HTF: 2
DHS: 1
RN: Ch 13

Introduction to Machine Learning

Machine Perception

An Example

Pattern Recognition Systems

The Design Cycle

Learning and Adaptation

Questions

- What is learning ?

- Is learning really possible?
  Can an algorithm really predict the future?

- Why learn?

- Is learning $\subset^?$ statistics ?

What is Machine Learning?

- "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

What else is Machine Learning?

- Data Mining
  - “The nontrivial extraction of implicit, previously unknown, and potentially useful information from data.”
    - W. Frawley, G. Piatetsky-Shapiro, C. Matheus
  - “..the science of extracting useful information from large data sets or databases.”
    - D. Hand, H. Mannila, P. Smyth
  - “Data-driven discovery of models and patterns from massive observational data sets.”
    - P. Smyth

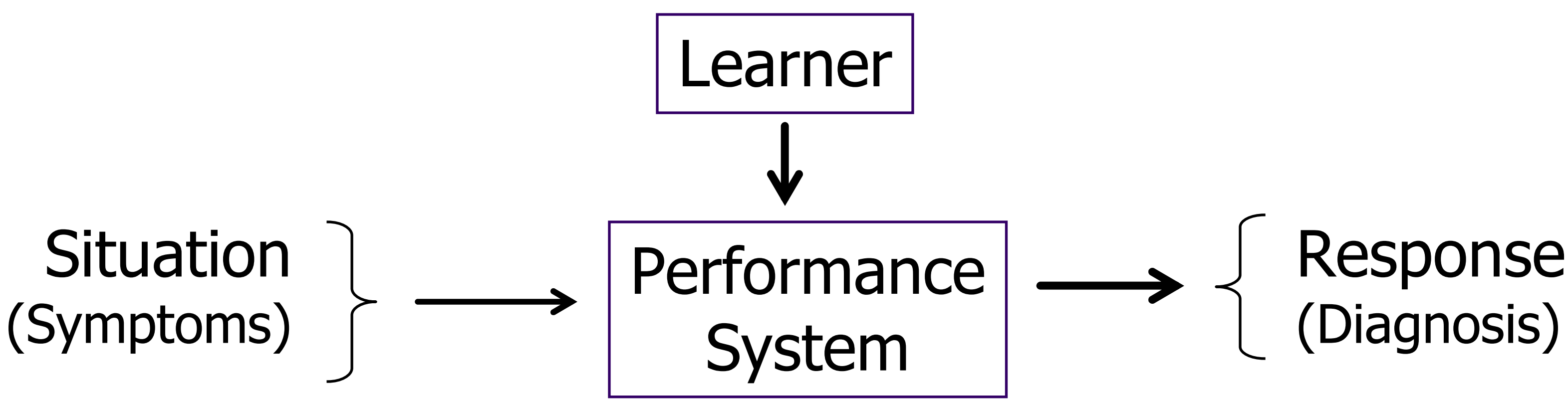
What is learning ?

- A_1: Improved performance ?

Performance System solves "Performance Task"
(Eg, Medical dx; Control plant; Retrieve webDocs; ...)

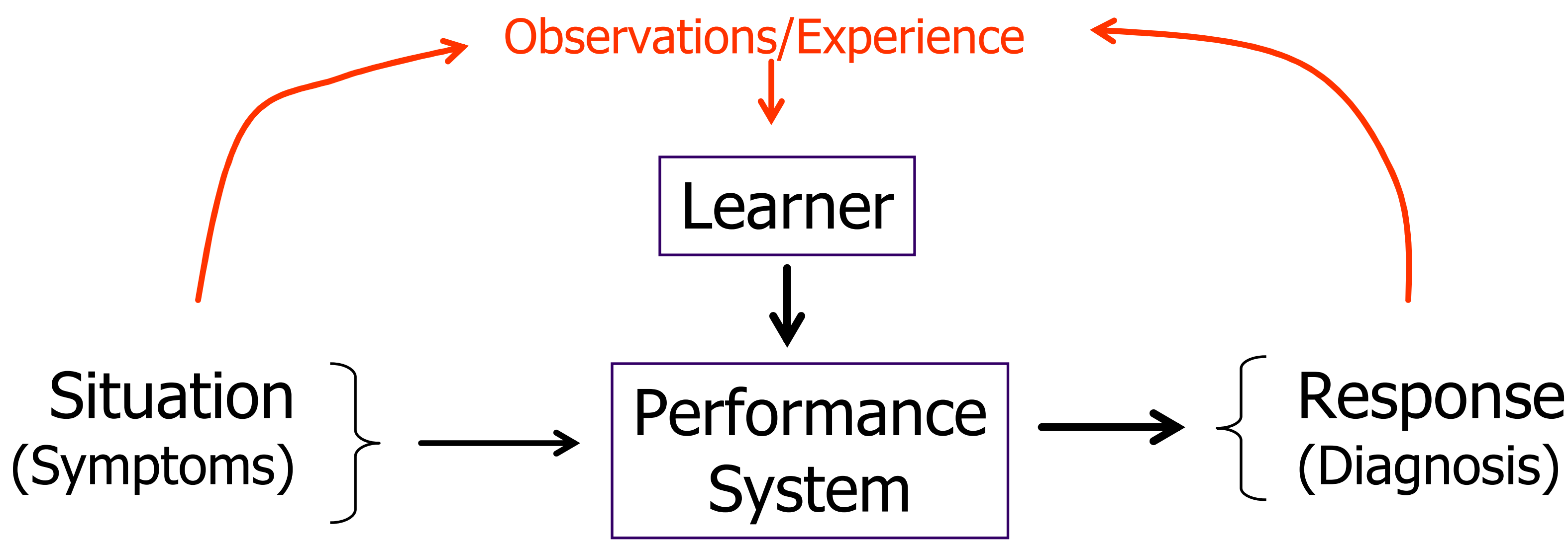
Learner makes Performance System "better“
More accurate; Faster; More complete; ...

(Eg, learn Dx/classification function, parameter setting, ...)



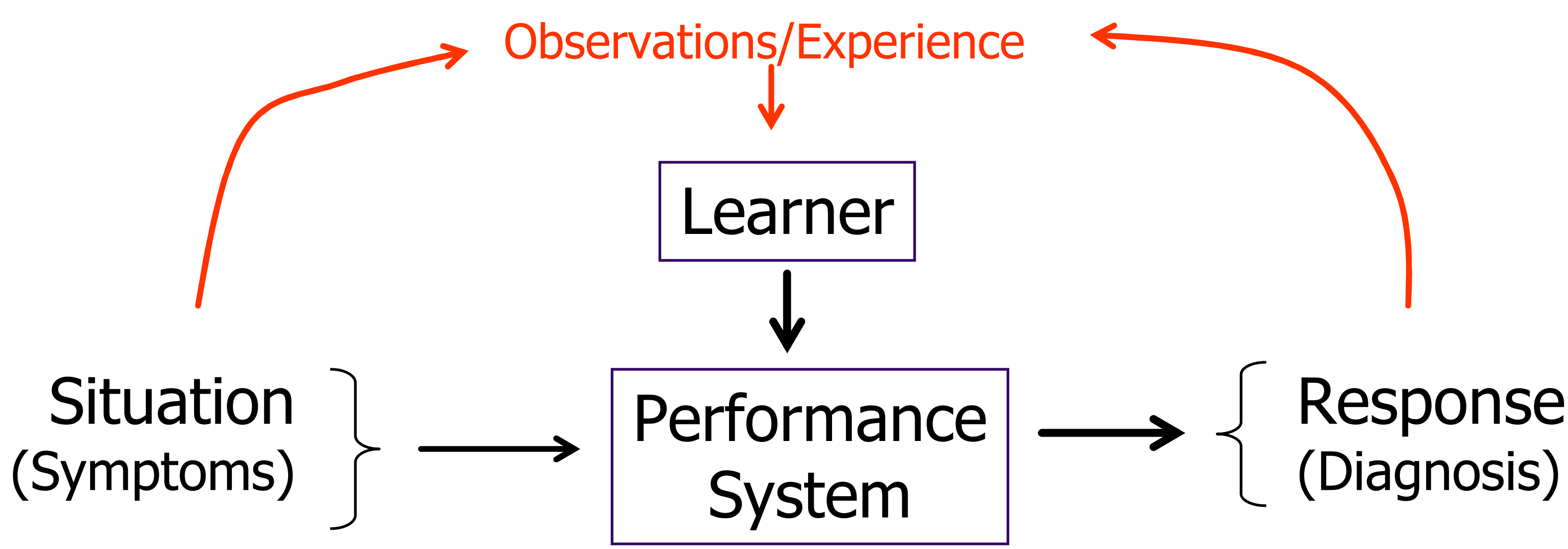
What is learning ? … con't

- A_1: Improved performance ?
  But… by re-programming? … faster CPU?
- A_2: Improved performance ?
  based on some "experience"



What is learning ? … con't

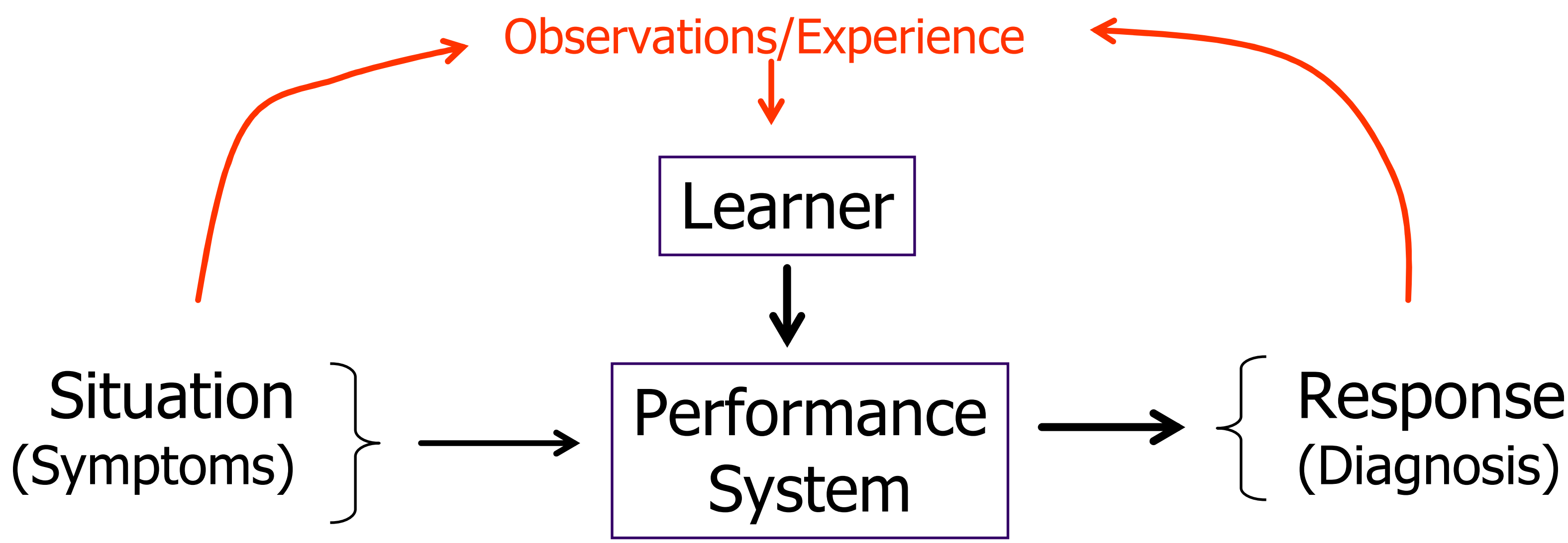
- A_2: Improved performance ?
  based on some "experience"
  but … simple memo-izing



What is learning ? … con't

- A_3: Improved performance
  based on **partial** "experience"
- Generalization (aka Guessing)

  deal with situations BEYOND training data



Learning Associations

- What things go together?
  - ?? Chips and beer?
- What is P(chips | beer) ?
  "The probability a particular customer will buy chips, given that s/he has bought beer."
- Estimate from data:
  - P(chips | beer) ≈ #(chips & beer) / #beer
  - Just count the people who bought beer ***and*** chips, and divide by the number of people who bought beer
- Not glamorous but… counting / dividing is learning!
- Is that all???

Learning to Perceive

Build a system that can recognize patterns:

- Speech recognition
- Fingerprint identification
- OCR (Optical Character Recognition)
- DNA sequence identification
- Fish identification
- …

Fish Classifier

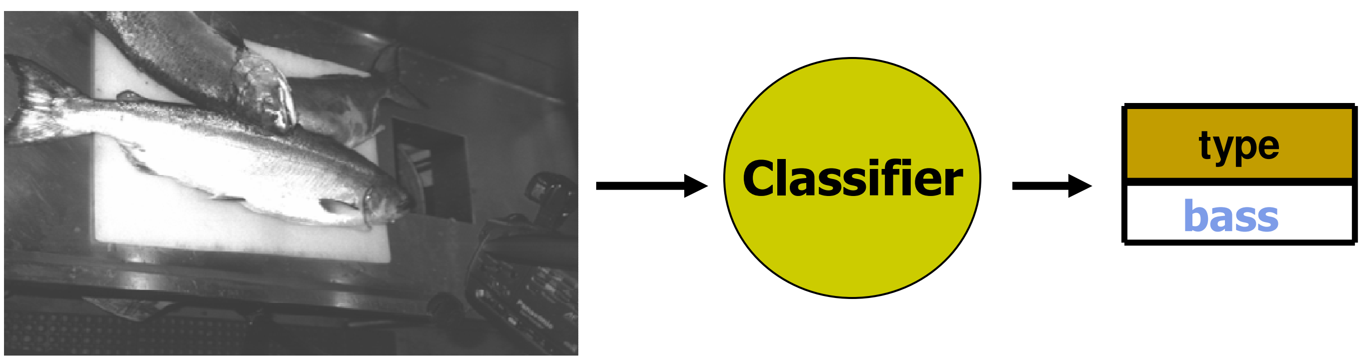
Sort Fish

into Species

Sea bass

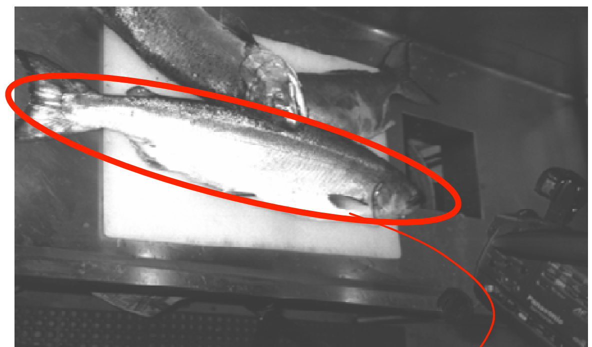
Salmon

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

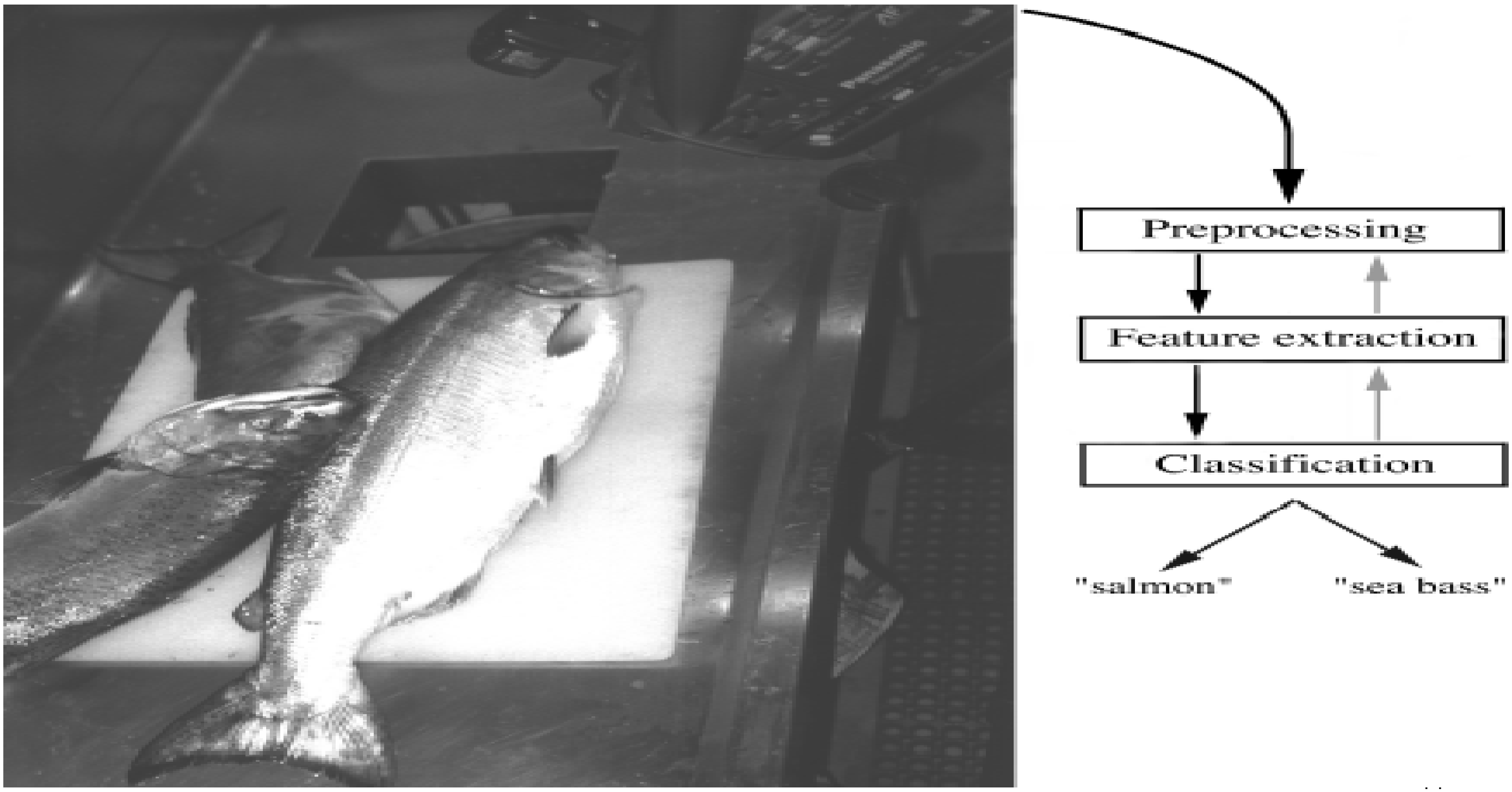
type
bass

Preprocessing

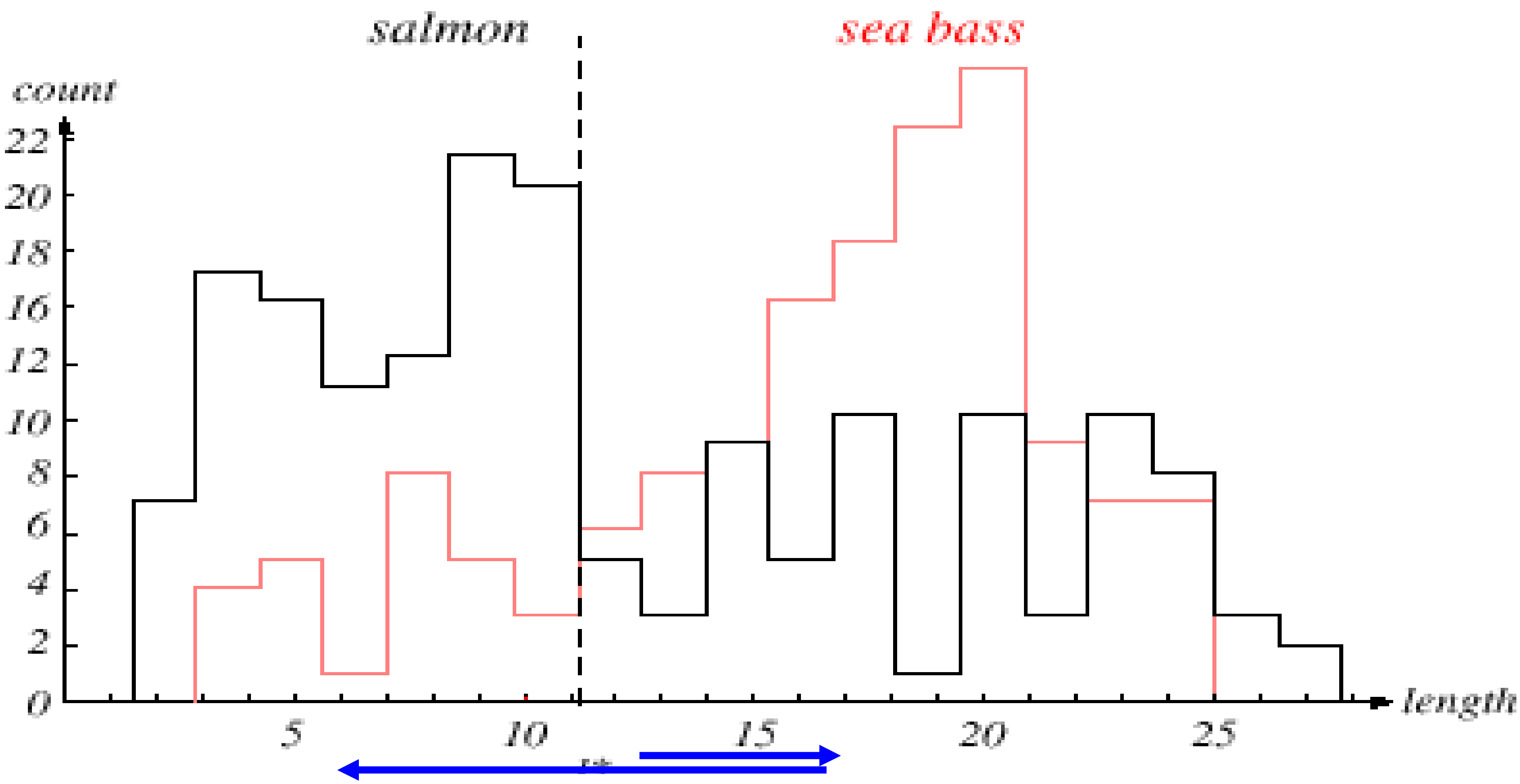
- Use *segmentation* to isolate
  - fish from background
  - fish from one another

- Send info about each single fish to *feature extractor*,
  … compresses data,
  into small set of features
- Classifier sees these features

Length	Wtdth	Pixel Bright	…	Light
50	10	2.8	…	Pale

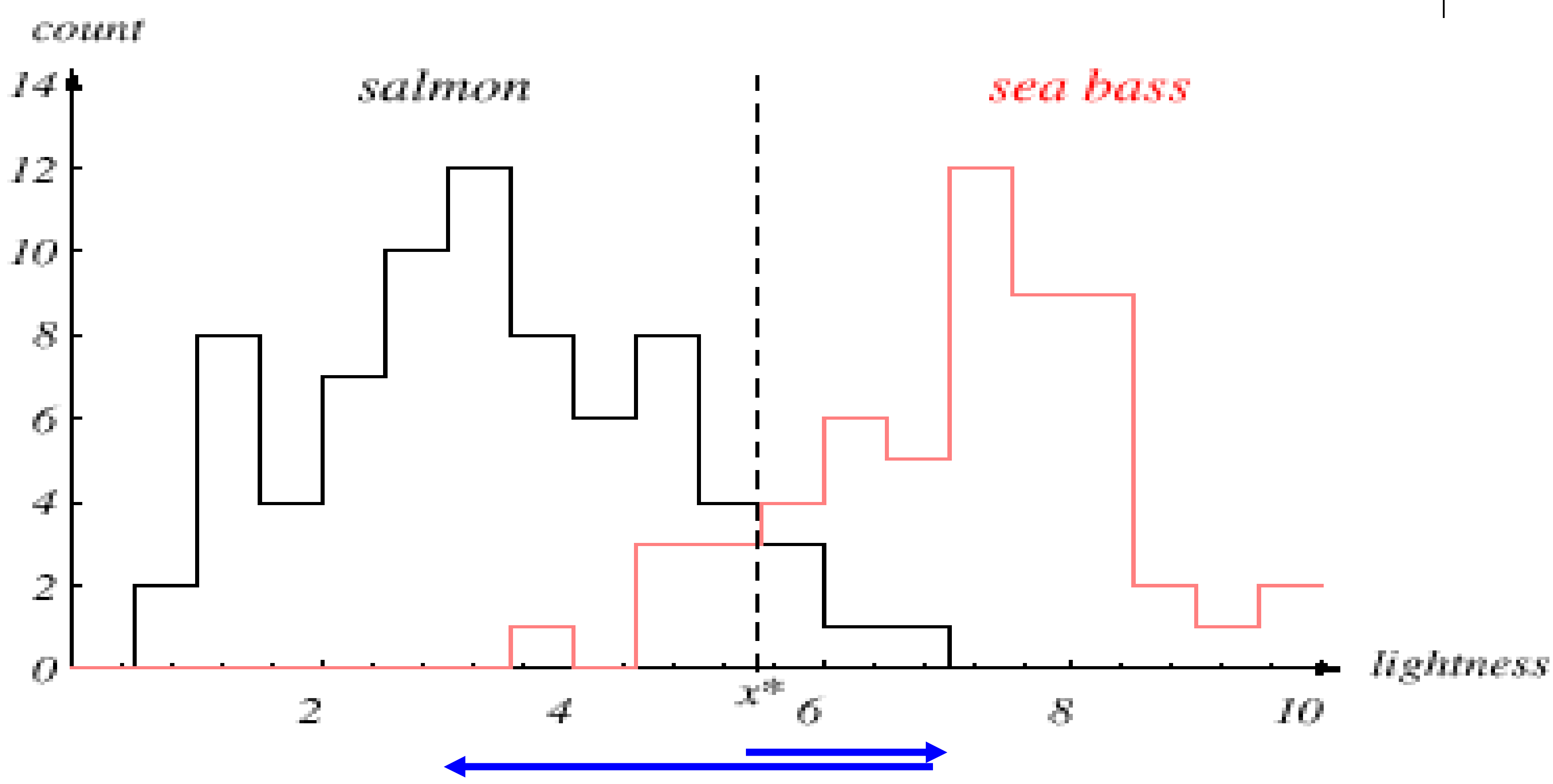

Preprocessing
Feature extraction
Classification
"salmon"
"sea bass"

Use “Length”?



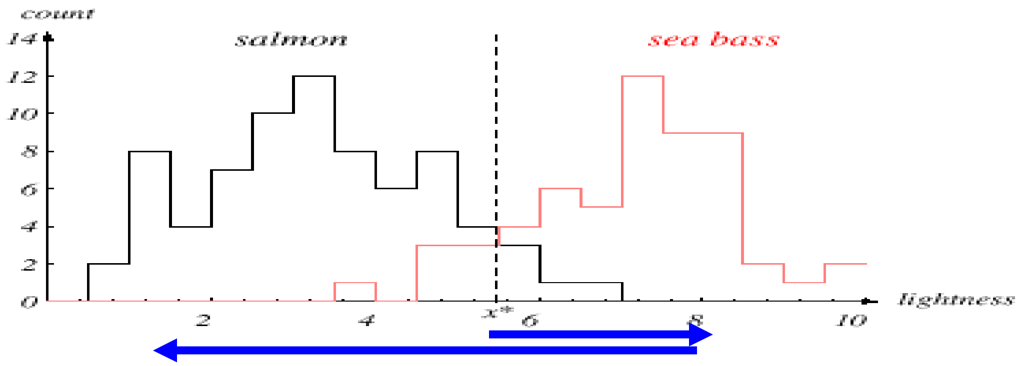
- Problematic… many incorrect classifications

Use "Lightness"?



- Better… fewer incorrect classifications
- Still not perfect

Where to place boundary?



- *Salmon Region* intersects *SeaBass Region*
  - ⇒ So no “boundary” is perfect
    - *Smaller* boundary ⇒ fewer SeaBass classified as Salmon
    - *Larger* boundary ⇒ fewer Salmon classified as SeaBass
- Which is best… depends on misclassification costs

Task of decision theory

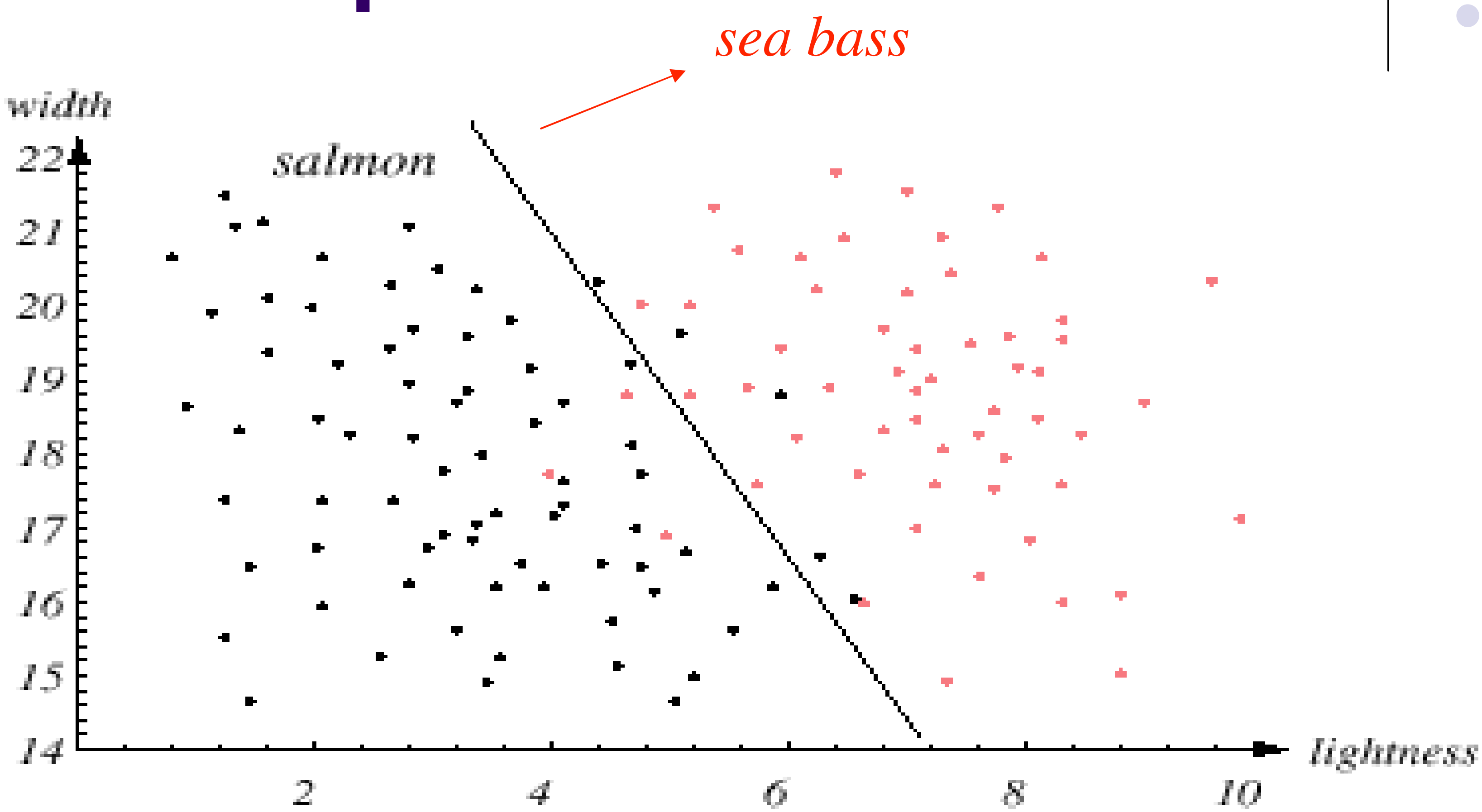
Why not 2 features?

- Use *lightness* and *width* of fish

Fish $\Rightarrow$ $x^T = [x_1, x_2]$

Lightness   Width

Use Simple Line ?



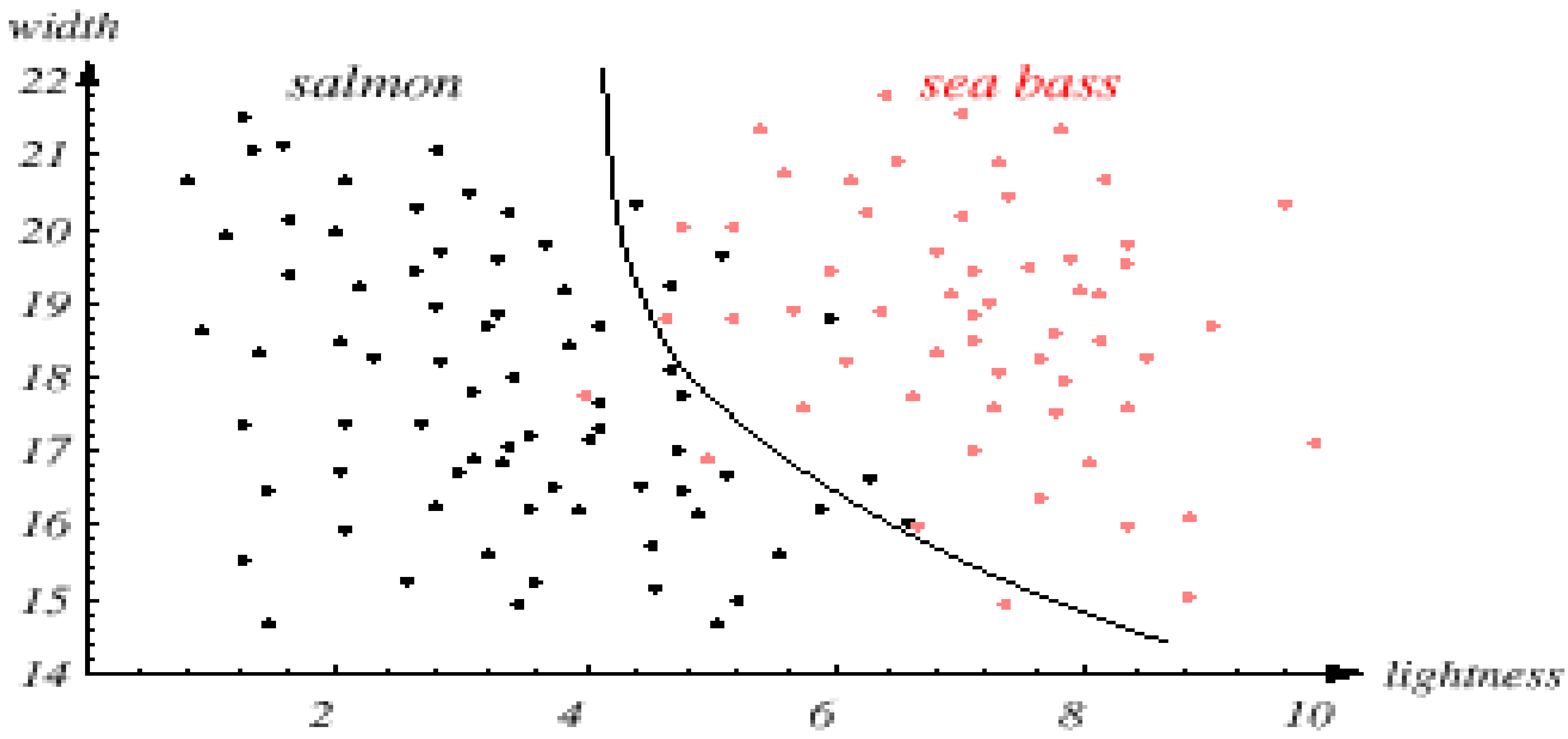
- Much better…
  very few incorrect classifications !

How to produce Better Classifier?

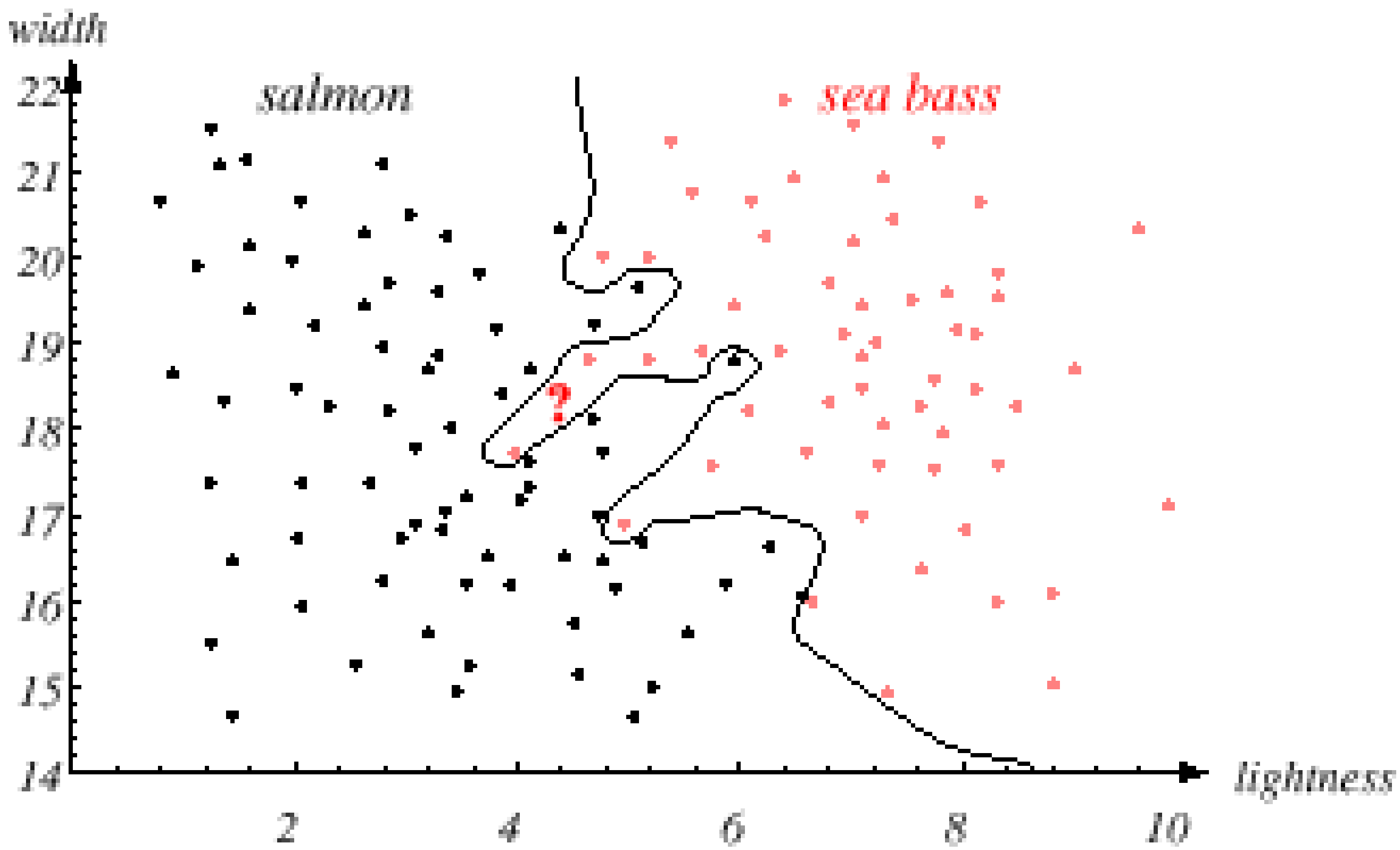
- Perhaps add other features?
  - Best: not correlated with current features
  - Warning: "noisy features" will ***reduce*** performance

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

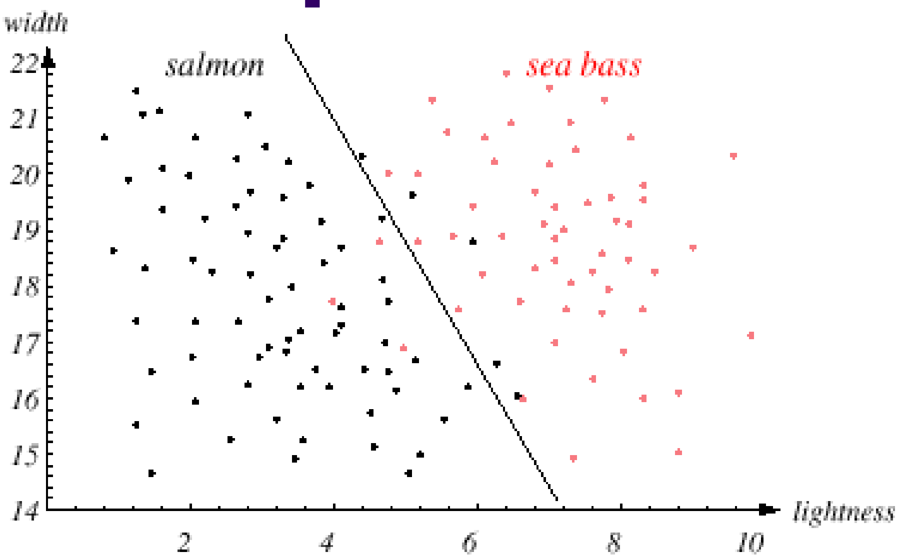
Simple (non-line) Boundary

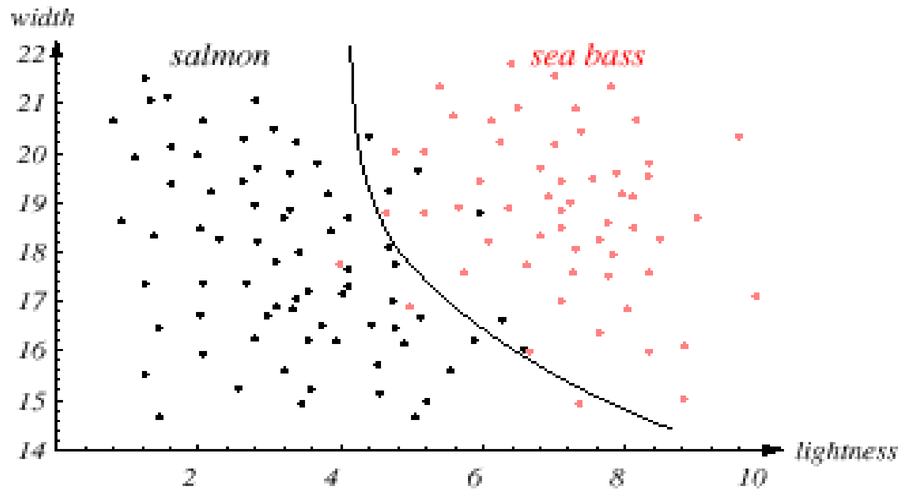


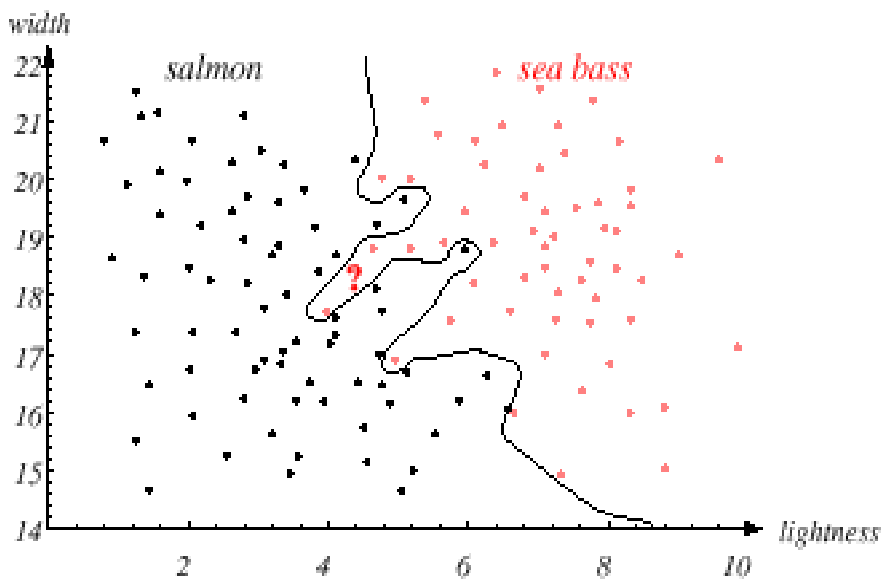
"Optimal Performance" ??



Comparison… wrt NOVEL Fish







Objective: Handle Novel Data

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

    $\neq$

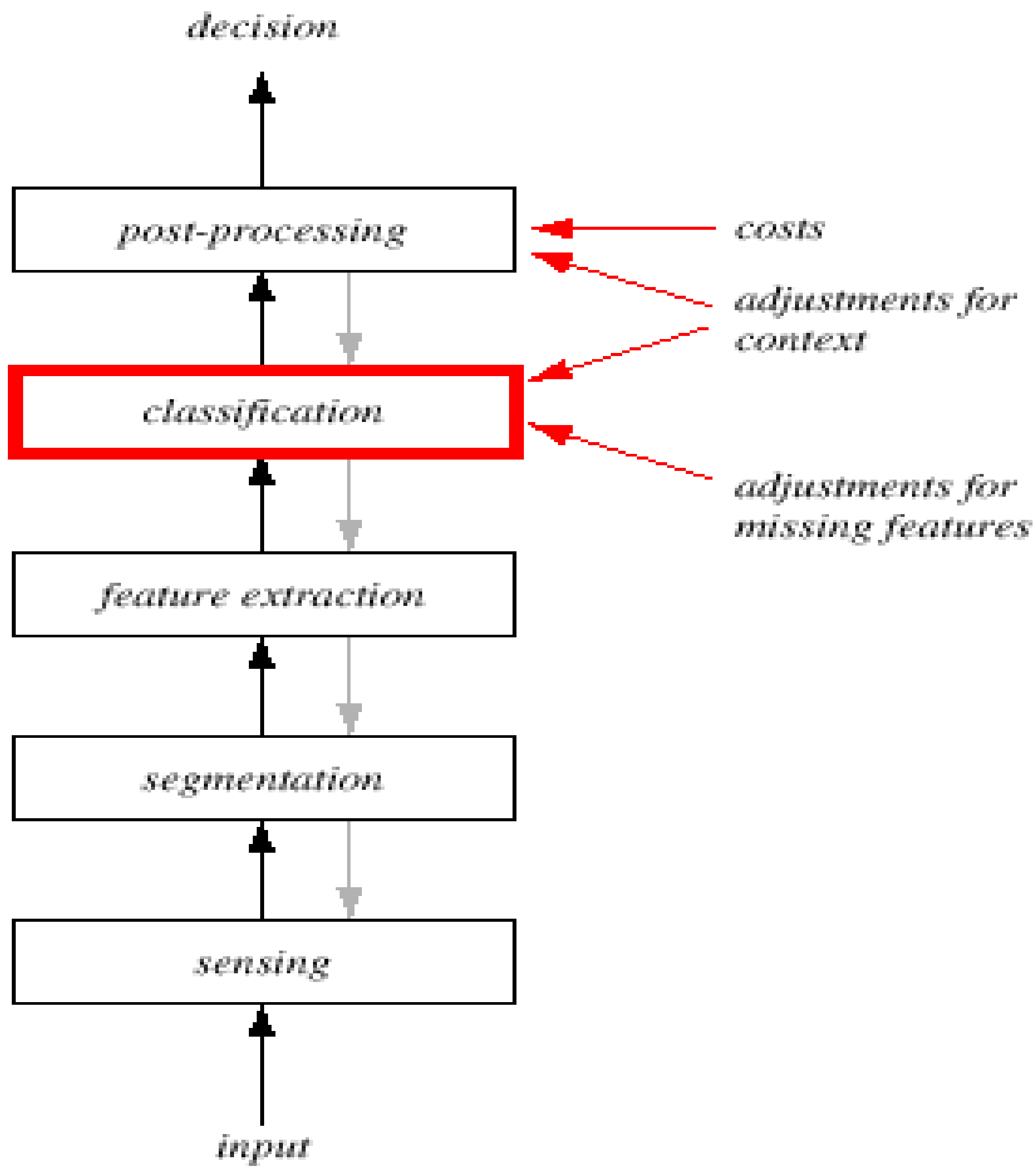
    Performance on NOVEL data

Issue of generalization!

Pattern Recognition Systems

- Sensing
  - Using transducer (camera, microphone, …)
  - PR system depends of the bandwidth
    - the resolution sensitivity distortion of the transducer

- Segmentation and grouping
  - Patterns should be well separated (should not overlap)

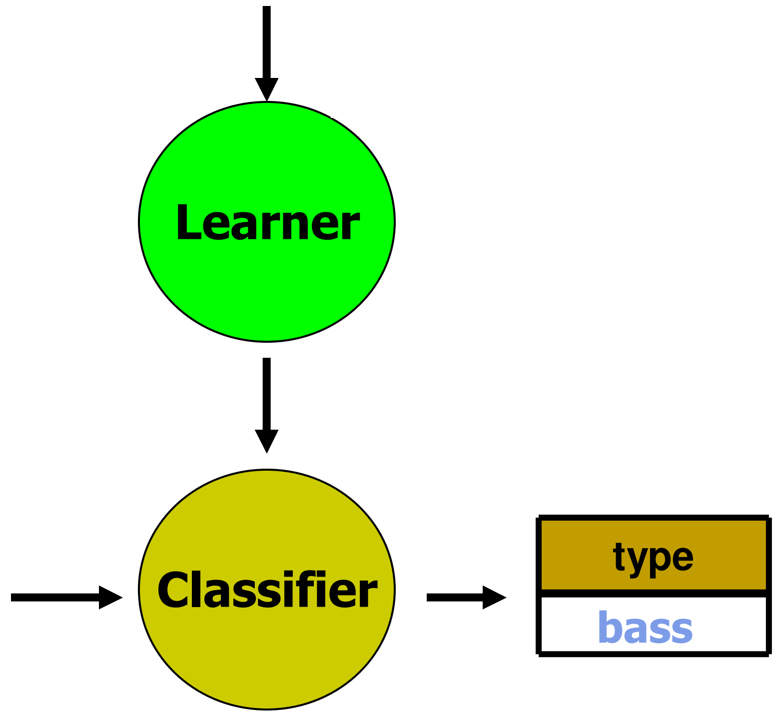

decision
post-processing
costs
adjustments for context
classification
adjustments for missing features
feature extraction
segmentation
sensing
input

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



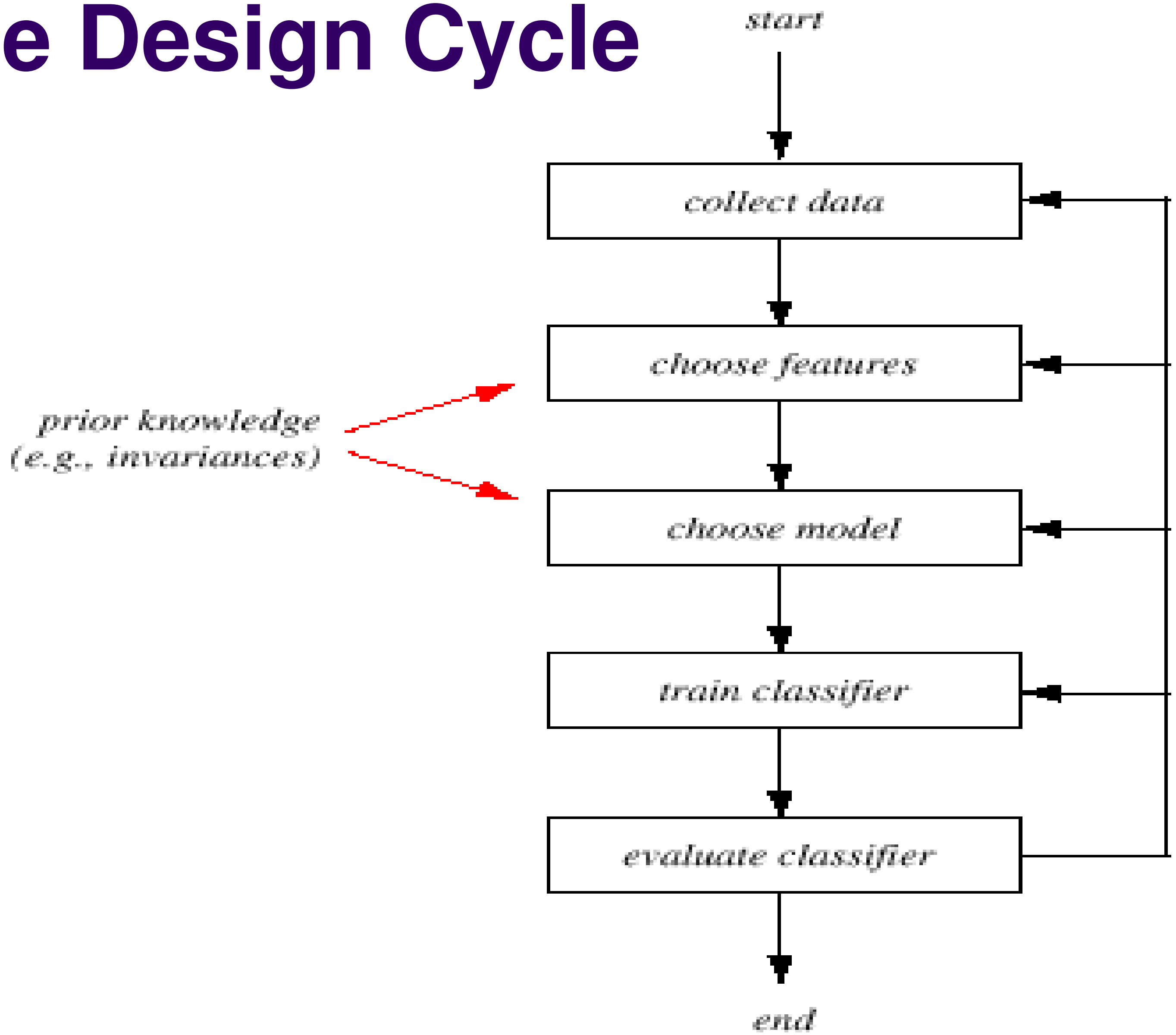
Width	Size	Eyes	…	Light
32	90	N	…	Pale

The Design Cycle

- Data collection
- Feature Choice
- Model Choice
- Training
- Evaluation

Computational Complexity

The Design Cycle

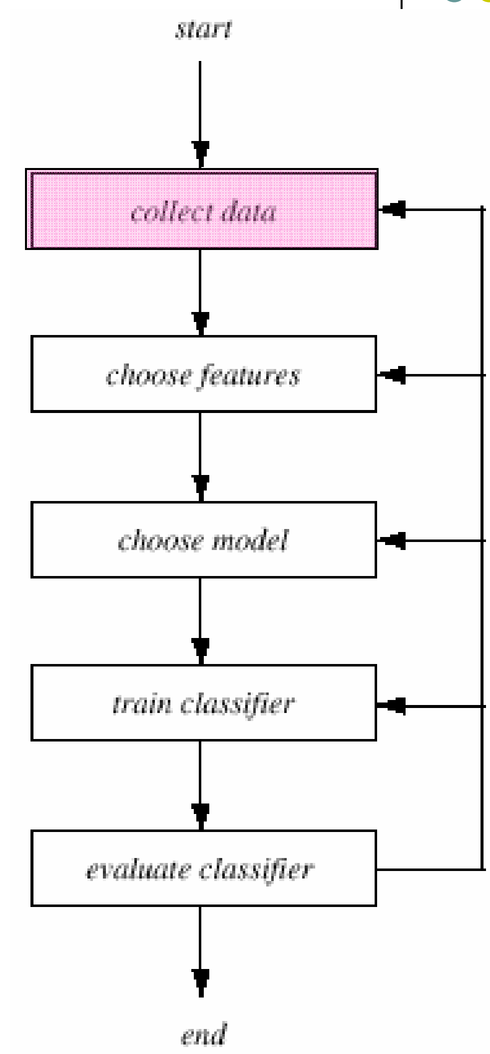


Computational Complexity

Data Collection

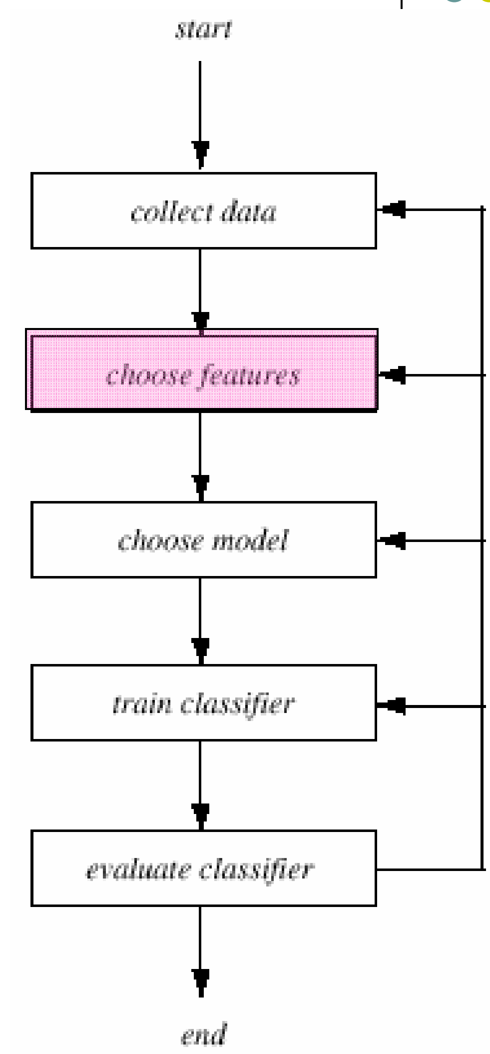
- Need set of examples

  for training and testing the system
- How much data?
  - sufficiently large # of instances
  - representative



Which Features?

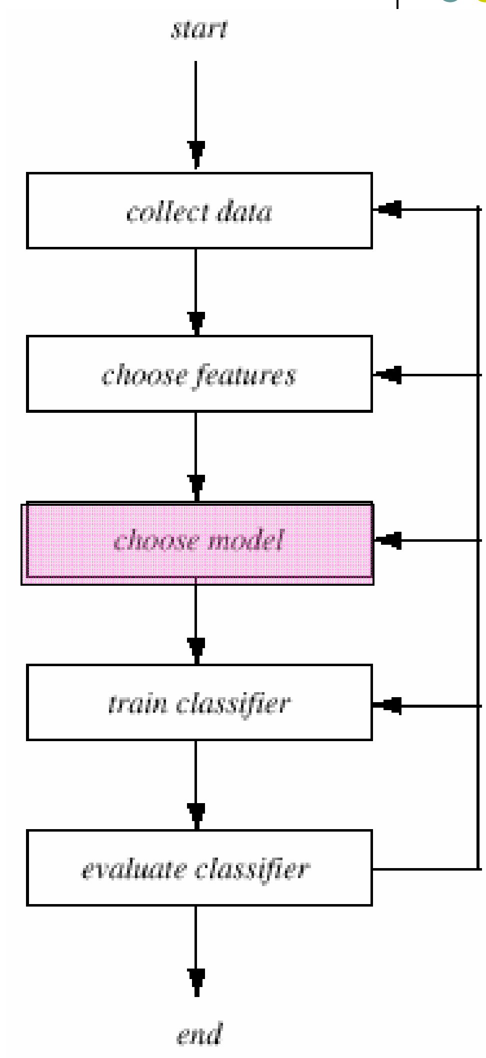
- Depends on characteristics of problem domain
- Ideally…
  - Simple to extract
  - Invariant to irrelevant transformation
  - Insensitive to noise



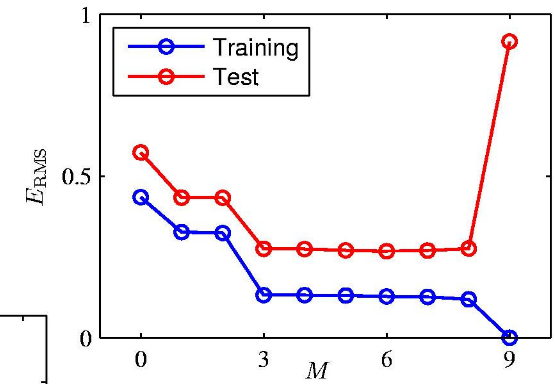
Which Model?

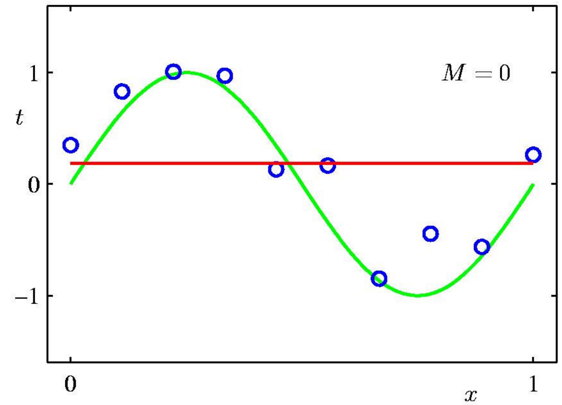
- Try one from simple class
  - Degree1 Poly
  - Gaussian
  - Conjunctions (1-DNF)
- If not good…

  try one from **yet** more complex class of models
  - Degree2 Poly
  - Mixture of 2 Gaussians
  - 2-DNF

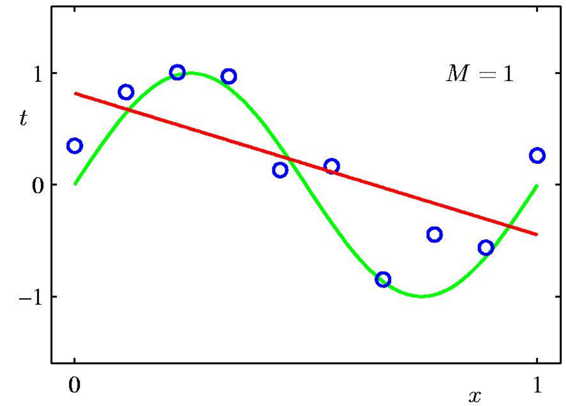


Which Model??

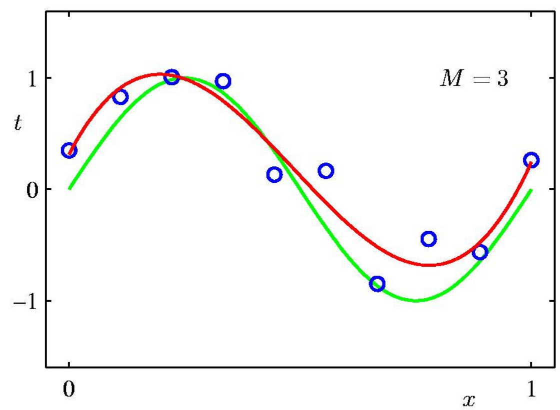




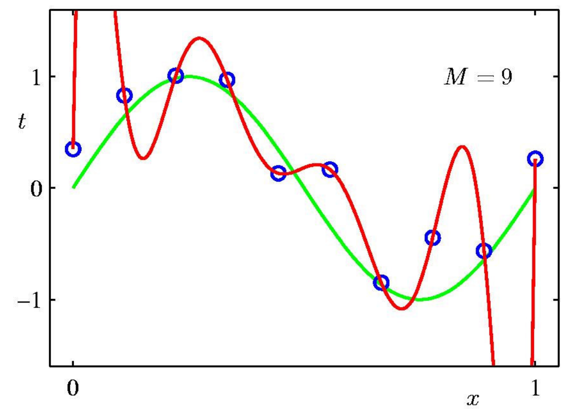
Constant (0)



Linear (1)



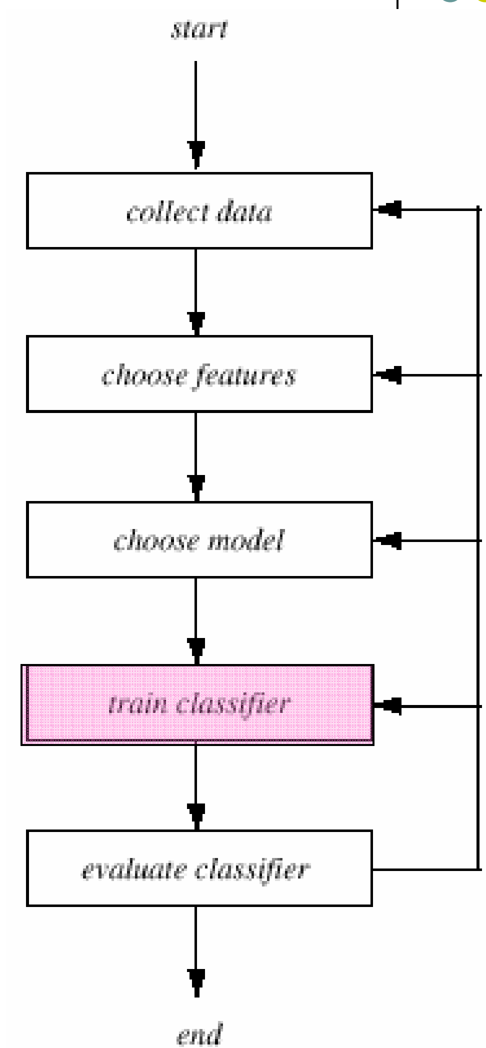
Cubic (3)



9th degree

Training

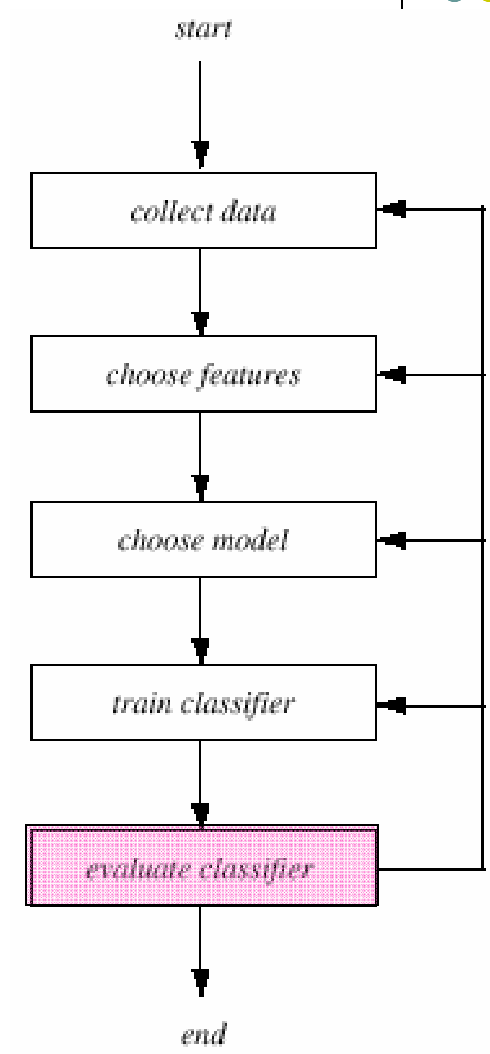
- Use data to obtain good classifier
  - identify best model
  - determine appropriate parameters
- Many procedures for training classifiers (and choosing models)



Evaluation

- Measure error rate

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



Computational Complexity

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

Learning and Adaptation

- Supervised learning
  - A teacher provides a category label for each pattern in the training set

- Unsupervised learning
  - System forms clusters or “natural groupings” of input patterns

Questions

- What is 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...

  Learning ≡ guessing;

  Guessing ⇒ might be wrong

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

Machine Learning studies …

Computers that use "*experiences*" to improve *performance* of some system

Computers that use "annotated data" to *autonomously* produce effective "rules"

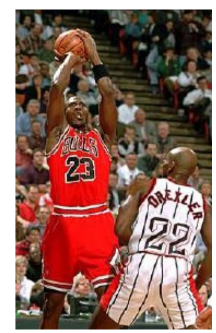
- to diagnose diseases
- to identify relevant articles
- to assess credit risk
- …

Successes: Mining Data Sets
Computer learns…

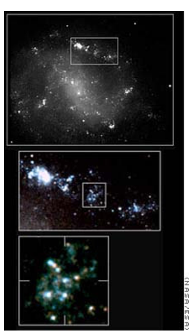
- to find ideal customers
  - Credit Card approval (AMEX)
    - Humans ≈50%; ML is >70% !

- to find best person for job
  - Telephone Technician Dispatch [Danyluk/Provost/Carr 02]
    - BellAtlantic used ML to learn rules to decide which technician to dispatch
    - Saved $10+ million/year

- to predict purchasing patterns
  - Victoria Secret (stocking)

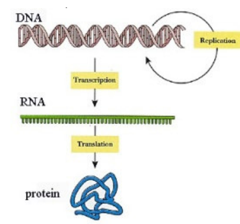


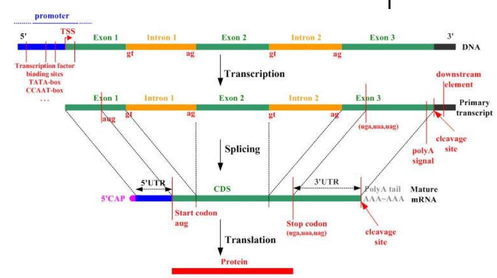
- to help win games
  - NBA (scouting)



- to catalogue celestial objects [Fayyad et al. 93]
  - Discovered 22 new quasars
  - >92% accurate, over tetrabytes

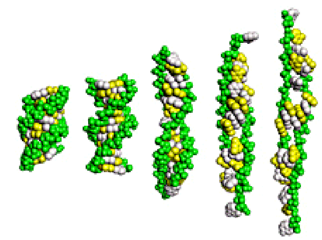
2: Sequential Analysis



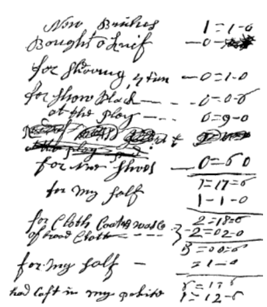


- **BioInformatics 1:** identifying genes
  - Glimmer [Delcher et al, 95]
  - identifies 97⁺% of genes, automatically!

- **BioInformatics 2:** Predicting protein function, …

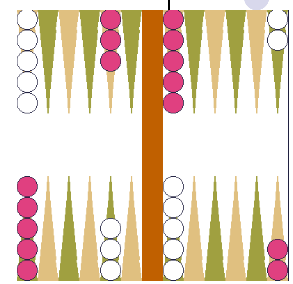


- **Recognizing Handwriting**

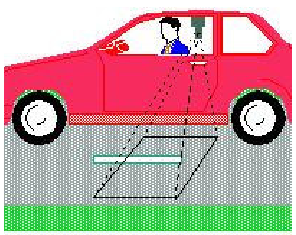


- **Recognizing Spoken Words**
  - **"How to wreck a nice beach"**

3: Control



- **TD-Gammon** (Tesauro 1993; 1995)
  - World-champion level play by **learning** …
  - by playing millions of games against itself!



- **Drive autonomous vehicles**
  - DARPA Grand Challenge (Thrun et al 2007)
- **Printing Press Control** (Evans/Fisher 1992)
  - Control rotogravure printer, prevent groves, ... specific to each plant
  - More complete than human experts
  - Used for 10+ years, reduced problems from 538/year to 26/year!

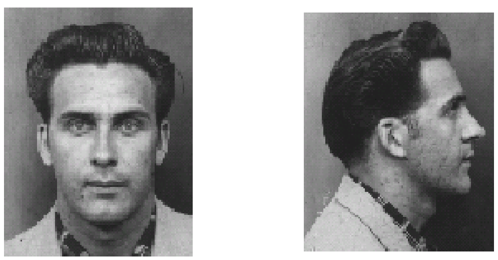
- **Oil refinery**
  - Separate oil from gas
  - … in 10 minutes (human experts require 1+ days)
- Manufacture nuclear fuel pellets (Leech, 86)
  - Saves Westinghouse >$10M / year
- **Adaptive** agents / user-interfaces

Growth of Machine Learning

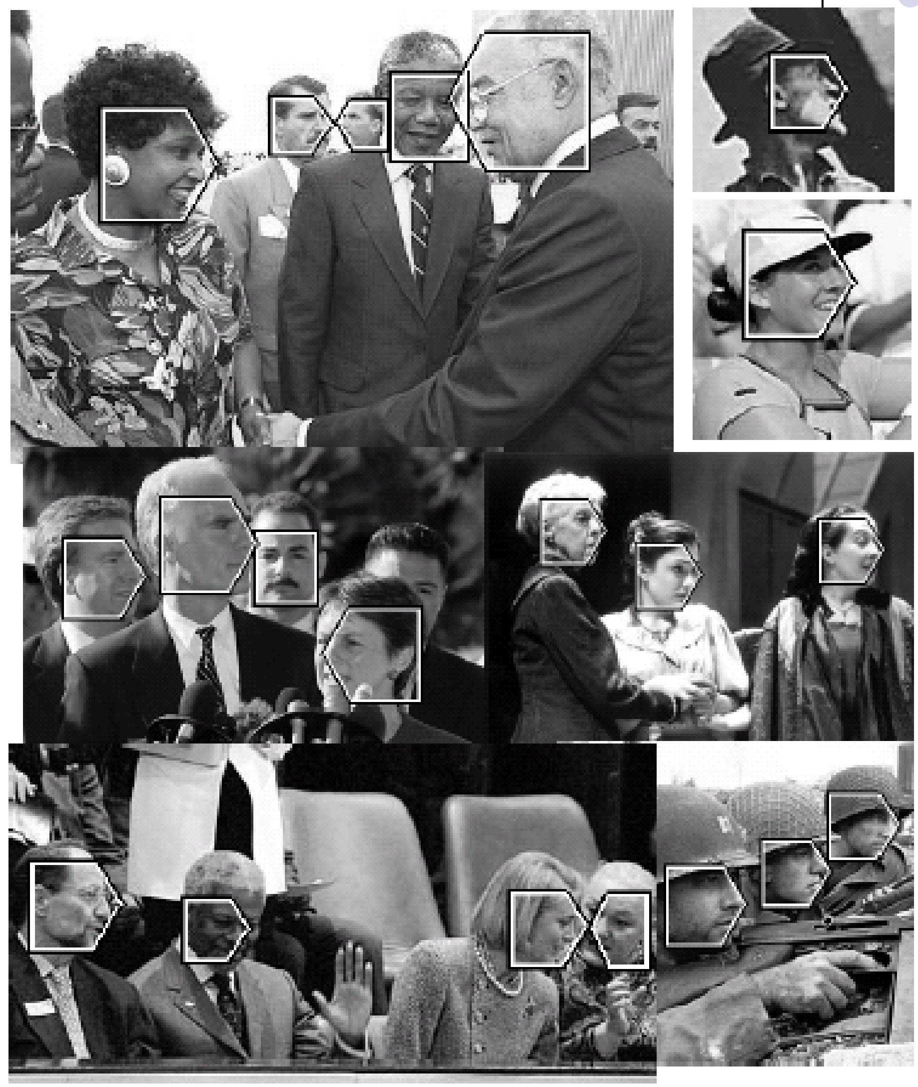
- Machine learning is preferred approach to
  - Speech recognition, Natural language processing
  - Computer vision
  - Medical outcomes analysis
  - Robot control
  - …

Object detection

(Prof. H. Schneiderman)



Example training images
for each orientation



Text classification

Company home page

vs

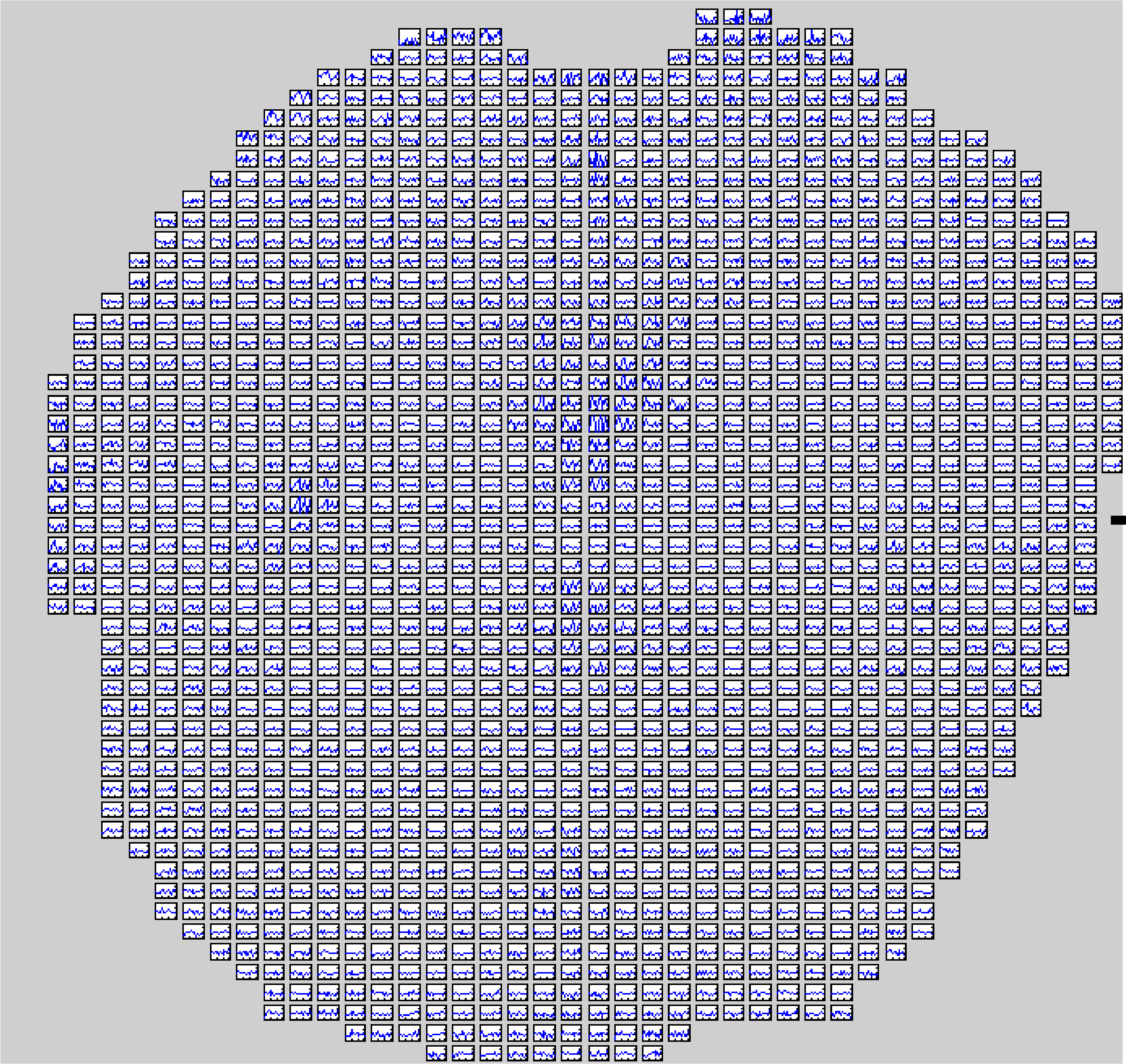
Personal home page

vs

Univeristy home page

vs

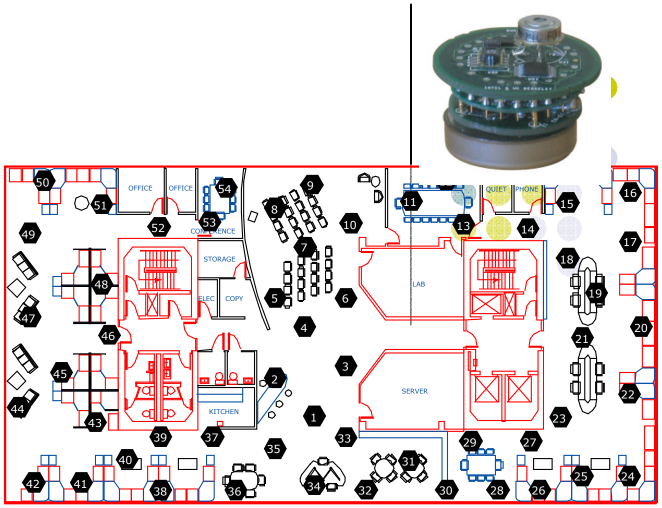
…

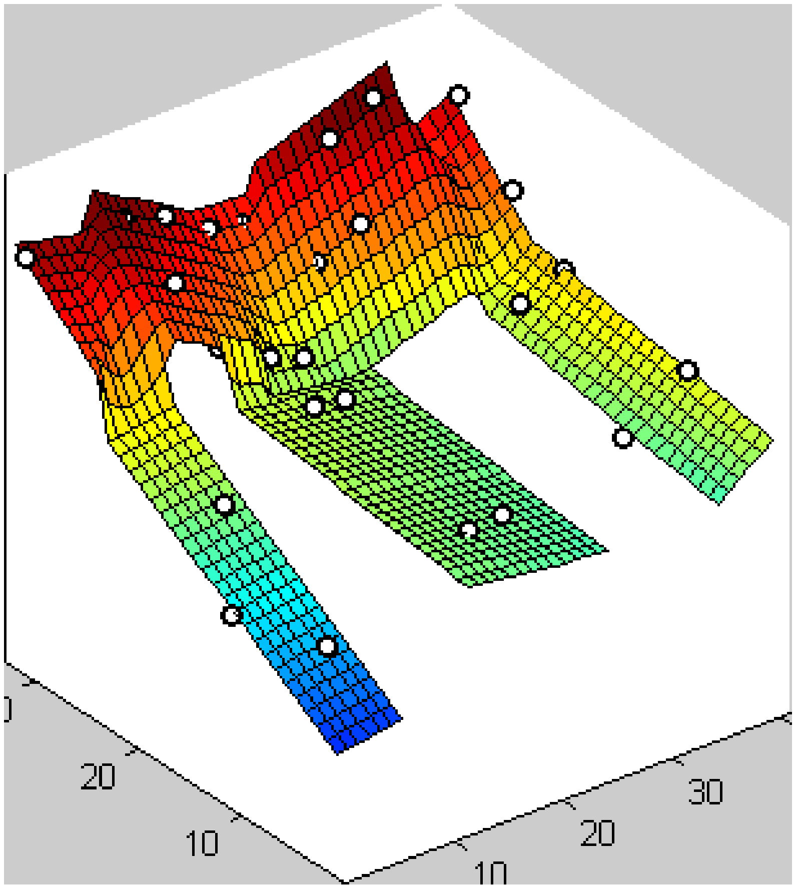


Reading a noun (vs verb)

[Rustandi et al., 2005]

Modeling sensor data





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

[Guestrin et al. ’04]

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 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…
- … 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, …
  - PAC-analysis, SRM, …
- Computational **power** is available
  - networks of fast machines
- Budding **industry** in many application areas
  - market analysis, adaptive process control, decision support, …
- Alberta Ingenuity Centre for Machine Learning

Questions

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

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

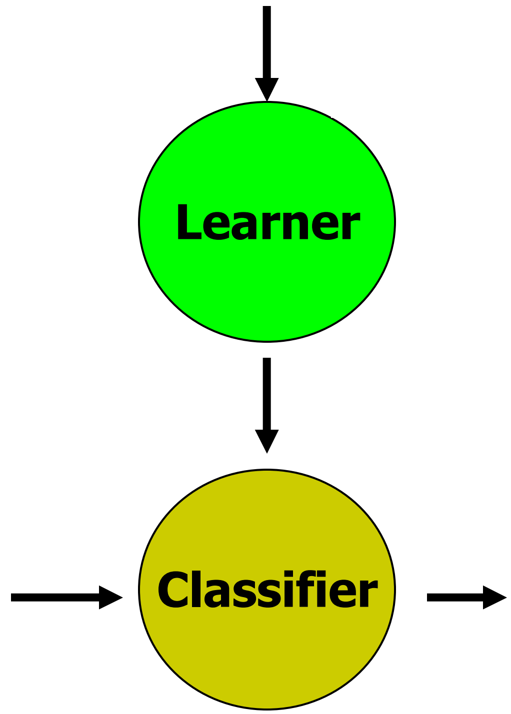
Statistics ≡

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

Training a Classifier

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

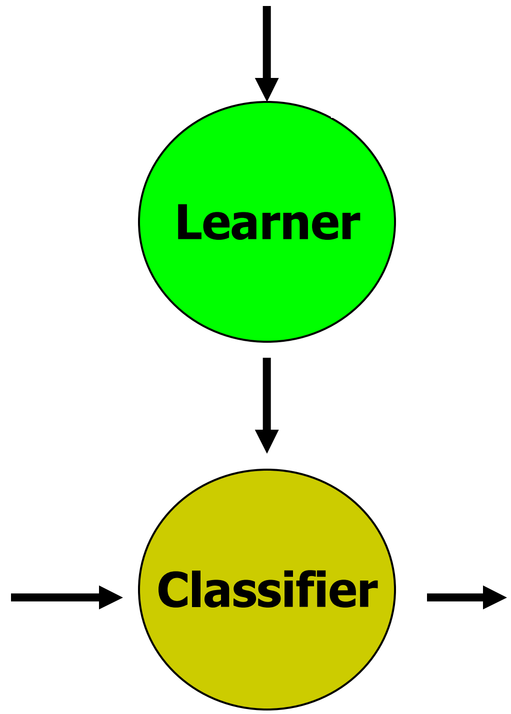


Width	Press.	Sore-Throat	…	Light
32	90	N	…	Pale

type
bass

Training a Regressor

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



Width	Size	Eyes	…	Light
32	90	N	…	Pale

size
19

Classification

- Input: "feature list"
  Output: "label"
  - Features can be *symbols*, *real numbers*, …
    - [age $\in \Re^+$, height $\in \Re^+$, weight $\in \Re^+$, gender$\in${M,F}, hair_colour, …]
  - Labels come from a (small) discrete set
    - L = { Icelander, Canadian }
- Output: *discriminant* function,
  mapping feature vectors to labels.
- We can learn this from data, in many ways.
    - ([27, 172, 68, M, brown, …], Canadian)
    - ([29, 160, 54, F, brown, …], Icelander)
    - …
- We can use it to *predict* the label of a new instance.
  - How good are our predictions?

Regression

- Input: “feature list”
  Output: “response”
  - Features can be symbols, real numbers, etc…
    - [age, height, weight, gender, hair_colour, …]
  - Response is *real-valued.*
    - *life_span* $\in \Re^+$
- We need a *regression* function that maps feature vectors to responses.
- We can learn this from data, in many ways.
    - ([27, 172, 68, M, brown, …], 86)
    - ([29, 160, 54, F, brown, …], 99)
    - …
- We can use it to *predict* the response of a new instance.
  - How good are our predictions?

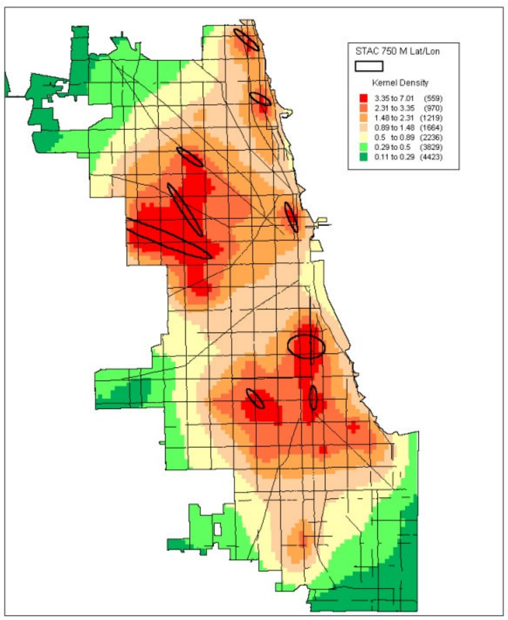
Pause:
Classification vs. Regression

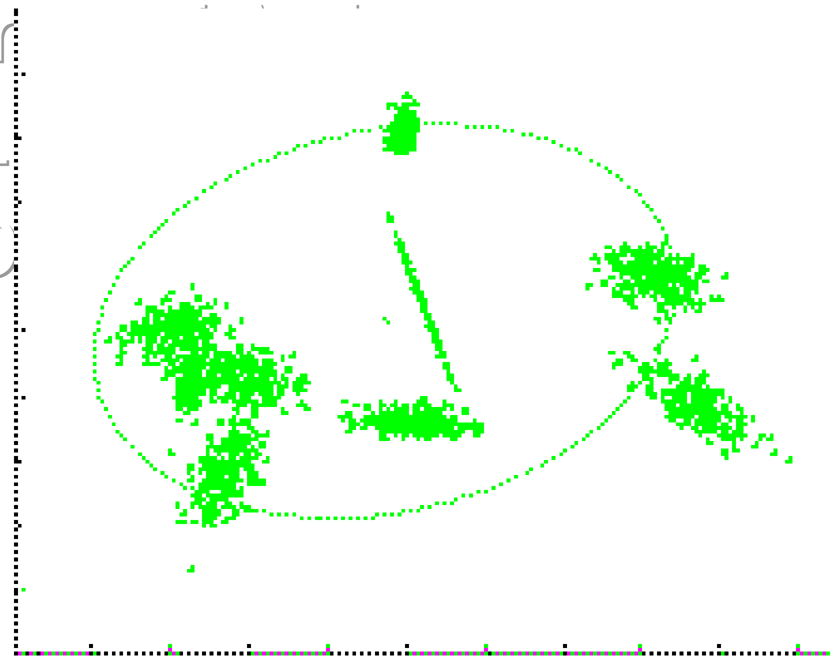
- Same: “Learn a function from labeled examples”
- Difference: Domain of label: small set vs $\Re$
  Why make the distinction?
  - Historically, they have been studied separately
  - The label domain can significantly impact what algorithms will work or not work

- Classification
  - “Separate the data”
- Regression
  - “Fit the data”

Other Types of Learning

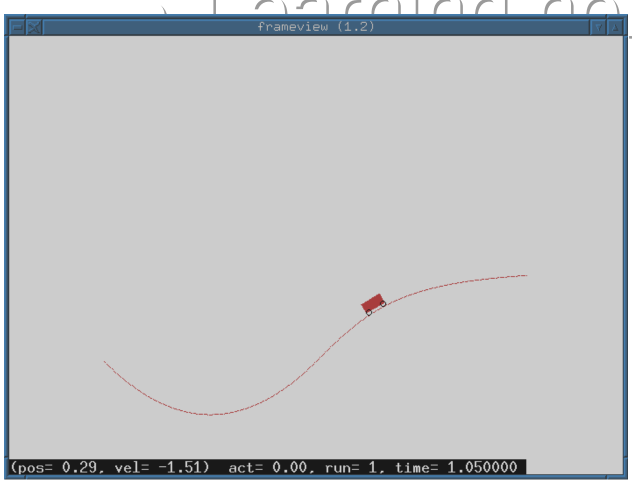
- Density Estimation
  - Learning Generative Model
  - Clustering

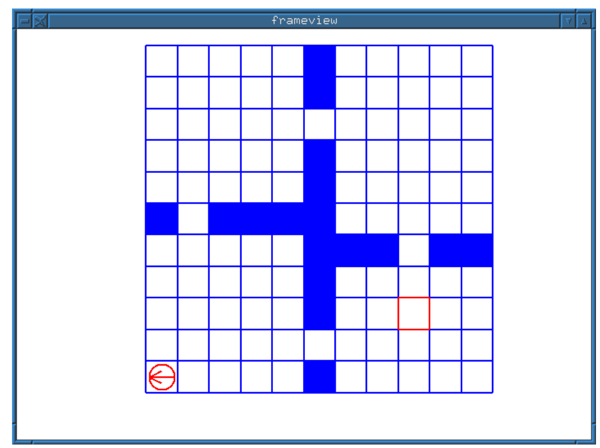




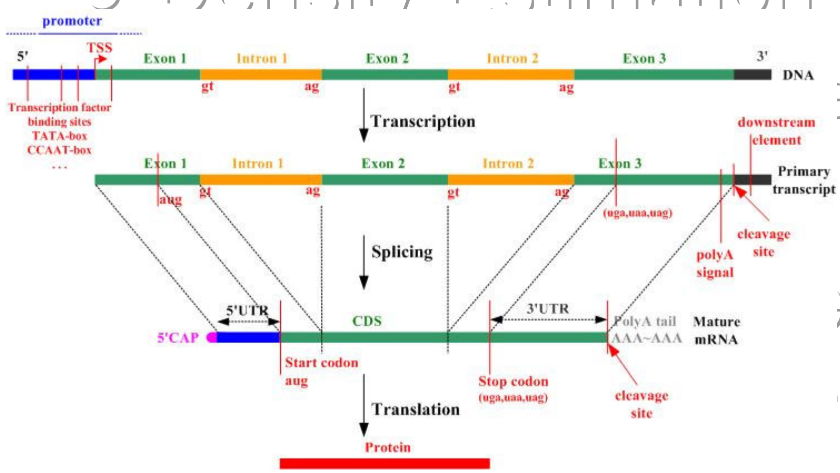
Other Types of Learning

- Density Estimation
  - Learning Generative Model
  - Clustering
- Learning Sequence of Actions
  - Reinforcement Learning
- Learning non-IID Data





Other Types of Learning



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

Other Types of Learning

- Density Estimation
  - Learning Generative Model
  - Clustering
- Learning Sequence of Actions
  - Reinforcement Learning
- Learning non-IID Data
  - Images
  - Sequences
  - …

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,
  - …

Relevant Disciplines

- Artificial intelligence
- Bayesian methods
- Computational complexity theory
- Control theory
- Information theory
- Philosophy
- Psychology and neurobiology
- Statistics
- ...

Summary

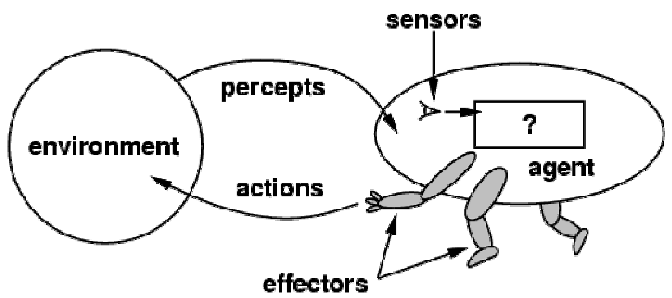
- Machine Learning is a **mature field**
  - solid theoretical foundation
  - many effective algorithms
- ML is *crucial* to large number of important **applications**
  - BioInformatics, WebReDesign, MarketAnalysis, Fraud Detection, …
- Fun: Lots of intriguing open questions!
- **Exciting time for Machine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

Unsupervised Learning

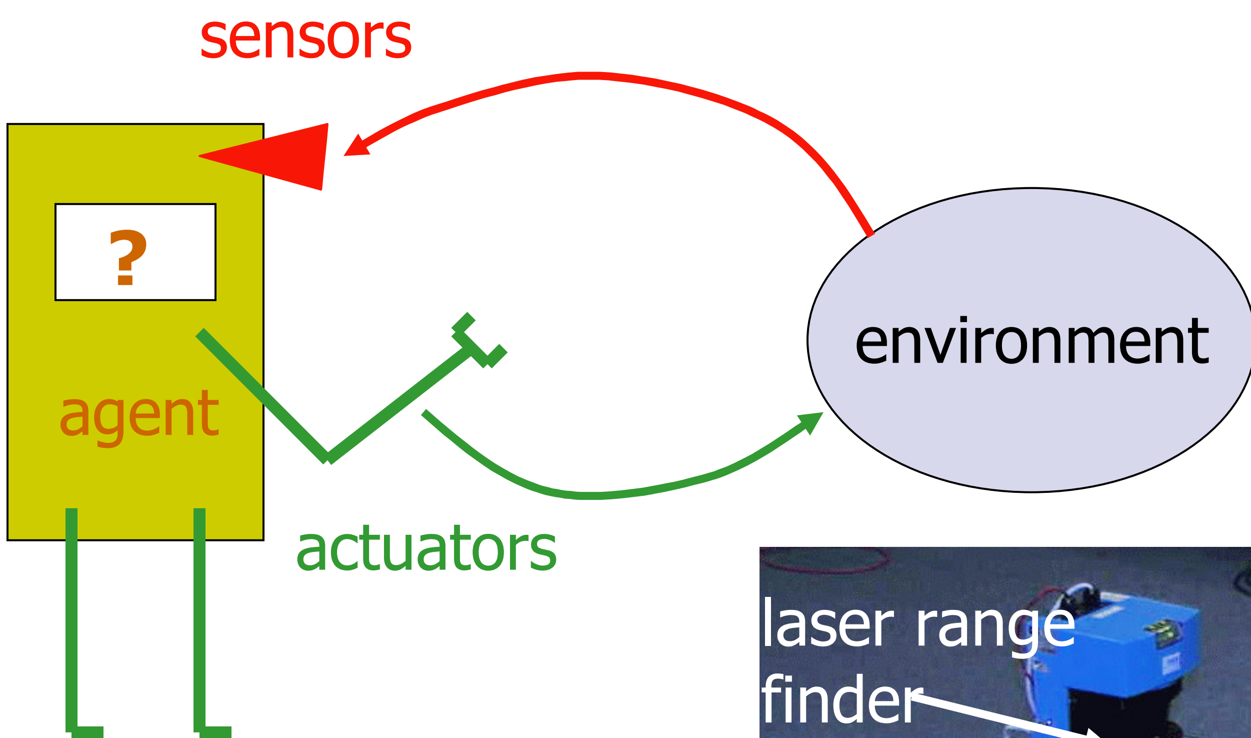
- Take clustering for example.
- Input: “features”   Output: “label”
  - Features can be symbols, real numbers, etc…
    - [age, height, weight, gender, hair_colour, …]
  - Labels are **not** given.
    (Sometimes |L| is known.)
- Each label describes a *subset of the data*
  - Clustering: group together examples that are “close”
    - … need to define “close”
    - Labels = “cluster centres”
- Here: cluster can be the end result
  (Not classification)
  - Subjective $\Rightarrow$ Evaluation is difficult

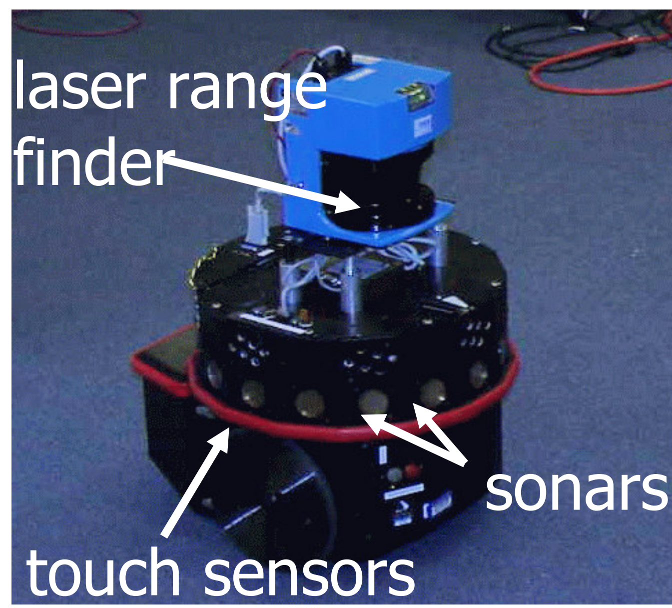
Reinforcement Learning

- Input: “observations”, “rewards”
  Output: “actions”
  - Observations may be real or discrete
  - Reward $\in \Re$
  - Actions may be real or discrete
- Think of …
  agent (“robot”) interacting with its environment
- On-going interaction
  At each time,
  - agent observes “observations”
  - Selects an actions
  - Receives a reward
- Agent can use Reinforcement Learning
  to improves its performance
  (ie, selecting actions that lead to better rewards)
  by analyzing past experience



Notion of an Agent

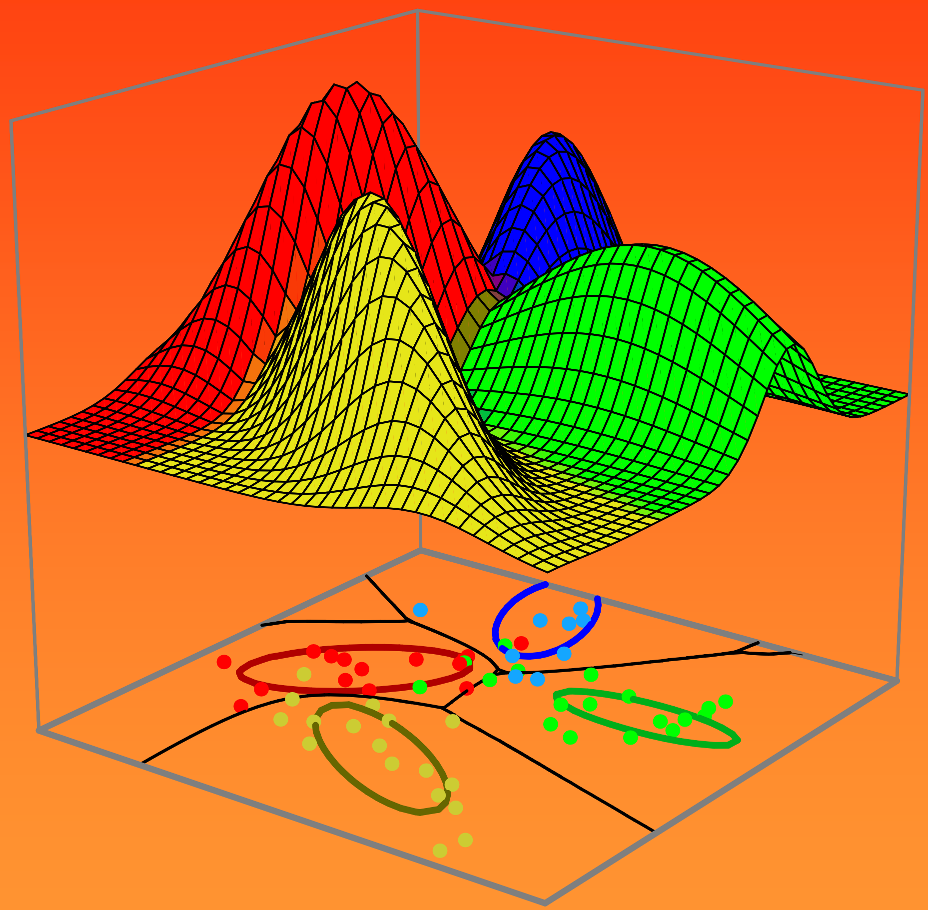




Source: robotics.stanford.edu/~latombe/cs121/2003/home.htm

Conclusion

- Machine Learning has many challenging sub-problems
- These sub-problems have be solved for many real-world problems!
- Many fascinating unsolved problems still remain



Pattern Classification

All materials in these slides were taken from
Pattern Classification (2nd ed) **by R. O. Duda, P. E. Hart and D. G. Stork, John Wiley & Sons, 2000**
with the permission of the authors and the publisher