Data Science & Machine Learning Lab

Introduction to Data Science & Machine Learning

Flavio Giobergia

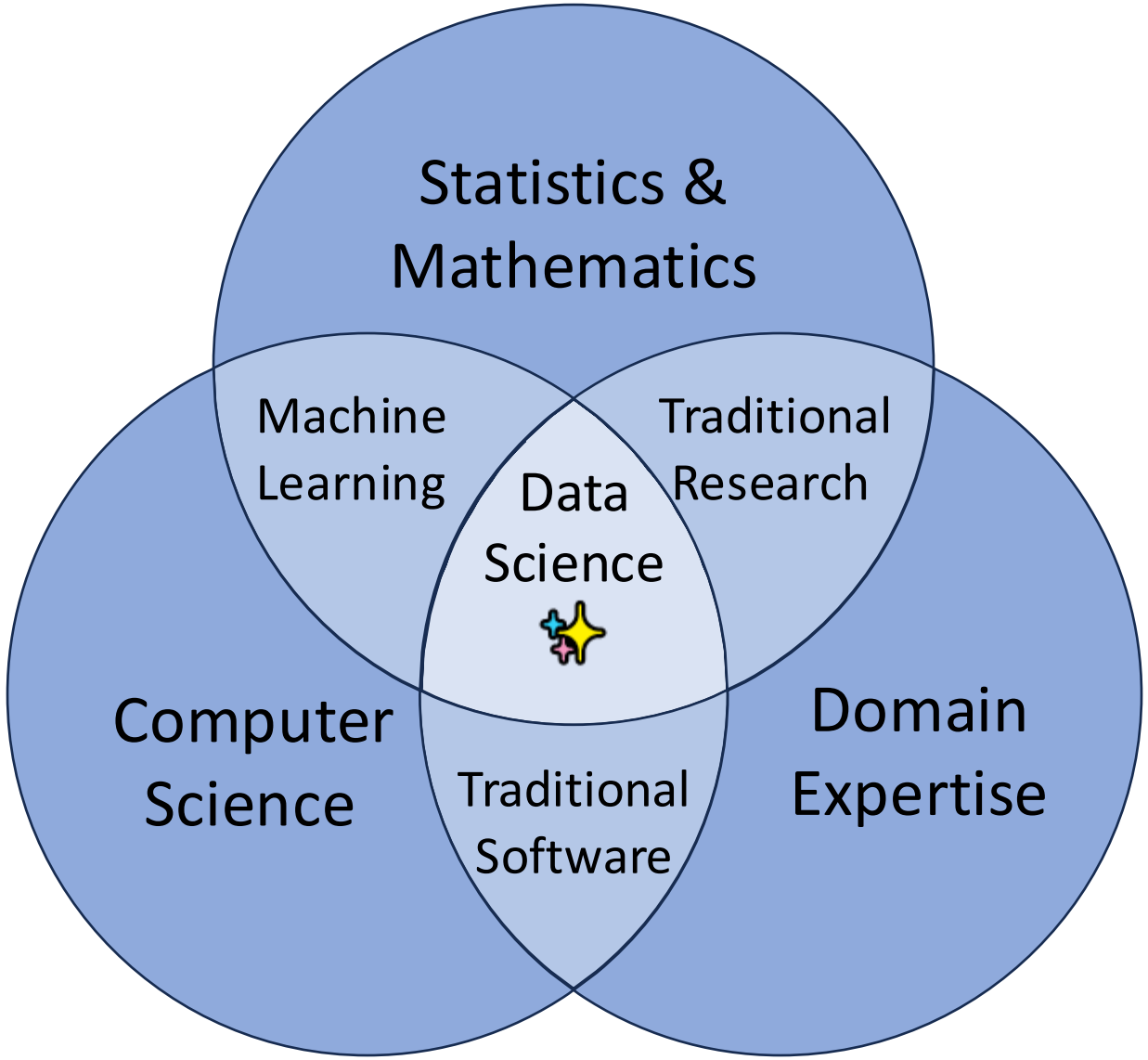
"The future of data analysis"

John W. Tukey, 1962

Princeton University and Bell Telephone Laboratories

All in all, I have come to feel that my central interest is in *data analysis*, which I take to include, among other things: procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data.

What is data science?



"Machine Learning"

Tom M. Mitchell, 1997

Definition: A computer program is said to **learn** from experience *E* with respect to some class of tasks *T* and performance measure *P*, if its performance at tasks in *T*, as measured by *P*, improves with experience *E*.

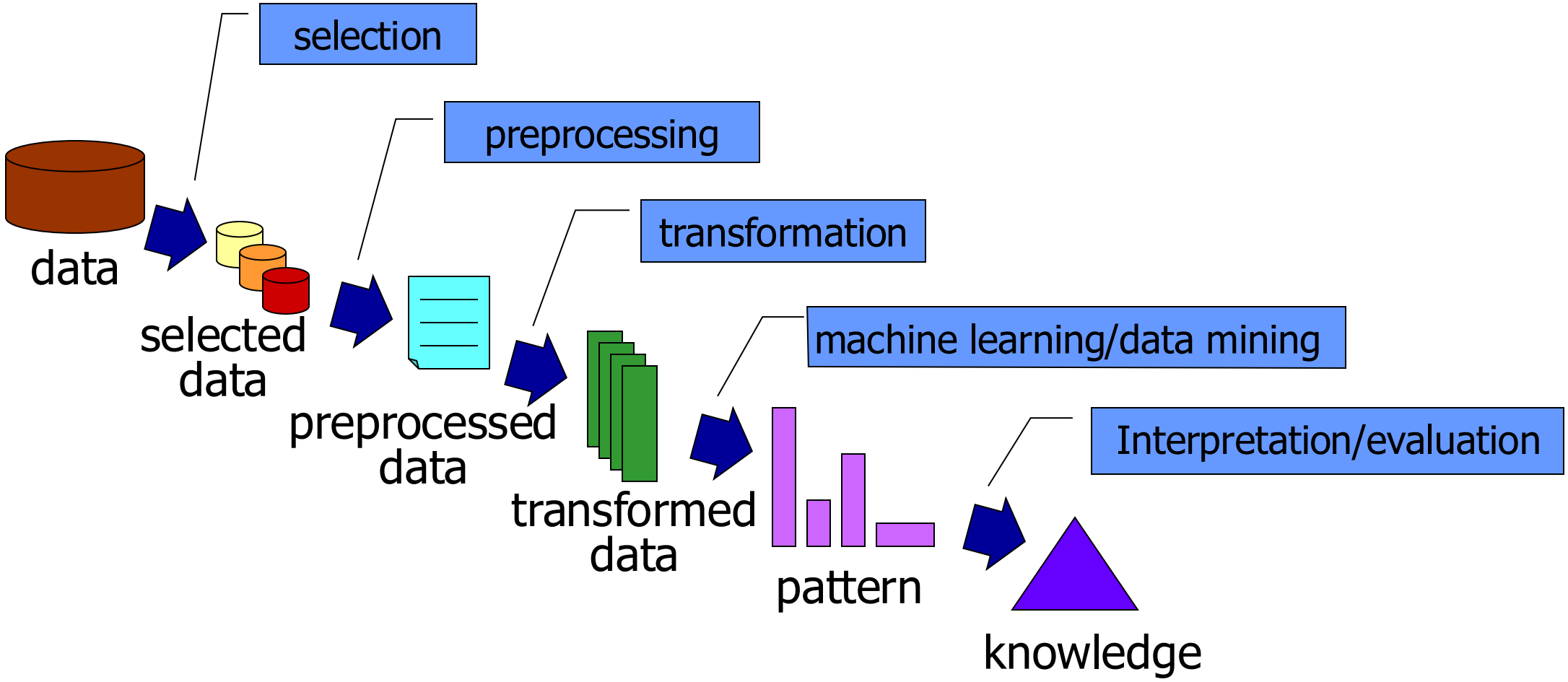
A *neural network* is said to learn from a *dataset of images* to *recognize pictures of dogs*, if its *accuracy* at predicting dogs improves after seeing the images.

Data mining

- Non trivial extraction of
  - implicit
  - previously unknown
  - potentially useful

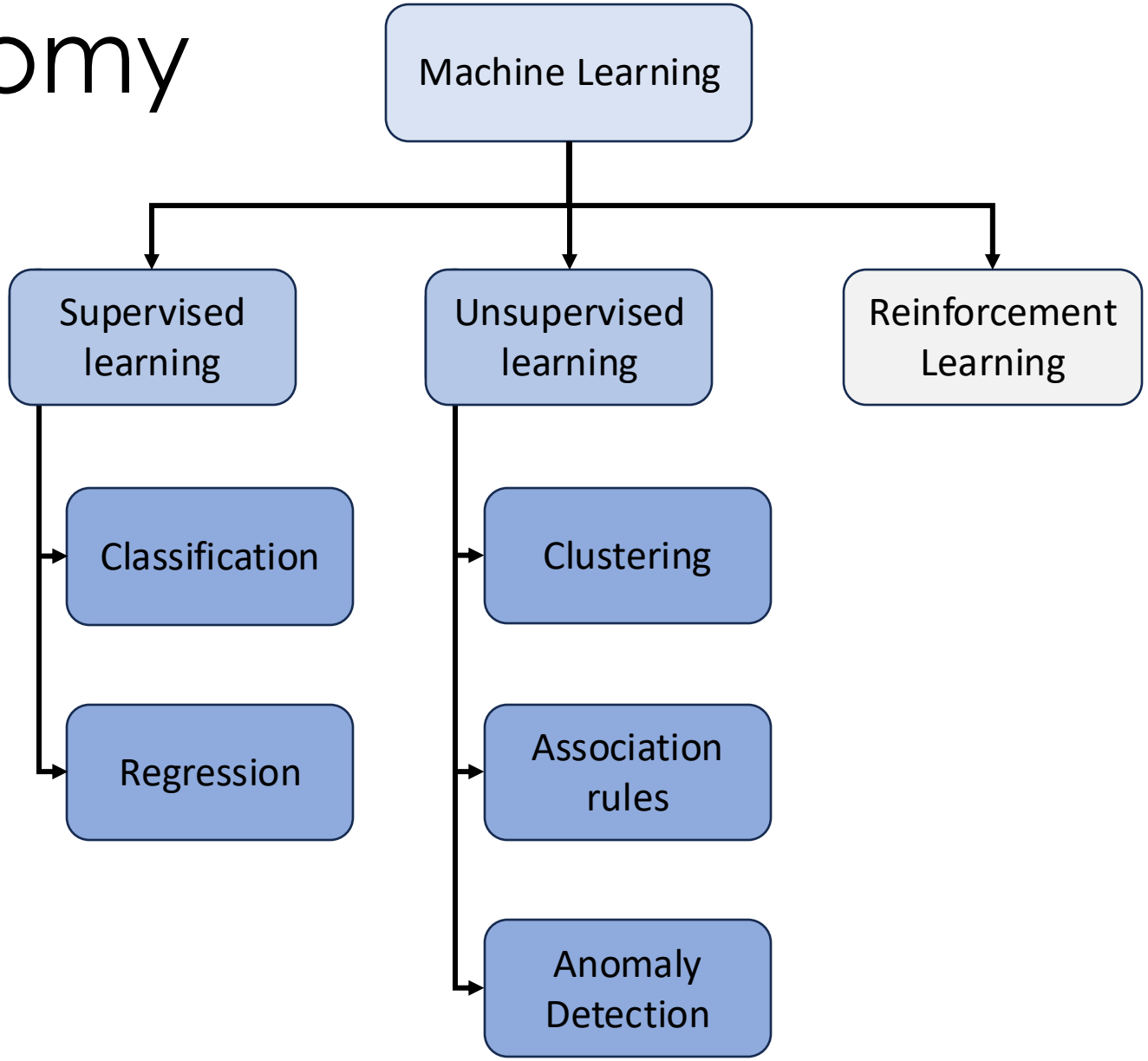
  information from available data
- Extraction is automatic
  - performed by appropriate algorithms
- Extracted information is represented by means of abstract models
  - denoted as *pattern*

Knowledge Discovery in Databases



A simplified taxonomy

- **Supervised Learning**
  - Inputs + targets → predict outputs
  - Examples: regression, classification
- **Unsupervised Learning**
  - No labels → find structure in data
  - Examples: clustering, dimensionality reduction
- **Reinforcement Learning**
  - Agent interacts with environment, learns by reward/punishment



Challenges in ML

- Data quality
  - Data will be missing, noisy, biased
- Scalability
  - big datasets, high dimensionality
- Interpretability
  - How do we understand black-box models?
- Production-level problems
  - Handle drifts, unlearn/forget data

Ethics & Responsible AI

- **Bias in data → biased predictions**
- **Fairness**: avoid discrimination
- **Privacy**: sensitive personal data
- **Transparency**: explainability, accountability
- **Impact**: decisions affecting lives (health, justice, finance)