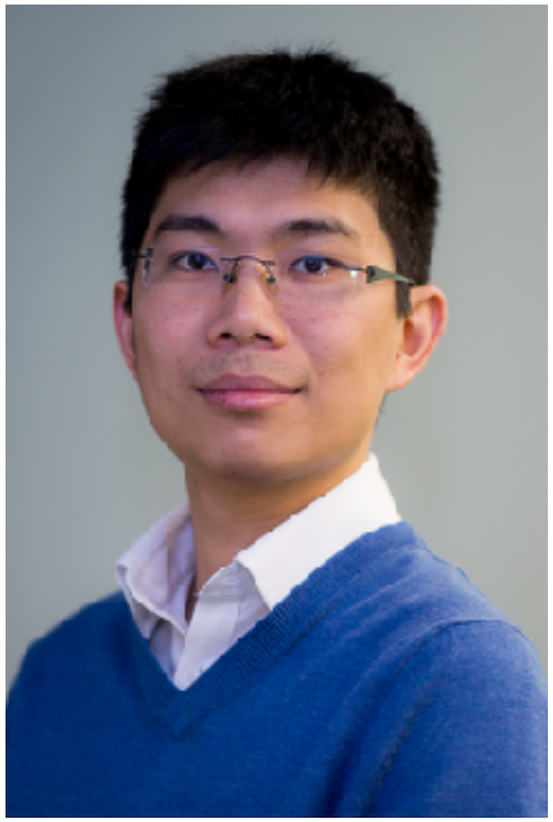
Kevin Wang

Sydney University
Mathematics Society

New Law 024, 4th
October, Thursday, 1-2pm

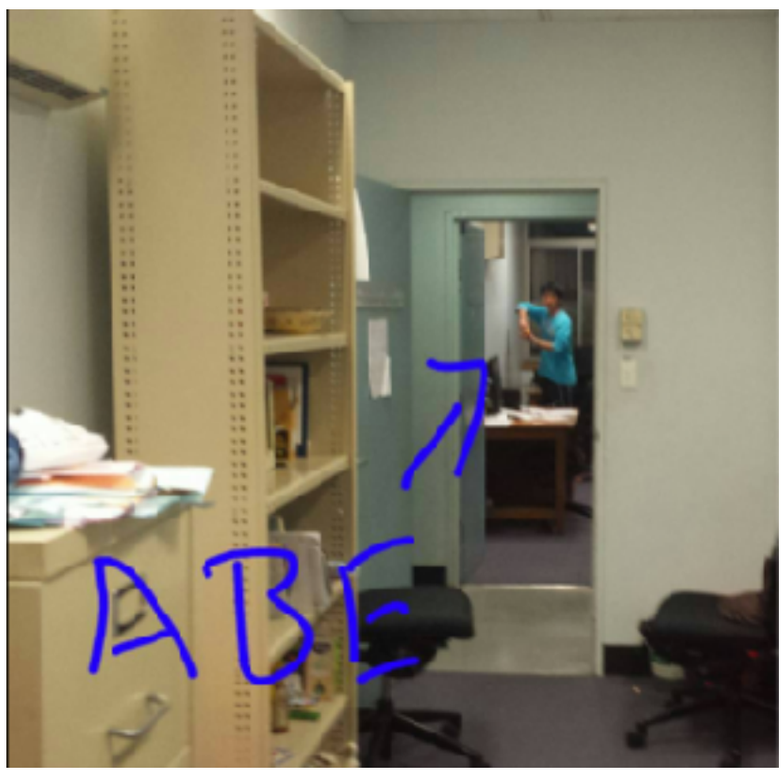
CRICINFORMATICS: How To Play Cricket With Your Mates While Pretending To Be Doing Research

A Quick Intro About Myself



- PhD candidate in Biostatistics/ Bioinformatics and Postgraduate Teaching Fellow at the School of Mathematics and Statistics

- Smashed way too many stuff in Carslaw playing office cricket



- I am here at the risk of being excommunicated by my supervisors

My Talk In One Slide

- Bioinformatics - an interdisciplinary field that uses mathematics, statistics and computer science tools to understanding biological data
- Cricinformatics - applying all those bioinformatics methods to cricket data
- Aim: ~~Shameless advertising to grab more students into our research group~~

WHAT IS CRICKET?

What Is Data?

- Each row is an observation and each column is a variable.

	A	B	C	D	E	F	G	H	I	J
1	Player	Career Start	Career End	Matches Play	Innings Batte	Not Outs	Runs Scored	Highest Innir	Highest Innir	Batting Avg
2	DG Bradman (1928-1948)	1928	1948	52	80	10	6996	334	334	99.94
3	MN Nawaz (2002-2002)	2002	2002	1	2	1	99	78*	78	99
4	VH Stollmeyer (1939-1939)	1939	1939	1	1	0	96	96	96	96
5	DM Lewis (1971-1971)	1971	1971	3	5	2	259	88	88	86.33
6	Abul Hasan (2012-2013)	2012	2013	3	5	3	165	113	113	82.5
7	RE Redmond (1973-1973)	1973	1973	1	2	0	163	107	107	81.5
8	BA Richards (1970-1970)	1970	1970	4	7	0	508	140	140	72.57
9	H Wood (1888-1892)	1888	1892	4	4	1	204	134*	134	68
10	TA Blundell (2017-2017)	2017	2017	2	3	1	136	107*	107	68
11	CS Dempster (1930-1933)	1930	1933	10	15	4	723	136	136	65.72

- The job of a statistician is to make data to sing a harmonious song and inform us of something useful.

In God We Trust
All Others Bring Data

Edwards Deming

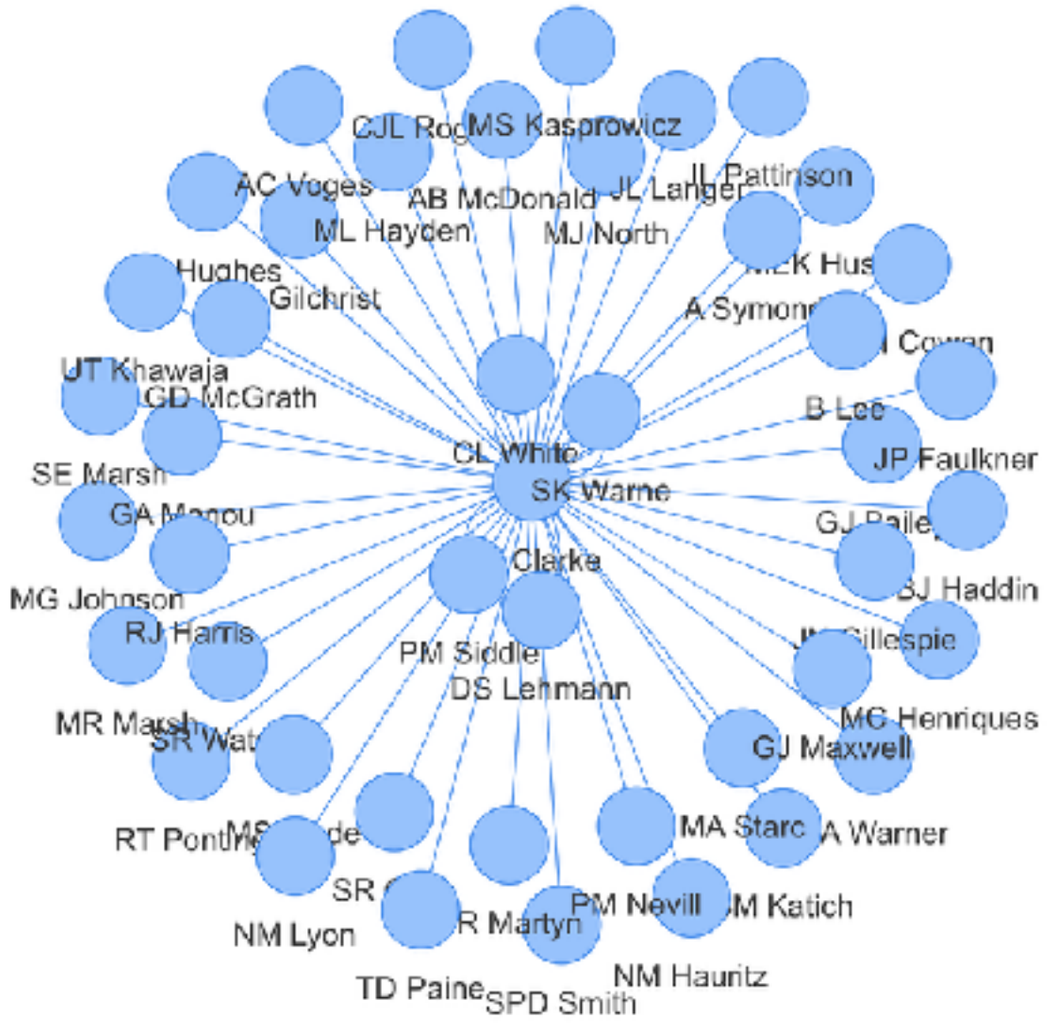
THE ASHES

PARTNERSHIP NETWORK ANALYSIS

Batting Partnership

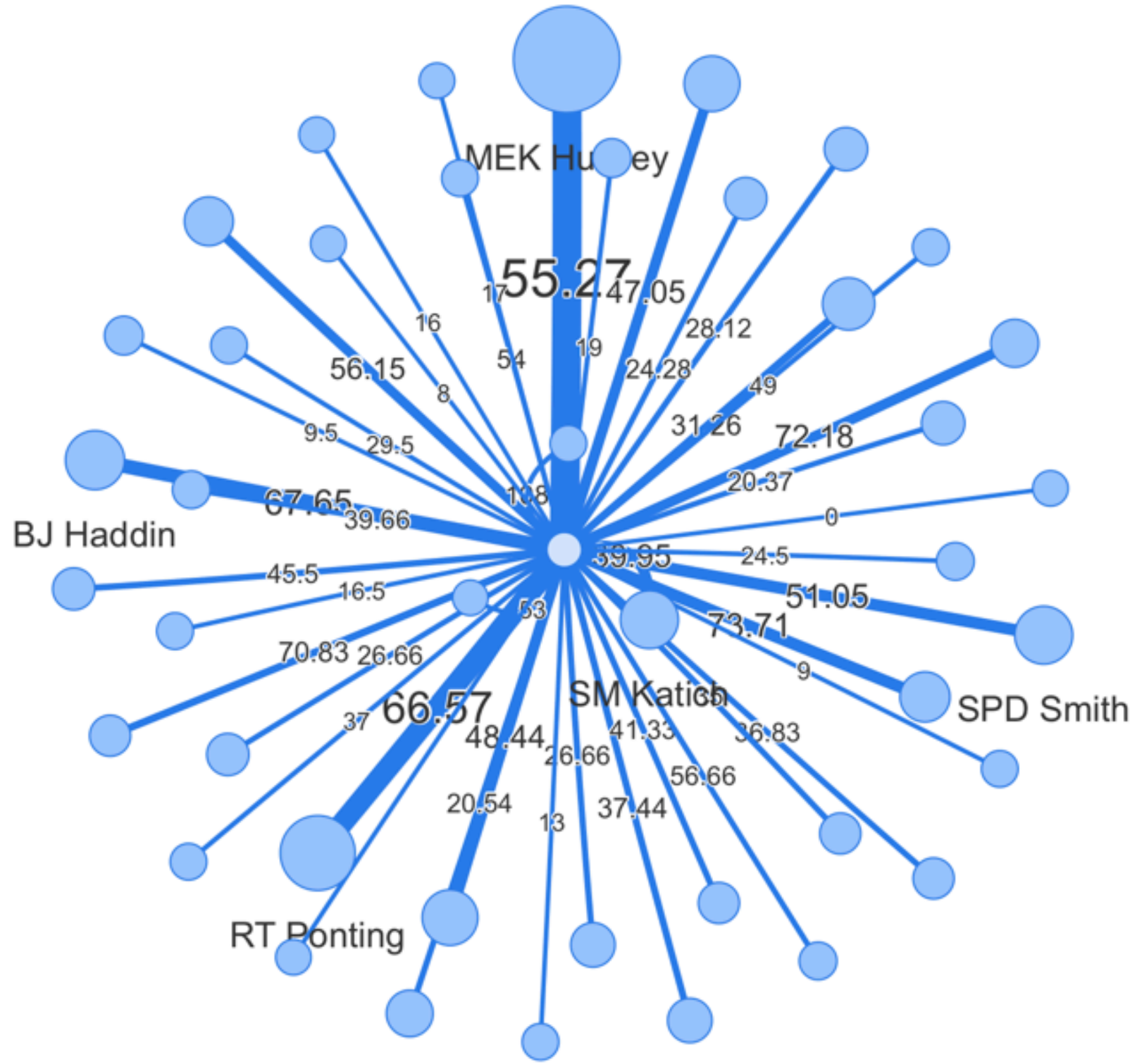
- In a batting Innings, two players must bat together in a **partnership** to score runs

- All Michael Clarke's batting partners



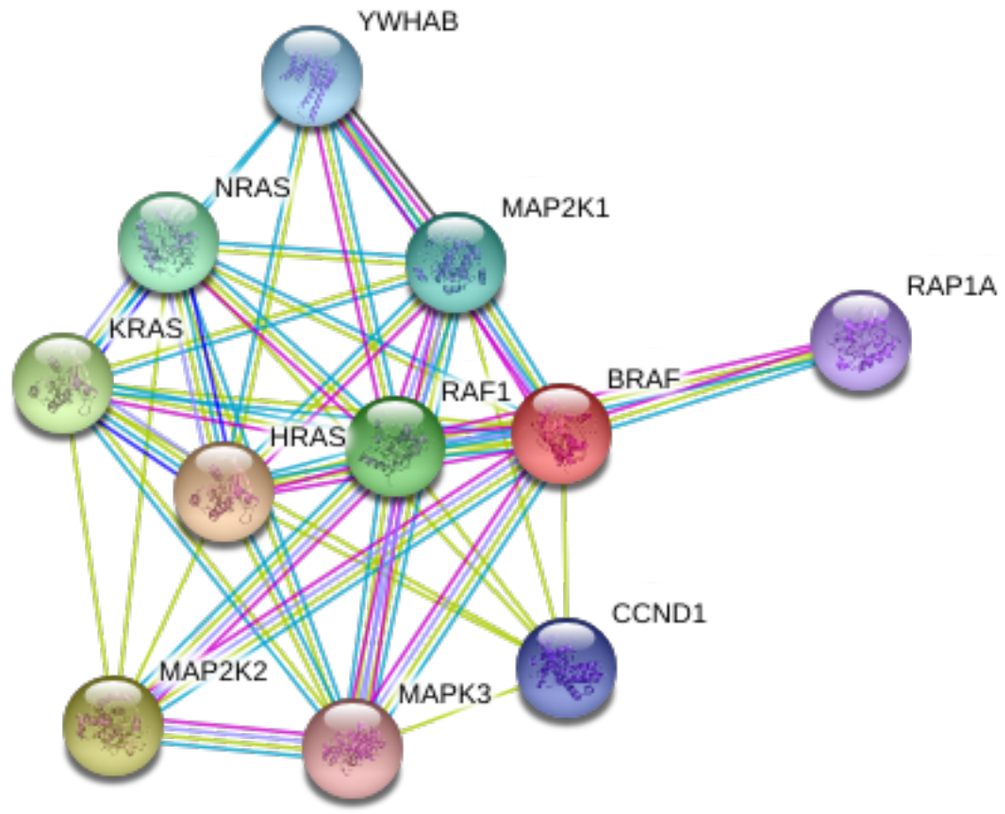
Network Visualisation

Good data visualisation can immediately bring out the important features of the data.



Relations To Bioinformatics

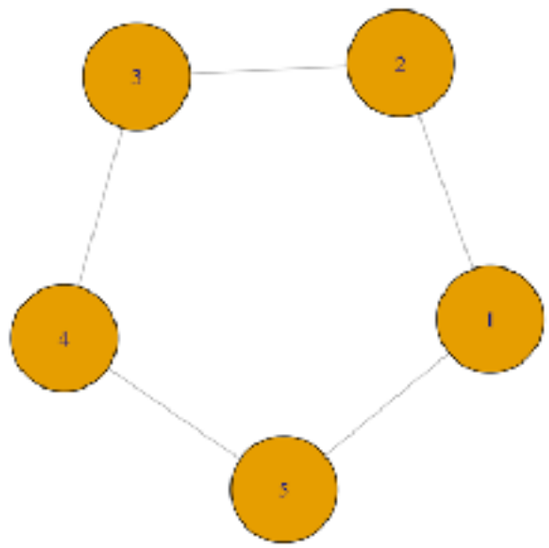
- Genes form networks to regulate different functions in a human body.



- Cancer = Uncontrollable cancer cell growth
- Often this growth is due to a disturbance in a key regulatory gene - the hub of the network.

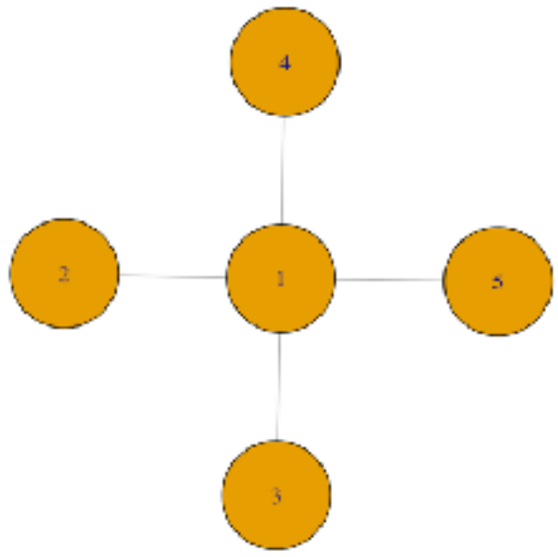
Network Centralisation

- **Node centralisation** defines the most important node in a network.



Centralisation = 0

- **Network centralisation** is the weighted average of node centralisations.



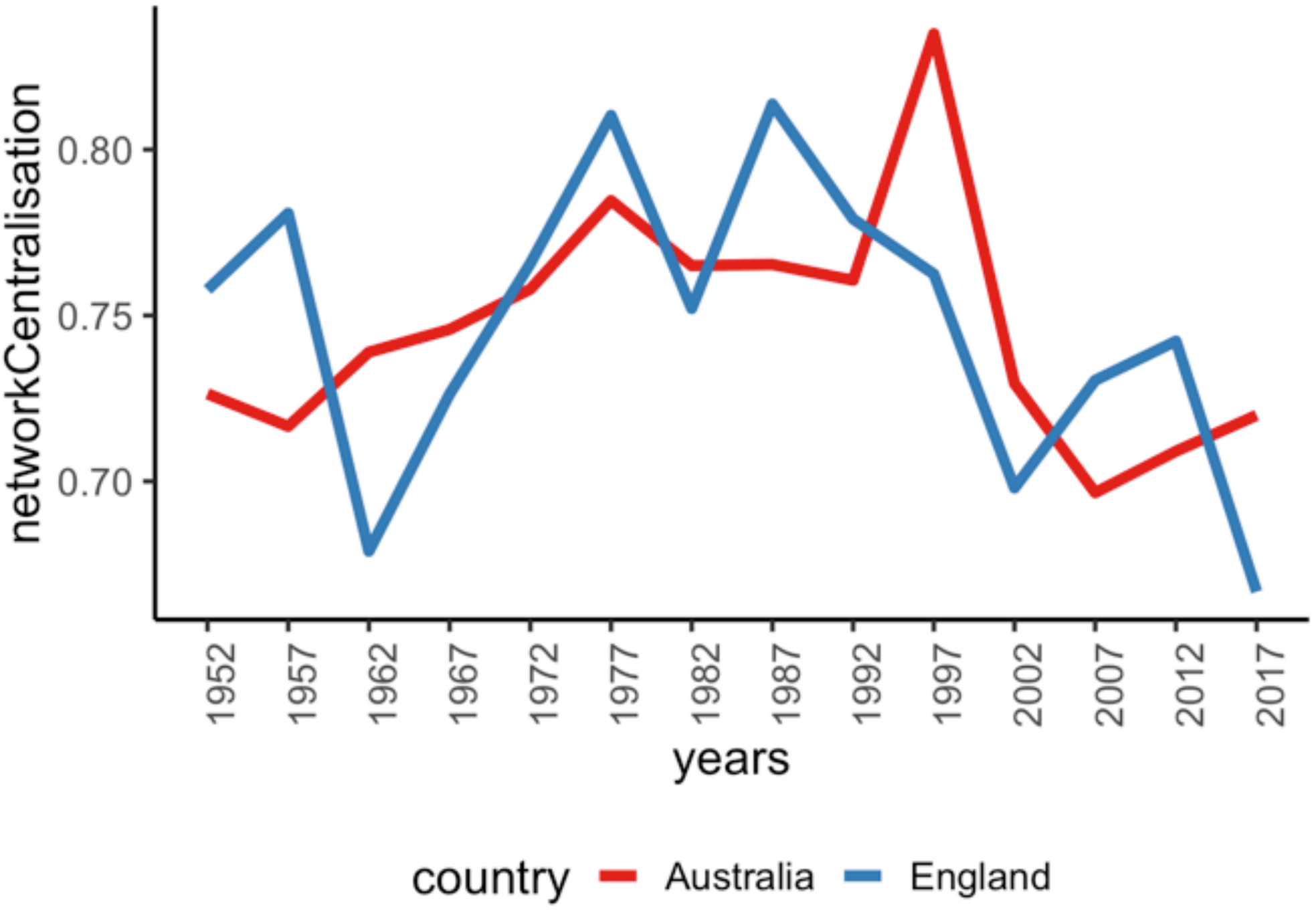
Centralisation = 0.67

Australia Vs England

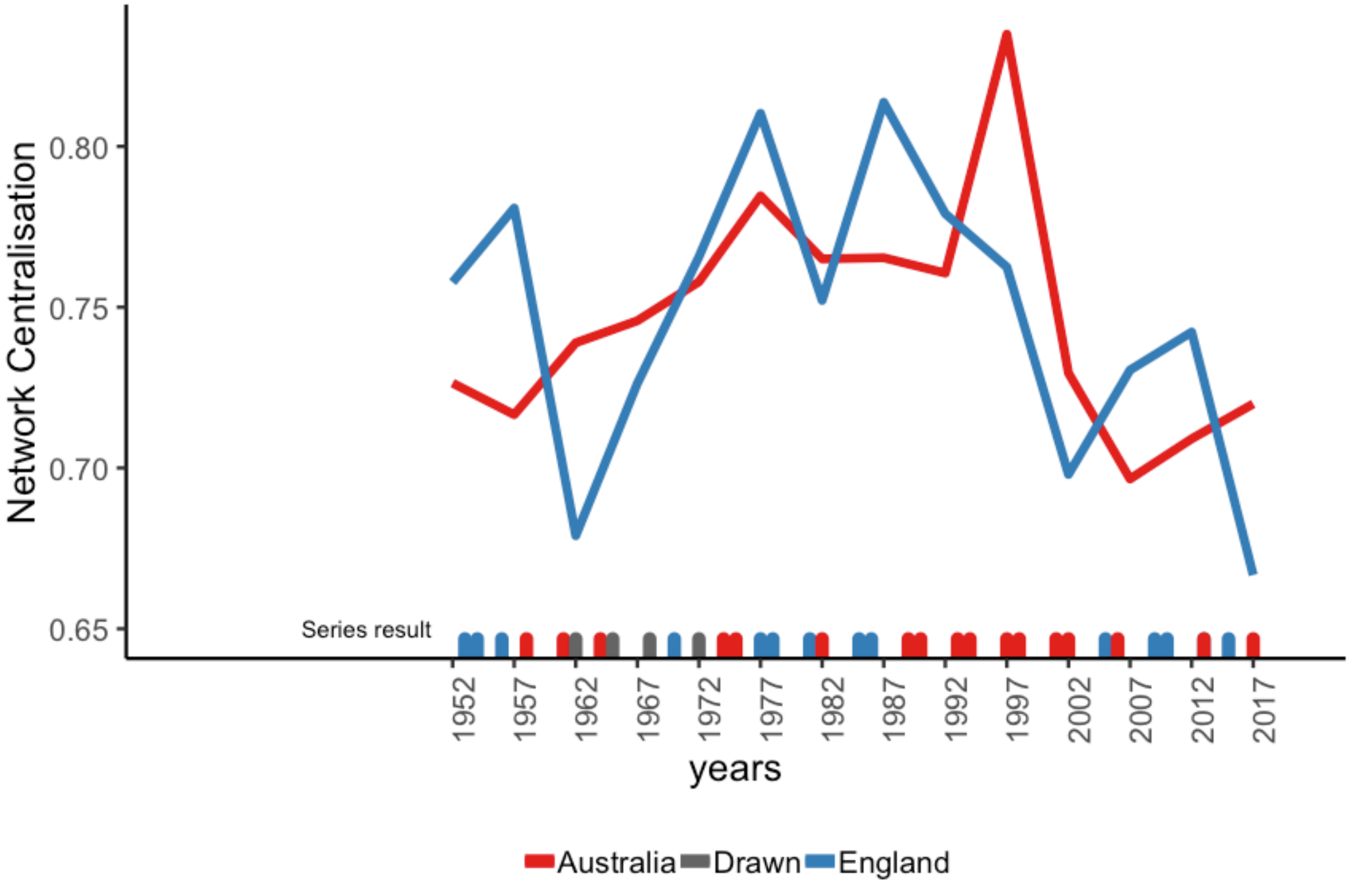
Networks can inform us of the batsmen lineage and team performance

https://gauss17gon.shinyapps.io/partnershipShiny/

Network Centralisation Of Australia And England



Network Centralisation Of Australia And England



ARE YOU A BATSMAN OR A BOWLER?

CLASSIFICATION, PREDICTION AND INTERPRETING ERRORS

Classification

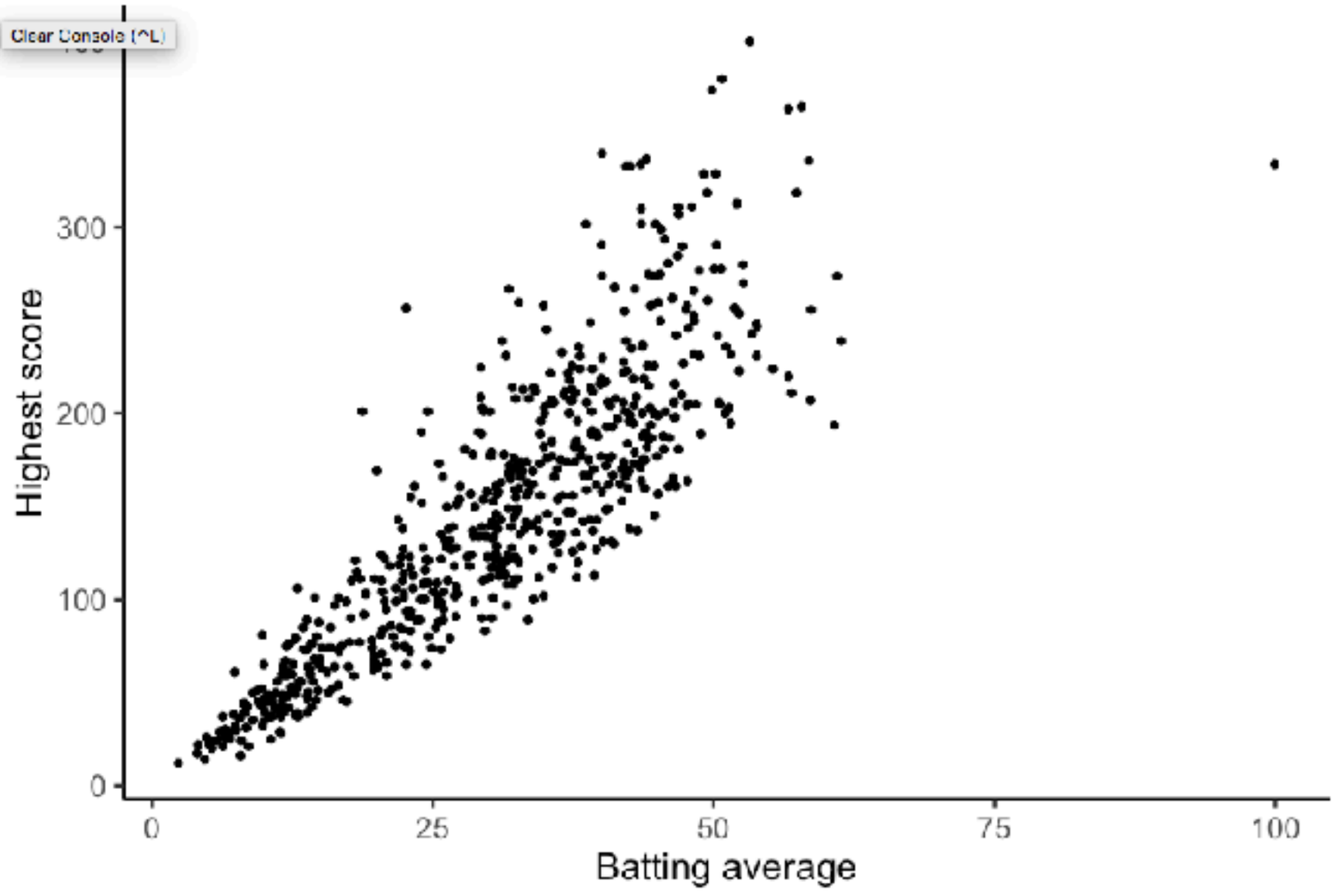
Suppose we want to build a mathematical relationship between:

- $y_i \in \{0,1\}$, denoting a player as a batsman or not a batsman.
- $X = (\boldsymbol{x_1}, \ldots, \boldsymbol{x_p}) \in \mathbb{R}^{n \times p}$, which is a data matrix, with row i denoting a cricket player and column j denoting a variable, e.g. batting average

This is called a **classification problem** and not at all trivial!

Let's Take A Step Back First

- What if we want to predict the **highest number of runs** by **batting average**?



Linear Regression

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

- β_0 and β_1 are both real numbers, to be estimated from the data
- ϵ_i are the random errors

Matrix Notation

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

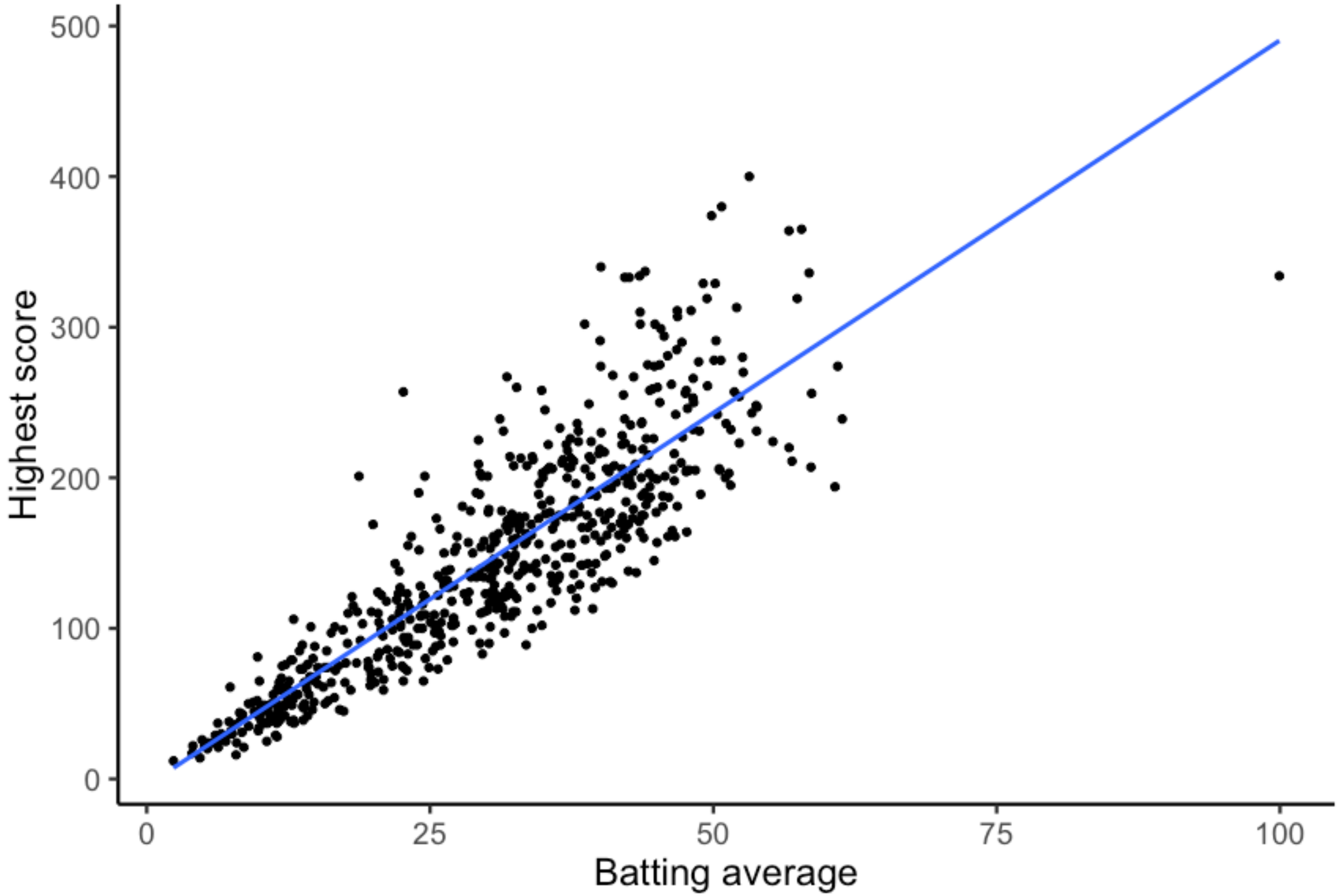
- $\boldsymbol{y}$, $\boldsymbol{\beta}$ and $\boldsymbol{\epsilon}$ are all vectors
- $X = (\mathbf{1}, x)$ is the data matrix with all entries in the the first column equal to 1

Minimisation Of Errors

- $||y - X\beta||^2$ calculates the sum of squared errors.
- We want to minimise this term with respect to β.
- The solution is:

$$\hat{\beta} = (X^\top X)^{-1} X^\top y$$

Linear Regression Line



Logistics Regression

- When a player is a batsman (1) or not (0), we don't have continuous response variable.
- Solution: transformations:

1. We model on the **probability** of a player being a batsman, $p_i \in [0,1]$.
2. We transform the probabilities so they can be modelled over $\mathbb{R}$:

$$\log\left(\frac{p}{1-p}\right) = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Interpretations

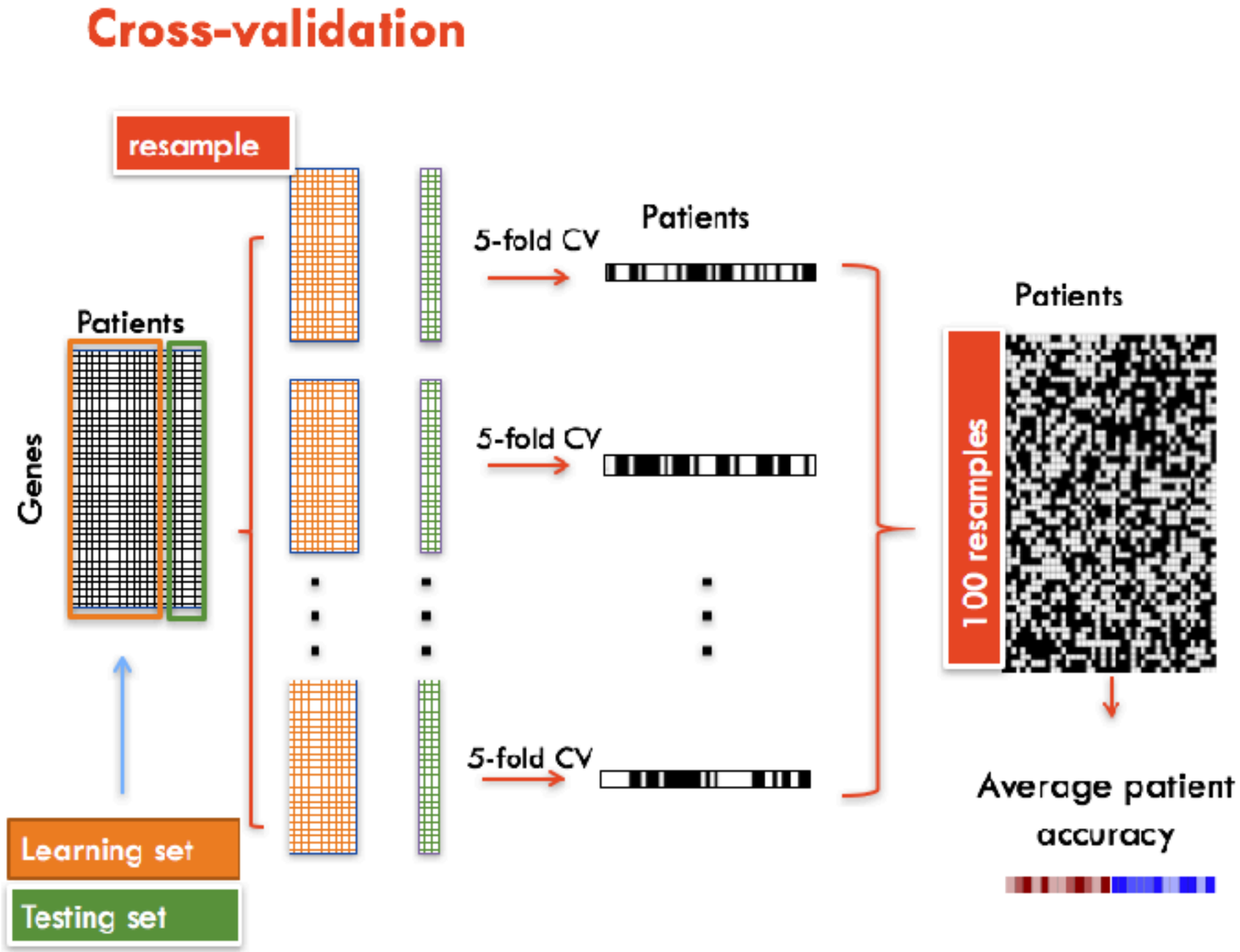
- β tells us about the strength and effect of each variable

	term	estimate	std.error	statistic	p.value
1	(Intercept)	-1.7	0.39	-4.4	0.000013
2	notOuts	-0.082	0.017	-4.7	0.0000022
3	battingAvg	0.11	0.021	5.1	2.7e-7
4	highestInningsScoreNum	0.0021	0.0034	0.62	0.54
5	ducksScored	-0.021	0.03	-0.68	0.5

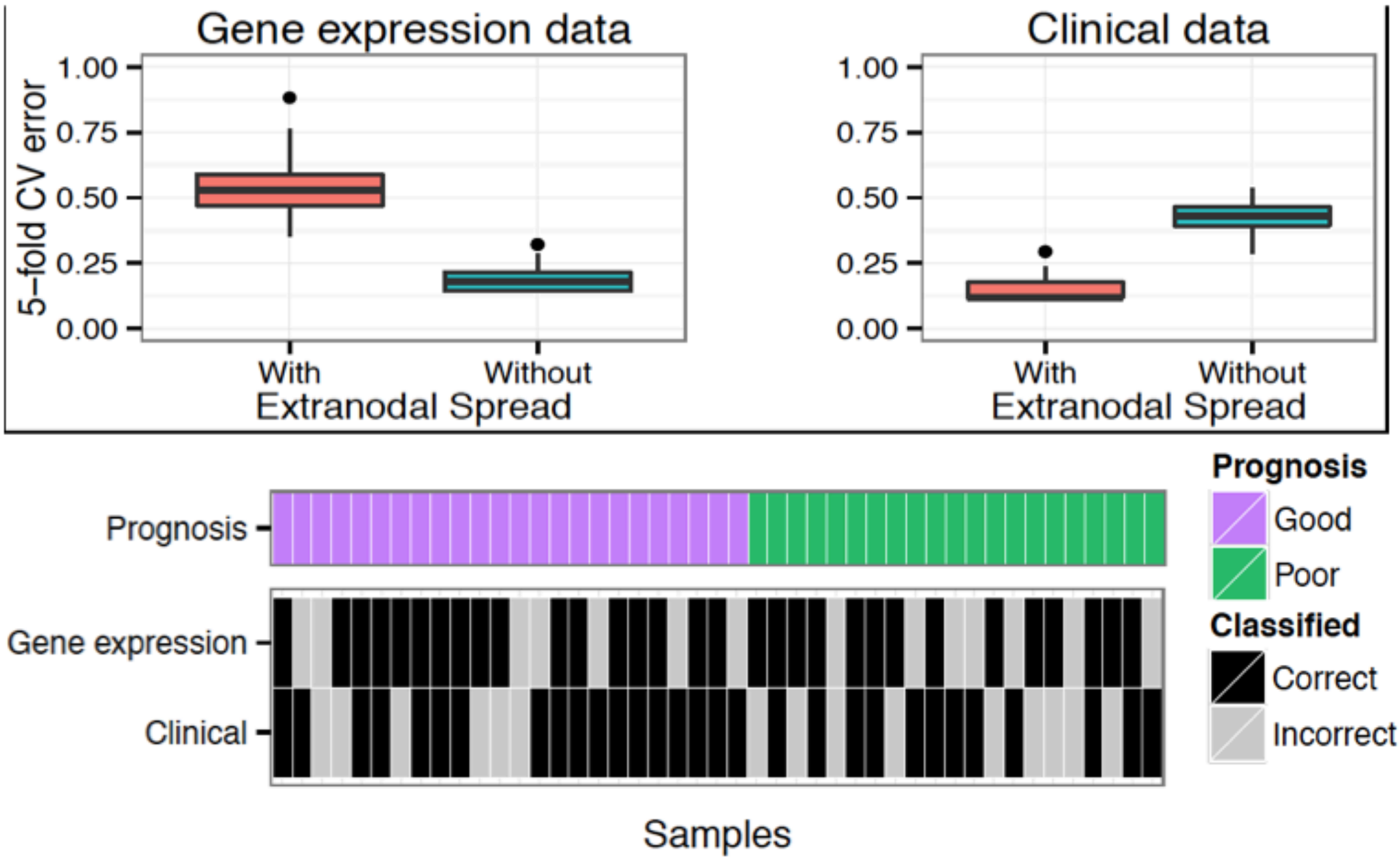
- p tells us about the probability of each observation being a 1 or 0

	player	notOuts	battingAvg	highestInningsScoreNum	ducksScored	isBowler	isBatsman	probBatsman	isAllrounder
1	MJ Clarke	22	49.10	329	9	Not bowler	1	0.901950262	FALSE
2	DG Bradman	10	99.94	334	7	Not bowler	1	0.999827751	FALSE
3	SK Warne	17	17.32	99	34	bowler	0	0.147570883	FALSE
4	GD McGrath	51	7.36	61	35	bowler	0	0.003311592	FALSE
5	GS Sobers	21	57.78	365	12	bowler	0	0.962297895	TRUE

Connection With Bioinformatics



Interpretation Of Errors



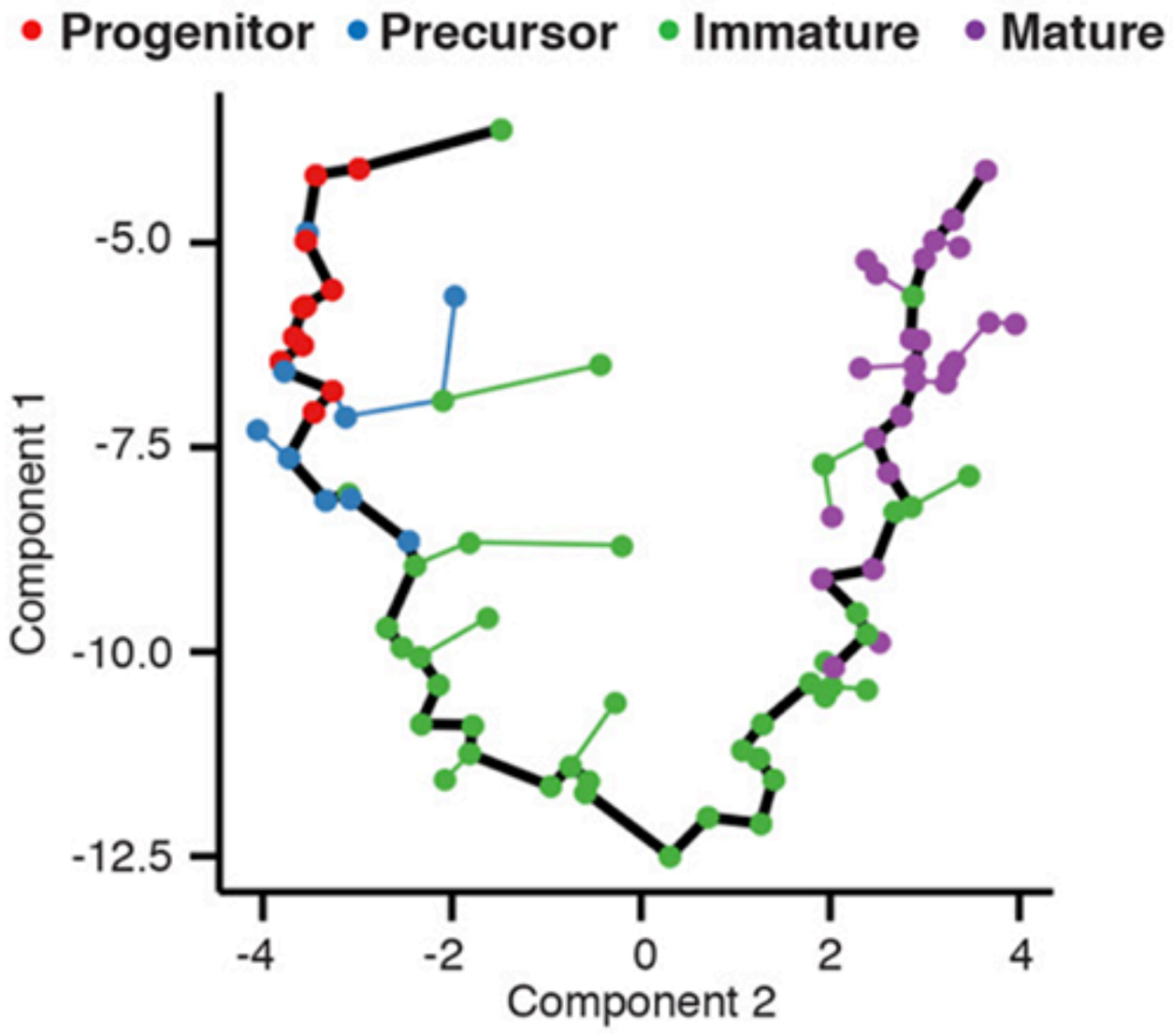
WILL YOU BECOME A GREAT CRICKET PLAYER?

TRAJECTORY ANALYSIS

Trajectory Analysis

- Biotechnology has improved to such a point that we can extract and analyse RNA materials within individual cells.

- One of the hottest question is how cells develop and knows their fate? What role does each gene play in this process?



(Way Too Complicated) Mathematics Of Cell Trajectories

- The idea is to find:

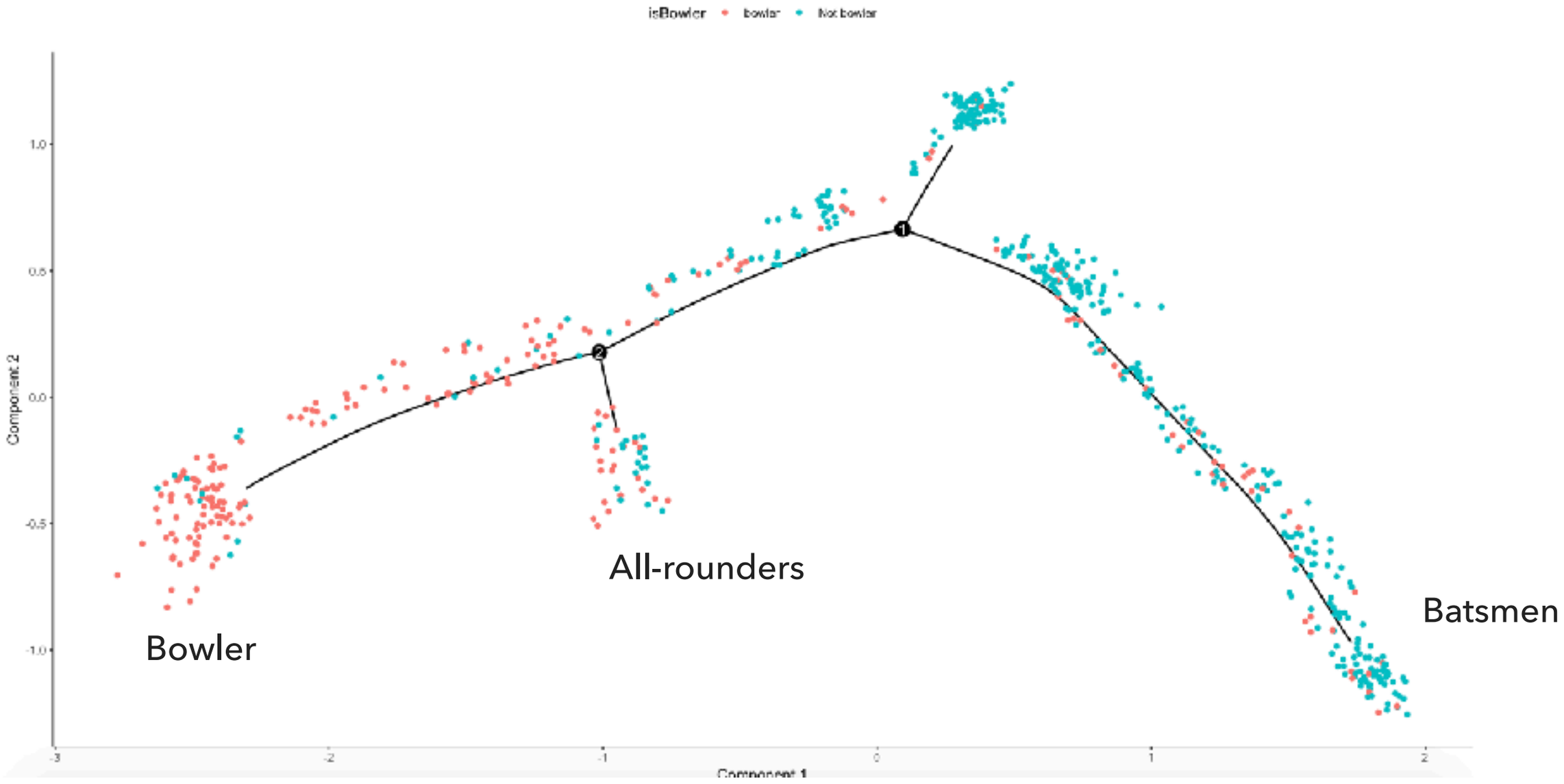
  1. a tree network $\mathcal{G}$
  2. a low-dimension representation Z of the original data X
  3. a function $f_{\mathcal{G}}$ that maps Z to X

s.t. we can preserve the similarities between individual cells in the original data.

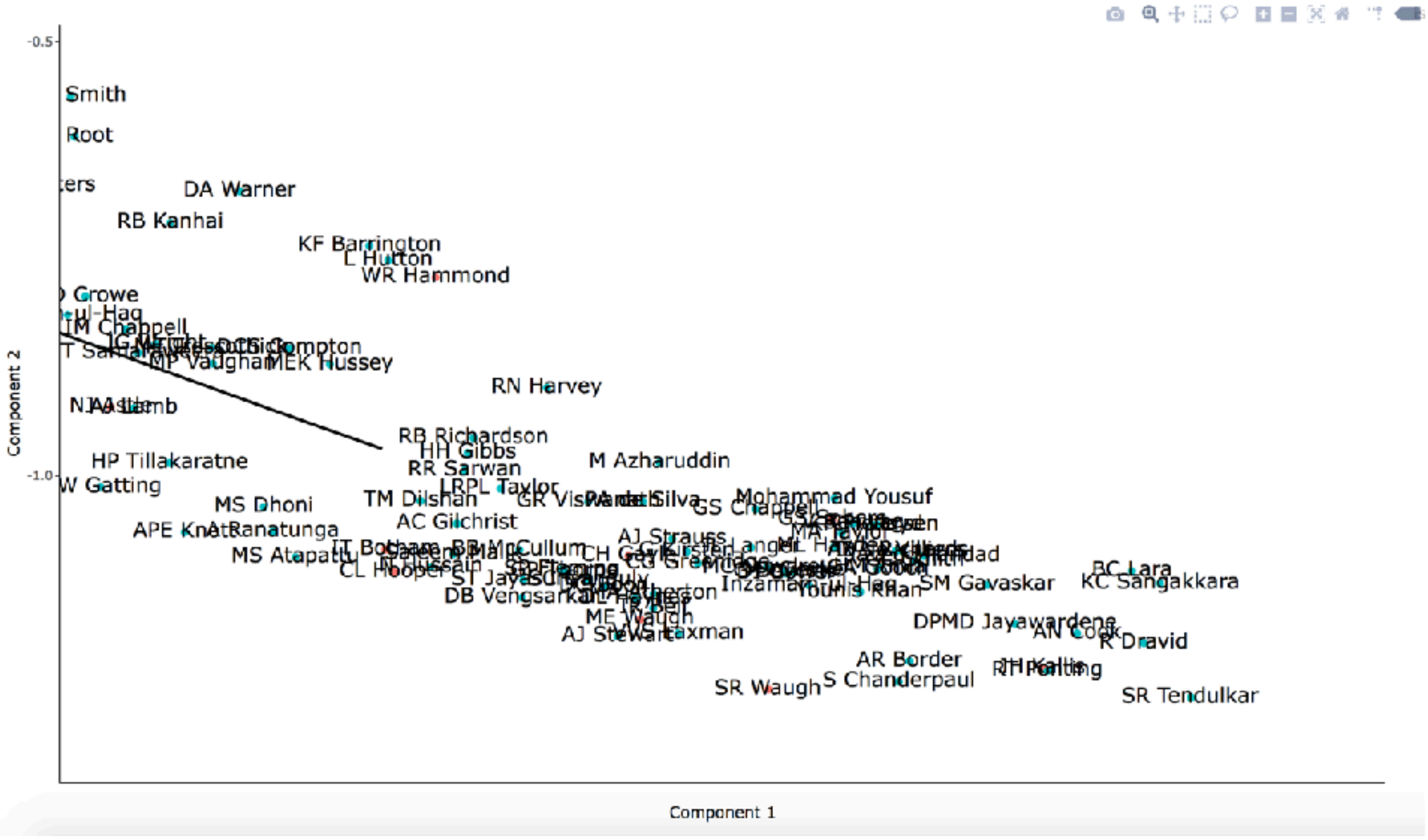
$$\min_{\mathcal{G}} \min_{f_{\mathcal{G}}} \min_{Z} \sum_{(V_i,V_j)\in E} w_{i,j} \| f_{\mathcal{G}}(z_i) - f_{\mathcal{G}}(z_j) \|^2$$

Who Cares About Maths? Show Me Pretty Pictures!

- We can think of each batsman as a cell with potential to develop further into their careers.



Who Cares About Maths? Show Me Pretty Pictures!



Final Words

- The biggest different between a mathematician and a statistician is the level of abstraction and removal of context.
- Even though these tools are very simple, but the data + context + programming make things harder.
- There is nothing stopping you from applying a tool, and it is actually quite fun to do so! But it has to make sense.
- Come and talk to me or the Sydney Bioinformatics Group to learn more!