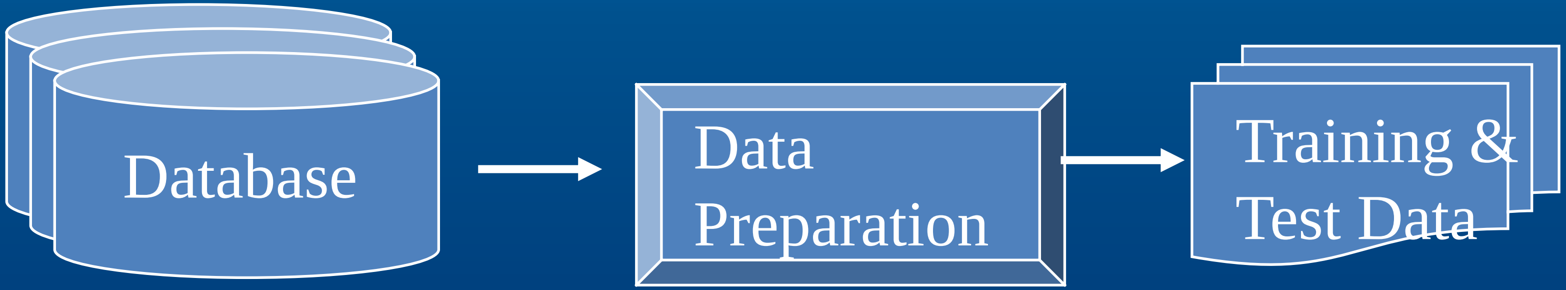
Big Data Preparation

Preamble to Machine Learning

- Broad Working definition:
  - organizing the data

    (aka ‘data wrangling’ or ‘data munging’)
  - cleaning, filtering, and transforming

Organizing Data into Input



Database

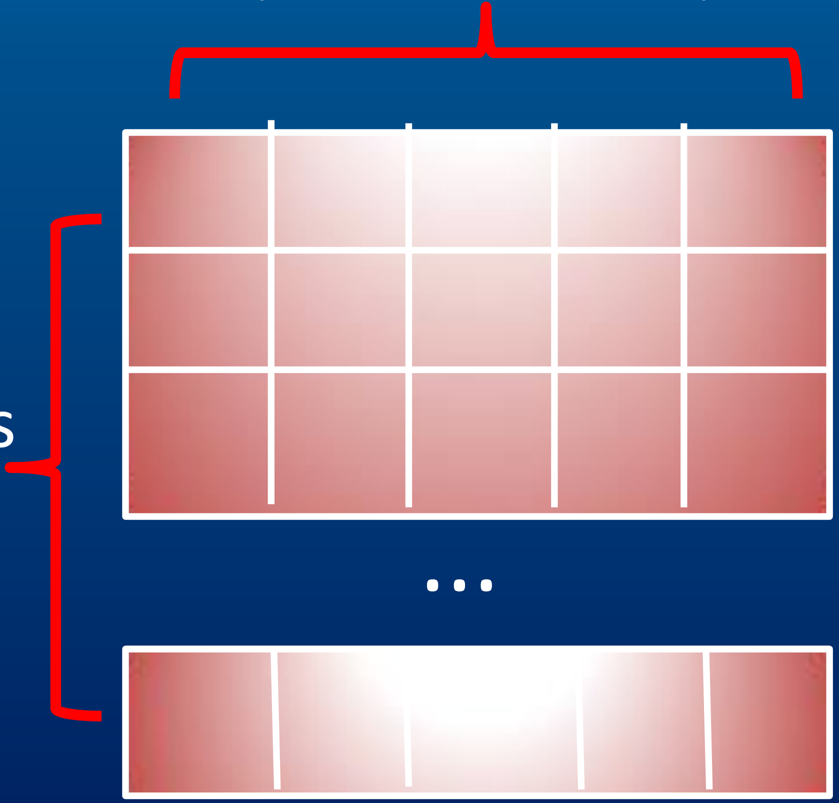
Data Preparation

Training & Test Data

Gather Relevant Data

DATA MATRIX

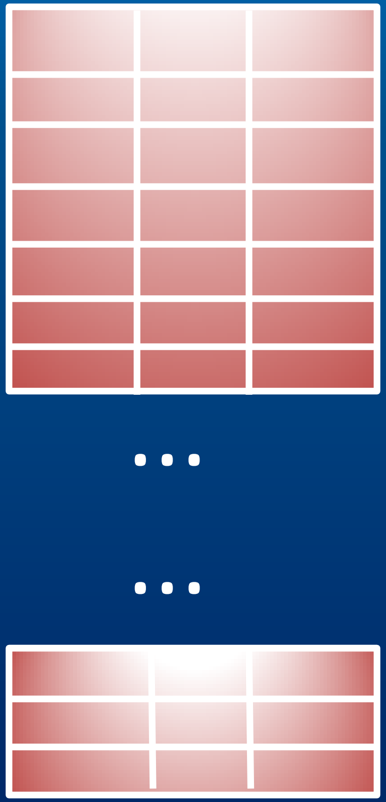
‘variables’, or ‘attributes’, ‘features’ (columns)



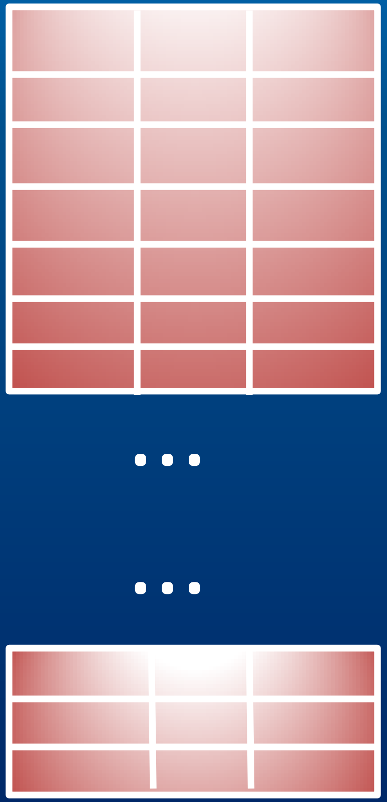
Instances
(rows)

N

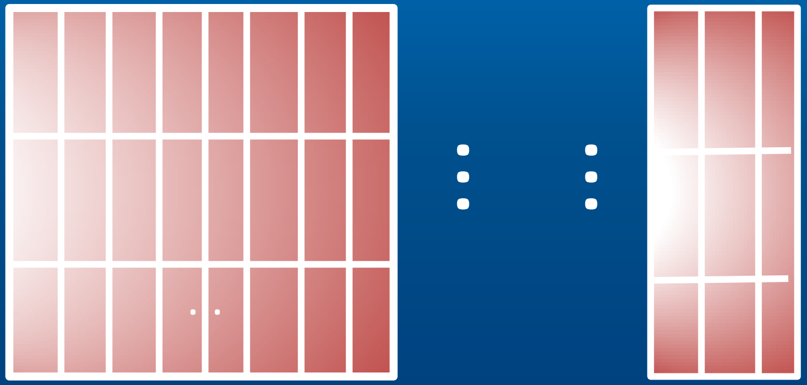
Large number of rows



Large number of rows

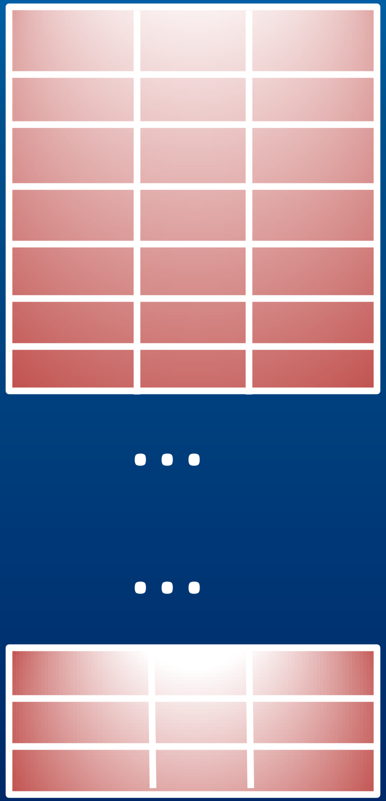


Large number of Columns

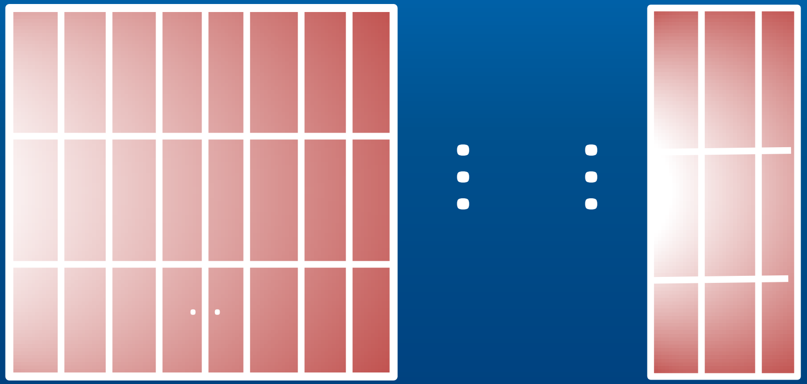


Assume: *data partitioned on rows, and 1 row fits in 1 computer memory*

Large number of rows



Large number of Columns



***Assume:** data partitioned on rows, and 1 row fits in 1 computer memory*

Data Matrix to Models

Instance	Customer	Item	Price	Date	Label/outcome
1	John	Acme Mower	100	Jan 2000	Used coupon
2	John	Acme Wrench	10	Sept 2000	Used coupon
3	Jane	Ace Mower	120	Mar 2003	No coupon
4	Jane	Ace Rake	20	Mar 2003	No coupon
5 …	Fred	Ace Hammer	15	July 2002	Used coupon

Data Matrix to Models

Instance	Customer	Item	Price	Date	Label/outcome
1	John	Acme Mower	100	Jan 2000	Used coupon
2	John	Acme Wrench	10	Sept 2000	Used coupon
3	Jane	Ace Mower	120	Mar 2003	No coupon
4	Jane	Ace Rake	20	Mar 2003	No coupon
5 …	Fred	Ace Hammer	15	July 2002	Used coupon

Model:

Coupon-use is function of Price, Customer, Item, Date, etc..

Data Matrix to Models

Instance	Customer	Item	Price	Date	Label/outcome
1	John	Acme Mower	100	Jan 2000	Used coupon
2	John	Acme Wrench	10	Sept 2000	Used coupon
3	Jane	Ace Mower	120	Mar 2003	No coupon
4	Jane	Ace Rake	20	Mar 2003	No coupon
5 …	Fred	Ace Hammer	15	July 2002	Used coupon

Model:

$$\mathit{coupon}_1 = F\,(\,\mathit{Price}_1\,,\,\mathit{Customer}_1\,,\,\mathit{Item}_1\,,\ \mathit{etc..})$$

$$\mathit{coupon}_2 = F\,(\,\mathit{Price}_2\,,\,\mathit{Customer}_2\,,\,\mathit{Item}_2\,,\ \mathit{etc..})$$

…

New Model: compare customers

Customer	Mower	Wrench	Rake	Hammer	…	(last item)
John	1	1	1	1		
Jane	1	0	0	0		
…						

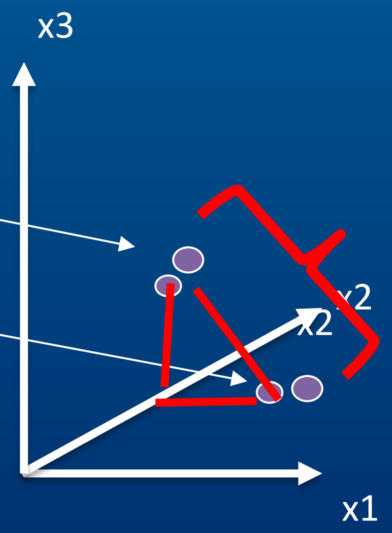
New Model: compare customers

Customer	Mower	Wrench	Rake	Hammer	…	(last item)
John	1	1	1	1		
Jane	1	0	0	0		
…						

Recode categorical items-bought as 1 column for each item

Each Row is now a vector

x1	x2	x3	....	
1	1.5	1.3		
2.2	1	.25		
1				



sometimes called
the ‘input space’

Data Preprocessing

Data Preprocessing

- – Cleaning & Filtering

Data Preprocessing

- – Cleaning & Filtering
- – Variable transformations

Data Preprocessing

- – Cleaning & Filtering
- – Variable transformations
- – Variable Selection

Cleaning Noise

– Entity Resolution and Record Linkage

e.g. Are these equal?

West Main Street

W Main St

Strategy:

use dictionaries and search possible matches

Cleaning Noise

– Entity Resolution and Record Linkage

e.g. Are these equal?

West Main Street

W Main St

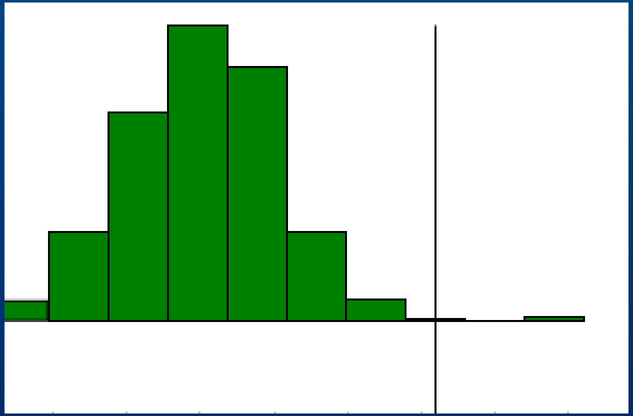
Strategy:

use dictionaries and search possible matches

Statistical Noise:

– Outliers

e.g. remove them,

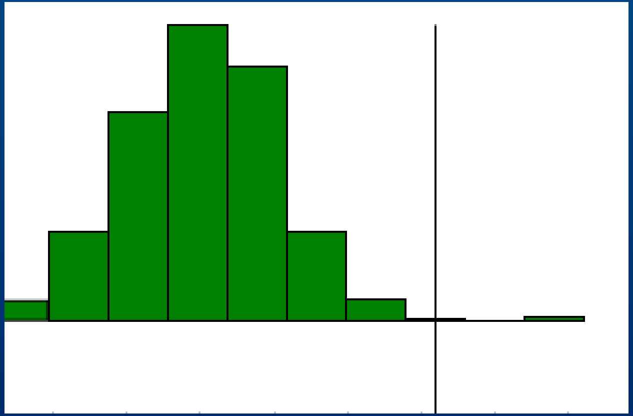


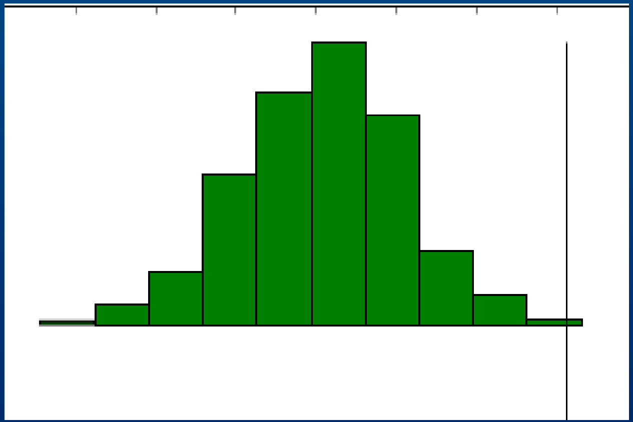
*mean + 3*std-devm*

Statistical Noise:

– Outliers

e.g. remove them, but cutoff is arbitrary





*mean + 3*std-devm*

Missing Data

- Not applicable

  e.g. spouse name depends on marital status

- Not Available

  unknown

  not entered

Missing Data

– Do missing cases depend on some other variable?

e.g. ‘CEOs’ don’t like to list their salary

Strategy: *get most common job titles for missing salaries*

Quick Approaches

- Delete instances

  and/or

- Delete attributes with high missingness

Quick Approaches

- – Leave as ‘NULL’ category
  - Some algorithms implementation handle NULL (ie Decision Trees)

Simple Imputation

- Use the attribute mean (by class)

Complicated Imputation

- Use a model (based on other attributes) to infer missing value

Not Simple Imputation

- Use a model (based on other attributes) to infer missing value

Best strategy depends on time vs accuracy tradeoffs

Variable Transformation

– and Feature Engineering

Variable Transformations

- **Combine attributes**

Variable Transformations

– **Combine attributes**

e.g. rates and ratios

– **Scaling data**

– **D**iscretize data

often more intuitive

Re-scaling

- Mean center

$$x_{new} = x - \mathrm{mean}(x)$$

- z-score

$$score = \frac{x - \mathrm{mean}(x)}{\mathrm{std}(x)}$$

- Scale to [0…1]

$$x_{new} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

- log scaling

$$x_{new} = \log(x)$$

Variable selection

- Heuristic methods:
  - remove variables with low correlations to outcome

Variable selection

- Heuristic methods:
  - remove variables with low correlations to outcome
  - step wise: add 1 variable at a time and test algorithm on samples

Summary

- Preparing data is based on statistical principles,

- But also heuristics