Python programming and machine learning

ILCB Summer School 2023 - Class 4/4

Thomas Schatz, August 31st 2023

The course material is available on the summer school amubox and at https://thomas.schatz.cogserver.net/teaching/

Tentative outline

Class 1

- What is machine learning and why should you care ?
- Case study introduction: classification of cat and dog images

Class 2

- Programming basics (in python)

Class 3

- Implementing classification of cat and dog images (with linear and nearest neighbor classification algorithms)

Class 4

- Testing our implementation
- General discussion

Actual outline

Class 1

- What is machine learning and why should you care ?
- Case study introduction: classification of cat and dog images

Class 2

- Programming basics (in python)

Class 3

- Programming basics (in python)

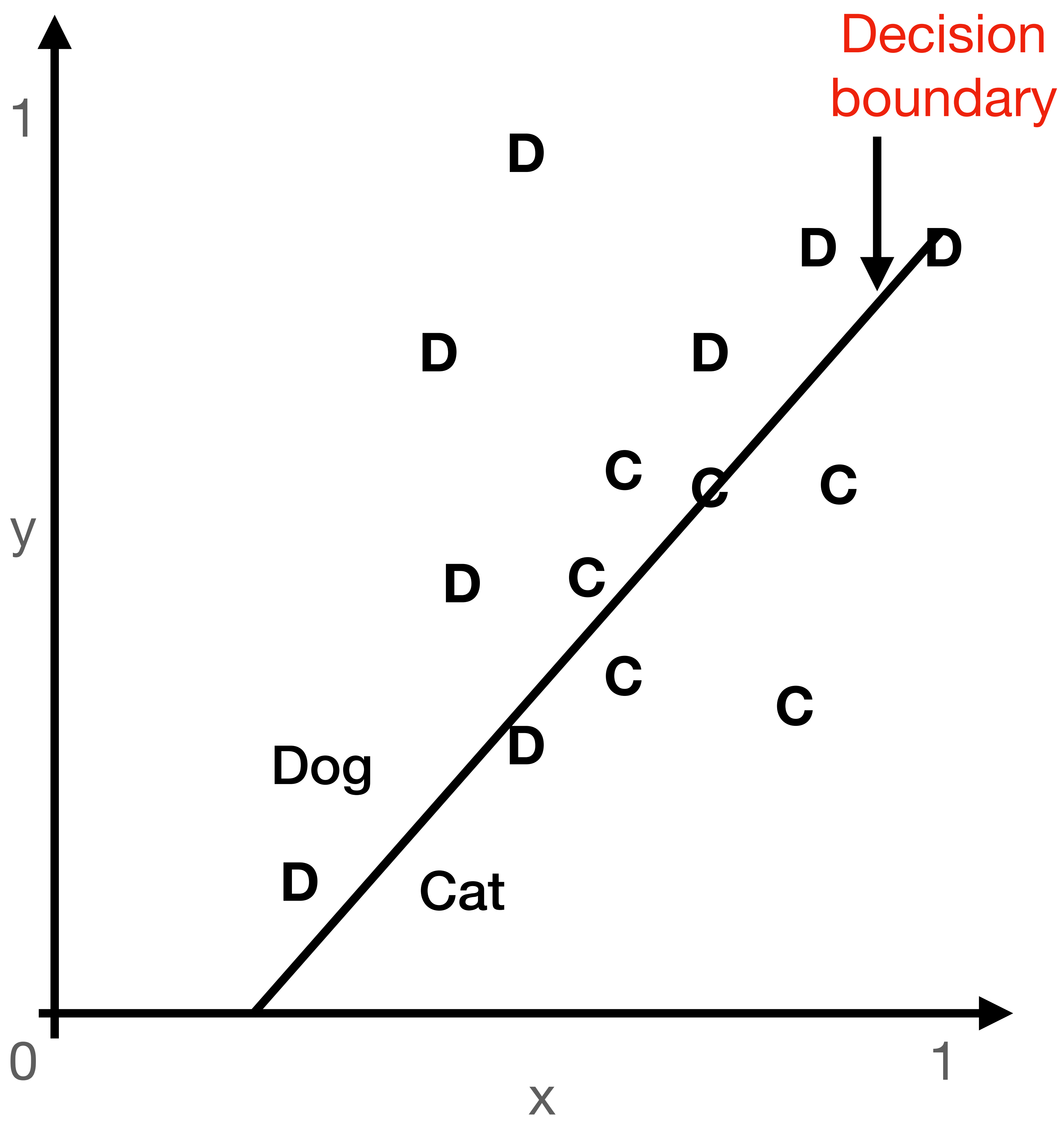
Class 4

- Implementing classification of cat and dog images (with linear and nearest neighbor classification algorithms)
- (Testing our implementation)
- General discussion (short)

Class 4

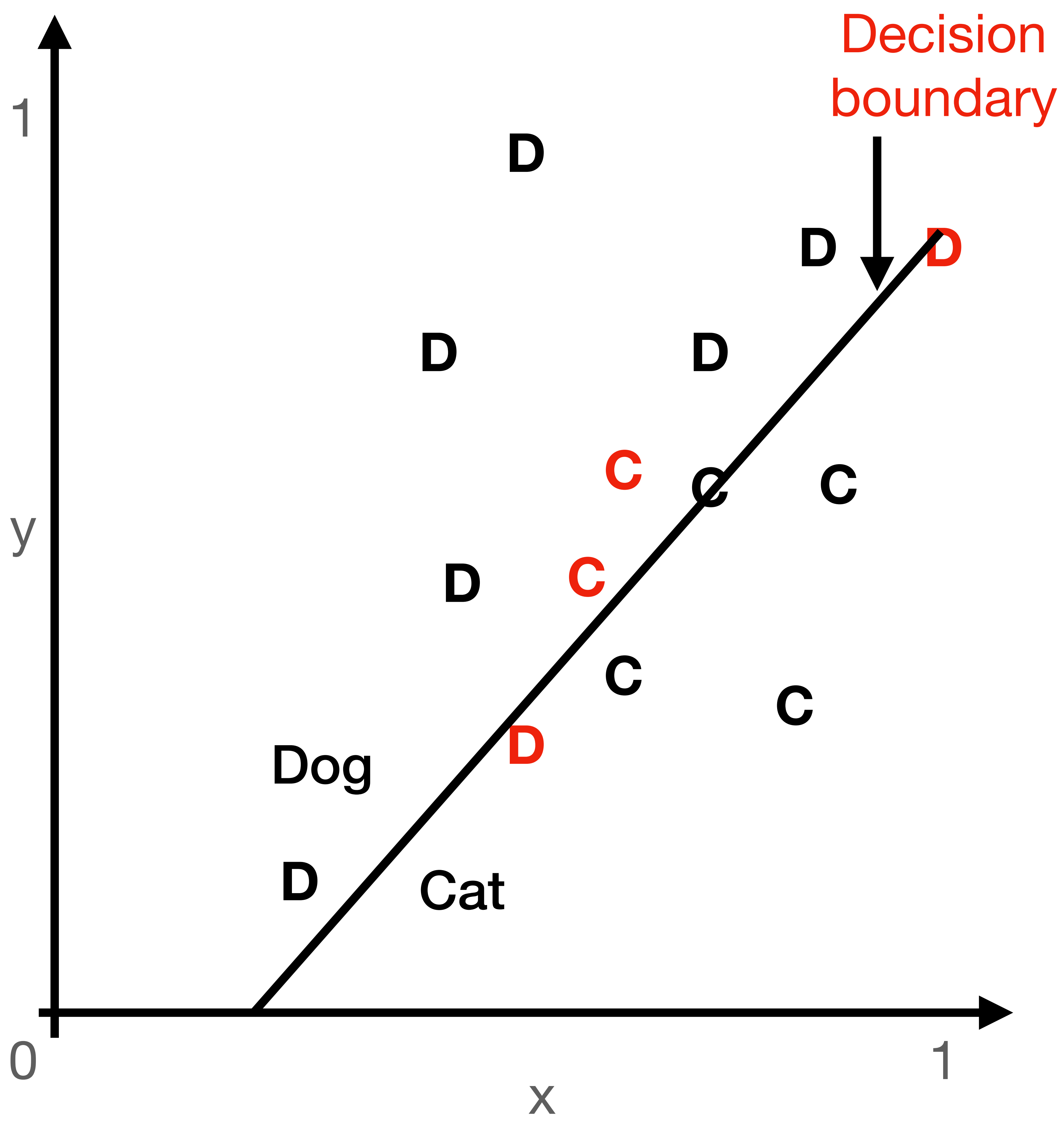
- Implementing classification of cat and dog images (with linear and nearest neighbor classification algorithms)
- (Testing our implementation)
- General discussion (short)

Case study: classification of cat and dog images



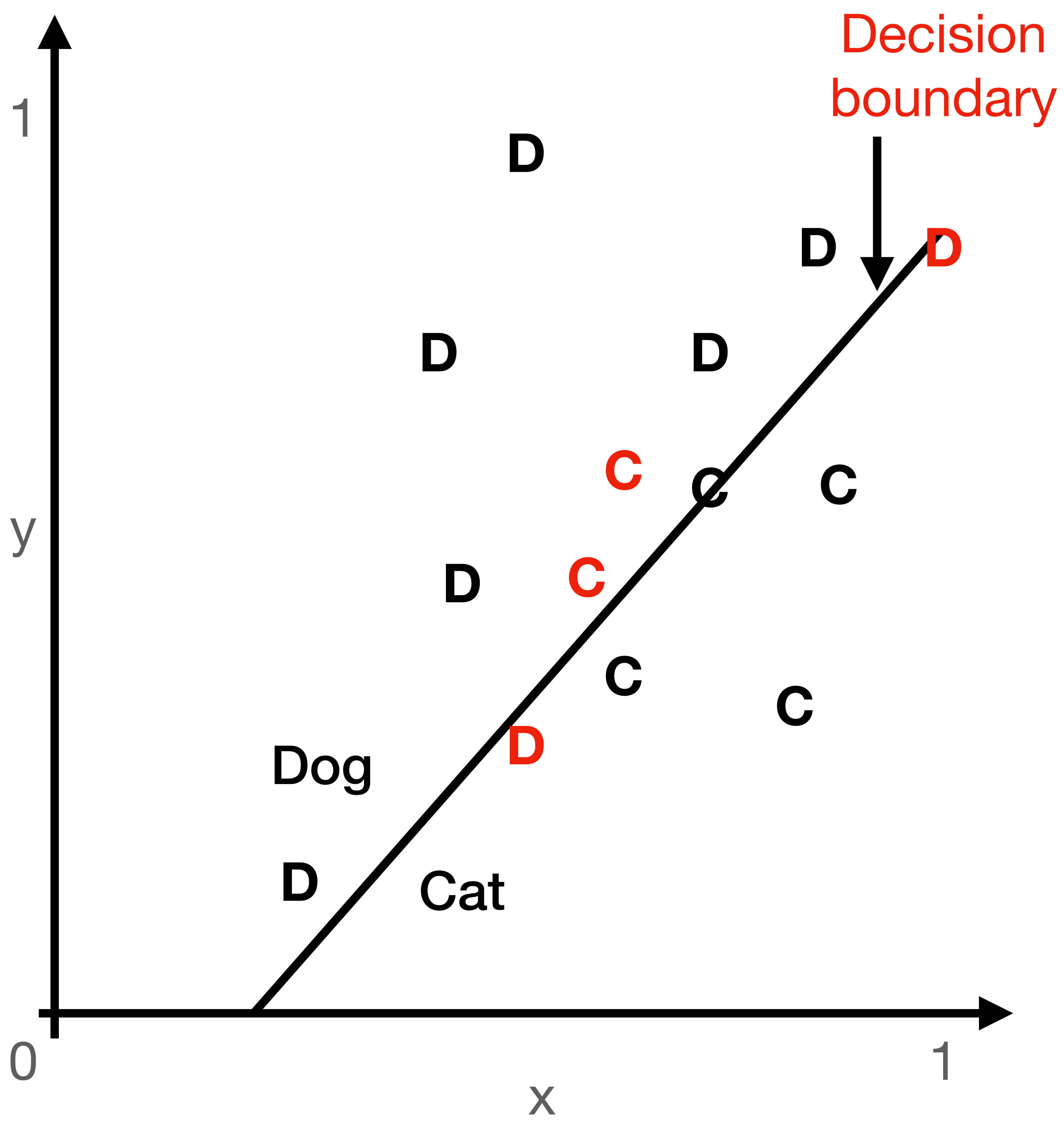
Linear classification

Case study: classification of cat and dog images



Linear classification

Case study: classification of cat and dog images

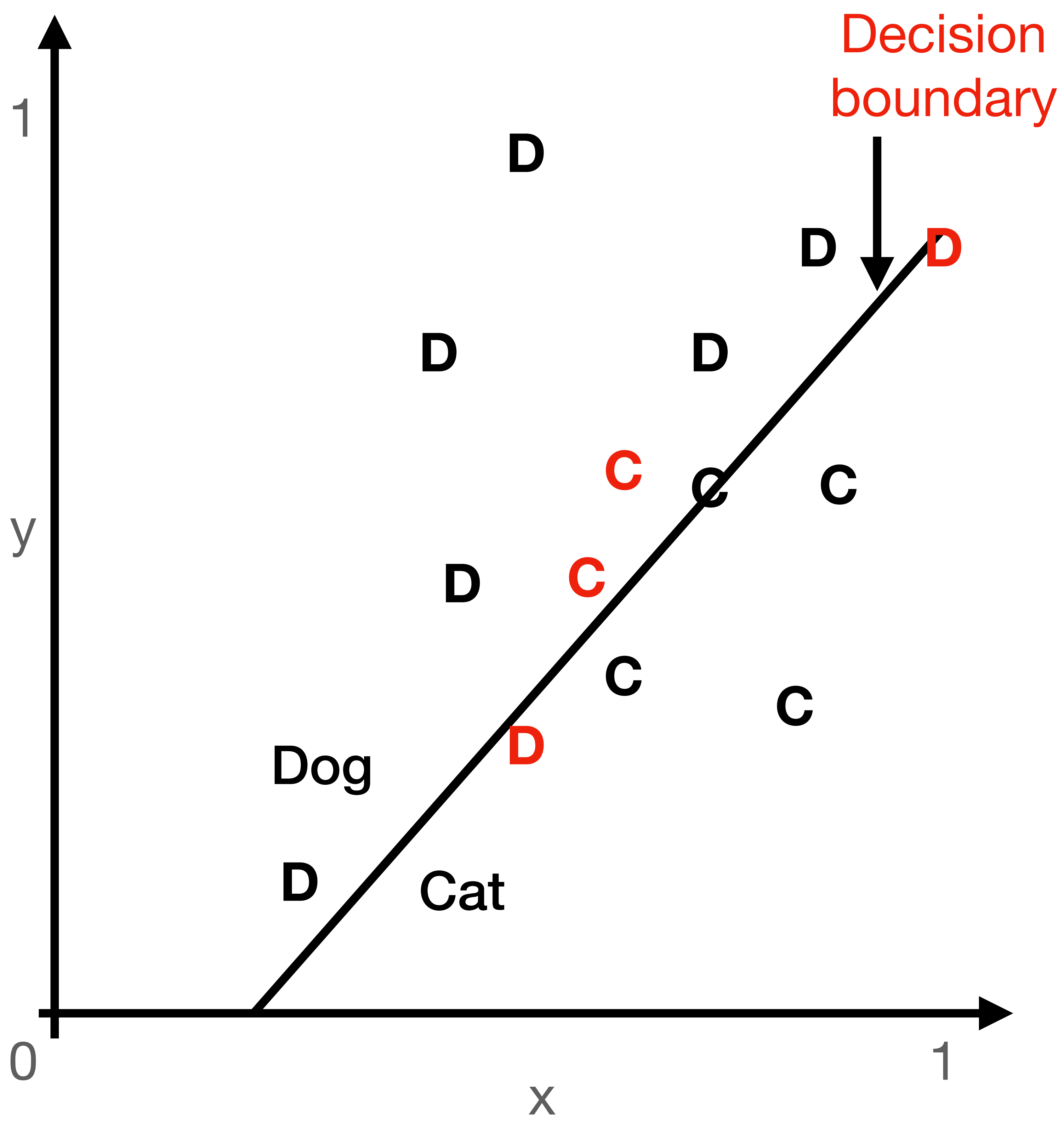


Which line/hyperplane?

“Empirical risk minimisation” idea:
pick the line that minimises
classification error on training set

Linear classification

Case study: classification of cat and dog images



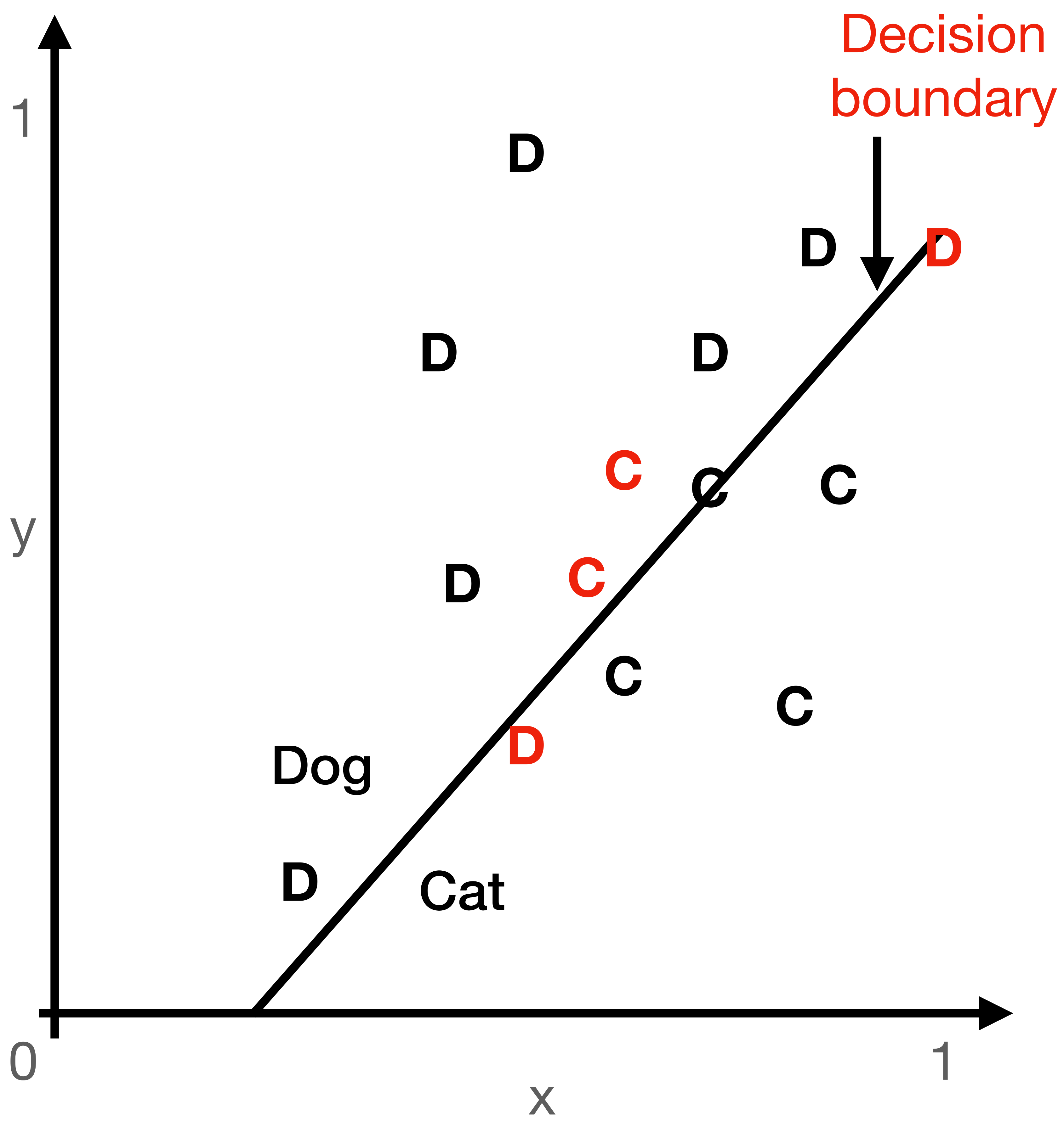
Which line/hyperplane?

"Empirical risk minimisation" idea:
pick the line that minimises
classification error on training set

Problem reduced to an optimisation problem

Linear classification

Case study: classification of cat and dog images

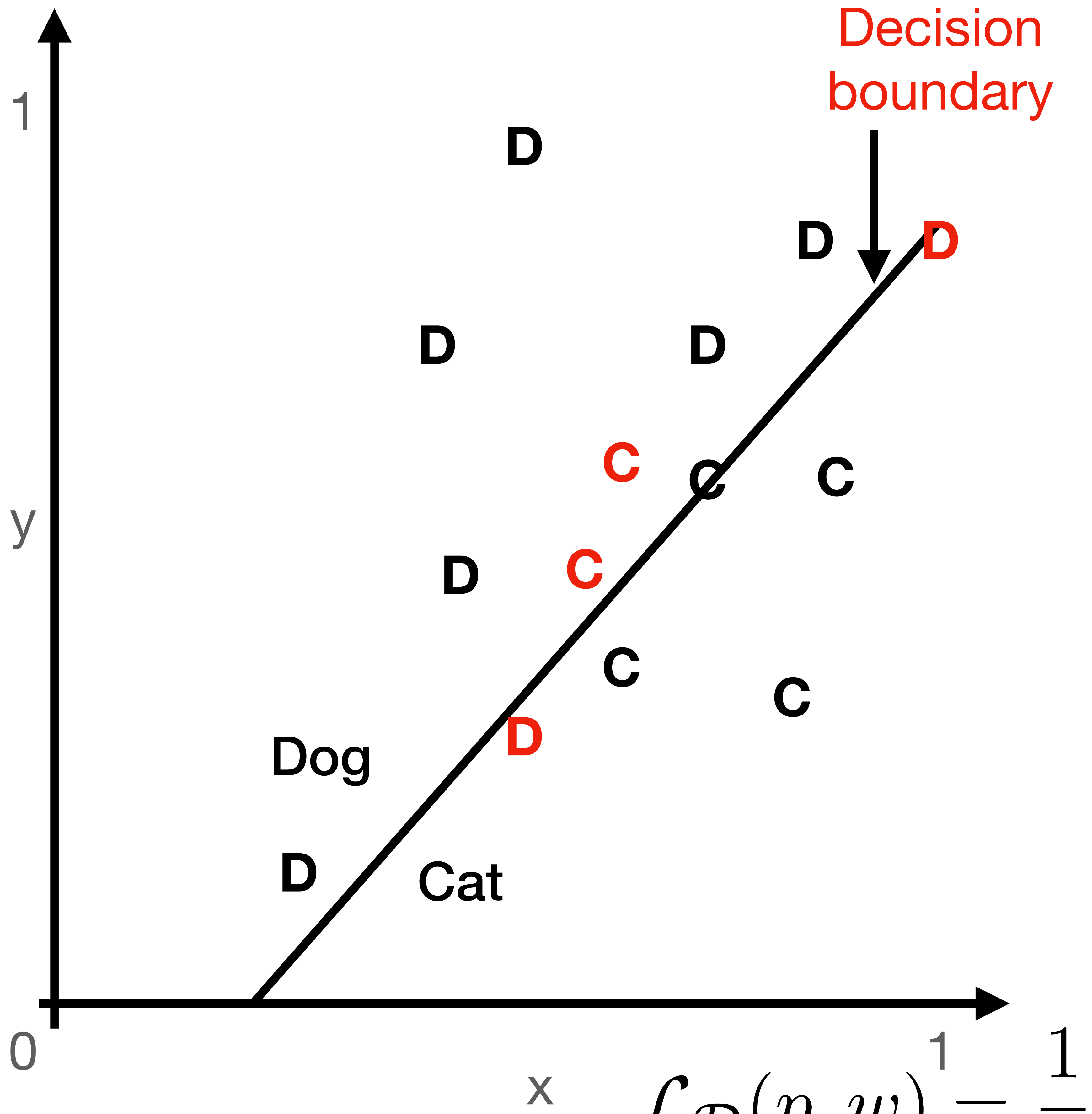


Formalisation of the optimisation problem (d=2)

$$h^* \in \arg\min_{h \in \mathcal{H}} \mathcal{L}_{\mathcal{D}}(h)$$

Linear classification

Case study: classification of cat and dog images



Formalisation of the optimisation problem (d=2)

$$h^* \in \arg\min_{h \in \mathcal{H}} \mathcal{L}_{\mathcal{D}}(h)$$

Training data

$$\mathcal{D} = (v_i, l_i)_{i=1}^n \in (\mathbb{R}^2 \times \{cat, dog\})^n$$

Parametrisation of a line

$$L_{p,w} = \{v \in \mathbb{R}^2 \mid (v-p)^T w = 0\}$$

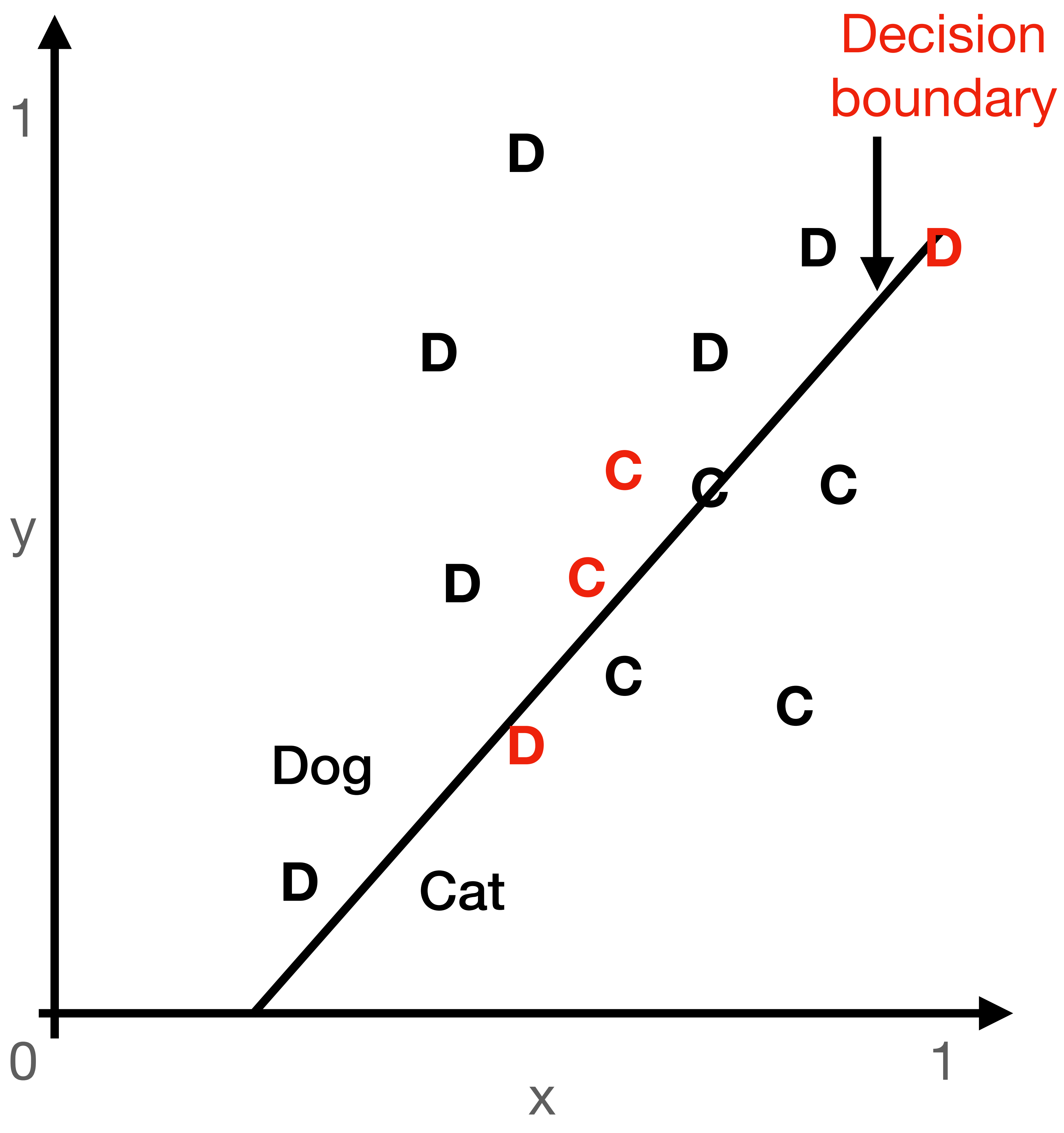
Hypothesis space

$$\mathcal{H} = \{(p, w) \in \mathbb{R}^2 \times (\mathbb{R}^2 \setminus \{(0,0)\})\}$$

Loss function

$$\mathcal{L}_{\mathcal{D}}(p, w) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{l_i=\text{cat}} \mathbf{1}_{(v_i-p)^T w > 0} + \mathbf{1}_{l_i=\text{dog}} \mathbf{1}_{(v_i-p)^T w < 0} + \frac{1}{2} \mathbf{1}_{(v_i-p)^T w = 0}$$

Case study: classification of cat and dog images



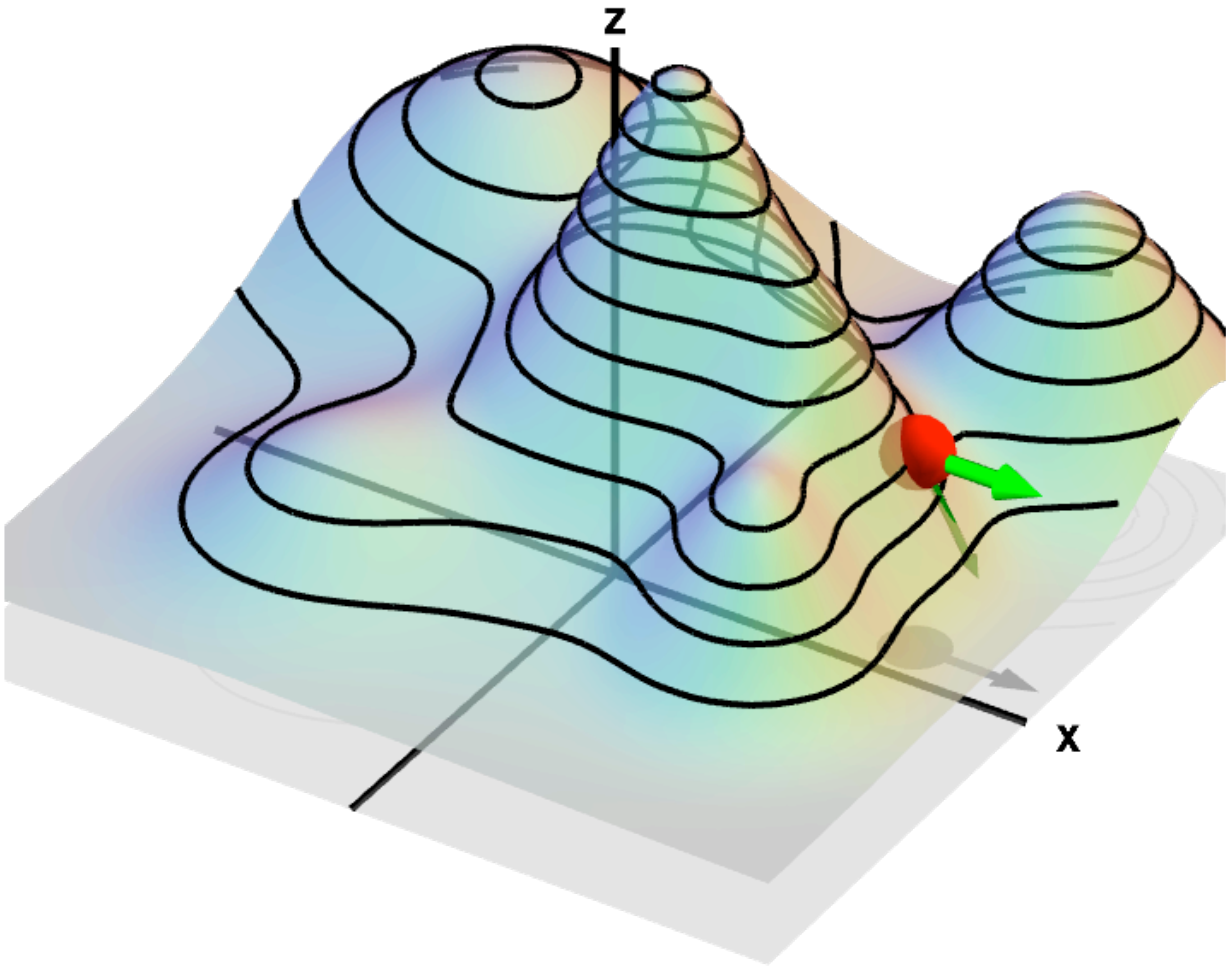
Search procedure?

- analytic?
- random?
- grid search?
- gradient descent: local search by starting from some point in parameter space and following slope (gradient) of loss function

Linear classification

Case study: classification of cat and dog images

Gradient descent



Gradient

$$\nabla f : x, y \mapsto \left(\frac{\partial f}{\partial x}(x, y), \frac{\partial f}{\partial y}(x, y)\right)$$

Figure: https://mathinsight.org/directional_derivative_gradient_introduction

Case study: classification of cat and dog images

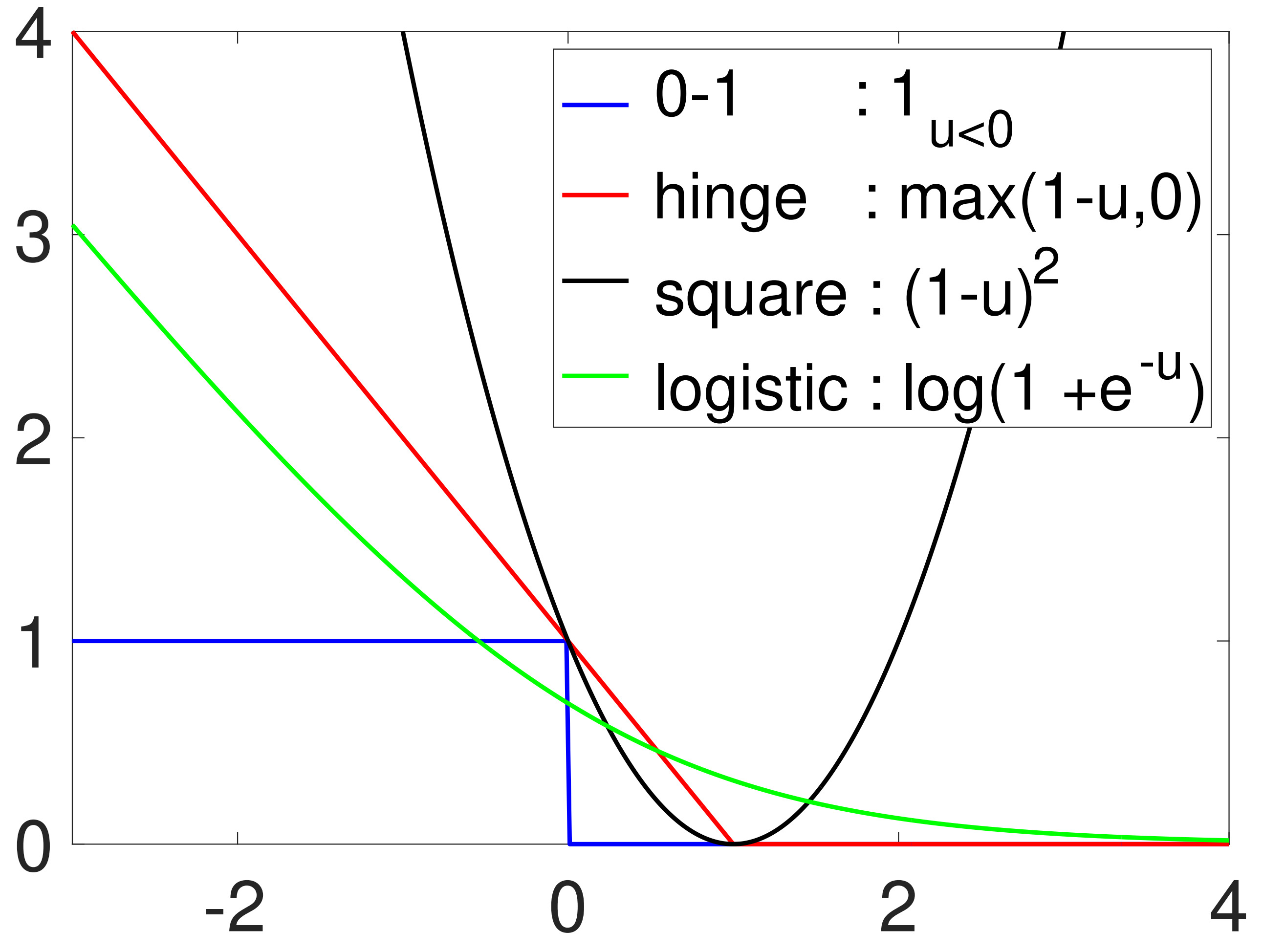
Loss function

$$\mathcal{L}_{\mathcal{D}}(p,w) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{l_i=\text{cat}}\mathbf{1}_{(v_i-p)^T w>0} + \mathbf{1}_{l_i=\text{dog}}\mathbf{1}_{(v_i-p)^T w<0} + \frac{1}{2}\mathbf{1}_{(v_i-p)^T w=0}$$

$$\Phi(x) = -\log\left(\sigma(x)\right)$$

$$\sigma(x) = \frac{1}{1+e^{-x}}$$

$$\lambda_{\mathcal{D}}(p,w) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{l_i=\text{cat}}\Phi\left(-(v_i-p)^T w\right) + \mathbf{1}_{l_i=\text{dog}}\Phi\left((v_i-p)^T w\right)$$



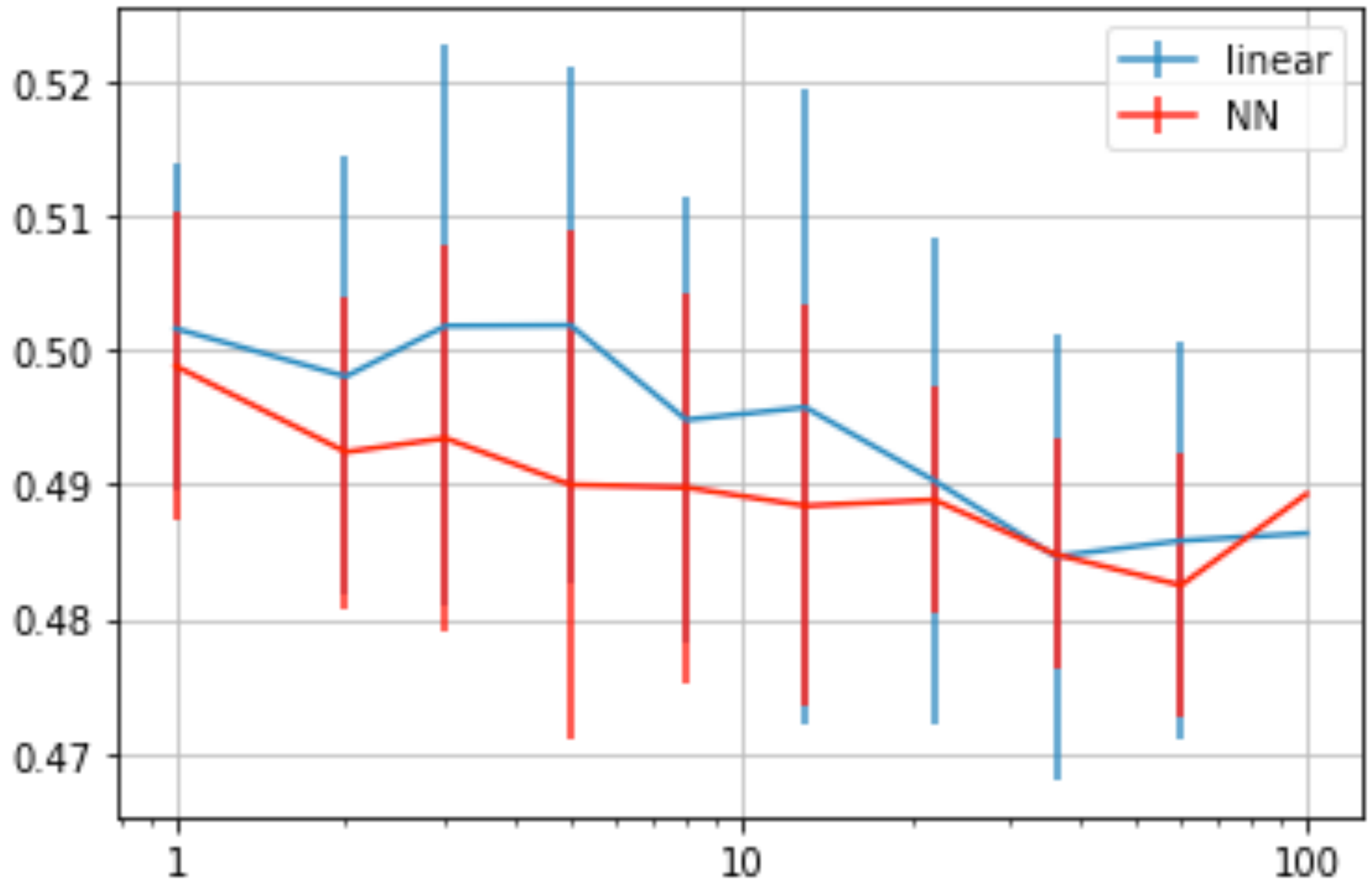
$$\nabla_w\lambda_{\mathcal{D}}(p,w) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{l_i=\text{cat}}\sigma\left((v_i-p)^T w\right)(v_i-p) - \mathbf{1}_{l_i=\text{dog}}\sigma\left(-(v_i-p)^T w\right)(v_i-p)$$

$$\nabla_p\lambda_{\mathcal{D}}(p,w) = \frac{1}{n}\sum_{i=1}^{n} -\mathbf{1}_{l_i=\text{cat}}\sigma\left((v_i-p)^T w\right)w + \mathbf{1}_{l_i=\text{dog}}\sigma\left(-(v_i-p)^T w\right)w$$

Figure: Learning Theory from First Principles, Bach (forthcoming)

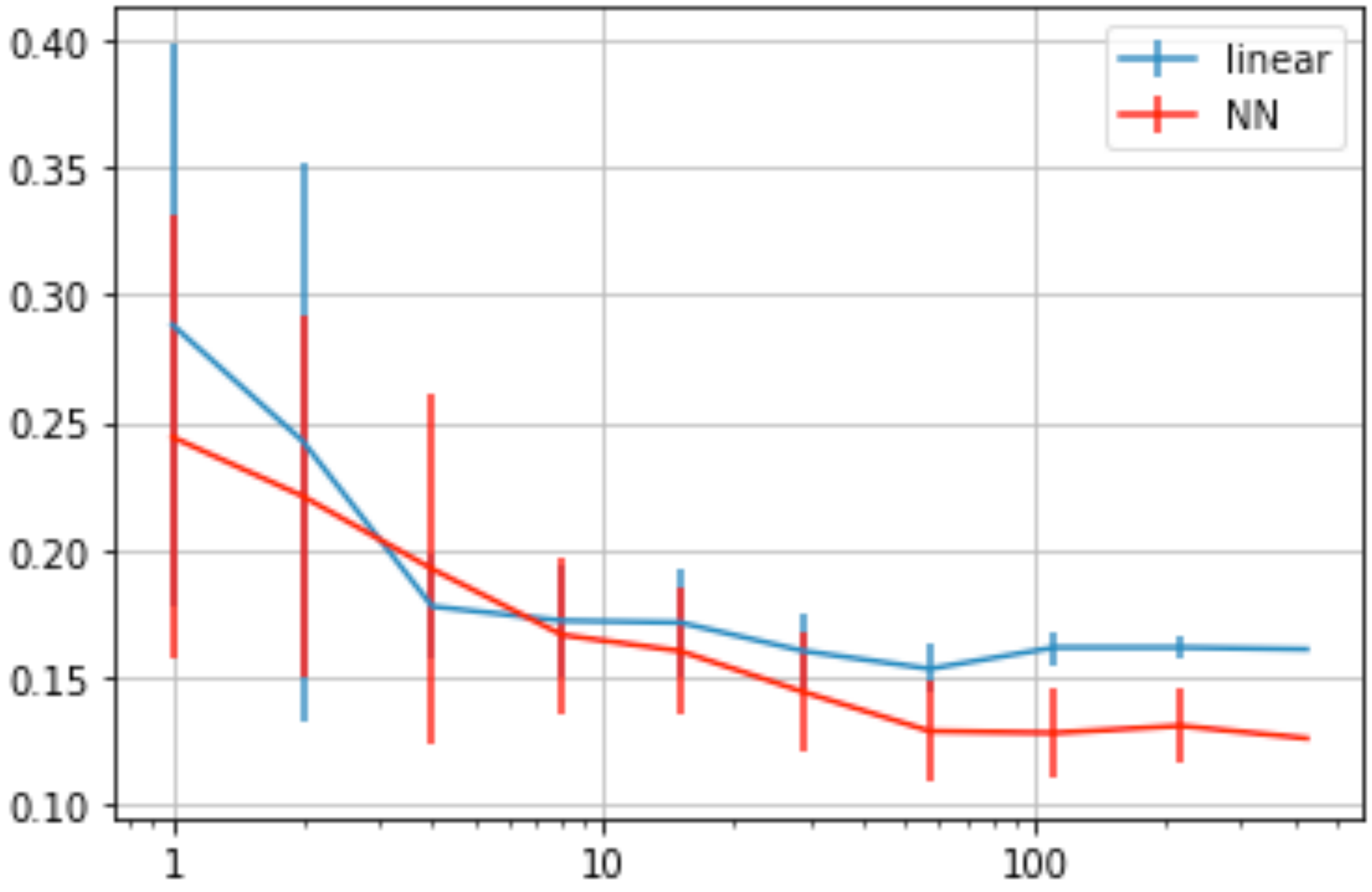
Evaluation of the classifiers

Classification error on novel images for nearest-neighbour and linear classifiers as a function of number of images of cats of dogs in the training set



Sanity check: classifier's performance on breast cancer data

https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_breast_cancer.html



How can we do better?

Three simple ideas

More computationally efficient procedure for searching good parameters with very large amount of training data:

stochastic gradient descent

Something like linear classification, but allowing decision boundaries that are more flexible than just straight lines?

neural networks

More data and more compute power

millions of images, training on GPUs, code optimisation…

Summary

- Machine learning may be relevant to cognitive scientists
  - Automatisation of time-consuming tasks; analysis of experimental data; development of models of cognitive and neural processes
- Machine learning is about designing and implementing computationally efficient statistical procedures
- Designed, implemented and evaluated two simple algorithms for (image) classification
  - Nearest neighbors
  - Linear classification using gradient descent with a fixed step-size
- Cats and dogs classification in images is hard, more computationally and statistically efficient procedures are needed

Brief discussion: ML beyond classification

We only considered a classification problem, but the scope of machine learning includes any generalisation (statistical) problem whose effective solution requires paying attention to computational costs

For example, inferring the phonemic inventory or the grammar of a language from raw (unannotated) recordings or videos (or more) of people speaking that language

Another example: combining representation learning and density estimation

Ramesh et al. (2022)

an espresso machine that makes coffee from human souls, artstation

panda mad scientist mixing sparkling chemicals, artstation

a corgi's head depicted as an explosion of a nebula

Brief discussion: ML beyond classification

The approach of seeking parameters that minimise a loss function using gradient descent is very general and often used in machine learning problems beyond classification

- Rule of thumb to understand a ML method at a basic level: is a loss (or reward) function being optimised? What is it? What are the parameters being optimised?

(Very) brief discussion: ML and cognition

Linear or nearest neighbour classifier:

- Pretty straightforward to implement using (biological) neural hardware instead of a my laptop

Next steps

- More computer science
  - MASCO students: continued in S1 (programming class)
- More statistics
  - MASCO students: continued in S1 (statistics class)
- Machine learning
  - MASCO students:
    - continued in S2 (machine learning class)
    - some elements in optional computational neuroscience class in S2
- Machine learning models of cognitive functions
  - MASCO students: likely to be S3 class on the topic in 2024

Thank you for your attention

The course material is available on the summer school amubox and at https://thomas.schatz.cogserver.net/teaching/