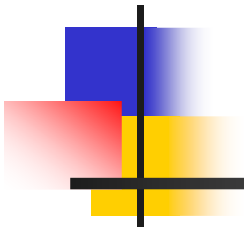


Machine Learning

Russ Greiner

Alberta Ingenuity Centre for Machine Learning

Department of Computing Science

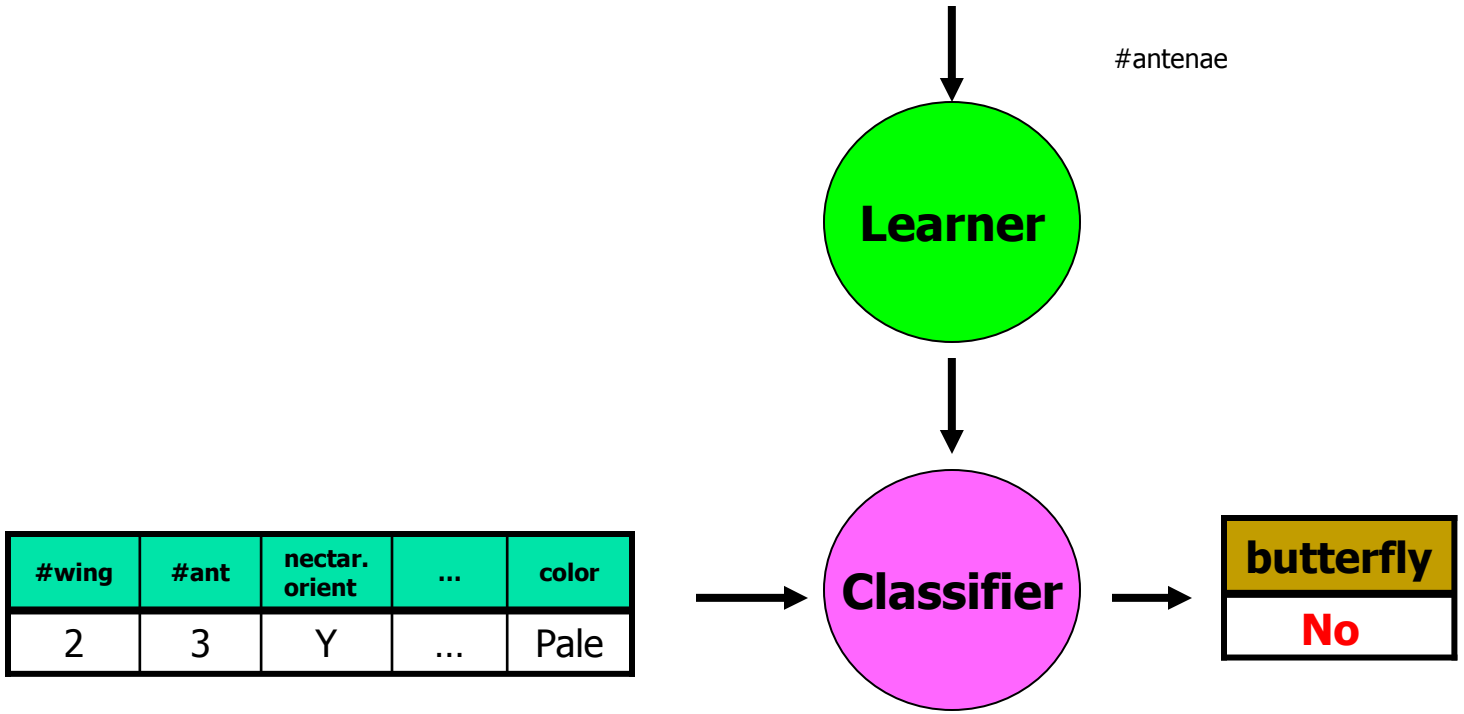
University of Alberta

aicml

Diagnosing Butterfly-itis

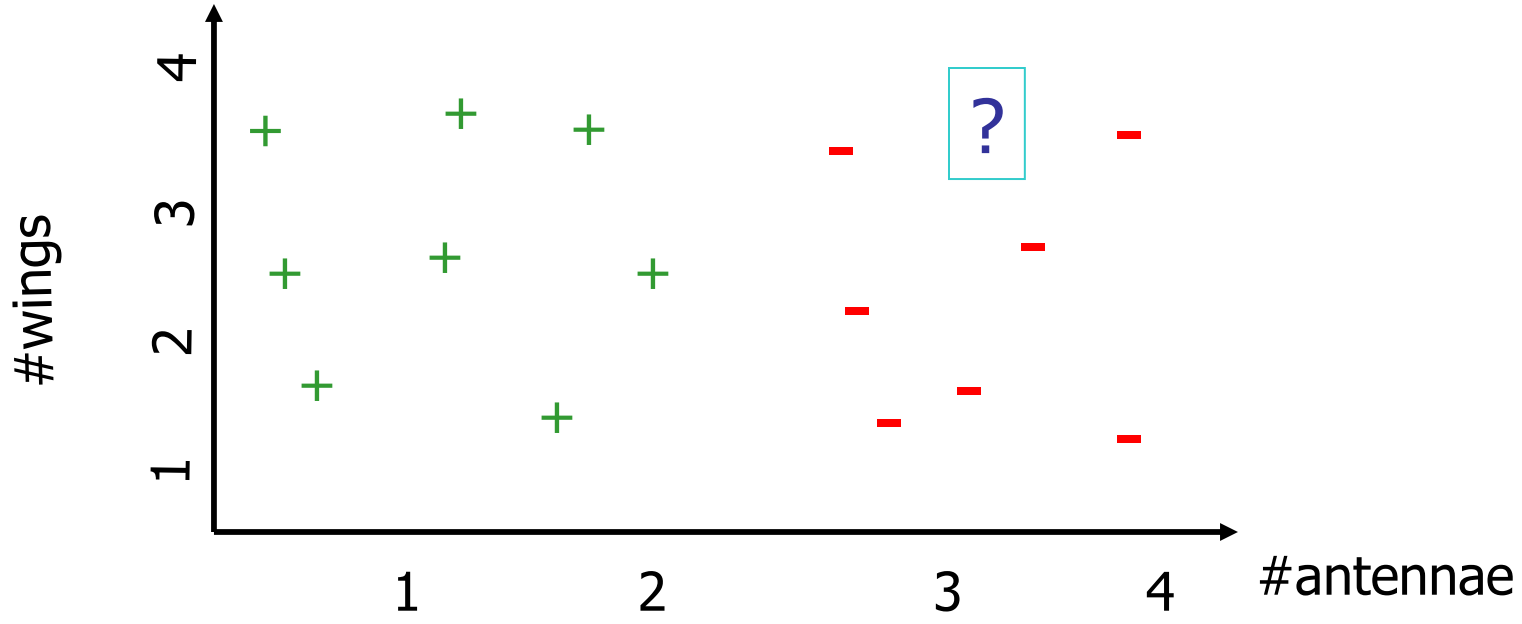
Classifying a Patient

#wing	#ant	nectar -orient	...	color	butterfly
3	5	Y	...	Pale	No
8	4	N	...	Clear	Yes
:	:			:	:
10	8	N	...	Pale	No



#wing	#ant	nectar. orient	...	color
2	3	Y	...	Pale

Visualizing Patient Data

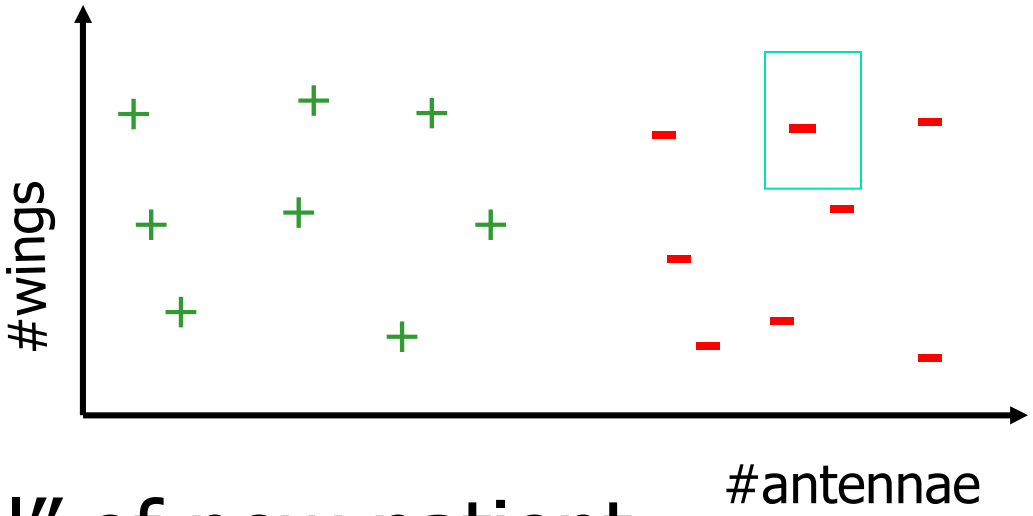


+ = yes − = No

- What about this new patient ?

This is learning…

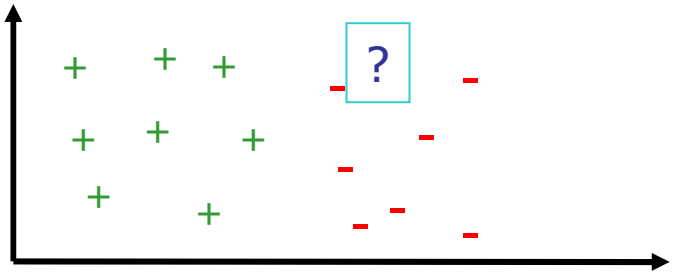
- Given data:



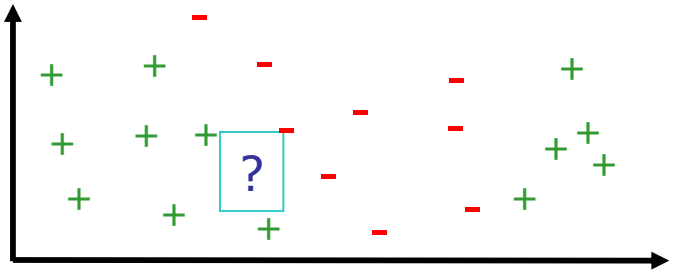
- Predicting "label" of new patient
  - Here: Negative "–" (not butterfly-it is)
- This is an ***EDUCATED GUESS***:
  - … not based on post-mortem, definitive test, …
  - use to decide on treatment, etc.

Challenges to Learning

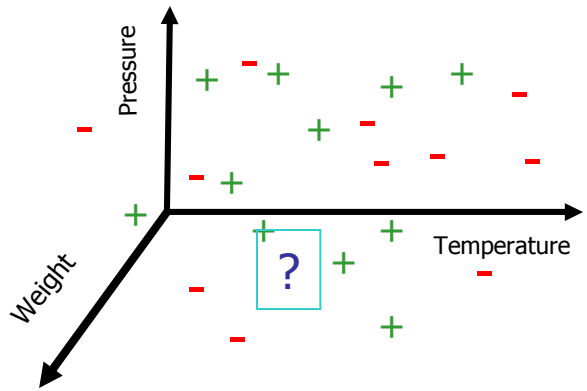
- Easy:

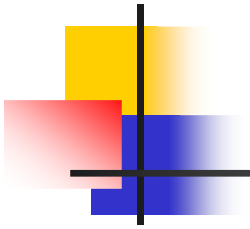


- Harder:



- High Dimension:



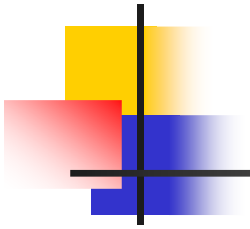


Machine Learning studies …

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
- …



Outline

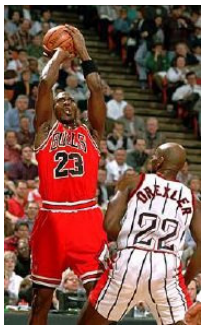
- Successes
  - Mining Data Sets
  - Sequential Analysis
  - Control
- Basic ideas
  - Foundations
  - Algorithms
  - Statistical Issues
- Current Research

Successes: Mining Data Sets Computer learns…

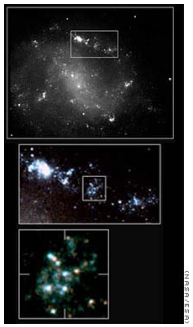
- to find ideal customers
  - Credit Card approval (AMEX)
    - Humans ≈50%; ML is >70% !

- 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)

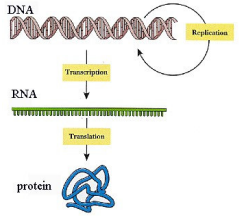


- to help win games
  - NBA (scouting)

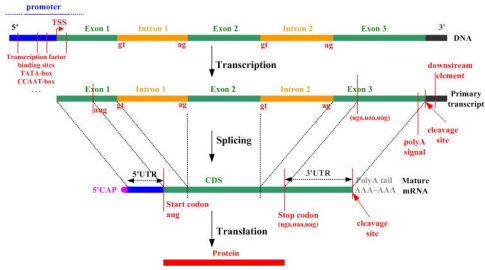


- to catalogue celestial objects [Fayyad et al. 93]
  - Discovered 22 new quasars
  - >92% accurate, over tetrabytes

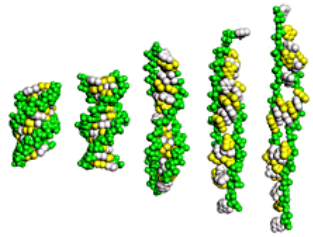
2: Sequential Analysis



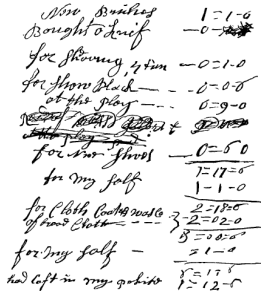
- **BioInformatics 1:** identifying genes
  - Glimmer [Delcher et al, 95]
  - identifies 97+% of genes, automatically!



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

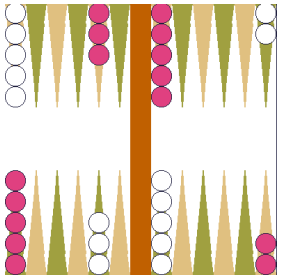


- **Recognizing Handwriting**

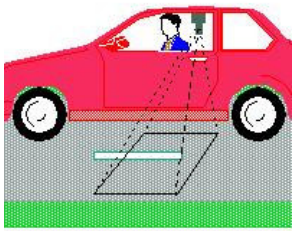


- **Recognizing Spoken Words**
  - **"How to wreck a nice beach"**

3: Control



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



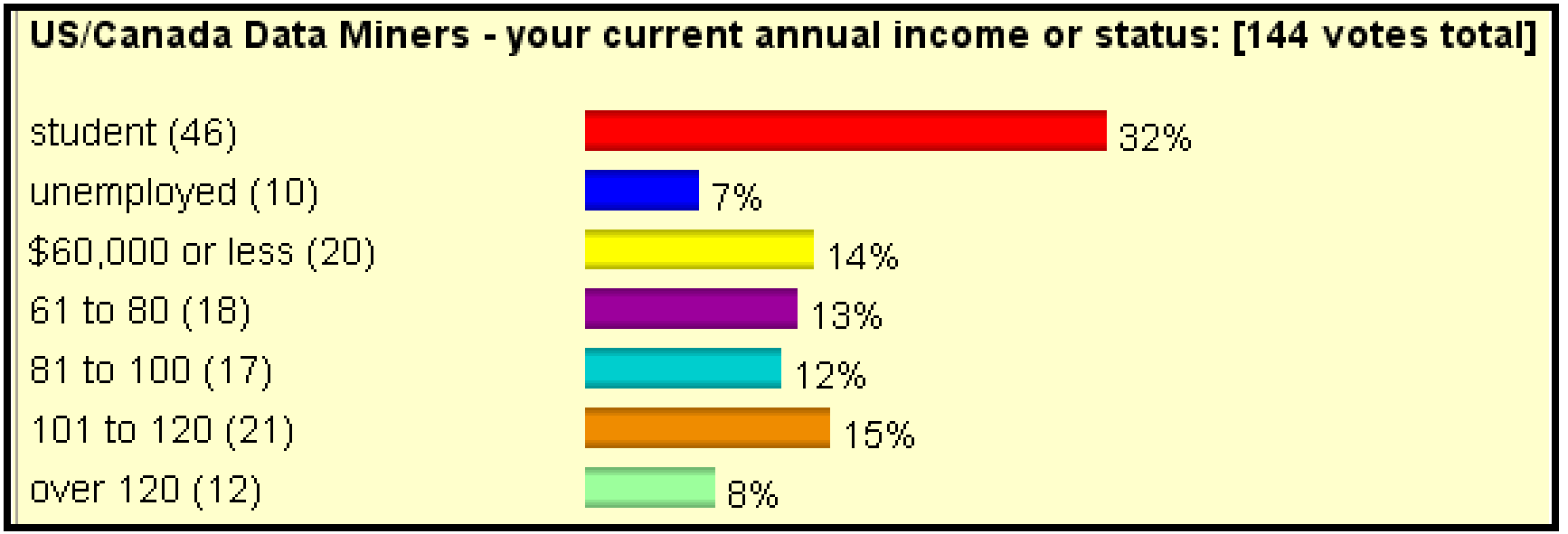
- **Drive autonomous vehicles** (Thrun 2005)
  - DARPA Grand Challenge

- **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!

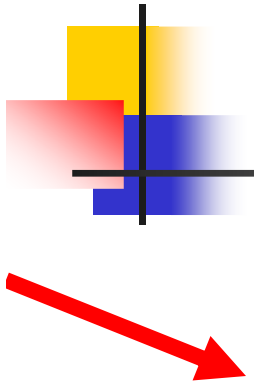
- **Oil refinery**
  - Separate oil from gas
  - … in 10 minutes (human experts require 1+ days)
- Manufacture nuclear fuel pellets (Leech, 86)
  - Saves Westinghouse >$10M / year
- **Adaptive** agents / user-interfaces

Salaries

Poll – KDD (US/Canada Data Miners Income/Status – 2004)



- Alberta Ingenuity Centre for Machine Learning

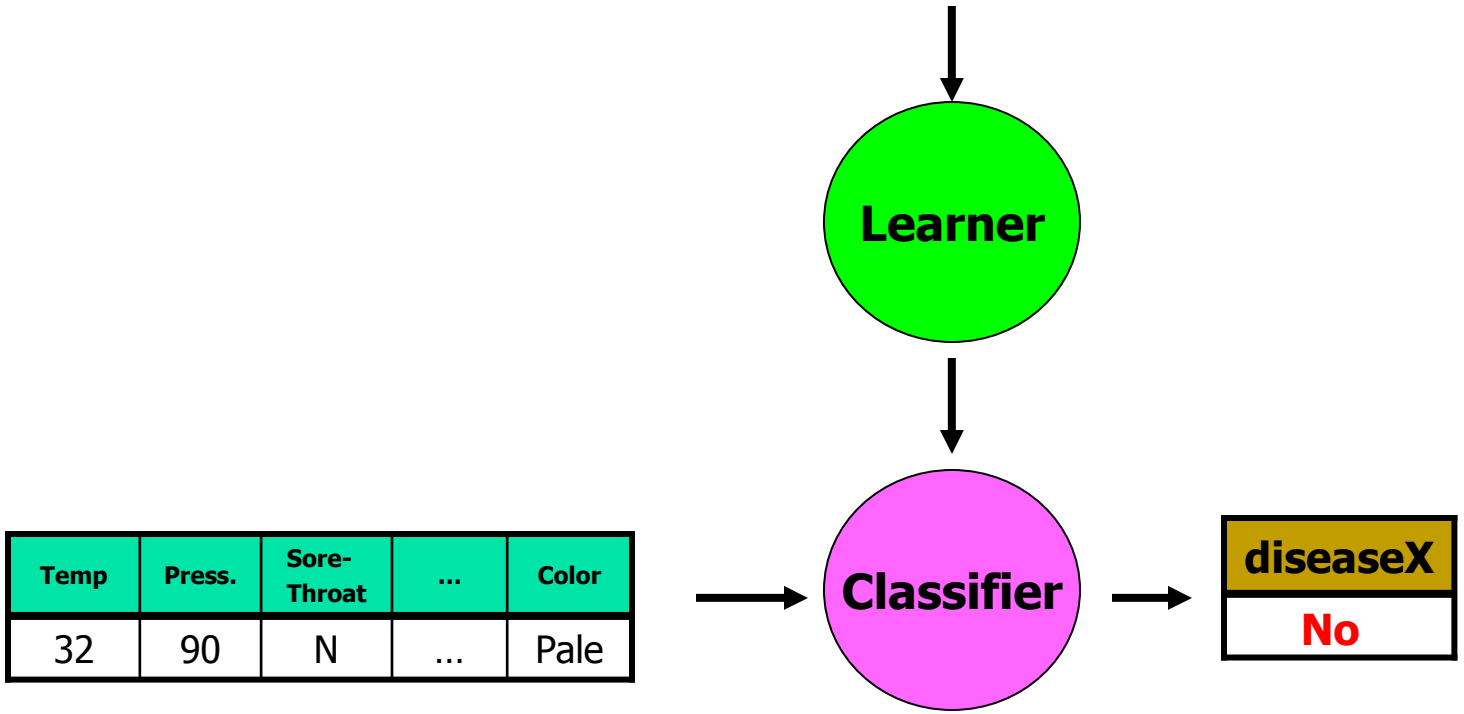


Outline

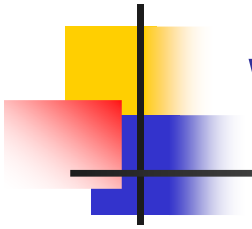
- Successes
- Basic ideas
  - Foundations
  - Algorithms
  - Statistical Issues
- Current Research

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



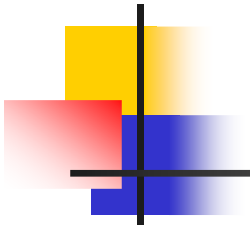
Temp	Press.	Sore-Throat	...	Color
32	90	N	...	Pale



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…

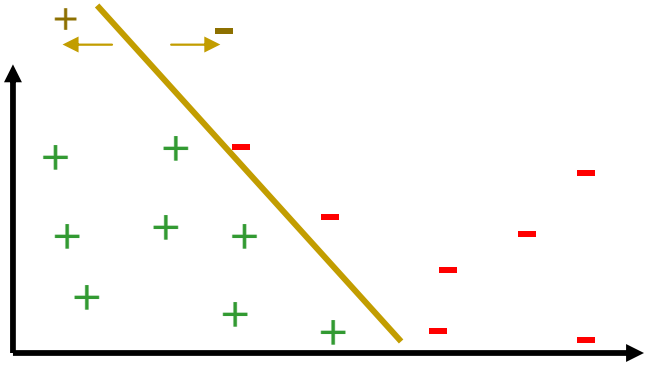


Why Machine Learning is especially relevant **now**!

- 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, …

Outline

- Successes
- Basic ideas
  - Foundations
  - Algorithms
    - Linear Separators
      - Support Vector Machines
    - Artificial Neural Nets
    - Decision Trees
    - Naïve Bayes
    - Nearest Neighbor, …
  - Statistical Issues
- Current Research
- UofA + AICML

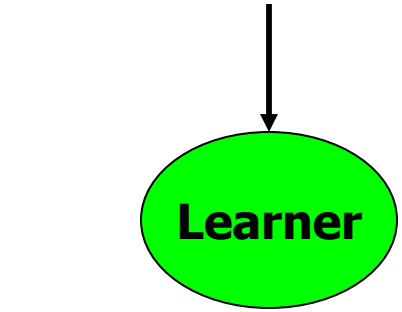


General Process

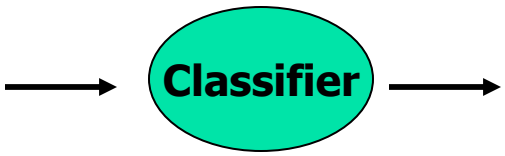
- Given "labeled data"

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



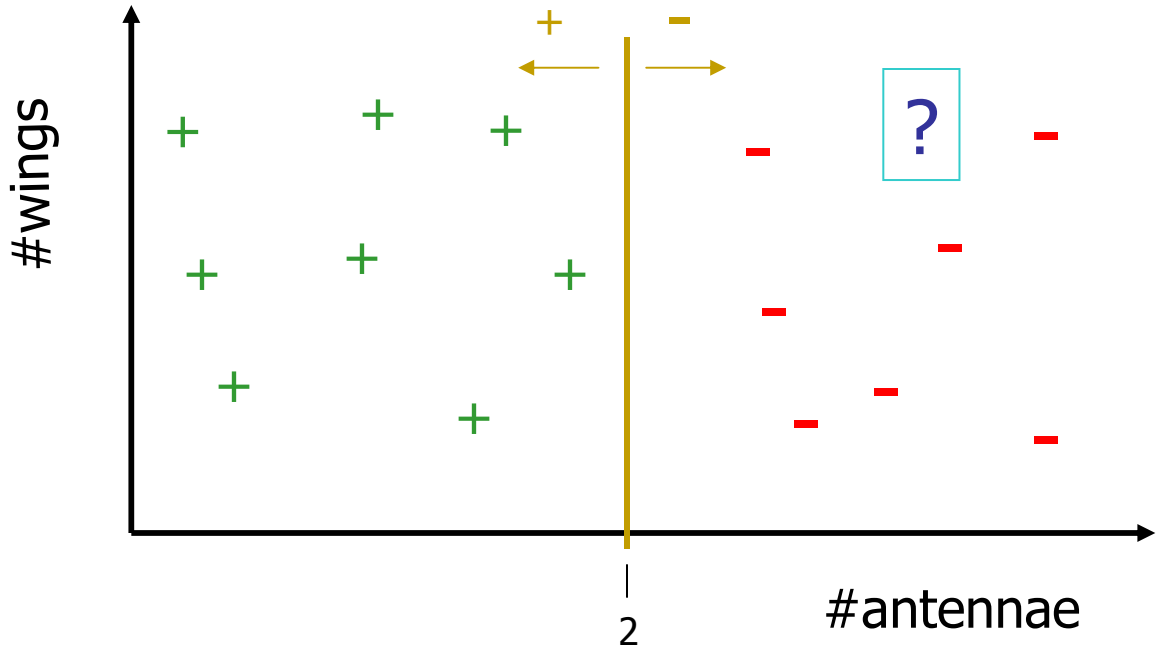
Temp	BP	Sore-Throat	...	Color	diseaseX
32	90	N	...	Pale	?



diseaseX
No

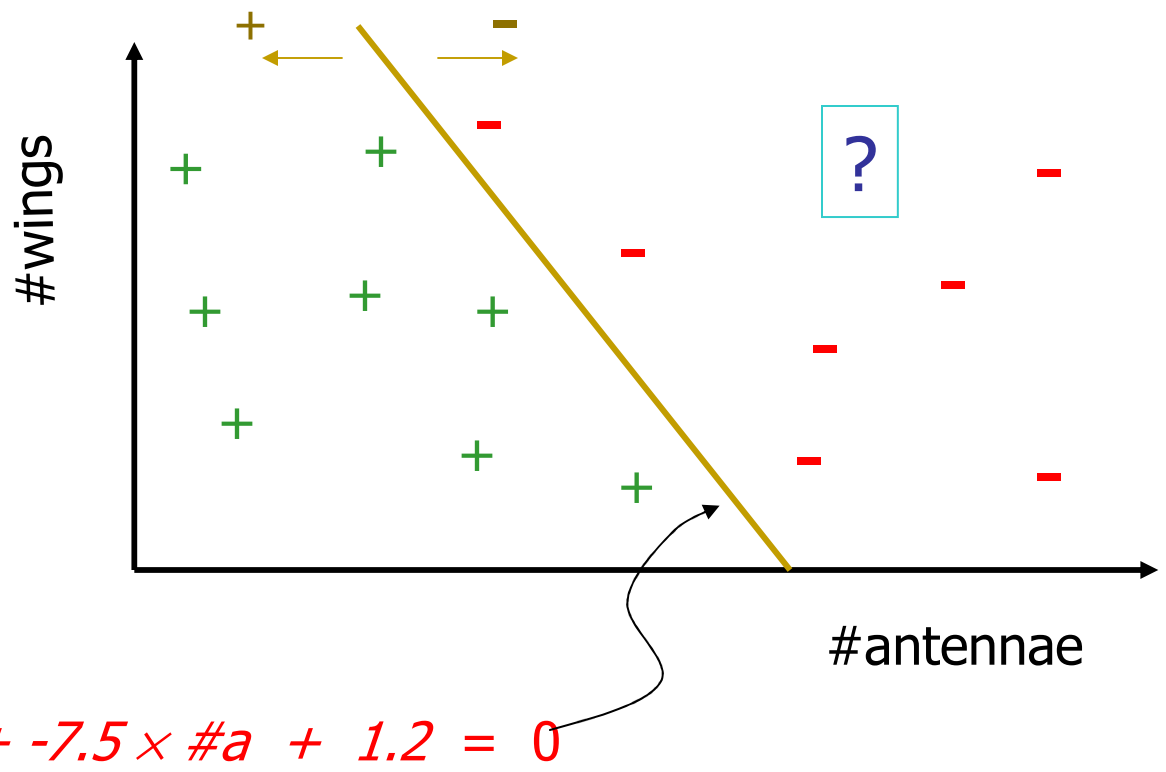
Alg 1: Linear Separators

- Draw "separating line"



- If #antennae ≤ 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)

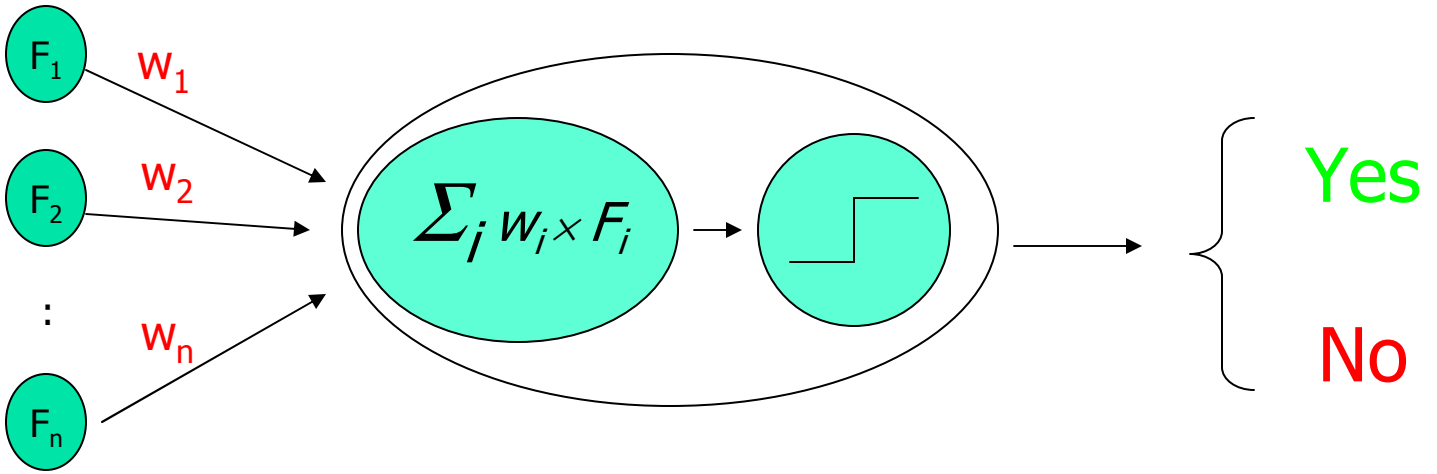
F_1	F_2	…	F_n	Class
35	95	…	3	No
22	80	…	-2	Yes
:	:		:	:
10	50	…	1.9	No

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

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

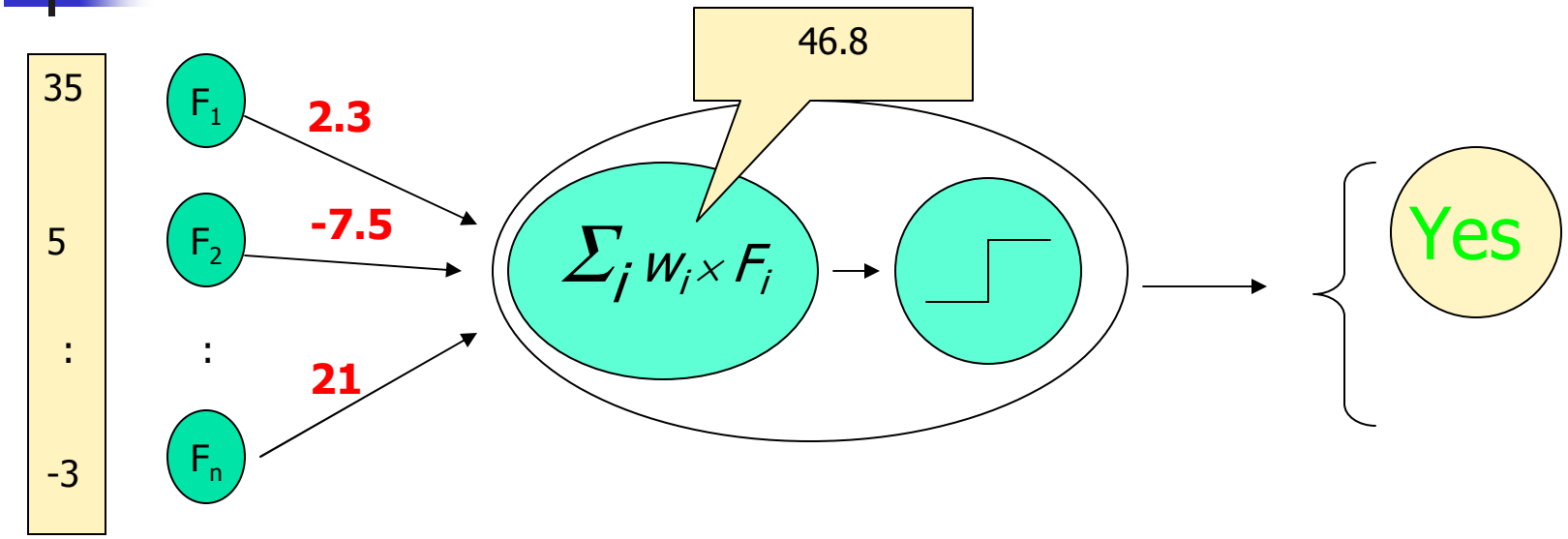
means $Class = Yes$

Linear Separator



Just view $F_0 = 0$, so w_0 …

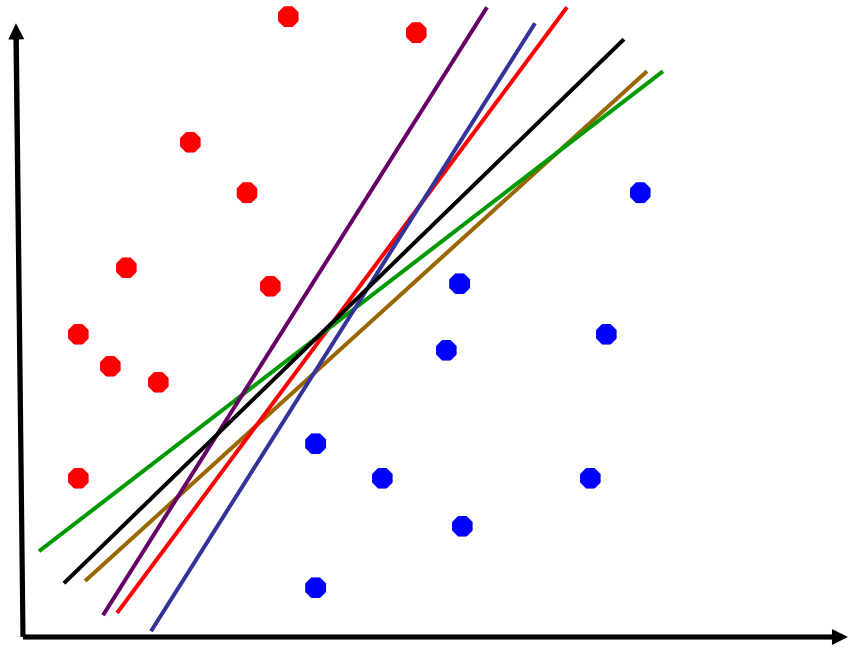
Linear Separator



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

Support Vector Machine (SMO)

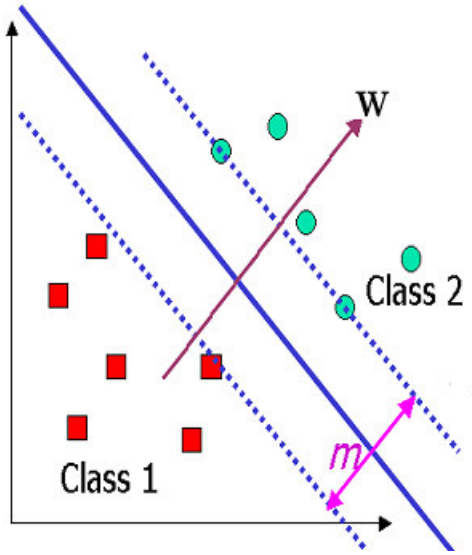
- Many linear separators …
- Which is best?



Support Vector Machine (SMO)

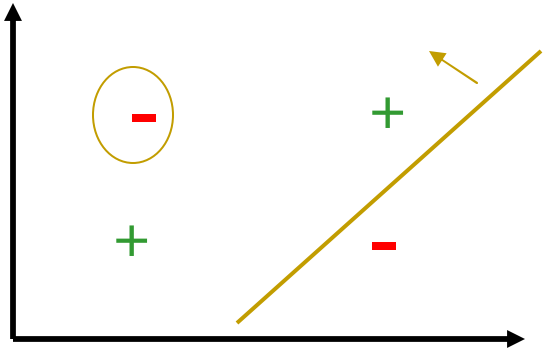
- Decision boundary should be as far away as possible from the data

⇒ maximize margin, *m*



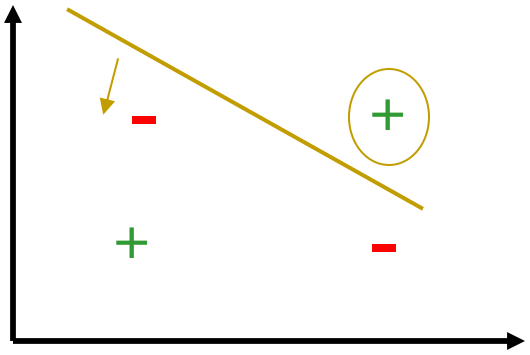
Linear Separators – Facts

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



Linear Separators – Facts

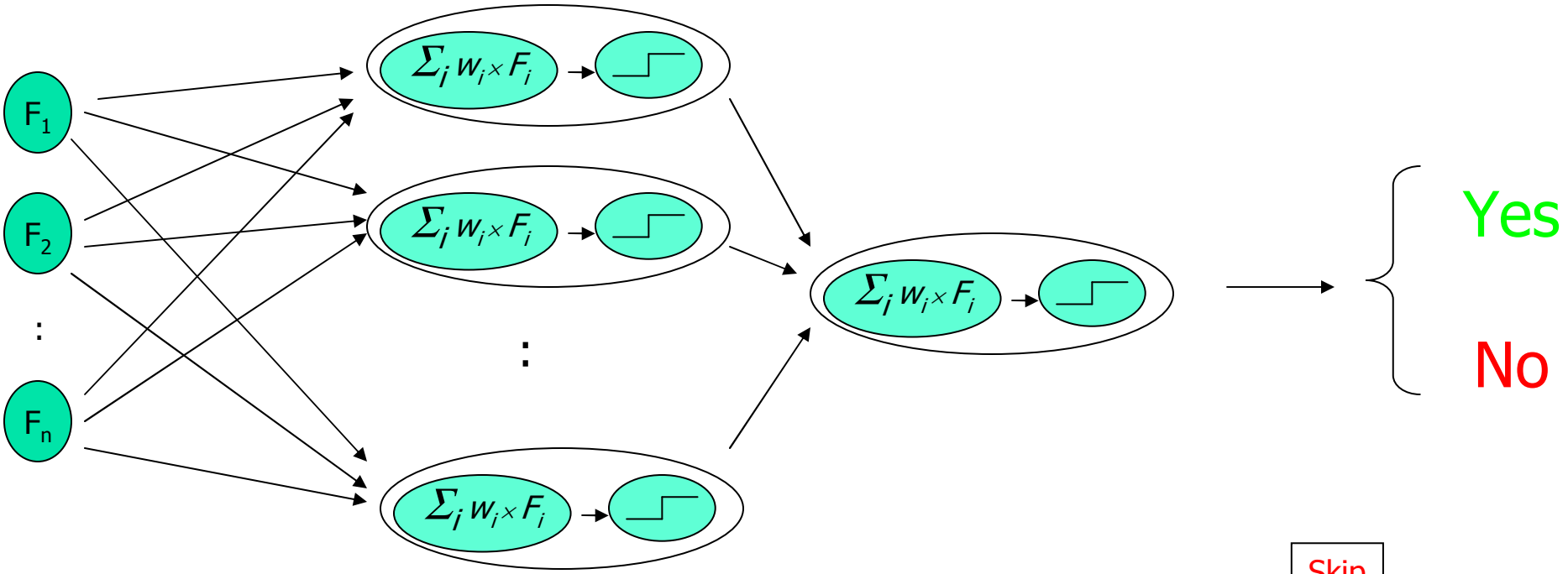
- GOOD NEWS:
  - If data is linearly separated,
  - Then FAST ALGORITHM finds correct $\{w_i\}$!
- But…



- Some "data sets" are
  NOT linearly separatable!

Alg 2: Artificial Neural Nets

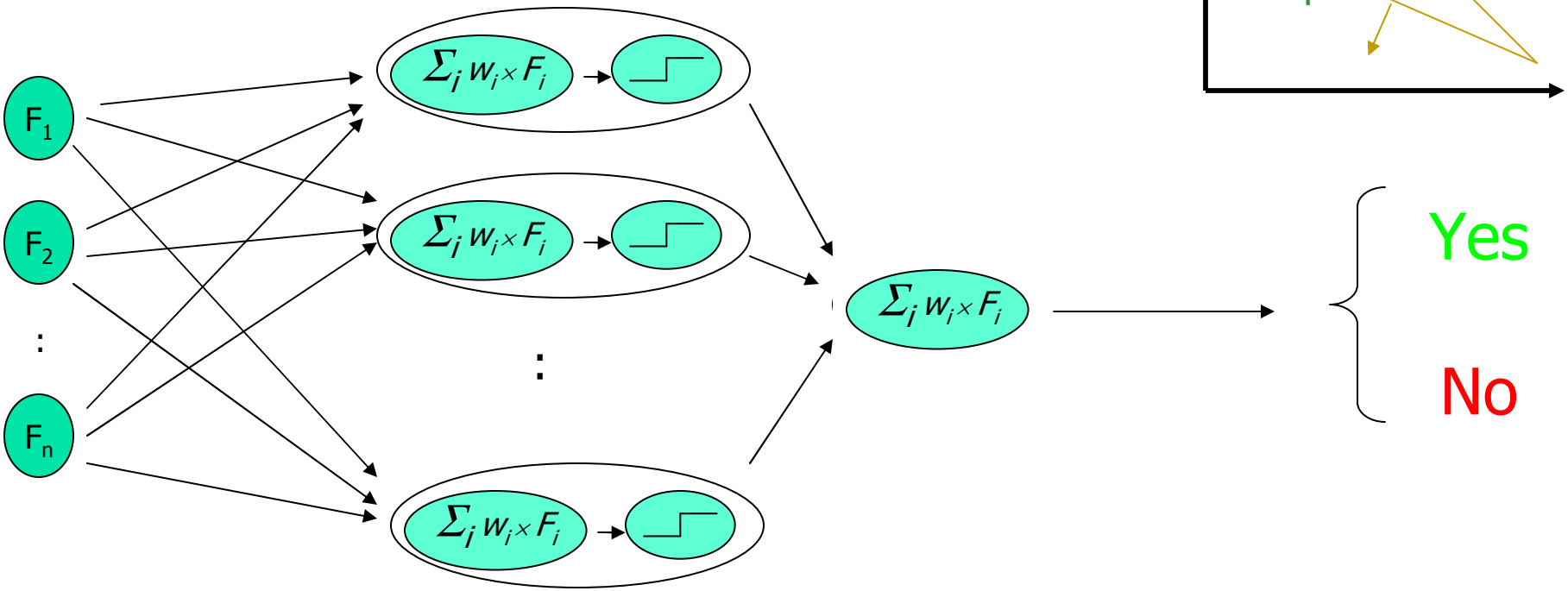
- Why not use *SET* of
  connected Linear Separators?



Skip

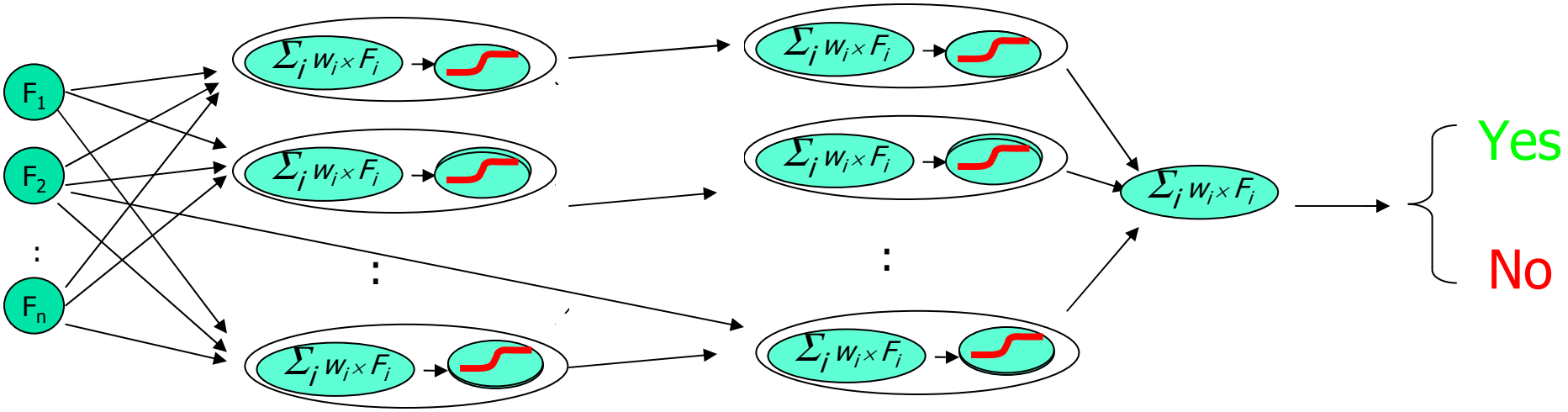
Artificial Neural Nets

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



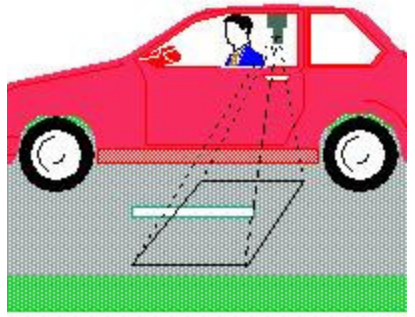
ANNs: Architecture

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



Uses of Artificial Neural Nets

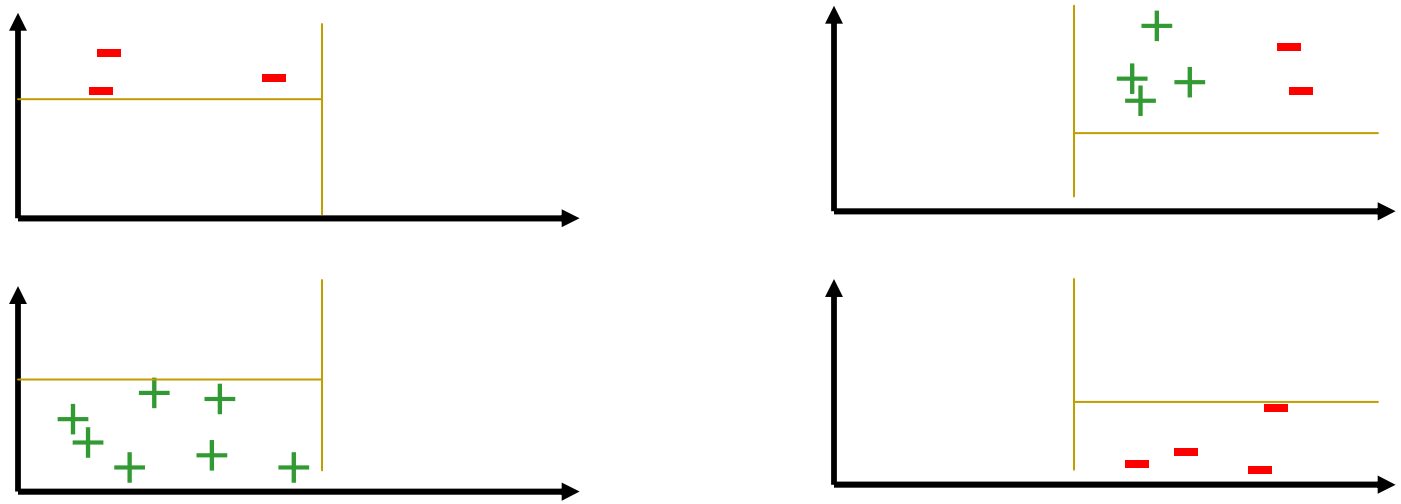
Learning to …



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

Algorithm 3: Decision Trees

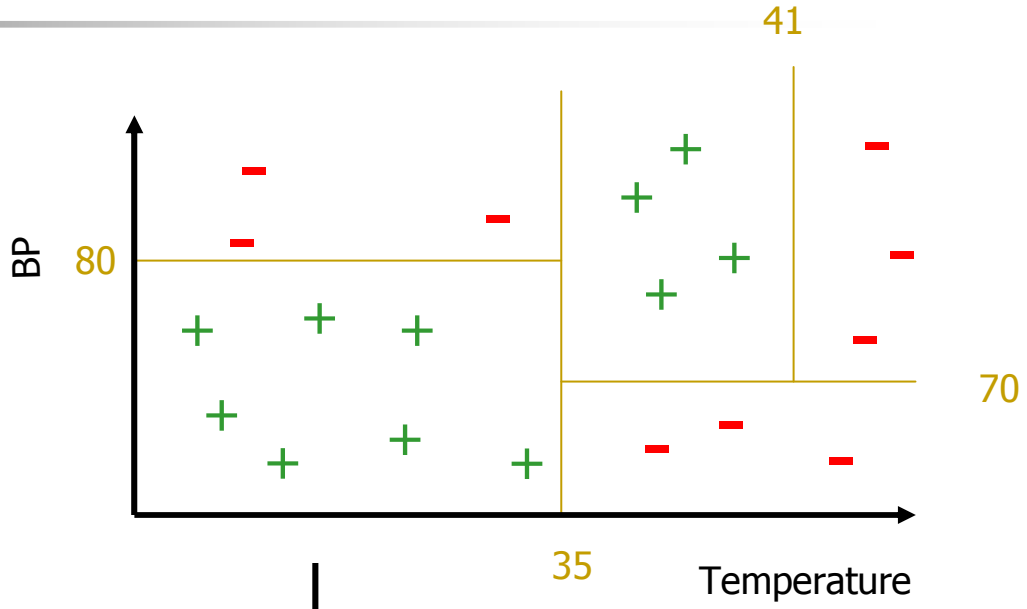
- Given data, decide on best *first* split



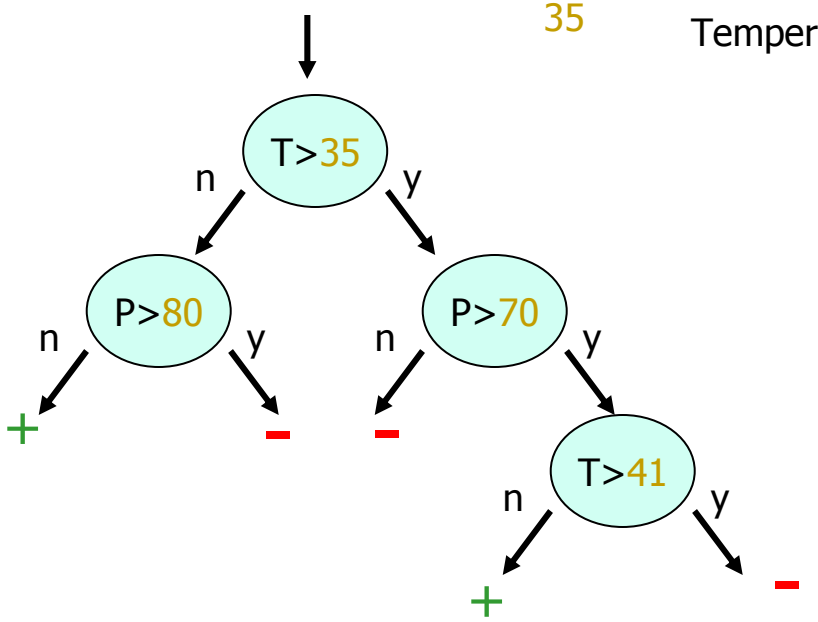
- Then consider each subset of data:
  - decide on its best split
- Recur… until "purity"

Alg 3: Decision Trees

- Consider data:

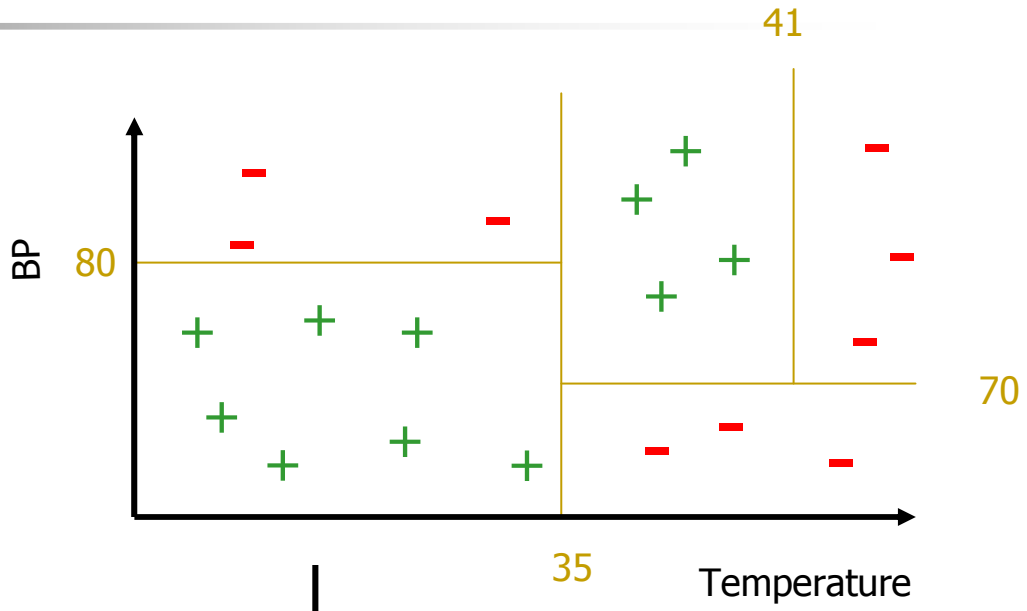


- "Hierarchical Split"
  - Divide and conquer

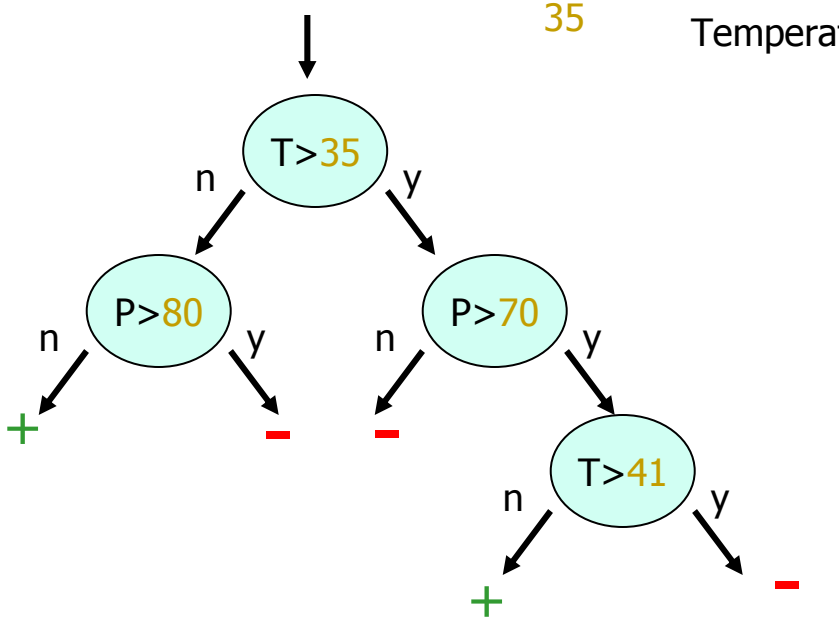


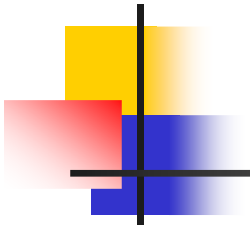
Alg 3: Decision Trees

- Partitioned data:



- "Hierarchical Split"
  - Divide and conquer





Issues ⇒ Demo

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

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

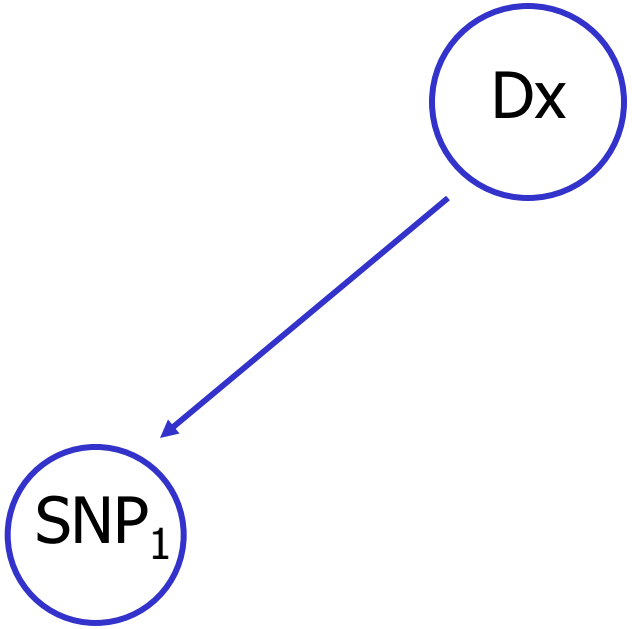
Alg 4: Naïve Bayes

- First-order stats:
  - P(SNP_1 = 1 | "Dx=+")
  - P(SNP_1 = 2 | "Dx=+")
  - P(SNP_1 = 3 | "Dx=+")

  +
  - P(SNP_1 = 1 | "Dx=–")
  - P(SNP_1 = 2 | "Dx=–")
  - P(SNP_1 = 3 | "Dx=–")

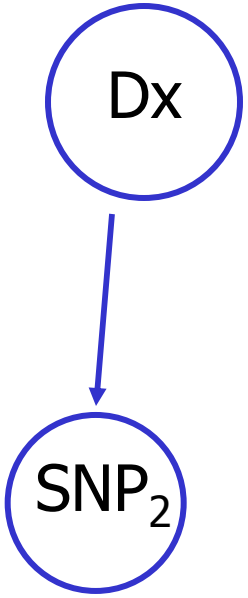
- Similarly for SNP_2, SNP_3, … SNP_{53}

Naïve Bayes, con't



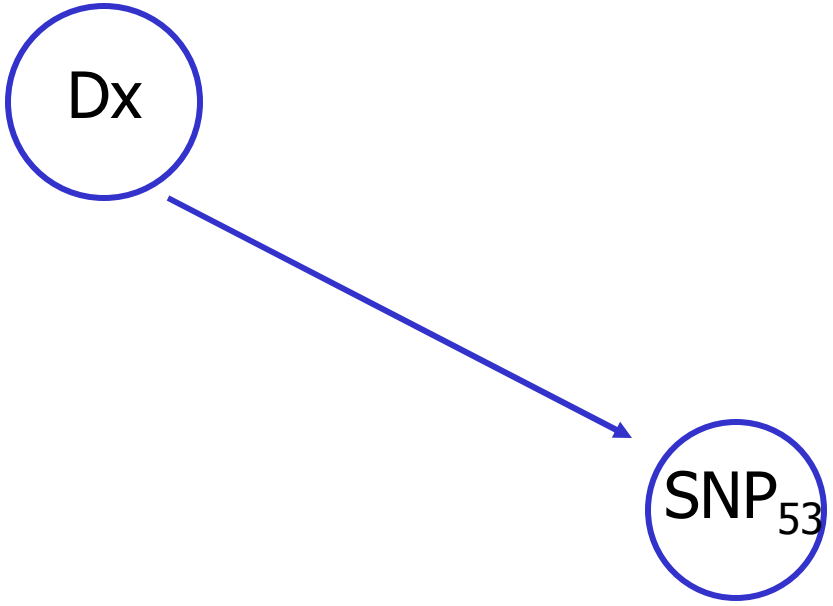
b	P(SNP$_1$=1 | Dx=b)	P(SNP$_1$=2| Dx=b)	P(SNP$_1$=3 | Dx=b)
+	0.05	0.92	0.03
--	0.80	0.19	0.01

Naïve Bayes, con't



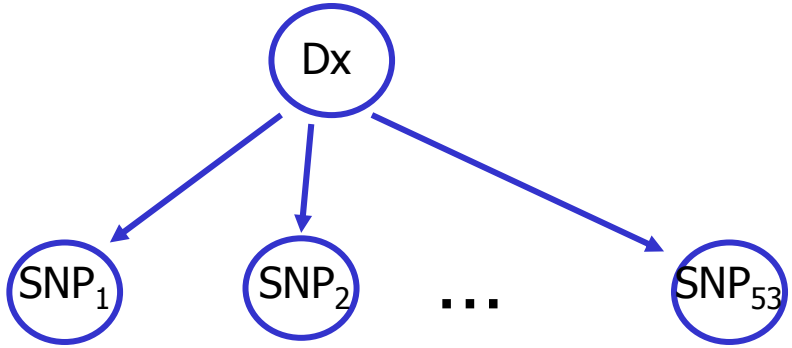
b	P(SNP$_2$=1 | Dx=b)	P(SNP$_2$=2 | Dx=b)	P(SNP$_2$=3 | Dx=b)
+	0.15	0.05	0.80
--	0.73	0.10	0.17

Naïve Bayes, con't



b	P(SNP$_{53}$=1 | Dx=b)	P(SNP$_{53}$=2 | Dx=b)	P(SNP$_{53}$=3 | Dx=b)
+	0.90	0.05	0. 05
--	0.70	0.20	0.10

Naïve Bayes, con't



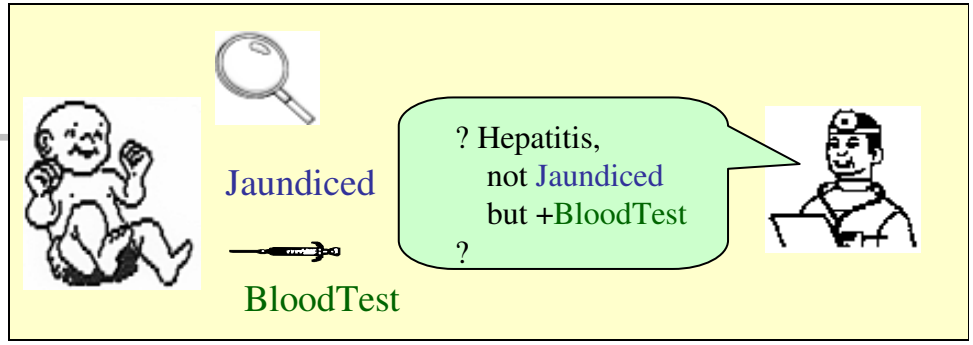
$$P(+b \mid s_1, s_2, \ldots, s_{53}) = \frac{1}{z} P(+b) \Pi_i P(SNP_i = s_i \mid +b)$$

$$P(-b \mid s_1, s_2, \ldots, s_{53}) = \frac{1}{z} P(-b) \Pi_i P(SNP_i = s_i \mid -b)$$

Answer: Take larger of

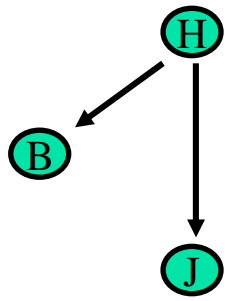
$$P(+b)\ \Pi_i P(SNP_i = s_i \mid +b)$$
$$P(-b)\ \Pi_i P(SNP_i = s_i \mid -b)$$

Classification



- Which is more likely: +h vs –h ?
- Given independencies:
  + values:

h	P(+b | h)	P(-b | h)
1	0.95	0.05
0	0.03	0.93



P(+h)	P(-h)
0.05	0.95

h	P(+j |h)	P(-j |h)
1	0.8	0.2
0	0.3	0.7

- argmax_h P(h | +b, –j)
  = argmax_h P(h) x P(+b | h) x P(–j | h)
  = argmax_h { 0.05 x 0.95 x 0.2, 0.95 x 0.03 x 0.7 }

-h as 0.0095 < 0.01995

"Naïve Bayes"

- Classification Task:

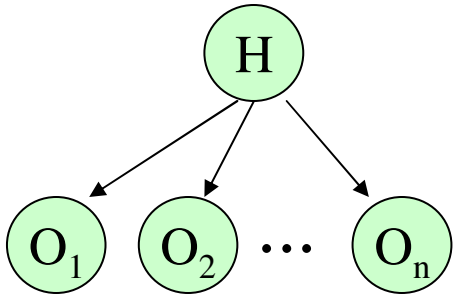
  Given $\{ O_1 = v_1, \ldots, O_n = v_n \}$

  Find h_i that maximizes $P(H = h_i \mid O_1 = v_1, \ldots, O_n = v_n)$

- Given

$$P(H = h_i)$$

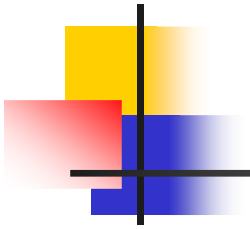
$$P(O_j = v_k \mid H = h_j)$$



$$\text{Independent:} \quad P(O_j \mid H, O_k, \ldots) = P(O_j \mid H)$$

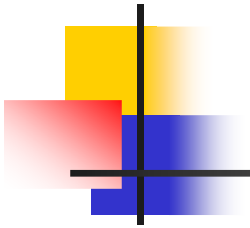
$$P(H = h_i \mid O_1 = v_1 \ldots, O_n = v_n) = \frac{1}{\alpha} P(H = h_i) \prod_j P(O_j = v_j \mid H = h_i)$$

- Find argmax $\{h_i\}$



Other Algorithms

- Nearest Neighbor
- 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. How to evaluate a classifier?
    5. Overfitting
    6. Computational Efficiency
    7. Imbalanced data (fraud detection)
    8. Non-IID tuples (stock market, temporal)
- Current Research

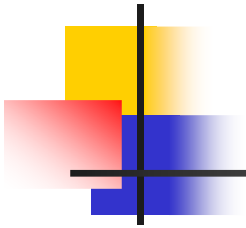
1. Goal of Learning?

	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?



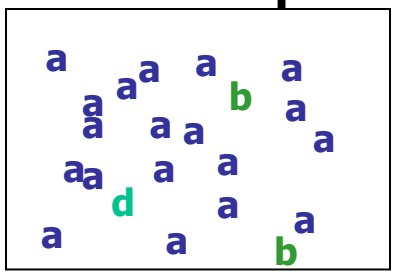
2. Why should Learning work?

- *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?

- Overall Population



	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 C that does well on S:
  - As S includes a b,

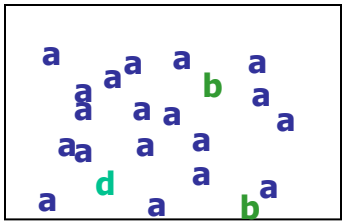
$$\begin{cases} C(a) = \text{No} \\ C(b) = \text{Yes} \end{cases}$$

- Notice d not is S

$\Rightarrow$ $C(d) = ??$

How good is Classifier *C*?

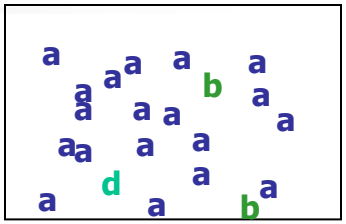
- To evaluate *C*
  - Draw new patient, $x \in$
  - Compute $C(x)$
  - Correct?



$$\begin{cases} C(\text{a}) = \text{No} \\ C(\text{b}) = \text{Yes} \end{cases}$$

- Given true distribution,
  - *expect* x = 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*
2. If *x* occurs rarely P(*x*) ≈ 0
   - doesn't matter if *C* is wrong!

a a a b a a

- Even good classifiers are wrong occasionally…

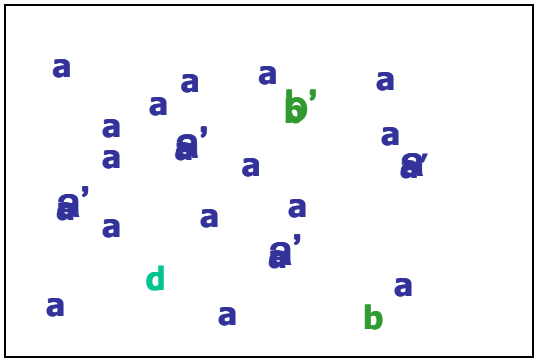
Populations

- 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...

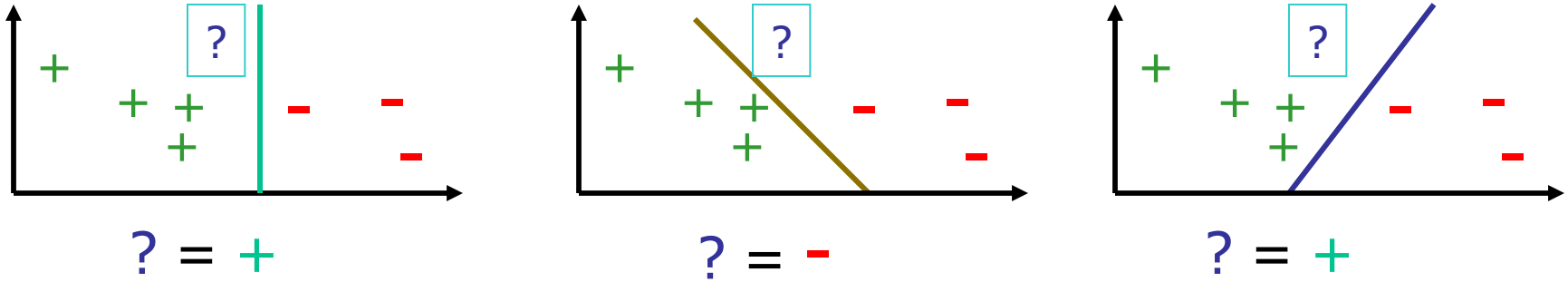


	F_1	F_2	...	F_n	Class
a =	35	95	...	3	No
a′ =	36	95	...	2.1	No
b =	22	80	...	-2	Yes
b′ =	22	78	...	-2.3	Yes
d =	10	50	...	1.9	No

- 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 ...

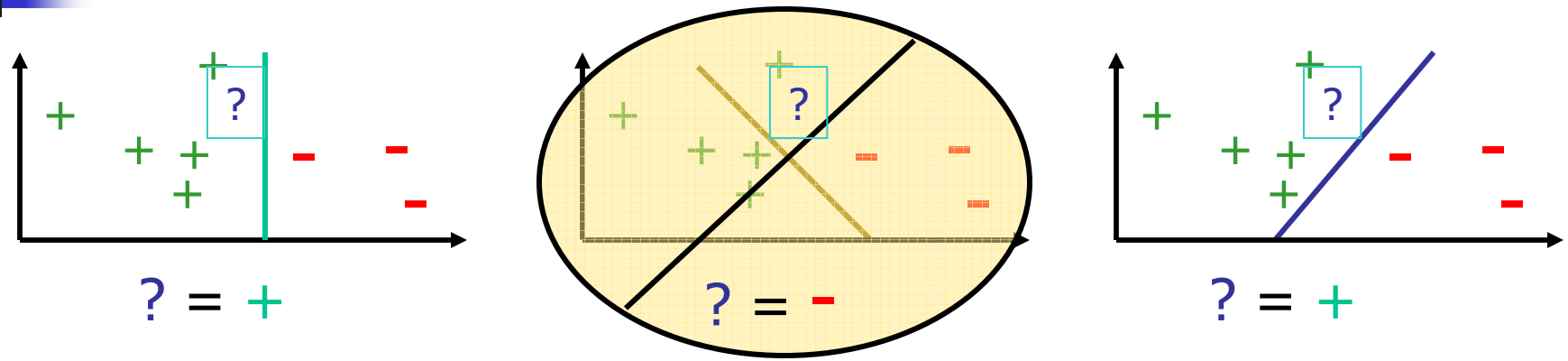
3. How much training data?

- What is best linear separator for…

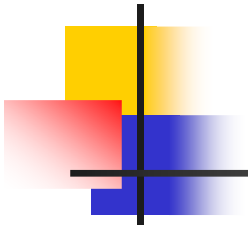


- Makes a difference: what is "?" ?
- Learning gets easier with more training data…

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. How to Evaluate a Classifier ?

Labeled Training Data

SNP1	SNP2	SNP3	...	SNP53	Dx?
G/A	C/C	T/T	...	T/C	**No**
A/A	C/C	A/T	...	T/T	**Yes**
A/A	C/T	A/A	...	T/T	**Yes**
:	:	:		:	**:**
G/A	C/T	A/A	...	T/T	**No**

TRAIN

SNP1	SNP2	SNP3	...	SNP53	Dx?
C/G	A/G	T/T	...	T/C	No
T/C	C/C	A/A	...	T/T	Yes
:	:	:		:	:
G/A	T/C	G/G	...	T/C	No

Learner

TEST

SNP1	SNP2	SNP3	...	SNP53
G/A	C/C	T/T	...	T/C
A/A	C/C	A/T	...	T/T
:	:	:		:
G/A	C/T	A/A	...	T/T

Classifier

Dx?
No
No
:
Yes

Training Set Error
... too optimistic

4. How to Evaluate a Classifier ?

Labeled Training Data

SNP1	SNP2	SNP3	...	SNP53	Dx?
G/A	C/C	T/T	...	T/C	**No**
A/A	C/C	A/T	...	T/T	**Yes**
A/A	C/T	A/A	...	T/T	**Yes**
:	:	:		:	**:**
G/A	C/T	A/A	...	T/T	**No**

TRAIN

SNP1	SNP2	SNP3	...	SNP53	Dx?
C/G	A/G	T/T	...	T/C	No
T/C	C/C	A/A	...	T/T	Yes
:	:	:		:	:
G/A	T/C	G/G	...	T/C	No

Learner

TEST

SNP1	SNP2	SNP3	...	SNP53
G/A	C/C	T/T	...	T/C
A/A	C/C	A/T	...	T/T
:	:	:		:
G/A	C/T	A/A	...	T/T

Classifier

Dx?
No
No
:
Yes

Training Set Error
... too optimistic

How to Evaluate a Classifier ?

Labeled Training Data

SNP1	SNP2	SNP3	...	SNP53	Dx?
G/A	C/C	T/T	...	T/C	**No**
A/A	C/C	A/T	...	T/T	**Yes**
A/A	C/T	A/A	...	T/T	**Yes**
:	:	:		:	**:**
G/A	C/T	A/A	...	T/T	**No**

TRAIN

SNP1	SNP2	SNP3	...	SNP53	Dx?
C/G	A/G	T/T	...	T/C	No
T/C	C/C	A/A	...	T/T	Yes
:	:	:		:	:
G/A	T/C	G/G	...	T/C	No

Learner

TEST

SNP1	SNP2	SNP3	...	SNP53
G/A	C/C	T/T	...	T/C
A/A	C/C	A/T	...	T/T
:	:	:		:
G/A	C/T	A/A	...	T/T

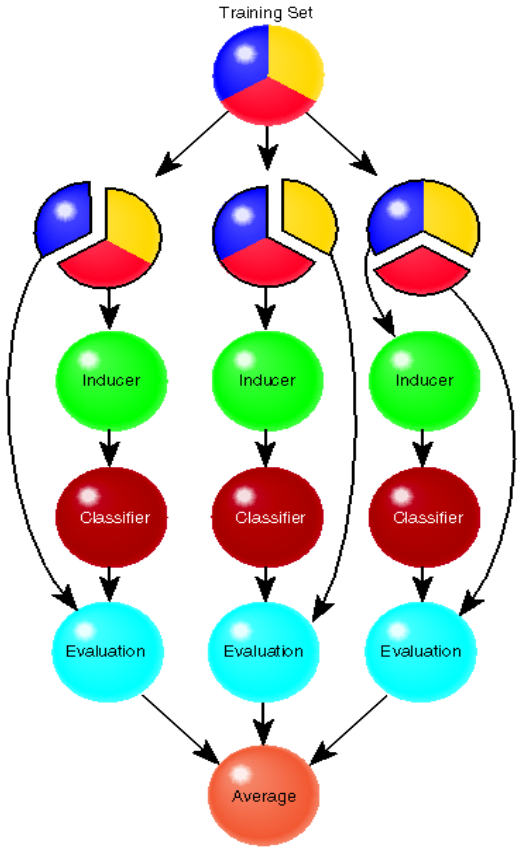
Classifier

Dx?
No
No
:
Yes

Simple Hold-out Set Error
... slightly pessimistic

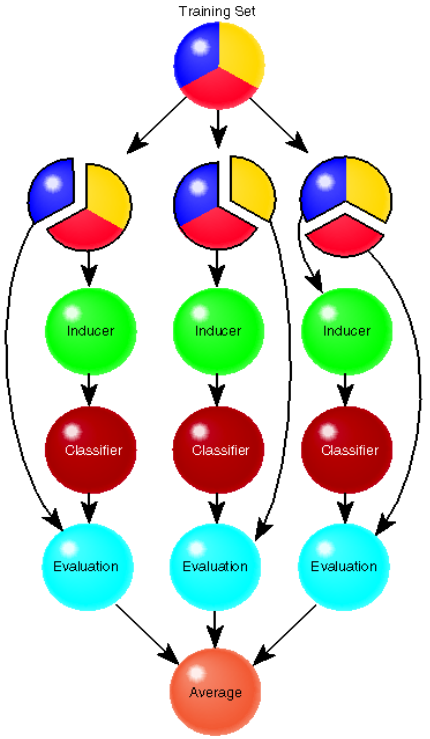
How to Evaluate a Classifier ?

- K-fold Cross Validation
  - Eg, K=3
- Not as pessimistic
  - every point is test example, once



Estimating Error: Cross Validation

- "Cross-Validation"



CV(data S, alg L, int k)
  Divide S into k disjoint sets $\{ S_1, S_2, \ldots, S_k \}$
  For i = 1..k do
    Run L on $S_{-1} = S - S_i$
    obtain $h_i := L(S_{-i})$
    Evaluate h_i on S_i
      $err_{S_i}(h_i) = 1/|S_i| \sum_{\langle x,t \rangle \in S_i} [h_i(x) - t]^2$
  Return Average $1/k \sum_i err_{S_i}(h_i)$

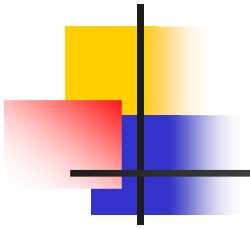
$\Rightarrow$ Less Pessimistic

as train on $(k - 1)/k$ $|S|$ of the data

Comments on Cross-Validation

- Every point used as
  Test 1 time, Training k – 1 times
- Computational cost for *k*-fold Cross-validation … linear in *k*
- Should use "balanced CV"
  If class c_i appears in m_i instances,
  insist each S_k include $\approx \frac{1}{k}\frac{m_i}{|S|}|$ such instances
- Use **CV(S, L, k)** as ESTIMATE of true error of L(S)
  Return L(S) and CV(S, L, k)
- Leave-One-Out-Cross-Validation k = m !
  - eg, for Nearest-Neighbor
- Notice different folds are correlated
  as training sets overlap: (k-2)/k unless k=2
- 5 × 2-CV
  - Run 2-fold CV, 5 times. . .

Can use CV to estimate parameters in general!



To Form k *Balanced* Folds

1. Partition the data S based on the class:
   - subset S_+ has all the positive instances,
   - subset S_- has all the negative instances.

2. Randomly partition each subset into k folds --
   - $S_+ = U \{ S_{+1}, \ldots, S_{+k} \}$
   - $S_- = U \{ S_{-1}, \ldots, S_{-k} \}$

3. $S_j = S_{+j} \cup S_{-j}$ for j=1..k

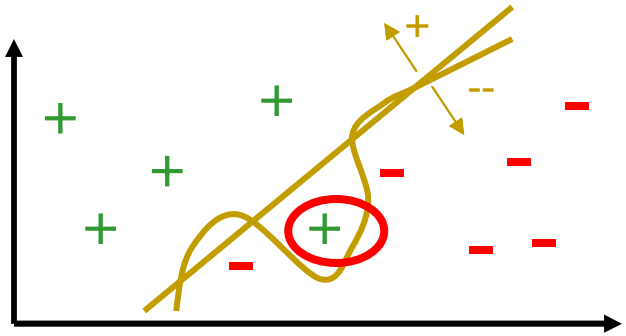
Summary of Classifier Results

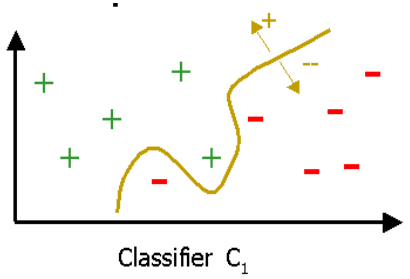
Classifier Accuracy	% Correct
Simple Logistic	65.85
Support Vector Machine	70.73
Decision Tree	68.29
Naïve Bayes	60.98
"Just say No"	*65.85*

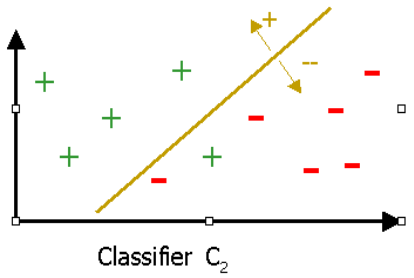
… using 51 SNPs with 10-fold cross validation

Why so bad? Overfitting!

- What is best separator for…
- Compare:





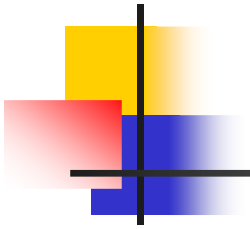


- Sometimes appropriate to IGNORE details of training data
  - perhaps some training data points are mislabeled !
  - … or **some features are irrelevant, misleading …**
- Solution:
  - Feature selection

5. "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

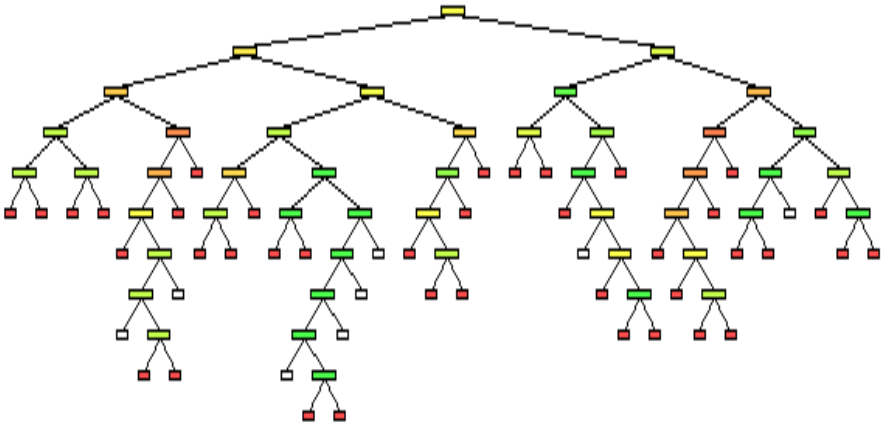


Example – continued

- 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) …

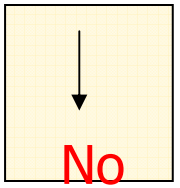
Decision Tree Results

- Standard decision tree learner:



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

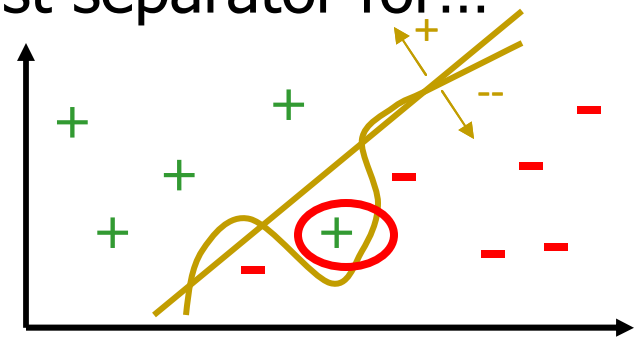
- Optimal decision tree:



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

Overfitting

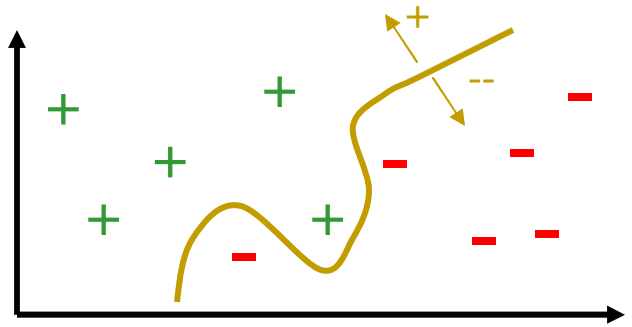
- Some features are not helpful
- Data often noisy
  - typos in recording, error in equipment, human error…
- What is best separator for…



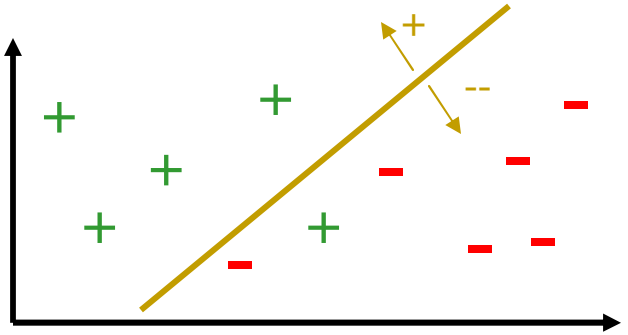
- Sometimes:
  Appropriate to IGNORE details of training data
  - Here: one training data point is mislabeled !
- Simpler hypothesis often better classifier!
  - eg, LINEAR Separator

Overfitting

- Compare…



Classifier C_1



Classifier C_2

- C_1 appears better (on training data) than C_2, but C_2 is actually better
- *Overfitting* !
- To address this… reduce dimensionality…

Reduce dimensions…

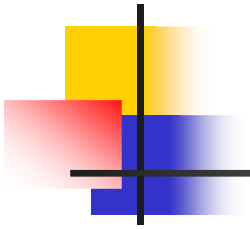
- Principle Component Analysis
- Feature Selection?
- Sort features based on *Information Gain*

$$I(A,C) = \sum_{c,a} P(c,a) \log \frac{P(c,a)}{P(c)\ P(a)}$$

  - Notice:
    - 0 if attribute A is NOT correlated with class C
    - Positive if correlated
  - Considers each attribute independent of others!

  Take top k features…

- …



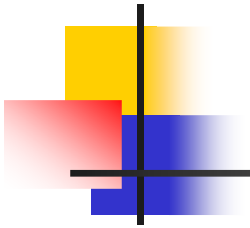
How many features?

- Perhaps try each value k=1,2,3, …
  and see how well each classifier does, on test set

- No!!
  Must NOT use the test set to help learner
  - by selecting the number of features…

  $\Rightarrow$ no longer unbiased!!!

- Test set only unbiased if you *never never never* do any *any any any* learning on the test data



Significance

- Found 8 most "informative" SNPs:
  - {gn_3001__xrcc3, gn_3040__cyp2d6_4, gn_2442__mlh1_1, gn_2469__brca2_12_13, gn_3012__rad51, gn_3010__nbs1, gn_961__brca1_5_a201g, gn_3002__xrcc3}
- Reasonable… associated with cell damage
- Classifier (using 8 snps) was **78%** accuracy
- Is this significant?
  … especially given our complicated process?
- Suppose NO signal in data.
  Trivial to get ≈**65.9%**

Permutation Test

- Randomly rearrange LABELs of data…
  … so no signal left …
- Run thru same algorithm
- Get results (CV)

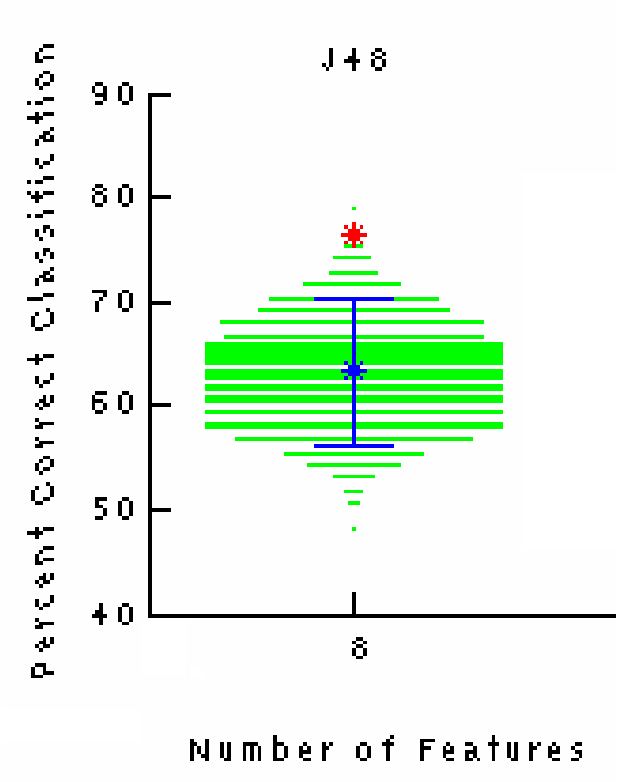
H8 = ???

	A	B	C	D	E	F	G	H
1	CTIP	955/BRCA1_1	1009/ESR1_1	1017/BCL2_1	2442/MLH1_1	2564/ERCC2_2	… … …	Label
2	LTP25	G/A	C/C	T/C	T/T	A/G	… … …	bleed
3	LTP26	G/A	C/C	T/C	T/T	G/G	… … …	bleed
4	LTP27	A/A	C/G	T/C	C/T	A/G	… … …	bleed
5	LTP28	G/A	C/C	T/T	C/C	A/G	… … …	noBleed
6	LTP29	A/A	C/C	T/C	T/T	G/G	… … …	bleed
7	LTP30	A/A	C/G	T/C	T/T	G/G	… … …	noBleed

Shuffle the labels!!

Significance

- One run might still do well.
  How about 4000 trials …
- Results of permutation tests:



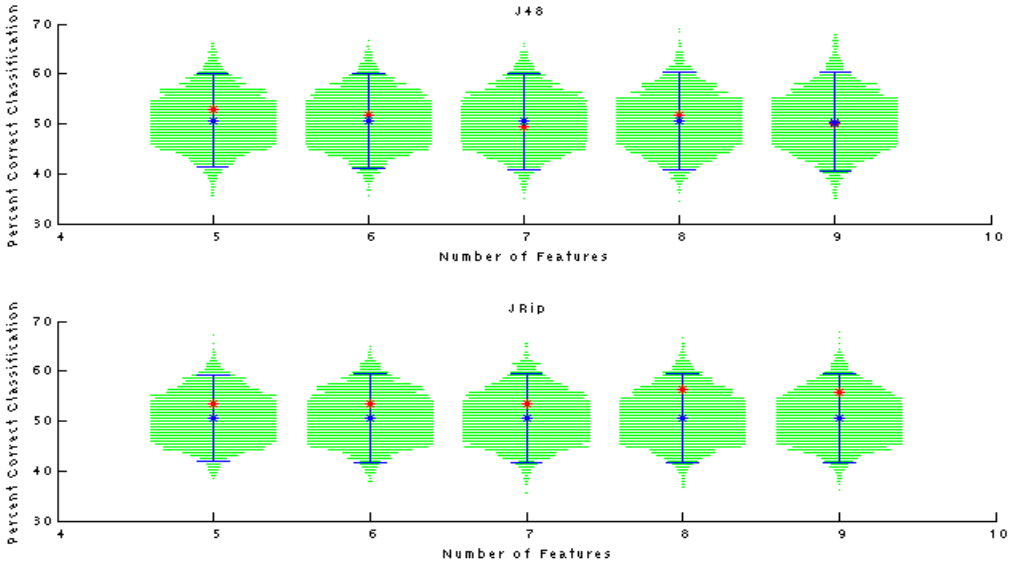
Conclusion
:
This
Works!

Prostate Study #1b

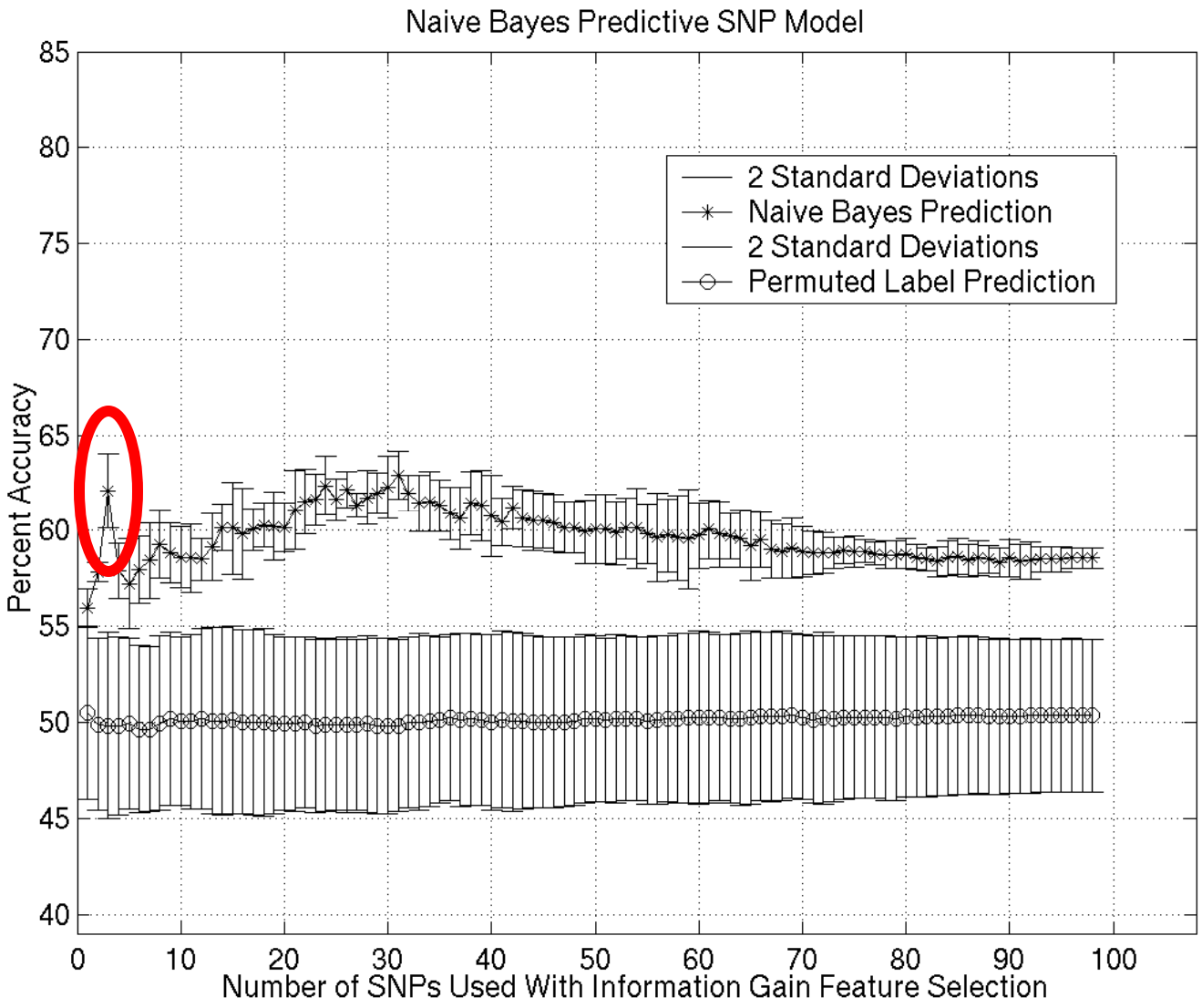
- Cancer vs. No Cancer

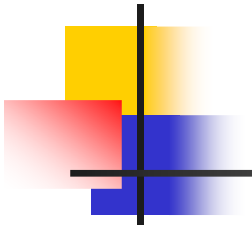
  … using same SNPs
- No correlation found!

  permutation tests result:



Accuracy… Permutation Tests



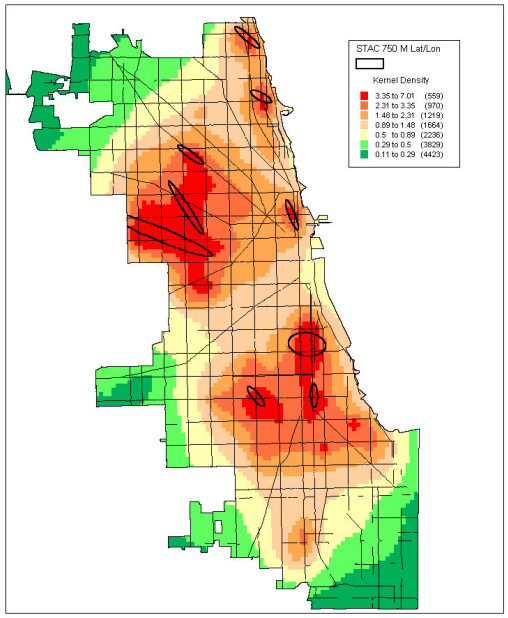


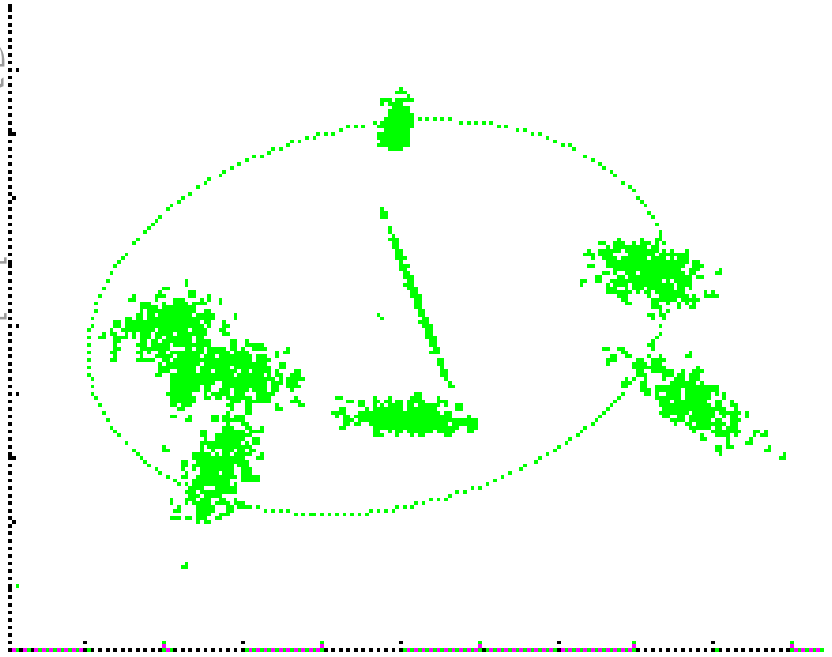
Outline

- Successful
- Basic ideas
  - Foundations
  - Algorithms
  - Statistical Issues
    1. Goal of Learning
    2. Why should Learning work?
    3. How much data is needed?
    4. How to Evaluate Classifier?
    5. Overfitting
    6. Computational Efficiency
    7. Imbalanced data (fraud detection)
    8. Non-IID tuples (stock market, temporal)
    9. Other types of learning
- Current research

Other Types of Learning

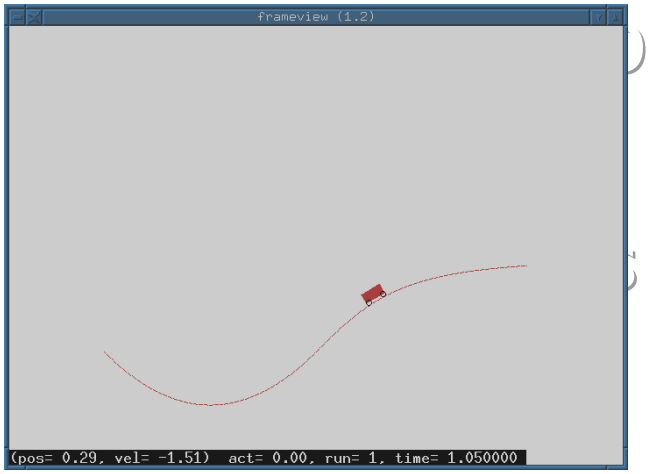
- Density Estimation
  - Learning Generative Model
  - Clustering

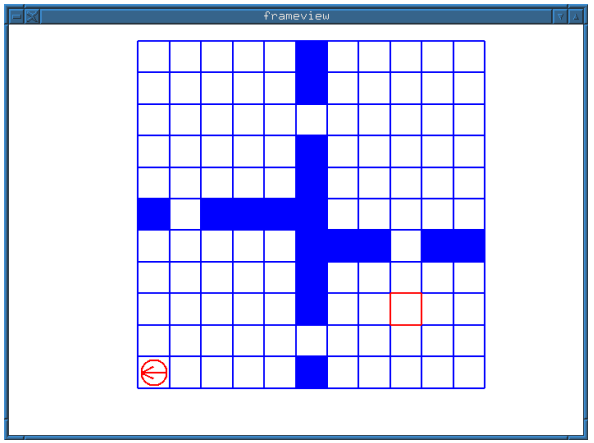




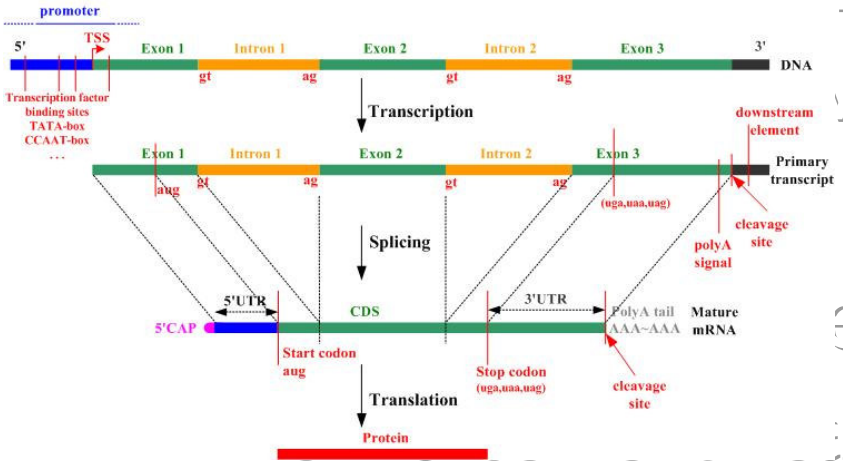
Other Types of Learning

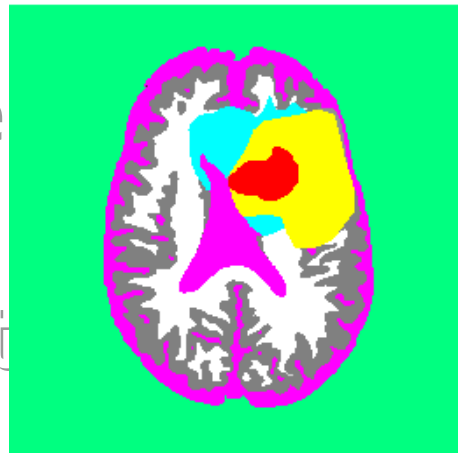
- Density Estimation
  - Learning Generative Model
  - Clustering
- **Learning Sequence of Actions**
  - **Reinforcement Learning**
  - on-IID Data



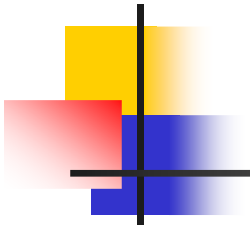


Other Types of Learning





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



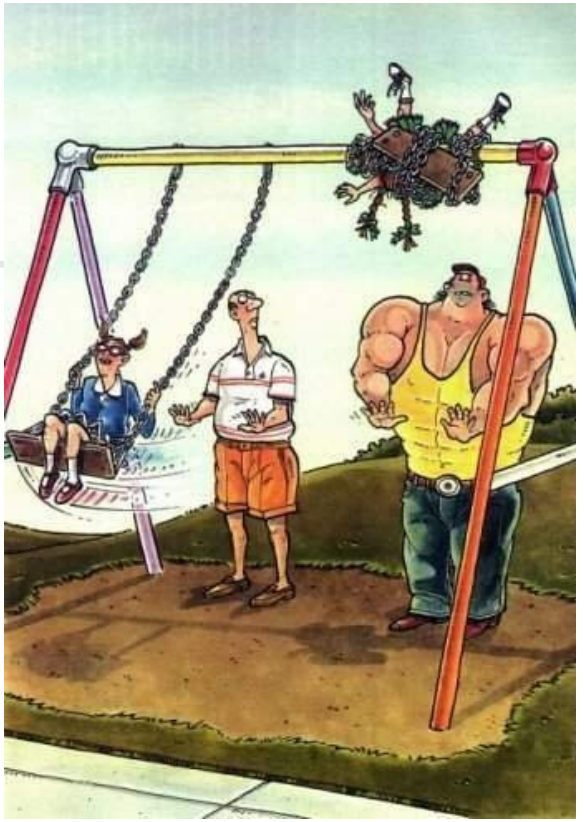
Other Types of Learning

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

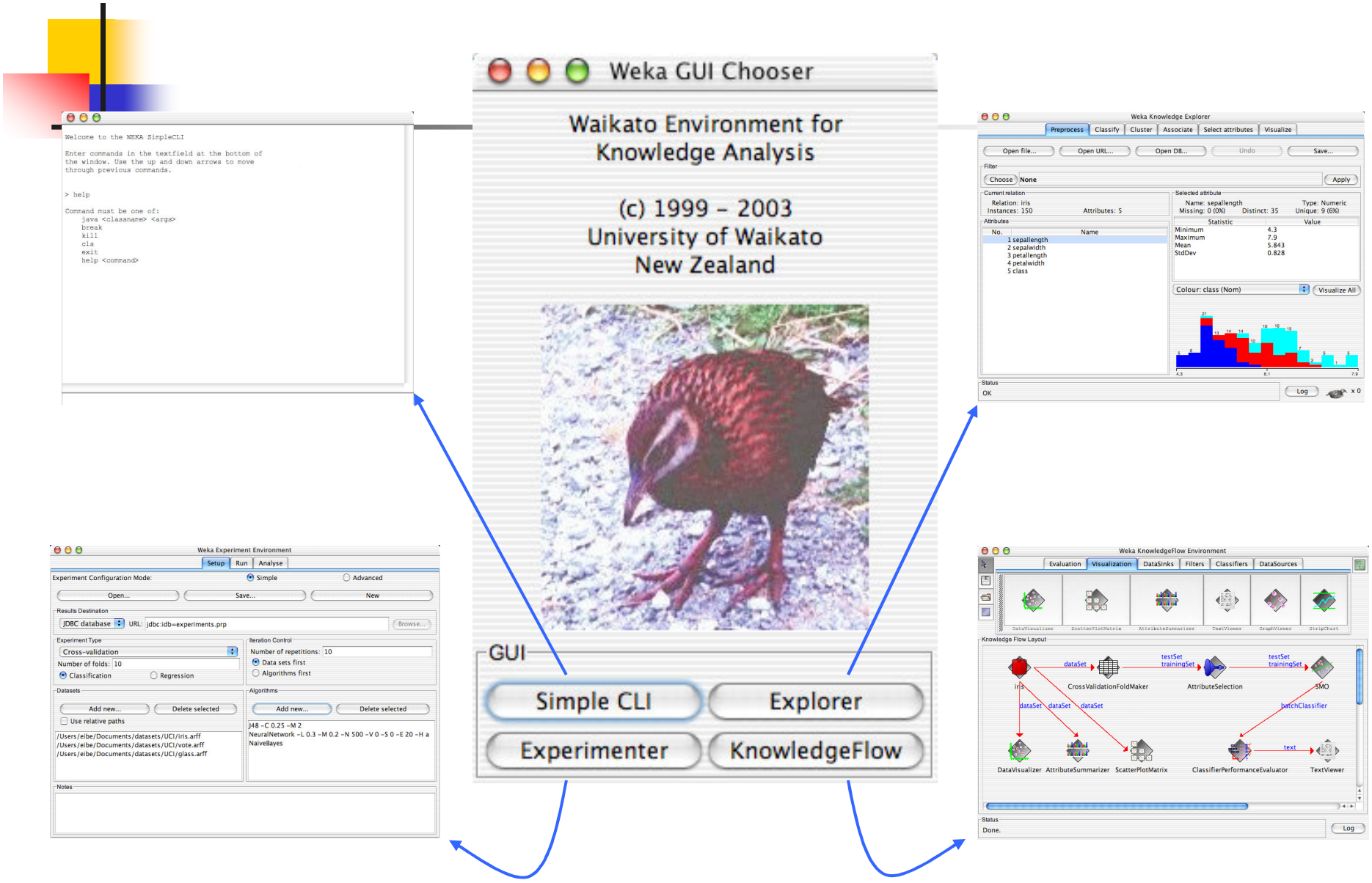
My Research I: Application Pull

- Brain Tumor Growth Prediction
- Proteome Analyst
- Human Metabolome Project
  - Inventory of ALL relevant small molecules …
- PolyomX
  - Patient-specific cancer treatment
- Understanding Microarray Data
- Complete-Web Recommendation System
  - Find/use "browsing patterns" to identify important *words*, then important *pages*…

My Research II: Technology Push



- Support-Vector Random Fields
- Budgeted Learning
- Computing Variance of Belief Net Response
  - Mixture Using Variance
- Learning Belief Nets
  - Learning Generative Structure
  - Learning Discriminative Structure
  - Learning Discriminative Parameters
- …


Weka GUI Chooser
Waikato Environment for Knowledge Analysis
(c) 1999 - 2003
University of Waikato
New Zealand
GUI
Simple CLI
Explorer
Experimenter
KnowledgeFlow
Weka Knowledge Explorer
Weka Experiment Environment
Weka KnowledgeFlow Environment

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**


Thank you!