UNION
COLLEGE

CSC321 Data Mining & Machine Learning

Prof. Nick Webb

webbn@union.edu

UNION COLLEGE

Data vs. information

- Society produces huge amounts of data
  - Sources: business, science, medicine, economics, geography, environment, sports, …
- Potentially valuable resource
- Raw data is useless: need techniques to automatically extract information from it
  - Data: recorded facts
  - Information: patterns underlying the data

UNION COLLEGE

Data mining

- Extracting
  - implicit,
  - previously unknown,
  - potentially useful

  information from data
- Needed: programs that detect patterns and regularities in the data
- Strong patterns $\Rightarrow$ good predictions
  - Problem 1: most patterns are not interesting
  - Problem 2: patterns may be inexact (or spurious)
  - Problem 3: data may be garbled or missing

UNION COLLEGE

Machine learning techniques

- *Algorithms for acquiring (structural) descriptions from examples*
- Structural descriptions represent patterns explicitly
  - Can be used to predict outcome in new situation
  - Can be used to understand and explain how prediction is derived (*may be even more important*)
- Methods originate from artificial intelligence, statistics, and research on databases

UNION COLLEGE

Learning Example

- Netflix: Predict how a viewer will rate a movie

- Based on what they have watched and rated we have:
  - A pattern
  - That is hard to describe mathematically
  - And there is data

UNION COLLEGE

Example

- Viewer, as a vector

Comedy	Horror	Drama	…	Steven Segal	Judy Dench	…	Length	Award	…
10	6	2	…	Yes	Yes	…	Short	Yes	…

- Movie, as a vector

Comedy	Horror	Drama	…	Steven Segal	Judy Dench	…	Length	Award	…
3	1	8	…	No	Yes	…	Long	No	…

UNION COLLEGE

To produce a rating

- Add up all the contributing features
- With associated weighting
- Result: Predicted weighting

- Where's the machine learning?

UNION COLLEGE

Machine Learning

- STARTS from the rating viewers give to movies
- Tries to find what factors are consistent with that rating
- Start by assigning RANDOM values to vectors
- Make small adjustments to vectors based on real rating
- For hundreds of thousands of ratings
- At the end, given a viewer vector, and a movie they haven't seen -> produce consistent rating

UNION COLLEGE

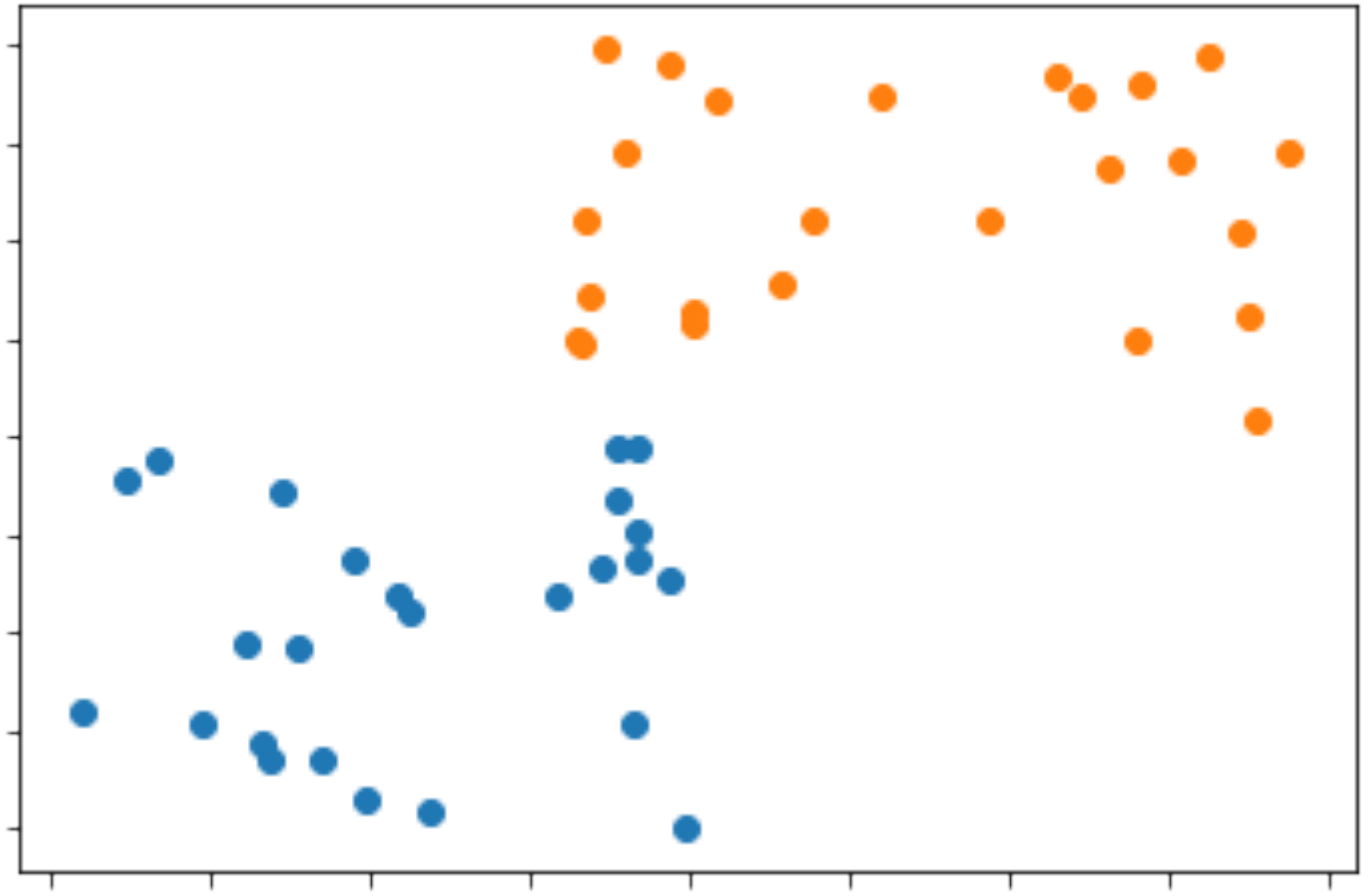
Regression vs. Classification

- Predicting a numeric score -> regression
- Can use (mostly) the same methods for classification -> Assigning a label

- Imagine some data with two classes

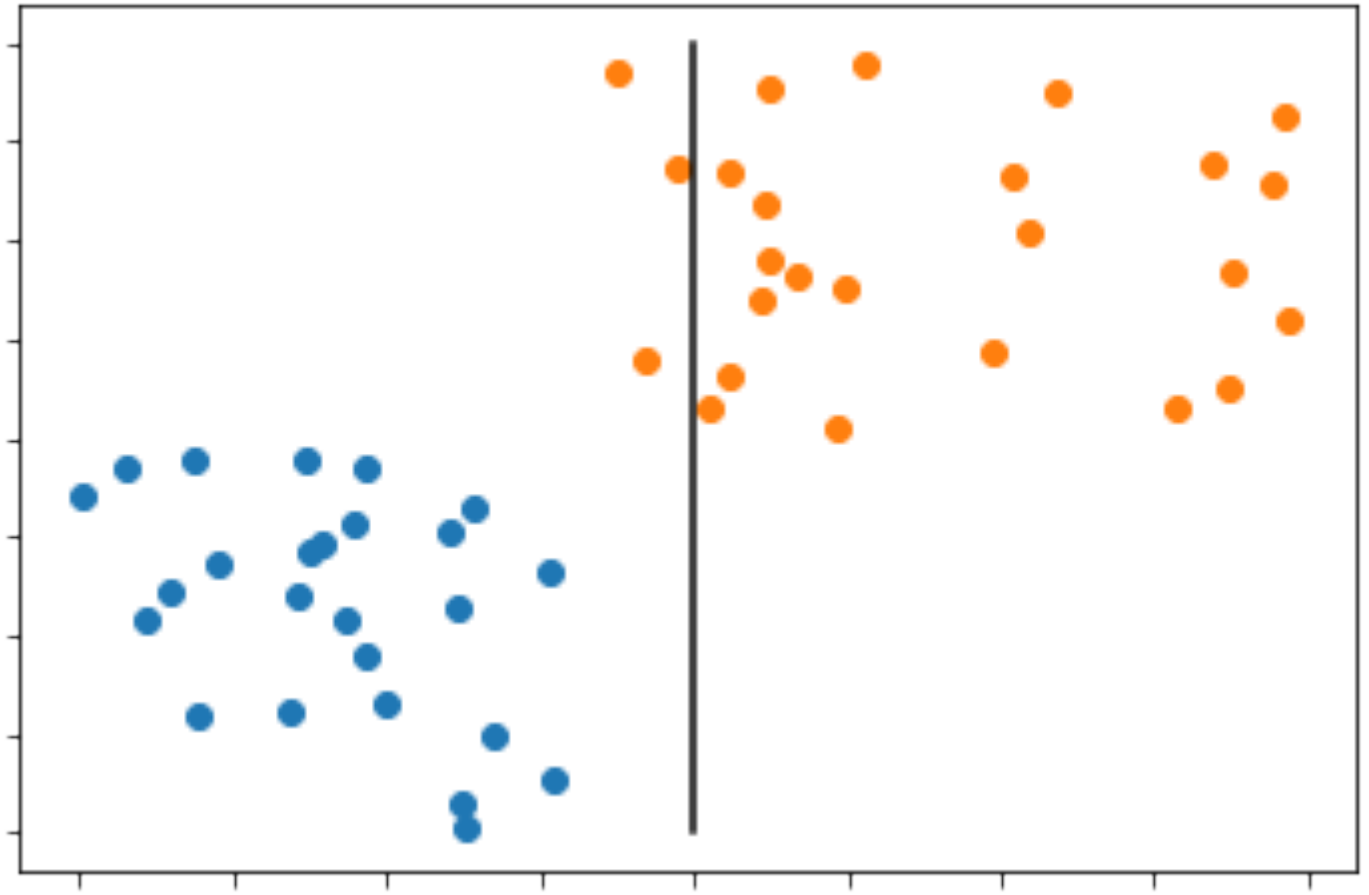
UNION COLLEGE

2 class data



UNION COLLEGE

Start with random separation



UNION COLLEGE

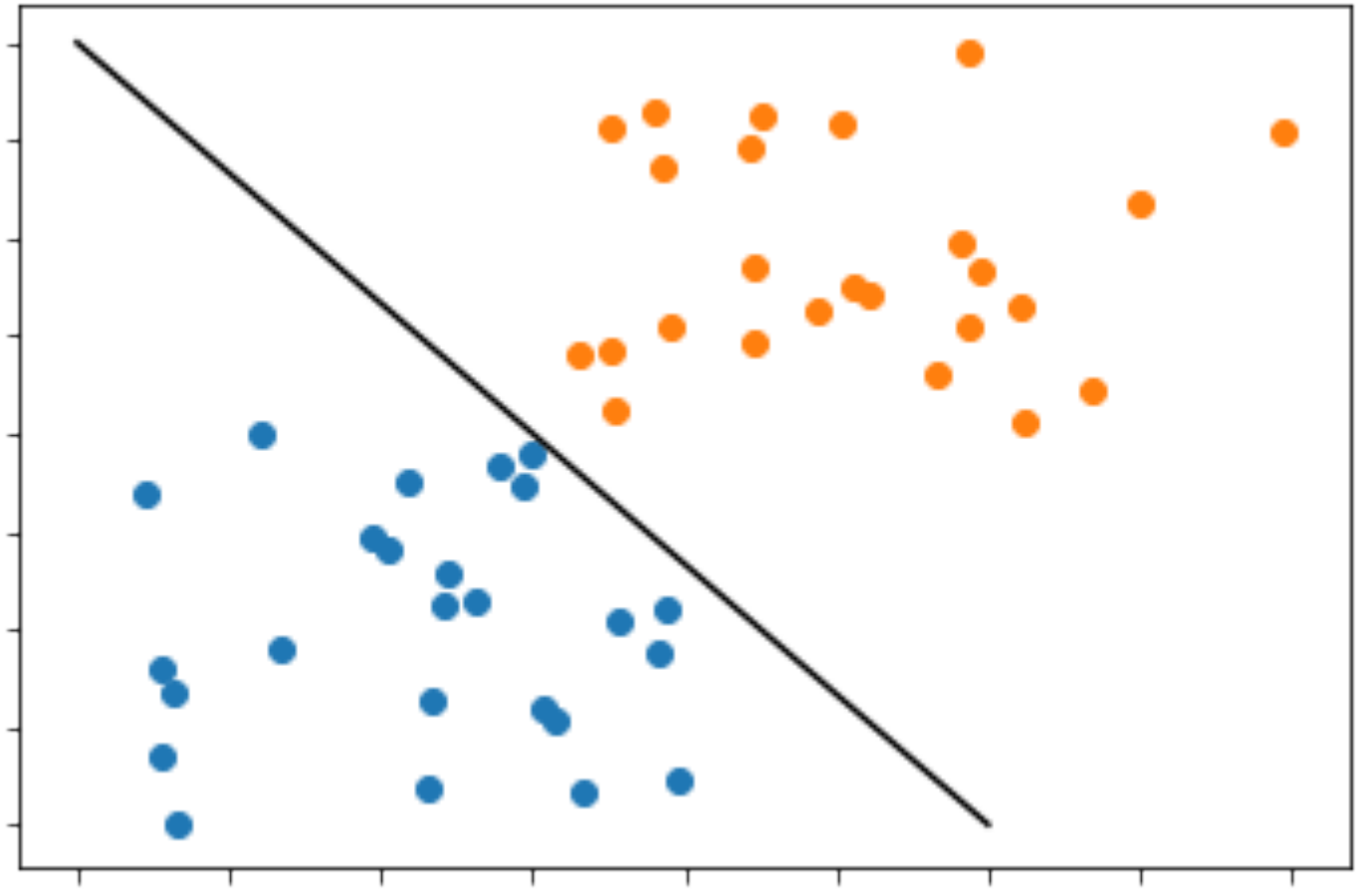
For each misclassified point

- Find the direction of misclassification
  - It should have been positive or
  - It should have been negative
- Adjust the weight on the feature VERY SLIGHTLY in that direction
- Repeat

- This is the perceptron learning algorithm

UNION
COLLEGE

Move line *incrementally*



UNION COLLEGE

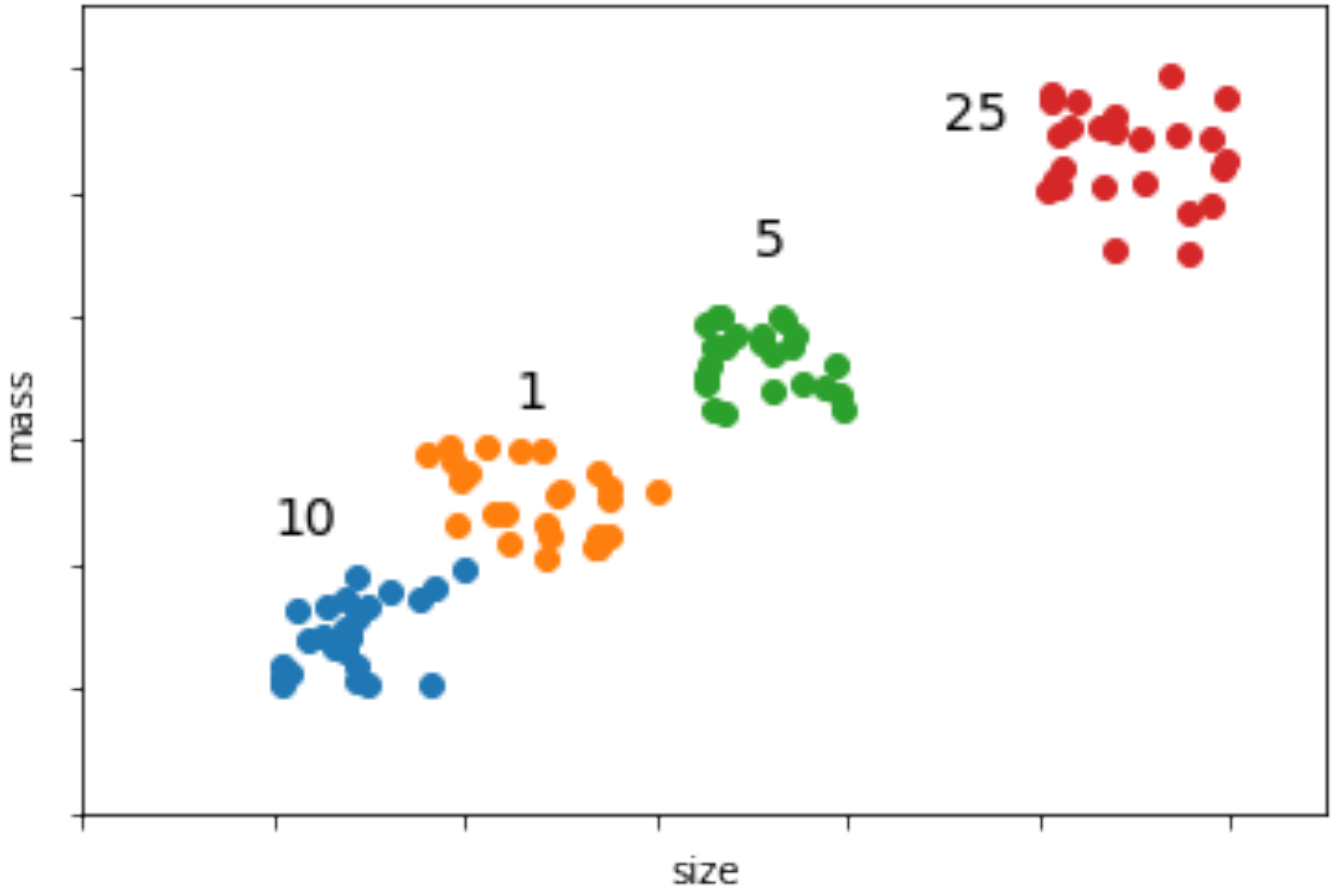
Learning problem

- Use a set of observations
- To uncover underlying process

- Supervised Learning
- Unsupervised Learning

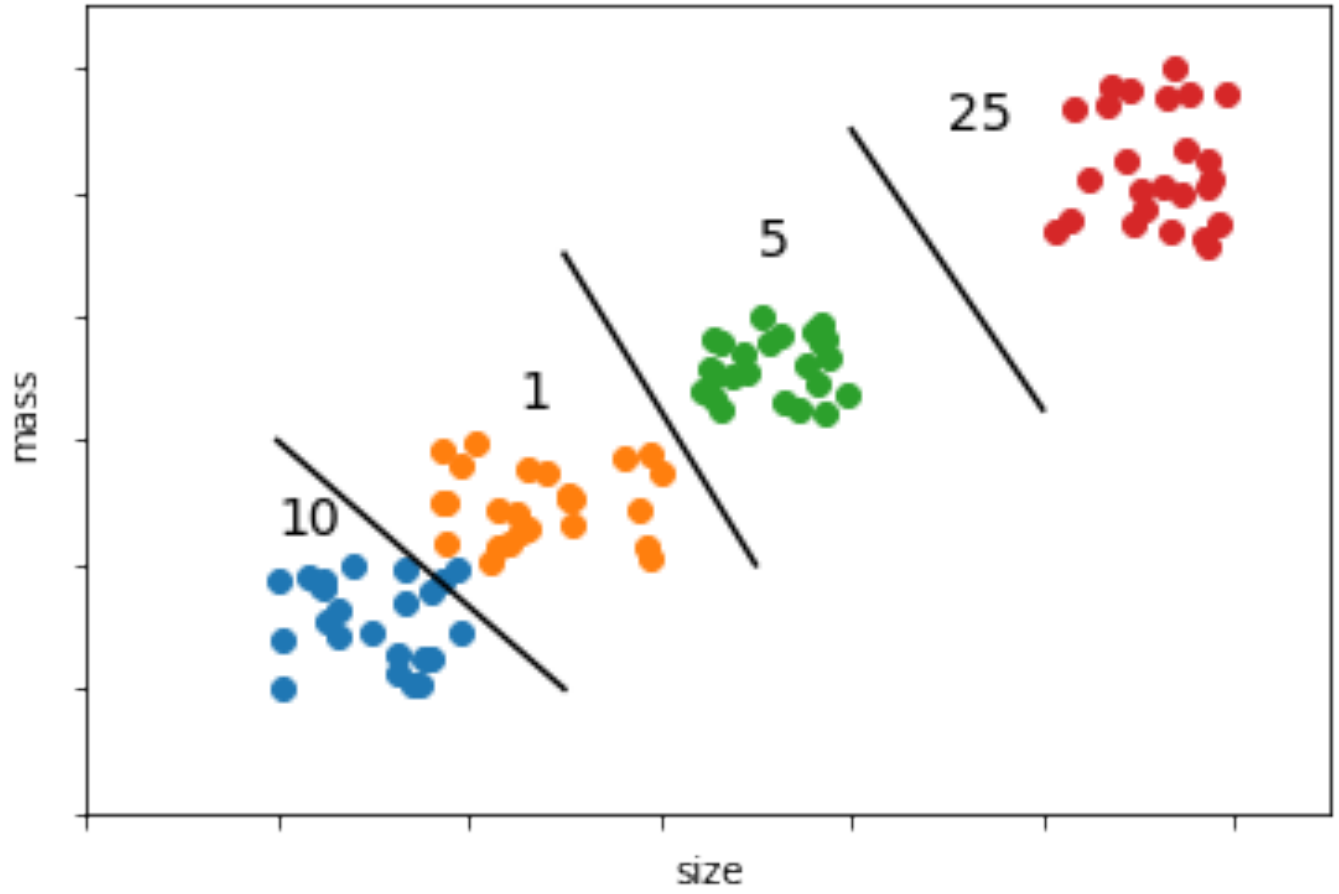
UNION COLLEGE

Supervised Learning



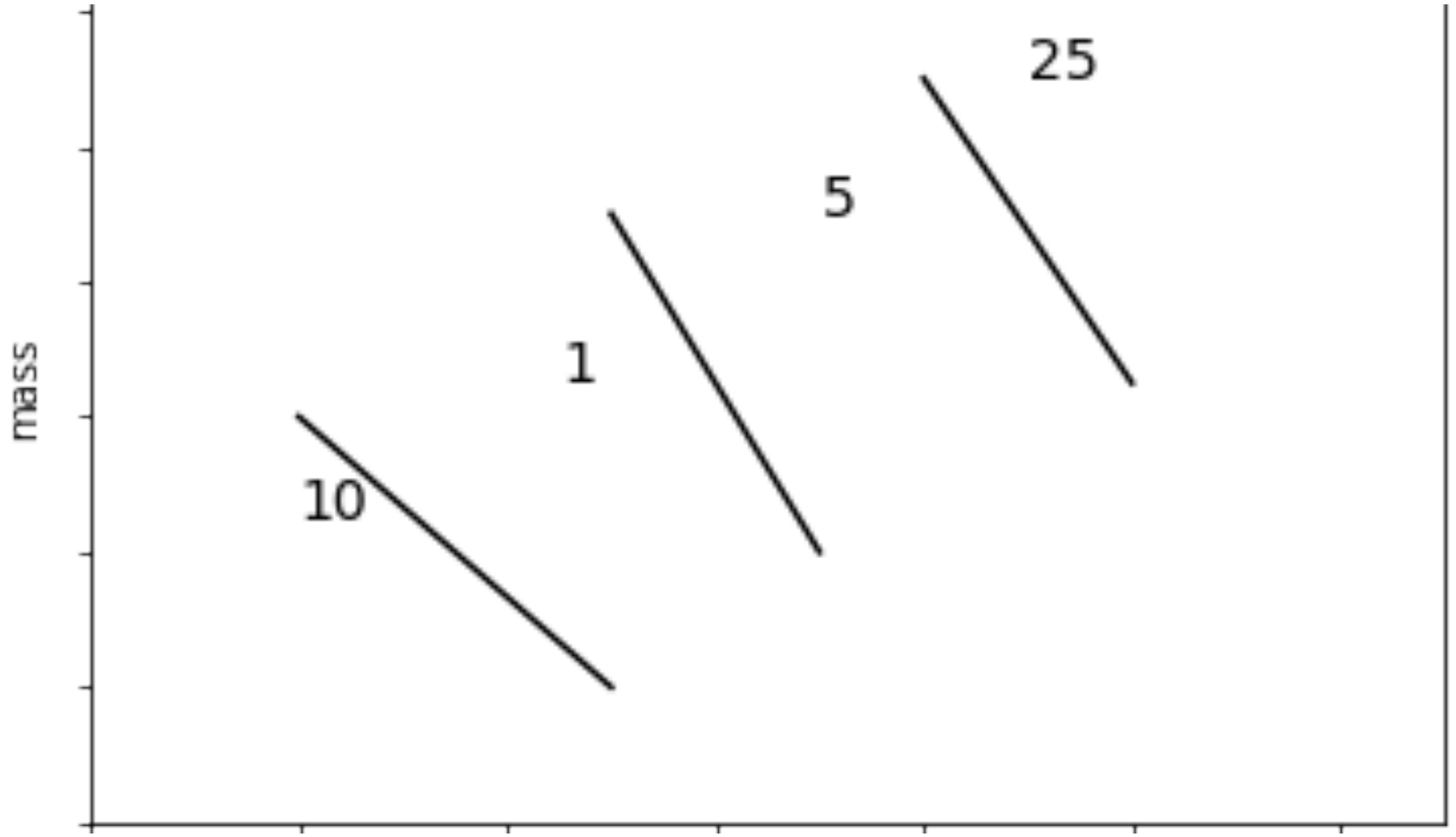
UNION
COLLEGE

Supervised Learning



UNION COLLEGE

Supervised Learning



UNION COLLEGE

Contact lens data

Age	Spectacle prescription	Astigmatism	Tear production rate	Recommended lenses
Young	Myope	No	Reduced	None
Young	Myope	No	Normal	Soft
Young	Myope	Yes	Reduced	None
Young	Myope	Yes	Normal	Hard
Young	Hypermetrope	No	Reduced	None
Young	Hypermetrope	No	Normal	Soft
Young	Hypermetrope	Yes	Reduced	None
Young	Hypermetrope	Yes	Normal	hard
Pre-presbyopic	Myope	No	Reduced	None
Pre-presbyopic	Myope	No	Normal	Soft
Pre-presbyopic	Myope	Yes	Reduced	None
Pre-presbyopic	Myope	Yes	Normal	Hard
Pre-presbyopic	Hypermetrope	No	Reduced	None
Pre-presbyopic	Hypermetrope	No	Normal	Soft
Pre-presbyopic	Hypermetrope	Yes	Reduced	None
Pre-presbyopic	Hypermetrope	Yes	Normal	None
Presbyopic	Myope	No	Reduced	None
Presbyopic	Myope	No	Normal	None
Presbyopic	Myope	Yes	Reduced	None
Presbyopic	Myope	Yes	Normal	Hard
Presbyopic	Hypermetrope	No	Reduced	None
Presbyopic	Hypermetrope	No	Normal	Soft
Presbyopic	Hypermetrope	Yes	Reduced	None
Presbyopic	Hypermetrope	Yes	Normal	None

UNION COLLEGE

Can construct if-then rules?

- Taking a look at the NONE recommendation, construct me a rule which is pretty good at determining if I should NOT wear contacts

UNION COLLEGE

Structural Descriptions

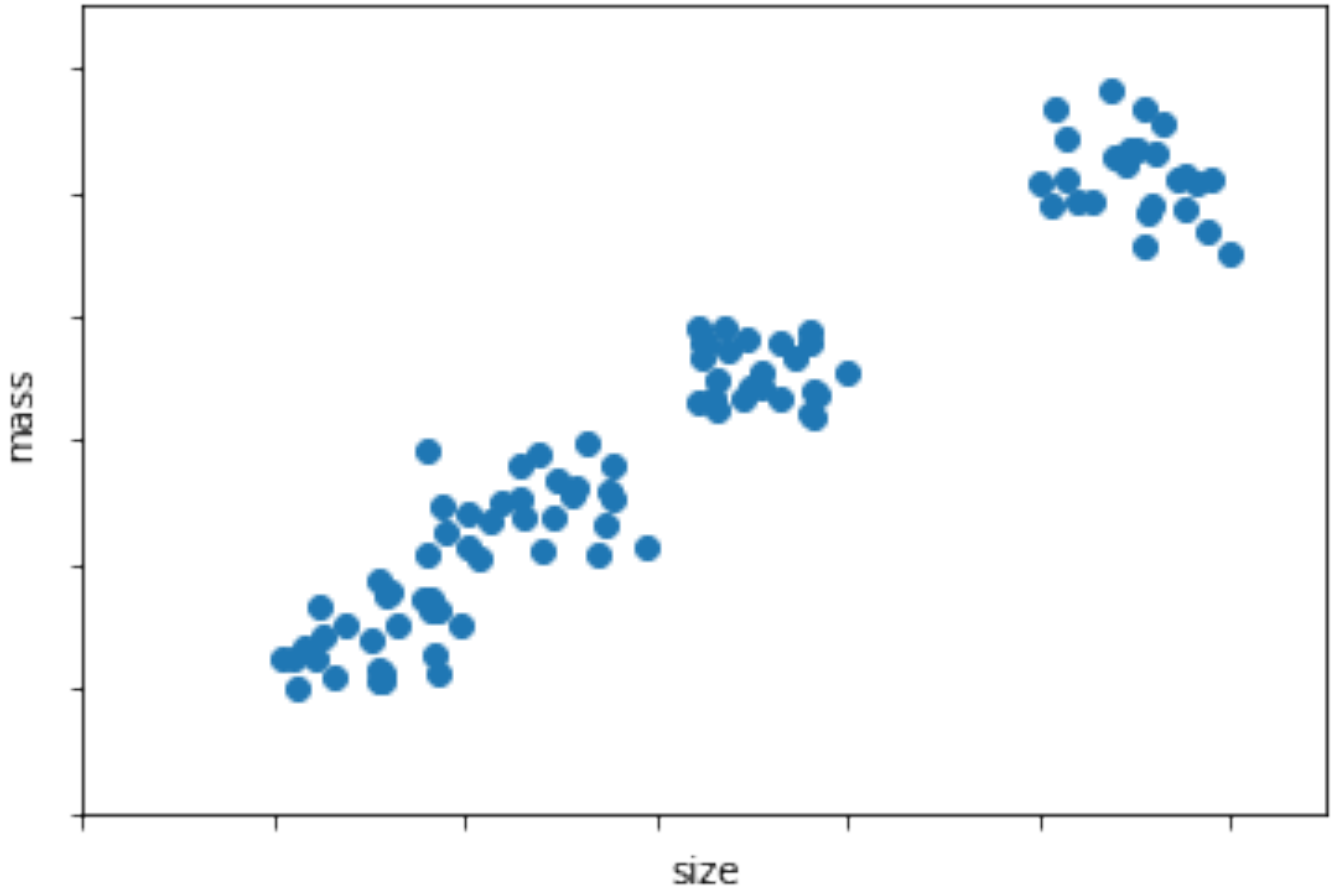
- Example: if-then rules

```
If tear production rate = reduced
   then recommendation = none
Otherwise, if age = young and astigmatic = no
   then recommendation = soft
```

Age	Spectacle prescription	Astigmatism	Tear production rate	Recommended lenses
Young	Myope	No	Reduced	None
Young	Hypermetrope	No	Normal	Soft
Pre-presbyopic	Hypermetrope	No	Reduced	None
Presbyopic	Myope	Yes	Normal	Hard
…	…	…	…	…

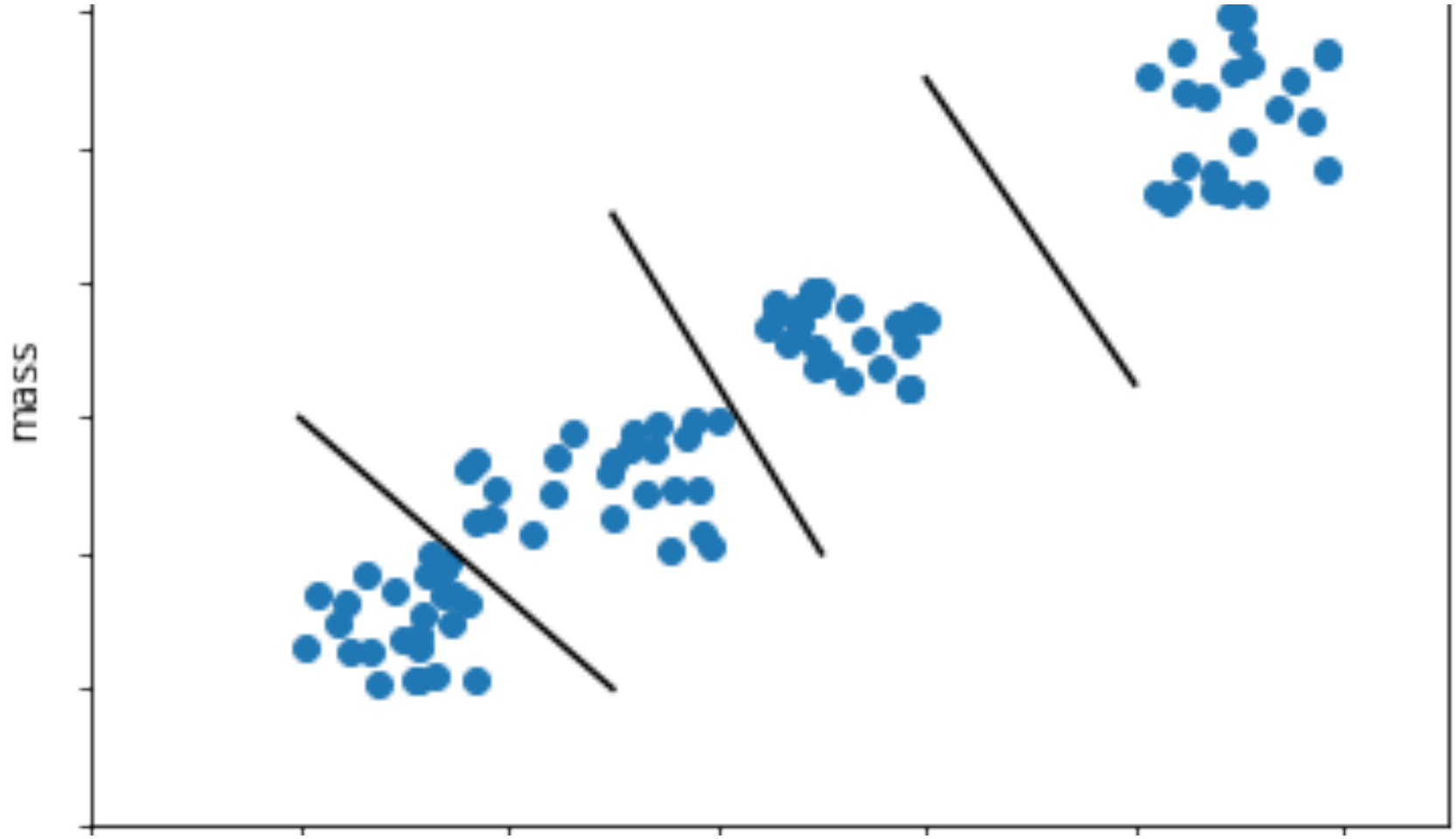
UNION COLLEGE

Unsupervised Learning



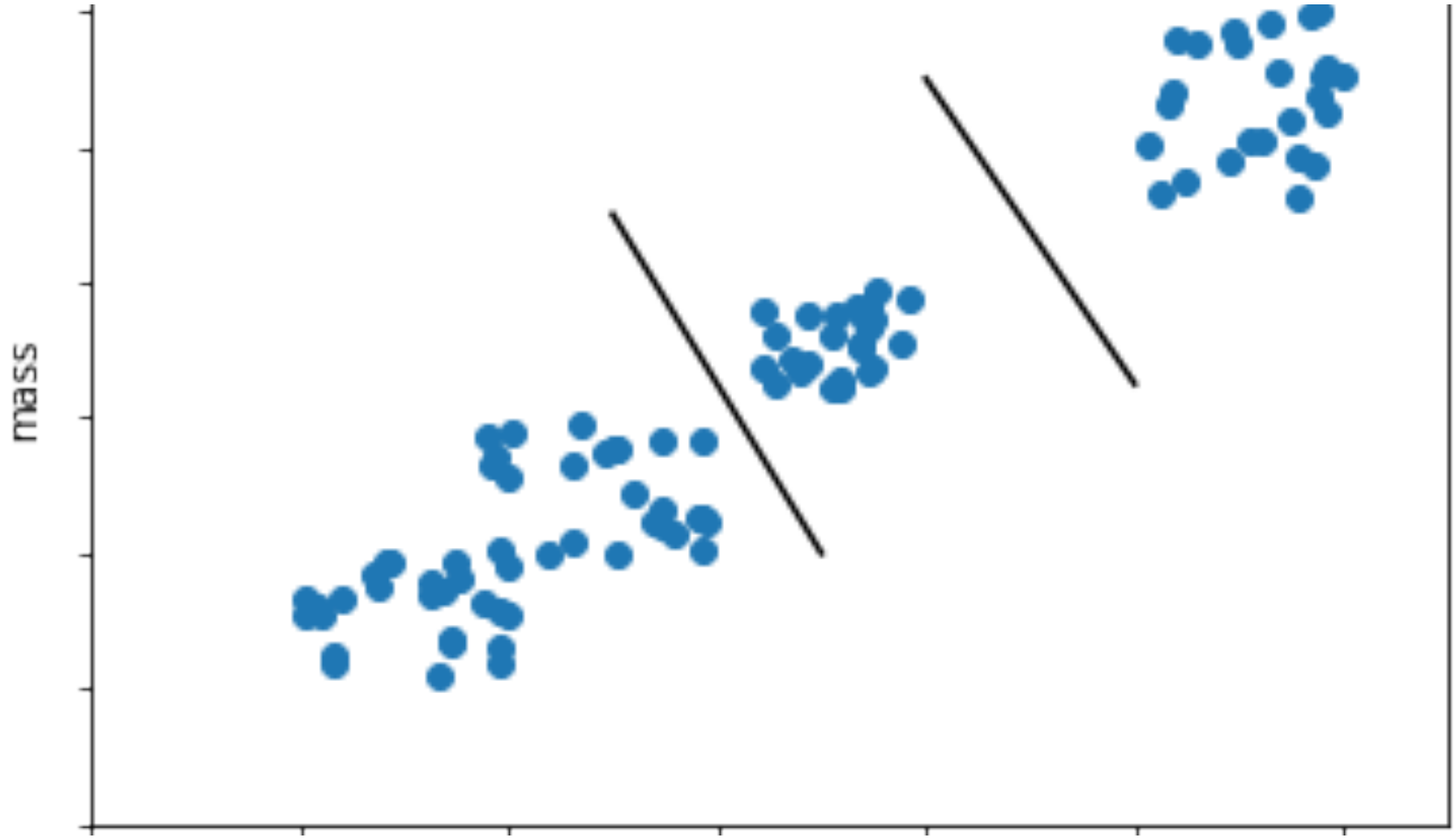
UNION
COLLEGE

Unsupervised Learning



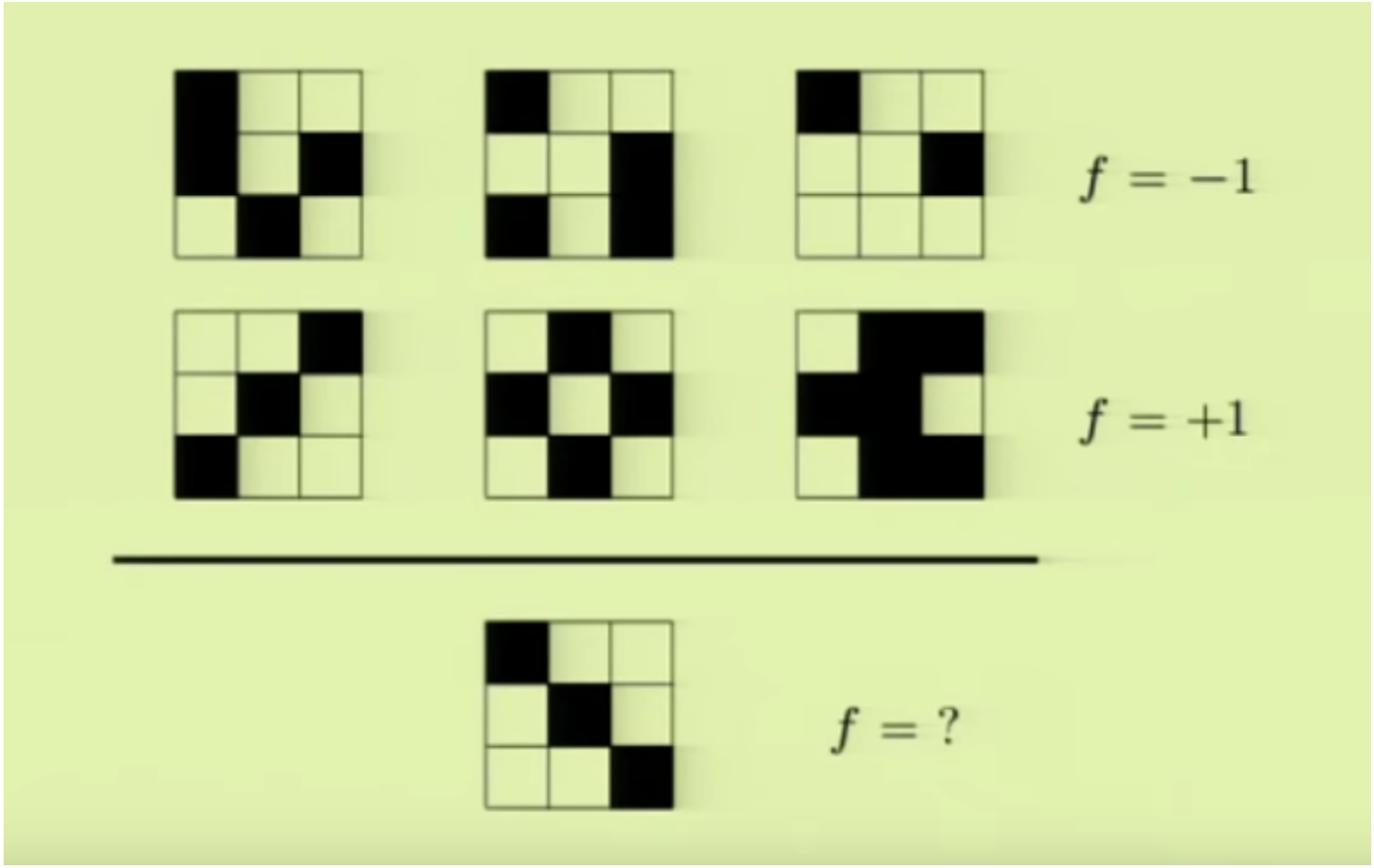
UNION
COLLEGE

Unsupervised Learning



UNION
COLLEGE

Learning Functions



UNION COLLEGE

Classification vs. Association

- Rules for playing a certain game

Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	Normal	False	Yes
…	…	…	…	…

```
If outlook = sunny and humidity = high then play = no
If outlook = rainy and windy = true then play = no
If outlook = overcast then play = yes
If humidity = normal then play = yes
If none of the above then play = yes
```

UNION COLLEGE

Classification vs. association rules

- Classification rule:
  - Predicts value of a given attribute (classification of an example)

```
If outlook = sunny and humidity = high
   then play = no
```

- Association rule:
  - Predicts value of arbitrary attribute (or combination)

```
If temperature = cool then humidity = normal
If humidity = normal and windy = false
   then play = yes
If outlook = sunny and play = no
   then humidity = high
If windy = false and play = no
   then outlook = sunny and humidity = high
```

UNION COLLEGE

Weather data with mixed attributes

- Some attributes have numeric values

Outlook	Temperature	Humidity	Windy	Play
Sunny	85	85	False	No
Sunny	80	90	True	No
Overcast	83	86	False	Yes
Rainy	75	80	False	Yes
...	...	...	...	...

```
If outlook = sunny and humidity > 83 then play = no
If outlook = rainy and windy = true then play = no
If outlook = overcast then play = yes
If humidity < 85 then play = yes
If none of the above then play = yes
```

UNION COLLEGE

Contact lens data

Age	Spectacle prescription	Astigmatism	Tear production rate	Recommended lenses
Young	Myope	No	Reduced	None
Young	Myope	No	Normal	Soft
Young	Myope	Yes	Reduced	None
Young	Myope	Yes	Normal	Hard
Young	Hypermetrope	No	Reduced	None
Young	Hypermetrope	No	Normal	Soft
Young	Hypermetrope	Yes	Reduced	None
Young	Hypermetrope	Yes	Normal	hard
Pre-presbyopic	Myope	No	Reduced	None
Pre-presbyopic	Myope	No	Normal	Soft
Pre-presbyopic	Myope	Yes	Reduced	None
Pre-presbyopic	Myope	Yes	Normal	Hard
Pre-presbyopic	Hypermetrope	No	Reduced	None
Pre-presbyopic	Hypermetrope	No	Normal	Soft
Pre-presbyopic	Hypermetrope	Yes	Reduced	None
Pre-presbyopic	Hypermetrope	Yes	Normal	None
Presbyopic	Myope	No	Reduced	None
Presbyopic	Myope	No	Normal	None
Presbyopic	Myope	Yes	Reduced	None
Presbyopic	Myope	Yes	Normal	Hard
Presbyopic	Hypermetrope	No	Reduced	None
Presbyopic	Hypermetrope	No	Normal	Soft
Presbyopic	Hypermetrope	Yes	Reduced	None
Presbyopic	Hypermetrope	Yes	Normal	None

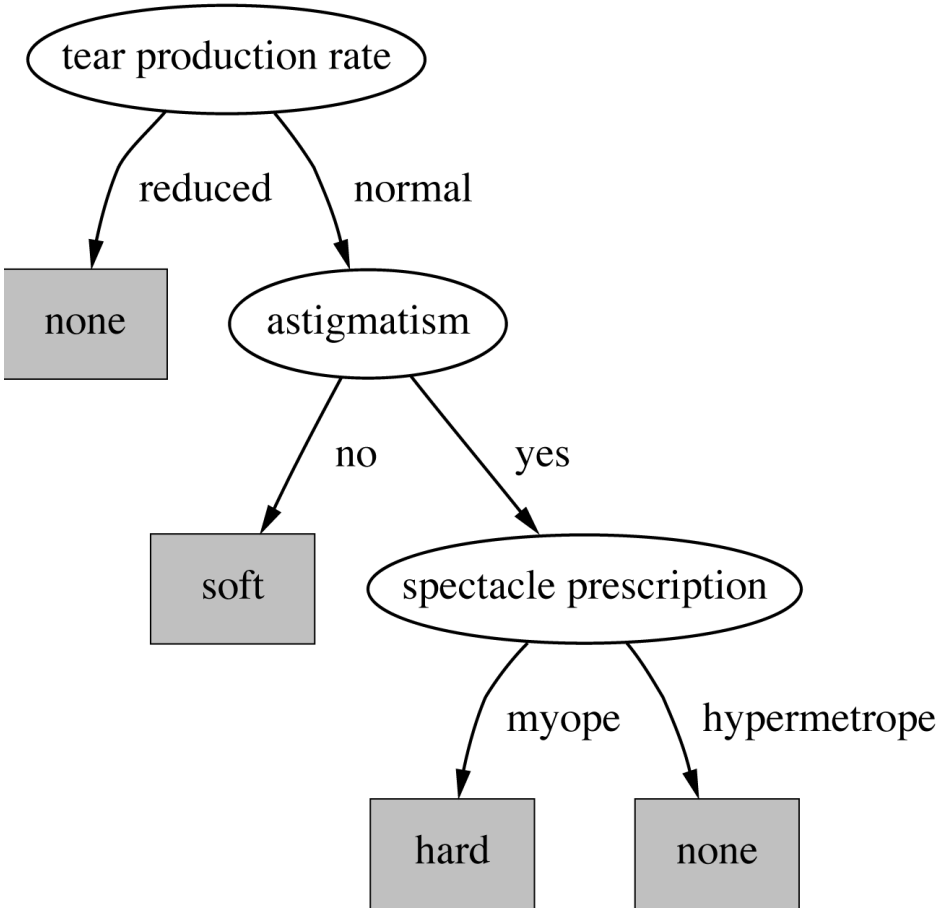
UNION COLLEGE

A complete and correct rule set

```
If tear production rate = reduced then recommendation = none
If age = young and astigmatic = no
    and tear production rate = normal then recommendation = soft
If age = pre-presbyopic and astigmatic = no
    and tear production rate = normal then recommendation = soft
If age = presbyopic and spectacle prescription = myope
    and astigmatic = no  then recommendation = none
If spectacle prescription = hypermetrope and astigmatic = no
    and tear production rate = normal then recommendation = soft
If spectacle prescription = myope and astigmatic = yes
    and tear production rate = normal then recommendation = hard
If age young and astigmatic = yes
    and tear production rate = normal then recommendation = hard
If age = pre-presbyopic
    and spectacle prescription = hypermetrope
    and astigmatic = yes then recommendation = none
If age = presbyopic and spectacle prescription = hypermetrope
    and astigmatic = yes then recommendation = none
```

UNION COLLEGE

A decision tree for this problem



UNION COLLEGE

Classifying iris flowers

	Sepal length	Sepal width	Petal length	Petal width	Type
1	5.1	3.5	1.4	0.2	Iris setosa
2	4.9	3.0	1.4	0.2	Iris setosa
…					
51	7.0	3.2	4.7	1.4	Iris versicolor
52	6.4	3.2	4.5	1.5	Iris versicolor
…					
101	6.3	3.3	6.0	2.5	Iris virginica
102	5.8	2.7	5.1	1.9	Iris virginica
…					

```
If petal length < 2.45 then Iris setosa
If sepal width < 2.10 then Iris versicolor
...
```

UNION COLLEGE

Predicting CPU performance

- Example: 209 different computer configurations

	Cycle time (ns)	Main memory (Kb)		Cache (Kb)	Channels		Performance
	MYCT	MMIN	MMAX	CACH	CHMIN	CHMAX	PRP
1	125	256	6000	256	16	128	198
2	29	8000	32000	32	8	32	269
…							
208	480	512	8000	32	0	0	67
209	480	1000	4000	0	0	0	45

- Linear regression function

```
PRP = -55.9 + 0.0489 MYCT + 0.0153 MMIN + 0.0056 MMAX
      + 0.6410 CACH - 0.2700 CHMIN + 1.480 CHMAX
```

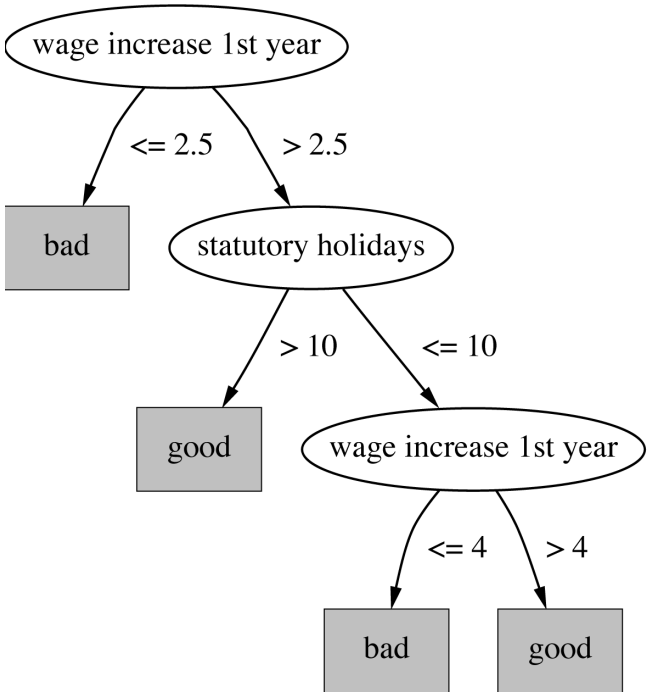
UNION COLLEGE

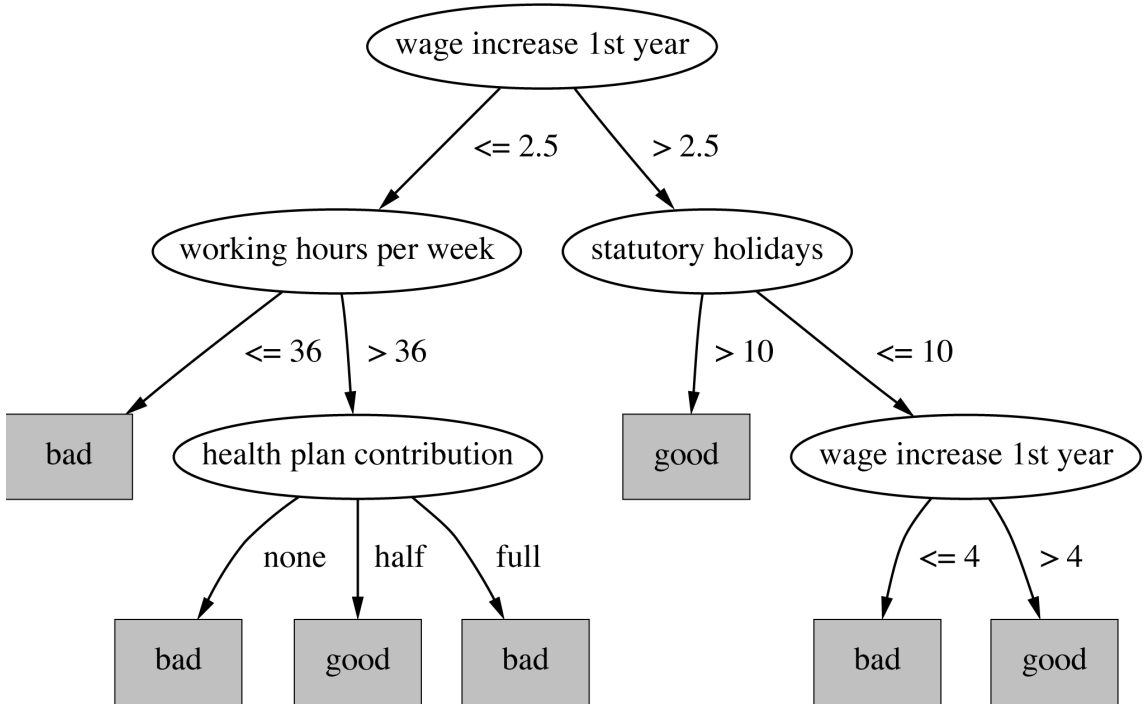
Data from labor negotiations

Attribute	Type	1	2	3	…	40
Duration	(Number of years)	1	2	3		2
Wage increase first year	Percentage	2%	4%	4.3%		4.5
Wage increase second year	Percentage	?	5%	4.4%		4.0
Wage increase third year	Percentage	?	?	?		?
Cost of living adjustment	{none,tcf,tc}	none	tcf	?		none
Working hours per week	(Number of hours)	28	35	38		40
Pension	{none,ret-allw, empl-cntr}	none	?	?		?
Standby pay	Percentage	?	13%	?		?
Shift-work supplement	Percentage	?	5%	4%		4
Education allowance	{yes,no}	yes	?	?		?
Statutory holidays	(Number of days)	11	15	12		12
Vacation	{below-avg,avg,gen}	avg	gen	gen		avg
Long-term disability assistance	{yes,no}	no	?	?		yes
Dental plan contribution	{none,half,full}	none	?	full		full
Bereavement assistance	{yes,no}	no	?	?		yes
Health plan contribution	{none,half,full}	none	?	full		half
Acceptability of contract	{good,bad}	bad	good	good		good

UNION COLLEGE

Decision trees for the labor data





UNION COLLEGE

Key language

- Data set
- Consists of instances
- Each instance contains attributes
- And often a class

UNION COLLEGE

Key issues

- How much data?
- What are the attributes?
- What are we predicting?
- Nominal? Numeric?
- Missing values?
- Domain knowledge?

UNION COLLEGE

Components of Learning

- Example: Apply for a credit card
- Bank does not have formula for creditworthiness

- BUT do have historical records, that show if customers were credit worthy or not

UNION COLLEGE

Applicant Information

- Age
- Gender
- Salary
- Years in residence
- Years in job
- Current debt…

- Assume that these factors have some relation to creditworthiness

UNION COLLEGE

Components

- Input
  - Customer application
- Output
  - Good/Bad customer
- Target function (ideal credit formula)
  - F: x -> y
- Data (historical records)
  - $(x_1,y_1), (x_2,y_2),\ldots,(x_n,y_n)$

UNION COLLEGE

Fielded applications

- The result of learning—or the learning method itself—is deployed in practical applications
  - Processing loan applications
  - Screening images for oil slicks
  - Electricity supply forecasting
  - Diagnosis of machine faults
  - Marketing and sales
  - Separating crude oil and natural gas
  - Reducing banding in rotogravure printing
  - Finding appropriate technicians for telephone faults
  - Scientific applications: biology, astronomy, chemistry
  - Automatic selection of TV programs
  - Monitoring intensive care patients

UNION COLLEGE

Processing loan applications

(American Express)

- Given: questionnaire with financial and personal information
- Question: should money be lent?
- Simple statistical method covers 90% of cases
- Borderline cases referred to loan officers
- But: 50% of accepted borderline cases defaulted!
- Solution: reject all borderline cases?
  - No! Borderline cases are most active customers

UNION COLLEGE

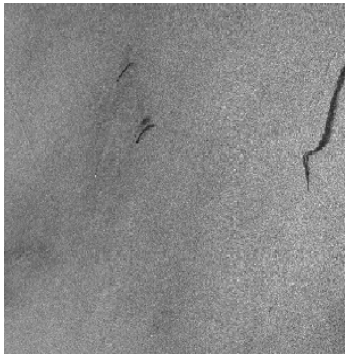
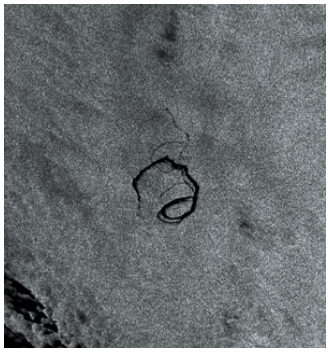
Enter machine learning

- 1000 training examples of borderline cases
- 20 attributes:
  - age
  - years with current employer
  - years at current address
  - years with the bank
  - other credit cards possessed,…
- Learned rules: correct on 70% of cases
  - human experts only 50%
- Rules could be used to explain decisions to customers

UNION COLLEGE

Screening images

- Given: radar satellite images of coastal waters
- Problem: detect oil slicks in those images
- Oil slicks appear as dark regions with changing size and shape
- Not easy: lookalike dark regions can be caused by weather conditions (e.g. high wind)
- Expensive process requiring highly trained personnel


UNION COLLEGE

Enter machine learning

- Extract dark regions from normalized image
- Attributes:
  - size of region
  - shape, area
  - intensity
  - sharpness and jaggedness of boundaries
  - proximity of other regions
  - info about background
- Constraints:
  - Few training examples—oil slicks are rare!
  - Unbalanced data: most dark regions aren't slicks

UNION COLLEGE

Load forecasting

- Electricity supply companies need forecast of future demand for power
- Forecasts of min/max load for each hour
  $\Rightarrow$ significant savings
- Given: manually constructed load model that assumes “normal” climatic conditions
- Problem: adjust for weather conditions
- Static model consist of:
  - base load for the year
  - load periodicity over the year
  - effect of holidays

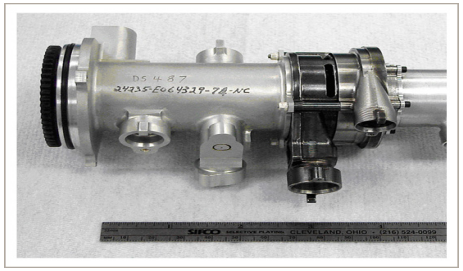
UNION COLLEGE

Enter machine learning

- Prediction corrected using “most similar” days
- Attributes:
  - temperature
  - humidity
  - wind speed
  - cloud cover readings
  - plus difference between actual load and predicted load
- Average difference among three “most similar” days added to static model

UNION COLLEGE

Diagnosis of machine faults



- Diagnosis: classical domain of expert systems
- Given: Fourier analysis of vibrations measured at various points of a device's mounting
- Question: which fault is present?
- Preventative maintenance of electromechanical motors and generators
- Information very noisy
- So far: diagnosis by expert/hand-crafted rules

UNION COLLEGE

Enter machine learning

- Available: 600 faults with expert's diagnosis
- Attributes augmented by intermediate concepts that embodied causal domain knowledge
- Expert not satisfied with initial rules because they did not relate to his domain knowledge
- Further background knowledge resulted in more complex rules that were satisfactory
- Learned rules outperformed hand-crafted ones

UNION COLLEGE

Marketing and sales I

- Companies precisely record massive amounts of marketing and sales data
- Applications:
  - Customer loyalty:
    identifying customers that are likely to defect by detecting changes in their behavior
    (e.g. banks/phone companies)
  - Special offers:
    identifying profitable customers
    (e.g. reliable owners of credit cards that need extra money during the holiday season)

UNION COLLEGE

Marketing and sales II

- Market basket analysis
  - Association techniques find groups of items that tend to occur together in a transaction (used to analyze checkout data)
- Historical analysis of purchasing patterns
- Identifying prospective customers
  - Focusing promotional mailouts (targeted campaigns are cheaper than mass-marketed ones)

UNION COLLEGE

Machine learning and statistics

- Historical difference (grossly oversimplified):
  - Statistics: testing hypotheses
  - Machine learning: finding the right hypothesis
- But: huge overlap
  - Decision trees
  - Nearest-neighbor methods
- Today: perspectives have converged
  - Most ML algorithms employ statistical techniques

UNION COLLEGE

Generalization as search

- Inductive learning: find a description that fits the data
- Simple solution:
  - enumerate the concept space
  - eliminate descriptions that do not fit examples
  - surviving descriptions contain target concept

- Example: rule sets as description language
  - Enormous, but finite, search space

UNION COLLEGE

Enumerating the concept space

- Search space for weather problem
  - 4 x 4 x 3 x 3 x 2 = 288 possible combinations
  - With 14 rules $\Rightarrow$ 2.7×10^{34} possible rule sets

  - Most practical algorithms use heuristic search that cannot guarantee to find the optimum solution

UNION COLLEGE

Data mining and ethics I

- Ethical issues arise in practical applications
- Anonymizing data is difficult
- Data mining often used to discriminate
  - E.g. loan applications: using some information (e.g. sex, religion, race) is unethical
- Ethical situation depends on application
  - E.g. same information ok in medical application
- Attributes may contain problematic information
  - E.g. area code may correlate with race

UNION COLLEGE

Data mining and ethics II

- Important questions:
  - Who is permitted access to the data?
  - For what purpose was the data collected?
  - What kind of conclusions can be legitimately drawn from it?
- Caveats must be attached to results
- Purely statistical arguments are never sufficient!