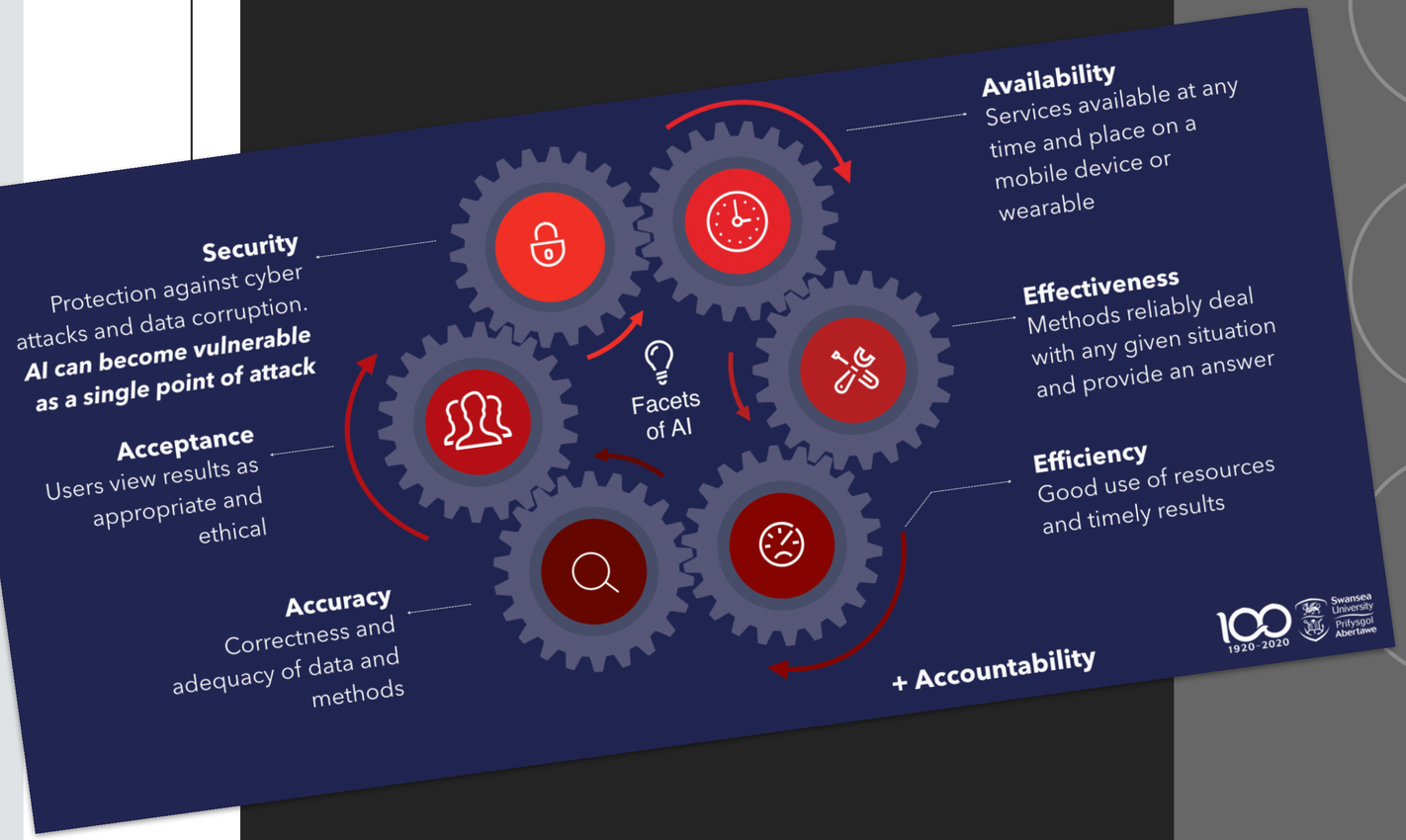
CSCM23 DESIGNING-IN TRUST, UNDERSTANDING, AND NEGOTIATION

1920~2020
Swansea University
Prifysgol Abertawe

RE³ APPROACH

CURRENT RESEARCH AIMS

RE³ = Reliable, Responsible, & Resilient



Learning & Reasoning

Combining machine learning with rule-based systems of reasoning, e.g., multi-agent systems.

Partial explainability

Counterfactuals

Explainability of decisions by automatic application of counterfactual reasoning.

Experimental explainability

Transparency

Use of tagged datasets and results to retain information about, e.g., the provenance, bias

Theory of propagation of bias.

Dynamic Assurances

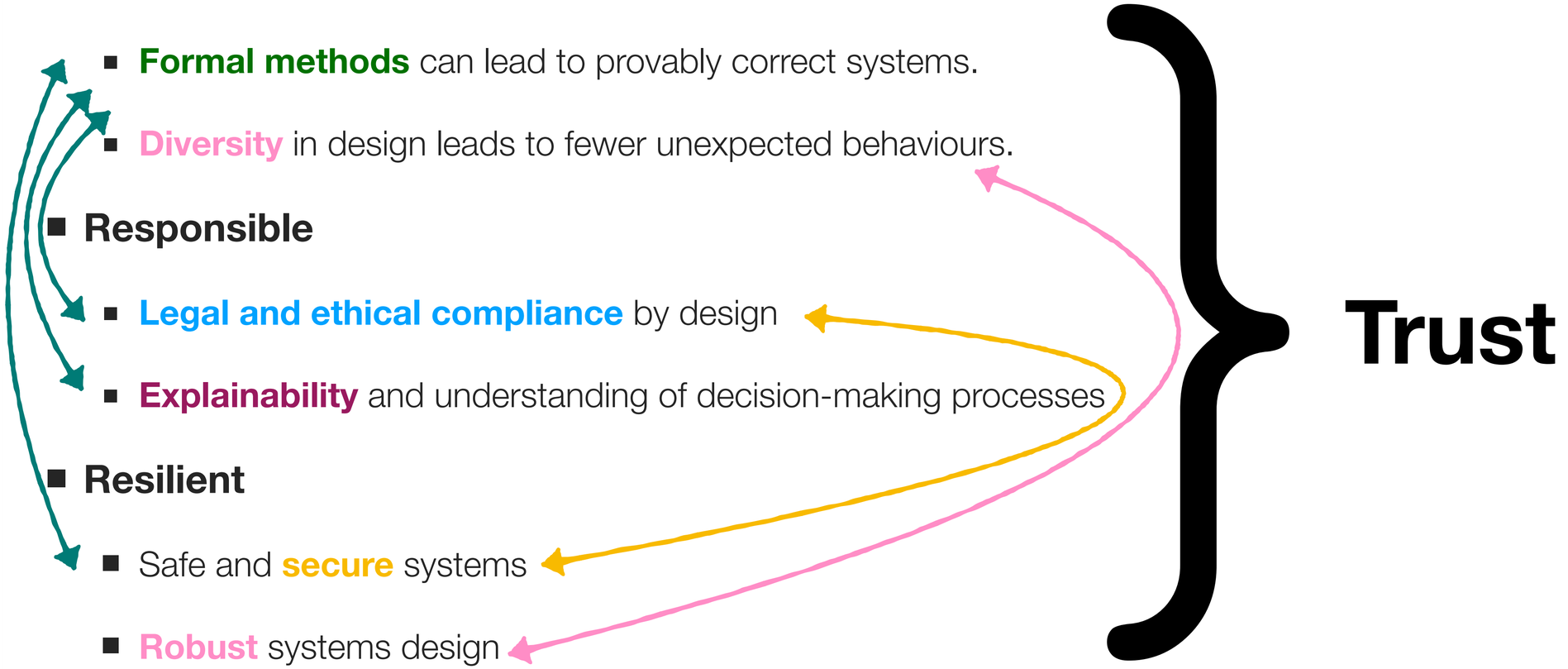
Autonomy requires new processes for deploying and monitoring systems that rely on automation and learning to ensure accuracy and long-term reliability in changing contexts.

Data governance, standards, regulation

OUTLOOK

SO WHAT SHOULD YOU EXPECT FROM THIS MODULE?

- **Reliable**
  - **Formal methods** can lead to provably correct systems.
  - **Diversity** in design leads to fewer unexpected behaviours.
- **Responsible**
  - **Legal and ethical compliance** by design
  - **Explainability** and understanding of decision-making processes
- **Resilient**
  - Safe and **secure** systems
  - **Robust** systems design



RE³ APPROACH

RELIABLE

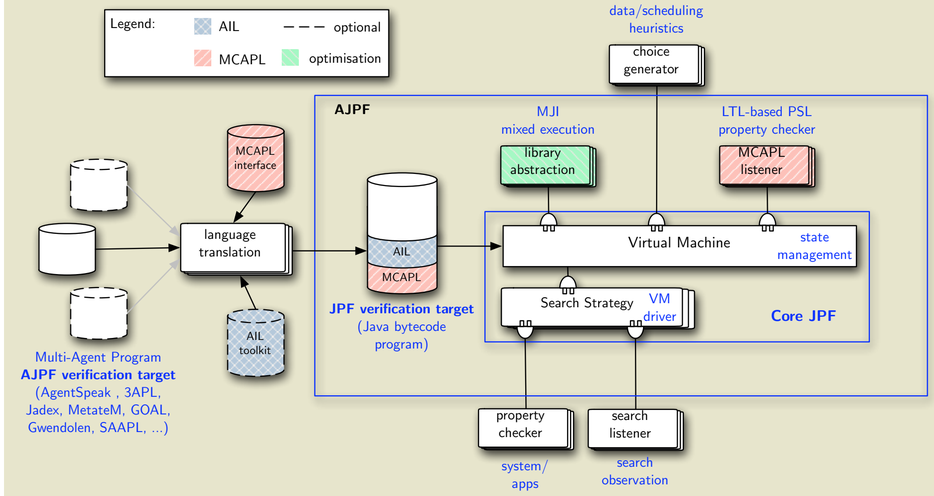
Accuracy

Efficiency

Availability

Explainability

Predictability



- Mechanisms of introducing bias into datasets and algorithms:
  - **Automatic:** Training on non-representative or otherwise inherently biased data
  - **Unintentional:** insufficiently diverse team/methods/ expertise increase the chance of bias introduced by ignorance.
  - **Deliberate:** biased or fake data introduced in any phase of design, implementation, or at runime.
- Can bias be qualified and quantified?

- BCS SIGiST – Summer 2017 Conference – Keynote: **Towards Verifiable and Ethical AI Systems**, London, 14 June 2017
- AI – Artificial Intelligence in Reality – **Engineering Verifiable Agent Programs**, BCS North London Branch, 25 November 2015
- **Modern Applications of Agent Technology – Where Mobility and Resources Actually Matter**, TU Clausthal, 1 February 2011
- R. Bordini, L. Dennis, B. Müller, M. Fisher. **Directions for Agent Model Checking**. Book chapter in: Specification and Verification of Multi-agent Systems. Pages 103-124, Springer. 2010.
- N. Bulling and B. Müller. **Expressing properties of resource-bounded systems: The logics RTL and RTL***. In J. Dix et al, eds, Computational Logic in Multi-Agent Systems. CLIMA X. Vol. 6214 of LNAI, pages 22-45, Springer 2010.
- Modelling and Verification of Resource-Bounded Multi-Agent Systems., Keynote at MATES – MOCA'09, 11 September 2009

GOOD DESIGN

Should lead to expected behaviour

GOOD DESIGN

Should try to eliminate unexpected behaviour

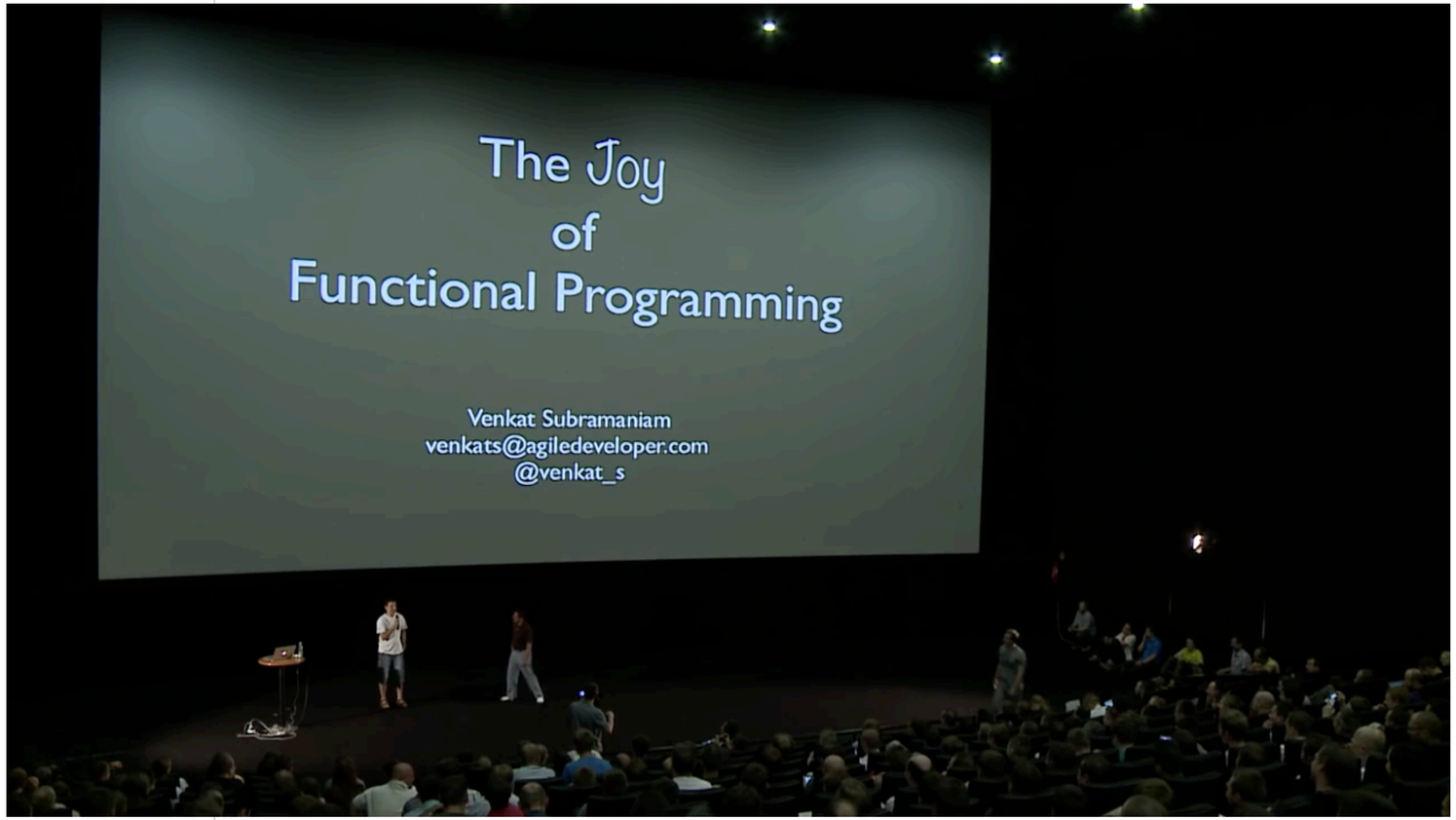
PRIVACY

GOOD DESIGN

Excerpt from a talk on good programming …

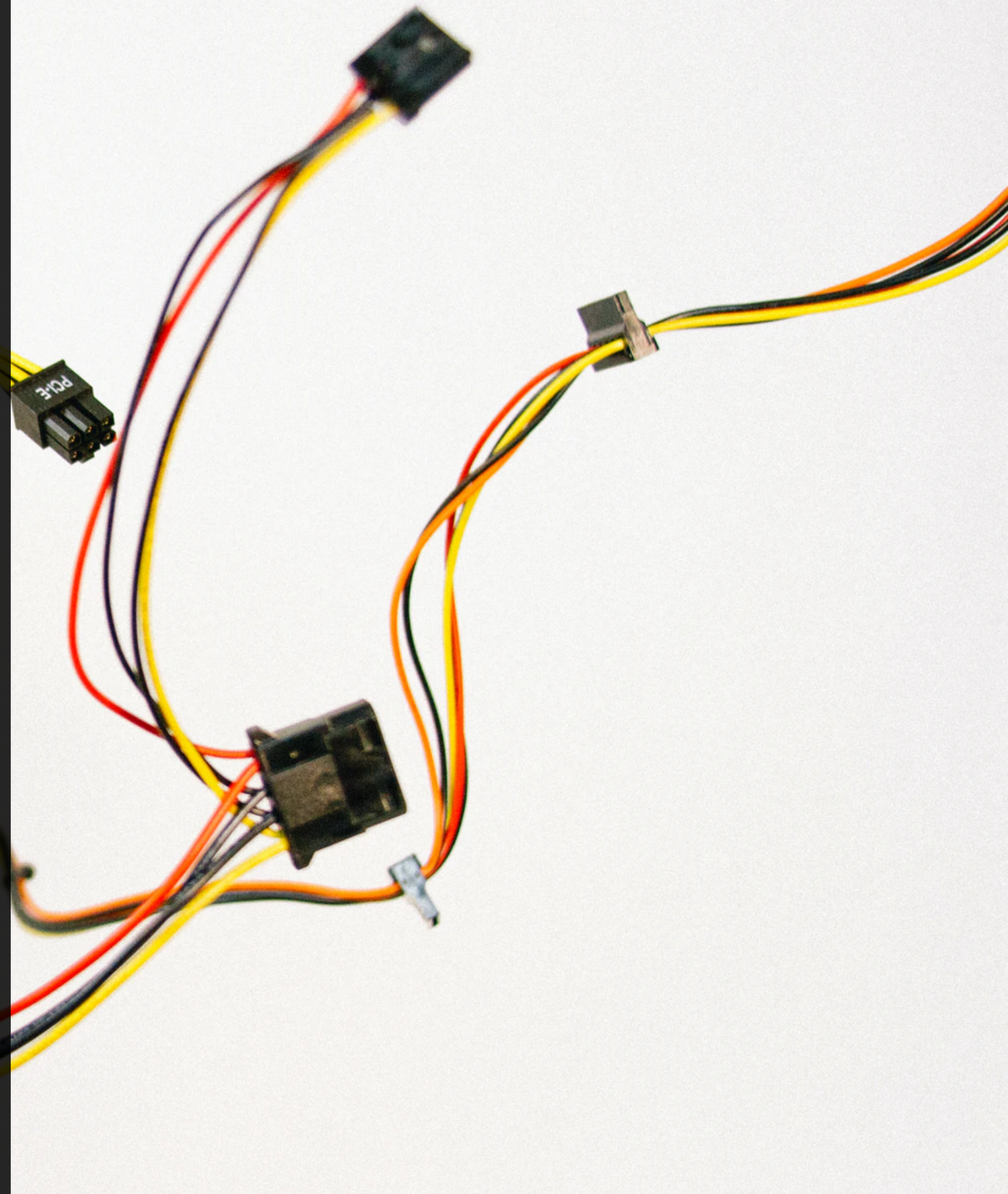
Bad design leads to bad products.

This is particularly true for privacy design.



LET'S DISCUSS

Your experience of trust in technology products and services.



PRIVACY

PRIVACY IS A HUMAN RIGHT

The right to privacy is explicitly stated under **Article 12** of the **1948 Universal Declaration of Human Rights**:

“No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence, nor to attacks upon his honor and reputation.”

- In the UK, human rights are protected by the **Human Rights Act 1998**. The Act gives effect to the human rights set out in the **European Convention on Human Rights**.
- **Article 8** - the right to respect for your family and private life, your home and your correspondence is one the rights protected by the Human Rights Act.
- This covers:
  - your sexuality
  - your body
  - personal identity and how you look and dress
  - forming and maintaining relationships with other people
  - how your personal information is held and protected

PRIVACY LAW: GDPR (25 MAY 2018)

The General Data Protection Regulation (GDPR) covers:

- Rules for business and organisations
- Rights for citizens

… in relation to personal data.

DPA 2018

- **Transposes the EU Data Protection Directive** 2016/680 (Law Enforcement Directive) **into domestic UK law.**
- **Covers exceptions for national security**
- **Covers ICO and duties, functions and powers including the enforcement provisions.**

- Balance of interests of **data subjects** with **data controllers**.
- Freedom to process data vs. privacy of individuals.
- **ICO**:

  “The **GDPR has direct effect across all EU member states** and has already been passed. This means organisations will still have to comply with this regulation and we will still have to look to the GDPR for most legal obligations.

  “However, the **GDPR gives member states limited opportunities to make provisions for how it applies in their country.** One element of the DPA 2018 is the details of these. It is therefore important the GDPR and the DPA 2018 are read side by side. “

PRIVACY

DPA 2018 / GDPR PRINCIPLES

DPA 2018 principles

1. **Lawfulness**
2. **Purpose**
3. **Data minimisation**
4. **Accuracy**
5. **Storage**
6. **Access**
7. **Security**
8. **Overseas transfer**
9. **Accountability**
10.

But how about Brexit?

Personal Data:

– is about a person who is alive and can be identified by that data.

Data Subject:

– is the individual that the data is about.

Processing:

– retrieving, holding, sorting, deleting.

The Data Controller:

– is the person who is responsible for the control of the data in a business or organisation.

Relevant Filing System:

– Readily accessible information about living individuals.

The Commissioner:

– is the person responsible for enforcing the law, including ensuring the owners of the data use good practice, and the individuals are aware of their rights.

PRIVACY

ALGORITHMIC: FEDERATED LEARNING

Avoid centralising training data on one machine or in a data centre.

Federated learning works without the need to store user data in the cloud.

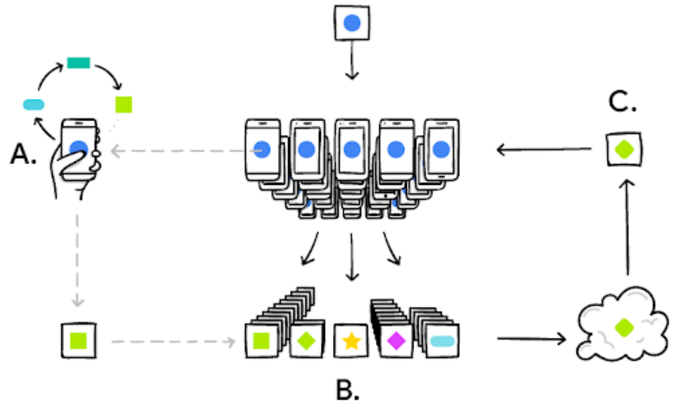
Applying Federated Learning requires machine learning practitioners to adopt new tools and a new way of thinking: model development, training, and evaluation with no direct access to or labeling of raw data, with communication cost as a limiting factor.

TensorFlow

TensorFlow supplies interfaces that facilitate federated learning tasks, such as federated training or evaluation with existing machine learning models implemented in TensorFlow.

Google AI

- "Federated Learning enables mobile phones to collaboratively learn a shared prediction model while keeping all the training data on device, decoupling the ability to do machine learning from the need to store the data in the cloud. This goes beyond the use of local models that make predictions on mobile devices […] by bringing model training to the device as well."



The phone personalises the model locally, based on your usage (A). Many users' updates are aggregated (B) to form a consensus change (C) to the shared model, after which the procedure is repeated.

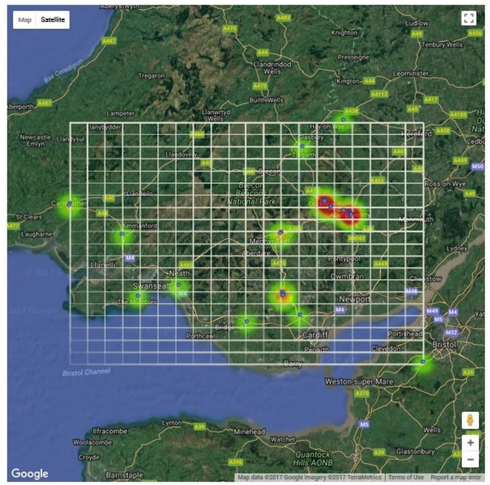
Latest flavour: agnostic federated learning, where the centralised model is optimised for any target distribution formed by a mixture of the client distributions.
This framework naturally yields a notion of fairness.

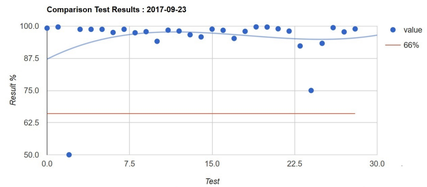
ETHICS

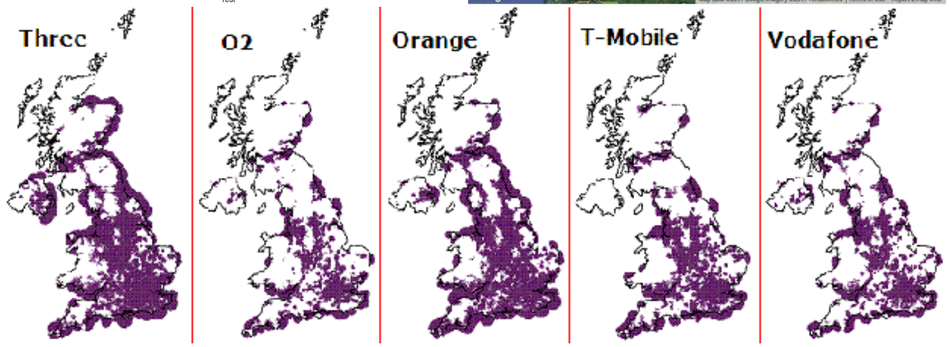
EXAMPLE: ASSISTED LIVING

A safety net for people affected by early-onset dementia.

Responsible & privacy-respecting







- Security threats and vulnerabilities resulting from deliberate introduction of bias.
- Reduction of reliability in changing contexts.
  - Monitoring & detection
  - Alerting & intervention
  - Prediction & prevention

ALGORITHMIC: DIFFERENTIAL PRIVACY

Differential privacy is a definition, not an algorithm.

For a given computational task T and a given value of ε there will be many differentially private algorithms for achieving T in an ε-differentially private manner.

It ensures that any sequence of outputs (responses to queries) is “essentially” equally likely to occur, independent of the presence or absence of any individual. Here, the probabilities are taken over random choices made by the privacy mechanism, and the term “essentially” is captured by a parameter, ε. A smaller ε will yield better privacy (and less accurate responses).

Differential privacy ensures that the same conclusions, e.g., smoking causes cancer, will be reached, independent of whether any individual opts into or opts out of the data set

- Differential privacy addresses the **paradox of learning nothing about an individual while learning useful information about a population**.
- Consider a medical database: smoking causes cancer
- Consider an insurance company: what are a smoker’s long-term medical costs?
  - Has the smoker been harmed by the analysis?
  - Perhaps — their insurance premiums may rise, if the insurer knows they smoke.
  - They may also be helped — learning of their health risks, they might enter a program to quit smoking.
  - **Has the smoker’s privacy been compromised?**
  - It is certainly the case that more is known about them after the study than was known before, but was their information “leaked”?

Data Cannot be Fully Anonymized and Remain Useful. Generally speaking, the richer the data, the more interesting and useful it is.

PRIVACY

AFR / LFR – LEGAL OR NOT?

Automated facial recognition has been found to be in compliance with the Human Rights Act in September 2019 (South Wales Police, High Court in Cardiff)

Use at World Cup in Russia, 2018 and Olympic Games in Pyeongchang, South Korea, 2018.

METROPOLITAN POLICE

Wherever we use it, we'll do so openly. That means we'll:

- tell people online where we're going to use LFR before any deployment
- publish the results of each deployment on the Met website
- provide information leaflets to give to the public
- place posters and signs in and around the area to make people aware the technology is being used
- make officers available to talk to members of the public to help explain what's happening and how LFR works