The opportunities of Machine Learning in Education

Adrián Pérez-Suay
Adrian.Perez@uv.es

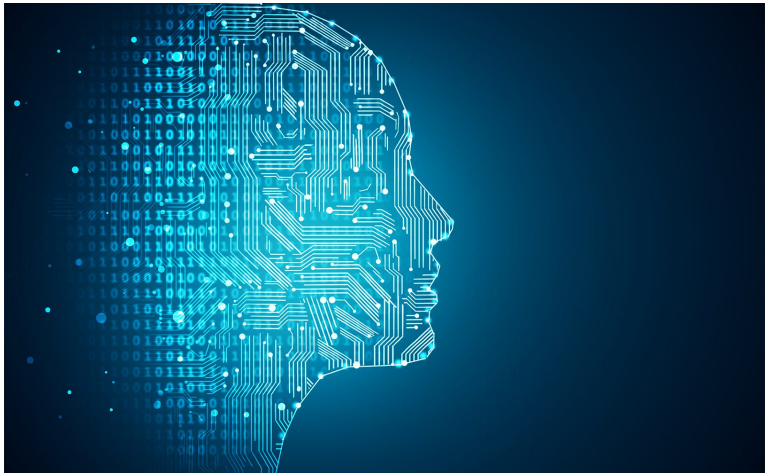
Universitat de València

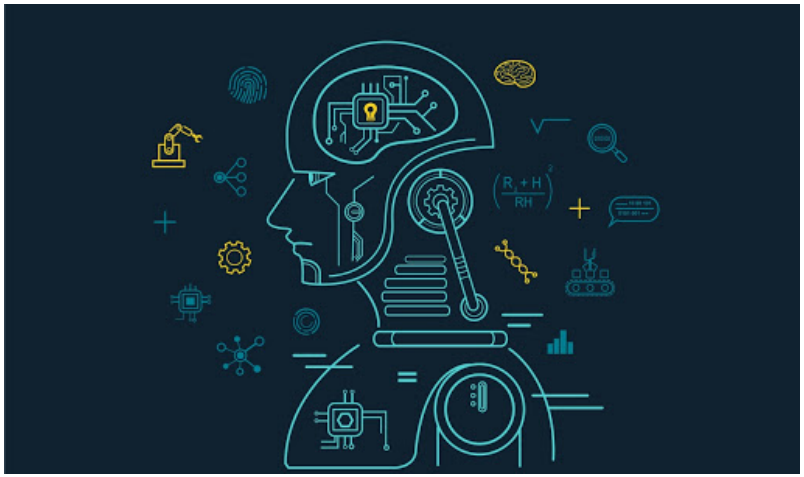
2020

Outline

Introduction

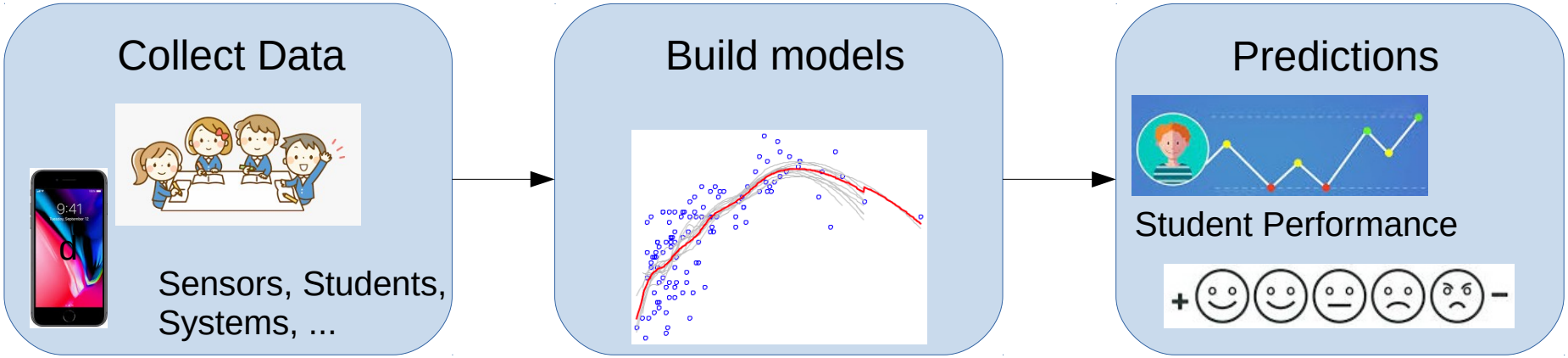

MACHINE
LEARNING





Introduction

Introduction



- Machine Learning (ML) is a wider part of Artificial Intelligence
- ML is an active field solving really challenging problems:
  - autonomous driving car, face recognition, natural language processing, ...
- ML builds models based on data
- Models are able to infer: regression, classification, fairness, causal inference, dimensionality reduction (PCA), ...

Regression

Regression

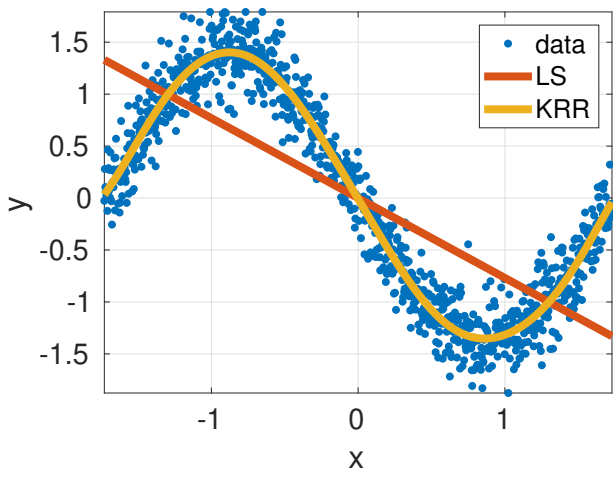
$$\mathcal{L} = \frac{1}{n} \sum_{i=1}^{n} V(f(\mathbf{x}_i), \mathbf{y}_i) + \lambda \, \Omega(\|f\|)$$

- $\mathbf{x}_i \in \mathcal{X}, \mathbf{y}_i \in \mathbb{R}$
- Linear model: $f := \hat{\mathbf{Y}} = \mathbf{X}\mathbf{W}$
- L2 norm for the errors, $V := \|\mathbf{Y} - \hat{\mathbf{Y}}\|_2^2$
- Tikhonov's regularization for smoothness, $\Omega(\|f\|) := \|\mathbf{W}\|_2^2$
- Closed-form-solution $\widehat{\mathbf{W}} = (\mathbf{X}^\top \mathbf{X} + \lambda \, \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{Y}$
- Prediction $\hat{\mathbf{Y}}_* = \mathbf{X}_* \widehat{\mathbf{W}}$

Regression: Kernel Ridge Regression

$$\mathcal{L} = \frac{1}{n}\sum_{i=1}^{n} V(f(\mathbf{x}_i), \mathbf{y}_i) + \lambda\ \Omega(\|f\|_{\mathcal{H}})$$

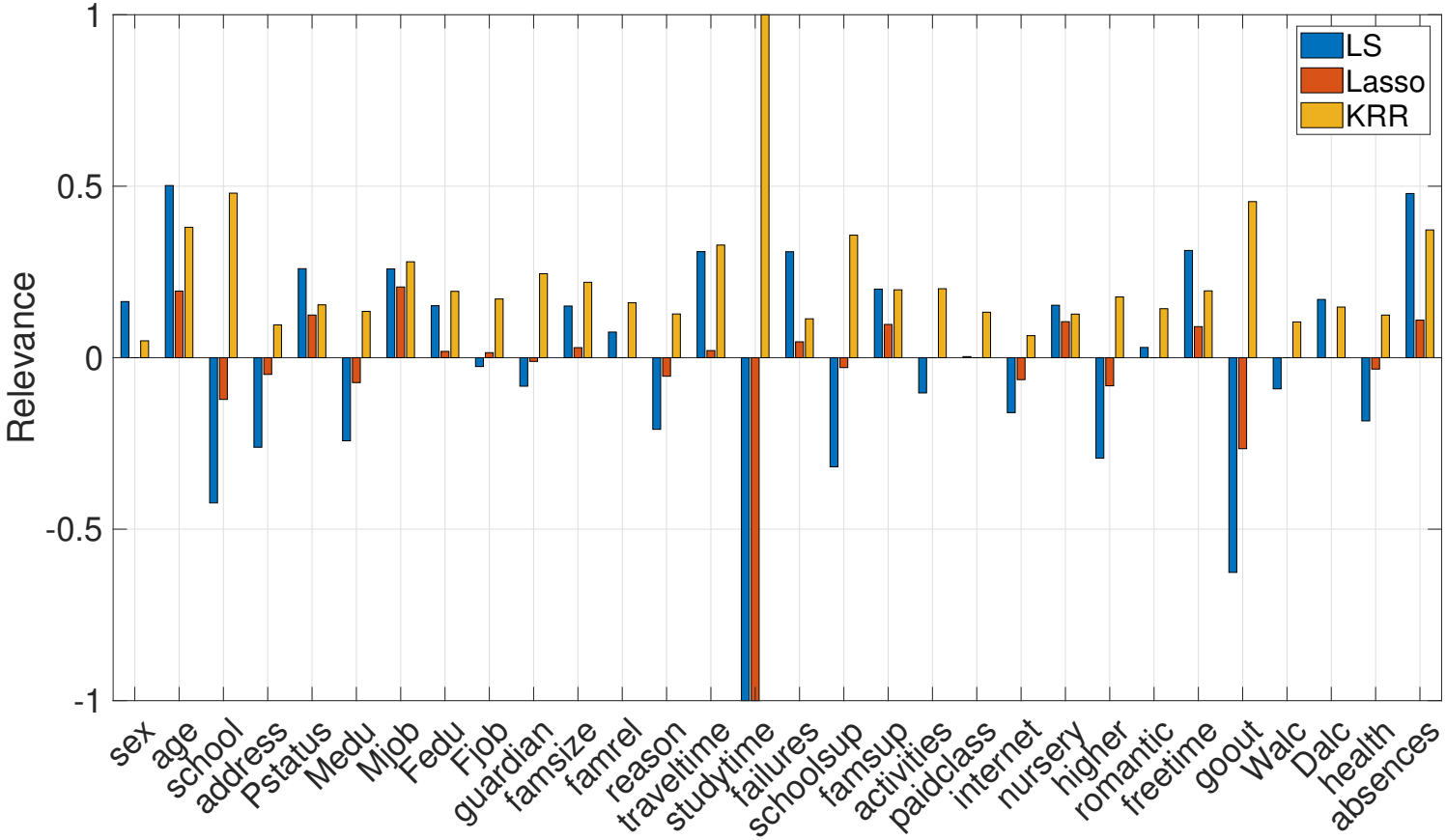
- Map to RKHS $\phi : \mathcal{X} \to \mathcal{H}$, $\mathbf{x}_i \to \phi(\mathbf{x}_i)$
- $k(\mathbf{x}, \mathbf{x}') = \langle \phi(x), \phi(x') \rangle$ E.g.
  - $k(\mathbf{x}_i, \mathbf{x}_j) = e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|_2^2}{2\sigma^2}}$ (RBF kernel)
  - $k(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^\top \mathbf{x}_j + c)^d$ (Polynomial)
  - Any others... must be PSD
- Representer Thm. [Schölkopf et al., 2001] $f(\cdot) = \sum_{i=1}^{n} \alpha_i k(\cdot, \mathbf{x}_i)$
- $\widehat{\mathbf{\Lambda}} = (\mathbf{K} + \lambda \mathbf{I})^{-1}\mathbf{Y}, \quad \hat{\mathbf{Y}}_* = \mathbf{K}(\mathbf{X}_*, \mathbf{X})\widehat{\mathbf{\Lambda}}$



Application: Students performance

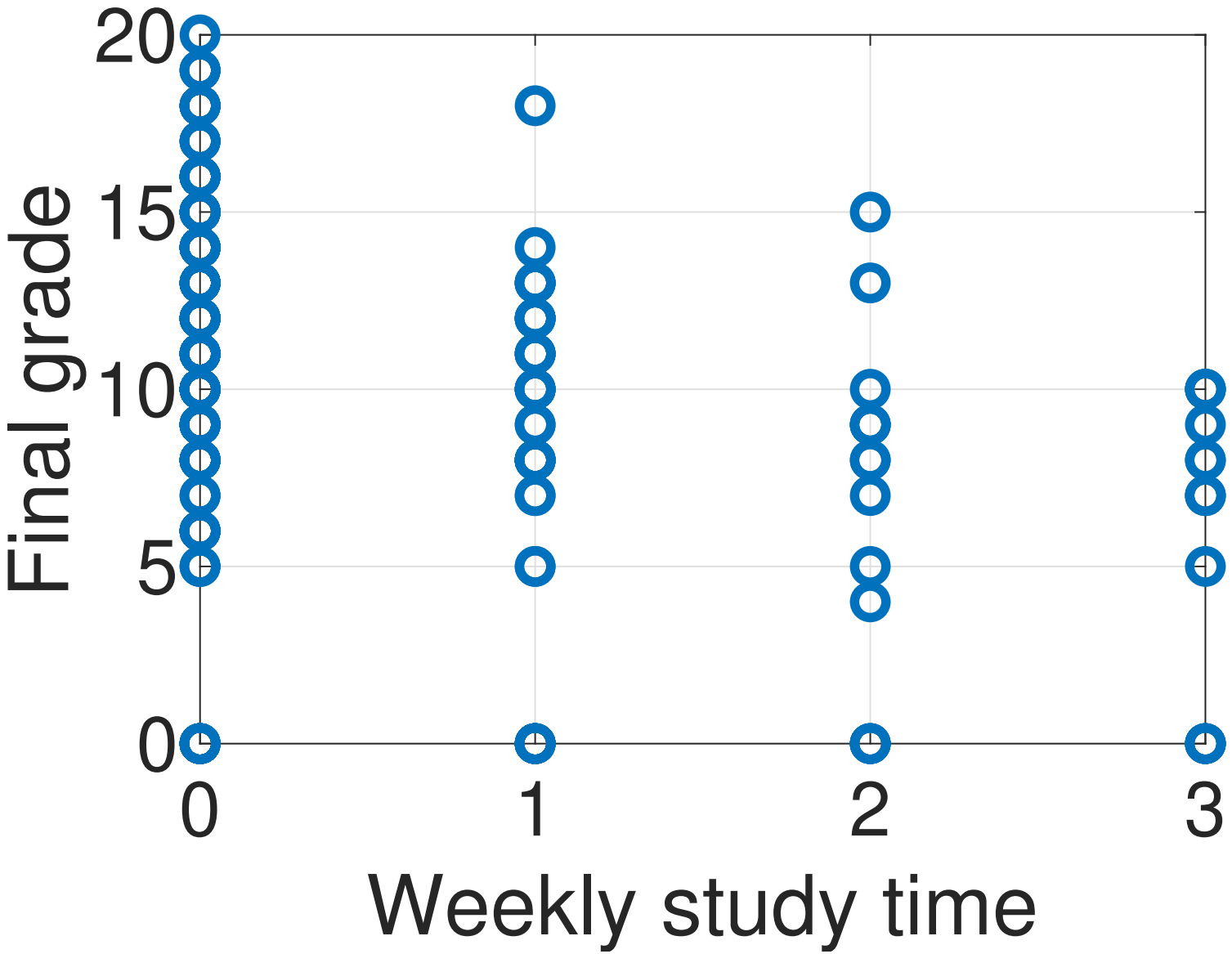
Attribute	Description (Domain)
sex	student's sex (binary: female or male)
age	student's age (numeric: from 15 to 22)
school	student's school (binary: *Gabriel Pereira* or *Mousinho da Silveira*)
address	student's home address type (binary: urban or rural)
Pstatus	parent's cohabitation status (binary: living together or apart)
Medu	mother's education (numeric: from 0 to 4[a])
Mjob	mother's job (nominal[b])
Fedu	father's education (numeric: from 0 to 4[a])
Fjob	father's job (nominal[b])
guardian	student's guardian (nominal: mother, father or other)
famsize	family size (binary: ≤ 3 or > 3)
famrel	quality of family relationships (numeric: from 1 – very bad to 5 – excellent)
reason	reason to choose this school (nominal: close to home, school reputation, course preference or other)
traveltime	home to school travel time (numeric: 1 – $<$ 15 min., 2 – 15 to 30 min., 3 – 30 min. to 1 hour or 4 – $>$ 1 hour).
studytime	weekly study time (numeric: 1 – $<$ 2 hours, 2 – 2 to 5 hours, 3 – 5 to 10 hours or 4 – $>$ 10 hours)
failures	number of past class failures (numeric: n if $1 \leq n < 3$, else 4)
schoolsup	extra educational school support (binary: yes or no)
famsup	family educational support (binary: yes or no)
activities	extra-curricular activities (binary: yes or no)
paidclass	extra paid classes (binary: yes or no)
internet	Internet access at home (binary: yes or no)
nursery	attended nursery school (binary: yes or no)
higher	wants to take higher education (binary: yes or no)
romantic	with a romantic relationship (binary: yes or no)
freetime	free time after school (numeric: from 1 – very low to 5 – very high)
goout	going out with friends (numeric: from 1 – very low to 5 – very high)
Walc	weekend alcohol consumption (numeric: from 1 – very low to 5 – very high)
Dalc	workday alcohol consumption (numeric: from 1 – very low to 5 – very high)
health	current health status (numeric: from 1 – very bad to 5 – very good)

Application: Students performance

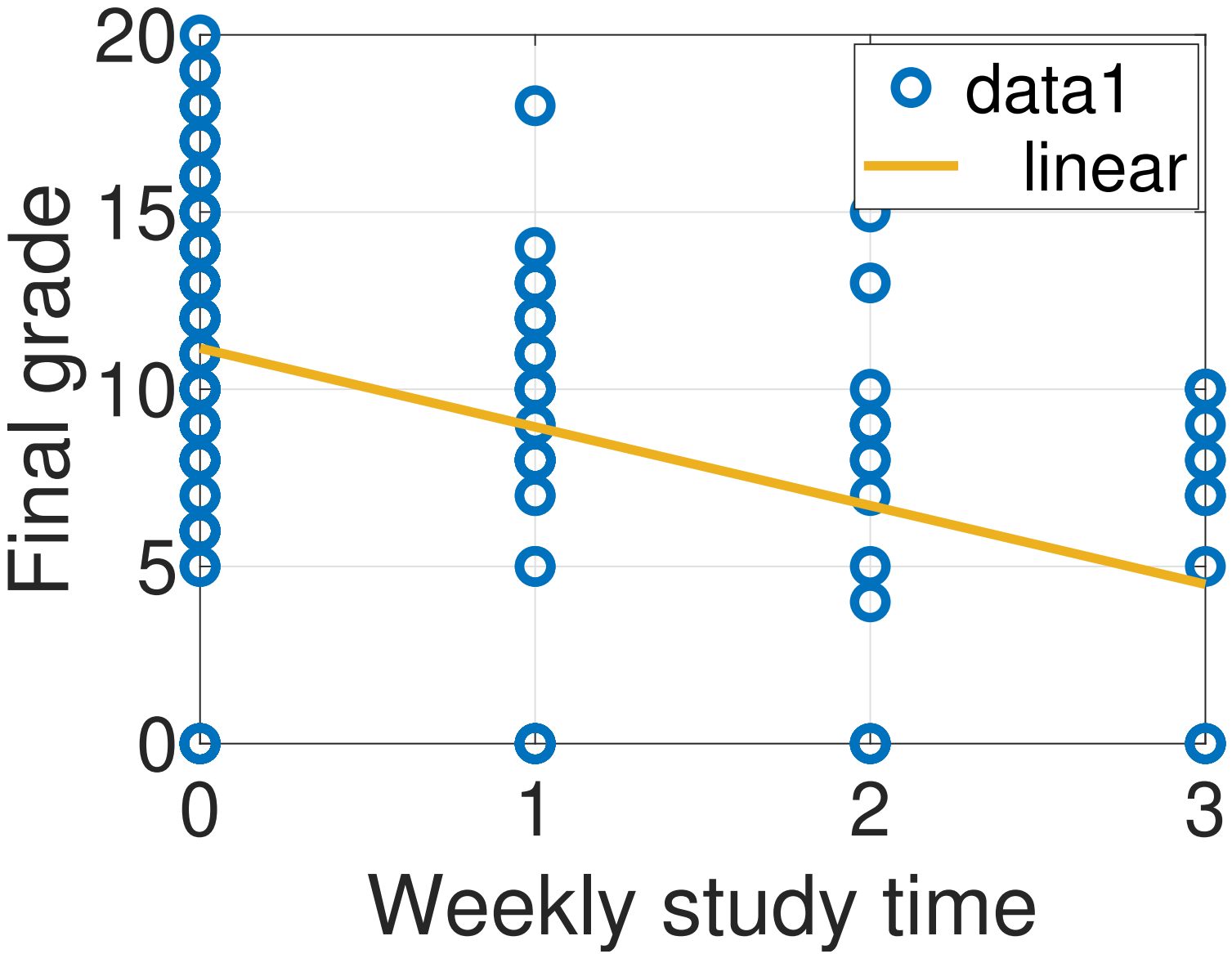


Method	RMSE
LS	4.42
LASSO	4.40
KRR	4.33

Application: Students performance



Application: Students performance



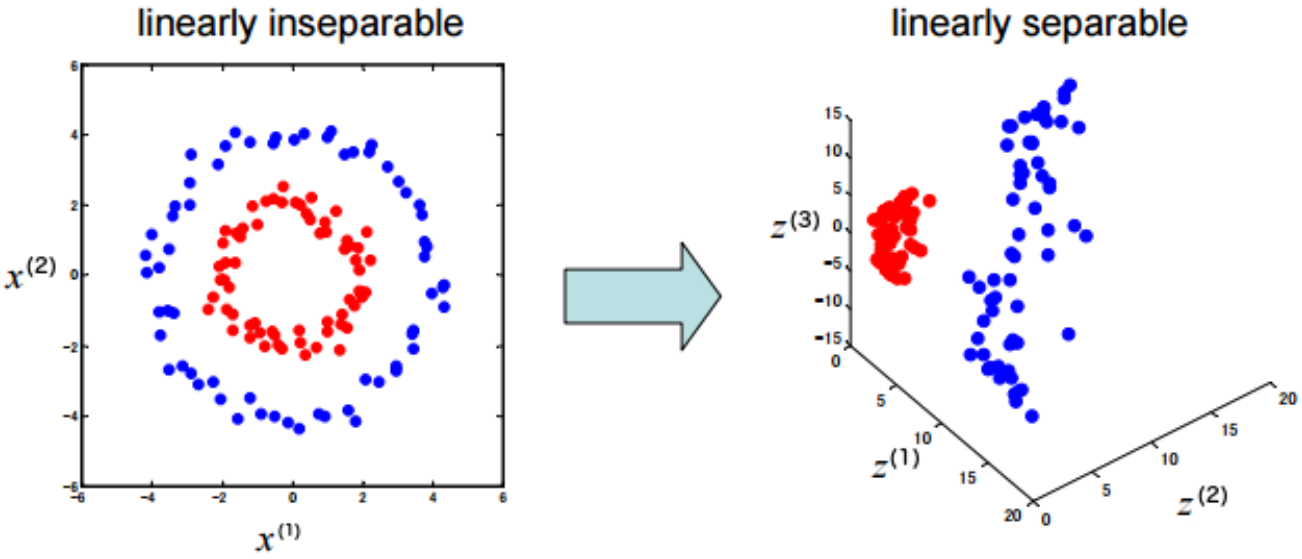
Classification

Classification

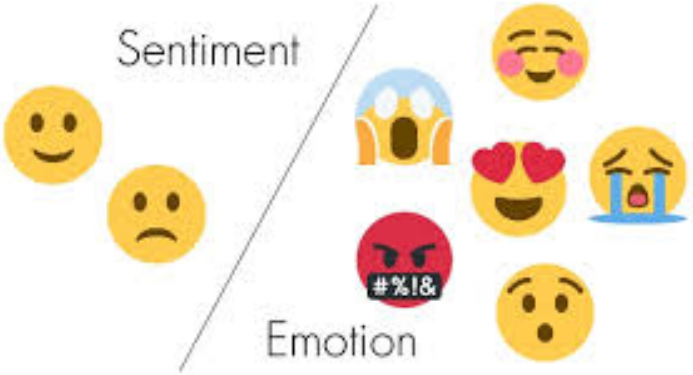
$$\mathcal{L} = \frac{1}{n}\sum_{i=1}^{n} V(f(\mathbf{x}_i), \mathbf{y}_i) + \lambda\ \Omega(\|f\|_{\mathcal{H}})$$

- Now $\mathbf{y}_i \in \{-1, 1\}$
- Map to RKHS $\phi : \mathcal{X} \to \mathcal{H}$, $\mathbf{x}_i \to \phi(\mathbf{x}_i)$
- $k(\mathbf{x}, \mathbf{x}') = \langle \phi(x), \phi(x') \rangle$ E.g.
- Representer Thm. [Schölkopf et al., 2001]
  $f(\cdot) = \sum_{i=1}^{n} \alpha_i k(\cdot, \mathbf{x}_i)$
- $\widehat{\mathbf{\Lambda}} = (\mathbf{K} + \lambda \mathbf{I})^{-1}\mathbf{Y}, \quad \hat{\mathbf{Y}}_* = \mathrm{sgn}(\mathbf{K}(\mathbf{X}_*, \mathbf{X})\widehat{\mathbf{\Lambda}})$

Classification

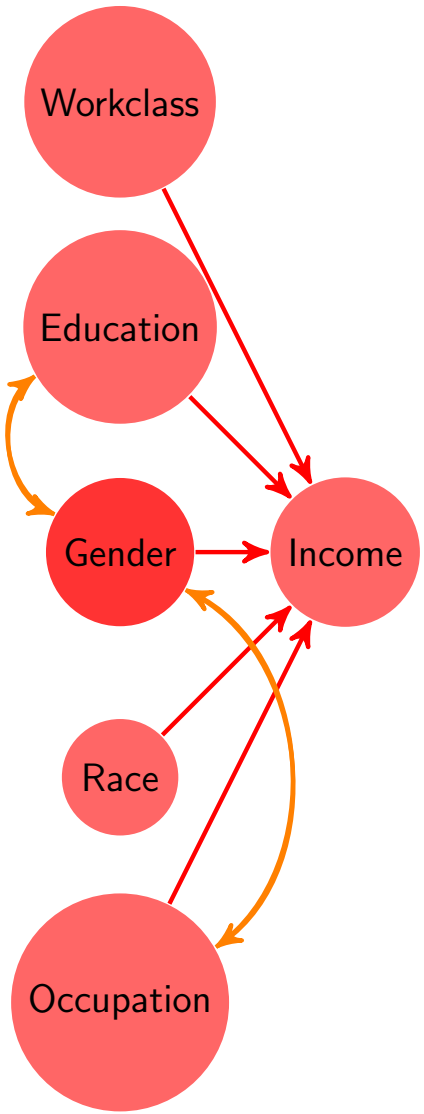


- Sentiment classification



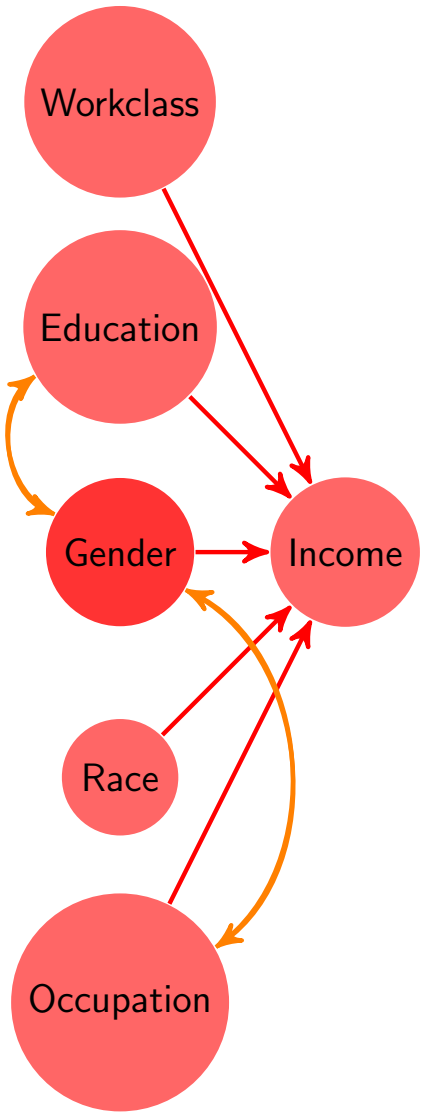
Fair Learning

Fair Learning



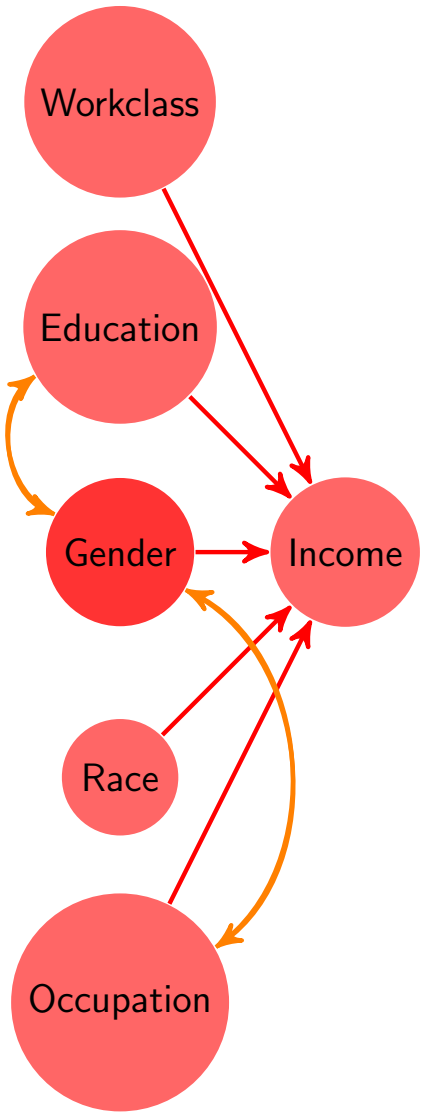
- Let's predict the income from some reasonable covariates
- Our company wants to be fair with the gender
- Removing the gender variable does not solve the problem
- Gender information is contained in other variables implicitly

Fair Learning



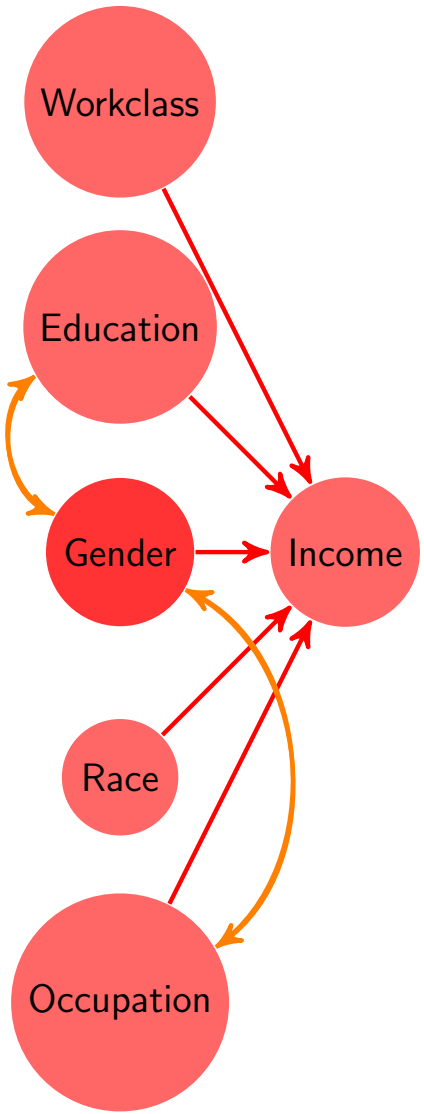
- Let's predict the income from some reasonable covariates
- Our company wants to be fair with the gender
- Removing the gender variable does not solve the problem
- Gender information is contained in other variables implicitly
- **How to avoid this omitted variable bias problem?**

Fair learning setup



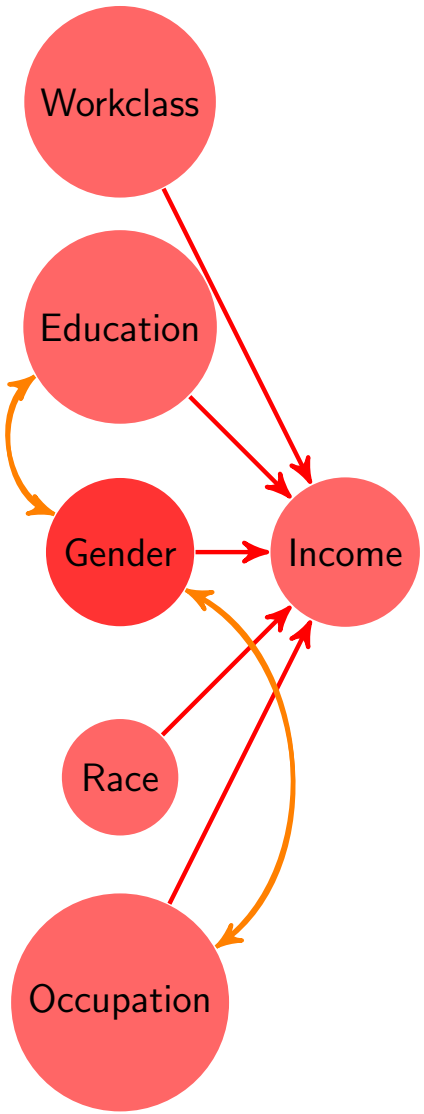
- **Goal:** Respect rules, ethics, laws; avoid disparate treatment

Fair learning setup



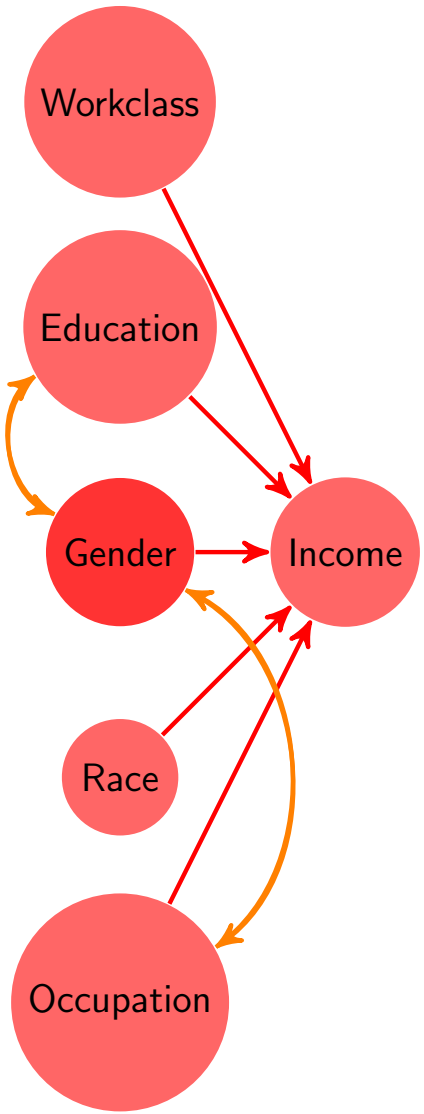
- **Goal:** Respect rules, ethics, laws; avoid disparate treatment
- **Premise:** Protected variables are worth including them

Fair learning setup



- **Goal:** Respect rules, ethics, laws; avoid disparate treatment
- **Premise:** Protected variables are worth including them
- **Definition**:
  "A prediction is said to be totally fair with respect to the sensitive features $\mathbf{S}$ *if and only if* $\widehat{\mathbf{Y}} \perp \mathbf{S}$*."*

Fair learning setup



- **Goal:** Respect rules, ethics, laws; avoid disparate treatment
- **Premise:** Protected variables are worth including them
- **Definition**:
  "A prediction is said to be totally fair with respect to the sensitive features $\mathbf{S}$ *if and only if* $\widehat{\mathbf{Y}} \perp \mathbf{S}$*."*
- **Idea:** Be accurate and insensitive to protected variables

Regularization framework for fair prediction

$$\mathcal{L} = \frac{1}{n} \sum_{i=1}^{n} V(f(\mathbf{x}_i), \mathbf{y}_i) + \lambda \; \Omega(\|f\|_{\mathcal{H}}) + \mu \; I(f(\mathbf{x}), \mathbf{s})$$

Regularization framework for fair prediction

$$\mathcal{L} = \frac{1}{n}\sum_{i=1}^{n} V(f(\mathbf{x}_i), \mathbf{y}_i) + \lambda\ \Omega(\|f\|_{\mathcal{H}}) + \mu\ \mathbf{HSIC}(\mathbf{f}(\mathbf{x}), \mathbf{s})$$

- Linear model: $f := \hat{\mathbf{Y}} = \mathbf{XW}$
- L2 norm for the errors, $V := \|\mathbf{Y} - \hat{\mathbf{Y}}\|_2^2$
- Tikhonov's regularization for smoothness, $\Omega(\|f\|^2) := \|\mathbf{W}\|_2^2$
- Estimate independence $\mathbf{I} = \mathbf{HSIC}$

Regularization framework for fair prediction

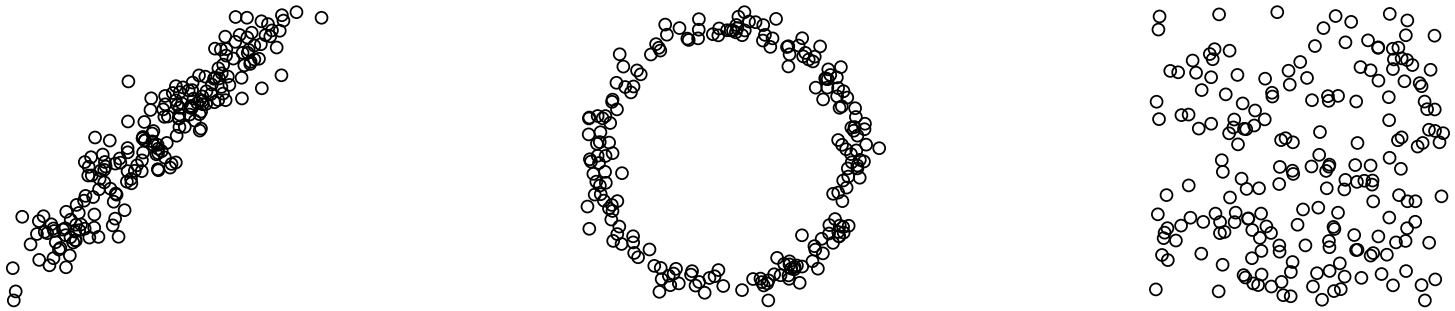
$$\mathcal{L} = \frac{1}{n}\sum_{i=1}^{n} V(f(\mathbf{x}_i), \mathbf{y}_i) + \lambda\, \Omega(\|f\|_{\mathcal{H}}) + \mu\, \mathbf{HSIC}(\mathbf{f}(\mathbf{x}), \mathbf{s})$$

- Linear model: $f := \hat{\mathbf{Y}} = \mathbf{XW}$
- L2 norm for the errors, $V := \|\mathbf{Y} - \hat{\mathbf{Y}}\|_2^2$
- Tikhonov's regularization for smoothness, $\Omega(\|f\|^2) := \|\mathbf{W}\|_2^2$
- Estimate independence $\mathbf{I} = \mathbf{HSIC}$

Proposal: HSIC [Pérez-Suay et al., 2017]

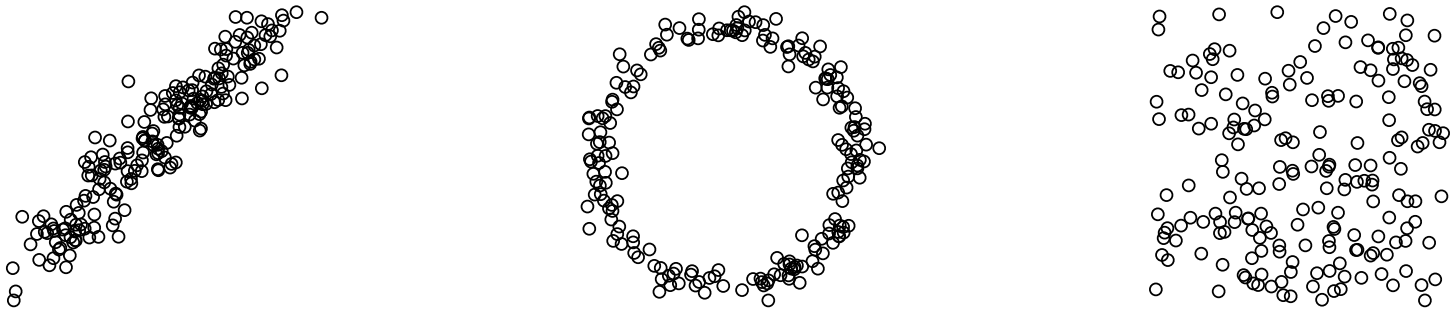
- ✓ Differentiable (allows closed-form-solutions)
- ✓ Captures higher order relations
- ✓ Multidimensional: $d_s, d_f \geq 1$
- ✓ Allows: $d_s \neq d_f$
- ✓ Easy implementation!

Dependence estimation



Pearson	0.94	0.04	-0.004

Dependence estimation



Pearson	0.94	0.04	-0.004
HSIC	0.09	0.03	0.002

Measuring dependence with kernels

- Define mapping functions $\phi : \mathcal{X} \to \mathcal{F}$ and $\psi : \mathcal{Y} \to \mathcal{G}$
- Define positive definite kernel functions:
  $k(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle$ and $l(\mathbf{y}_i, \mathbf{y}_j) = \langle \psi(\mathbf{y}_i), \psi(\mathbf{y}_j) \rangle$
- Mapped data: $\mathbf{X} \to \Phi \in \mathbb{R}^{n \times n_{\mathcal{F}}}$ and $\mathbf{Y} \to \Psi \in \mathbb{R}^{n \times n_{\mathcal{G}}}$
- Cross-covariance between mapped data: $\mathcal{C}_{\phi(x)\psi(y)} = \Phi^\top \Psi$
- The Hilbert-Schmidt norm is now:

$$\text{HSIC} = \|\Phi^\top \Psi\|^2 = \frac{1}{n^2} \text{Tr}((\Phi^\top \Psi)^\top (\Phi^\top \Psi)) = \frac{1}{n^2} \text{Tr}(\Phi \Phi^\top \Psi \Psi^\top) = \frac{1}{n^2} \text{Tr}(\mathbf{KL})$$

[Gretton et al., 2005]

Regularization framework for fair prediction

$$\mathcal{L} = \|\mathbf{Y} - \hat{\mathbf{Y}}\|_2^2 + \lambda \, \|\mathbf{W}\|_2^2 + \mu \frac{1}{n^2} \mathrm{tr}(\hat{\mathbf{Y}}\hat{\mathbf{Y}}^\top \mathbf{S}\mathbf{S}^\top)$$

- Linear fair regression:

$$\widehat{\mathbf{W}} = (\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} + \lambda \, \mathbf{I} + \frac{\mu}{n^2} \, \tilde{\mathbf{X}}^\top \tilde{\mathbf{S}} \tilde{\mathbf{S}}^\top \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^\top \mathbf{Y}, \quad \hat{\mathbf{Y}}_* = \mathbf{X}_* \widehat{\mathbf{W}}$$

Regularization framework for fair prediction

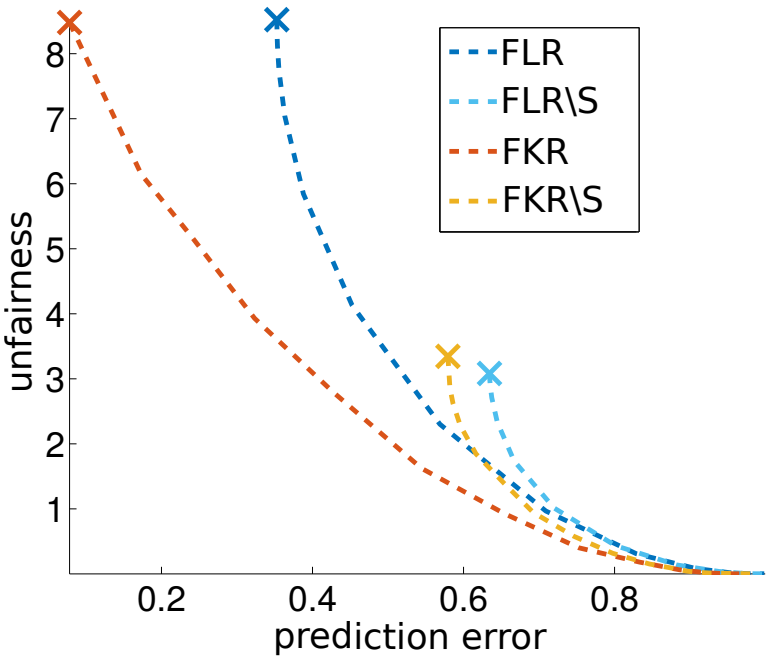
$$\mathcal{L} = \|\mathbf{Y} - \hat{\mathbf{Y}}\|_2^2 + \lambda \, \|\mathbf{\Phi}^\top \mathbf{\Lambda}\|_2^2 + \mu \frac{1}{n^2} \mathrm{tr}(\tilde{\mathbf{K}} \tilde{\mathbf{K}}_S)$$

- Linear fair regression:

$$\widehat{\mathbf{W}} = (\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} + \lambda \, \mathbf{I} + \frac{\mu}{n^2} \, \tilde{\mathbf{X}}^\top \tilde{\mathbf{S}} \tilde{\mathbf{S}}^\top \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^\top \mathbf{Y}, \quad \hat{\mathbf{Y}}_* = \mathbf{X}_* \widehat{\mathbf{W}}$$

- Kernel (nonlinear) fair regression:

$$\widehat{\mathbf{\Lambda}} = (\tilde{\mathbf{K}} + \lambda \mathbf{I} + \frac{\mu}{n^2} \tilde{\mathbf{K}} \tilde{\mathbf{K}}_S)^{-1} \mathbf{Y}, \quad \hat{\mathbf{Y}}_* = \mathbf{K}(\mathbf{X}_*, \mathbf{X}) \widehat{\mathbf{\Lambda}}$$



- unfairness $= \mathrm{HSIC}(\hat{\mathbf{Y}}, \mathbf{S})$
- Removing the sensitive variable is worse
- Kernel better in accuracy than linear
- Fairness-accuracy nice tradeoffs

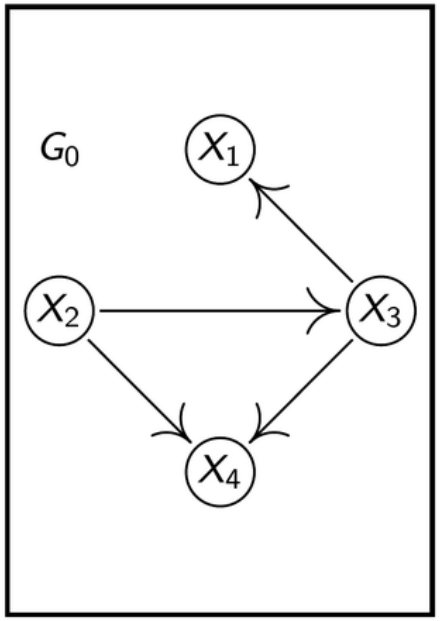
Causal Inference

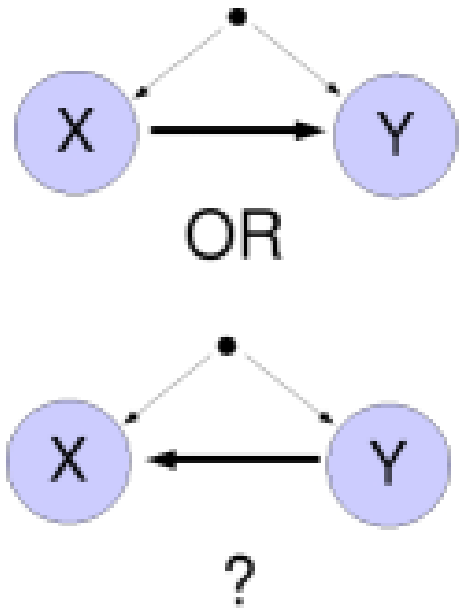
Causal inference ANM

[Hoyer08] scheme in Additive Noise Model (ANM)
ANM is a particular case of Structural Equation Models

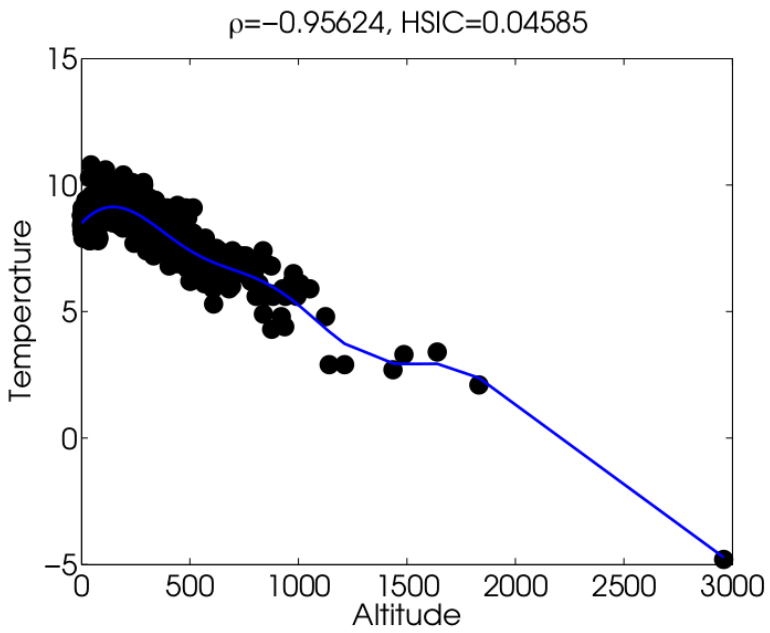
$$X_1 = f_1(X_3) + N_1$$
$$X_2 = N_2$$
$$X_3 = f_3(X_2) + N_3$$
$$X_4 = f_4(X_2, X_3) + N_4$$

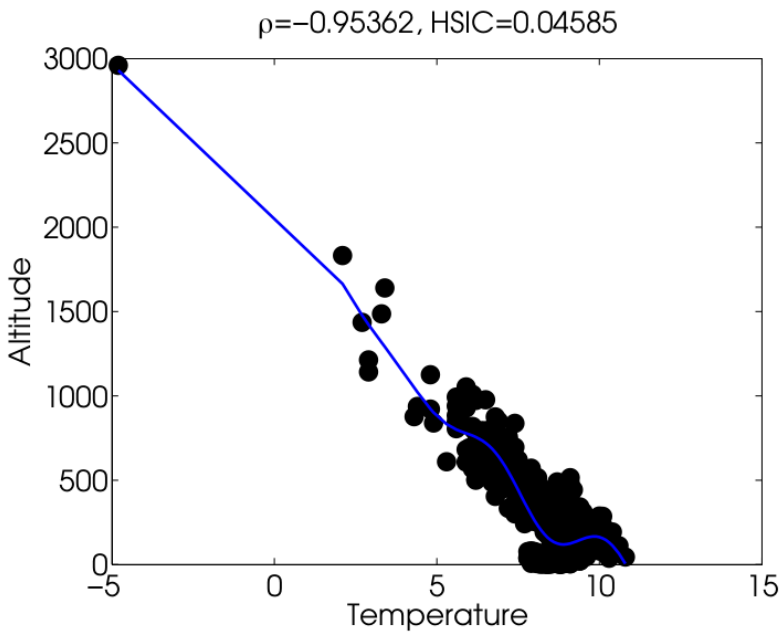
- N_i jointly independent
- G_0 has no cycles





Hoyer scheme



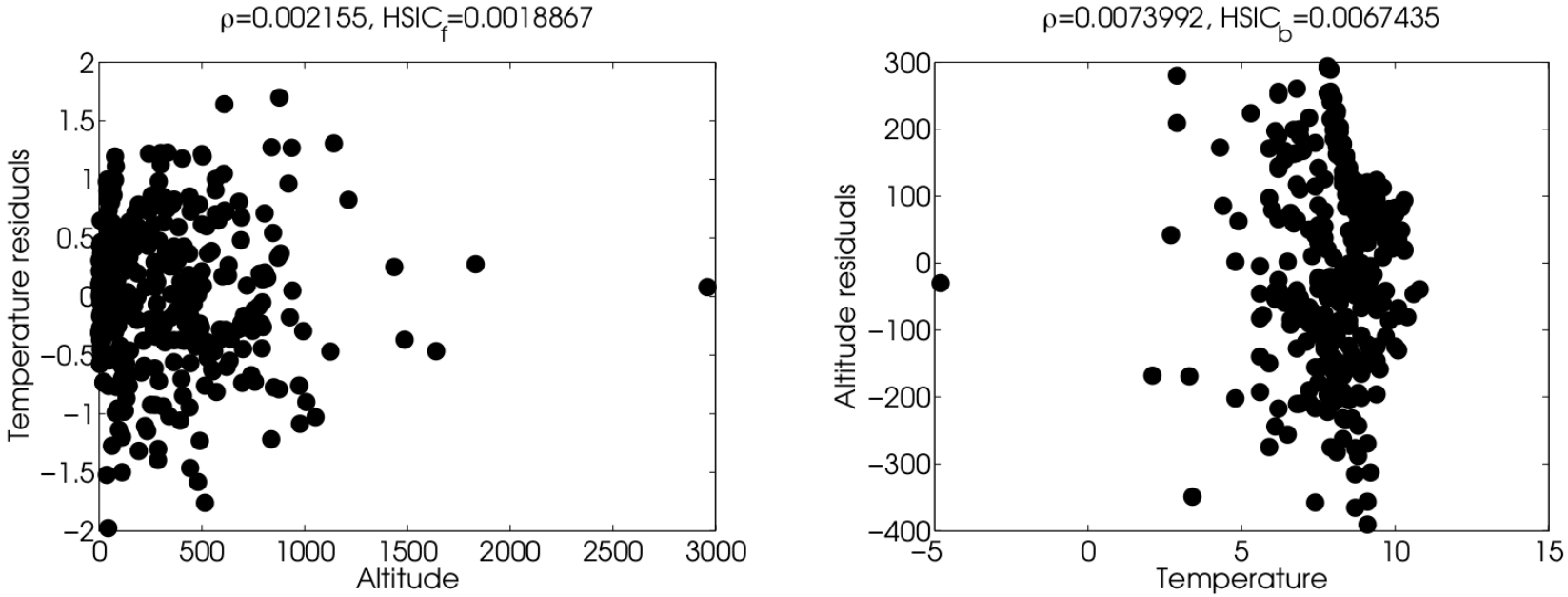


Step 1: build two models:

- Forward model: $\hat{y} = f(x)$
- Backward model: $\hat{x} = b(y)$

[Hoyer et al., 2009]

Hoyer scheme

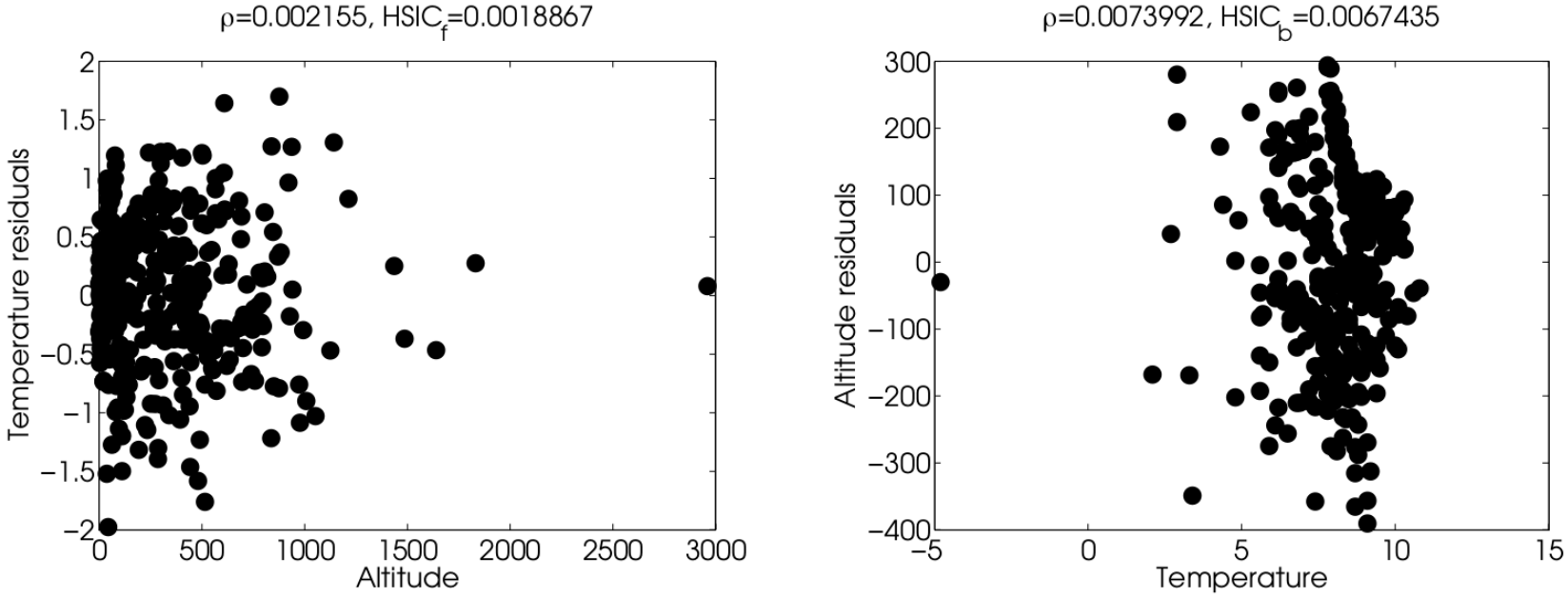


Step 2: check if residuals are independent of the inputs:

- Forward residuals: $(r_f = y - \hat{y}) \perp x$
- Backward residuals: $(r_b = x - \hat{x}) \perp y$

[Hoyer et al., 2009]

Hoyer scheme

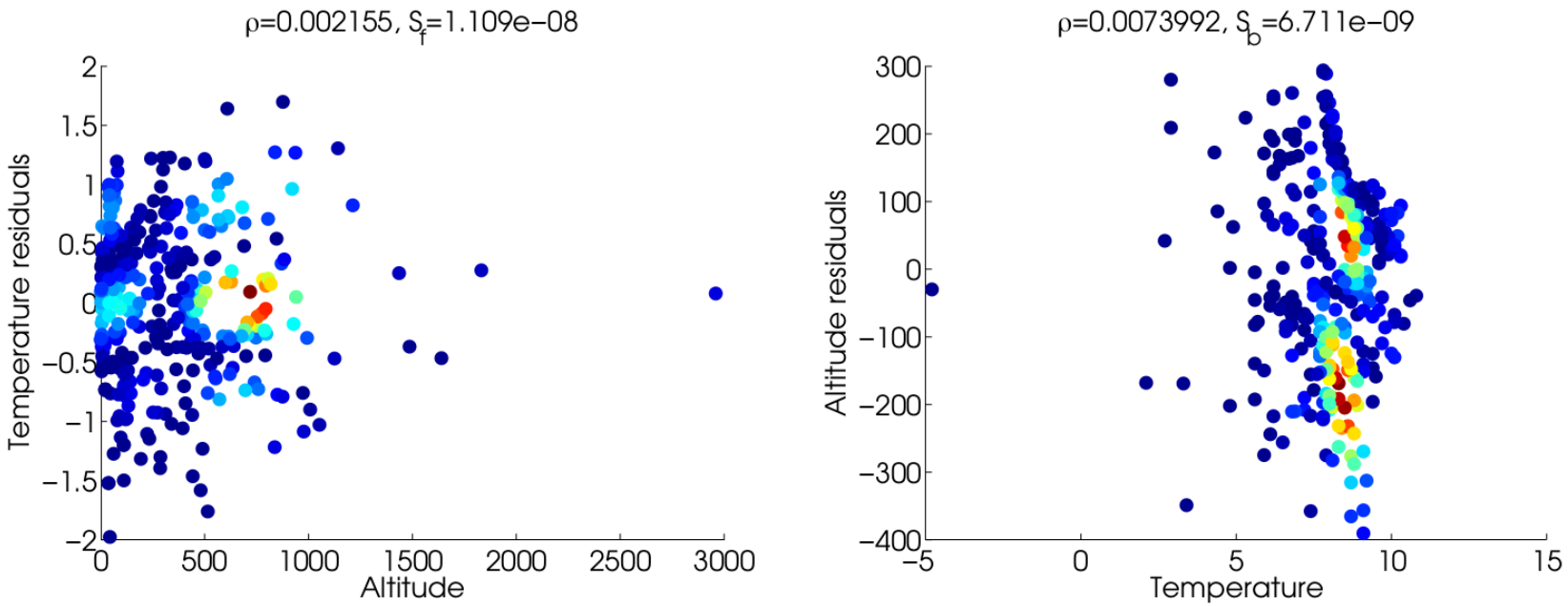


Step 3: the direction of causation is the most independent:

- Criterion: $\hat{C} := \mathrm{HSIC}(x, r_f) - \mathrm{HSIC}(y, r_b)$
- e.g.: If $\hat{C} < 0$ then $x \to y$

[Hoyer et al., 2009]

SHSIC



Step 4: the direction of causation is the most iid and sensitive:

- Criterion: $\hat{C}_s := (S_b^y + S_b^r) - (S_f^x + S_f^r)$
- e.g.: If $\hat{C} < 0$ then $x \to y$

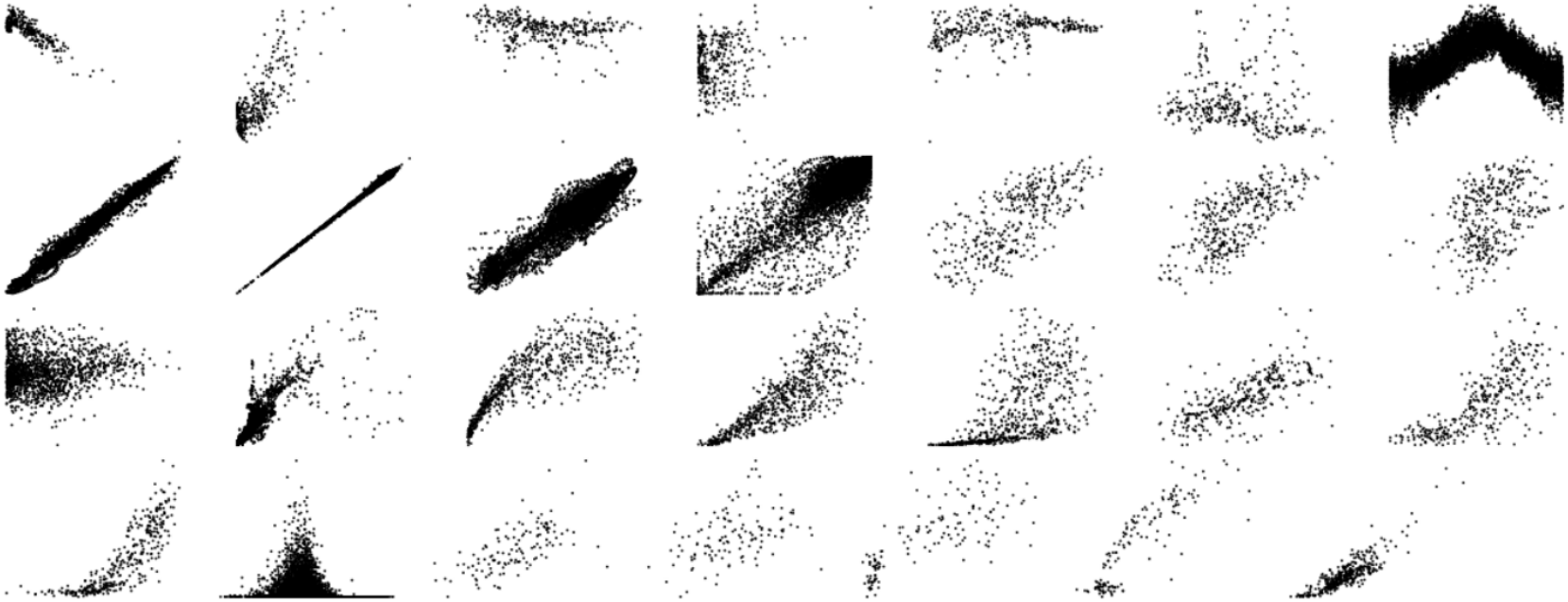
[Perez-Suay, 2016]

SHSIC

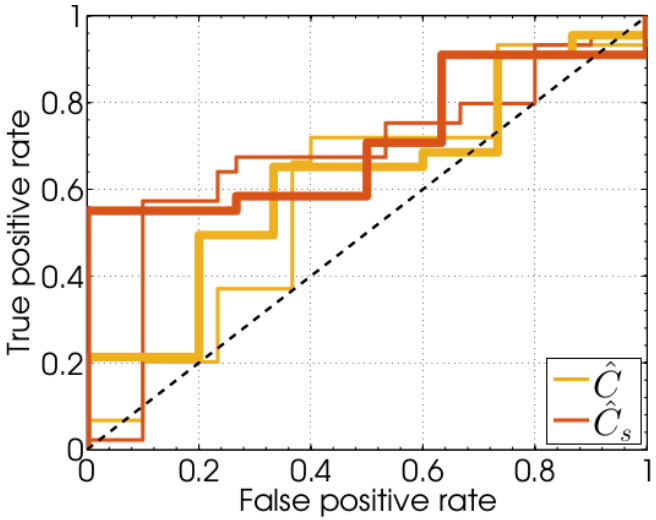
Cause Effect Pairs (CEP) contains annotated 100 pairs[1]

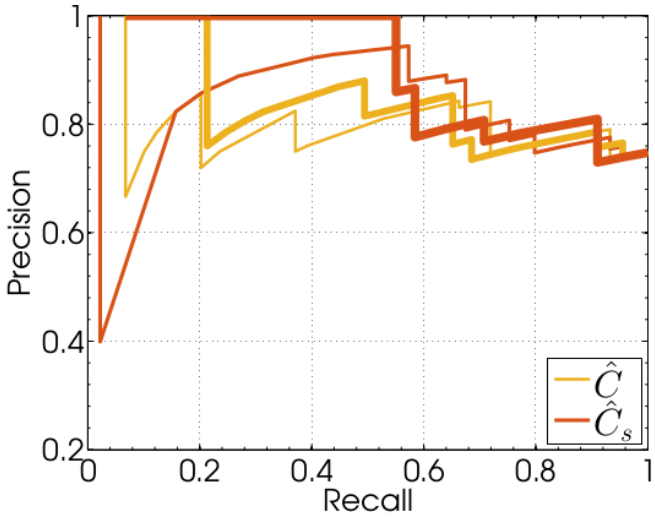
Unidimensional and GRS variables only (28 out of 100)

Random forests for regression, HSIC for dep.estim.

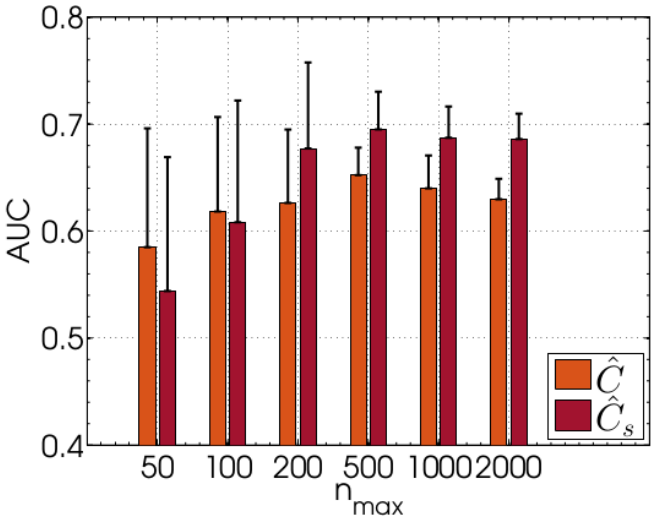


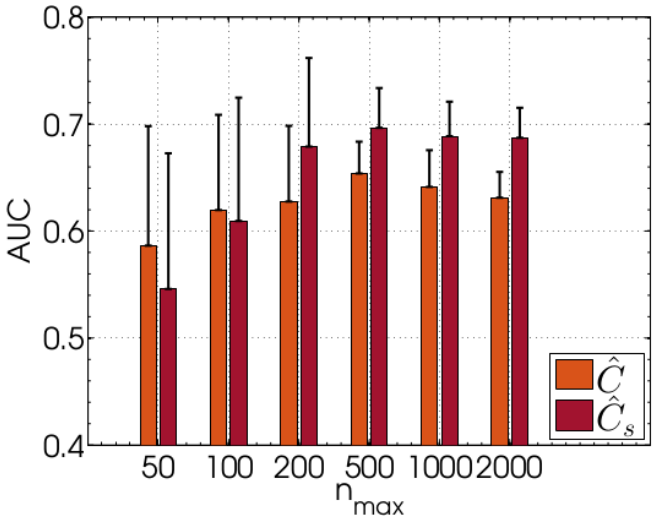
SHSIC





Robustness to few samples





Applications

- Collect Educational data with known causal directions
- Apply developed methods
- Develop novel methods to solve particular problems

Concluding remarks

Conclusions

- 4 different paradigms in ML (regression, classification, fairness, causality)
- There are others which we can explore (depending on the problem)
- Main research lines:
  - Collect data (and share it), databases of meaningful ME problems
  - Apply developed methods in particular problems (methods ready-to-use)
  - Get ME data + apply ML + infer + validate models + study
- Interested on the application of AI over ME problems

Conclusions

- 4 different paradigms in ML (regression, classification, fairness, causality)
- There are others which we can explore (depending on the problem)
- Main research lines:
  - Collect data (and share it), databases of meaningful ME problems
  - Apply developed methods in particular problems (methods ready-to-use)
  - Get ME data + apply ML + infer + validate models + study
- Interested on the application of AI over ME problems

Thanks for your attention!!

Bibliography

References I

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. (2005).
Measuring statistical dependence with hilbert-schmidt norms.
In Jain, S., Simon, H. U., and Tomita, E., editors, *Algorithmic Learning Theory*, pages 63–77, Berlin, Heidelberg. Springer.

Hoyer, P. O., Janzing, D., Mooij, J. M., Peters, J., and Schölkopf, B. (2008).
Nonlinear causal discovery with additive noise models.
In Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L., editors, *NIPS21, 2008*, pages 689–696. Curran Associates, Inc.

Pérez-Suay, A. and Camps-Valls, G. (2018).
Sensitivity maps of the hilbert-schmidt independence criterion.
Appl. Soft Comput., 70:1054–1063.

References II

Pérez-Suay, A. and Camps-Valls, G. (2019).
Causal inference in geoscience and remote sensing from observational data.
IEEE Trans. Geoscience and Remote Sensing, 57(3):1502–1513.

Pérez-Suay, A., Laparra, V., Mateo-Garcia, G., Muñoz-Marí, J., Gómez-Chova, L., and Camps-Valls, G. (2017).
Fair kernel learning.
In *ECML PKDD 2017*, LNCS. Springer.

Schölkopf, B., Herbrich, R., and Smola, A. J. (2001).
A generalized representer theorem.
In Helmbold, D. and Williamson, B., editors, *Computational Learning Theory*, pages 416–426, Berlin, Heidelberg. Springer Berlin Heidelberg.