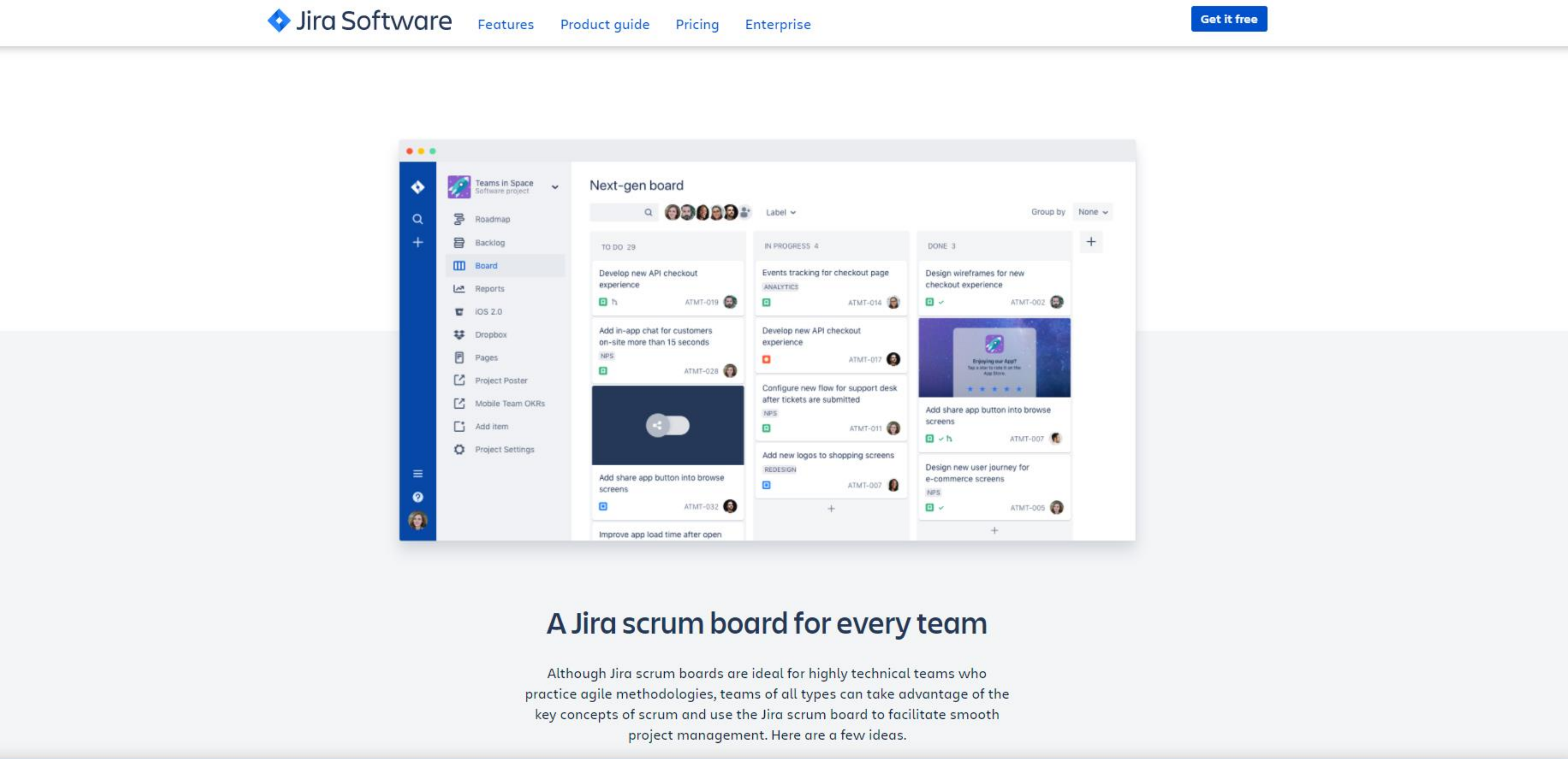
1. String values in years need to be replaced
2. Multiple values separated be `;` need to be splitted

…

→

- *Prioritize task*
- *Create tickets in Jira*
- *Team members pick the tickets and solve them*

Jira - Kanban


Jira Software
Features
Product guide
Pricing
Enterprise
Get it free
Teams in Space
Software project
Roadmap
Backlog
Board
Reports
iOS 2.0
Dropbox
Pages
Project Poster
Mobile Team OKRs
Add item
Project Settings
Next-gen board
Label
Group by
None
TO DO 29
Develop new API checkout experience
ATMT-019
Add in-app chat for customers on-site more than 15 seconds
NPS
ATMT-028
Add share app button into browse screens
ATMT-032
Improve app load time after open
IN PROGRESS 4
Events tracking for checkout page
ANALYTICS
ATMT-014
Develop new API checkout experience
ATMT-017
Configure new flow for support desk after tickets are submitted
NPS
ATMT-011
Add new logos to shopping screens
REDESIGN
ATMT-007
DONE 3
Design wireframes for new checkout experience
ATMT-002
Enjoying our App?
Add share app button into browse screens
ATMT-007
Design new user journey for e-commerce screens
NPS
ATMT-005

A Jira scrum board for every team

Although Jira scrum boards are ideal for highly technical teams who practice agile methodologies, teams of all types can take advantage of the key concepts of scrum and use the Jira scrum board to facilitate smooth project management. Here are a few ideas.

#ACCELERATE

It's your turn :

- Preprocess your raw data and export it to a pickle file
- Push your work to your repo

5. Descriptive Analytics

#ACCELERATE

“Asking the right question is half of the answer”

#ACCELERATE

It’s your turn:
What are the descriptive questions that you will answer ?

Think about what you want to do before you start doing it. Keep the original goal in mind

#ACCELERATE

My analytics question

General:

- Total number of answers
- Geographical distributions
- Missing answers

Skills:

- Frequency of each skill
- How are the skills correlated with each others

Jobs:

- Frequency of each job
- How are the jobs correlated with each others

Relation:

- How are the skills correlated with the jobs
- What is the specificity of each skill to a job

Levels of descriptive analytics

1. Stats or summary tables
2. Visualizations
3. Unsupervised learning (e.g. clustering)

Assignment:

Complete the setup, preprocessing & descriptive analytics phase

Wrap Up

#ACCELERATE

Wrap-up:
Today you have learned about

- *Build a business case*
- *Find suitable data sources*
- *Verify legal rights*
- *Cookie-cutting your directory structure*
- *Track your project via Git*
- *Explore and preprocess data*
- *Collaborate with your team using Jira*
- *Planning framework for your descriptive analytics*

Questions?

#ACCELERATE

#ACCELERATE