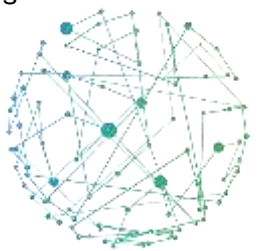
HM Government

A Guide to Data Analytic Techniques and their Use in Fighting Fraud and Improper Payments

Richard Sangster

Advantages of using data

- Can be used to prevent or detect fraud
- Examine more cases (whole populations)
- Prevent human harm from fraud
- Identify Hidden Patterns / Connections
- Deal with different types of fraud taxonomies
- Connect to other data sets for broader insight
- Recovery of Assets from Fraud
- Identify Control Weaknesses for Correction
- Faster checks and information corroboration
- Fill in gaps (fraud by omission)
- Improved user experience
- Others



- Keeps pace with increasingly data driven economies
- Reduces siloed working (fraud occurs in gaps)
- Repeatable methodologies (efficiency)

Background

Fraud is a massive problem, especially for the public sector.

- Our functions and structures make us attractive targets.
- We don’t have the freedom a private enterprise does to act.
- Legacy systems are not helpful.

Extraordinary spending

Diverse functions

Huge customer base

Limited Actions

Your fraud problem

What is the aim here:

- Measuring likelihood of fraud & error?
- Getting more information on your applicants?
- Identifying, reclaiming, and preventing fraud and error?

A project:

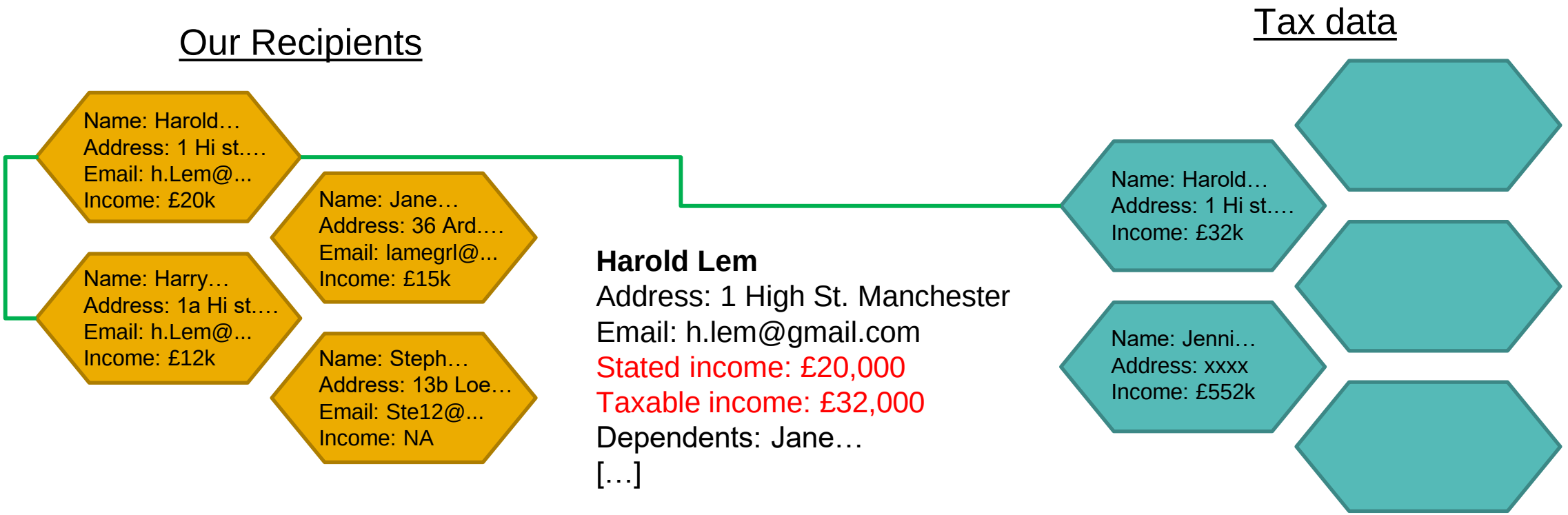
Identifying fraud risks and how you can spot instances.

Number crunching

Working with fraud investigators to confirm findings and reclaim/prevent fraud & error.

Entity Resolution

Also called data matching, this is the use of algorithms to compare records within or across datasets in order to identify when multiple records refer to the same entity.



Entity resolution – example process

Raw Data

RecID	Surname	GivenName	Street	Suburb	Postcode	State	DateOfBirth
a1	Smith	John	42 Miller St	O'Connor	2602	A.C.T.	12-11-1970
a2	Neighan	Joanne	Brown Pl	Dickson	2604	ACT	8 Jan 1968
a3	Meyer	Marie	3/12-14 Hope Cnr	SYDNEY	2050	NSW	01-01-1921
a4	Smithers	Lyn	Browne St	DIXON	2012	N.S.W.	13/07/1970
a5	Nguyen	Ling	1 Milli Rd	Nrth Sydeny	2022	NSW	10/08/1968
a6	Faulkner	Christine	13 John St	Glebe	2037	NSW	02/23/1981
a7	Sandy	Robert	RMB 55/326 West St	Stuart Park	2713	NSW	7/10/1970

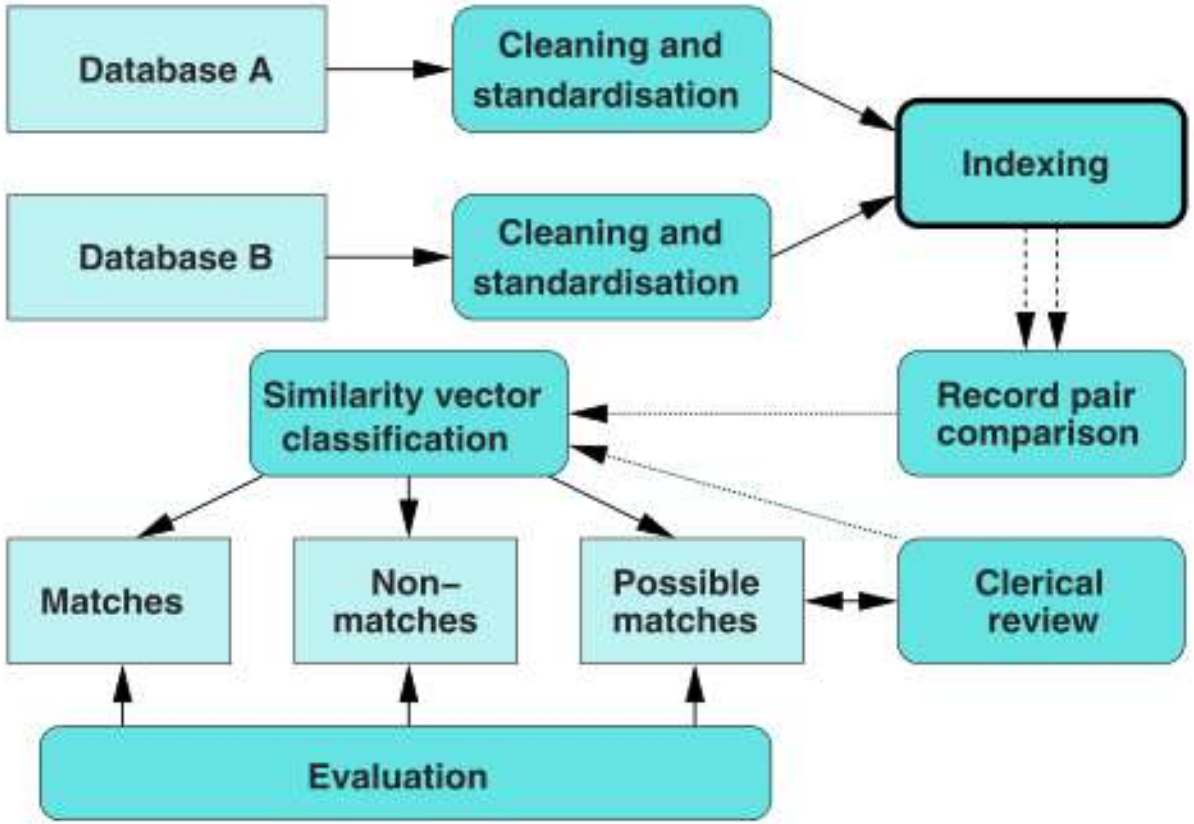
Cleaned and Standardised Data

RecID	GivenName	Surname	Gender	StrPrefix	StrNum	StrName	StrType	Suburb	Postcode	State	BDay	BMonth	BYear
a1	john	smith	m		42	miller	street	oconnor	2602	act	12	11	1970
a2	joanne	neighan	f			brown	place	dickson	2604	act	8	1	1968
a3	mary	meier	f	3	12-14	hope	corner	sydney	2050	nsw	1	1	1921
a4	lynette	smithers	f			browne	street	dixon	2012	nsw	13	7	1970
a5	ling	nguyen	?		1	milli	road	north sydney	2022	nsw	10	8	1968
a6	christine	faulkner	f		13	john	street	glebe	2037	nsw	23	2	1981
a7	robert	sandy	m	rmb 55	326	west	street	stuart park	2713	nsw	7	10	1970

Real world data is messy and very rarely has unique identifiers you can rely on.

1. Cleaning and standardisation; make sure al the data is recorded in the same way.
2. Indexing; algorithms that filter the comparisons being made.
3. Record pair comparison; detailed examination of records indexing highlighted.
4. Similarity vector classification: final estimation of likelihood two records belong to one entity.

Entity resolution – example process



Real world data is messy and very rarely has unique identifiers you can rely on.

1. Cleaning and standardisation; make sure al the data is recorded in the same way.
2. Indexing; algorithms that filter the comparisons being made.
3. Record pair comparison; detailed examination of records indexing highlighted.
4. Similarity vector classification: final estimation of likelihood two records belong to one entity.

Entity resolution

You need to understand it well, because if you are the one with the fraud risk chances are you will not be the one doing it.

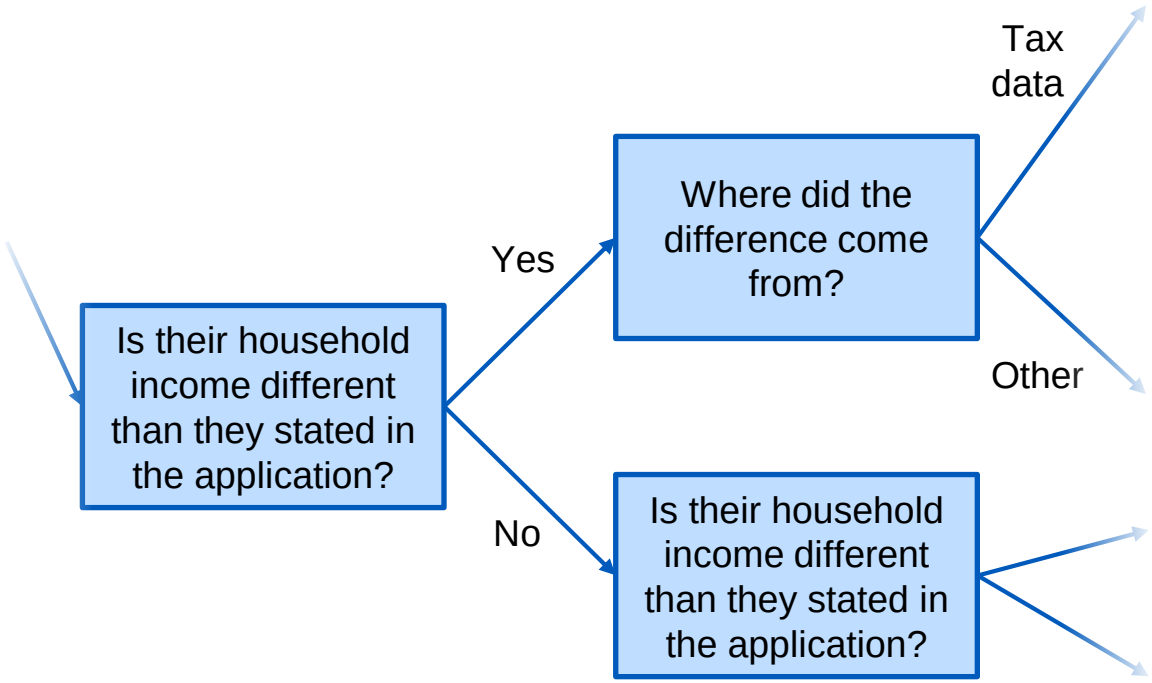
- Organise and clean our own data.
- Bring insights from other datasets into our analysis.
- Indexing algorithms have different strengths and limitations, that you need to consider against your data.
- Doesn't (by itself) let us do anything with these insights at scale.

Fill in gaps (fraud by omission)

Reduces siloed working (fraud occurs in gaps)

Connect to other data sets for broader insight

Rules based analysis – Expert systems



Take the knowledge and experience of your fraud experts, and build an evaluation tool based on their insights.

- Fairly approachable.
- Makes existing case evaluation more scalable.

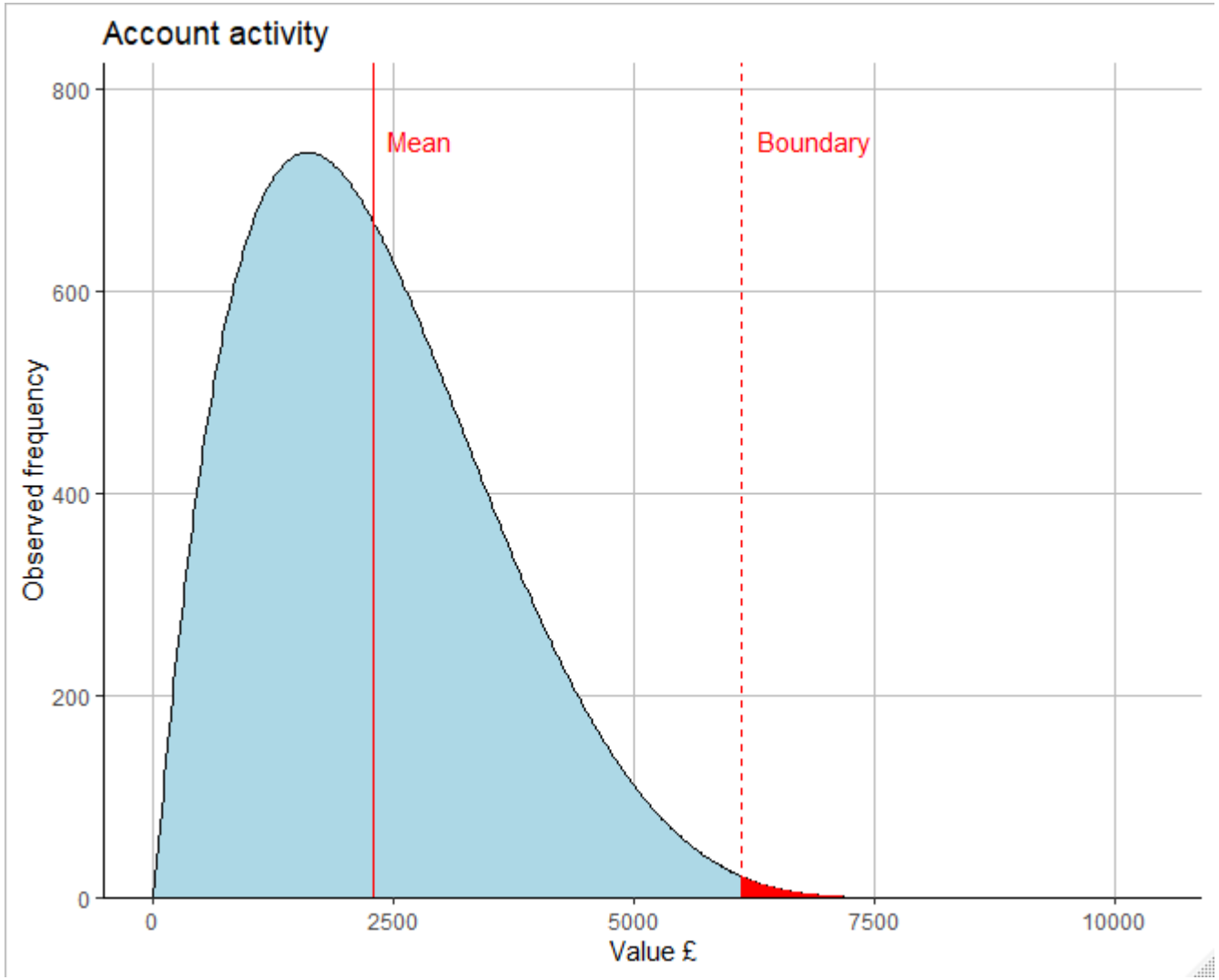
- Needs that expert knowledge.
- ‘Hard’ rules can quickly become outdated.

Stochastic evaluation

Using statistics to evaluate the data and set dynamic rules.

- Evaluate the data to identify outliers more easily.
- Able to fine tune the boundary.
- It reacts to changing behaviour.

In the example, how do we select the accounts for fraud investigators to review?

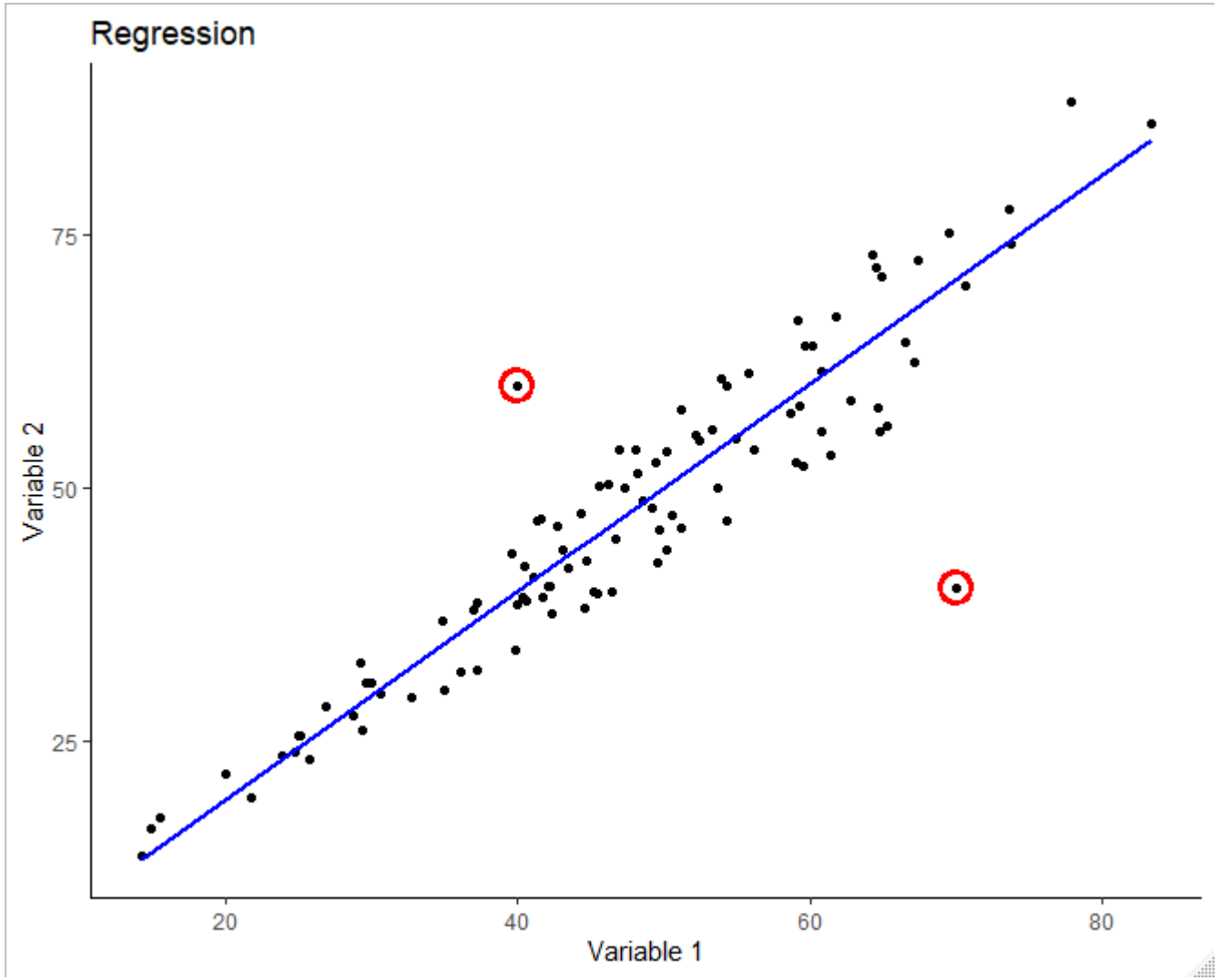


Regression

A process to estimate the relationship between variables. Used a lot for forecasting.

- Explore relationships between variables.
- Identify outliers for review.

- You have to consider the types of regression that are appropriate for your data.
- Regression alone doesn't prove relationships.



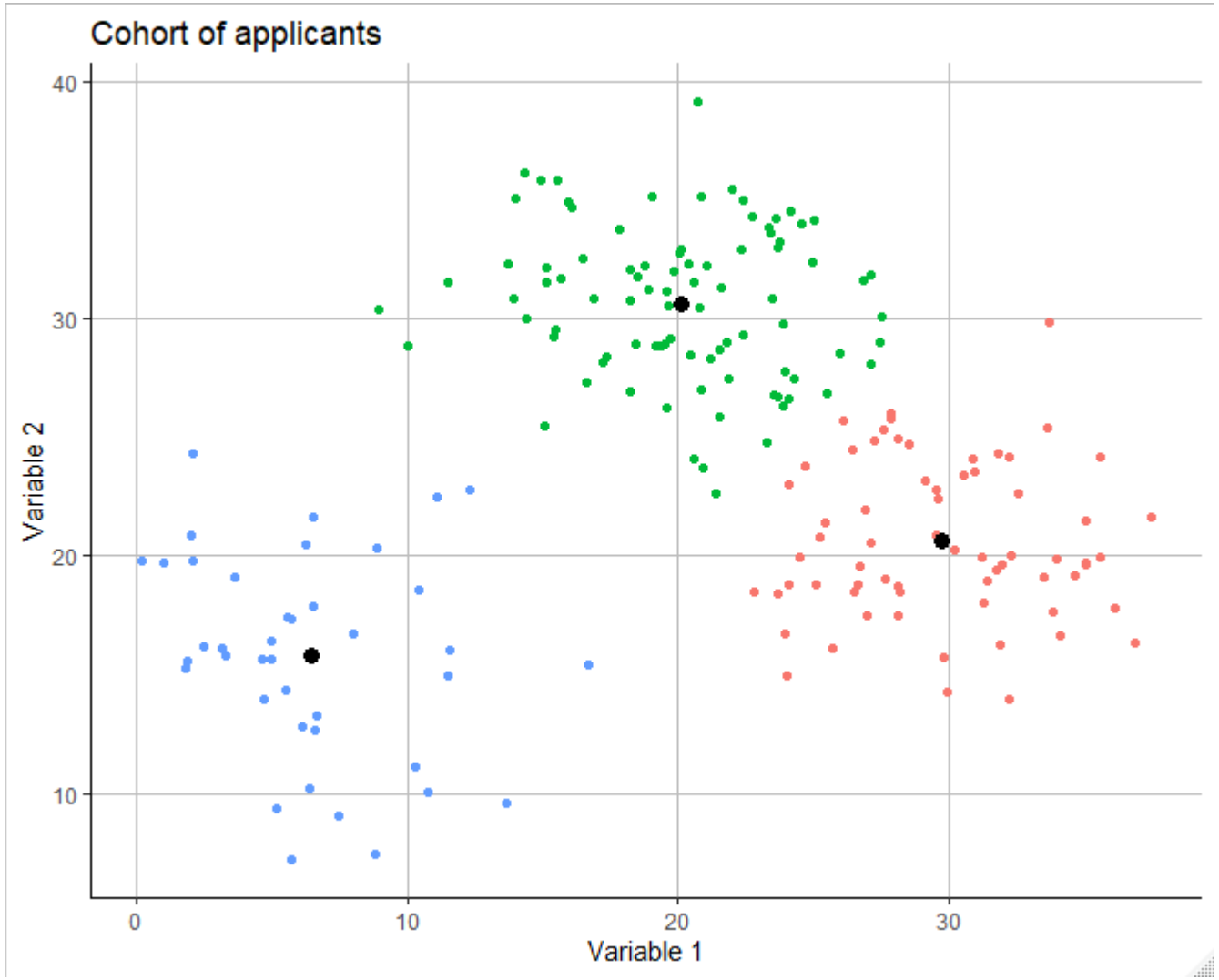
Clustering

Grouping similar entities together by identifying areas of high density in a data space.

- Identify anomalous points.
- Useful for predictive modelling.
- Can be done using a lot of variables.

But how to validate this?

- Using test data.
- Manually reviewing the results.



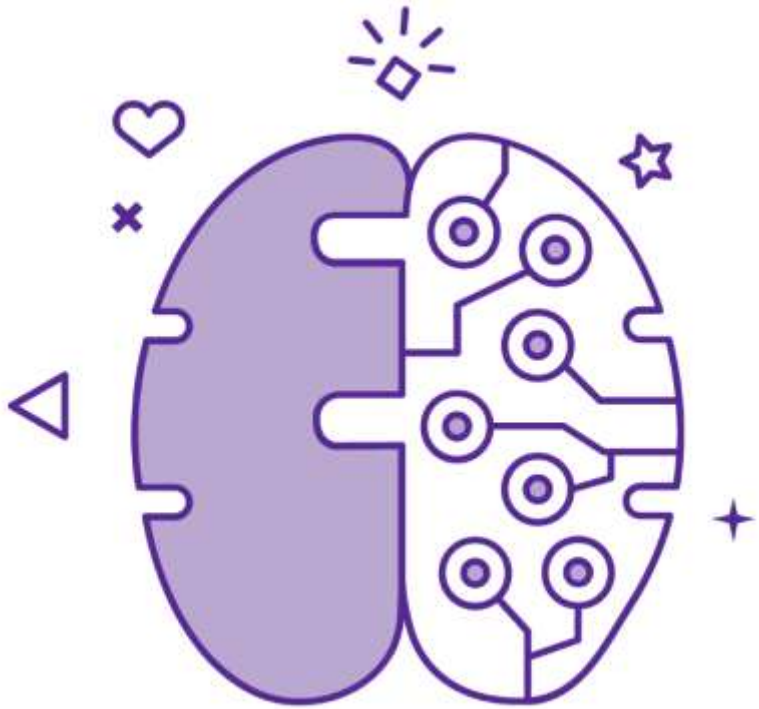
The blue yonder…

Machine learning

At its most basic, this takes a statistical model and designs it to optimise itself. Usually needs training data where you already know which cases are fraud/error, but can be used for data exploration as well.

Deep learning

Sophisticated decision making tools. Very powerful and adaptable, but difficult to build and potentially controversial to use for counter fraud & error in the public sector.



HM Government

Thank you

Richard Sangster
richard.sangster@officeforlifesciences.gov.uk