EARL 2015

EFFECTIVE APPLICATIONS OF THE R LANGUAGE

LONDON 14 - 16 SEPTEMBER

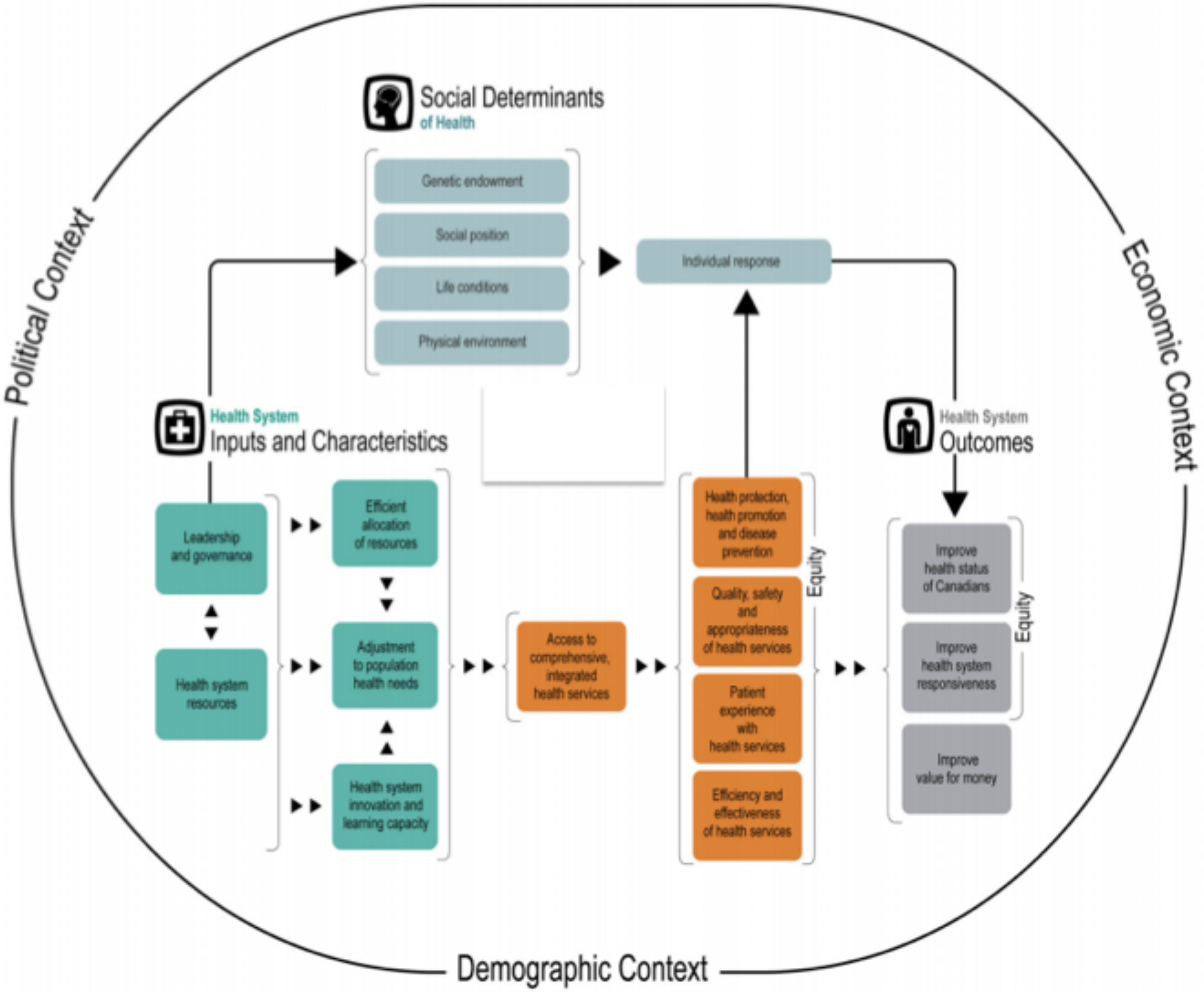
Building Interactive Visualizations with Shiny to Explore Data from Social and Health Care in UK

Armando Vieira

Summary

- The challenge
- The inputs and the outputs
- The predictive model
- Visualizations with Shiny and Google Motion Charts
- A random walk on graphs and causality

Armando Vieira – Data Scientist @dataAI
Armando@dataAI.uk



Armando Vieira – Data Scientist @dataAI
Armando@dataAI.uk

The Inputs

Social – Economics

- ✓ Demographics
- ✓ Economic Deprivation
- ✓ Obesity rates / Sports habits

Health

- ✓ Readmissions
- ✓ Ashma & Diabetes prevalence
- ✓ Cardiovascular diseases

Surveys

- ✓ ASCOF satisfaction
- ✓ Waiting times & service at PCT

Costs

- ✓ Annual Budget
- ✓ Costs per 1000 patients (PCT)

Armando Vieira – Data Scientist @dataAI
Armando@dataAI.uk

The Outputs

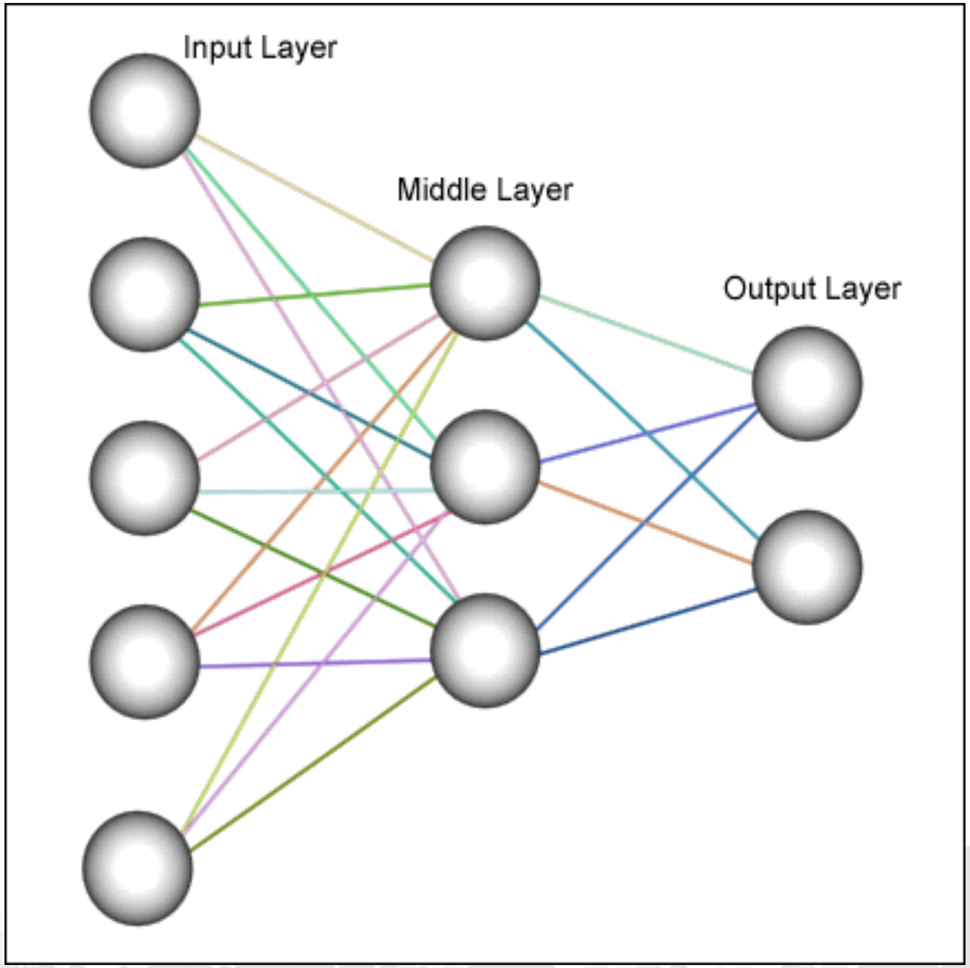
Health Score =
Life Expectancy . (1-Infant Mortality) . Satisfaction PCT Survey

Stress Score =
Growth Elder Pop . Growth Pop . Waiting Times PCT survey

Armando Vieira – Data Scientist @dataAI
Armando@dataAI.uk

Predictive model

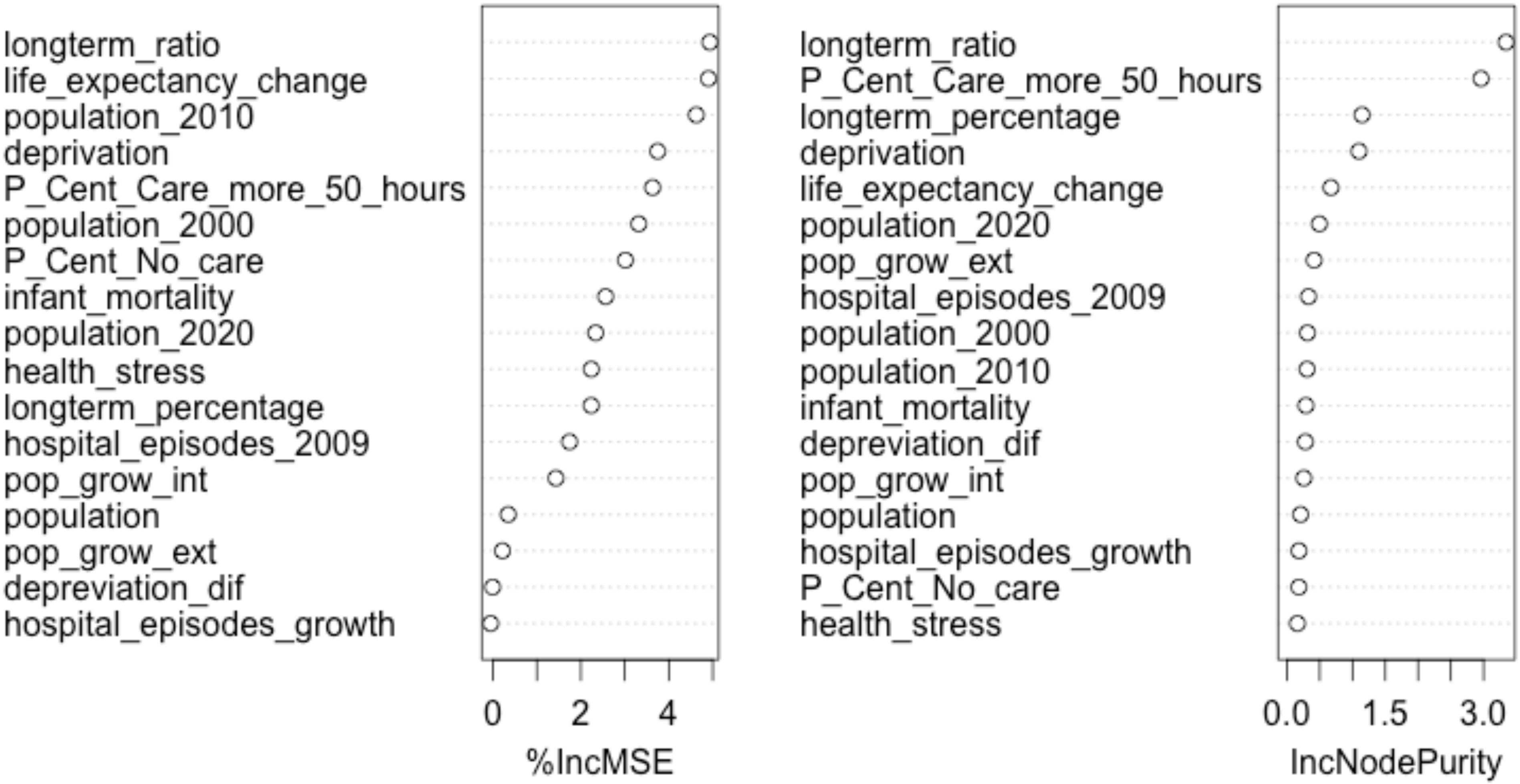
- % no care unpaid
- % > 50 hours unpaid
- Deprivation index
- Deprivation Δ
- Hospital episodes
- Hospital episodes Δ
- Infant mortal
- Life Expectancy Δ
- Long term illness
- Population
- Population internal
- Population external
- Population >65
- Population >65 Δ
- Readmission score
- Sports
- Obesity
- Diabetes



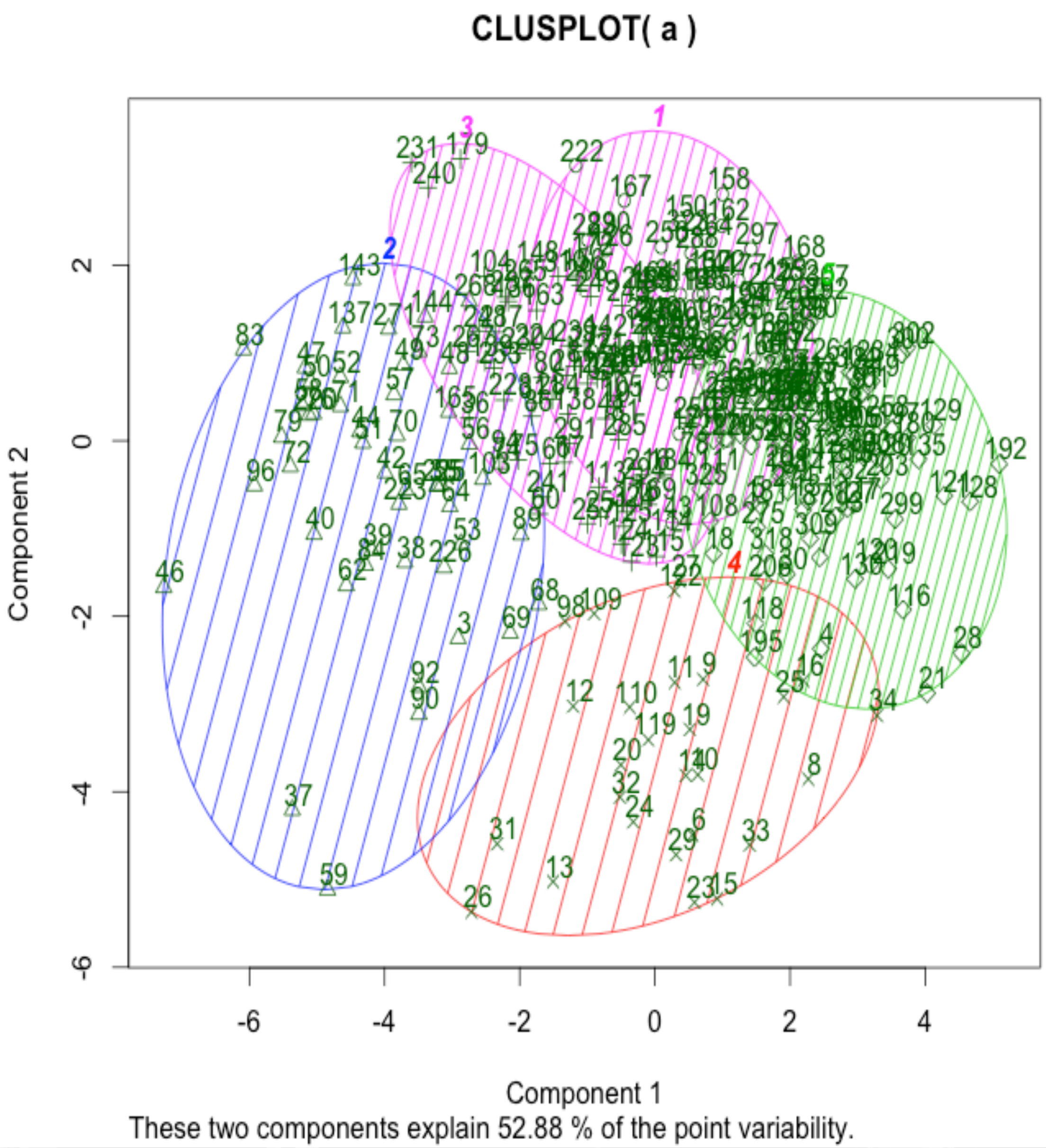
- **Health Score**
- **Stress Score**

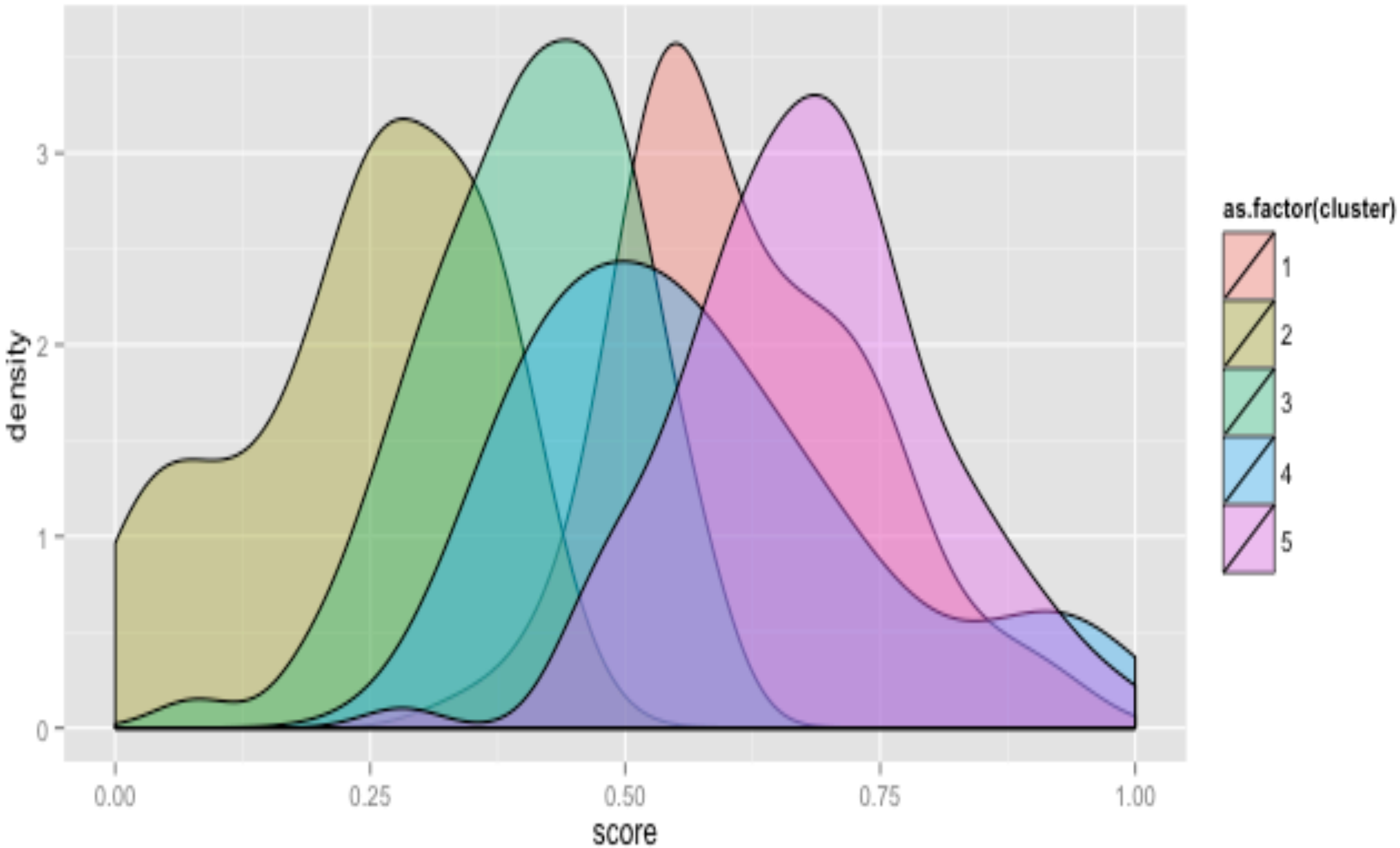
Armando Vieira – Data Scientist @dataAI
Armando@dataAI.uk

Importance in Health Score



Armando Vieira – Data Scientist @dataAI
Armando@dataAI.uk


CLUSPLOT(a)
Component 2
Component 1
These two components explain 52.88 % of the point variability.


as.factor(cluster)
1
2
3
4
5
density
3
2
1
0
0.00
0.25
0.50
0.75
1.00
score

Online demos

1. Google Motion Charts

2. Shiny + Leaflet maps

Armando Vieira – Data Scientist @dataAI
Armando@dataAI.uk

Two districts, two stories

Chiltern

- ✓Higher Health Score
- ✓Stable population
- ✓Healthiest
- ✓High satisfaction score

Liverpool

- ✓Low Health Score
- ✓Highest Hospital Episodes
- ✓Economical deprivated
- ✓High percentage unpaid social care

Armando Vieira – Data Scientist @dataAI
Armando@dataAI.uk

Conclusions I

The **Health Score** is:

- ✓ Higher for less deprivated areas
- ✓ Lower for long term illness

Not related to:

- ✓ Health stress
- ✓ Infant mortality rate
- ✓ % of older population
- ✓ Population size

Armando Vieira – Data Scientist @dataAI
Armando@dataAI.uk

Conclusions II

The **Stress Score** is:

- ✓ Higher for richer districts
- ✓ Higher for regions with large % of population > 65

Not related to:

- ✓ % Long term illness
- ✓ Long term disability rate

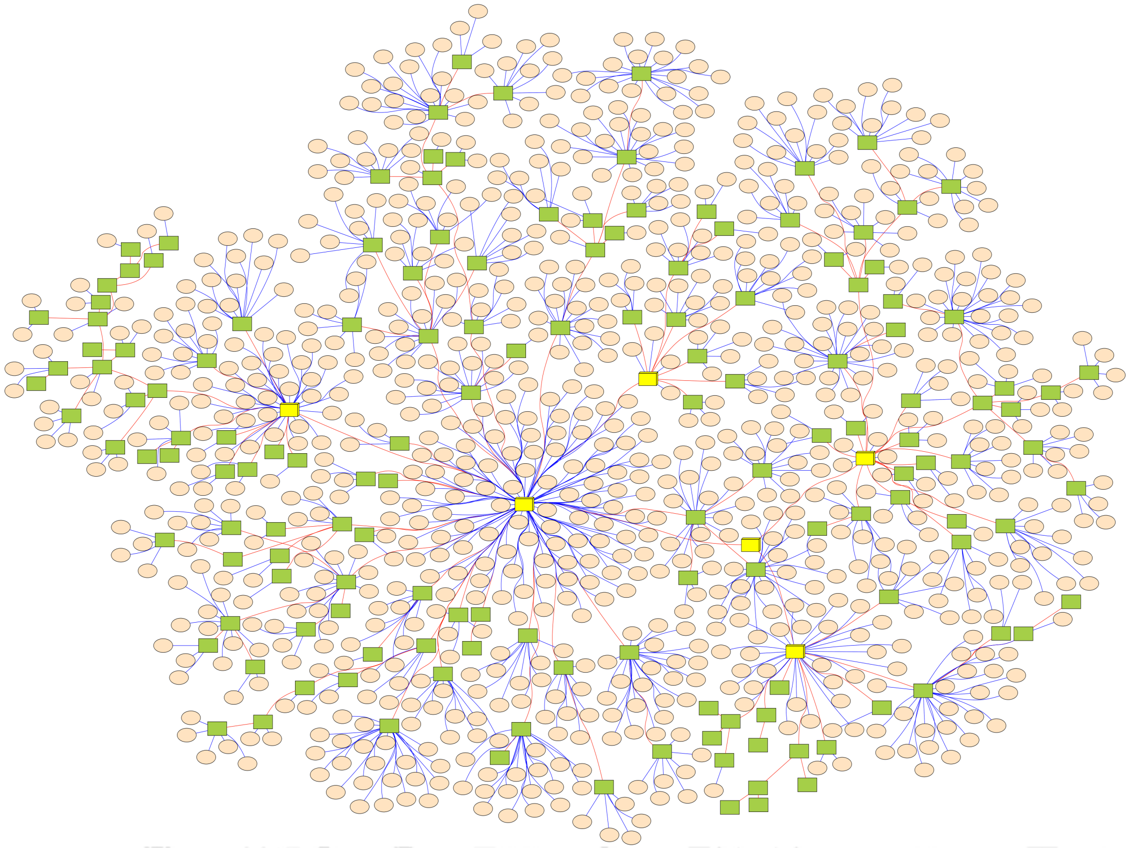
Armando Vieira – Data Scientist @dataAI
Armando@dataAI.uk

Why R?

With R and Shiny we can easily deploy interactive visualizations dashboards for powerful data exploration

Armando Vieira – Data Scientist @dataAI
Armando@dataAI.uk

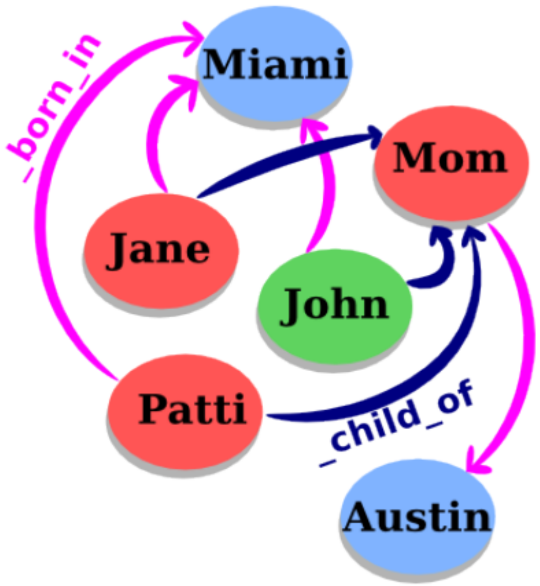
A random walk on graphs and causality



Armando Vieira – Data Scientist @dataAI
Armando@dataAI.uk

Multi-relational data

- Data is structured as a graph
- Each node = an entity
- Each edge = a relation/fact
- A relation = (*sub*, *rel*, *obj*):
    - *sub* =*subject*,
    - *rel* = *relation type*,
    - *obj* = *object*.
- Nodes w/o features.



Armando Vieira – Data Scientist @dataAI
Armando@dataAI.uk

Example

"Who influenced J.K. Rowling?"

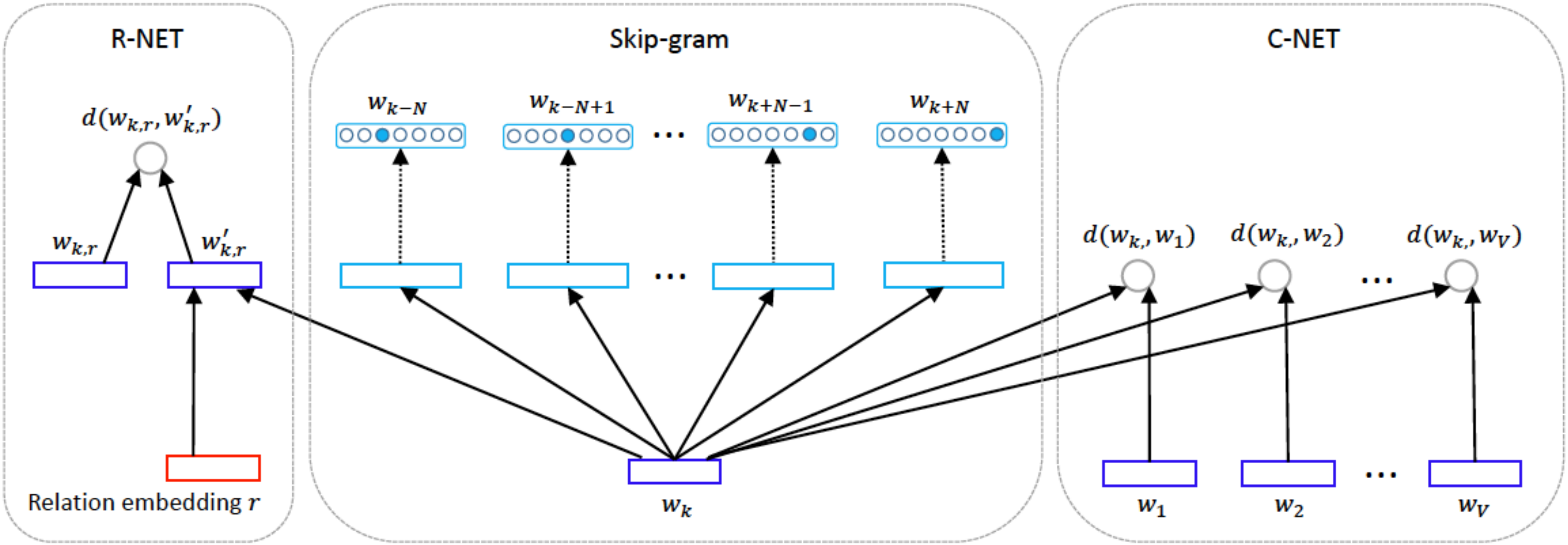
J. K. Rowling `_influenced_by`

- G. K. Chesterton
- J. R. R. Tolkien
- C. S. Lewis
- Lloyd Alexander
- Terry Pratchett
- Roald Dahl
- Jorge Luis Borges
- Stephen King
- Ian Fleming

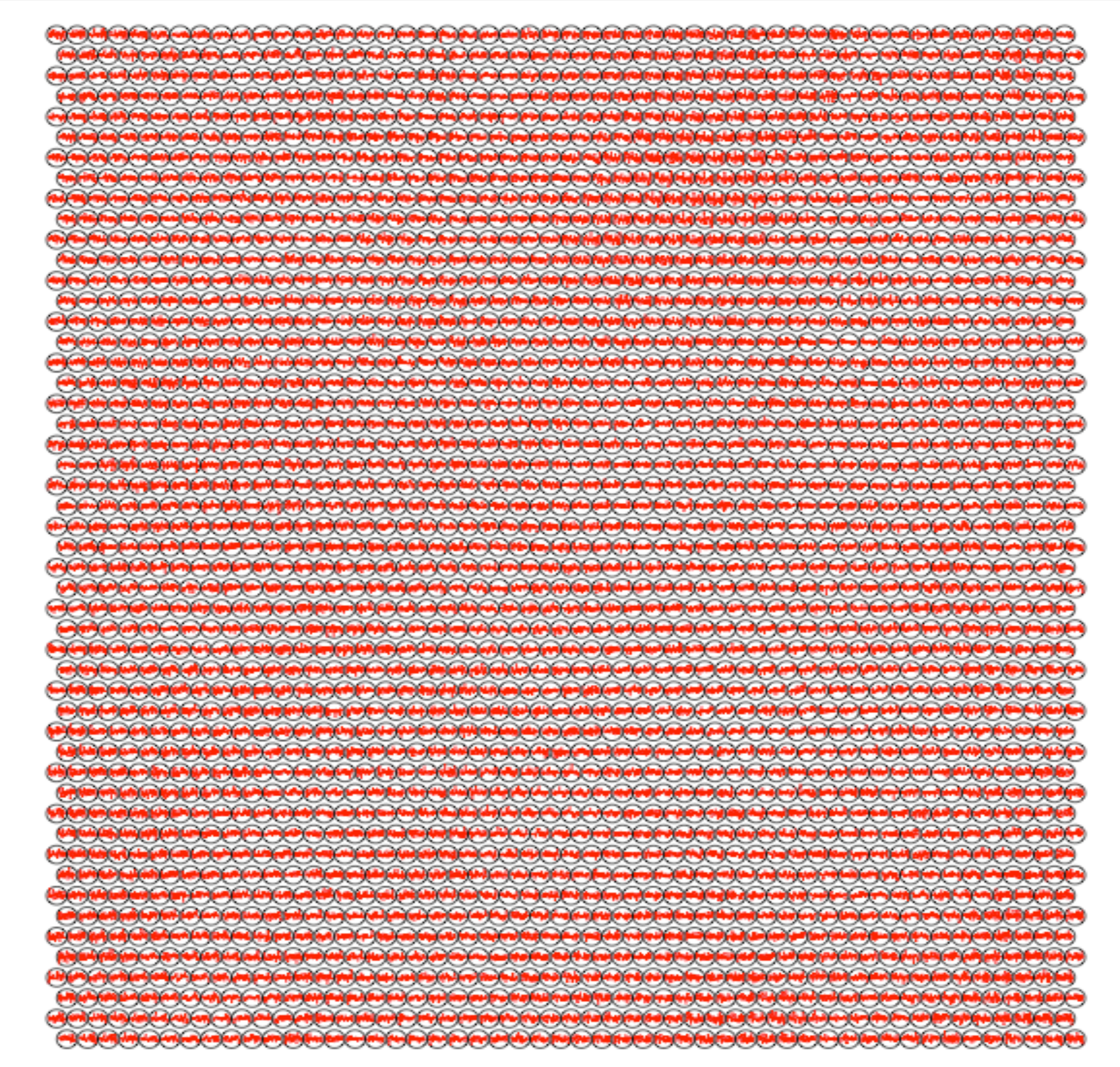
Green=Train Blue=Test Black=Unknown

Armando Vieira – Data Scientist @dataAI
Armando@dataAI.uk

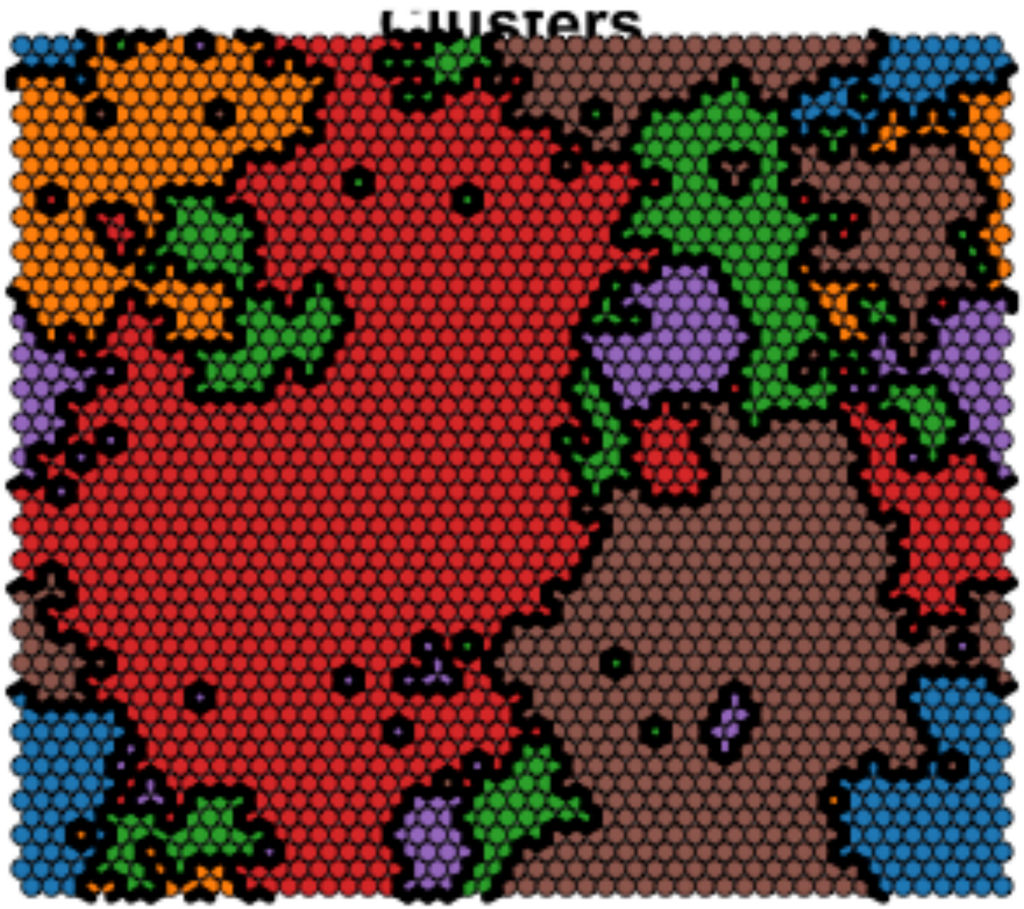
Learning embeddings



Armando Vieira – Data Scientist @dataAI
Armando@dataAI.uk

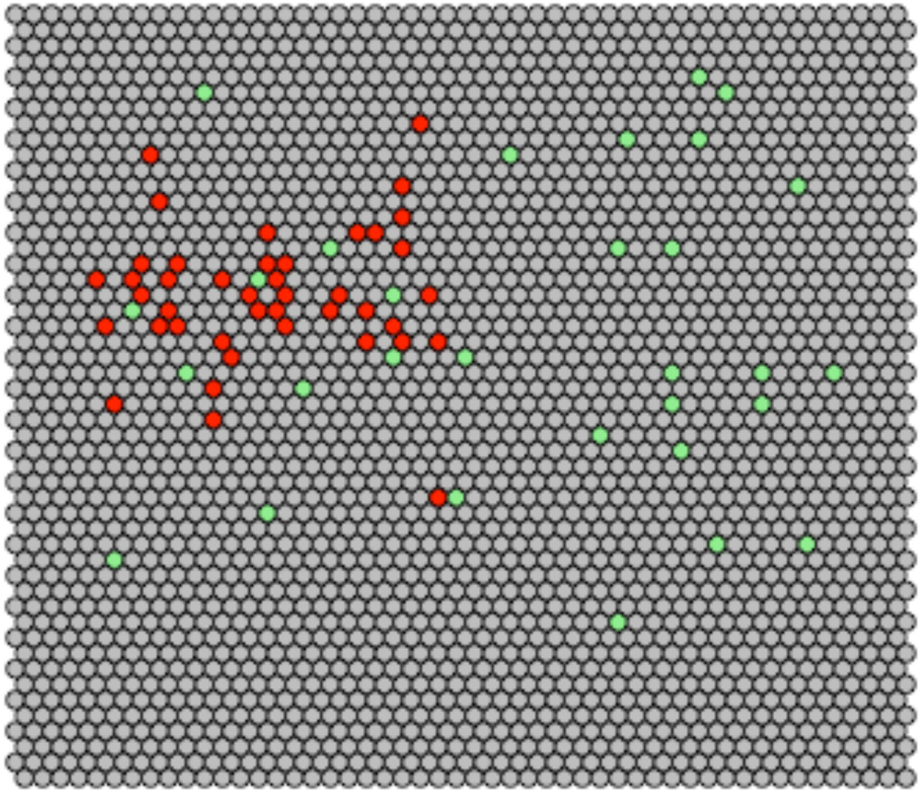


Armando Vieira – Data Scientist @dataAI
Armando@dataAI.uk

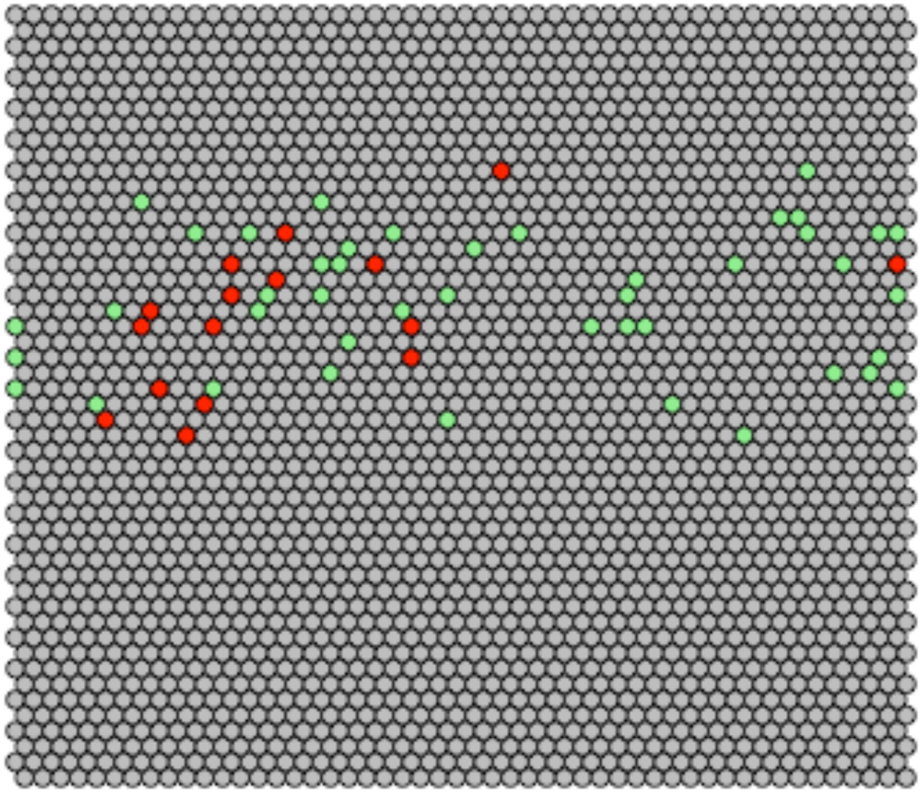

Clusters

Each disease has an unique fingerprint

Lung cancer



Ovary cancer



Armando Vieira – Data Scientist @dataAI
Armando@dataAI.uk

Causality?

- "More police in precincts with higher crime; does that mean that police cause crime?"
  - Policy decision: should we add more police to a given district?

- "Lots of people die in hospitals, are hospitals bad for your health?"
  - Policy decision: should I go to hospital for treatment?

- "Advertise more in December, sell more in December." But what is the causal impact of ad spending on sales?
  - Policy decision: how much should I spend on advertising?

Armando Vieira – Data Scientist @dataAI
Armando@dataAI.uk

counterfactuals, confounding variables

- "If I go to hospital will be better off than I would have been if I didn't go?"
- Sales = f(advertising) + other stuff
  - Xmas a confounding variable

Armando Vieira – Data Scientist @dataAI
Armando@dataAI.uk

Beware of inferences

- The problem with doing inferences on data originated from unknown processes is related to the (implicit) assumption that the system and interactions of variables are in equilibrium.

Armando Vieira – Data Scientist @dataAI
Armando@dataAI.uk

How much we should be worried?

- Economics
  - Experiment to determine policy change for population
  - Impact of treatment on population
  - Selection bias – are random samples really random?

- Business
  - Impact on advertisers who choose to use new feature or service
  - Impact of treatment on those who choose to be treated
  - Not necessarily worried about selection bias (but may be worried about early adopter bias)

Armando Vieira – Data Scientist @dataAI
Armando@dataAI.uk