COMP6940: BIG DATA AND DATA VISUALISATION

LECTURE #4: DATA ANALYSIS AND TASKS

INZAMAM RAHAMAN

AGENDA

1. Types/Definitions of data analysis
2. Machine Learning and Types of Machine Learning
3. Descriptive and Inferential Analysis
    1. Hypothesis Testing
    2. Clustering
4. Predictive Analytics Methods
    1. Classification
    2. Regression

DEFINITIONS OF DATA ANALYSIS

- After data has been collected, clean, and integrated, data must then be analysed to extract value and to inform decisions
- Depending on the data available and the problem context, there are different types of data analysis tasks that would need to be performed
- Taxonomising the problem is important to help decide what methods and approaches can be used
- In this lecture, we will be concerned primarily with diagnosing the problem at hand
  - Techniques will be covered in detail in future lectures or other classes
- Variable types:
  - Covariates: remember these are the columns in your data
  - Dependent variable: a variable that you are trying to model in terms of other variables; also called target
  - Independent variable: one of several variables that are used to model another variable
  - Identifying your variables is important first step

DEFINITIONS OF DATA ANALYSIS

- **Descriptive Analysis**: concerned with determining high-level trends and patterns within the data. Runs the gamut from measures of central tendency to more sophisticated approaches like clustering and market-basket analysis. Some descriptive analysis done with EDA
- **Inferential Analysis**: concerned with drawing and testing overarching conclusions about a population using a random sample from that population. Confirmatory Data Analysis/Hypothesis testing falls under this category. Bayesian Inference also fits here (covered in Computational Statistics)
- **Predictive Analysis**: concerned with using historical data to try and predict future state. Often makes use of machine learning and data mining techniques. Output from descriptive and inferential analysis helps.
- **Prescriptive Analysis**: concerned with prescribing actions. Uses results of descriptive, inferential, and predictive analysis. Involves techniques from operations research (optimisation, computer simulation, mathematical modelling)
- **Causal Analysis/Inference**: concerned with determining causal effects in observation or quasi-experimental settings

ML SLIDES

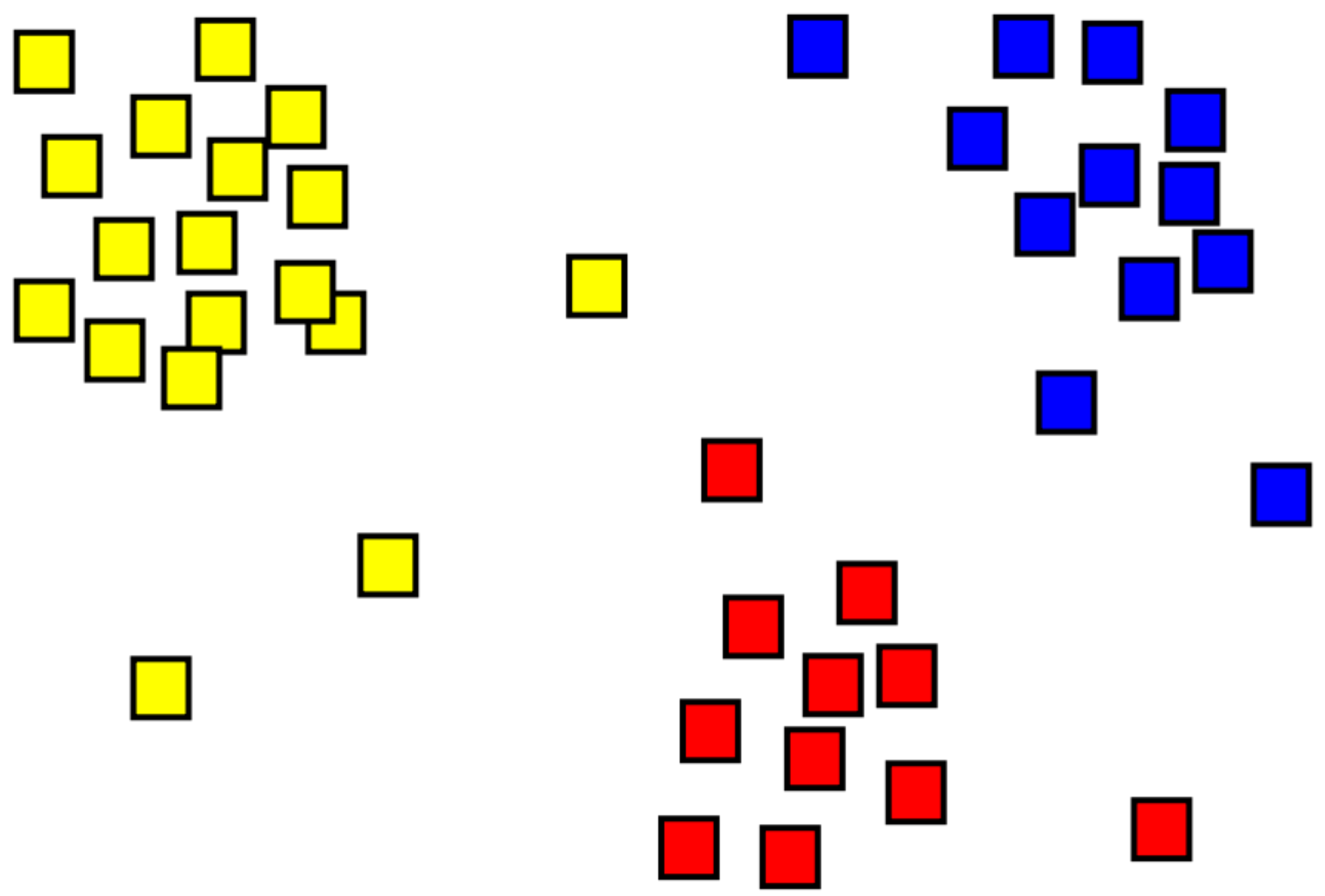
DESCRIPTIVE ANALYSIS

- Reporting of high-level statistics is important in descriptive analysis
- Measures of central tendency: mean, mode, median
    - Should also check percentiles (25th and 75th)
- Measures of spread: IQR, std. dev, variance
- Distribution properties: skewness and kurtosis
- Histograms and outlier analysis are important
    - Median and IQR more informative than mean and std. dev when data has significant and impactful outliers
- Also more sophisticated techniques involved

DESCRIPTIVE ANALYSIS

CLUSTERING

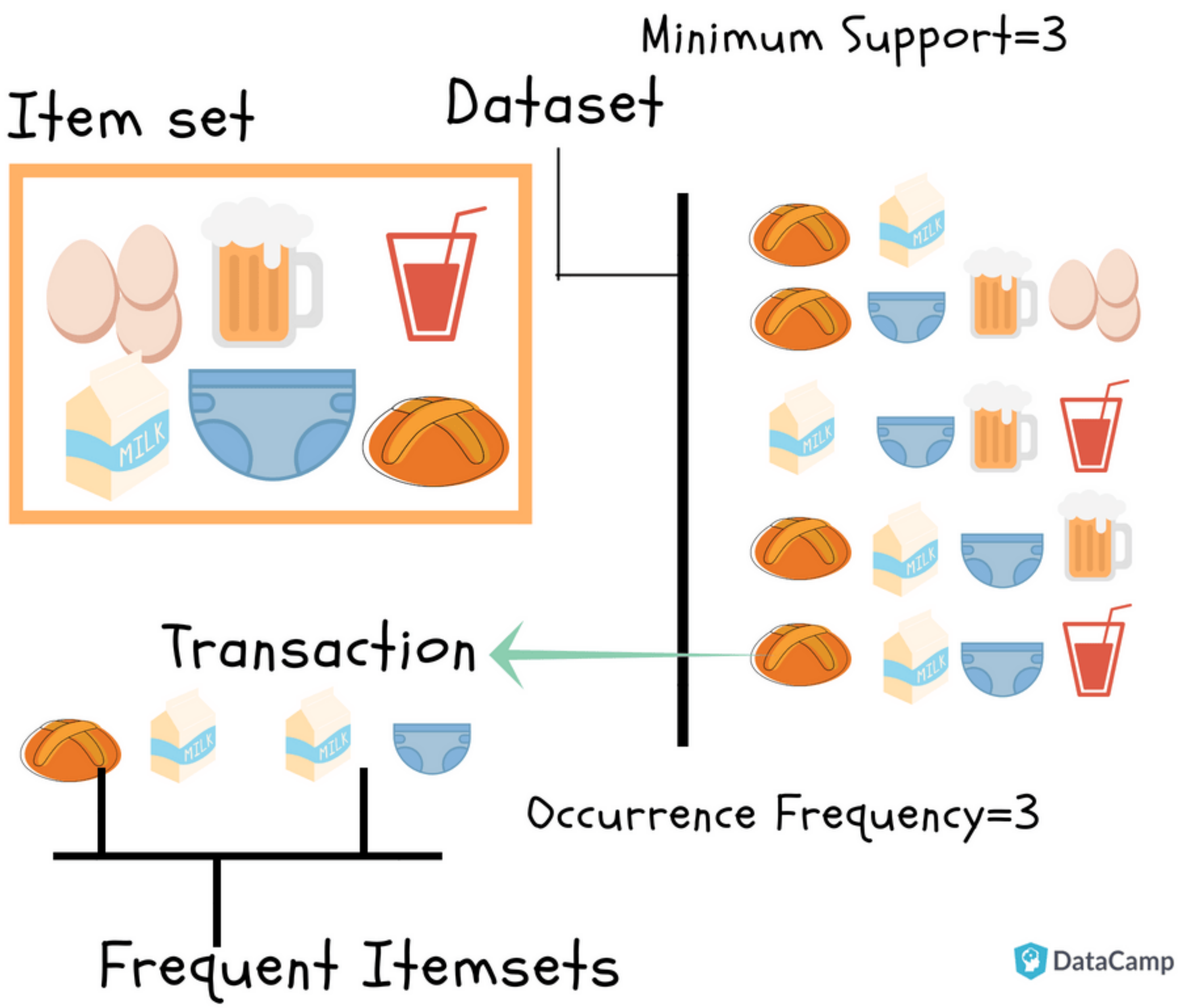
- Concerned with finding grouping in the data with similar covariate profiles
- Useful for finding sub-groups in your data for problems like customer profiling and segmentation
- Used in bioinformatics to find different groupings in data
- Different clustering methods that use different notions of similarity to segment space into clusters



DESCRIPTIVE ANALYSIS

MARKET-BASKET ANALYSIS

- Used to find items that are associated with one another
- Quintessential case is grocery cart:
  - Find products that are commonly bought together
- Can be used to create data-driven association rules
  - E.g. strawberries **AND** blueberries **=>** almond milk
- Can be used for many other contexts
- Can be used to build rule-driven recommendation systems

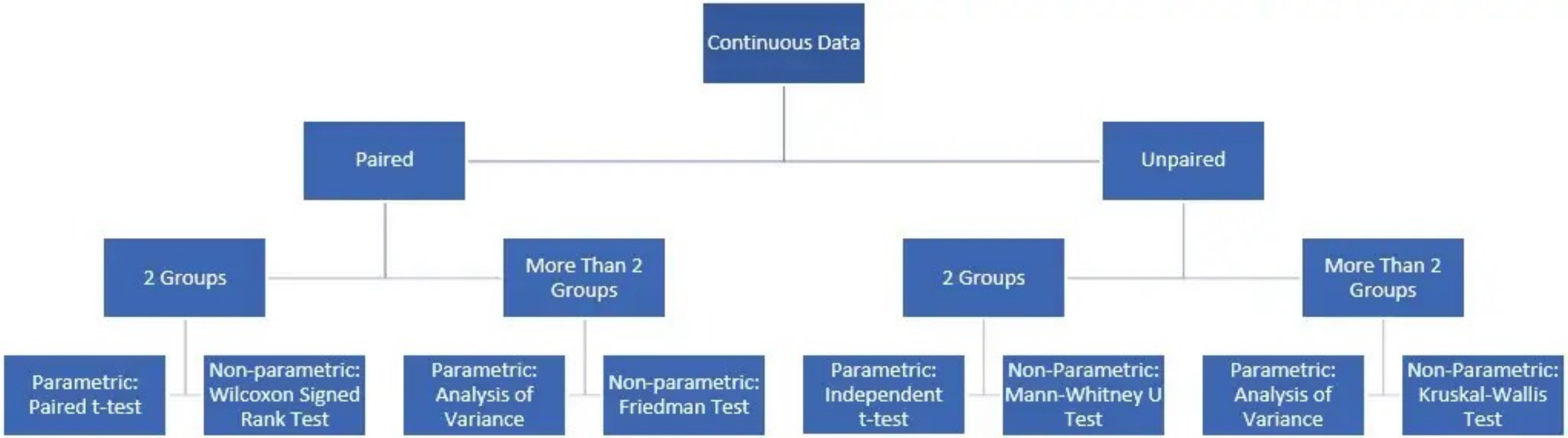


INFERENTIAL ANALYSIS

- Concerned with drawing general conclusions about populations from random samples
  - Usually want to test some hypothesis
- Important concepts:
  - H_a: alternative hypothesis - what we set out to adduce evidence for
  - H_0: null hypothesis - “no effect” hypothesis
  - P-value: probability of seeing a result as extreme as observed under H_0
  - Significance level: a p-value under which we accept our H_a . If p-value is below significance level, result is statistically significant
  - Type I error: incorrectly rejecting H_0
  - Type II error: incorrectly rejecting H_a
- In Python, statsmodel package is useful

INFERENTIAL ANALYSIS

TWO-SAMPLE TESTS



INFERENTIAL ANALYSIS

MORE THAN TWO SAMPLES

- When you have more than two samples - three or more with associated factor “levels”
    - E.g. imagine want compare yield obtained from three different fertilisers used in an experiment
- Methods: one-way ANOVA, two-way ANOVA, MANOVA
- After determining if differences are statistically significant, need to run a post-test such as a Tukey’s test to determine which differences are significant

INFERENTIAL ANALYSIS

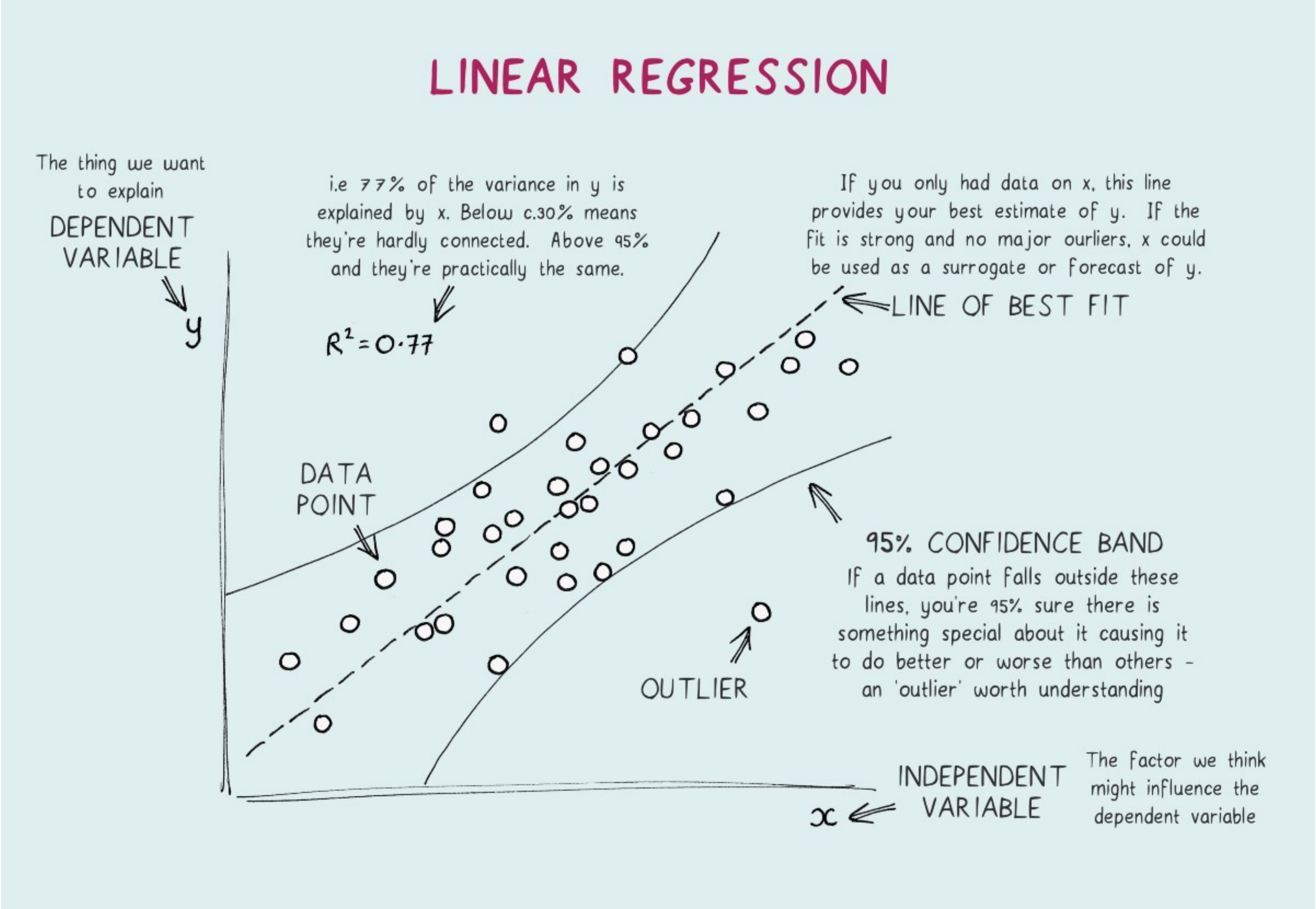
MORE THAN TWO SAMPLES

- When you have more than two samples - three or more with associated factor “levels”
  - E.g. imagine want compare yield obtained from three different fertilisers used in an experiment
- Methods: one-way ANOVA, two-way ANOVA, MANOVA
- After determining if differences are statistically significant, need to run a post-test such as a Tukey’s test to determine which differences are significant

INFERENTIAL ANALYSIS

REGRESSION ANALYSIS

- Setting: independent variables against dependent variable
- Fit best-fit line of independents against dependent
  - $y = mx + c$
- Check coefficients to see effects of independents on dependents
  - H_0: no effect
  - H_a: effect



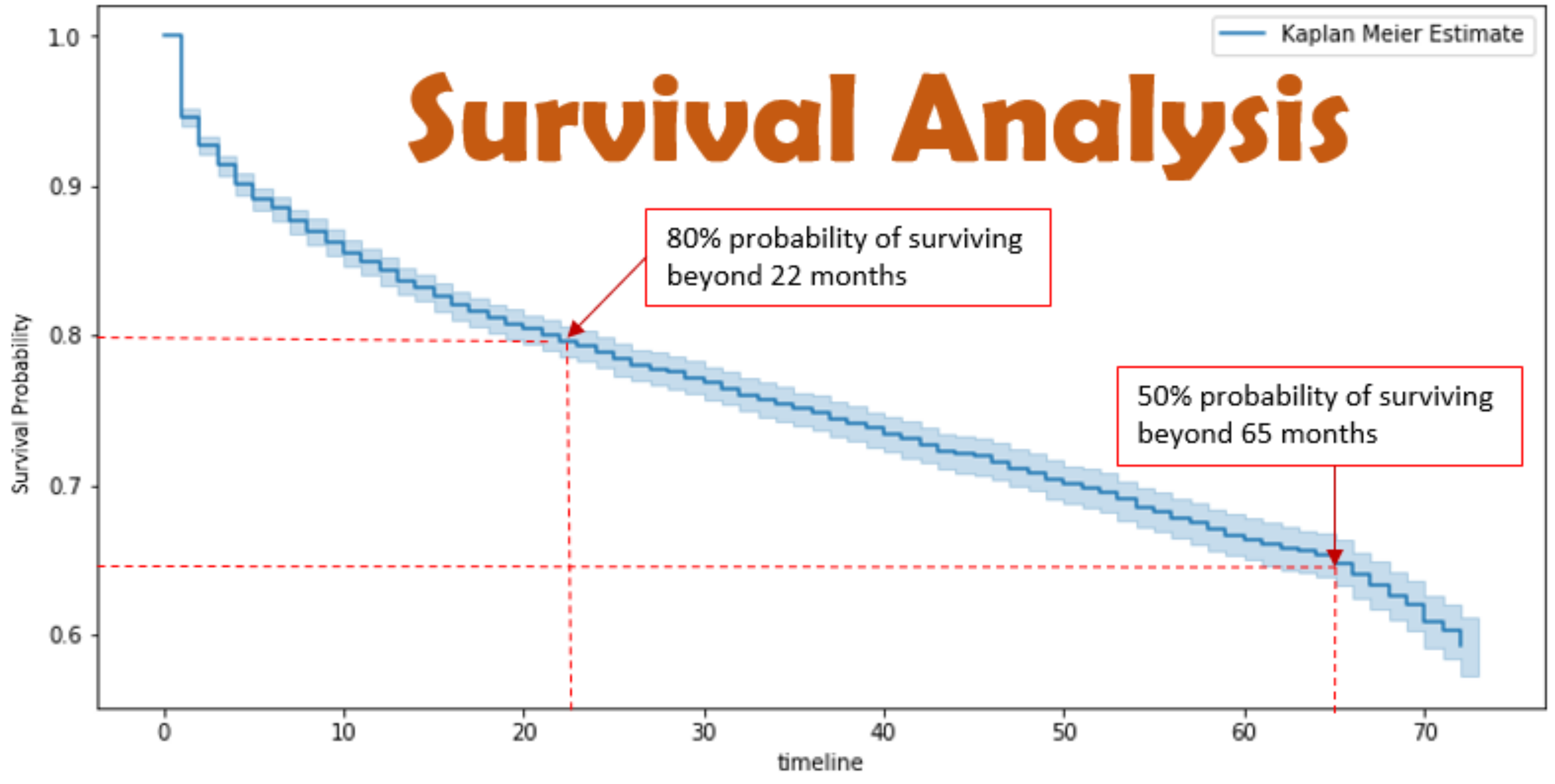
INFERENTIAL ANALYSIS

REGRESSION ANALYSIS

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  GRADE   R-squared:                       0.416
Model:                            OLS   Adj. R-squared:                  0.353
Method:                 Least Squares   F-statistic:                     6.646
Date:                Wed, 02 Nov 2022   Prob (F-statistic):            0.00157
Time:                        17:12:47   Log-Likelihood:                -12.978
No. Observations:                  32   AIC:                             33.96
Df Residuals:                      28   BIC:                             39.82
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
GPA            0.4639      0.162      2.864      0.008       0.132       0.796
TUCE           0.0105      0.019      0.539      0.594      -0.029       0.050
PSI            0.3786      0.139      2.720      0.011       0.093       0.664
const         -1.4980      0.524     -2.859      0.008      -2.571      -0.425
==============================================================================
Omnibus:                        0.176   Durbin-Watson:                   2.346
Prob(Omnibus):                  0.916   Jarque-Bera (JB):                0.167
Skew:                           0.141   Prob(JB):                        0.920
Kurtosis:                       2.786   Cond. No.                         176.
==============================================================================
```

INFERENTIAL ANALYSIS
SURVIVAL ANALYSIS

- Used to estimate the survival function
- Time to event
- Can be used to determine hazard ratios that estimate relative odds of suffering from event



INFERENTIAL ANALYSIS

CHI-SQUARED TEST

- Chi-squared test is used to compare counts between two samples
- Used for goodness-of-fit evaluations
- Can be used for non-count samples in some cases
- Can also be used to test for independence of variables

INFERENTIAL ANALYSIS
CHI-SQUARED TEST

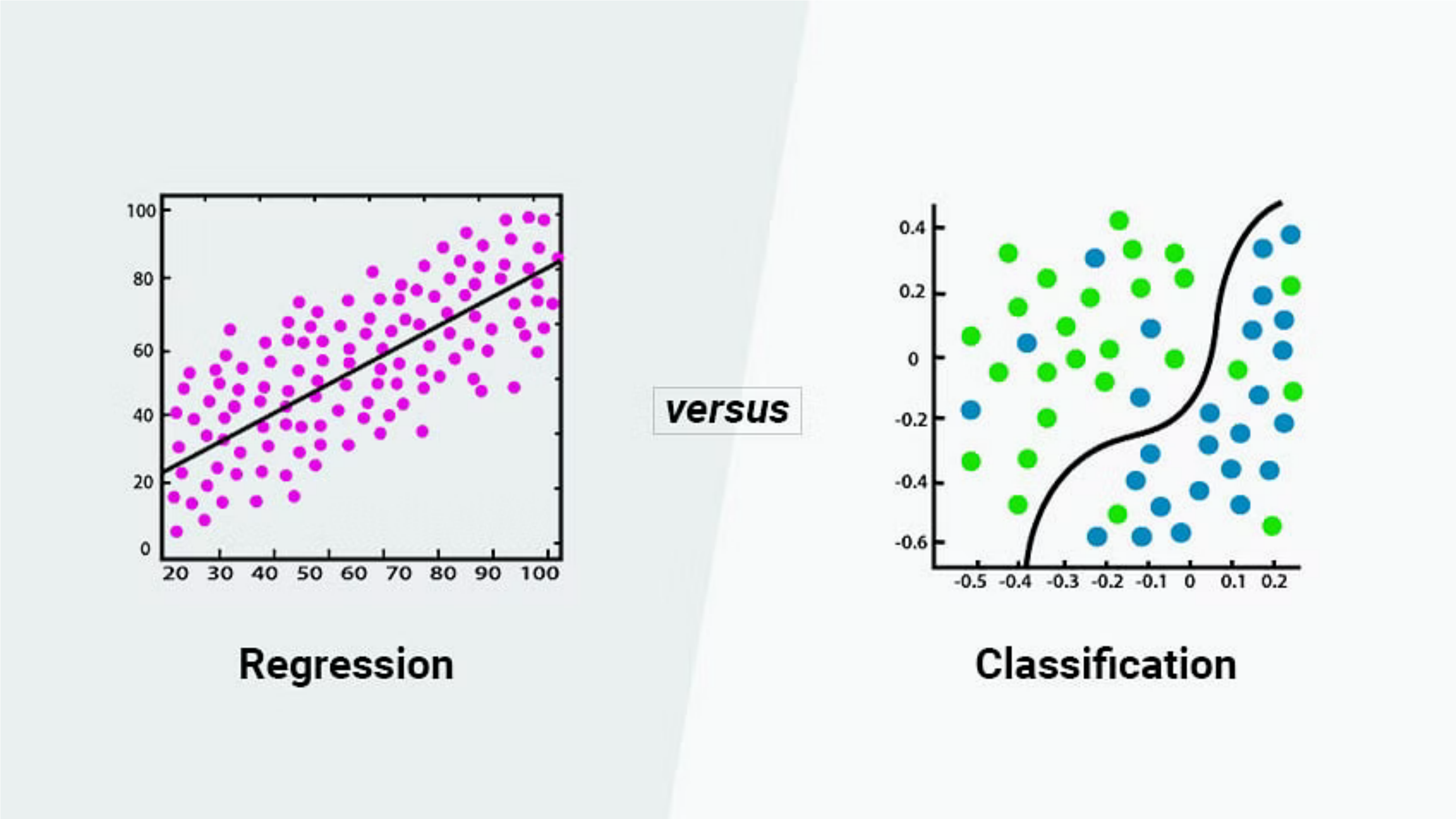
- Chi-squared test is used to compare counts between two samples
- Used for goodness-of-fit evaluations
- Can be used for non-count samples in some cases
- Can also be used to test for independence of variables

PREDICTIVE ANALYSIS
SETTING

- Consider that we have pairs of independent variables-dependent variables
- Want to use independent variables to predict dependent variables for future cases
    - i.e. want to predict the future by modelling past relationships
- Two main types of predictive analysis problems
    - Regression: continuous and/or infinite target (e.g. predict house selling price using square feet, location, num bedrooms, num bathrooms, etc…)
    - Classification: discrete finite target (e.g. use blood work results [ESR, WBC, CBC, CRP] to determine if a patient has sepsis or not)

PREDICTIVE ANALYSIS
SETTING

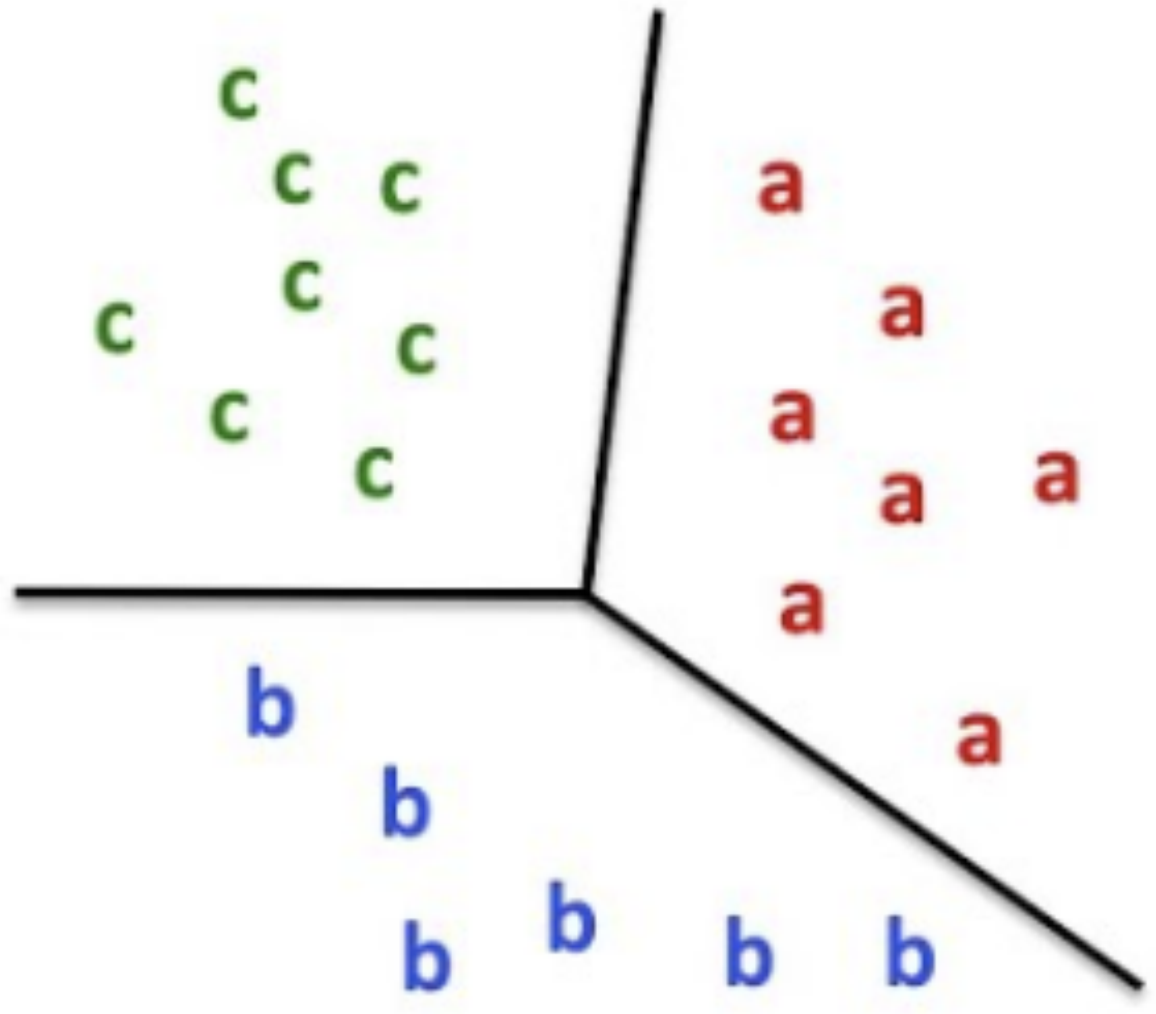
- Consider that we have pairs of independent variables-dependent variables
- Want to use independent variables to predict dependent variables for future cases
  - i.e. want to predict the future by modelling past relationships
- Two main types of predictive analysis problems
  - Regression: continuous and/or infinite target (e.g. predict house selling price using square feet, location, num bedrooms, num bathrooms, etc…)
  - Classification: discrete finite target (e.g. use blood work results [ESR, WBC, CBC, CRP] to determine if a patient has sepsis or not)

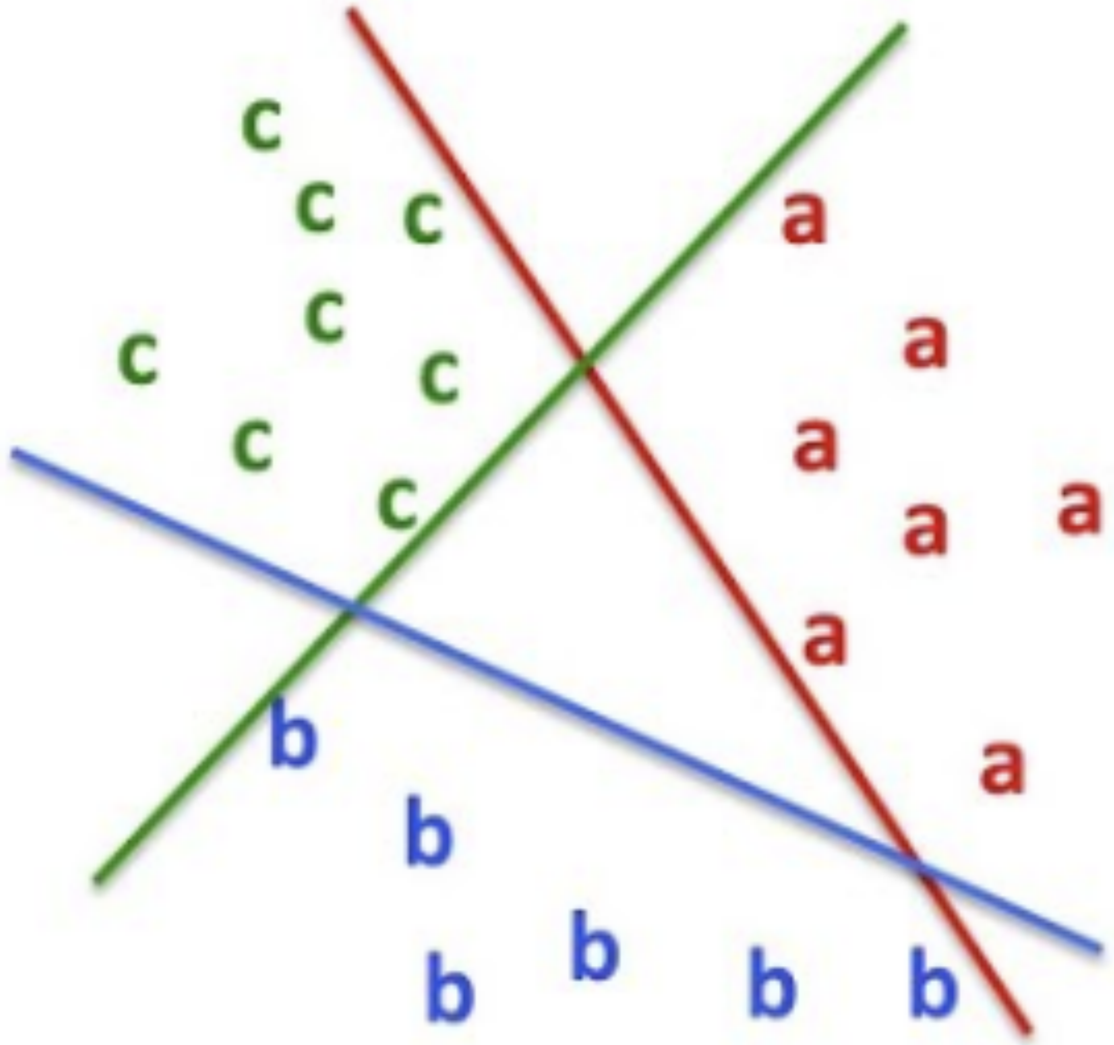


100
80
60
40
20
0
20 30 40 50 60 70 80 90 100
versus
0.4
0.2
0
-0.2
-0.4
-0.6
-0.5 -0.4 -0.3 -0.2 -0.1 0 0.1 0.2
Regression
Classification

Multi-class vs. Binary classification

- Multi-class:
  - classes mutually exclusive:
    - instance is either a or b or c
    - even if it's an outlier
  - NB, kNN, DT, logistic
- Binary:
  - one-vs-rest:
    - {a} vs {not a}, {b} vs {not b}
  - classes may overlap
    - instance can be both a and b
    - can be in none of the classes
  - SVM, logistic, perceptron





PREDICTIVE ANALYSIS

MODEL DRIFT

- Predictive analytics implicitly models $P(Y|X)$
- However, the world around a model can change, leading to changes in the distributions implicitly used and learnt by the machine learning model - this causes model drift
- Types of model drift
  - $P(Y|X)$ - concept drift
  - $P(X)$ - covariate drift
  - $P(Y)$ - target drift
- Different models are more robust to different types of drift
- Concept drift usually only handled by re-training the model :(

PREDICTIVE ANALYSIS
MEASURING PERFORMANCE

- We typically use model metrics with nice mathematical properties
  - MSE, LogLoss, F1, AUC, MAPE, etc…
- However, these don’t necessarily translate to impact!
- When developing model, simulate impact of model on something people actually care about
  - Usually money earned or money saved

PRESCRIPTIVE ANALYSIS

- Will cover these techniques in OR course
- Concerned with informing decisions directly
- Can involve non-trivial computer simulations to assess outcomes and values of different actions
    - Be careful with computer simulations, they can be “doomed to succeed”

CAUSAL ANALYSIS

- Most of the above techniques use associations
- Associations are necessary but NOT SUFFICIENT for determining causation
- Different techniques are sometimes needed to determine causal relationships
- Casual relations are useful for designing interventions to achieve certain outcomes
- Important in medicine in particular
- No formal course and out of scope for this course :(

QUESTIONS?