BAYER

An End-to-End Data Science Project

Deena Gergis
Lead Data Scientist @ Bayer

#ACCELERATE

Workshop overview:

Session 1
Preparation
10.04.2022

Start with the business problem, find data source, preprocess data, set up team process and tech

Session 2
Analytics
17.04.2022

Analyze and understand your data. Gain insights and prepare for the predictive modeling

Session 3
Machine learning
x.05.2022

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

Session 4
Production
x.05.2022

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:

- **Form a team of 3 members**
- **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

#ACCELERATE

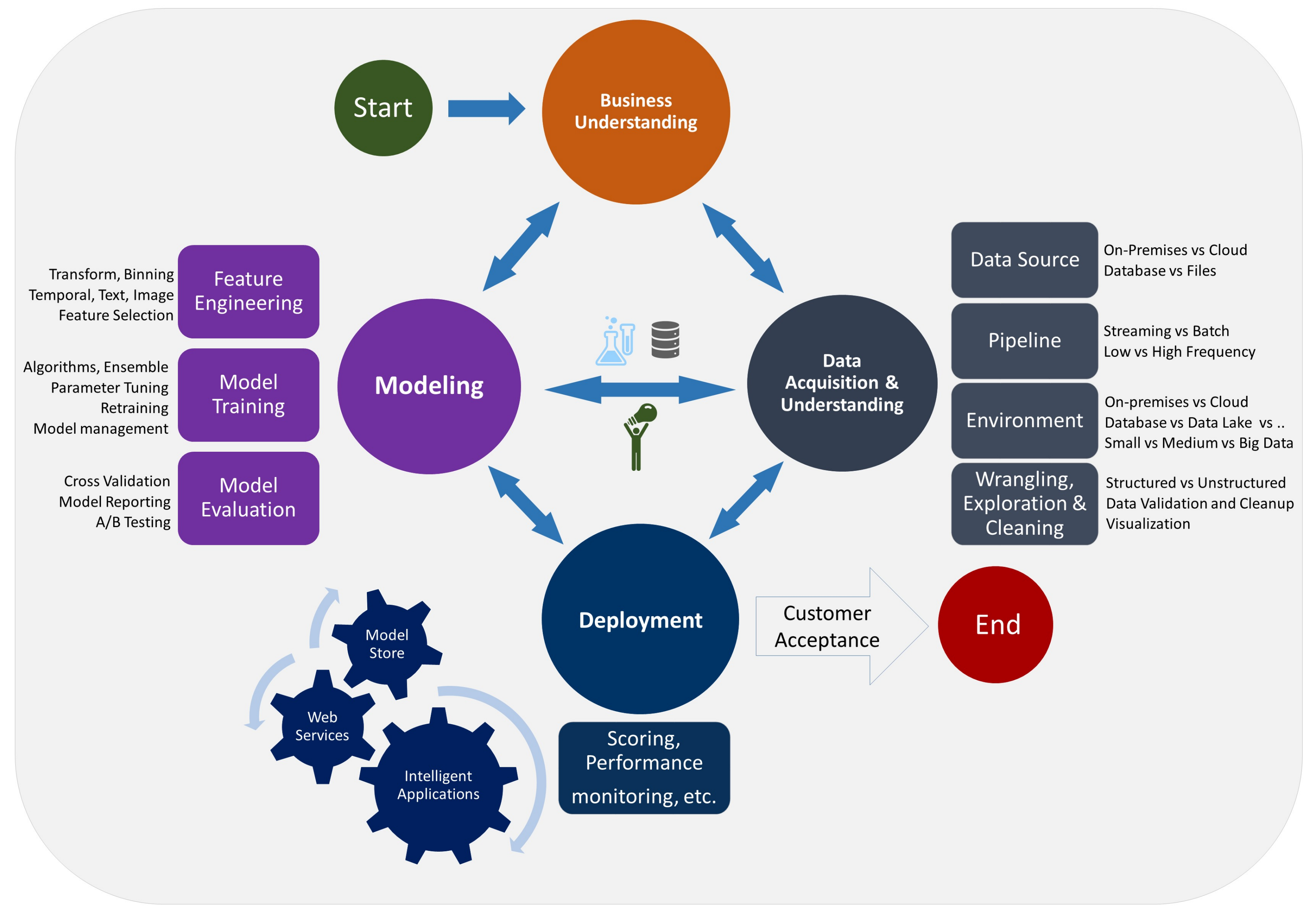
Session 1: Recap

#ACCELERATE

Session 1: Main points

- *Build a business case*
- *Find suitable data sources*
- *Verify legal rights*
- *Track your project via Git*
- *Explore and preprocess data*
- *Collaborate with your team using Kanban*

Session 1: Main points



#ACCELERATE

Session 1: Main points

- *Form your **team** and create your **Kanban** board*
- *Create your **project directory** and track in a new **GitHub** repo*
- ***Preprocess** your raw data and export it to a pickle file*
- *Complete your **descriptive analytics** part – understand your data and get insights to be used in the modelling*

Commercial Data Science

Article:
5 things I wish I knew about real-life AI
https://www.linkedin.com/pulse/5-things-i-wish-knew-real-life-ai-deena-gergis/

Podcast: Beyond Coding
https://www.facebook.com/100046924503697/posts/511122910461855/

Webinar: ApplAI - Ain Shams
https://www.facebook.com/100046924503697/posts/370643447843136/

Session 2:
Descriptive Analytics

Part 1.
Insights

“Asking the right question is half of the answer”

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

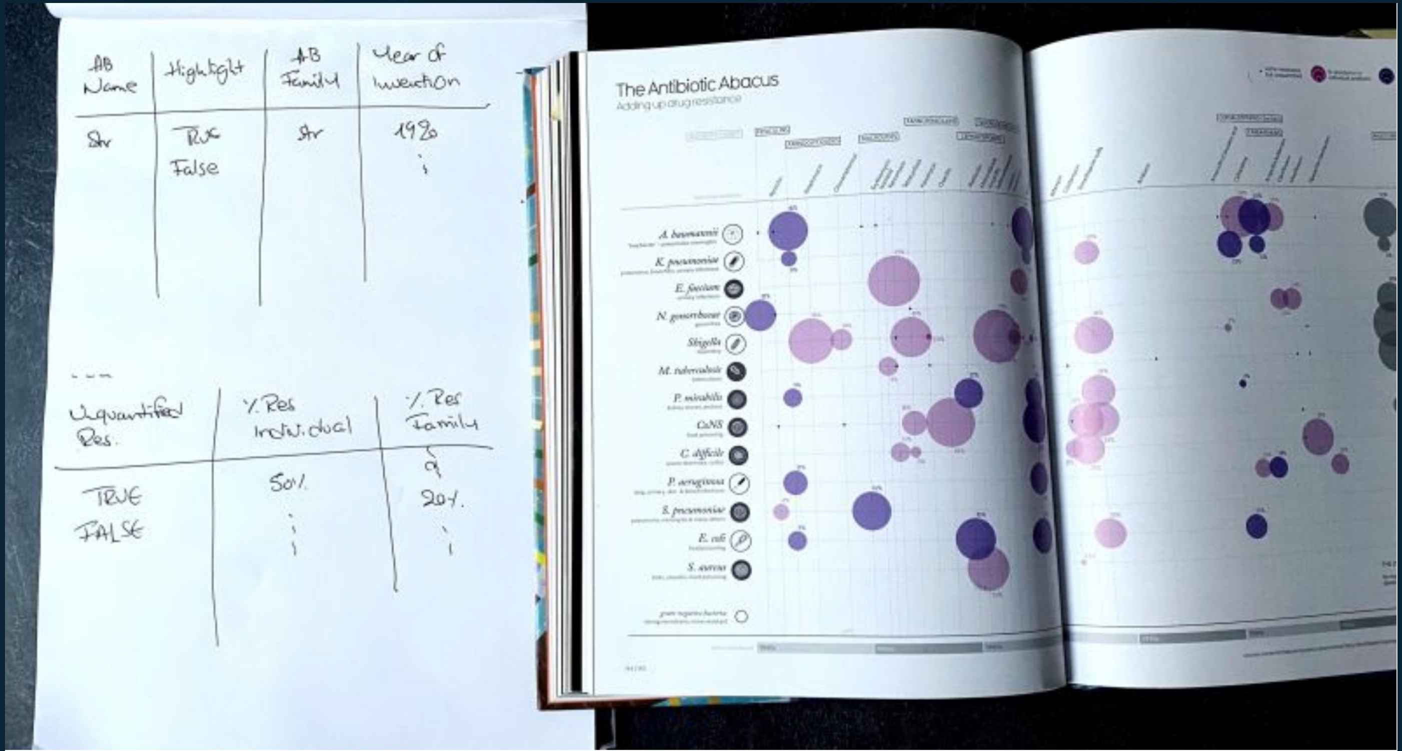
Levels of descriptive analytics

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

How to improve your data visualization skills?

https://www.linkedin.com/feed/update/urn:li:activity:6838041403960942593/

https://www.youtube.com/watch?v=lYlhJnqNNIA



AB Name	Highlight	AB Family	Year of Invention
Str	True False	Str	1920 ⋮

Unquantified Res.	% Res Individual	% Res Family
TRUE FALSE	50% ⋮	20% ⋮

#ACCELERATE

Part 2. Unsupervised to Supervised

Unsupervised to Supervised

T-SNE

Stands for t-distributed stochastic neighbor embedding.
Nonlinear dimensionality reduction technique

Agglomerative Clustering

Recursively merges the pair of clusters that minimally increases a given linkage distance

Silhouette metric

The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation)

Part 3.
Data manipulation

Data selection

Responses:

- Select responses within reasonable ranges

Classes:

- Drop non-relevant classes (e.g. Senior executive)
- Merge close classes (e.g. Scientist & Researcher)
- Split vague classes (e.g. Backend developer)

Features:

- Create new features (e.g. Skills groups)
- Drop irrelevant features (e.g. Platforms)

#ACCELERATE

Till next time:

- Complete and enhance your descriptive analytics pipeline
- Start with the predictive analytics (X: Skills , Y: Jobs)

Questions?

#ACCELERATE