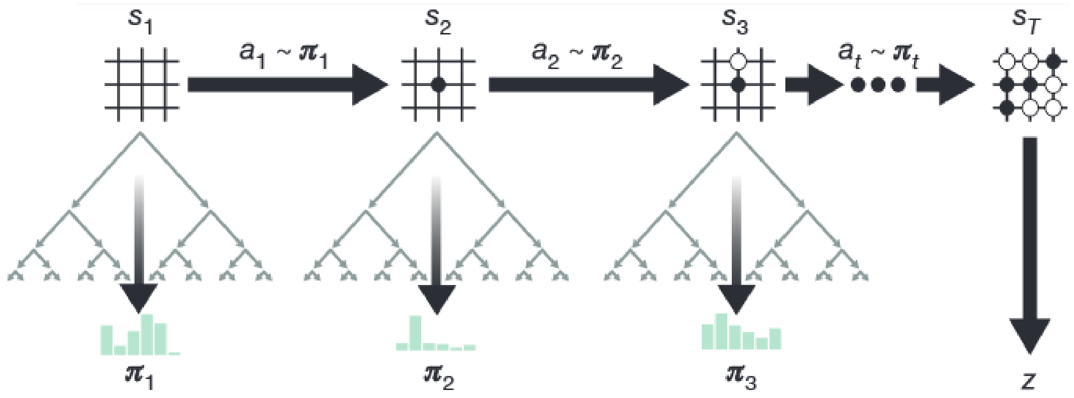
Too Big To Fail: Building Robust Intelligent Systems with Causal Machine Learning

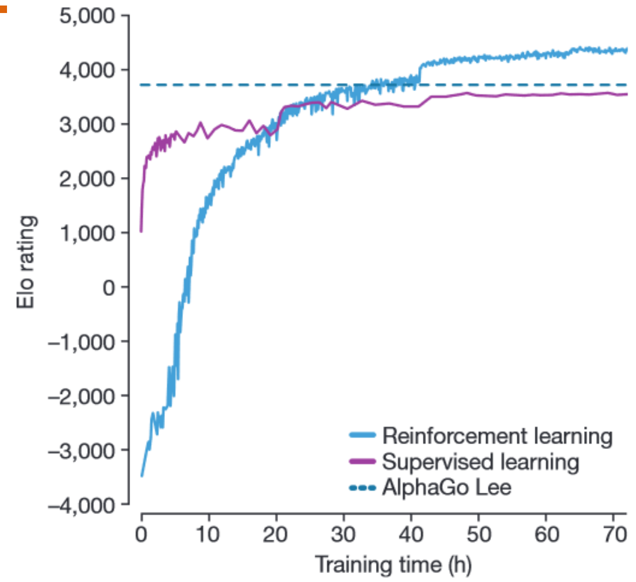
Christian Medeiros Adriano
System Analysis and Modeling Group
Hasso Plattner Institute at the University of Potsdam

October 27, 2021

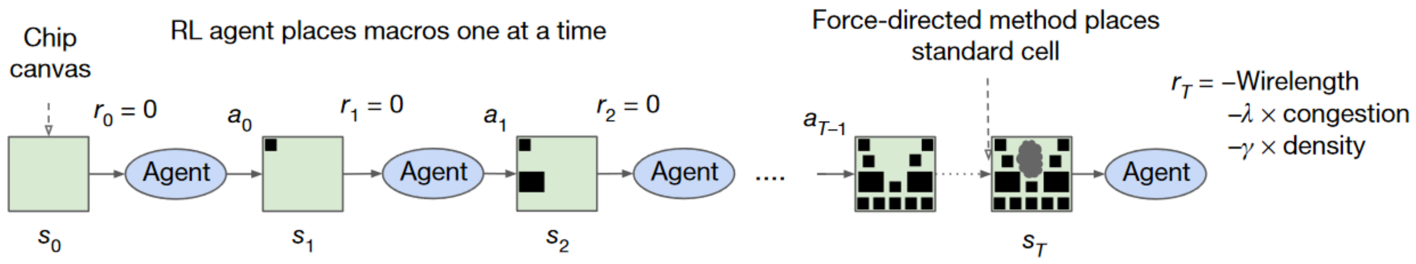
Context – The Progress of Reinforcement Learning

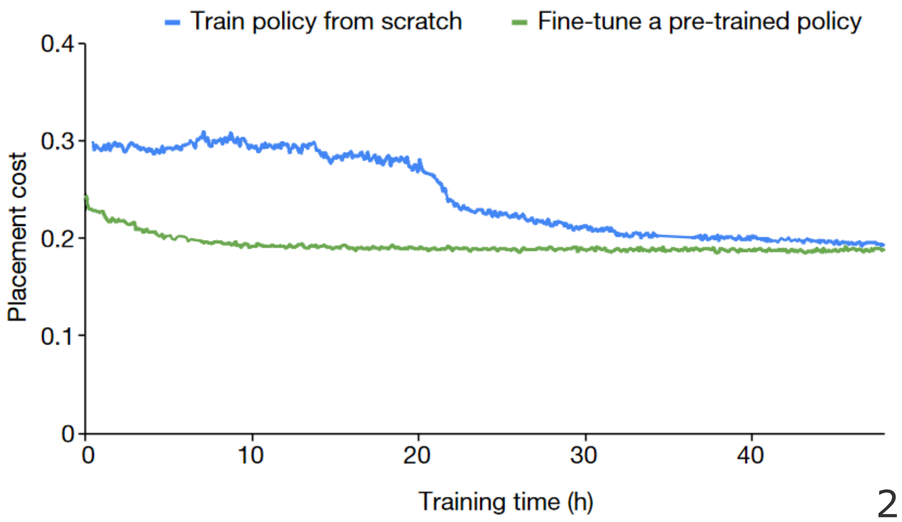
Alpha Go-Zero learned to win without supervision, only by self-play **[DeepMind 2017]**





RL agent designed optimal layouts for TPU chip circuit **[Google 2021]**





However
AI systems are not being deployed

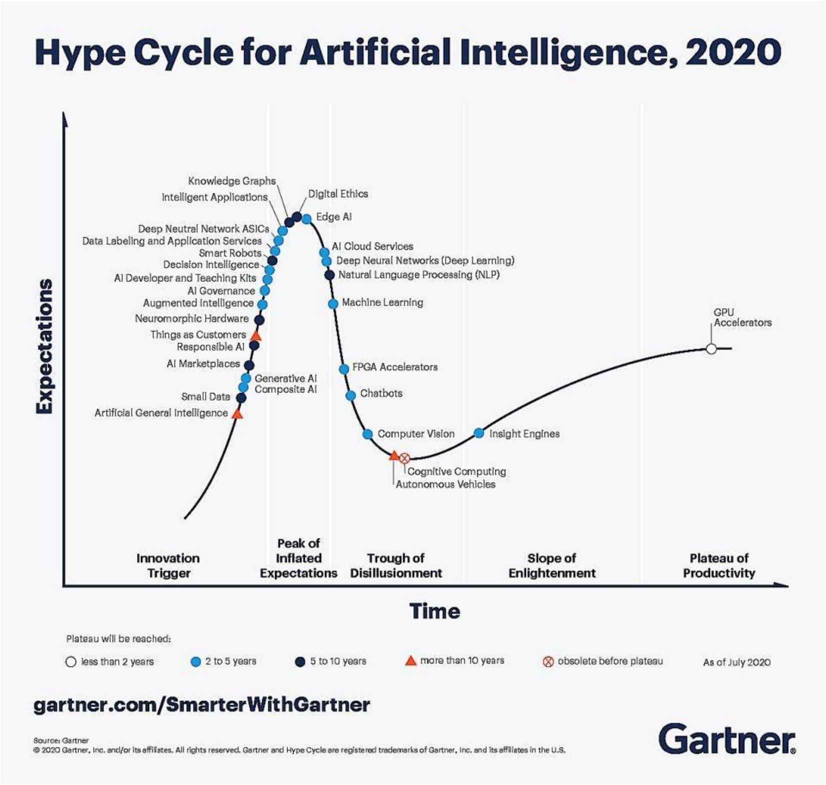
- **55%** of companies surveyed haven't deployed a machine learning model **[Algorithmia 2020]**

- **72%** that began AI pilots before 2019 haven’t deployed a single system yet **[Capgemini 2020]**

Reason? systems cannot **<u>adapt</u>** to more complex and evolving realities - <u>adversarial environments</u>

Problem?
In practice: Lack of Robustness in Production
In research: Lack of Generalizability

[Jordan 2019], [D’Amour et al. 2020]



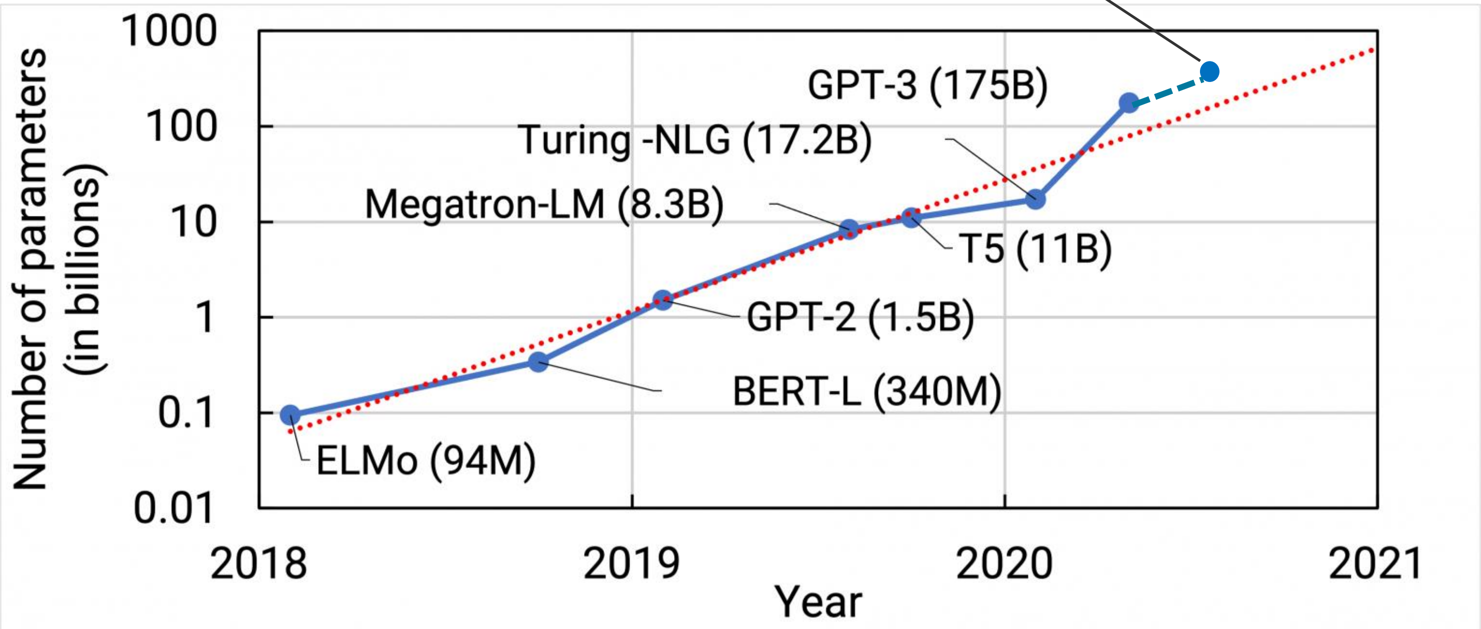
[Gartner 2020]

Smoking Gun-1 [Thompson et al. 2021] [NVIDIA 2021]

1- A bug in GPT3 was discovered but team decided not to fix because of the cost of retraining

- Alpha Go ~ 35 million USD
- GPT3 ~ 3 million USD

MT-NLP contains 530 *billion parameters*

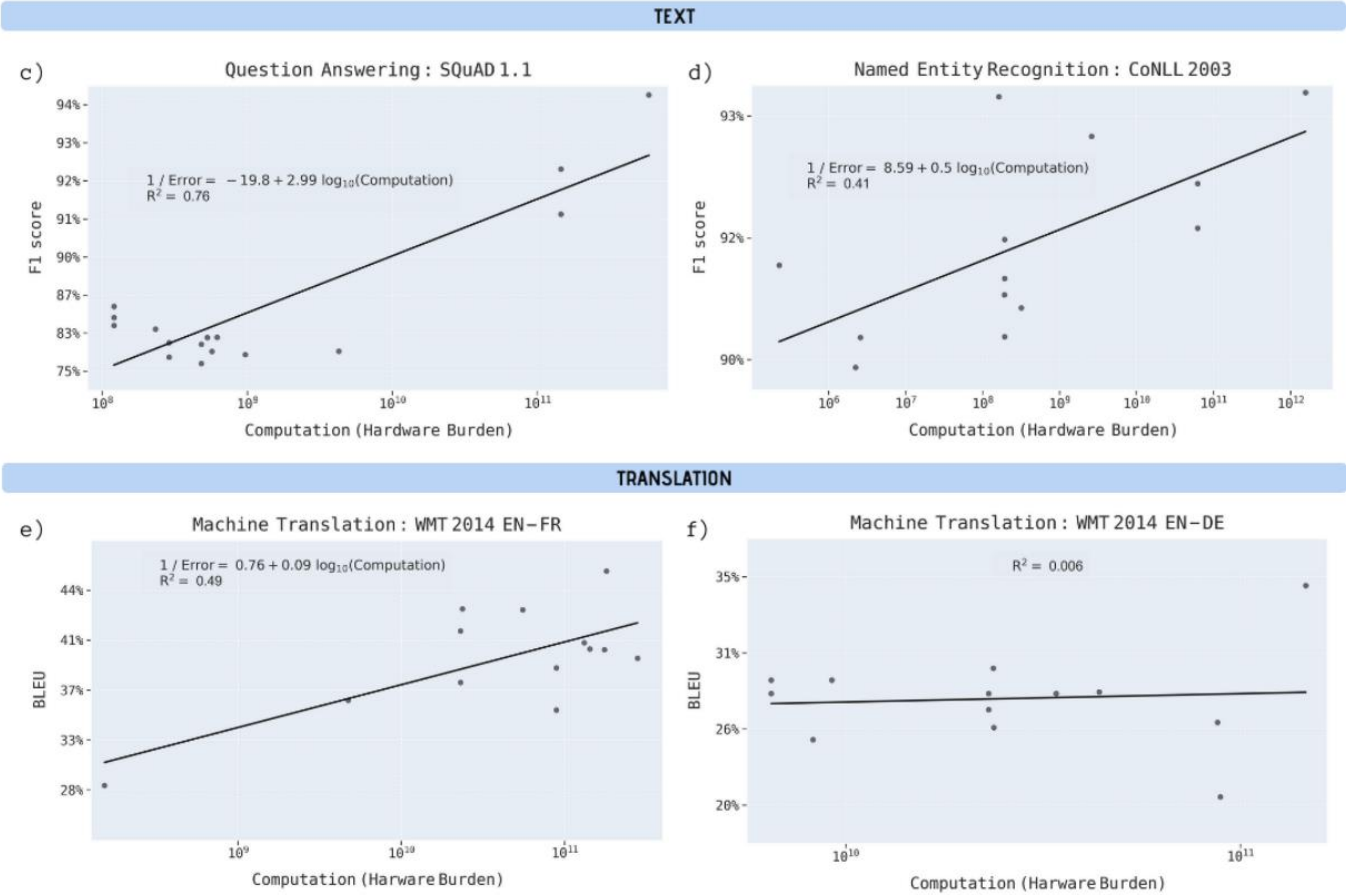


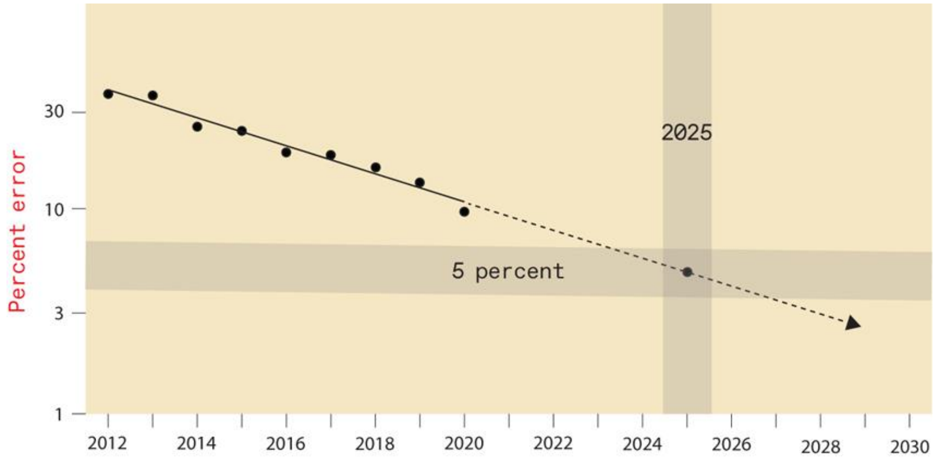
Source: [NVIDIA 2021]

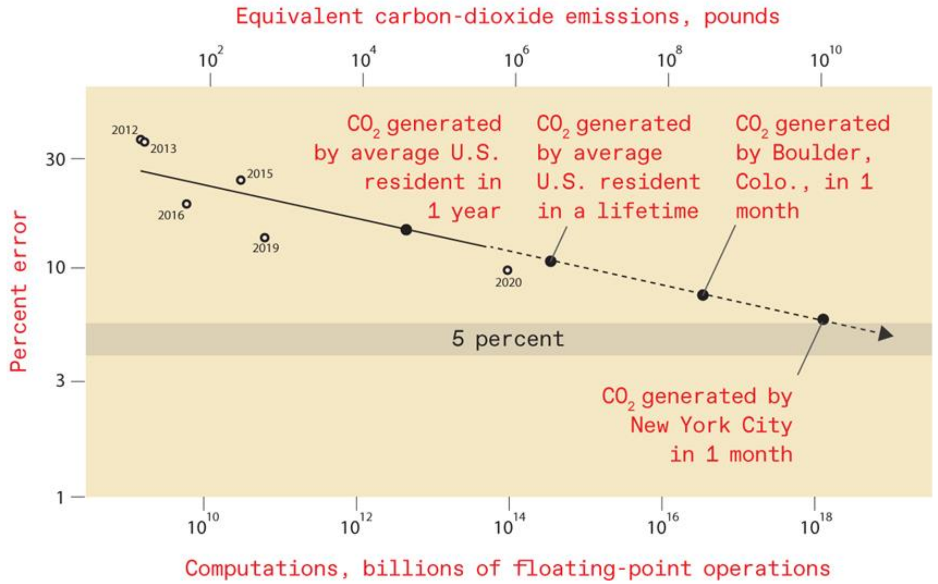
source: Thompson et al. 2021, Deep Learning Diminishing Returns, IEEE Spectrum

Smoking Gun-2 [Thompson et al. 2021]

2- Exponential environment cost for linear gains







Thompson et al. 2021, Deep Learning Diminishing Returns, IEEE Spectrum

Smoking Gun-3 [Thompson et al. 2021]

3- Even smaller models (few millions of parameters), the costs are already too high or business timeline too short

- "A large **European supermarket chain** recently abandoned a deep-learning-based system ... because they judged that the **cost of training and running the system would be too high**."

Thompson et al. 2021, Deep Learning Diminishing Returns, IEEE Spectrum

How do we current think about robustness? Bias-Variance Trade-off Intuition

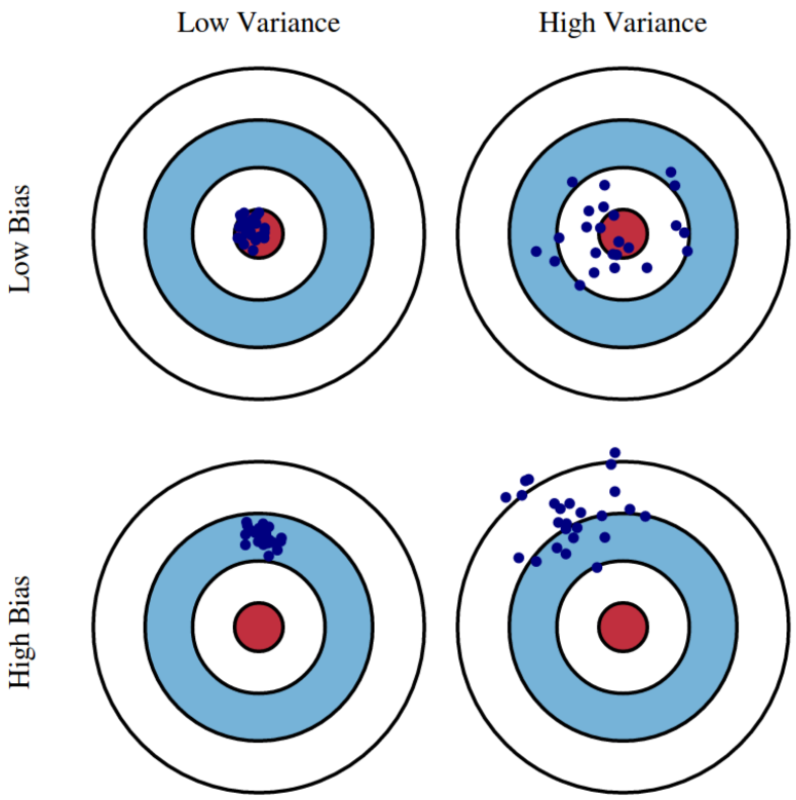


Fig 1: Graphical illustration of bias and variance.
Source: http://scott.fortmann-roe.com/docs/BiasVariance.html

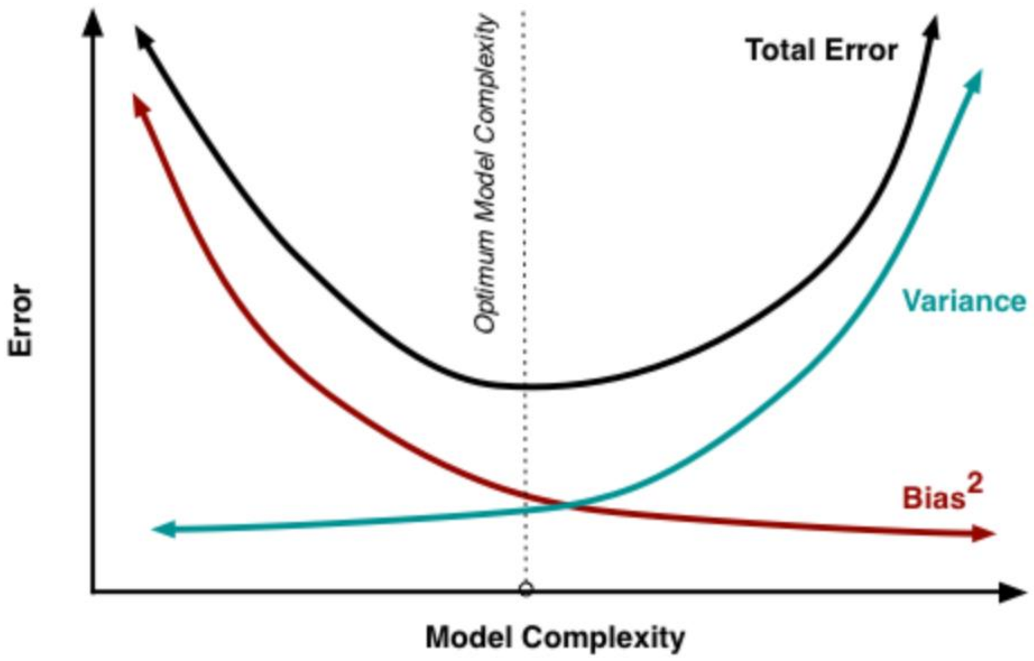


Fig 2: The variation of Bias and Variance with the model complexity. This is similar to the concept of overfitting and underfitting. More complex models overfit while the simplest models underfit.
Source: http://scott.fortmann-roe.com/docs/BiasVariance.html

Bias Variance Trade-off

Bias-Variance Decomposition (irreducible error)

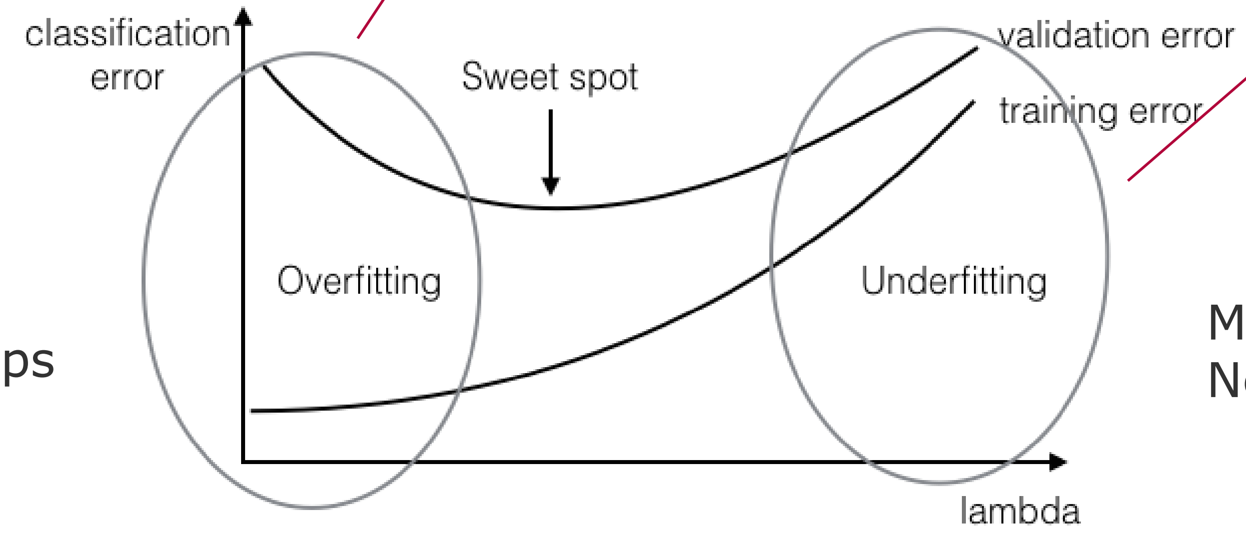
Prediction

Ground truth

$$\underbrace{E_{\mathbf{x},y,D}\left[\left(h_D(\mathbf{x})-y\right)^2\right]}_{\text{Expected Test Error}} = \underbrace{E_{\mathbf{x},D}\left[\left(h_D(\mathbf{x})-\bar{h}(\mathbf{x})\right)^2\right]}_{\text{Variance}} + \underbrace{E_{\mathbf{x},y}\left[\left(\bar{y}(\mathbf{x})-y\right)^2\right]}_{\text{Noise}} + \underbrace{E_{\mathbf{x}}\left[\left(\bar{h}(\mathbf{x})-\bar{y}(\mathbf{x})\right)^2\right]}_{\text{Bias}^2}$$

[Weinberger 2018]

More data helps



More data does not help
Need better model!

Lambda = strength of regularization
It pulls the model away from local minima

How should we be thinking about robustness? Essential Sources of Lack of Robustness

Hidden confounders + Selection Bias

Simpson’s and Berkson’s paradoxes

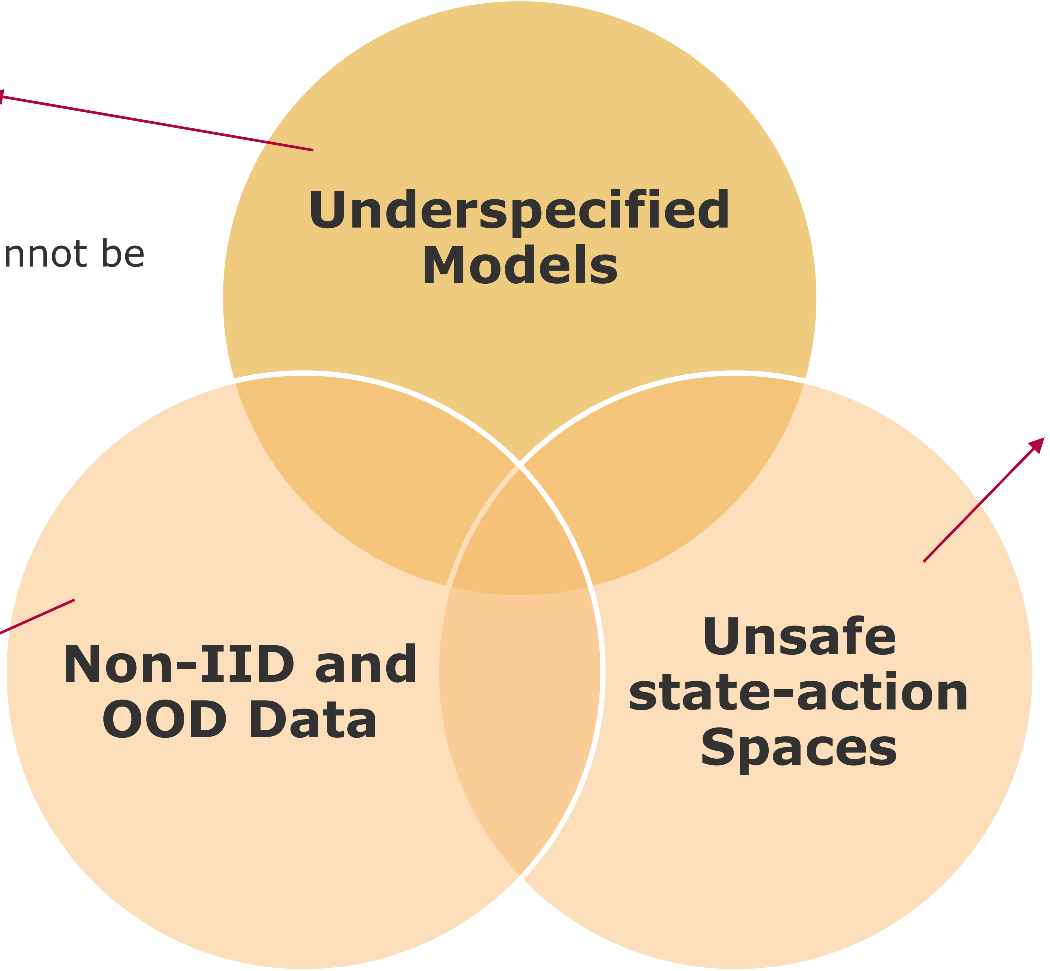
Shortcut learning in Neural Nets

This goes beyond overfitting, as it cannot be solved with more or better data!

Real-world is non-stationary

Predictions affects the data generation process

Modeling better recommender systems is not enough, because uncertainty grows wildly when extrapolating out-of-distribution



Wrong predictions can spur unsafe actions that can lead to unsafe states.

Sensitivity analysis and testing on hold-out-sets are ad hoc approaches cannot guarantee safety.

“Program testing can be used to show the presence of bugs, but never to show their absence!” — Edsger W. Dijkstra

How should we be thinking **deeper** about robustness?
Nature of the Problem - Structure vs Frequency

by Christian M. Adriano

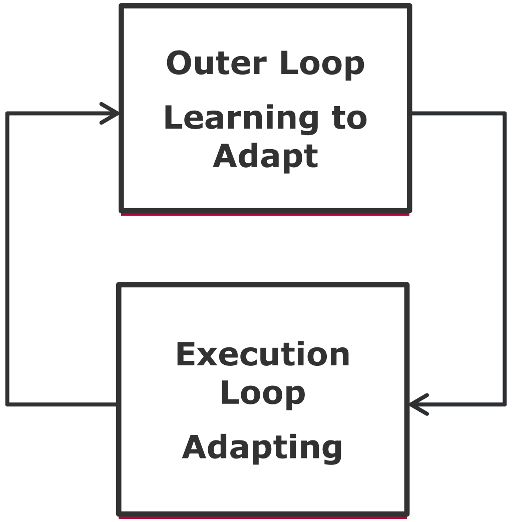
Degree of structure	Sparse (causal associations)	High (accidental associations)
Strong (well-constrained)	1 **Physical laws** *circuit design* *protein folding* *autonomous driving* [Google 2021]	2 **Artificial laws** *games* *simulators* *software systems* [DeepMind 2017]
Medium (under-specified)	4 **Adversarial laws** *Non-IID and OOD* *dynamical* *systems* [Ghahremni, Adriano & Giese, 2018]	3 **Human laws** *human-in-the loop* *image & text* *interfaces* [Adriano 2018]

Frequency of positive events

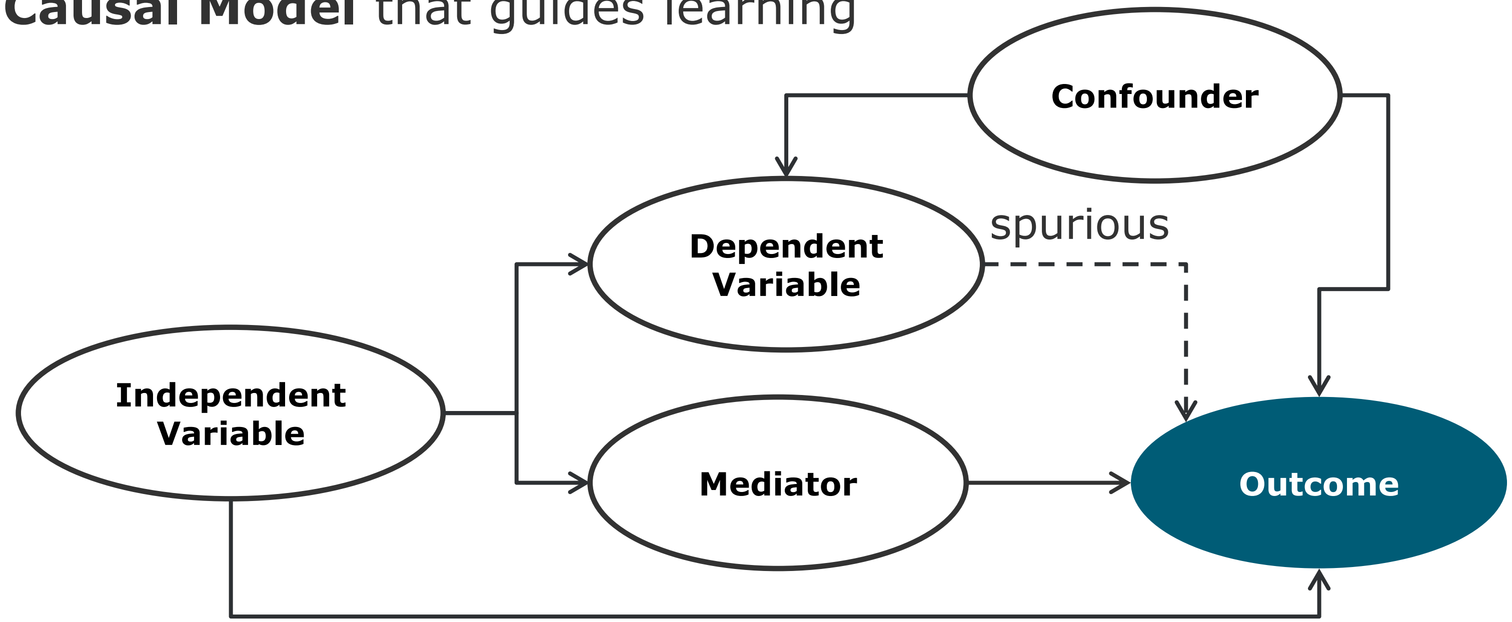
This is the space we have been working

How are we approaching robustness?

An **Outer Loop** that keeps learning a representation of the world



A **Causal Model** that guides learning



Multiple Mechanisms
&
Intervention Options

Concrete Challenges and Solution

Catastrophic forgetting

Solution: uncover the causal structure by mapping hidden confounders and invariants

Sample efficiency

Solution: generative models (model-based RL, replay-buffer, digital-twins) to hallucinate hypothetical adversarial realities

Delayed rewards

Solution: continuous learning (transfer, meta, curriculum learning) to train for new, modified or more complex tasks

For all solutions we need a model that can recommend **interventions** that can generate adversarial situation (non-IID, OOD, possibly unsafe)

Typology of Interventions for Robustness

Accidental Shift (changes outside the causal path)

- **how**? Choice of intervention creates no path or blocks the path to the outcome variable.
- **why**? Uncover latent confounders, detect spurious correlations, disentangle accidental and essential attributes, necessary and sufficient causes

Essential Shift (changes in the causal path)

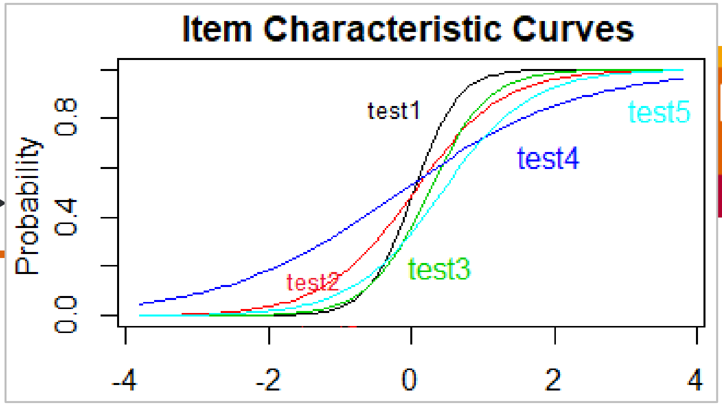
- **how**? Choice of intervention creates one or more paths to the outcome variable. Sensitivity analysis can be used.
- **why**? Test the direction and magnitude of causality, and the independence between interventions and mechanisms

Mechanism Shift (changes in mechanisms)

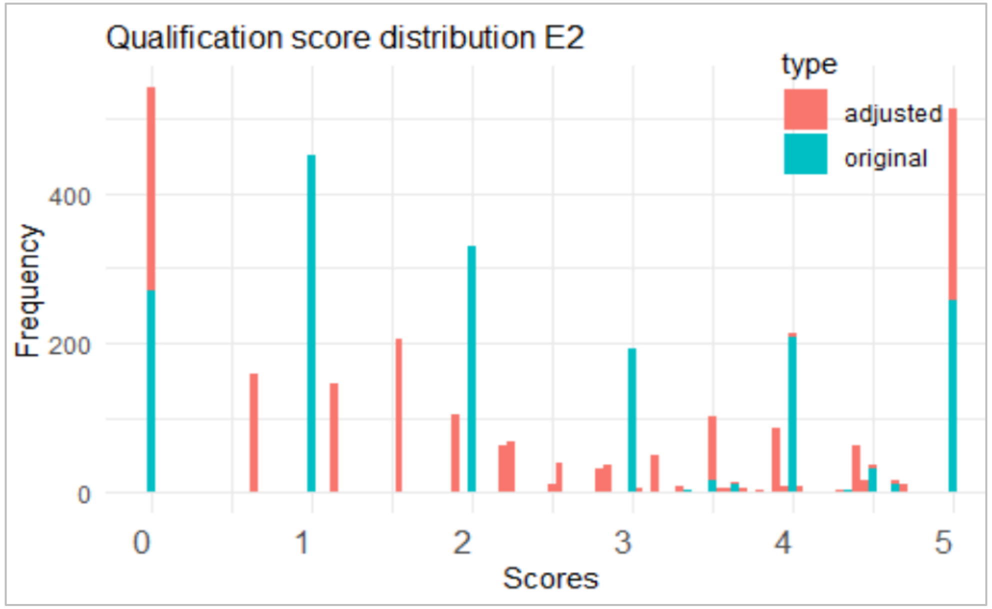
- **how**? Causal paths to the outcome variable depends on the magnitude or value of the intervention, i.e., violation of the mechanism independence assumption. Domain shifts can be used.
- **why**? Uncover the invariant mechanisms across domains

Results of Interventions

Intervention-Accidental Shift Changes in Score Distribution-E2

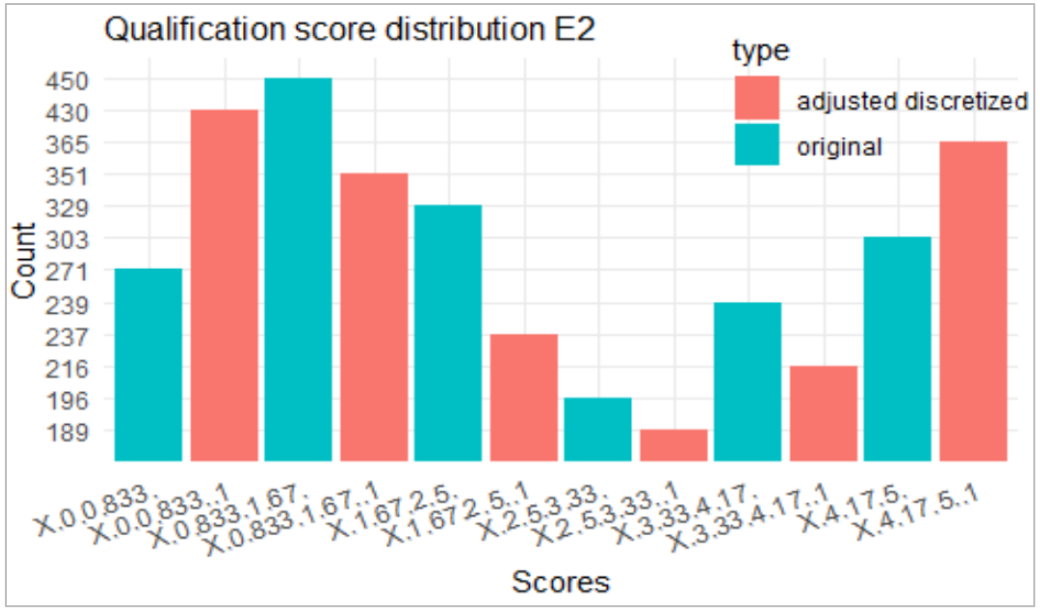


Adjusted Score from Item Response Theory Model: fits a logistic model based on difficulty of programming tests



Statistically distinct? **YES**
Kruskal-Wallis chi-squared = 1787,
df = 61, p-value < 2.2e-16

Entropy values distinct? **NO**
Change ($\frac{adjusted-original}{original}$) = -0.53%

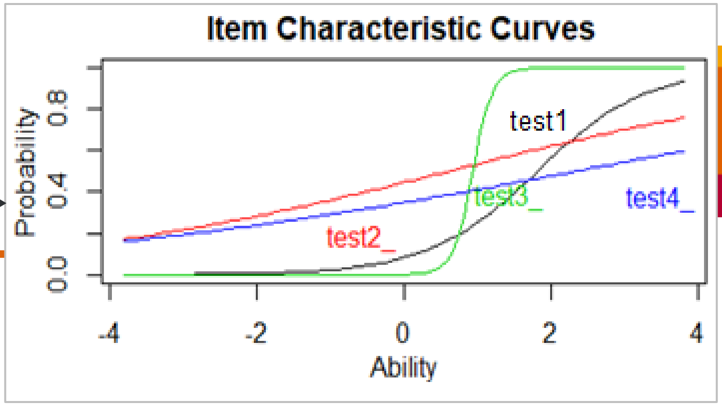


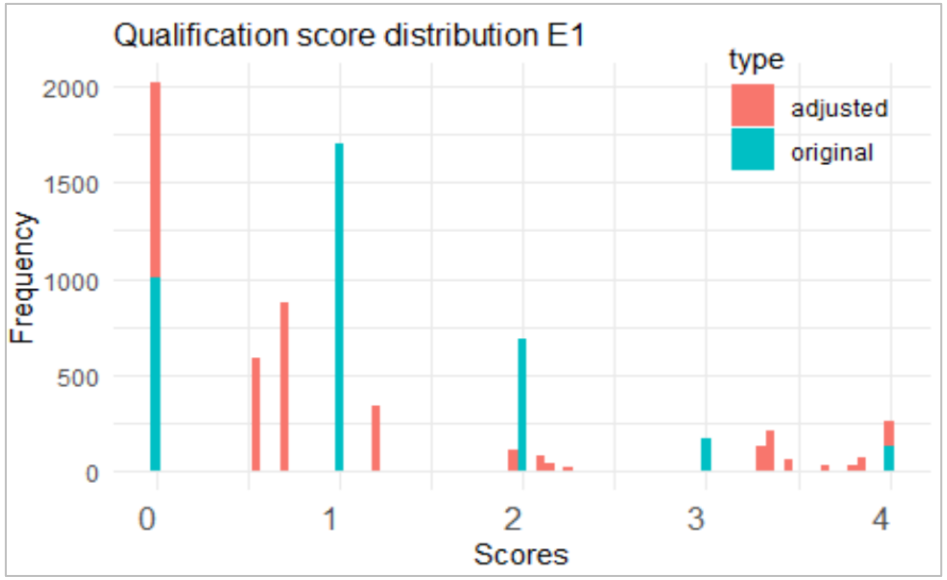
Statistically distinct? **POSSIBLY NOT**
Kruskal-Wallis chi-squared = 5,
df = 5, p-value = 0.4159

Entropy values distinct? **NO**
Change ($\frac{adjusted-original}{original}$) = 0.04%

Intervention-Accidental Shift
Changes in Score Distribution – E1

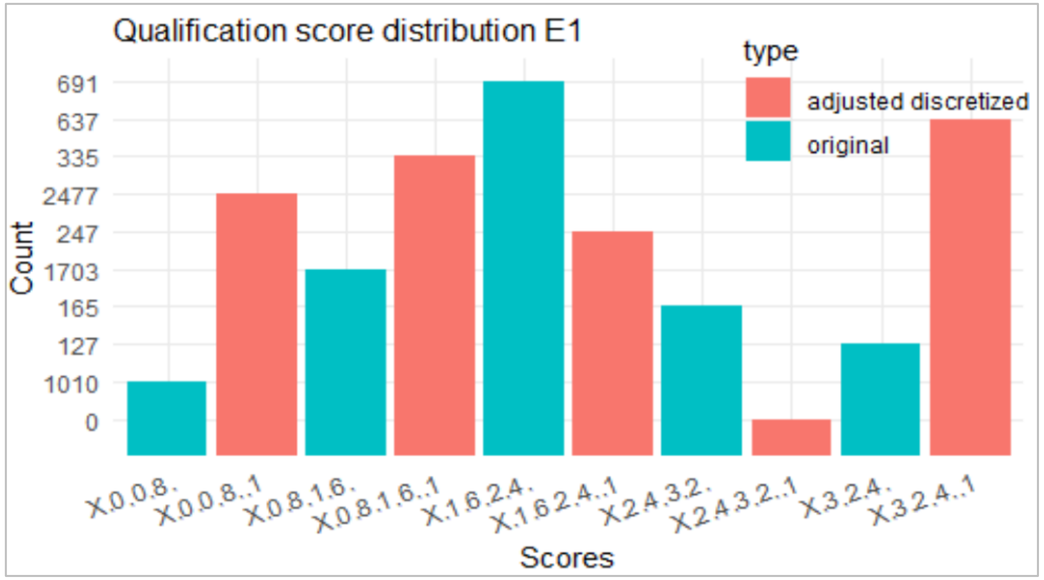
Adjusted Score from Item Response Theory Model: fits a logistic model based on difficulty of programming tests





Statistically distinct? **YES**
Kruskal-Wallis chi-squared = 3695,
df = 15, p-value < 2.2e-16

Entropy values distinct? **NO**
Change ($\frac{adjusted-original}{original}$) = -1.04%

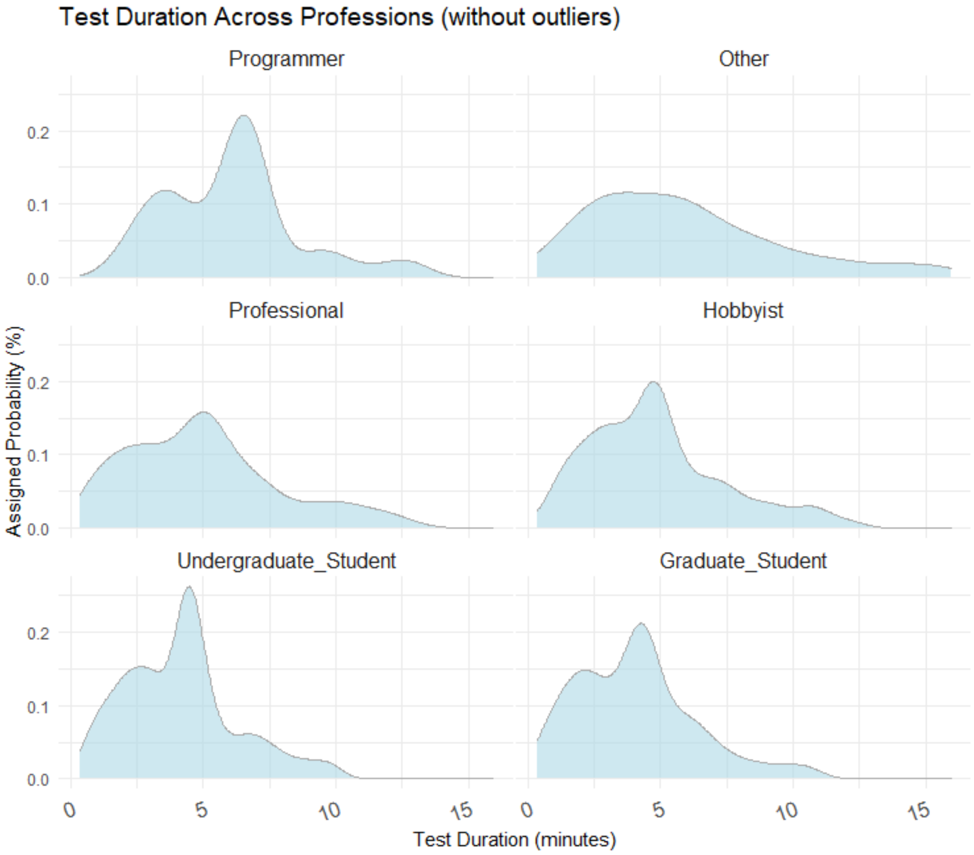


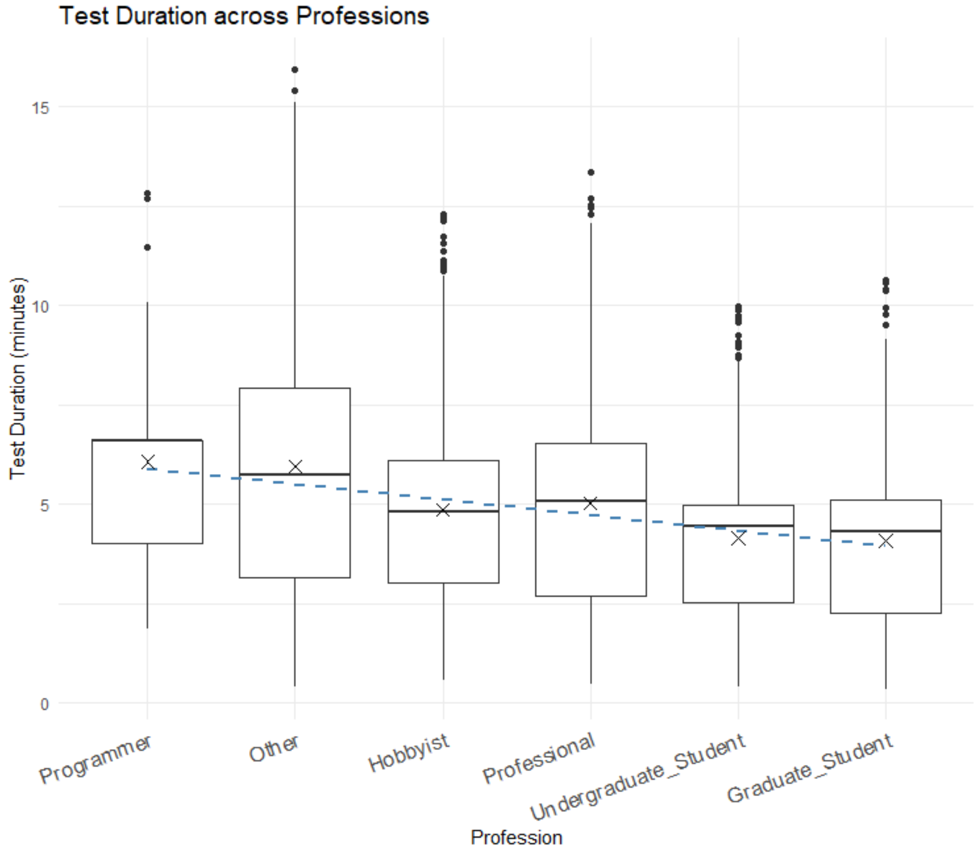
Statistically distinct? **POSSIBLY NOT**
Kruskal-Wallis chi-squared = 4,
df = 4, p-value = 0.406

Entropy values distinct? **YES**
Change ($\frac{adjusted-original}{original}$) = -24.13%

Trade-off

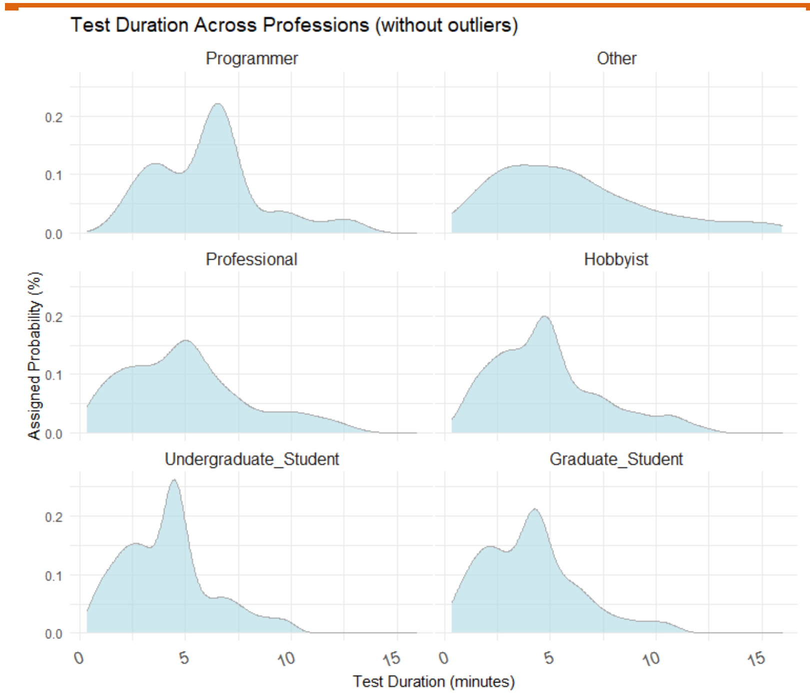
Intervention - Essential Shift
Changes in Task duration across professions

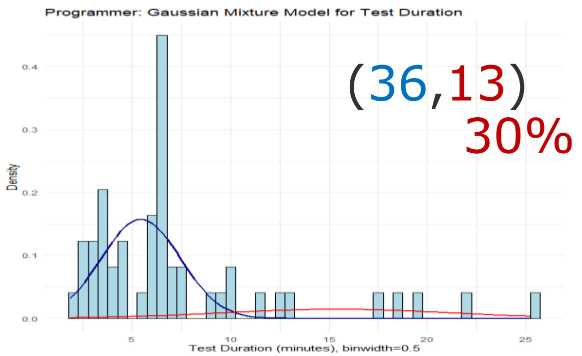


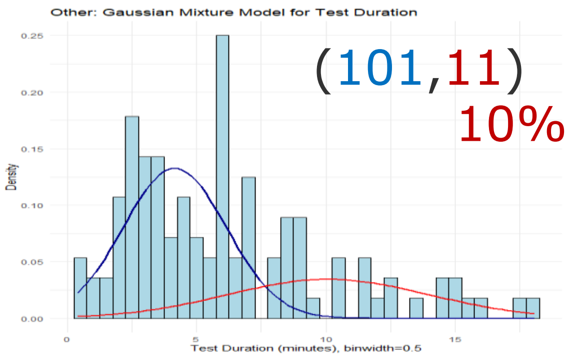


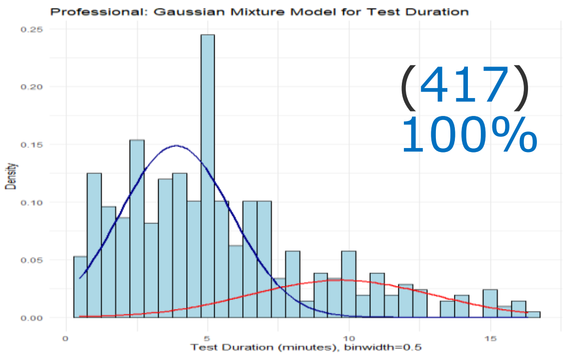
Gaussian Mixture Model by Profession

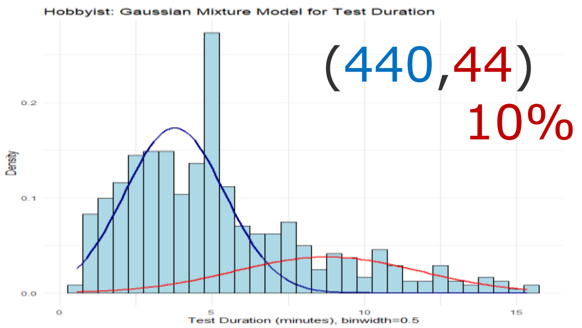
Proportion (Blue,Red), Blue ϵ fast-cluster, Red ϵ slow-cluster

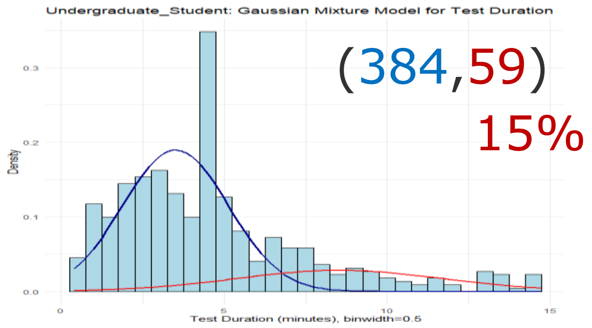


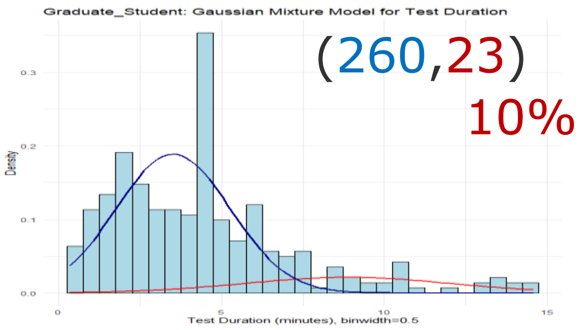




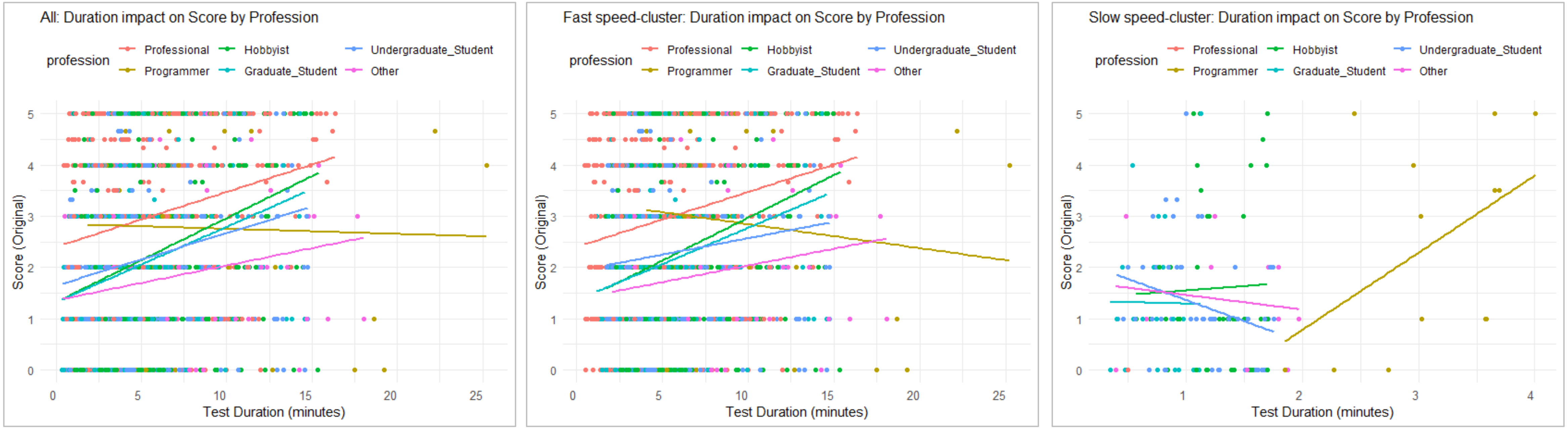




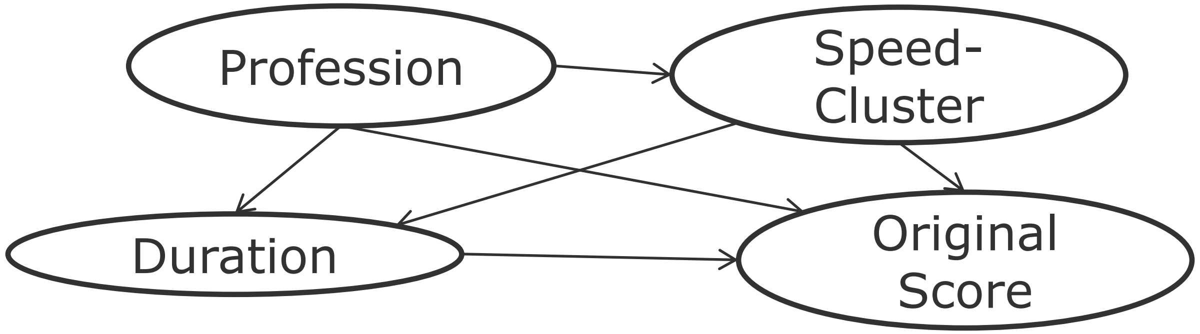




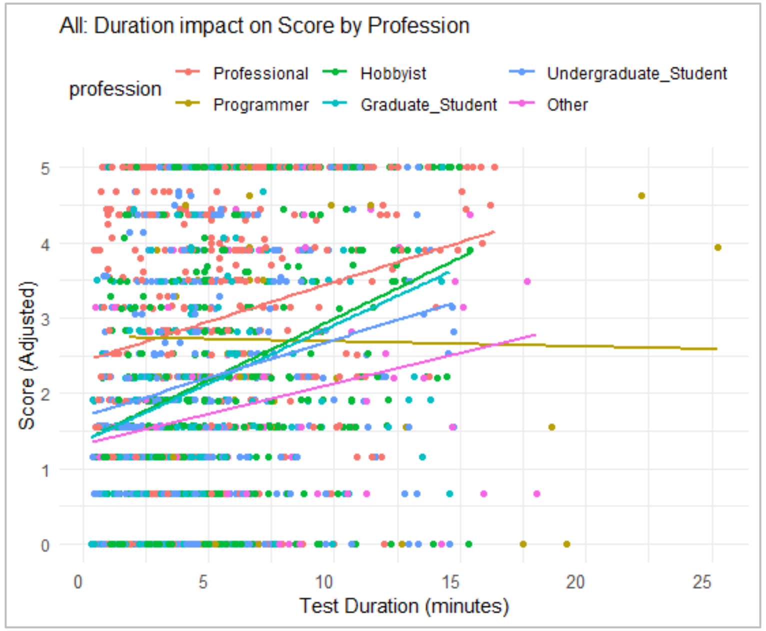
Effect of duration on **original score** by speed-cluster

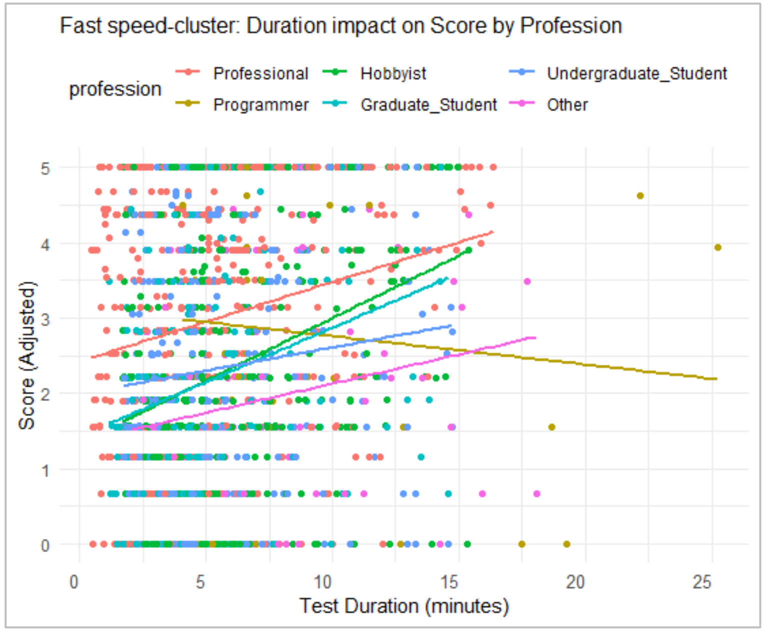


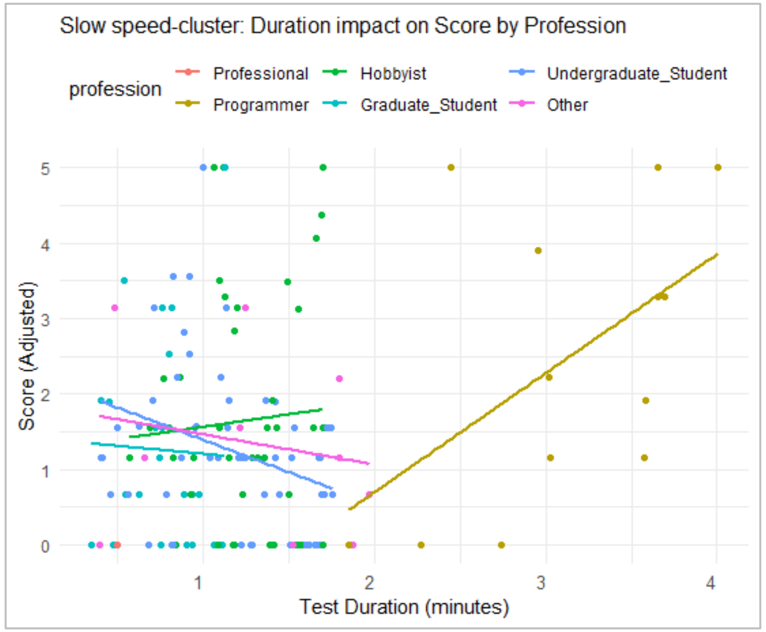
Mixture Model membership is a confounder of the effect of task duration on score.



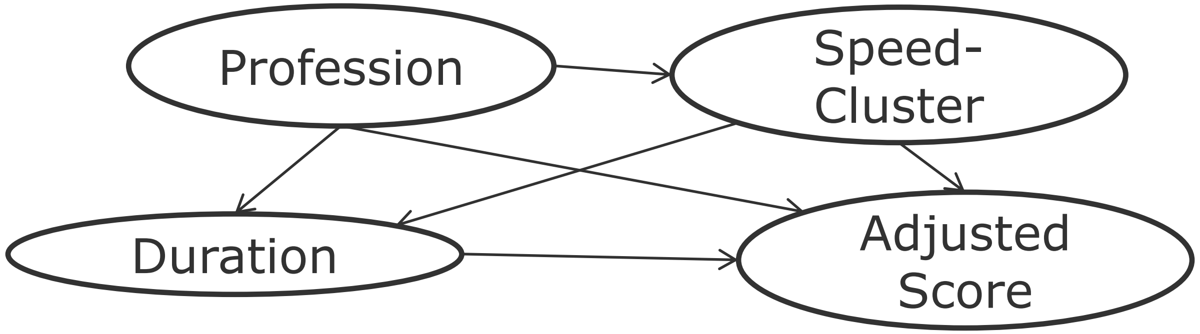
Effect of duration on **adjusted score** by speed-cluster



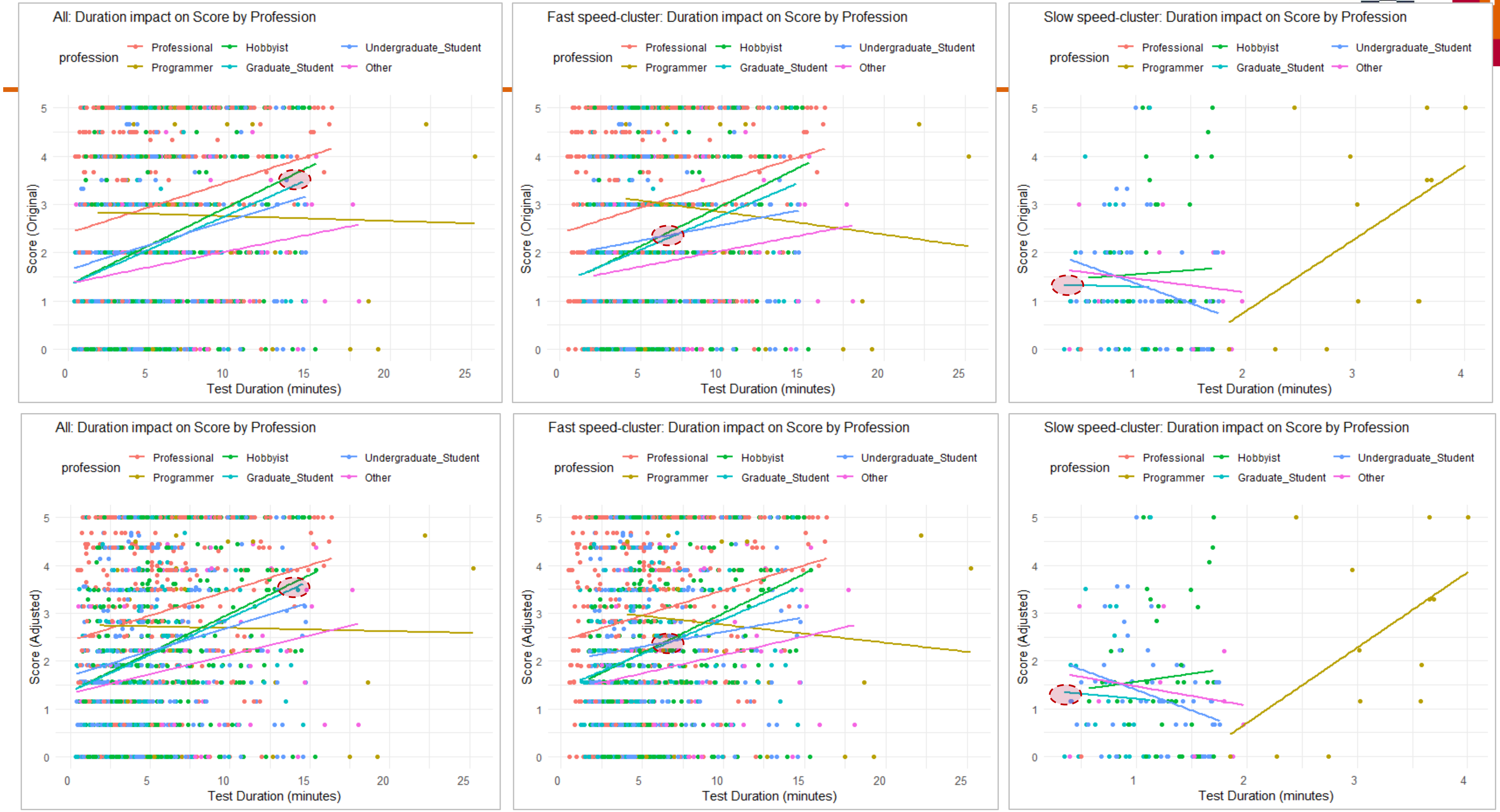




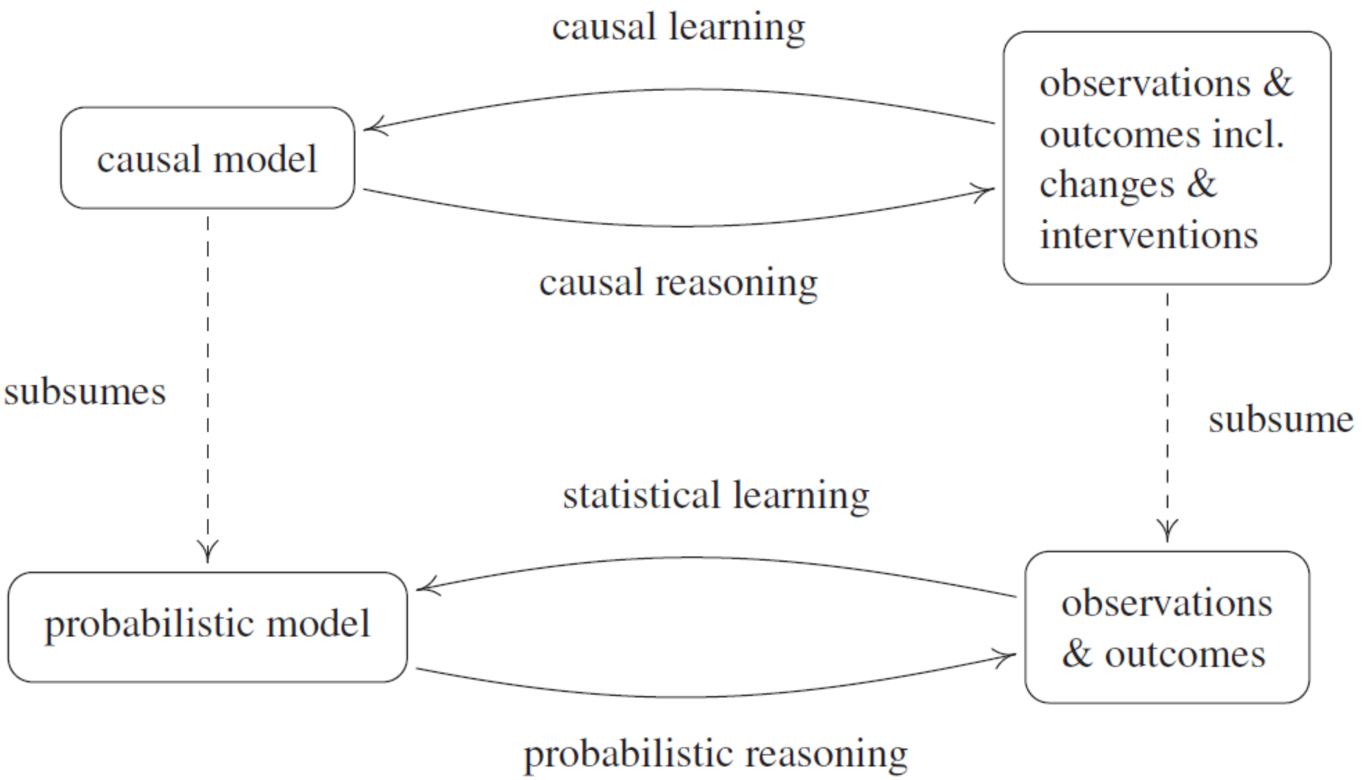
Mixture Model membership is a <u>confounder</u> of the effect of task duration on score.



Small variations in association (original vs adjusted score).

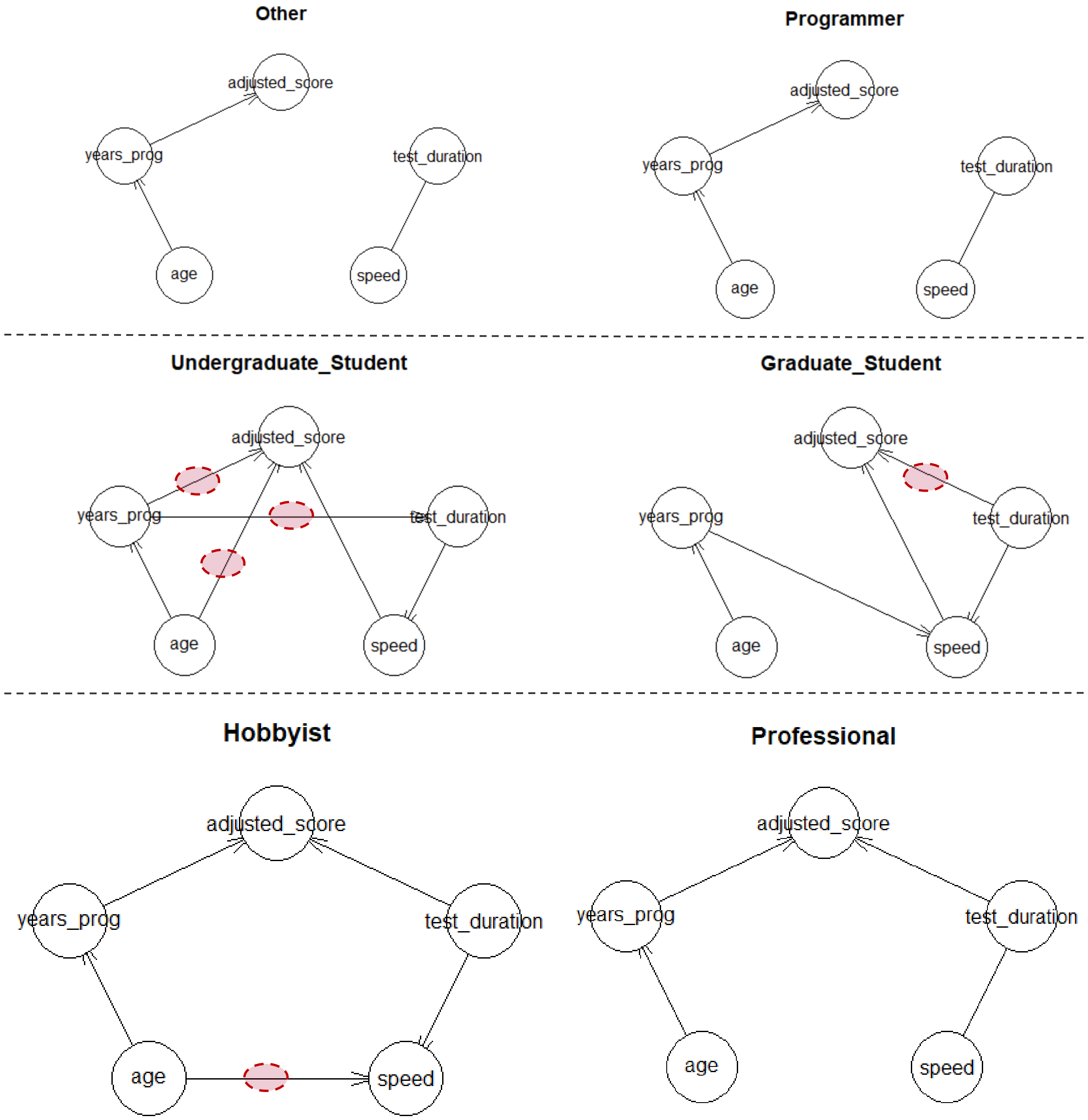


Intervention – Mechanism Shift
Overall approach to causal inference



source: Peters, J., Janzing, D., & Schölkopf, B. (2017). *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press.

Causal Graphs by Profession: Constraint-Based Method [Glymour et al 2019]



Consequence for Planning Interventions

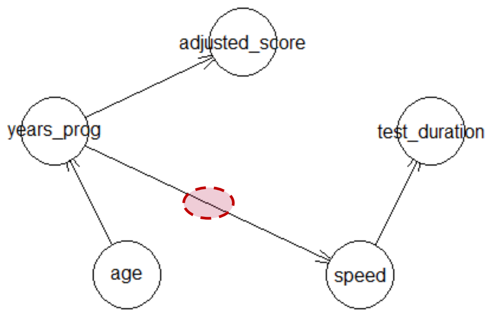
1- No need to distinguish programmers among the Others

2- If the only information is that the person is a student, then can only rely on interventions that change the speed relative to the average students.

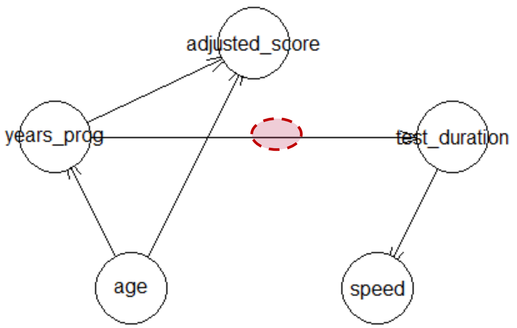
3- Irrelevant non-invariant, because speed is not a valid intervention for this group

Causal Graphs by Profession: Score-Based Method [Glymour et al 2019]

Other



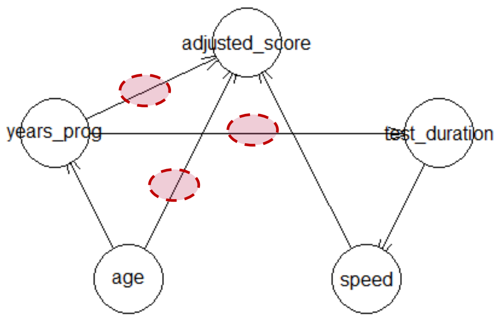
Programmer



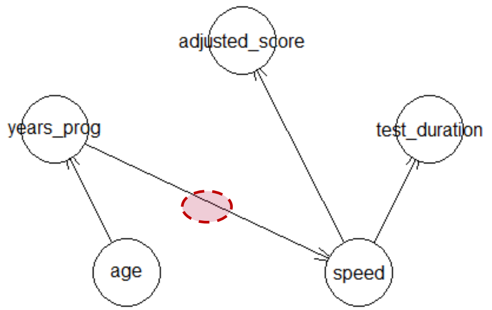
Consequence for Planning Interventions

1- No need to distinguish programmers among the Others

Undergraduate_Student

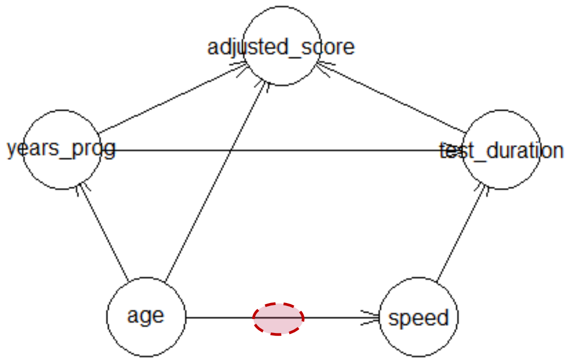


Graduate_Student

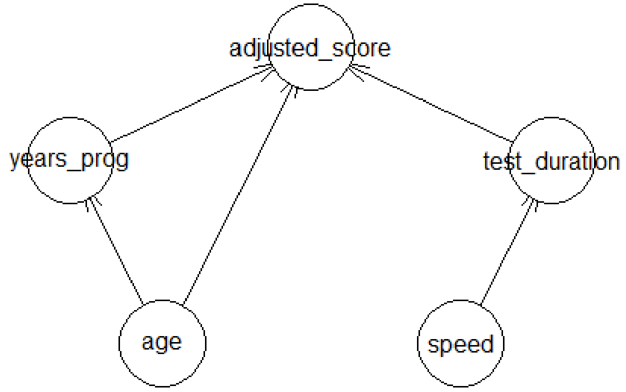


2- If the only information is that the person is a student, then can only rely on interventions that change the speed relative to the average students.

Hobbyist



Professional

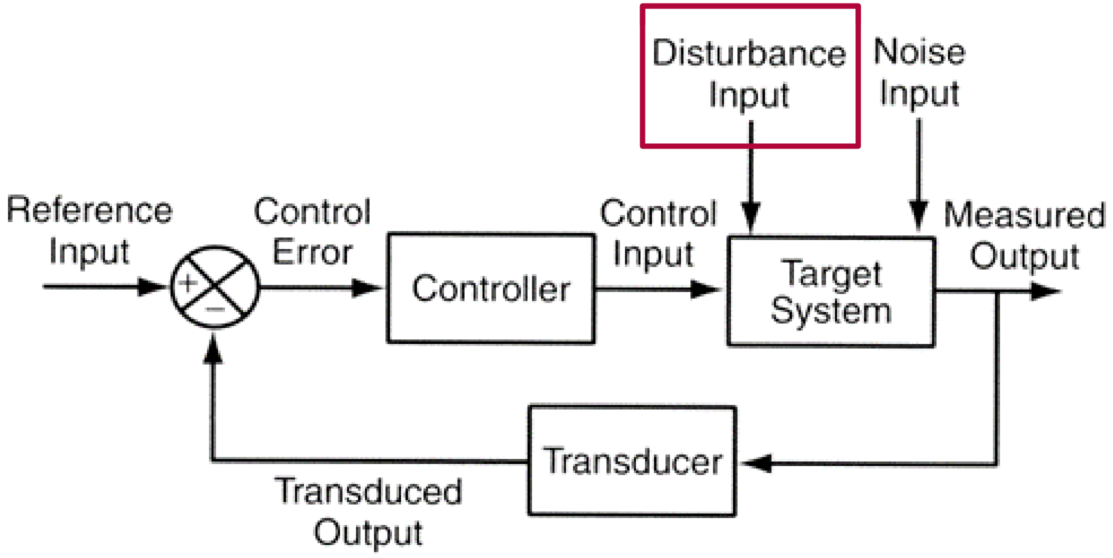


3- Irrelevant non-invariant, because speed is not an valid intervention for this group

Infrastructure to run causal system experiments

Feedback loop models

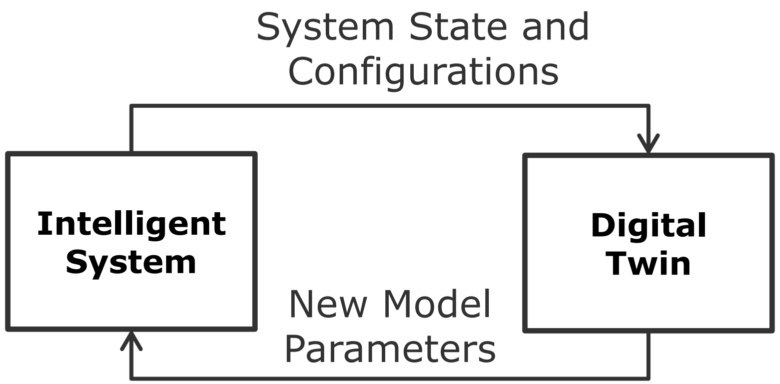
"When solving a problem of interest, do not solve a more general problem as an intermediate step. Try to get the answer that you really need but not a more general one." **Vladimir Vapnik**



Simulation models

"Thinking is acting in an imagined space" **Konrad Lorenz**

"Perception is a generative act" – [Gross et al. 1999]
" Consciousness is a controlled hallucination " - [Seth et al. 2000]



Take-aways

What are effective and plausible environment changes?

- Change outcome distribution forced **Accidental Shifts** (adjusted score)
- Change input distribution forced **Essential Shifts** (profession)
- Change features forced **Mechanism Shift** (speed membership)

What is the **lack of robustness** detected after environment changes?

- Reversal or cancelling of effects (Simpson's and Berkson's paradoxes)
- Weak and non-significant effects (close to zero)

What are the model invariants?

- <u>Hidden confounders</u> that entail spurious correlations (structural)
- Discovery of relationships that are environment specific (accidental)

“**There is no causation without manipulation**” (Rubin 1975) (Holland 1986)
- We need to design “system experiments”.

“**All models are wrong, some are useful**” – (George Box 1976)
- Models must be continuously updated to cope with a changing environment

END

References (1)

[Algorithmia 2020] The State of Enterprise ML, https://info.algorithmia.com/hubfs/2019/Whitepapers/The-State-of-Enterprise-ML-2020/Algorithmia_2020_State_of_Enterprise_ML.pdf

[Meyer 1997] Meyer, B. (1997). *Object-oriented software construction* (Vol. 2, pp. 331-410). Englewood Cliffs: Prentice hall.

[Capgemini 2020] The AI-Powered Enterprise, https://www.capgemini.com/gb-en/research/the-ai-powered-enterprise/

[Chen et al. 2021] Chen, L., et al., 2021, Decision Transformer: Reinforcement Learning via Sequence Modeling, https://arxiv.org/pdf/2106.01345.pdf

[Cochran & Chambers 1965] W. G. Cochran and S. Paul Chambers, 1965, The Planning of Observational Studies of Human Population, in Journal of the Royal Statistical Society. Series A (General) , 1965, Vol. 128, No. 2, pp. 234-266

[D'Amour et al. 2020] D'Amour, A. et al., 2020, Underspecification presents challenges for credibility in modern machine learning, https://arxiv.org/pdf/2011.03395.pdf

[DeepMind 2017] Silver, D., et al., 2017, Mastering the game of Go without human knowledge, Nature, pp. 354-371, doi:10.1038/nature24270. https://deepmind.com/blog/article/alphago-zero-starting-scratch

References (2)

[Franks et al 2019] Franks, A., D'Amour, A., & Feller, A. (2019). Flexible sensitivity analysis for observational studies without observable implications. *Journal of the American Statistical Association*.

[D'Amour 2019] D'Amour, A. (2019). On multi-cause causal inference: Impossibility, sensitivity, and the promise of proxies. In International Conference on Artificial Intelligence and Statistics, pp. Forthcoming
Wang, Y. and D. M. Blei (2018). The blessings of multiple causes. arXiv preprint arXiv:1805.06826

[Gao et al. 2021] Gao, C., et al., 2021, Advances and Challenges in Conversational Recommender Systems: A Survey, https://arxiv.org/pdf/2101.09459.pdf

[Gartner 2020] https://www.forbes.com/sites/louiscolumbus/2020/10/04/whats-new-in-gartners-hype-cycle-for-ai-2020/?sh=723a6e57335c

[Google 2021] Mirhoseini, A., et al., 2021, A graph placement methodology for fast chip design. *Nature,* https://doi.org/10.1038

[Janer et al. 2021] Janer, M., et al., 2021, Reinforcement learning as one big sequence modeling problem, https://arxiv.org/pdf/2106.02039.pdf

[Jordan 2019] Jordan, M., 2019, Artificial Intelligence—The Revolution Hasn't Happened Ye, MIT Press, https://hdsr.mitpress.mit.edu/pub/wot7mkc1/release/9

[Popper 1962] Popper, K., 1962, *Conjectures and refutations: The growth of scientific knowledge*. Routledge.

[Vogel et al. 2018] Vogel, T., et al., 2018, mRUBiS: an exemplar for model-based architectural self-healing and self-optimization*, in SEAMS '18, pp. 101–107. https://doi.org/10.1145/3194133.3194161*

[Suton & Barto 2018] Sutton, R. & Barto, A., 2018, Reinforcement Learning, An Introduction, MIT Press

[Weinberger 2018] Weinberger K., 2018, Lecture Notes https://www.cs.cornell.edu/courses/cs4780/2018fa/lectures/lecturenote12.html

Backup Slides

Link between Robustness and Reproducibility: Principled Engineering + Counterfactual Models

Fundamental inquiry: If my model performs better (whatever better means), can I reproduce it in similar contexts (whatever similar means)? Something that is not reproducible has already failed a very simple test of robustness – generalizability.

My insight: This requires forward and backward reasoning.

- The forward reasoning is a set of principles that should be part of an engineering body of knowledge discipline (principled engineering)
- The backward reasoning relies on explaining the outcomes via associations derived from a causal mechanism (counterfactual models)

Principled Engineering = is a set of methods to guide design decisions at various levels of granularity and constrained by well-specified requirements and concrete implications

Counterfactual Models = allow to explain the outcome of mechanism by answering what-if questions

The apparent paradox
Occam-Razor versus Elaborate Hypotheses

The Ockham's-Razor or The Principle of Parsimony

Simpler models (theories) are preferable because they require fewer conditions to explain a phenomenon [Duigan 2021]. With fewer fundamental conditions [Schaffer 2015], there are higher chances that the model will generalize over many instances (variations) of the same generative data process (phenomenon).

"Make your hypotheses elaborate" principle – Sir Ronald Fisher

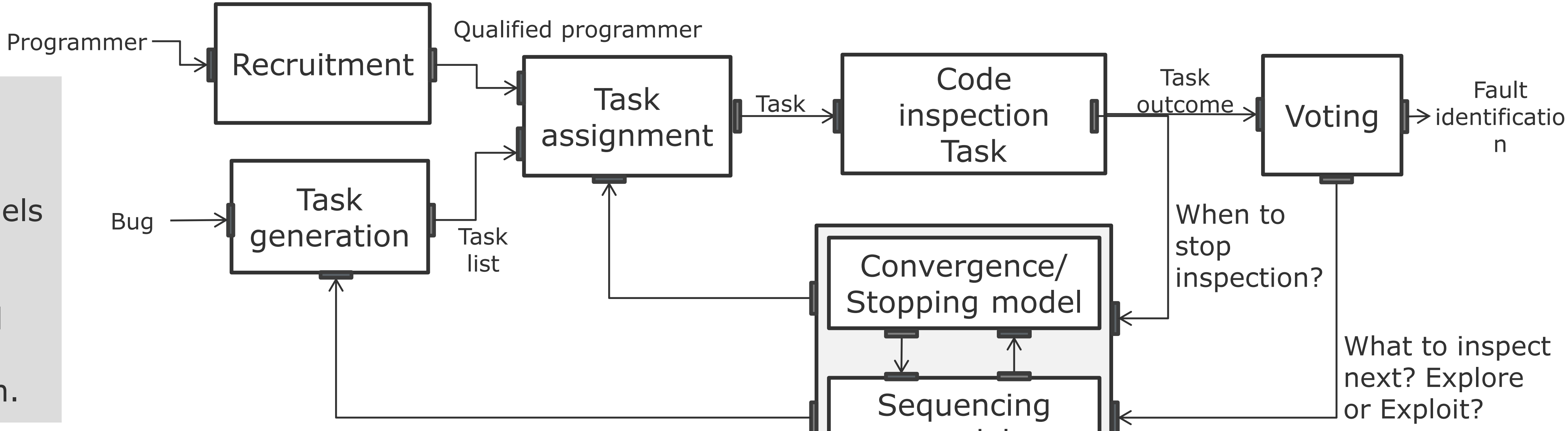
"…one should envisage as many different consequences of its (theory) truth as possible, and plan observational studies to discover whether each of these consequences is found to hold " (explanation to Fisher's answer to a question about the Occam-Razor principle, see section 5 in [Cochran & Chambers 1965]). This agrees with the Falsification Principle [Popper 1962].

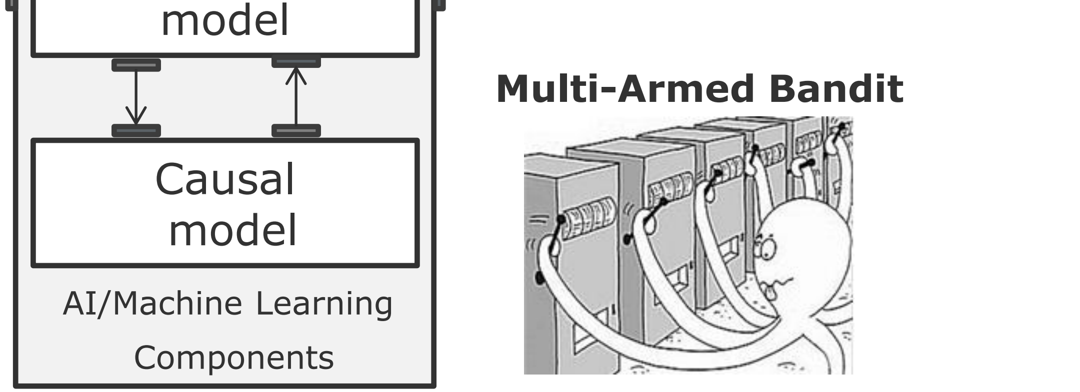
Hence, there is no paradox.

- The Ockham's-Razor principle aims at the **internal mechanisms** (as simple as possible) that still generate correct predictions. This is important to prevent that failing explanations can always salvage by ad hoc hypotheses, which would prevent any model to be falsified
- The elaborate hypotheses principle aims at the **generalizability of predictions** (as many instances as possible)

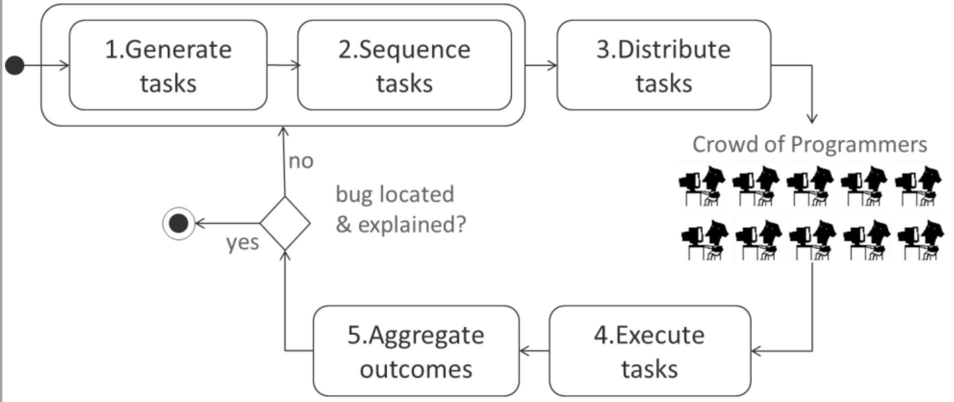
How to Optimally Allocate Bug Inspection Tasks to Minimize Cost and Maximize Accuracy?

<u>Approach</u>: causal and sequential decision models decide which tasks to generate and who should execute them.

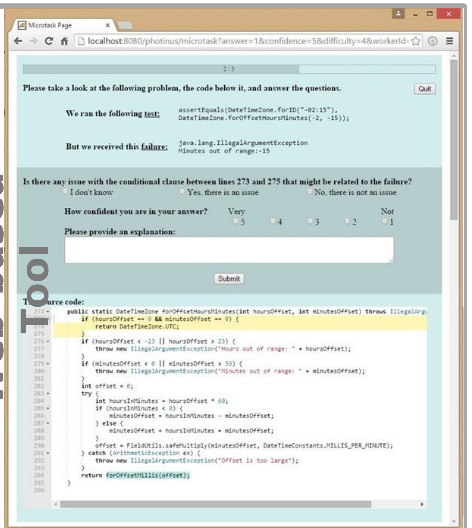




Overview of Crowdsourcing Inspection Tasks



Web-based Tool



Limit of the Falsification Principle

The escape to this conundrum is to be guided by a more fundamental goal of falsification [Popper 1945].

Note however, that falsification is not silver bullet either, because one cannot guarantee a unique mapping between generative processes (phenomenon), explanations (hypotheses) and models.

Definitions for Robustness

Bertrand Meyer [Meyer 1997] definitions:

Correctness : The ability of software products to perform their exact tasks, as defined by their specification.

Robustness : The ability of software systems to react appropriately to abnormal conditions.

Reliability : A concern encompassing correctness and **robustness**.

What are the abnormal conditions and how to detect and measure them?

To answer that in the context of Machine Learning models, we need to look at what is abnormal from the perspective of the user of predictions. The abnormal correspond to many categories of bias (next)

Essential and Accidental Changes

Accidental is a problem caused by the technology, the method, hence epistemological in nature.

Essential is problem inherent to the object, hence ontological in nature.

Philosophical Groundings

For more into these topics see Kant immanence concepts and Aristotle essential and accidental properties, which George Lakoff summarizes "make the thing what it is, and without which it would be not *that* kind of thing" [Wikipedia 2020]

Software systems Groundings

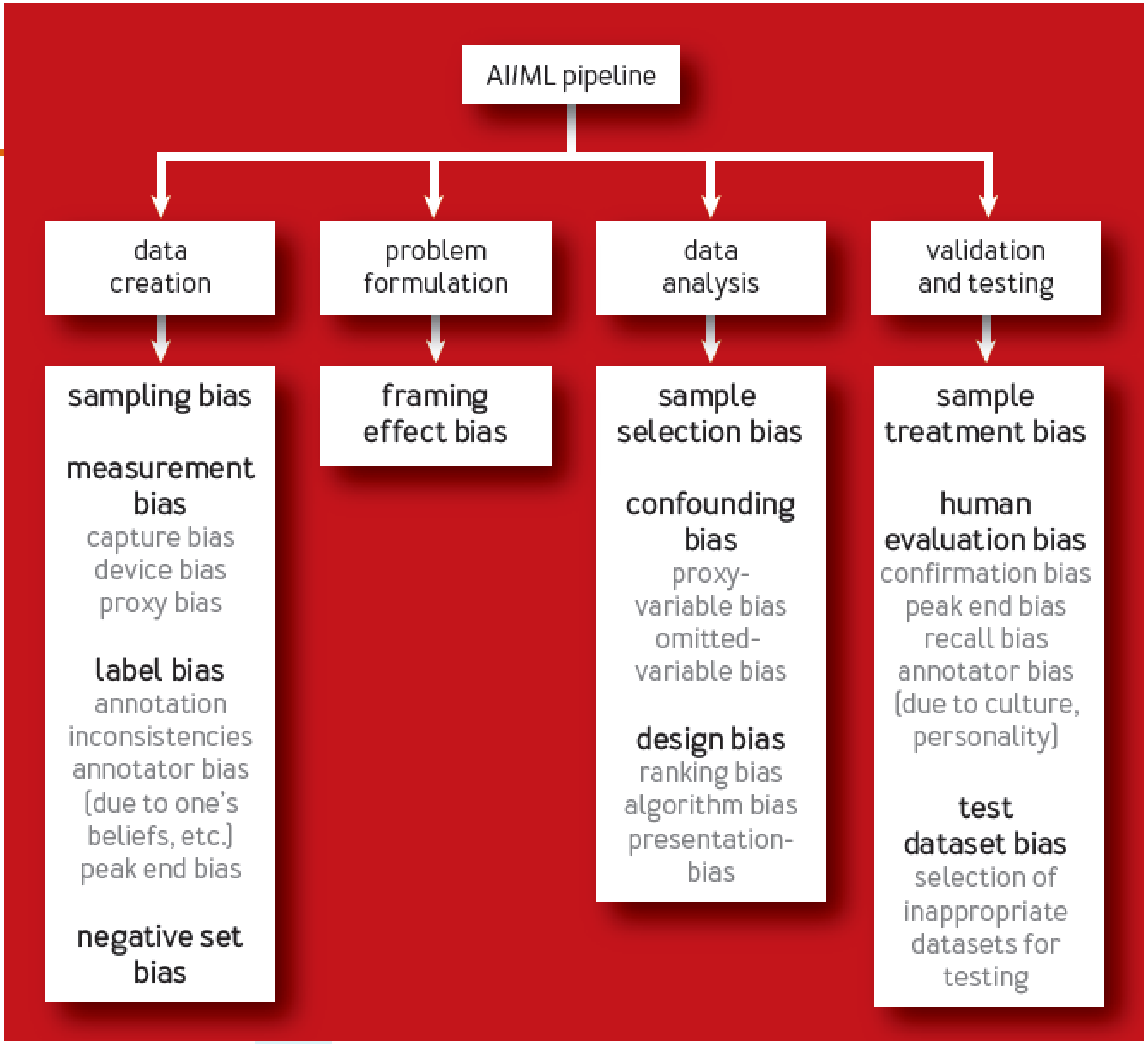
Fred Brooks in seminal paper No Silver-Bullet – Essence and Accidents in Software, proposed four characteristics that make developing software difficult: **changeability**, **invisibility**, **complexity**, and **conformity**.

Because they are essential, their effects can only be mitigated, not eliminated.

Taxonomy of Biases

- There are many reasons for an engineer to have a wrong model of the world (figure-1)
- These biases also impact users in very diverse ways.
- I am more interested on bias sample selection bias and confounding bias (under the data analysis)
- Before we delve into these bias, we need to answer the question, why simply getting more data does not solve the bias problem?
- The Reason: the bias-variance trade-off (next)

FIGURE 1: TAXONOMY OF BIAS TYPES ALONG THE AI PIPELINE



[Srinivasan & Chander 2021]

Implications to predictive models

Goal: Generalize data associations as predictive patterns

Assumptions: good data and observable patterns

Reality: sparse data and hidden states

- Sparse data (Essential limitation, cannot eliminate with better prediction models)
- Latent patterns (Accidental, can eliminate with better models)
  - Source – Misspecification

Not enough data or bad tunning of a model can make the concept drift more severe, as models might present strong bias (insensitive to crucial features) or high variance (too sensitive to noise).

- Under-specification (leads to bias-underfitting)
- Over-specification (leads to variance-overfitting)

Sources of Sparsity and Unobservability

Changes in the Data Generation Process:

- Covariate Shift (change in data distribution)
- Domain Shift (change in the state space)
- Concept Drift (change in the associations)

These changes are independent of the model, but the model might make the problem worse.

Goal: A robust model should have structures and conditions in place to mitigate the effect of these changes on the performance of the model.

Plausible Changes -> Sparsity + Observability -> Model performance

Robustness approaches

Robustness approaches involve simulating, measuring, and identifying the situations (e.g., a given environment change) in which the prediction models will not be robust.

In the next slides, I will detail three families of these robustness models:

- **Generative Models** – rely on methods to generate data that can simulate the environmental conditions that challenge prediction model robustness
- **Structural Models** – rely on methods to capture hidden and observable states and their associations, which allow to generate hypotheses about spurious correlations
- **Validation Models** – rely on methods to measure the outcome of model under various environmental conditions

Generative Models

Anticipate the effect of changes (approximate changes if nonstationary process, determine performance envelopes of performance to detect a systematic trend that will breach the envelope).

Data augmentation (Model-Based Simulation, Data Transformations)

Oversampling

Probability Weighting

Data Splitting

Train-Test-Validation Split

Cross-Validation

Validation models (1)

Models of validation allow to measure of the performance of the prediction models. Because these measurements consist of well-defined metrics, it their outcome also allows to compare more models.

Entropy-based methods

Information Criteria are methods based on entropy

WAIC is the most modern method and currently preferred over other IC methods like AIC, BIC, DIC.

Pareto-Smooth LOOC also a modern method, which produce comparable results to WAIC. The best practice is to always execute both methods to check for inconsistencies.

Robustness approaches
Validation models (2)

Definition: Intervention Models determine how to modify the inputs in meaningful ways to discover the frontier when the prediction models start producing predictions with unacceptable accuracy.

Sensitivity Analysis tests if changes in the putative causes (inputs) should be accompanied by expected changes in the effects (outcomes). This requires a proper definition of causes effects and the mechanisms that connect both. Essentially, this allow to test the model w.r.t. to the sensitivity to unobserved confounders [Franks, D'Amour & Feller 2019][Wang & Blei 2018]

Transductive Tests [Chapelle et al. 2009] consists of measuring the performance of the prediction model with datasets, that were not used during training/validation, but that still resemble to the same distribution of the training data.

Inductive Tests consists of measuring the performance of the prediction model with datasets, that were not used during training/validation, but that still resemble to the same distribution of the training data.

Ablation Studies

Structural models

Model-free methods do not allow to know what went wrong but preclude assumptions about the unknowns. Robustness requirements are concerned to avoid harm from catastrophic failure (extreme events). In this case the robustness requirements involve fail-safe or degraded performance.

Model-based methods make strong assumptions about the unknowns which could be justifiable when robustness requirements assume stationarity or smooth nonstationary changes. i.e., no abrupt changes that invalidate the past completely, for instance, one expect that fundamental model properties like the Markov property and Causal Markov condition will hold.

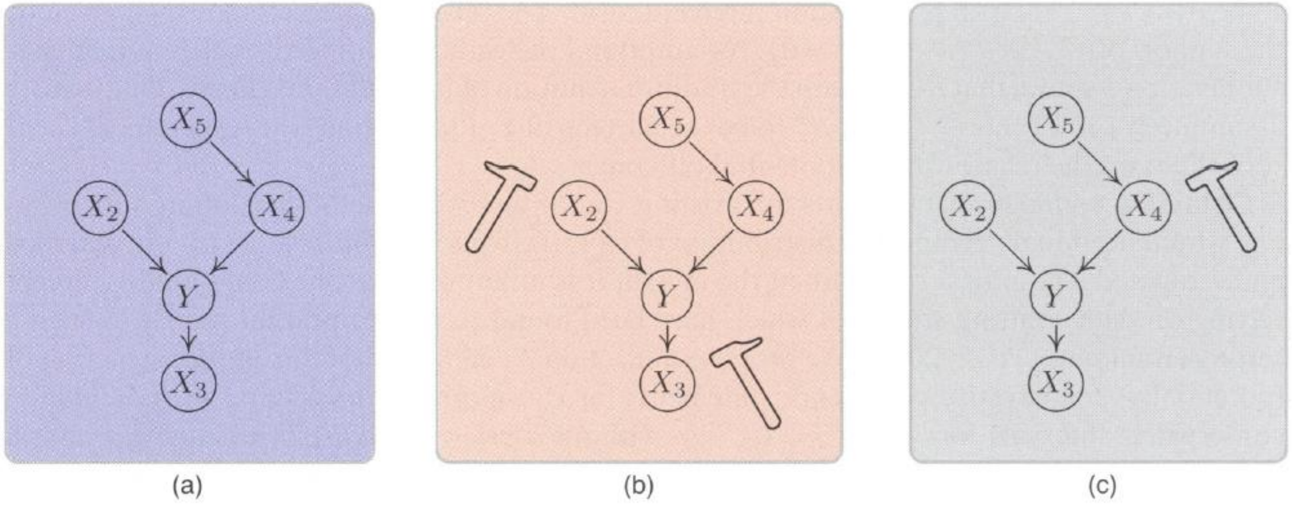
- **Generative models** approximate the process (phenomenon) that generates the data.
- **Latent models** approximate the hidden states and their relationships with the observable states, e.g., Hidden Markov Models, Partially Observable Markov Decision Processes, Causal Models.
- **Model invariant methods** aims at discovering elements of the model (usually the internal associations) that do not change significantly across environments.

Model Invariance Discovery Methods

Empirical Risk Minimization (bias-variance trade-off tunning)

Invariant Risk Minimization (does not assume causal mechanisms)

Invariant Causal Prediction



[Peters 2016]

Requirements:

Environments should present shifts that are large enough to expose the effect of latent confounder, but not too large that the causal mechanism is invalidated.

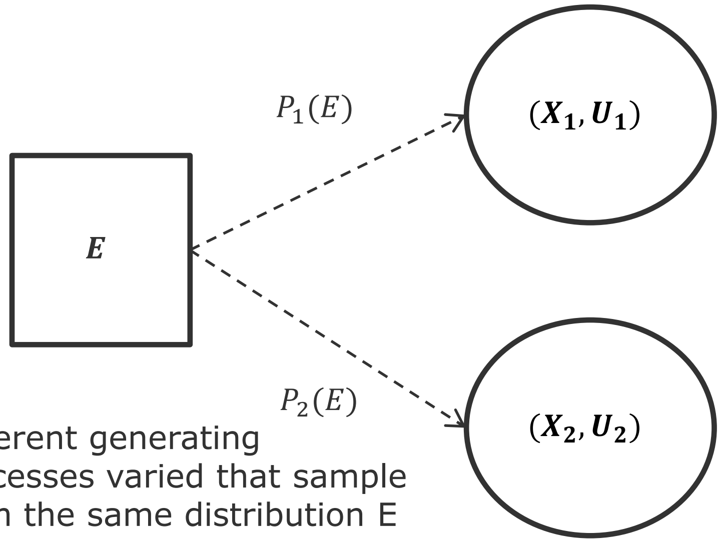
Strong reliance on the how on the correctness of model, i.e., the model presenting all the covariate coefficients that correspond to the effect of these covariates in the treatment assignment.

So, because one might not obtain perfect ignorability, we can mitigate that by applying adjustment and reweighting techniques before we fit the multi-variate regression

Multi-Environments Setting

The data generation process changes across environments E_i

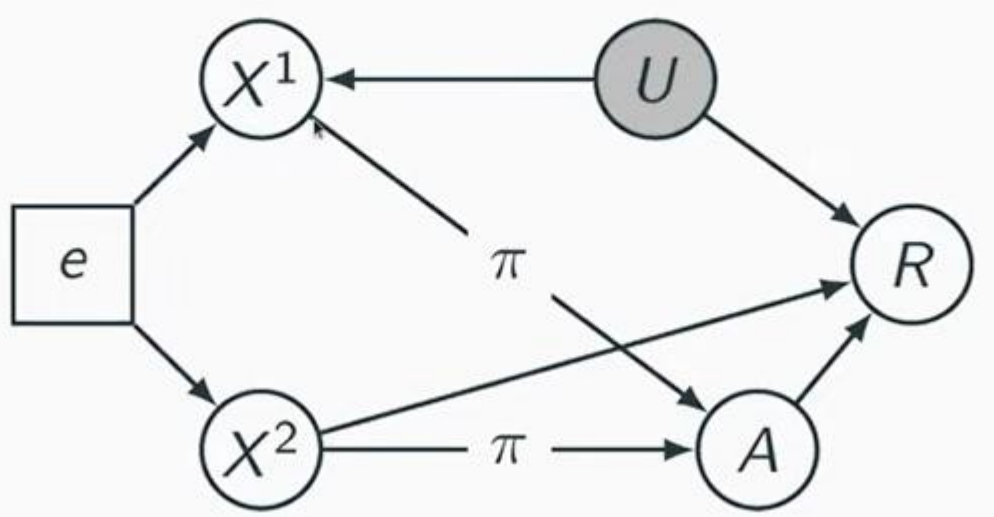
This means that each environment produces a different observable contexts X_i and unobservable contexts U_i.



Different generating processes varied that sample from the same distribution E

Robust Policy π

- Observed contexts X
- Unobserved contexts U
- Actions $A \in \{a_1, \ldots, a_k\}$
- Reward R
- Environments $\mathcal{E} = \{e_1, \ldots, e_L\}$



Sampling procedure

1. Random contexts are drawn: $(X_i, U_i) \sim \mathbb{P}^e_{(X,U)}$
2. Policy selects action: $A_i \sim \pi(A_i|X_i)$
3. Reward is drawn: $R_i \sim \mathbb{P}_{R|X_i,U_i,A_i}$

Goal Learn a policy π that maximizes the worst-case expected reward

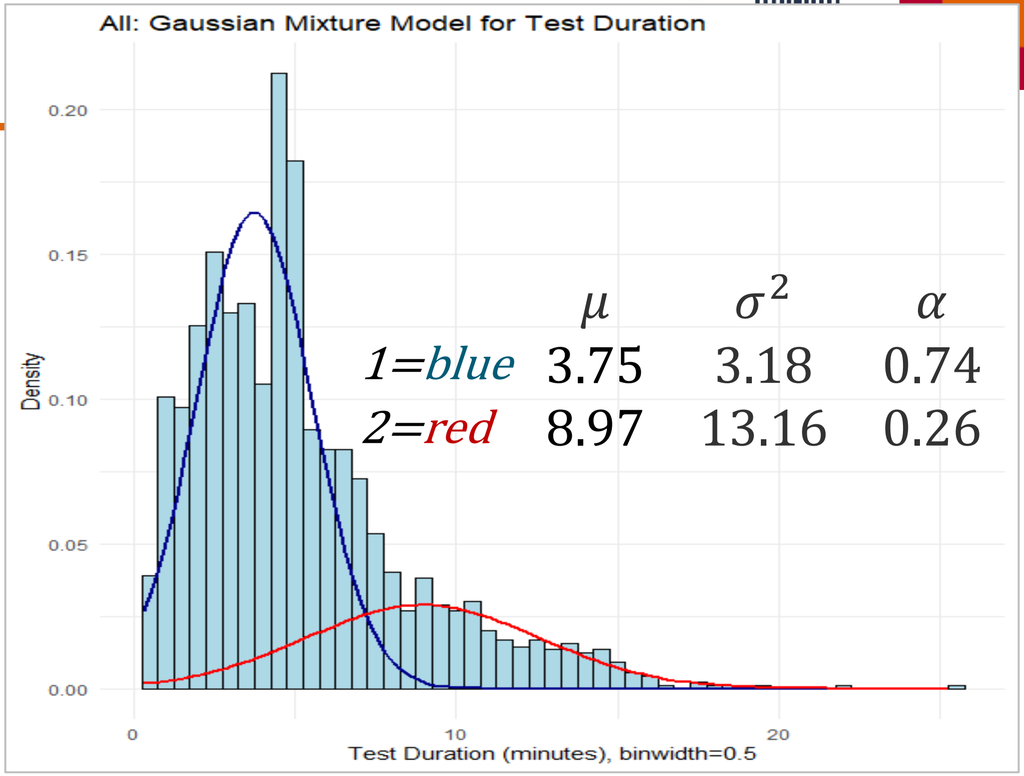
$$V^{\mathcal{E}}(\pi) := \inf_{e \in \mathcal{E}} \mathbb{E}^{\pi,e}[R]$$

Gaussian Mixture Model with the Expectation Maximization algorithm

E-STEP $\mathcal{N}(x_i|\mu_{k_j},\sigma^2_{k_j})$ α_{k_j}

$$P(x_i \in k_j|x_i) = \frac{P(x_i|x_i \in k_j)P(k_j)}{P(x_i)}$$

$$\sum_{k=1}^{K} \alpha_k \mathcal{N}(x_i|\mu_k,\sigma^2_k)$$



M-STEP

$$\mu_k = \frac{\sum_i^N P(x_i \in k_j|x_i)x_i}{\sum_i^N P(x_i \in k_j|x_i)} \quad \sigma^2_k = \frac{\sum_i^N P(x_i \in k_j|x_i)(x_i - \mu_k)^2}{\sum_i^N P(x_i \in k_j|x_i)} \quad \alpha_k = \frac{\sum_i^N P(x_i \in k_j|x_i)}{N}$$

Initialize prior using K-Means, which will give us:

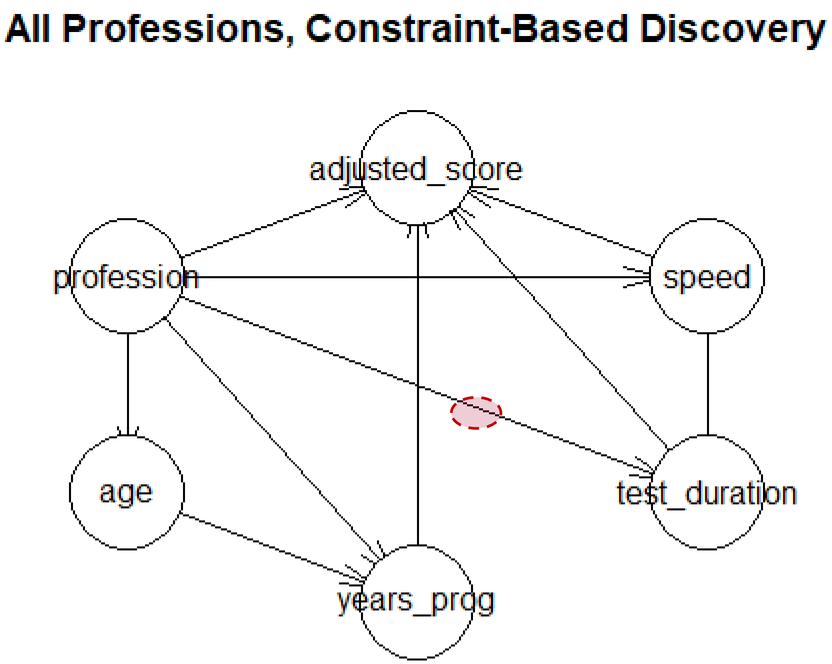
	μ	σ^2	α
1=blue	3.68	2.75	0.77
2=red	10.1	7.73	0.33

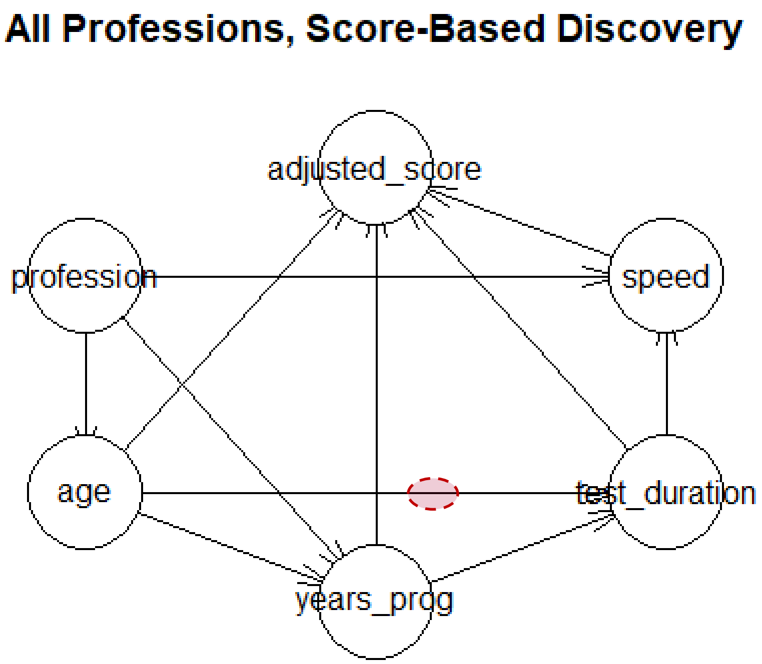
Loop between E-Step and M-Step until convergence, i.e., $\Delta\mu_k < 10^{-6}$

Membership to cluster *kj*

$$E[z_{i,j}] = \frac{P(x = x_i|\mu = \mu_j)}{\sum_{m=1}^{k} P(x = x_i|\mu = \mu_m)}$$

Causal graphs (adjusted score)





Non-Invariant Associations

Causal graphs (adjusted score)