Day 3

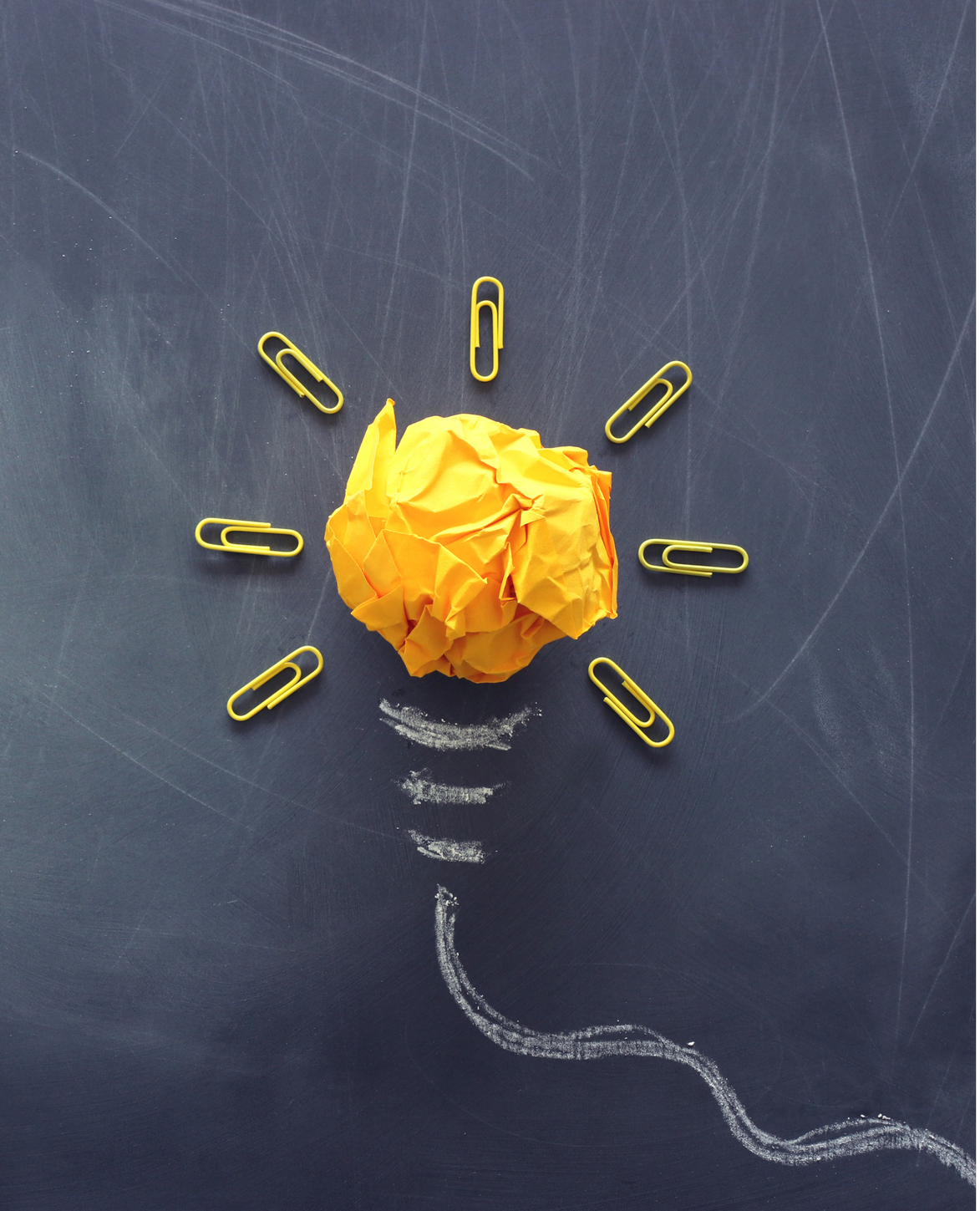
The Project Lifecycle (Part 1)

Overview

Day 3

- Case Studies
- Project Design
  - Project Planning
  - Problem Formulation
  - Data Extraction & Procurement
  - Data Analysis

- Guest Lecture - Professor Sabina Leonelli



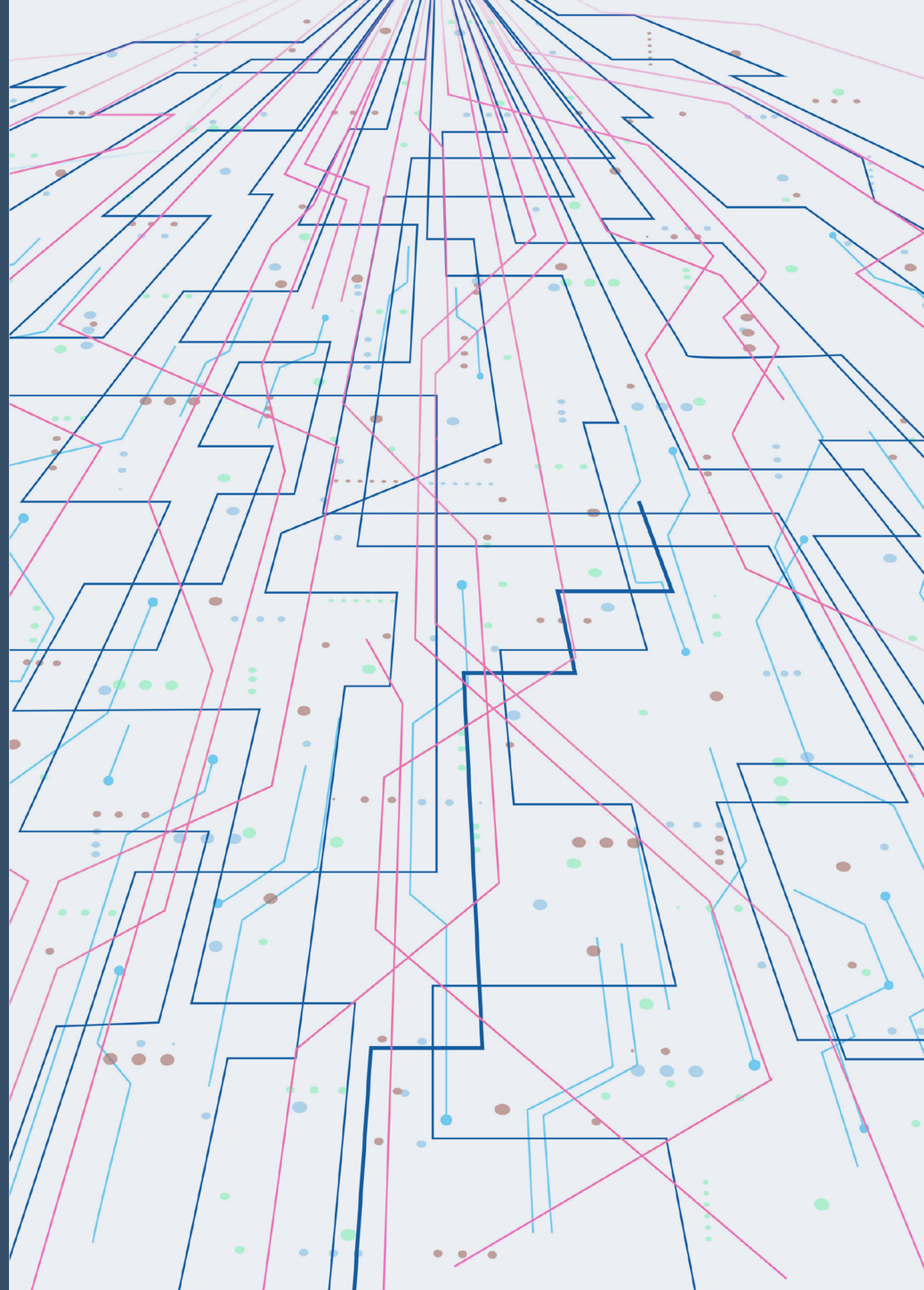
Day 3

Learning Objectives

- Gain a high-level understanding of the (project) design stage.
- Explore the activities that are associated with this stage, focusing on salient ethical, social, and legal issues.

Day 3

Case Studies



Case Study 1

Predicting Risk of Reoffending

You are a project team responsible for developing a predictive risk assessment tool that can support sentencing decisions by judges in criminal courts. The tool will take data about a defendant and feed this into an algorithm that predicts a risk score, between 1 and 5 that is presented to the judge alongside additional case evidence (e.g., witness testimony). This score will represent the likelihood of reoffending, and, therefore, inform the sentencing decision made by the judge. For those that are discharged, the system will also receive feedback about whether the defendant goes on to reoffend.

Case Study 2

Recommending Courses

You work for an EdTech company and need to develop a recommender system that will be sold to schools to augment careers advice for students considering university courses. The system will ask each student to answer a series of questions, and will then provide an ordered list of recommended courses (linked to the respective university) that it “believes” are good options for the student. The system will also use satisfaction survey results and obtained degree results from those students who used the system previously as a way of calibrating and adjusting its recommendations (i.e., learning).

Case Study 3

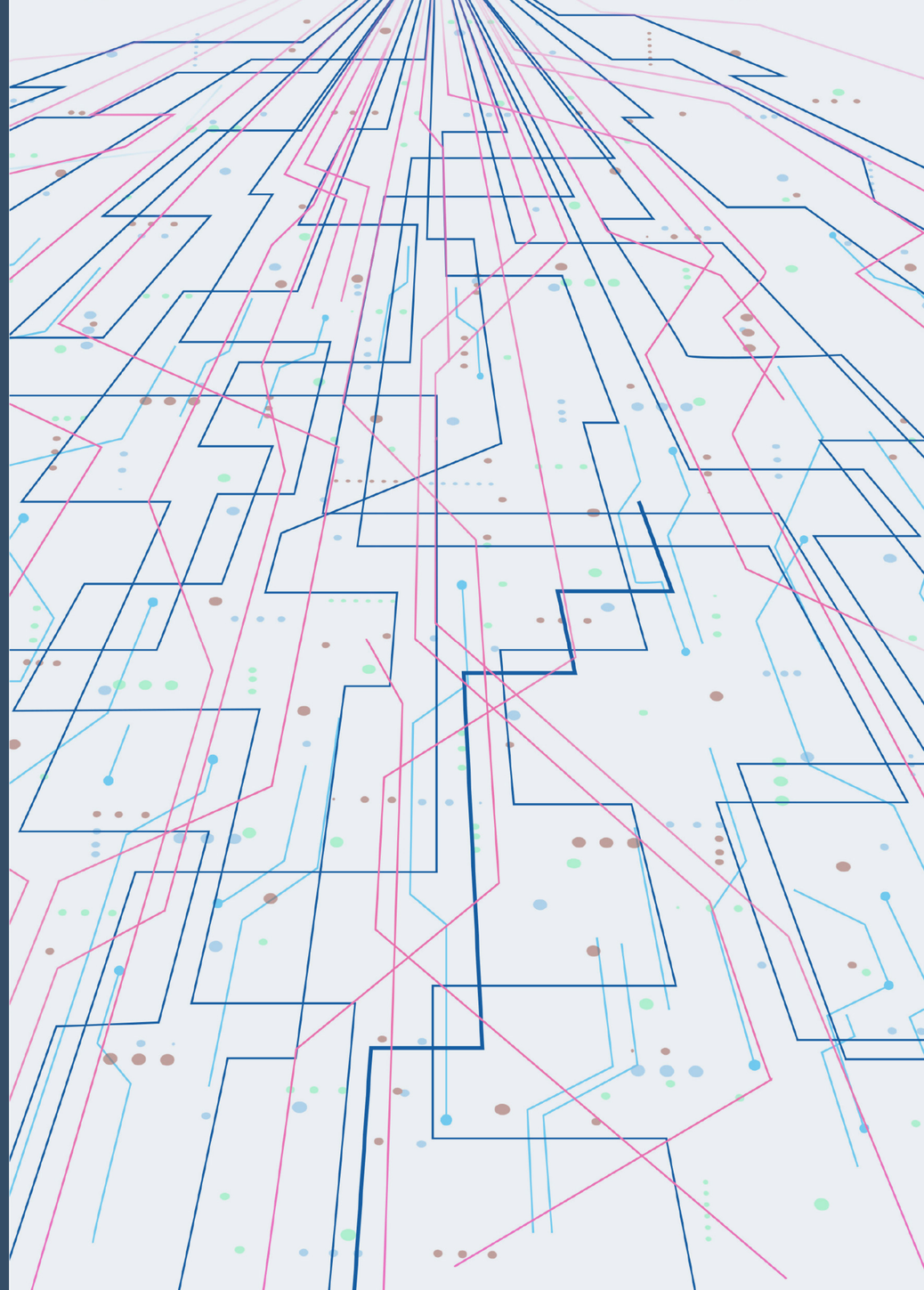
Classifying Hate Speech

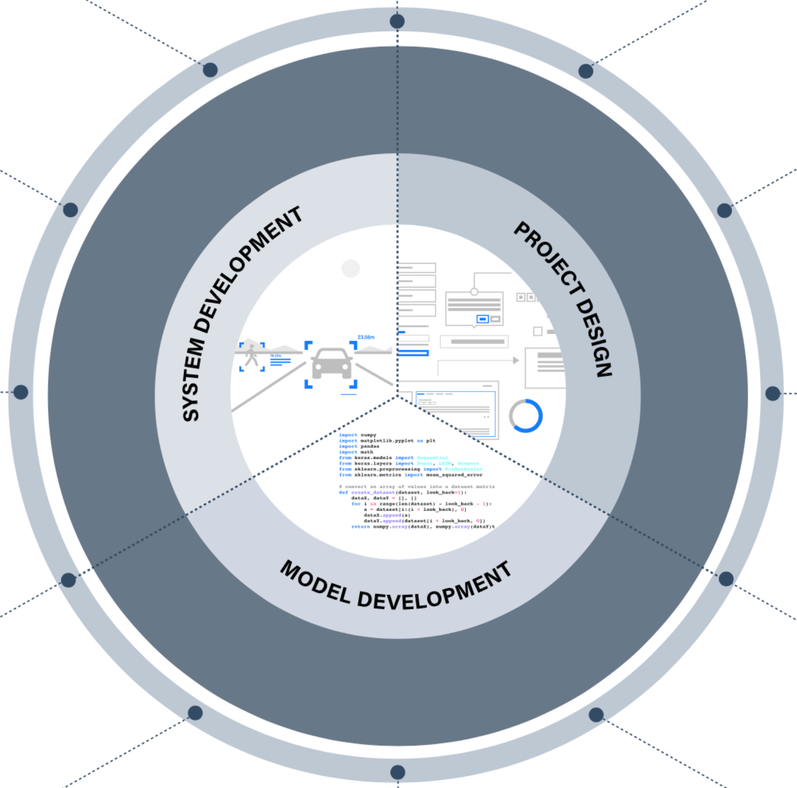
You are a team of social data scientists employed as consultants for a social media company. You have been tasked with reducing the levels of hate speech on the company's platform by developing a classifier that can flag potential instances of hate speech for review by human moderators. The tool will automatically review every post submitted to the platform, but will only flag those that are likely to represent an instance of hate speech, based on whether they exceed some likelihood threshold. In addition to the textual content contained within the post, your tool can also use a variety of other input sources to improve its decision-making, including feedback from the human moderators that help improve the accuracy of the model over time.

Breakout Groups

Day 3

(Project) Design





Project Planning

Problem Formulation

Data Extraction & Procurement

Data Analysis

Preprocessing & Feature Engineering

Model Selection & Training

Model Testing & Validation

Model Reporting

Model Production-alisation

User Training

System Use & Monitoring

Model Updating & Deprovisioning

(Project) Design

Project Planning

In October 2021, the Financial Times reported that facial recognition cameras were being used in UK schools to scan the faces of “thousands of British pupils in school canteens” to automate the process of taking payment for lunches. The managing director of CRB Cunninghams—the company that developed the system sold to schools—claimed that “In a secondary school you have around about a 25-minute period to serve potentially 1,000 pupils. So we need fast throughput at the point of sale.”

(Project) Design

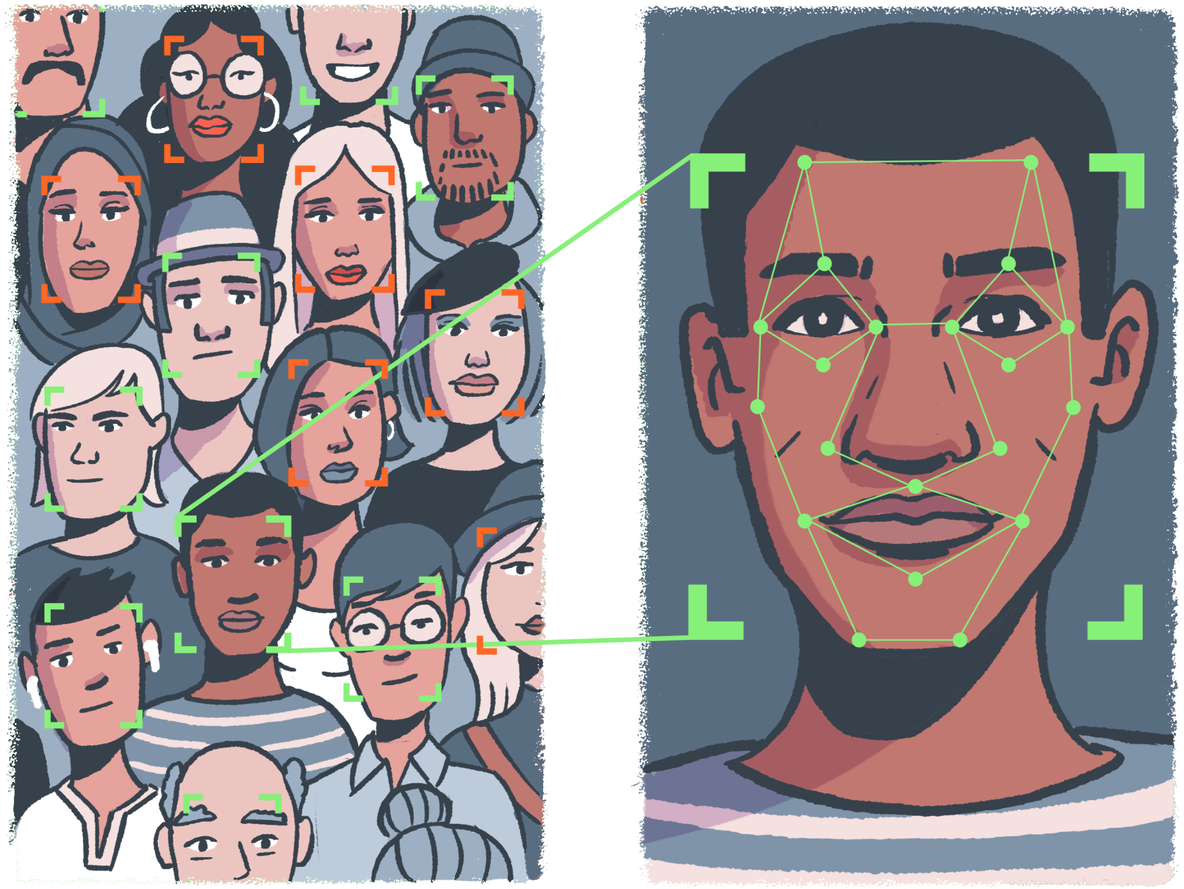
Project Planning

Does this seem like a plausible justification for the design, development, and deployment of an automated facial recognition system?

(Project) Design

Project Planning

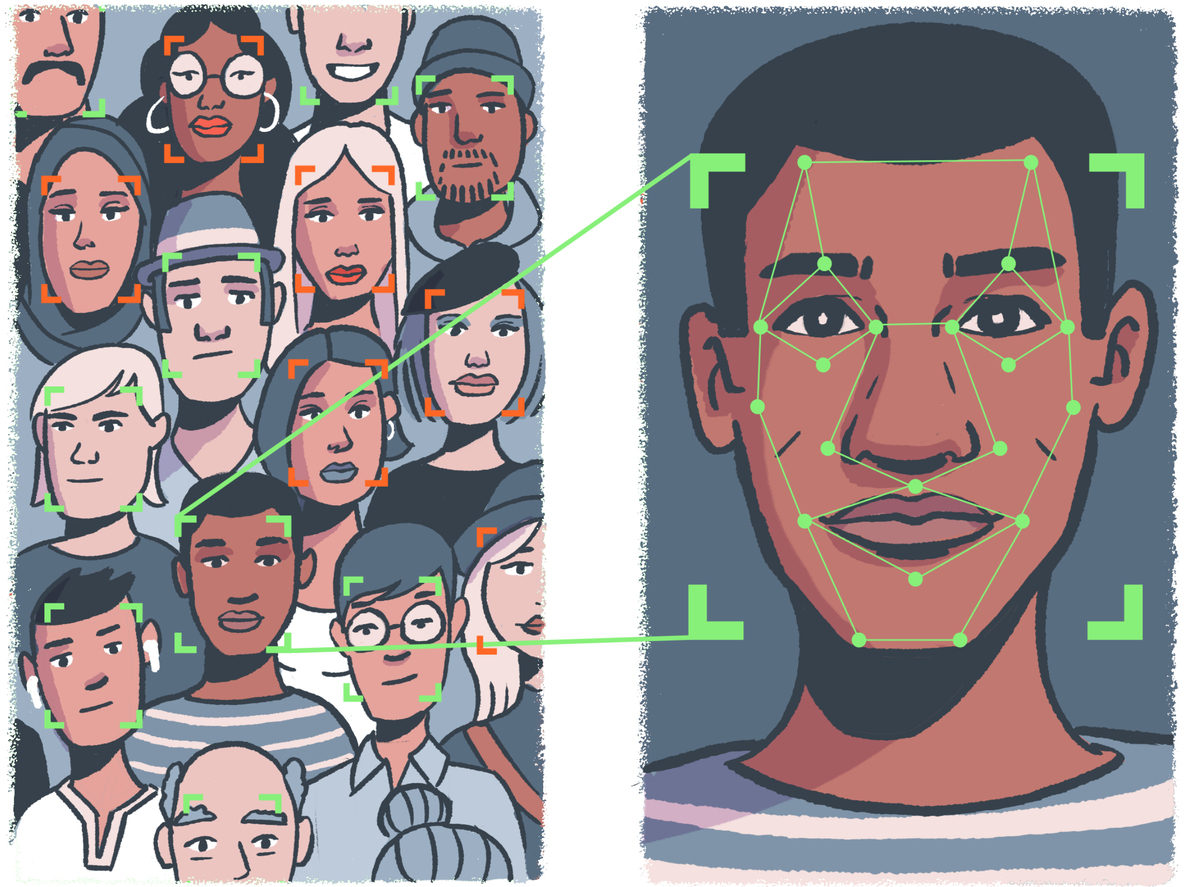
Project Planning involves a variety of preliminary steps, including asking whether a system should even be developed.



(Project) Design

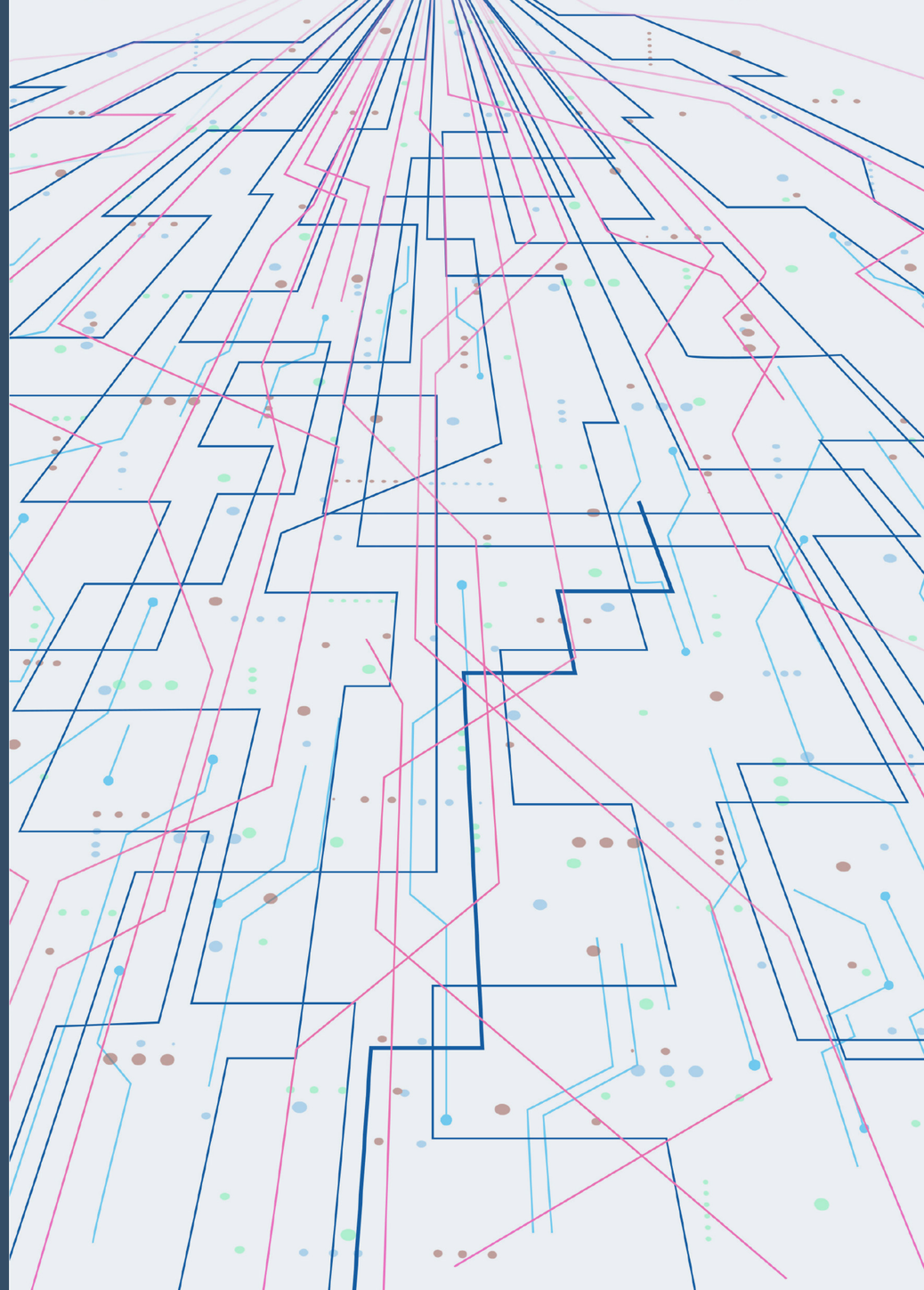
Project Planning

- An assessment of whether developing or using data-driven technology is the right approach given available resources and data, existing technologies and processes already in place, the complexity of the use-contexts involved, and the nature of the policy or social problem that needs to be solved.
- An analysis of user needs in relation to the prospective model or system
- Identification and mapping of key stages in the project to support project governance and business tasks (e.g. scenario planning).
- A contextual assessment of the target domain and of the expectations, norms, and requirements that derive therefrom.
- Wider impact assessments (e.g. equality impact assessments, data protection impact assessments, human rights impact assessment, bias assessment).



Activity 6

Privilege Walk



Breakout Groups

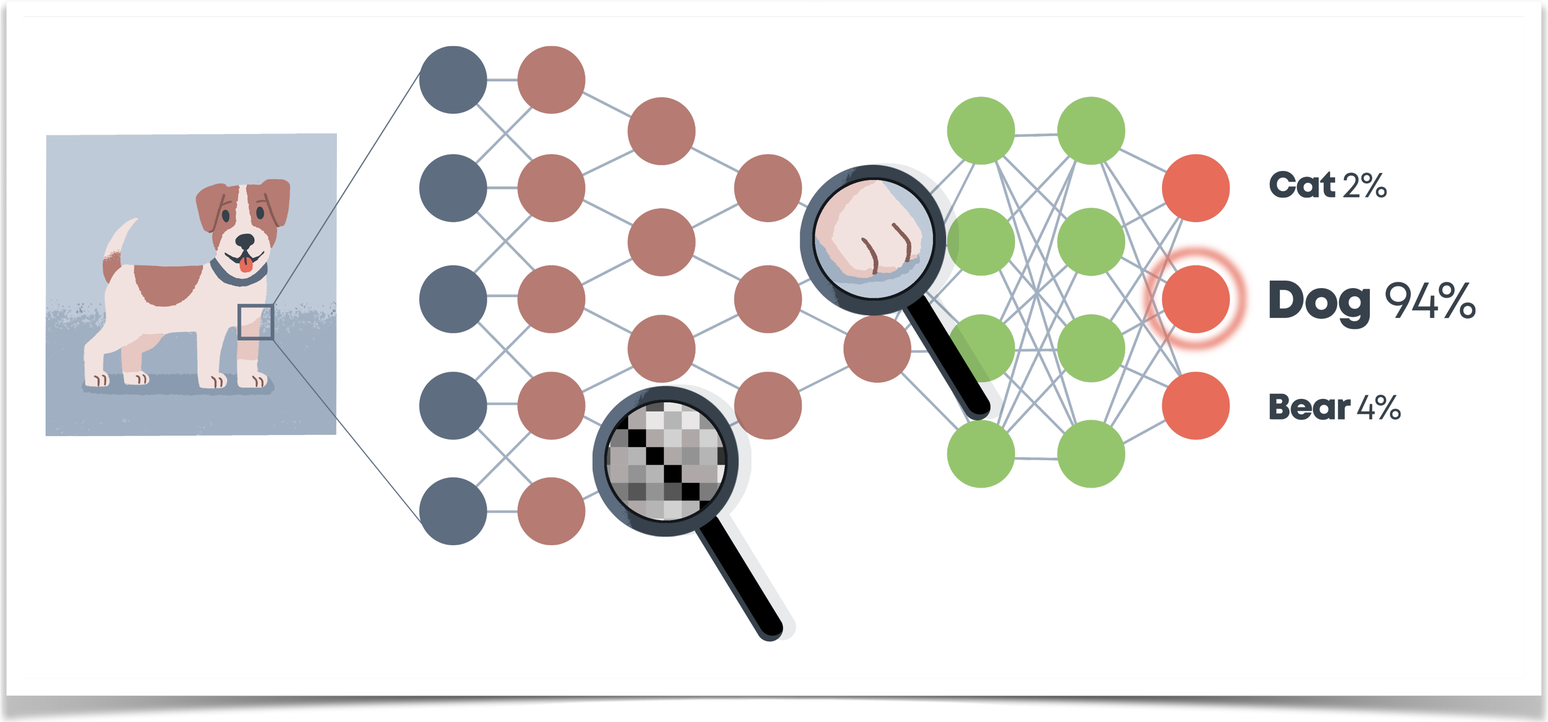
Plenary

(Project) Design

Problem Formulation

To refer to a *well-defined computational process* (or a higher-level abstraction of the process) that is carried out by the algorithm to map inputs to outputs.

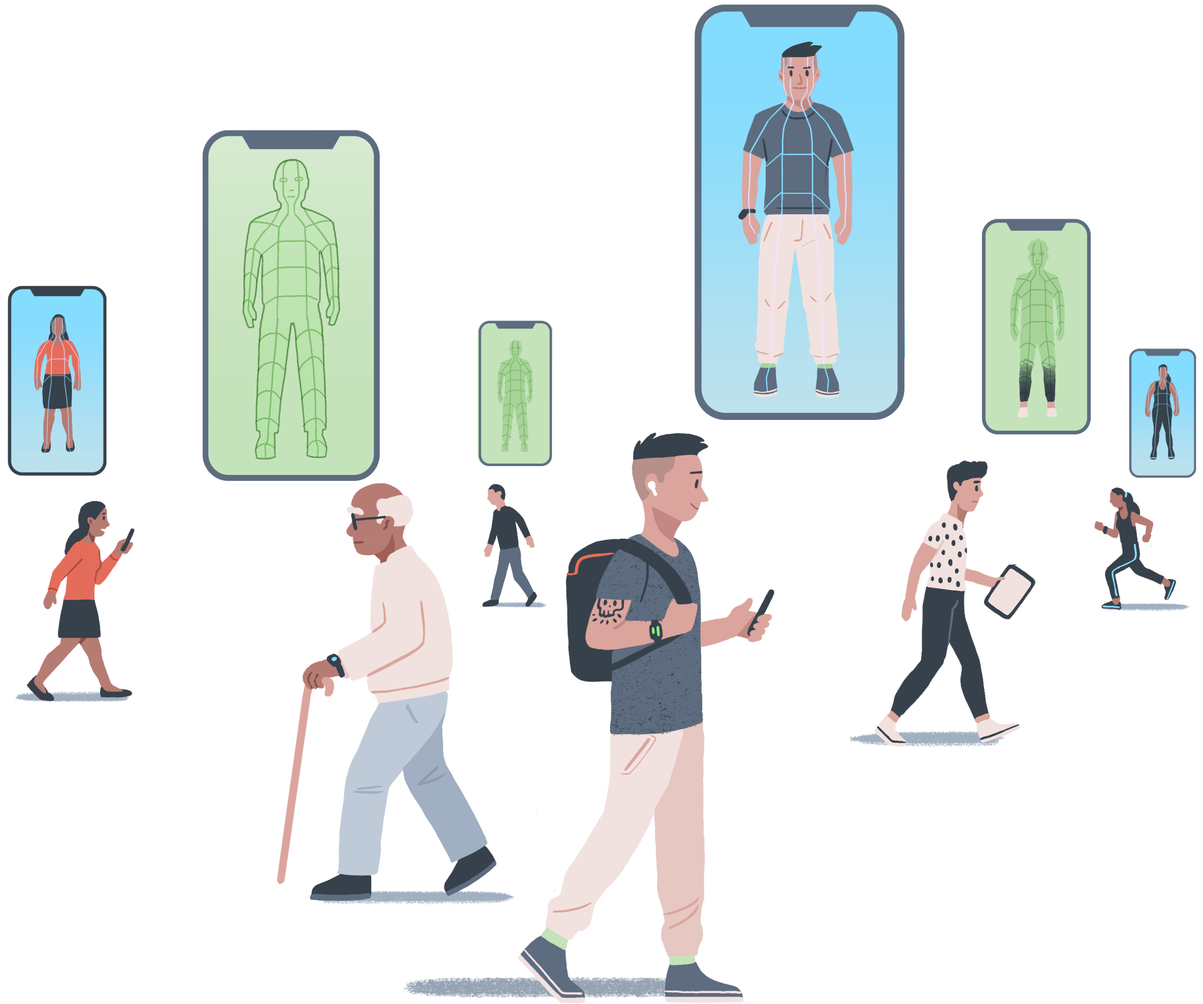
To refer to the wider practical, social, or policy issue that will be addressed through the translation of that issue into the aforementioned mathematical or computational framing.



(Project) Design

Data Extraction or Procurement

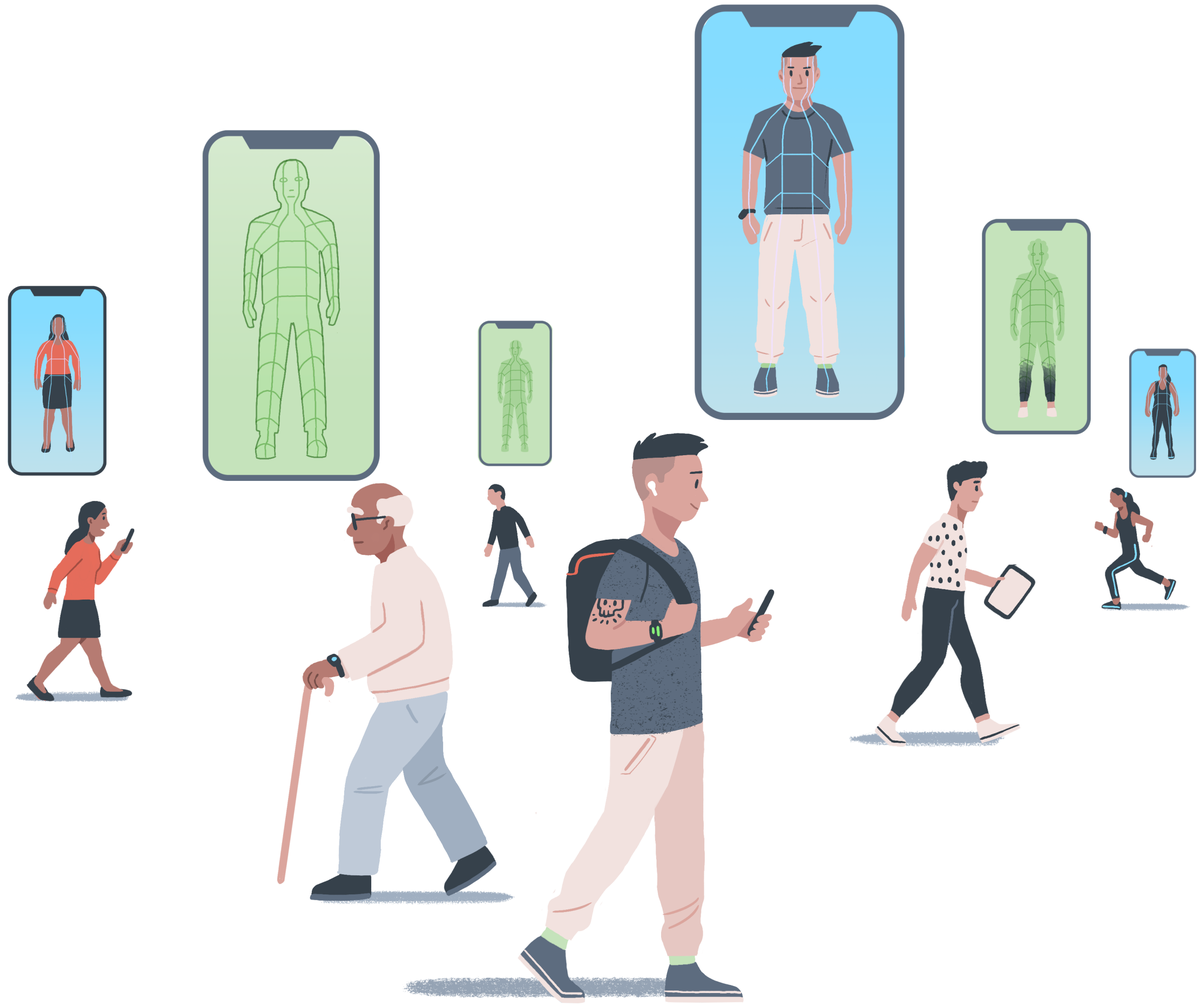
Ideally, the project team should have a clear idea in mind (from the planning and problem formulation stages) of what data are needed prior to collection, extraction, or procurement.



(Project) Design

Data Extraction or Procurement

Where data is procured, questions about provenance arise (e.g. legal issues, concerns about informed consent of human data subjects).



(Project) Design

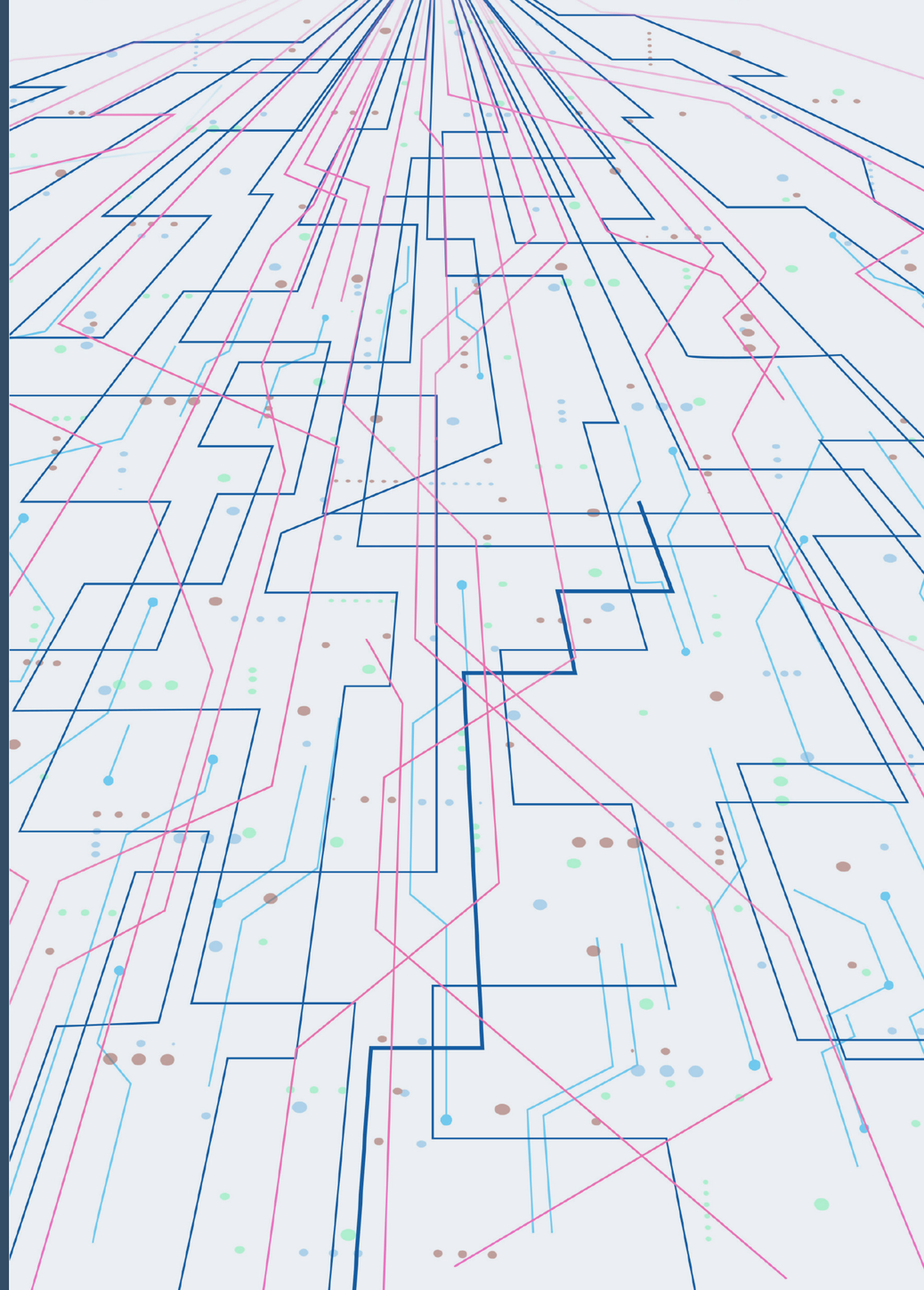
Data Extraction or Procurement

The FAIR principles for scientific data management and stewardship were developed, as a means to improve the Findability, Accessibility, Interoperability, and Reuse of research data and digital assets.

Visit The Turing Way for more information on reproducible data science.

Activity 7

Developing Case Studies



Breakout Groups

(Project) Design

Data Analysis

Exploratory data analysis can involve:

- Describing the dataset and important variables
- Cleaning the dataset
- Identifying missing data and outliers, and deciding how to handle them
- Provisional analysis of any relationships between variables
- Uncovering possible limitations of the dataset (e.g. class imbalances) that could affect the project

notebooks.gesis.org

File Edit View Run Kernel Tabs Settings Help

Filter files by name

/ chapter4 / project_design /

Name	Last Modified
activity4-1...	4 minutes ago
activity4-2...	4 minutes ago
activity4-3...	4 minutes ago
covid_pati...	4 minutes ago
covid_pati...	4 minutes ago
data_analy...	4 minutes ago
data_extra...	4 minutes ago
generative...	4 minutes ago
planning.md	4 minutes ago
problem.md	4 minutes ago
synthetic_...	4 minutes ago

data_analysis.ipynb

Download GitHub Binder Markdown git Python 3 (ipykernel)

COVID-19 Hospital Data

For the purpose of this section we have created a synthetic dataset that contains 27308 records for fictional patients who were triaged (and possibly admitted) to a single hospital for treatment of COVID-19.

The dataset has been designed with this pedagogical task in mind. Therefore, although we relied upon plausible assumptions when developing our generative model, the data are not intended to be fully representative of actual patients. Our methodology for generating this dataset can be found here.

Importing Data

First of all, we need to import our data and the software packages that we will use to describe, analyse, and visualise the data. The following lines of code achieve this by importing a series of software packages and then loading a csv file `covid_patients_syn_data.csv` into a DataFrame `df` using the `pd.read_csv` command from the Pandas package.

[1]:
```
# The following lines import necessary packages and renames them

import numpy as np
import pandas as pd
import matplotlib.pylab as plt
import seaborn as sns

# This line imports data from a csv file

df = pd.read_csv('covid_patients_syn_data.csv')
```

Describing the Data

Once we have imported our data, we will then want to identify what variables there are, what their typical values are, and also assess a variety of other summary statistics. We can use several commands to help us describe our dataset and get a quick overview.

First, we can use the `shape` attribute to list the number of rows and columns in our dataset. The output (44252, 12) means that there are 44252 rows and 12 columns.

[2]:
```
df.shape
```

[2]:
```
(44252, 12)
```

Second, we can use the `head` attribute to return the first 5 rows of our dataset, which can be useful if you want to see a small sample of values for each variable.

[3]:
```
df.head()
```

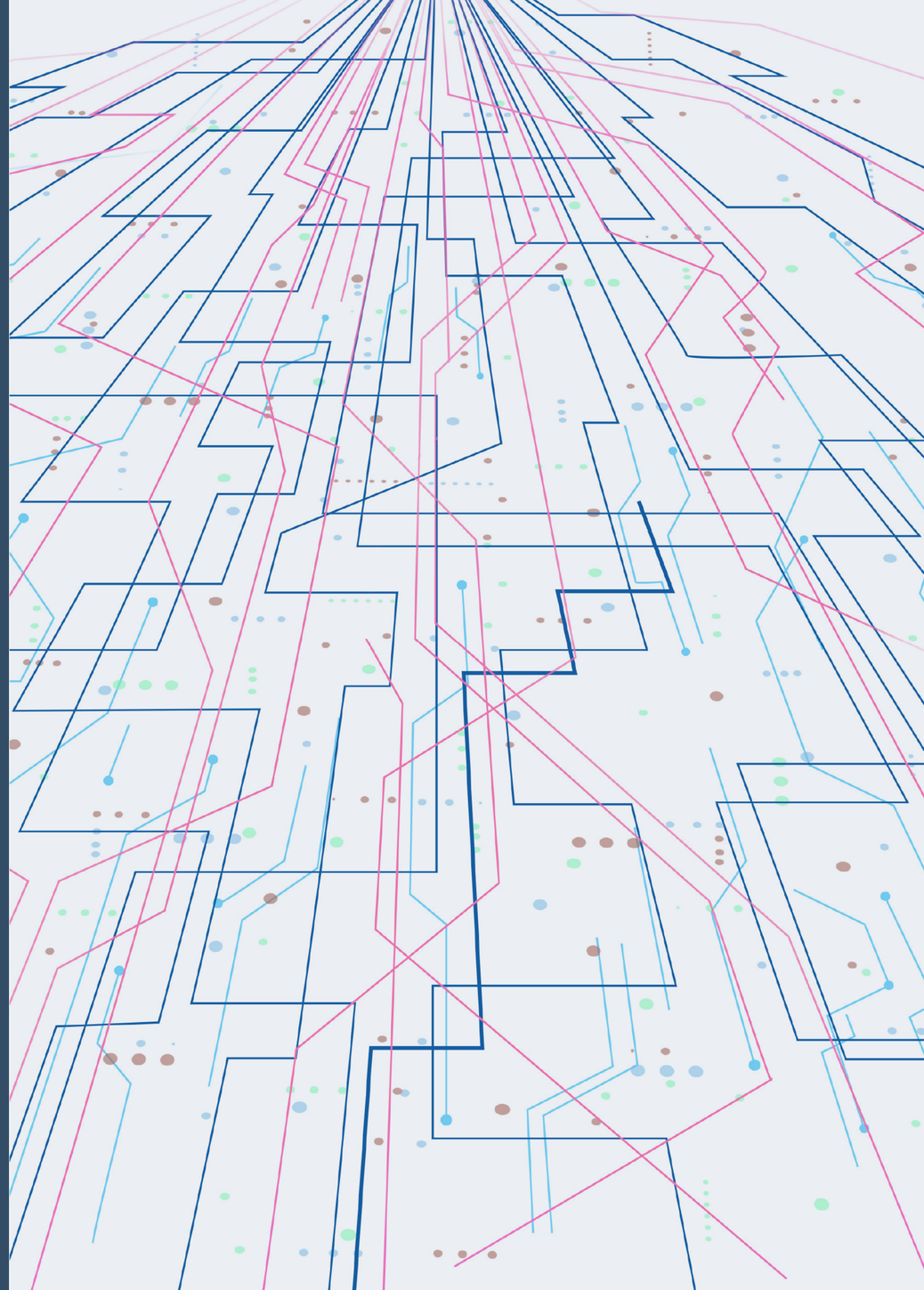
[3]:

	Unnamed: 0	nhs_id	site_id	age	sex	ethnicity	height	weight	admitted	admission_date	intrusive_ventilation	died
0	0	031d28a9	UHJ_43643	84	M	White	1.768606	72.135416	False	NaN	False	False
1	11495	9bfa8fb8	UHJ_43643	46	M	White	1.547297	81.346290	True	2020-04-25	False	False
2	11510	306432ae	UHJ_43643	38	M	White	1.743599	68.299982	False	NaN	False	False

Simple 0 1 Python 3 (ipykernel) | Idle Mem: 167.05 / 8192.00 MB Mode: Command Ln 1, Col 1 data_analysis.ipynb

Activity 8

Identifying Missing Data



Breakout Groups

Plenary

Thank you!

See you at the Guest Lecture!

The Alan Turing Institute

Dr Christopher Burr
cburr@turing.ac.uk
@chrisdburr