



Machine Learning

Russ Greiner

Alberta Ingenuity Centre for Machine Learning

Department of Computing Science

University of Alberta

Diagnosing Butterfly-itis

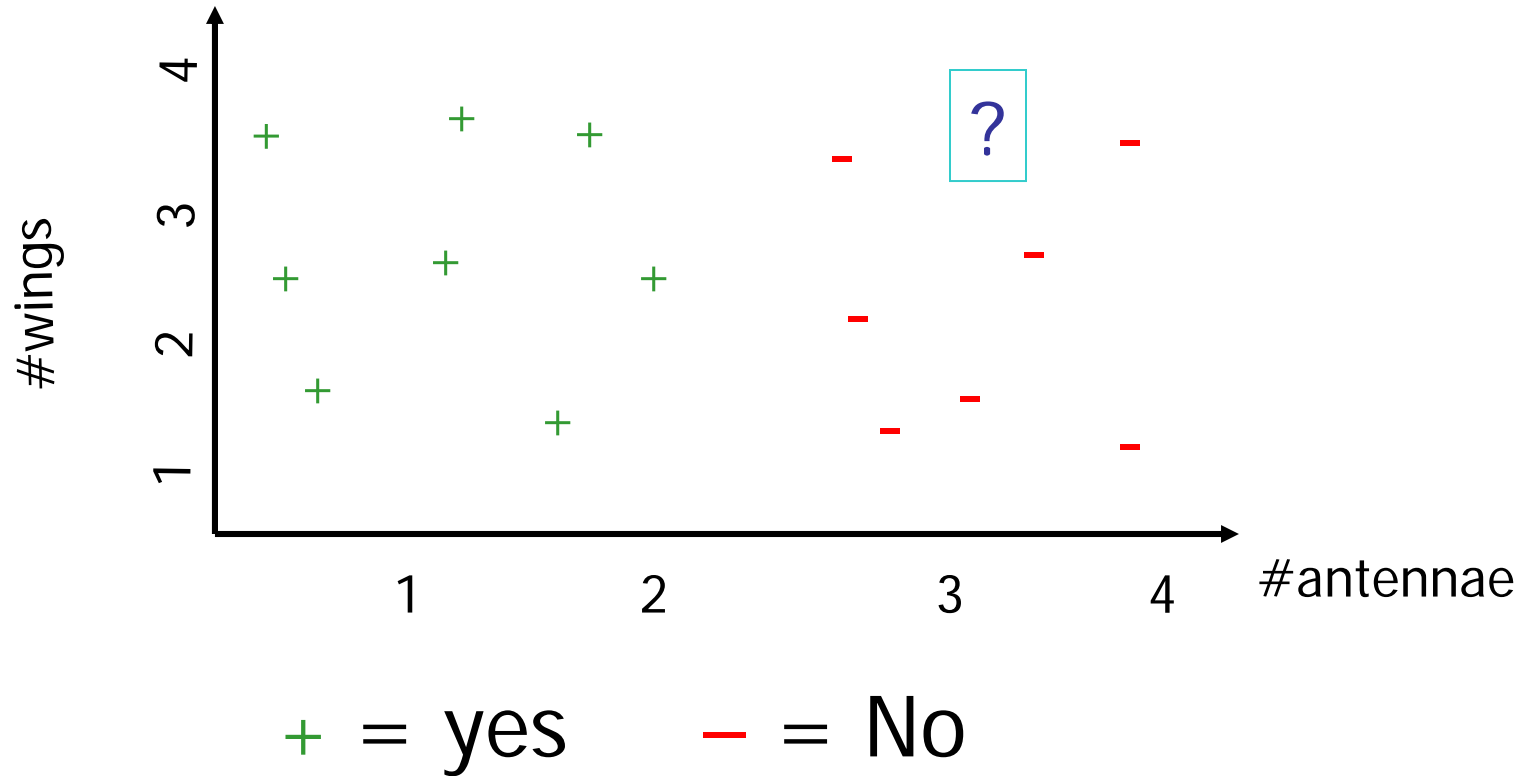




Data from Previous Patients

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

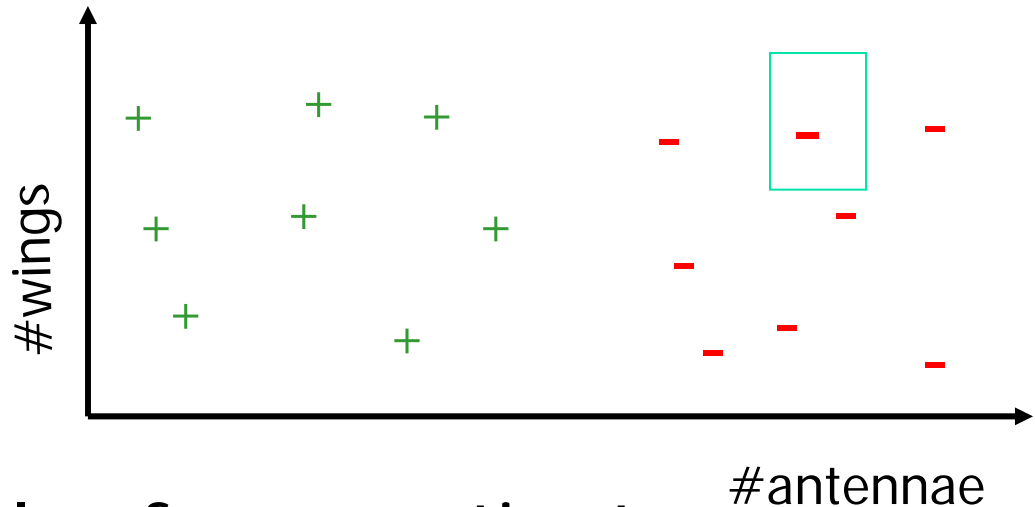
Visualizing Patient Data



- What about this new patient ?

This is learning...

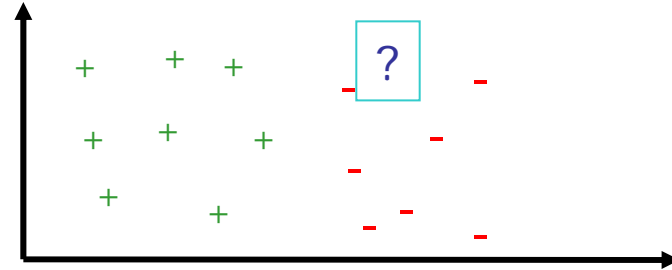
- Given data:



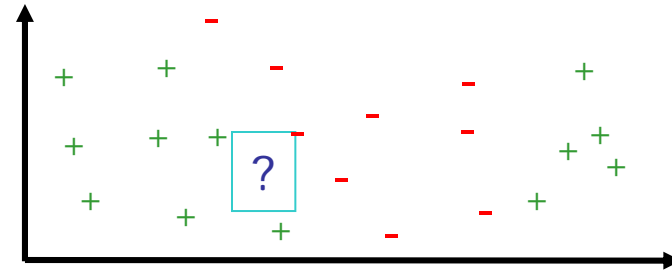
- Predicting "label" of new patient
 - Here: Negative "-" (not butterfly-it is)
- This is an ***EDUCATED GUESS***:
 - ... not based on post-mortem, definitive test, ...
 - use to decide on treatment, etc.

Challenges to Learning

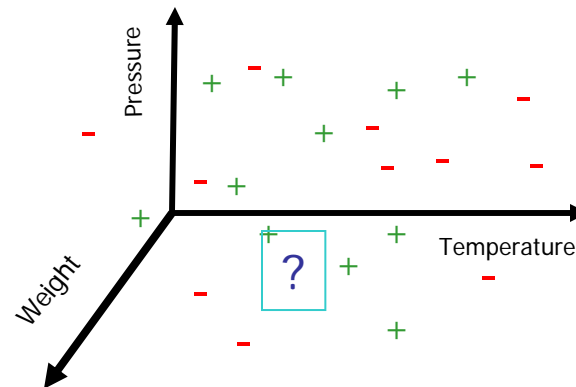
- Easy:



- 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 \approx 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

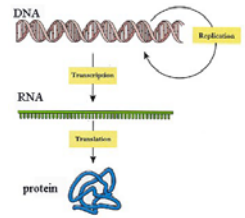


- to predict purchasing patterns
 - Victoria Secret (stocking)
- to help win games
 - NBA (scouting)



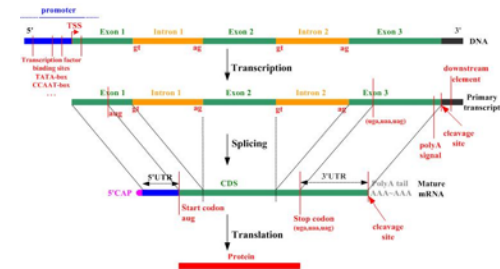
- to catalogue celestial objects [Fayyad et al. 93]
 - Discovered 22 new quasars
 - >92% accurate, over tetrabytes

2: Sequential Analysis

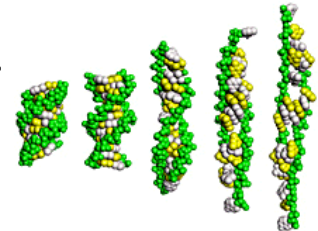


- **BioInformatics 1:** identifying genes

- Glimmer [Delcher et al, 95]
- identifies 97+% of genes, automatically!



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



- **Recognizing Handwriting**

Now, bushes 1-1-0
brought to life -0-0-0
for skimming, yim -0-1-0
for whom black -0-0-0
at the play -0-9-0
~~for my self~~
for my self -0-5-0
for my self 1-1-0
for cloth, bush, road 2-1-0
of my self 3-2-0
for my self 5-0-0
for my self 3-1-0
had left in my garden 1-12-0

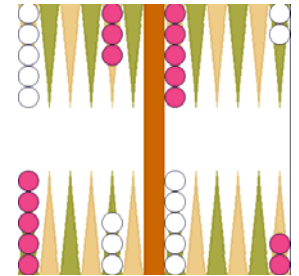
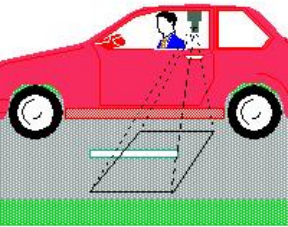
- **Recognizing Spoken Words**

- "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 grooves, ... 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



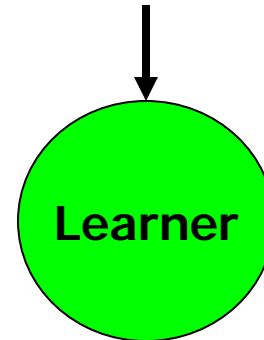


Outline

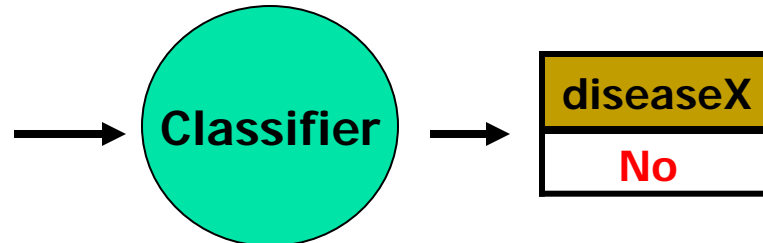
- Successes
 - Basic ideas
 - Foundations
 - Algorithms
 - Statistical Issues
- 

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



Why Learn?

Why not just “program it in”?

Appropriate Classifier ...

- ... is not known

Medical diagnosis... Credit risk... Control plant...

- ... is too hard to “engineer”

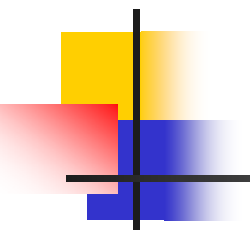
Drive a car... Recognize speech...

- ... changes over time

Plant evolves...

- ... user specific

Adaptive user interface...



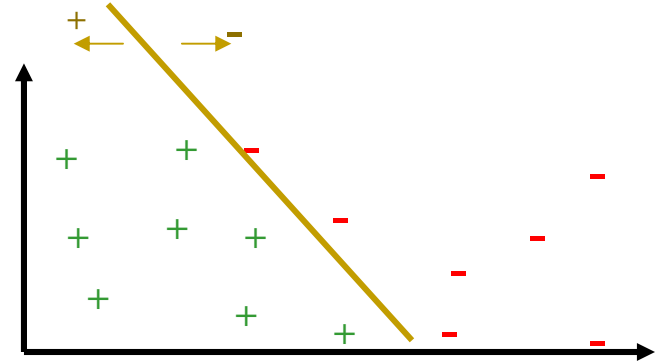
Why Machine Learning is especially relevant **now!**

- 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



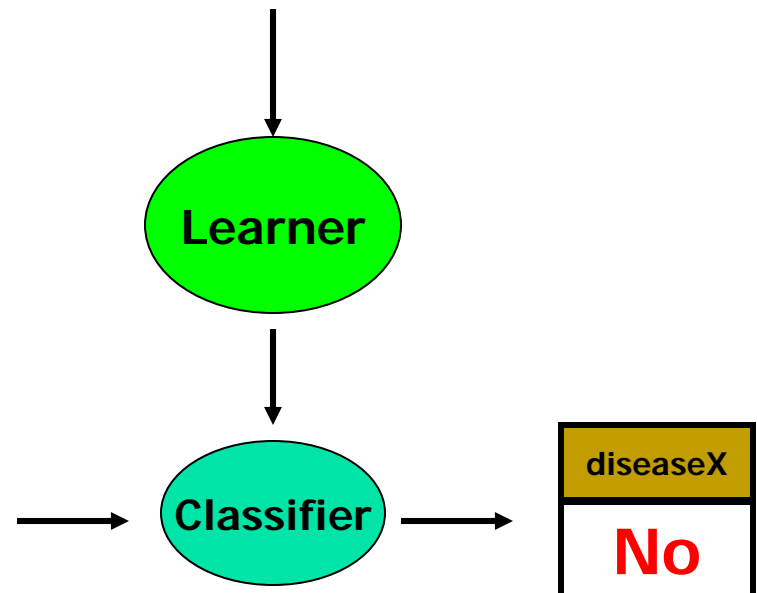
General Process

- Given "labeled data"

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

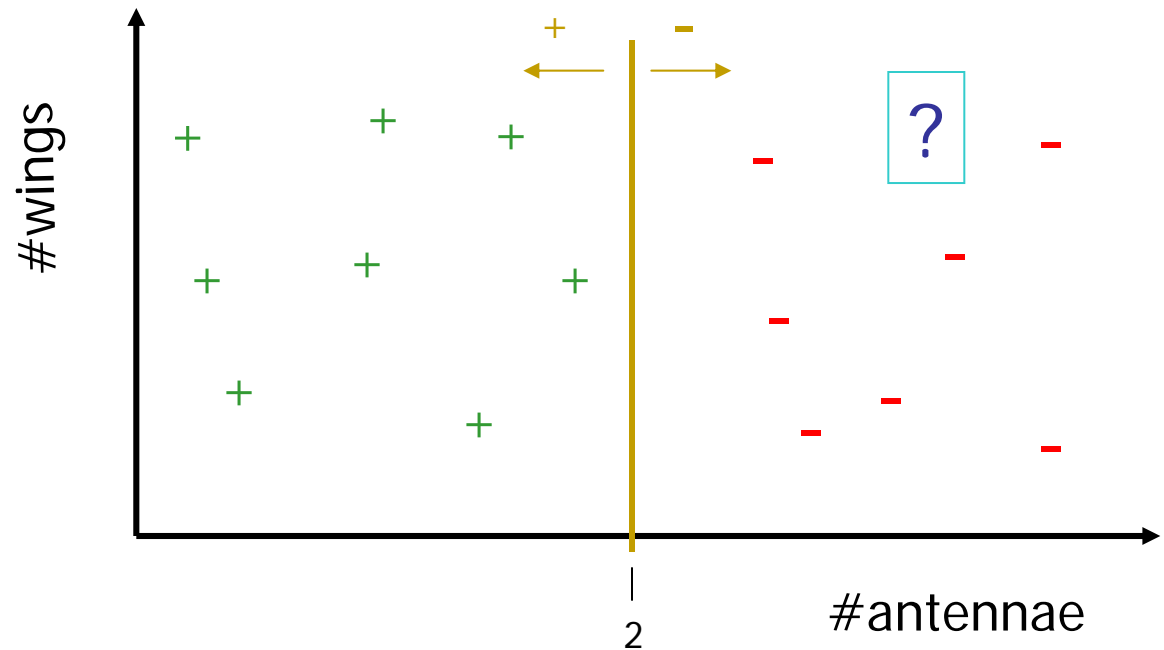
- Learn CLASSIFIER, that can predict label of *NEW* instance

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



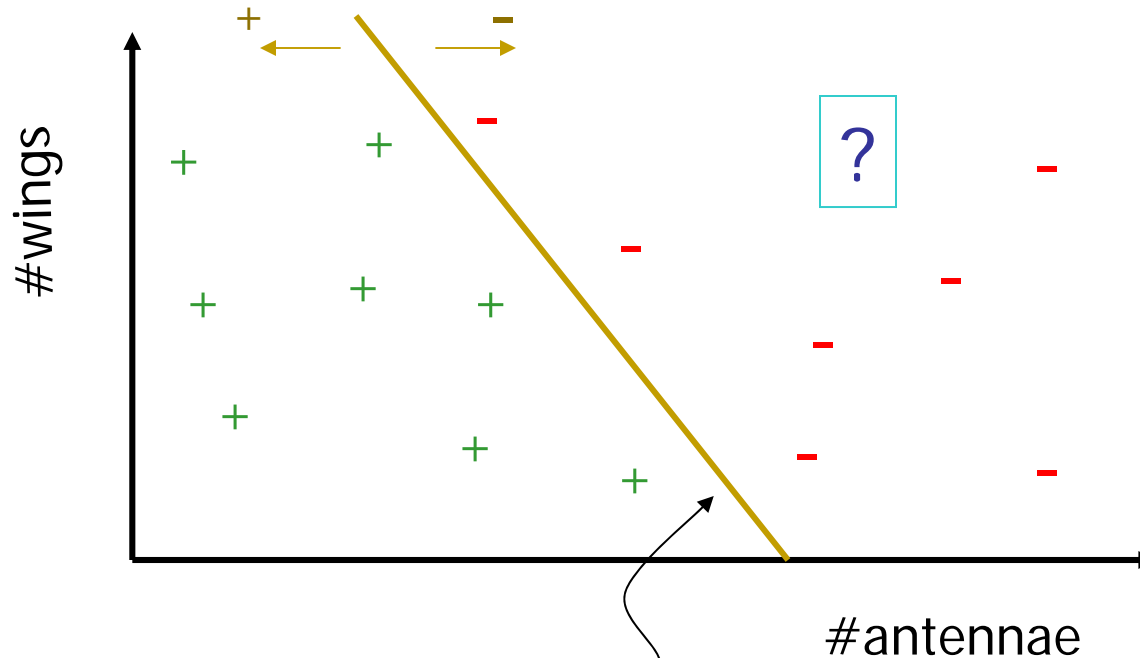
Alg 1: Linear Separators

- 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 $2.3 \times \#Wings + 7.5 \times \#antennae + 1.2 > 0$ 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

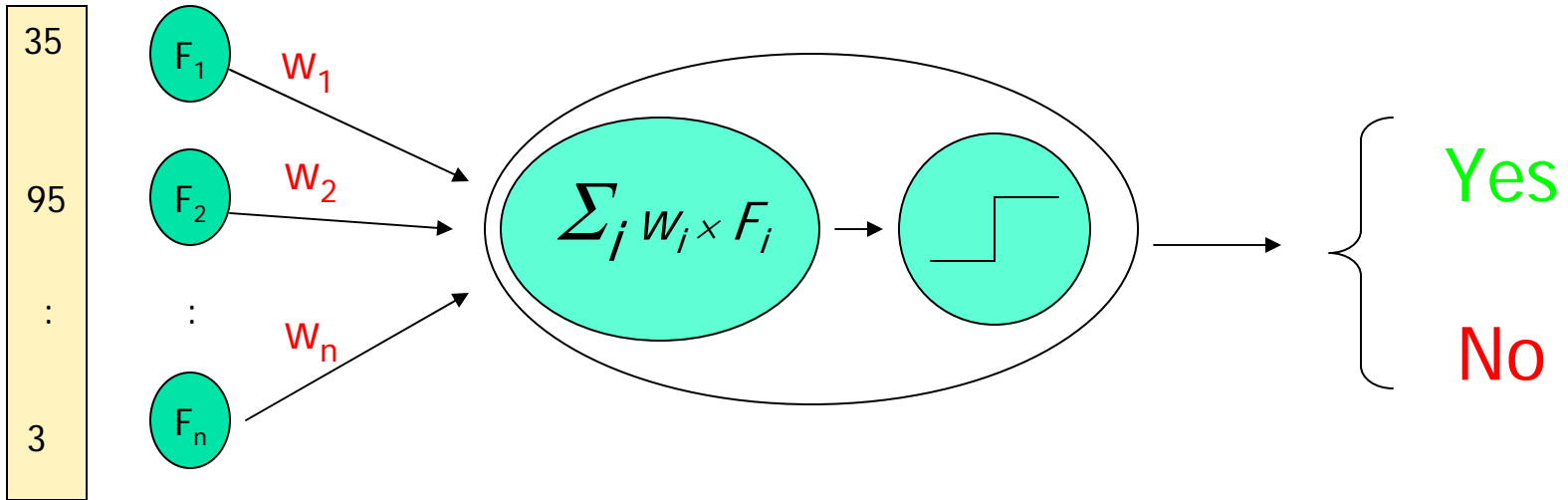
- find “weights” $\{W_1, W_2, \dots, W_n, W_0\}$
such that

$$W_1 \times F_1 + \dots + W_n \times F_n + W_0 > 0$$

means

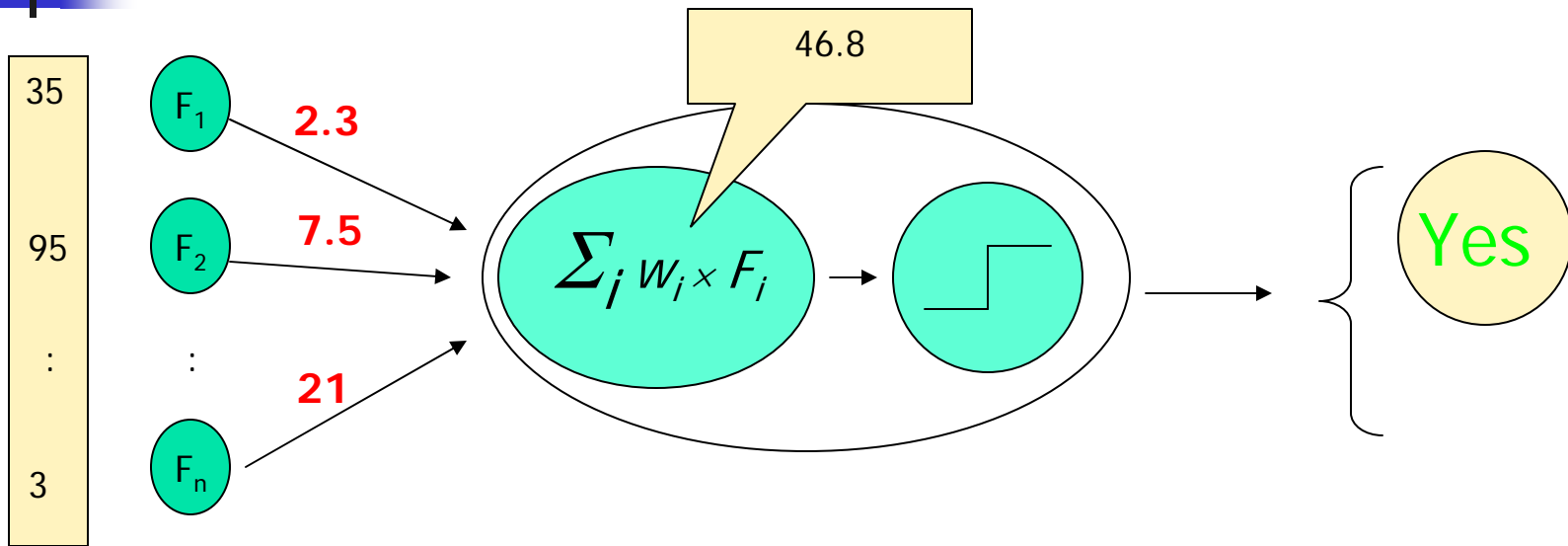
Class = Yes

Linear Separator



Just view $F_0 = 0$, so $w_0 \dots$

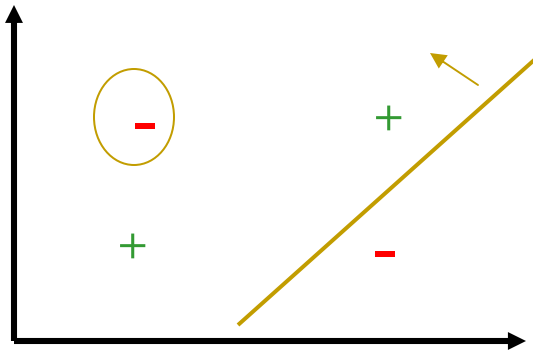
Linear Separator



- 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...

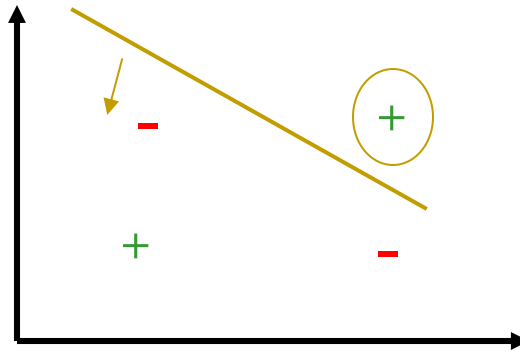


Linear Separators – Facts

- GOOD NEWS:

- If data is linearly separated,
- Then **FAST ALGORITHM** finds correct $\{w_i\}$!

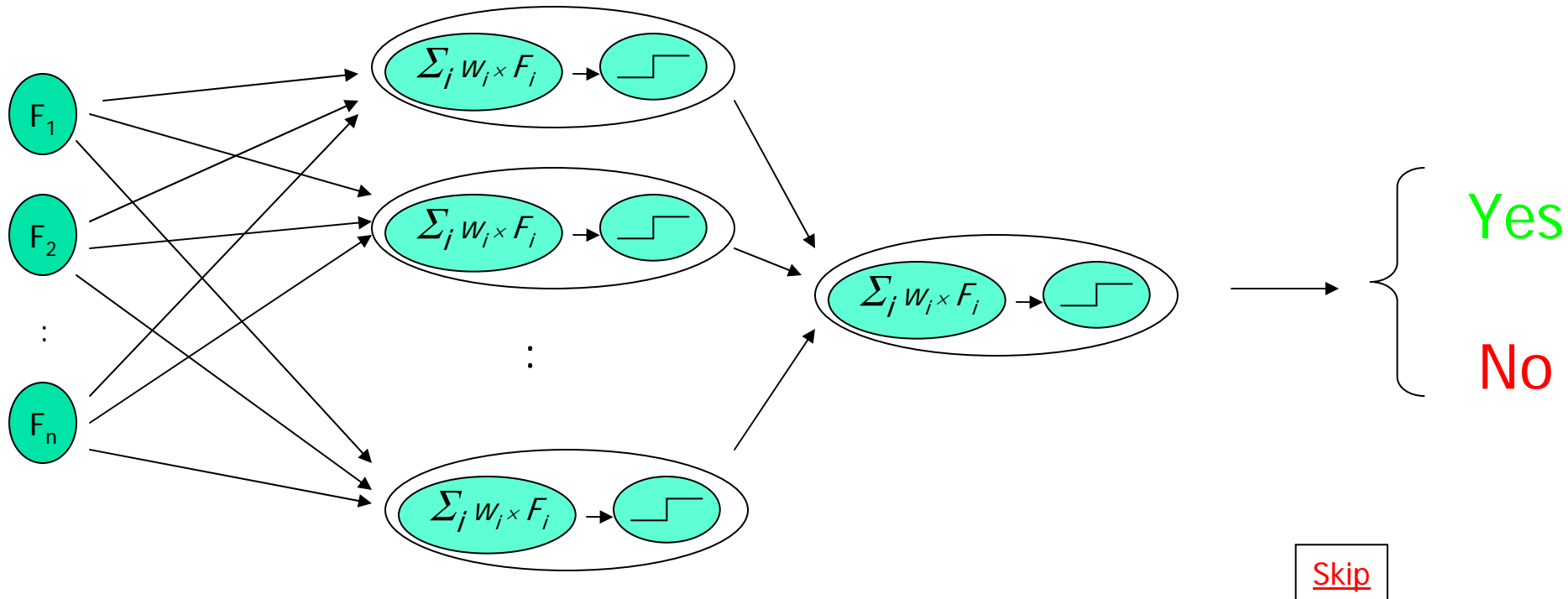
- But...



- Some “data sets” are NOT linearly separatable!

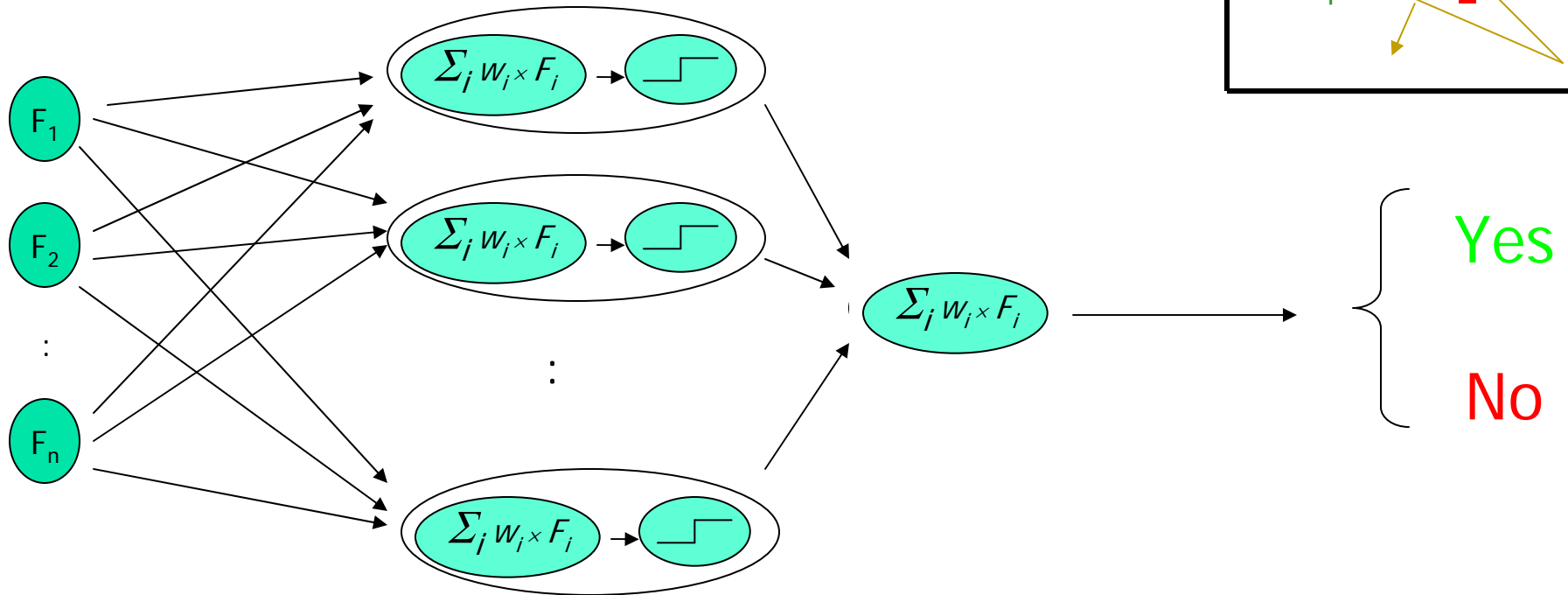
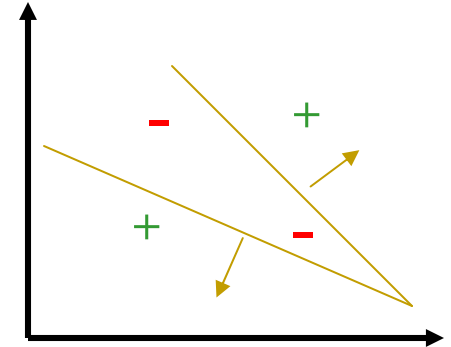
Alg 2: Artificial Neural Nets

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



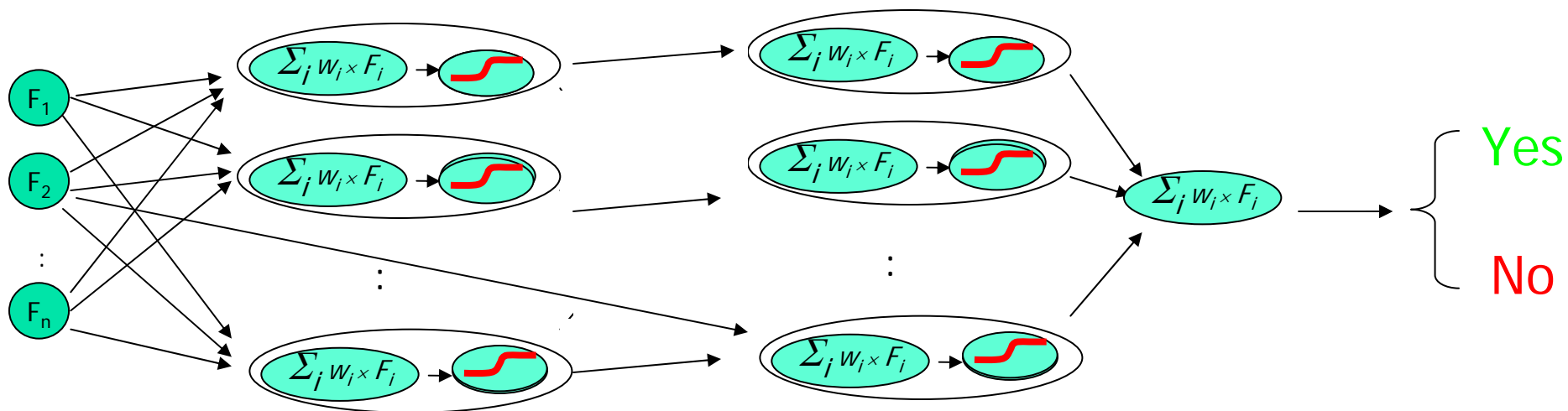
Artificial Neural Nets

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



ANNs: Architecture

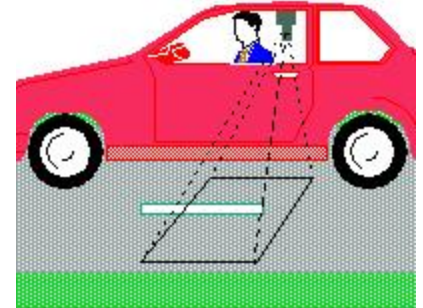
- Different # of layers
- Different structures
 - what's connected to what..
- Different "squashing function"



Uses of Artificial Neural Nets

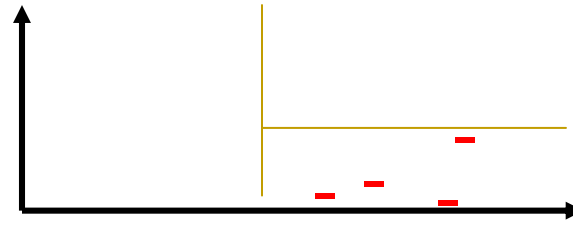
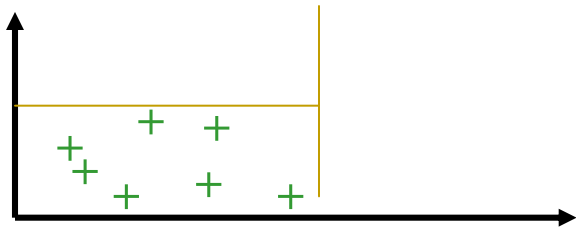
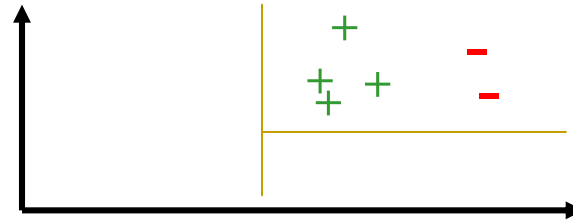
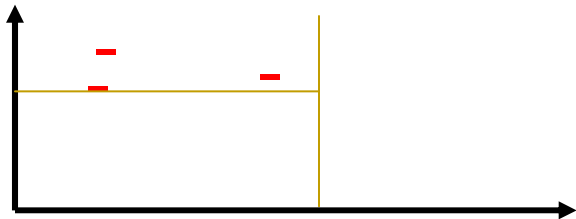
Learning to ...

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



Algorithm 3: Decision Trees

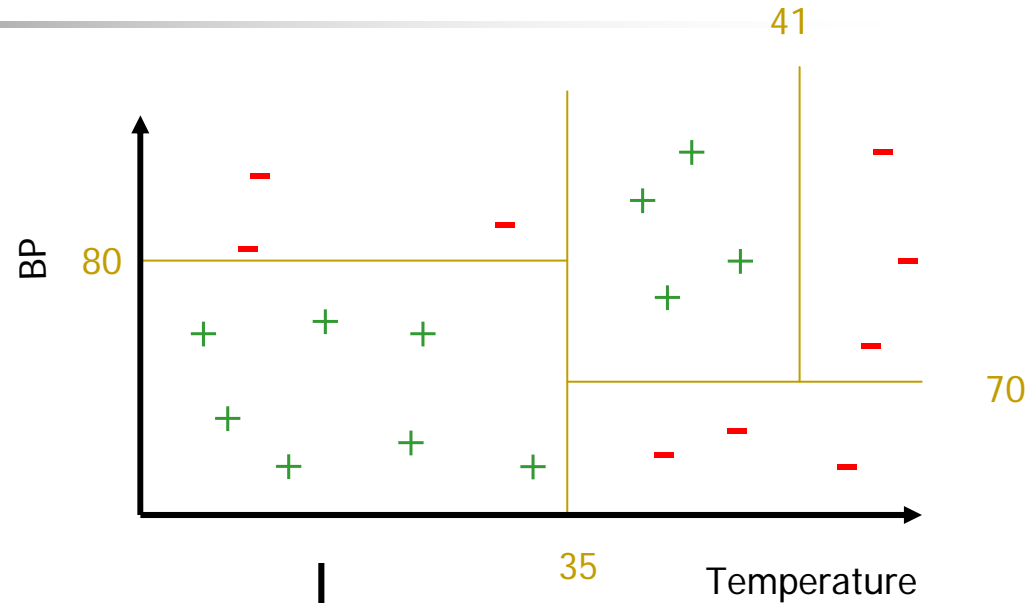
- Given data, decide on best *first* split



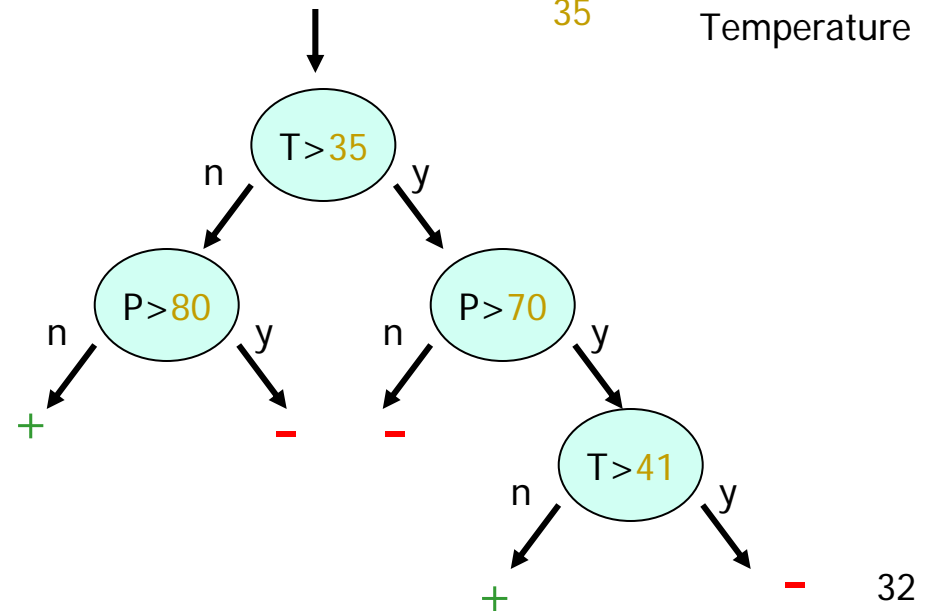
- Then consider each subset of data:
 - decide on its best split
- Recur... until "purity"

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)

<http://www.cs.ualberta.ca/~aixplore>

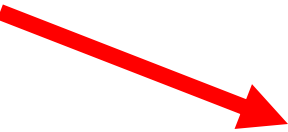


Other Algorithms

- 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 \mathcal{R}



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?

a =

b =

d =

F ₁	F ₂	...	F _n	Class
35	95	...	3	No
22	80	...	-2	Yes
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?



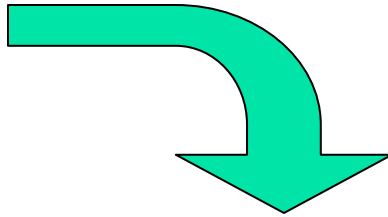
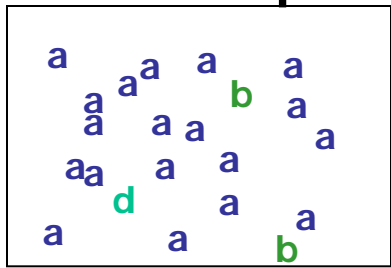
2. Why should Learning work?

- *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?

- Overall Population



a =

b =

d =

	F ₁	F ₂	...	F _n	Class
a =	35	95	...	3	No
b =	22	80	...	-2	Yes
d =	10	50	...	1.9	No

- Draw sample $S =$ a a a b a a

- Learn classifier \mathcal{C} that does well on S :

- As S includes $a b$,

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

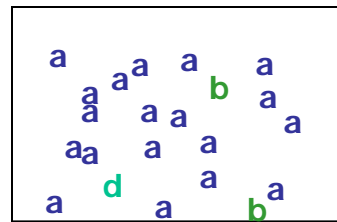
- Notice d not in S

$\Rightarrow \mathcal{C}(d) = ??$

How good is Classifier \mathcal{C} ?

- To evaluate \mathcal{C}

- Draw new patient, $x \in$
- Compute $\mathcal{C}(x)$
- Correct?



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

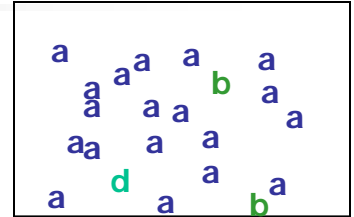
- Given true distribution,

- *expect* $x = a$... or $x = b$
 - Here: $\mathcal{C}(x)$ is correct!
- Otherwise, $\mathcal{C}(x)$ may be wrong.
 - But this is rare!

Why should Learning work?

Consider a new patient, x ...

1. If x occurs a LOT $P(x) \gg 0$
 - x probably appears in S
 - As C does well on S ,
 C gives correct answer on x
 2. If x occurs rarely $P(x) \approx 0$
 - doesn't matter if C is wrong!
- Even good classifiers are wrong occasionally...

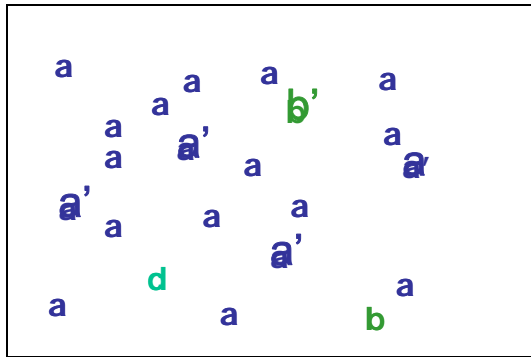


Populations

- 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...



a =

a' =

b =

b' =

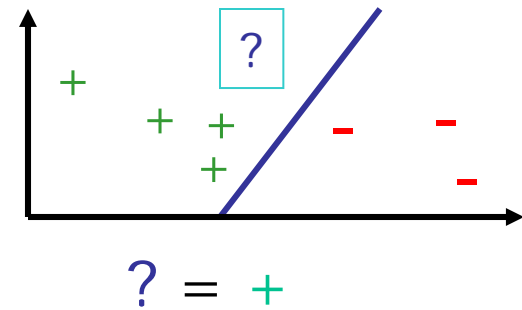
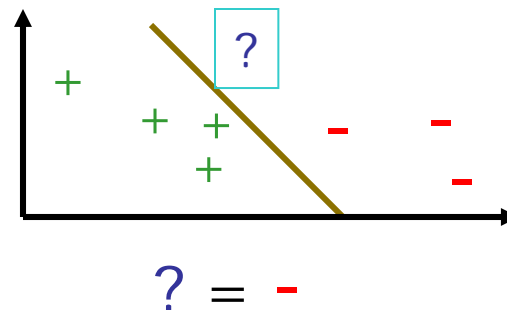
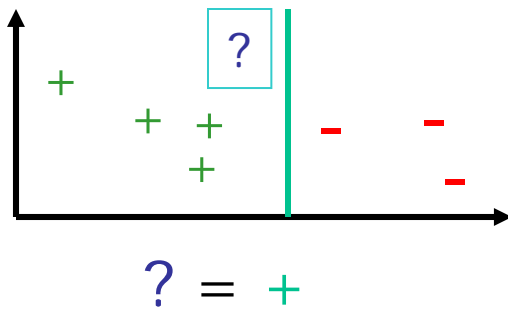
d =

F_1	F_2	...	F_n	Class
35	95	...	3	No
36	95	...	2.1	No
22	80	...	-2	Yes
22	78	...	-2.3	Yes
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 ...

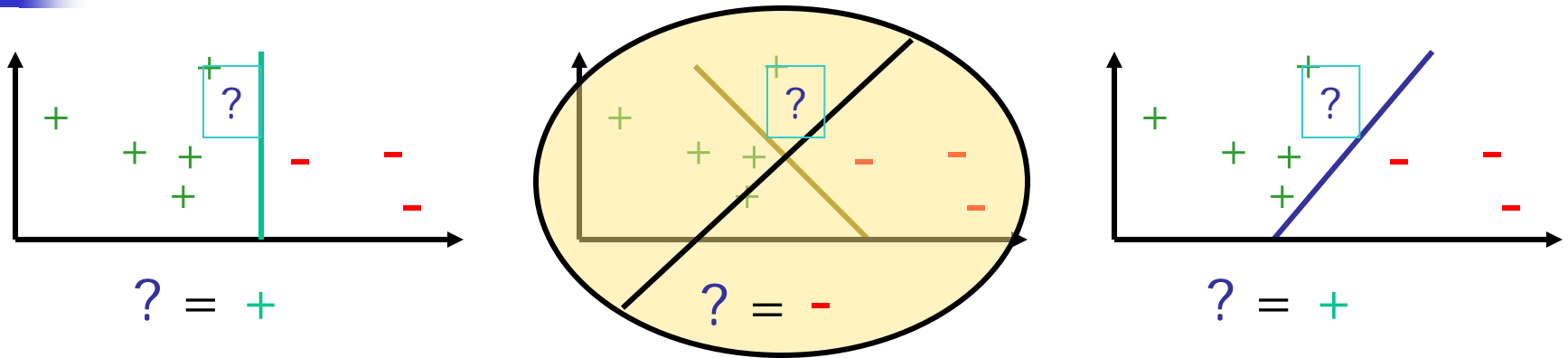
3. How much training data?

- 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



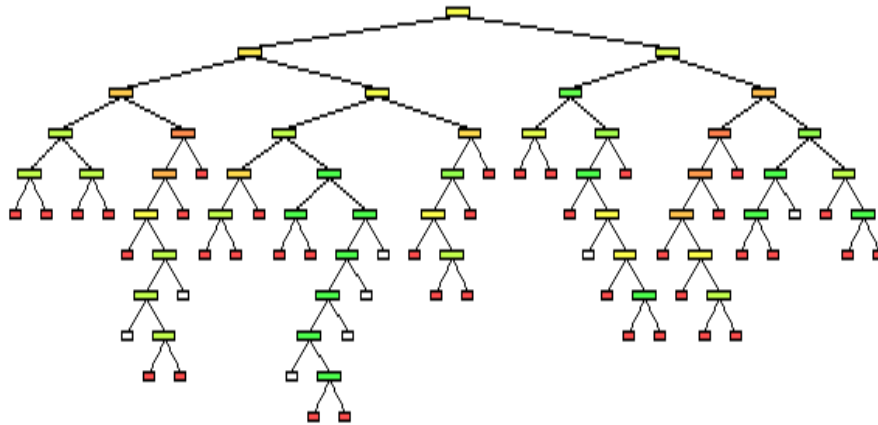


Example – continued

- 25% have **butterfly-itis**
- $\frac{1}{2}$ of patients have $F_1 = 1$
 - Eg: “odd birthday”
- $\frac{1}{2}$ 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

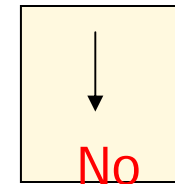
- Standard decision tree learner:



- Error Rate:

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

- Optimal decision tree:

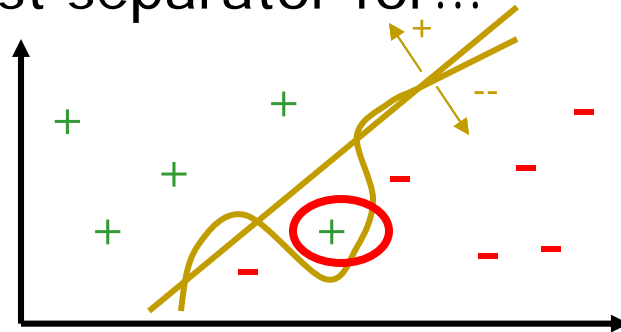


- Error Rate:

- Train data: 25%
- New data: 25%

Overfitting

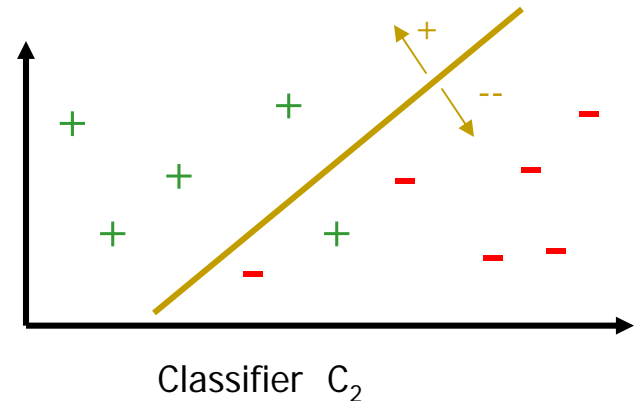
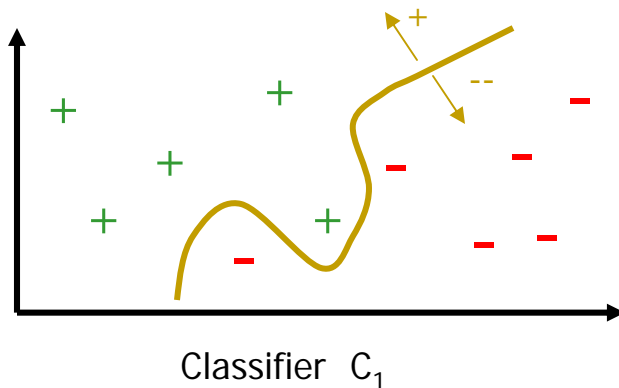
- Some features are not helpful
- Data often noisy
 - typos in recording, error in equipment, human error...
- What is best separator for...



- 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!*

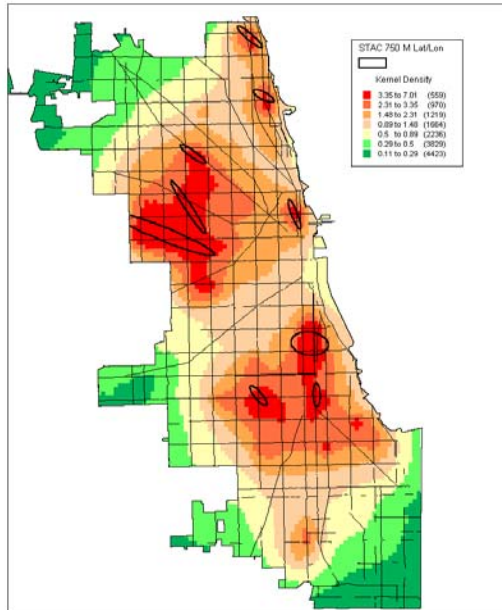


Outline

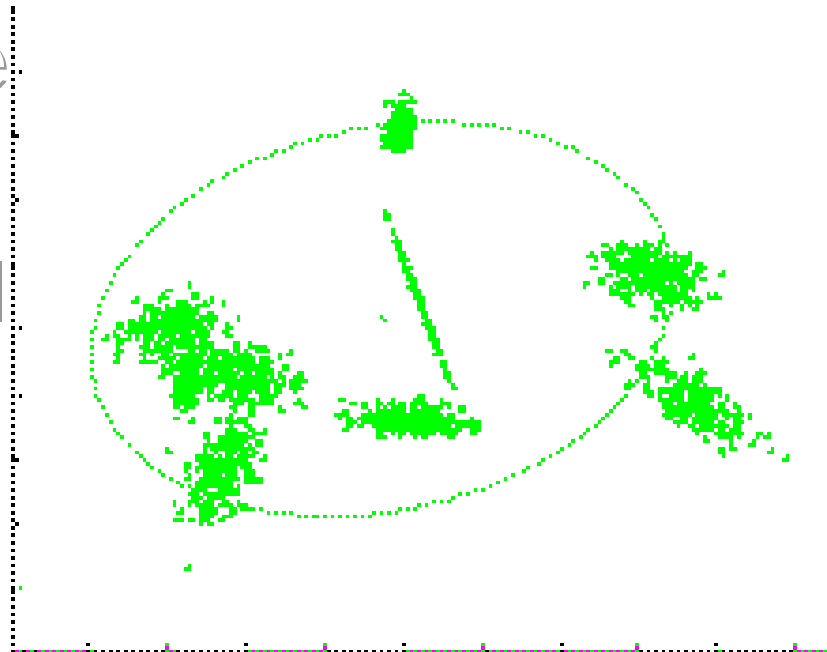
- Successes
 - Basic ideas
 - Foundations
 - Algorithms
 - Statistical Issues
 1. Goal of Learning
 2. Why should Learning work?
 3. How much data is needed?
 4. Overfitting
 5. Computational Efficiency
 6. Imbalanced data (fraud detection)
 7. Non-IID tuples (stock market, temporal)
- 

Other Types of Learning

- Density Estimation
 - Learning Generative Model
 - Clustering

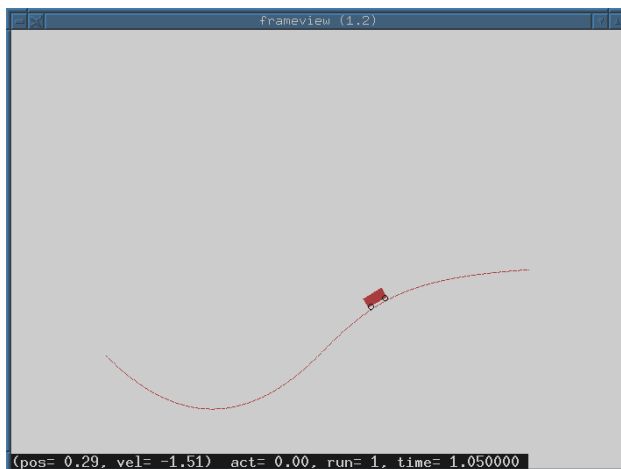


equation-11

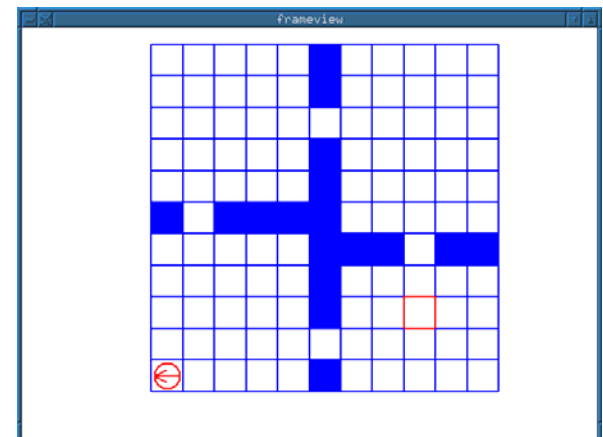


Other Types of Learning

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

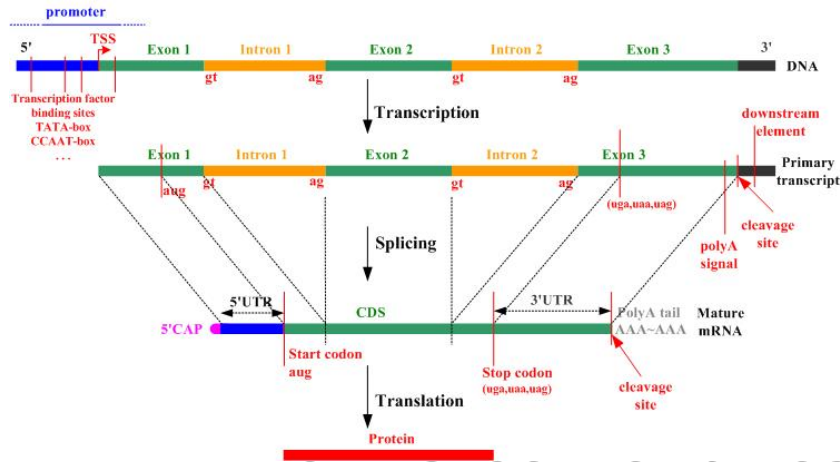


on-ID Data



Other Types of Learning

Density Estimation



ive Mode

e of Acti

arning



■ 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?
