ML Conf Berlin

DOMINO

Nikolay Manchev
Head of Data Science for EMEA

Twitter: @nikolaymanchev
https://www.linkedin.com/in/nikolaymanchev/

Underspecified pipelines: Why good models underperform in production

December 2021

Frontiers of Data Science

Machine Learning models often exhibit poor performance when applied to real-world problems. Surprisingly, this effect is observed even when rigorous validation procedures have been followed, and the model generalisation measured under laboratory conditions seems acceptable.

Recent research conducted by Google identifies underspecified pipelines as the main culprit for models that score well on hold-out sets but perform poorly once operationalised.

In this advanced talk we will go over the research results, explain the theoretical foundations of the underspecification effect, discuss several case studies using ensemble and deep learning models, and provide suggestions for training models with credible inductive biases.

Nikolay Manchev
Head of Data Science for EMEA

INCREASING MODEL PERFORMANCE BY ADDRESSING PIPELINE UNDERSPECIFICATION

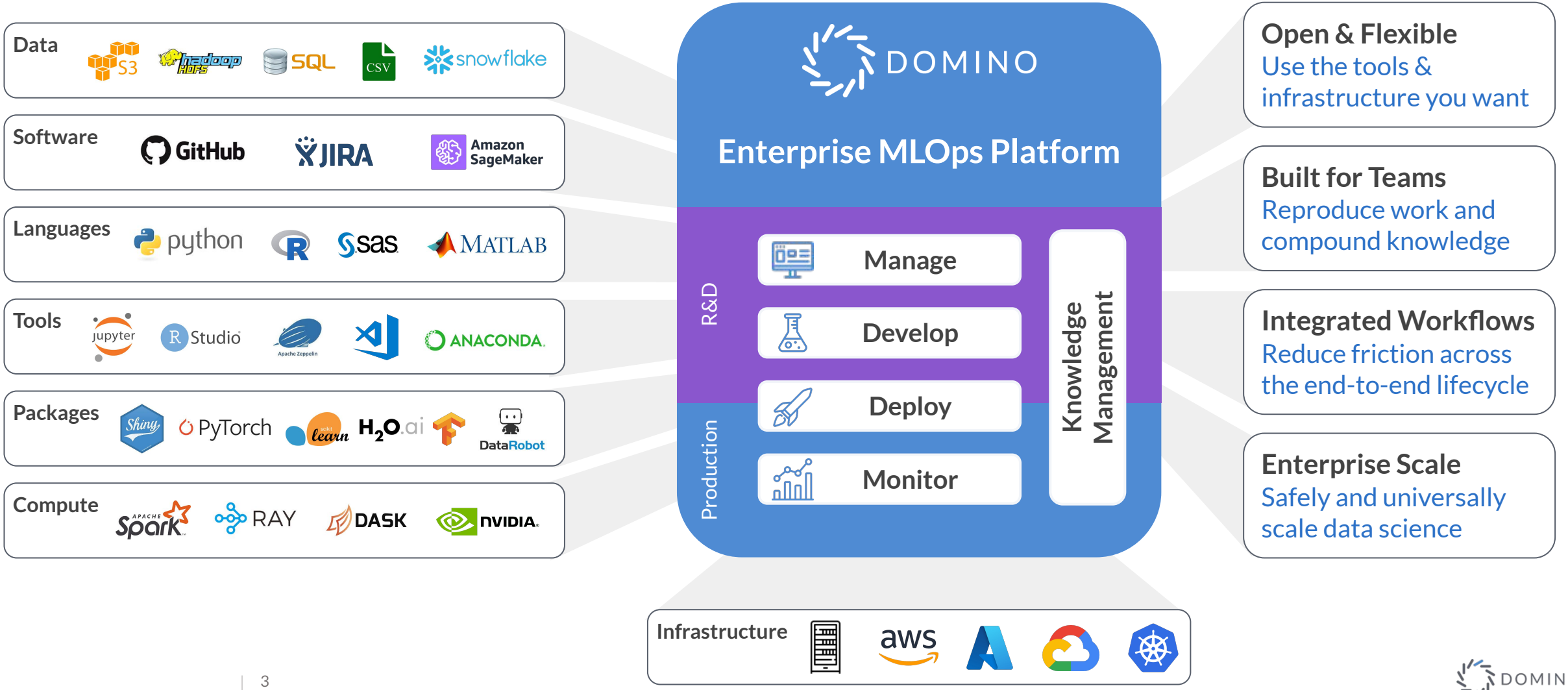
Feb 2020

DOMINO

The Domino Enterprise MLOps Platform

Trusted by 20% of the Fortune 100

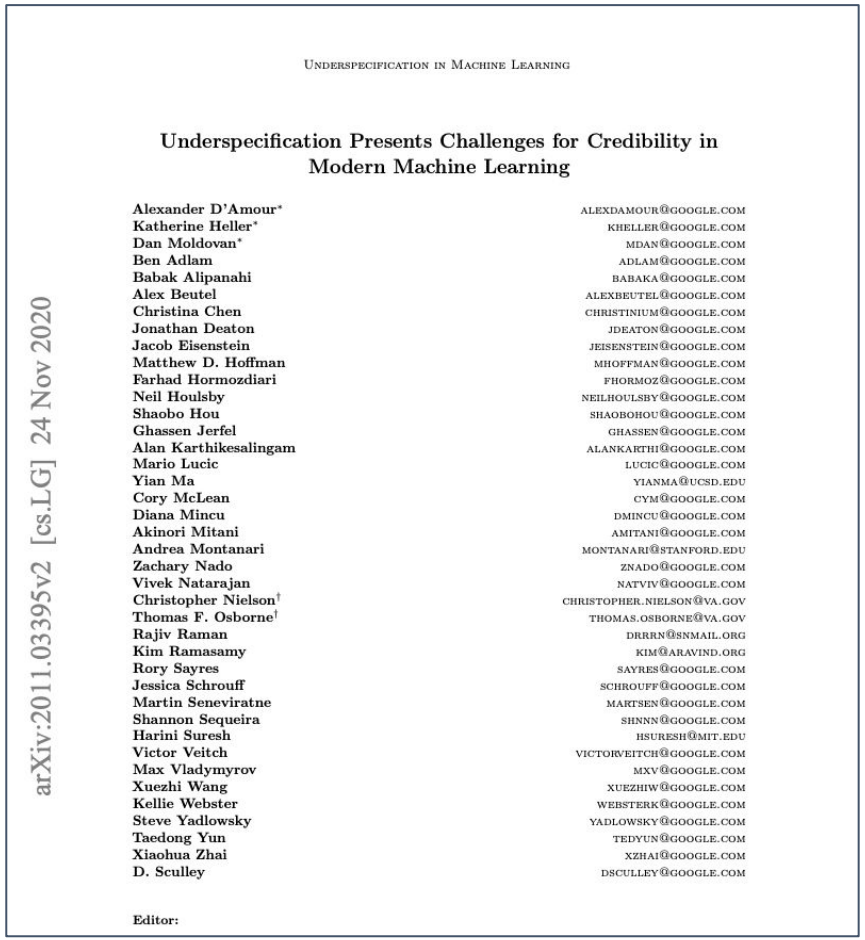
2 of largest	2 of top-5	3 of top-5	4 of top-10
global banks	health insurers	ratings agencies	pharma companies



Talk Details

Based on the following Google paper

- Released on 24th Nov
- Available on arXiv https://arxiv.org/pdf/2011.03395.pdf



Underspecification Presents Challenges for Credibility in Modern Machine Learning

Alexander D'Amour* ALEXDAMOUR@GOOGLE.COM
Katherine Heller* KHELLER@GOOGLE.COM
Dan Moldovan* MDAN@GOOGLE.COM
Ben Adlam ADLAM@GOOGLE.COM
Babak Alipanahi BABAKA@GOOGLE.COM
Alex Beutel ALEXBEUTEL@GOOGLE.COM
Christina Chen CHRISTINIUM@GOOGLE.COM
Jonathan Deaton JDEATON@GOOGLE.COM
Jacob Eisenstein JEISENSTEIN@GOOGLE.COM
Matthew D. Hoffman MHOFFMAN@GOOGLE.COM
Farhad Hormozdiari FHORMOZ@GOOGLE.COM
Neil Houlsby NEILHOULSBY@GOOGLE.COM
Shaobo Hou SHAOBOHOU@GOOGLE.COM
Ghassen Jerfel GHASSEN@GOOGLE.COM
Alan Karthikesalingam ALANKARTHI@GOOGLE.COM
Mario Lucic LUCIC@GOOGLE.COM
Yian Ma YIANMA@UCSD.EDU
Cory McLean CYM@GOOGLE.COM
Diana Mincu DMINCU@GOOGLE.COM
Akinori Mitani AMITANI@GOOGLE.COM
Andrea Montanari MONTANARI@STANFORD.EDU
Zachary Nado ZNADO@GOOGLE.COM
Vivek Natarajan NATVIV@GOOGLE.COM
Christopher Nielson† CHRISTOPHER.NIELSON@VA.GOV
Thomas F. Osborne† THOMAS.OSBORNE@VA.GOV
Rajiv Raman DRRRN@SNMAIL.ORG
Kim Ramasamy KIM@ARAVIND.ORG
Rory Sayres SAYRES@GOOGLE.COM
Jessica Schrouff SCHROUFF@GOOGLE.COM
Martin Seneviratne MARTSEN@GOOGLE.COM
Shannon Sequeira SHNNN@GOOGLE.COM
Harini Suresh HSURESH@MIT.EDU
Victor Veitch VICTORVEITCH@GOOGLE.COM
Max Vladymyrov MXV@GOOGLE.COM
Xuezhi Wang XUEZHIW@GOOGLE.COM
Kellie Webster WEBSTERK@GOOGLE.COM
Steve Yadlowsky YADLOWSKY@GOOGLE.COM
Taedong Yun TEDYUN@GOOGLE.COM
Xiaohua Zhai XZHAI@GOOGLE.COM
D. Sculley DSCULLEY@GOOGLE.COM

Editor:

DOMINO

Agenda

- Definition of underspecification
- Example 1: SIR epidemiological model (+ demo)
- Theoretical analysis of underspecification
- Example 2: Underspecified XGBoost model (demo)
- Underspecification in Deep Learning
- Detecting underspecification via stress testing
- Mitigation strategies

Underspecification

Definition

"An ML pipeline is underspecified when it can return **many predictors with equivalently strong held-out performance** in the training domain"

...

"Predictors returned by underspecified pipelines are often treated as equivalent based on their training domain performance, but we show here that such predictors can behave very differently in deployment domains. "

...

"This ambiguity can lead to instability and poor model behavior in practice, and is a distinct failure mode from previously identified issues arising from structural mismatch between training and deployment domains."

Multiple predictors

Simple optimisation problem

$$X = \{x_1, x_2, \ldots, x_N\}^T \quad y = \{y_1, y_2, \ldots, y_N\}^T$$

$$\hat{y} = Xw$$

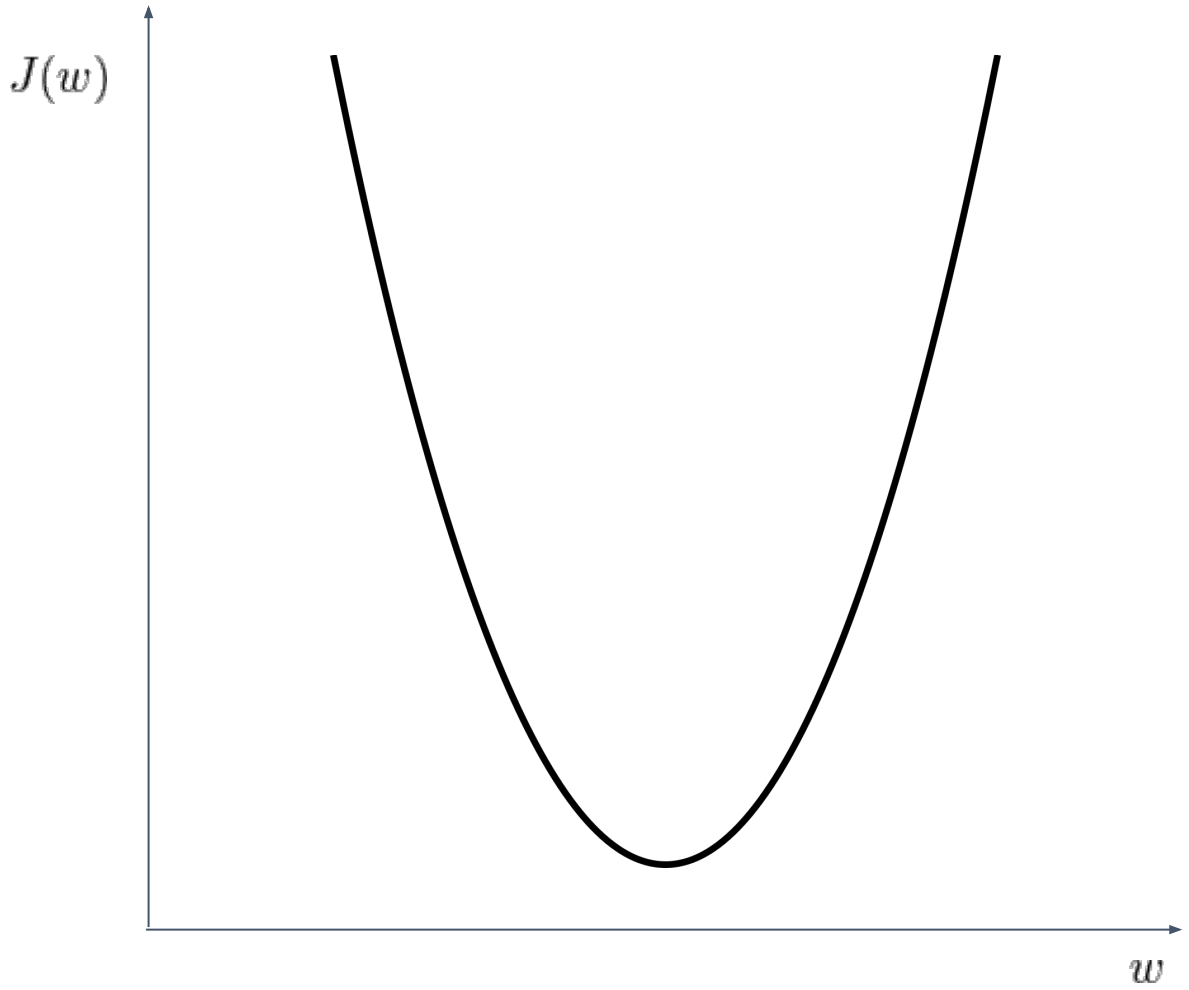
$$J(w) = \frac{1}{2N}\sum_{i=1}^{N}(\hat{y}_i - y)^2$$

Assuming no closed-form solution -> GD

```
repeat until convergence{
```

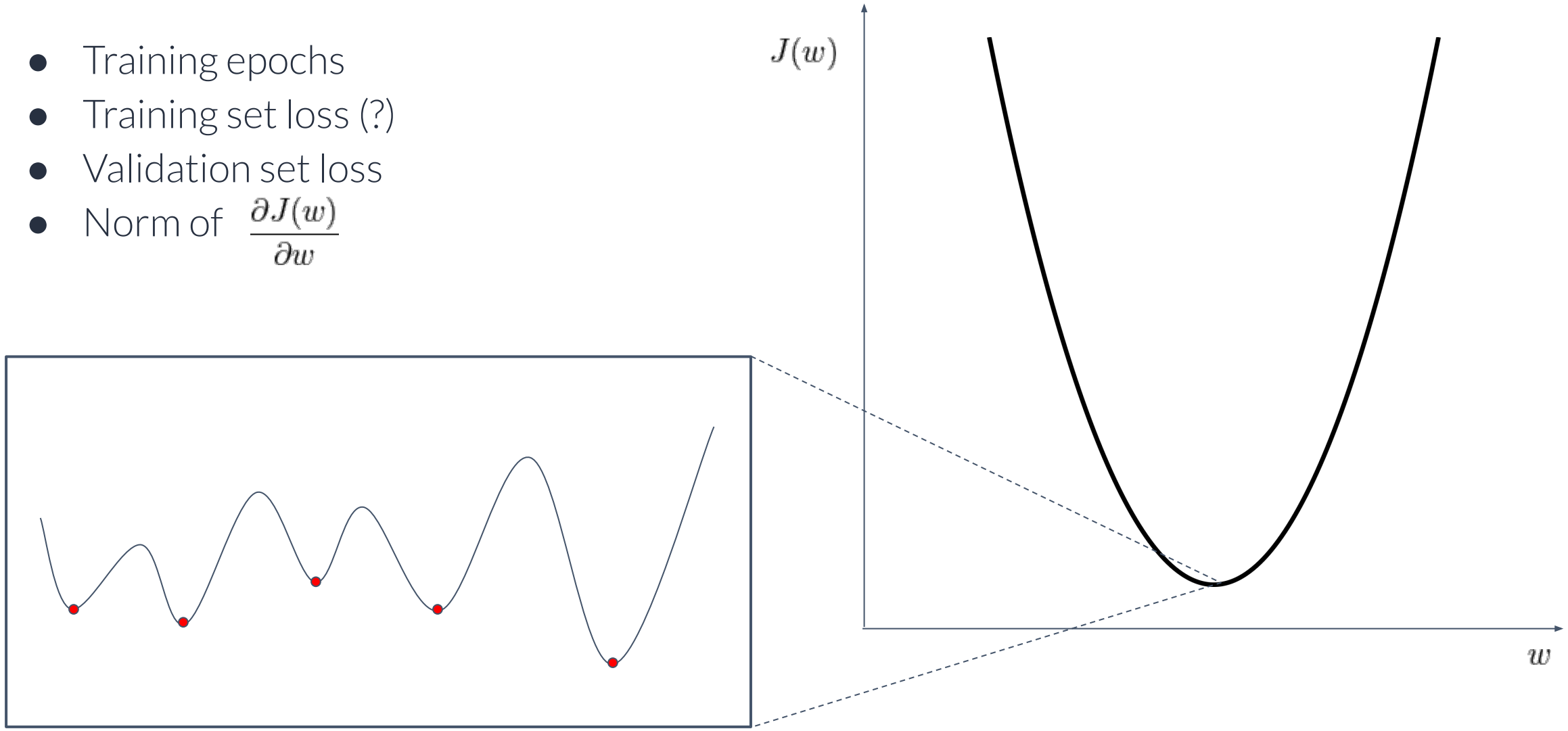
$$w := w - \alpha\frac{\partial J(w)}{\partial w}$$

```
}
```



Convergence criteria

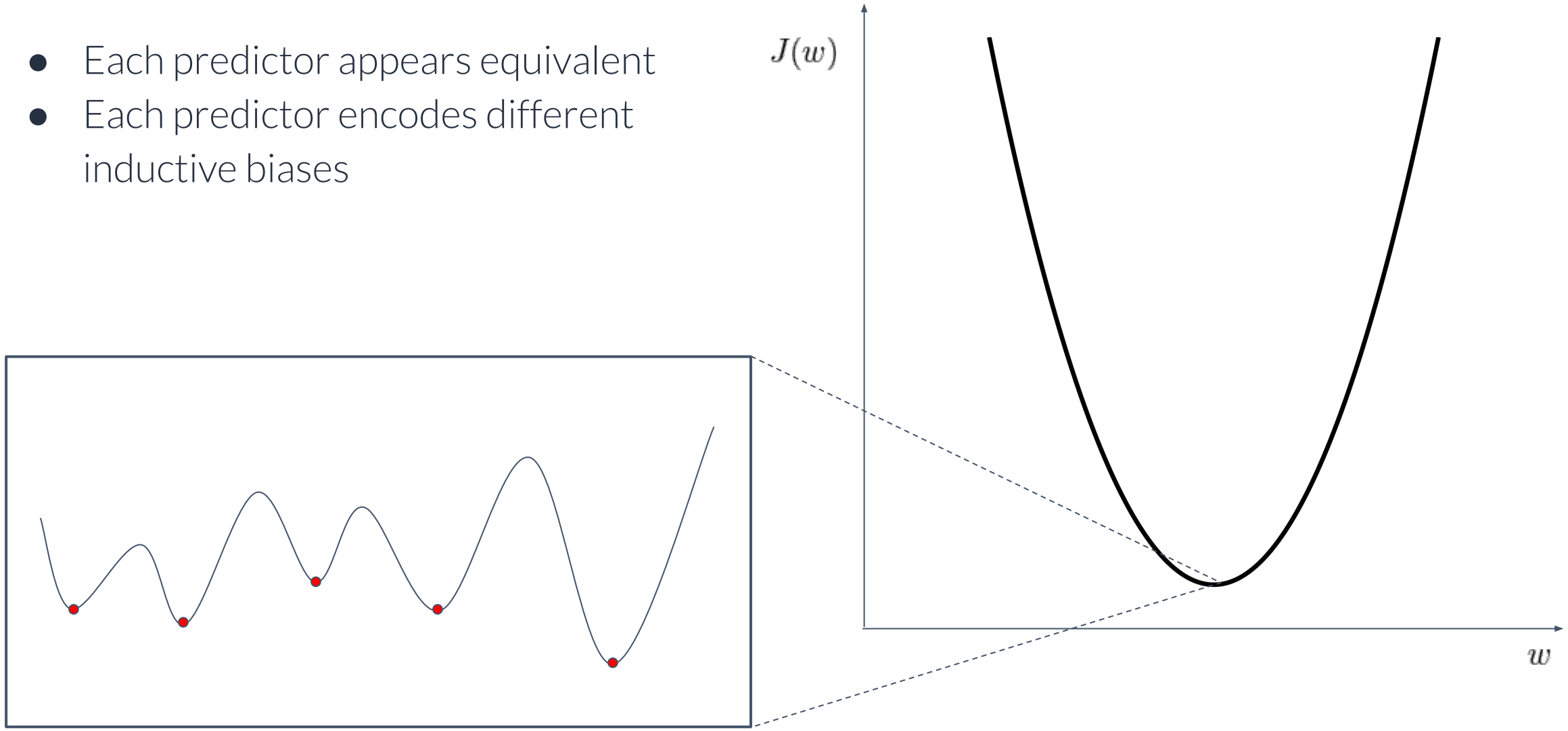
- Training epochs
- Training set loss (?)
- Validation set loss
- Norm of $\frac{\partial J(w)}{\partial w}$



DOMINO

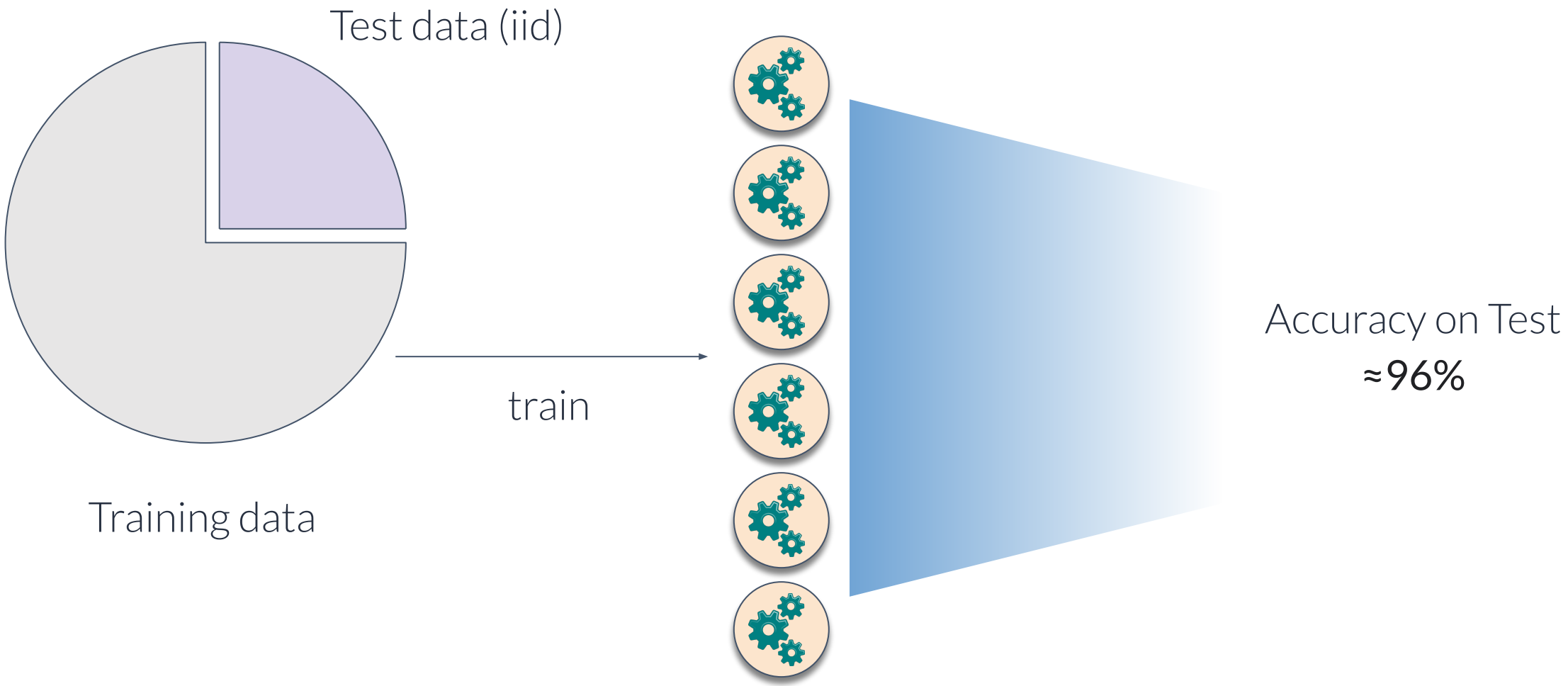
The problem

- Each predictor appears equivalent
- Each predictor encodes different inductive biases



DOMINO

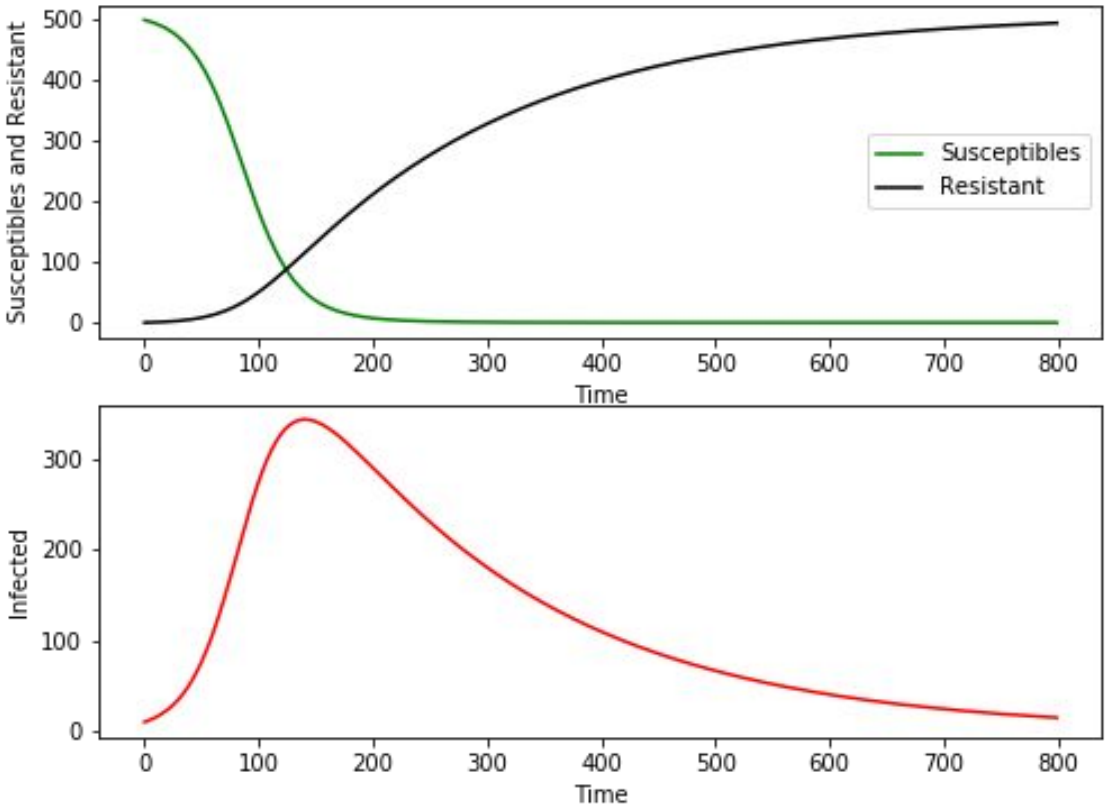
The standard recipe



DOMINO

Example 1: SIR model

States in an SIR epidemic model image by Arthur Goldberg - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=95460351



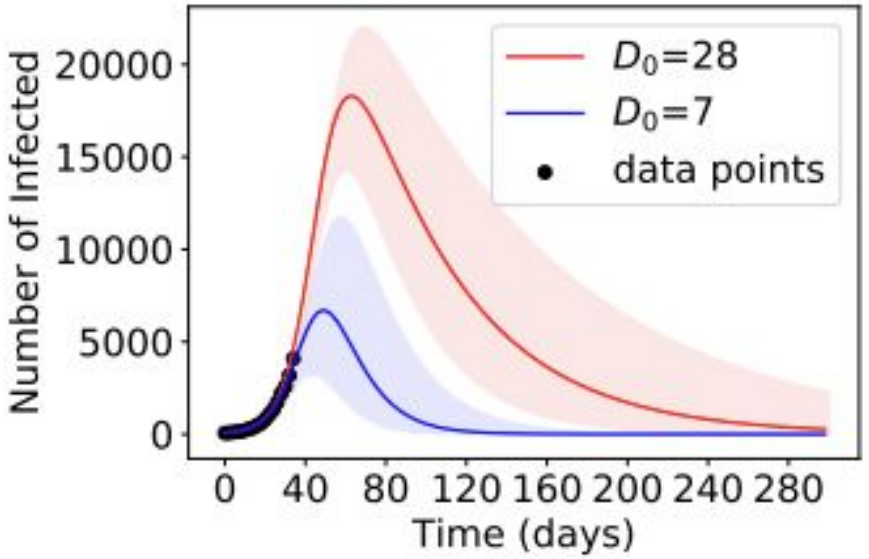
- SIR is a very simple epidemiological model
  - **S** - number of susceptible individuals
  - **I** - number of infectious individuals
  - **R** - number of resistant individuals

$$\frac{\partial I}{\partial t} = \frac{\beta I S}{N} - \frac{I}{D}$$

DOMINO

Image from A. D'Amour, K. Heller, D. Moldovan, et al
https://arxiv.org/pdf/2011.03395.pdf

Example 1: SIR model



Simulation of a forecasting task

- Generate trajectory from a full time course T, but only train on
- During the early stages T_{obs} is small
- There are many (β, D) that fit T_{obs}
- Arbitrary choices in the learning process determine which set of observation-equivalent parameters are returned by the learning algorithm
- Epidemiological modelling relies on domain knowledge

Theoretical analysis

Regression problem

$$\{(\boldsymbol{x}_i, y_i)\}_{i<n}, y_i \in \mathbb{R}, \boldsymbol{x}_i \in \mathbb{R}^d$$

$$f(x) = \Theta^T \sigma(W\boldsymbol{x}), W \in \mathbb{R}^{N\times d}, \boldsymbol{w}_i \sim U, \|\boldsymbol{w}_i\|_2 = 1$$

$$\min \quad \|\Theta\|$$

$$\text{s.t.} \quad f_\tau(\boldsymbol{x}_i) = y_i, \forall i$$

$$R(W, Q) = \mathbb{E}_{(X,Y)\sim Q}(Y - \hat{\Theta}(W)\sigma(WX))^2$$

$$\exists \quad W_1 \perp\!\!\!\perp W_2 :$$

$$R(W_1, P) \approx R(W_2, P)$$

$$f_{W_1}(x) \perp f_{W_2}(x), x \in P\Delta$$

Underspecification in Deep Learning

Demonstrate underspecification

- Construct an ensemble of predictors by perturbing parts of the pipeline
- Confirm that models in the ensemble have near-equivalent iid performance
- Evaluate the ensemble on one or more application-specific stress tests
  - variability in stress test performance indicates underspecification

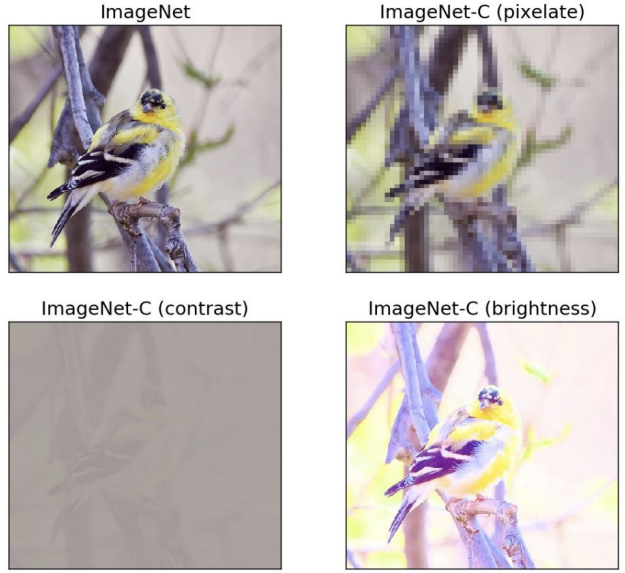
Stress test

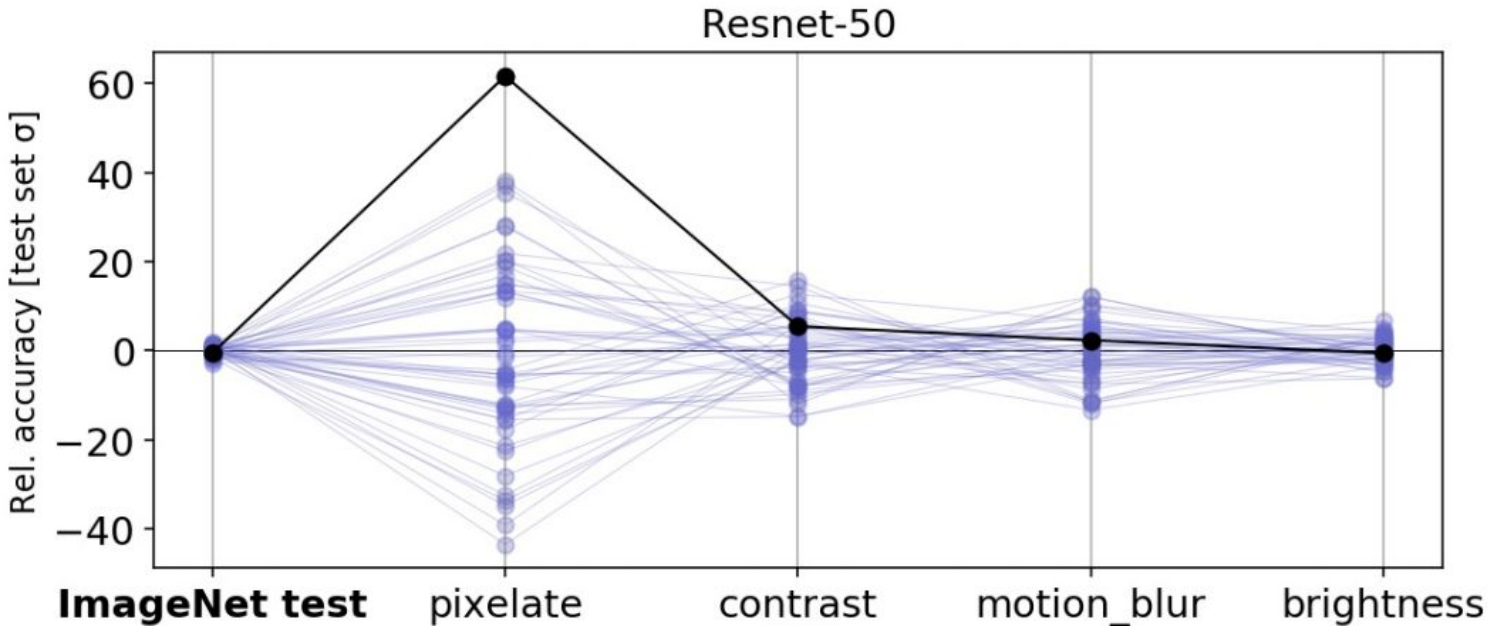
- Empirical evaluations that probe the model's inductive biases on practically relevant dimensions
- Evaluations that probe a predictor by observing its outputs on specifically designed inputs

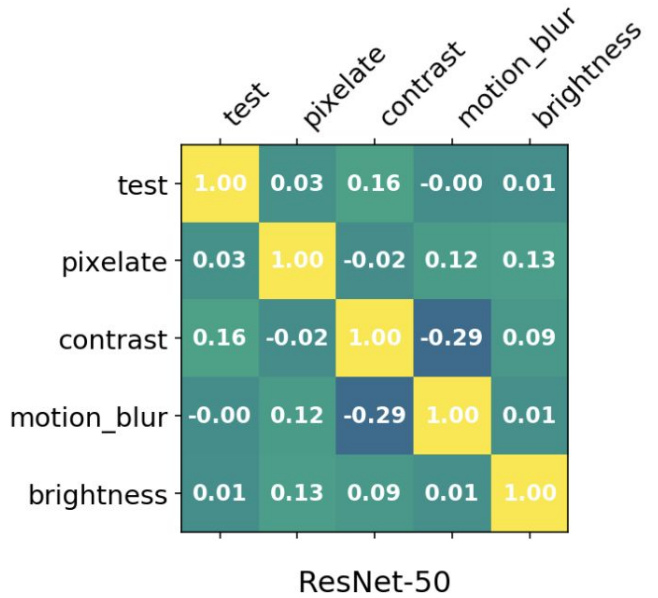
ImageNet test

An ensemble of 50 ResNet-50 models

- Trained on ImageNet
- Identical pipelines, different seed
- 75.9% ± 0.11 top-1 accuracy
- Tested on ImageNet-C*







* Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In International Conference on Learning Representations, 2019.

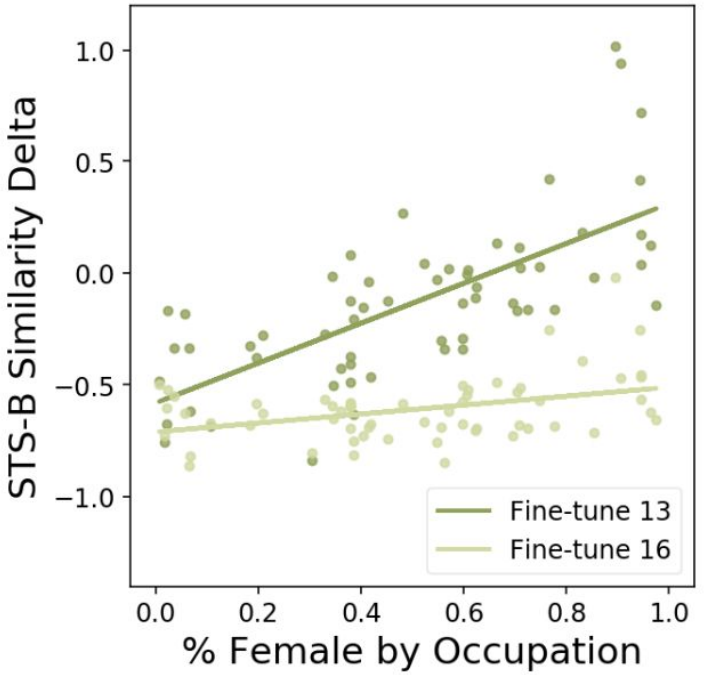
DOMINO

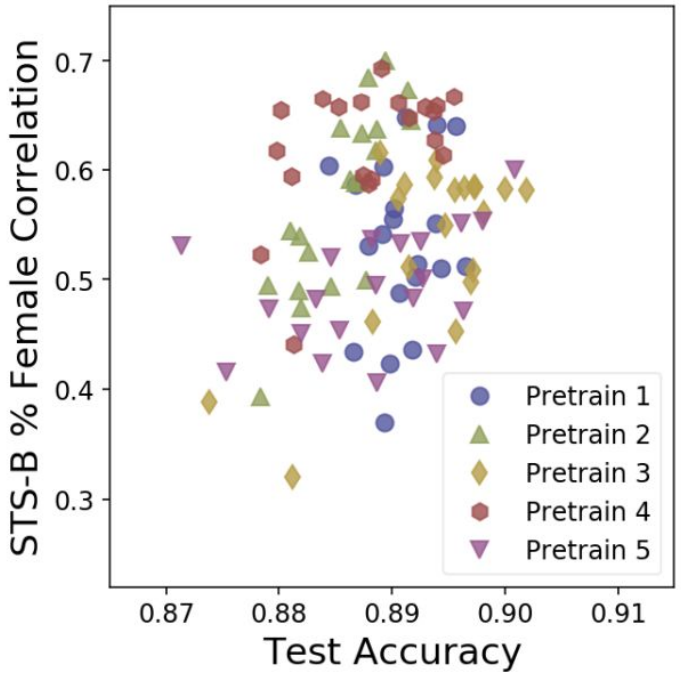
NLP test

Pretrain unsupervised / fine-tune using labeled data

- train 5 instances of BERT
  - Wikipedia and BookCorpus data
  - 340 million parameters
- STS task
  - fine-tuning BERT checkpoints on the STS-B benchmark
  - consistent accuracy matching Devlin et al. (2019)
  - measure reliance on gendered correlations

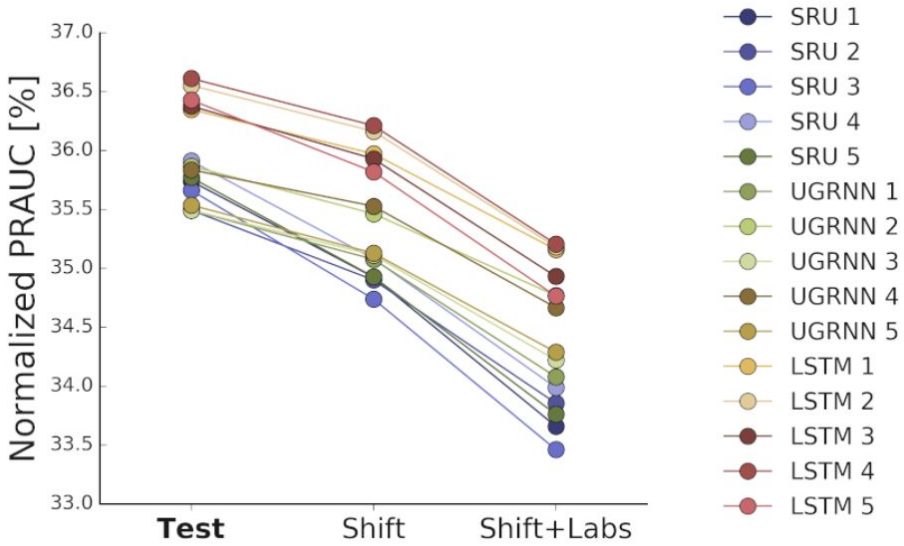
Conclusion: Reliance on gender correlation is affected by random initialisation





Other tests

- RNN on Electronic Health Record (EHR) data to predict predict onset of acute kidney injury
  - disentangle physiological signals from operational factors
  - embedding layers + 3 layer-stacked RNN + a dense layer for prediction
  -
- LSTM
- ...



DOMINO

Orchestrating stress tests

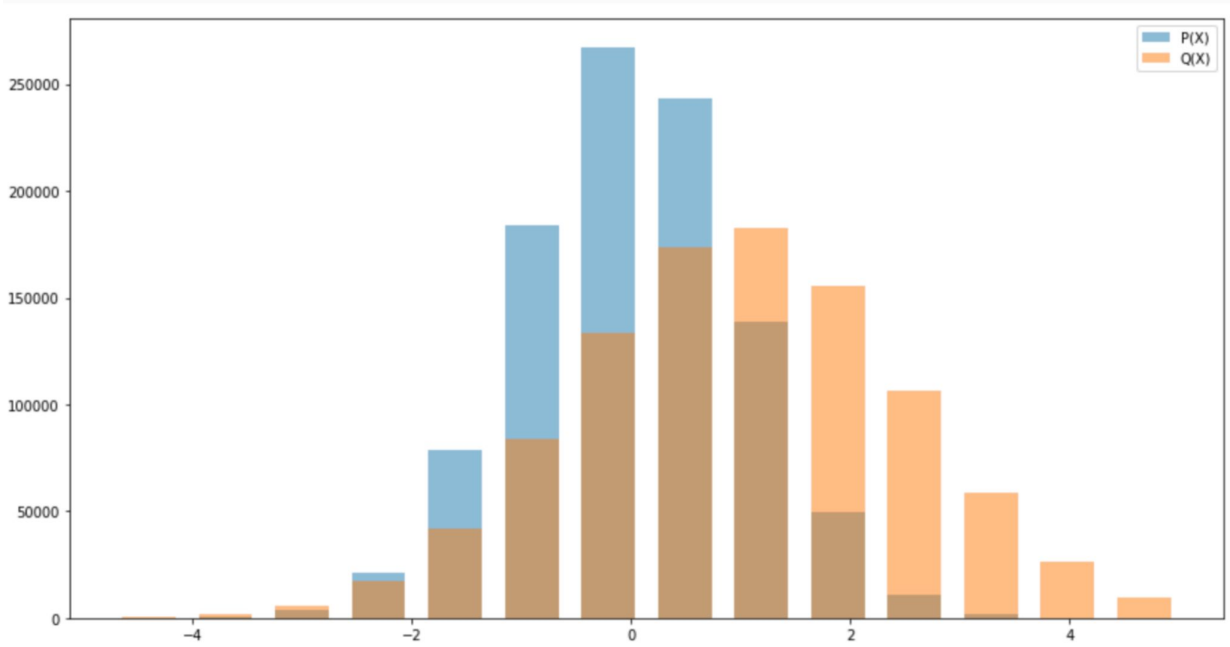
Three types suggested in the paper

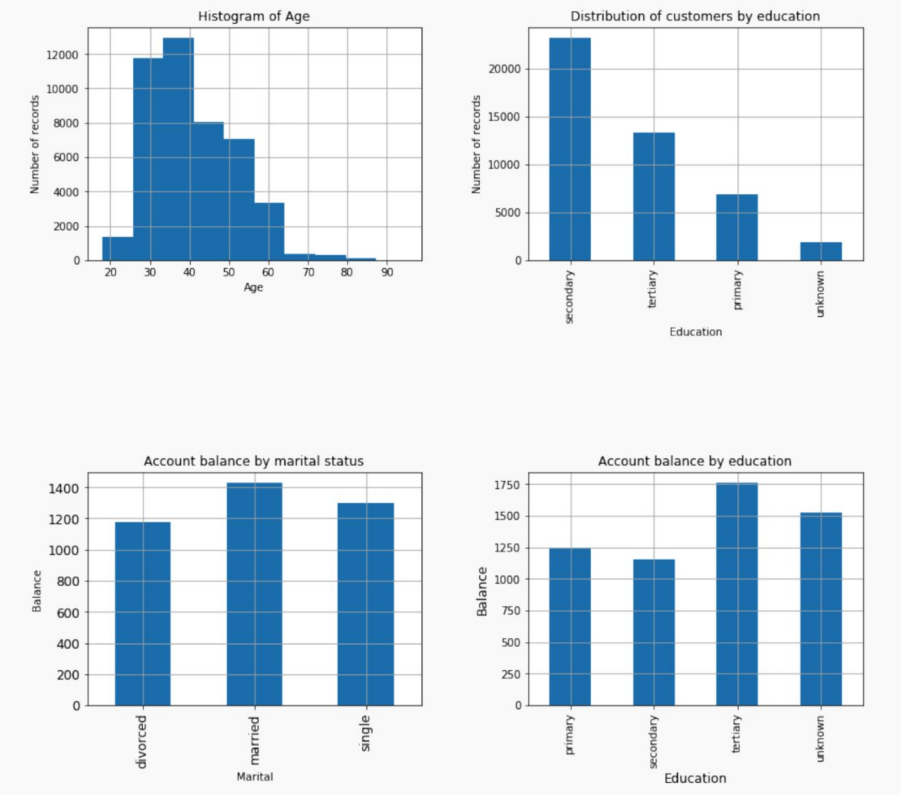
- **Stratified Performance Evaluations**
  - stratify based on feature A
  - calculate a performance metric and compare across strata
- **Shifted Performance Evaluations**
  - define a new data distribution $P' \neq P$
  - evaluate performance on test
- **Contrastive Evaluations**
  - localized analysis of particular inductive biases
  - check if a particular modification of the input x causes the output of the model to change in unexpected ways

"We're trying to make our stress tests more realistic."

Constant monitoring

- **Monitor model performance**
- **Monitor for concept drift**





DOMINO

Summary

- Underspecification is a **key failure mode** for machine learning models to encode generalizable inductive biases
- Distinct from generalization failures due to structural mismatch between training and deployment domains.
- Decisions are **determined by arbitrary choices** such as the random seed used for parameter initialization
- Extreme complexity of modern ML models ensures that some aspect of the model will almost certainly be underspecified
- Specifying **selection criteria** or constraints on F
- Designing stress tests is a major challenge
- Monitor your models

Thank You