Intro

Data Mining

Edgar (Paco) Roman-Rangel.
edgar.roman@itam.mx

Department of Computer Science.
Instituto Tecnológico Autónomo de México, ITAM.

Outline

Definition

Data mining:
a.k.a. Knowledge discovery from data (KDD).

More definitions

What is Data Mining?

- Process by which **information** and patterns of interest are discovered in (big) **data**.
- Those patterns often help the decision making process.
- Between the intersection of AI and statistics.
- Challenges: scalability, high dimensionality, heterogeneity, secrecy and intellectual property, distribution limitations, noise, etc.
- Tasks: predictive (classification and regression), descriptive (correlation, segmentation, anomaly detection).

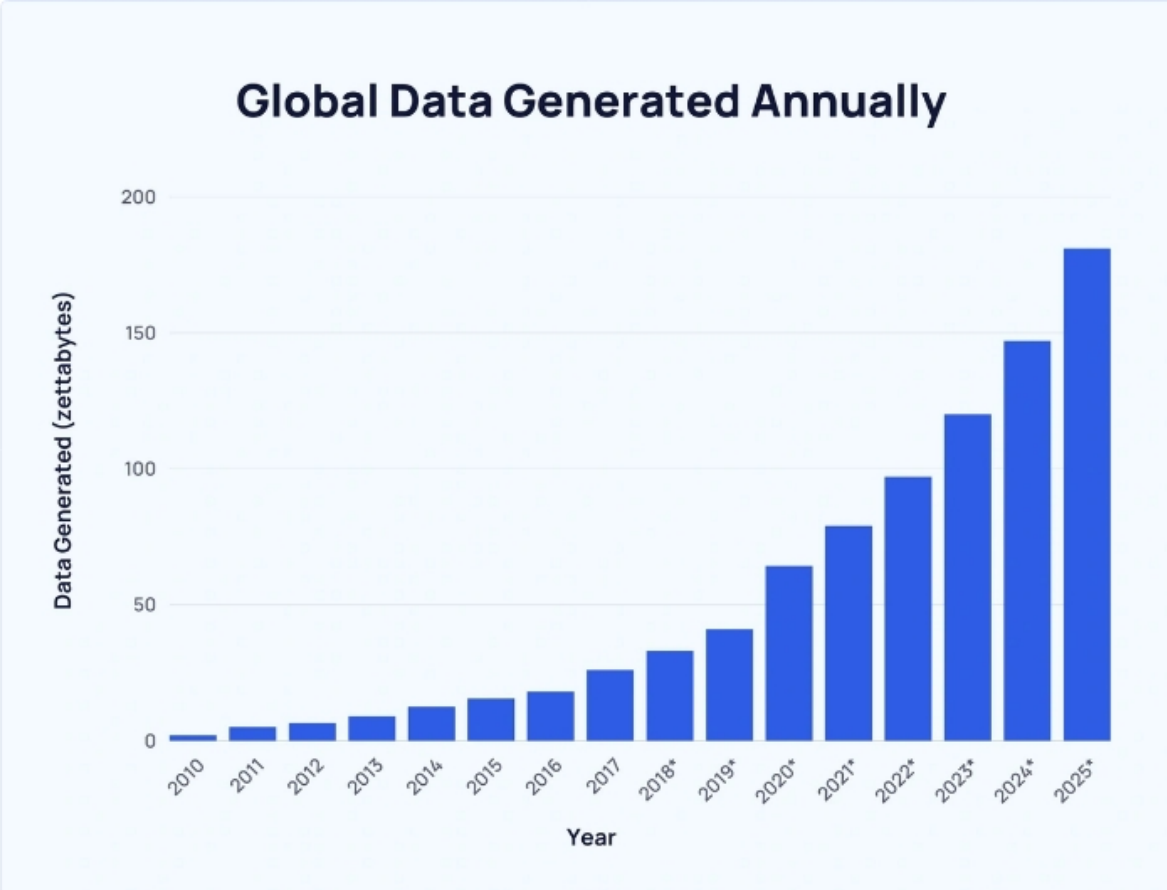
Data Mining and information

We often say *we live in the age of information*,

Rather, we actually live in the "age of data".

Amount of data

According to the latest estimates, **402.74 million terabytes** of data are created each day.



Data miners

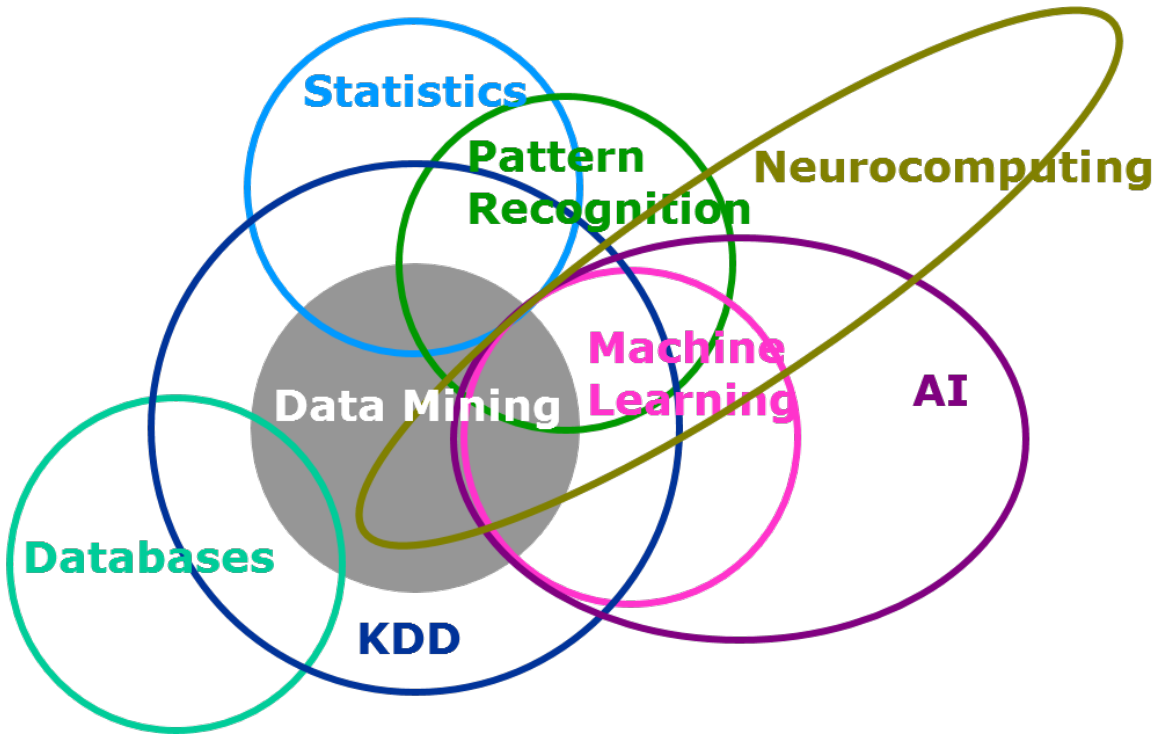
Our task: To turn large collections of data into knowledge.

Data types that can be mined

- Tabular data: think of a table, where each row is a sample and each column is a variable.
- Images: where each data point is formed by a set of variables arranged in matrix form. Each sample is a matrix.
- Time series: each sample is a series of time-dependent variables. Could be uni- or multi-variate. It can also be real- or integer-valued.
- Transactional: where each data point is a transaction, listing a set of co-occurring items, e.g., supermarket tickets.
- Graphs: where the database describe relationships between individuals. The structure consists of vertices and edges.

Data Mining vs Machine Learning

Data Mining is not quite the same as Machine Learning.



Related areas

- **Data Science:** broad area, dealing with processing and understanding data for decision making.
- **Data Mining:** a part of the data science process, in charge of extracting relevant patterns (information) from data.
- **Machine Learning:** set of algorithms that learn to solve problems based on experiences.
- **Deep Learning:** sub area of machine learning based on the use of deep neural networks.

Data Mining vs Machine Learning

Coming back to Data Mining.

Data Mining	**Machine Learning**
Extract information from data	Learn relation between data
Discover patterns	Learn models for data
Born in the 30's	Born in the 50's
Focused on data	Focused on algorithms

Data, information, and knowledge

- Data: just a set of records. No meaning associated to them.
- Information: filtered data. Identification of relevant patterns.
- Knowledge: analysis of information. Information plus context. Allows to make informed decisions.
- Wisdom: mastering of the decision making process.

We say that "data might contain information".

About this course

You can think of this course as a complement of machine learning, were we will focus on mining data (squeeze, massage, torture) to get information that can feed ML models.

We will also work with Deep Learning models, comparing tow forms of data representation: feature engineered vs automatically learned.

Q&A

Thank you!

edgar.roman@itam.mx