AMSI SUMMER SCHOOL 2017
Modelling and Analysis of Big Data

Lecturer: Kerrie Mengersen, Professor of Statistics
Tutor: Miles McBain, Consultant Statistician

In collaboration with colleagues from BRAG and ACEMS
Acknowledgement: some content was written for QUT FL MOOC

QUT

ACEMS
AUSTRALIAN RESEARCH COUNCIL CENTRE OF EXCELLENCE FOR MATHEMATICAL AND STATISTICAL FRONTIERS

Plan

1. Let's talk about "Big Data"
2. Methods for modelling and analysis of big data
3. Digging deeper: (1) classification
4. Digging deeper: (2) regression
5. Digging deeper: (3) clustering
6. Digging deeper: (4) dimension reduction
7. Case study: recommender systems
8. From the learning to the doing: tips and tricks

QUT

ACEMS
AUSTRALIAN RESEARCH COUNCIL CENTRE OF EXCELLENCE FOR MATHEMATICAL AND STATISTICAL FRONTIERS

Examples

- Web searches: the Internet is massive, so a search query can result in a large number of very different objects. Hence the searches are grouped into common topics and then presented to the user, so we can more easily scroll through the topics to find what we want. This grouping is a form of clustering, and is commonly achieved using different types of k-means clustering algorithms.

- Marketing: with a portfolio of many products, instead of developing an individual marketing plan for each product, we could find groups, or clusters, of products with common features and develop marketing plans for each group of products.

- Gene expression analysis: clusters of genes can provide information about common cell types and biological functions.

Examples

- Anomaly detection: events or objects with similar features can be clustered and those events or objects that are markedly different can form a different cluster. These anomalies can then be inspected to identify reasons for the differences.
  For example: in disease mapping, the anomalies might be spikes in incidence of the disease as particular locations. It might be discovered that these have a common environmental exposure such as arsenic in the soil or a common socioeconomic pattern such as poverty or poor access to health services.

- Health risk profiling: a group of patients can be allocated to groups according to their risk of an adverse outcome. Those patients who are clustered into a high risk group can then be treated first. In many hospitals, the 'triage' of patients in an emergency ward is an example of risk clustering.

Clustering approaches

Common approaches:

- K-means
- Agglomerative clustering
- Mixture models

K-means

Partition observations into a fixed number (k) of clusters; each observation belongs to a specific cluster, based on similar properties.

Three steps:

1. Start the algorithm by choosing the number of clusters (k) and setting k points in the sample space. These points are chosen arbitrarily or according to some rule, and will be the initial cluster centres (centroids).
2. Allocate each observation to the closest cluster centroid.
3. Recalculate the cluster centroid as the mean, or average, of the observations that have been allocated to it.

Repeat steps 2 and 3 until the allocations stabilize.

The aim is to locate means and allocate observations to minimise within-cluster variation.

K-means

Given a set of observations

$$(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$$

where each observation is a *d*-dimensional real vector,
partition the *n* observations into *k* (≤ *n*) sets

$$\mathbf{S} = \{S_1, S_2, \ldots, S_k\}$$

to minimize the within-cluster sum of squares (WCSS)
(sum of distance functions of each point in the cluster to the K center).

So the objective is to find:

$$\arg\min_S \sum_{i=1}^{k} \sum_{x \in S_i} ||\boldsymbol{x} - \boldsymbol{\mu_i}||^2$$

where $\boldsymbol{\mu}_i$ is the mean of points in S_i.

This also leads to a Voronoi partition.

QUT

ACEMS
AUSTRALIAN RESEARCH COUNCIL CENTRE OF EXCELLENCE FOR MATHEMATICAL AND STATISTICAL FRONTIERS

Voronoi partitions

https://en.wikipedia.org/wiki/Voronoi_diagram

A **Voronoi diagram** is a partitioning of a plane into regions based on distance to points in a specific subset of the plane. That set of points (called seeds, sites, or generators) is specified beforehand, and for each seed there is a corresponding region consisting of all points closer to that seed than to any other. These regions are called Voronoi cells.

Eg: consider a group of shops in a city. Suppose we want to estimate the number of customers of a given shop. Assume that customers go to the nearest shop. In this case the Voronoi cell of a given shop can be used for giving a rough estimate on the number of potential customers going to this shop (which is modeled by a point in our city).

We can measure the distance between points using a range of distance metrics

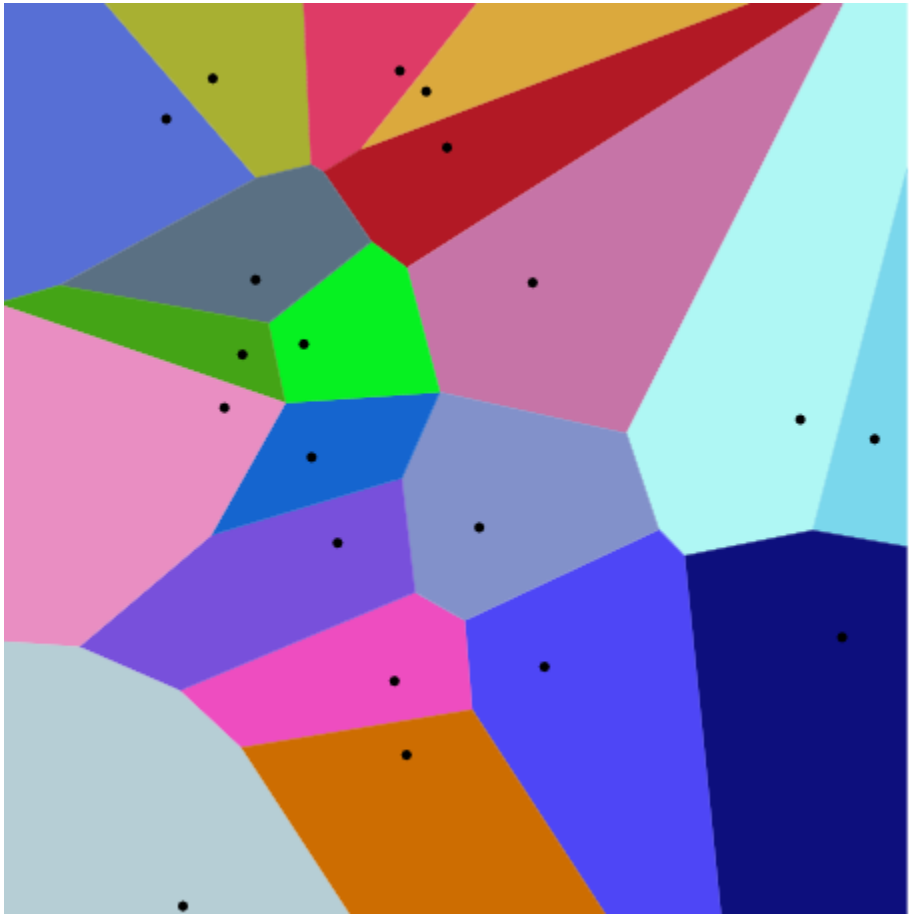
Examples of metrics:

L2 (Euclidean) $d[(a_1, a_2), (b_1, b_2)] = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2}$

L1 (Manhattan) $d[(a_1, a_2), (b_1, b_2)] = |a_1 - b_1| + |a_2 - b_2|$

QUT

ACEMS
AUSTRALIAN RESEARCH COUNCIL CENTRE OF EXCELLENCE FOR MATHEMATICAL AND STATISTICAL FRONTIERS



Euclidean partition

Manhattan partition

https://en.wikipedia.org/wiki/Voronoi_diagram

QUT

ACEMS

AUSTRALIAN RESEARCH COUNCIL CENTRE OF EXCELLENCE FOR MATHEMATICAL AND STATISTICAL FRONTIERS

Choosing k

- Sometimes there is a reason for specifying a certain number of clusters: this might be based on knowledge of the underlying biological or physical system in a scientific application, an economic or marketing rationale in business, and so on.

- If there is no such reason, then the k-means model can be run with different values of k, and the results compared.

Comparing results

- As k increases, the average distance between the observations and their cluster centroids decreases, which means that the observations within a cluster will be more similar.

- Sometimes there is a clear point at which the increase in k results in much less improvement in within-cluster similarity. The analyst can choose k to be at this point.

- Cross-validation can be used to provide a robust estimate of k. Cross-validation entails fitting the k-means algorithm to a subset of the data (called the training set) and then applying the clusters to the remaining data (the test set). This is particularly useful if the eventual aim is to allocate new observations to the clusters.

- The analyst can also gain insight into the effect of changing k by monitoring how particular observations are allocated to different groups.

Example

https://www.edureka.co/blog/implementing-kmeans-clustering-on-the-crime-dataset/

Crime data in 50 US states, per 100,000 people in a year

row.names	Murder	Assault	UrbanPop	Rape
Alabama	13.2	236	58	21.2
Alaska	10.0	263	48	44.5
Arizona	8.1	294	80	31.0
Arkansas	8.8	190	50	19.5
California	9.0	276	91	40.6
Colorado	7.9	204	78	38.7
Connecticut	3.3	110	77	11.1
Delaware	5.9	238	72	15.8
Florida	15.4	335	80	31.9
Georgia	17.4	211	60	25.8
Hawaii	5.3	46	83	20.2
Idaho	2.6	120	54	14.2
Illinois	10.4	249	83	24.0
Indiana	7.2	113	65	21.0
Iowa	2.2	56	57	11.3

https://www.edureka.co/blog/implementing-kmeans-clustering-on-the-crime-dataset/

Results:

	crime$cluster	Murder	Assault	UrbanPop	Rape
Alabama	4	13.2	236	58	21.2
Alaska	4	10	263	48	44.5
Arizona	4	8.1	294	80	31
Arkansas	3	8.8	190	50	19.5
California	4	9	276	91	40.6
Colorado	3	7.9	204	78	38.7
Connecticut	2	3.3	110	77	11.1
Delaware	4	5.9	238	72	15.8
Florida	4	15.4	335	80	31.9

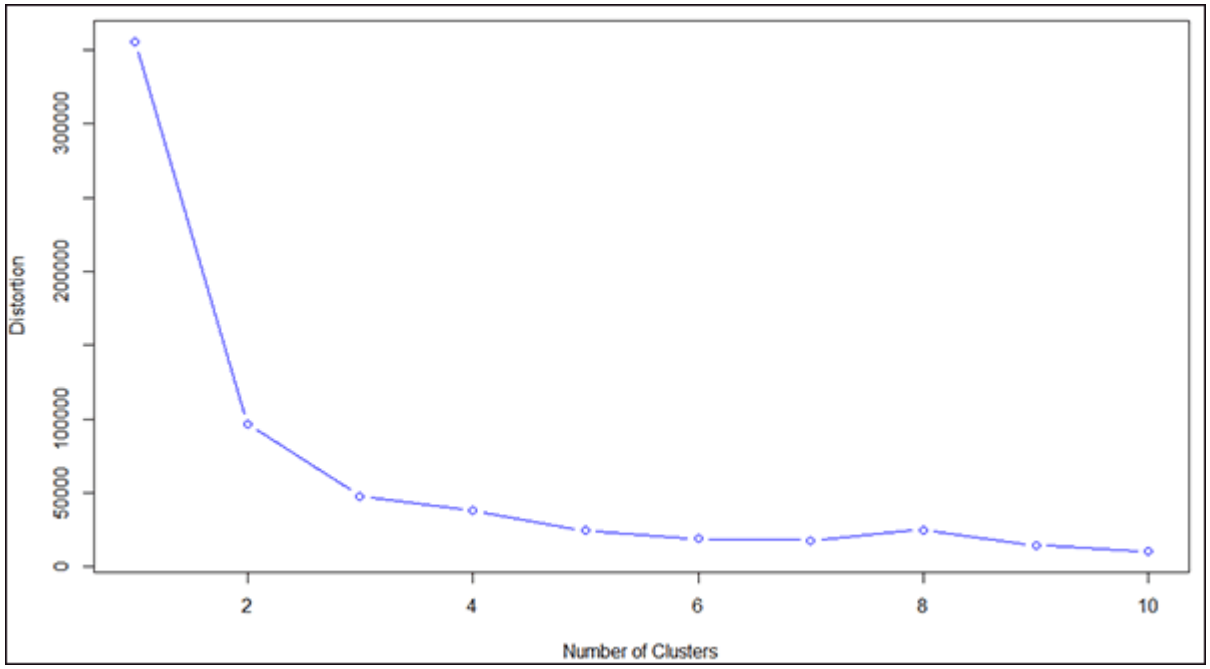
No. clusters: 5
Total SS: 355808
Within SS: 4548, 2286, 1480, 3653
Total Within SS: 28240
Between SS: 327568
Size of clusters: 10, 9, 14, 10, 7

QUT

ACEMS
AUSTRALIAN RESEARCH COUNCIL CENTRE OF EXCELLENCE FOR MATHEMATICAL AND STATISTICAL FRONTIERS

https://www.edureka.co/blog/implementing-kmeans-clustering-on-the-crime-dataset/

Compare distortion (via within SS) for various values of k:



QUT

ACEMS
AUSTRALIAN RESEARCH COUNCIL CENTRE OF EXCELLENCE FOR
MATHEMATICAL AND STATISTICAL FRONTIERS

https://www.edureka.co/blog/implementing-kmeans-clustering-on-the-crime-dataset/

k=4:

Cluster centres:	Murder	Assault	UrbanPop	Rape
Texas	4.74	104.85	62.96	16.10
Louisiana	10.90	219.92	71.71	25.95
South Carolina	13.37	284.50	46.25	25.05
New Mexico	11.04	298.00	77.60	32.68

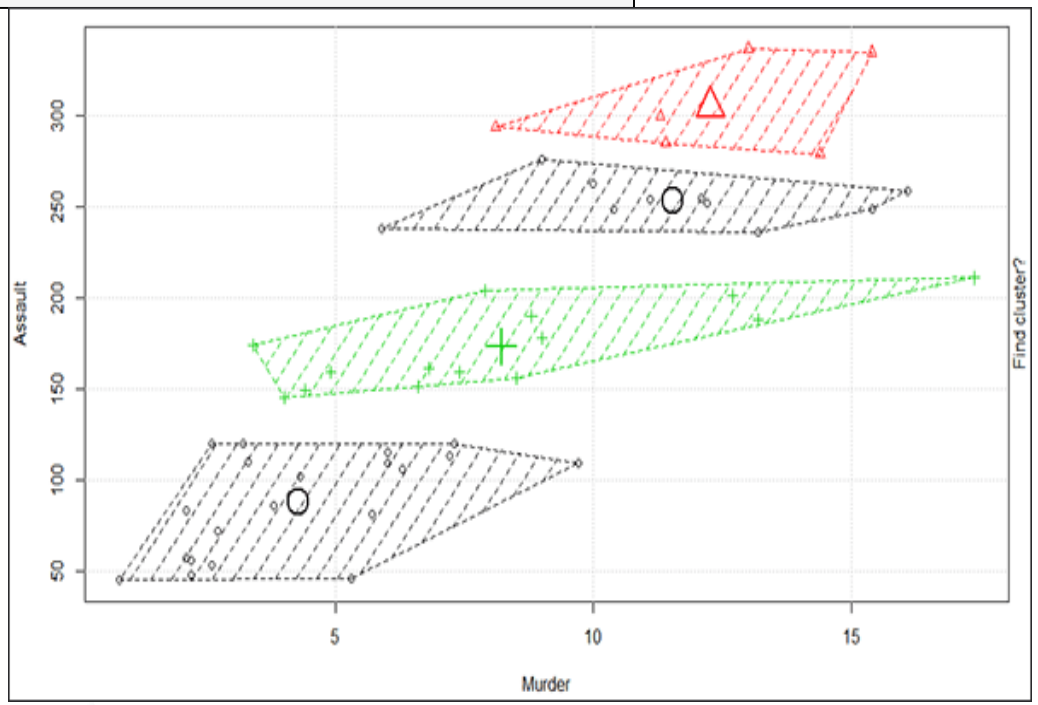
	Cluster Assign	Murder	Assault	UrbanPop	Rape
Alabama	2	13.2	236	58	21.2
Alaska	3	10	263	48	44.5
Arizona	4	8.1	294	80	31
Arkansas	2	8.8	190	50	19.5
California	4	9	276	91	40.6
Colorado	2	7.9	204	78	38.7
Connectic	1	3.3	110	77	11.1
Delaware	2	5.9	238	72	15.8
Florida	4	15.4	335	80	31.9
Georgia	2	17.4	211	60	25.8
Hawaii	1	5.3	46	83	20.2

https://www.edureka.co/blog/implementing-kmeans-clustering-on-the-crime-dataset/

k=4:

Cluster centres:	Murder	Assault	UrbanPop	Rape
Texas	4.74	104.85	62.96	16.10
Louisiana	10.90	219.92	71.71	25.95
South Carolina	13.37	284.50	46.25	25.05
New Mexico	11.04	298.00	77.60	32.68

	Cluster Assign	Murder	Assault	UrbanPop	Rape
Alabama	2	13.2	236	58	21.2
Alaska	3	10	263	48	44.5
Arizona	4	8.1	294	80	31
Arkansas	2	8.8	190	50	19.5
California	4	9	276	91	40.6
Colorado	2	7.9	204	78	38.7
Connectic	1	3.3	110	77	11.1
Delaware	2	5.9	238	72	15.8
Florida	4	15.4	335	80	31.9
Georgia	2	17.4	211	60	25.8
Hawaii	1	5.3	46	83	20.2



QUT

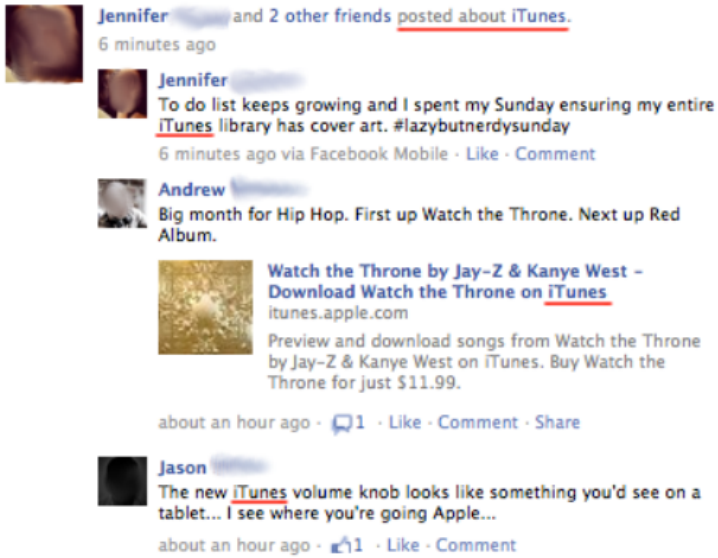
ACEMS
AUSTRALIAN RESEARCH COUNCIL CENTRE OF EXCELLENCE FOR MATHEMATICAL AND STATISTICAL FRONTIERS

Clustering big data

http://www.cse.nd.edu/Fu_Prize_Seminars/jain/slides.pdf

Clustering Users on Facebook

- ~300,000 status updates per minute on tens of thousands of topics
- Cluster users based on topic of status messages

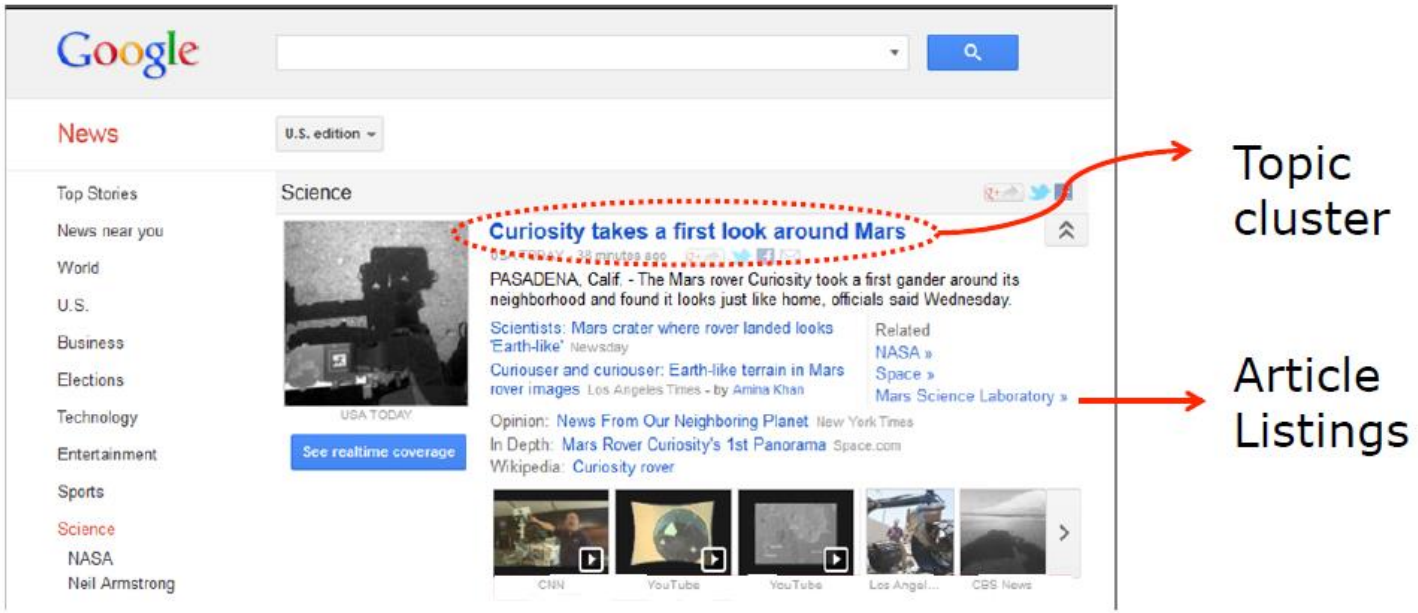


http://www.insidefacebook.com/2011/08/08/posted-about-page/
http://searchengineland.com/by-the-numbers-twitter-vs-facebook-vs-google-buzz-36709

Clustering big data

http://www.cse.nd.edu/Fu_Prize_Seminars/jain/slides.pdf

Clustering Articles on Google News

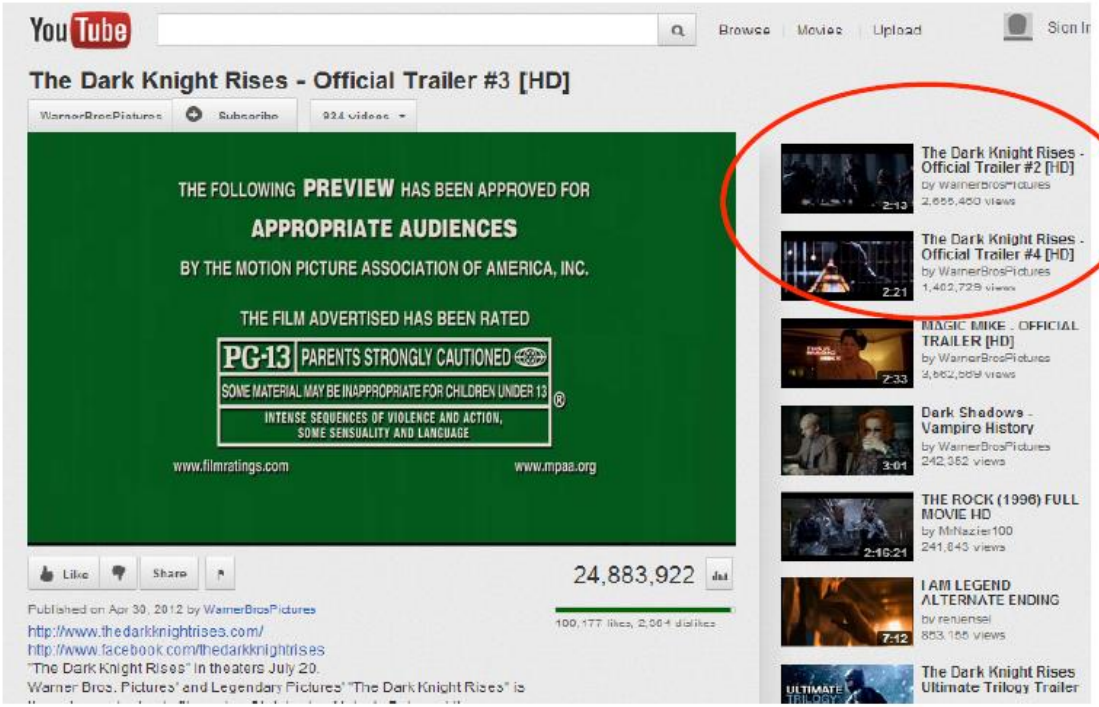


http://blogoscoped.com/archive/2006-07-28-n49.html

Clustering big data

http://www.cse.nd.edu/Fu_Prize_Seminars/jain/slides.pdf

Clustering Videos on Youtube



- Keywords
- Popularity
- Viewer engagement
- User browsing history

http://www.strutta.com/blog/blog/six-degrees-of-youtube

Clustering big data

http://www.cse.nd.edu/Fu_Prize_Seminars/jain/slides.pdf

Clustering for Efficient Image retrieval

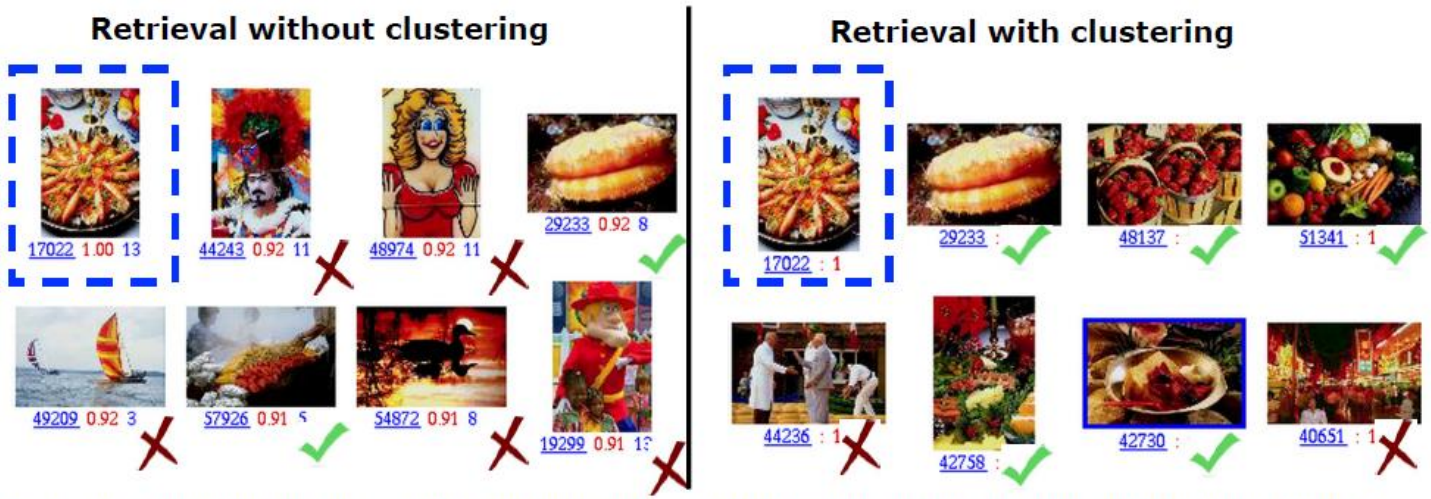


Fig. 1. Upper-left image is the query. Numbers under the images on left side: image ID and cluster ID; on the right side: Image ID, matching score, number of regions.

Retrieval accuracy for the "food" category (average precision):

Without clustering: **47%** With clustering: **61%**

Chen et al., "CLUE: cluster-based retrieval of images by unsupervised learning," IEEE Tans. On Image Processing, 2005.

When k-means clustering fails

https://dzone.com/articles/when-k-means-clustering-fails

When groups have very different sizes or data are categorical:

- Partitioning Around Medoids (**pam**)
- **clara** when the dataset is very large (and pam is slow)

While this [failure of k-means] may not be news to long-time clustering gurus it was a little sobering for us. It taught us once again that just because you've heard of some fancy algorithm and are using R's implementation of that algorithm does not guarantee that you'll get the results you expect. You always have to inspect results visually.

clara

Consider sub-datasets of fixed size (sampsize) such that the time and storage requirements become linear in *n* rather than quadratic.

- Each sub-dataset is partitioned into k clusters using the same algorithm as in pam.
- Once k representative objects have been selected from the sub-dataset, each observation of the entire dataset is assigned to the nearest medoid.
- The mean (equivalent to the sum) of the dissimilarities of the observations to their closest medoid is used as a measure of the quality of the clustering. The sub-dataset for which the mean (or sum) is minimal, is retained.

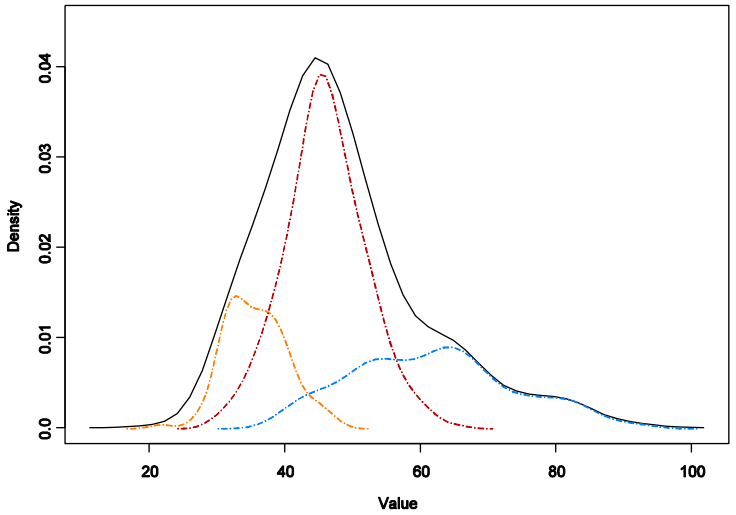
A further analysis is carried out on the final partition.

- Each sub-dataset is forced to contain the medoids obtained from the best sub-dataset until then.
- Randomly drawn observations are added to this set until sampsize has been reached.

https://stat.ethz.ch/R-manual/R-devel/library/cluster/html/clara.html

Mixture Models

The observed values
are observations from
a mixture of distributions



Eg, phenotypes from
3 genotypes: qq, qQ, QQ

Eg, for mixture of K=3 Normals: $\theta = (\mu,\sigma)$

$$y \sim p_1\, N(\mu_1, \sigma_1^2) + p_2\, N(\mu_2, \sigma_2^2) + p_3\, N(\mu_3, \sigma_3^2)$$

QUT

ACEMS
AUSTRALIAN RESEARCH COUNCIL CENTRE OF EXCELLENCE FOR
MATHEMATICAL AND STATISTICAL FRONTIERS

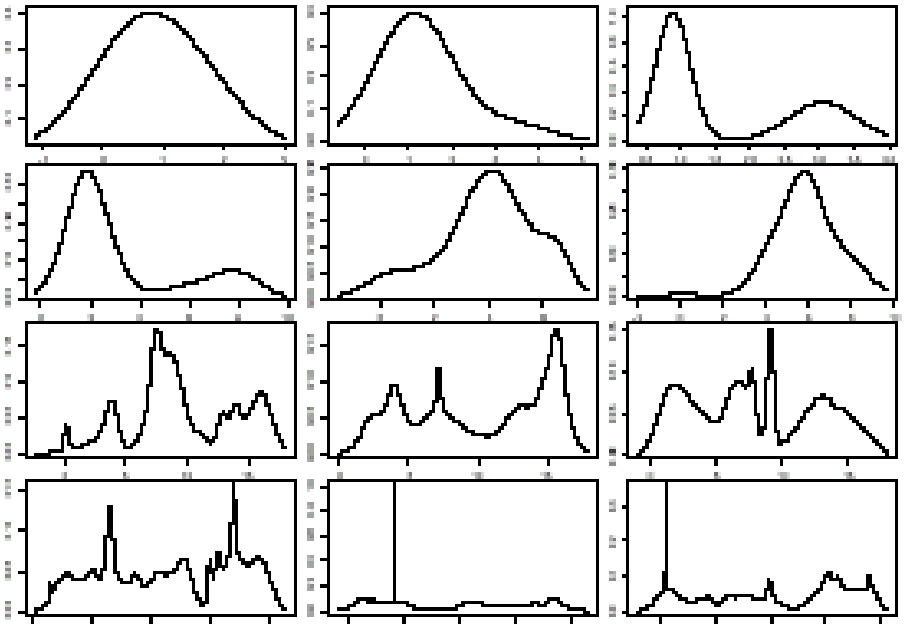


FIGURE 1. Some normal mixture densities for $K = 2$ *(first row)*, $K = 5$ *(second row)*, $K = 25$ *(third row)* and $K = 50$ *(last row)*.

Mixture models: Frequentist approach

Expectation-Maximisation (EM)

- expectation (E) step: computes the expectation of the log-likelihood evaluated using the current estimate for the parameters
- maximization (M) step: computes parameters maximizing the expected log-likelihood found on the *E* step.

For large datasets, use factor mixtures:

Lee, McLachlan, Pyne (2016), in *Big Data Analytics* by Pyne, Rao and Rao, Springer, 2016.

QUT

ACEMS
AUSTRALIAN RESEARCH COUNCIL CENTRE OF EXCELLENCE FOR MATHEMATICAL AND STATISTICAL FRONTIERS

Mixture model approaches

Other two-step approaches

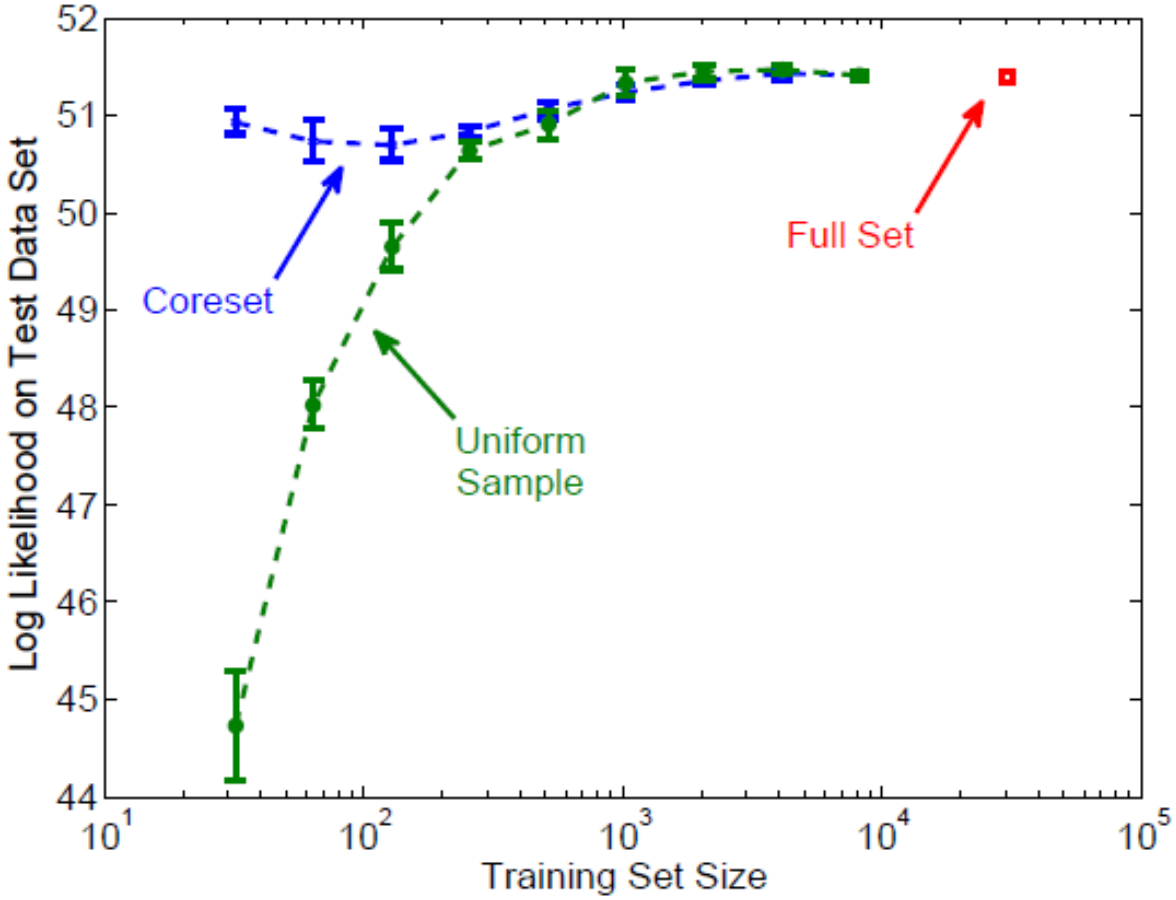
- Example: employ a two-step approach for very large datasets: (i) compress data by clustering the observations into a medium number of groups and representing each group by a triple of sufficient statistics (mean vector, covariance matrix, no. observations); (ii) estimate the mixture using by applying an adapted EM algorithm to the sufficient statistics; (iii) classify observations to clusters by maximum posterior probability of component membership (Steiner & Hudec, 2007).

- Example: use coresets: weighted subsets of the data (https://las.inf.ethz.ch/files/feldman11scalable-long.pdf)

Gaussian mixtures admit coresets of size *independent* of the size of the data set. More precisely, we prove that a weighted set of $O(dk^3/\varepsilon^2)$ data points suffices for computing a $(1+\varepsilon)$-approximation for the optimal model on the original n data points. Moreover, such coresets can be efficiently constructed in a map-reduce style computation, as well as in a streaming setting.

QUT

ACEMS
AUSTRALIAN RESEARCH COUNCIL CENTRE OF EXCELLENCE FOR MATHEMATICAL AND STATISTICAL FRONTIERS



(a) *MNIST*

QUT

ACEMS
AUSTRALIAN RESEARCH COUNCIL CENTRE OF EXCELLENCE FOR MATHEMATICAL AND STATISTICAL FRONTIERS

Bayesian approaches

$y \sim \Sigma_{j=1:k}\, p_j\, N(\mu_j, \sigma_j^2)$

$\mu \sim$ Normal

$\sigma \sim$ Uniform

$p \sim$ Dirichlet(α_1,.., α_K)

$f(p;\alpha) \propto \Pi\, p_j^{\alpha j-1}$; setting α=1 for all j gives the Uniform.

QUT

ACEMS
AUSTRALIAN RESEARCH COUNCIL CENTRE OF EXCELLENCE FOR MATHEMATICAL AND STATISTICAL FRONTIERS

Latent variable approach

- Associate with each y_i another variable T_i that identifies the component of the mixture to which that y_i belongs.
(Note that we don't observe the T's.)

- We can then 'break down' the likelihood:

$$y_i \mid T_i = T \sim \mathrm{N}(y \mid \mu_{Ti}, \sigma_{Ti}^2)$$

now just a univariate problem

- A typical prior for T is the multinomial or categorical distribution
$T_i \sim Multi(p_1,\ldots,p_K)$

Gibbs sampling for mixtures

0. *Initialisation:* Choose $\underline{p}^{(0)}$ and $\underline{\theta}^{(0)}$ arbitrarily

For t=1,…

1.1 Allocate observations to components:
Generate $T^{(t)}$ for each observation

1.2 Generate new weights for the components:
Generate $\underline{p}^{(t)}$

1.3 Generate new parameters for each component:
Generate $\underline{\theta}^{(t)}$

R packages for clustering

Cran Task View: Cluster analysis and finite mixture models

https://cran.r-project.org/web/views/Cluster.html

- Hierarchical clustering
- Partitioning clustering
- Model-based clustering
- Bayesian estimation
- Other estimation methods
- Other clustering algorithms
- Cluster-wise regression (e.g. time series, mixtures of regressions, growth mixtures)
- Additional functionality

Plan

1. Let's talk about "Big Data"
2. Methods for modelling and analysis of big data
3. Digging deeper: (1) classification
4. Digging deeper: (2) regression
5. Digging deeper: (3) clustering
6. Digging deeper: (4) dimension reduction
7. Case study: recommender systems
8. From the learning to the doing: tips and tricks

QUT

ACEMS
AUSTRALIAN RESEARCH COUNCIL CENTRE OF EXCELLENCE FOR MATHEMATICAL AND STATISTICAL FRONTIERS

Dimension reduction

Common approaches:

- Principal Component Analysis (PCA)
- Factor Analysis (FA)
- Page Rank

PCA

- Principal components analysis, or PCA, seeks to find a set of orthogonal axes such that the first axis, or first principal component, accounts for as much variability as possible and subsequent axes are chosen to maximize variance while maintaining orthogonality with previous axes.

- Principal components are typically computed either by a singular value decomposition of the data matrix or an eigenvalue decomposition of a covariance or correlation matrix.

https://www.r-bloggers.com/big-data-pca-50-years-of-stock-data/

Eigenvalue analysis

Principal components. Population PCA for the random vector $x = (X_1, \cdots, X_d)^T$ first produces a measure of the variability of x by finding the linear combination $e^T x$ that has maximal normalized variance $Var\left(e^T x\right) / \|e\|^2$. Let Σ denote the covariance matrix of x, then e_1, the first eigenvector, is

$$e_1 = \underset{e:\|e\|=1}{\operatorname{argmax}} \{e^T \Sigma e\} \tag{2.1}$$

and the first eigenvalue and the first principal component (PC_1) are

$$\lambda_1 = e_1{}^T \Sigma e_1, \quad PC_1 = e_1{}^T x.$$

The second eigenvector e_2, second eigenvalue λ_2, and second PC are obtained in the same way except e_2 is found by maximizing (2.1) over e orthogonal to e_1. To obtain e_k, λ_k and PC_k, (2.1) is maximized over e orthogonal to $e_1, \cdots, e_{k-1}$. This process produces the principal components $PC_1, \cdots, PC_d$ that capture much of the variability of x in the sense that $Var(PC_j) = \lambda_j$ and $\sum_{j=1}^{d} \lambda_j = \sum_{j=1}^{d} Var(X_j)$.

http://www.stat.wisc.edu/~doksum/papers/BigData.pdf

QUT

ACEMS
AUSTRALIAN RESEARCH COUNCIL CENTRE OF EXCELLENCE FOR MATHEMATICAL AND STATISTICAL FRONTIERS

Example

- Eg: Stock market data for open, high, low, close, and adjusted close from 1962 to 2010: 9.2 million observations of daily data for 2800 stocks.

```
summary(stockPca)
Importance of components: Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
Standard deviation          2.0756 0.8063 0.1976 0.0454 0.0018
Proportion of Variance      0.8616 0.1300 0.0078 0.0004 6.7E-5
Cumulative Proportion       0.8616 0.9917 0.9995 0.9999 1.0000
```

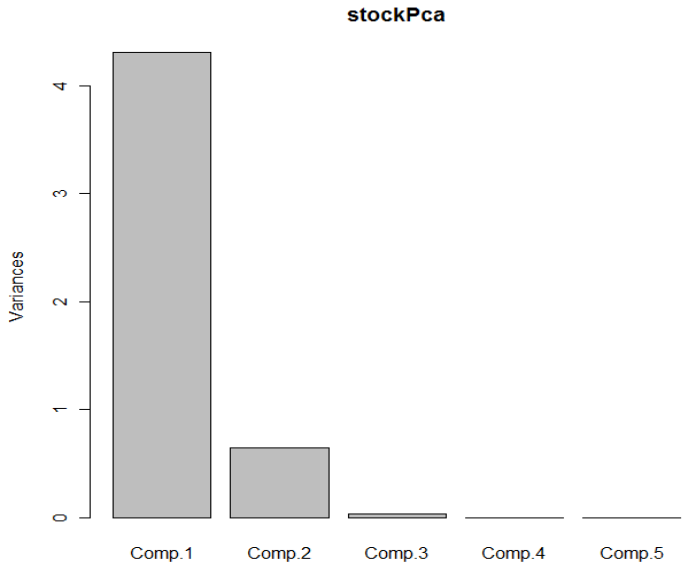
```
loadings(stockPca)
Loadings:                 Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
stock_price_open          -0.470 -0.166 0.867
stock_price_high          -0.477 -0.151 -0.276 0.410 -0.711
stock_price_low           -0.477 -0.153 -0.282 0.417 0.704
stock_price_close         -0.477 -0.149 -0.305 -0.811
stock_price_adj_close     -0.309 0.951
```

https://www.r-bloggers.com/big-data-pca-50-years-of-stock-data/

QUT

ACEMS
AUSTRALIAN RESEARCH COUNCIL CENTRE OF EXCELLENCE FOR MATHEMATICAL AND STATISTICAL FRONTIERS

Example



https://www.r-bloggers.com/big-data-pca-50-years-of-stock-data/

See also
http://www.bigdatanews.com/profiles/blogs/principal-component-analysis-using-r

https://www.analyticsvidhya.com/blog/2016/03/practical-guide-principal-component-analysis-python/

QUT

ACEMS
AUSTRALIAN RESEARCH COUNCIL CENTRE OF EXCELLENCE FOR MATHEMATICAL AND STATISTICAL FRONTIERS

PageRank

- Aims to rank web pages based on their hyperlinks (i.e., links between the pages).
- This algorithm underpins the search engine Google, and variations of the algorithm are now used for every online search engine.
- A hyperlink from page x to page y is defined as a vote, by page x, for page y. Votes casted by pages that are themselves "important" weigh more heavily and help to make other pages more "important". This is exactly the idea of rank prestige in social networks.
- The computation of PageRank values of the Web pages can be done using the power iteration method, which produces a principal eigenvector with an eigenvalue of 1. The iteration ends when the PageRank values do not change much (e.g, the sum of the absolute values of the residuals are less than a specified threshold).

See also https://en.wikipedia.org/wiki/PageRank
http://www.math.cornell.edu/~mec/Winter2009/RalucaRemus/Lecture3/lecture3.html

Plan

1. Let's talk about "Big Data"
2. Methods for modelling and analysis of big data
3. Digging deeper: (1) classification
4. Digging deeper: (2) regression
5. Digging deeper: (3) clustering
6. Digging deeper: (4) dimension reduction
7. Case study: recommender systems
8. From the learning to the doing: tips and tricks

QUT

ACEMS
AUSTRALIAN RESEARCH COUNCIL CENTRE OF EXCELLENCE FOR MATHEMATICAL AND STATISTICAL FRONTIERS

Overview

1. What *is* it?
2. What kind of data are used to create it?
3. What algorithms are used for it?
4. What statistical inferences can we draw from it?

Example of a recommender system

Recent (2016) paper:

Deep Neural Networks for YouTube Recommendations
by Paul Covington, Jay Adams, Emre Sargin

http://static.googleusercontent.com/media/research.google.com/en//pubs/archive/45530.pdf

- Recommending YouTube videos is extremely challenging due to *scale, freshness and noise.*
- The system is built on Google Brain which was recently open sourced as TensorFlow. deep neural network architectures using largescale distributed training.
- The models learn approximately one billion parameters and are trained on hundreds of billions of examples.

QUT

ACEMS
AUSTRALIAN RESEARCH COUNCIL CENTRE OF EXCELLENCE FOR MATHEMATICAL AND STATISTICAL FRONTIERS

Example

In order to facilitate searching and browsing of this massive database of videos, YouTube provides several features for information retrieval and recommendation. Users can use the search engine to search for specific videos, watch charts for Most Viewed, Top favorited, Most Liked and Most Discussed videos of the day, week, month and all-time. Apart from this, YouTube provides most popular videos in over 14 categories like Music, Sports, Education etc. It also shows users videos that their friends shared on social networking websites like Google+, Facebook etc. or liked using Google's +1. Lastly, it also recommends videos to signed-in users based on their browsing activity on the website.

http://dushyantarora.yolasite.com/resources/597A_dushyant_project_report.pdf

QUT

ACEMS
AUSTRALIAN RESEARCH COUNCIL CENTRE OF EXCELLENCE FOR MATHEMATICAL AND STATISTICAL FRONTIERS

Example

YouTube maintains the co-visitation counts (number of times videos were co-watched within a session of 24 hours) for each pair of videos (v_i,v_j). Now given that a user watched a video v_i, it calculates a relatedness score for all the videos that were co-watched with v_i. The relatedness score is calculated as:

$$r(v_i, v_j) = \frac{c_{ij}}{f(v_i, v_j)}$$

where c_i, c_j are the total occurrence counts for videos v_i and v_j and c_{ij} is the co-visitation count. $f(v_i,v_j)$ is a normalization function usually set as c_j, which favors less popular videos to more popular ones. All the videos with a relatedness score above a threshold are ranked and added to related videos set R_i for v_i. Now given a seed set S of videos watched by a user, the set of related videos R_i is calculated for all videos $v_i \in S$. These videos are now further ranked using a linear combination of signals like view count, ratings, commenting, favoriting and user's watch history to generate a list of recommendation videos for the user. A similar recommendation algorithm is used to implement "Customers who bought this also bought" feature provided on Amazon's website

http://dushyantarora.yolasite.com/resources/597A_dushyant_project_report.pdf

QUT

ACEMS
AUSTRALIAN RESEARCH COUNCIL CENTRE OF EXCELLENCE FOR MATHEMATICAL AND STATISTICAL FRONTIERS

Class discussion

- Content-based filtering collaboration
- Collaborative-based filtering collaboration

- Neural networks
- Boosted regression trees

Example: ACEMS students Amy Cook and Yasin Abbasi were semi-finalists for the 2015 Hilti Big Data Analytics Competition

QUT

ACEMS
AUSTRALIAN RESEARCH COUNCIL CENTRE OF EXCELLENCE FOR MATHEMATICAL AND STATISTICAL FRONTIERS

Plan

1. Let's talk about "Big Data"
2. Methods for modelling and analysis of big data
3. Digging deeper: (1) classification
4. Digging deeper: (2) regression
5. Digging deeper: (3) clustering
6. Digging deeper: (4) dimension reduction
7. Case study: recommender systems
8. From the learning to the doing: tips and tricks

QUT

ACEMS
AUSTRALIAN RESEARCH COUNCIL CENTRE OF EXCELLENCE FOR MATHEMATICAL AND STATISTICAL FRONTIERS

Big data analysis in the cloud

Informative reference site:

https://cloud.google.com/blog/big-data/

Doing analysis in the cloud:

Over to Miles…

MOOCS

Our moocs developed by QUT, ACEMS and FutureLearn

https://www.futurelearn.com/programs/big-data-analytics

QUT

ACEMS
AUSTRALIAN RESEARCH COUNCIL CENTRE OF EXCELLENCE FOR MATHEMATICAL AND STATISTICAL FRONTIERS

Case Studies

Healthy People

Sustainable Environments

Prosperous Societies

Big Data Analytics, Big Models, New Insights

QUT

ACEMS
AUSTRALIAN RESEARCH COUNCIL CENTRE OF EXCELLENCE FOR MATHEMATICAL AND STATISTICAL FRONTIERS

1. Healthy People

An unparalleled opportunity for medicine

Evidence base (overall)

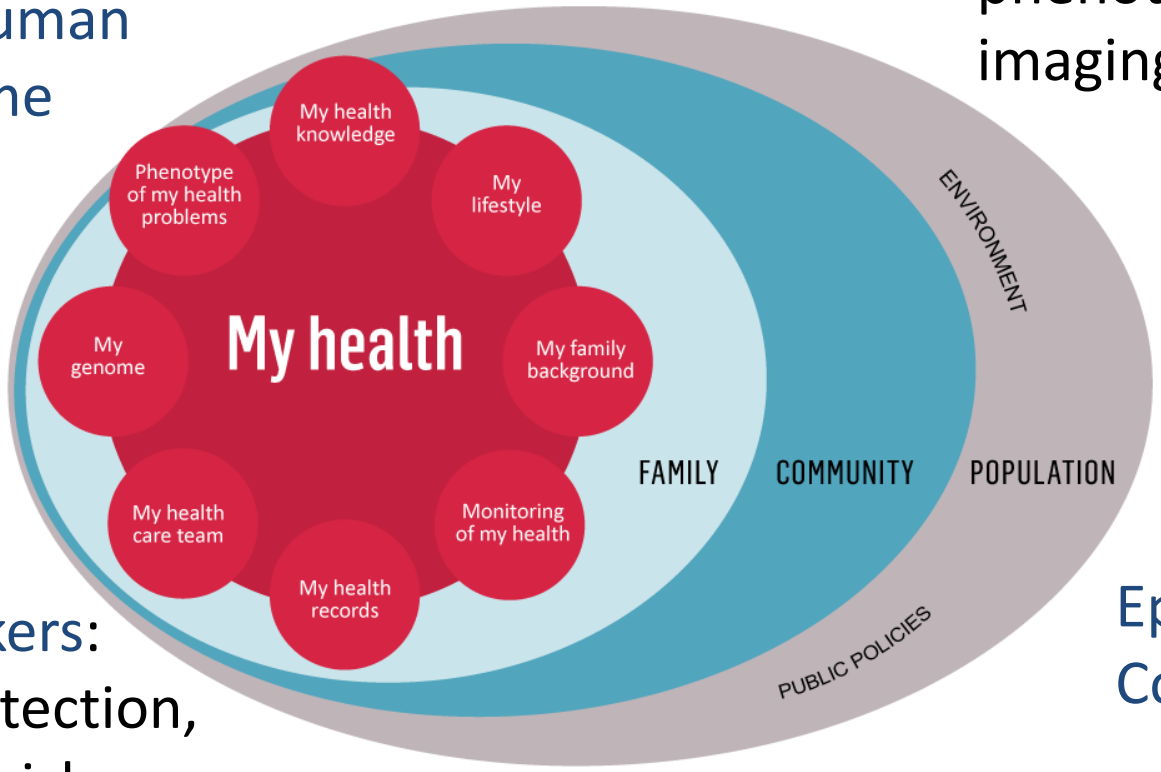
+

Individual patients (personalised)

Big Data?

Multiplex approach:
"omics" technologies
phenotype studies,
imaging, in-vivo studies

The Human Genome



Biomarkers:
early detection,
disease risk,
disease evolution,
response and toxicity
to a given treatment

Epidemiology
Community Health

The Human Microbiome:
Ecology + health + medicine

QUT

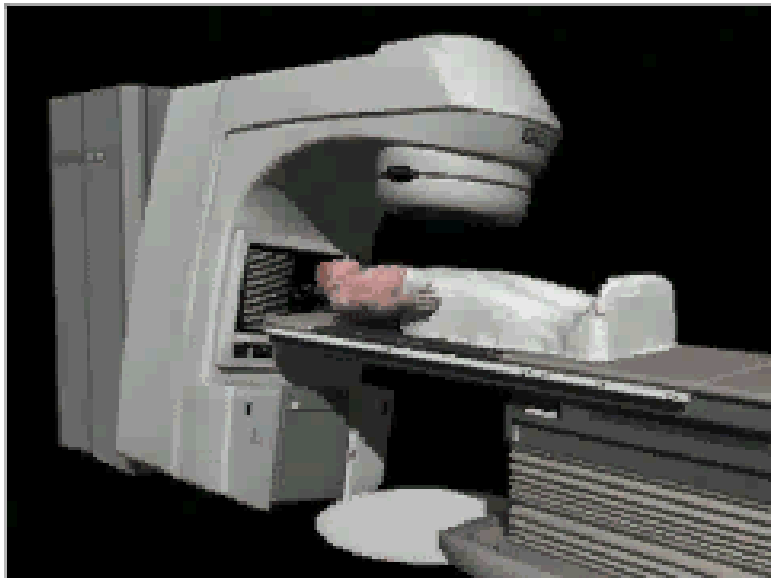
ACEMS
AUSTRALIAN RESEARCH COUNCIL CENTRE OF EXCELLENCE FOR MATHEMATICAL AND STATISTICAL FRONTIERS

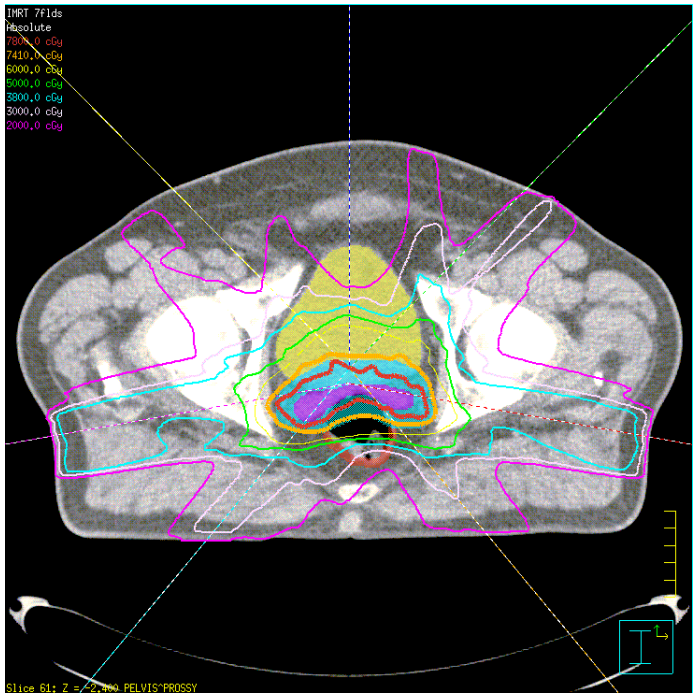
Bayesian statistics?

Imaging

Personalised Medicine

Deep Brain Stimulation for Parkinson's Disease



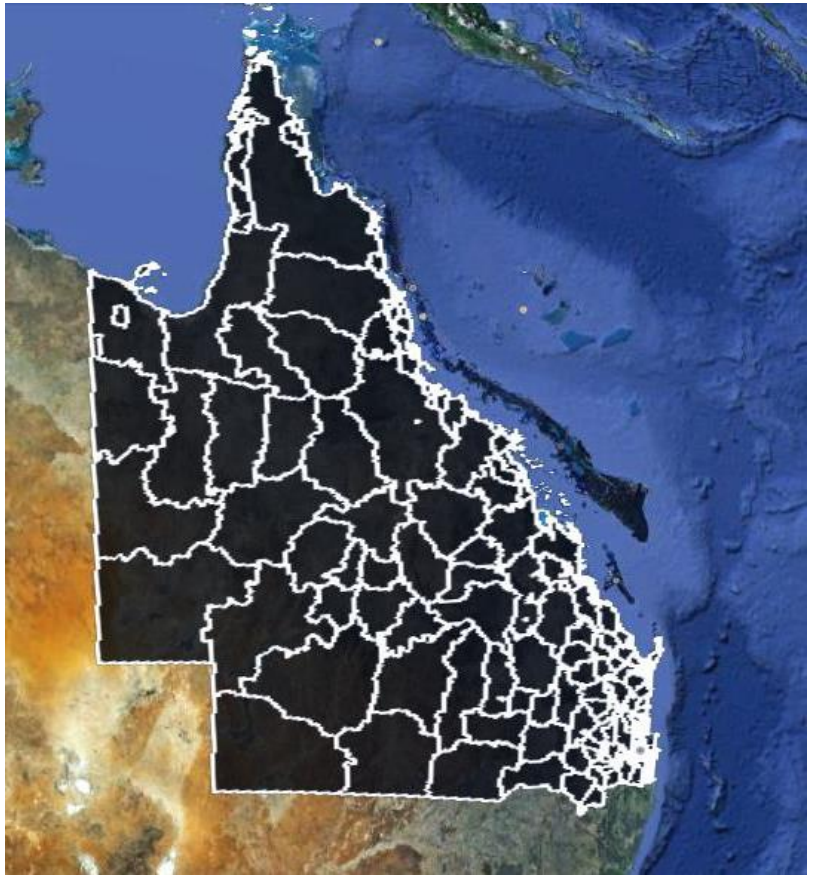


QUT

ACEMS

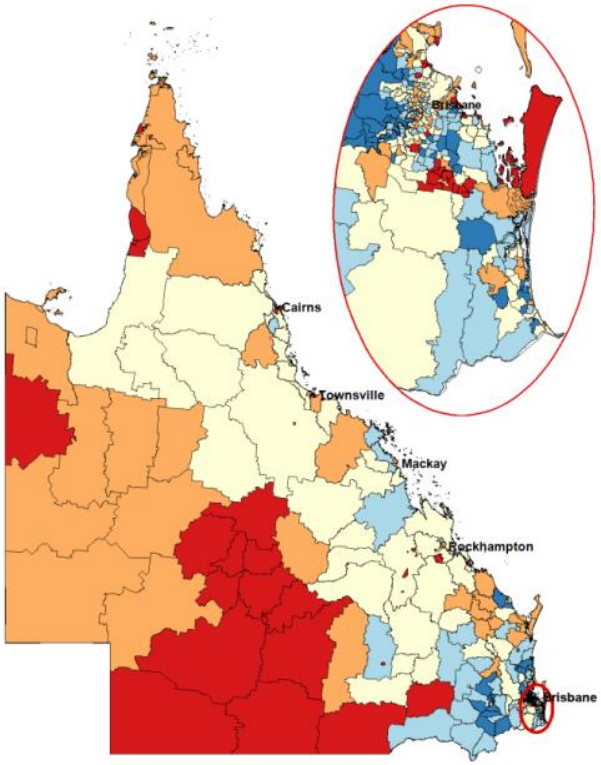
AUSTRALIAN RESEARCH COUNCIL CENTRE OF EXCELLENCE FOR MATHEMATICAL AND STATISTICAL FRONTIERS

Does “place” impact on cancer survival?



QUT

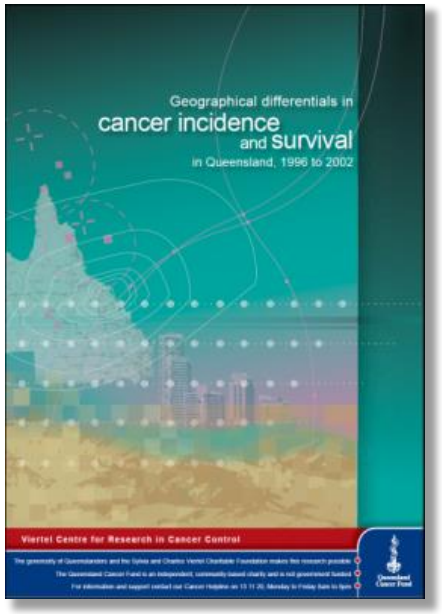
ACEMS
AUSTRALIAN RESEARCH COUNCIL CENTRE OF EXCELLENCE FOR
MATHEMATICAL AND STATISTICAL FRONTIERS



QUT

ACEMS
AUSTRALIAN RESEARCH COUNCIL CENTRE OF EXCELLENCE FOR
MATHEMATICAL AND STATISTICAL FRONTIERS

So what?



QUT

ACEMS
AUSTRALIAN RESEARCH COUNCIL CENTRE OF EXCELLENCE FOR
MATHEMATICAL AND STATISTICAL FRONTIERS

2. Prosperous Societies

"Smart Cities" = Supercomputing

2. Prosperous Societies

“Airports of the Future”

2. Prosperous Societies

Meteorology = Supercomputing

Australia's current Bureau of Meteorology 'Raijin' (Fujitsu-built) supercomputer

- 57,472 Intel 2.6GHz Xeon Sandy Bridge cores in 3596 compute nodes
- around 160TB main memory
- peak performance of 1.2 petaflops
- cost $50 million to build and $12 million a year to run

New: US$60M Cray

Based on the specifications of Raijin but:

- 16 times faster
- peak performance of 1.6 petaflops.

(Our first supercomputer was bought in 1998 and was less powerful than an iPhone 5.)

QUT

ACEMS
AUSTRALIAN RESEARCH COUNCIL CENTRE OF EXCELLENCE FOR MATHEMATICAL AND STATISTICAL FRONTIERS

Why?

Data requirements

- Current data: 1TB per day, expected to increase by 30 percent every 18 months to two years
- Other data: eg, Himwari 8 (new Japanese-owned satellite) a billion observation points per day + ground systems & radar systems

Models

- Stochastic (systems-based, data-based, probabilistic, almost-real-time)

Outputs

- Combine with a next-generation forecasting and warning system and upgraded satellites.
- Provide weather forecasts for a 6 km square region of choice to a mobile phone.
- Aim for 1km grid by 2020 for city models (currently 4km grid)
- Updates: 24 times daily by 2020 (currently 4 times daily)

QUT

ACEMS
AUSTRALIAN RESEARCH COUNCIL CENTRE OF EXCELLENCE FOR MATHEMATICAL AND STATISTICAL FRONTIERS

A 2020 Vision:
Linking Australia's weather to the World

QUT

ACEMS
AUSTRALIAN RESEARCH COUNCIL CENTRE OF EXCELLENCE FOR
MATHEMATICAL AND STATISTICAL FRONTIERS

Meteorology is not alone!

United Nations Global Working Group: Big Data in Official Statistics

"The statistical community has the obligation of exploring the use of new data sources, such as Big Data, to meet the expectation of the society for enhanced products and improved and more efficient ways of working."

8 task teams: "linking Big Data and the Sustainable Development Goals",…, "Mobile phone data", "Satellite imagery", "Social media data"

Members: Australia, Bangladesh, Cameroon, China, Colombia, Denmark, Egypt, Indonesia, Italy, Mexico, Morocco, Netherlands, Oman, Pakistan, Philippines, Tanzania, UAE, USA, UNSD,UNECE, UNESCAP, UN Global Pulse, ITU, OECD, World Bank, Eurostat, GCC-stat

http://unstats.un.org/unsd/bigdata/

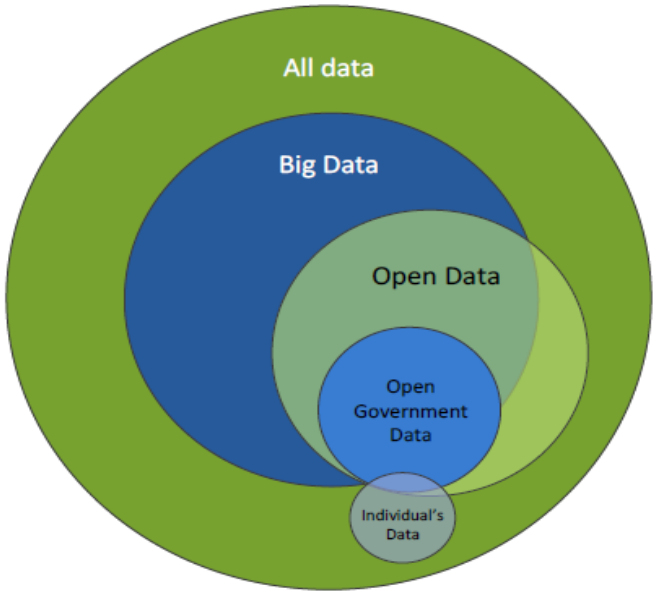
QUT

ACEMS
AUSTRALIAN RESEARCH COUNCIL CENTRE OF EXCELLENCE FOR MATHEMATICAL AND STATISTICAL FRONTIERS

From Big Data to Open Data

Australian Productivity Commission Issues Paper (April 2016): Data Availability and Use

- 2016: Australian Productivity Comm2014 Financial System Inquiry (the Murray Inquiry) – recommended review of the benefits and costs of increasing the availability and improving the use of data
- 2015 Harper Review of Competition Policy – recommended considering ways to improve individuals' ability to access their own data to inform consumer choices.



Global Institute (2013).

QUT

ACEMS
AUSTRALIAN RESEARCH COUNCIL CENTRE OF EXCELLENCE FOR MATHEMATICAL AND STATISTICAL FRONTIERS

3. Sustainable Environments

Crown of thorns starfish pest 'the greatest impact on Great Barrier Reef'
By Kirsty Nancarrow
Posted Fri 14 Jun 2013

Jan 29 2014
Concerned scientists demand rejection of permit for dumping dredge spoil in Great Barrier Reef's waters

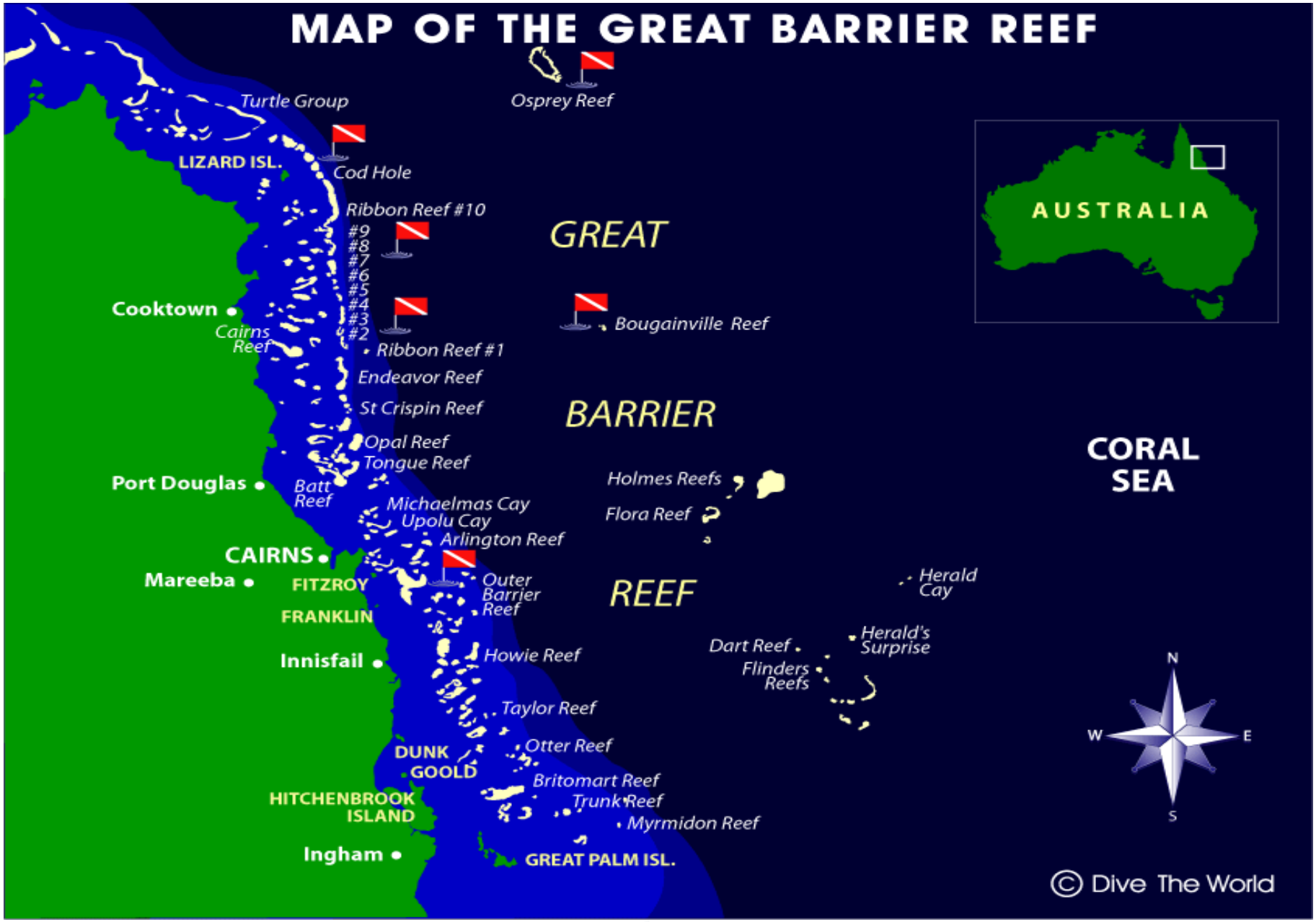
Cyclone damage devastating to the Great Barrier Reef

Sep 2 2013
Thousands gather for Rally for the Reef

The world's largest living structure
Almost 3000 reefs
Over 2,600 km

QUT

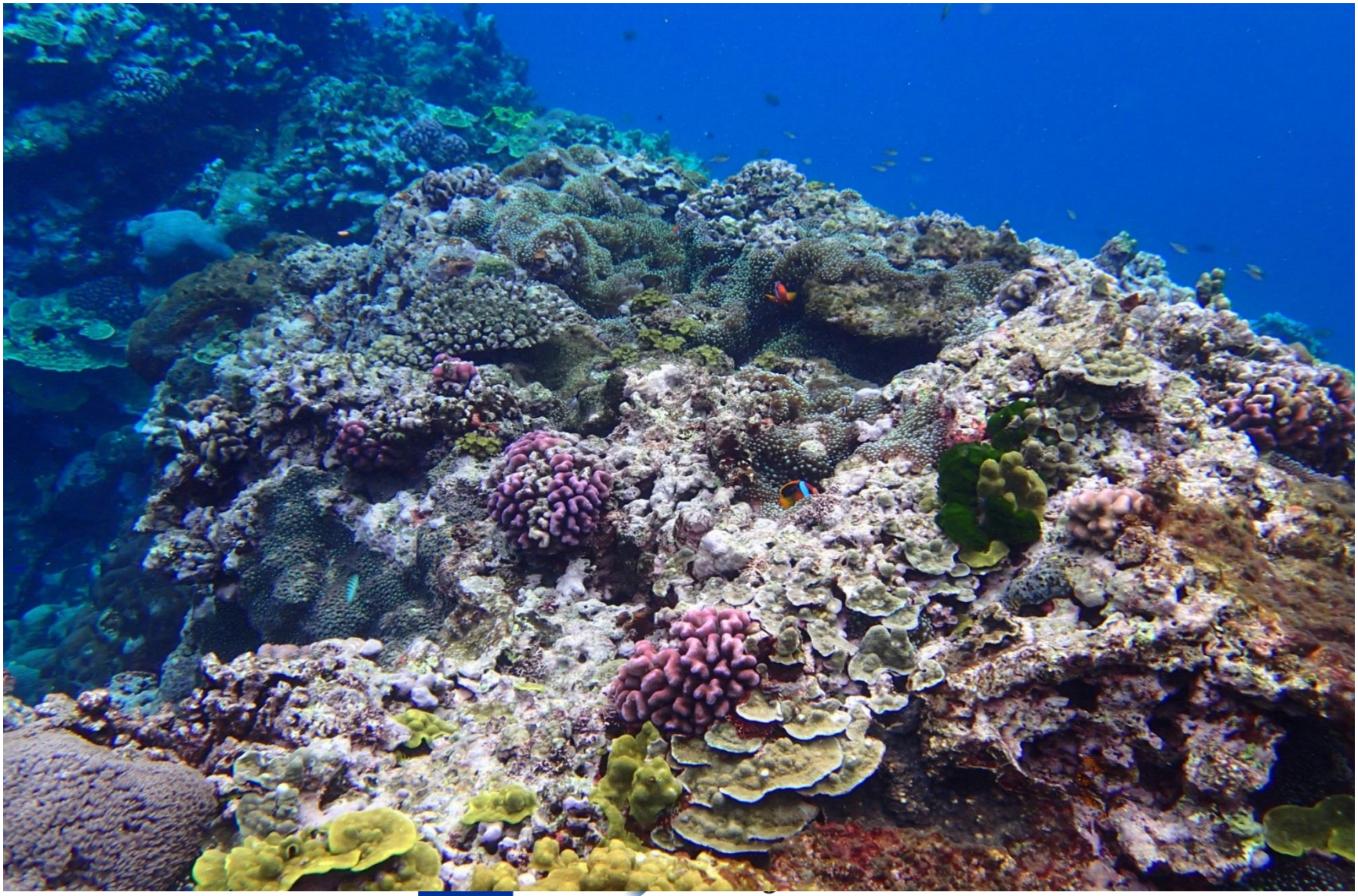
ACEMS
AUSTRALIAN RESEARCH COUNCIL CENTRE OF EXCELLENCE FOR MATHEMATICAL AND STATISTICAL FRONTIERS



QUT

ACEMS
AUSTRALIAN RESEARCH COUNCIL CENTRE OF EXCELLENCE FOR MATHEMATICAL AND STATISTICAL FRONTIERS

Data source 1: observations

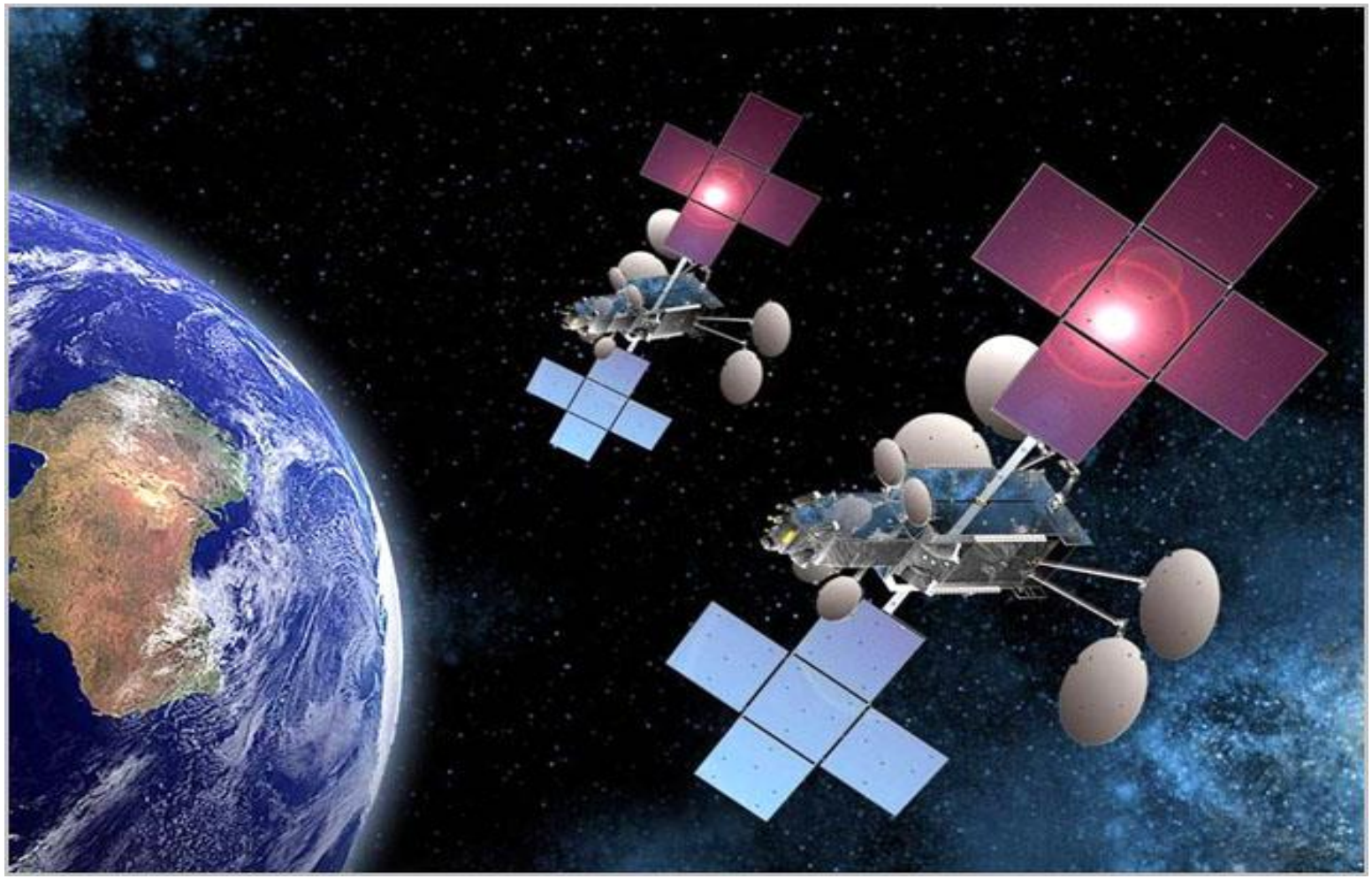


Data source 2: sensors

QUT

ACEMS
AUSTRALIAN RESEARCH COUNCIL CENTRE OF EXCELLENCE FOR
MATHEMATICAL AND STATISTICAL FRONTIERS

Data source 3: images



delimiter.com.au

QUT

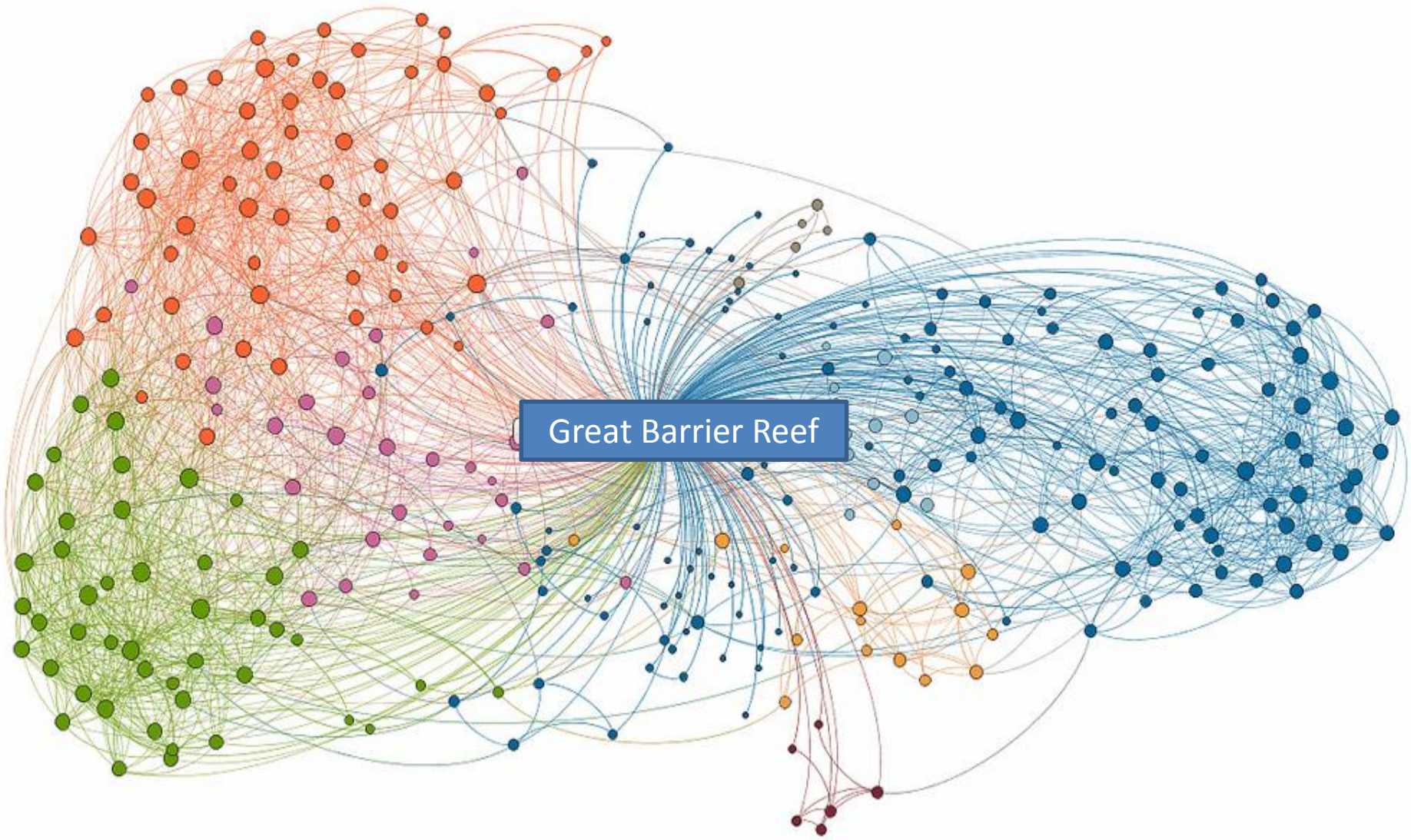
ACEMS
AUSTRALIAN RESEARCH COUNCIL CENTRE OF EXCELLENCE FOR
MATHEMATICAL AND STATISTICAL FRONTIERS

Data source 4: Internet

12,200,000 results (0.47 seconds)
Search Results
Media Room - Great Barrier Reef Marine Park Authority
www.gbrmpa.gov.au/media-room
by GBMP Authority - 2011
Media releases and announcements. Read more on Latest ... Reef in Brief is the Great Barrier Reef Marine Park Authority's regular e-newsletter. ... Social media.
Tourism on the Great Barrier Reef - GBRMPA
www.gbrmpa.gov.au › Managing the Reef › How the Reef is managed
by GBMP Authority - 2011 - Cited by 2 - Related articles
Commercial marine tourism extends throughout the Great Barrier Reef,and ... The Great Barrier Reef Marine Park Authority (GBRMPA) and the ... Social media.
Great Barrier Reef: Queensland Government launches ...
www.abc.net.au/news/2015-07-02/queensland...reef.../6588712
Jul 1, 2015 - She launched a new social media tourism campaign in Cairns, ... "The Great Barrier Reef is ours to protect and share," Ms Palaszczuk said.
Fight For The Reef - Essential Media Communications
www.essentialmedia.com.au/case_studies/fight-for-the-reef/
The Great Barrier Reef, one of the natural wonders of the world, is under threat ... paid advertising, digital and social media including supporter journey plans to ...

QUT

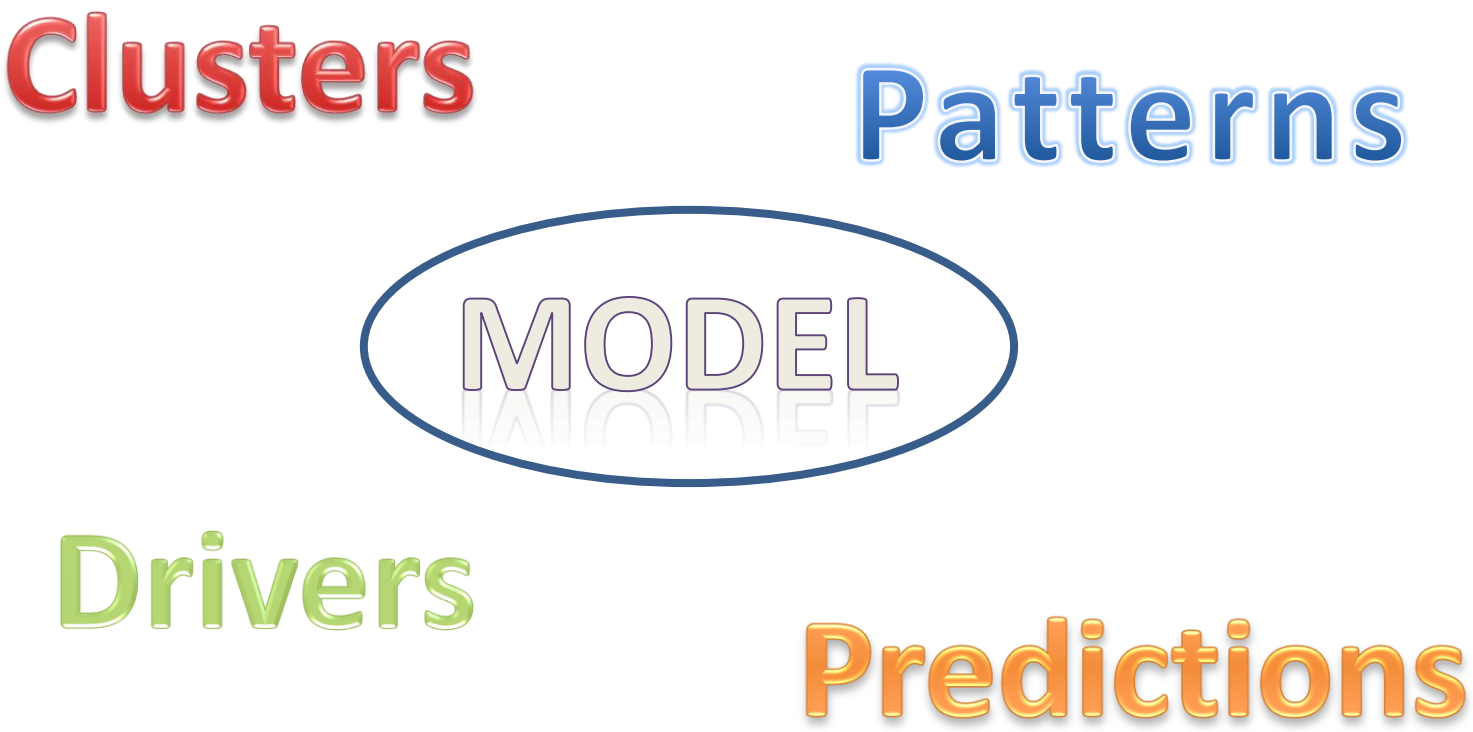
ACEMS
AUSTRALIAN RESEARCH COUNCIL CENTRE OF EXCELLENCE FOR MATHEMATICAL AND STATISTICAL FRONTIERS



QUT

ACEMS
AUSTRALIAN RESEARCH COUNCIL CENTRE OF EXCELLENCE FOR
MATHEMATICAL AND STATISTICAL FRONTIERS

Bayesian statistics and big data



QUT

ACEMS
AUSTRALIAN RESEARCH COUNCIL CENTRE OF EXCELLENCE FOR
MATHEMATICAL AND STATISTICAL FRONTIERS

Hierarchical Bayesian model

- **Data Model**: Pr(data | process and data parameters)
  - How data are observed given underlying biological and environmental processes
- **Process Model**: Pr(process | process parameters)
  - Potential processes given inputs regarding biology / ecology
- **Parameter Model**: Pr(data and process parameters)
  - Prior distribution to describe uncertainty in detectability, exposure, growth …
- **The posterior distribution of the process (and parameters) is related to the prior distribution and data by:**

Pr(*process, parameters* | *data*) $\propto$
Pr(*data* | *process, parameters*) Pr(*process* | *parameters*) Pr(*parameters)*

QUT

ACEMS
AUSTRALIAN RESEARCH COUNCIL CENTRE OF EXCELLENCE FOR MATHEMATICAL AND STATISTICAL FRONTIERS

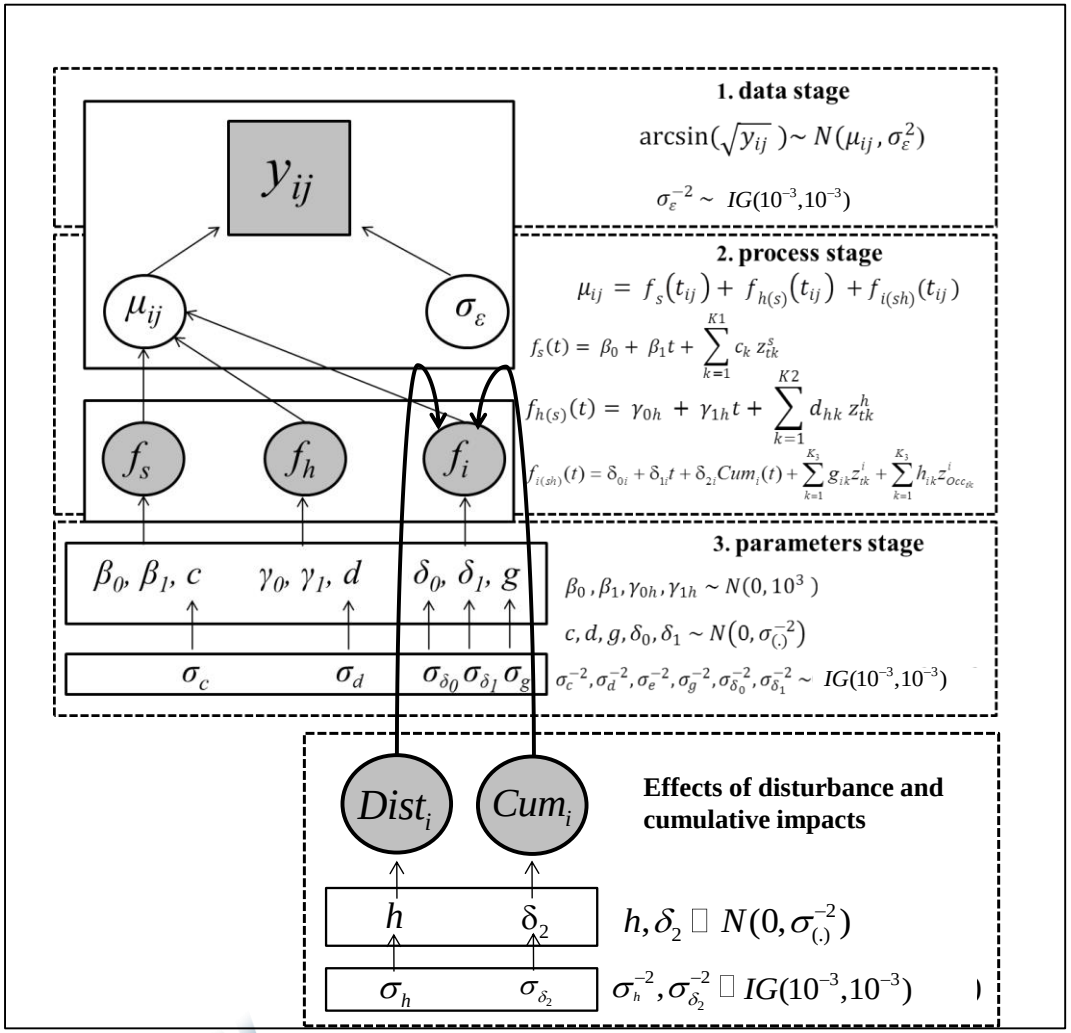
Hierarchical modelling

Data

Process

Parameters

Drivers



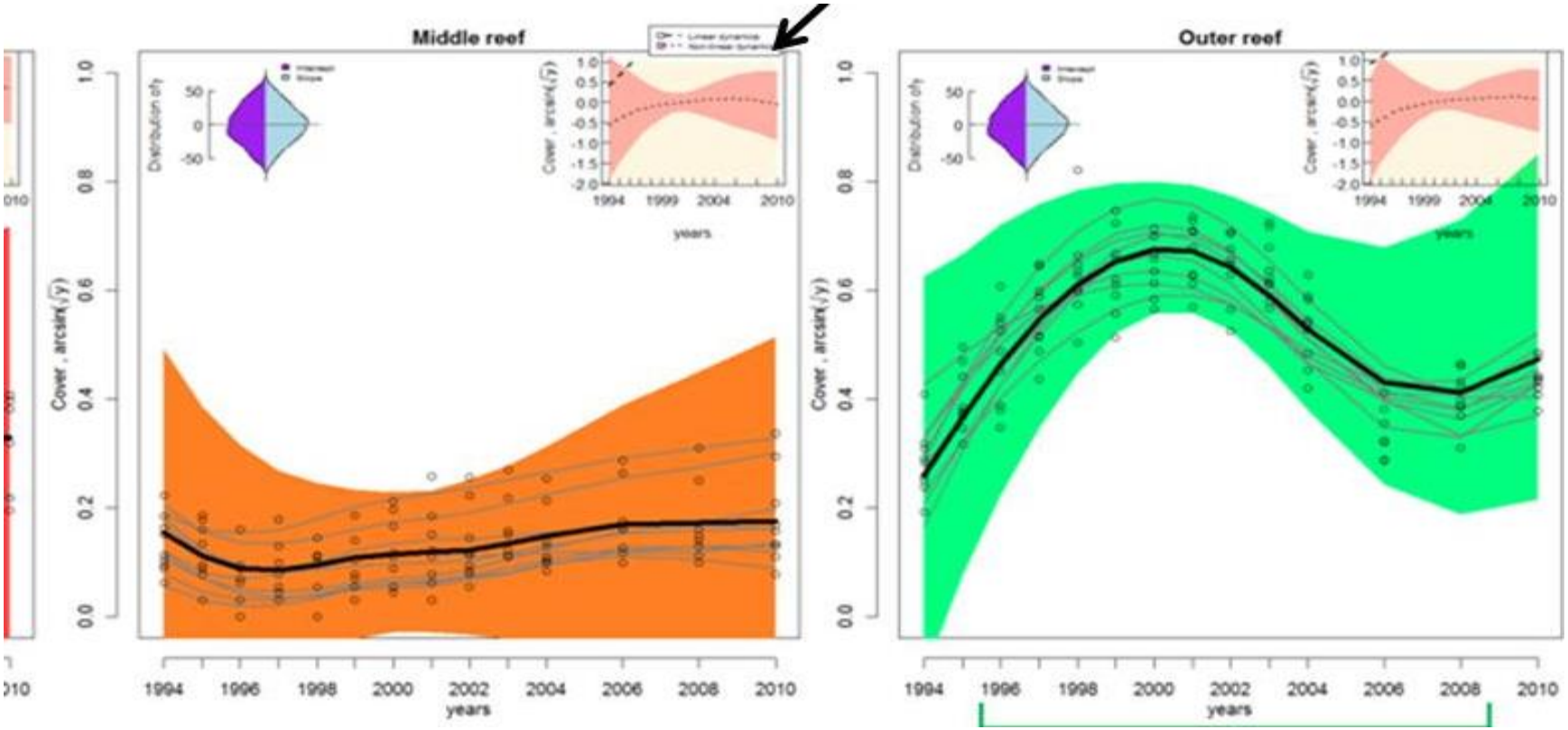
QUT

ACEMS
AUSTRALIAN RESEARCH COUNCIL CENTRE OF EXCELLENCE FOR MATHEMATICAL AND STATISTICAL FRONTIERS

What are we learning?

Inner reef

Outer reef



QUT

ACEMS
AUSTRALIAN RESEARCH COUNCIL CENTRE OF EXCELLENCE FOR
MATHEMATICAL AND STATISTICAL FRONTIERS

QUT

ACEMS

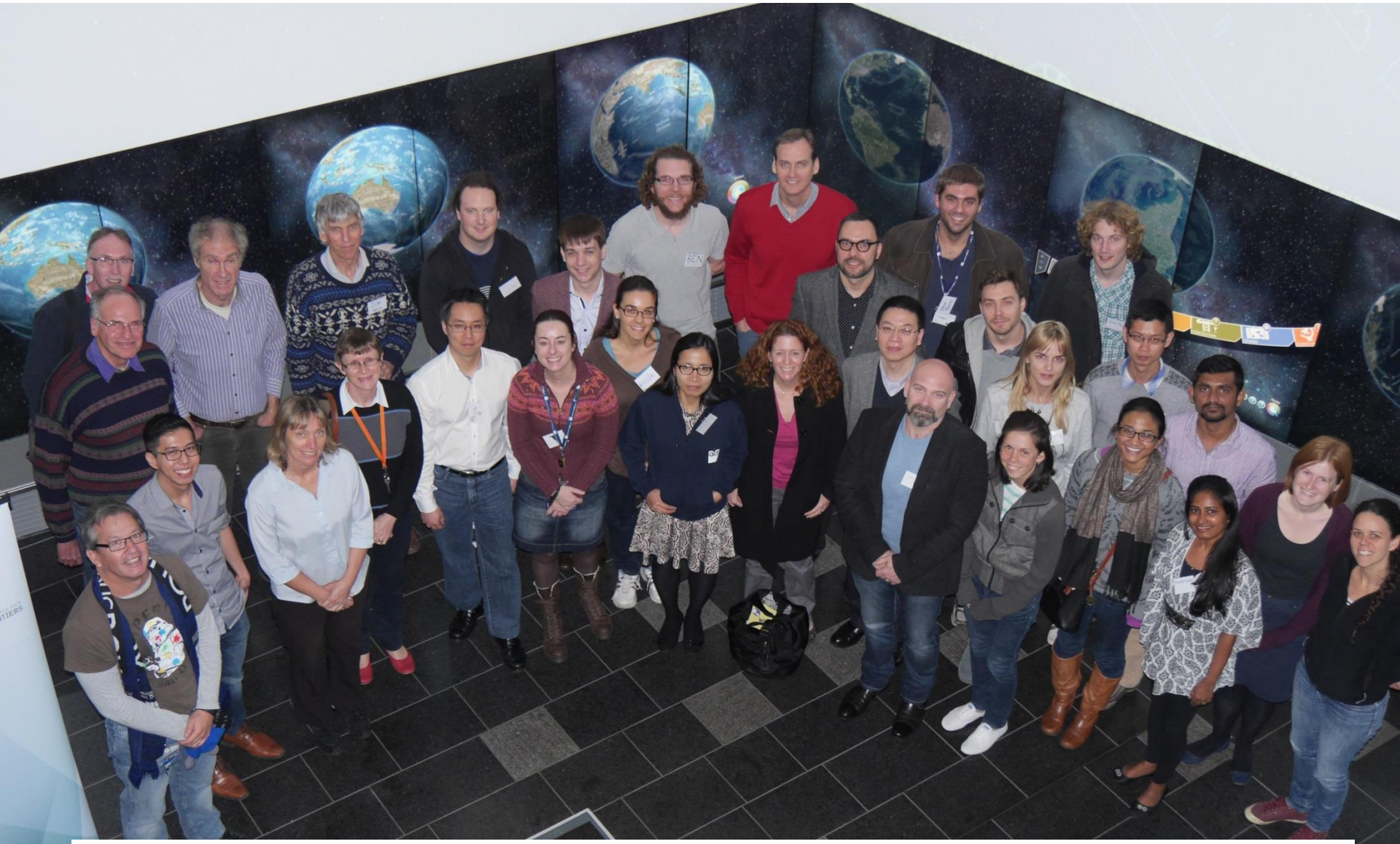
AUSTRALIAN RESEARCH COUNCIL CENTRE OF EXCELLENCE FOR
MATHEMATICAL AND STATISTICAL FRONTIERS

Recap

1. Let's talk about "Big Data"
2. Methods for modelling and analysis of big data
3. Digging deeper: (1) classification
4. Digging deeper: (2) regression
5. Digging deeper: (3) clustering
6. Digging deeper: (4) dimension reduction
7. Case study: recommender systems
8. From the learning to the doing: tips and tricks

QUT

ACEMS
AUSTRALIAN RESEARCH COUNCIL CENTRE OF EXCELLENCE FOR MATHEMATICAL AND STATISTICAL FRONTIERS



With acknowledgements to our awesome QUT & ACEMS teams and collaborators!