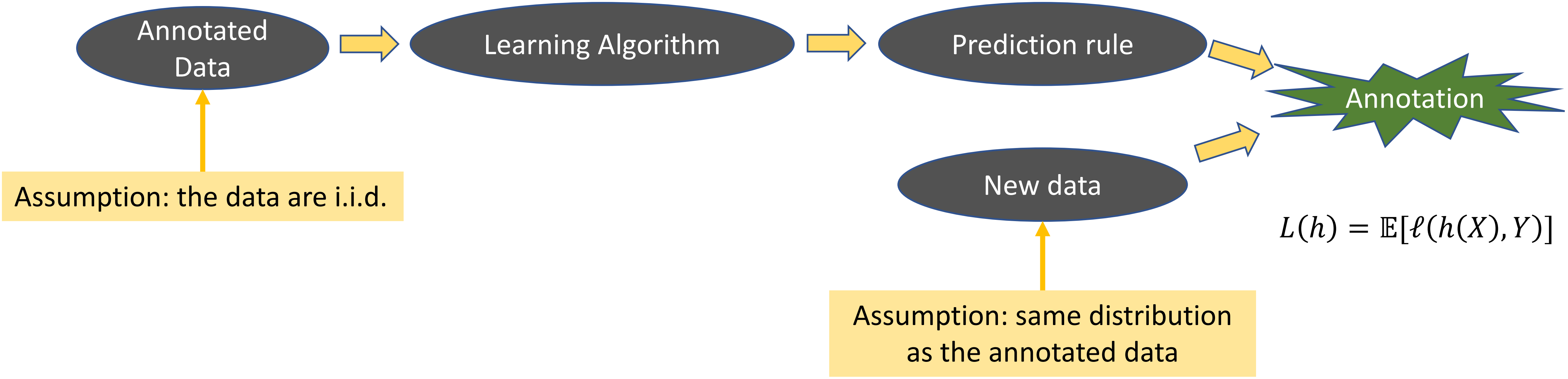
From ML-A to ML-B

Theory gets tighter

Yevgeny Seldin

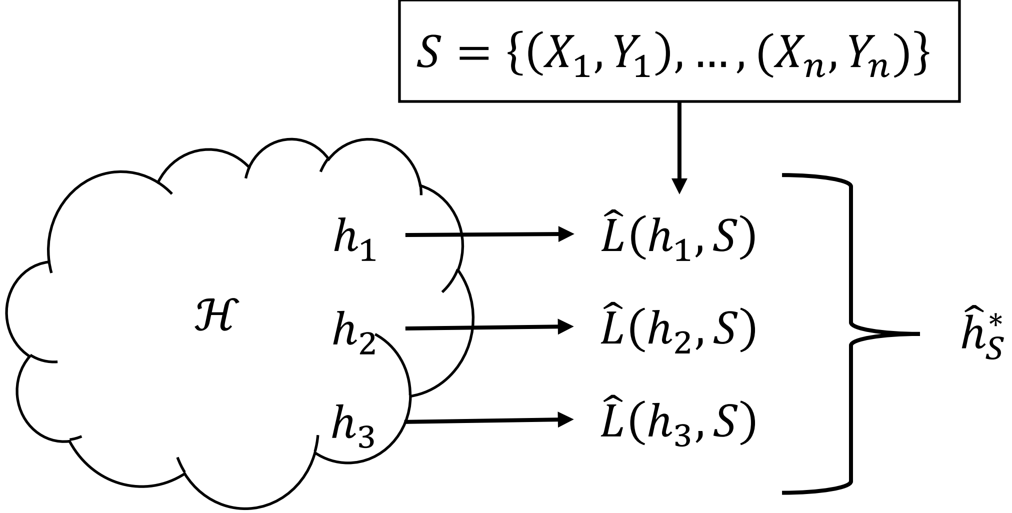
The same "Classical" Supervised Learning



$$L(h) = \mathbb{E}[\ell(h(X), Y)]$$

- With the same assumptions
- Primary quantity of interest - $L(h)$ – unknown
- Known: $\hat{L}(h, S)$
- Major question: what can we say about $L(h)$ based on $\hat{L}(h, S)$?

ML-A, a quick reminder



- What can we say about $L(\hat{h}_S^*)$?

$$\mathbb{P}\left(L(\hat{h}_S^*) \geq \hat{L}(\hat{h}_S^*, S) + \varepsilon\right) \leq \mathbb{P}\left(\exists h \in \mathcal{H}:\ L(h) \geq \hat{L}(h, S) + \varepsilon\right)$$

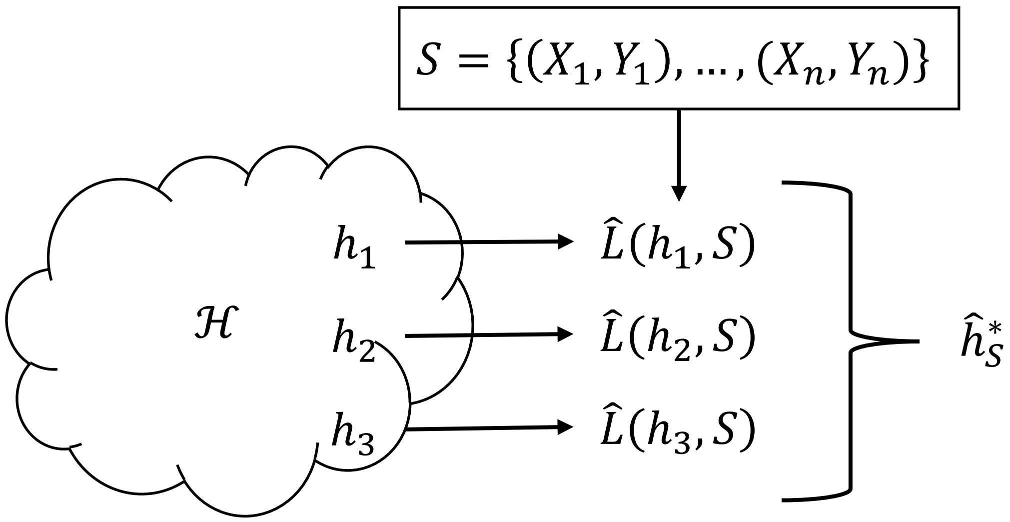
$$\leq \sum_{h \in \mathcal{H}} \mathbb{P}\left(L(h) \geq \hat{L}(h, S) + \varepsilon\right)$$

$$\leq \sum_{h \in \mathcal{H}} e^{-2n\varepsilon^2} = \underbrace{M}_{\substack{\text{Selection} \\ \text{(Union bound)}}} \times \underbrace{e^{-2n\varepsilon^2}}_{\substack{\text{Concentration} \\ \text{(Hoeffding)}}} = \delta$$

- Occam's razor (countable $\mathcal{H}$): $\mathbb{P}\left(\exists h \in \mathcal{H}: L(h) \geq \hat{L}(h, S) + \sqrt{\frac{\ln \frac{1}{\pi(h)\delta}}{2n}}\right) \leq \delta$
  - Based on bounding $\mathbb{P}\left(\exists h \in \mathcal{H}:\ L(h) \geq \hat{L}(h, S) + \varepsilon_h\right)$
  - $\varepsilon_h = \sqrt{\frac{\ln \frac{1}{\pi(h)\delta}}{2n}}$

From ML-A to ML-B

- We are in the same setting



- With the same assumptions:
  - $\{(X_1, Y_1), \dots, (X_n, Y_n)\}$ are i.i.d.
  - New data points come from the same distribution
- We will:
  - Derive tighter and practically useful bounds
  - Learn how to control selection from uncountable $\mathcal{H}$

ML-A vs. ML-B

ML-A

- Concentration
  - Hoeffding's inequality
  - $\mathbb{P}\left(p \geq \hat{p}_n + \sqrt{\frac{\ln\frac{1}{\delta}}{2n}}\right) \leq \delta$
  - "Slow rate"
- Selection
  - Occam's razor
    - Selection from countable $\mathcal{H}$
    - Tool: union bound

ML-B

- Concentration
  - kl-inequality
  - $\mathbb{P}\left(p \geq \hat{p}_n + \sqrt{\frac{2\hat{p}_n \ln\frac{1}{\delta}}{n}} + \frac{2\ln\frac{1}{\delta}}{n}\right) \leq \delta$
  - "Fast rate"
  - Bernstein's inequalities – "fast rate" based on small variance
- Selection
  - VC analysis
    - Selection from uncountable $\mathcal{H}$
    - Tools: bound on *effective selection* (which is countable) + a union bound
  - PAC-Bayesian analysis
    - Selection from uncountable $\mathcal{H}$
    - Tools: *active avoidance of selection* (by randomization) + *change of measure inequality* (a continuous substitute to the union bound; union bound is a special case when the selection is discrete)

Weighted Majority Votes

- ML-A:
  - Random Forests – majority vote of decision trees
  - Majority vote often performs better than individual classifiers
    - Cancellation of errors effect

- ML-B:
  - PAC-Bayesian analysis of generalization power of the weighted majority vote
  - PAC-Bayesian weight tuning for weighted majority votes
  - Boosting – targeted construction of ensembles with anticorrelated errors