Elemental Learning Theory

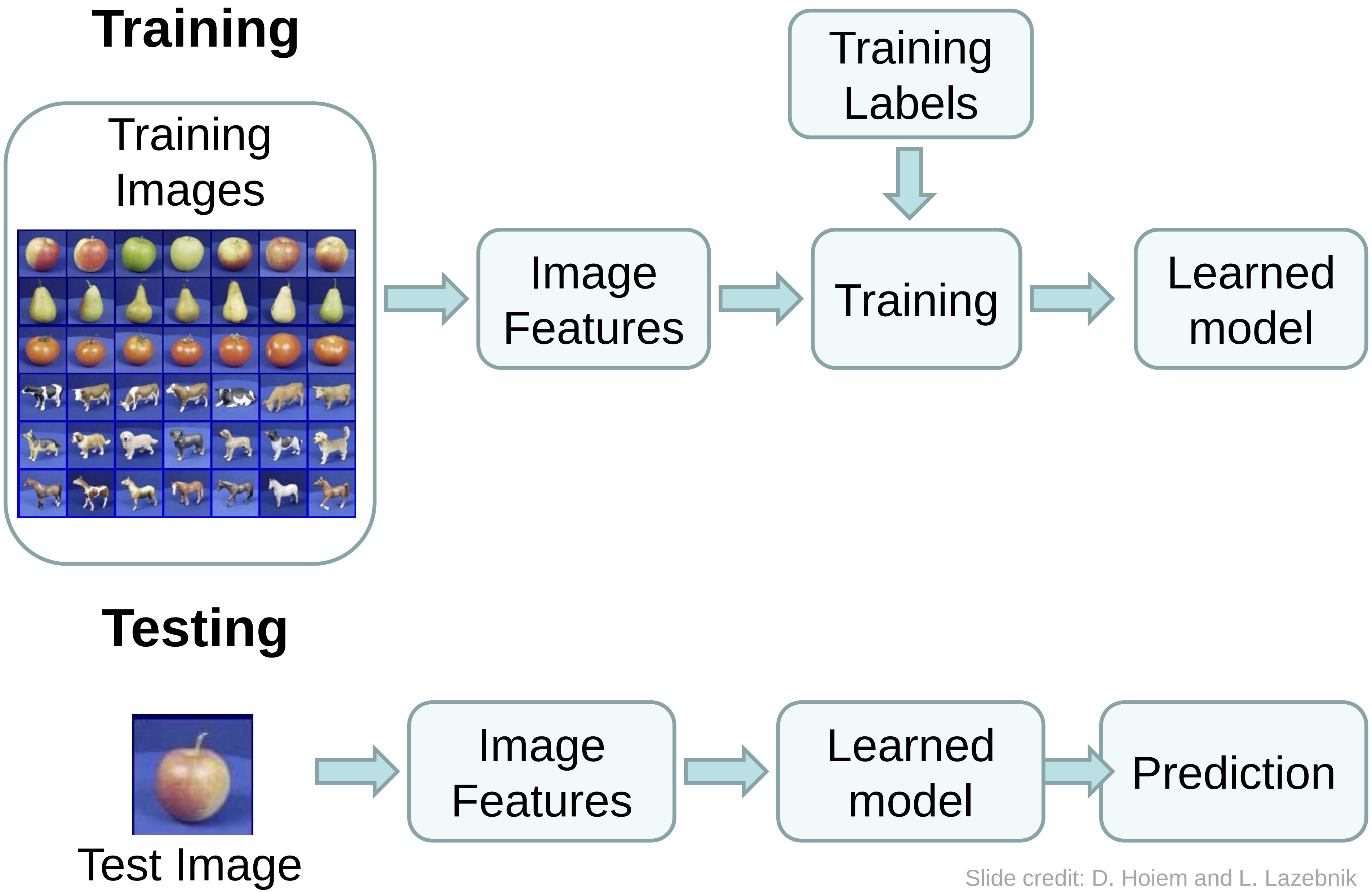
Alexander G. Ororbia II

Introduction to Machine Learning

CSCI-635

9/15/2023

Goal: Construct a Pipeline



Slide credit: D. Hoiem and L. Lazebnik

The Functional Machine Learning Framework

We are engaged in a form of function approximation

$$y = f(\mathbf{x}; \boldsymbol{\theta})$$

output

prediction function

feature vector

parameters (the “brain”)

Note that this is a parametric form of learning (as opposed to non-parametric learning)

- ***Test-time inference***: apply a prediction function to a feature representation of the image to get the desired output:

f() = "*apple*"

f() = "*tomato*"

f() = "*cow*"

Learning Algorithm *A*

- Uses training values for the target function to induce a hypothesized definition that fits these examples and hopefully generalizes to unseen examples
- In statistics, learning to approximate a continuous function is called ***regression***
  - Discrete functions = ***classification*** or categorization
- Attempts to minimize some measure of error (***loss function***) such as ***mean squared error***

ML Terminology

- Regression
  - Predict a numerical value t given some input
    - Learning algorithm has to output function $f: \mathrm{R}^n \rightarrow \mathrm{R}$
      - where n =no of input variables (or D)
- Classification
  - If t value is a label (categories): $f: \mathrm{R}^n \rightarrow \{1,..,k\}$
- Ordinal Regression
  - Discrete values, ordered categories

Spectrum of supervision

Less **More**

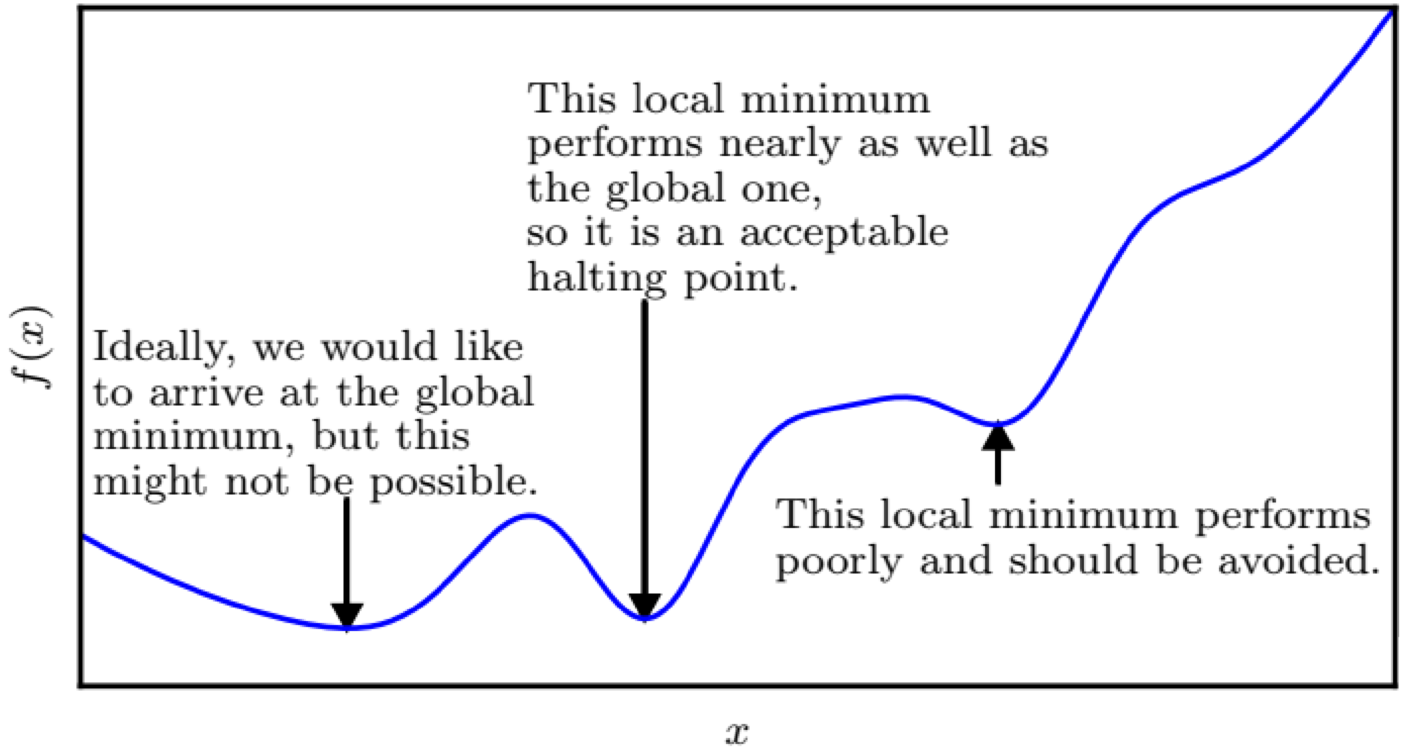
Unsupervised

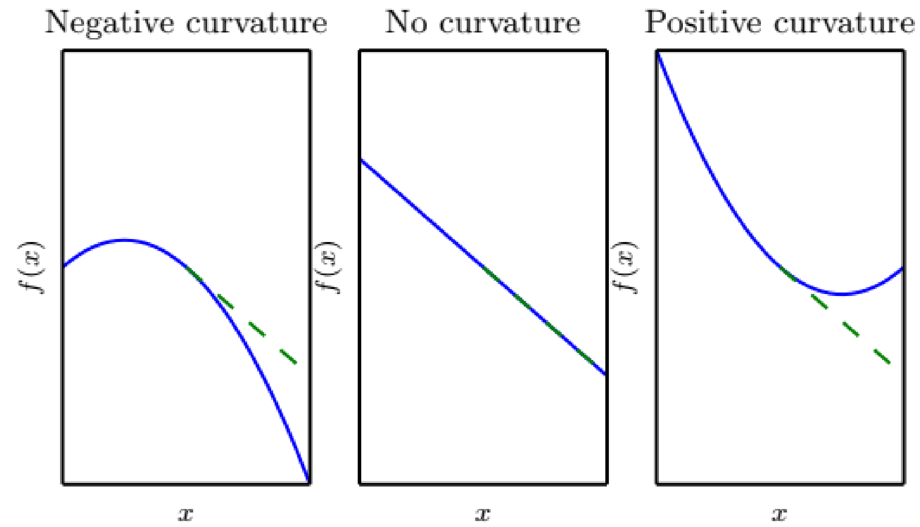
“Weakly” supervised

Fully supervised

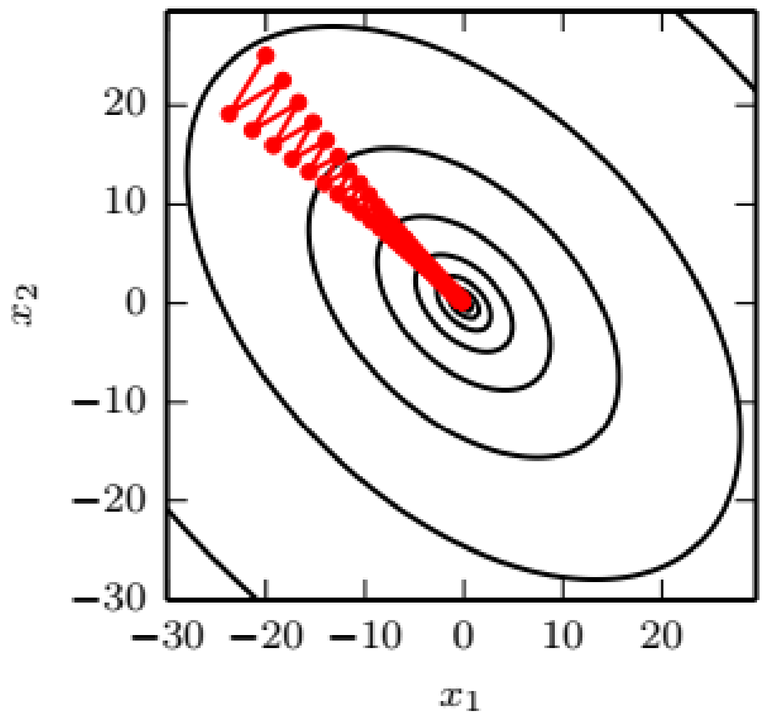
Definition depends on task

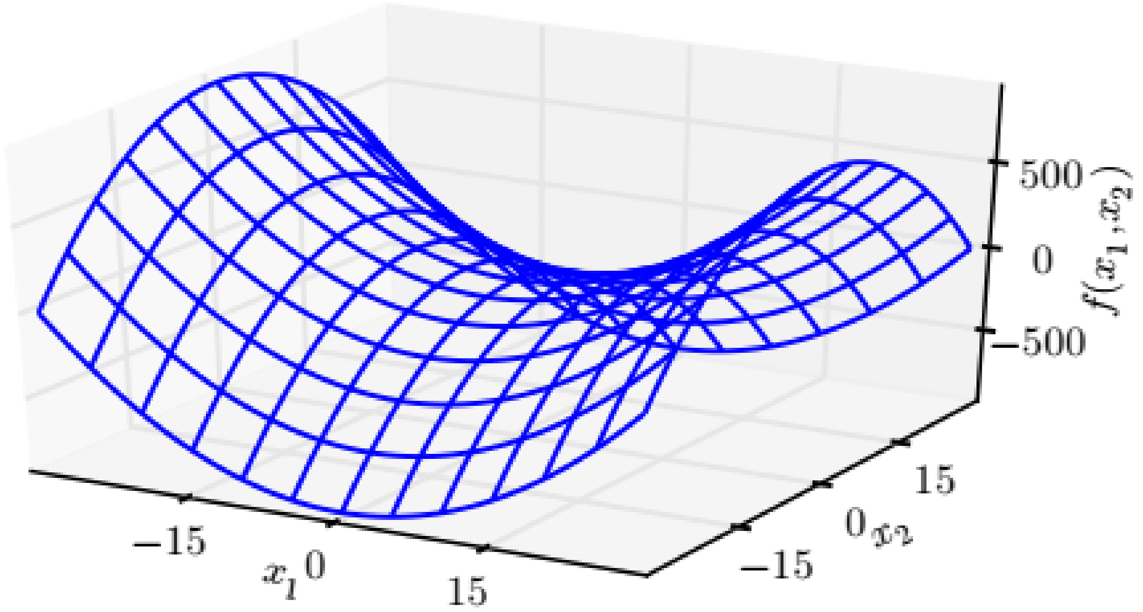
Approximate Optimization





Beware the Challenges of Optimization!





Gradient descent might not always be best choice for optimization!

Evaluation of Learning Systems

- ***Experimental***
  - Conduct controlled cross-validation experiments to compare various methods on variety of benchmark datasets
  - Gather data on performance, e.g. test accuracy, training-time, testing-time
  - Analyze differences for statistical significance (frequentist measures)
- ***Theoretical***
  - Analyze algorithms mathematically and prove theorems about their:
    - Computational complexity
    - Ability to fit training data
    - Sample complexity (number of training examples needed to learn an accurate function)

Usefulness of statistical learning theory

- Provides intellectual justification that machine learning algorithms can work
- But rarely used in practice with deep learning
- This is because:
  - The bounds are loose
  - Also difficult to determine capacity of deep learning algorithms

Generalization

Resting on the IID assumption!

- Challenge of ML is generalization
  - Perform well on previously unseen outputs
- ML training algorithm reduces training error, which is a task of optimization
- What differentiates ML from optimization is that we want test error (generalization error) to be low as well
- Generalization error definition
  - Expected error on a new input
  - Expectation wrt distribution encountered in practice

Generalization

Training set (labels known)

Test set (labels unknown)

- How well does a learned model generalize from the data it was trained on to a new test set?

Slide credit: L. Lazebnik

Factors of Generalization

Bias-Variance Tradeoff

- Components of generalization error
  - **Bias:** how much the average model over all training sets differ from the true model?
    - Error due to inaccurate assumptions/simplifications made by the model
  - **Variance:** how much models estimated from different training sets differ from each other
  - **(Irreducible) Noise**
- **Underfitting:** model is too "simple" to represent all the relevant class characteristics
  - High bias and low variance
  - High training error and high test error

You did bad!

- **Overfitting:** model is too "complex" and fits irrelevant characteristics (noise) in the data
  - Low bias and high variance
  - Low training error and high test error

You tried too hard!

The Bias-Variance Trade-off

$$E(MSE) = noise^2 + bias^2 + variance$$

Unavoidable error

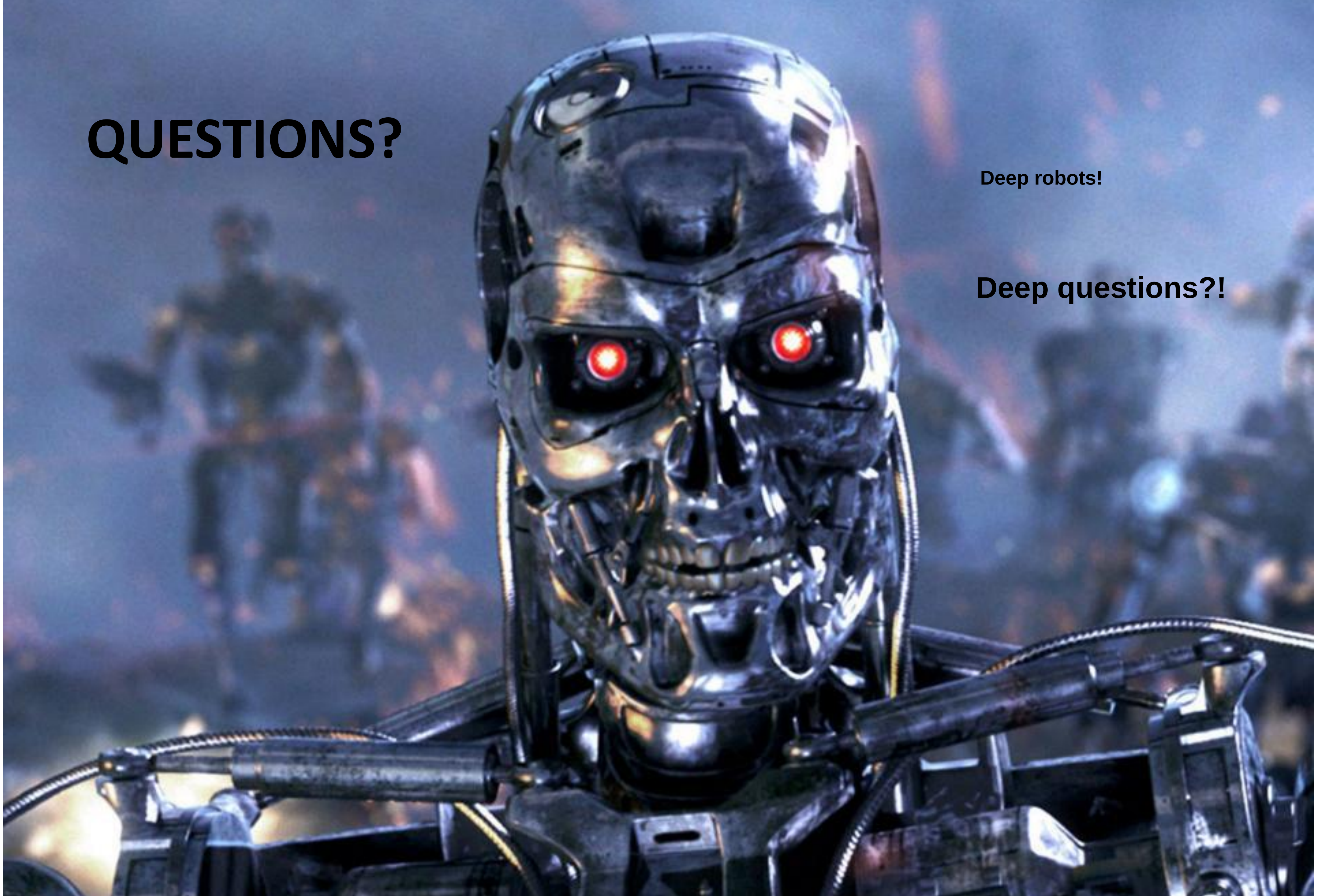
Error due to incorrect assumptions

Error due to variance of training samples

See the following for explanations of bias-variance:

- http://www.inf.ed.ac.uk/teaching/courses/mlsc/Notes/Lecture4/BiasVariance.pdf
- Bishop's "Neural Networks" book

QUESTIONS?



Deep robots!

Deep questions?!