I I .AI

BULLETPROOF.AI

Machine Learning behind the Scenes

Pitfalls and Origin of Bias

Martin Rehak

AI
Disrupts Finance

- Immediate decisions, anytime
- Better decisions & pricing drive competition
- New markets

- **Immediate convenience**

BULLETPROOF.AI

Security solutions for AI, machine learning and automated statistical decisions

AI Models make critical business decisions in split seconds, every second of the day

How Secure, Fair and Robust is your Machine Learning System?

Artificial Intelligence is like an army of 5-year old kids.

(paraphrased from Alex Stamos)

Alex Stamos @alexstamos

Having access to the world's best machine learning is like having access to 10 billion five-year-olds.

If your task is "move that huge pile of bricks" then 10B kids are super helpful, but you can't ask them "build the Taj Mahal".

Alex Stamos @alexstamos · Apr 25

Replying to @alexstamos

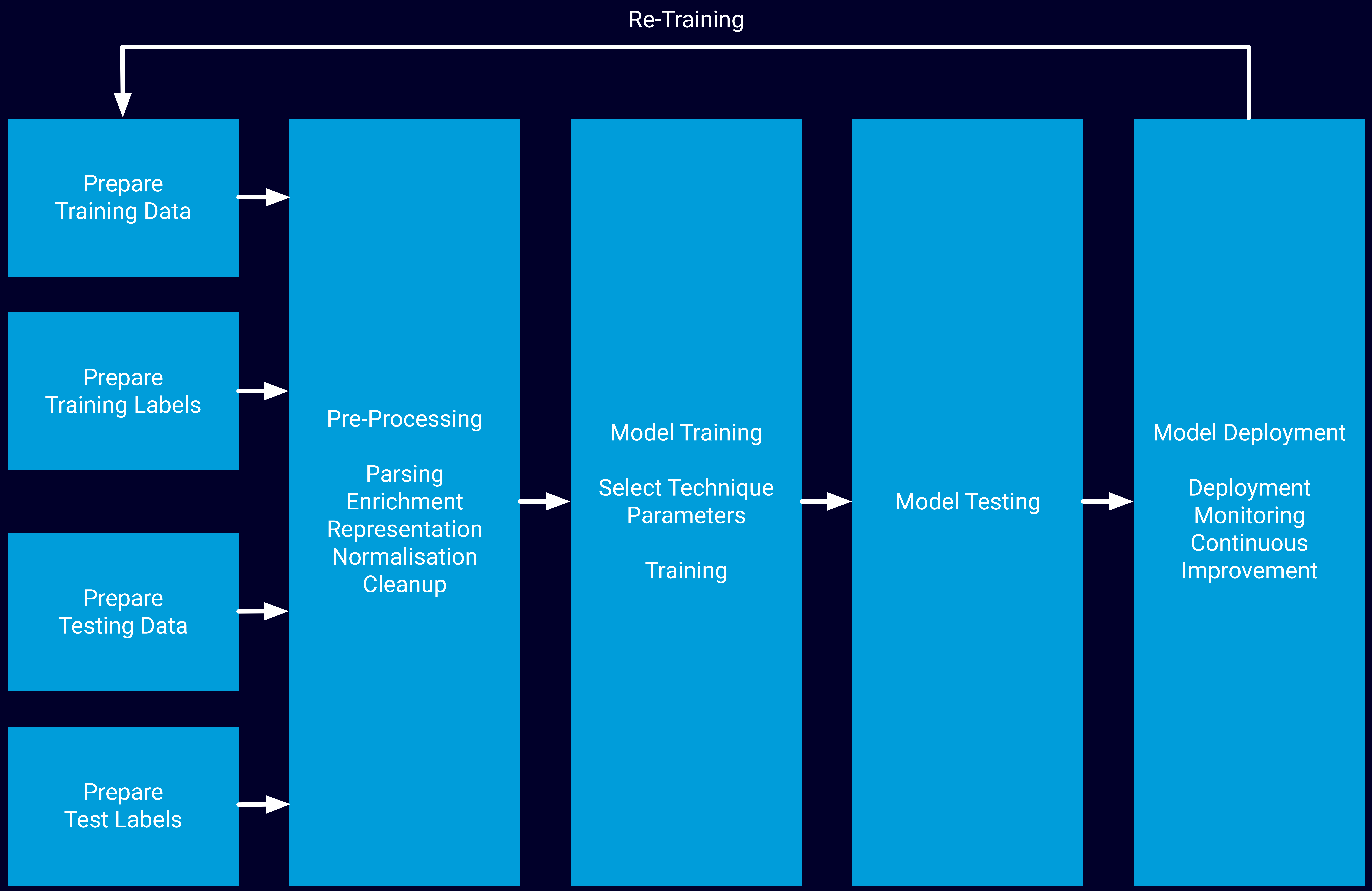
So yes, now that humans are looking at an example of a harmful video, it is trivial to pick out ML strategies to detect it. Telling computers "find all videos where people are being hurt" against an infinite search space of possibilities is AGI-hard.

3 26 148

Alex Stamos @alexstamos · Apr 25

One of the problems here is that tech execs like to say "we will fix it with AI" while thinking "...in five years and $1B" and the media hears "...next month" and the actual ML engineering director thinks:

How to manage the army of kids?


Re-Training
Prepare
Training Data
Prepare
Training Labels
Prepare
Testing Data
Prepare
Test Labels
Pre-Processing
Parsing
Enrichment
Representation
Normalisation
Cleanup
Model Training
Select Technique
Parameters
Training
Model Testing
Model Deployment
Deployment
Monitoring
Continuous
Improvement

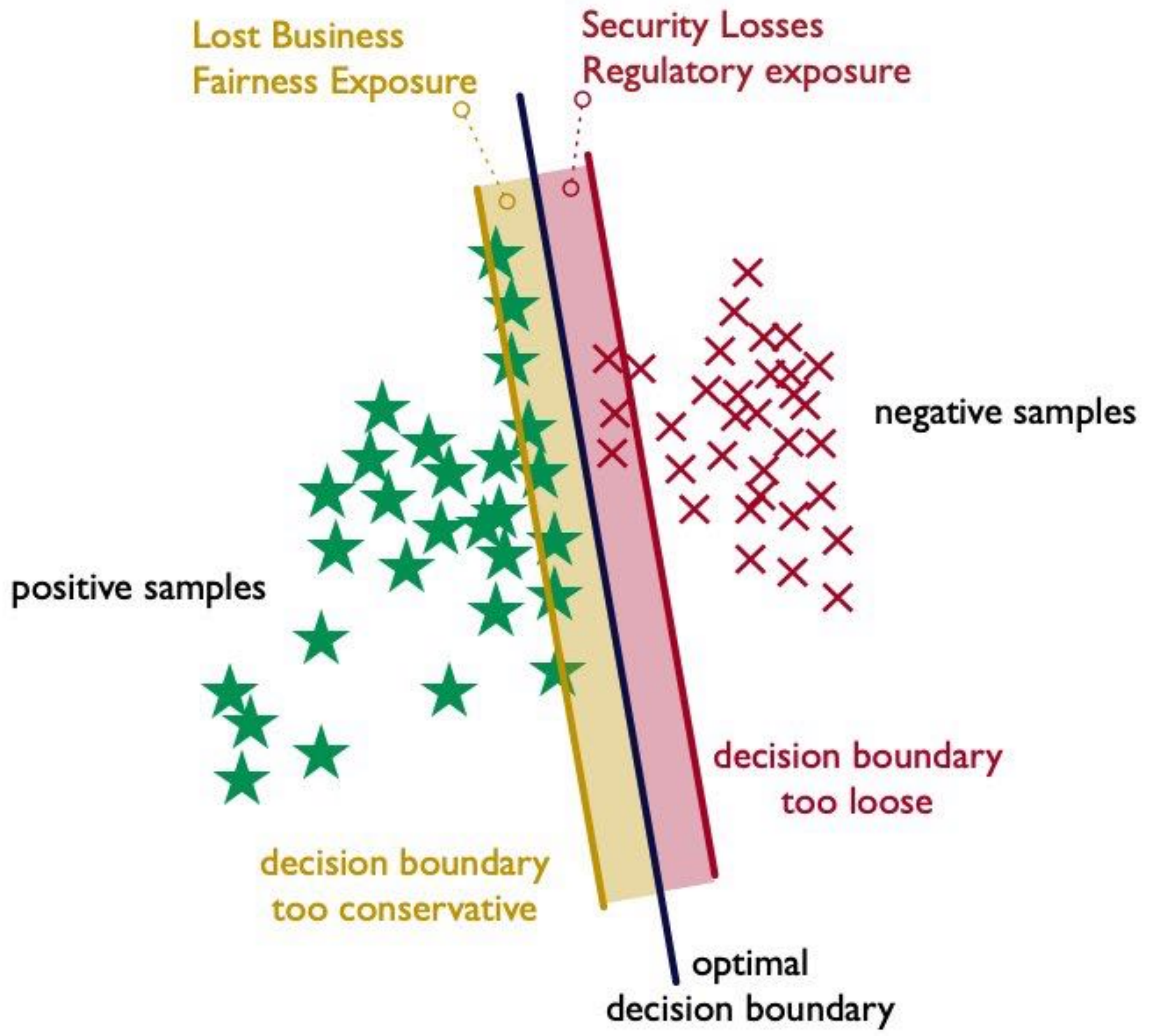
ML Time Investment

70% Data preparation

20% Labeling

9% Representation and Pre-Processing

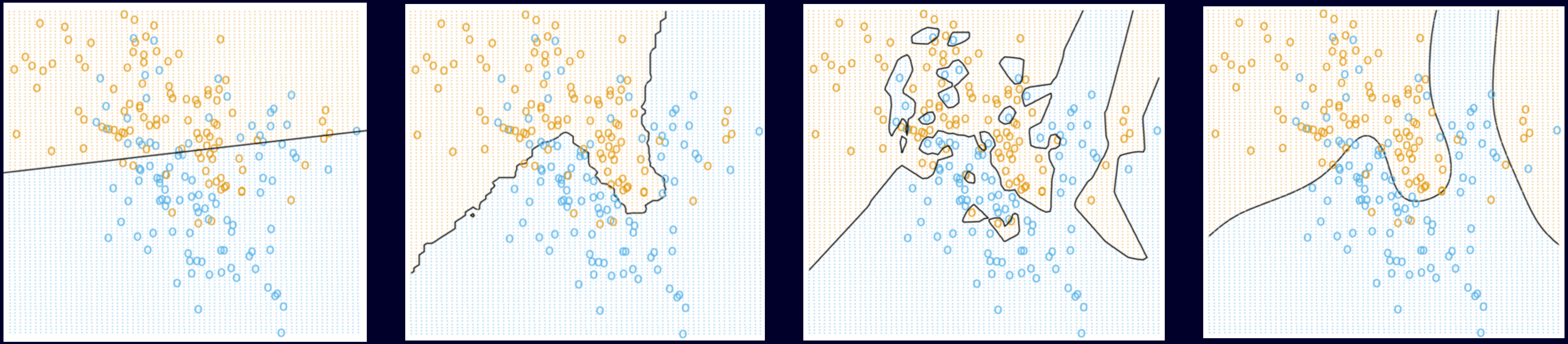
1% Model Training



Decision Boundary

- **Facebook effect:** posts on the edge of acceptable use policy get the highest engagement score, regardless of what the actual policy is.
- **Margin impact:** Business next to the decision boundary is less competitive and brings higher margins

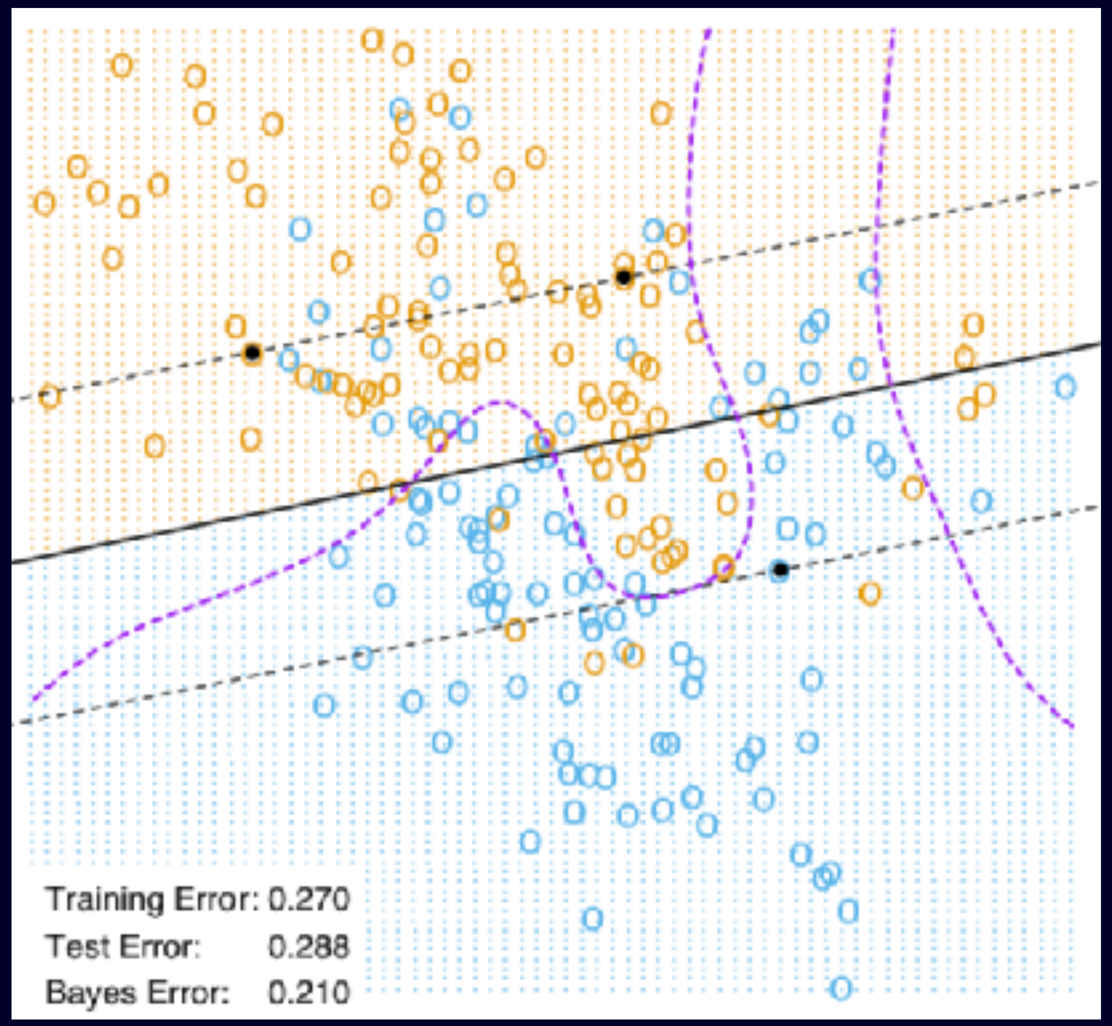
Algorithm Classes - Local vs. Global



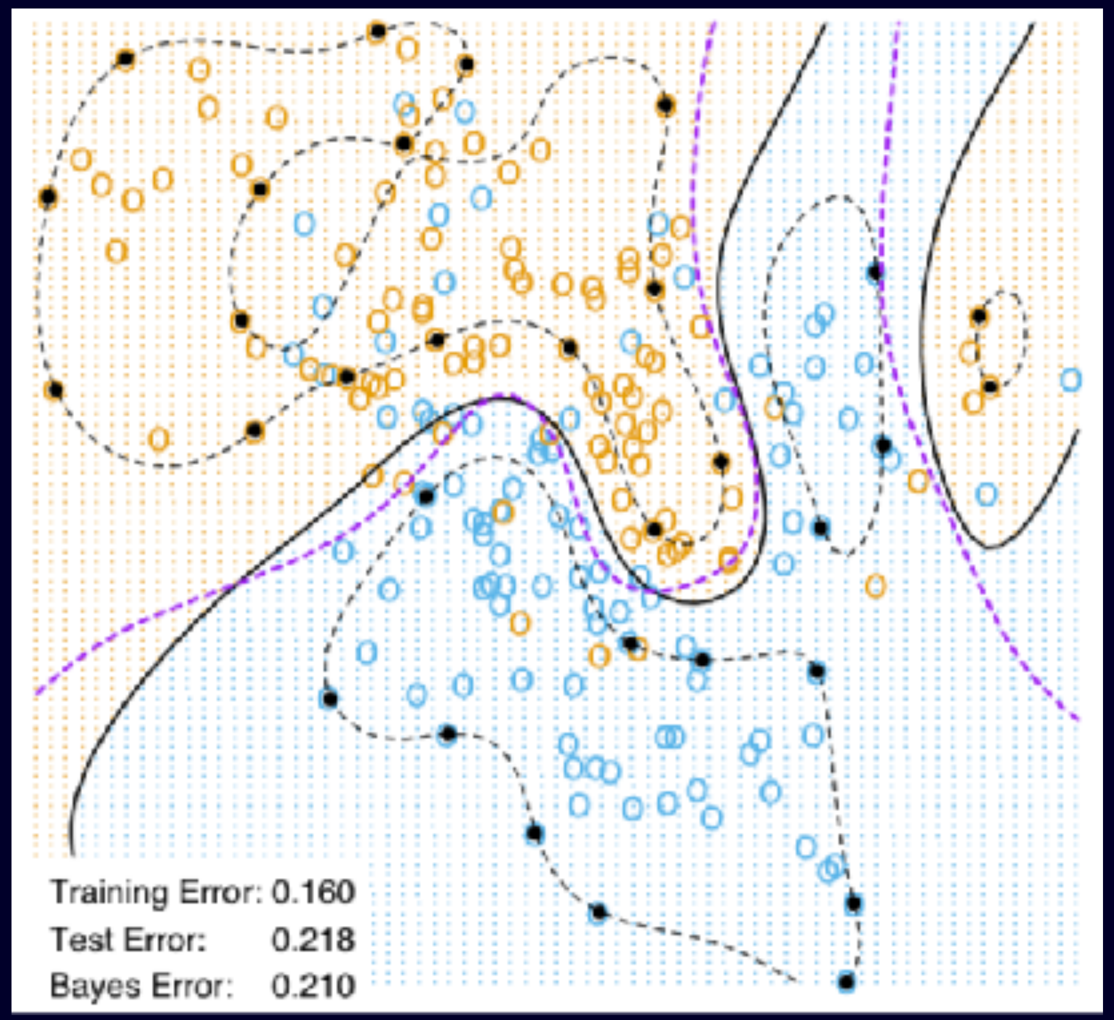
Linear Regression | 15-NN Classifier | 1-NN Classifier | Bayes Classifier

From Hastie et al.: The Elements of Statistical Learning, 2nd ed., 2008

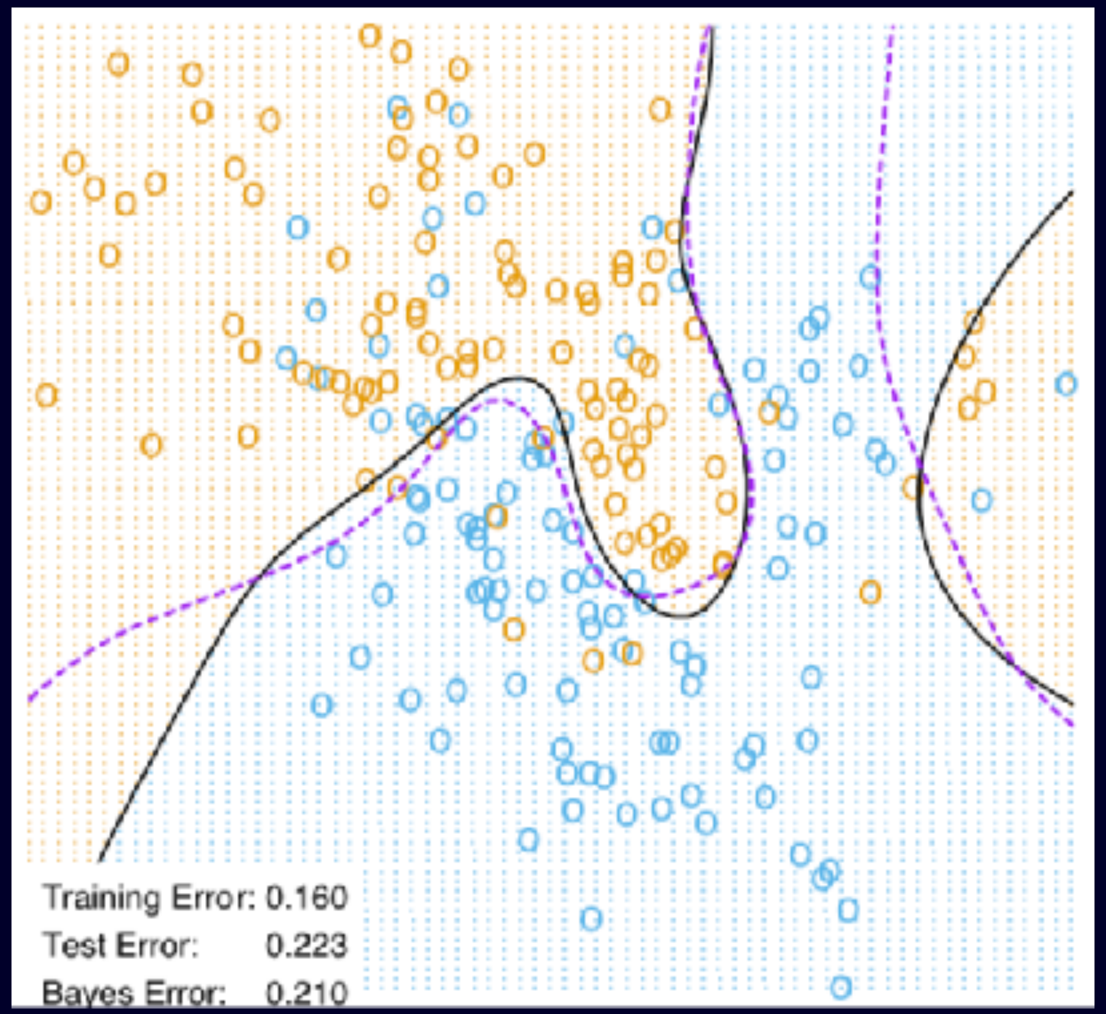
SVMs, Neural Nets and Random Forests



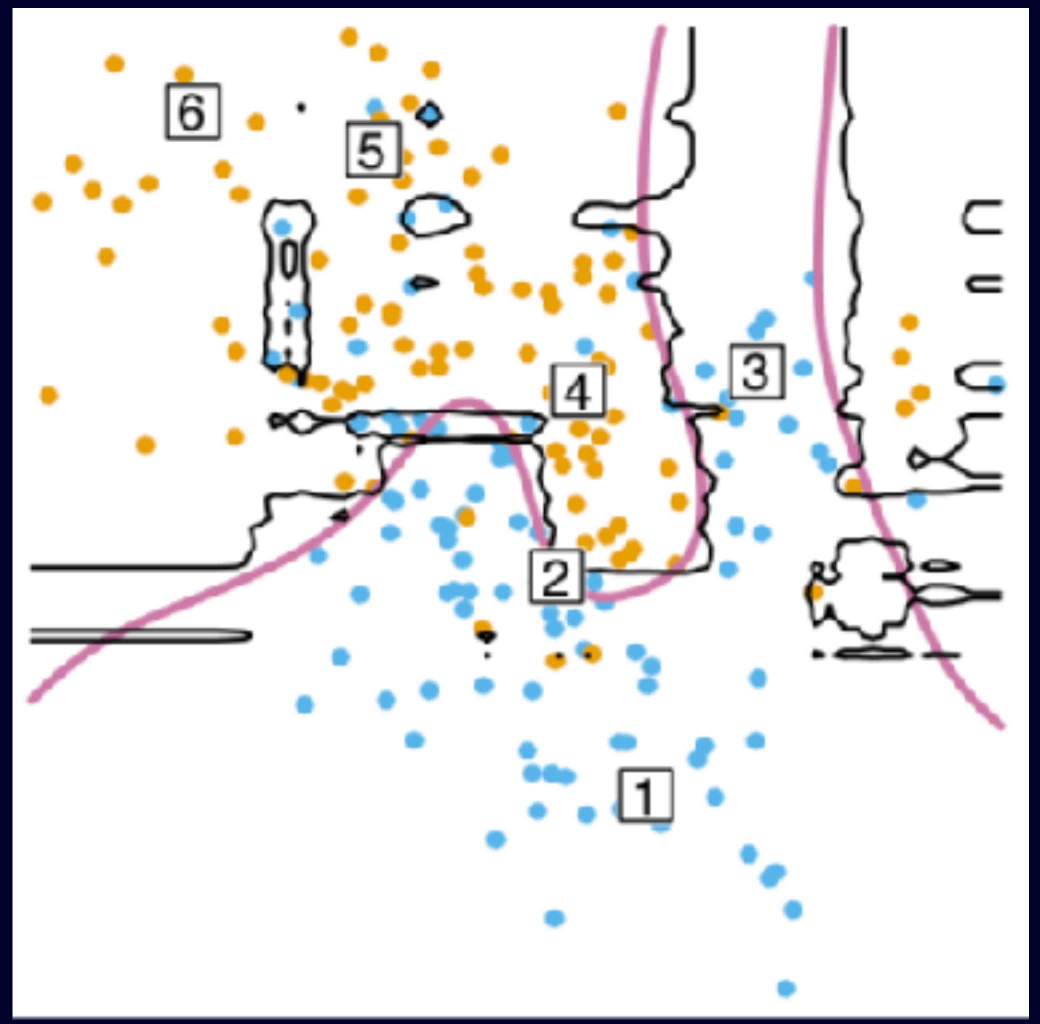
Linear SVM



SVM + Radial Kernel



Neural network

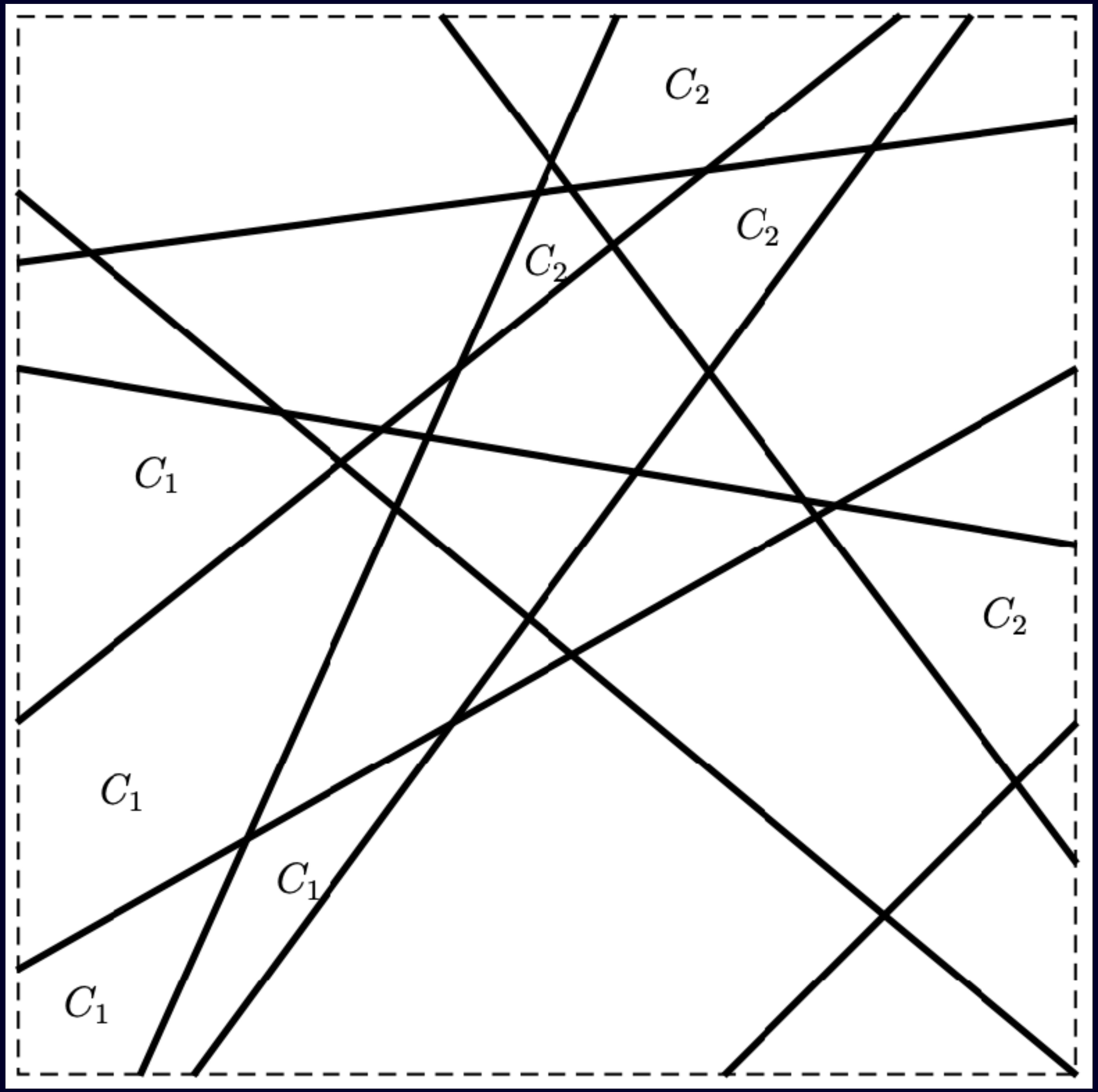


Random Forest

From Hastie et al.: The Elements of Statistical Learning, 2nd ed., 2008

The curse of dimensionality

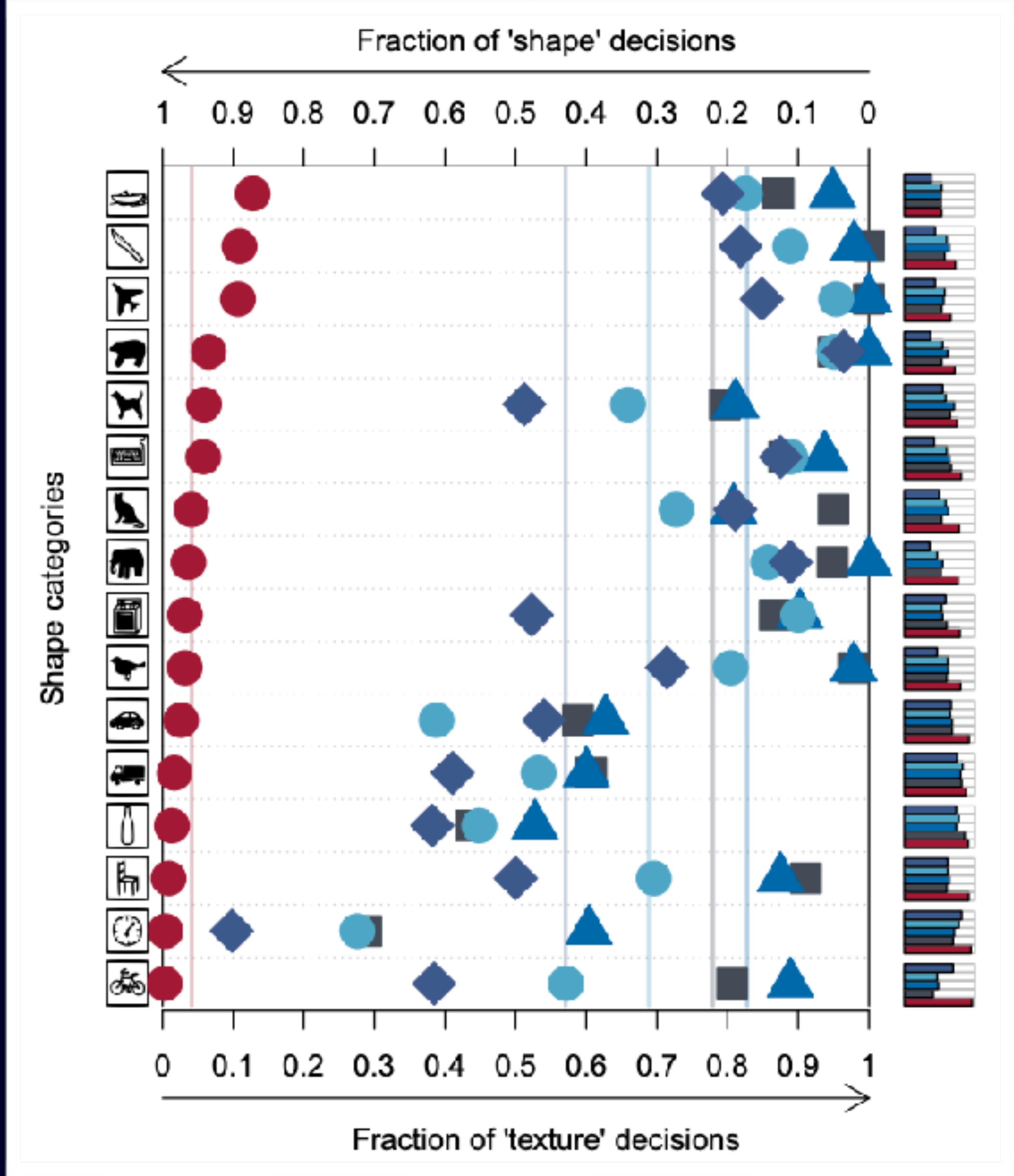
- With increasing dimension, properties of the space change dramatically:
- Eucleidian distance no longer has much meaning
- We are **always** just a tiny step away from a mistake - in some dimension(s)



Deep Networks and Details

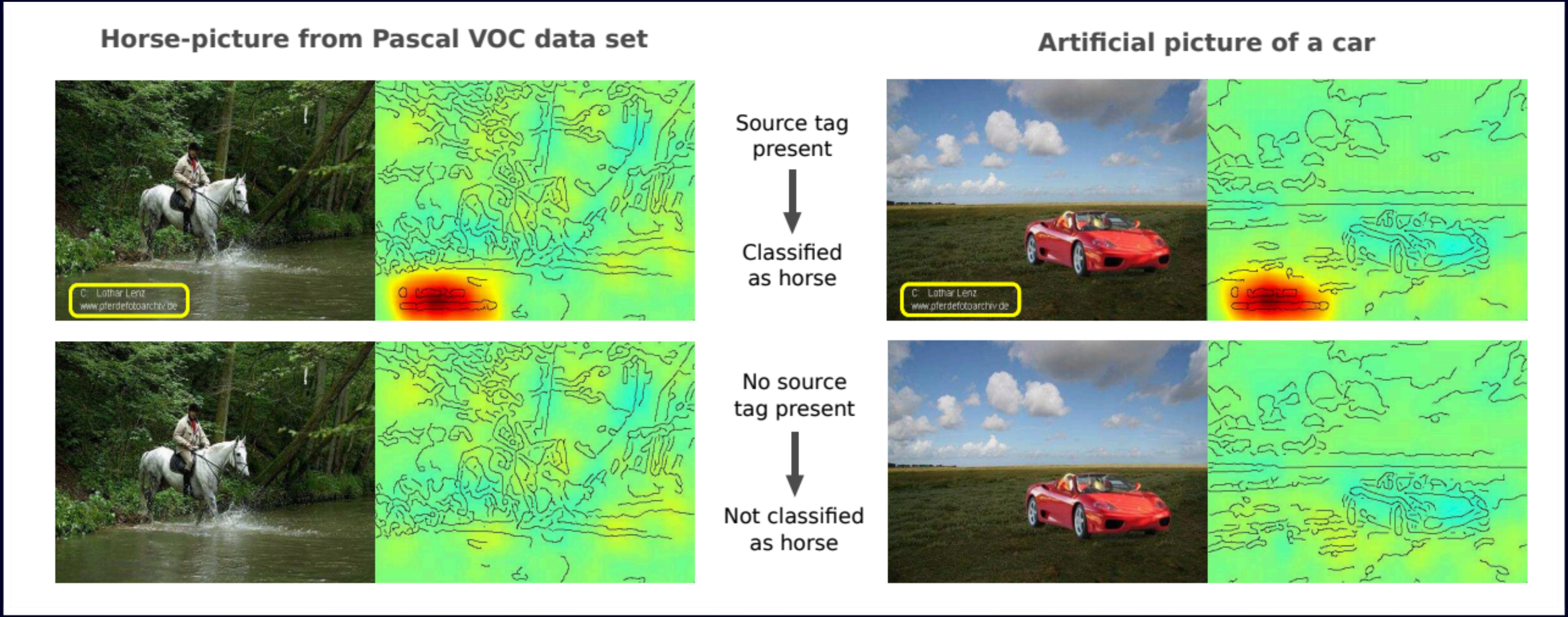
- Deep learning methods exhibit strong preference for detail at the expense of high-level concept extraction

cat with elephant texture | car with clock texture | bear with bottle texture



Geirhos et al.: ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness, ICLR 2019

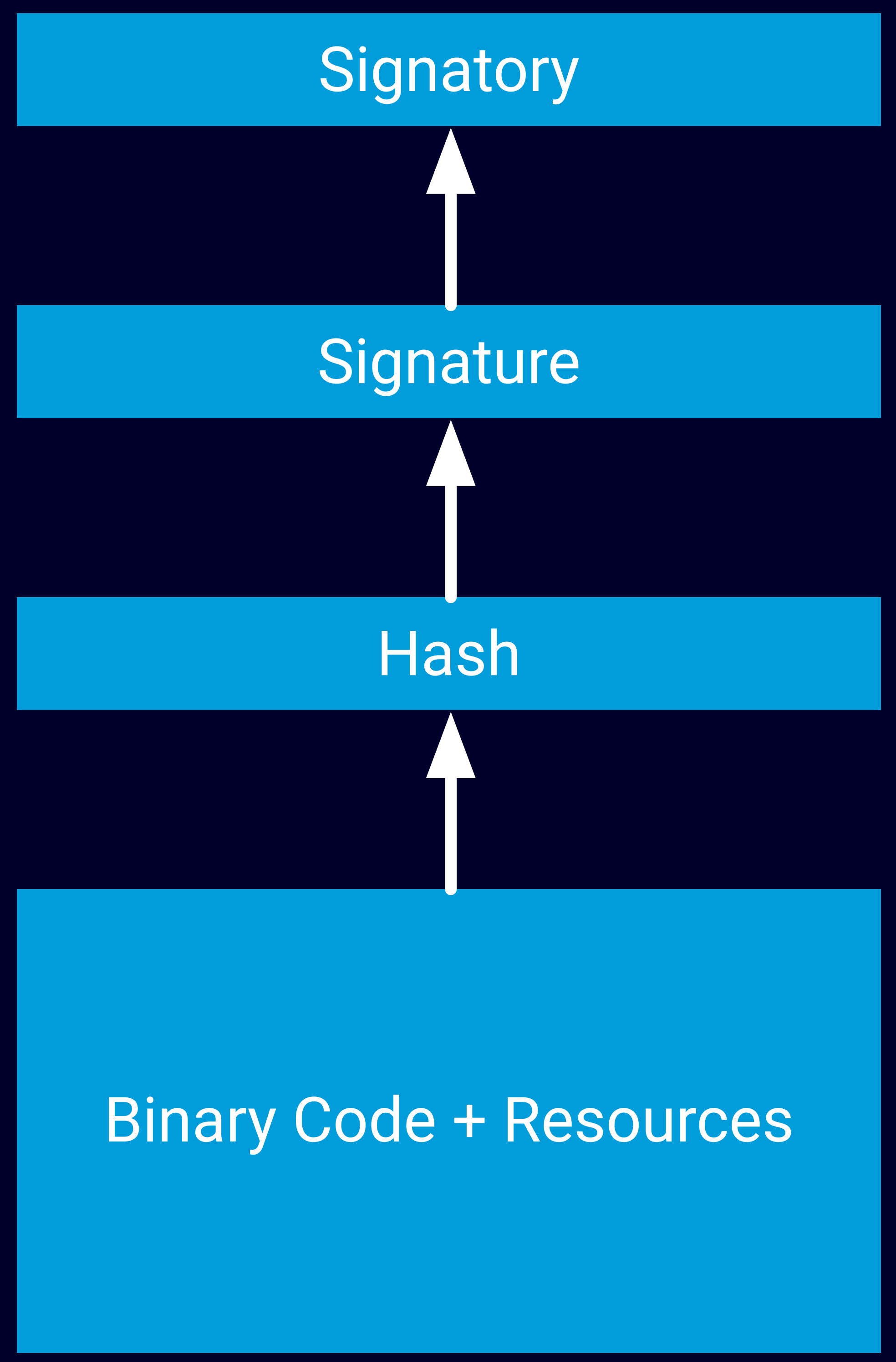
Deep Networks and Details



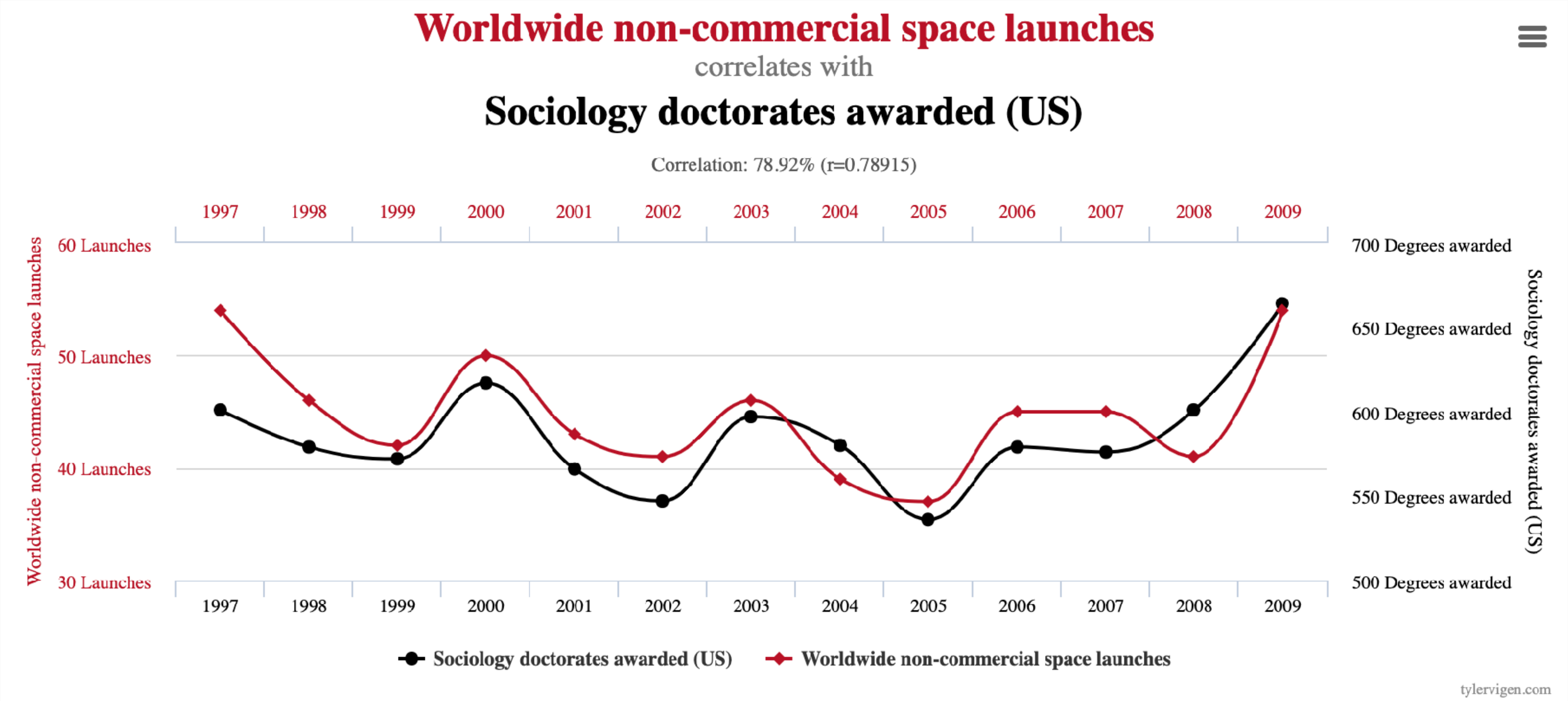
Lapuschkin et al. "Unmasking Clever Hans Predictors and Assessing What Machines Really Learn", Nature Communications, 2019.

Example - Signed Executables

- Identification of relationships is hard:
  - Executable is hashed
  - Hash is signed (PKI)
  - Signature is from the right signer
  - Revocation



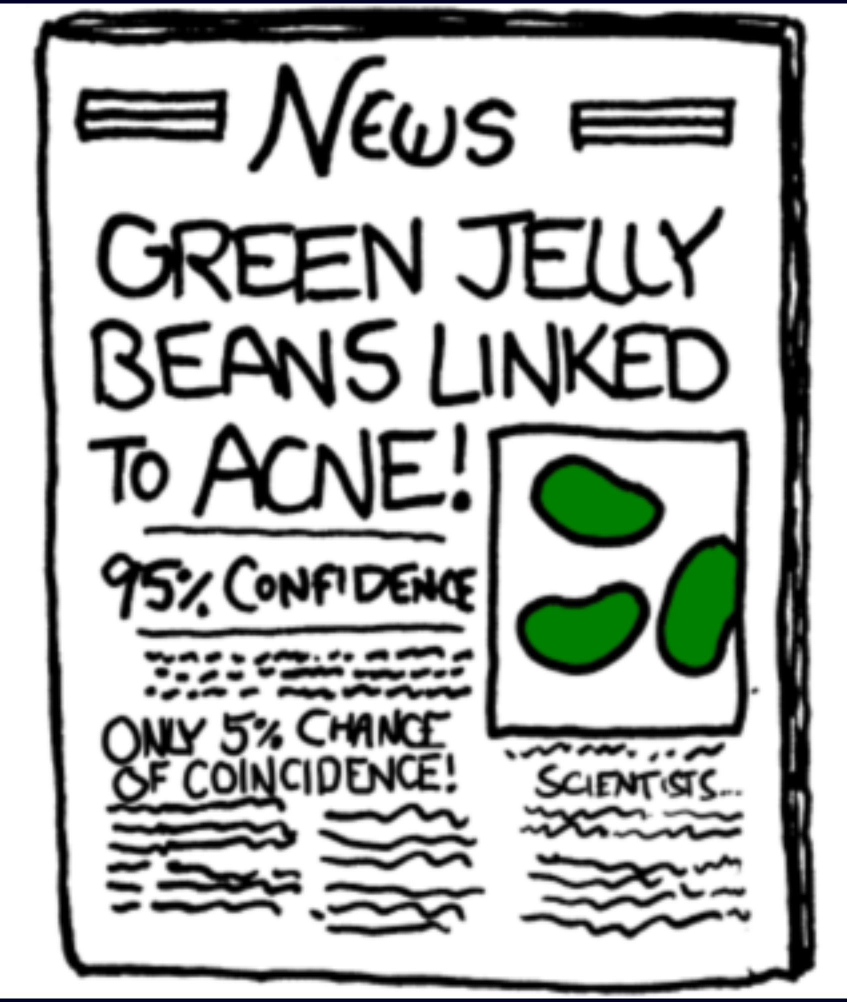
- Do you have good training samples for all combination of errors?
  - Hash-Code mismatch
  - Bad signature
  - No certificate
  - Certificate/signature mismatch
  - ...



Data sources: Federal Aviation Administration and National Science Foundation

Overfitting

- **With enough features, you can always find relationship with any label set**.
- Models with huge dimensions and low training data richness effectively perform p-value hacking.
- Training can formulate arcane, super-complex hypothesis to achieve perfect performance on the training set.
- But testing set would save us, right?
- Not always:
  - Artefacts present in both testing/training set.
  - Information leakage from cross-validation.
  - Bias in the data.



Amazon HR system

REUTERS BUSINESS NEWS

OCTOBER 9, 2018 / 11:12 PM / UPDATED 11 HOURS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

SAN FRANCISCO (Reuters) - Amazon.com Inc's (AMZN.O) machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

- Text analysis: Huge number of features available to the system.
- Problem: System refuses to hire women candidates (based on the past decisions).
- Fix 1: Explicit sex/gender field removed.
- Fix 2: The system then started using his/hers salutations - clean-up.
- Fix 3: Sports, schools and other hard-to-remove features surfaced…
- Project canceled.

Overfitting Consequences

- Overfitting breaks the classifier ability to generalise and turns it into a memory system.
  - Can be actually useful for specific applications, such as malware family detection - classifier is a fuzzy "hash" function.
- Don’t expect any predictive capability from an overfitted classifier.
- Is overfitting really a problem?
  - House number as a criteria for credit
  - Specific user-agent makes the loan accepted
  - Exact value of salary used in the criteria


Any good news?

Scientific Approach

- Use scientific approach to the problem.
- Before building a classifier, formulate a **hypothesis**.
  - Hypothesis should postulate a relationship between the features and the label.
  - Training process selects the features that predict/explain the labels.
- Training set richness (size/diversity) limits the complexity of the relationship that can be correctly identified.
- If you don't have enough training data, reduce the feature set or break-down the problem.

Divide and Conquer

- Breaking down the problem often yields more stable solution:
  - **Ensemble** methods offer strong ways how to build a collective classifierd
  - **Specialised classifiers** can tackle well defined part of the problem, with their output being used as input for other classifiers - more efficient use of training set & features
  - **Dedicated classifiers** addressing part of the problem can be simpler: e.g. fraud vs. non-intetional default

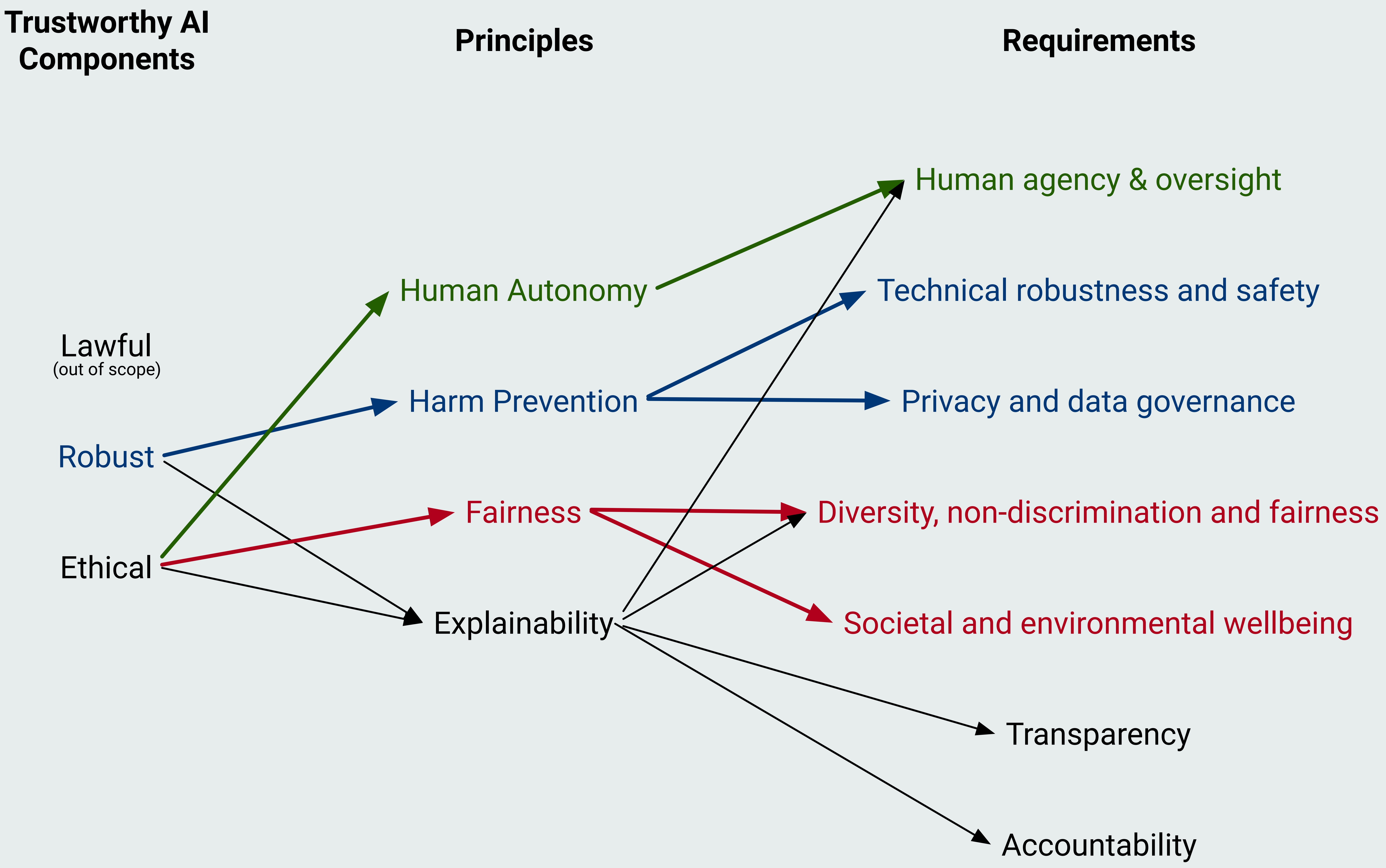
Series of Classifiers

- Limited/adjustable autonomy
  - Combine simple, easy to understand classifiers with sophisticated ones
  - Simple classifiers used as policy **guardrails** - define the set of strategies allowed by the user.
  - Sophisticated classifier can optimise within the safe bounds defined by guardrails
  - Automated reaction or escalation to human in case of breach
  - Frequently used in trading context


How can we control AI?

EU Trustworthy AI Guidelines

- Issued in April 2019
- Independent informal guidelines
- Include assessment checklist
- Formal AI regulation would be premature
- Sector-specific regulation should be applied if appropriate
- Piloting, Revised version scheduled for 2020



Main Relationships between Components, Principles and Requirements - Grossly Oversimplified

Requirements: Design Phase

Principles	Requirements	Detailed Requirements	Checklist
Human Autonomy respect	**Human agency & oversight**	Fundamental rights	Does the system operation negatively affect fundamental human rights?
		Human agency	Are the users empowered to make informed decisions in their interaction with the system? Does the system’s fully automated decision significantly impact the user, including legal effects?
		Human oversight	Does the system include appropriate human oversight mechanism using the appropriate approach - human-in-the-loop, human-on-the-loop or human-in-command?
Harm Prevention		Dual-Use system	Can the system be mis-used by malicious actors?

Requirements: Design Phase

Principles	Requirements	Detailed Requirements	Checklist
Transparency & Explainability		Effects on organisation	What is the algorithm’s effect on organisational culture, decision-making process and business model?
		Communication	Is user aware of the nature of the system, limitations and conditions of use? Are the limitations accurately described? Is there a human-based fallback?
Fairness	Diversity, non-discrimination and fairness	Stakeholder participation	Have stakeholders affected by the system been appropriately informed and consulted?

Requirements: Design Phase

Principles	Requirements	Detailed Requirements	Checklist
Fairness		Stakeholder Participation	Have stakeholders affected by the system been appropriately informed and consulted?
Fairness	Societal and environmental wellbeing	Sustainability, environmental friendliness	Is the system adoption and usage environmentally friendly? E.g. Does it replace a more labour/energy/material intensive process? Does it indirectly incite higher energy consumption?
		Social Impact	Have you considered the system's (mostly) indirect impact on social well-being and user's emotions?
		Society & Democracy	Have you assessed the effects of the system on democratic process and political institutions?

Requirements: Design Phase

Principles	Requirements	Detailed Requirements	Checklist
Explainability	Accountability	Minimisation of negative impacts and their reporting	Is there an appropriate process for internal and external reporting of negative system impacts, ensuring protection of reporters? Are the reports effectively used to improve the system?
		Trade-offs	Have the tradeoffs between the above-listed non-functional requirements (and functional requirements) been properly acknowledged and documented? Accountability of decision makers and ongoing tradeoff-management process in place.
		Ability to redress	Is there an appropriate redress mechanism with corresponding capacity?

Requirements: Implementation & Train

Table 1

Principles	Requirements	Detailed Requirements	Checklist
Harm Prevention	**Technical robustness and safety**	Fallback solution	Do you have a fallback plan in place to address attacks, wrong decisions or other failures? Do you have a failure impact model?
Harm Prevention	**Privacy and data governance**	Privacy & data protection	Do you protect explicitly or implicitly stored information about the users? Do you do this in all lifecycle stages? Do you follow the least-information principle?
		Data quality and integrity	Is the data collected accurate-enough for the purpose of the classification task? Do you protect the system from adversarial manipulation?
		Access control to data	Do you follow need-to-store and need-to-access approach to data access management?

Requirements: Implementation & Train

Principles	Requirements	Detailed Requirements	Checklist
Transparency & Explainability		Traceability	Document the data sets, processes and tools used to build the classifier and reach the decision. Logging design.
Fairness	**Diversity, non-discrimination and fairness**	Unfair bias avoidance	Are the decisions taken by the system fair and unbiased? Have precautions been taken to eliminate pre-existing bias in the training data or the process being replaced?
		Accessibility, universal design	Is the system accessible and usable by all relevant groups according to age, gender, abilities or characteristics?
Explainability	**Accountability**	Auditability	Can the system be audited by authorised third-parties?

Requirements: Empirical & Runtime

Table 1

Principles	Requirements	Detailed Requirements	Checklist
Harm Prevention	**Technical robustness and safety**	Security - AML resilience	Consider the possible attacks, nature of vulnerabilities and the threat model of the system?
			Have you verified system behaviour under realistic deliberate attack?
			Have you designed, deployed and tested appropriate security mechanism?
			Have you verified environmental assumptions and verified the effects of breached assumptions and ynexpected situations?

Requirements: Empirical & Runtime

Principles	Requirements	Detailed Requirements	Checklist
Harm Prevention	**Technical robustness and safety**	Accuracy	Does the system reliably produce the decisions with sufficient accuracy for the given application?
			Can you detect inaccuracies before they cause harm, either individually or systematically?
		Reliability	Reliability - can the system be trusted in a wide range of situations, and have you identified all features and their lineage correctly?
		Reproducibility	Reproducibility - Will the system exactly reproduce its behaviour under the same circumstances?

Requirements: Empirical & Runtime

Principles	Requirements	Detailed Requirements	Checklist
Harm Prevention	**Privacy and data governance**	Data quality and integrity	Is the data collected accurate-enough for the purpose of the classification task? Do you protect the system from adversarial manipulation?
Transparency & Explainability		Explainability	Can you explain the decision taken by the system to humans? Reason about tradeoffs with accuracy. Emphasise explainability for decisions with major impact on people's lives.
Fairness	**Diversity, non-discrimination and fairness**	Unfair bias avoidance	Have you empirically assessed the system bias for known bias risks and for unknown bias that may have been introduced while building the system?

And in practice?

Measuring and Assessing AI

- **Implementation-agnostic** assessment, based on frequent measurement
  - **Data-centric** - assess the training/testing/validation/production data
    - Ratio between model complexity (features and method) and data richness
- Empirical
  - Bring your **own samples** & distributions for testing
- Better **Stress-Testing**
  - Test fine-grained hypothesis (automotive decline or organised attack)
- **Continuous** measurement of production system performance

Machine Learning Makes Us Safer

- ML provides more precise and **individual decisions**
- ML also comes with a set of finer-grained, more **individual risk** measurements

- ML enables more **frequent** model updates and lower obsolescence risk
- ML brings faster innovation for better **resilience** against attacks