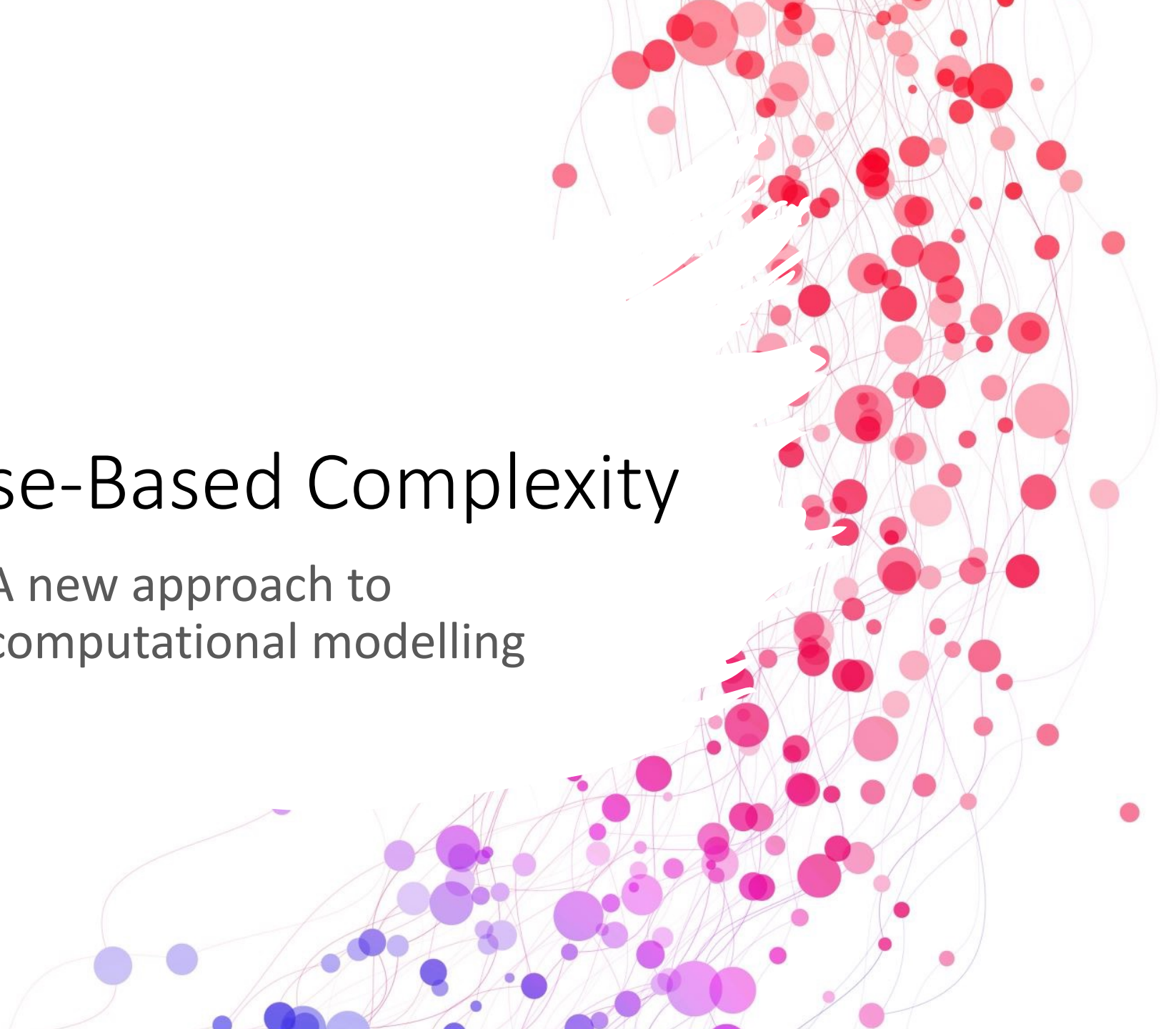
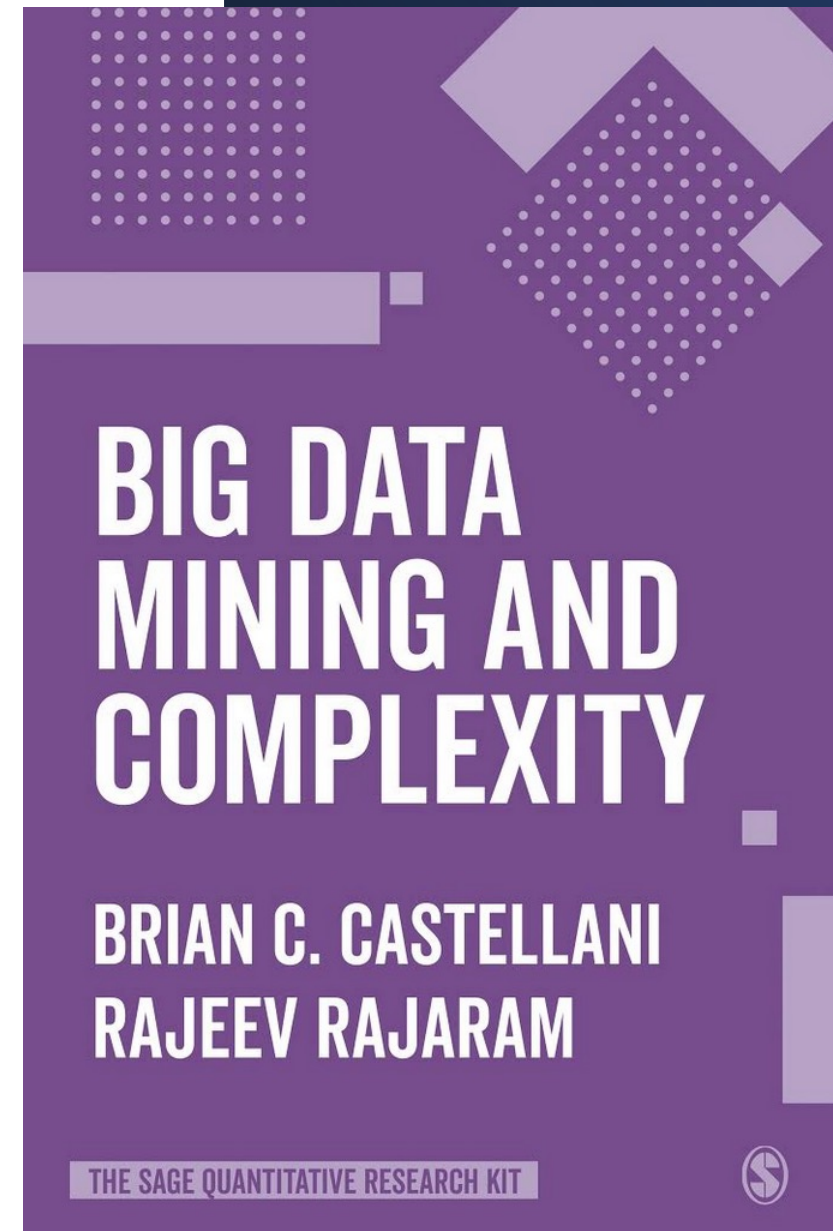


# Case-Based Complexity

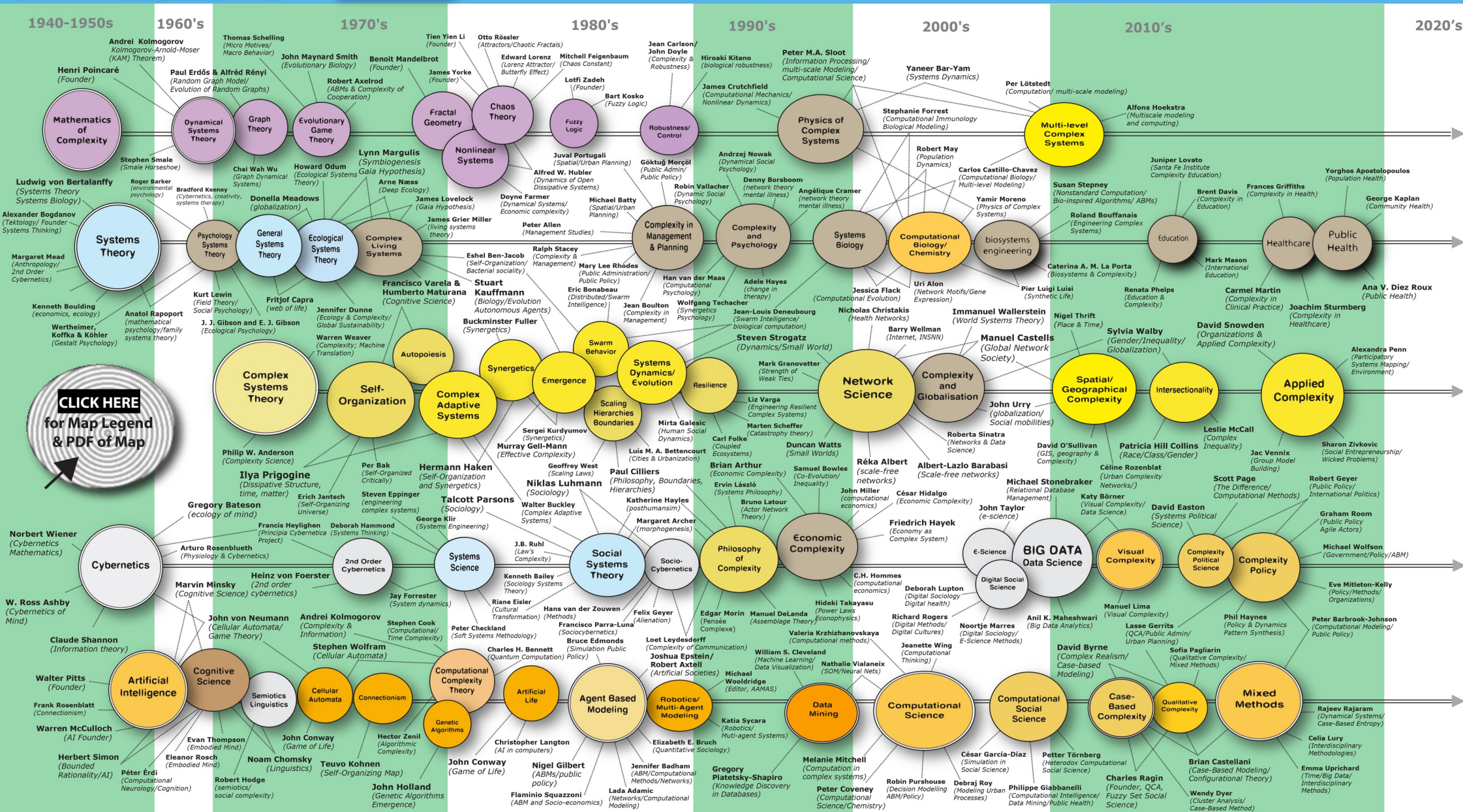
A new approach to  
computational modelling



- **PRIMARY TEXT:** Castellani and Rajaram (2019). *Data Mining Big Data: A Complex and Critical Perspective*. SAGE Quantitative Methods Kit. **READ: Chapters 6 and 7.**
- Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern recognition letters*, 31(8), 651-666.

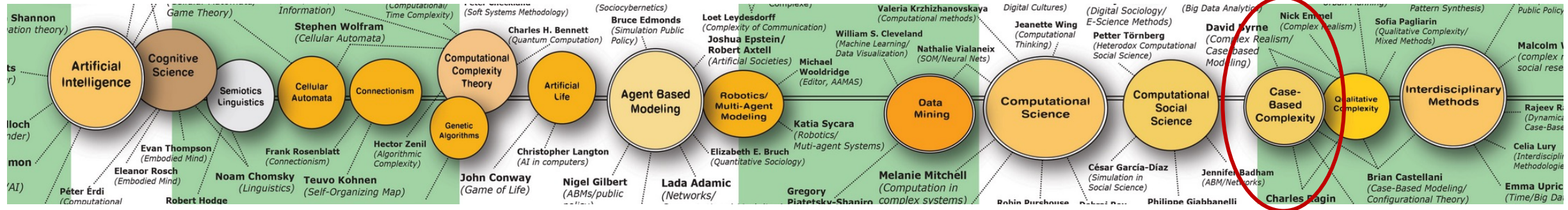








- Within the world(s) of computational modelling and interdisciplinary mixed methods, *case-based complexity* constitutes one of the major methodologies for modelling complex social systems or, more generally, social complexity.



A close-up of the map of the complexity sciences

## Types of Case-Based Complexity

### The SAGE Handbook of Case-Based Methods



Edited by  
David Byrne  
and Charles C. Ragin



SECOND EDITION

### **COMPLEXITY THEORY AND THE SOCIAL SCIENCES**

*The State of the Art*

David Byrne and  
Gillian Callaghan



# Basic Tenets

- 1 The case and its trajectory across time/space are the focus of study, not the individual variables or attributes of which it is comprised.
- 2 Cases and their trajectories are treated as composites (profiles), comprised of an interdependent, interconnected sets of causal conditions, variables, factors or attributes.
- 3 And, finally, cases and their relationships and trajectories are the methodological equivalent of complex systems – that is, they are emergent, self-organising, non-linear, dynamic, network-like and so on – and therefore should be studied as such.



# Case-Based Modeling Additional Tenets

- 1 Cases and their trajectories are dynamically evolving across time/space and, therefore, should be explored to identify their major and minor trends.
- 2 In turn, these trends should be explored in the aggregate for key global-temporal patterns, as in the case of spiralling sources and saddles.
- 3 The social interactions amongst cases are also important, as are the hierarchical social contexts in which these relationships take place.
- 4 And, finally, the complex set of relationships amongst cases is best examined using the tools of network science and simulation.

# What is a case?

- a **case c** is simply an abstract description of the qualitative and quantitative characteristics of some object under study.
- Cases can be individuals in a dataset, nodes in a network, interacting agents in simulation, groups being clustered, organisations, cities, countries and so forth.





# Defining a case mathematically

---

The state of a case  $c_i$  is described by its profile, as measured by one or more variables  $x_{ij}$ , where  $j$  can take on integer values from 1 to  $k$ . In other words, each case has a profile that can be described by the values taken on by the  $k$  variables  $(x_{i1}, x_{i2}, \dots, x_{ik})$ . This arrangement of the variables in the form of a row is called a row vector. Hence,  $c_i = (x_{i1}, x_{i2}, \dots, x_{ik})$  denotes the profile of the  $i$ th case  $c_i$ .

We envision a large database  $D$  consisting of row vectors  $c_i = (x_{i1}, x_{i2}, \dots, x_{ik})$ , where each element  $x_{ij}$  is a measured variable for some **case profile**  $c_i(t)$ , as defined for a particular instant of time  $t$ . Suppressing the dependence on time  $t$ , one can represent such a database  $D$  in matrix form as shown below:

$$D = \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix} = \begin{bmatrix} x_{11} & \dots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{nk} \end{bmatrix}. \quad (6.1)$$

<div> <div></div> <div>Area Code</div> </div>	Income	Employment	Health						
	People in income deprivation (%)	Working-age people in employment deprivation (%)	GP-recorded chronic condition (rate per 100) ⚠	Limiting long-term illness (rate per 100)	Premature death (rate per 100,000)	GP-recorded mental health condition (rate per 100) ⚠	Cancer incidence (rate per 100,000)	Low birth weight (live single births less than 2.5kg) (%)	Children aged 4-5 who are obese (%) ⚠
	16	10	14.3	22.7	382.4	23.2	611.9	5.5	11.8
Isle of Anglesey	Vector								
Gwynedd									
Conwy	15	10	12.9	20.6	375.6	23.7	593.6	5.1	11.4
Denbighshire	17	11	14.7	21.8	397.4	28.4	639.3	6.1	12.2
Flintshire	12	8	14.1	19.7	358.1	23.1	647.8	5.4	11.2
Wrexham	15	9	14.3	21.5	393.7	24.3	637.3	6.4	12.4
Powys	11	7	12.8	18.8	309.1	19.0	579.8	4.7	10.5
Ceredigion	12	8	12.7	20.0	322.4	19.9	545.5	4.8	10.5
Pembrokeshire	15	10	13.1	20.5	345.8	22.1	606.1	5.2	12.5
Carmarthenshire	15	11	13.9	23.7	365.5	20.0	602.6	5.4	12.8

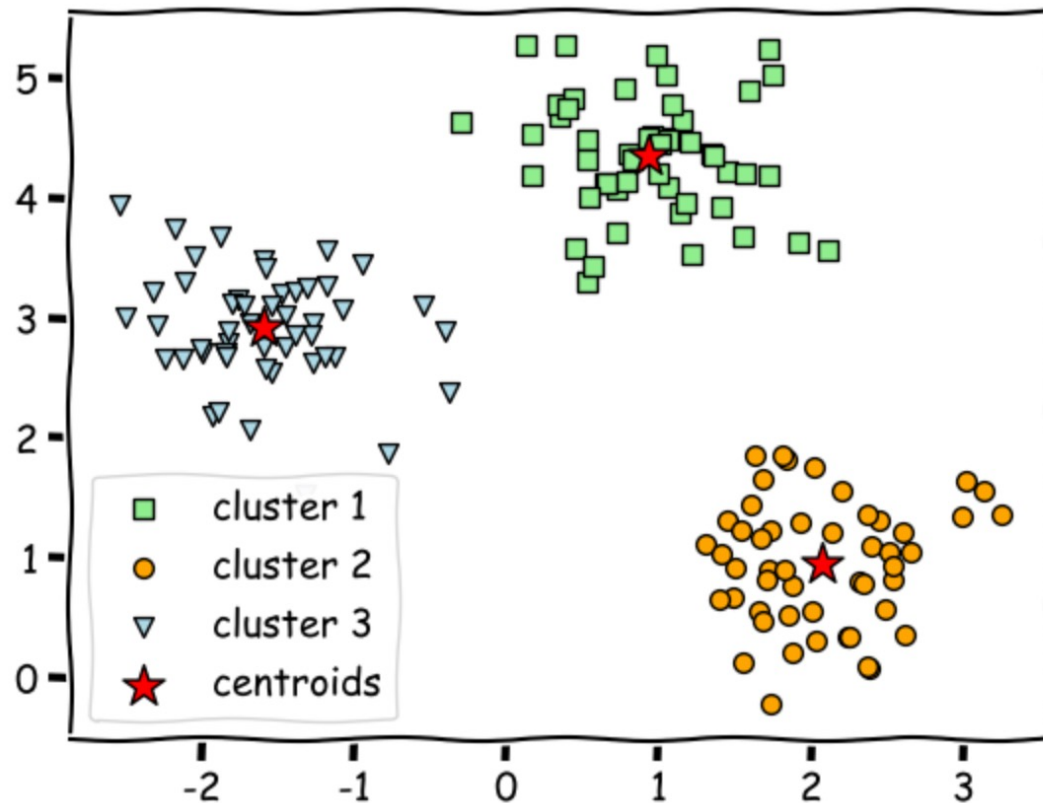
<https://statswales.gov.wales/Catalogue/Community-Safety-and-Social-Inclusion/Welsh-Index-of-Multiple-Deprivation/WIMD-Indicator-data-2019/indicatordata-by-localauthority>

# Classification and Clustering

- **Clustering and Classification** are mathematical techniques that let us group together cases that have similar profiles – as well as position them away from groups of cases with different profiles.
- While both approaches often use the same mathematical algorithms, **the former uses a training set** of known case-based clusters to arrive at its results, while **the latter is largely exploratory**, seeking to identify groupings that may or may not be known.



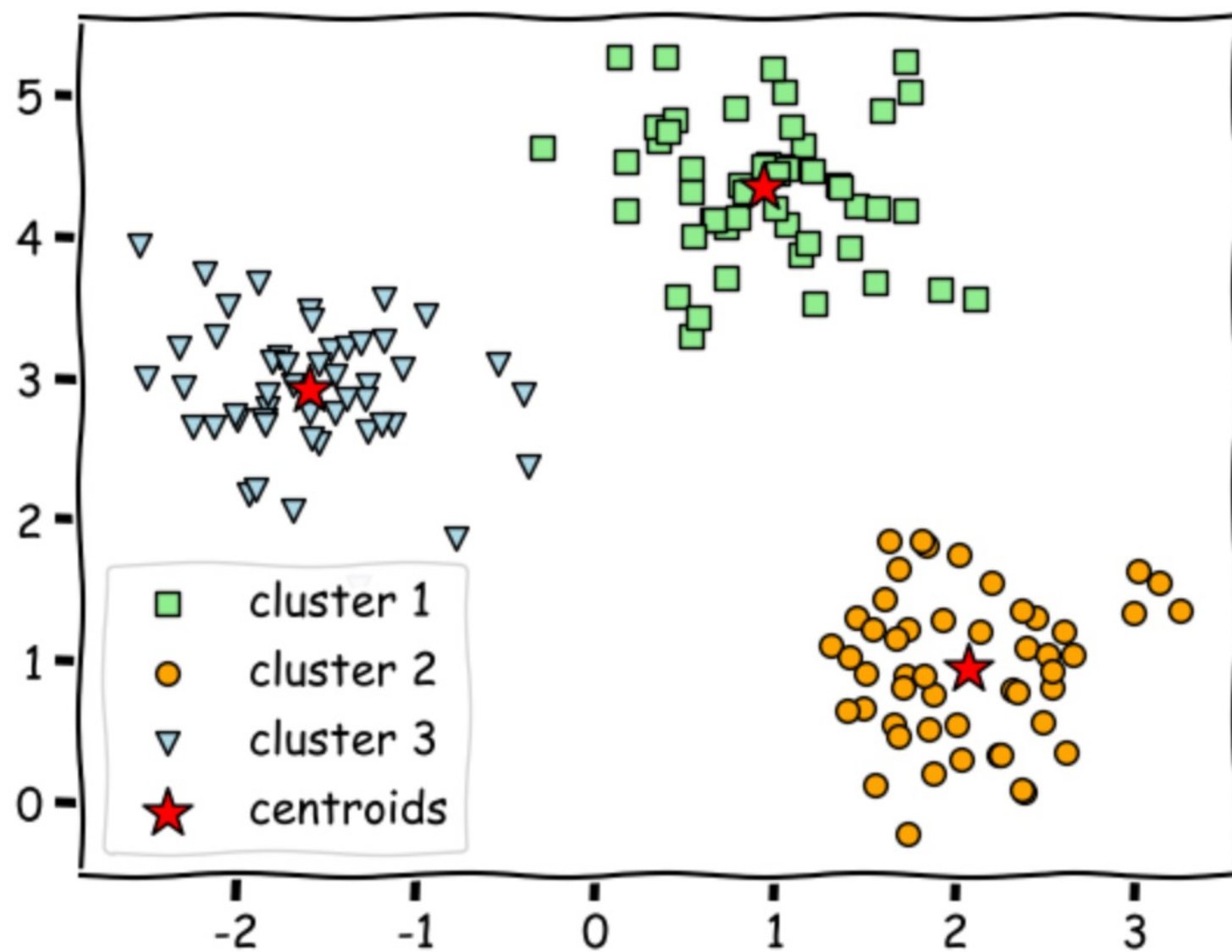
# Ch. 7. Clustering



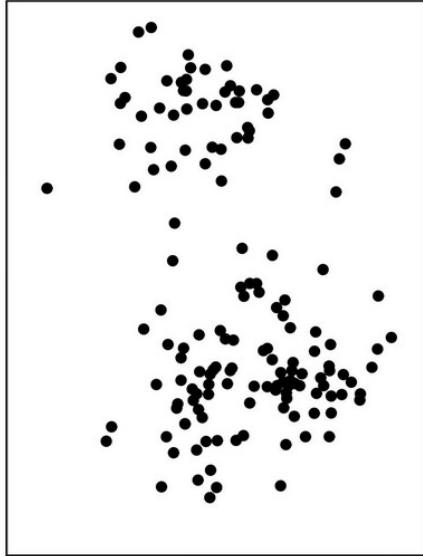
## Partitioning methods

Here, the number of clusters  $K$  is predetermined, and the cases  $c_i$  are iteratively assigned and reassigned to the clusters using a non-optimal or greedy algorithm. There are two types of partitioning methods.

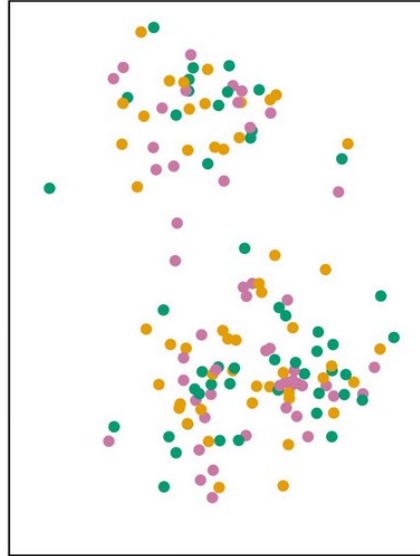
The first type is called *error minimization algorithms*. A perfect example is the  $K$ -means algorithm, where we start with  $K$  cluster centers, assign all cases based on the nearest center, recalculate the means of the cluster and keep iterating until the error minimization criterion is satisfied. This is a *gradient descent* method, because it can be mathematically proved that the sum of squares error actually decreased from one iteration to the next. However, it is not a globally optimal routine. The time complexity for  $T$  iterations with  $K$ -means,  $N$  cases and  $k$  variables per case is given to be  $O(T * K * k * N)$ . The linearity of time complexity is one of the main reasons why this algorithm is very popular.



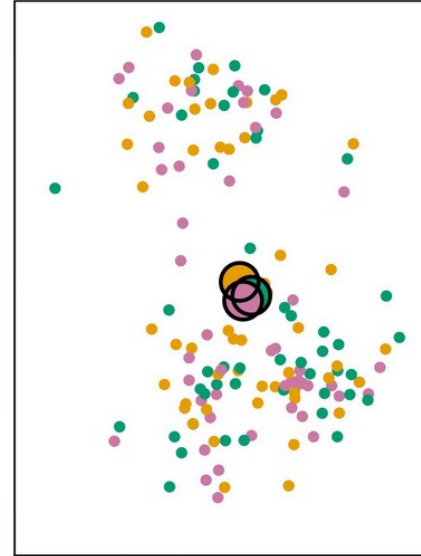
Data



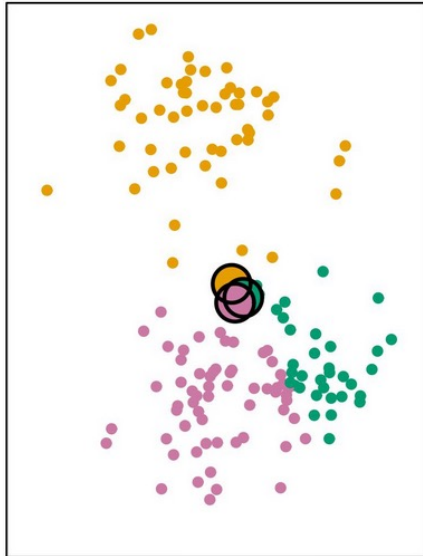
Step 1



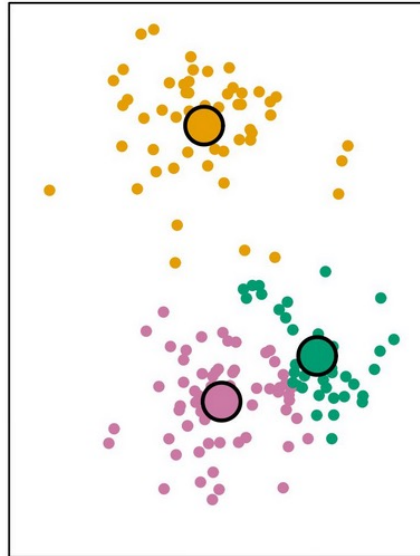
Iteration 1, Step 2a



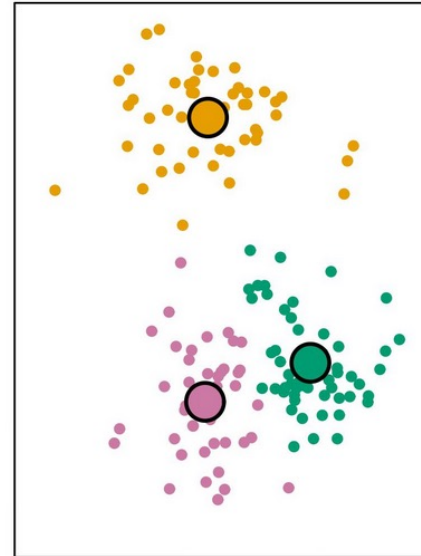
Iteration 1, Step 2b



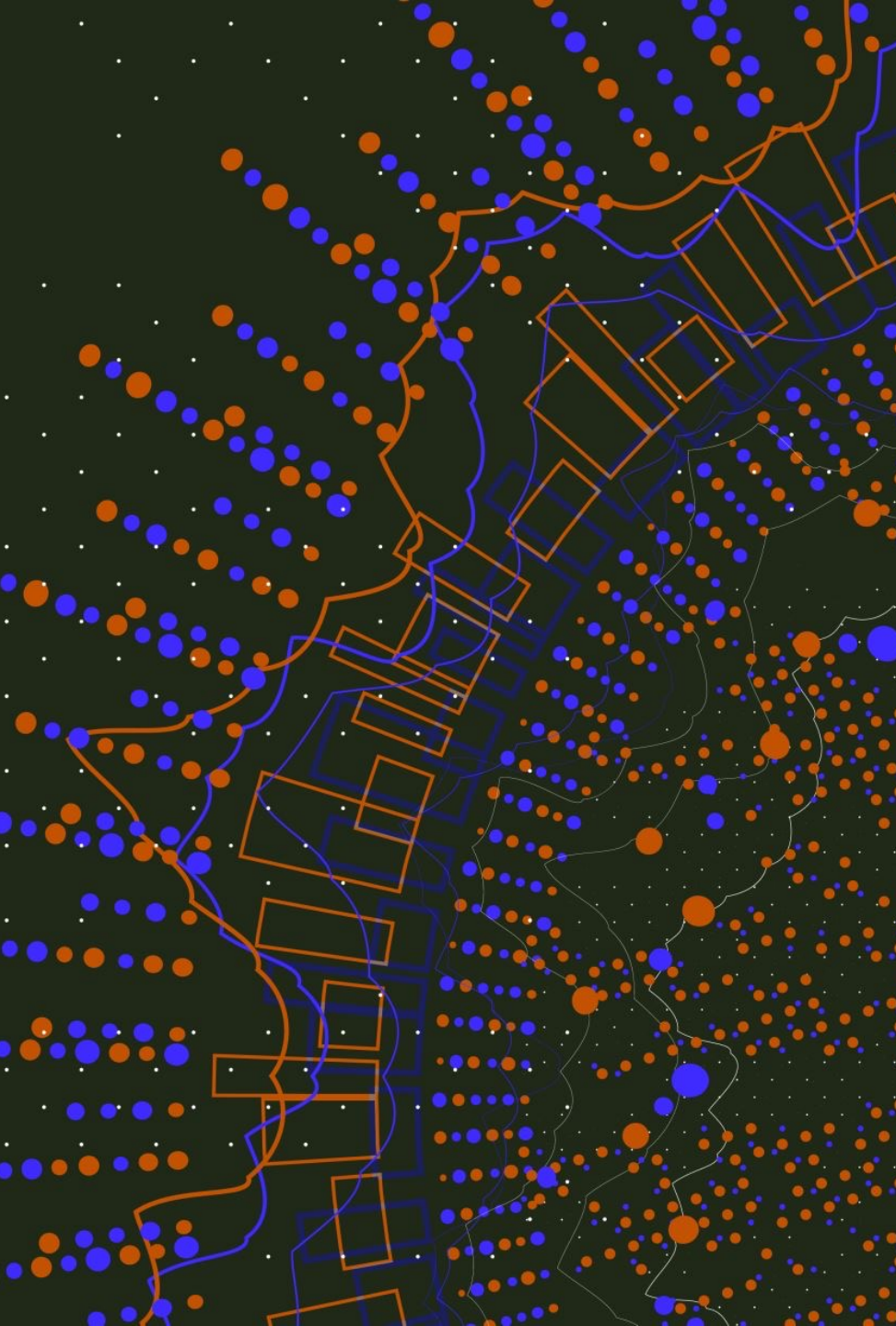
Iteration 2, Step 2a



Final Results



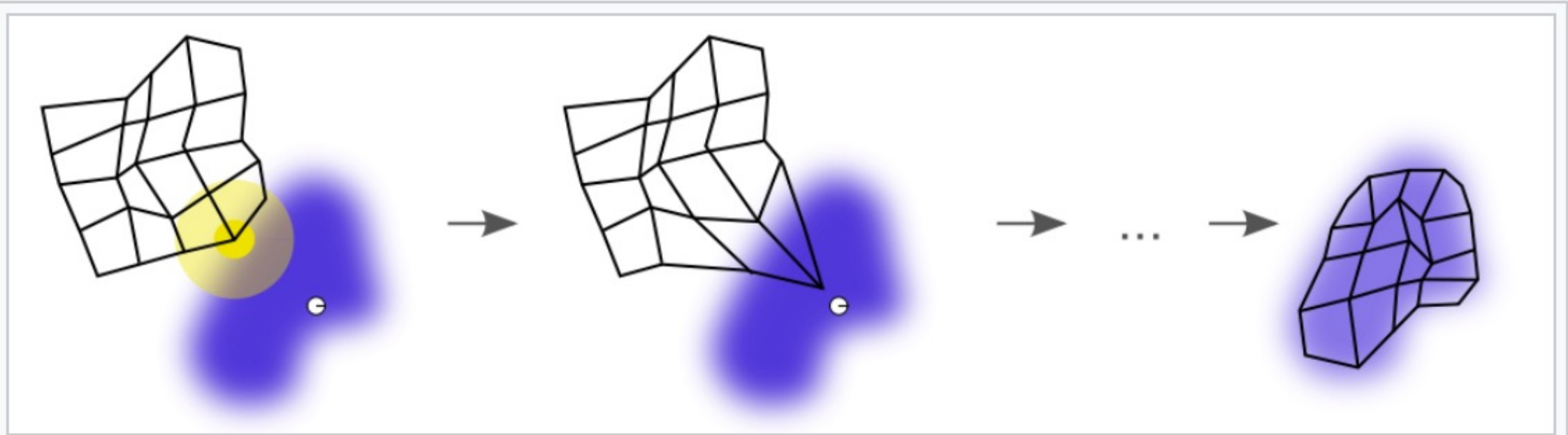




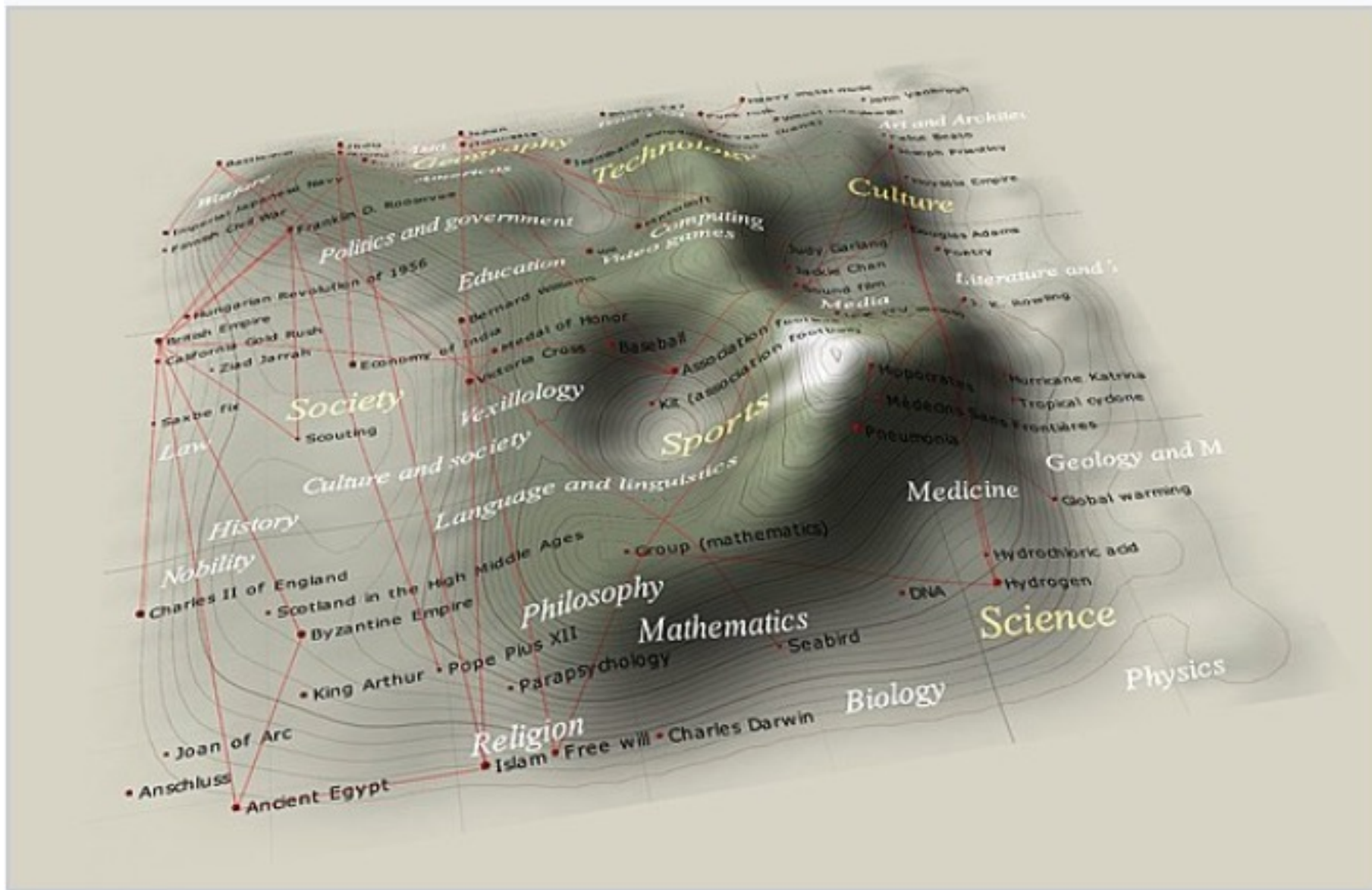
# Using AI

---

- A **self-organizing map (SOM)** or **self-organizing feature map (SOFM)** is an [unsupervised machine learning](#) technique.
- It is used to produce a [low-dimensional](#) (typically two-dimensional) representation of a higher dimensional data set while preserving the [topological structure](#) of the data.
- For example, a data set with  $p$  variables measured in  $n$  observations could be represented as clusters of observations with similar values for the variables.
- These clusters then could be visualized as a two-dimensional "map" such that observations in proximal clusters have more similar values than observations in distal clusters.
- This can make high-dimensional data easier to visualize and analyse.



An illustration of the training of a self-organizing map. The blue blob is the distribution of the training data, and the small white disc is the current training datum drawn from that distribution. At first (left) the SOM nodes are arbitrarily positioned in the data space. The node (highlighted in yellow) which is nearest to the training datum is selected. It is moved towards the training datum, as (to a lesser extent) are its neighbors on the grid. After many iterations the grid tends to approximate the data distribution (right).



Cartographical representation of a self-organizing map ([U-Matrix](#)) based on Wikipedia featured article data (word frequency). Distance is inversely proportional to similarity. The "mountains" are edges between clusters. The red lines are links between articles.





Exploring complex data from a case-based perspective

## Build the Model

1. Build Database and Import Cases
2. Cluster Cases

## Test the Model

3. The Computer's turn
4. Compare and Visualise Results

## Extend the Model

5. Simulate Interventions
6. Predict New Cases

## Export Results

7. Generate Report

beta version  
release 2019

COMPLEX-IT is a web-based and downloadable software tool designed to increase your access to the tools of computational social science (i.e., artificial intelligence, micro-simulation, predictive analytics). It does this through a user friendly interface, with quick access to introductions on concepts and methods; and with directions to richer detail and information for those who want it.

The result is a seamless and visually intuitive learning environment for exploring your complex data -- from data classification and visualisation to exploring simulated interventions and policy changes to data forecasting.

**You don't need any technical expertise to start using COMPLEX-IT, all that is required is a data set you want to explore, and a curious mind!**



DOWNLOAD  
VERSION



WEB  
VERSION

USER  
RESOURCES

Video Tutorials  
Step-by-step User Guide  
Additional Readings

Meet the team

Brian Castellani



Corey Schimpf




Michael Ball



Peter Barbrook-Johnson



## Exploring comorbid depression and physical health trajectories: A case-based computational modelling approach

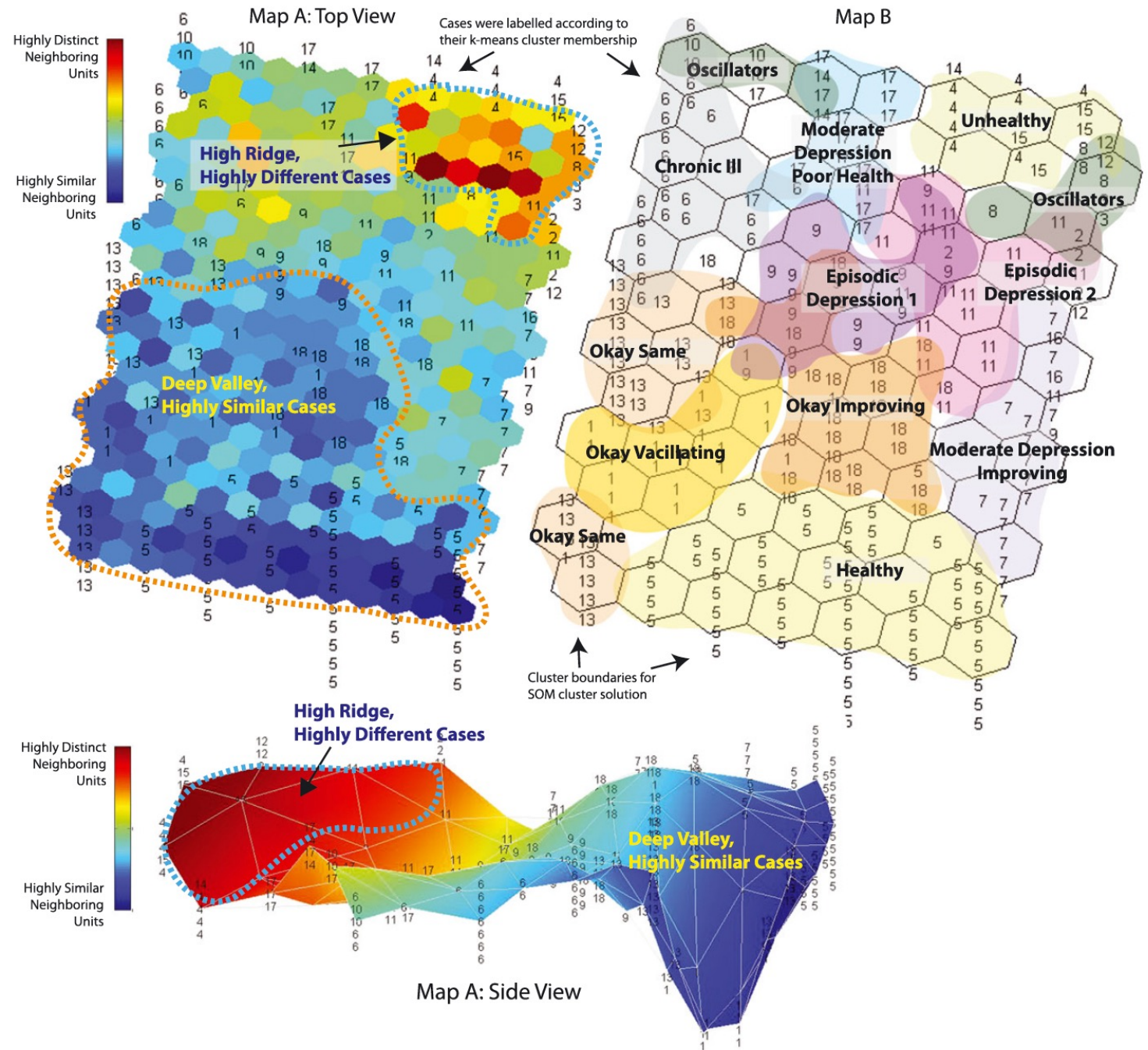
Brian Castellani PhD, Professor of Sociology<sup>1,6</sup>  |

Frances Griffiths MD PhD, Professor of Medicine<sup>2,3</sup> |

Rajeev Rajaram PhD, Associate Professor<sup>4</sup> | Jane Gunn MD PhD, Professor of Medicine<sup>5</sup>

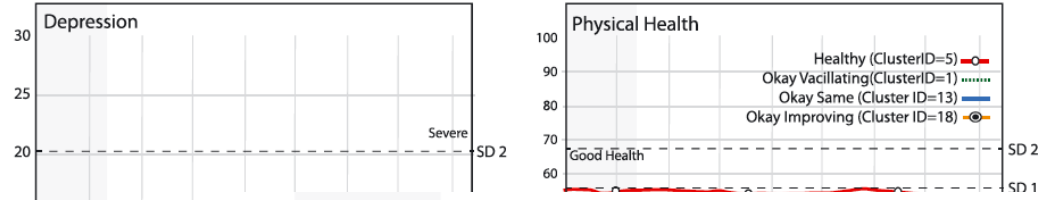
### Abstract

While comorbid depression/physical health is a major clinical concern, the conventional methods of medicine make it difficult to model the complexities of this relationship. Such challenges include cataloguing multiple trends, developing multiple complex aetiological explanations, and modelling the collective large-scale dynamics of these trends. Using a case-based complexity approach, this study engaged in a richly described case study to demonstrate the utility of computational modelling for primary care research. N = 259 people were subsampled from the *Diamond* database, one of the largest primary care depression cohort studies worldwide. A global measure of depressive symptoms (PHQ-9) and physical health (PCS-12) were assessed at 3, 6, 9, and 12 months and then annually for a total of 7 years. Eleven trajectories and 2 large-scale collective dynamics were identified, revealing that while depression is comorbid with poor physical health, chronic illness is often low dynamic and not always linked to depression. Also, some of the cases in the unhealthy and oscillator trends remain ill without much chance of improvement. Finally, childhood abuse, partner violence, and negative life events are greater amongst unhealthy trends. Computational modelling offers a major advance for health researchers to account for the diversity of primary care patients and for developing better prognostic models for team-based interdisciplinary care.





Healthy to Okay Health Group



Overall Poor Health Group

