Statistical Machine Learning (BE4M33SSU)
Lecture 1.

Czech Technical University in Prague

Course format

Instructors: Jan Drchal, Vojtech Franc and Jakub Paplhám

Structure: 1 lecture & 1 seminar per week (6 credits $\approx$ 150-180 hours)

Seminars: Solving theoretical assignments. Explaining and discussing homeworks.

Homeworks:

- Submit a Python code solving the assignment. Automatic evaluation.

Challenge:

- Solve an open problem using a real-world data.

Assesment:

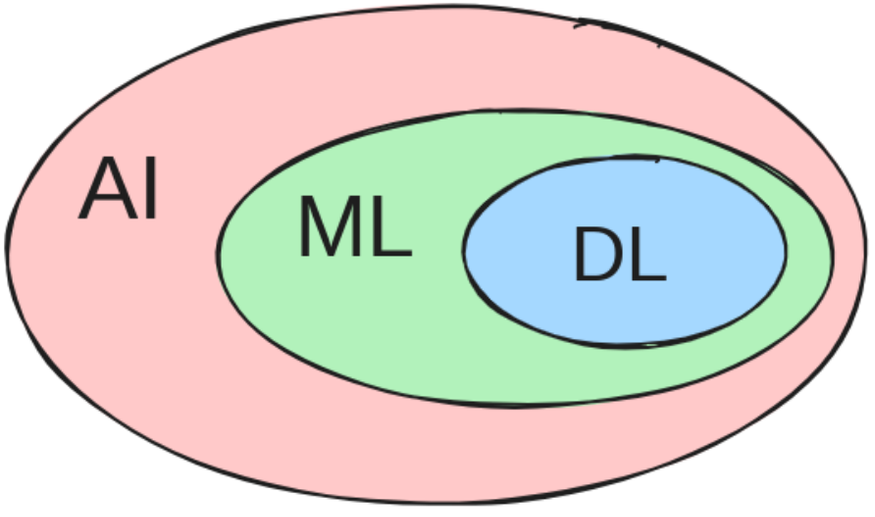
- Passing requires at least 50% of points in both seminars and exam.
- Final grade = 40% homework + 60% written exam (+ bonus points possible)

Prerequisites:

- Probability Theory and Statistics (A0B01PSI)
- Pattern Recognition and Machine Learning (AE4B33RPZ)
- Optimization (AE4B33OPT)

More details: https://cw.fel.cvut.cz/wiki/courses/be4m33ssu/start

What is AI, ML and DL?

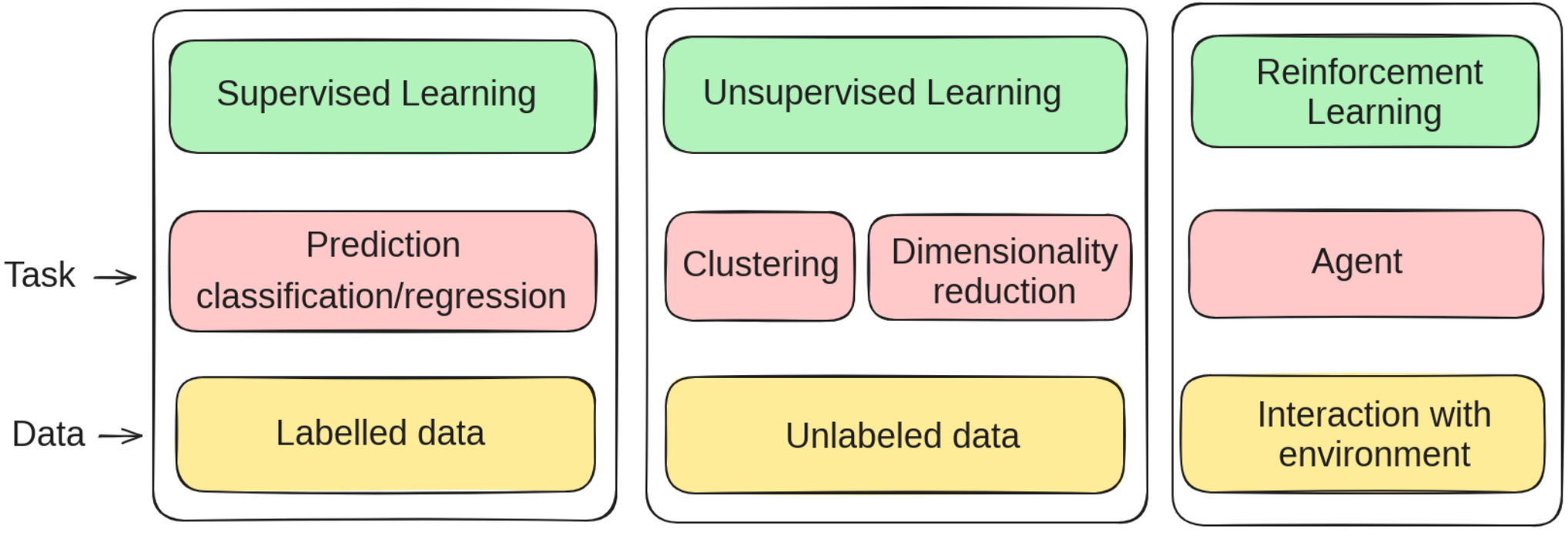


Artificial Intelligence: Techniques for creating computer systems that replicate aspects of human intelligence.

Machine Learning: A subset of AI that uses statistical algorithms to learn from data and perform tasks without being explicitly programmed.

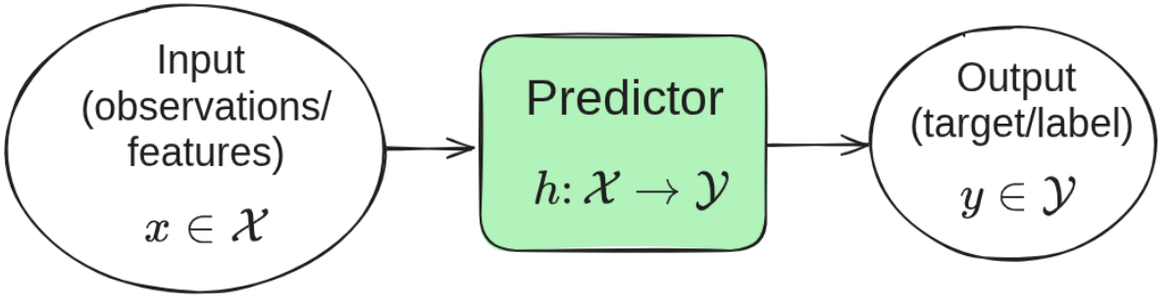
Deep Learning: An advanced branch of machine learning that leverages large-scale neural networks to process vast data and uncover intricate patterns.

Branches of Machine Learning



This course focuses primarily on Supervised Learning, with additional coverage of Unsupervised Learning.

Prediction Problems



Examples of prediction problems:

Classification when $\mathcal{Y} = \{1, \ldots, Y\}$

- spam detection
- credit scoring
- malware detection
- image classification
- speech classification
- text classification
- . . .

Regression when $\mathcal{Y} = \mathbb{R}$

- house price prediction
- stock price prediction
- energy consumption
- sales forecasting
- age prediction
- . . .

Course Goals and Topics

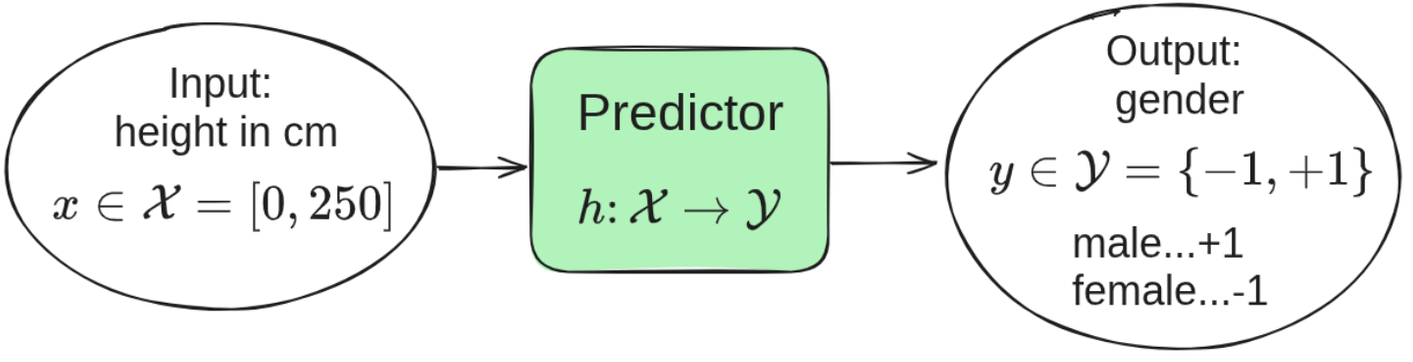
Main Objectives:

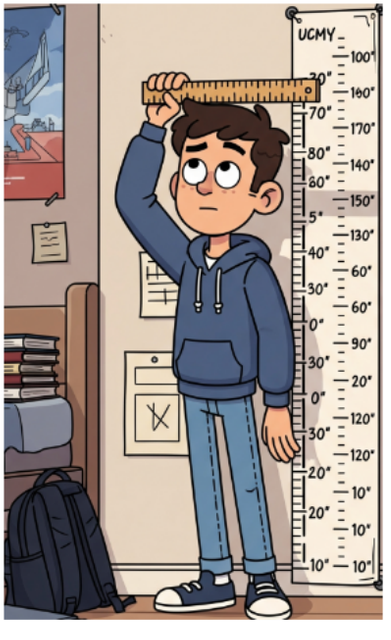
1. Introduce fundamental concepts and their theoretical foundations.
2. Present key models for classification and regression, and demonstrate how they can be learned using these concepts.

Topics Covered:

- Empirical risk minimization
- Probably Approximately Correct learning
- VC-dimensions
- Supervised learning of deep networks
- Stochastic Gradient Descend
- Deep Convolutional Neural Networks
- Support Vector Machines
- Ensembling
- Generative Learning
- Bayesian Learning
- Expectation-Maximization algorithm
- Hidden Markov Models

Example: Gender prediction





Data

x	y
179.5	+1
161.5	+1
176.9	+1
157.7	-1
184.0	+1
167.4	+1
175.1	+1

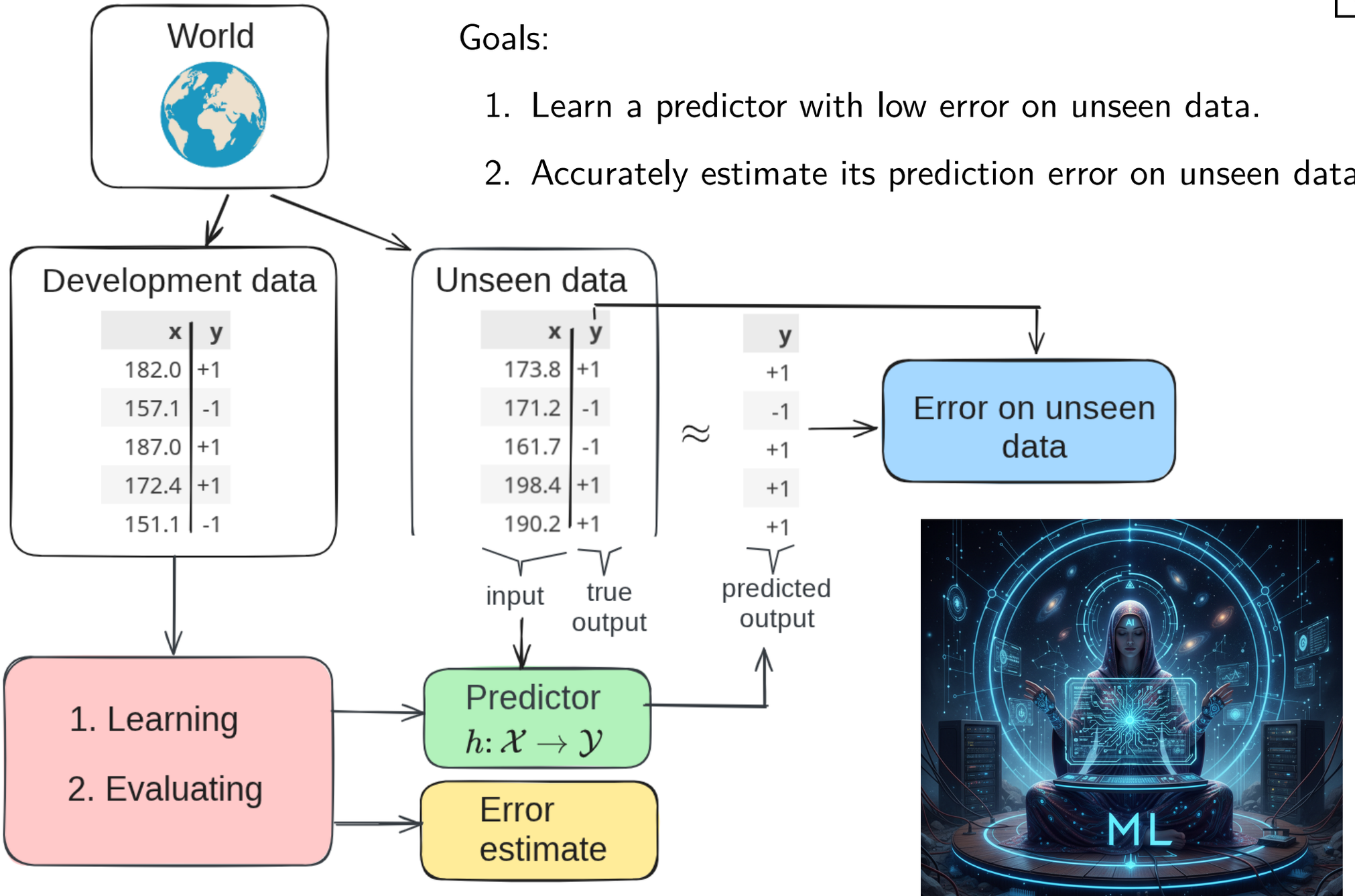
height [cm]

gender
male... +1
female... -1

Predictor Learning: Setup and Goals

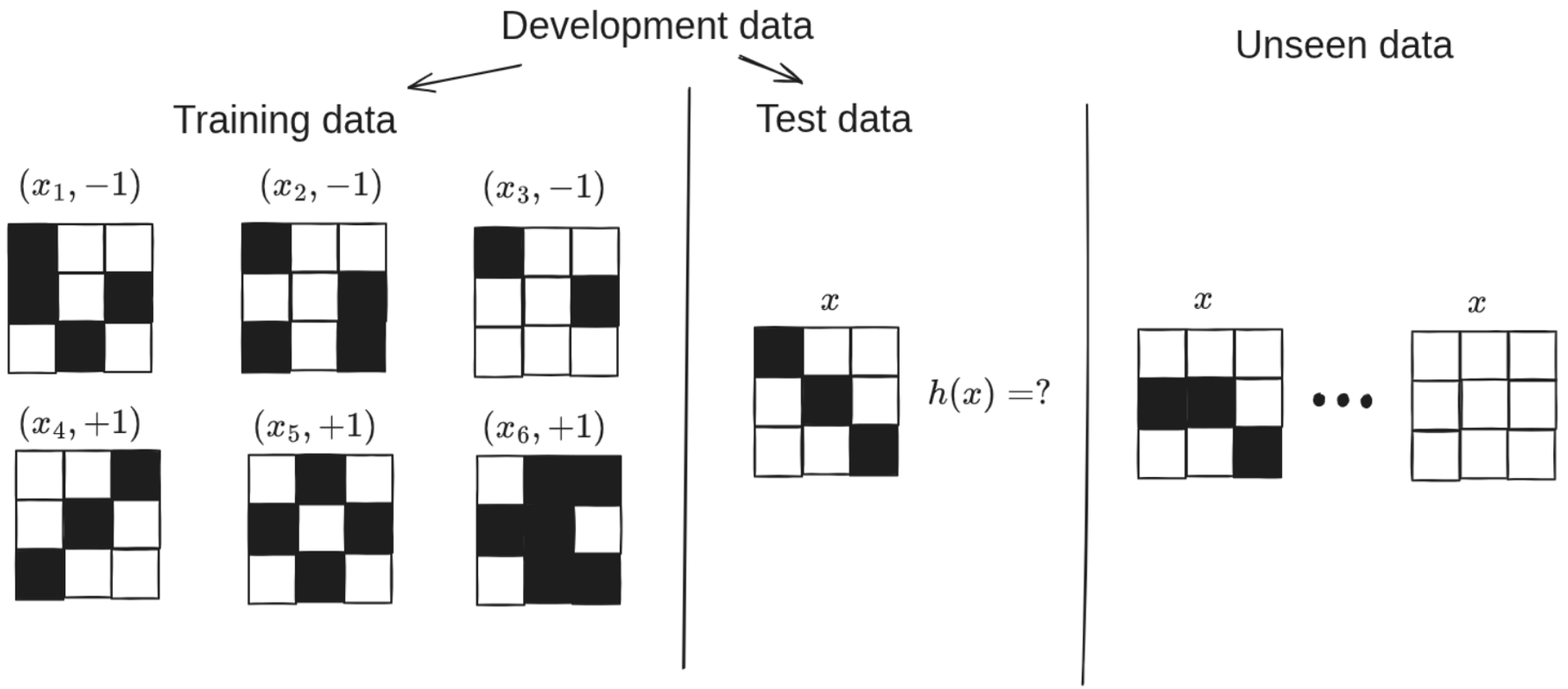
Goals:

1. Learn a predictor with low error on unseen data.
2. Accurately estimate its prediction error on unseen data.



Is Learning from Data Always Feasible?

Example: Try to learn $h\colon \{0,1\}^{3\times 3} \to \{+1,-1\}$ from the following data:

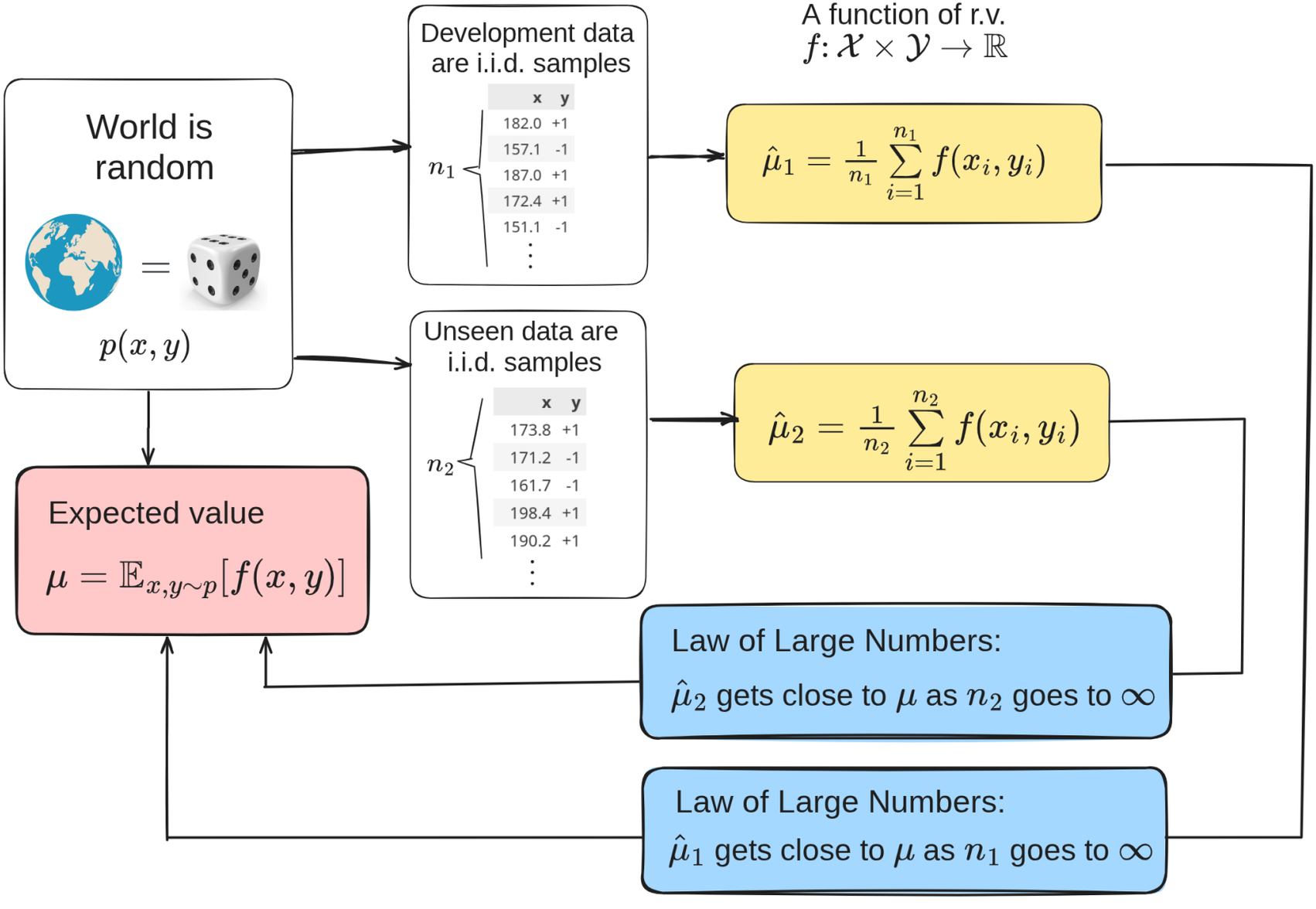


In the first block (Machine Learning Theory), we ask:

- Under what conditions is learning from data possible?
- How can we design a algorithms with guaranteed succeess?
- How is performance on training/test data related to performance on unseen data?

Picture adopted from: Abu-Mostafa et al: Learning from data. AMLBook, 2012.

Assumption that makes learning feasible: Independent and identically distributed data



Identically Distributed: Each data point comes from the same probability distibution.
Independent: The outcome of one data point does not affect the outcome of another.

Assumption that makes learning feasible: Independent and identically distributed data

- **Data:** We observe samples

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

drawn from independent and identically distributed (i.i.d.) random pairs

$$(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$$

- **Identically distributed:**

$$p(X_1 = x, Y_1 = y) = p(X_2 = x, Y_2 = y) = \cdots = p(X_n = x, Y_n = y), \quad \forall x \in \mathcal{X}, y \in \mathcal{Y}.$$

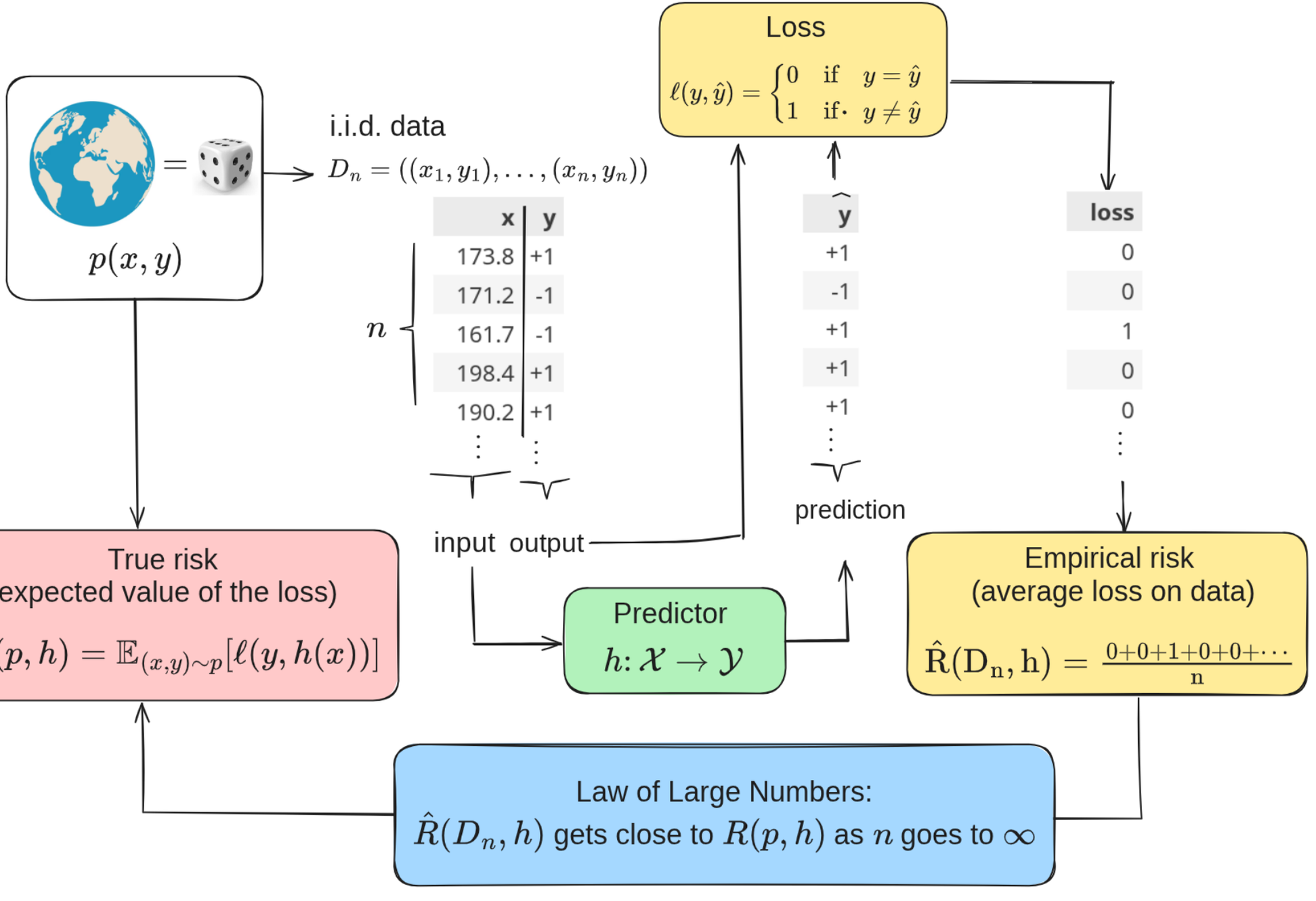
Remark: We will use $p(x, y)$ as shorthand for $p(X = x, Y = y)$.

- **Independent:** The occurrence of one sample does not affect another:

$$p\big((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\big) = \prod_{i=1}^{n} p(x_i, y_i).$$

Quiz: In the context of gender prediction, give an example of when the samples are (i) not independent, and (ii) not identically distributed.

Prediction error



Prediction error

- **Loss function** $\ell\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ penalises wrong predictions, i.e. $\ell(y, \hat{y})$ is the loss for predicting $\hat{y} = h(x)$ when y is the true output.

  Example: 0/1-loss

$$\ell(y, \hat{y}) = \begin{cases} 0 & \text{if} \quad y = \hat{y} \\ 1 & \text{if} \quad y \neq \hat{y} \end{cases}$$

- **Empirical risk (training/test error)** evaluates the performance of a a predictor $h\colon \mathcal{X} \to \mathcal{Y}$ on given data $D_n = ((x_1, y_1), \ldots, (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$:

$$\hat{R}(D_n, h) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, h(x_i))$$

- **True risk (a.k.a. true error, expected loss)** evaluates the performance of a predictor $h\colon \mathcal{X} \to \mathcal{Y}$ on unseen data:

$$R(p, h) = \int \sum_{y \in \mathcal{Y}} \ell(y, h(x))\; p(x, y)\; \mathrm{d}x = \mathbb{E}_{(x,y)\sim p}\Big[\ell(y, h(x))\Big]$$

- **Law of Large Numbers:** Provided the data D_n are i.i.d. samples drawn from $p(x, y)$, the empirical risk $\hat{R}(D_n, h)$ converges (in probability) to the true risk $R(p, h)$:

$$\hat{R}(D_n, h) \xrightarrow{p} R(p, h)$$

The (Bayes) optimal predictor

- **The goal** is to find a predictor $h\colon \mathcal{X} \to \mathcal{Y}$ which minimizes the true risk

$$R(p,h) = \int \sum_{y\in\mathcal{Y}} \ell(y,h(x))\; p(x,y)\; \mathrm{d}x$$

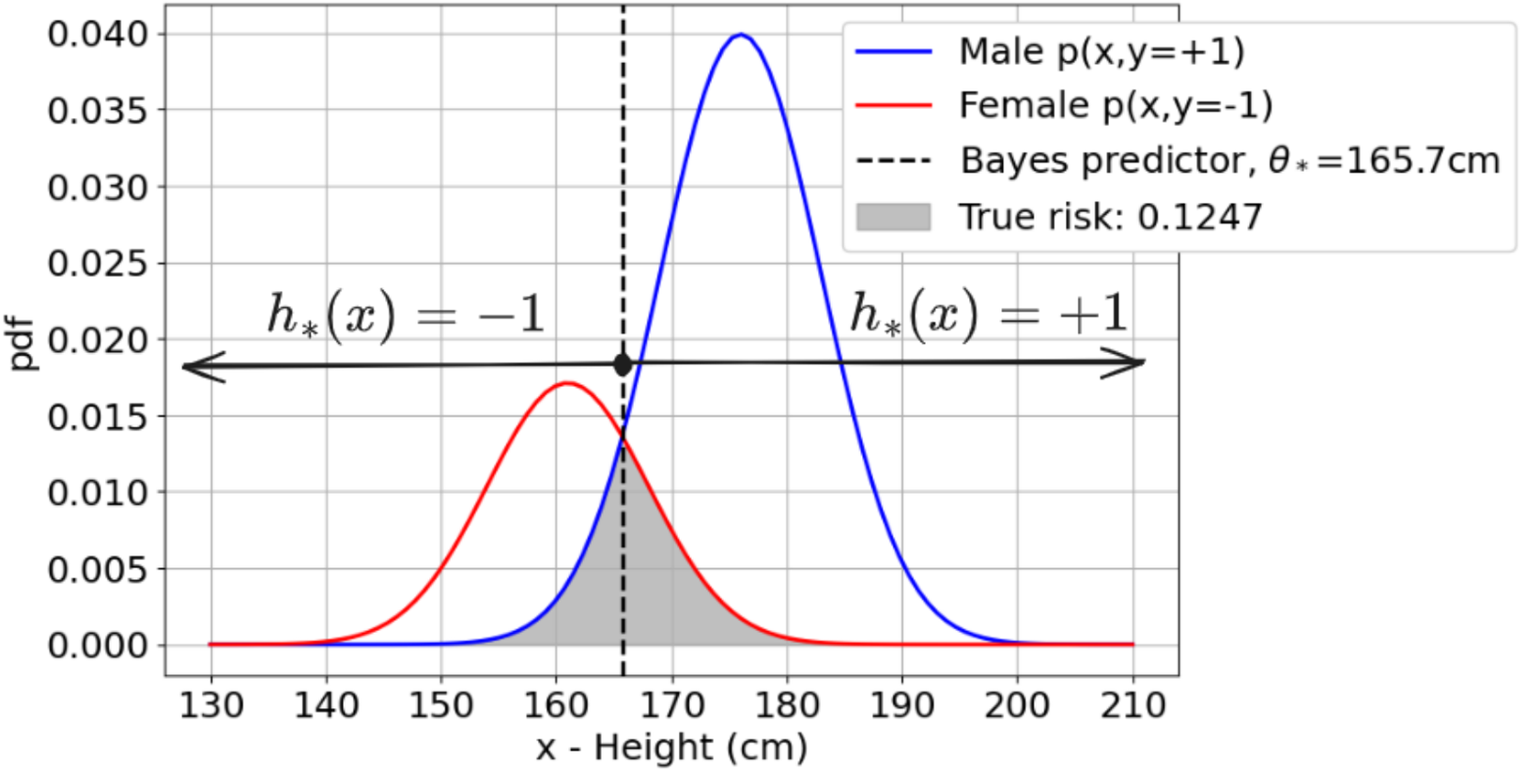
- **The optimal (a.k.a. Bayes) predictor:** minimizing $R(p,h)$ reads

$$\begin{aligned} h_* \in \operatorname*{arg\,min}_{h\in\mathcal{Y}^{\mathcal{X}}} R(p,h) \quad \Rightarrow \quad h_*(x) &= \operatorname*{arg\,min}_{\hat{y}\in\mathcal{Y}} \sum_{y\in\mathcal{Y}} p(x,y)\ell(y,\hat{y}) \\ &= \operatorname*{arg\,min}_{\hat{y}\in\mathcal{Y}} \sum_{y\in\mathcal{Y}} p(y \mid x)\ell(y,\hat{y}) \end{aligned}$$

Example: the (Bayes) optimal predictor

$$\mathcal{X} = \mathbb{R}\,, \mathcal{Y} = \{+1, -1\}\,, \quad p(x,y) = p(y)\,\frac{1}{\sqrt{2\pi}\sigma}\,e^{-\frac{1}{2\sigma^2}(x-\mu_y)^2}\,, \quad \ell(y,y') = \begin{cases} 0 & \text{if} \quad y = y' \\ 1 & \text{if} \quad y \neq y' \end{cases}$$

$\mu_+ = 176\,\text{cm}$
$\mu_- = 161\,\text{cm}$
$\sigma = 7\,\text{cm}$
$p(+1) = 0.7$
$p(-1) = 0.3$



- Bayes predictor: $h_*(x) = \arg\max_{y \in \mathcal{Y}} p(y \mid x) = \text{sign}(x - \theta_*)$
- True risk: $R(p, h_*) = \int_{-\infty}^{\theta_*} p(x,+1)dx + \int_{\theta_*}^{\infty} p(x,-1)dx$

Summary

Key concepts introduced:

- Prediction problem
- I.I.D. data assumption
- True Risk/Error, empirical risk/error, and loss function
- Bayes optimal predictor