HTF: 1, 2
DHS: 1
RN: 13

Introduction to Machine Learning

(Cmput 466 / 551)

R Greiner
Department of Computing Science
University of Alberta

Skip

Questions

- What is (machine) learning ?
- Is learning really possible?
  Can an algorithm really predict the future?
- Why learn?
- Is learning $\subset^{?}$ statistics ? … or other discipline?

What is machine learning (ML)?

- Merriam-Webster:
  "learning" ≡ "To gain knowledge or understanding of
  or skill in by study, instruction, or experience"

- Machine Learning = Learning by machines (computers)

- ML techniques:
  algorithms that enable the machines to improve its performance at some task through experience
- Tasks: recognition, diagnosis, prediction, data mining, segmentation, planning, robot control, …

What is Machine Learning (ML)?

- “Machine learning is programming computers to optimize a performance criterion using example data or past experience.”
  - Alpaydin
- “The field of machine learning is concerned with the question of how to construct computer programs that automatically improve with experience.”
  - Mitchell
- “…the subfield of AI concerned with programs that learn from experience.”
  - Russell & Norvig
- “The goal of machine learning is to develop methods that can automatically detect patterns in data, and then to use the uncovered patterns to predict future data or other outcomes of interest.”
  - Murphy

Typical Taxonomy for ML

- Learning to **Predict**
  - Supervised learning
  - Semi-supervised learning
- Learning to **Model**
  - Unsupervised learning
  - Clustering
  - Generative Models
- Learning to **Control**
  - Reinforcement learning

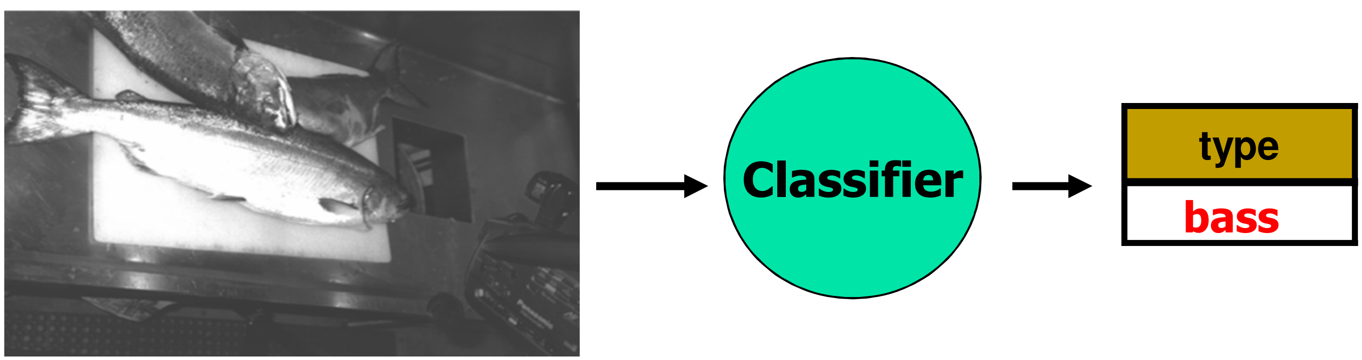
Fish Classifier

Sort Fish

into Species

- **Sea bass**
- **Salmon**

using optical sensing



Problem Analysis

- Extract *features* from sample images:
  - Length
  - Width
  - Average pixel brightness
  - Number and shape of fins
  - Position of mouth
  - …

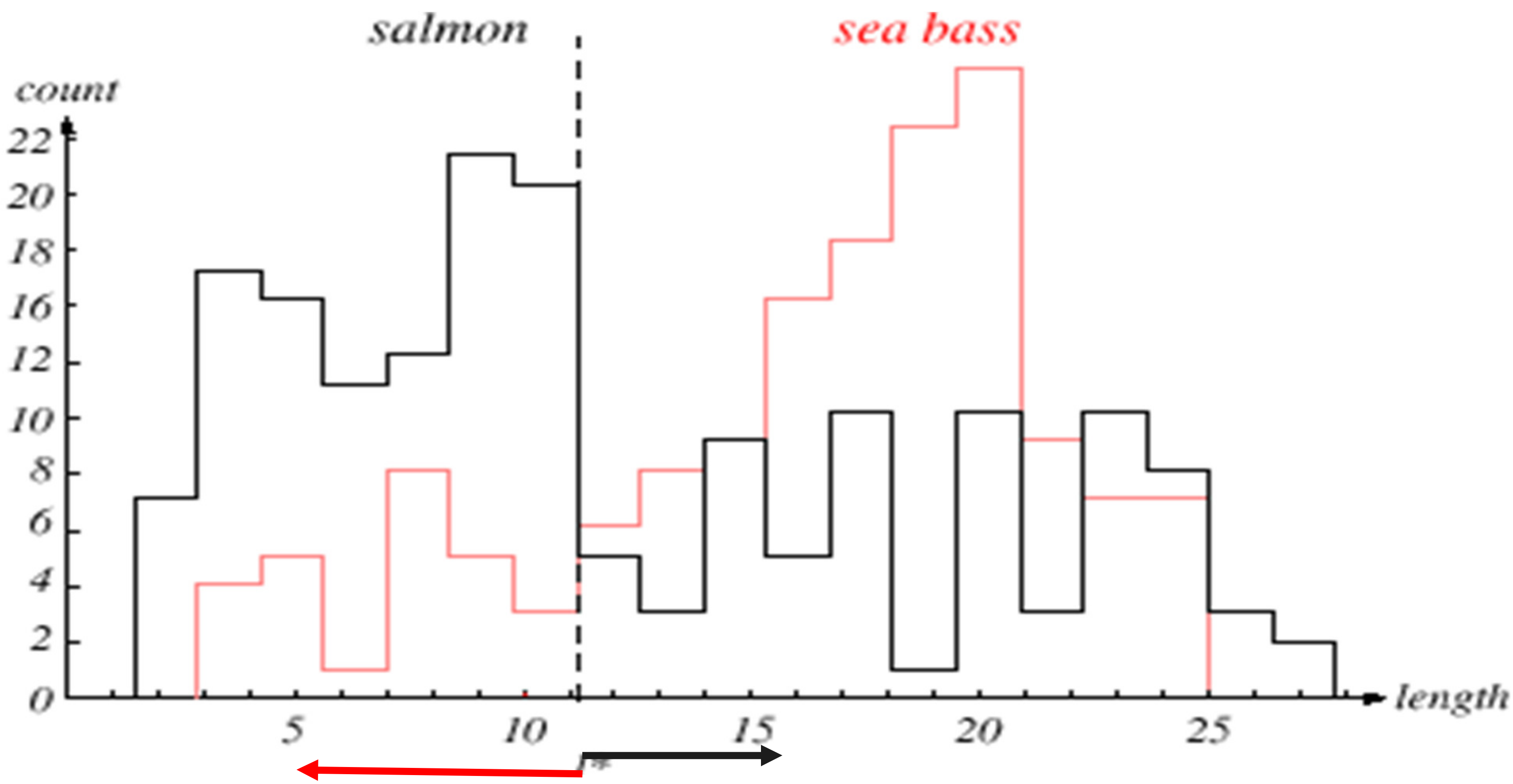
[L=50, W=10, PB=2.8, #fins=4, MP=(5,53), …]

Length	Wtdth	Pixel Bright	…	Light
50	10	2.8	…	Pale

Classifier

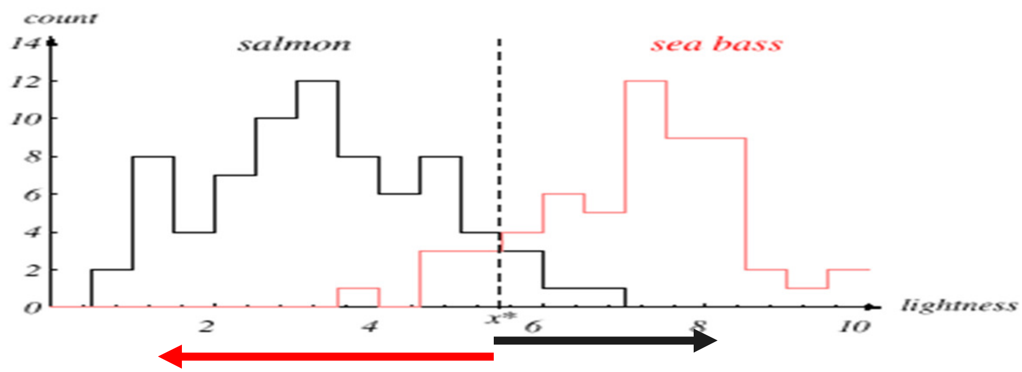
type
bass

Use "Length"?



- Problematic… many incorrect classifications

Where to place boundary?

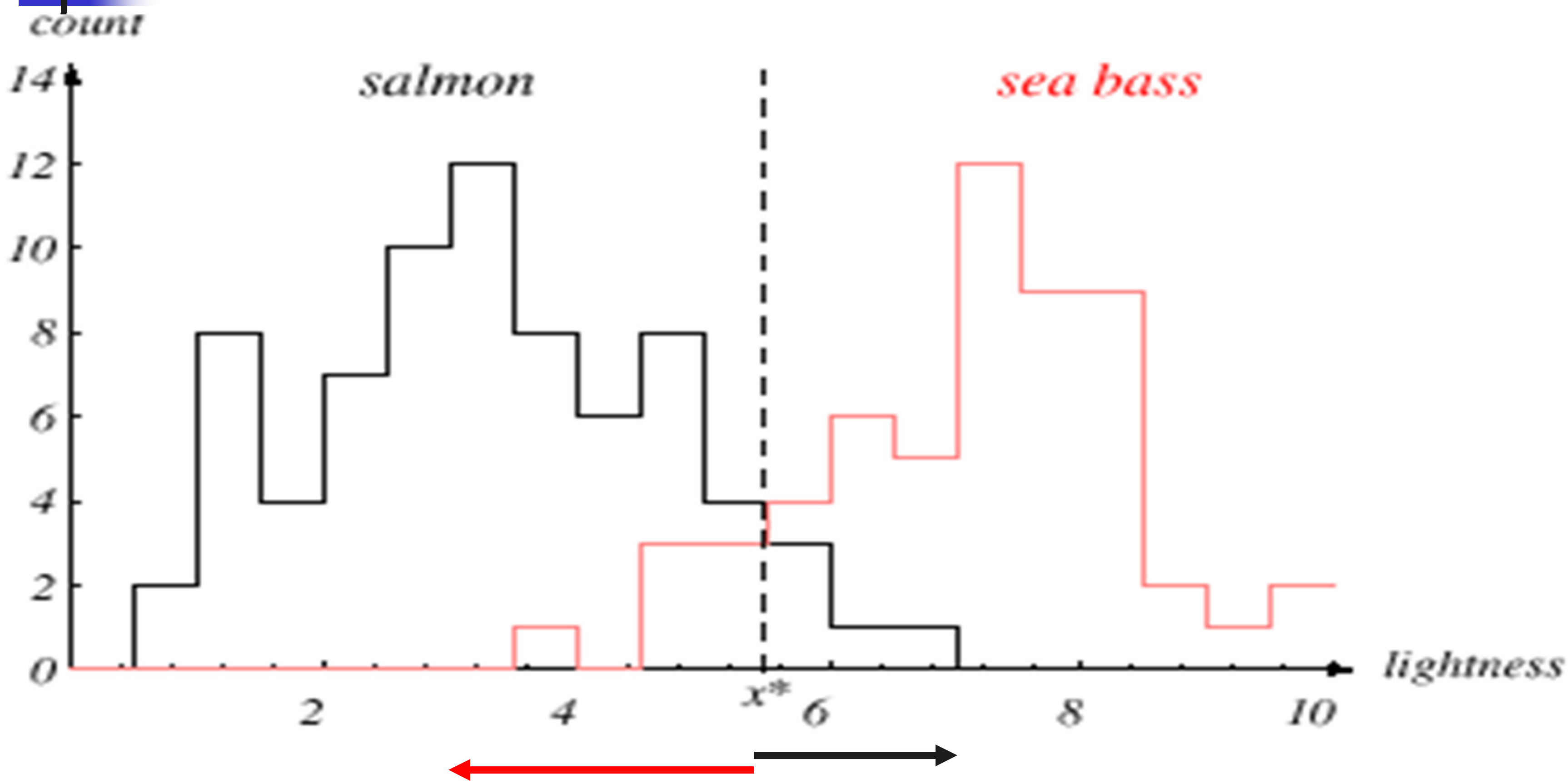


- *Salmon Region* intersects *SeaBass Region*

  $\Rightarrow$ So no "boundary" is perfect
  - *Smaller* boundary $\Rightarrow$ fewer SeaBass classified as Salmon
  - *Larger* boundary $\Rightarrow$ fewer Salmon classified as SeaBass
- Which is best… depends on misclassification costs

Task of decision theory

Use “Lightness”?



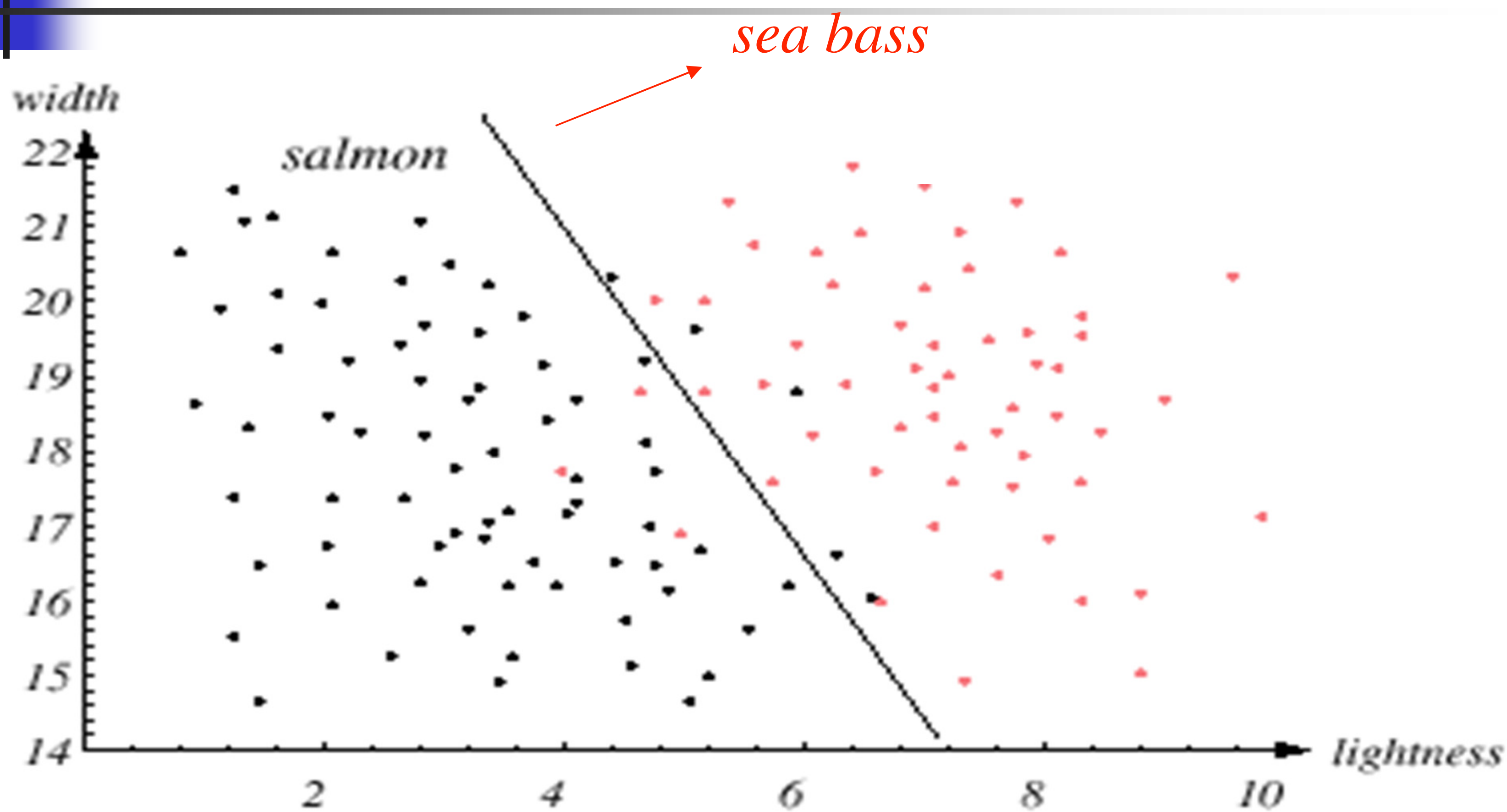
- Better… fewer incorrect classifications
- Still not perfect

Why not 2 features?

- Use *lightness* and *width* of fish

$$\text{Fish} \longrightarrow x^T = [\underbrace{x_1}_{\textbf{Lightness}}, \underbrace{x_2}_{\textbf{Width}}]$$

Use Simple Line ?



- Much better…
  very few incorrect classifications !

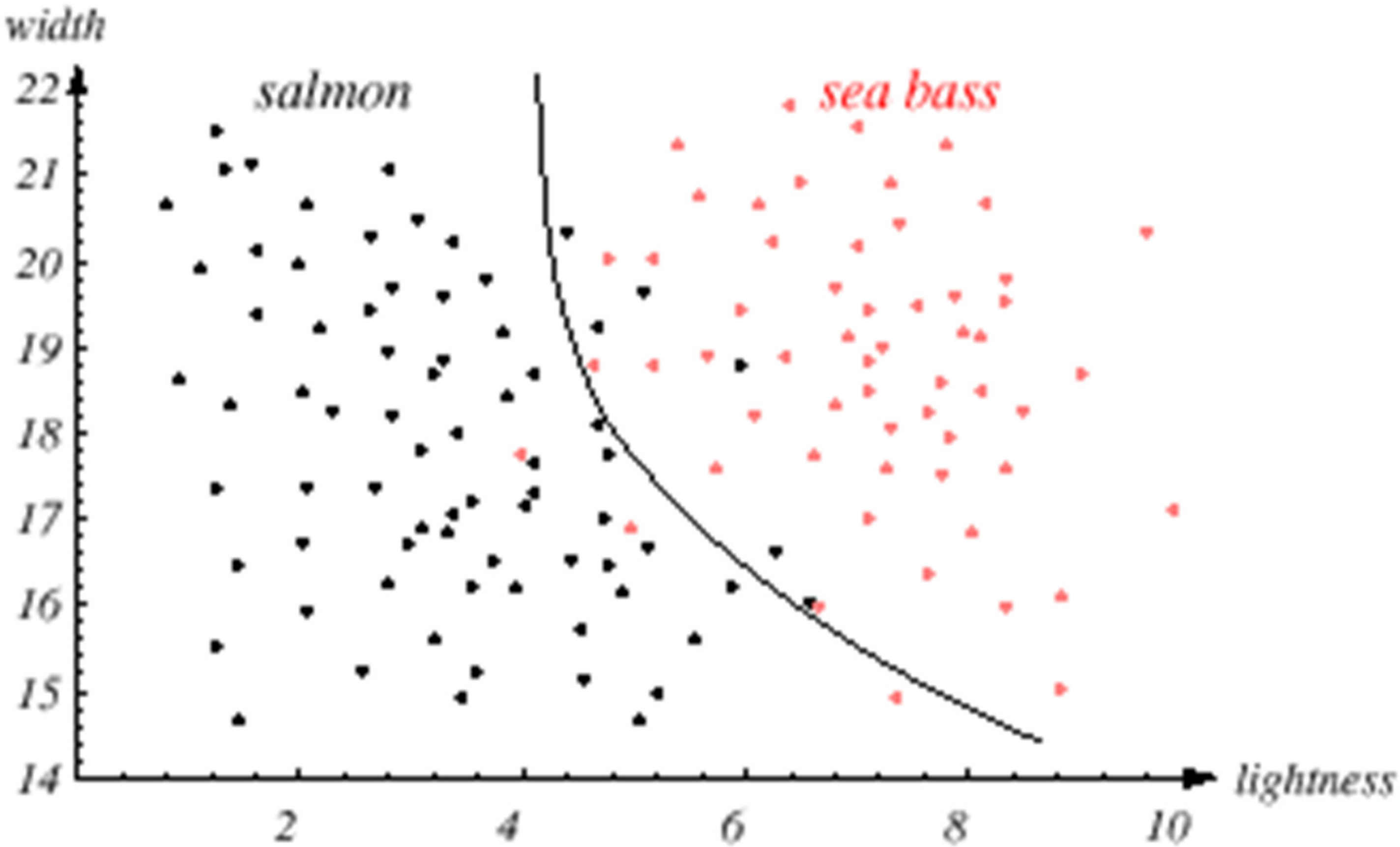
How to produce Better Classifier?

- Perhaps add other features?
  - Best: correlated with labels

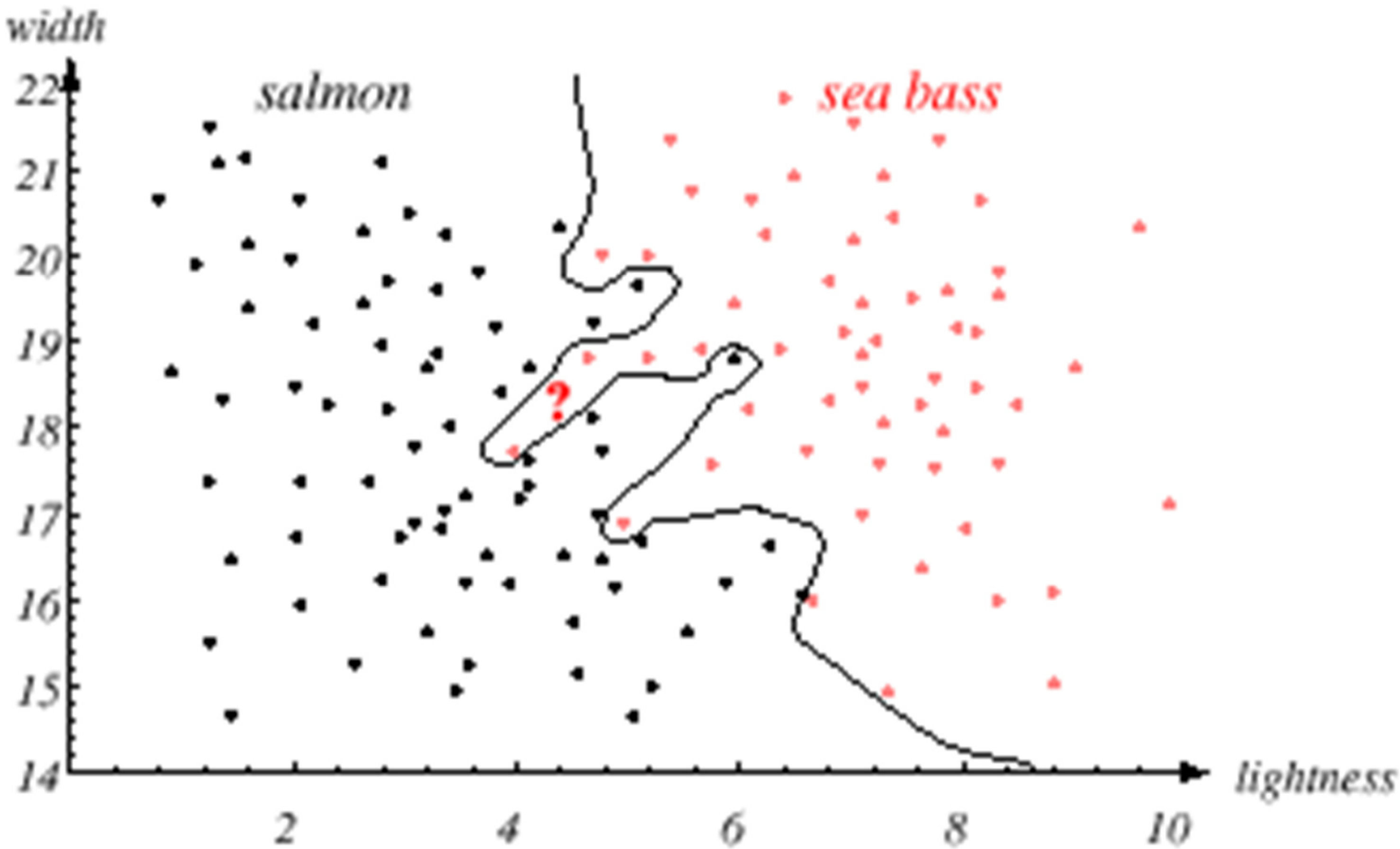
    … but not correlated with current features
  - Warning: "noisy features" will ***reduce*** performance

- Best decision boundary ≡
  one that provides optimal performance
  - Not necessarily STRAIGHT LINE
  - For example …

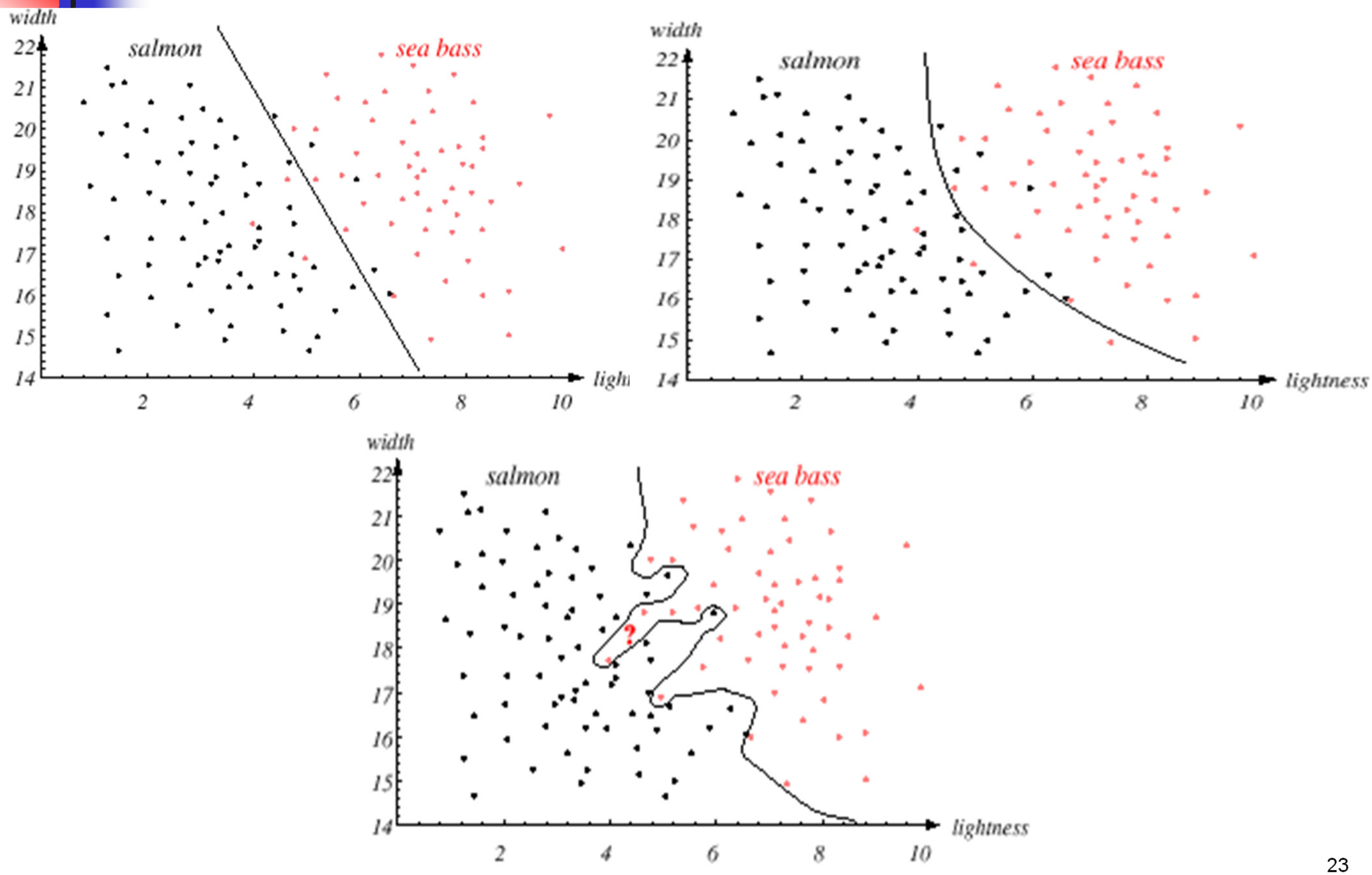
Simple (non-line) Boundary



"Optimal Performance" ??



Comparison… wrt NOVEL Fish



Objective: Handle Novel Data

Goal:

- Optimal performance on *NOVEL* data
- Performance on TRAINING DATA

  $\neq$

  Performance on NOVEL data

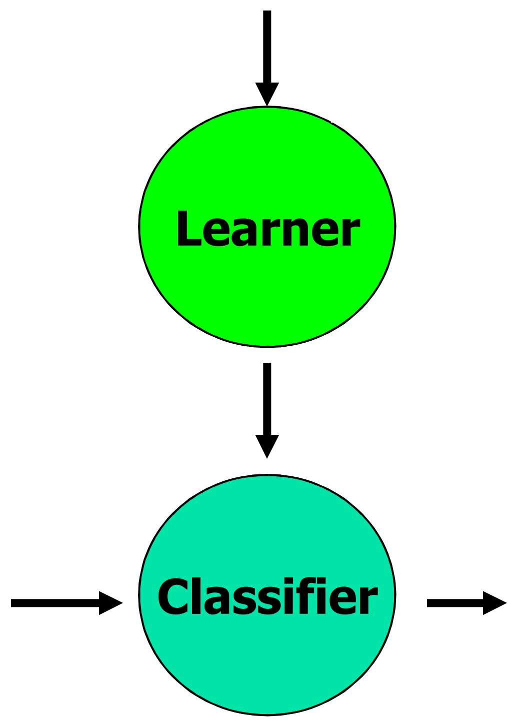
Issue of generalization!

Machine Learning Steps

- Feature extraction
  - Discriminative features
  - Want useful features
    - Here: INVARIANT wrt translation, rotation, scale
- Classification
  - Using feature vector (provided by feature extractor) to assign given object to a *category*
- Post Processing
  - Exploit context (information not in the target pattern itself) to improve performance

Training a Classifier

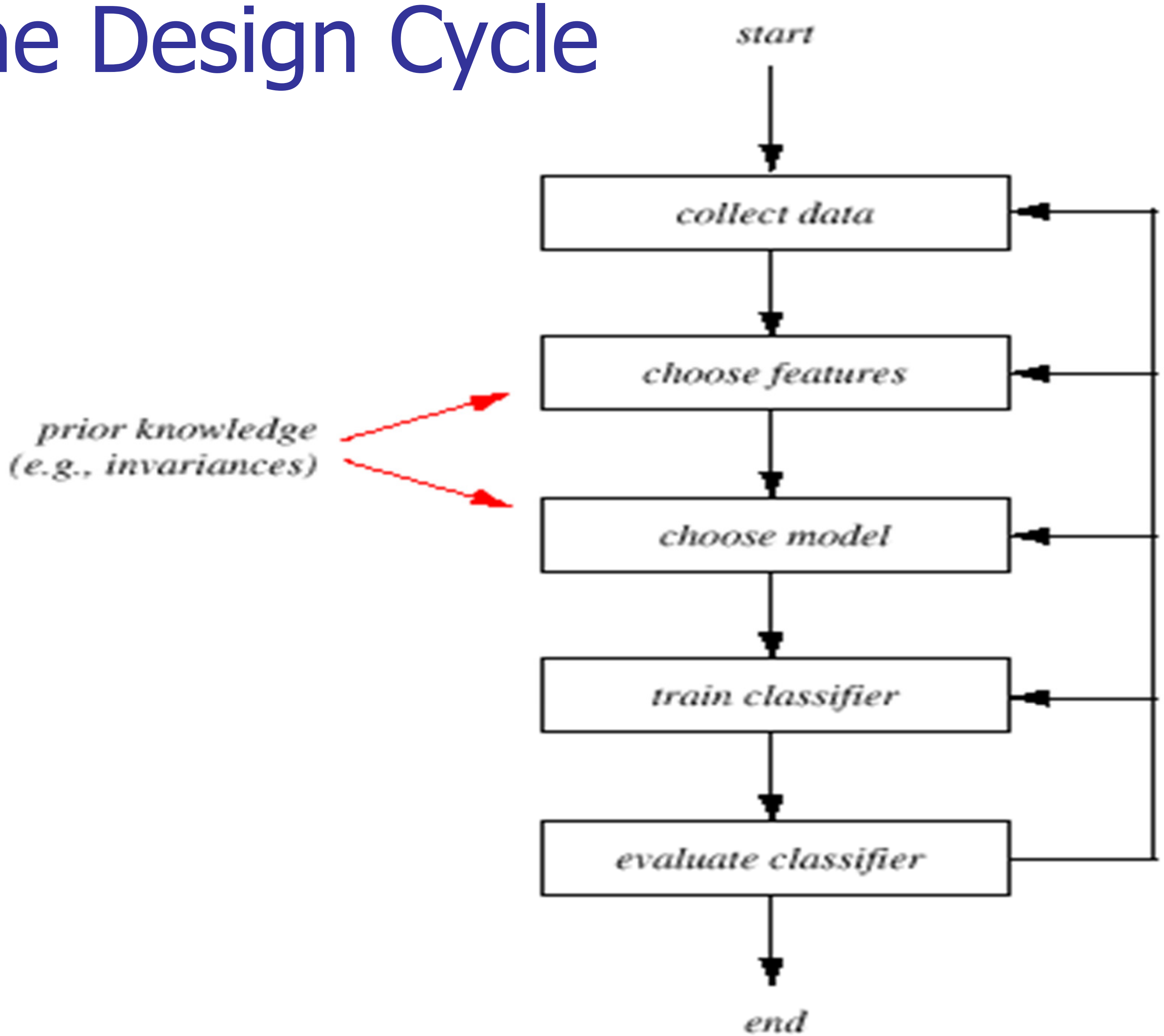
Width	Size.	Eyes	…	Light	type
35	95	Y	…	Pale	bass
22	110	N	…	Clear	salmon
:	:			:	:
10	87	N	…	Pale	bass



Width	Size	Eyes	…	Light
32	90	N	…	Pale

type
bass

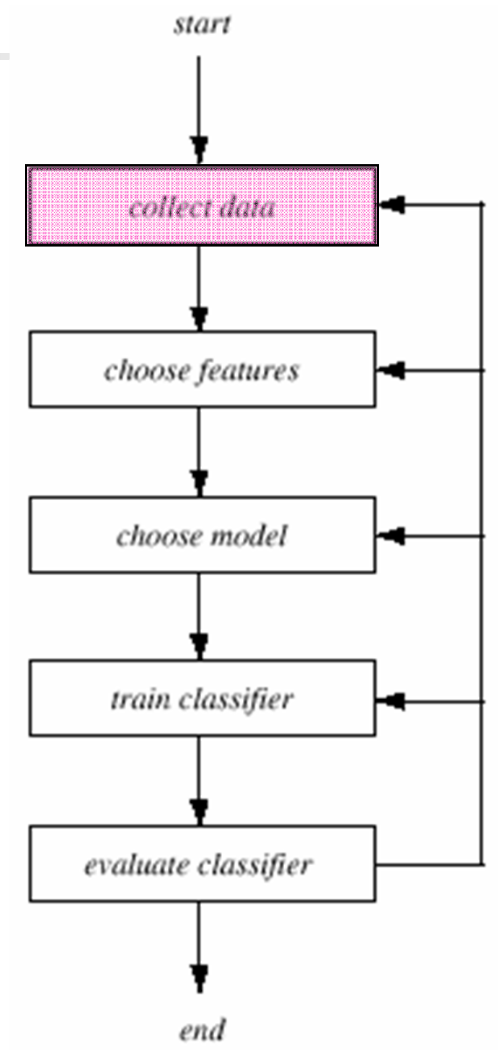
The Design Cycle



… Computational Complexity

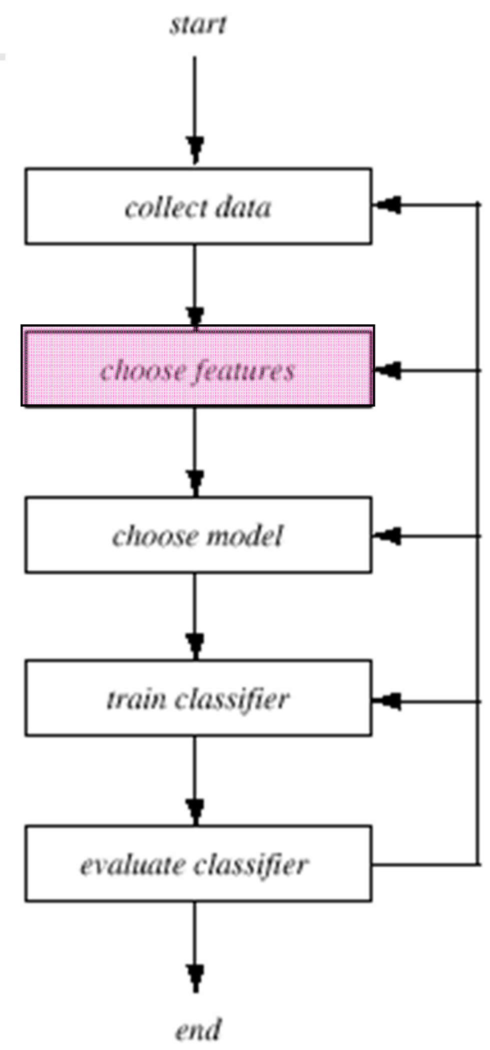
Data Collection

- Need set of (labeled) examples for training the system … and evaluation
- How much data?
  - sufficiently large # of instances
  - representative



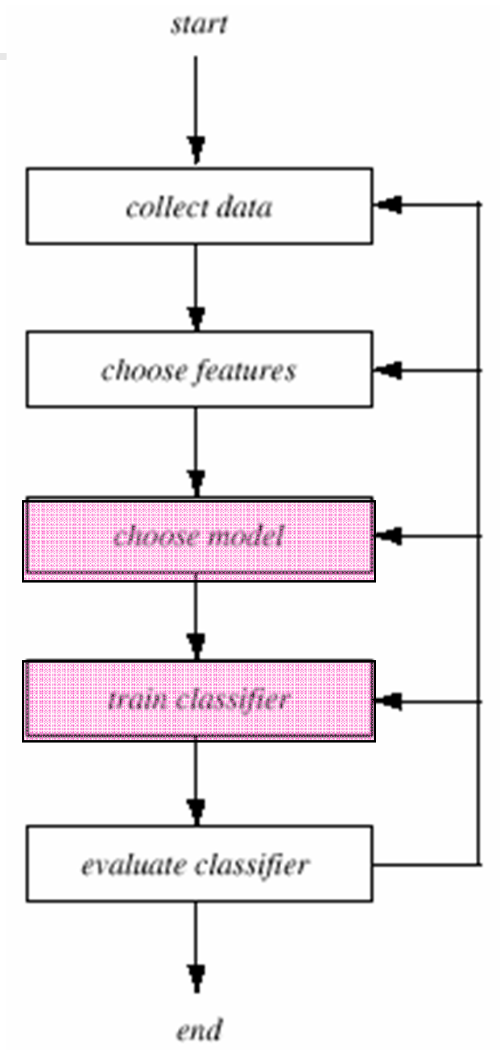
Which Features?

- Depends on characteristics of problem domain
- Ideally…
  - Relevant to label !
  - Simple to extract
  - Insensitive to noise
  - … invariant to irrelevant transformation



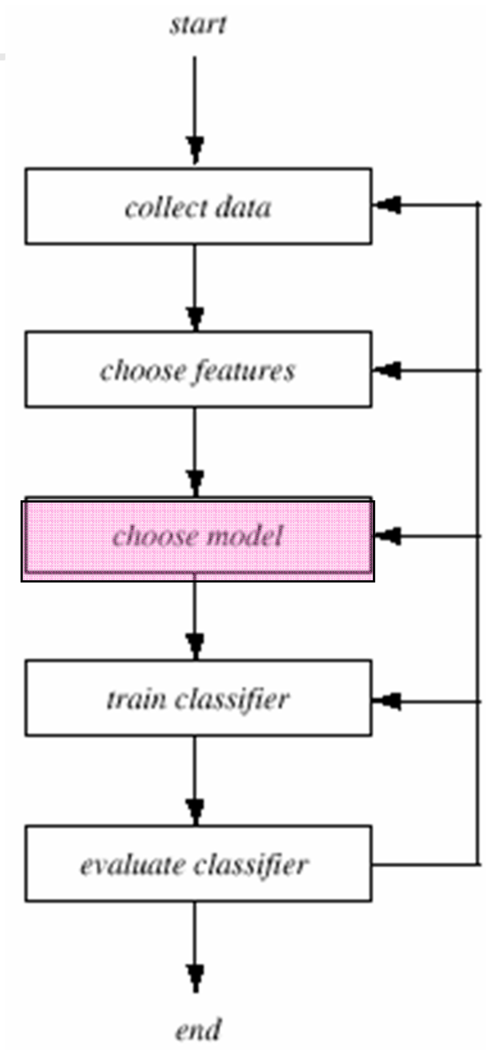
Training

- Use data to obtain good classifier
  - identify best model (classifier)
  - determine appropriate parameters
- Many procedures for training classifiers

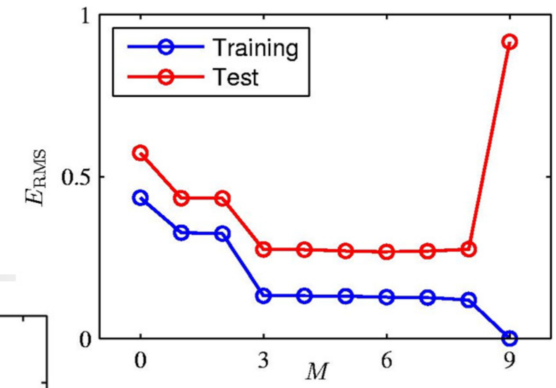


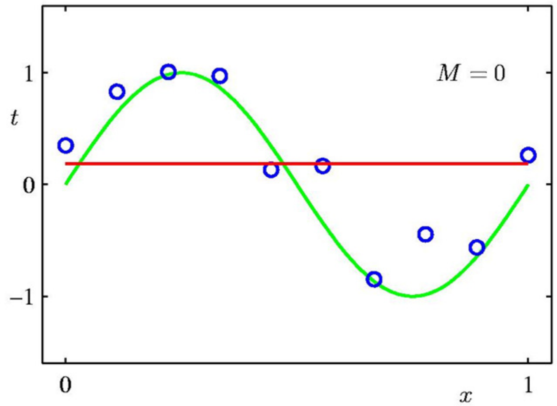
Which Model?

- Based on known properties?
- Try one from simple class
  - Degree-1 Poly
  - Gaussian
  - Conjunctions (1-DNF)
- If not good…
  try one from **yet** more complex class of models
  - Degree-2 Poly
  - Mixture of 2 Gaussians
  - 2-DNF

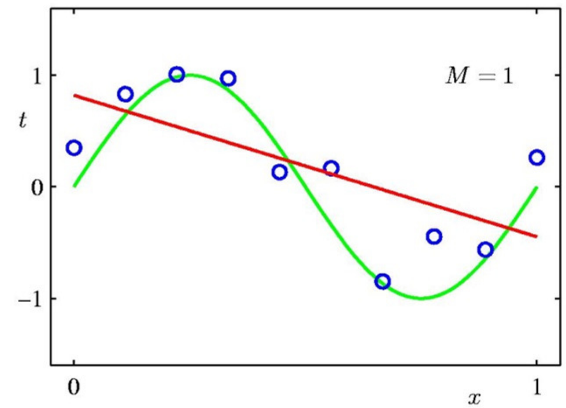


Which Model??

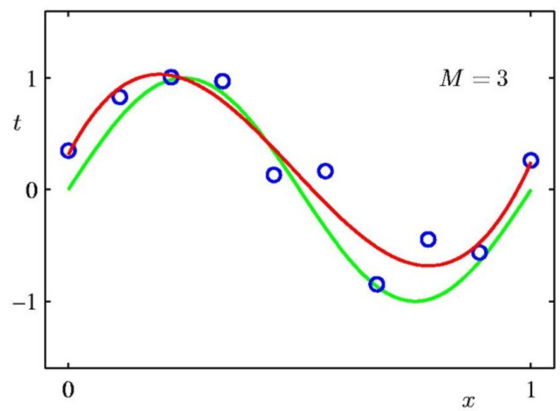




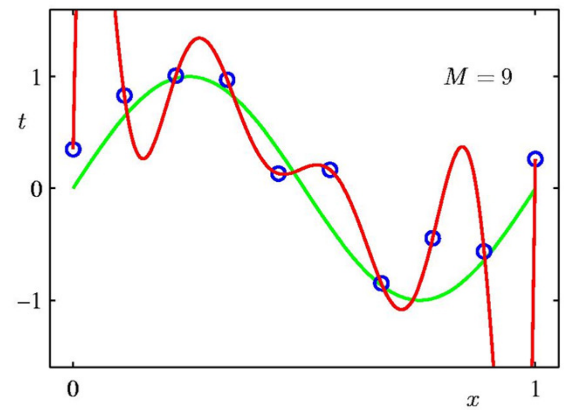
Constant (0)



Linear (1)



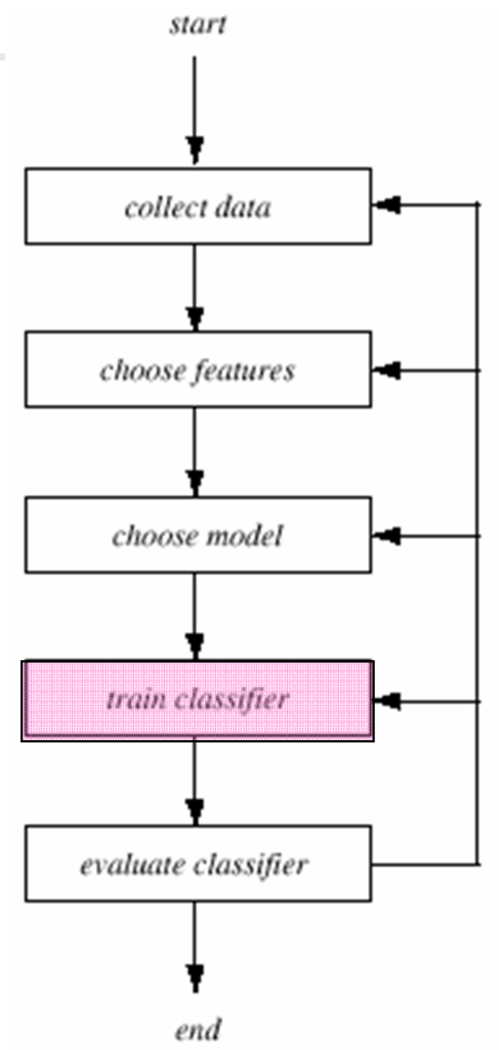
Cubic (3)



9^{th} degree

Training

- Often many procedures for training classifiers (for given model)
- Exact / Approximate
- Efficiency
- Robustness
- ...

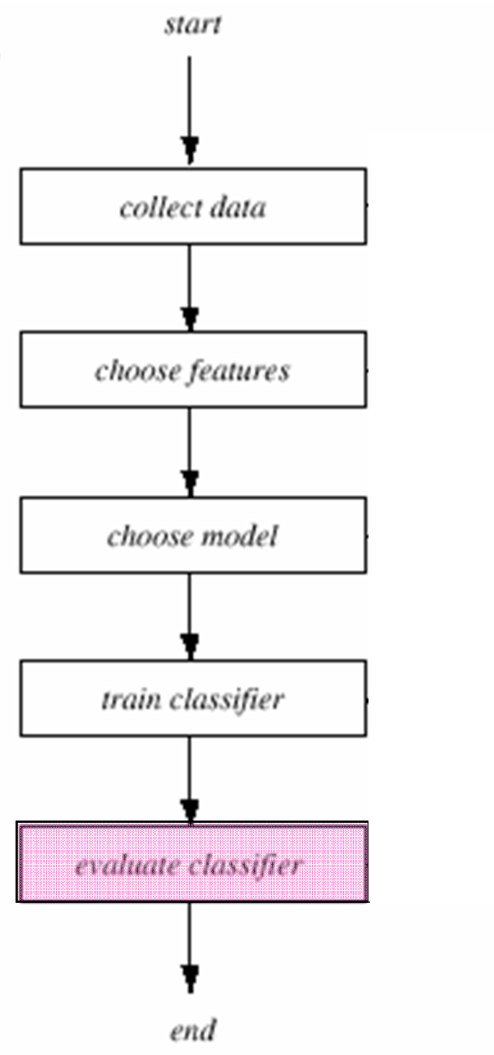


Evaluation

- Measure error rate

  ≈ performance

- May suggest switching…
  - from one set of features to another one
  - from one model to another



Issues wrt Supervised Learning

- Computational Complexity
  - How algorithm scales as function of number of features, patterns or categories?
  - Trade-off between computational ease and performance?

- Complex INPUT
- Complex OUTPUT

Complex Input: Webpage classification (text)

Company home page

vs

Personal home page

vs

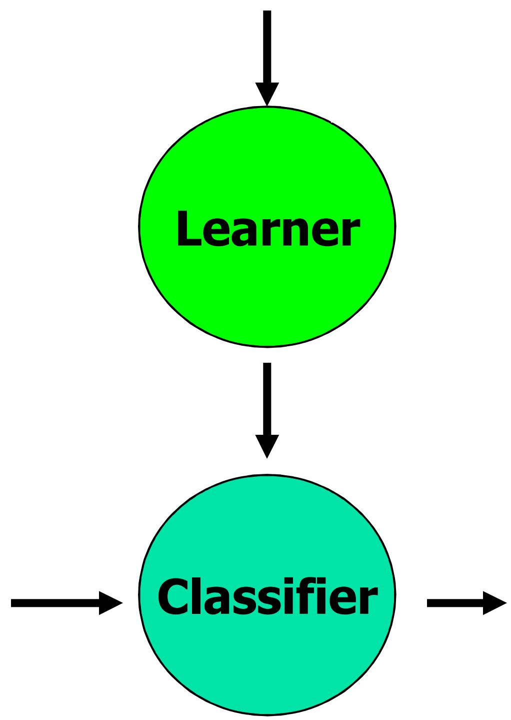
University home page

vs

…

Training a Classifier

Width	Size	Sore Throat	…	Light	type
35	95	Y	…	Pale	bass
22	110	N	…	Clear	salmon
:	:			:	:
10	87	N	…	Pale	bass

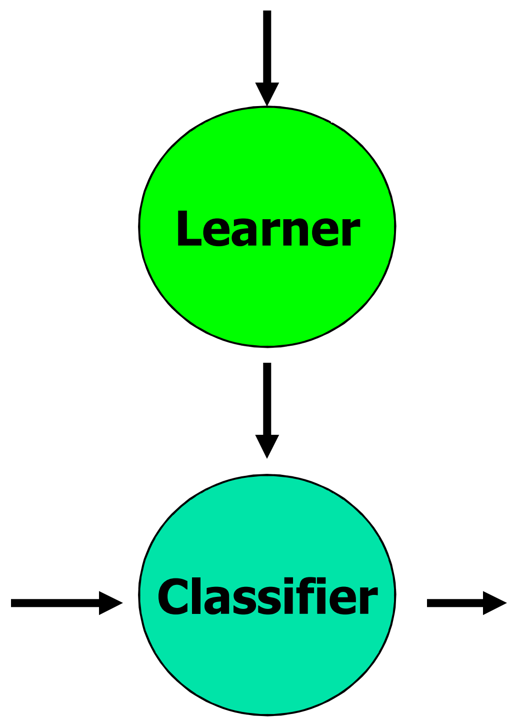


Width	Size	Sore Throat	…	Light
32	90	N	…	Pale

type
bass

Training a Regressor

Width	Size	Sore Throat	…	Light	size
35	95	Y	…	Pale	22
22	110	N	…	Clear	18
:	:			:	:
10	87	N	…	Pale	33

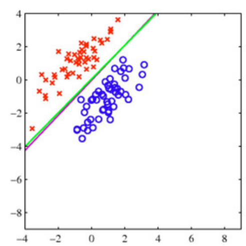


Width	Size	Sore Throat	…	Light
32	90	N	…	Pale

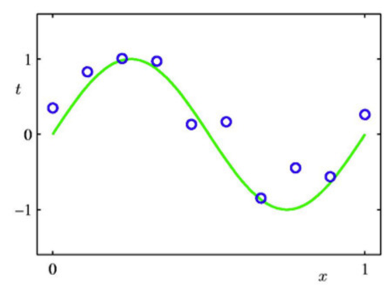
size
19

Classification vs. Regression

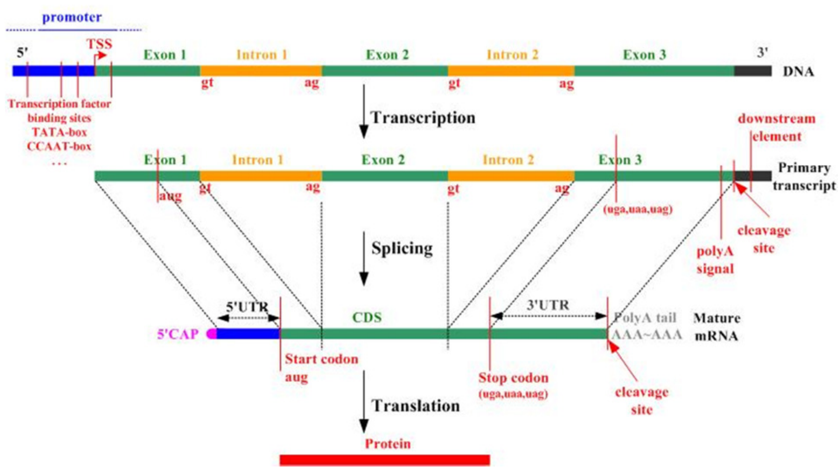
- Same: "Learn a function from labeled examples"
- Different: Domain of label: small set {Y, N, …} vs $\Re$
  Why make the distinction?
  - Historically, they have been studied separately
  - The type of label can determine whether an algorithm will work or not work

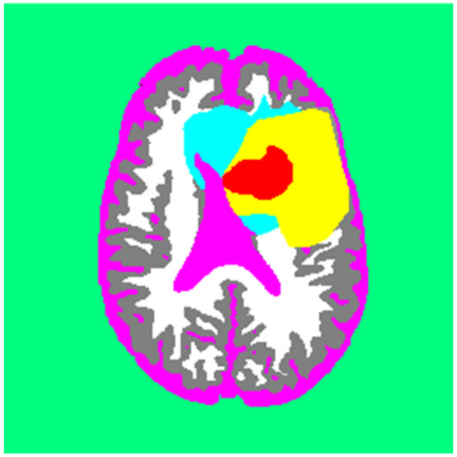


- Classification
  - "Separate the data"
- Regression
  - "Fit the data"



Learning Complex Labels

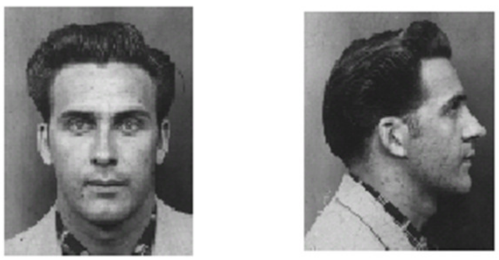




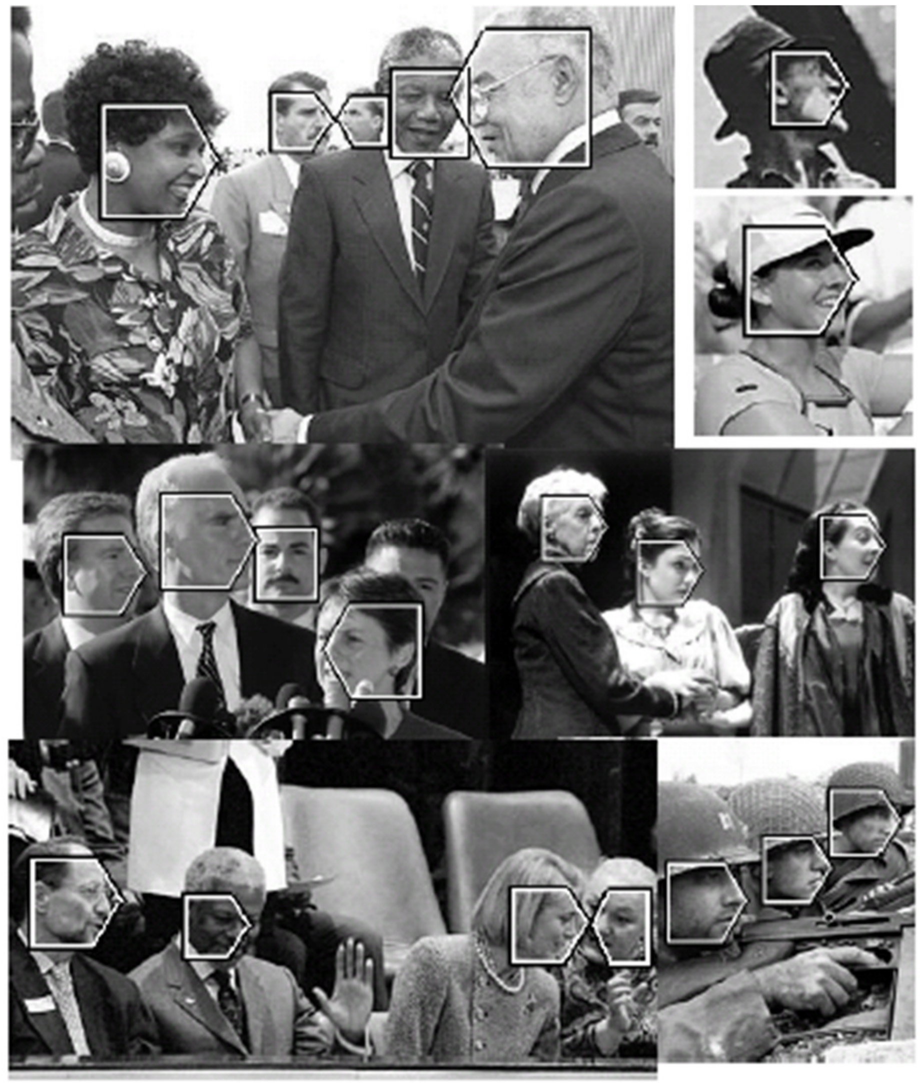
- Learning non-IID Data
  - Sequences
  - Images
  - …

Object detection

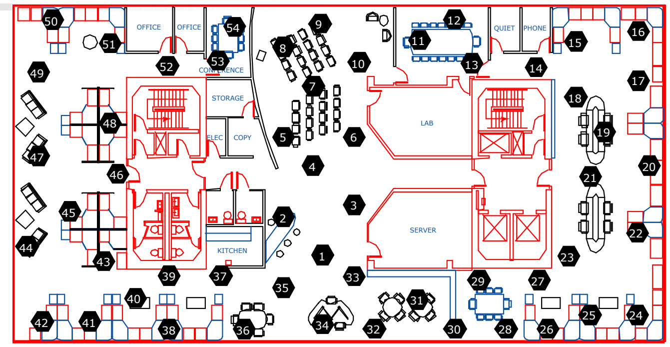
(Prof. H. Schneiderman)

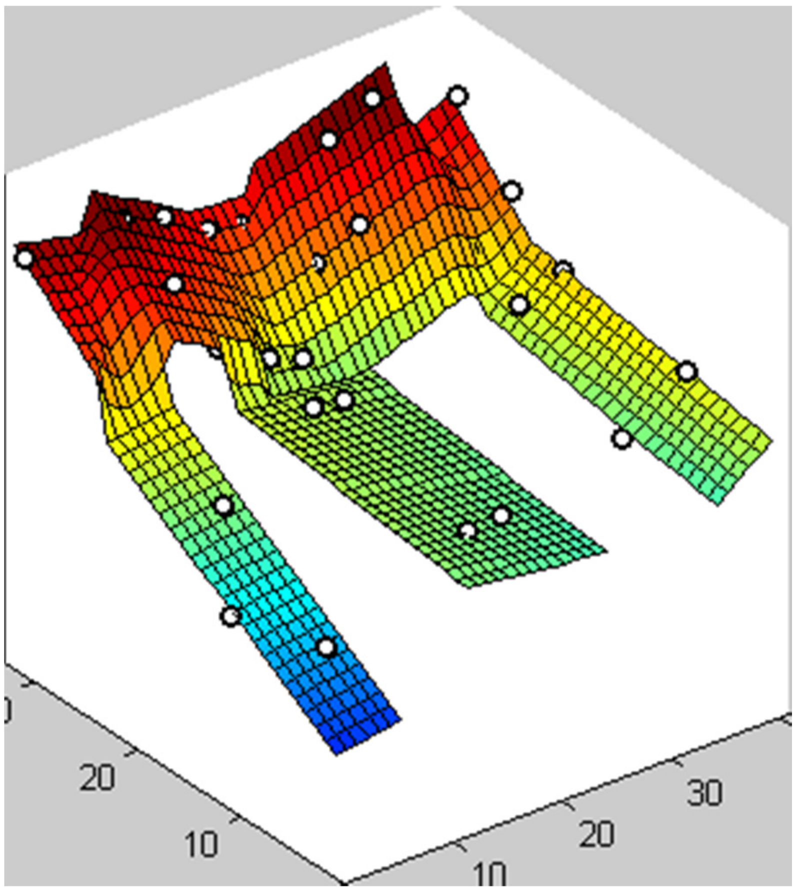


Example training images
for each orientation



Modeling sensor data





[Guestrin et al. ’04]

- Measure temperatures at some locations
- Predict temperatures throughout the environment

Type of Supervised Learning

- Supervised learning
  - A "teacher" provides a label for **each** instance in the training set
  - Categorical: Classification
    Numeric label: Regression
- Semi-supervised learning
  - Teacher provides label for **SOME** of the instances…
  - The unlabeled instances still provide information about $p(x)$ [but not about $p(y|x)$]

Typical Taxonomy for ML

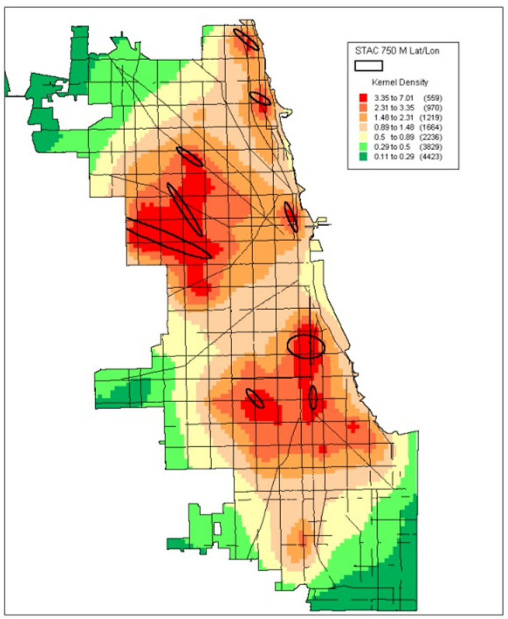
- Learning to Predict
  - Supervised learning
  - Semi-supervised learning
- Learning to Model
  - Unsupervised learning
  - Clustering
  - Generative Models
- Learning to Control
  - Reinforcement learning

Learning to Model: Unsupervised Learning

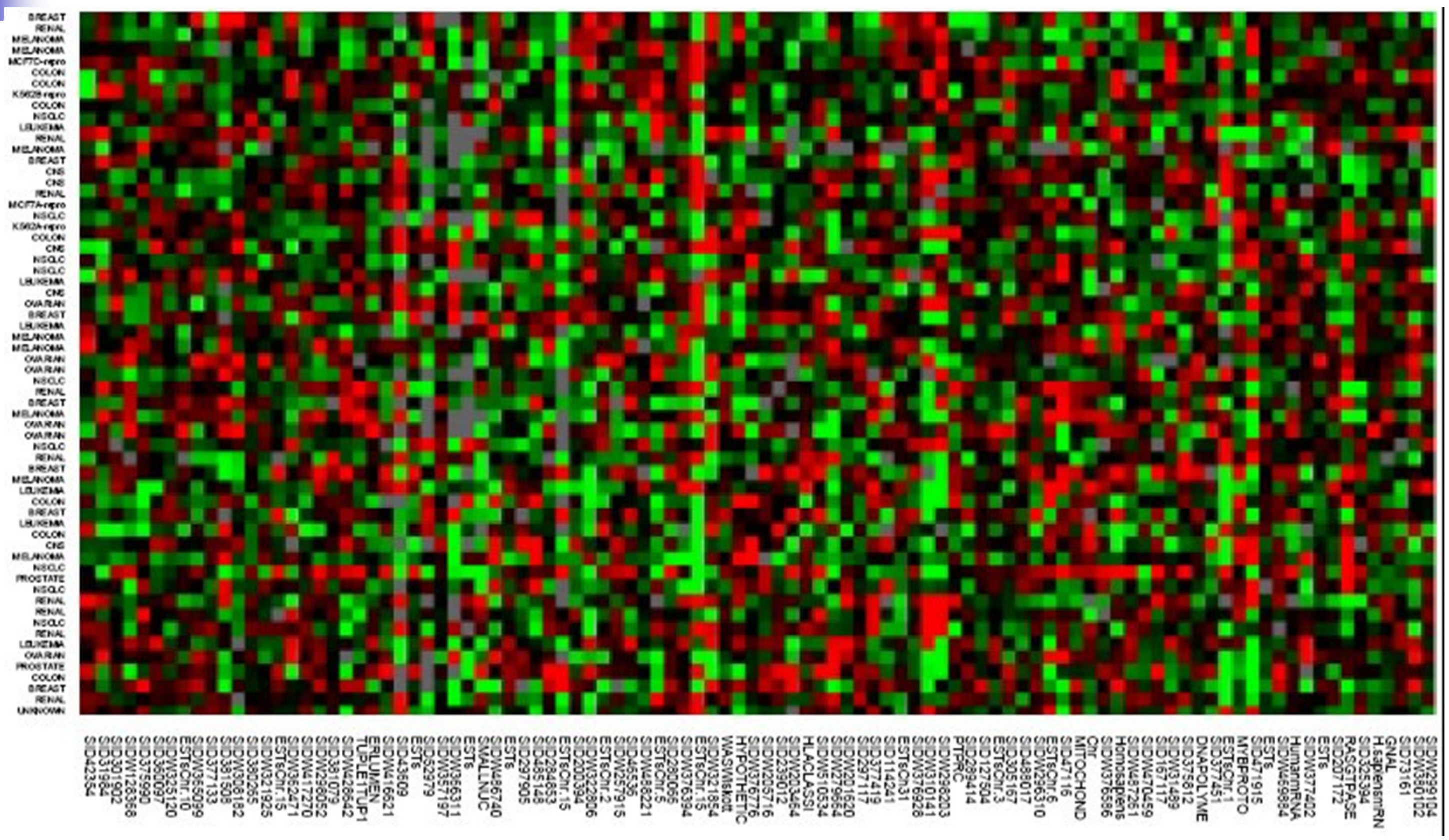
- Training data is just unlabeled instances
  - Not “(instance, label)” pairs
- Applications:
  - (Pre-processing for supervised learning)
  - Dimensionality reduction
  - Data compression
  - Outlier detection
  - Segmentation/clustering
  - Probability density estimation
  - …

Density Estimation

- Learning Generative Model
- Clustering



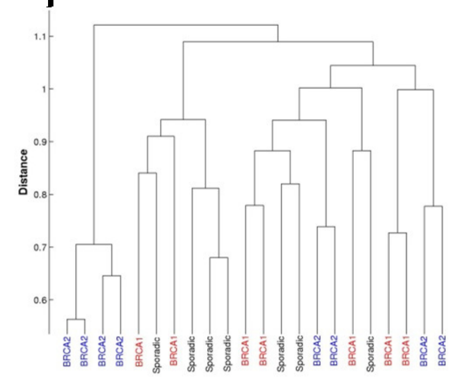
Example: Unsupervised Learning

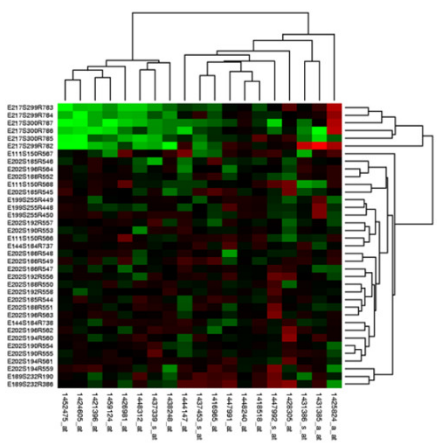


DNA microarray data (taken from [HTF])

Examples: Unsupervised Learning

- Applications wrt DNA microarray
  - Clustering: Group genes or samples into similar expression profiles
    - Hierarchical clustering: Dendogram
  - Bi-clustering:
    Subset of genes exhibiting
    similar expression pattern wrt subset of samples
  - Dimension reduction
  - …





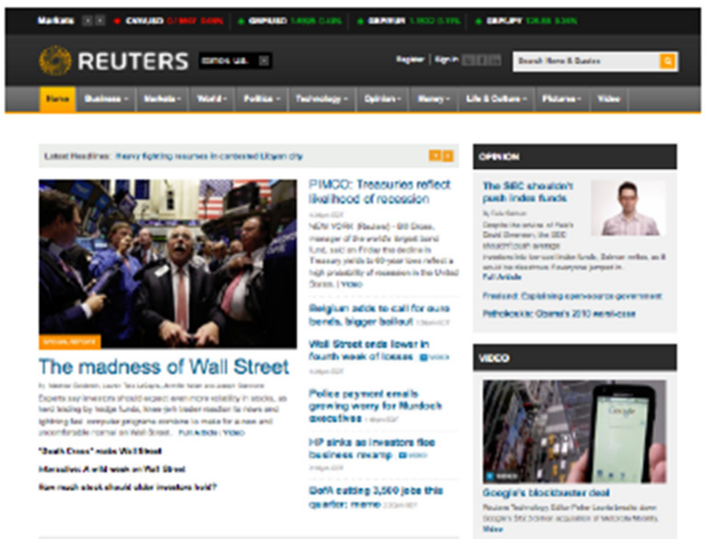
http://www.biomedcentral.com/1471-2105/5/126/figure/F1?

Finding Structure in Data

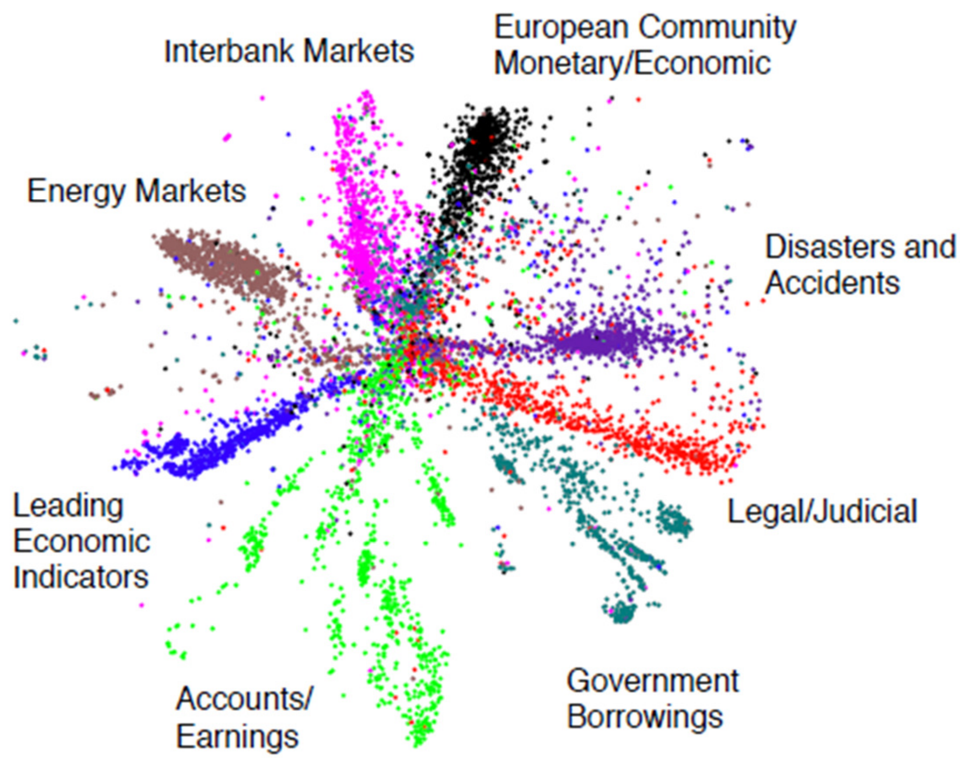
$$P(\mathbf{x}) = \frac{1}{\mathcal{Z}} \sum_{\mathbf{h}} \exp\left[\mathbf{x}^\top \mathbf{W} \mathbf{h}\right]$$

Vector of word counts on a webpage

Latent variables: hidden topics



804,414 newswire stories



from Ruslan Salakhutdinov

Finding Structure in Data

Collaborative Filtering/
Matrix Factorization/
Product Recommendation

Hierarchical Bayesian Model

Rating value of user i for item j

Latent user feature (preference) vector

Latent item feature vector

$$r_{ij}|\mathbf{u}_i, \mathbf{v}_j, \sigma \sim \mathcal{N}(\mathbf{u}_i^\top \mathbf{v}_j, \sigma^2),$$

$$\left.\begin{aligned} \mathbf{u}_i|\sigma_u &\sim \mathcal{N}(\mathbf{0}, \sigma_u^2 \mathrm{I}), \\ \mathbf{v}_j|\sigma_v &\sim \mathcal{N}(\mathbf{0}, \sigma_v^2 \mathrm{I}). \end{aligned}\right\}$$

Latent variables that we infer from observed ratings.

- Infer latent variables and make predictions using Markov chain Monte Carlo (MCMC).

Finding Structure in Data

Collaborative Filtering/
Matrix Factorization/
Product Recommendation

Learned ``genre''

Netflix dataset:
480,189 users
17,770 movies
Over 100 million ratings.

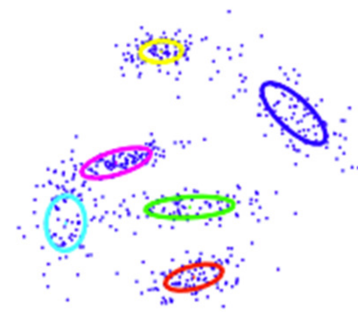
Fahrenheit 9/11
Bowling for Columbine
The People vs. Larry Flynt
Canadian Bacon
La Dolce Vita

Independence Day
The Day After Tomorrow
Con Air
Men in Black II
Men in Black

Friday the 13th
The Texas Chainsaw Massacre
Children of the Corn
Child's Play
The Return of Michael Myers

- Part of the wining solution in the Netflix contest (1 million dollar prize).

Unsupervised Learning

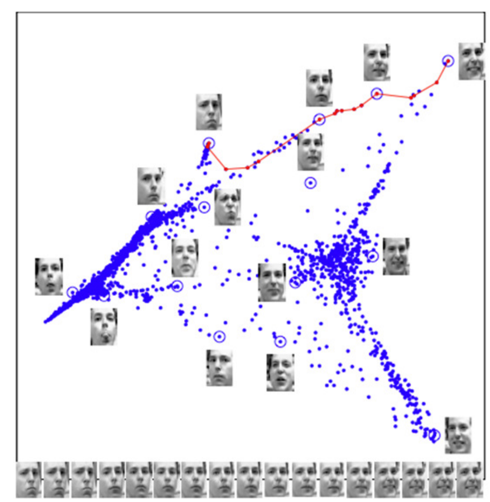


The goal is to construct statistical model that finds useful representation of data:

- Clustering
- Dimensionality reduction
- Modeling the data density
- Finding hidden causes (useful explanation) of the data

Unsupervised Learning can be used for:

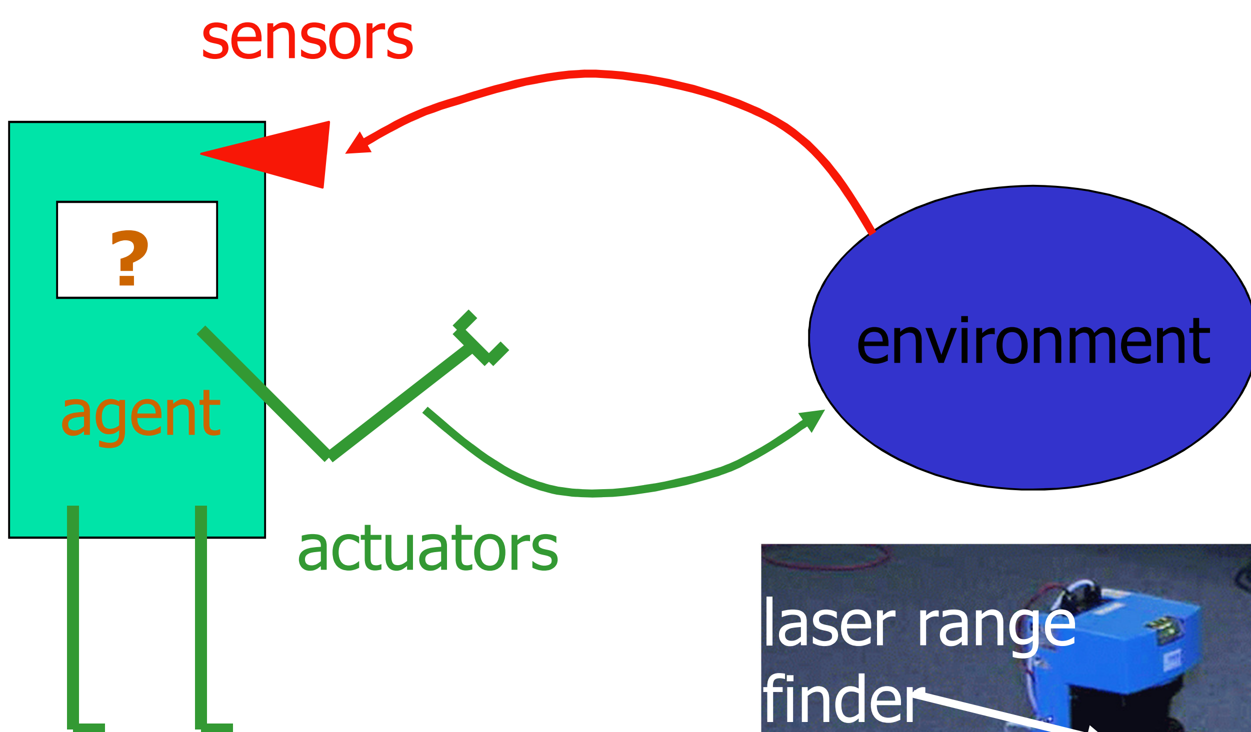
- Structure discovery
- Anomaly detection / Outlier detection
- Data compression, Data visualization
- Used to aid classification/regression tasks

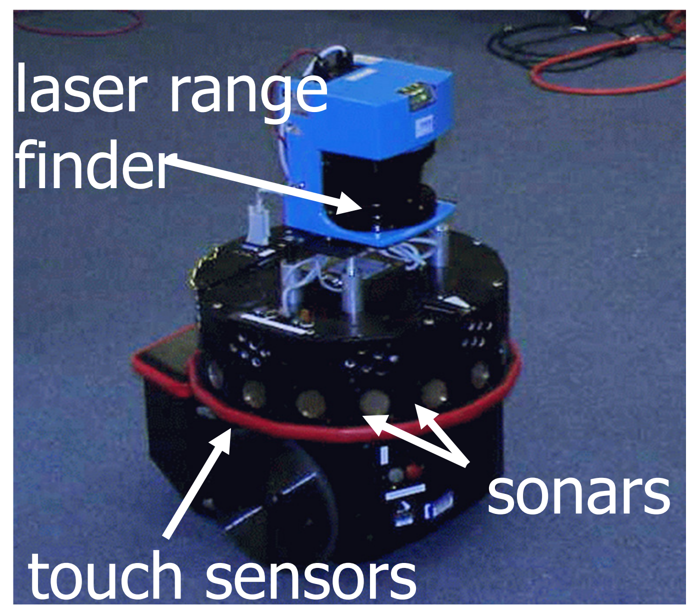


Typical Taxonomy for ML

- Learning to Predict
  - Supervised learning
  - Semi-supervised learning
- Learning to Model
  - Unsupervised learning
  - Clustering
  - Generative Models
- Learning to Control
  - Reinforcement learning

Notion of an Agent





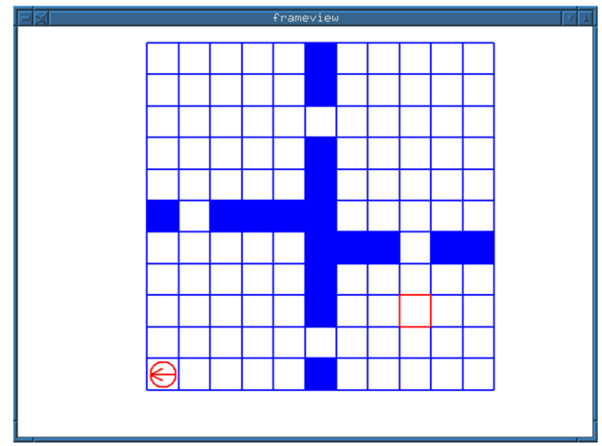
Source: robotics.stanford.edu/~latombe/cs121/2003/home.htm

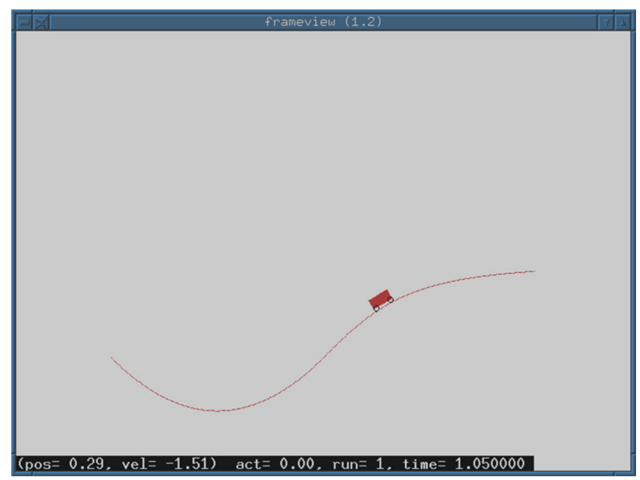
Learning to Control: Reinforcement Learning

- Consider
  - playing backgammon…
  - driving a car …
  - controlling a factory…
- Not a single action, but a SEQUENCE of actions
- Agents learn “behavior” through trial-and-error interactions with a dynamic environment
  - No teacher telling the agent wrong or right
  - But… a critic that gives a reward / penalty for the agent’s action
- Real world applications:
  - Robotics
  - Combinatorial search problems, such as games
  - Industrial manufacturing
  - …

Learning to act

- Reinforcement learning
- An agent
  - Makes sensor observations
  - Must select action
  - Receives rewards
    - positive for “good” states
    - negative for “bad” states





[Ng et al. ’05]

Questions

- What is (machine) learning ?
- Is learning really possible?
  Can an algorithm really predict the future?
- Why learn?
- Is learning $\subset^?$ statistics ?

2: Is Learning Possible?

Is learning possible?

Can an algorithm really predict the future?

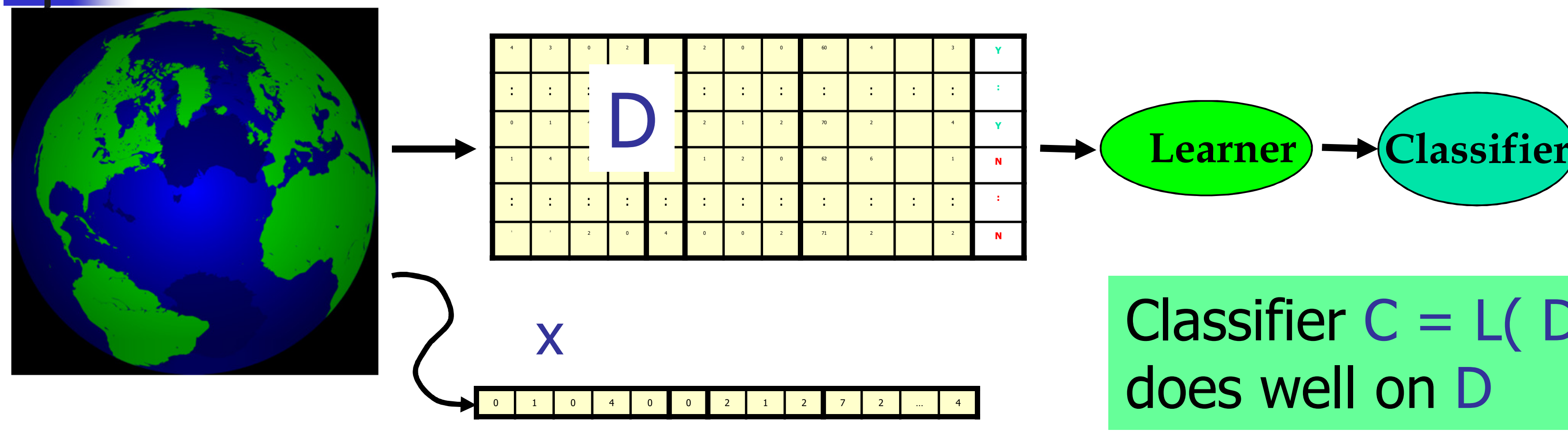
- No... not perfectly …

  Learning ≡ guessing;

  Guessing ⇒ might be wrong

- But...
  - Can do "best possible" (Bayesian)
  - Can USUALLY do CLOSE to optimally
- Empirically…

Why should Learning work?



Classifier $C = L(D)$ does well on D

If $x \in D$

$\Rightarrow$ L has seen $\approx x$

$\Rightarrow$ $C = L(D)$ does well on x

- x is common $\Rightarrow x \in D \Rightarrow$ $C(x)$ is $\approx$ correct
- x is NOT common $\Rightarrow$ Small penalty

Growth of Machine Learning

Machine learning is preferred approach to

- Speech recognition
- Natural language processing
- Computer vision
- Medical outcomes analysis
- Robot control
- …

…This trend is accelerating

Questions

- What is (machine) learning ?
- Is learning really possible?
  Can an algorithm really predict the future?
- Why learn?
- Is learning $\subset^?$ statistics ?

Why Learn?
Why not just "program it in"?

Appropriate Classifier ...

- … is not known
  Medical diagnosis… Credit risk… Control plant…
- … is too hard to "engineer"
  Drive a car… Recognize speech…
- … changes over time
  Plant evolves… Trending news …
- … user specific
  Adaptive user interface…

Why Machine Learning is especially relevant now!

- Growing flood of online **data**
  - customer records, telemetry from equipment, scientific journals, …
- Recent progress in **algorithms** and **theory**
  - SVM, Reinforcement Learning, Boosting, Gaussian Processes, …
  - Deep Learning
  - PAC-analysis, Bayesian Models, SRM, …
- Computational **power** is available
  - networks of fast machines
- Budding **industry** in many application areas
  - market analysis, adaptive process control, decision support, …

(Alberta Innovates Centre for Machine Learning)

ALBERTA INNOVATES CENTRE FOR MACHINE LEARNING

Questions

- What is (machine) learning ?
- Is learning really possible?
  Can an algorithm really predict the future?
- Why learn?
- Is learning $\subset^?$ statistics ? … or other discipline?

4. Is learning $\subset^?$ statistics?

Statistics ≡

- Use examples to identify best model
- Use model for predictions (labels of new instances, ...)

- Like Machine Learning, statistics also…
  - Deals with required # of samples, quality of output, ...
  - Over discrete / continuous,
    parameterized/not,
    complete/partial,
    frequentist/bayesian,
    ...

- But Machine Learning also …
  - deals with COMPUTATIONAL ISSUEs
  - different focus/frameworks
    (on-line, reinforcement, ...)
  - embraces MULTI-Variate predictions

Disciplines Relevant to ML…

- Statistics
  - Bayesian methods
  - Information theory
- Computing Science
  - Artificial intelligence
  - Computational complexity theory
- Control theory
- Philosophy
- Psychology and neurobiology
- …
- + applications of ML
  - … business, agriculture, web search, ad placement, sports, …
  - and everything else

Issues wrt Learning

- What is measure of improvement ?
  “accuracy/effectiveness”, “efficiency”, ...
- What is feedback ?
  Supervised, Delayed Reinforcement, Unsupervised
- What is representation of to-be-improved component?
  Rules, Decision Tree, Bayesian net, Neural net, ...
- What prior information is available?
  “Bias”, space of hypotheses, background theory, ...
- What statistical assumptions?
  - Stationarity (iid), Markovian, ...
  - "Noisy" or Clean,
  - ...

Summary

- Machine Learning is a **mature field**
  - solid theoretical foundation
  - many effective algorithms
- ML is *crucial* to large number of important **applications**
  - BioInformatics, WebReDesign, MarketAnalysis, Fraud Detection, …
- Fun: Lots of intriguing open questions!
- **Exciting time for Machine Learning**