Machine Learning

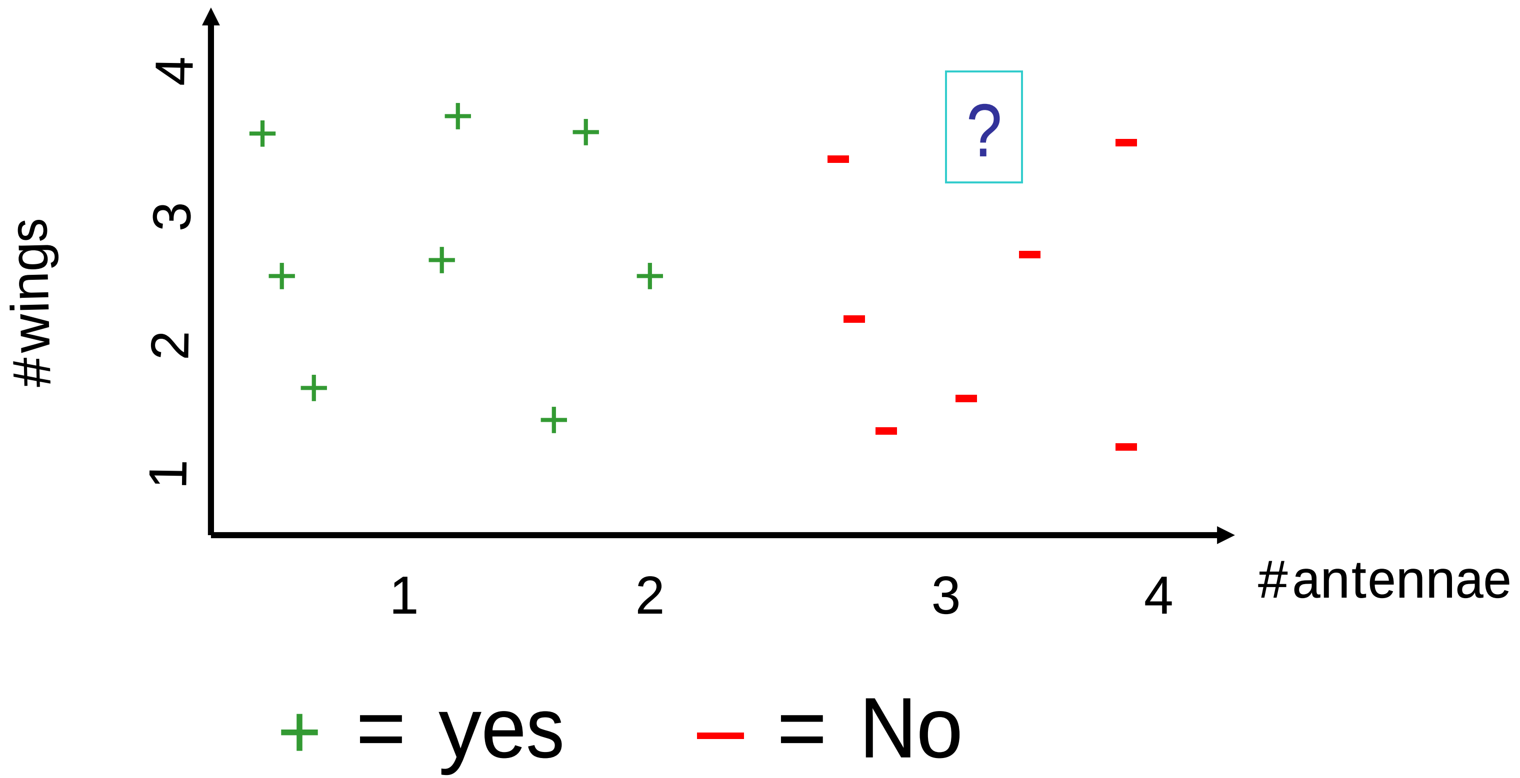
Russ Greiner
Alberta Ingenuity Centre for Machine Learning
Department of Computing Science
University of Alberta

Diagnosing Butterfly-itis

Data from Previous Patients

#wings	#antennae	...	nectar-orient.
2	1	...	++

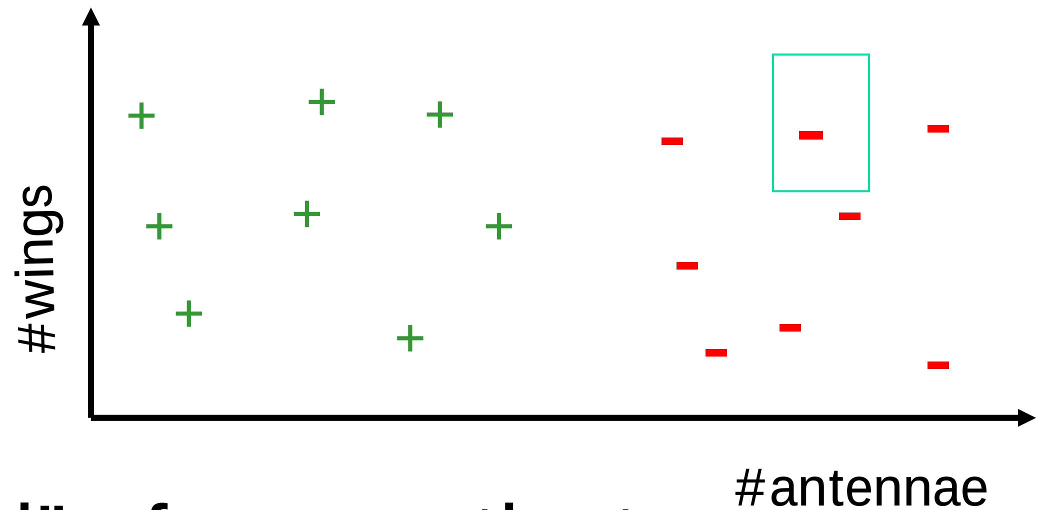
Visualizing Patient Data



- What about this new patient ?

This is is learning...

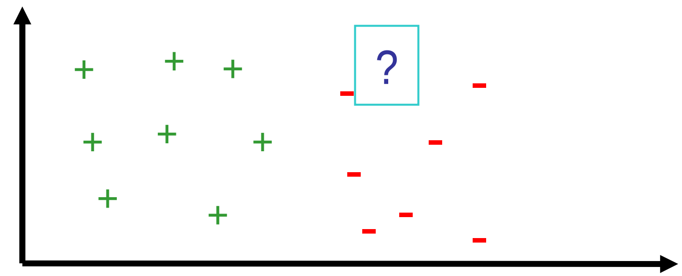
- Given data:



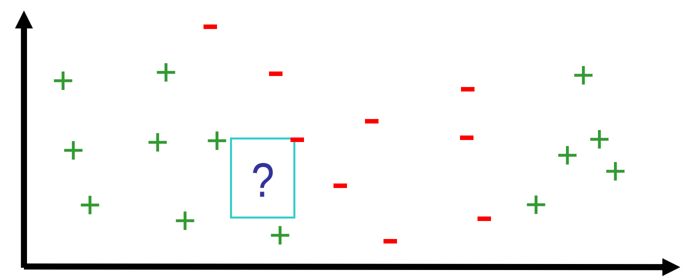
- Predicting “label” of new patient
  - Here: Negative “–” (not butterfly-it is)
- This is an ***EDUCATED GUESS***:
  - ... not based on post-mortem, definitive test, ...
  - use to decide on treatment, etc.

Challenges to Learning

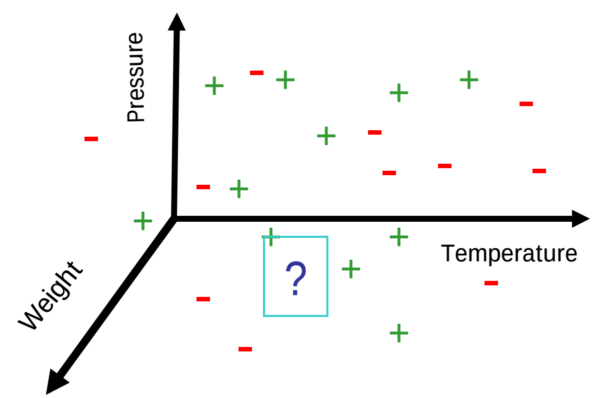
- Easy:



- Harder:



- High Dimension:



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...

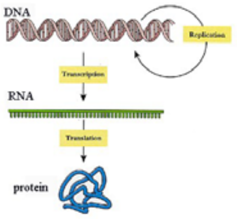
- to find ideal customers
  - Credit Card approval (AMEX)
    - Humans ≈50%; ML is >70% !

- 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)

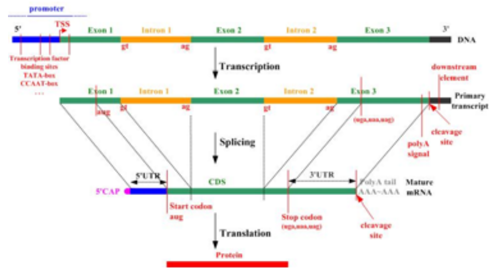
- to help win games
  - NBA (scouting)

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

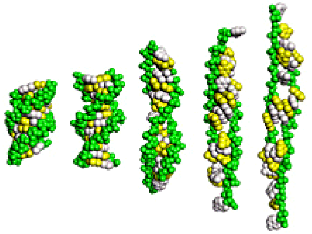


2: Sequential Analysis

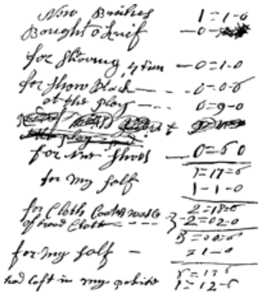
- **BioInformatics 1:** identifying genes
  - Glimmer [Delcher et al, 95]
  - identifies 97^{+}% of genes, automatically!



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

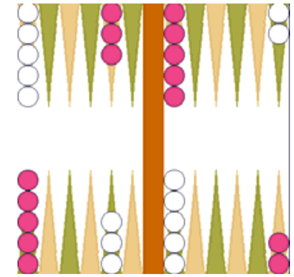


- **Recognizing Handwriting**

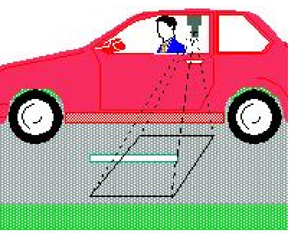


- **Recognizing Spoken Words**
  - **“How to wreck a nice beach”**

3: Control



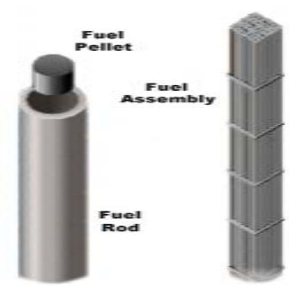
- **TD-Gammon** (Tesauro 1993; 1995)
  - World-champion level play by **learning** ...
  - by playing millions of games against itself!



- **Drive autonomous vehicles** (Thrun 2005)
  - DARPA Grand Challenge

- **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)



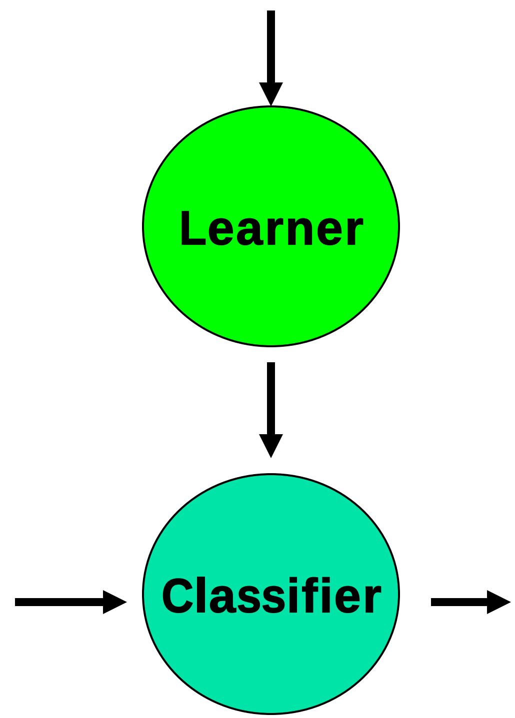
- Manufacture nuclear fuel pellets (Leech, 86)
  - Saves Westinghouse >$10M / year
- **Adaptive** agents / user-interfaces

Outline

- Successes
- Basic ideas
  - Foundations
  - Algorithms
  - Statistical Issues

Learning is ... Training a Classifier

Temp.	Press.	Sore Throat	...	Colour	diseaseX
35	95	Y	...	Pale	No
22	110	N	...	Clear	Yes
:	:			:	:
10	87	N	...	Pale	No



Temp	Press.	Sore-Throat	...	Color
32	90	N	...	Pale

diseaseX
No

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

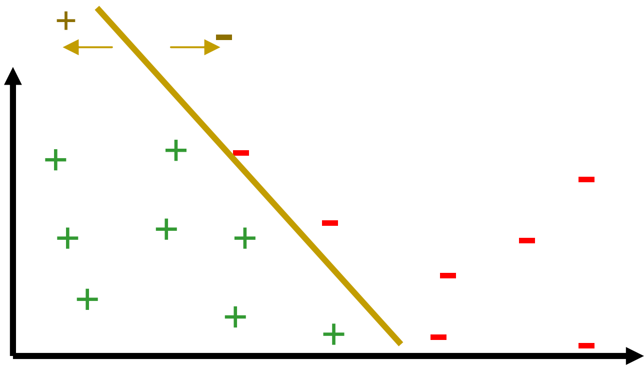
  Adaptive user interface...

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

Outline

- Successes
- Basic ideas
  - Foundations
  - Algorithms
    - Linear Separators
    - Artificial Neural Nets
    - Decision Trees
    - Nearest Neighbor, Naïve Bayes, Support Vector Machines, Reinforcement Learning, ...
  - Statistical Issues

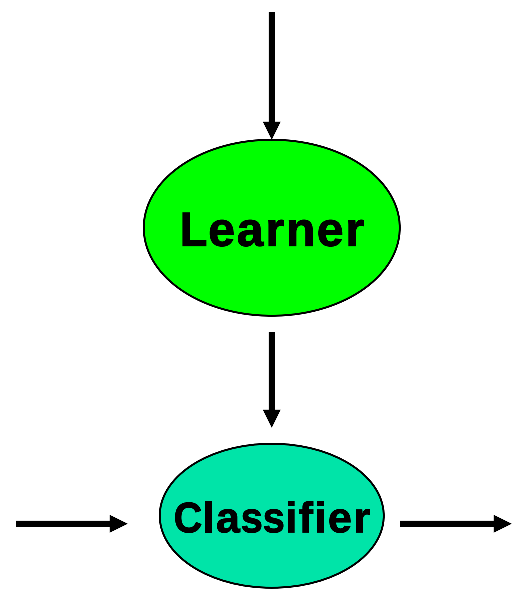


General Process

- Given “labeled data”

Temp.	BP.	Sore Throat	…	Colour	diseaseX
35	95	Y	…	Pale	No
22	110	N	…	Clear	Yes
:	:			:	:
10	87	N	…	Pale	No

- Learn CLASSIFIER, that can predict label of *NEW* instance

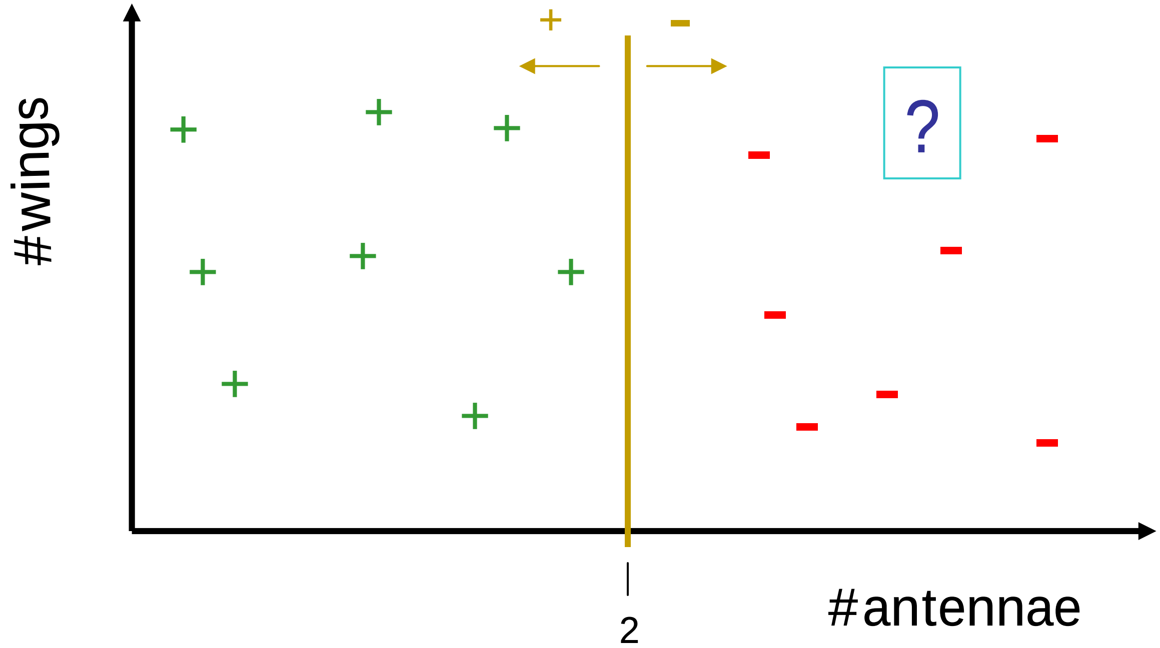


Temp	BP	Sore-Throat	…	Color	diseaseX
32	90	N	…	Pale	?

diseaseX
No

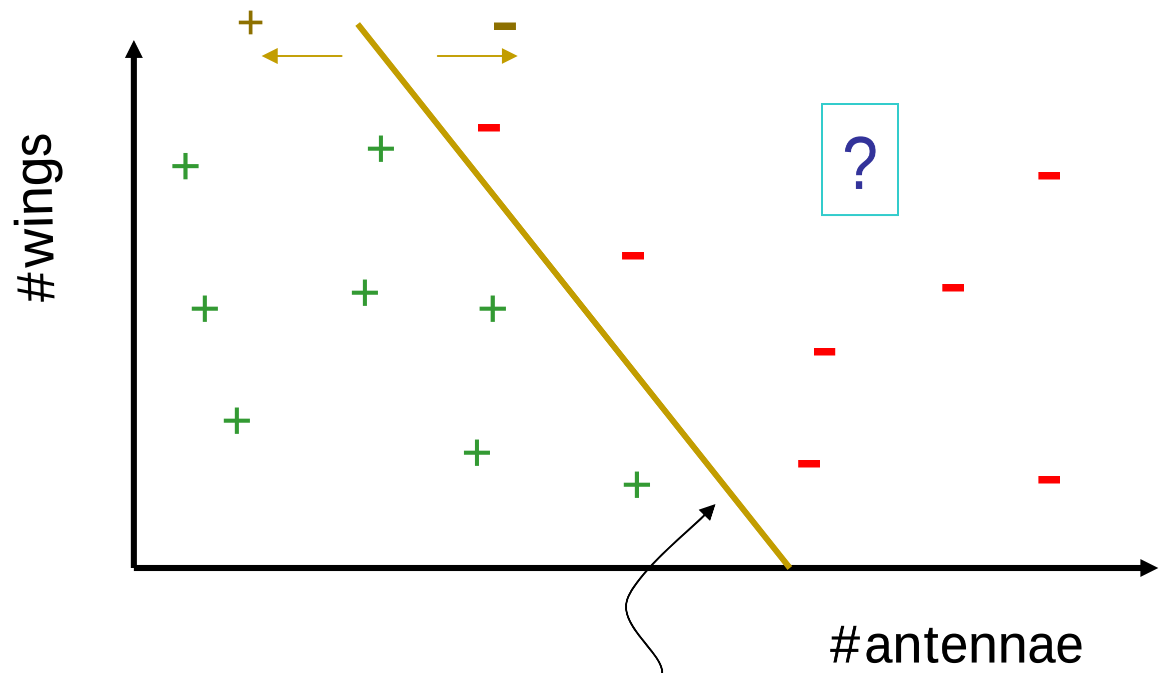
Alg 1: Linear Separators

- Draw "separating line"



- If #antennae $\leq$ 2, then butterfly-itis
- So ? is Not butterfly-itis.

Can be “angled”...



$2.3 \times \#w + 7.5 \times \#a + 1.2 = 0$

- If $\boxed{2.3} \times \#Wings + \boxed{7.5} \times \#antennae + \boxed{1.2} > 0$
  then butterfly-itis

Linear Separators, in General

- Given data (many features)

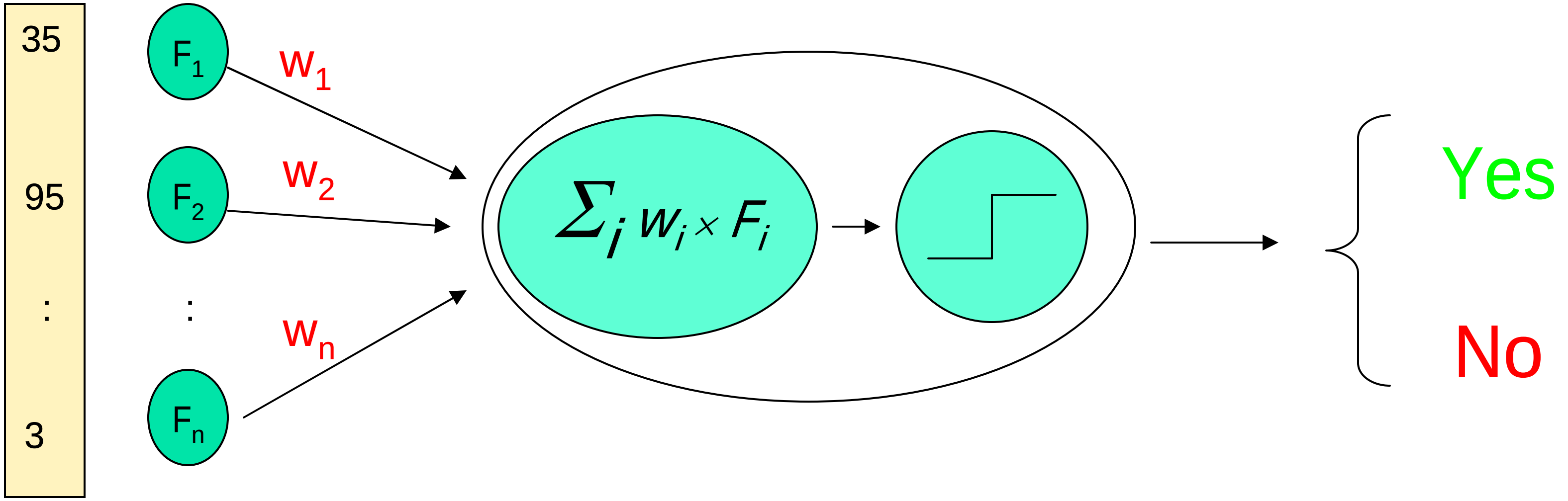
F_1	F_2	...	F_n	Class
35	95	...	3	No
22	80	...	-2	Yes
:	:		:	:
10	50	...	1.9	No

- find "weights" $\{w_1, w_2, \ldots, w_n, w_0\}$ such that

$$\boxed{w_1} \times F_1 + \ldots + \boxed{w_n} \times F_n + \boxed{w_0} > 0$$

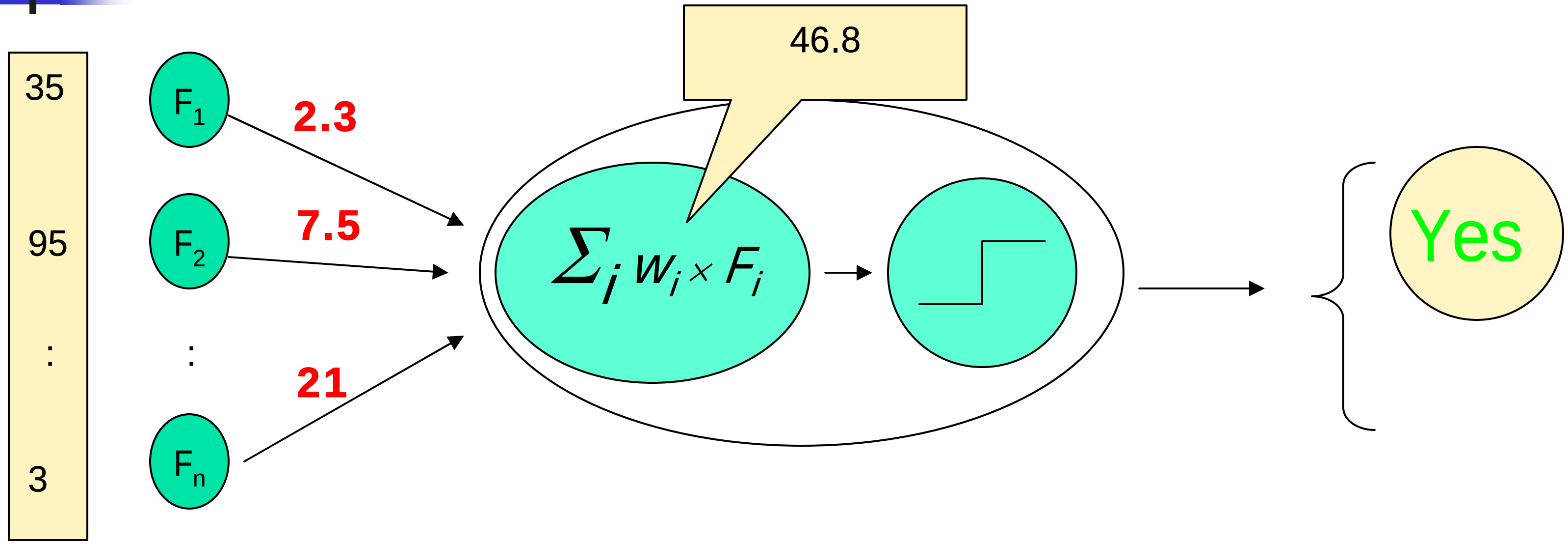
means $\text{Class} = \text{Yes}$

Linear Separator



Just view $F_0 = 0$, so w_0 ...

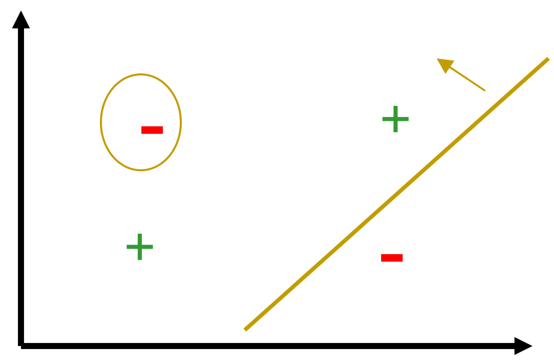
Linear Separator



- Challenge:
  - Given labeled data, find “correct” $\{w_i\}$
- “Perceptron”

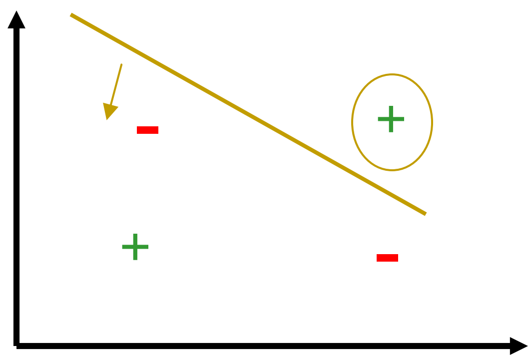
Linear Separators – Facts

- GOOD NEWS:
  - If data is linearly separated,
  - Then FAST ALGORITHM finds correct $\{w_i\}$!
- But...



Linear Separators – Facts

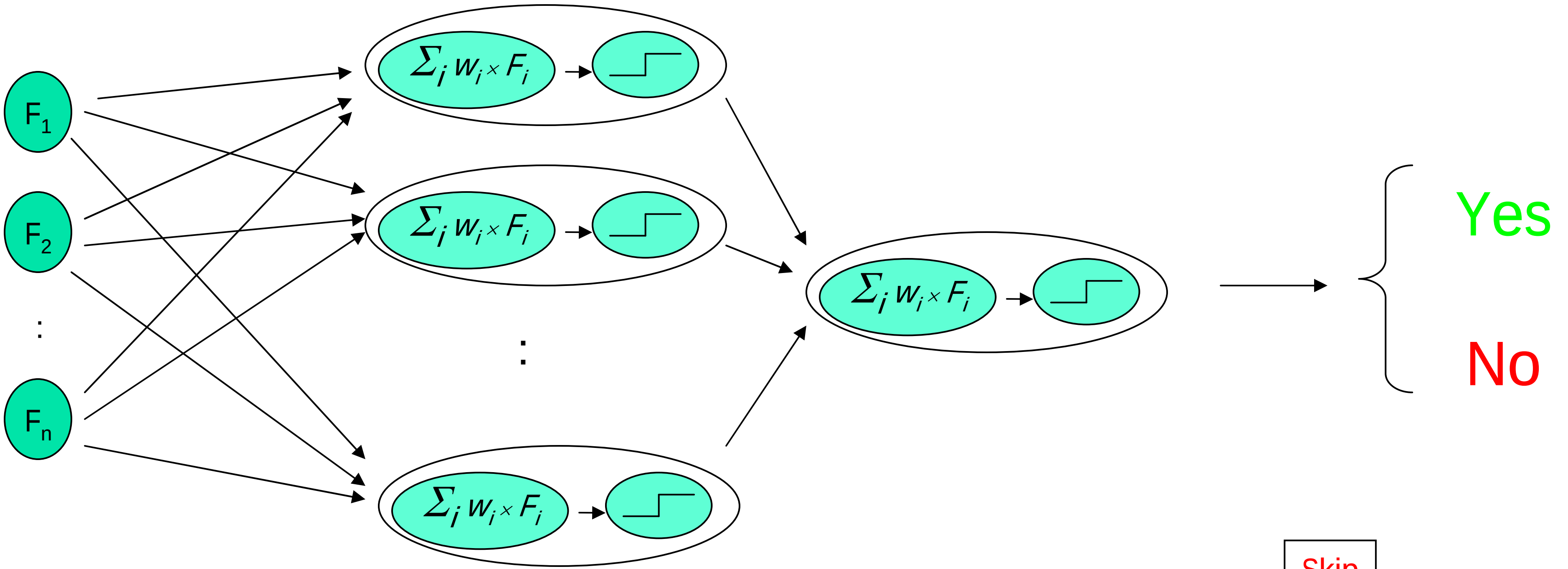
- GOOD NEWS:
  - If data is linearly separated,
  - Then FAST ALGORITHM finds correct $\{w_i\}$!
- But...



- Some "data sets" are
  NOT linearly separatable!

Alg 2: Artificial Neural Nets

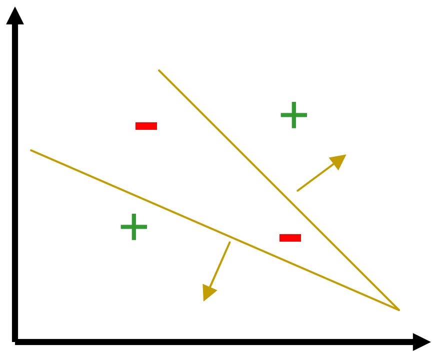
- Why not use *SET* of **connected Linear Separators**?

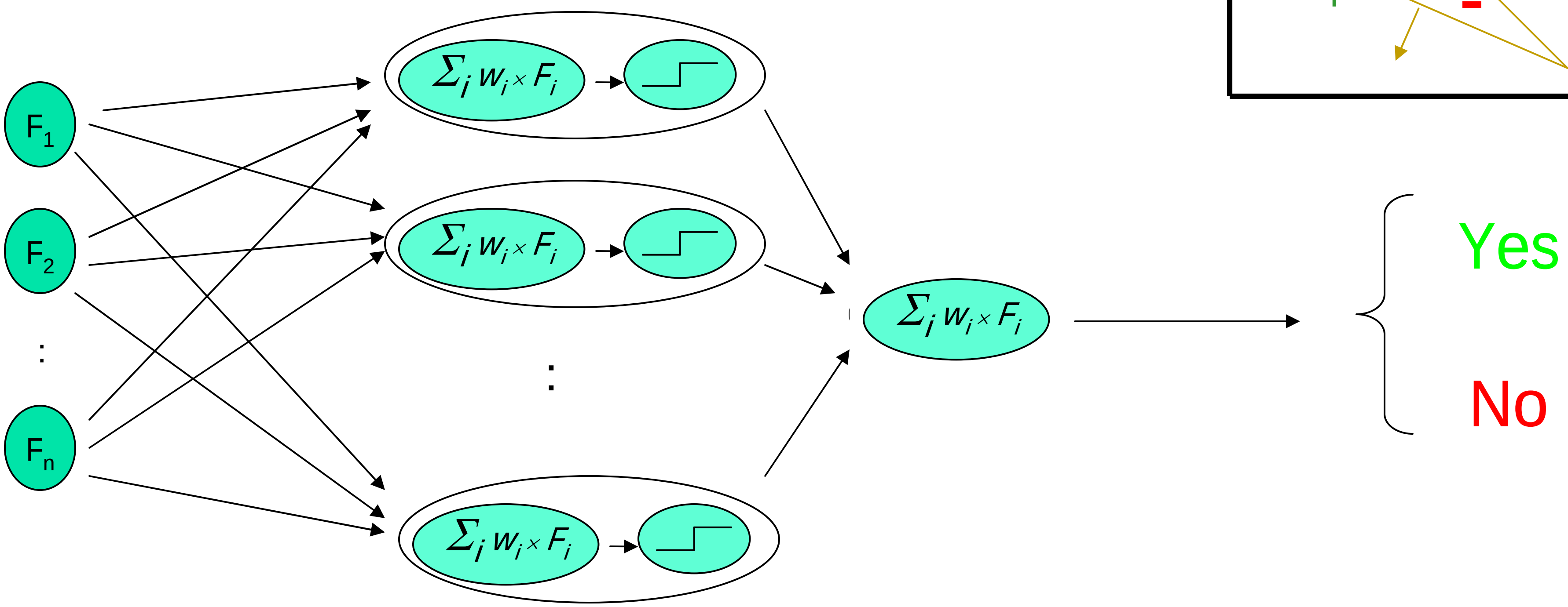


Skip

Artificial Neural Nets

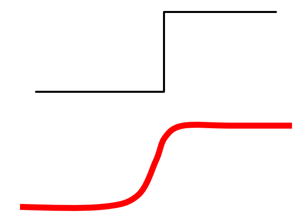
- Can Represent *ANY* classifier!
  - w/just 1 "hidden" layer...
  - in fact...

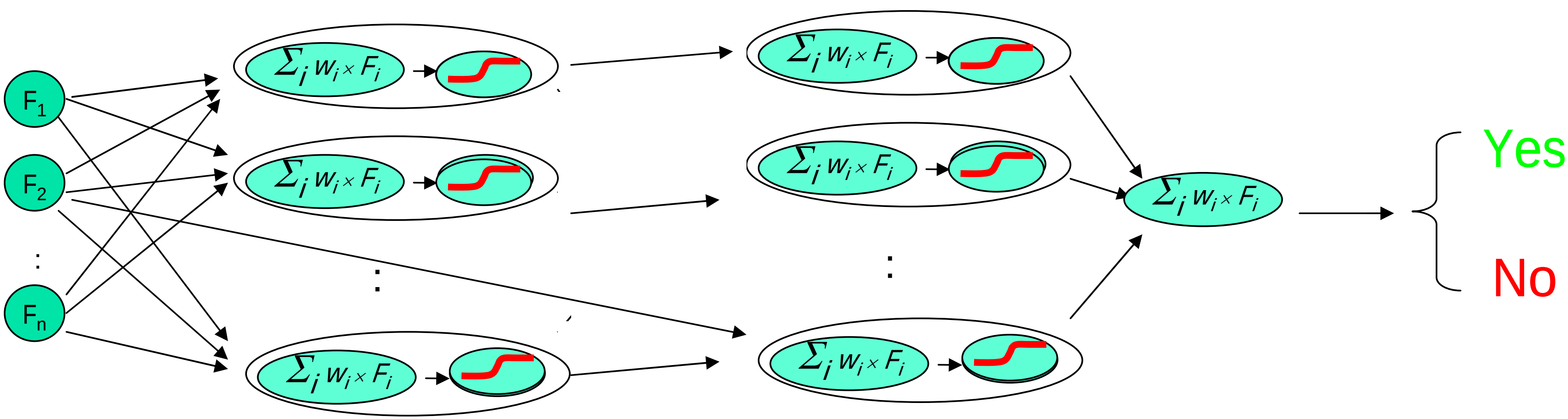




ANNs: Architecture

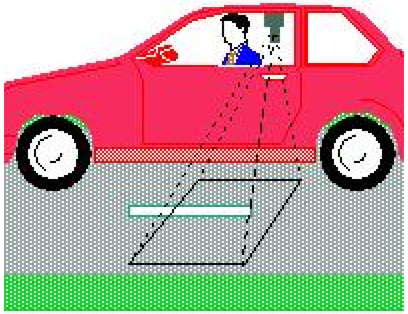
- Different # of layers
- Different structures
  - what's connected to what..
- Different "squashing function"





Uses of Artificial Neural Nets

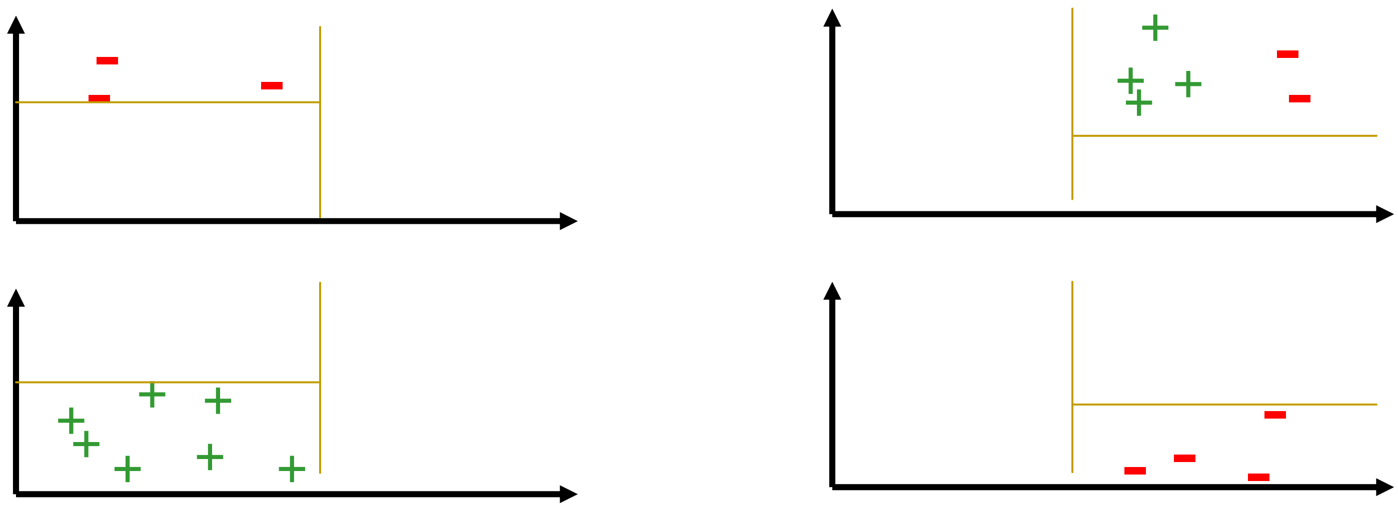
Learning to ...



- drive a car
- assess credit risk
- pronouncing words (NETtalk)
- recognize handwritten characters
- control plant
- ...

Algorithm 3: Decision Trees

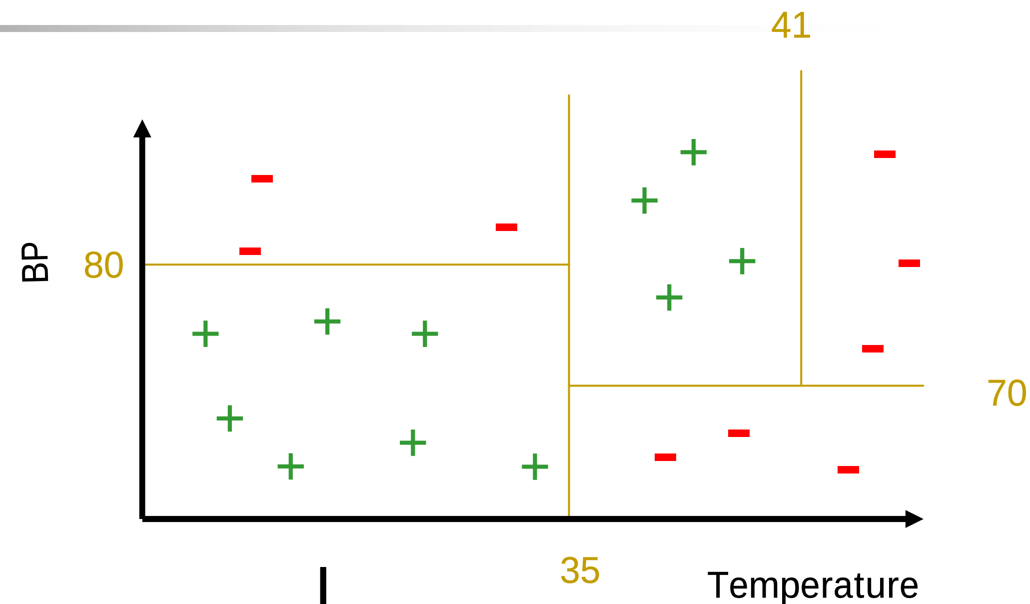
- Given data, decide on best *first* split



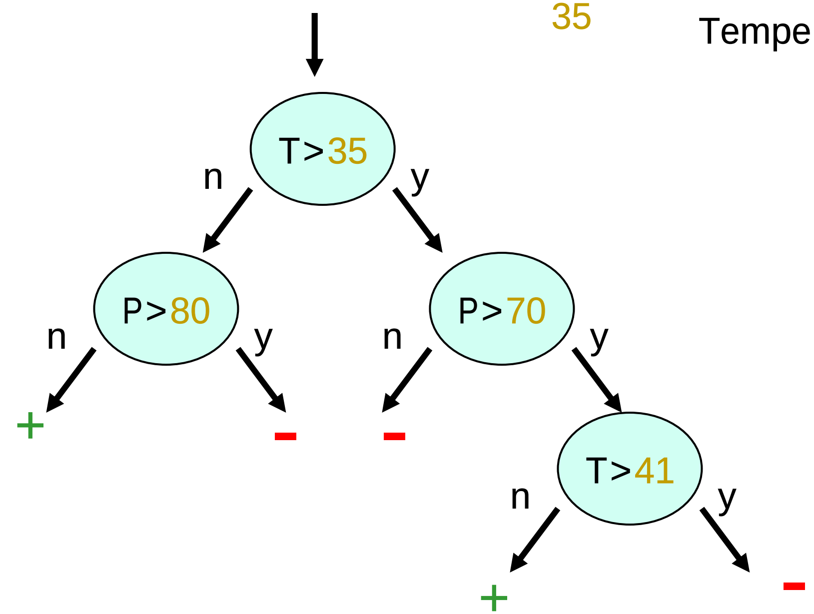
- Then consider each subset of data:
  - decide on its best split
- Recur… until "purity"

Alg 3: Decision Trees

- Partitioned data:



- “Hierarchical Split”
  - Divide and conquer



Issues $\Rightarrow$ Demo

- Issues:
  - How to split?
  - When to stop?
  - Avoid overfitting
  - Real vs Discrete

- AIxploratorium!
  http://www.cs.ualberta.ca/~aixplore

Other Algorithms

- Nearest Neighbor
- Support Vector Machines
  - Find *BEST* line between + and –'s
- Naïve Bayes
  - Probabilistic model:
    What is *chance* that datapoint is + vs – ?
- Learning "Ensembles"
  - Ways to combine "ok" classifiers, to be better
  - Boosting, Bagging, Stacking, ...
- More than just + vs – ...
  - {Ok, MildSick, AverageSick, VerySick}
  - Real values $\Re$

Outline

- Successes
- Basic ideas
  - Foundations
  - Algorithms
  - Statistical Issues
    1. Goal of learning
    2. Why should Learning work?
    3. How much data is needed?
    4. Overfitting
    5. Computational Efficiency
    6. Imbalanced data (fraud detection)
    7. Non-IID tuples (stock market, temporal)

1. Goal of Learning?

	F$_1$	F$_2$	...	F$_n$	Class
a =	35	95	...	3	No
b =	22	80	...	-2	Yes
d =	10	50	...	1.9	No

- If goal of learning is just

  score well on *training data* ...
- *Trivial:* just memorize data!

  { a is No b is Yes d is No }
- Instead: want to do well on
  - *NEW UNSEEN data*
  - On e =

F$_1$	F$_2$	...	F$_n$	Class
32	90	...	-3	??

- How can learning possibly succeed?

2. Why should Learning work?

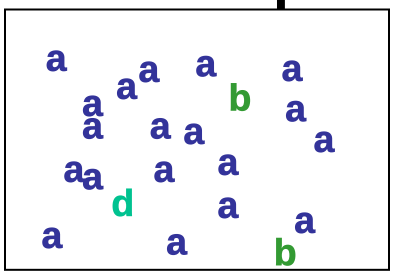
- *Rare is rare*
  - If patient type is *common*, then it is in sample
    - If in sample, classifier “gets” it
  - If patient type NOT *common*, then ... so what?
    - Classifier will be wrong, but penalty is small
- Overfitting can be prevented
- More data is better

Skip details

Why should Learning work?

	F_1	F_2	...	F_n	Class
a =	35	95	...	3	No
b =	22	80	...	-2	Yes
d =	10	50	...	1.9	No

- Overall Population



- Draw sample S = $\boxed{\text{a a a b a a}}$
- Learn classifier C that does well on S:
  - As S includes a b,

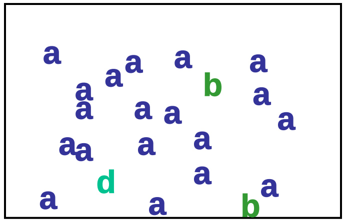
$$\left\{ \begin{array}{l} C(\text{a}) = \text{No} \\ C(\text{b}) = \text{Yes} \end{array} \right\}$$

- Notice d not is S

$\Rightarrow \quad C(\text{d}) = ??$

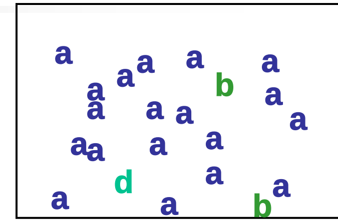
How good is Classifier C ?

$$\left\{ \begin{array}{l} C(\text{a}) = \text{No} \\ C(\text{b}) = \text{Yes} \end{array} \right\}$$

- To evaluate C
  - Draw new patient, $x \in$ 

  - Compute $C(x)$
  - Correct?
- Given true distribution,
  - *expect* $x = \text{a}$ ... or x = b
    - Here: $C(x)$ is correct!
  - Otherwise, $C(x)$ may be wrong.
    - But this is rare!

Why should Learning work?



Consider a new patient, *x* ...

1. If *x* occurs a LOT   P(*x*) >> 0
   - *x* probably appears in *S*
   - As *C* does well on *S*,
     C gives correct answer on *x*

a a a b a a

2. If *x* occurs rarely   P(*x*) ≈ 0
   - doesn't matter if *C* is wrong!

- Even good classifiers are wrong occasionally...

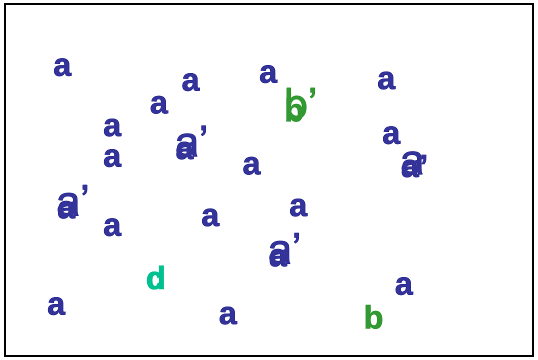
Populations

- Train a "Feline classifier" *FC* using
  - Pets in my neighborhood,
- *FC* should do well on
  - household cats +
  - household dogs –

- *FC* will probably be WRONG wrt
  - Tigers
- Not surprising: *FC* was NOT trained on them!

Similar Patients...

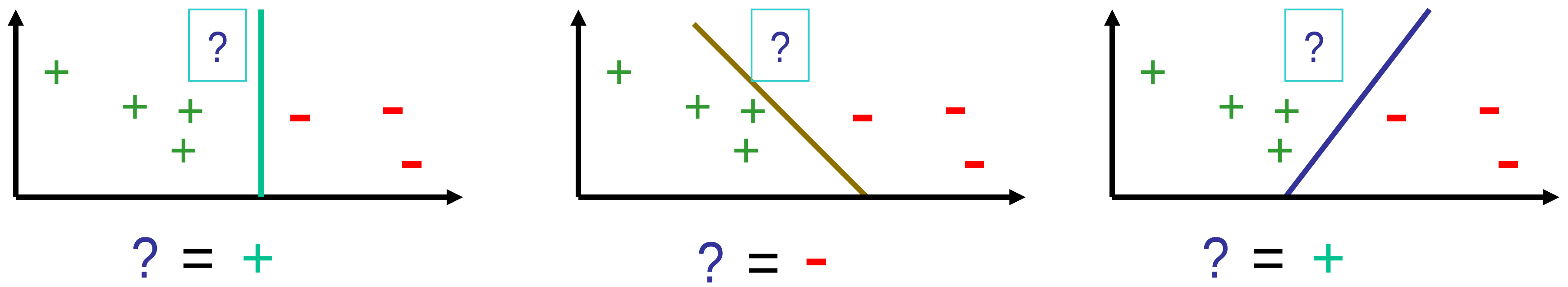
	F_1	F_2	...	F_n	Class
a =	35	95	...	3	No
a’ =	36	95	...	2.1	No
b =	22	80	...	-2	Yes
b’ =	22	78	...	-2.3	Yes
d =	10	50	...	1.9	No



- So far: assume many IDENTICAL patients
  - Same values for each feature
- More realistic: *Similar* patients...
- Same idea:
  - if need to classify x, and $x \sim u$ where $u \in S$,
  - then $C(u) \approx C(x)$ and probably correct ...

3. How much training data?

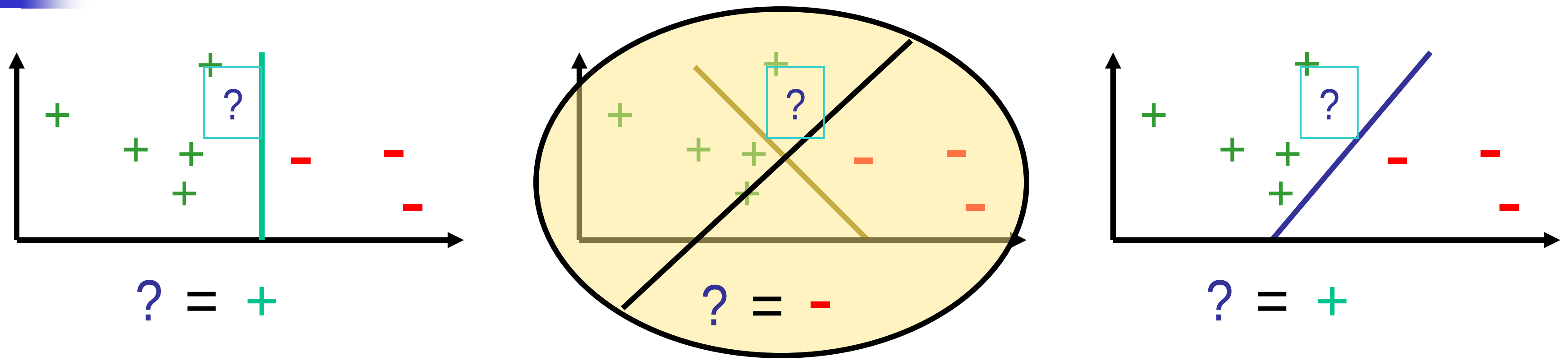
- What is best linear separator for...



- Makes a difference: what is "?" ?

- Learning gets easier with more training data...

More data helps...



- Suppose next training point is...
- Eliminates 2nd option...
  - Leaving only ? = +

Learnability Theory

Can **QUANTIFY** how many *training instances* are needed, as function of

- Hypothesis space
  - Linear Separators, Decision Trees, ...
- Accuracy required
- Chance of being completely wrong

- (Think of Hypothesis Testing...)

4. “Overfitting”

- Spse we used the WRONG features:
  - whether birthday was odd/even,
  - whether SSN was odd/even
  - whether car license odd/even
  - ...

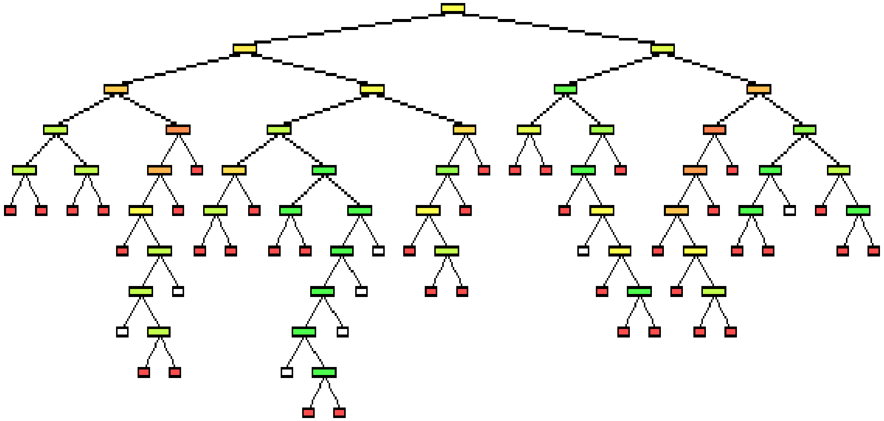
- Here: NO correlation between
  - butterfly-itis and
  - any (combination) of feature
- Best classifier:
  - Ignore features; just use majority class

Example – continued

- 25% have butterfly-itis
- ½ of patients have $F_1 = 1$
  - Eg: “odd birthday”
- ½ of patients have $F_2 = 1$
  - Eg: “even SSN”
- ... for 10 features
- Decision Tree results
  - over 1000 patients (using these silly features) ...

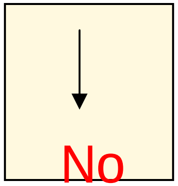
Decision Tree Results

- Standard decision tree learner:



- Error Rate:
  - Train data: 0%
  - New data: 37%

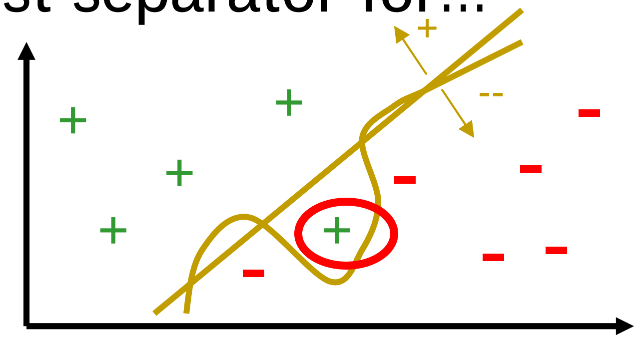
- Optimal decision tree:



- Error Rate:
  - Train data: 25%
  - New data: 25%

Overfitting

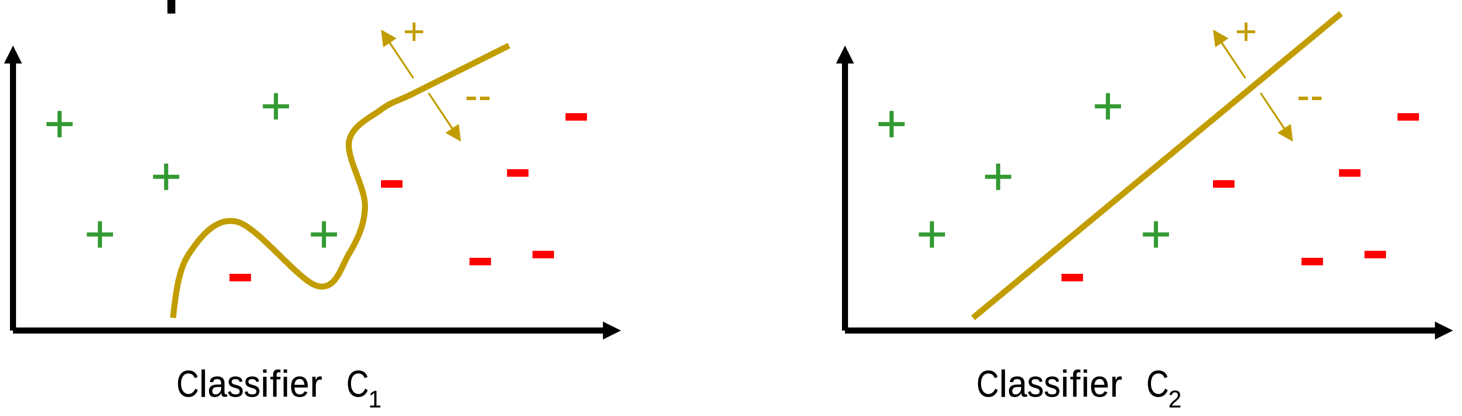
- Some features are not helpful
- Data often noisy
  - typos in recording, error in equipment, human error...
- What is best separator for...



- Sometimes:
  Appropriate to IGNORE details of training data
  - Here: one training data point is mislabeled !
- Simpler hypothesis often better classifier!
  - eg, LINEAR Separator

Overfitting

- Compare...



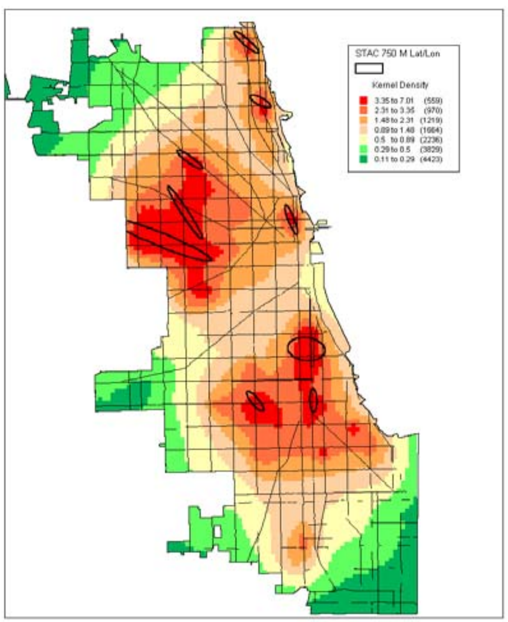
- C_1 appears better (on training data) than C_2, but C_2 is actually better
- *Overfitting* !

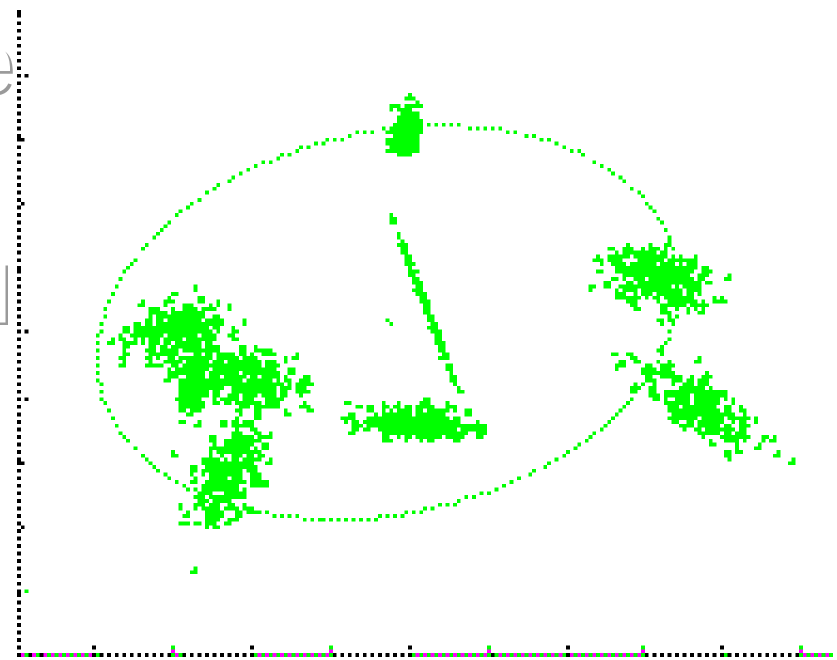
Outline

- Successes
- Basic ideas
  - Foundations
  - Algorithms
  - Statistical Issues
    1. Goal of Learning
    2. Why should Learning work?
    3. How much data is needed?
    4. Overfitting
    5. Computational Efficiency
    6. Imbalanced data (fraud detection)
    7. Non-IID tuples (stock market, temporal)

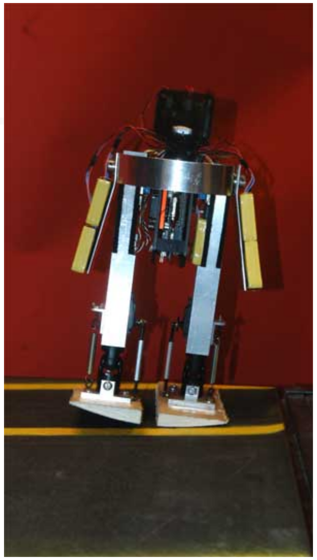
Other Types of Learning

- Density Estimation
  - Learning Generative Model
  - Clustering

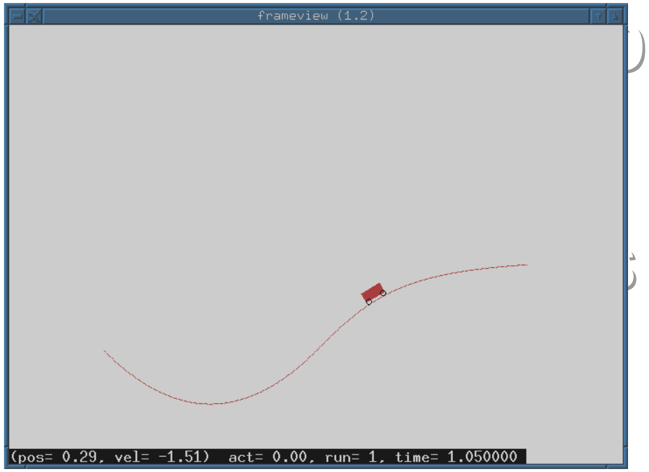




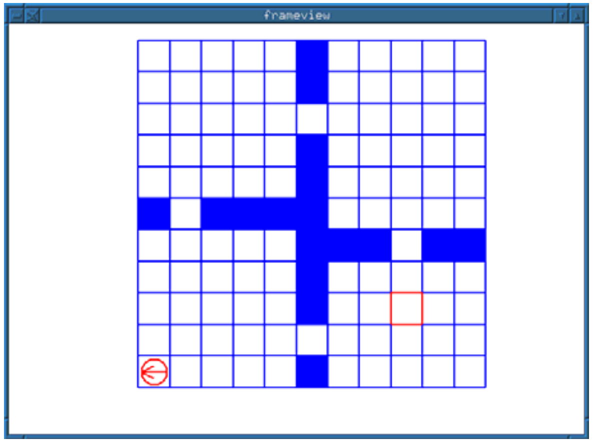
Other Types of Learning



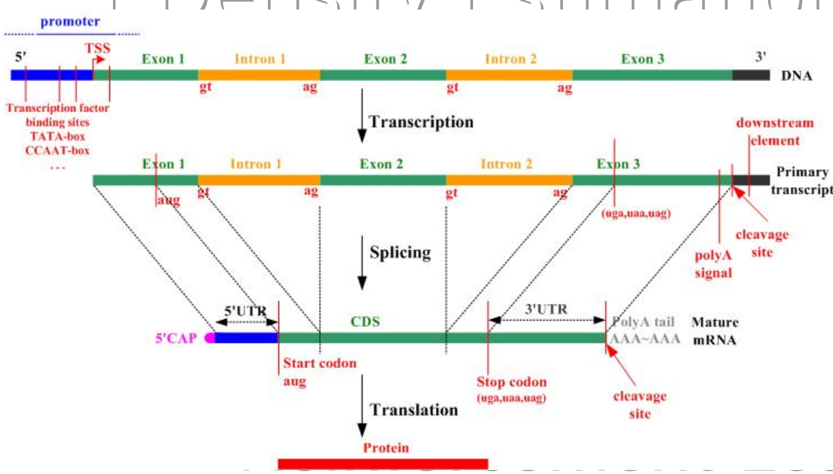
- Density Estimation
  - Learning Generative Model
  - Clustering
- Learning Sequence of Actions
  - Reinforcement Learning

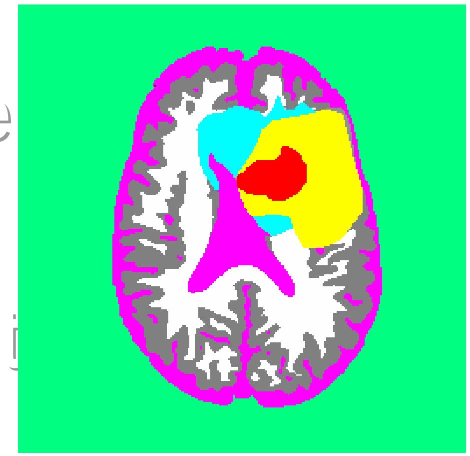


on-IID Data



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
  - ...

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**

Questions?