Challenges and Ethical Concerns

Challenges

- **Data diversity** - we can mine pretty much every single data type, but what if every data type is combined in the same problem?
- **Domain Knowledge** - Few are the techniques able to incorporate available domain knowledge in the process - feature generation is probably the new path to achieve this, but it is still a challenge;
- **Automation** - AutoML tools aim to automate the whole kdd process, but they are still far from being able to do so;
- **Engineering Principles** - Data Science is still a new field, and it is still lacking engineering principles, such as **standard methodologies**.

Ethical Concerns

Given the increasing extension of data science applications, it is important to consider the ethical implications of the data science process, and **how it can affect the society**, particularly how it can affect the **citizens' rights**.

- **Our opinion should not matter** because there are a **set of laws** established to defend the citizens' rights - **EU GDPR (EU General Data Protection Regulation)** is an example of such laws; The GDPR recognizes **two central entities**:
  - The **data subject** - a person, whose data is being collected and processed when/if he gives his **consent**;
  - The **data controller** - the entity that determines the **purposes and means of the processing of personal data**;

Personal Data

Personal data is *any information relating to the data subject, which allows for his direct or indirect identification, in particular, any identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person* - Art. 4 of the GDPR.

Data Processing

Data processing is *any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organization, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction* - Art. 4 of the GDPR.

Data Protection Principles

Art. 5 of the GDPR establishes the following principles:

- Data processing must be **lawful, fair and transparent**, **limited to its purpose**, minimal, **adequate and necessary** to the purpose defined;
- Data must be **accurate** and **kept up to date**, and **kept for no longer than necessary**;
- Data **integrity and confidentiality** must be ensured;

Lawfulness of Processing

Art. 6 of the GDPR establishes the following conditions for the lawfulness of processing:

- **Consent** - the data subject has given consent to the processing of his personal data for one or more specific purposes;
- **Contract** - processing is necessary for the performance of a contract to which the data subject is party or in order to take steps at the request of the data subject prior to entering into a contract;
- **Legal Obligation** - processing is necessary for compliance with a legal obligation to which the controller is subject;
- **Vital Interests** - processing is necessary in order to protect the vital interests of the data subject or of another natural person;
- **Public Interest** - processing is necessary for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller;

- **Legitimate Interests -** processing is necessary for the purposes of the legitimate interests pursued by the controller or by a third party, except where such interests are overridden by the interests or fundamental rights and freedoms of the data subject which require protection of personal data, in particular where the data subject is a child.

Prohibition

Art. 9 of the GDPR establishes that is prohibited the processing of personal data revealing:

- **Racial/ethnic origin**;
- **Political opinions**;
- **Religious/philosophical beliefs**;
- **Trade union membership**;
- **Genetic data**;
- **Biometric data**.

Simple Rules for Data Science

1. Acknowledge that data is people and can be harmful;
2. Recognize that privacy is more than a binary variable;
3. Guard against the re-identification of your data;
4. Practice ethical sharing;
5. Consider the strengths and limitation of your data. Big does not automatically mean better ;) ;
6. Debate the tough ethical choices;
7. Develop a code of conduct for your organization, research community, or industry;
8. Design your data and systems for auditability;
9. Engage with the broader consequences of data and analysis practices;
10. Know when to break these rules.

    Transfer Learning - the process of transferring knowledge from one domain to another, and it is a way to incorporate domain knowledge in the process.

    Data Leakage - when information from outside the training dataset is used to create the model, and it is a way to incorporate domain knowledge in the process.