COMP5310 - Week 3

Data Mining and Hypothesis Testing

A Tale of a Statistical Study

In 1747, a doctor named James Lind did an experiment on board a British ship:

- selected 12 scurvy patients to compare the effectiveness of common cures
- divided them into groups of two
  - Each group got the same sailor diet, but with different additions.
  - E.g. two drank seawater, two were given two spoonsful of vinegar three times a day, … two were given 2 oranges and a lemon each day
- He reported: "… the most sudden and visible good effects were perceived from the use of oranges and lemons"

Image via Pixabay.

"The medical establishment continued to believe that scurvy was caused by disruptions of the digestive system caused by the sailors' hard work and bad diet, and that it could be cured by "fizzy drinks" …"
https://medium.com/science-uncovered/data-mining-a-plague-not-a-cure-b30ec520d00e

Experimental Design is Important

- What do you think about Lind's approach? x
- Did he use a sound experimental design? x
- Is his analysis based on two patients valid and convincing? x
- What could he have improved? a lot
- What could the medical establishment have done with his experimental data? no idea.

Goals for this Week

- High level overview of statistical tests and data mining (not a deep dive) ✓✓
- Provide some tools for selecting appropriate statistical tests for evaluating a predictive model, and justifying the choice of tool ✓✓
- Help you seek details of how to use a statistical method or tool in the data analytic process ✓✓

could do data mining on previous work to deliver new and improved insights.


THE UNIVERSITY OF
SYDNEY

Types of Statistical Studies

Types of Statistical Studies

- **Observational Study**
  - Simply observing what happens
  - Records information about subjects without applying any treatments to subjects (passive participation of researcher)

Passive participants

- **Experimental Study**
  - Records information about subjects while applying treatments to subjects and controlling study conditions to some degree (active participation of researcher)

variables are used and tuned, etc.

Observational Studies

- Sample Survey
  - provide information about a population based on a sample at a specific point in time

1. Retrospective Studies
  - collect information about sample on specific outcomes that have happened
  - generally cheap and can be completed more rapidly than prospective studies
  - have problems due to inaccuracies in data due to memory recall errors.
  - Example Study: Tanning and Skin Cancer
    - The observational study involving 1,500 people
    - Selected a group of people who had skin cancer and another group who did not have skin cancer
    - Asked all participants whether they used tanning beds
      - Wanted to see if there was an association between tanning beds and skin cancer prevalence

Observational Studies (cont'd)

2. Prospective Studies

- follow the sample into the future to observe outcomes
- in prospective studies, subjects can keep careful records of their daily activities.
- subjects can also be instructed to avoid certain activities which may bias the study; while this reduces some of the problems of retrospective studies, the potential influences of confounding variables may not be completely controlled.
- Example Study: Average Computer Time vs Blood Pressure
  - Enroll 100 individuals in the observational study.
  - Ask them to keep track of the computer time they spend each day.
  - Measure blood pressure.

Difference b/w causation and causality

→ recall ECON 2206 from UNSW

- Observational Studies only establish correlation, not causality

Experimental Studies

- Strong hypotheses, sample size for desired power and controlled data collection per specified protocols
- Establish causality
- e.g. randomized control trials, A/B tests, drug trials
  - previous “Average Computer Time vs Blood Pressure” as experimental study:
    - 1. Control Group (computer time max 30 minutes)
    - 2. Treatment Group (computer time of at least 2 hours)
    - From the 100 individuals, 50 subjects randomly assigned to each group
    - Factor: average computer time; response: measure blood pressure of each group
- But it has its flaws too
  - slow and expensive
  - selection bias of trial participants
  - publication bias of positive results

bias can be introduced based on the selection

John Snow and the 1854 Cholera Outbreak

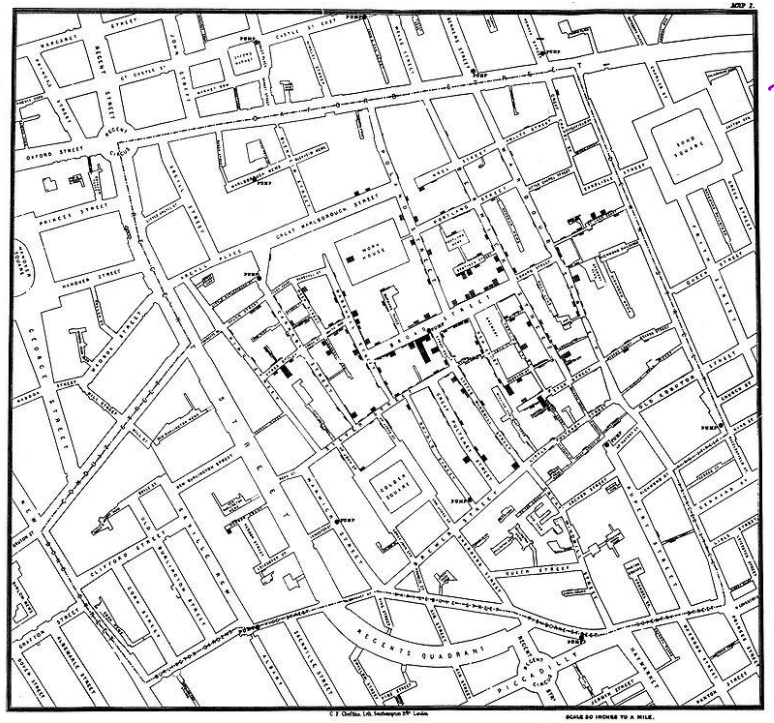


Image from Wikipedia Commons (CC3.0)

- London Cholera epidemic of 1854
- Founding event of epidemiology
- https://www.pastmedicalhistory.co.uk/john-snow-and-the-1854-cholera-outbreak/

Took another 11 years to confirm and understand causes for cholera.

perhaps causality?

correlation of living close to water pump shows increase in cholera.
↳ Though there were alot of flaws in this study.


THE UNIVERSITY OF
SYDNEY

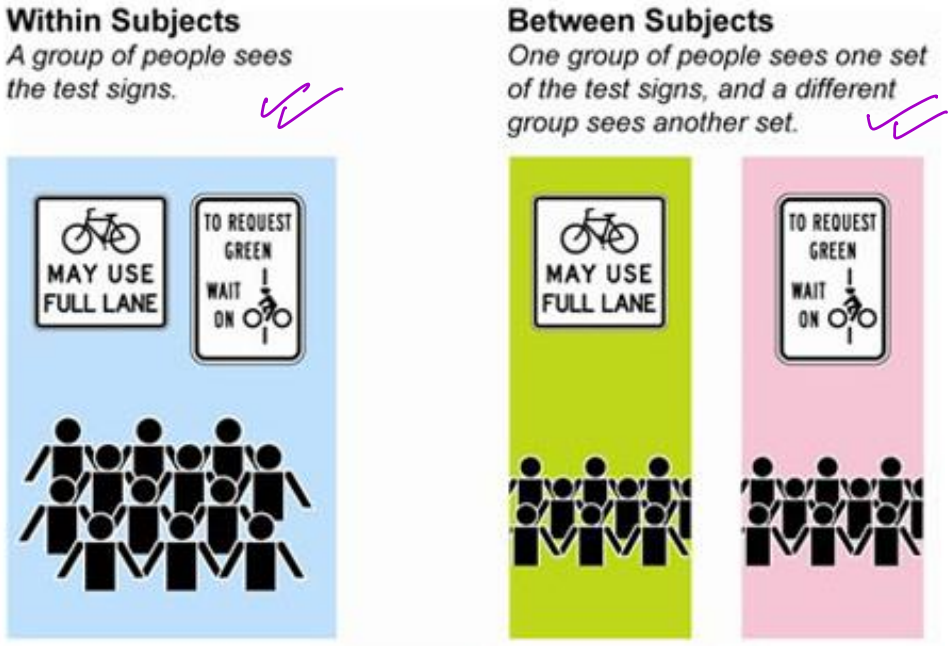
Qualitative study...

Till we include data
etc.

Setting up an Experiment

Experiment Design

- ***Between subjects***: Each subject sees one and only one condition
- ***Within subjects***: Subjects see more than one or all conditions



http://www.fhwa.dot.gov/publications/research/safety/pedbike/11035/004.cfm

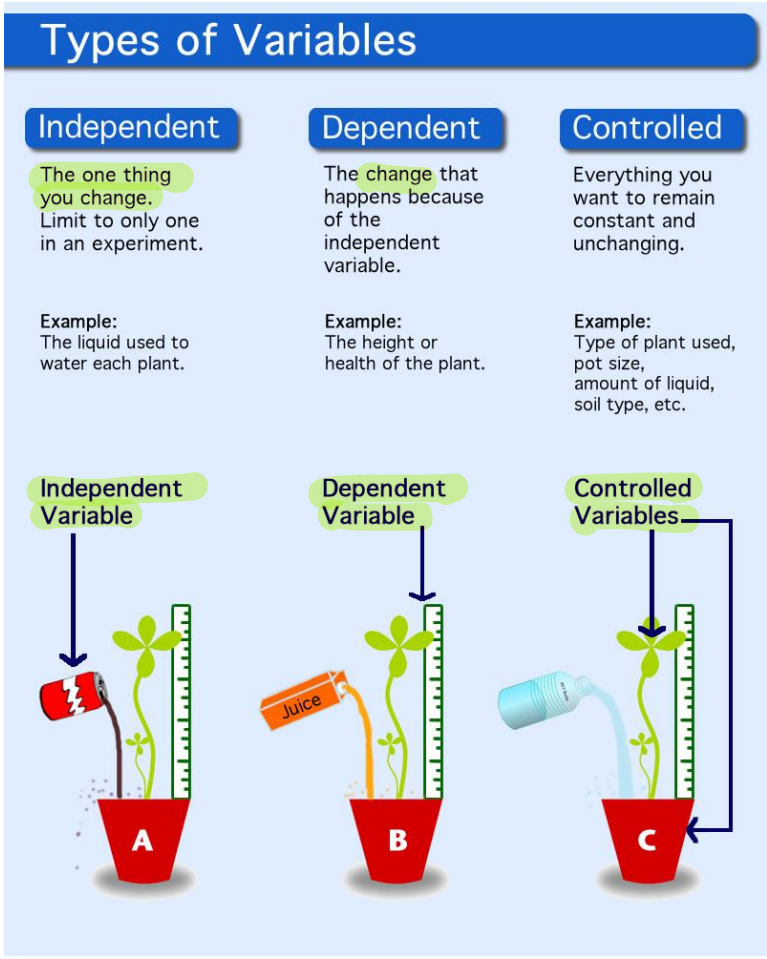
Year 9 science lol

Types of Variables

- ***Dependent variable*** is the measure of interest
- ***Independent variable*** is manipulated to observe the effect on dependent variable
- ***Controlled variables*** are materials, measurements and methods that don't change

y – outcome

$x_1 x_2 x_3 \ldots$ – predictors



http://edtech2.boisestate.edu/angelacovil/506/procedure.html

Research Question

- Research question (Q):
  - Asks whether the independent variable has an effect
  - “If there is a change in the independent variable, will there also be a change in the dependent variable?”
- Null hypothesis (H_0):
  - The assumption that there is no effect
  - “There is no change in the dependent variable when the independent variable changes.”


THE UNIVERSITY OF
SYDNEY

Hypothesis Testing

Hypothesis Testing

not population

- We use it to specify whether to accept or reject a claim about a population depending on the evidence provided by a sample of data.
- A hypothesis test examines two opposing hypotheses about a population parameter (e.g. the mean):
  - **The null hypothesis**
  - **The alternative hypothesis**
- The null hypothesis represents our initial assumption about the parameter, and we collect evidence to possibly reject the null hypothesis in favour of the alternative hypothesis
  - Example: determine whether the mean of a population differs significantly (this has a special meaning) from a specific value or from the mean of another population.

Testing Reliability with p-Values

we can call if 95% or 99% confidence interval.

- Most tests calculate a p-value measuring observation extremity
- Compare to significance level threshold α
  - α is the probability of (wrongly) rejecting H_0 given that it is true
  - aka Type I error rate (false positive)
  - Commonly use α of 5% or 1%

Truth	Decision: Accept H_0	Decision: Reject H_0
H_0 (no difference)	Right Decision	Type I Error
H_1 (difference exists)	Type II Error	Right Decision

P-value	Indicates	Reject H_0?
≤α	Strong evidence against the null hypothesis	Yes
>α	Weak evidence against the null hypothesis	No
=α	Marginal	NA

remember

Not every test result is correct

- P=0.05 will erroneously reject H_0 5% of the time
- Perform enough tests and you will get a false result (p-hacking)
- Good science:
  - Determine hypotheses before looking at data
  - Perform hypothesis-agnostic data cleaning
  - Remember that p-values do not replace common sense
- http://faculty.washington.edu/dwhm/2016/03/09/the-arbitrary-magic-of-p-0-05/

Increase the power of a significance test

- Obtain a larger sample
- Larger N means more reliable statistics
- Less likely to have errors
  - Type I: Reject true H_0
  - Type II: Fail to reject false H_0

Tips and Tricks

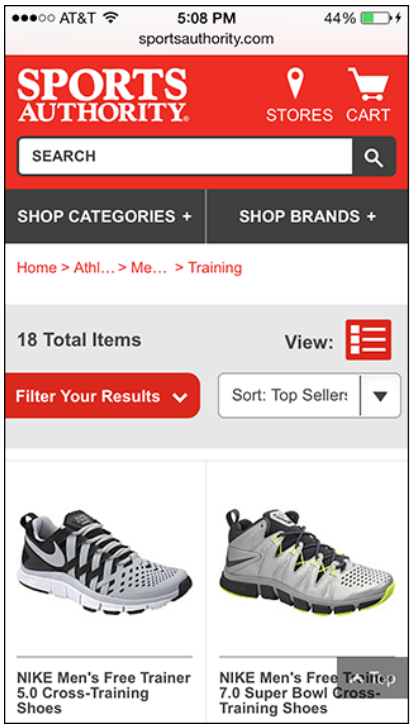
- Statistical hypothesis testing ensures results are reliable
- Experimental design includes:
  - Formulating a research question and null hypothesis
  - Designing and running experiments
  - Analysing results using appropriate statistics
- Use textbooks and documentation to find the right stats
- Sample representatively; Report p-value; Don’t hack p-value
- Report precision, recall, f-score and significance

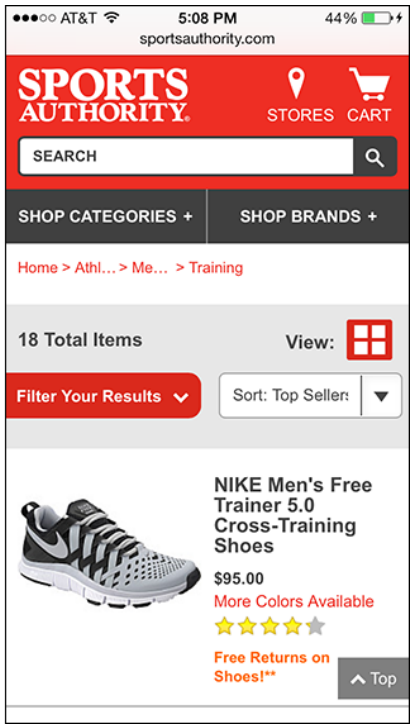

THE UNIVERSITY OF
SYDNEY

Testing which Approach is better Between Subjects

Scenario: Comparing Visual Layouts

Grid view





List view

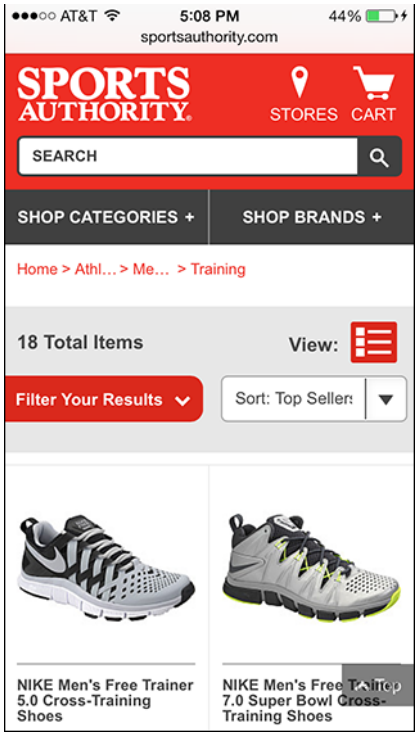
https://www.nngroup.com/articles/image-vs-list-mobile-navigation/

Research Question

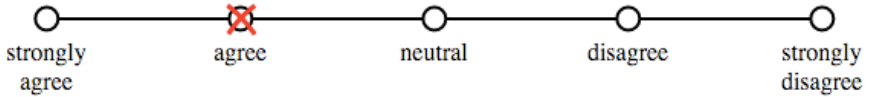
Do users prefer grid or list view?

Data/Measurement: User Ratings of Layouts

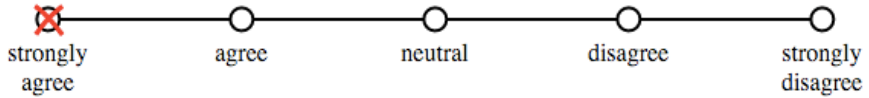
Example response from User Group A



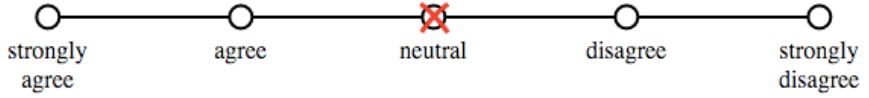
Page is easy to use.



Page gives good overview.

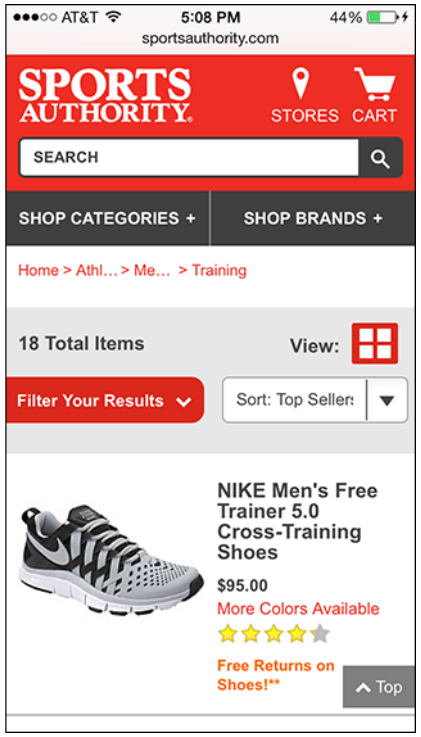


Page gives sufficient detail.

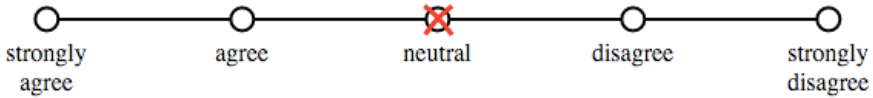


Data/Measurement: User Ratings of Layouts

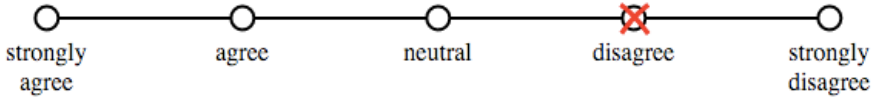
Example response from User Group B



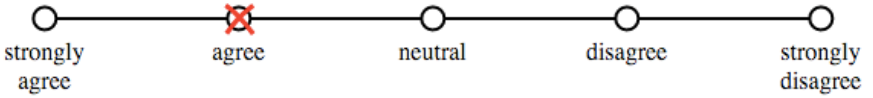
Page is easy to use.



Page gives good overview.



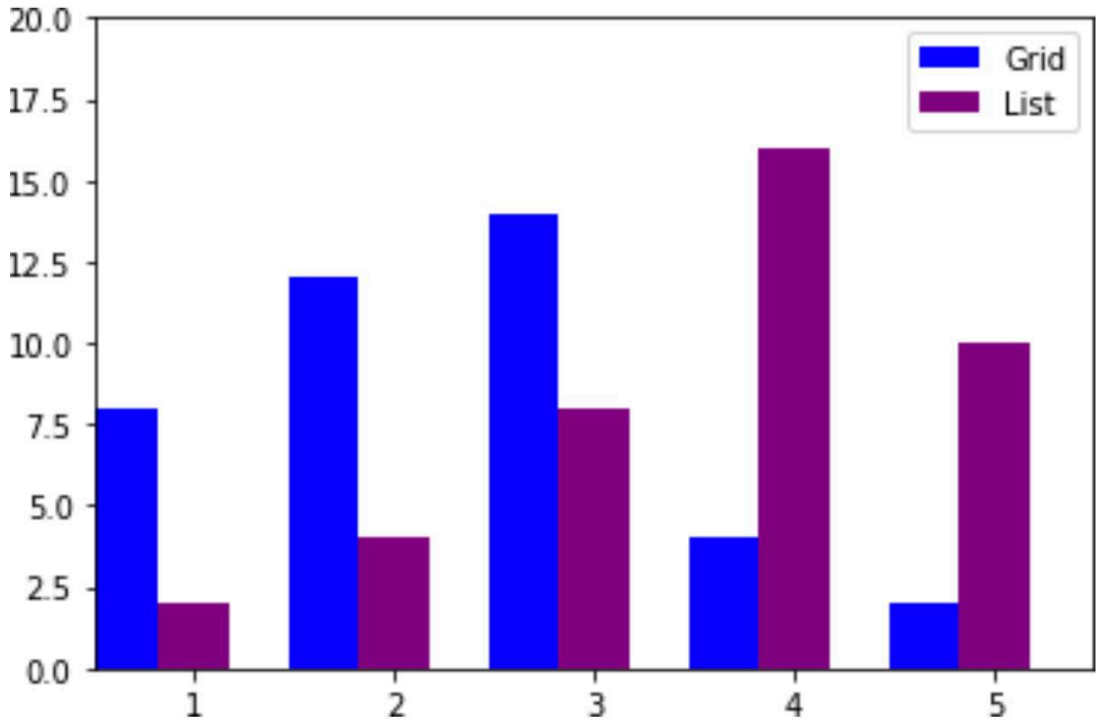
Page gives sufficient detail.



Generate Ratings Data

- We assume different subject groups for each condition.
- Each subject sees one of the layouts and is asked to rate on a 5-point Likert scale how strongly he agree or disagree with the statement:
  - Question to subjects: Page gives a good overview?
  - 1=strongly agree; 2=agree; 3=neutral; 4=disagree; 5=strongly disagree
- Outcome:
  - G_data= [1,3,3,2,4,2,3,3,1,5,2,3,4,2,1,3,2,2,1,3,2,3,4,2,1, 3,2,2,1,3,1,3,3,2,4,2,3,3,1,5]
  - L_data= [4,5,2,4,4,3,5,4,3,5,1,4,5,3,4,4,2,3,4,5,1,4,5,3,4,4,2,3,4,5,4,5,2,4,4,3,5,4,3,5]
- G_data corresponds to ratings from users that see the grid view.
- L_data corresponds to ratings from users that see the list view.

Visualise Ratings Data



Setup: Comparing Two Versions of a Display

- Subjects are users of the display (or summary, interface, etc)
  - Dependent variable is user rating (or comprehension, etc)
  - Independent variable is the version of the display
- Q: Do users prefer grid view?
- H_0: Grid and list data are drawn from the same distribution

Significance: Unpaired Student's t-Test

- Tests the null hypothesis that two population means are equal
- Assumes
  - The samples are independent
  - Populations are normally distributed
  - Standard deviations are equal
- Note
  - Multiply two-tailed p-value by 0.5 for one-tailed p-value (e.g., to test A>B, rather than A>B OR A<B)
- http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_ind.html#scipy.stats.ttest_ind

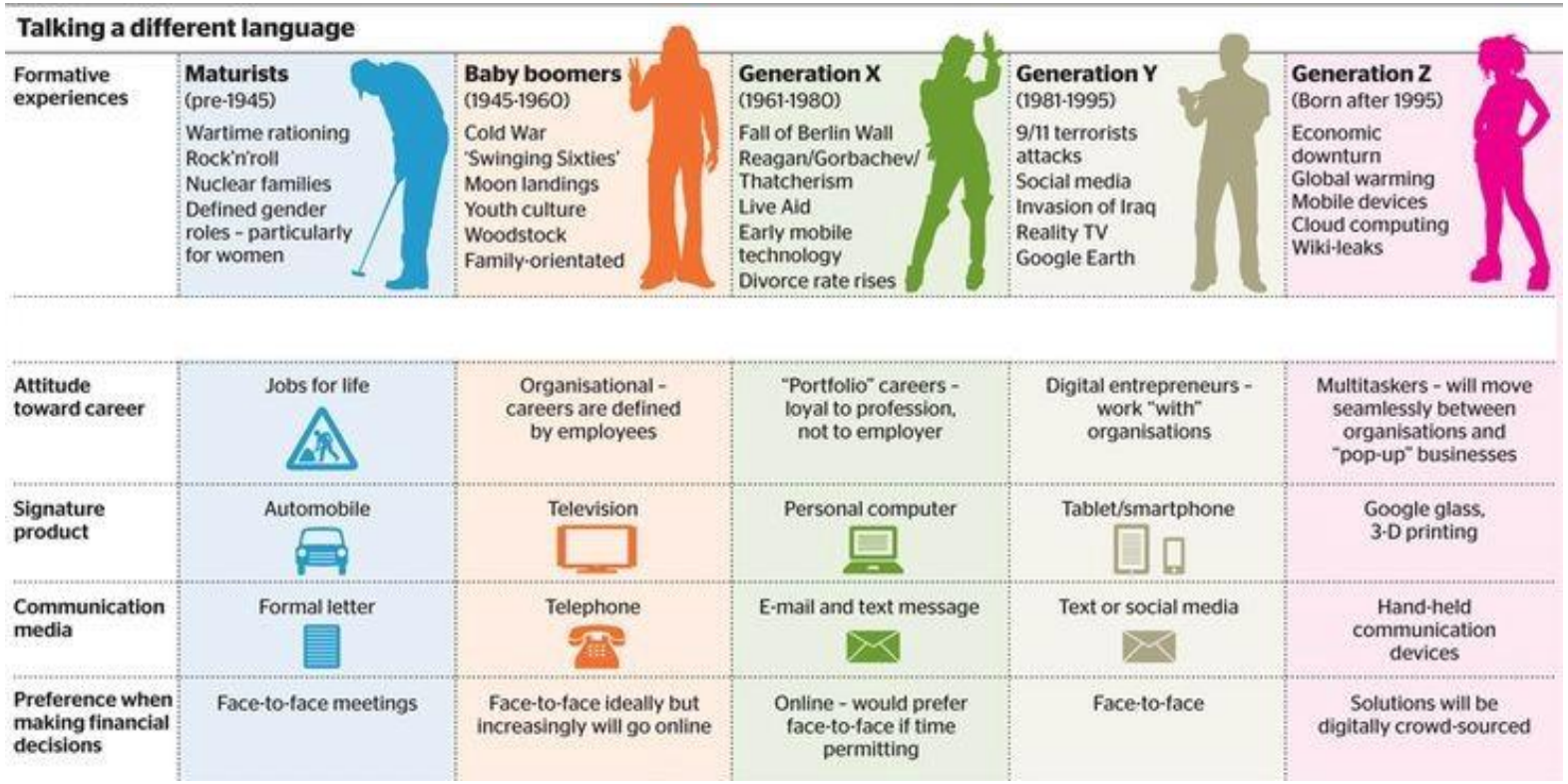
Significance: Mann-Whitney U Test

- Nonparametric version of unpaired t-test
- Assumes
  - The samples are independent
  - The data is at least ordinal
- Note
  - N should be at least 20
- http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html#scipy.stats.mannwhitneyu


THE UNIVERSITY OF
SYDNEY

Testing Whether Groups Differ

Scenario: Mobile Use by Generation



Talking a different language

	Maturists (pre-1945)	Baby boomers (1945-1960)	Generation X (1961-1980)	Generation Y (1981-1995)	Generation Z (Born after 1995)
Formative experiences	Wartime rationing Rock'n'roll Nuclear families Defined gender roles - particularly for women	Cold War 'Swinging Sixties' Moon landings Youth culture Woodstock Family-orientated	Fall of Berlin Wall Reagan/Gorbachev/ Thatcherism Live Aid Early mobile technology Divorce rate rises	9/11 terrorists attacks Social media Invasion of Iraq Reality TV Google Earth	Economic downturn Global warming Mobile devices Cloud computing Wiki-leaks
Attitude toward career	Jobs for life	Organisational - careers are defined by employees	"Portfolio" careers - loyal to profession, not to employer	Digital entrepreneurs - work "with" organisations	Multitaskers - will move seamlessly between organisations and "pop-up" businesses
Signature product	Automobile	Television	Personal computer	Tablet/smartphone	Google glass, 3-D printing
Communication media	Formal letter	Telephone	E-mail and text message	Text or social media	Hand-held communication devices
Preference when making financial decisions	Face-to-face meetings	Face-to-face ideally but increasingly will go online	Online - would prefer face-to-face if time permitting	Face-to-face	Solutions will be digitally crowd-sourced

https://ihumanmedia.com/2015/09/14/gen-x-millennials-vs-baby-boomer-real-estate-baby-work-travel-politics-shopping/

Research Question

Does mobile use differ across generations?

Data/Measurement: Survey of Mobile Use

- May be collected by survey or user data
- Dependent variable: Number of texts per day
- Independent variable: Generation {B,X,M}

Texting Survey

What year were you born?

_____

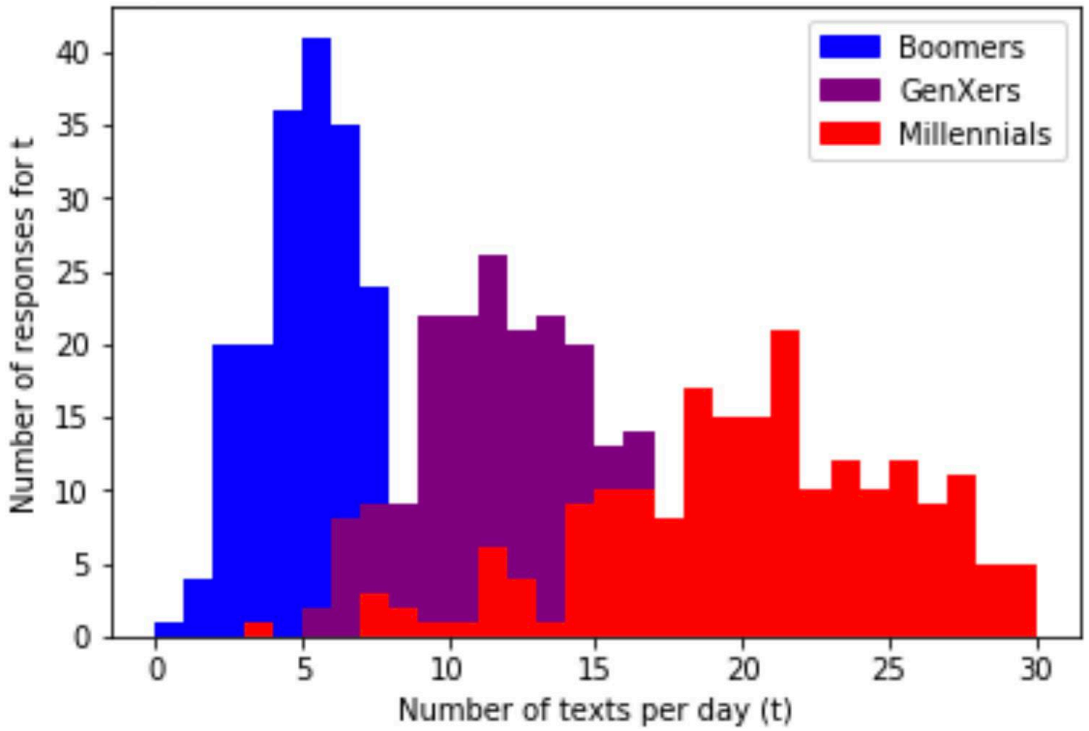
How many texts do you send per day?

_____

Generate “Mobile Usage” Data

- Imagine we conducted a survey of 200 baby boomers (B – born 1945-1960), 200 generation Xers (X – born 1961-1980) and 200 millennials (M – born 1981-1995)
- For the purposes of this exercise, let’s generate some simulated samples. We assume:
  - Baby Boomers send 5 texts per day on average with standard deviation 2
  - Gen Xers send 12 texts per day on average with standard deviation 3
  - Millennials send 20 texts per day on average with standard deviation 5

Visualise “Mobile Usage” Data



Setup: Comparing Behaviour across Groups

- Subjects are rows of data
  - Dependent variable is number of texts per day
  - Independent variable is generation {B,X,M}
- Q: Is there any difference in mobile usage between groups?
- H_0: Group means (or medians for nonparametic methods) are the same

Significance: Analysis of Variance (ANOVA)

- Tests the null hypothesis two or more groups have the same population mean
- Assumes:
  - The samples are independent ✓
  - Populations are normally distributed ✓
  - Standard deviations are equal ✓ not all the assumptions here are holding.

- http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.f_oneway.html#scipy.stats.f_oneway

Significance: Kruskall-Wallis H-test

- Nonparametric version of ANOVA
  - doesn't assume your data comes from a particular distribution such as normal distribution
- Assumes:
  - The samples are independent
- Note:
  - Not recommended for samples smaller than 5
  - Not as statistically powerful as ANOVA

Parametric test	Non-parametric test
• ANOVA • Student's t-test	• Kruskall-Wallis H-test • Mann-Whitney test

- Both ANOVA and Kruskall-Wallis H-test are extensions of the Mann-Whitney test and Unpaired Student's t-test used to compare the means of more than two populations.
  - http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.kruskal.html#scipy.stats.kruskal

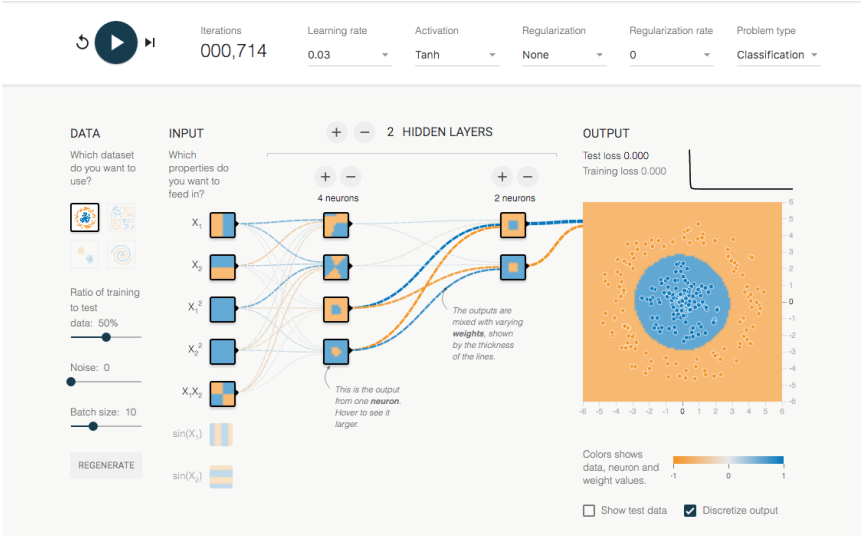

THE UNIVERSITY OF
SYDNEY

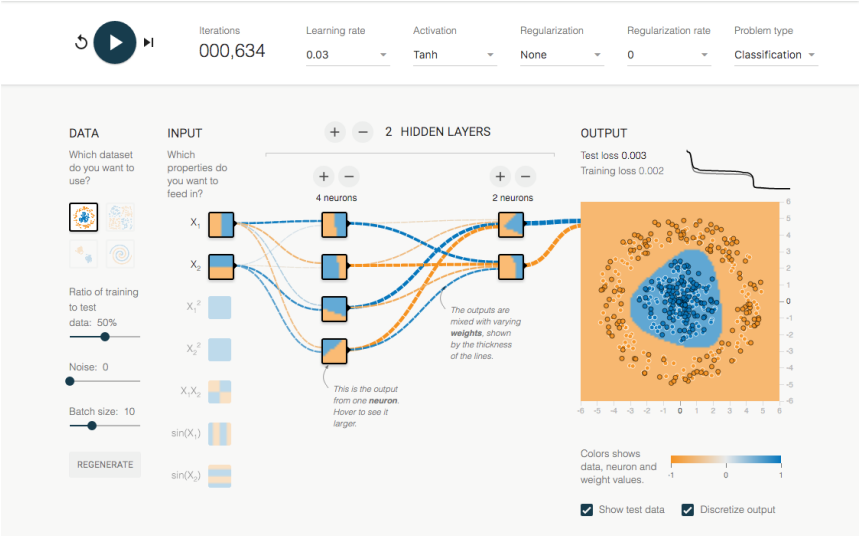
Model Evaluation

Testing which Approach is better Within Subjects

What does this mean?

Scenario: Comparing Classifiers





http://playground.tensorflow.org/

Research Question

Does my new model perform better?

Task: Spam/no-Spam Detection

- Let’s assume our classifiers predict whether an email is:
  - 1 (spam)
  - 0 (no-spam)

- Features are words, eg:

  Bitcoin_up, iphone.14.Pro, winner, P.a.Y.p.a.l, ph4rMa, v1agrA, Settlement4U

Measurement: Model Evaluation

how accurate is if it's spam or not spam?

- Need to measure accuracy of system output `S`
- Compare to gold-standard labelling `G`
- Define evaluation measure: score(`S`, `G`)
- http://scikit-learn.org/stable/modules/model_evaluation.html#model-evaluation

Measurement: Accuracy, Precision, Recall, F1

g just stands for gold-standard level.

	s=1	s=0
g=1	*TP* (true positives)	*FN* (false negatives)
g=0	*FP* (false positives)	*TN* (true negatives)

simple specificity/sensitivity table.

- Accuracy: *(TP+TN) / N*
  % correct over all instances
- Precision: *TP / (TP+FP)*
  % correct system predictions
- Recall: *TP / (TP+FN)*
  % correct gold labels
- F1: 2PR / (P+R)
  Harmonic mean of Precision and Recall

Evaluating Classifier Accuracy: Holdout & Cross-Validation Methods

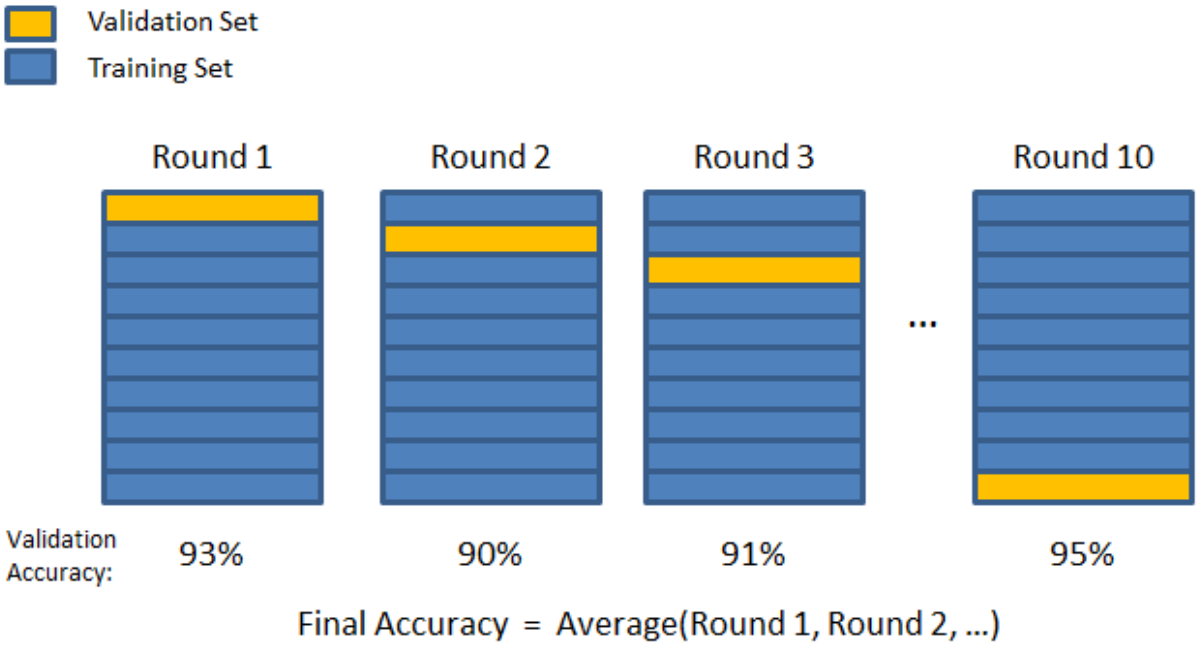
- **Holdout method**
  - splits the data randomly into two independent sets
    - **Training set** (e.g., 2/3) for model construction
    - **Test set** (e.g., 1/3) for accuracy estimation (also: *Validation Set*)
  - **Random sampling**: a variation of holdout
    - Repeat holdout *k* times; accuracy = avg. of the accuracies obtained

don't get confused by the terminologies

- **Cross-validation** (*k*-fold, where *k*=10 is most popular)
  - Randomly partition the data into *k* mutually exclusive subsets, each approx. equal size
  - Leave-One-Out is a particular form of cross-validation:
    - *k* folds where *k* = # of tuples, for small sized data

LOOCV

Data: Cross Validation



https://chrisjmccormick.wordpress.com/2013/07/31/k-fold-cross-validation-with-matlab-code/

Setup: Comparing Classifiers

Dep var:
– measure of accuracy

Indep var:
– Algorithm, features

- Subjects correspond to cross-validation folds
  - Dependent variable is some measure of accuracy (precision, recall, F1, …)
  - Independent variable is the algorithm, feature set, etc
- Q: Is my shiny, new model better?
- H_0: Accuracy is not better for new model
- http://sci2s.ugr.es/keel/pdf/algorithm/articulo/dietterich1998.pdf

Student t-tests aren't just testing two population means, whether they are equal.

Significance: Paired Student's t-test

- Tests the null hypothesis that two population means are equal
- Assumes
  - ***The samples are paired (e.g. before and after treatment)***
  - Populations are normally distributed
  - Standard deviations are equal
- Note
  - Multiply two-tailed p-value by 0.5 for one-tailed p-value
    (to test A>B, rather than A>B OR A<B)
- http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_rel.html#scipy.stats.ttest_rel

Significance: Paired tests for non-parametric data

- Nonparametric version of paired t-test
- Assumes
  - The samples are paired
  - Data is at least ordinal
- Note
  - Often used for ordinal data, e.g., Likert ratings
  - N should be large, e.g., ≥20

http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.wilcoxon.html#scipy.stats.wilcoxon


THE UNIVERSITY OF
SYDNEY

Data Mining

Types of Statistical Studies: Post-hoc Analysis

Post-hoc analysis = data mining

Post-hoc Analysis

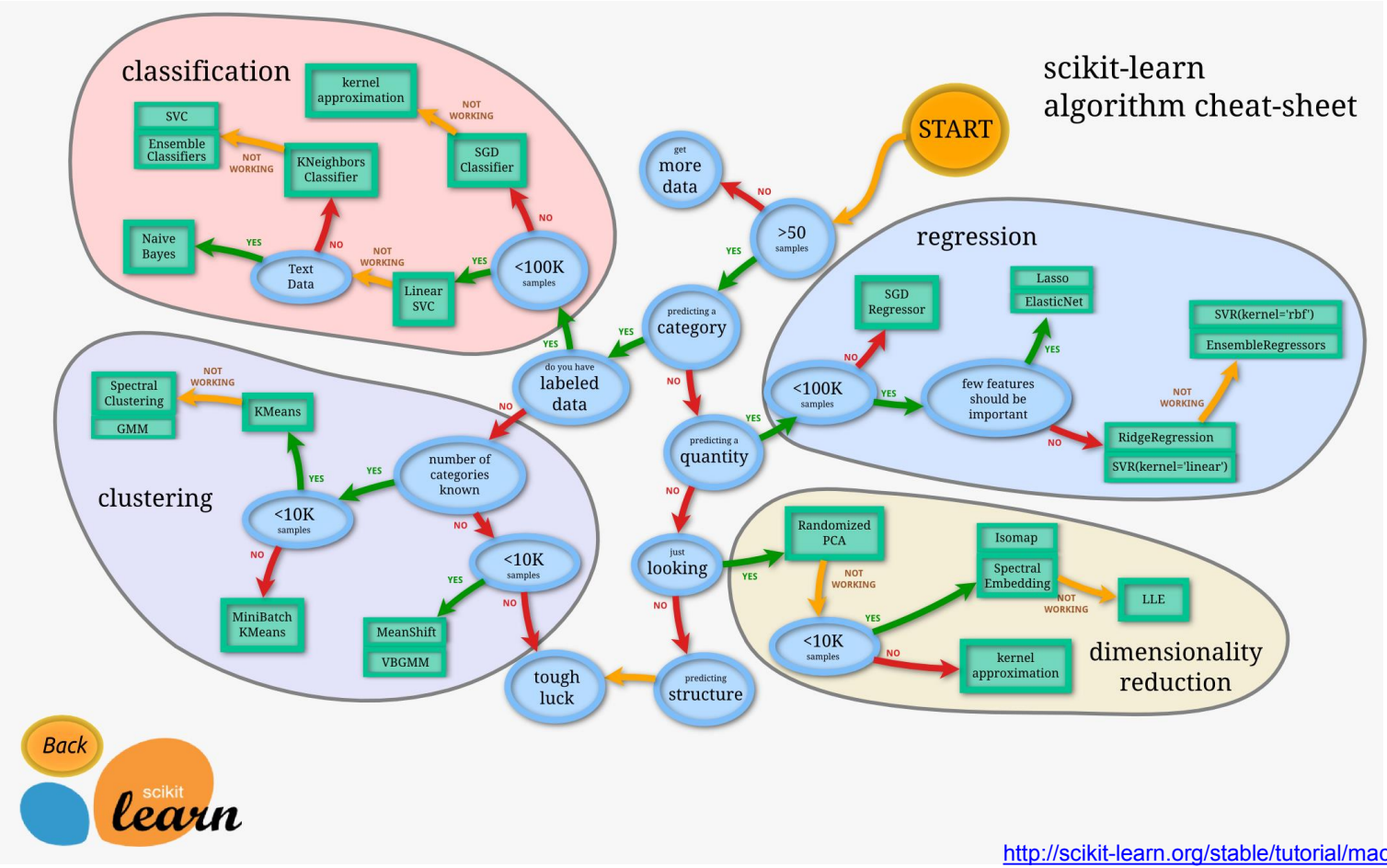
- testing hypotheses formulated after data collected
- aka data dredging OR **data mining**
- Test a wider range of hypotheses faster and cheaper
- May discover surprising patterns

- Be careful with our inferences
  - testing multiple hypotheses
  - control for family-wise error rate
- Confirm findings using experiments

What is Data Mining?

- Extraction of correlations and patterns from data
  https://en.wikipedia.org/wiki/Data_mining
- We'll focus on **unsupervised** machine learning techniques
  - Dimensionality reduction
  - ***Association rule mining***
  - Clustering
  - Outlier detection
  - Etc.
- Textbooks often include some supervised learning as well
- Grew out of database community, often business-oriented

Machine Learning Map from Scikit-learn



http://scikit-learn.org/stable/tutorial/machine_learning_map/

Not THAT kind of Data Mining!

- Sometimes refers to *p* hacking and other bad science, e.g.:
  - Deriving hypotheses from data exploration
  - Drawing unrepresentative samples to support a hypothesis
  - Making multiple comparisons to get a significant *p*-value

- https://en.wikipedia.org/wiki/Data_dredging

How a Cornell Scientist Turned Shoddy Data Into Viral Studies About How We Eat

Data Mining – A Cautionary Tale

https://www.buzzfeednews.com/article/stephaniemlee/brian-wansink-cornell-p-hacking#.bmxLG1XPpN

"When Siğirci started working with him, she was assigned to analyze a dataset from an experiment that had been carried out at an Italian restaurant. Some customers paid $8 for the buffet, others half price. Afterward, they all filled out a questionnaire about who they were and how they felt about what they'd eaten.

Somewhere in those survey results, the professor was convinced, there had to be a meaningful relationship between the discount and the diners. But he wasn't satisfied by Siğirci's initial review of the data.

"***I don't think I've ever done an interesting study where the data 'came out' the first time I looked at it***," he told her over email."

https://ericbrown.com/data-mining-cautions.htm


THE UNIVERSITY OF
SYDNEY

Association Rule Mining

Association Analysis

regression problem? → correlation vs. causation

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Market-basket transactions

TID: Transaction Identifier

Items: Transaction item set

How can businesses improve sales by analysing customer purchase data?

Slides adapted from Tan et al. Introduction to data mining.
http://www-users.cs.umn.edu/~kumar/dmbook/
http://www-users.cs.umn.edu/~kumar/dmbook/dmslides/chap6_basic_association_analysis.pdf

Association Rule Mining

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Market-basket transactions

TID: Transaction Identifier

Items: Transaction item set

- Predict occurrence of an item based on other items in the transaction, eg:

  {Diaper} → {Beer}

  {Milk,Bread} → {Eggs,Coke}

  {Beer,Bread} → {Milk}

- Note that arrows indicate co-occurrence, not causality

Definition: Itemset

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Market-basket transactions

TID: Transaction Identifier

Items: Transaction item set

- An **itemset** is a collection of one or more items
  {Milk,Bread,Diaper}
- A **k-itemset** is an itemset containing k items

Definition: Frequent Itemset

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Market-basket transactions

TID: Transaction Identifier

Items: Transaction item set

- **Support count** (σ) is the itemset frequency

$$s(\{\text{Milk,Diaper,Beer}\}) = 2$$

- **Support** (s) is the normalised itemset frequency

$$s = \frac{s(\{\text{Milk,Diaper,Beer}\})}{|T|} = \frac{2}{5}$$

- A **frequent itemset** has $s \geq min\_support$

Definition: Association Rule

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Market-basket transactions

TID: Transaction Identifier

Items: Transaction item set

- An **association rule** is an implication of the form X→Y where X and Y are itemsets
  {Milk,Diaper}→{Beer}

- **Confidence** (c) measures how often Y occurs in transactions with X

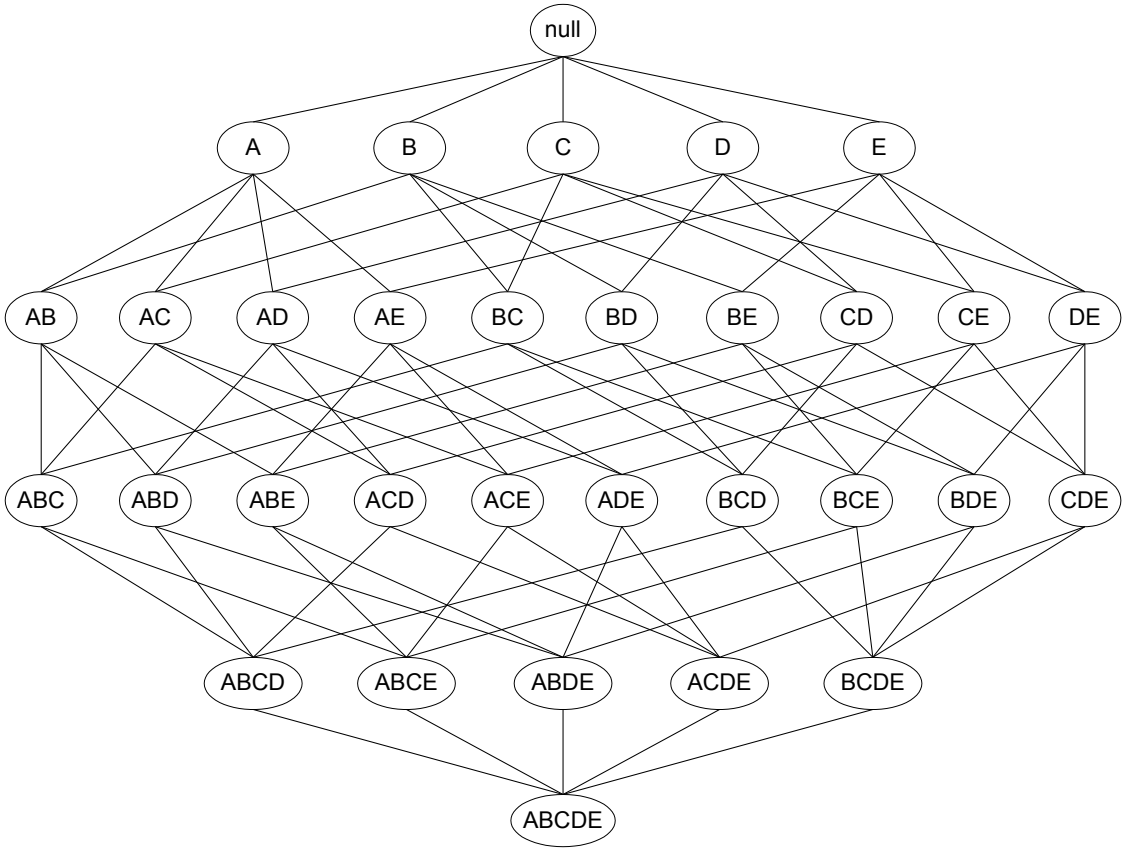
$$c = \frac{s(\{\text{Milk,Diaper,Beer}\})}{s(\{\text{Milk,Diaper}\})} = \frac{2}{3}$$

Mining Association Rules

1. **Frequent itemset generation**
   - Generate all itemsets with $s \geq min\_support$
2. **Rule generation**
   - Generate high-confidence rules from each frequent itemset
   - Each rule is a binary partitioning of a frequent itemset

Easy! But brute force enumerate is computationally prohibitive..

There are 2^d Candidate Itemsets!



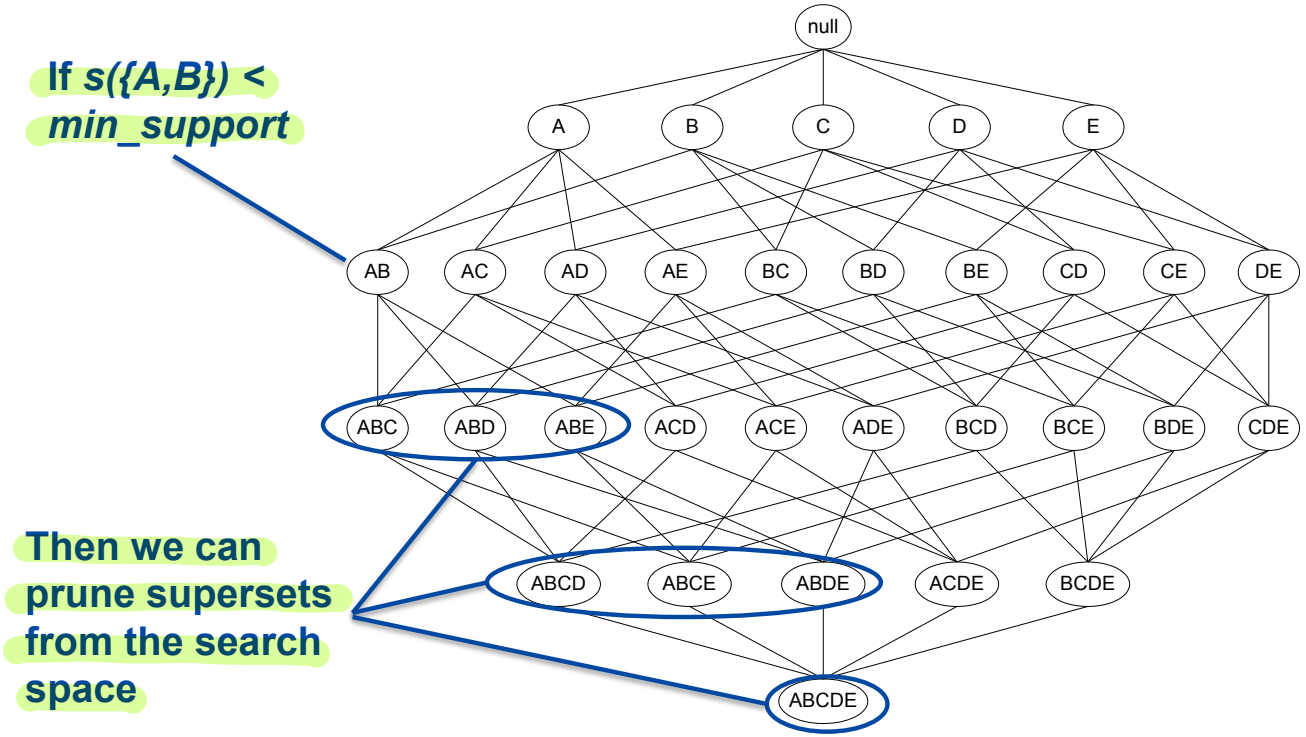
Enumeration of 2^5 candidate itemsets for {A,B,C,D,E}

Reducing the Number of Candidates

- **Apriori Principle**
  If an itemset is frequent, then all of its subsets are also frequent

- Conversely
  If an itemset is infrequent, then its supersets are also infrequent

Pruning the 2^d Candidate Itemsets



Apriori Algorithm for Generating Frequent Itemsets

While the list of (*k*-1)-itemsets is non-empty:

- Generate candidate *k*-itemsets
- Identify and keep frequent k-itemsets

Python DEMO

Create Initial 1-Itemsets

Add each item to the initial list of candidate itemsets

```
def createC1(dataset):
    "Create a list of candidate item sets of size one."
    c1 = []
    for transaction in dataset:
        for item in transaction:
            if not [item] in c1:
                c1.append([item])
    c1.sort()
    #frozenset because it will be a ket of a dictionary.
    return list(map(frozenset, c1))
```

Sort and return as list of sets

Identify Itemsets that meet the support threshold

```
def scanD(dataset, candidates, min_support):
    "Returns all candidates that meets a minimum support level"
    sscnt = {}
    for tid in dataset:
        for can in candidates:
            if can.issubset(tid):
                sscnt.setdefault(can, 0)
                sscnt[can] += 1

    num_items = float(len(dataset))
    retlist = []
    support_data = {}
    for key in sscnt:
        support = sscnt[key] / num_items
        if support >= min_support:
            retlist.insert(0, key)
        support_data[key] = support
    return retlist, support_data
```

Calculate support counts for each candidate

normalised itemset frequency.

threshold.

Check whether candidates meet threshold

Generate the next list of candidates

go back to 15:47. also code up Apriori Algorithm

(k-1)-itemsets

Iterate through all pairs of itemsets

```
def aprioriGen(freq_sets, k):
    "Generate the joint transactions from candidate sets"
    retList = []
    lenLk = len(freq_sets)
    for i in range(lenLk):
        for j in range(i + 1, lenLk):
            L1 = list(freq_sets[i])[:k - 2]
            L2 = list(freq_sets[j])[:k - 2]
            L1.sort()
            L2.sort()
            if L1 == L2:
                retList.append(freq_sets[i] | freq_sets[j]) # | is set union
    return retList
```

Check whether pairs differ by a single item

A|B returns the union of A and B

Generate all Frequent Itemsets

```
def apriori(dataset, min_support=0.5):
    "Generate a list of candidate item sets"
    C1 = createC1(dataset)
    D = list(map(set, dataset))
    L1, support_data = scanD(D, C1, min_support)
    L = [L1]
    k = 2
    while (len(L[k - 2]) > 0):
        Ck = aprioriGen(L[k - 2], k)
        Lk, supK = scanD(D, Ck, min_support)
        support_data.update(supK)
        L.append(Lk)
        k += 1

    return L, support_data
```

Initialise L with frequent 1-itemsets

While the list of (k-1)-itemsets is non-empty:

Generate candidate k-itemsets

Identify frequent k-itemsets

Keep frequent k-itemsets

Identify rules that meet the confidence threshold

Frequent itemset (rule components)

Possible consequences (RHS of rule)

Rule accumulator

```
def calc_confidence(freqSet, H, support_data, rules, min_confidence=0.7):
    "Evaluate the rule generated"
    pruned_H = []
    for conseq in H:
        conf = support_data[freqSet] / support_data[freqSet - conseq]
        if conf >= min_confidence:
            #print(freqSet - conseq, '--->', conseq, 'conf:', conf)
            rules.append((freqSet - conseq, conseq, conf))
            pruned_H.append(conseq)
    return pruned_H
```

Calculate confidence

Return consequences that pass the confidence threshold

Add rule to accumulator if $c \geq min\_confidence$

Recursively Evaluate Rules

Need at least 1 item for LHS

Generate candidate consequence itemsets

Update rules and return consequences that pass confidence threshold

```
def rules_from_conseq(freqSet, H, support_data, rules, min_confidence=0.7):
    "Generate a set of candidate rules"
    m = len(H[0])
    if (len(freqSet) > (m + 1)):
        Hmp1 = aprioriGen(H, m + 1)
        Hmp1 = calc_confidence(freqSet, Hmp1,  support_data, rules, min_confidence)
        if len(Hmp1) > 1:
            rules_from_conseq(freqSet, Hmp1, support_data, rules, min_confidence)
```

Recurse with new consequence candidates

Mine all Association Rules

```
def generateRules(L, support_data, min_confidence=0.7):
    """Create the association rules
    L: list of frequent item sets
    support_data: support data for those itemsets
    min_confidence: minimum confidence threshold
    """
    rules = []
    for i in range(1, len(L)):
        for freqSet in L[i]:
            H1 = [frozenset([item]) for item in freqSet]
            #print("freqSet", freqSet, 'H1', H1)
            if (i > 1):
                rules_from_conseq(freqSet, H1, support_data, rules, min_confidence)
            else:
                calc_confidence(freqSet, H1, support_data, rules, min_confidence)
    return rules
```

For each k
For each k-itemset

Initial consequence candidates

Recursively evaluate rules if k>1

Evaluate H1 only

Apriori_python

- There are also third party libraries available which implement the apriori algorithm directly
- Example:

```
from apriori_python import apriori
dataset =   [['A', 'C', 'D'], ['B', 'C', 'E'], ['A', 'B', 'C','E'],['B', 'E']]
freqItemSet, rules = apriori(dataset, minSup=0.7, minConf=0.0)
for r in rules:
      print('{} ==> {} (c={})'.format(*r))
```


THE UNIVERSITY OF
SYDNEY