ETICHIS & PRIVACY

Anna Monreale

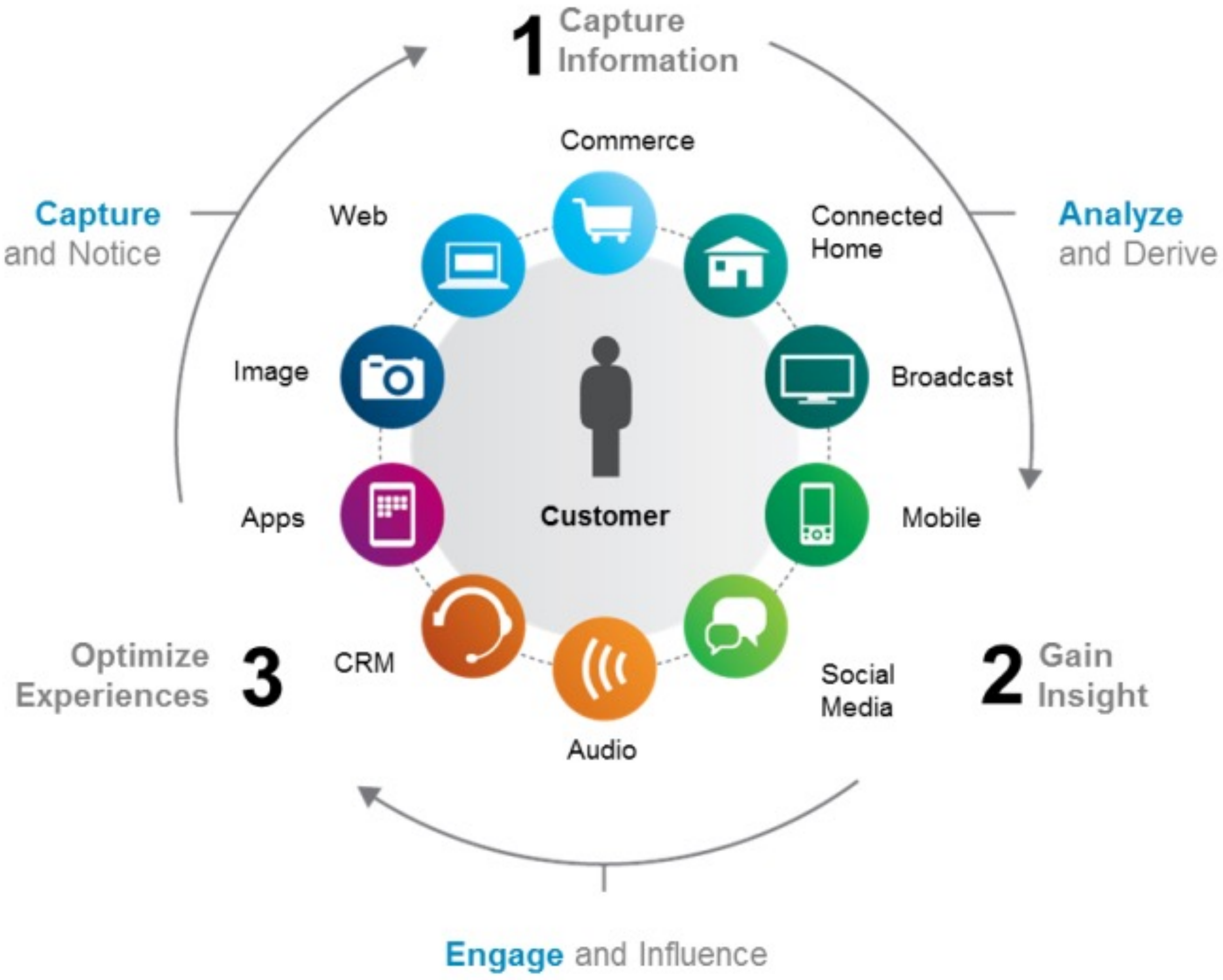
Università di Pisa

Knowledge Discovery and Delivery Lab
(ISTI-CNR & Univ. Pisa)
www-kdd.isti.cnr.it

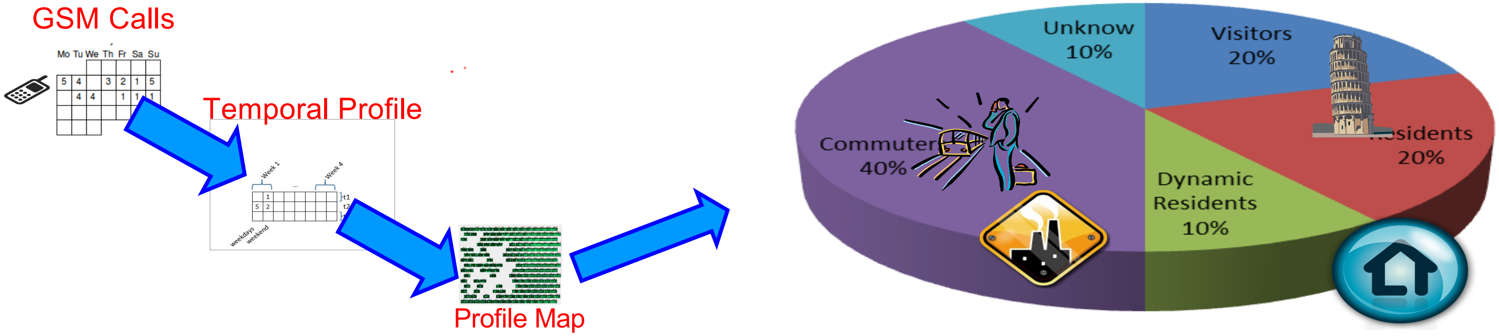
Our digital traces ….

- We produce an unthinkable amount of data while running our daily activities.
- How can we manage all these data? Can we get an added value from them?

Big Data: new, more carefully targeted financial services



A Sociometer based on Mobile Phone Data for Real Time Demographics



GSM Calls

Temporal Profile

Profile Map

ISTITUTO DI SCIENZA E TECNOLOGIE DELL'INFORMAZIONE "A. FAEDO"

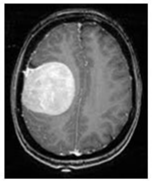
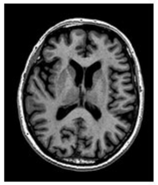
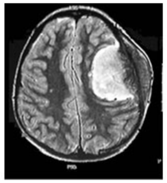
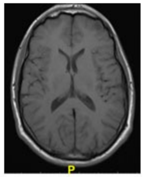
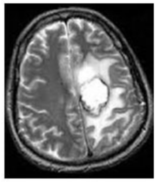
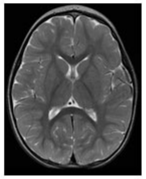
KDD Lab

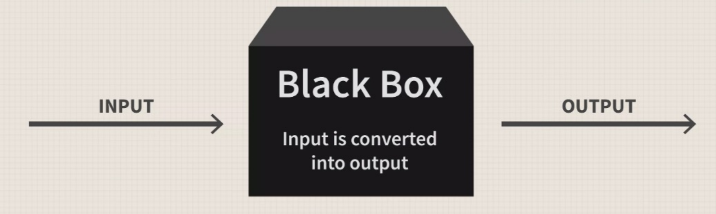
UNIVERSITÀ DI PISA

Istat

WIND Più vicini.

AI in healthcare

Brain Tumor Image	Brain Non Tumor Image
	
	
	

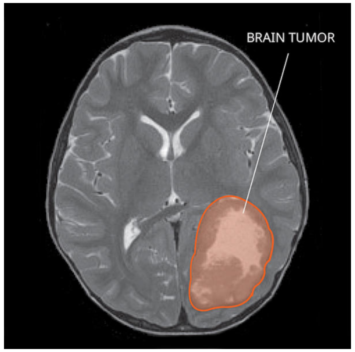


INPUT

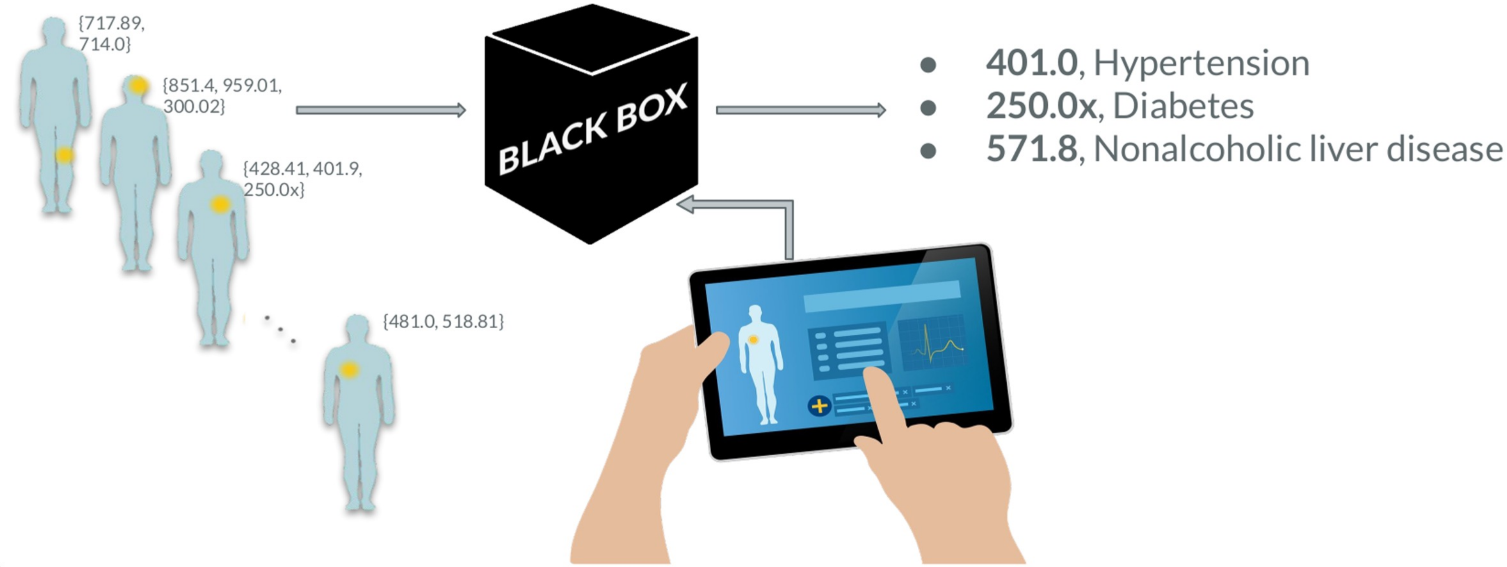
Black Box

Input is converted into output

OUTPUT



AI in healthcare



{717.89, 714.0}

{851.4, 959.01, 300.02}

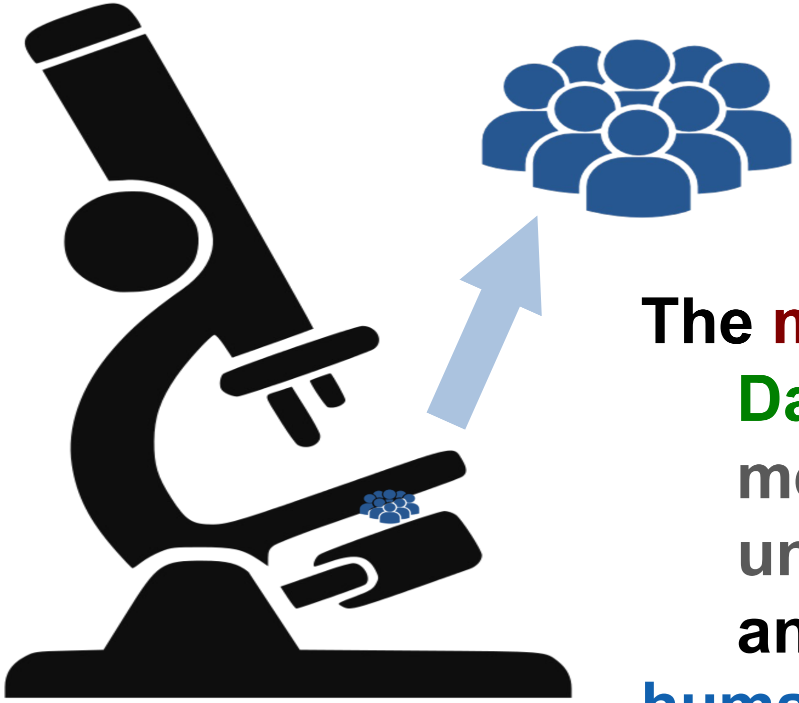
{428.41, 401.9, 250.0x}

{481.0, 518.81}

BLACK BOX

- **401.0**, Hypertension
- **250.0x**, Diabetes
- **571.8**, Nonalcoholic liver disease

AI, Big Data Analytics & Social Mining



The main tool for a Data Scientist to measure, understand, and possibly predict human behavior

Artificial Intelligence: what is it now?

From **encoding** intelligent behavior

↓

To **discovery** and **capture**
intelligent behavior from **data**

Especially (but not only) **personal data**

Artificial Intelligence
=
Collective Intelligence!!

- **Learning from many examples**

- Provide **support for decision making**
  - Enabling nowcasting, what-if simulations based on big data analytics & modeling

Learning from experience

- Data mining & machine learning + big data are the **fulcrum of AI**
- Big data = record the (human) experience
- IoT will facilitate this trend

Data Scientist needs to take into account ethical and legal aspects and social impact of data science & AI

EU Ethics Guidelines for AI – (2019)

Human-centric approach: AI as a means, not an end

Trustworthy AI as our foundational ambition, with three components

Lawful AI

complying with all applicable laws and regulations

Ethical AI

ensuring adherence to ethical principles and values

Robust AI

perform in a **safe, secure** and **reliable** manner, both form technical and a social perspective, with safeguards to foresee and prevent unintentional harm

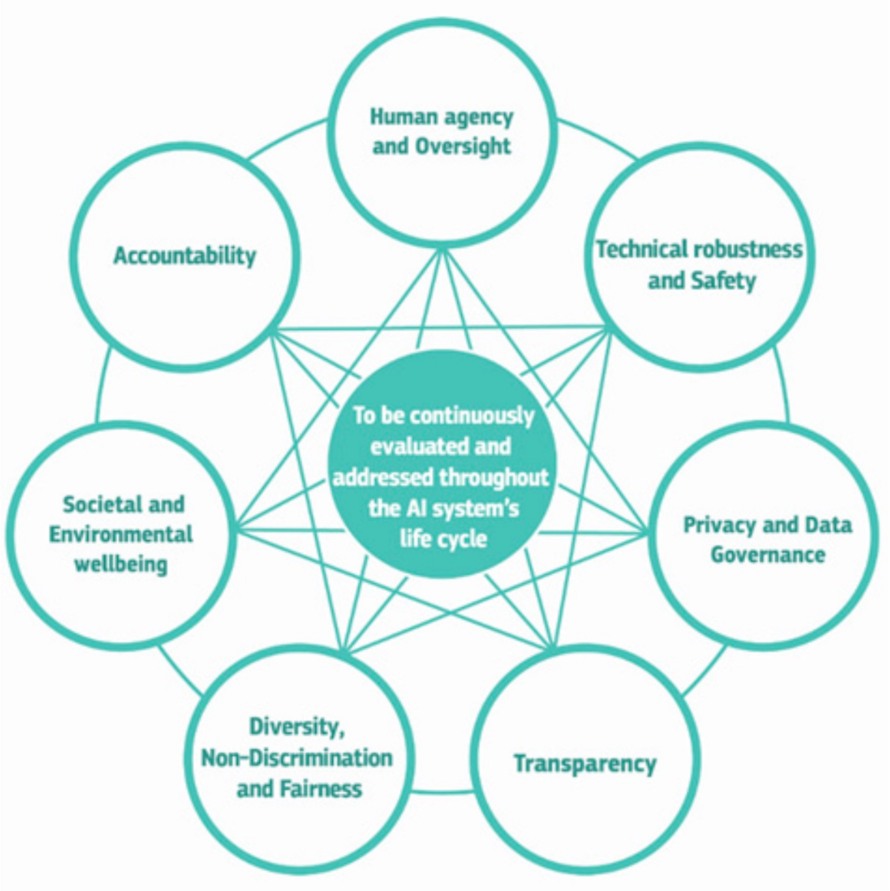
Requirements

1. **Human agency and oversight**
   - Fundamental rights
   - Human agency
   - Human oversight

2. **Technical robustness**
   - Resilience to attack and security
   - Safety
   - Accuracy
   - Reliability and reproducibility

3. **Privacy and data governance**
   - Privacy and data protection
   - Quality and integrity of data
   - Access to data

4. **Transparency**
   - Traceability
   - Explainability


Human agency and Oversight
Technical robustness and Safety
Privacy and Data Governance
Transparency
Diversity, Non-Discrimination and Fairness
Societal and Environmental wellbeing
Accountability
To be continuously evaluated and addressed throughout the AI system's life cycle

Requirements

5. Diversity, non-discrimination and fairness

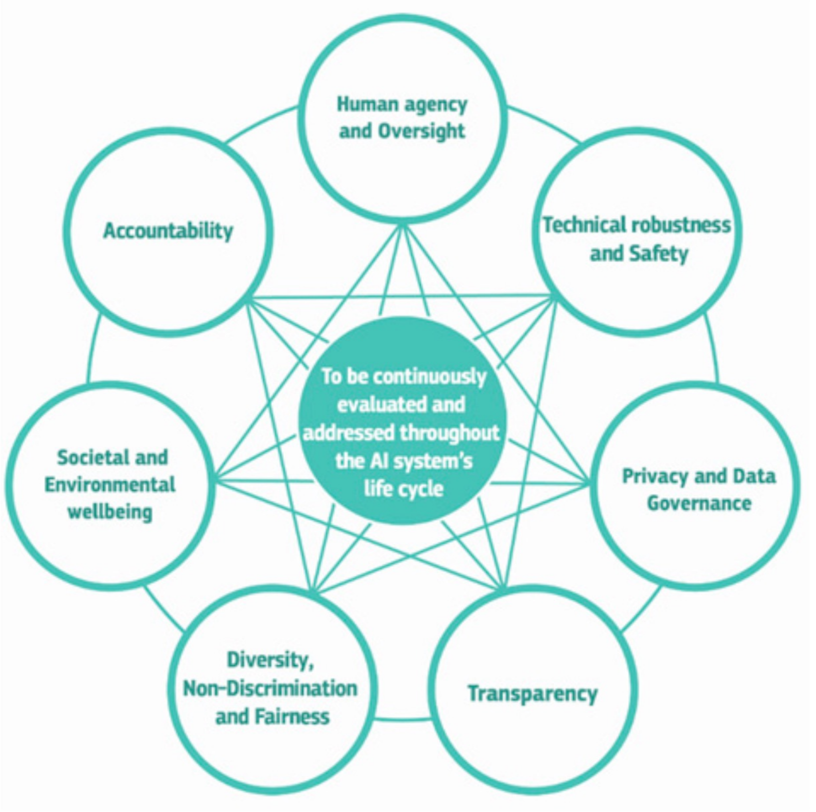
- Avoidance of unfair bias
- Accessibility and universal design
- Stakeholder Participation

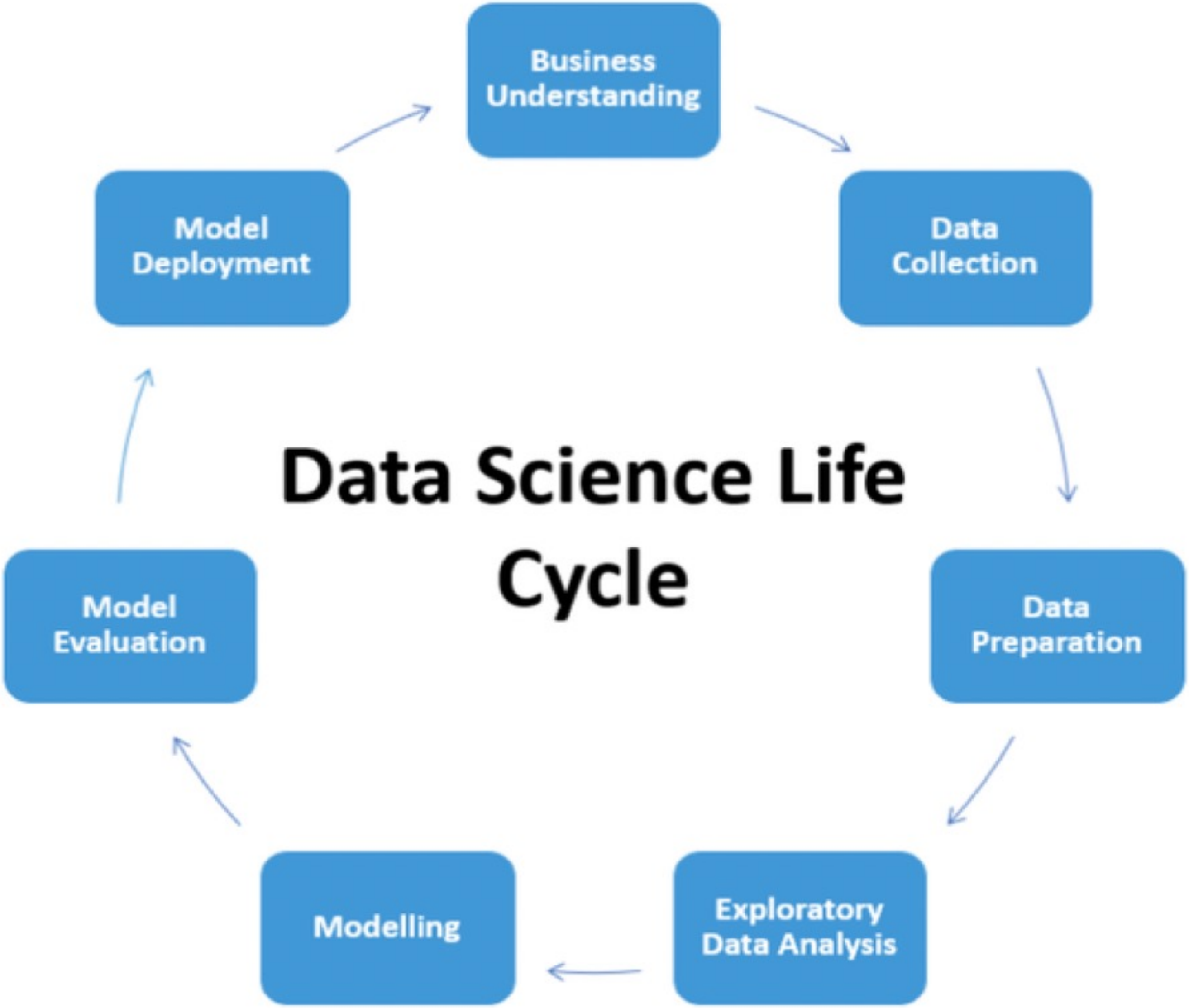
6. Societal and environmental well-being

- Sustainable and environmentally friendly AI
- Social impact
- Society and Democracy

7. Accountability

- Minimisation and reporting of negative impacts
- Auditability
- Trade-offs


Human agency and Oversight
Technical robustness and Safety
Privacy and Data Governance
Transparency
Diversity, Non-Discrimination and Fairness
Societal and Environmental wellbeing
Accountability
To be continuously evaluated and addressed throughout the AI system's life cycle



Data Science Life Cycle

Business Understanding

Data Collection

Data Preparation

Exploratory Data Analysis

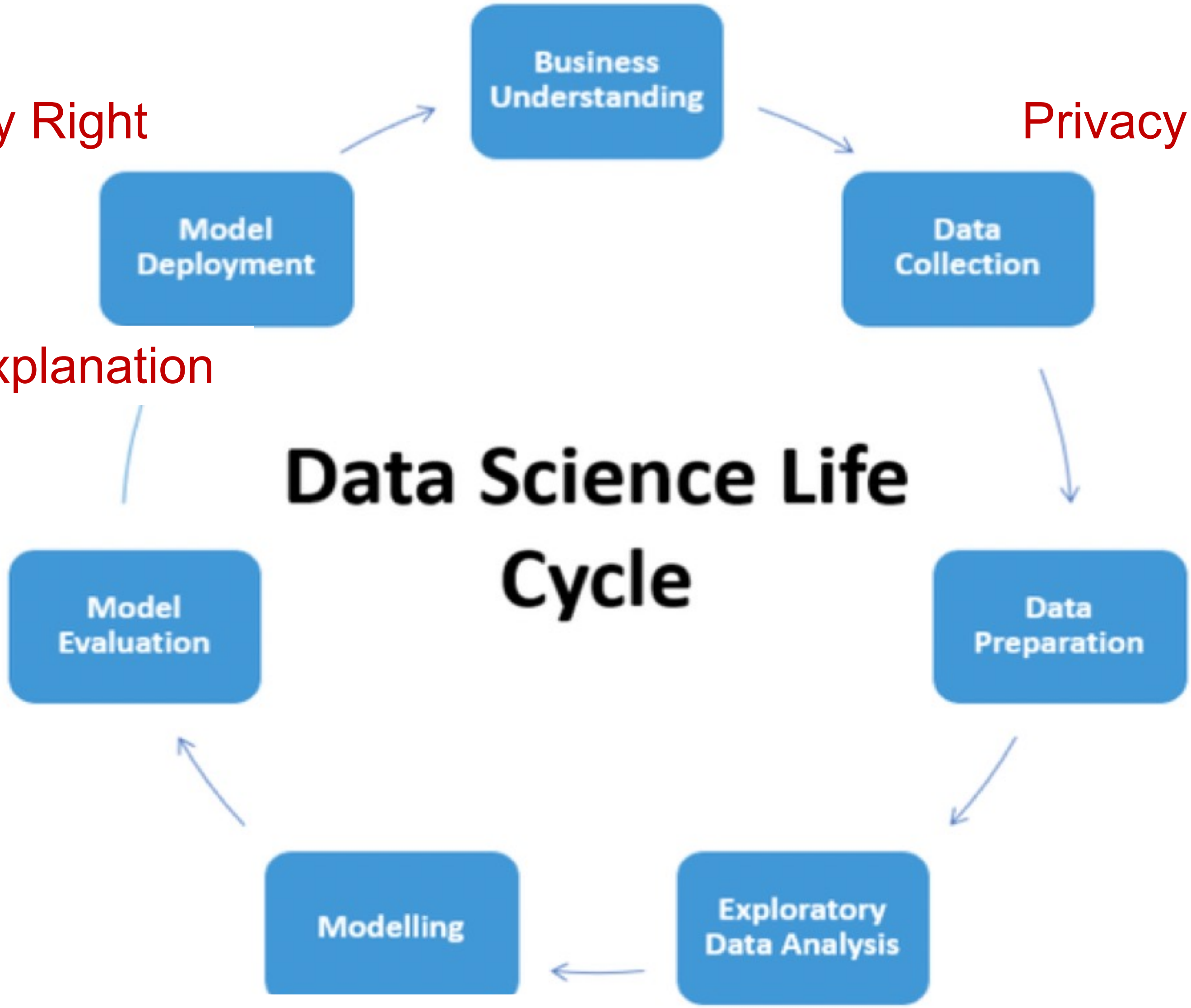
Modelling

Model Evaluation

Model Deployment

Privacy Right

Right of Explanation



Data Science Life Cycle

Business Understanding

Data Collection

Data Preparation

Exploratory Data Analysis

Modelling

Model Evaluation

Model Deployment

Privacy Right

Privacy Right

Right of Explanation

Privacy Right

Right of Explanation

PRIVACY & DATA PROTECTION

EU Legislation for protection of personal data

- European directives:
  - Data protection directive (95/46/EC)
  - ePrivacy directive (2002/58/EC) and its revision (2009/136/EC)
  - General Data Protection Regulation (May 2018)

http://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=IT

EU: Personal Data

- **Personal data** is defined as any information relating to an identity or **identifiable** natural person.

- An **identifiable person** is one who can be identified, **directly or indirectly**, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity.

Personal Data

- Your name
- Home address
- Photo
- Email address
- Bank details
- Posts on social networking websites
- Medical information,
- Computer or mobile IP address
- Mobility traces
- ………

Sensitive Data

- Sensitive personal data is a specific set of “**special categories**” that must be treated with extra security
  - Racial or ethnic origin
  - Political opinions
  - Religious or philosophical beliefs
  - Trade union membership
  - Genetic data
  - Biometric data

EU Directive (95/46/EC) and GDPR

- **GOALS**:
  - protection protection of individuals with regard to the **processing** of personal data
  - the free movement of such data
  - User control on personal data
- The term “process” covers anything that is done to or with personal data:
  - collecting
  - recording
  - organizing, structuring, storing
  - adapting, altering, retrieving, consulting, using
  - disclosing by transmission, disseminating or making available, aligning or combining, restricting, erasing, or destroying data.

Anonymity according to 1995/46/EC

- The principles of protection must apply to any information concerning an identified or identifiable person;
- To determine whether a person is identifiable, account should be taken of **all the means likely reasonably to be used** either by the controller or by any other person to identify the said person
- **The principles of protection shall not apply to data rendered anonymous in such a way that the data subject is no longer identifiable**

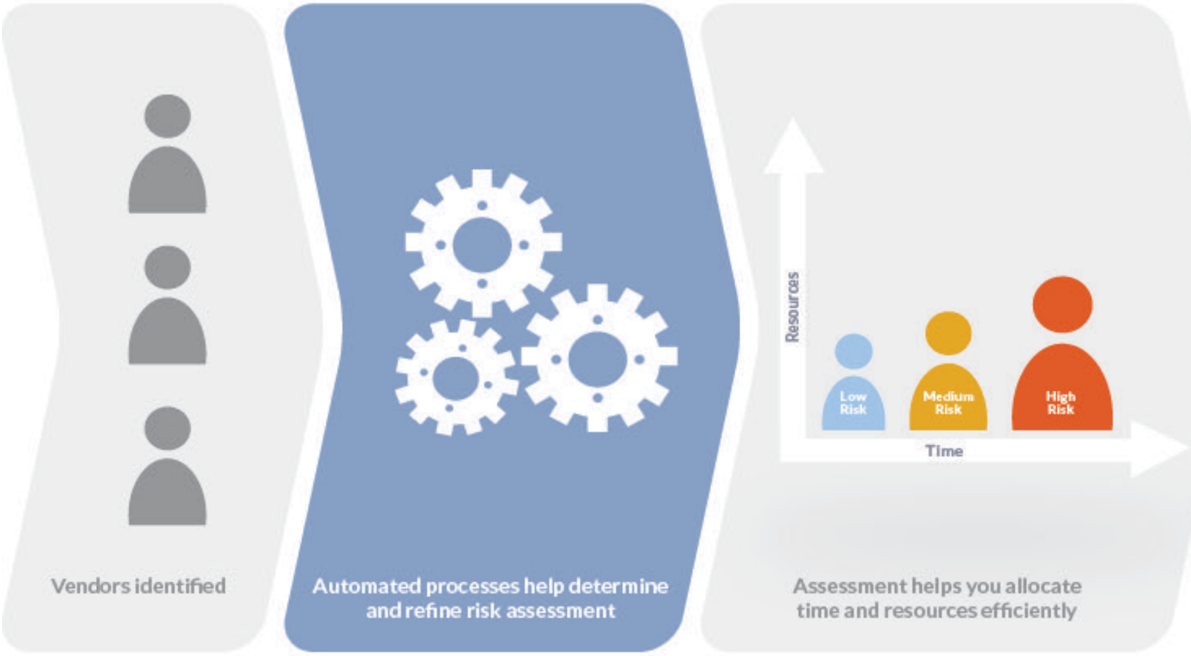
Privacy by Design Principle

- **Privacy by design** is an approach to protect privacy by inscribing it into the **design specifications** of information technologies, accountable business practices, and networked infrastructures, from the very start

- Developed by Ontario's Information and Privacy Commissioner, Dr. Ann Cavoukian, in the 1990s
  - as a response to the growing threats to online privacy that were beginning to emerge at that time.

Privacy Risk Assessment

- GDPR requires that data controllers maintain an updated report on the privacy risk assessment on perosnal data collected



Vendors identified

Automated processes help determine and refine risk assessment

Assessment helps you allocate time and resources efficiently

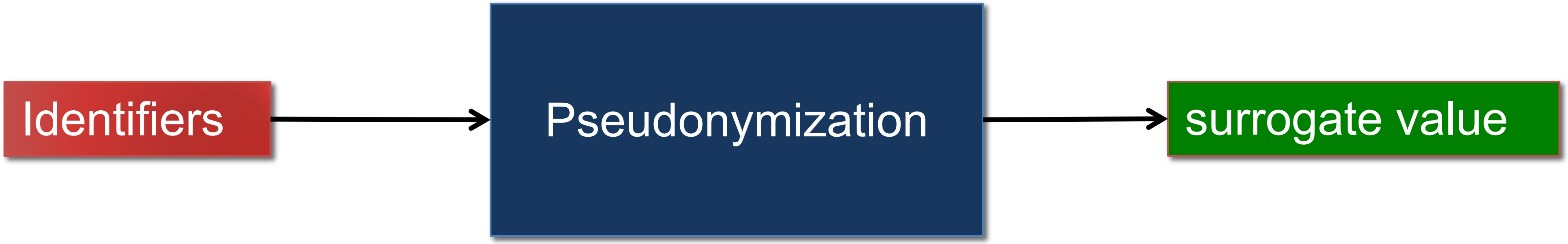
PSEUDONYMIZATION & ANONYMIZATION

Anonymization vs Pseudonimization

- Pseudonymization and Anonymization are two distinct terms often confused
- Anonymized data and pseudonymized data fall under very different categories in the regulation
- **Anonymization guarantees data protection** against the (direct and indirect) data subject re-identification
- **Pseudonymization substitutes the identity** of the data subject in such a way that additional information is required to re-identify the data subject

Pseudonymization

Substitute an **identifier** with a surrogate value called **token**



Substitute unique names, fiscal code or any attribute that identifies uniquely individuals in the data

Example of Pseudonymization

Name	Gender	DoB	ZIP Code	Diagnosis
Anna Verdi	F	1962	300122	Cancro
Luisa Rossi	F	1960	300133	Gastrite
Giorgio Giallo	M	1950	300111	Infarto
Luca Nero	M	1955	300112	Emicrania
Elisa Bianchi	F	1965	300200	Lussazione
Enrico Rosa	M	1953	300115	Frattura

ID	Gender	DoB	ZIP CODE	DIAGNOSIS
11779	F	1962	300122	Cancro
12121	F	1960	300133	Gastrite
21177	M	1950	300111	Infarto
41898	M	1955	300112	Emicrania
56789	F	1965	300200	Lussazione
65656	M	1953	300115	Frattura

Properties of a Surrogate Value

- Irreversible without private information
- Distinguishable from the original value

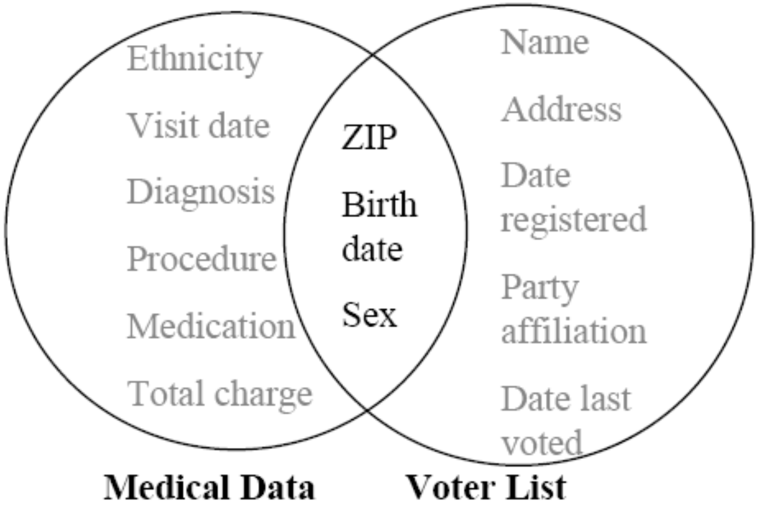
Is Pseudonymization enough for data protection?

Pseudonymized data are still Personal Data!!

Massachussetts' Governor

- Sweeney managed to re-identify the medical record of the governor of Massachussetts
  - MA collects and publishes sanitized medical data for state employees (microdata) left circle
  - voter registration list of MA (publicly available data) right circle

- looking for governor's record
- join the tables:
  - **6 people had his birth date**
  - **3 were men**
  - **1 in his zipcode**



Latanya Sweeney: k-Anonymity: A Model for Protecting Privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10(5): 557-570 (2002)

Linking Attack

Governor: **birth date = 1950,** CAP = **300111**

ID	Gender	YoB	ZIP	DIAGNOSIS
1	F	1962	300122	Cancer
2	F	1960	300133	Gastritis
3	M	1950	300111	Heart Attack
4	M	1955	300112	Headache
5	F	1965	300200	Dislocation
6	M	1953	300115	Fracture

Which is the disease of the Governor?

Making data anonymous

K-anonymity

Governor: Birth Date = **1950,** CAP = **300111**

ID	Gender	YoB	ZIP	DIAGNOSIS
1	F	[1960-1956]	300***	Cancer
2	F	[1960-1956]	300***	Gastritis
3	M	[1950-1955]	30011*	Heart Attack
4	M	[1950-1955]	30011*	Headache
5	F	[1960-1956]	300***	Dislocation
6	M	[1950-1955]	30011*	Fracture

Which is the disease of the Governor?