

Explainable Empirical Risk Minimization for Trustworthy AI

Alex(ander) Jung
Assistant Professor for Machine Learning
Department of Computer Science
Aalto University



Outline

- Empirical Risk Minimization
- What is an Explanation?
- Measuring Explainability
- Explainable Empirical Risk Minimization

Outline

- Empirical Risk Minimization
- What is an Explanation?
- Measuring Explainability
- Explainable Empirical Risk Minimization

ML Principle (informal)


fit **model** to **data** to make **accurate**
predictions or forecasts !

4, 5, 6, 7, 8, ?

1. element 2. 3. 4. 5. 6.

4, 5, 6, 7, 8, ?

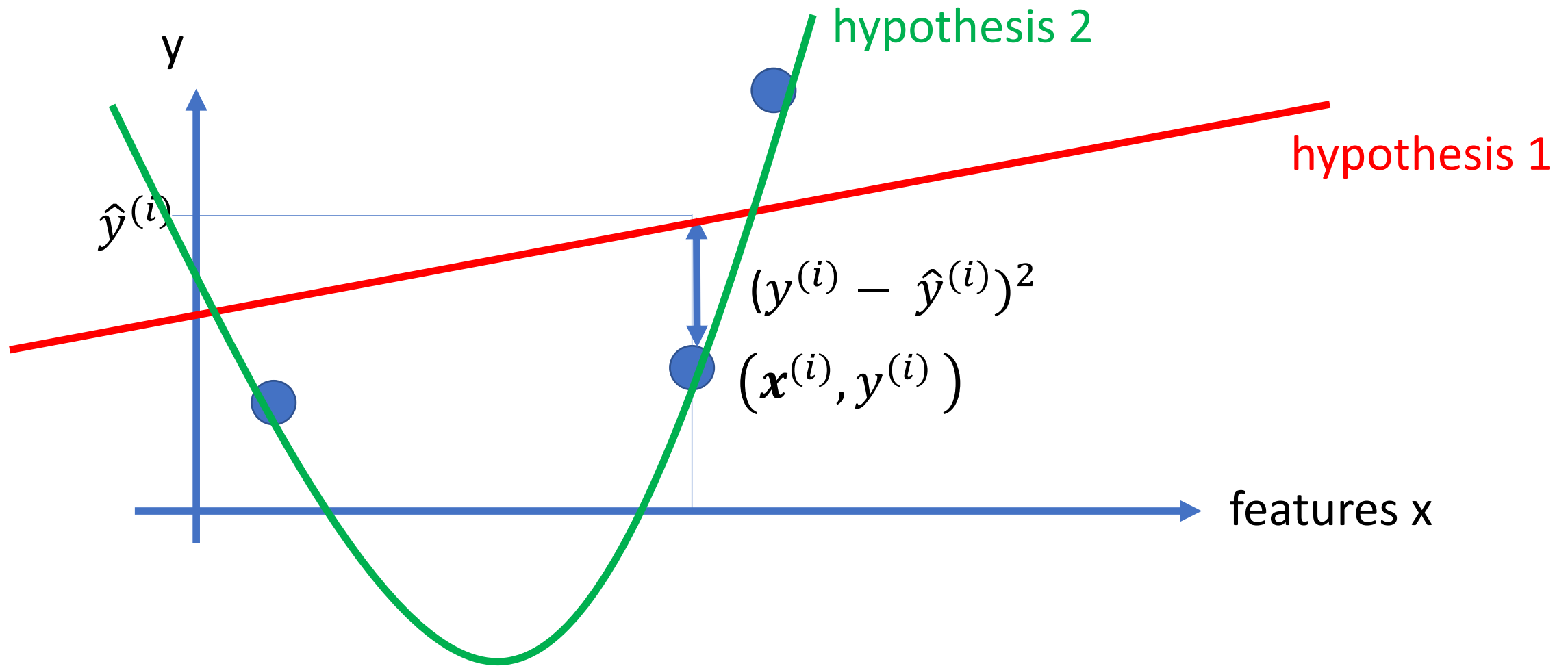
“data point”



ML Principle (more formal)

learn hypothesis out of a hypothesis space
(model) that allows **to predict label** of a
data point **from** its **features**

Empirical Risk Minimization



Empirical Risk Minimization

$$\hat{h} \in \operatorname{argmin}_{h \in \mathcal{H}} \hat{L}(h|\mathcal{D})$$

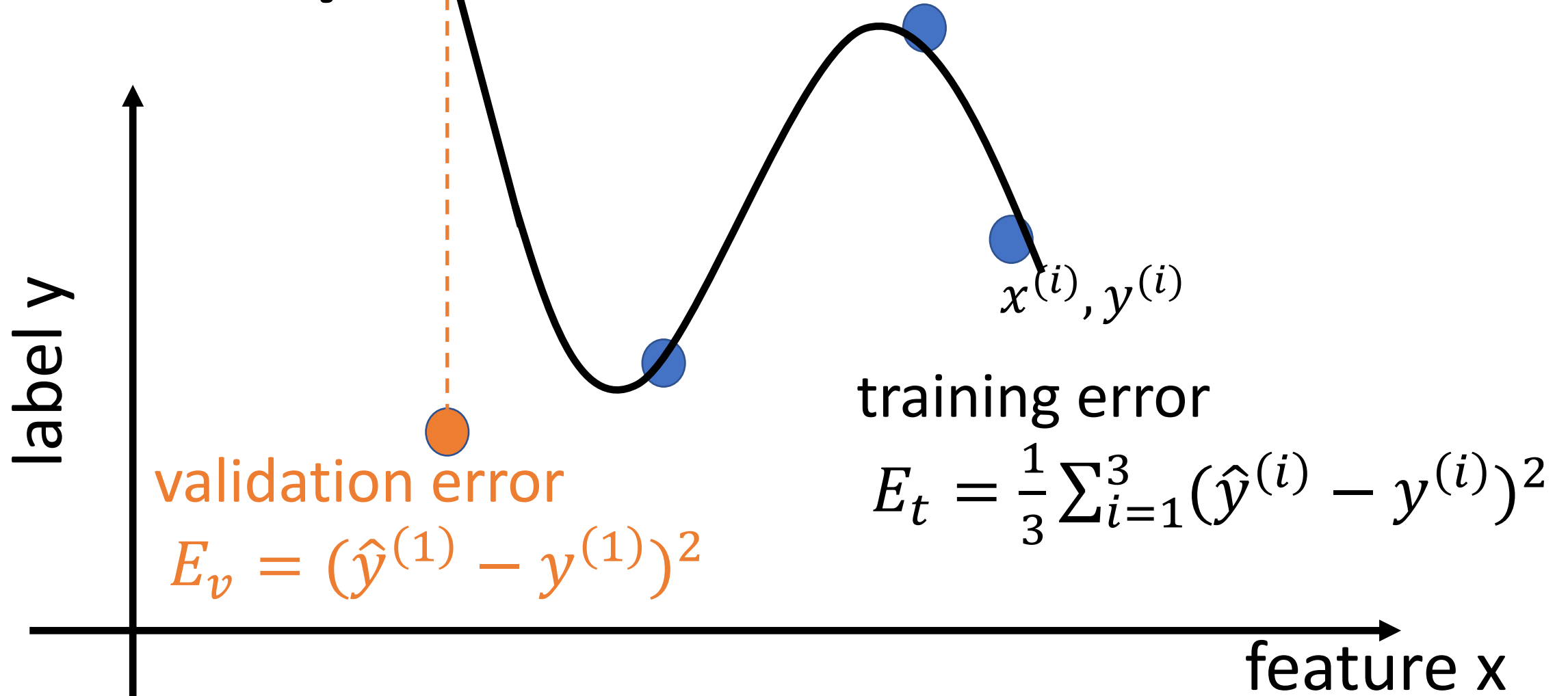
(2.16) $\stackrel{=}{=} \operatorname{argmin}_{h \in \mathcal{H}} (1/m) \sum_{i=1}^m L(\underbrace{(\mathbf{x}^{(i)}, y^{(i)})}_{\text{data}}, h).$

loss

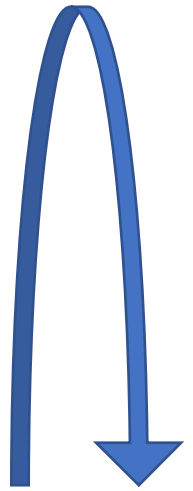
model

data

Always Validate !

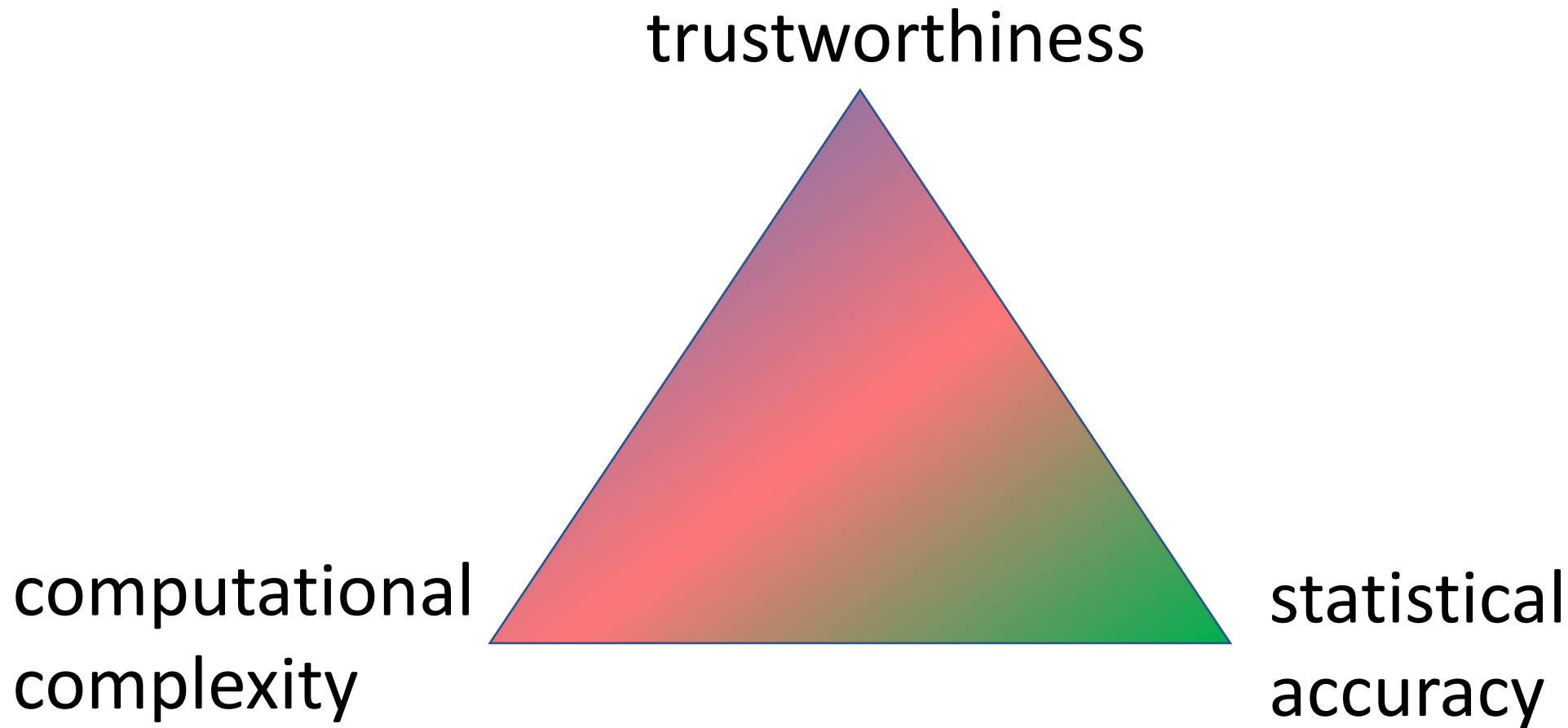


Life-Cycle of ML



- learn hypothesis $h(x)$ via ERM (“**train**”)
- apply $h(x)$ to new data (“**validate**”)
- **adapt** ERM design choices and repeat

Design Choices: Data, Model, Loss.



- Human agency and oversight
- Technical robustness and safety
- Privacy and data governance
- **Transparency**
- Diversity, non-discrimination and fairness
- Societal and environmental wellbeing
- Accountability



<https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html>

Explainability.

*“...Technical explainability requires that the decisions made by an AI system can be **understood and traced by human beings**. Moreover, trade-offs might have to be made between enhancing a system's explainability (which may reduce its accuracy) or increasing its accuracy (at the cost of explainability)...”*

Two Key Questions

- what is an explanation ?
- how to measure explainability ?

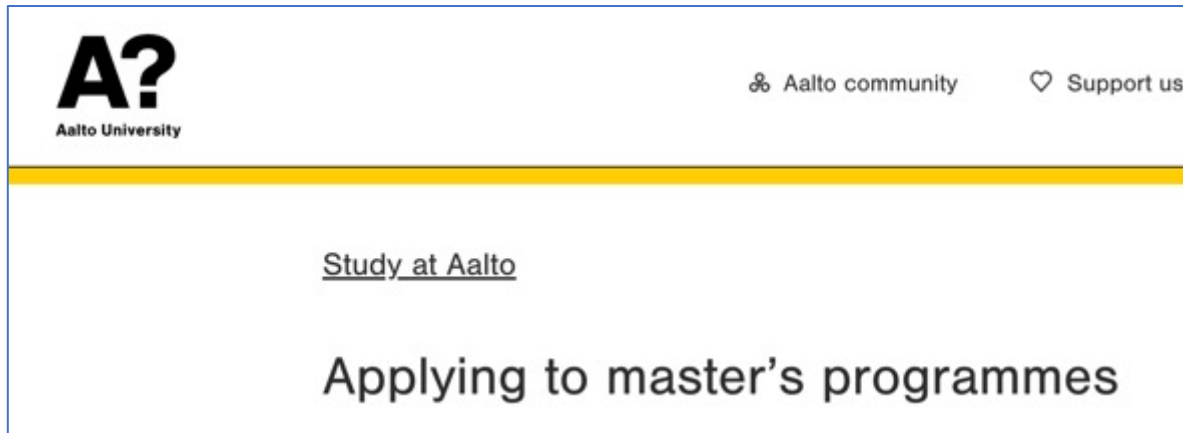
Outline

- Empirical Risk Minimization
- What is an Explanation?
- Measuring Explainability
- Explainable Empirical Risk Minimization

ISO/IEC TR 24028

*“...An explanation is always an **attempt to communicate understanding**. The effectiveness of an explanation can be improved by tailoring...”*

Premium Version of Explanations ...



Among my students,

explaining a ML method could amounts to

- specification of **data** format and source
- specification of **model** (hypothesis space)
- specification of **loss** function

Explaining Entire ML Method.

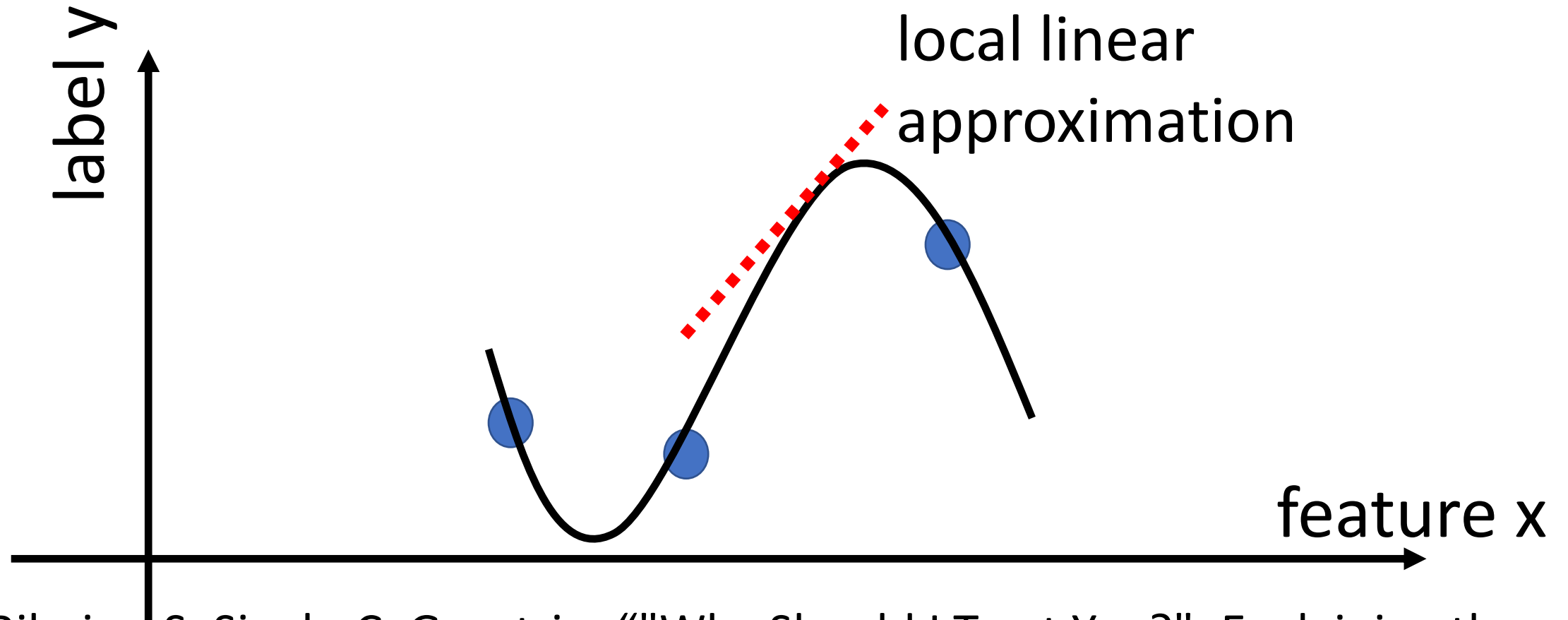
“linear regression learns a linear hypothesis by minimizing the average squared error on training set”

Explaining Individual Predictions.

provide information about how prediction $h(x)$ is computed for given data point with features x

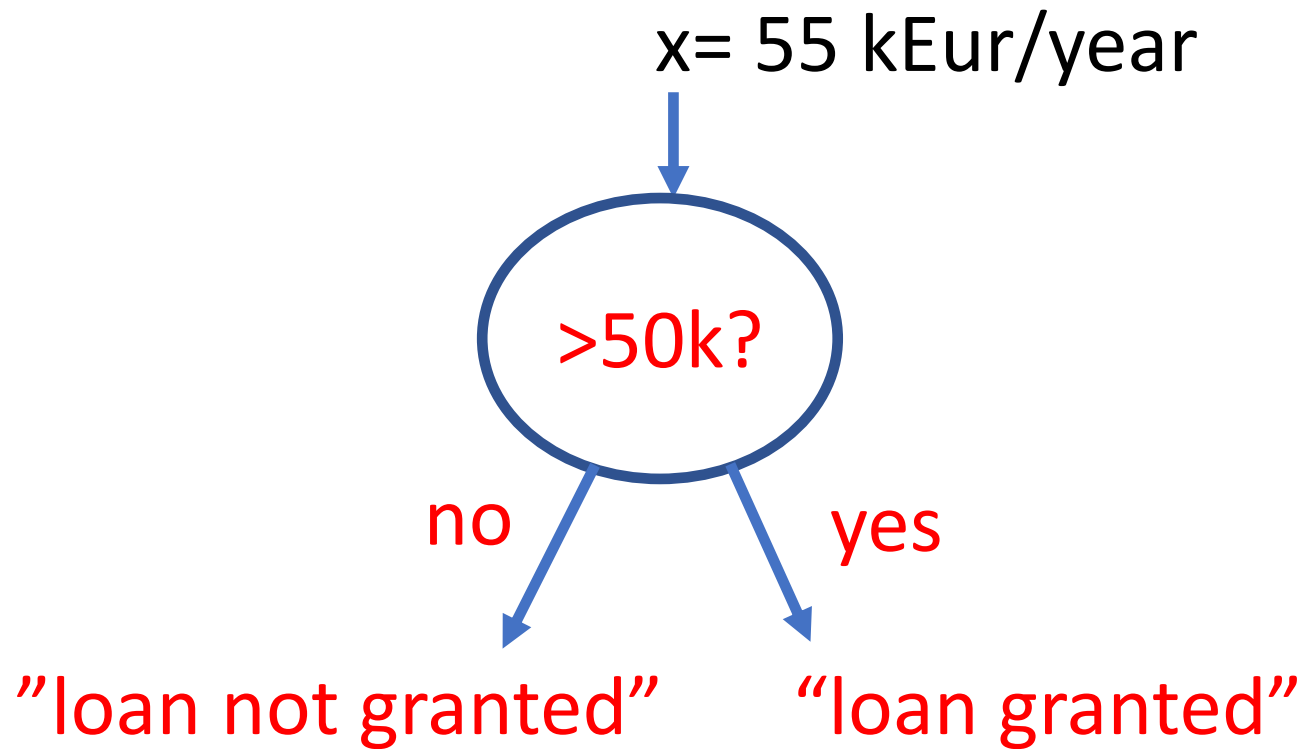
e.g., “the prediction is obtained since $x=4$ for this data point and we use a linear hypothesis $h(x) = w_1 * x_1 + w_2 * x_2$ with weights $w_1 = 10$ and $w_2=4$ ”

LIME - Local Interpretable Model-Agnostic

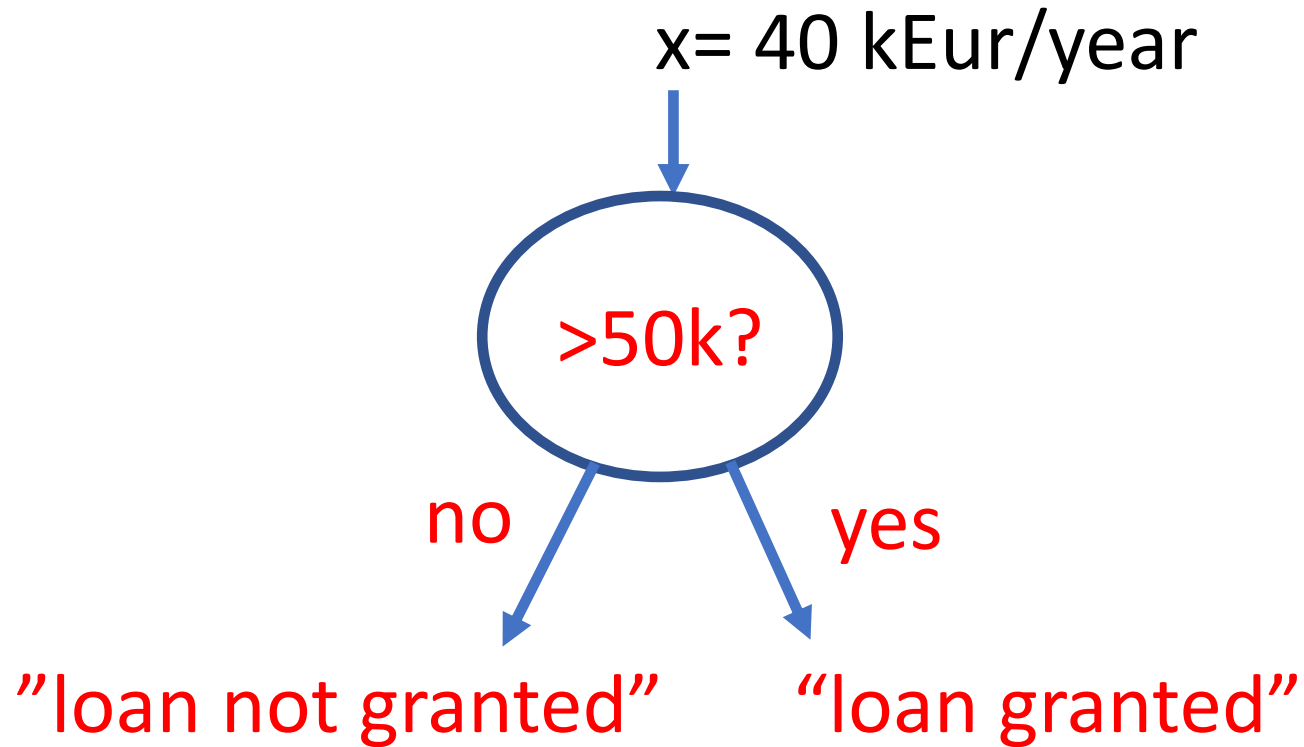


M. Ribeiro, S. Singh, C. Guestrin, ““Why Should I Trust You?": Explaining the Predictions of Any Classifier”, *arXiv e-prints*, 2016.

Explaining Decision Tree Prediction.



Explaining via Counterfactual.



if your salary would be higher than 50 kEur, then the loan would have been granted

Explaining a Prediction.

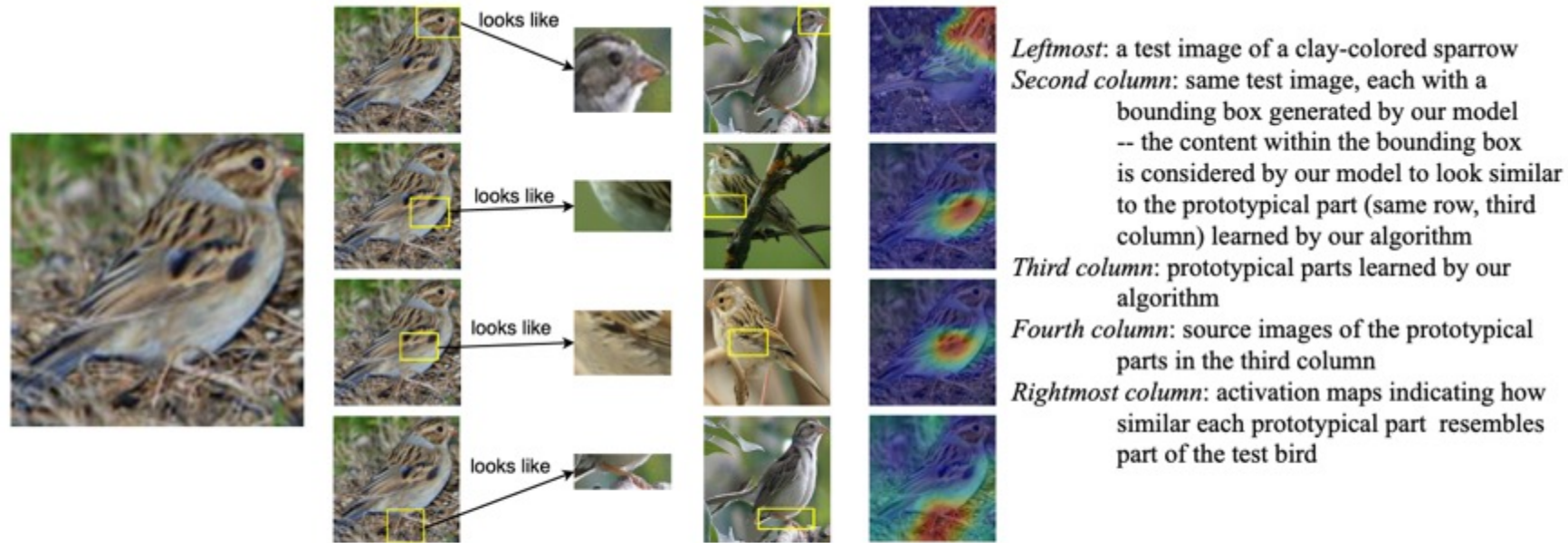
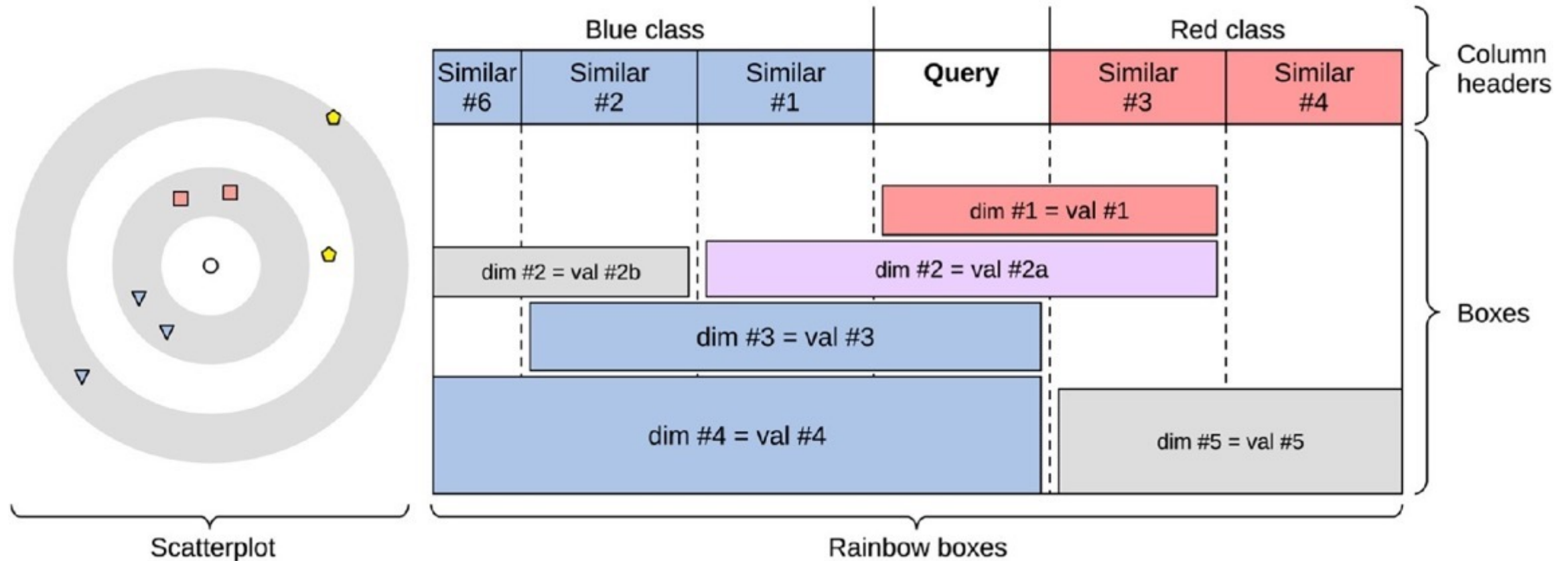


Figure 1: Image of a clay colored sparrow and how parts of it look like some learned prototypical parts of a clay colored sparrow used to classify the bird's species.

Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, Jonathan K. Su **"This Looks Like That: Deep Learning for Interpretable Image Recognition", Neurips 2019**

Case-Based Reasoning.



Lamy et.al., "Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach," Artificial Intelligence in Medicine, Volume 94, 2019.

Towards a Definition.

“ explanation is some artefact “e” that is revealed to a user “u” who is also served the prediction $\hat{y} = h(\mathbf{x})$ for a data point with features \mathbf{x} ”

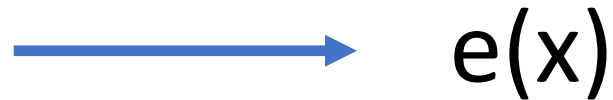
A Precise Definition.

since we serve explanations for predictions on unlabelled data, explanation is a (stochastic) function of features only,

data point



features x, y



$e(x)$

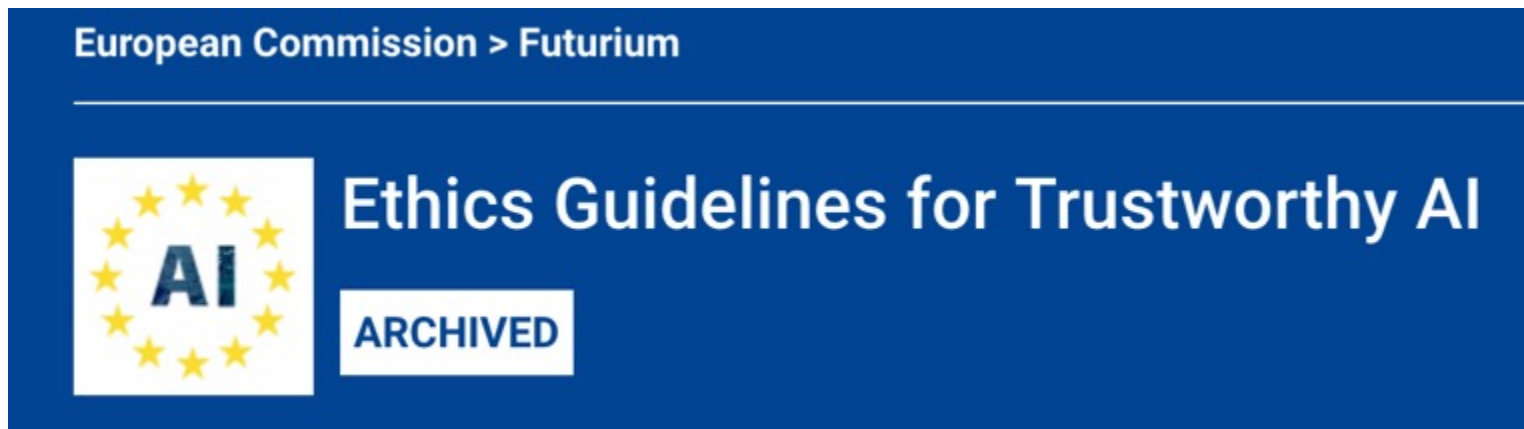
restrict function $e(.)$ to belong to feasible set \mathcal{F} (similar to a hypothesis space!)

Outline

- Empirical Risk Minimization
- What is an Explanation?
- Measuring Explainability
- Explainable Empirical Risk Minimization

Explainability is Subjective.

*“... explanation should be timely and **adapted** to the expertise of the **stakeholder** concerned (e.g. layperson, regulator or researcher)....”*



Adapting to User

SEO Basics: What are user signals?

4 October 2017 | [13 Comments](#) | Tags [Google Analytics](#), [SEO basics](#), [Webmaster tools](#)

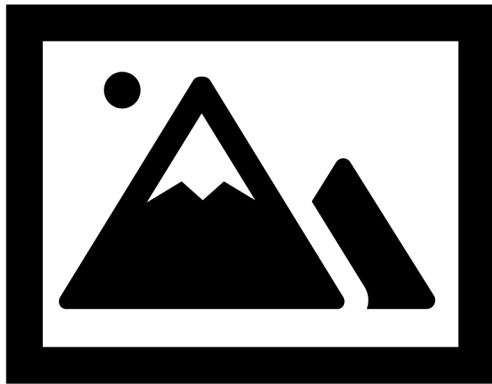


“User signals are behavioral patterns.... The most important user signals are the bounce rate and the click-through rate (CTR)”

<https://yoast.com/what-are-user-signals/>

User Signal.

data point



features x ,
label y



user 1



user 2



user 3

User Brain Signal.



[Products](#) [Services](#) [Applications](#) [Science](#) [About us](#)



[← GO BACK TO BLOG](#)

NEUROTECHNOLOGY - SCIENCE & RESEARCH

**What is BCI? An
introduction to brain-
computer interface using
EEG signals**



User Psychological Signal



What do you see ?

<https://www.tutordale.com/what-do-you-see-pictures-psychology/>

User Signal via Interpretable Representation (Features)

*“...Lime explains those classifiers in terms of **interpretable representations (words)**, even if that is not the representation actually used by the classifier....”*

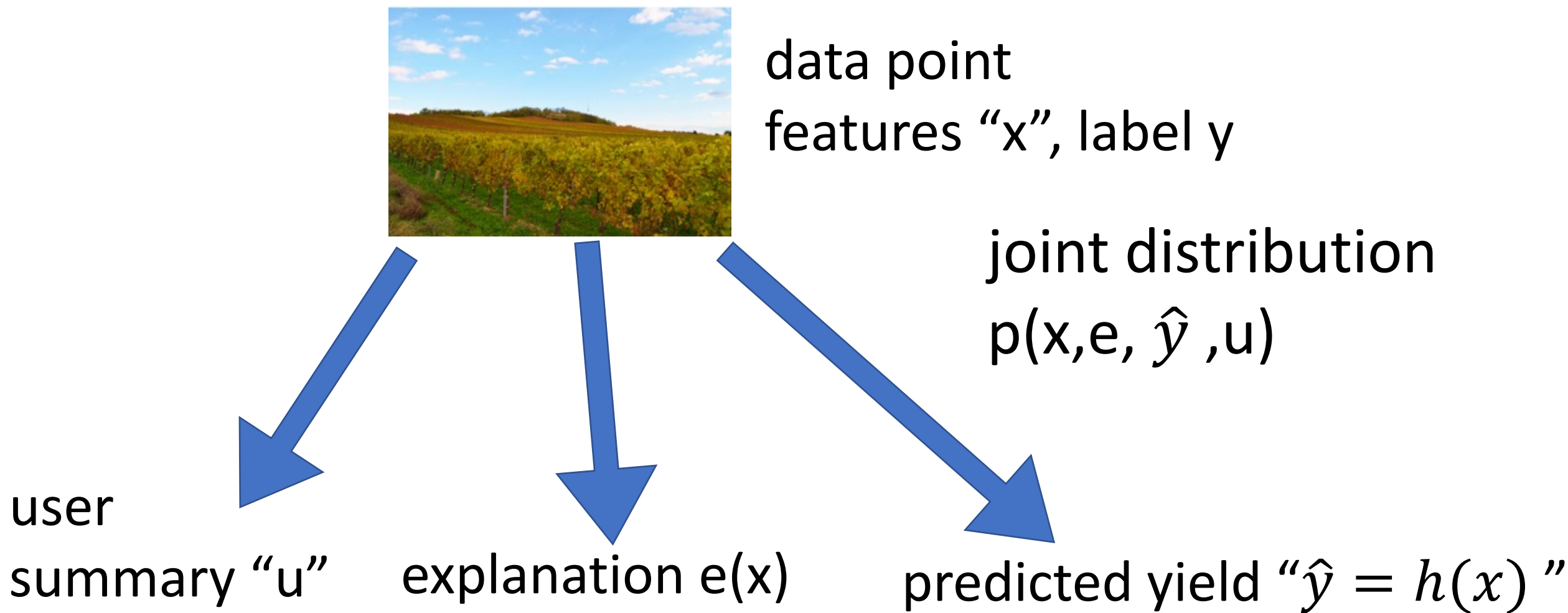
<https://homes.cs.washington.edu/~marcotcr/blog/lime/>

Abstract User Signal.

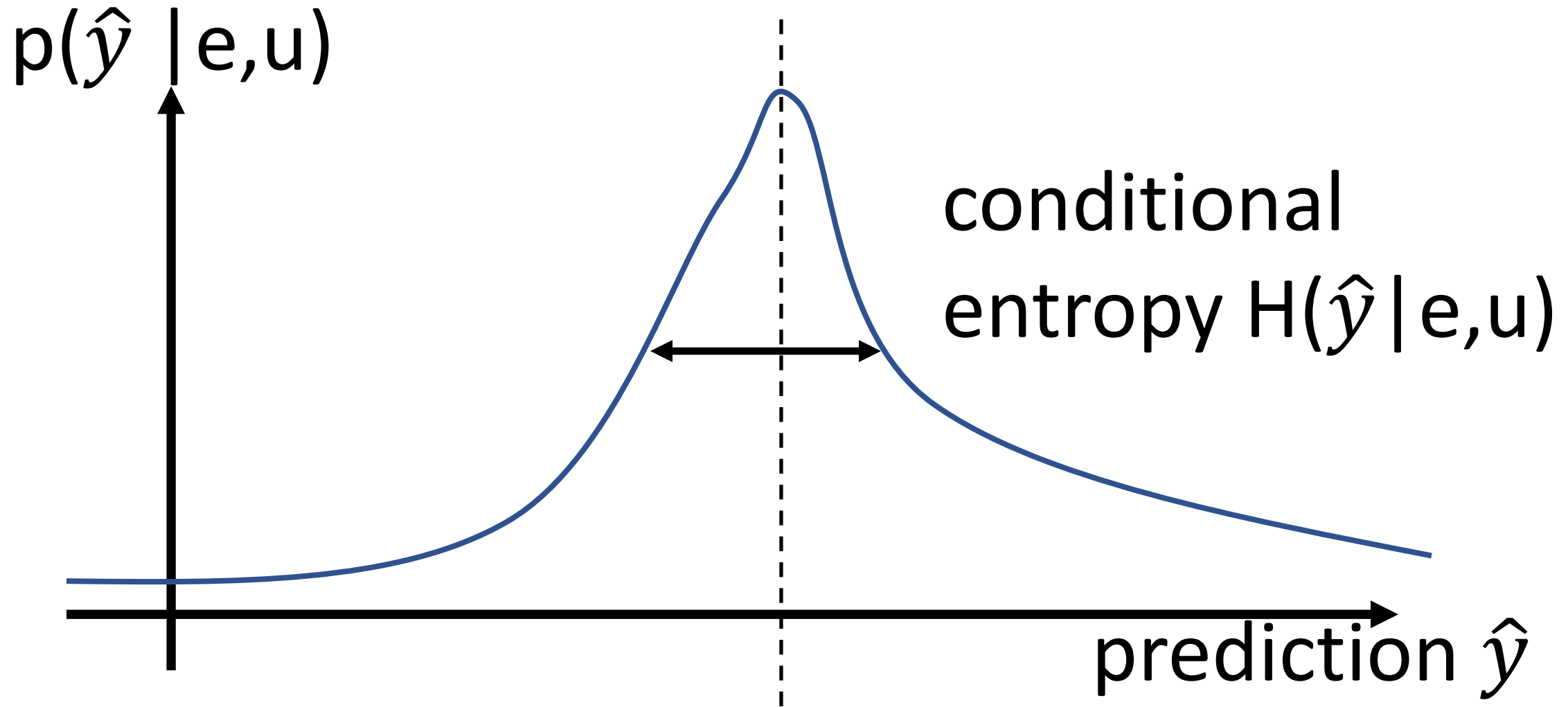
some user-specific quantity e associated with a data point

might interpret e as **user-specific feature** or label

A Simple Probabilistic Model



Explainability = Predictability



My Information-Theory Slide.

conditional entropy

$$I(e; \hat{y}|u) = H(\hat{y}|u) - H(\hat{y}|e, u)$$

conditional mutual information

see Chapter 8 of

**T. Cover, J. Thomas, “Elements of Information Theory”,
Wiley, 2005**

Computing Explanations

$$I(e^*; \hat{y}|u) = \sup_{e \in \mathcal{F}} I(e; \hat{y}|u)$$

← set of "allowed"
explanations

optimal explanation varies for different users u !
personalized explanations !

Towards an Algorithm.

$$I(e^*; \hat{y}|u) = \sup_{e \in \mathcal{F}} I(e; \hat{y}|u)$$

- estimate $h(\hat{y}|e, u)$ using i.i.d. training set $(x^{(1)}, u^{(1)}, \hat{y}^{(1)}) \dots (x^{(m)}, u^{(m)}, \hat{y}^{(m)})$
- choose tractable explanation space \mathcal{F}
- apply your favourite solver

The Story so far...

- measure (lack of) explainability via $H(\hat{y}|e, u)$
- construct map $e(x)$ to minimize $H(\hat{y}|e, u)$
- IDEA: skip explanation and minimize $H(\hat{y}|u)$ by learning simpler (interpretable) predictor $\hat{y} = h(x)$

Outline

- Empirical Risk Minimization
- What is an Explanation?
- Measuring Explainability
- Explainable Empirical Risk Minimization

Recall the ERM Principle

$$\hat{h} \in \operatorname{argmin}_{h \in \mathcal{H}} \hat{L}(h|\mathcal{D})$$

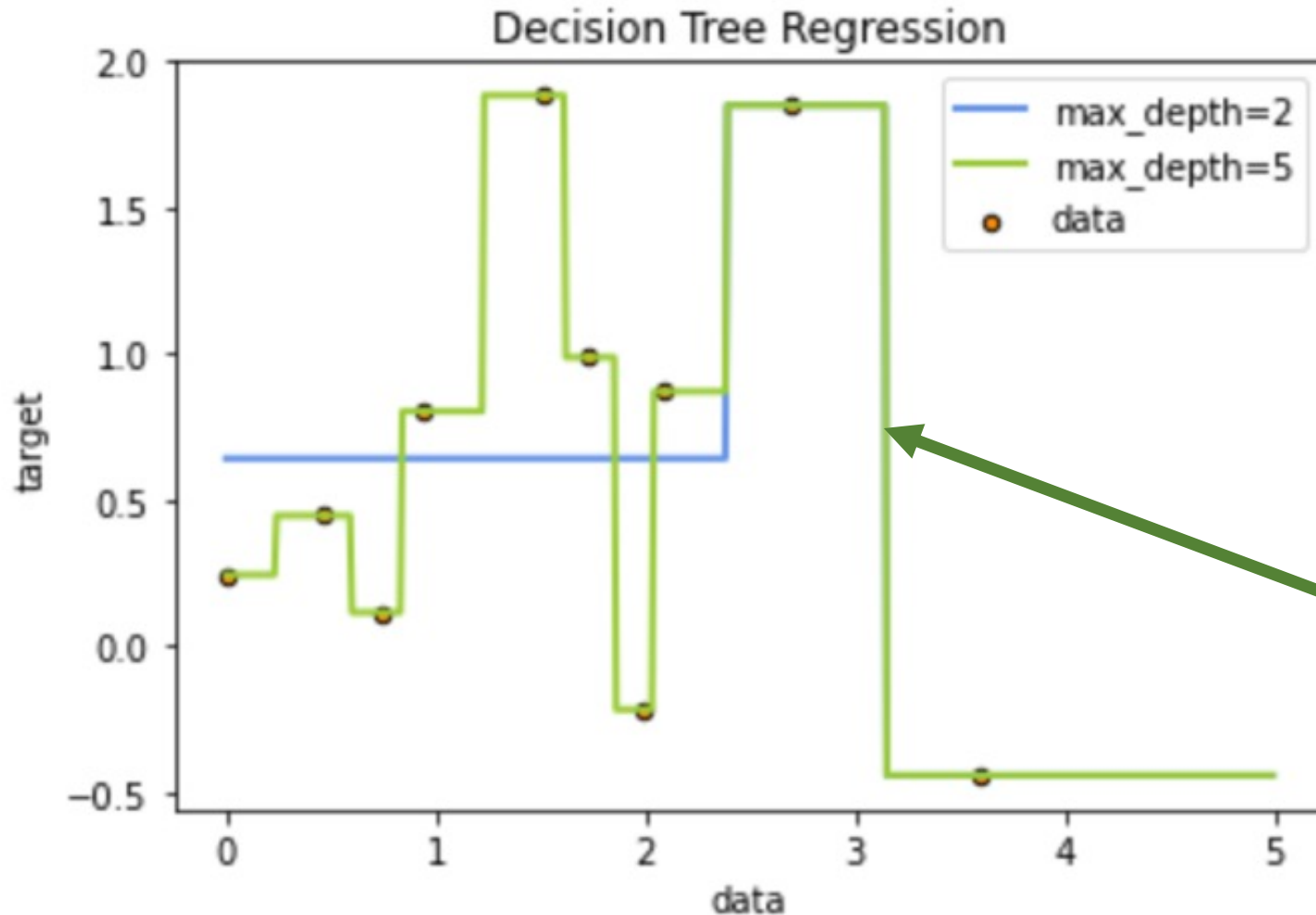
(2.16) $\stackrel{=}{=} \operatorname{argmin}_{h \in \mathcal{H}} (1/m) \sum_{i=1}^m L(\underbrace{(\mathbf{x}^{(i)}, y^{(i)})}_{\text{data}}, h).$

loss

model

data

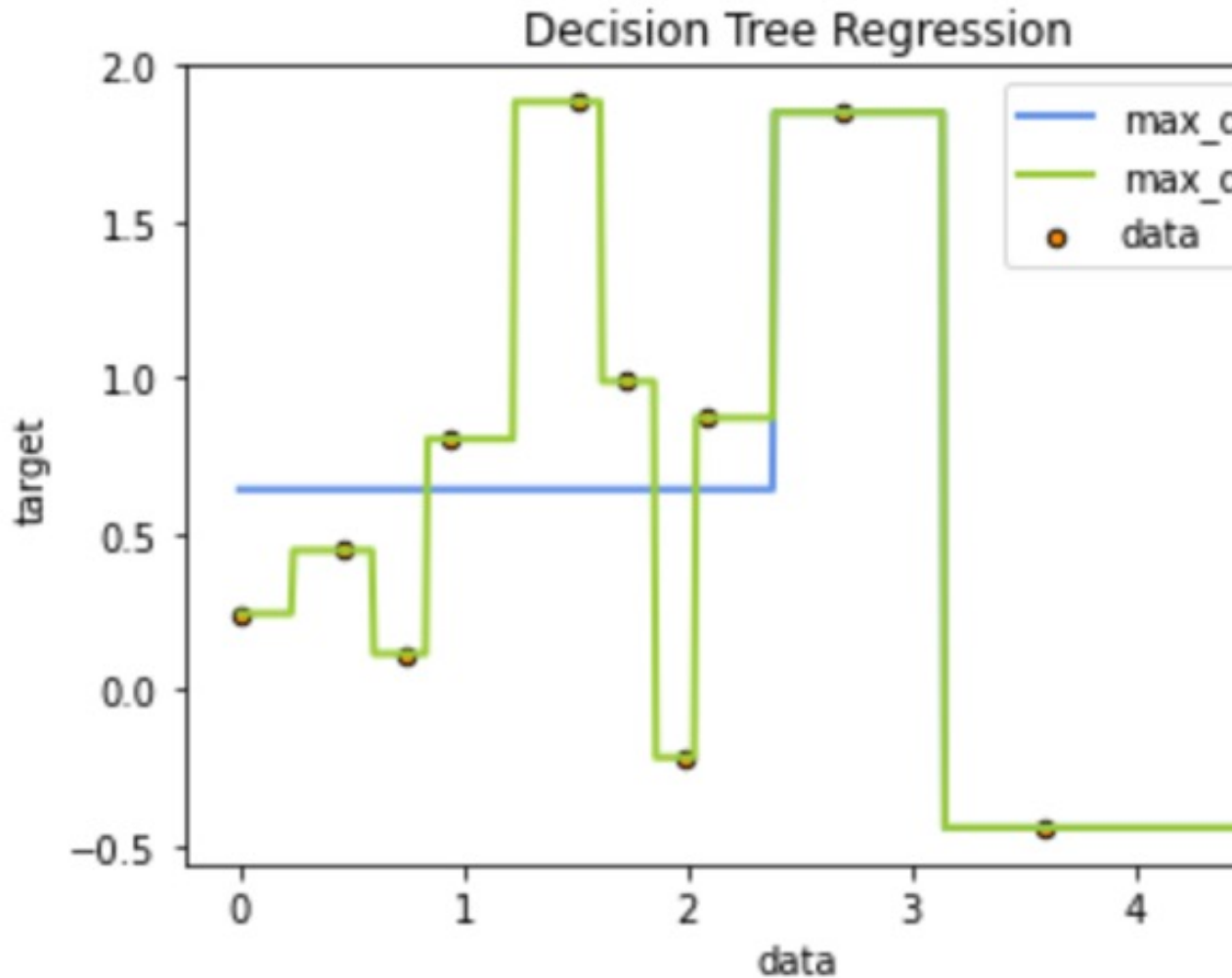
Overfitting.



do you like this learnt hypothesis which achieves ZERO empirical risk ?

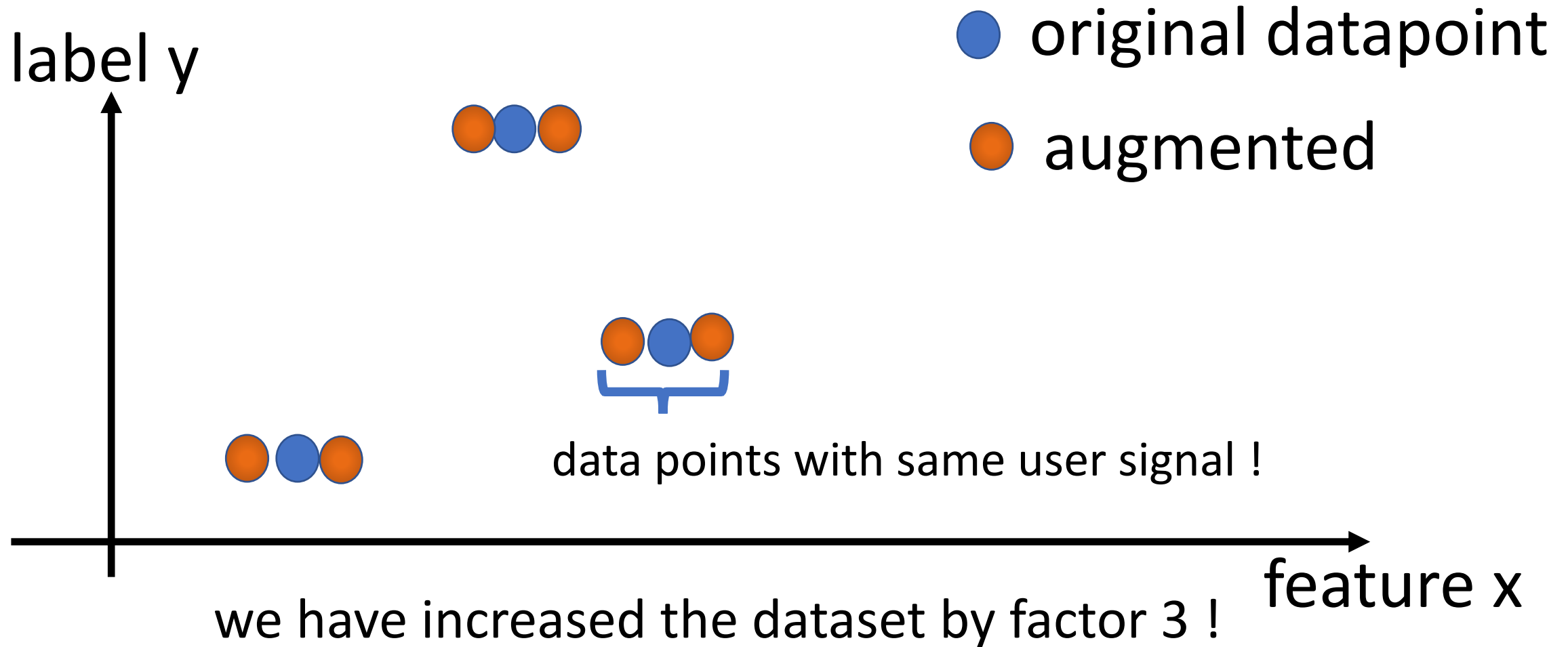


Avoid Overfitting by Regularization.



learnt hypothesis should be nearly constant for data points whose feature values are within distance 0.5

Regularization via Augmentation.

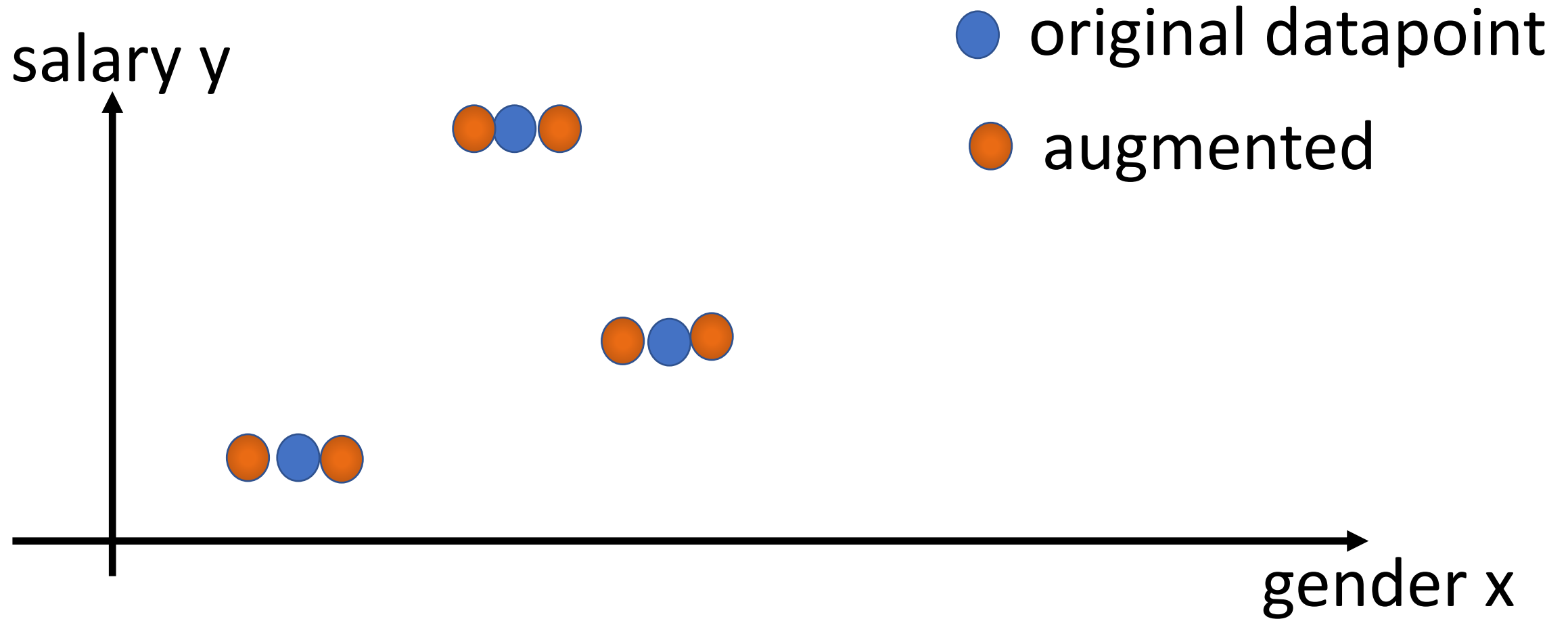


- Human agency and oversight
- Technical robustness and safety
- Privacy and data governance
- Transparency
- **Diversity, non-discrimination and fairness**
- Societal and environmental wellbeing
- Accountability

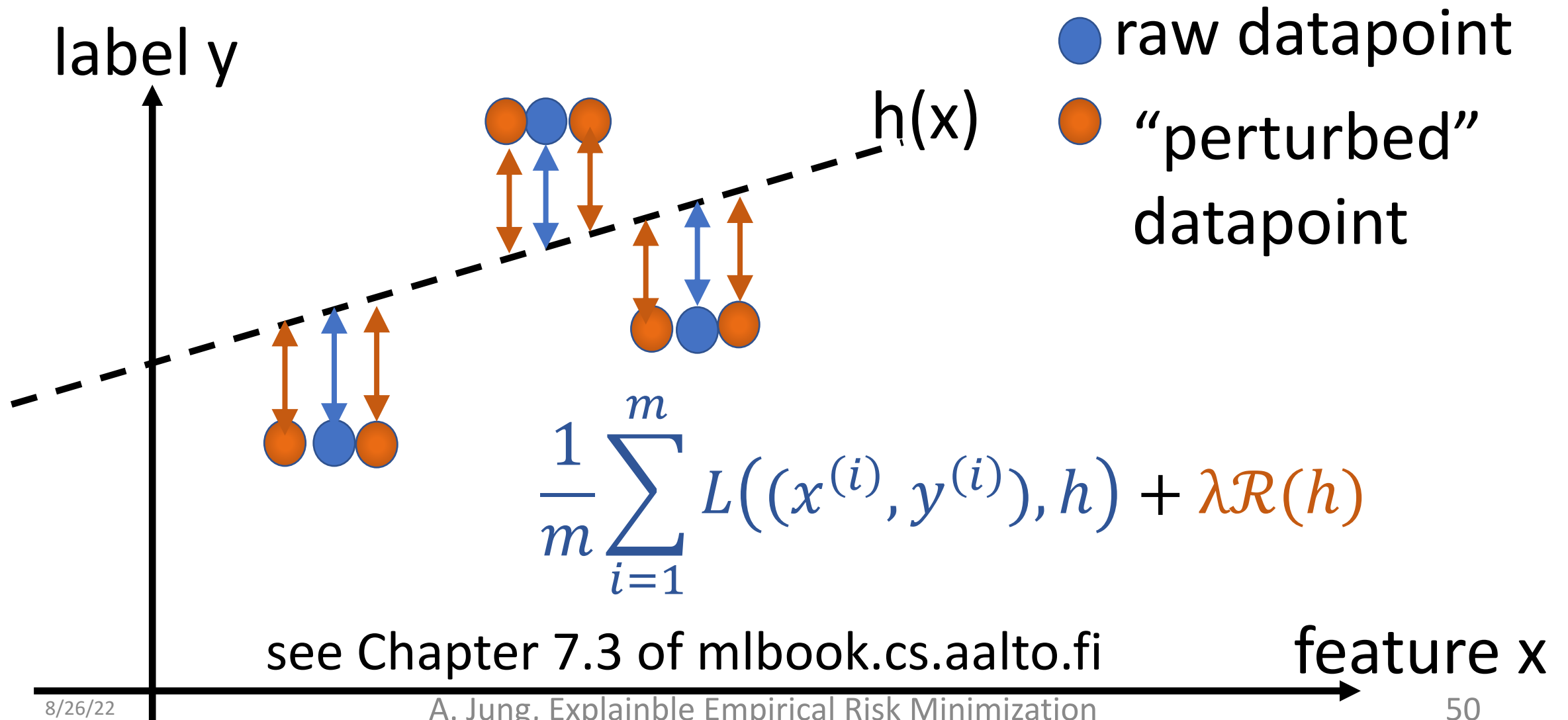


<https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html>

Fairness via Augmentation.



Regularization = Implicit Data Aug.



Explainable ERM (EERM)

$$\min_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^m L((x^{(i)}, y^{(i)}), h) + \lambda H(h|u)$$

- $H(h|u)$ measures (lack of) subj. explainability
- $h(x)$ similar for data points with similar user signal u
- EERM design choices: \mathcal{H} and loss L

Regularization = Implicit Pruning!

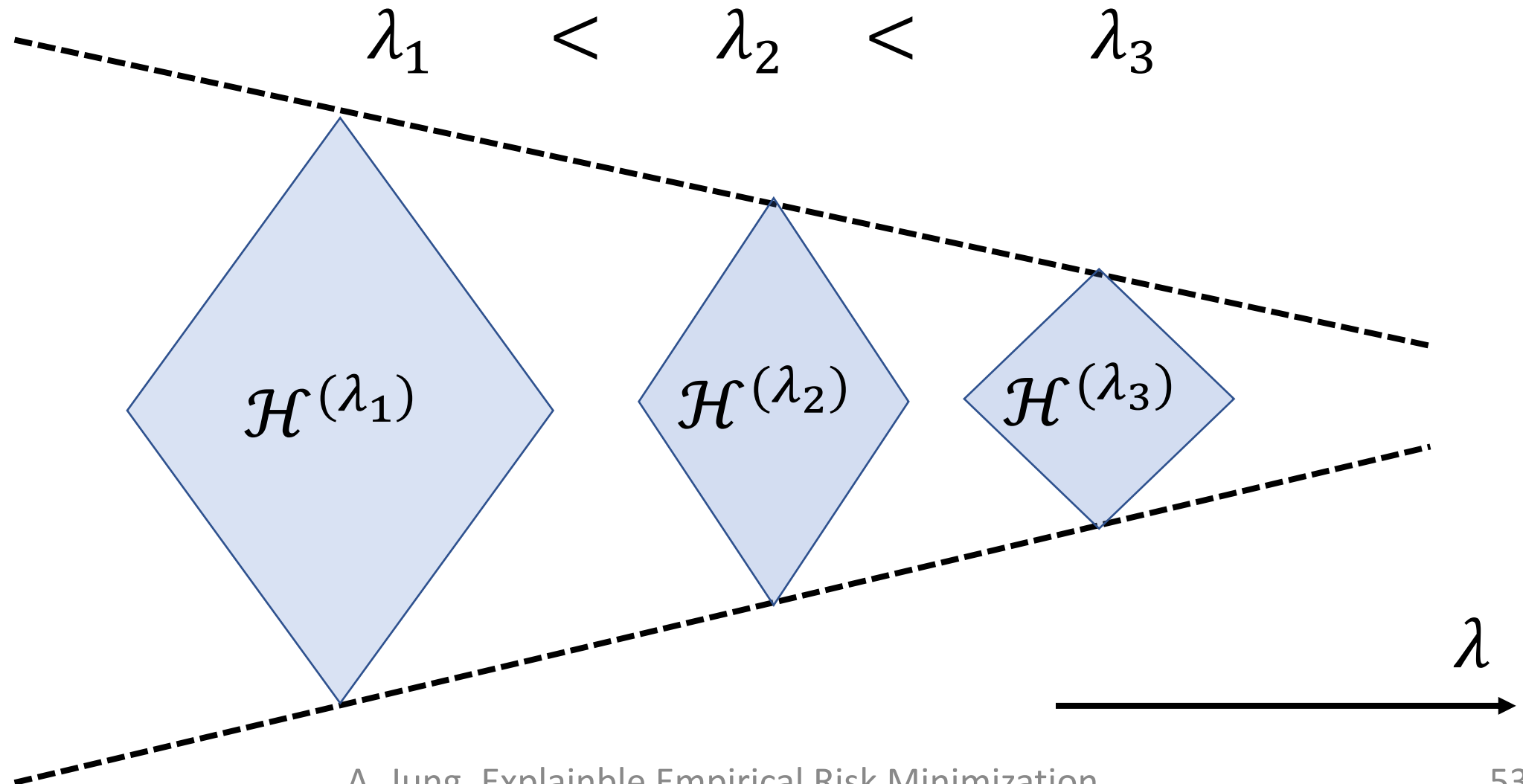
$$\min_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^m L((x^{(i)}, y^{(i)}), h) + \lambda H(h|u)$$

equivalent to

$$\min_{h \in \mathcal{H}^{(\lambda)}} \frac{1}{m} \sum_{i=1}^m L((x^{(i)}, y^{(i)}), h)$$

with pruned (interpretable?) model $\mathcal{H}^{(\lambda)} \subset \mathcal{H}$

EERM = Soft Interpret. Model Selection



Explainable Linear Regression

Algorithm 1 Explainable Linear Regression

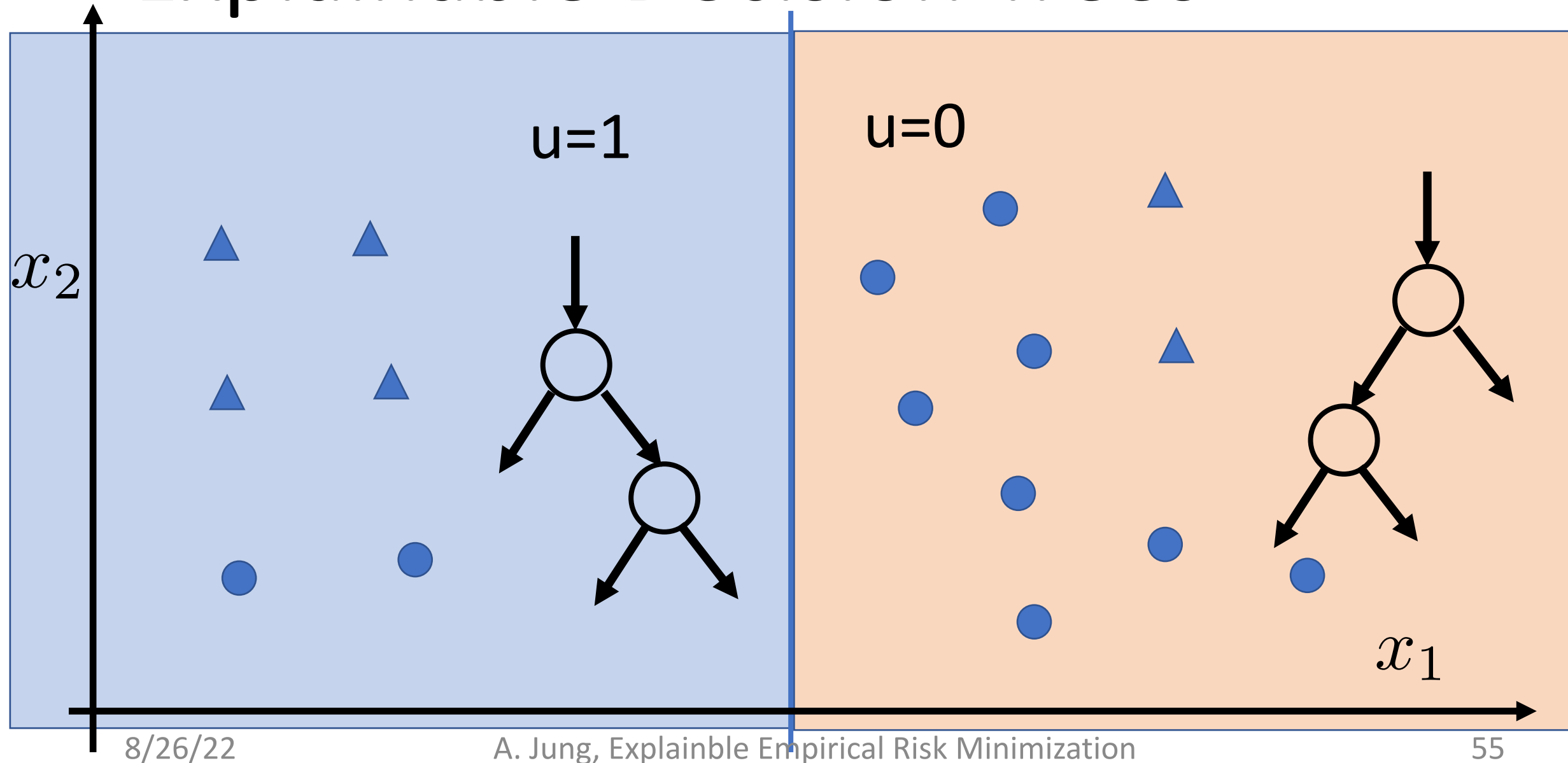
Input: explainability parameter λ , training set \mathcal{D} (see (5))

1: solve

$$\begin{aligned} \hat{\mathbf{w}} \in \operatorname{argmin}_{\alpha \in \mathbb{R}, \mathbf{w} \in \mathbb{R}^n} & \sum_{i=1}^m \underbrace{\left(y^{(i)} - \mathbf{w}^T \mathbf{x}^{(i)} \right)^2}_{\text{empirical risk}} \\ & + \lambda \underbrace{\left(\mathbf{w}^T \mathbf{x}^{(i)} - \alpha u^{(i)} \right)^2}_{\text{subjective explainability}} \end{aligned} \quad (19)$$

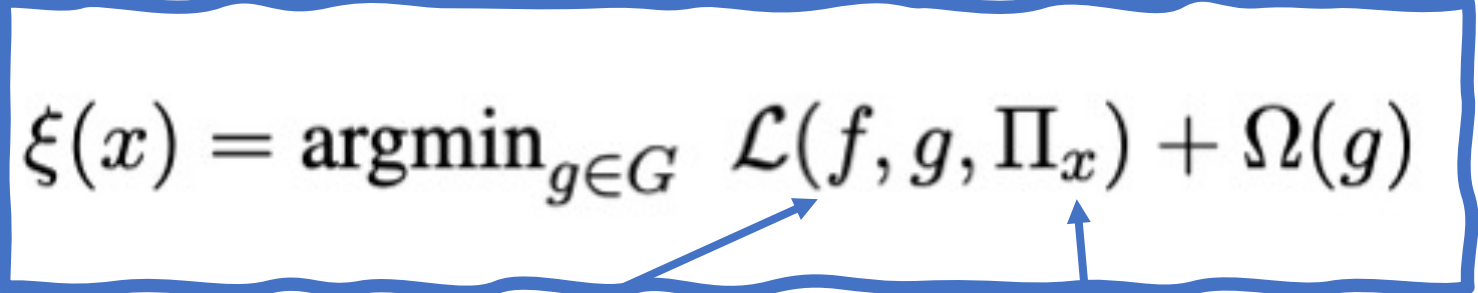
Output: $h^{(\lambda)}(\mathbf{x}) := \mathbf{x}^T \hat{\mathbf{w}}$

Explainable Decision Trees



EERM vs. LIME

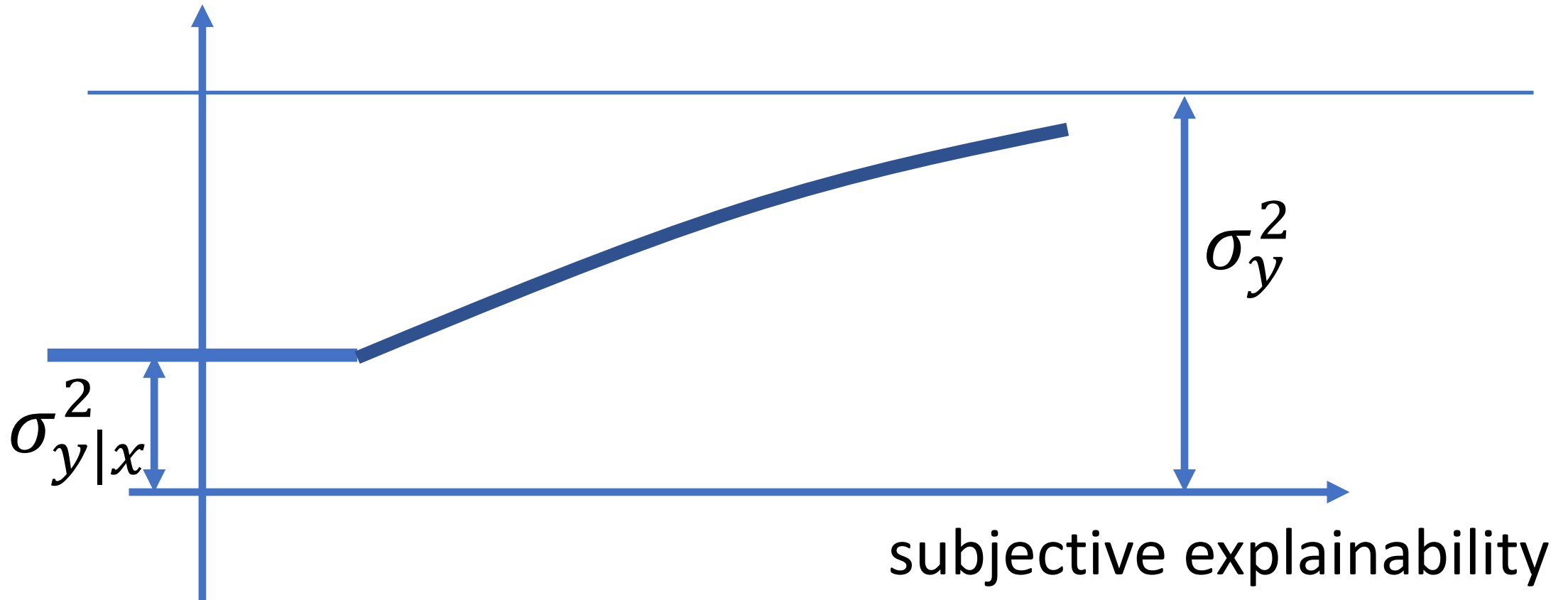
$$\min_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^m L((x^{(i)}, y^{(i)}), h) + \lambda H(h|u)$$


$$\xi(x) = \operatorname{argmin}_{g \in G} \mathcal{L}(f, g, \Pi_x) + \Omega(g)$$

- EERM and LIME essentially solve a regularized ERM
- LIME solves separate regularized ERM for each feature value x
- “empirical risk” in LIME based on faithfulness to given ML method

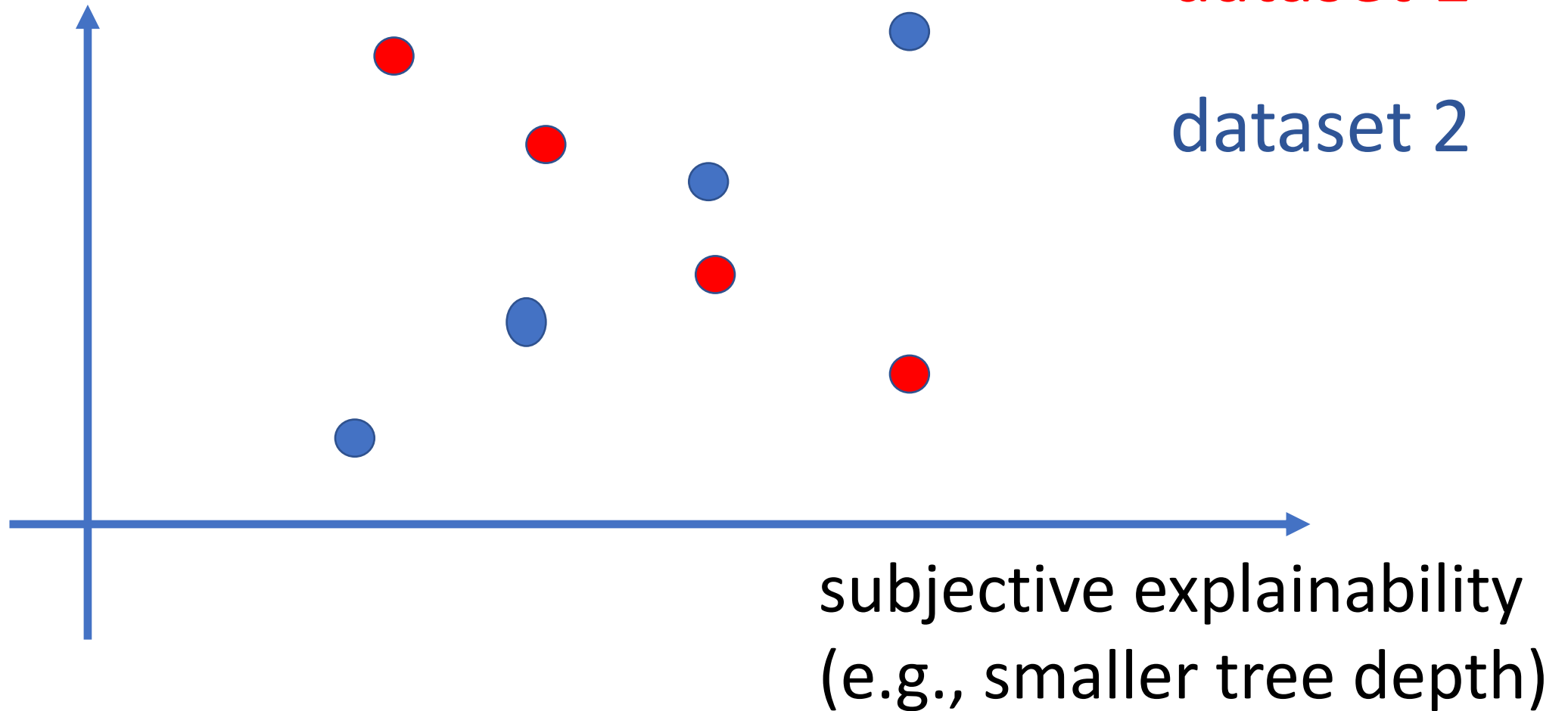
Explainability vs. Risk - Ideal

risk (expected loss)



Explainability vs. Risk - Practical

risk (expected loss)



To Summarize ...

- identify explainability with predictability
- subj. expl. = conditional entropy of predictions
- require user signal to define “subjective”
- EERM uses subj. explain. to regularize ERM
- special case: expl. lin.reg and expl. decision trees

EERM for Trustworthy AI

- **Human agency and oversight**
 - **Technical robustness and safety**
 - **Privacy and data governance**
 - **Transparency**
 - **Diversity, non-discrimination and fairness**
 - **Societal and environmental wellbeing**
 - **Accountability**
- $+ \lambda H(h|u)$
- <https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html>

References

- W.J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, “Definitions, methods, and applications in interpretable machine learning”, PNAS, Vol. 116, No. 44, 2019
- M. T. Ribeiro, S. Singh, and C. Guestrin.. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. SIGKDD, 2016.
- AJ and P. Nardelli, "An Information-Theoretic Approach to Personalized Explainable Machine Learning," in *IEEE Signal Processing Letters*, vol. 27, pp. 825-829, 2020, doi: 10.1109/LSP.2020.2993176.
- L. Zhang, G. Karakasidis, A. Odnoblyudova, L. Dogruel, AJ, “Explainable Empirical Risk Minimization”, 2020. <https://arxiv.org/abs/2009.01492>

References (ctd)

- Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, Jonathan K. Su “This Looks Like That: Deep Learning for Interpretable Image Recognition”, Neurips 2019
- Lamy et.al., “Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach,” Artificial Intelligence in Medicine, Volume 94, 2019.
- AJ, “Machine Learning: The Basics,” Springer, Singapore, 2022.