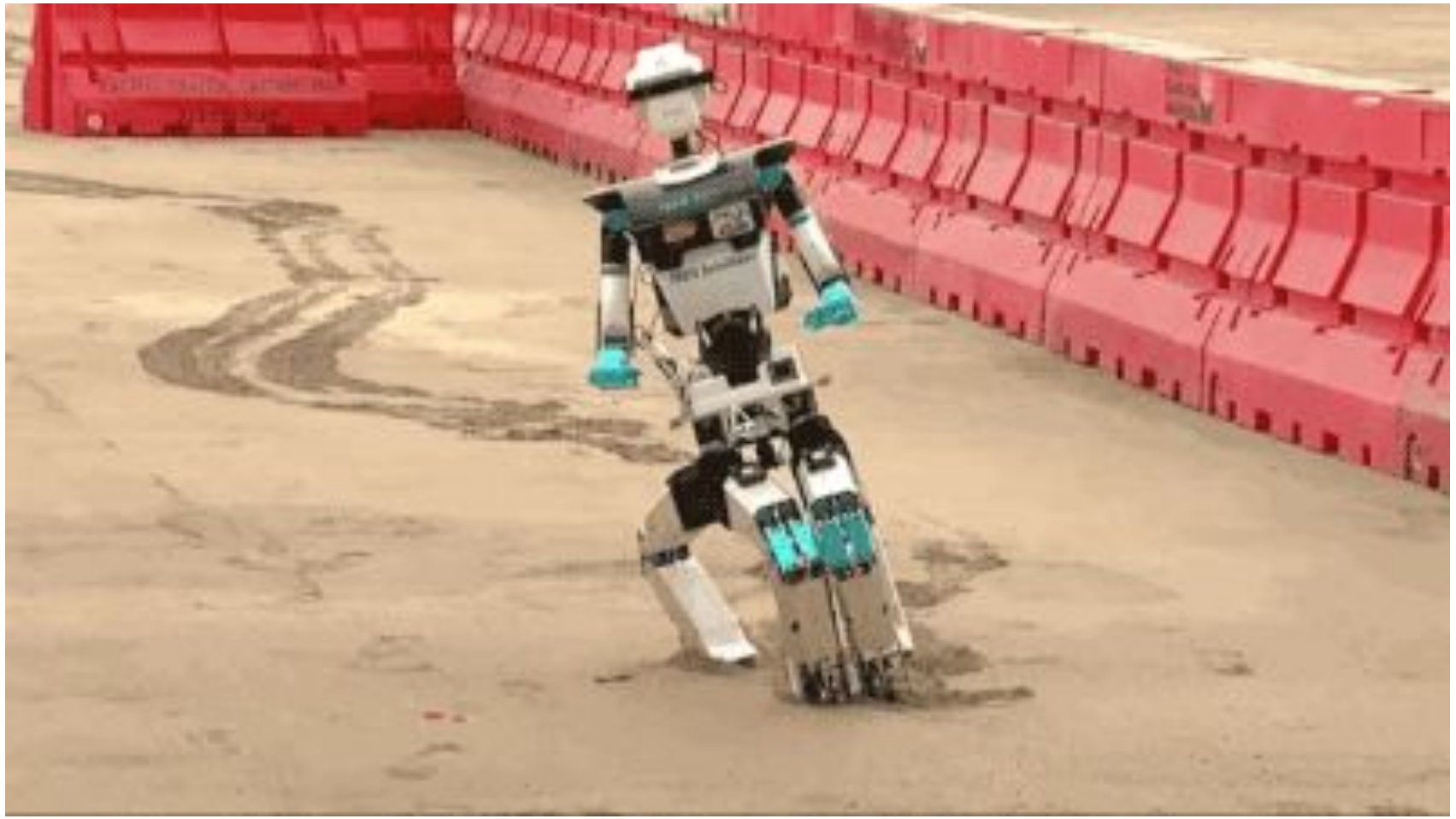
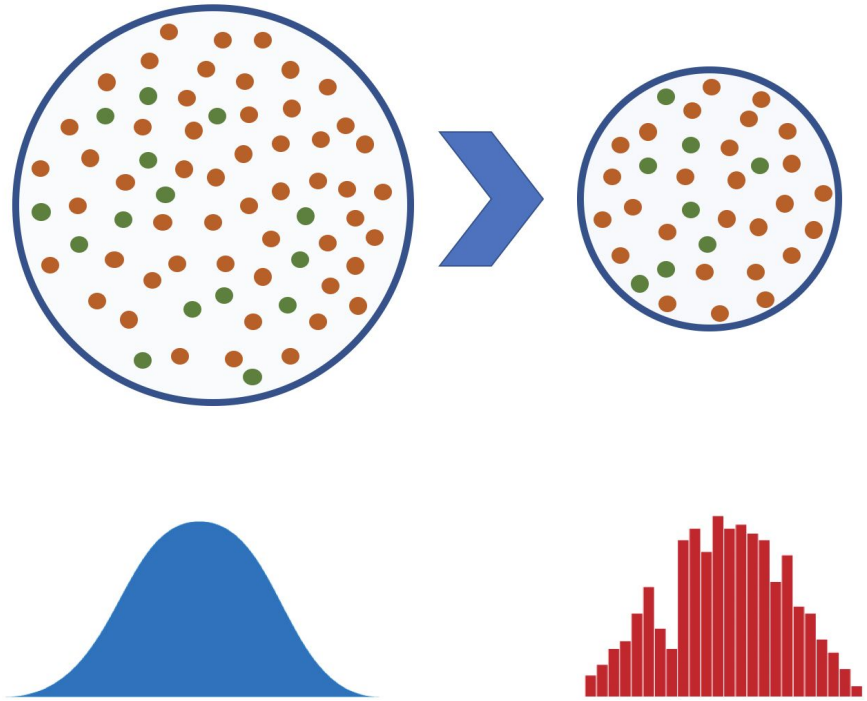
1. OVERFITTING: THE BIG LIMITATION IN AI



EMPIRICAL RISK MINIMIZATION

We are talking about "empirical risk minimization", that is to say we have **a dataset** which we consider being **drawn from an underlying data distribution**, from which we would like to learn something. **We do not know the true, underlying distribution** (the true, underlying relationships), so we estimate the model that best represents the relationships presented in the available empirical data. In order to do so, ***we minimize the "empirical risk", that is we select the model that performs best on the empirical data we have obtained.***

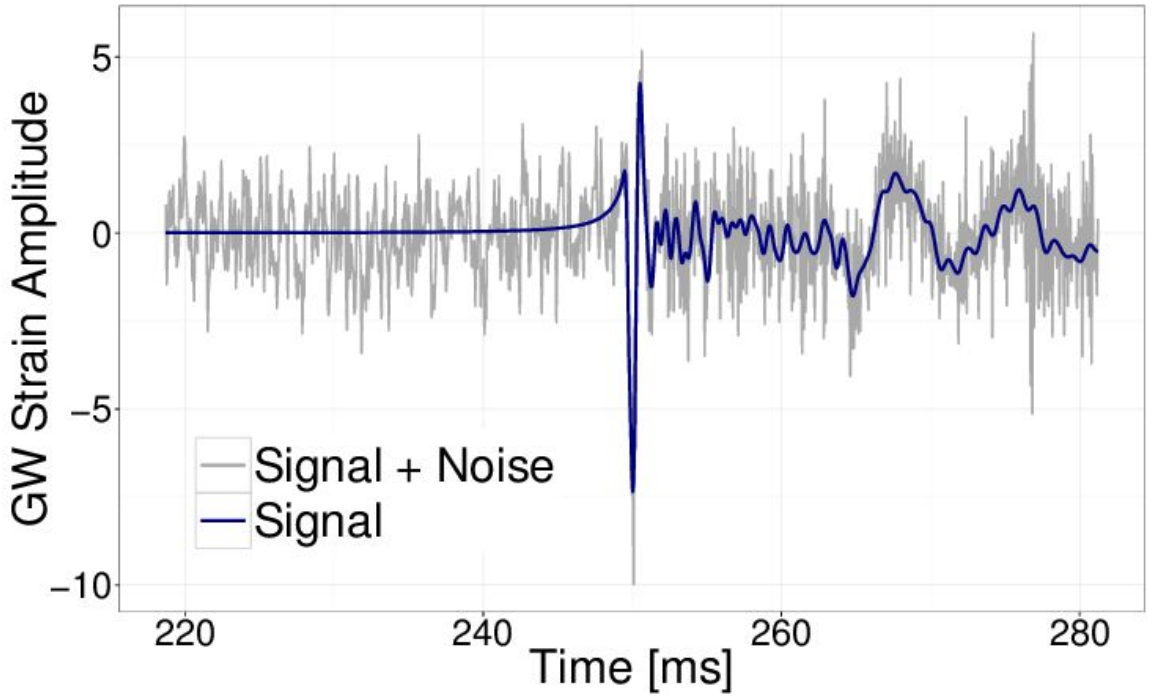


“It is difficult to do predictions, especially about the future!”

- attributed to Niels Bohr (and some others)

“NOT THAT EASY...“

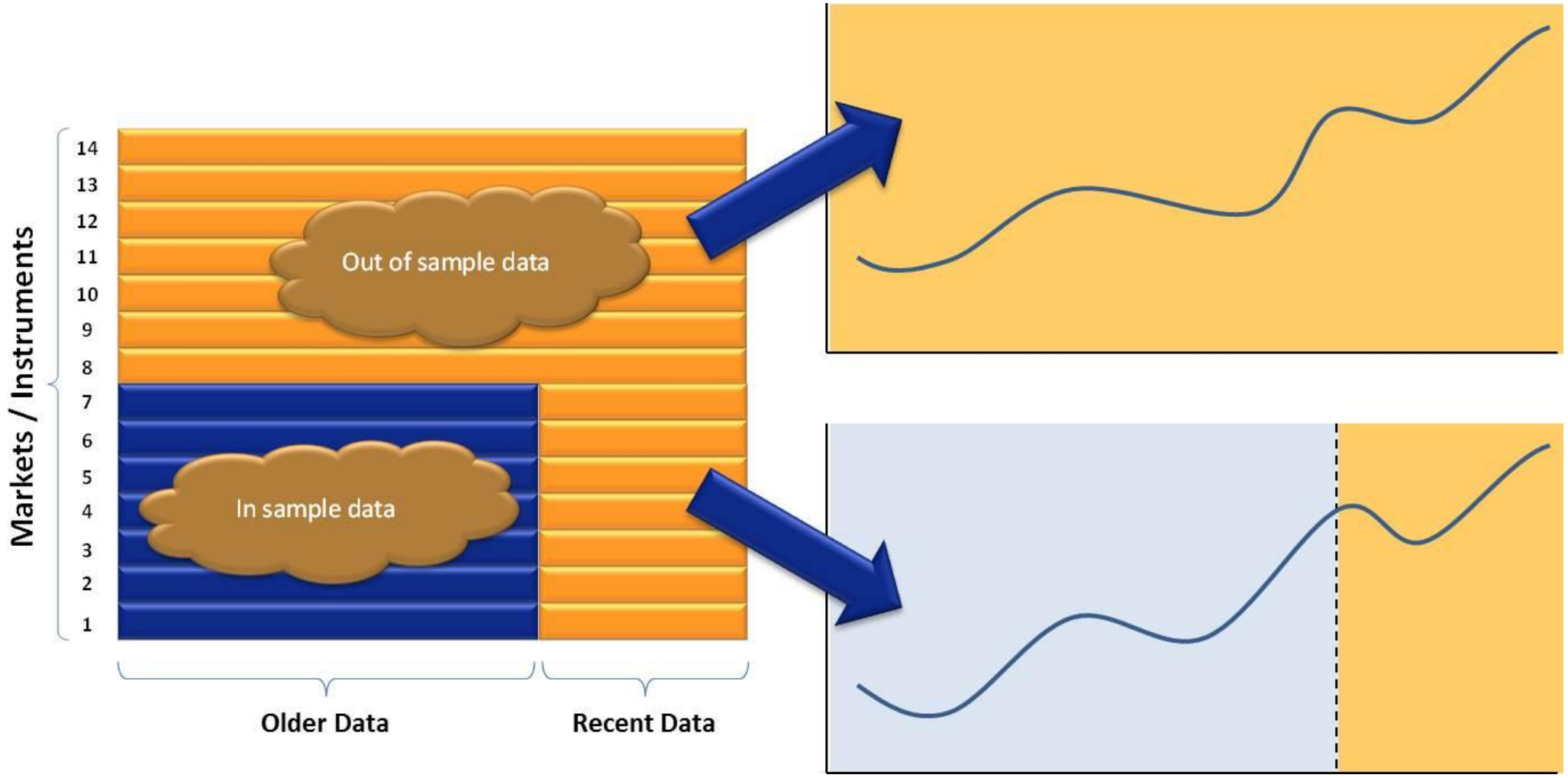
SAMPLING HAS ITS DRAWBACKS



Powered by AI Partners

"DON'T JUST FIT"

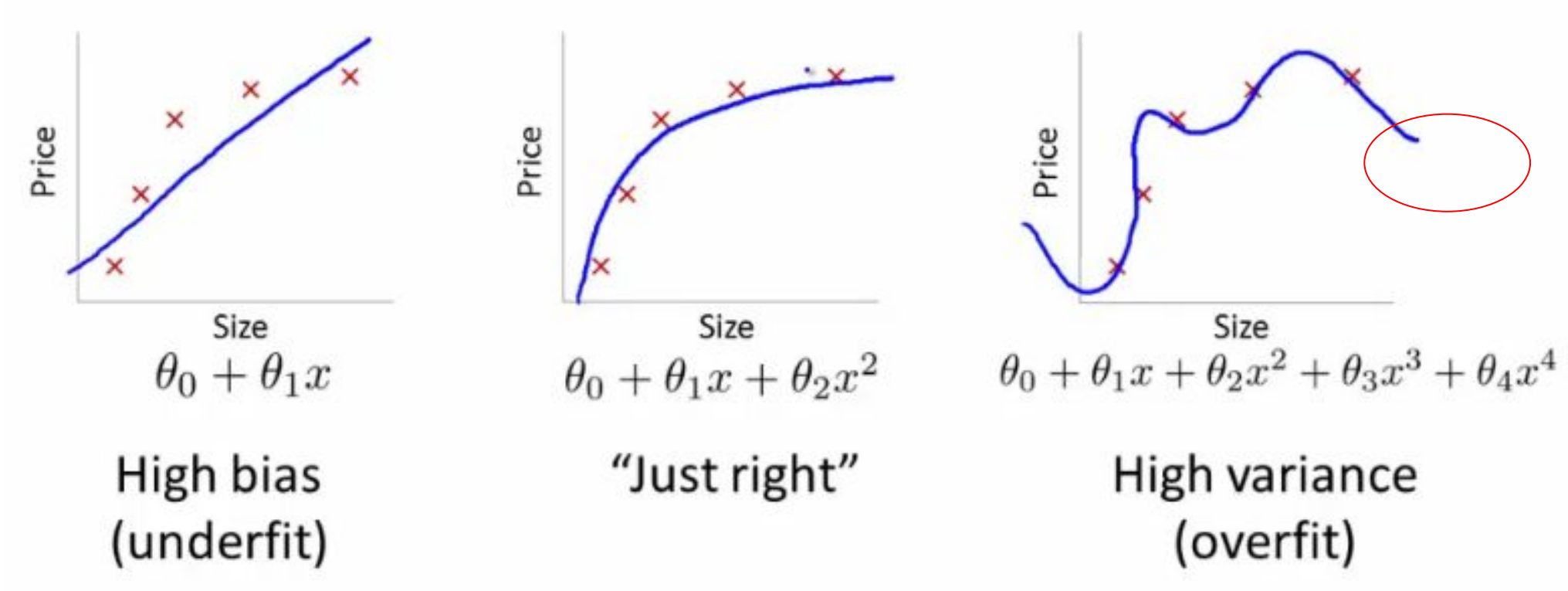
ALWAYS VALIDATE AND TEST!



Powered by AI Partners

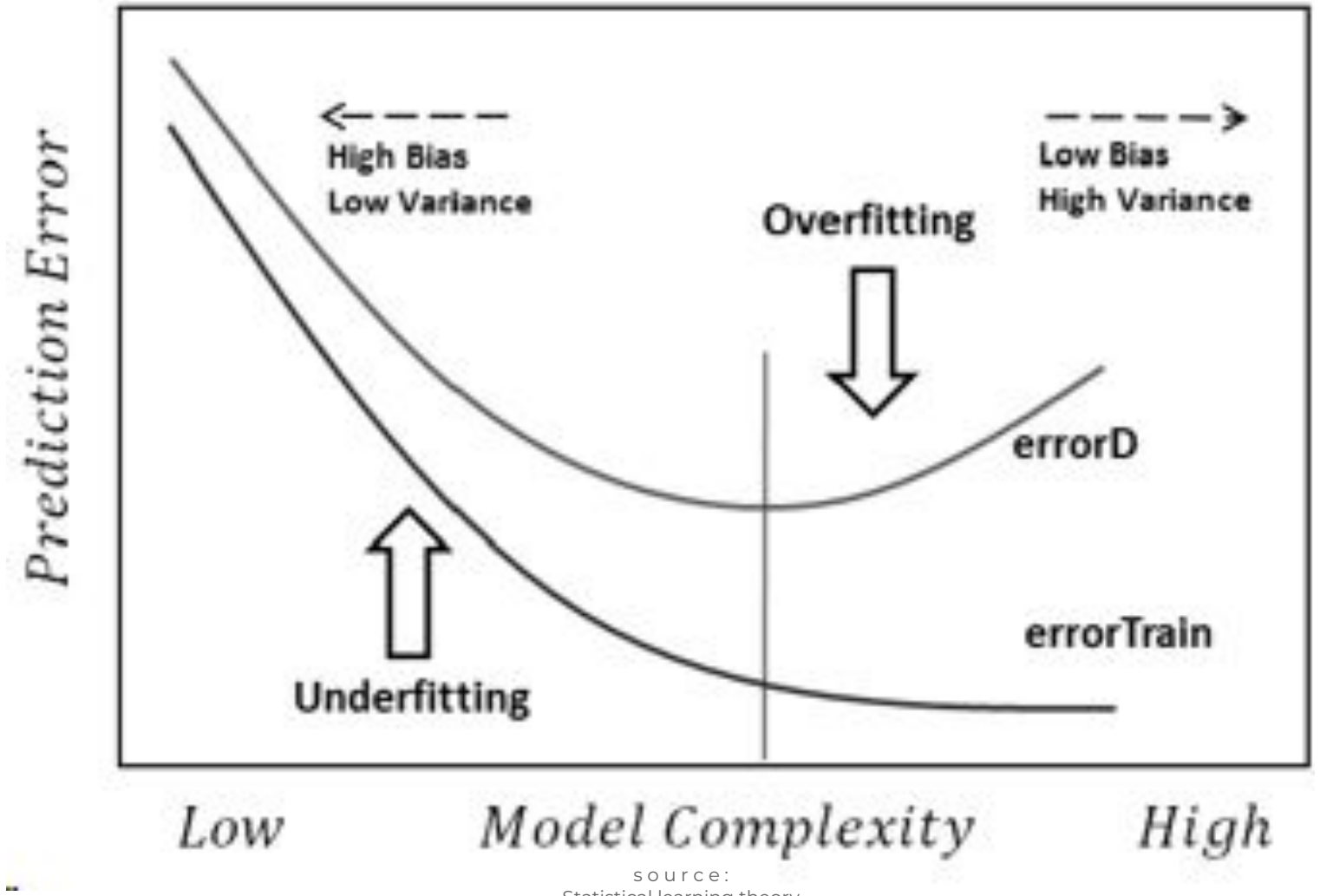
“DEGREES OF FREEDOM“

OVERFITTING IS DANGEROUS



“DON’T JUST FIT”

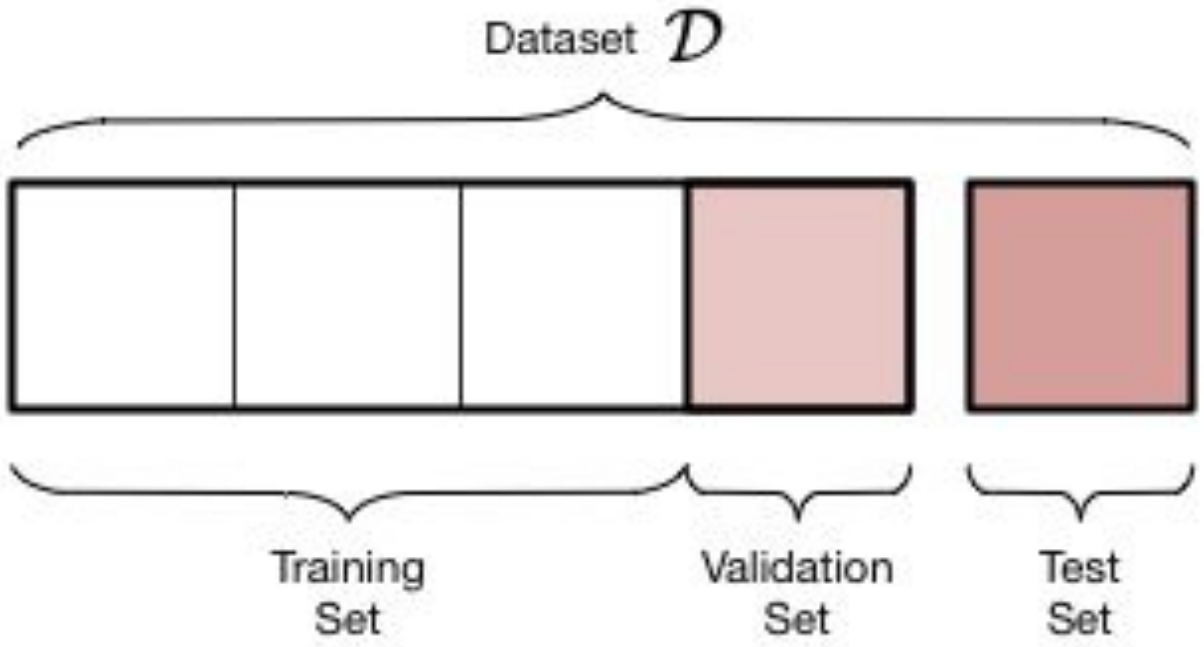
ALWAYS VALIDATE AND TEST!



source:
Statistical learning theory
Vapnik–Chervonenkis dimension

“DON‘T CONTAMINATE THE PROCESS!“

UNBIASED ESTIMATES ARE TO BE PRESERVED!

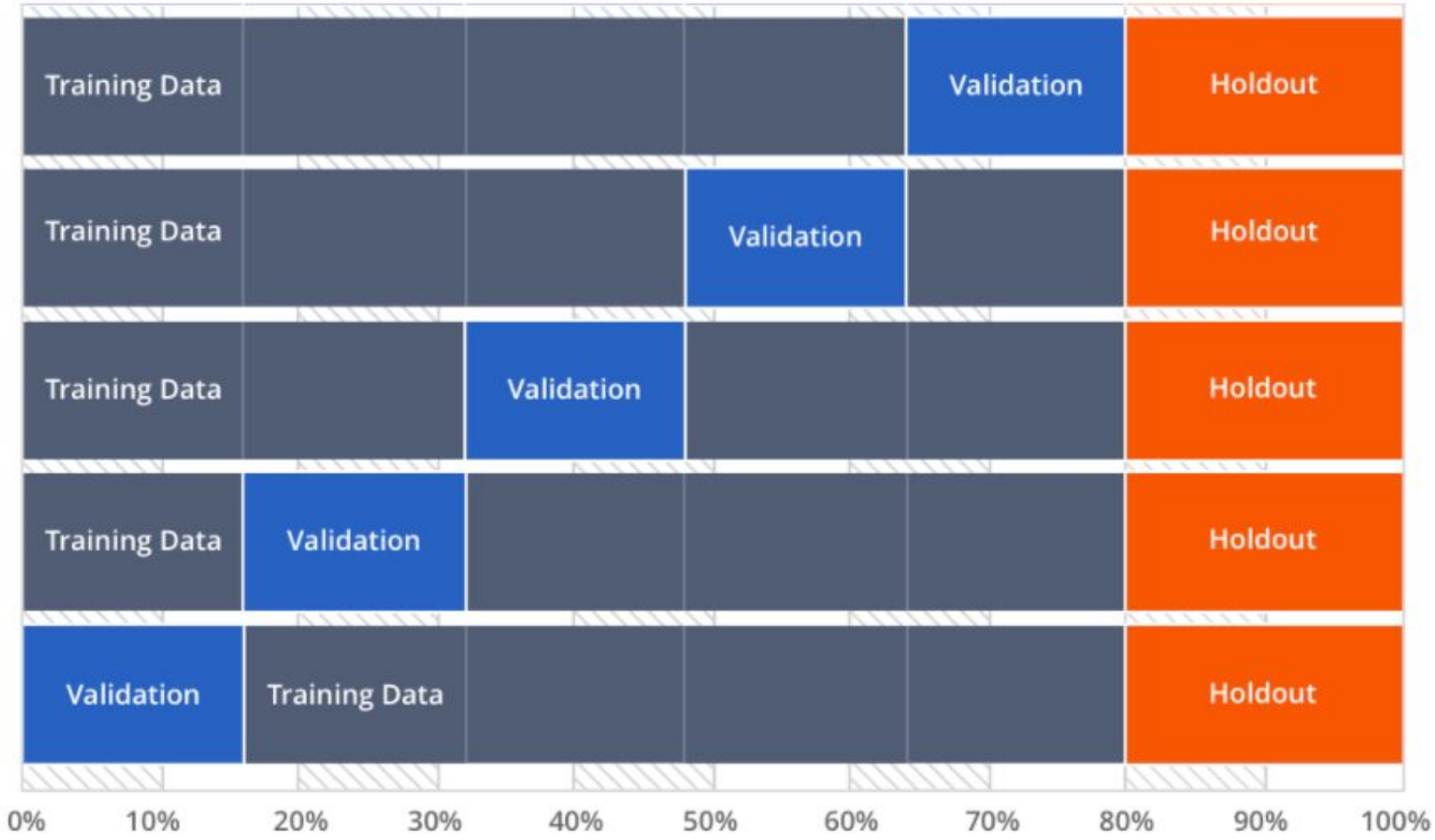


- **If we make decisions based on a held-ot data, it’s information content will influence our model**. It is no longer an unbiased estimate of performance!
- **We need TWO held out datasets,** one we use for decisions, the other WE NEVER TOUCH, just before deployment, as an estimate for performance.
- If we use a dataset, “accidental contamination” can occur. (by normalization, by hyperparameter optimization, by feature selection, …)
- **NEVER TOUCH TEST**

source:
Accidental contamination
Selection bias

“DON’T JUST FIT”

ALWAYS VALIDATE AND TEST!



“NOT ENOUGH INFORMATION”

SMALL DATASETS VIOLATE BASIC ASSUMPTIONS

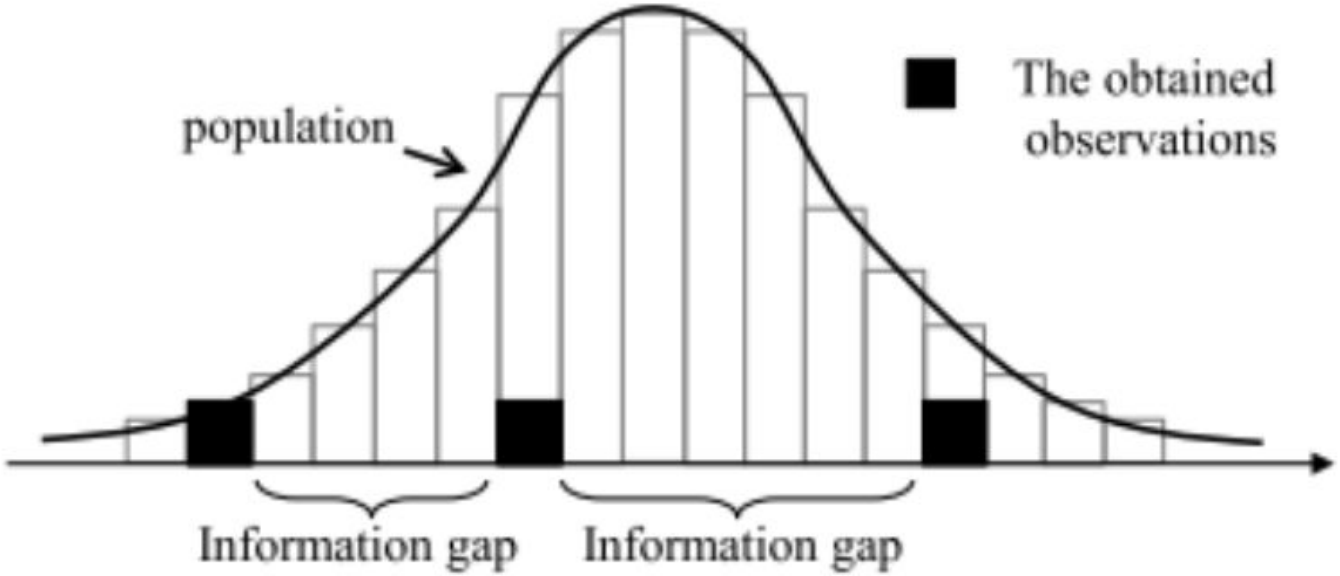
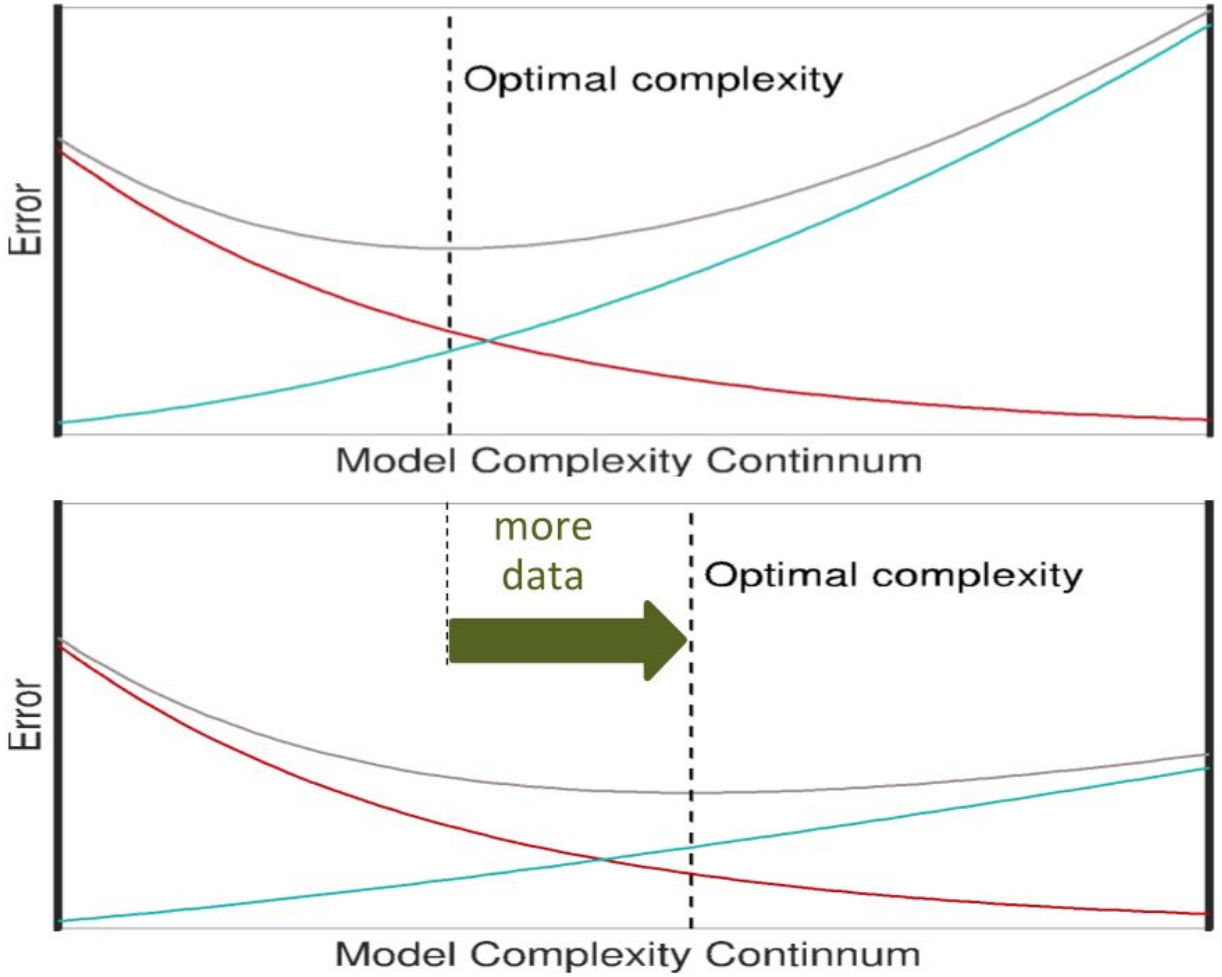


Figure 2. The distribution of a small dataset relative to its population [6]

source:
‘Handling a Small Dataset Problem in Prediction Model by employ Artificial Data Generation Approach: A Review’

“DON’T JUST FIT!”

...AND ADD DATA!

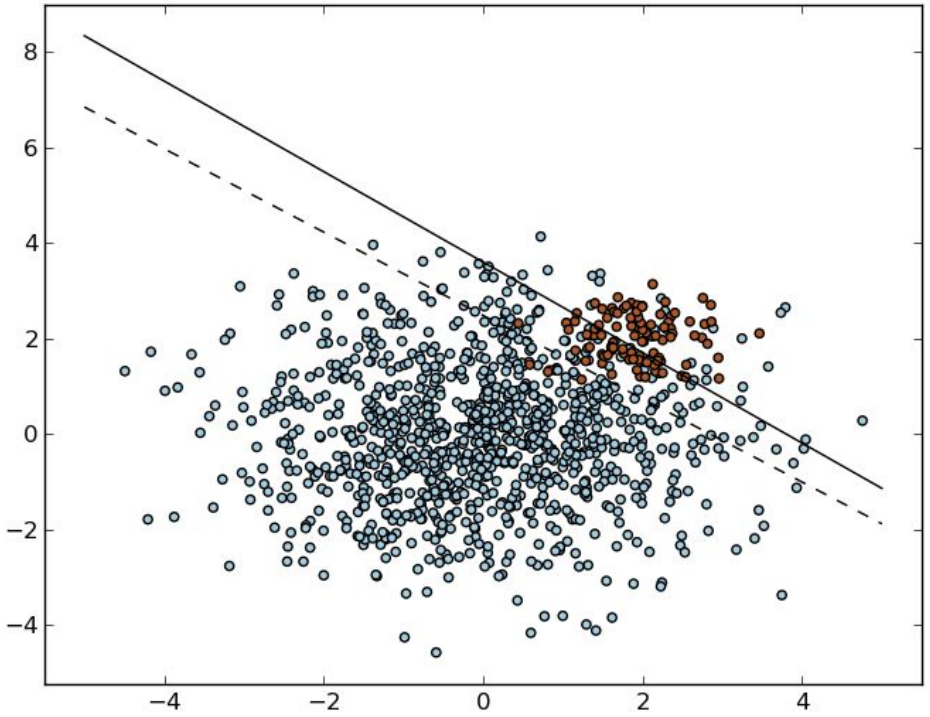


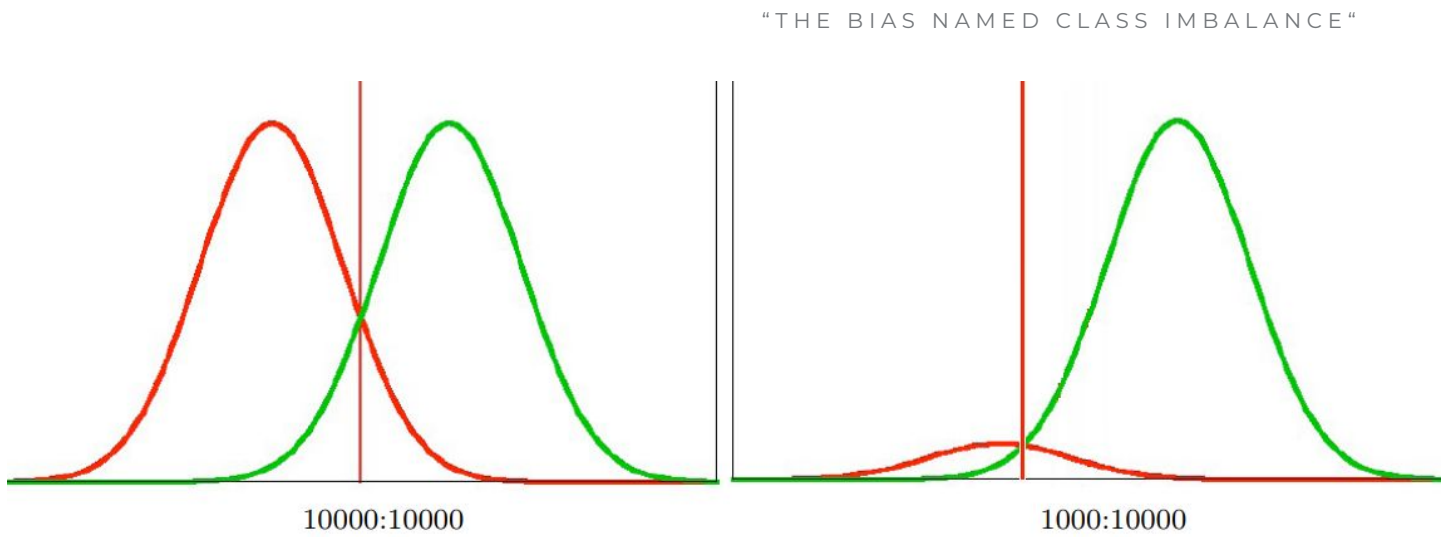
source:
“Lecture series of Michael C. Mozer at DeepLearn2017 Bilbao”

THE CASES OF SMALL DATA (DATA QUANTITY)

“THE BIAS NAMED CLASS IMBALANCE“

CASE I. - WE DON’T HAVE ENOUGH OF ONE THING

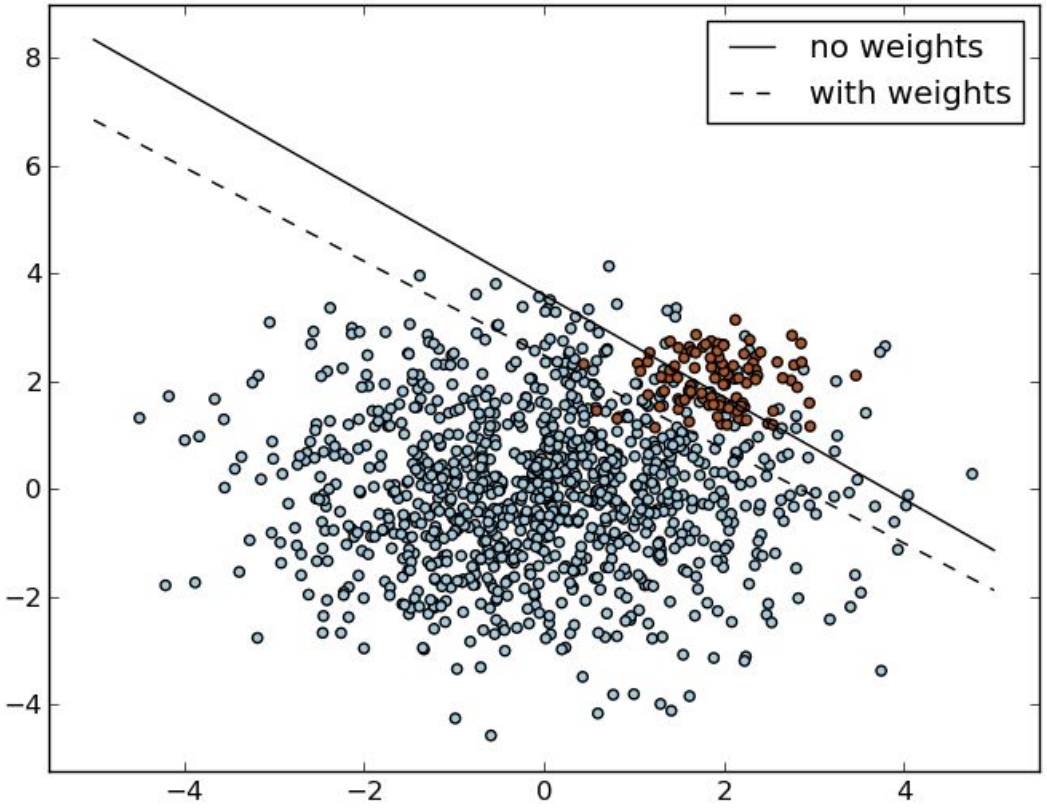




source:
[“Classification in imbalanced datasets”]

“DATA IS NOT CREATED EQUAL“

SOLUTION 1. - “COST SENSITIVE LEARNING”



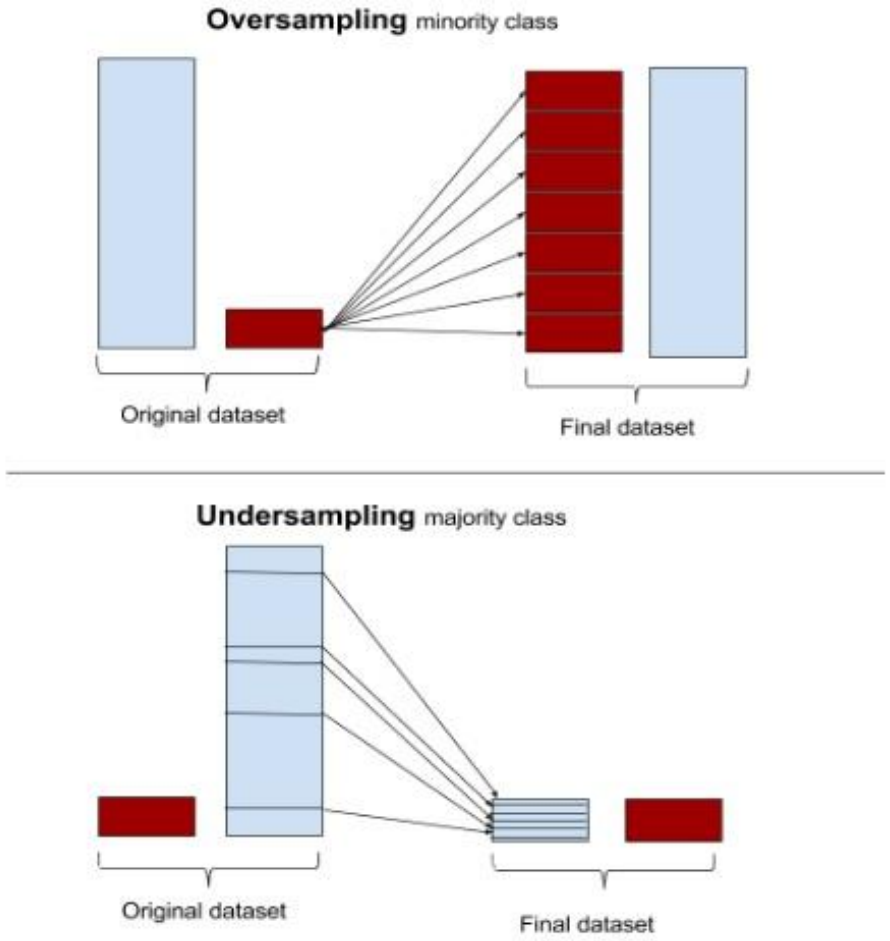
Actual \ Predicted	Category-A	Category-B
Category-A	90	0
Category-B	10	0

We can try to **modify our objective** / cost calculation to accommodate the fact, that making an error on the minority class is a “more serious issue”.

source:
“Cost sensitive learning and the class imbalance problem”

“BIASED COINS FOR THE WIN“

SOLUTION 2. - “SAMPLING”



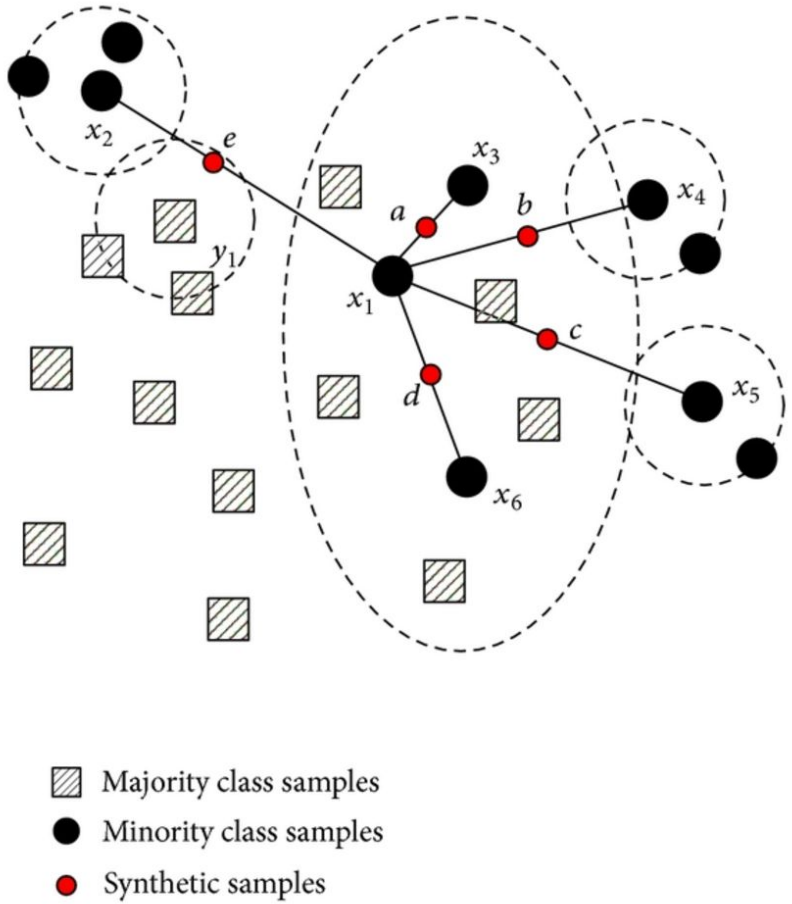
- **Oversampling**:
  - Repeatedly use some of the minority class data points
  - Good question is: Which ones?
  - Can we be more intelligent then random choice?

- **Undersampling:**
  - Choose only some of the majority class data points
  - Reduces the overall dataset, **not recommended**

source:
Undersampling and oversampling questions

"IF YOU DON'T HAVE IT, CREATE IT"

SOLUTION 3. - DATA SYNTHESIS



- **Create new data points! (SMOTE)**

  "First it finds the n-nearest neighbors in the minority class for each of the samples in the class. Then it draws a line between the the neighbors an generates random points on the lines."

- **...and add some noise! (ADASYN)**

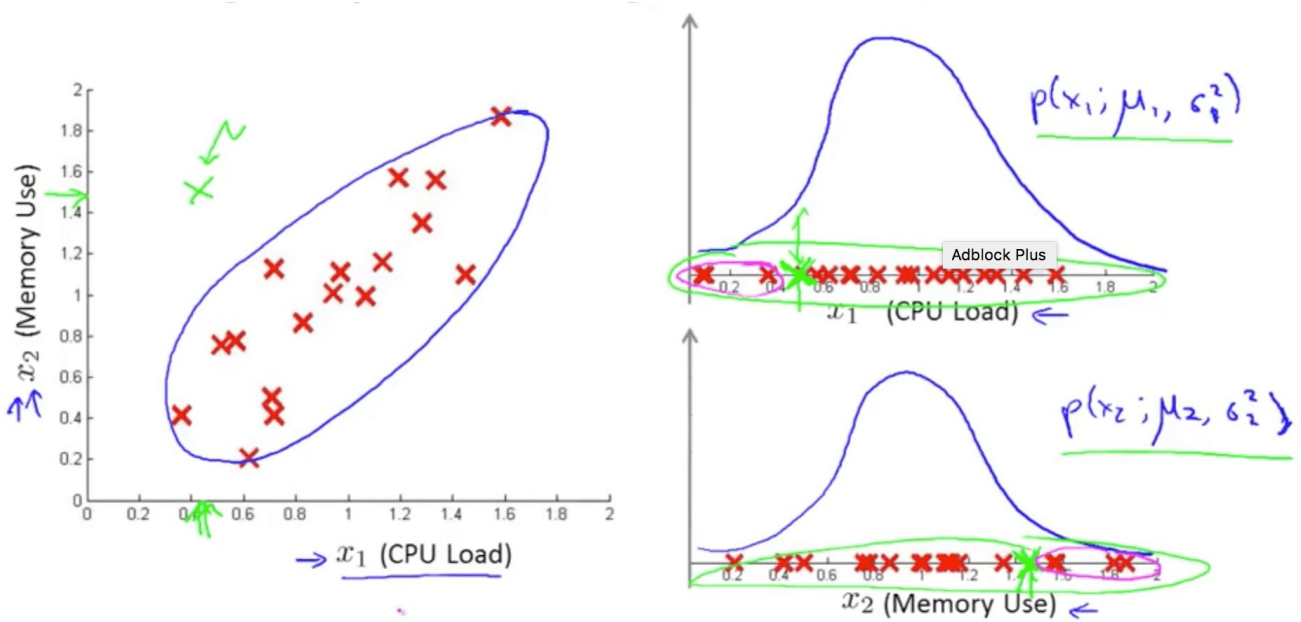
  "After creating those sample it adds a random small values to the points thus making it more realistic. In other words instead of all the sample being linearly correlated to the parent they have a little more variance in them i.e they are bit scattered."

- **...and use clusters! (Cluster Based Oversampling)**

source:
SMOTE and ADASYN
"Clustering and Learning from Imbalanced Data"

“WELL IF YOU LOOK IT JUST WAY“

SOLUTION 4. - RECAST PROBLEM!

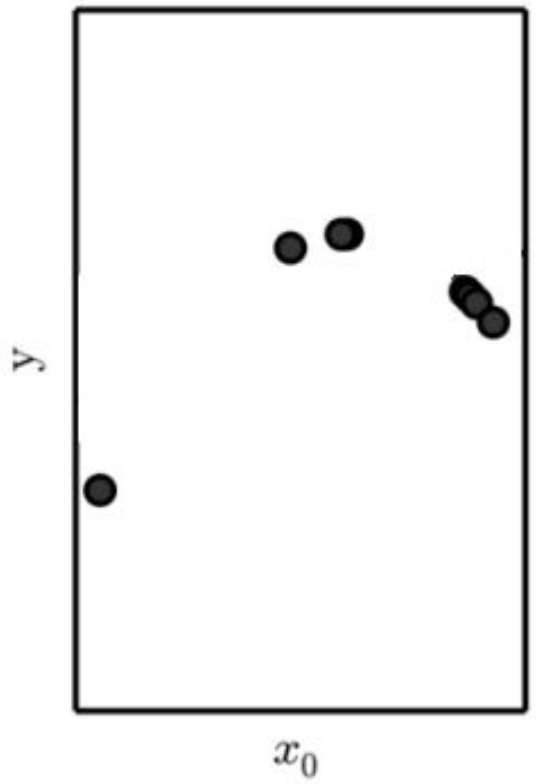


- If the minority class points are so rare, they can be considered “exceptions”, or **“anomalies”**
- There are tools for “one class” classification (eg.: “One class SVM” and “Isolation forests”)
- But if we basically get a good **probabilistic model** of the majority class distribution, we are done.
- This will lead us to **“representation learning”**

source:
Classification based outlier detection techniques
Anomaly Detection using the Multivariate Gaussian Distribution
Ritchie Ng: Anomaly detection

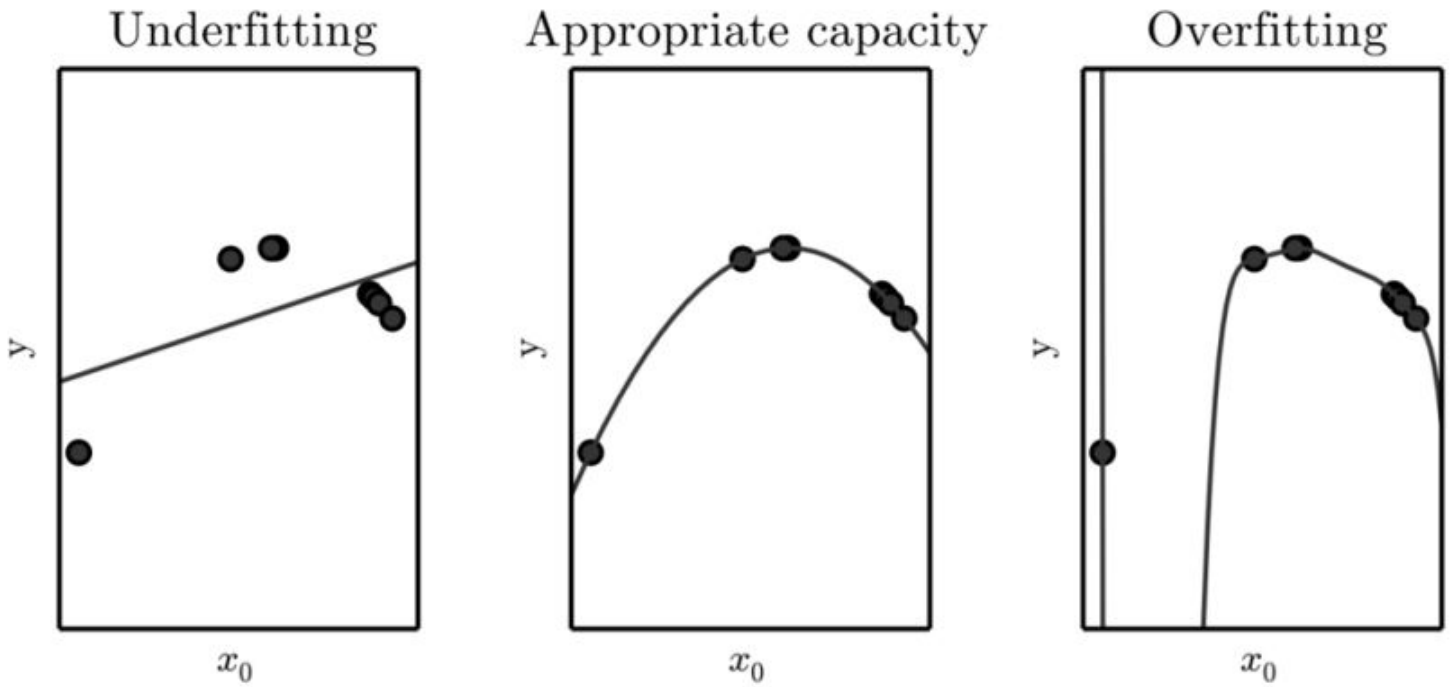
CASE II. - WE DON’T HAVE ENOUGH ANYTHING

SIMPLY: NOT ENOUGH DATA



“NOT ENOUGH IN WHAT SENSE”

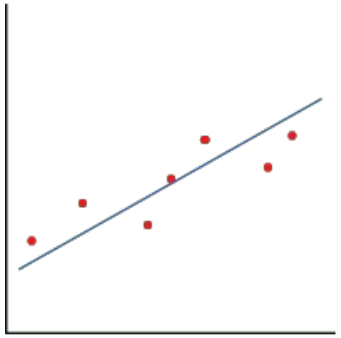
CONNECTION WITH OVERFITTING



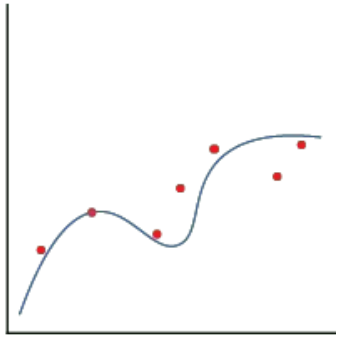
source:
Overfitting - Wikipedia

“TRY THE CLASSICS FIRST”

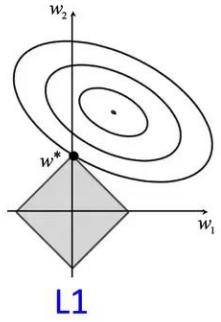
FIRST TRY - CLASSIC METHODS FOR STABILITY



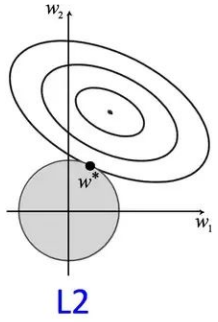
Simple model



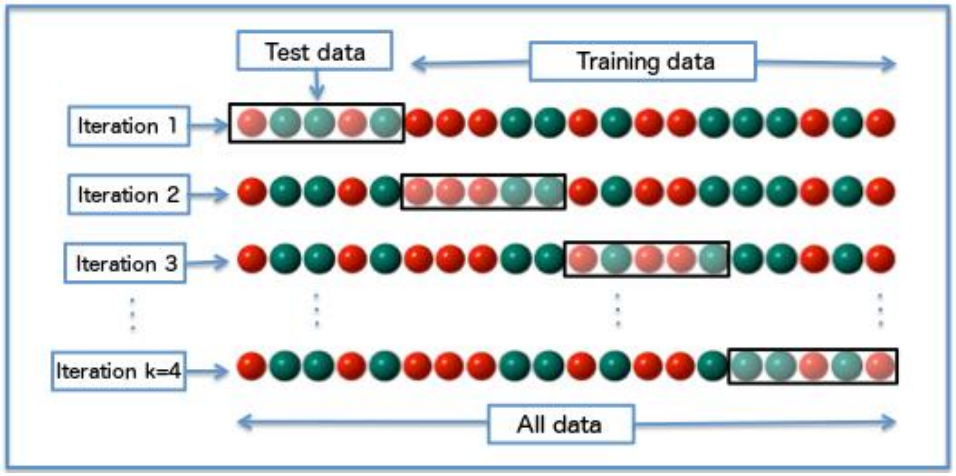
Complex model



L1



L2

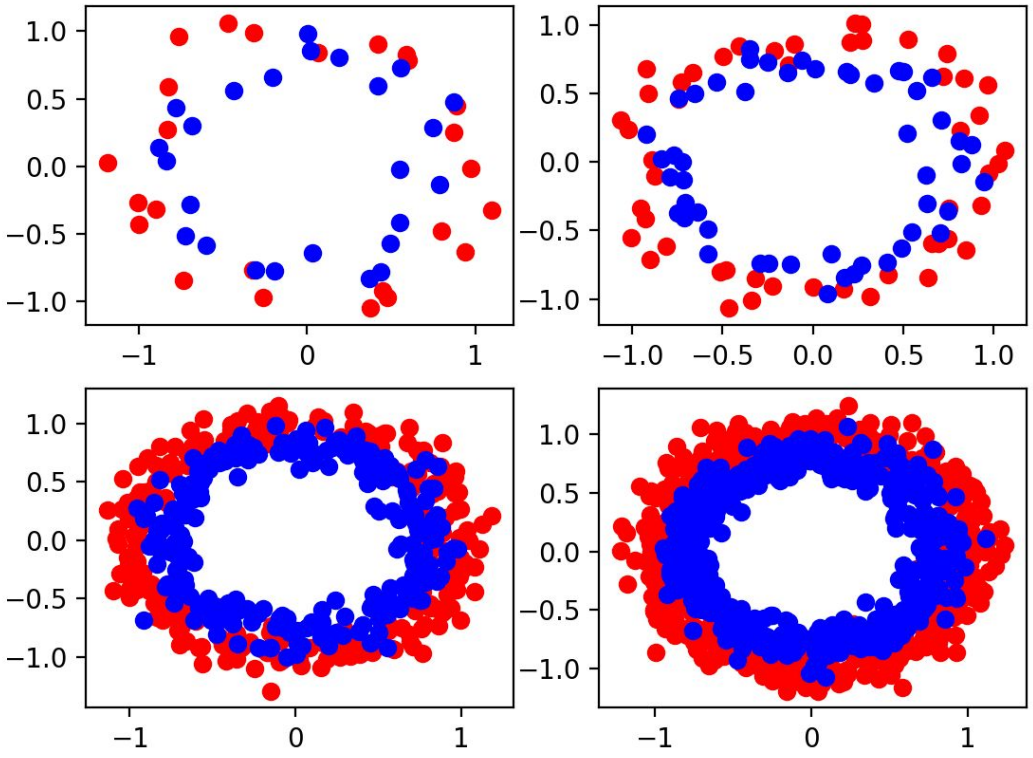


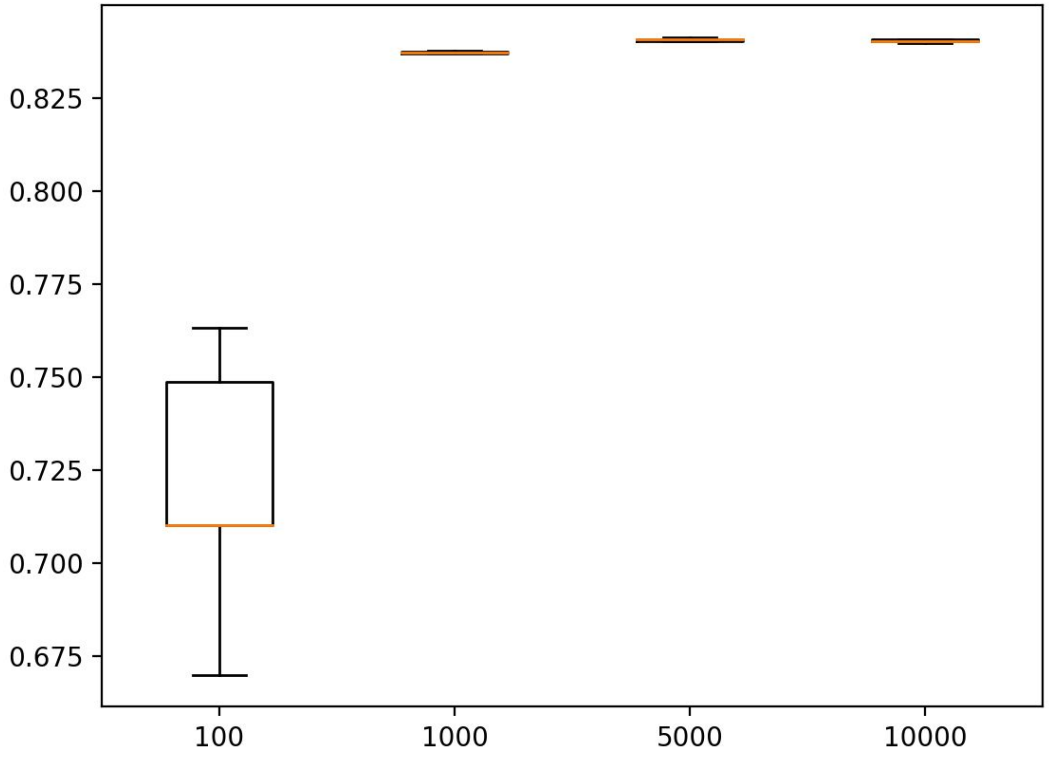
- **Modify the model:**
  - Use a simple model
    - We are often forced to use a complex one since the data itself is complex (dimensions, non-linearity…)
  - Use special models (eg. SUFTware)
- **Modify the objective:**
  - Add regularization term (Capacity control)
  - Use “max margin” objective (like in SVMs)
- **Modify the training:**
  - Use cross-validation for getting a bit more out of the data
  - Use PU Learning

Powered by AI Partners

“THE MORE THE MERRIER”

HOW MUCH IS “ENOUGH” FOR A SMALL NEURAL NET?



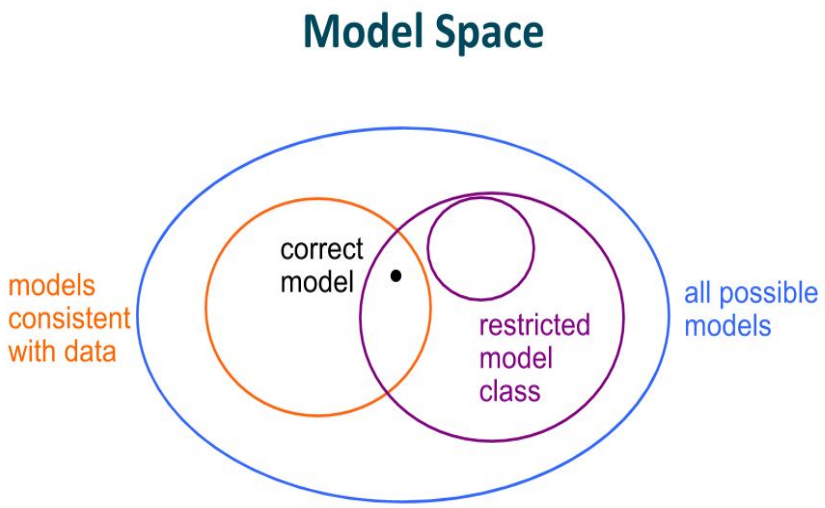


"...two inputs, 25 nodes in the hidden layer, and one output..."

source:
Jason Brownlee: “Impact of dataset size on deep learning model skill and performance estimates”

“THE MORE THE MERRIER”

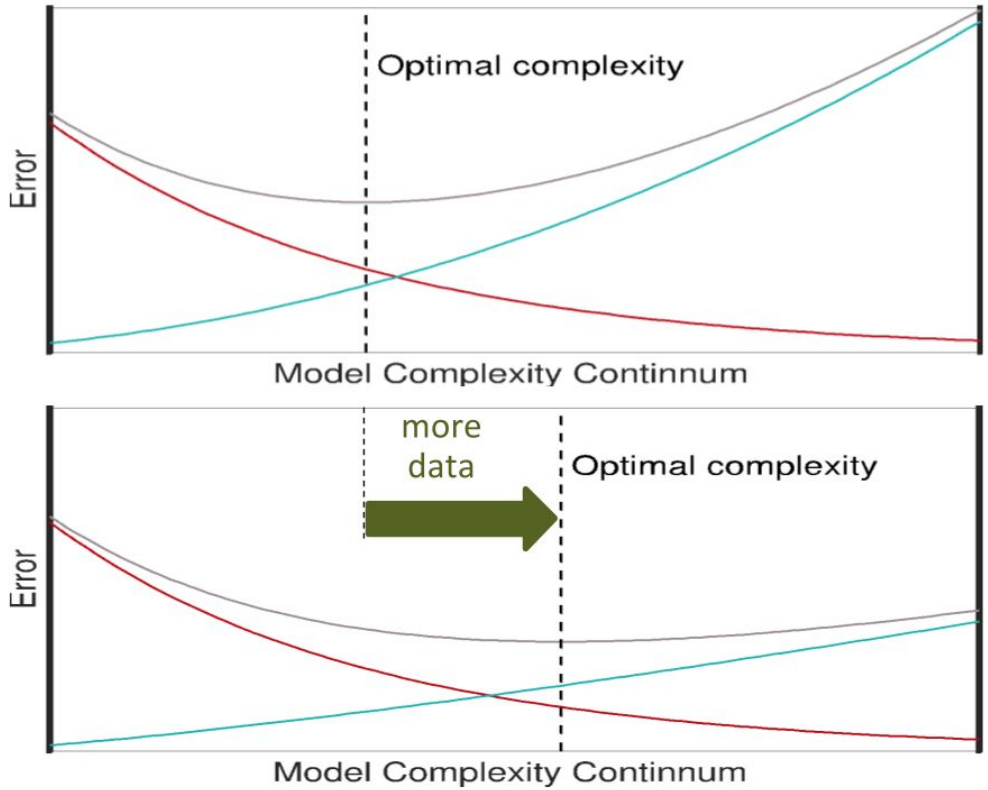
REMARK: ADDING MORE DATA ACTS AS “REGULARIZER”



Restricting model class can help

Or it can hurt

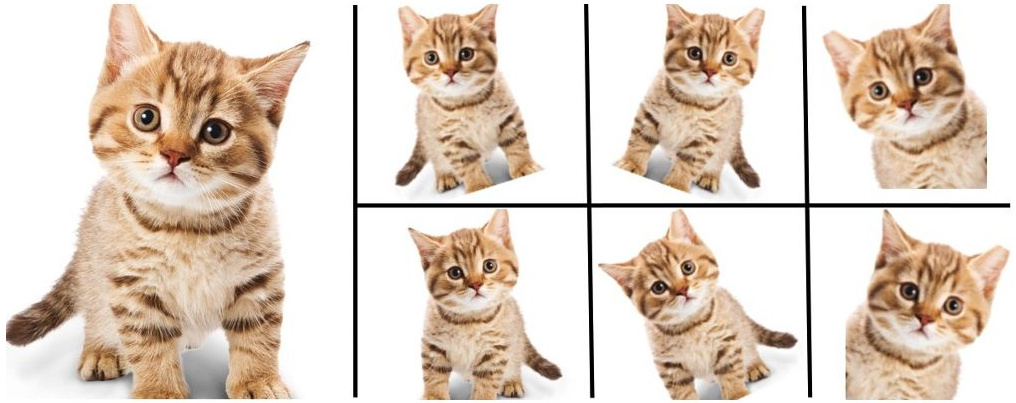
Depends on whether restrictions are domain appropriate



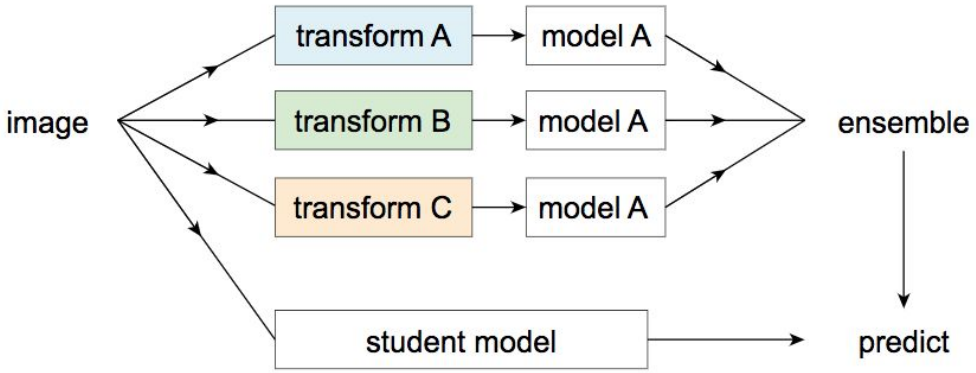
source:
“Lecture series of Michael C. Mozer at DeepLearn2017 Bilbao”

“THE MORE THE MERRIER”

GET MORE “DATA” 1. - GENERATE OR AUGMENT



Enlarge your Dataset

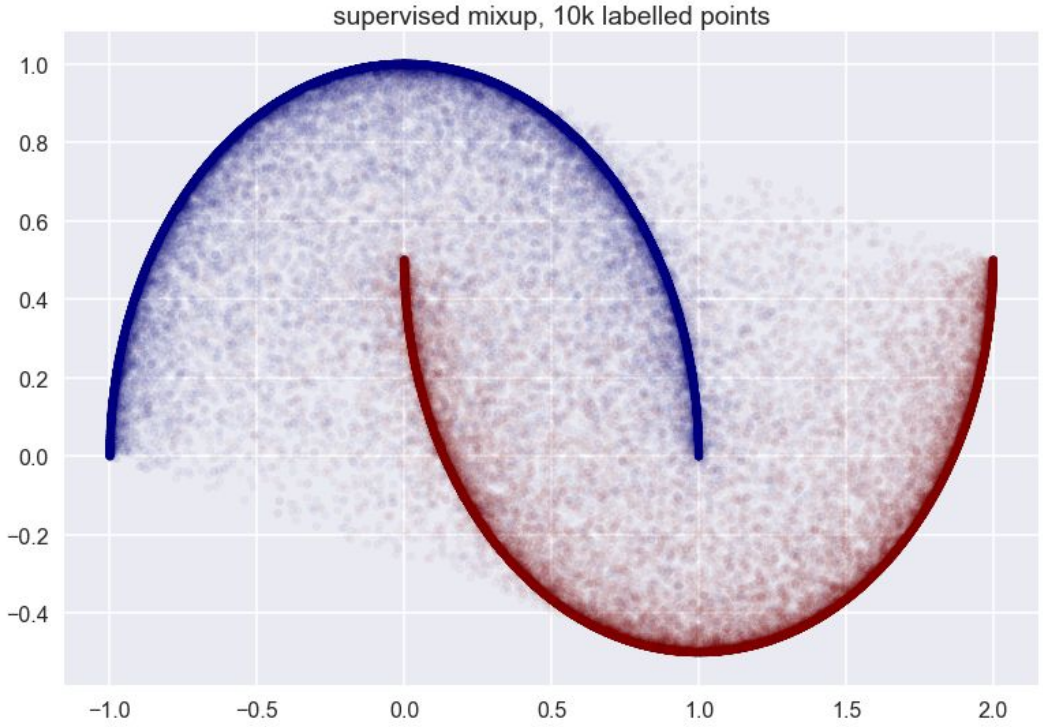


- **Data augmentation:**
  - Use simple operations to modify the data
    - Images: rotate, mirror, crop,...
    - **MUST be realistic for the domain distribution**
- **“Self labeling”:**
  - Transform data, train sub-classifiers, use them on new data, add predictively labelled data to original.
- **Weak supervision:**
  - Can be, that labels will be noisy - crowdsourcing

source:
Data augmentation - How to use Deep Learning when you have limited data?”
“Data distillation: Towards omni-supervised learning”
“A brief introduction to weakly supervised learning”
“The Quiet Semi-Supervised Revolution”

“LEARN THE DISTRIBUTION”

GET MORE “DATA” 1.1 - “MIXUP”



The idea of “Mixup”:

Contribution Motivated by these issues, we introduce a simple and data-agnostic data augmentation routine, termed *mixup* (Section 2). In a nutshell, *mixup* constructs virtual training examples

$$\tilde{x} = \lambda x_i + (1 - \lambda) x_j,$$
$$\tilde{y} = \lambda y_i + (1 - \lambda) y_j,$$

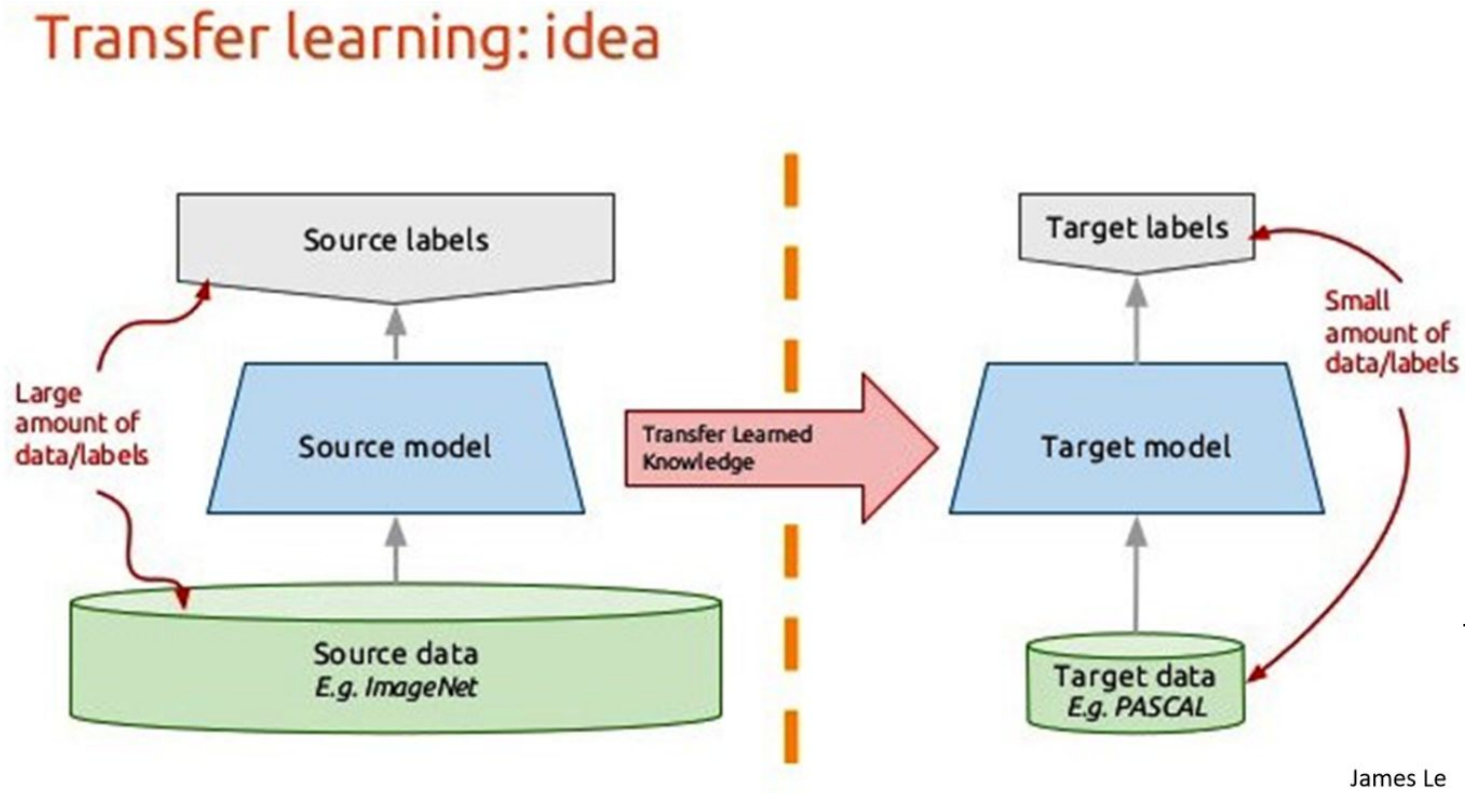
where (x_i, y_i) and (x_j, y_j) are two examples drawn at random from our training data, and $\lambda \in [0, 1]$. Therefore, *mixup* extends the training distribution by incorporating the prior knowledge that linear interpolations of feature vectors should lead to linear interpolations of the associated targets. *mixup* can be implemented in a few lines of code, and introduces minimal computation overhead.

Approximates a whole distribution!

source:
Mixup: Beyond empirical risk minimization
Mixup: Data-dependent Data Augmentation (analysis by inFERENCe)”

"THE KNOWLEDGE RESIDES WITHIN"

GET MORE "DATA" 2. - TRANSFER IT! (COMPRESSED)



source:
"Transfer Learning - Machine Learning's Next Frontier"
""Data Augmentation in Emotion Classification Using Generative Adversarial Networks"

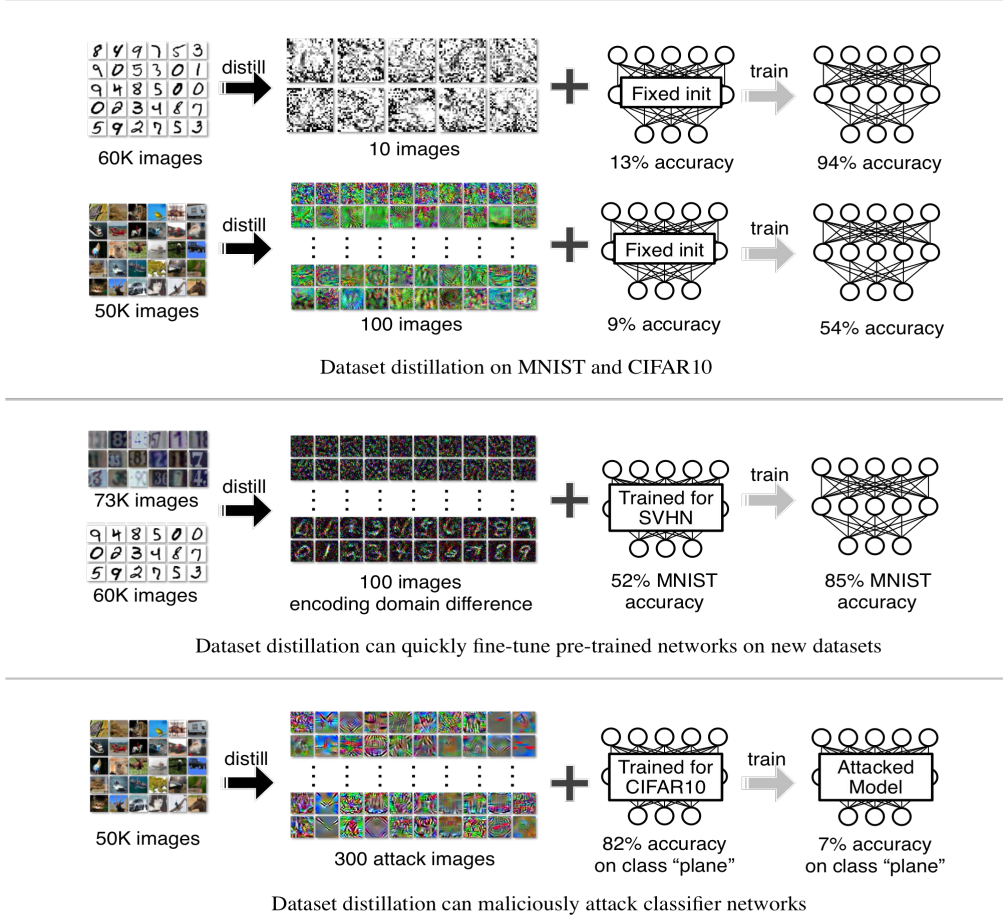
- **Transfer learning!**
  - A **HUGE** topic in itself (with more and more sophisticated methods for preventing **"catastrophic forgetting"**)
  - **We have to see, that models are "storing" data, albeit compressed.**
  - There are **plenty of pre-trained models available, USE THEM!**

- What model to "transfer"?
  - Notion of "learning a whole representation space" (see eg.: Mixup method)
  - **GANs or VAEs** are generally strong candidates (+ few labeled data case)

Powered by AI Partners

“NOT ALL DATA IS EQUAL”

SIDENOTE: DATASET DISTILLATION

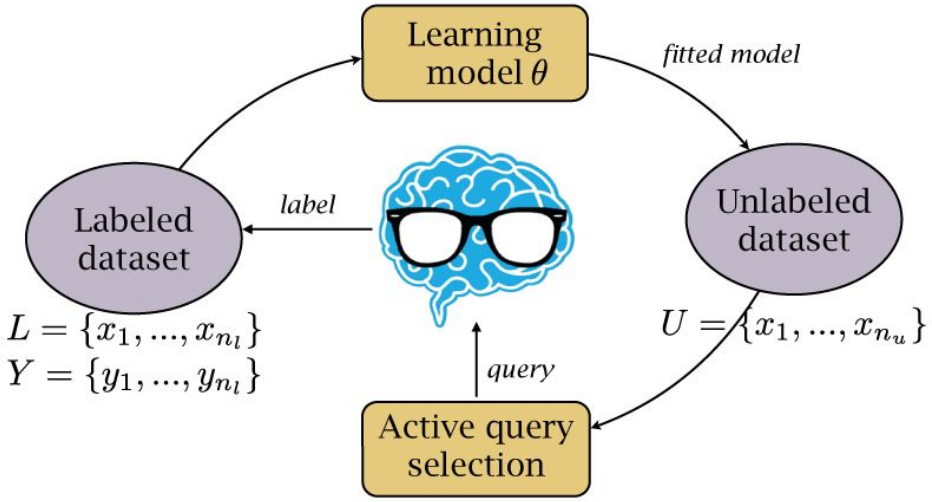


“...The idea is to synthesize a small number of data points that do not need to come from the correct data distribution, but will, when given to the learning algorithm as training data, approximate the model trained on the original data. For example, we show that it is possible to compress **60,000 MNIST training images into just 10 synthetic distilled images (one per class) and achieve close to original performance** with only a few steps of gradient descent, given a particular fixed network initialization”

source:
Dataset Distillation

“PLEASE”

GET MORE “DATA” 3. - ASK FOR IT! :-)



- **Crowdsource!**
  - Amazon Mechanical Turk
  - or CrowdFlower.
- **Design a learning loop!**
  - Continuous, **Online learning**
  - There are key points worth asking for (margin, adversarial examples)
    -> **Active learning**
    -> **Building Models via Comparisons**

source:
“Adversarial sampling for active learning”
“Attacking machine learning with adversarial examples”
“ModAL - Active learning with Keras”

Powered by AI Partners

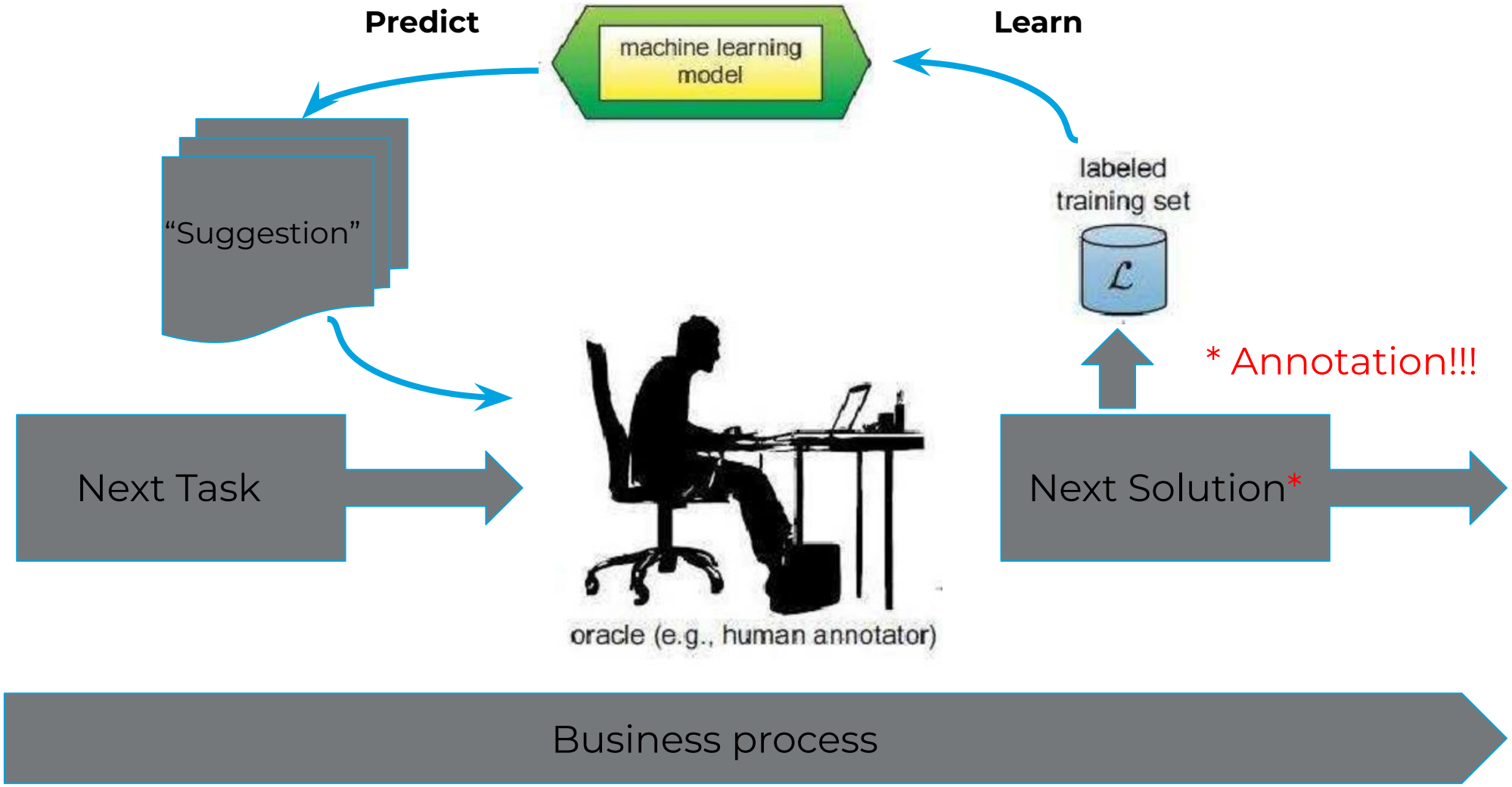
CURRENT LIMITATION OF AI:

requires huge amounts of annotated data -> annotation is a big business

Why big tech pays poor Kenyans to teach self-driving cars
Data Annotation: The billion dollar business behind AI breakthroughs

"ANNOTATION AS A SIDE EFFECT"

PROCESS INTEGRATION



Powered by AI Partners

“NOT JUST A SIDE PRODUCT”

DATA EXHAUST - GOLD, WE THROW AWAY

Data exhaust can be annotation too!

“A company that provides IT solutions probably has plenty of data around the transactions executed between their clients and business website. Info related to the initial sign-up and monthly payments can be considered the **primary data** that easily fits into your big data plans. **Supporting details** such as what device customers used, the time of day those transactions took place, and the navigational pattern that led them to checkout is the **secondary data or exhaust** many organizations don’t even bother with."

Data Exhaust – What Is It And Why Should IT Care?

source:
Data exhaust

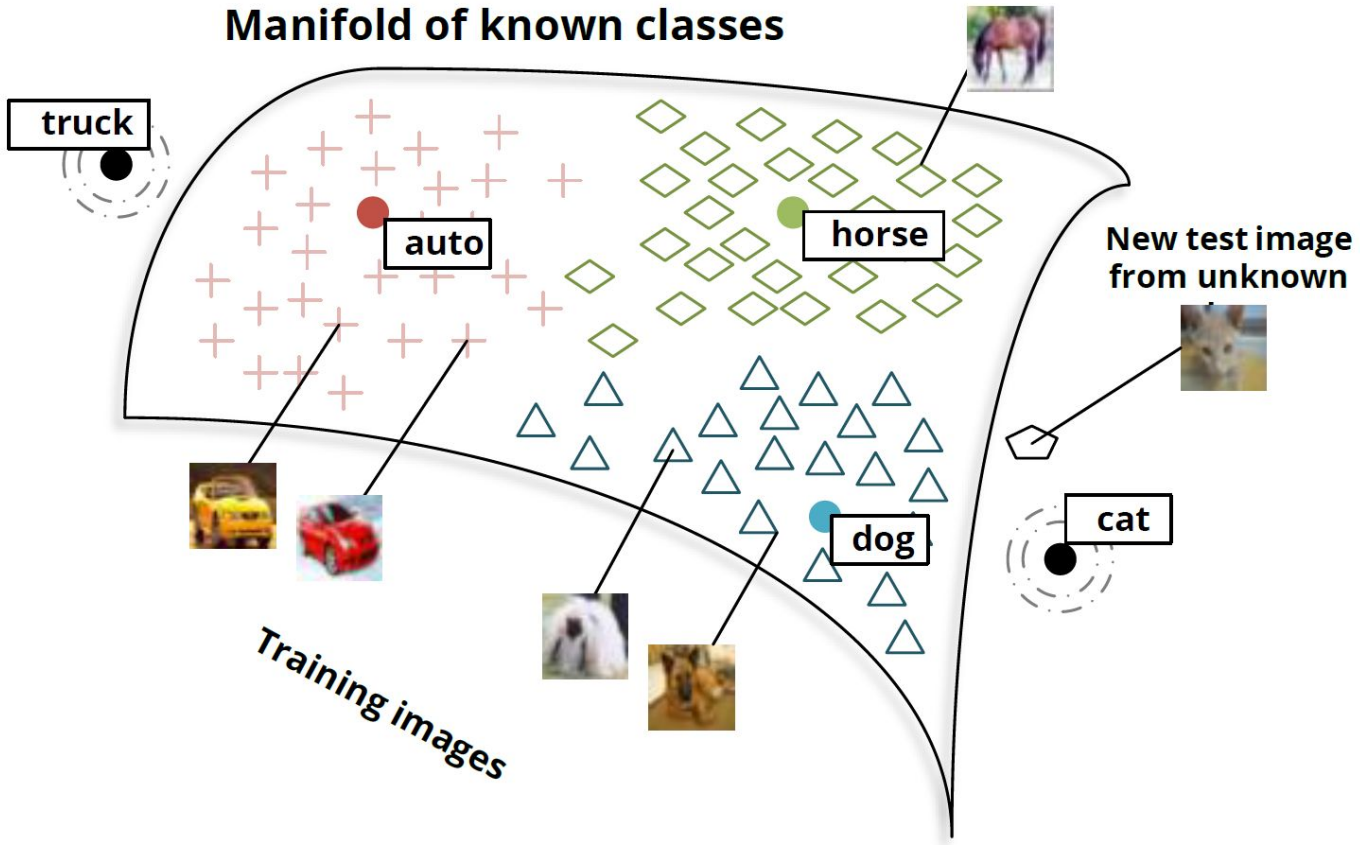
WHAT ARE THE CURRENT LIMITS OF ML?

Limitation 2:

“APPROXIMATION”

“THE REAL UNKNOWN”

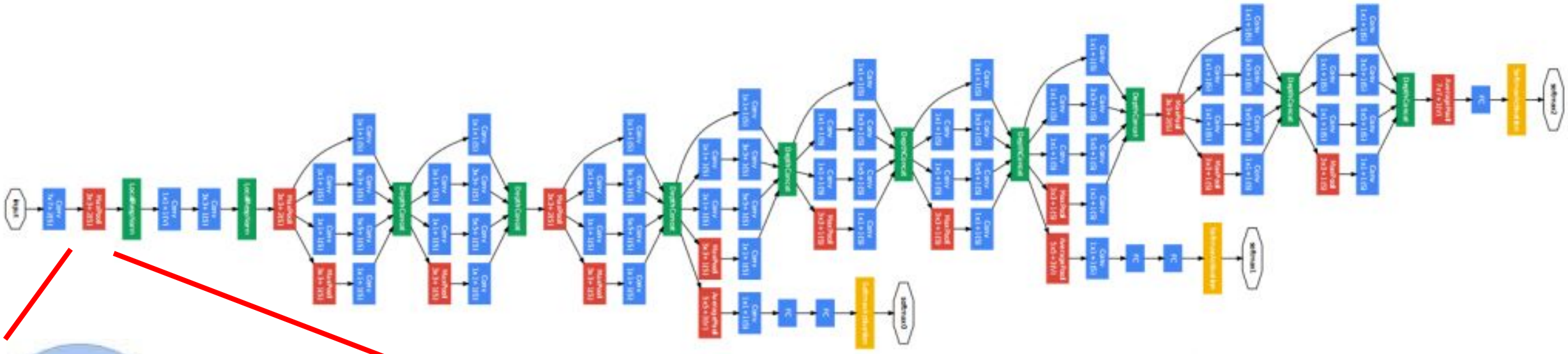
CAN NOT HANDLE UNKNOWN CLASSES!

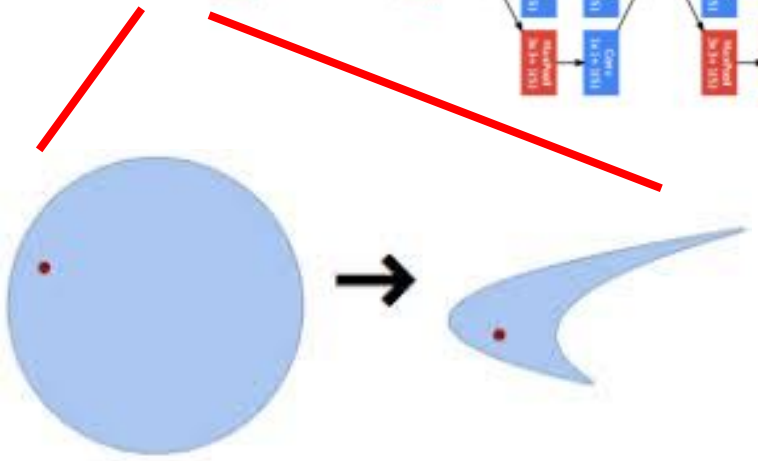


SOURCE:
Deep learning, NLP and representations

“BLACK BOXES”

THE MODEL IS NOT INTERPRETABLE





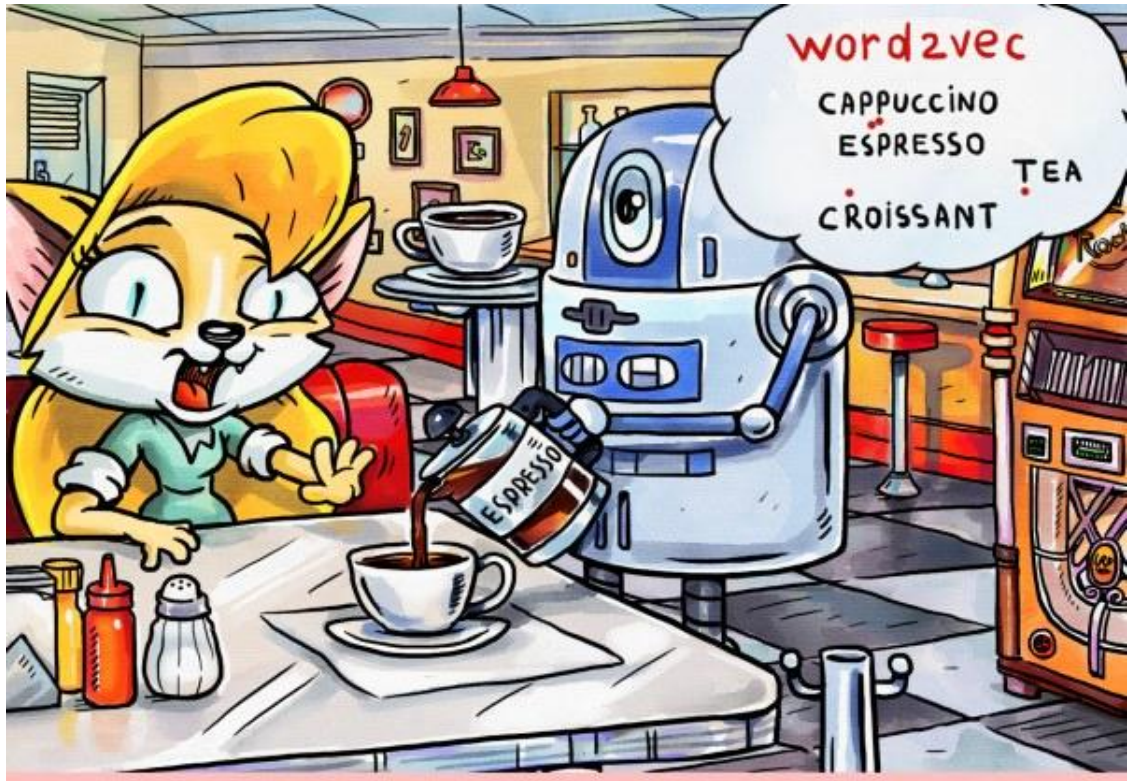
MANY LAYERS + NON-LINEAR TRANSFORMATIONS !

$$\{x, y\} \to \{x\,y, x + y\}$$

SOURCE:
Neural image captioning for mortals

"CLOSE ENOUGH"

REPRESENTATIONS ARE NOT "CRISP"



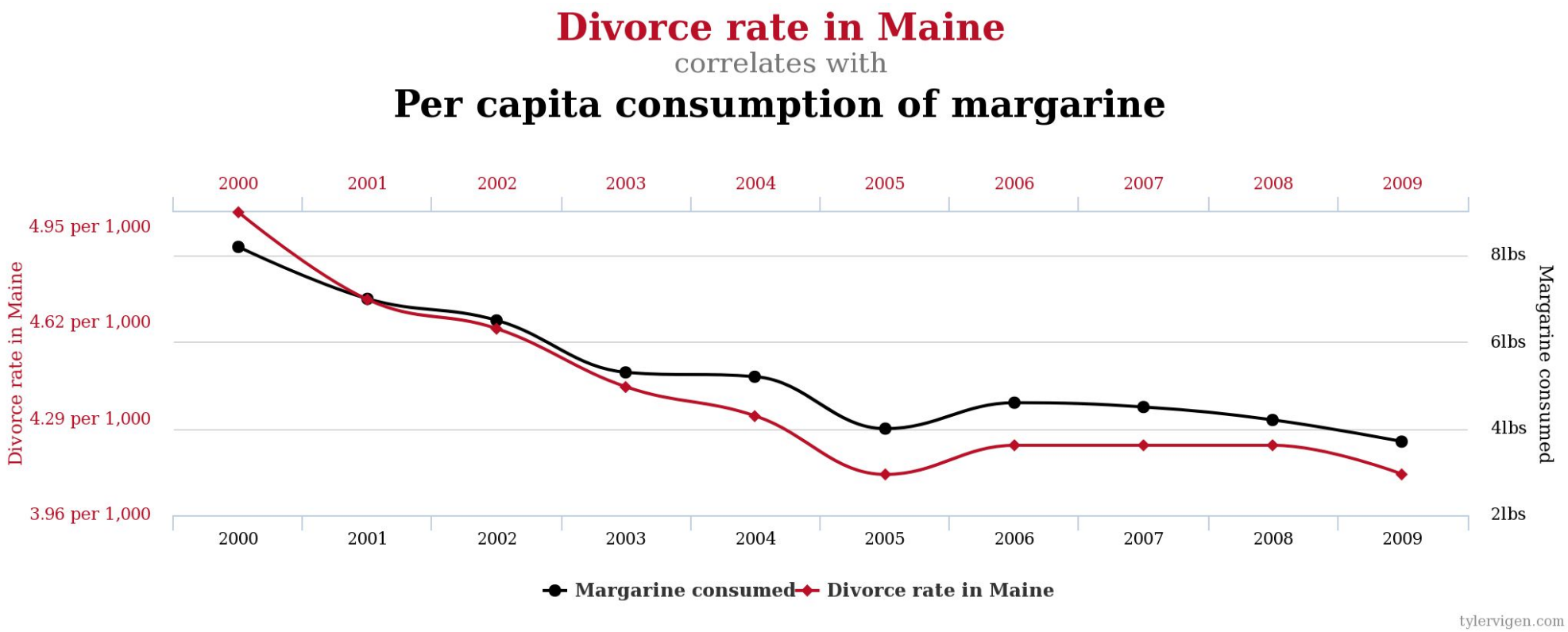
- Espresso? But I ordered a cappuccino!
- Don't worry, the cosine distance between them is so small that they are almost the same thing.

WHEN IT FAILS, IT FAILS BIG TIME - WHY?

SOURCE:
GOOGLE APOLOGIZES FOR PHOTO APP'S RACIST BLUNDER"

“REPEAT IT ENOUGH AND IT BECOMES TRUE?”

IT HAS TO RELY ON CORRELATIONS



SOURCE:
“Spurious correlations website”

Powered by AI Partners

“REPEAT IT ENOUGH AND IT BECOMES TRUE?”

IT HAS TO RELY ON CORRELATIONS

Left: A man is holding a dog in his hand
Right: A woman is holding a dog in her hand
Image: @SouperSarah

SOURCE:
“DO NEURAL NETWORK DREAM OF ELECTRIC SHEEP?”

“WATCH THE TRAIL!”

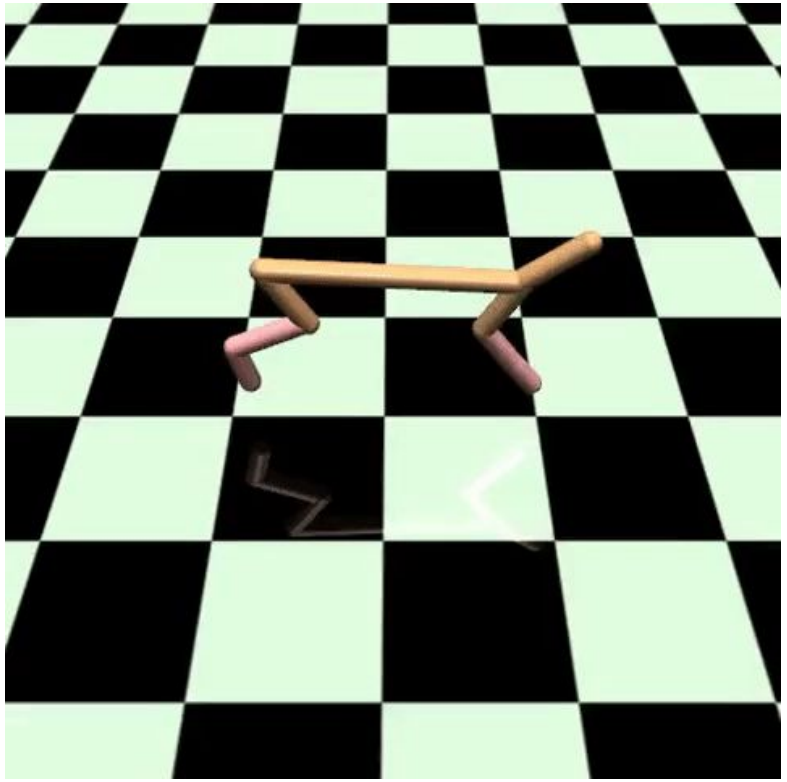
LEARNS FROM RAW INPUT - NO OBJECT CONCEPT

POSSIBLE “CURE”:
“GOOGLE’S AI WIZARD UNVEILS A NEW TWIST ON NEURAL NETWORKS”

Powered by AI Partners

“BUT IT JUST WORKS, ISN’T IT?”

NO CONCEPT OF A “GOOD SOLUTION”



SOURCE
“Deep reinforcement learning does not work yet

Limitation 3:

“BIAS”

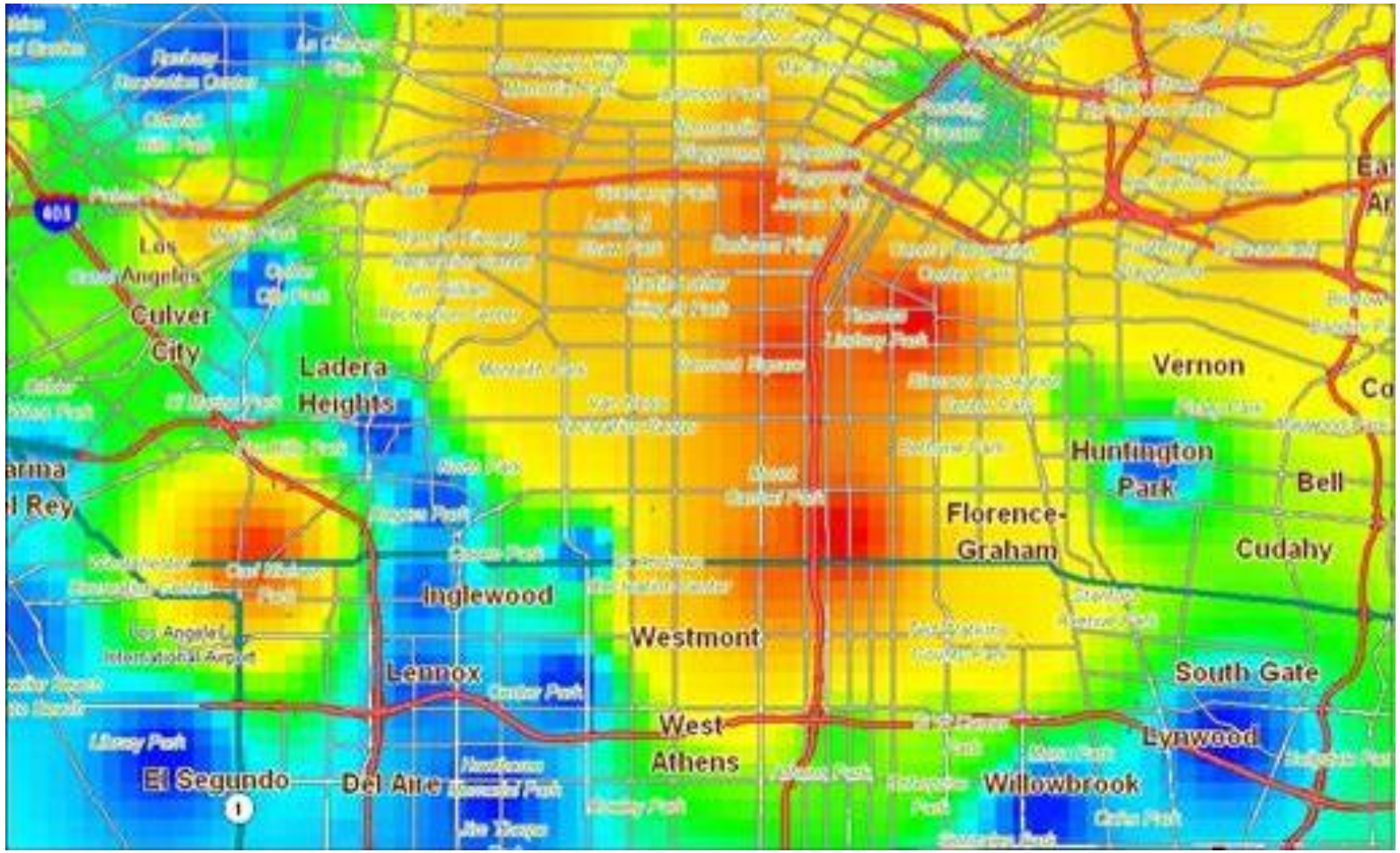
“FOR THE COMMON(?) GOOD (?)?”

THE DANGEROUS CASE OF PREDICTING CRIME

Palantir

“FOR THE COMMON(?) GOOD (?)?”

THE DANGEROUS CASE OF PREDICTING CRIME

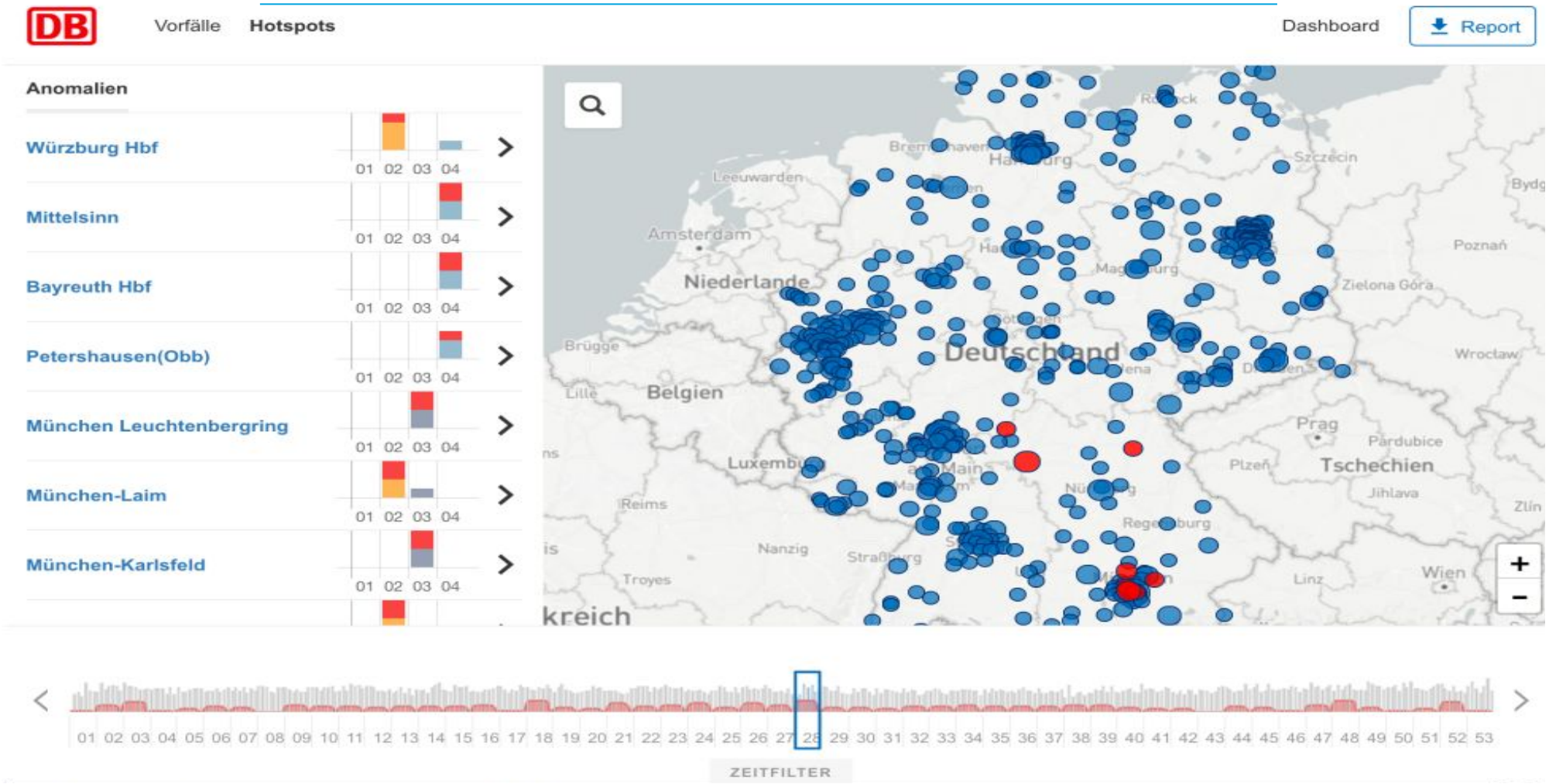


SOURCE
In Los Angeles, using statistics to predict crime

Powered by AI Partners

NEURON SOLUTIONS

PREDICTIVE POLICING FOR THE GERMAN RAILWAYS



Powered by AI Partners

NEURON SOLUTIONS

“FOR THE COMMON(?) GOOD (?)?”

FREE WILL? - BIAS!

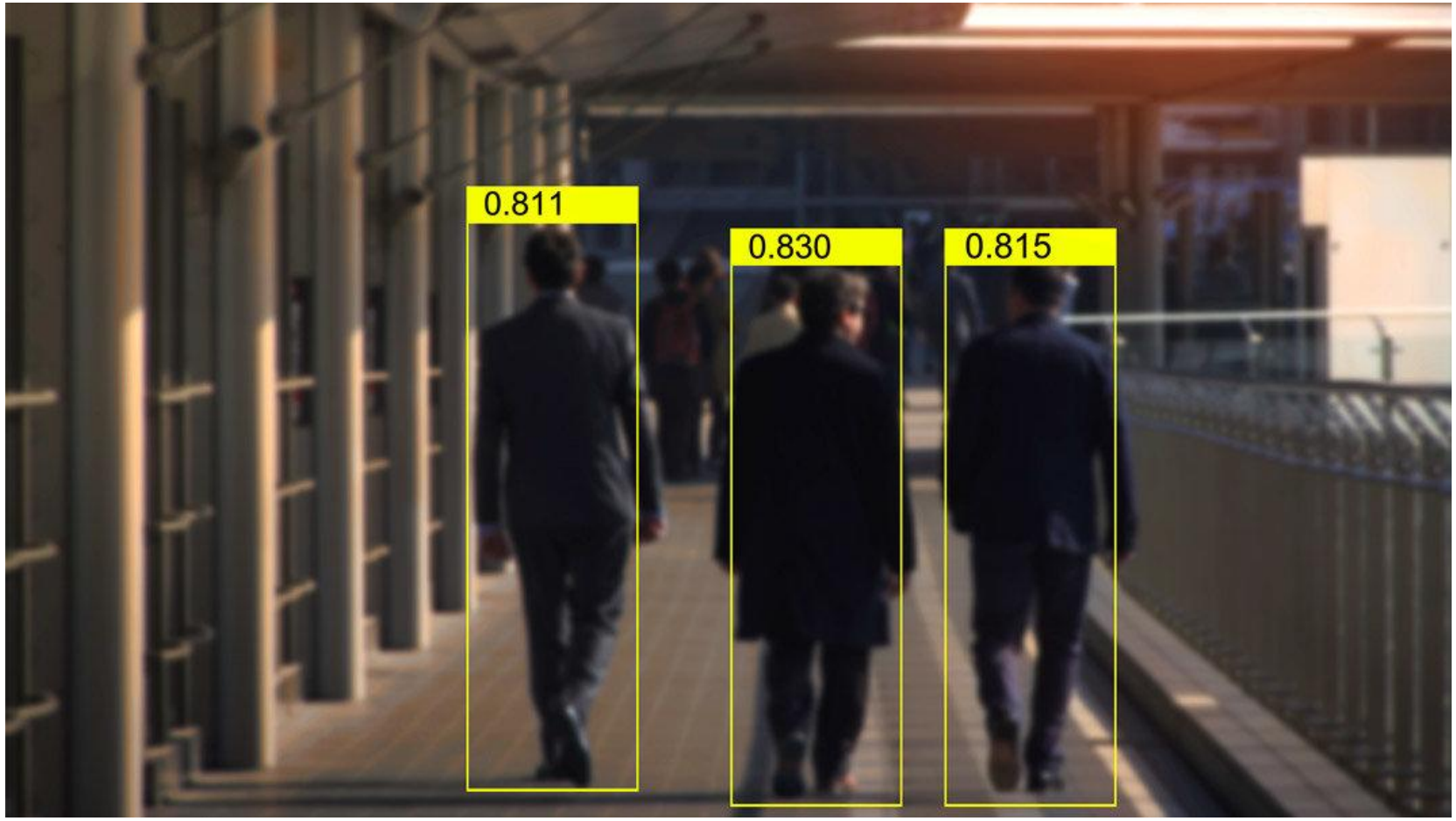
The film posed the dilemma of **predictability** vs. **free will**. This is a valid philosophical concern.

See more on this topic in Harari’s The myth of freedom.

But what if the problem arises on an even more trivial level?

“EMPIRICAL???”

LEARNS OUR BIAS AND PREJUDICE (DATA!!!)



SOURCE:
“Predictive policing is a scam that perpetuates systemic bias”
“Beyond GIGO: how "predictive policing" launders racism, corruption and bias to make them seem empirical”

Limitation 3:

“ATTACKABILITY”

“JUST SAY WHAT I SAY!”

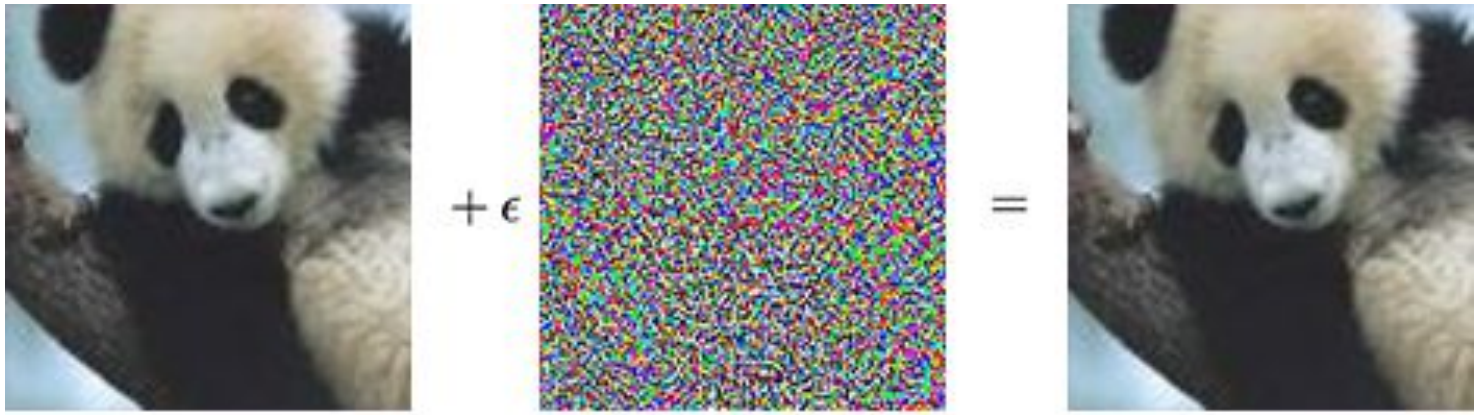
CAN BE ACTIVELY TAUGHT TO BIAS, PREJUDICE

SOURCE:
“ MICROSOFT DELETES RACIST GENOCIDAL TWEETS FROM AI CHATBOT”
“MAN IS TO COMPUTER PROGRAMMER AS WOMAN IS TO HOMEMAKER”

Powered by AI Partners

“JUST WATCH MY HAND!”

IT CAN BE CONFUSED WITH “ADVERSARIAL EXAMPLES”



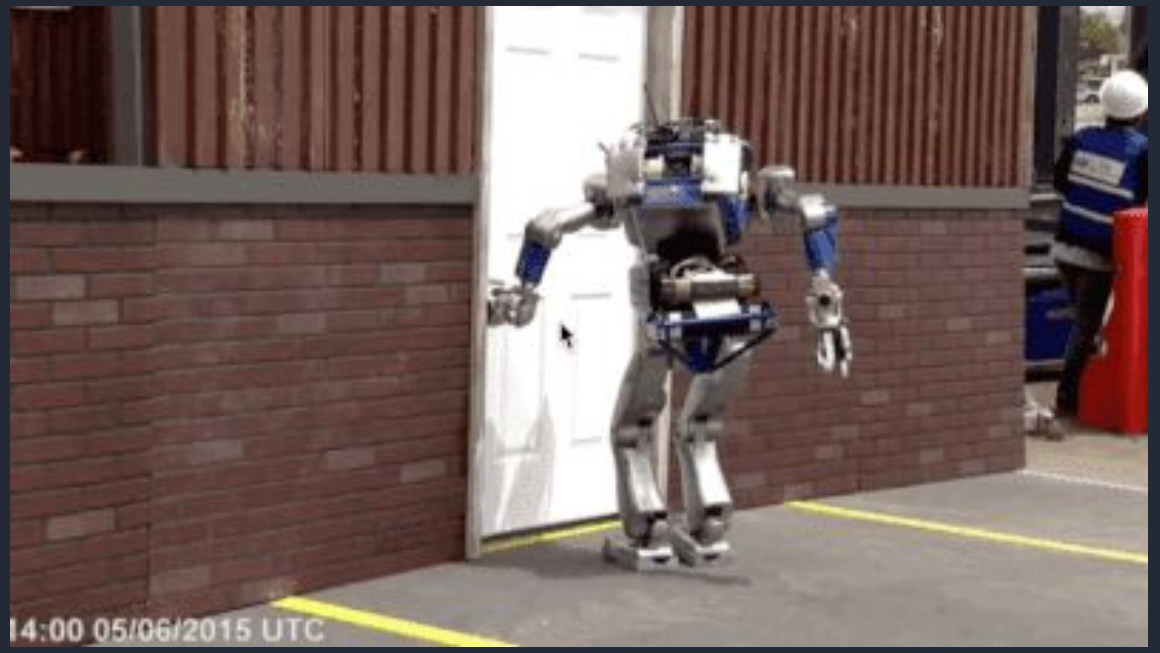
SOURCE:
“”ATTACKING MACHINE LEARNING WITH ADVERSARIAL EXAMPLES

Conclusion:

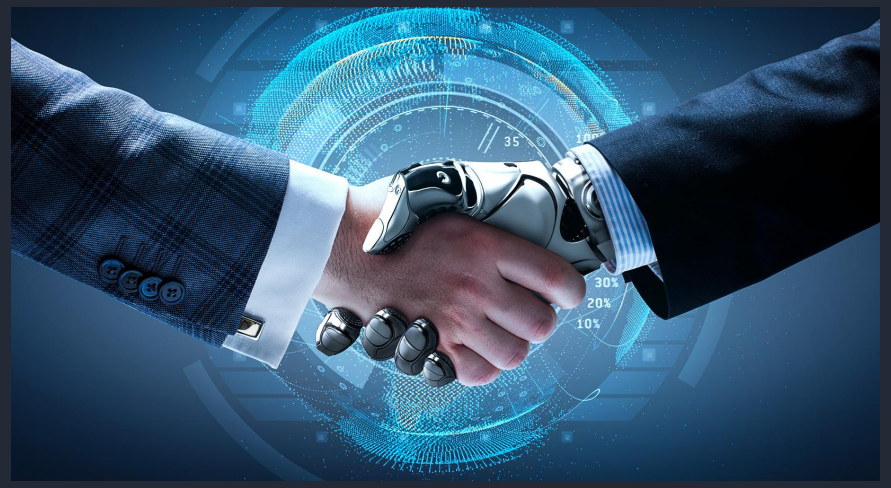
“COOPERATIVE SYSTEMS”

“INEFFECTIVE + FAILURE PRONE”

REPLACING YOUR WORKFORCE?



DO NOT!



DON’T REPLACE, AUGMENT!

COOPERATIVE SYSTEMS ARE MINIMIZING RISK

Artificial intelligence VS Augmented intelligence

Powered by AI Partners

NEURON SOLUTIONS