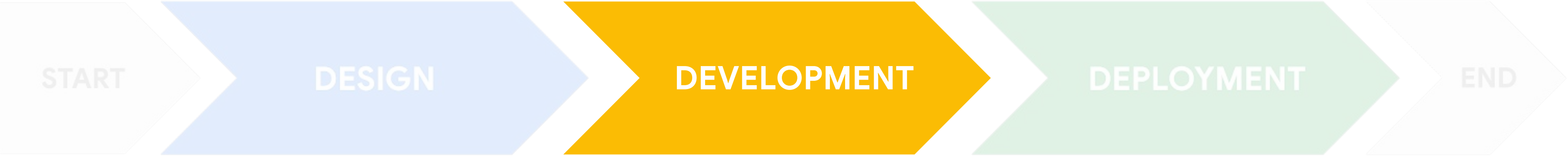
What Data Will Be Collected?

Responsible AI: Human-Centered Design



Course 1

Fundamentals of TinyML

- **What** am I building?
- **Who** am I building this for?
- What are the **consequences** for the user if it ***fails***?

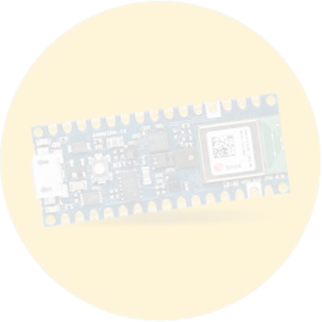
Course 2

Applications of TinyML

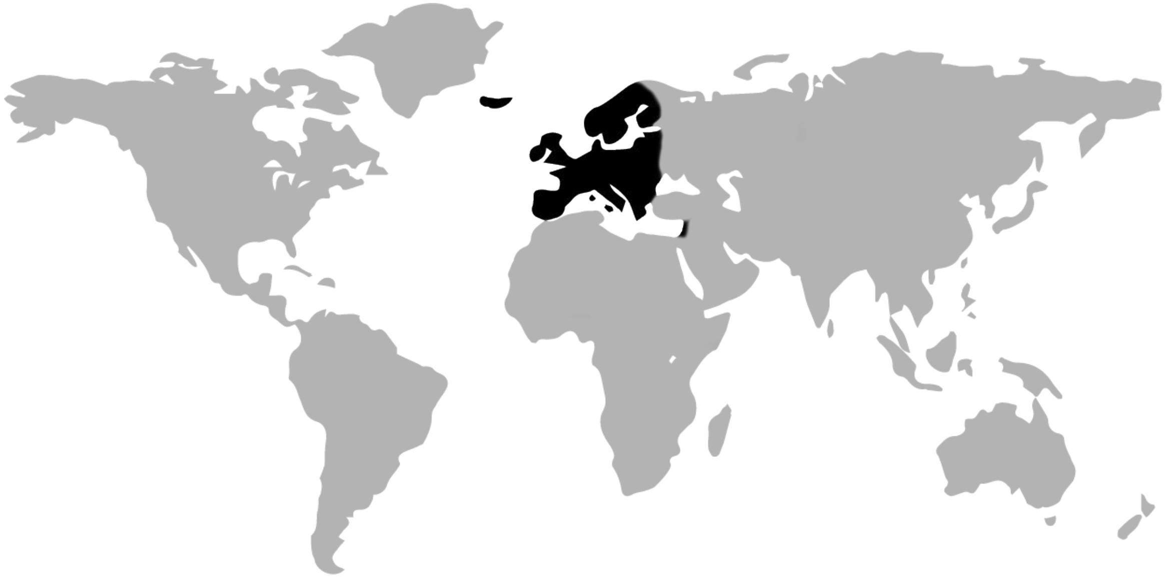
- **What data will be collected to train the model?**
- Is the dataset biased?
- How can we ensure the model is fair?

Course 3

Deploying TinyML



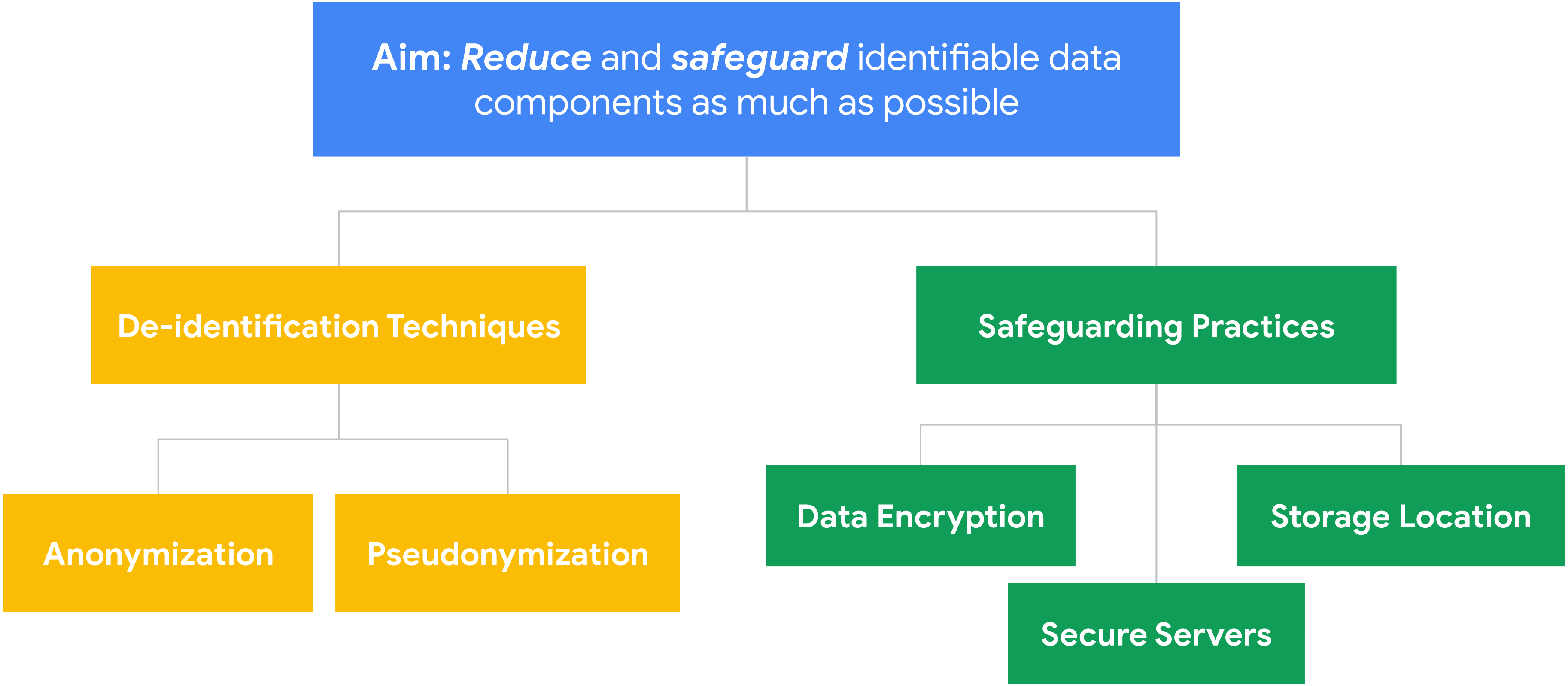
Data **Laws** and **Regulations**



3 Key Features of Data Protection

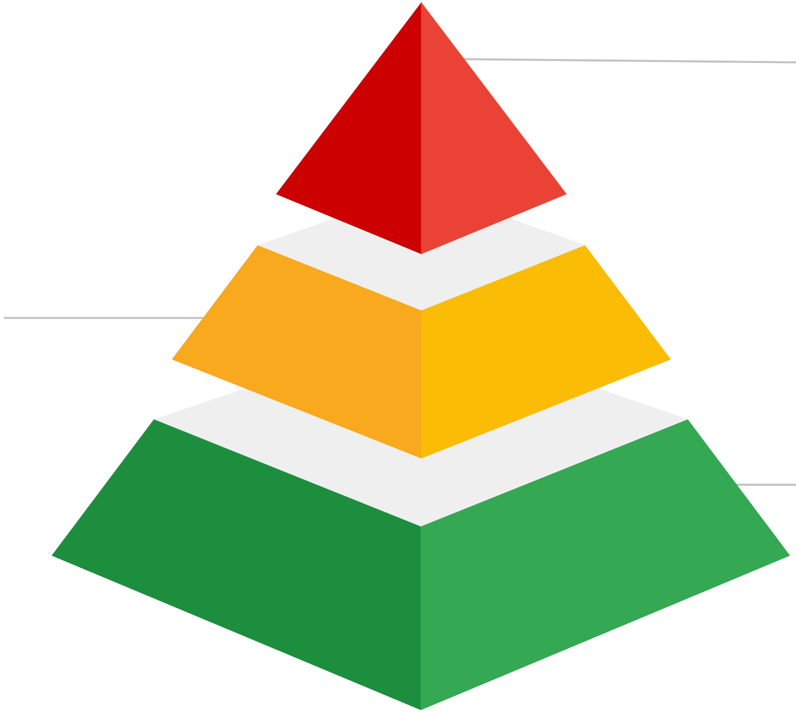
- Identifiability
- Data Minimization
- Notice and Consent

1. Identifiability



Aim: ***Reduce*** and ***safeguard*** identifiable data components as much as possible

Levels of Data Classification



Sensitive Data

Genetic, biometric and health data, as well as data that reveals race, ethnicity, political or religious affiliation etc.

Personal Data

Data that includes name and surname, home address, location, email address, IP address etc.

Non-personal Data

Generalized data (e.g. age range), aggregated statistics, data collected by government bodies (e.g. census data), etc.

2. Data minimization

Aim: ***Limit*** data collection and duration of storage to **only** what is **required** to fulfill a specific purpose

- **How long** will I need the data to achieve the purpose?
- Is there **unnecessary data** that can be deleted?
- How often should I periodically **review** data **and delete** what isn’t needed?

Right to be Forgotten (**Erasure**)

Subjects have the **right* to request that their data be erased** by the data controller, as soon as possible.

**This right may be overridden in some cases.*

What should data collectors do?

- Provide subjects with **clear information** and **practical ways** to make a request for data erasure

3. Notice and Consent

Aim: Prepare ***clear notice*** and ***consent*** communication to data subjects

- **Notice** ensures that subjects are ***aware*** of the intended data practices
- **Consent** ensures that subjects are only ***implicated*** in those practices ***if they want to be***.

Necessary Conditions of **Informed Consent**

Informed	Subject has sufficient knowledge and comprehension of the matter to enable an enlightened decision	No lies, deceit, or partial disclosure
Voluntary	Subject freely chooses to give consent	No coercion, inappropriate pressure or influence
Competent	Subject has the decisional capacity to offer consent	No children, adults deemed mentally incompetent

How will data be collected?

Mechanisms for Data Collection

1. Crowdsourcing
2. Product users
3. Paid contributors (mechanical turk)

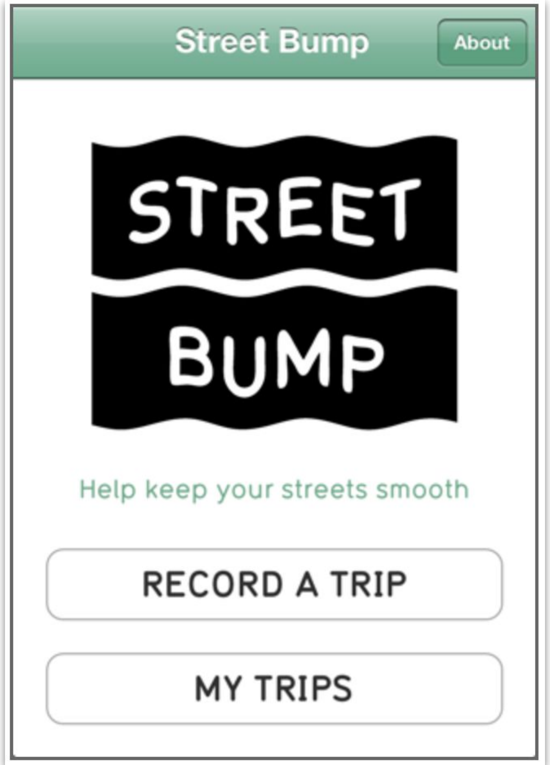
> “...we need to ask which people are excluded. Which places are less visible? What happens if you live in the shadow of big data sets?”

Kate Crawford
Principal Researcher at Microsoft and Professor at NYU Tandon School of Engineering

Source: *“The Hidden Biases in Big Data,” Harvard Business Review, April 2013.*

Data Collection: **Product Users**

Does the demographic of product users accurately represent the population?



Open Datasets

Open source, publicly available datasets to foster innovation and healthy competition!

Accent

23% United States English, **8%** England English, **5%** India and South Asia, **4%** Australian English, **3%** Canadian English, **2%** Scottish English, **1%** Irish English, **1%** Southern African, **1%** New Zealand English

Age

23% 19–29, **14%** 30–39, **10%** 40–49, **6%** < 19, **4%** 50–59, **4%** 60–69, **1%** 70–79

Source: https://commonvoice.mozilla.org/en/datasets