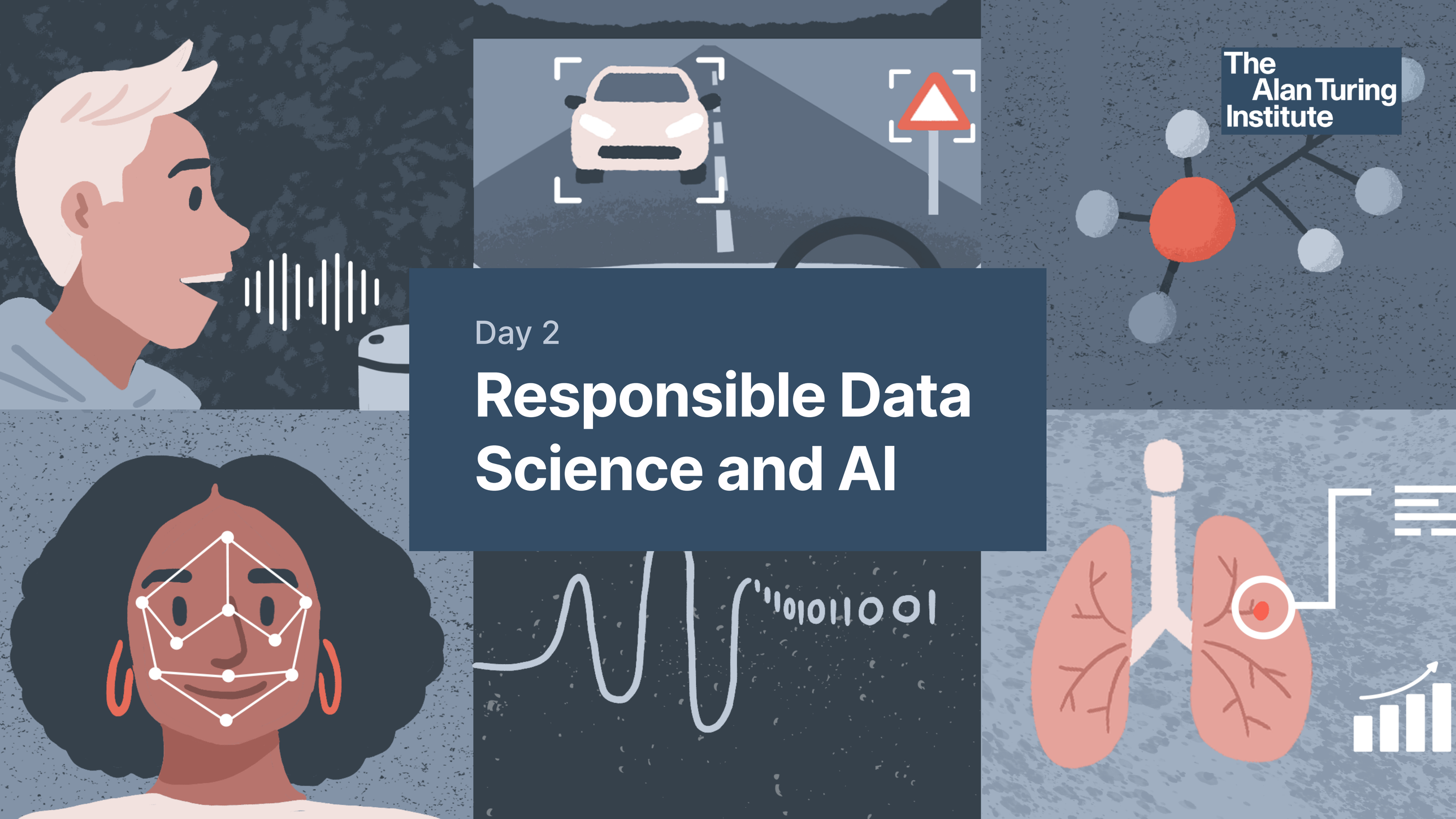

The Alan Turing Institute
Day 2
Responsible Data Science and AI
10101001

Recap

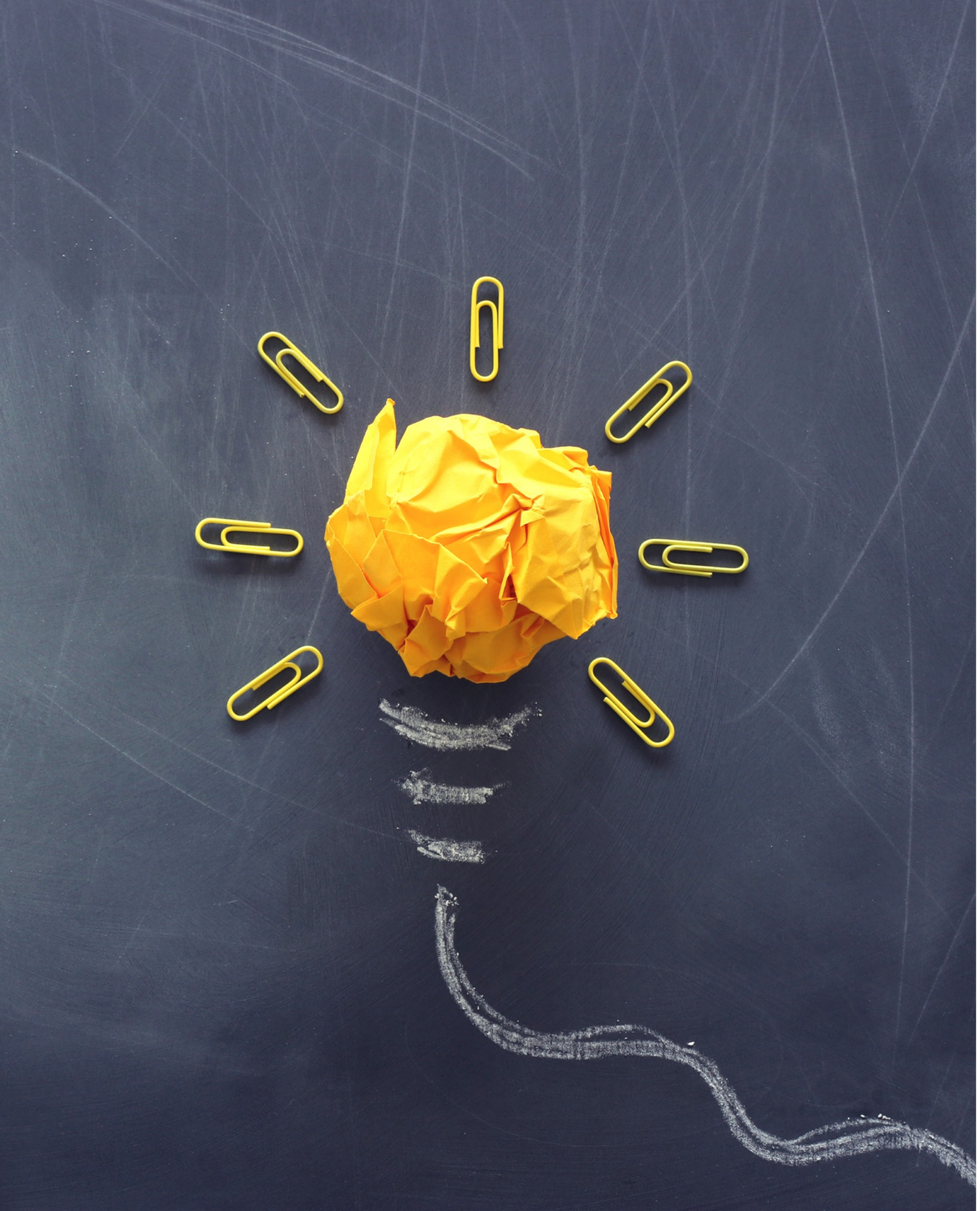
Day 1

- What is RRI?
  - Interactive approach to research and innovation
  - Emphasis on ethical acceptability of projects
  - Recognition of science and technology as socially embedded
- Principles grounded in felt needs and challenges
- Procedural Approach to RRI
  - Risk and Impact Assessment
  - Stakeholder Engagement
- Exploring agents, subjects, and values

Overview

Day 2

- Responsible Data Science and AI
- The Project Lifecycle
- Roles and Responsibilities
- Understanding Bias



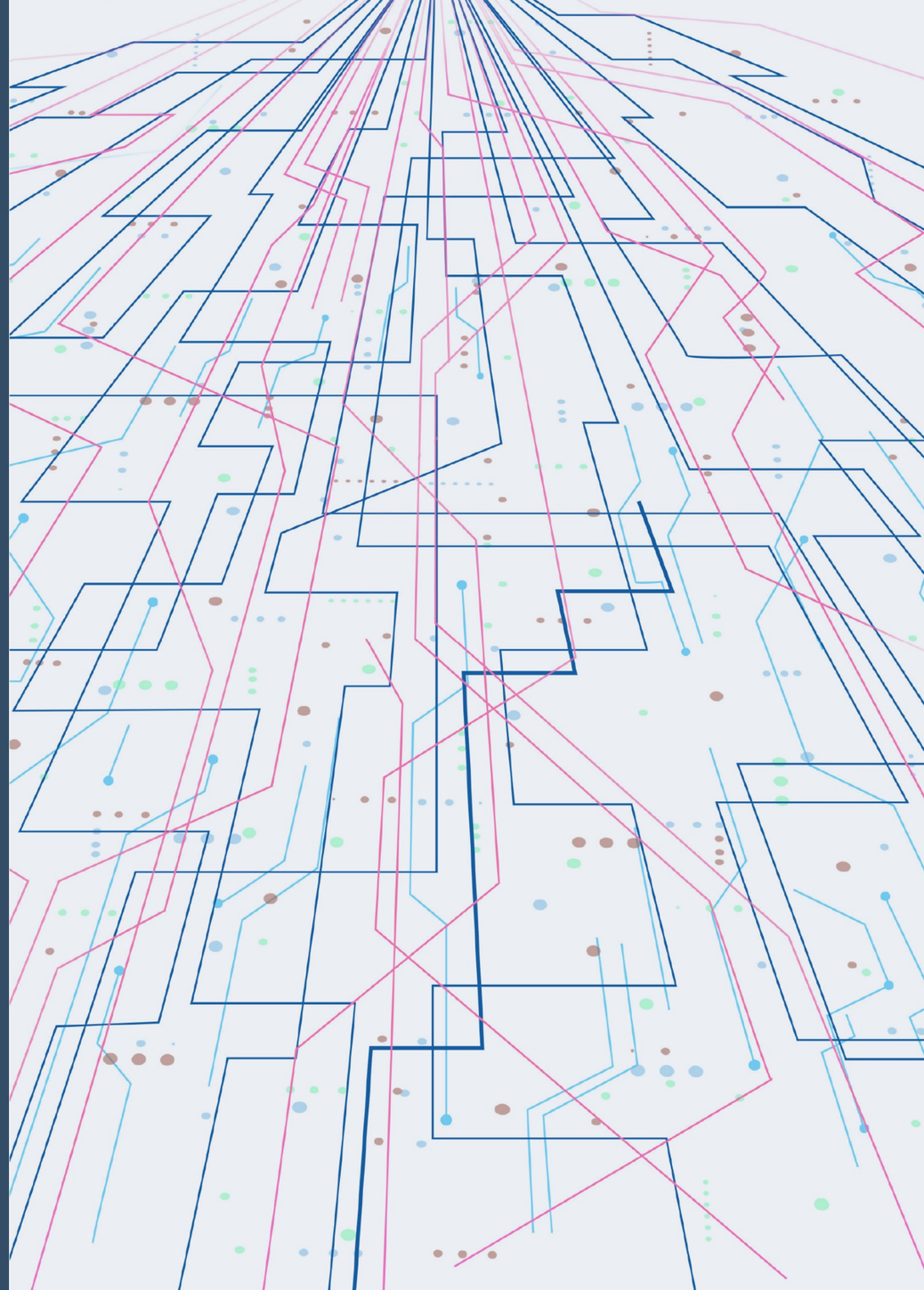
Day 2

Learning Objectives

- Explore what differentiates responsible data science and AI from responsible research and innovation more generally.
- Examine a model of a typical project lifecycle to better appreciate why individual responsibility is often insufficient in the context of data science and AI.
- Understand the differences between social, statistical, and cognitive biases, and why they all matter for responsible data science and AI.

Day 2

Responsible Data Science and AI



What separates responsible data science and AI from responsible research and innovation more generally?

Predicting Risk

Avon and Somerset Police and Bristol City Council developed a sophisticated predictive risk tool that was used, among other things, to predict the risk of children suffering sexual abuse. But, the Bristol Cable reported that many children were falsely flagged as being at risk, and that the tool was developed using dozens of public sector databases, including schools, housing, NHS records, and even credit scores from Experian.

Racist Photo Cropping Tool

Twitter was forced to apologise after many users reported that the automated tool for cropping images on the social media platform showed a racial bias towards faces of white people over faces of black people. According to Twitter, one source of the issue was the use of a “saliency algorithm” that was trained on human eye-tracking data.

Lethal Autonomous Weapons

Turkish company STM manufactures the Kargu-2—an attack drone that can operate autonomously by using machine learning and real-time image processing to identify targets. According to a UN security council report this drone was reported to have been used to “remotely engage” and “hunt down” logistics convoys and retreating forces in the Libyan civil war during 2019.

Predicting Risk

Avon and Somerset Police and Bristol City Council developed a sophisticated predictive risk tool that was used, among other things, to predict the risk of children suffering sexual abuse. But, the Bristol Cable reported that many children were falsely flagged as being at risk, and that the tool was developed using dozens of public sector databases, including schools, housing, NHS records, and even credit scores from Experian.

Predicting Risk

Avon and Somerset Police and Bristol City Council developed a sophisticated predictive risk tool that was used, among other things, to predict the risk of children suffering sexual abuse. But, the Bristol Cable reported that many children were falsely flagged as being at risk, and that the tool was developed using dozens of public sector databases, including schools, housing, NHS records, and even credit scores from Experian.

Racist Photo Cropping Tool

Twitter was forced to apologise after many users reported that the automated tool for cropping images on the social media platform showed a racial bias towards faces of white people over faces of black people. According to Twitter, one source of the issue was the use of a “saliency algorithm” that was trained on human eye-tracking data.

Lethal Autonomous Weapons

Turkish company STM manufactures the Kargu-2—an attack drone that can operate autonomously by using machine learning and real-time image processing to identify targets. According to a UN security council report this drone was reported to have been used to “remotely engage” and “hunt down” logistics convoys and retreating forces in the Libyan civil war during 2019.

Racist Photo Cropping Tool

Twitter was forced to apologise after many users reported that the automated tool for cropping images on the social media platform showed a racial bias towards faces of white people over faces of black people. According to Twitter, one source of the issue was the use of a “saliency algorithm” that was trained on human eye-tracking data.

Predicting Risk

Avon and Somerset Police and Bristol City Council developed a sophisticated predictive risk tool that was used, among other things, to predict the risk of children suffering sexual abuse. But, the Bristol Cable reported that many children were falsely flagged as being at risk, and that the tool was developed using dozens of public sector databases, including schools, housing, NHS records, and even credit scores from Experian.

Racist Photo Cropping Tool

Twitter was forced to apologise after many users reported that the automated tool for cropping images on the social media platform showed a racial bias towards faces of white people over faces of black people. According to Twitter, one source of the issue was the use of a “saliency algorithm” that was trained on human eye-tracking data.

Lethal Autonomous Weapons

Turkish company STM manufactures the Kargu-2—an attack drone that can operate autonomously by using machine learning and real-time image processing to identify targets. According to a UN security council report this drone was reported to have been used to “remotely engage” and “hunt down” logistics convoys and retreating forces in the Libyan civil war during 2019.

Lethal Autonomous Weapons

Turkish company STM manufactures the Kargu-2—an attack drone that can operate autonomously by using machine learning and real-time image processing to identify targets. According to a UN security council report this drone was reported to have been used to “remotely engage” and “hunt down” logistics convoys and retreating forces in the Libyan civil war during 2019.

Reflection and Deliberation

Normative Principles

What is the role of normative principles?

They serve to distil values and norms into action-guiding constraints for reflection and deliberation. However, they are often insufficient to determine actions without further specification.

Consider the principle of ‘respect for autonomy’.

Reflection and Deliberation

SAFE-D Principles

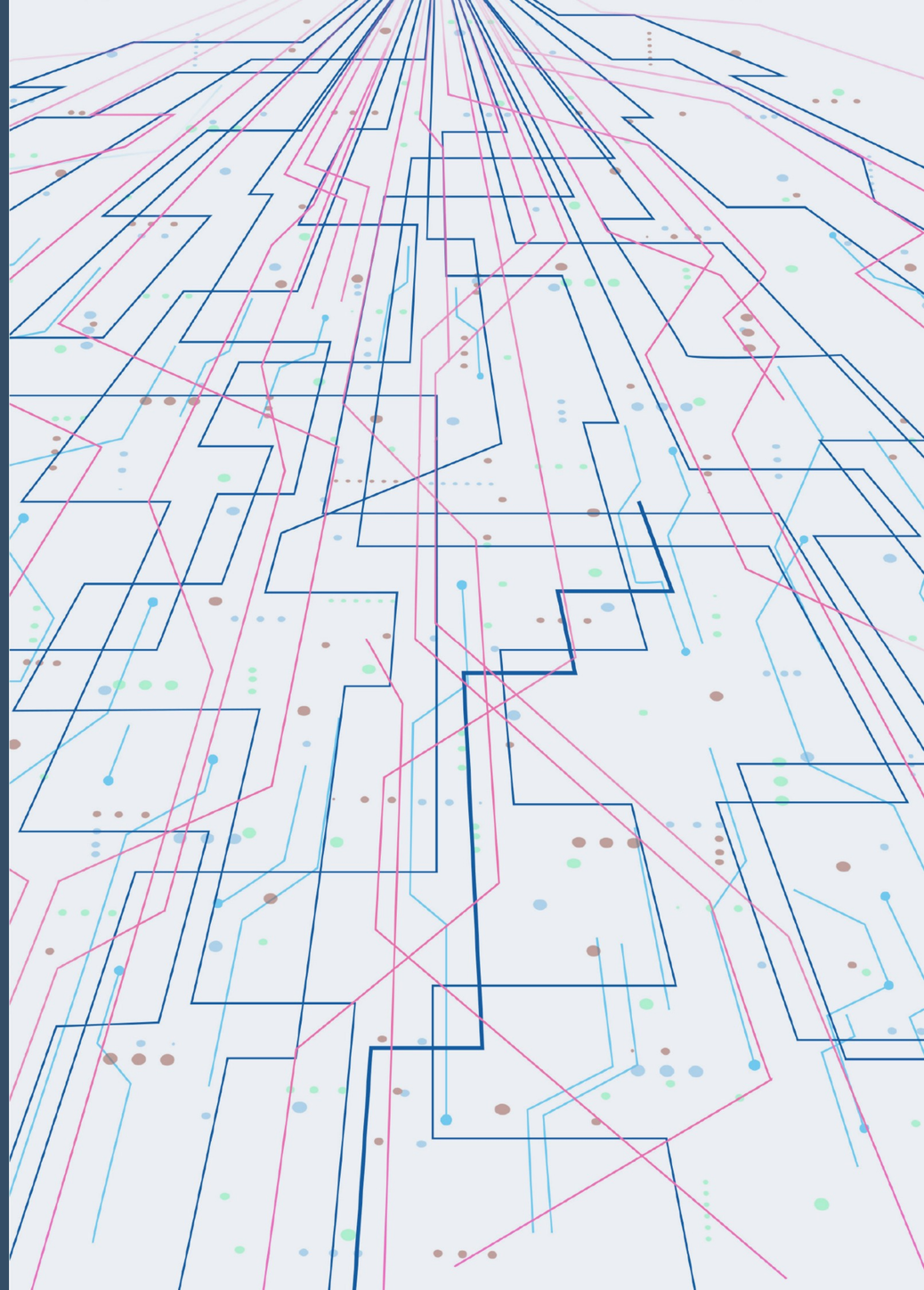
The following principles, known as the SAFE-D principles have been designed to support the responsible governance of data science and AI:

- Sustainable
- Accountable
- Fair
- Explainable
- Data (Quality, Integrity and Protection)

But specifying and operationalising them can only be done in conjunction with stakeholders and domain experts.

Activity 3

Specifying Principles



Breakout Groups

Plenary

SAFE-D Principles

Sustainability

Sustainability requires the outputs of a project to be:

- safe, secure, robust, and reliable
- informed by ongoing consideration of the risk of exposing individuals to harms even after the system has been deployed and the project completed—a long-term (or sustainable) form of safety.

SAFE-D Principles

Accountability

Accountability can refer to transparency of processes and associated outcomes that enable people to understand how a project was conducted (e.g., project documentation), or why a specific decision was reached. But it can also refer to broader processes of responsible project governance that seek to establish clear roles of responsibility where full transparency may be inappropriate (e.g., confidential projects).

SAFE-D Principles

Fairness

Fairness is inseparably connected with legal conceptions of equality and justice, which may emphasise a variety of features such as non-discrimination, equitable outcomes, or procedural fairness through bias mitigation.

However, these notions serve as a subset of broader normative considerations pertaining to social justice, socioeconomic capabilities, diversity and inclusivity.

SAFE-D Principles

Explainability

Explainability is a key condition for autonomous and informed decision-making in situations where data-driven systems interact with or influence human judgement and choice behaviour.

Explainability goes beyond the ability to merely interpret specific aspects of a project (e.g., interpreting the parameters of a model); it also depends on the ability to provide an accessible and relevant information base about the processes behind the outcome.

SAFE-D Principles

Data Quality, Integrity, Protection and Privacy

‘Data Quality’ captures the static properties of data, such as whether they are (a) relevant to and representative of the domain and use context, (b) balanced and complete in terms of how well the dataset represents the underlying data generating process, and (c) up-to-date and accurate as required by the project.

SAFE-D Principles

Data Quality, Integrity, Protection and Privacy

‘Data Integrity’ refers to more dynamic properties of data stewardship, such as how a dataset evolves over the course of a project lifecycle. In this manner, data integrity requires (a) contemporaneous and attributable records from the start of a project (e.g., process logs; research statements), (b) ensuring consistent and verifiable means of data analysis or processing during development, and (c) taking steps to establish findable, accessible, interoperable, and reusable records towards the end of a project’s lifecycle.

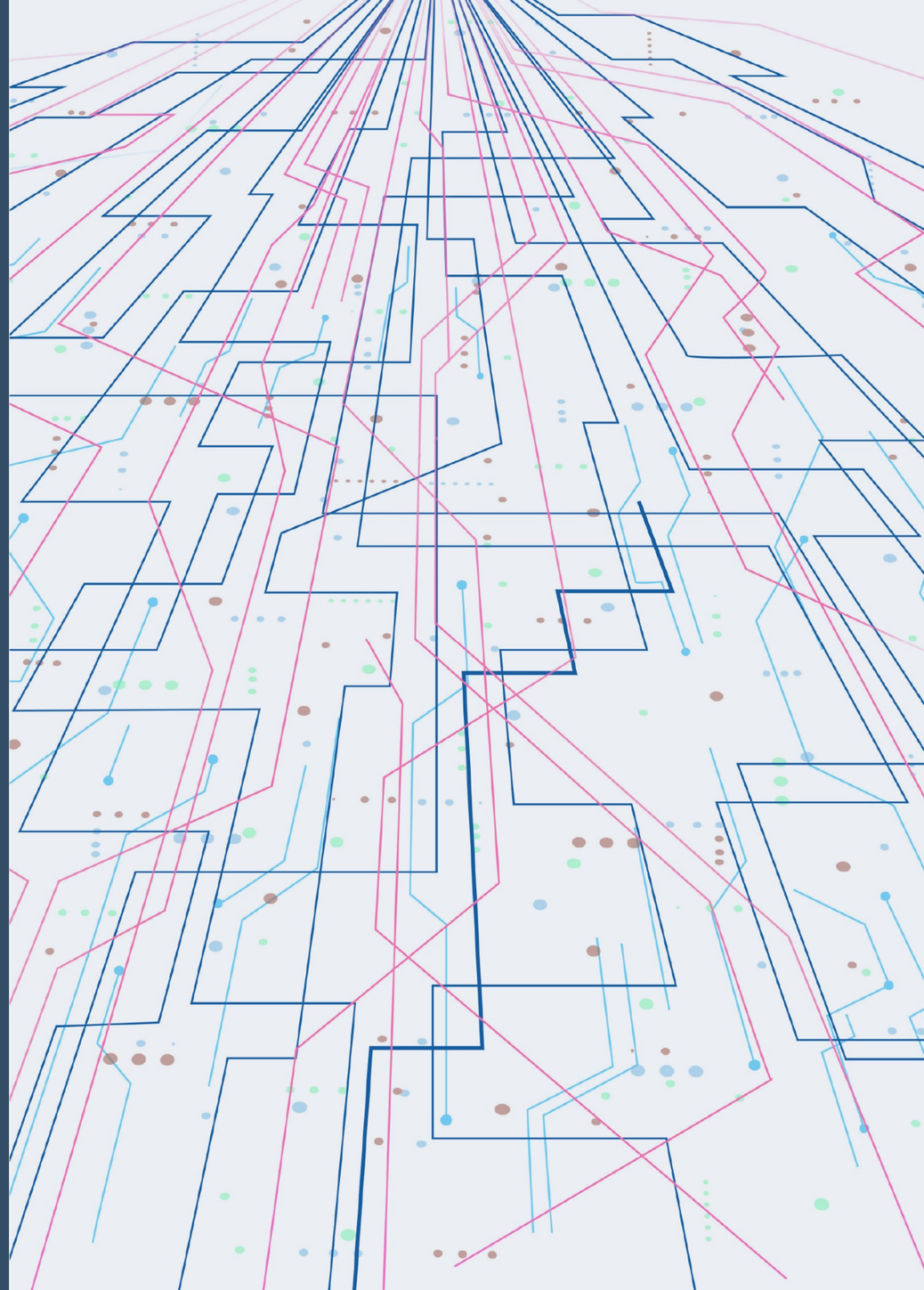
SAFE-D Principles

Data Quality, Integrity, Protection and Privacy

‘Data protection and privacy’ reflect ongoing developments and priorities as set out in relevant legislation and regulation of data practices as they pertain to fundamental rights and freedoms, democracy, and the rule of law. For example, the right for data subjects to have inaccurate personal data rectified or erased.

Activity 4

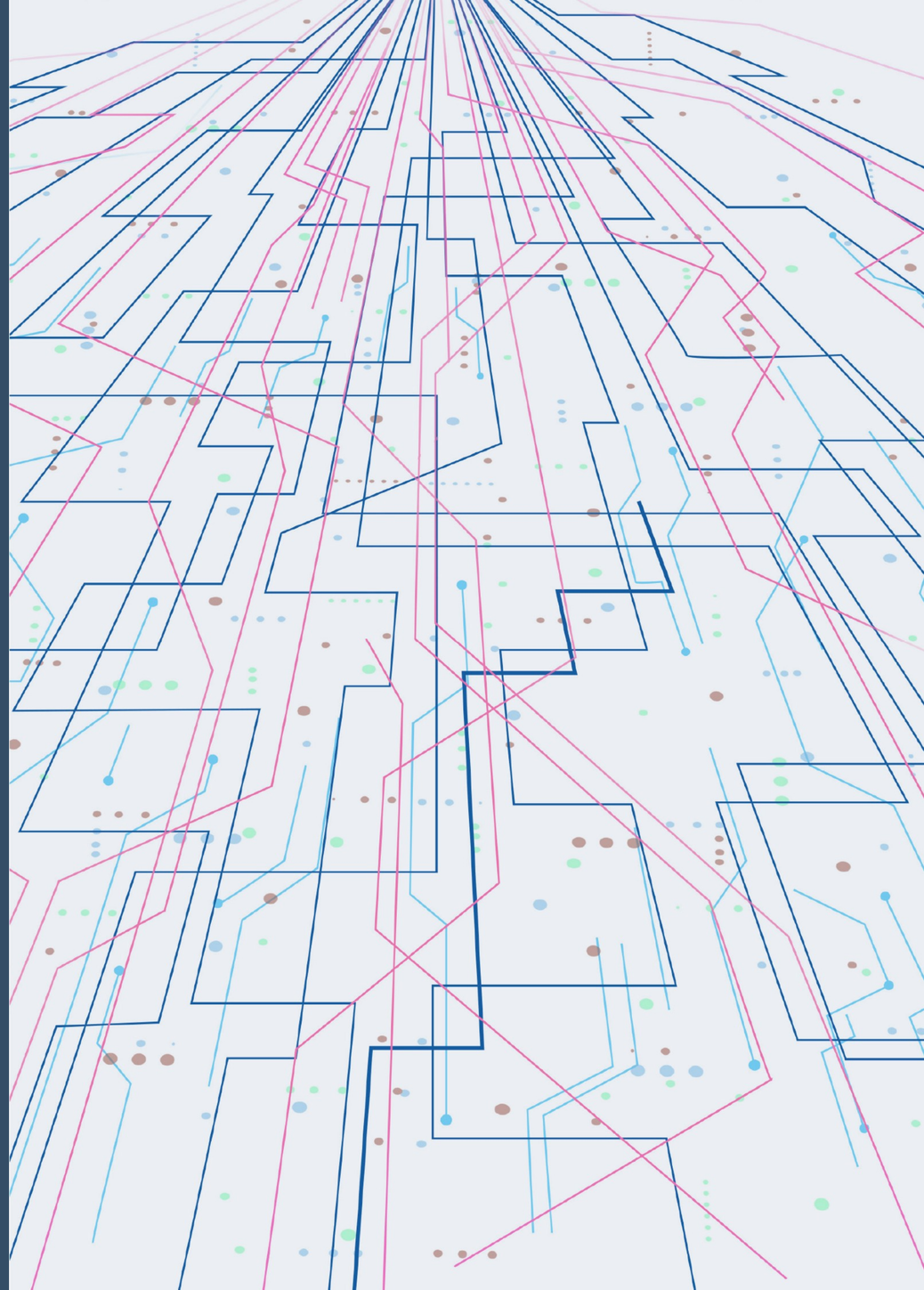
Contributing to Collaborative Projects

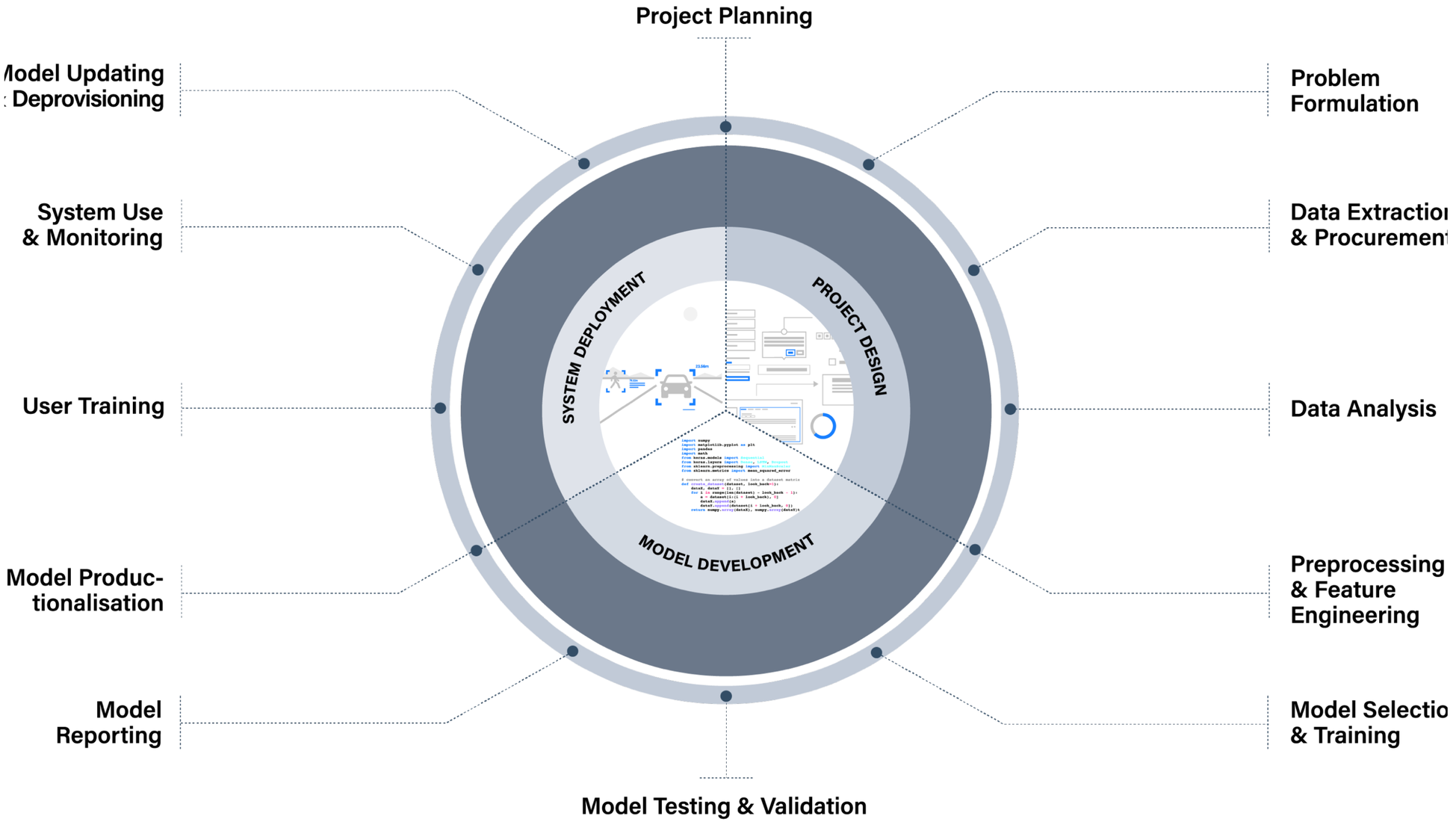


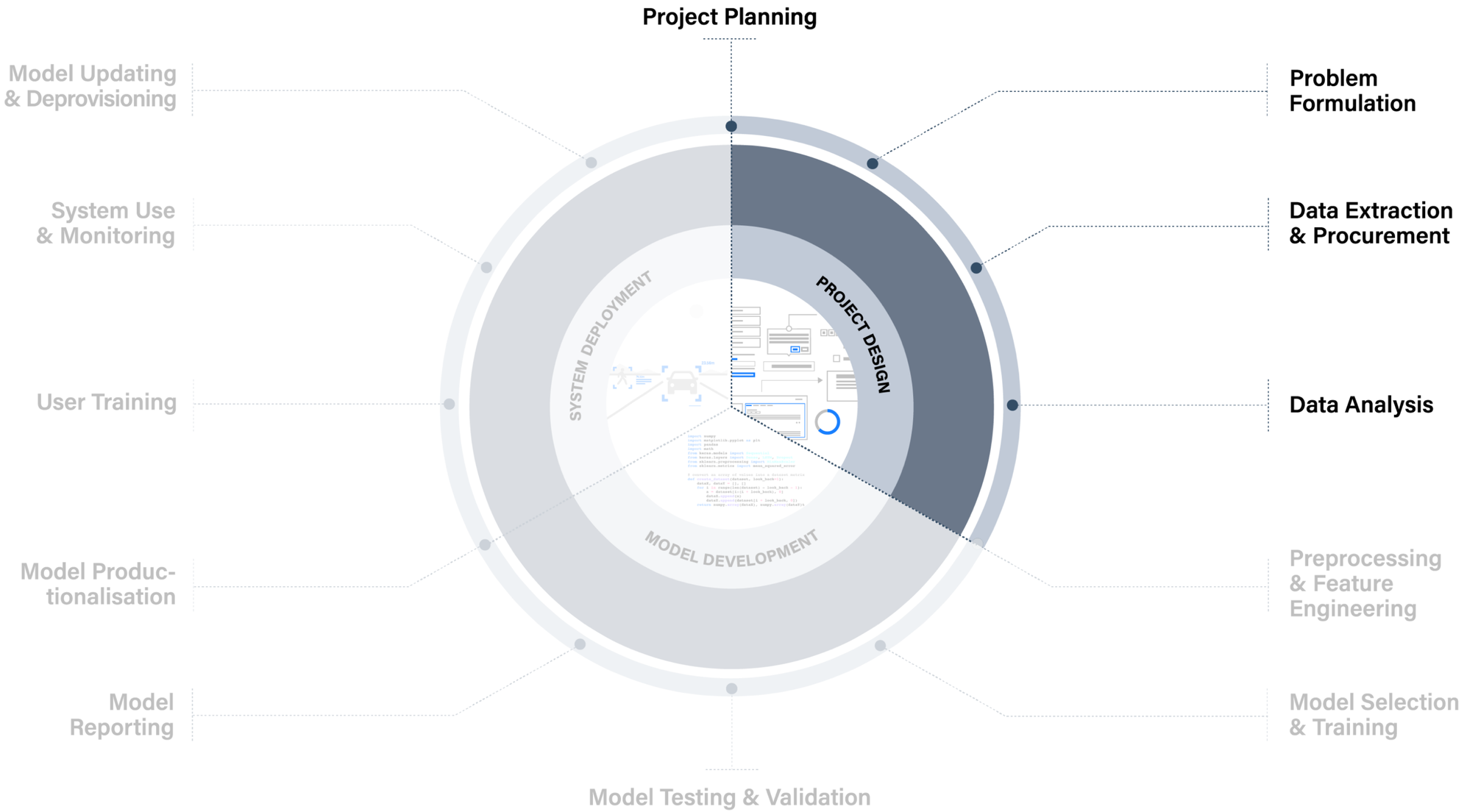
Lunch

Day 2

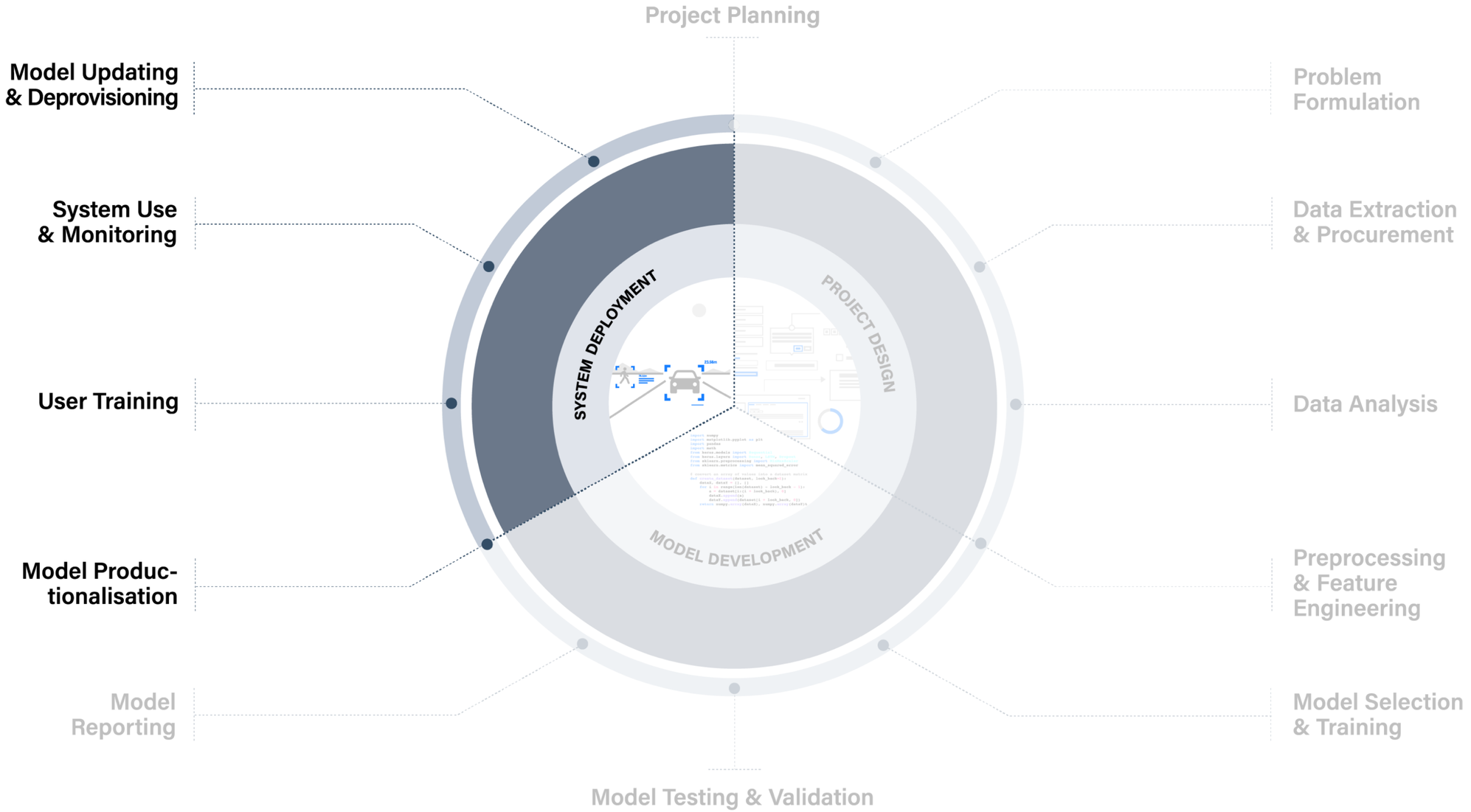
Introducing the Project Lifecycle

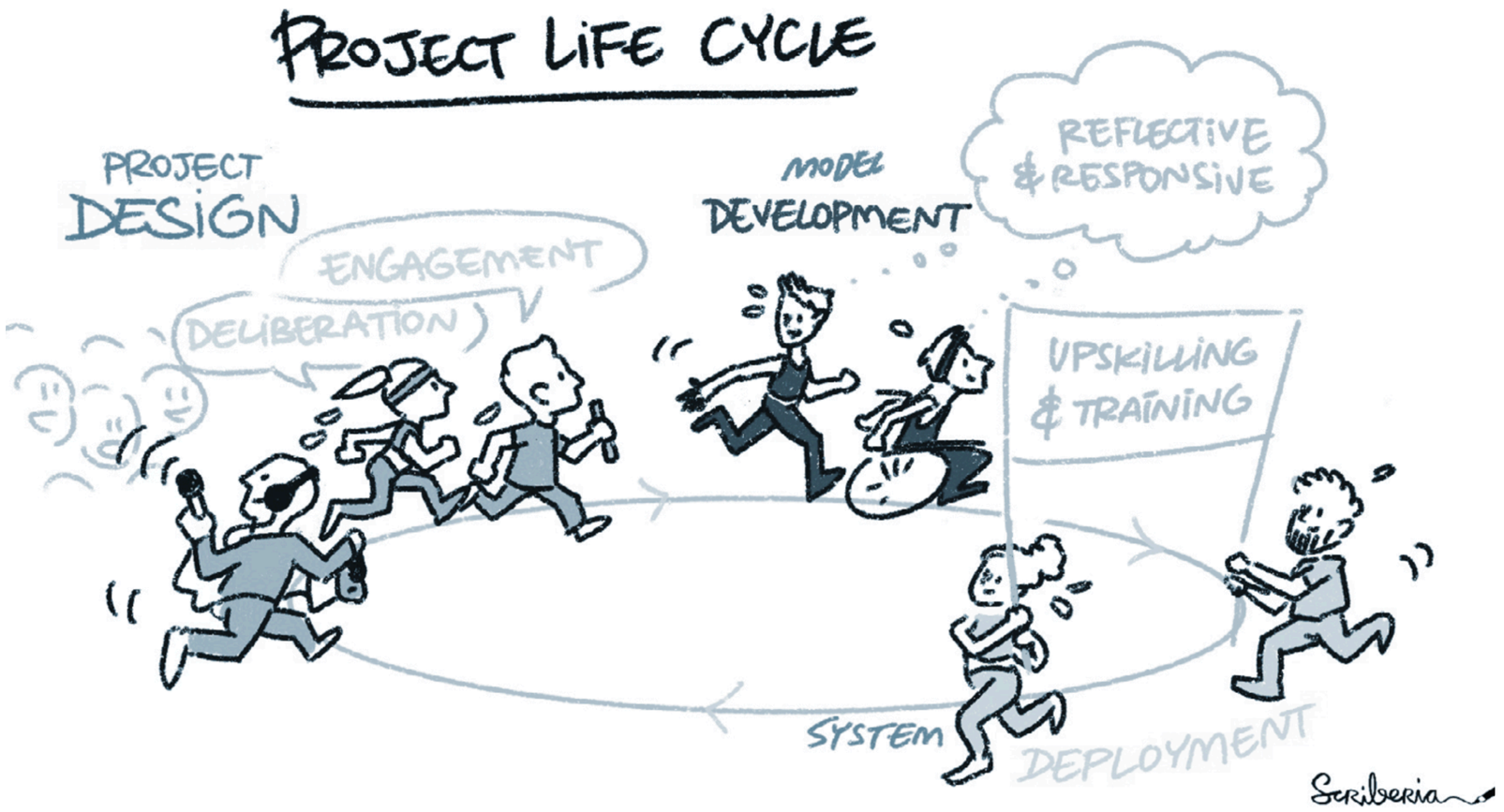



Project Planning
Problem Formulation
Data Extraction & Procurement
Data Analysis
Preprocessing & Feature Engineering
Model Selection & Training
Model Testing & Validation
Model Reporting
Model Production-tionalisation
User Training
System Use & Monitoring
Model Updating & Deprovisioning
PROJECT DESIGN
MODEL DEVELOPMENT
SYSTEM DEPLOYMENT


Project Planning
Model Updating & Deprovisioning
System Use & Monitoring
User Training
Model Productionalisation
Model Reporting
Model Testing & Validation
Problem Formulation
Data Extraction & Procurement
Data Analysis
Preprocessing & Feature Engineering
Model Selection & Training
SYSTEM DEPLOYMENT
PROJECT DESIGN
MODEL DEVELOPMENT


Project Planning
Model Updating & Deprovisioning
Problem Formulation
System Use & Monitoring
Data Extraction & Procurement
User Training
Data Analysis
Model Productionalisation
Preprocessing & Feature Engineering
Model Reporting
Model Selection & Training
Model Testing & Validation
SYSTEM DEPLOYMENT
PROJECT DESIGN
MODEL DEVELOPMENT


Project Planning
Model Updating & Deprovisioning
System Use & Monitoring
User Training
Model Productionalisation
Model Reporting
Model Testing & Validation
Problem Formulation
Data Extraction & Procurement
Data Analysis
Preprocessing & Feature Engineering
Model Selection & Training
SYSTEM DEPLOYMENT
PROJECT DESIGN
MODEL DEVELOPMENT



Collective

Roles and Responsibilities

Within a team, individual roles and responsibilities interconnect to form a collective responsibility.

Issues upstream will cascade down to effect downstream actions.

Project members should not be reduced to their functional role as this limits their creativity and reflective contributions.

Situating Responsibility

Situated Project Lifecycle

Projects don’t operate in a vacuum.

The project lifecycle presupposes a context in which it is situated.

Responsibility requires a contextual awareness of the norms, values, and expectations of the culture in which the project operates.

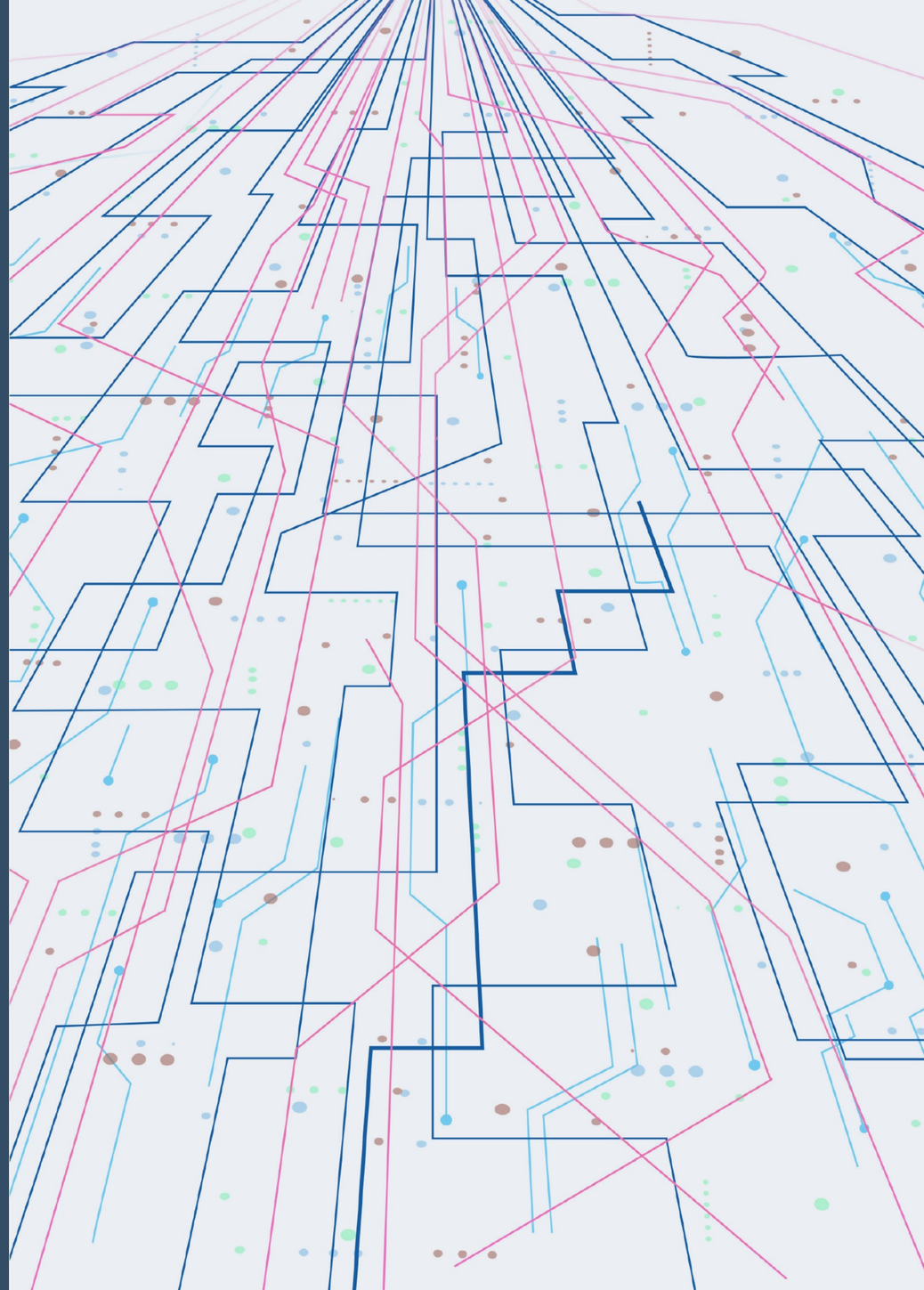
Situating Responsibility

Team Dynamics

- Who is responsible for the glue work in your team?
- Who receives praise and credit?
- How do you ensure everyone has a voice?
- How does this credit affect subsequent opportunities?

Day 2

Understanding Bias

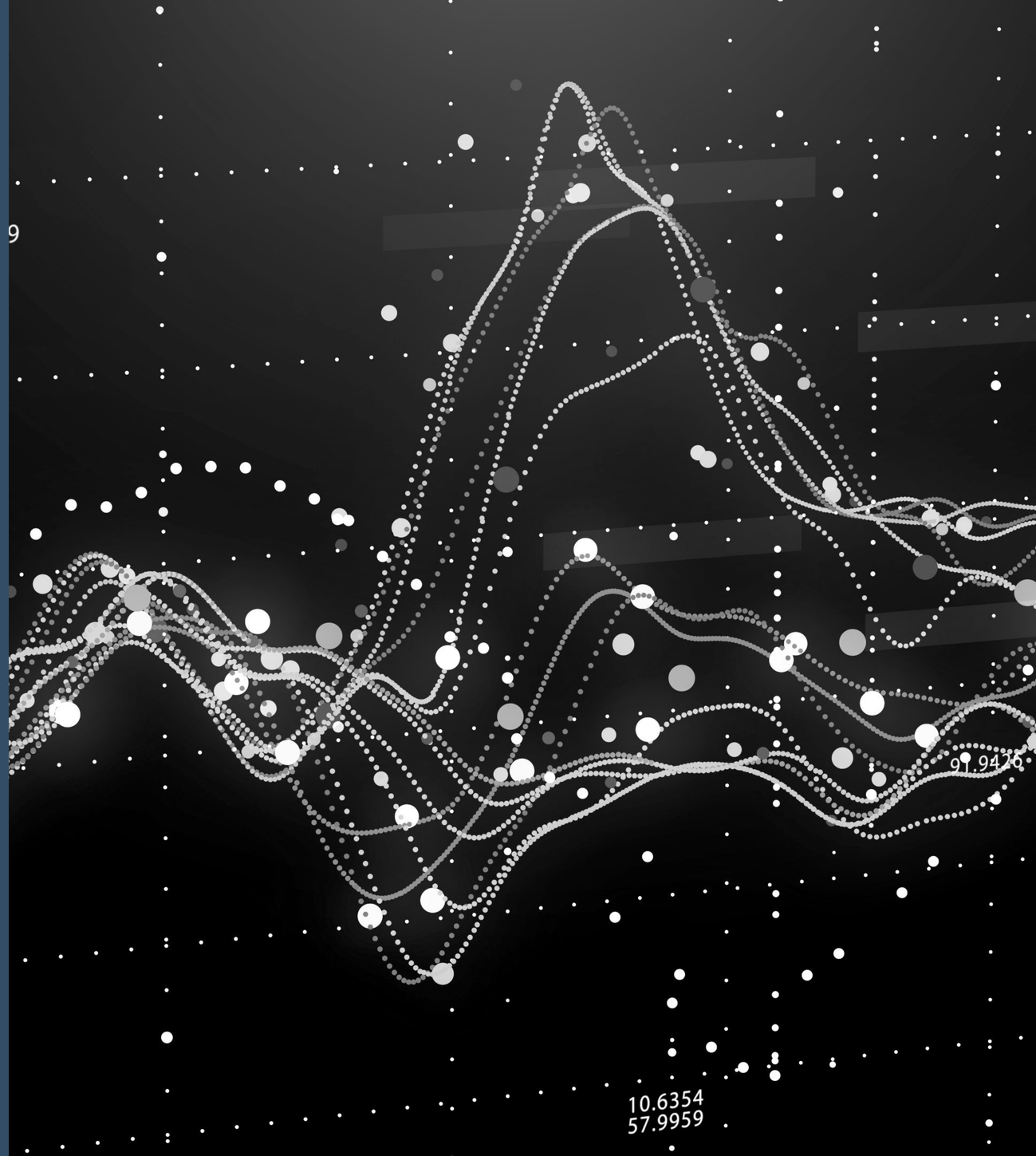


Understanding Bias

More than Just Statistics

Three concepts of bias:

1. Social Bias
2. Statistical Bias
3. Cognitive Bias



72%

Mothers who had to reduce work hours due to insufficient childcare during the pandemic

Source: Pregnant Then Screwed (2020) Childcare, Covid and Career: The true scale of the crisis facing working mums

Understanding Bias

Social Bias

Amazon’s recruitment tool that perpetuated bias in hiring against women.

This algorithmic system learned to perpetuate a bias to prefer male candidates to female candidates because this reflected past hiring decisions.

Understanding Bias

Social Bias

Predictive policing that use geospatial data to try to learn associations between places, events, and historical crime rates.

The attempt to predict where and when crimes are more likely to happen can create a positive feedback loop, which results in over-policing that may exacerbate tensions between communities and police.

Understanding Bias

Social Bias

Clinical decision support systems can contribute to existing forms of racial bias in access to healthcare.

A study conducted in the US found that an algorithm that used health costs as a proxy for health needs was “less likely to refer black people than white people who were equally sick to programmes that aim to improve care for patients with complex medical needs”. (Obermeyer et al., 2019)

Understanding Bias

Statistical Bias

Statistical definitions of ‘bias’ will typically make reference to one of the following features (Aronson, 2018):

- Systematicity: bias arises from a systematic process, rather than a random or chance process.
- Truth: a realist assumption that the deviation is from a true state of the world
- Error: the bias reflects an error, perhaps due to sampling or measurement
- Deviation (or Distortion): a quantity in which the observed result is taken to differ from the actual result were there no bias.
- Affected elements: the study elements that may be affected by the bias include the conception, design, and conduct of the study, as well as the collection, analysis, interpretation, and representation of the data
- Direction: the deviation is directional, as it can be caused by both an under- or over-estimation

Understanding Bias

Cognitive Bias

Our decision-making and judgement is affected by a wide variety of psychological vulnerabilities.

Cognitive Biases

Judgement and Decision-Making

Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.

Which is more probable?

A. Linda is a bank teller.
B. Linda is a bank teller and is an active campaigner for women's rights.

Tversky, A and Kahneman, D. (1982) Judgments of and by representativeness

Go to www.menti.com
and use the code
4086 5326

Cognitive Biases

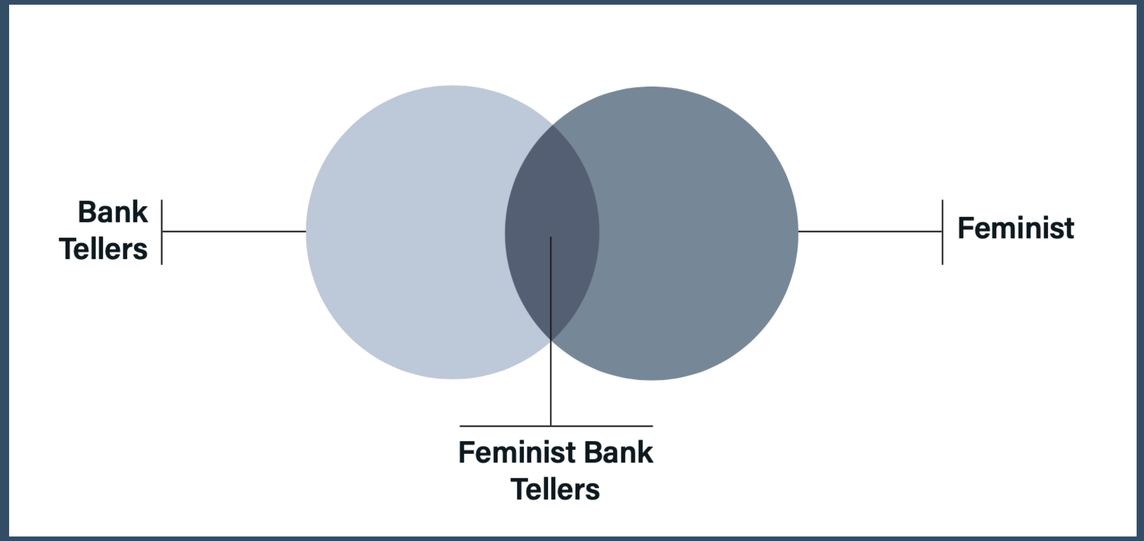
Judgement and Decision-Making

The correct answer is (A). Did you get it right?

If you got it wrong, you have just committed what is known as the ‘conjunction fallacy’.

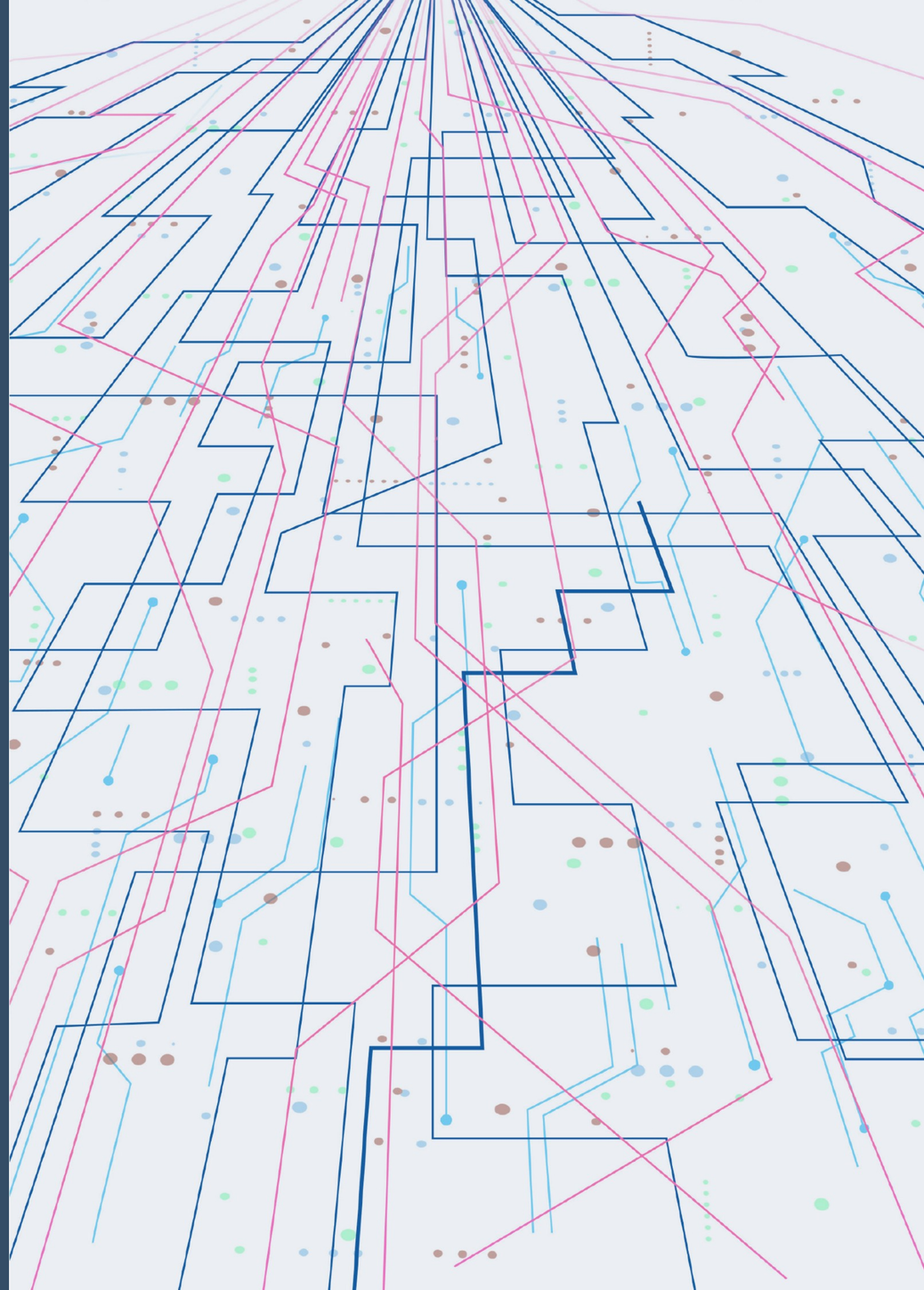
When Tvserky and Kahneman posed this question to a group of 88 undergraduate students, only 15 got the correct answer.

Tversky and Kahneman attributed this systematic error to what is known as the representativeness heuristic.



Activity 5

Bias Cards



Breakout Groups

Plenary

Day 3

Tomorrow

- The Project Lifecycle (Project Design)
  - Project Planning
  - Problem Formulation
  - Data Extraction and Procurement
  - Data Analysis
- Guest Lecture (Professor Sabina Leonelli)

Thank you!

See you tomorrow!

The Alan Turing Institute

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