How to be Bayesian in the Big Data Era

Kerrie Mengersen
Professor of Statistics, QUT
ARC Centre of Excellence in Mathematical and Statistical Frontiers: Big Data, Big Models, New Insights

In collaboration with Antonietta Mira
Professor of Statistics, USI

QUT

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

Introductions

Bayesian Research and Applications Group

↓

Faculty of Science and Engineering

↓

Institute for Future Environments

↓

Queensland University of Technology

↓

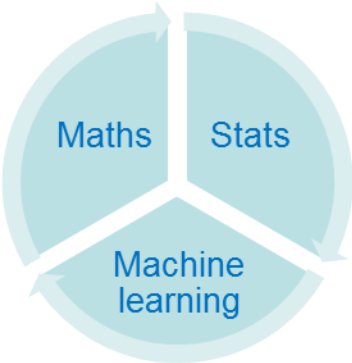
Centre of Excellence in Mathematical and Statistical Frontiers:
Big Data, Big Models, New Insights

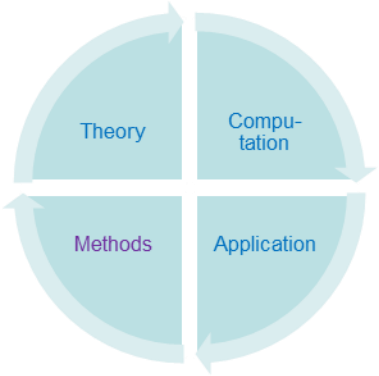
↓

Australia

QUT

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





Australian Research Council Centre of Excellence
Mathematical & Statistical Frontiers:
Big Data, Big Models, New Insights

7 year horizon

6 Universities

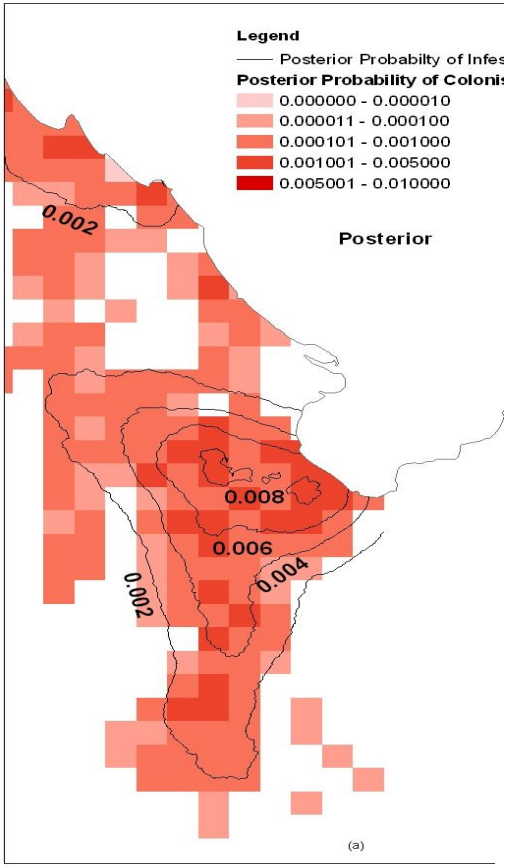
7 Partner Organisations

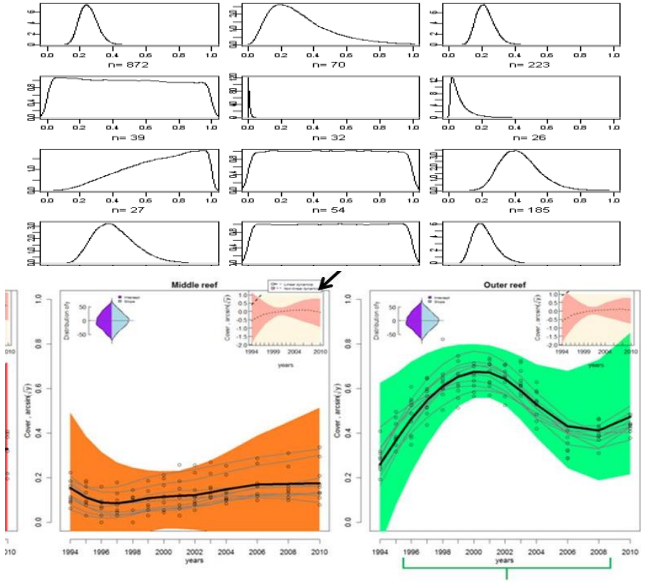
50 Professorial Investigators, >50 Researchers&students

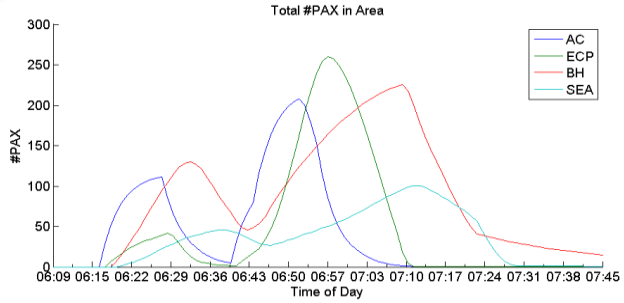
QUT

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

Our Interests at QUT: Modelling and Analysis (Big Data, Bayesian Approaches)





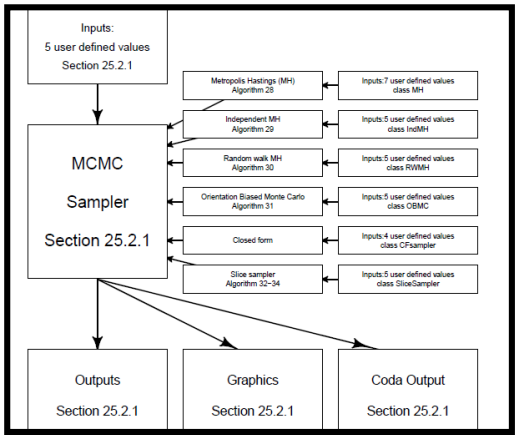


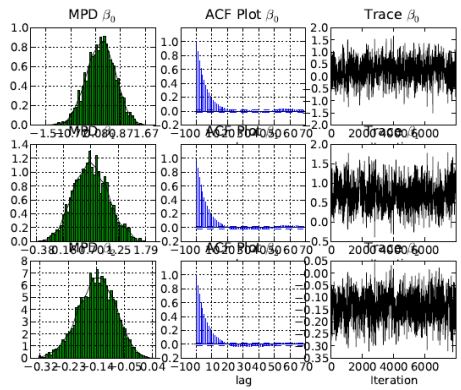
- Combining data sources
- Modelling with uncertainty
- Using prior information
- Probabilistic prediction
- Risk stratification
- Complex systems models

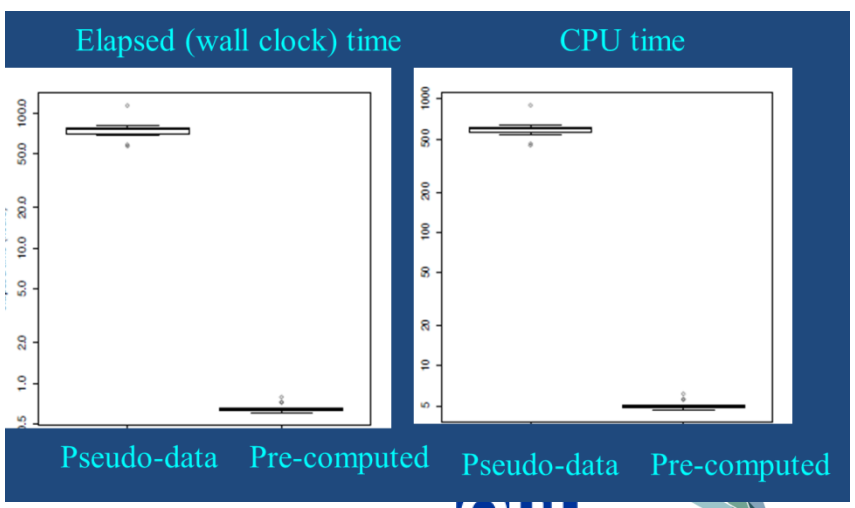
QUT

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

Our Interests at QUT: Fast computation





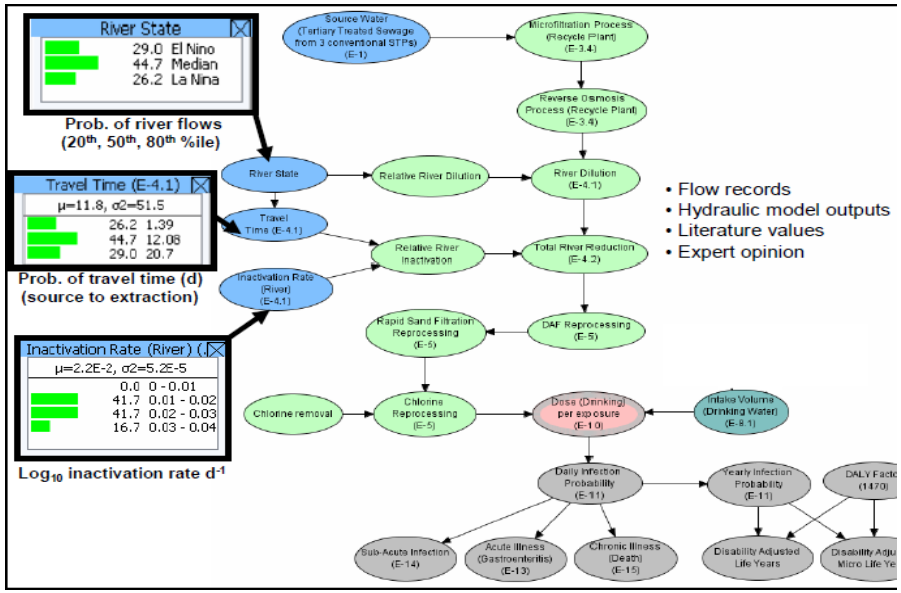


- Algorithms
- Approximations
- Ensembles
- Big data environments
- Platforms
- HPC

QUT

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

Our Interests at QUT: Systems approaches



Indicator	Category	Farm	Factory	Market	Rating
Economic	Commodity prices	0.8	0.6	0.6	0.7
	Legal and administrative environment	0.0	0.1	0.8	0.3
	Access to capital and labour	0.6	0.8	0.8	0.7
	Profitability	0.3	0.8	0.7	0.6
	Workforce capabiility	0.1	0.8	0.6	0.5
	Economic sustainability rating	0.4	0.6	0.7	0.6
Social	Lifestyle and community	0.0	0.5	0.1	0.3
	Health and well being	0.6	0.9	0.8	0.7
	Value and contribution	0.1	0.6		0.6
	Product, safety and production	0.8	0.0	0.0	0.1
	Social relevance	0.6	0.8	0.8	0.7
	Social sustainability rating	0.4	0.6	0.3	0.5
Environment	Energy, effluent and water	0.6	0.2	0.2	0.3
	Materials, suppliers and transport	0.2	0.8		0.6
	Products and services	0.8	0.8	0.2	0.6
	Biodiversity	0.2	0.8	0.6	0.6
	Compliance	0.2	0.0	0.0	0.1
	Environment sustainability rating	0.4	0.5	0.2	0.4
Dairy Industry sustainability rating		0.4	0.6	0.4	0.5

- *G*raphical, probabilistic modelling of complex processes
- Multiple perspectives and stakeholders
- Varied information sources

QUT

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

Why a statistician?

David Dunson:

I would describe statistics as the science of variability, meaning that the main goal of statistics is to develop methods and algorithms for the mathematical exploration, elicitation and control of variability, and the uncertainty it generates. Inference and uncertainty quantification are at the core of statistics and they have generated correlated siblings like prediction, testing, controlling for dependence, confounding, randomization.

Model-driven approach

Data

Data-driven approach

QUT

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

“How to be Bayesian…”

Bayes’ Theorem:

$$Pr(\theta|y) = Pr(y|\theta)\, P(\theta) \,/\, Pr(y)$$

y: knowns – data,

θ: unknowns – model parameters, models, missing data, predictions, latent variables, etc

QUT

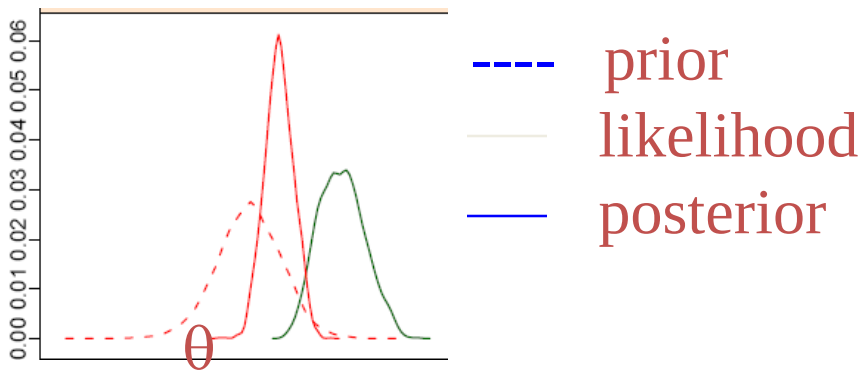
ACEMS
AUSTRALIAN RESEARCH COUNCIL CENTRE OF EXCELLENCE FOR
MATHEMATICAL AND STATISTICAL FRONTIERS

Bayesian modelling

Prior (Existing understanding) → Data (Likelihood) (Model) → Posterior (Updated understanding)

$p(\theta)$ $\quad$ $p(Y|\theta)$ $\quad$ $p(\theta|Y)$

$$P(\theta|Y) = P(Y|\theta)\, P(\theta) / P(Y)$$



QUT

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

Frequentist versus Bayesian

Frequentist: estimate θ through the likelihood: $p(Y|\theta)$

How likely is Y for given values of θ ?

Use moment estimators or maximum likelihood.

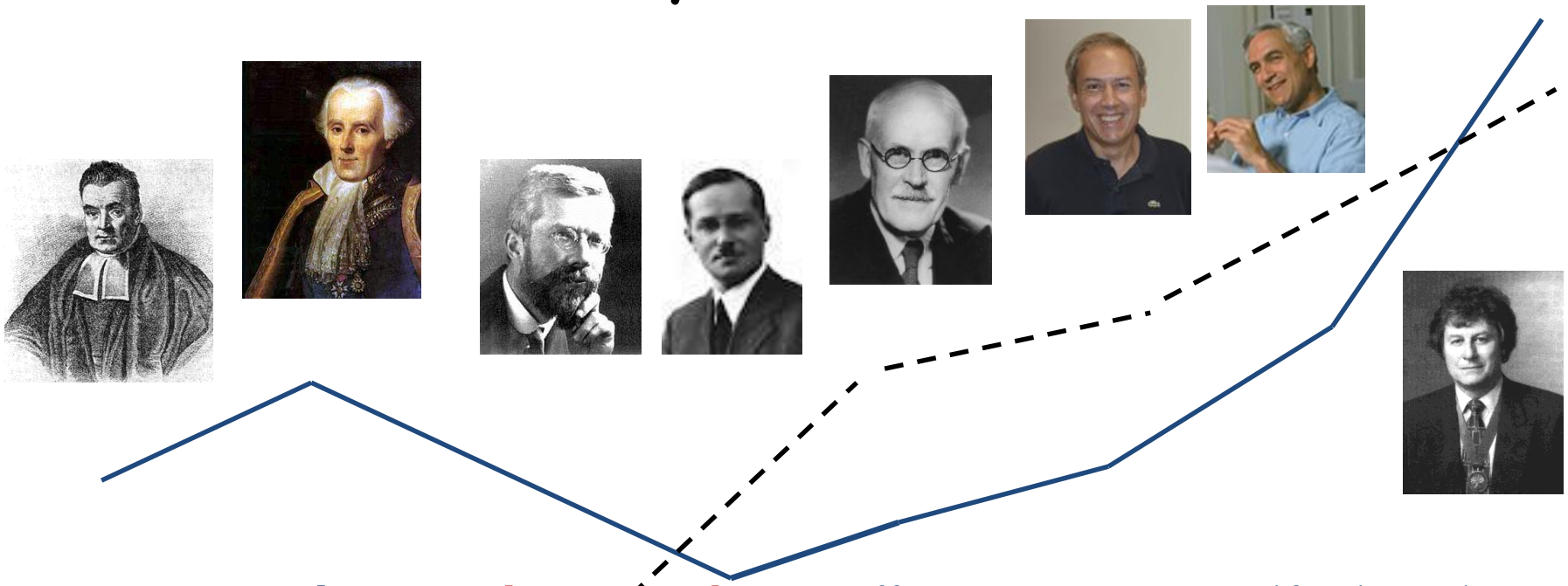
Bayesian: estimate θ through the posterior: $p(\theta|Y)$

Obtain whole distribution for θ given the data

QUT

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

Modern Bayesian Statistics



Bayes	Laplace	Boole Venn	Fisher Neyman	Jeffreys	Geman Geman	Gelfand Smith	Today's Bayesians
1763	1812	1838	1930's	1950's	1980's	1990's	2000's

"Probability Theory" | " Inverse Probability" | "Bayesian Analysis"

QUT

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

Why Bayes?

"A model-based revolution"

- Sir Adrian Smith, DG Knowledge & Innovation

Bayesian methods allow us to:

- Think differently about estimating and interpreting unknown parameters

  "what are possible values of this parameter?"
- Combine prior information with the data

  "what else do I know about this parameter and model?"
- Describe many sources of uncertainty in the model

  "how sure am I about the inputs and outputs of my model?"
- Describe complex systems using hierarchical or multi-level models

QUT

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

Why Bayes?

Bayesian computational methods (such as MCMC and approximations) allow us to:

- Use non-standard distributions
- Fit very complex models
- Obtain a very wide variety of estimates
- Make a very wide range of inferences, based directly on posterior probabilities

QUT

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

Bayes in practice

Statistical Science, 2014, Volume 29 (1)
https://projecteuclid.org/euclid.ss/1399645721

- Estimating the Distribution of Dietary Consumption Patterns
  Raymond J. Carroll; 2 - 8
- Estimation of HIV Burden through Bayesian Evidence Synthesis
  Daniela De Angelis, Anne M. Presanis, Stefano Conti, and A. E. Ades; 9 - 17
- Bayesian Estimation of Population-Level Trends in Measures of Health Status
  Mariel M. Finucane, Christopher J. Paciorek, Goodarz Danaei, and Majid Ezzati; 18 - 25
- How Bayesian Analysis Cracked the Red-State, Blue-State Problem
  Andrew Gelman; 26 - 35
- From Science to Management: Using Bayesian Networks to Learn about *Lyngbya*
  Sandra Johnson, Eva Abal, Kathleen Ahern, and Grant Hamilton; 36 - 41
- Experiences in Bayesian Inference in Baltic Salmon Management
  Sakari Kuikka, Jarno Vanhatalo, Henni Pulkkinen, Samu Mäntyniemi, and Jukka Corander; 42 - 49
- Finding the Most Distant Quasars Using Bayesian Selection Methods
  Daniel Mortlock; 50 - 57
- Bayesian Population Projections for the United Nations
  Adrian E. Raftery, Leontine Alkema, and Patrick Gerland; 58 - 68
- Search for the Wreckage of Air France Flight AF 447
  Lawrence D. Stone, Colleen M. Keller, Thomas M. Kratzke, and Johan P. Strumpfer; 69 - 80
- Galaxy Formation: Bayesian History Matching for the Observable Universe
  Ian Vernon, Michael Goldstein, and Richard Bower; 81 - 90

QUT

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

“…In the Big Data Era”

What *is* “big data”?

“inconveniently large”

https://www.youtube.com/watch?v=PI7SLOovO5c

QUT

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

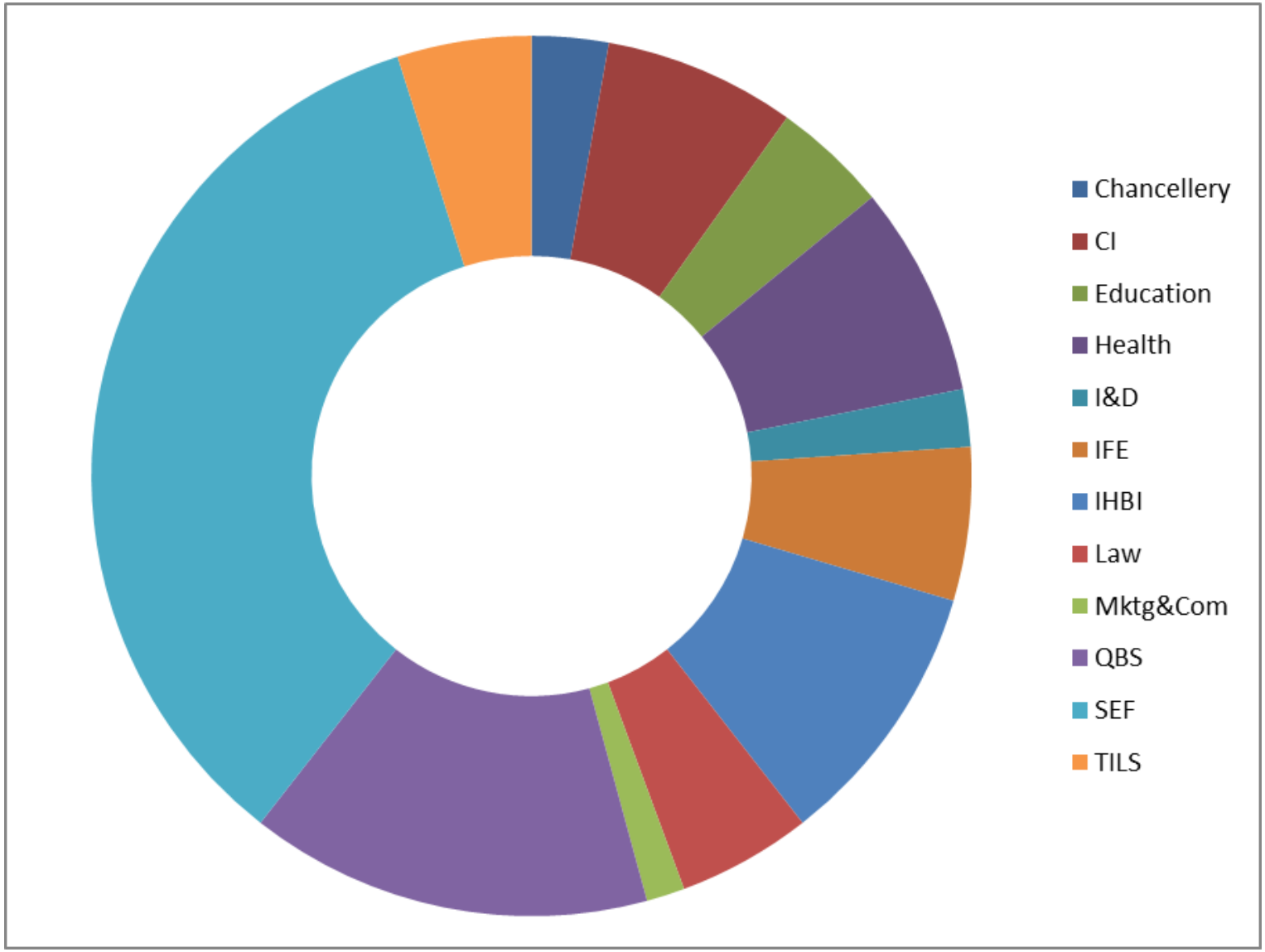
Tableau Top Trends in 2015

1. The Cloud: revenues for top 50 public cloud providers increased by 47% in Q4 of 2013 to $6.2Billion. Amazon, Google, Terradata
2. Personal data cleansing tools: free up ~80% analyst time
3. New SQL tools for scale, flexibility, speed
4. Hadoop, massively parallel processing
5. "big data lake"
6. Big data ecosystem: new ways to store data, clean it, add content, bring in social media, analyse it, visualise it
7. Internet of Things: car sends data back to manufacturer – Ford, GE, Rolls Royce

QUT

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

“Big Data” at QUT



Supercomputing at QUT: Flashlite

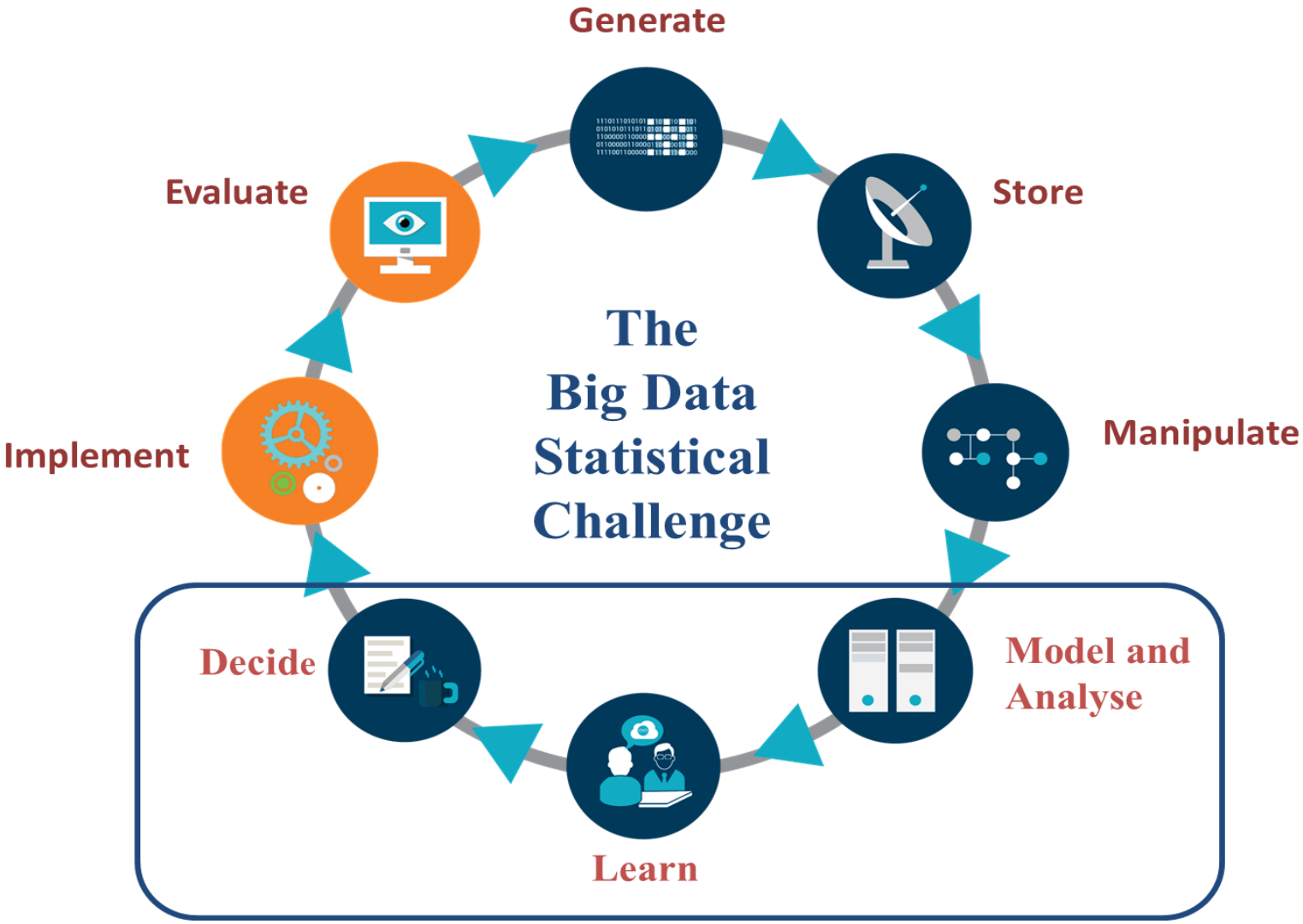
"Maximises Input Output Operations Per Second (IOPS) while achieving competitive FLOPS ratings and high performance networking; uses ScaleMP's vSMP software, that provide seamless access to data regardless of location"

- 68 x compute nodes
- dual rail 56Gbps Mellanox infiniband fabric
- non-blocking within groups of 24 nodes
- 2:1 blocking factor between groups of 24 nodes
- ScaleMP vSMP software that aggregates multiple nodes into "super nodes" with larger memory/CPU/disk/IO than the individual nodes
- ROCKS cluster management software
- Torque + Maui batch system
- 150+ TB of high performance storage connected via NFS into IB fabric.

QUT

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

What is the role of Statisticians in the Big Data Era?



What is the role of *Bayesian* Statistics in the Big Data Era?

"A model-based revolution"

1. Principled analyses:
   decision-focused, model-based
2. Multi-scale modelling:
   local/process/system, known/unknowns
3. Priors:
   updating, regularisation, merging information
4. Rich distribution-based outputs:
   estimates, predictions, probabilities, uncertainty
5. Model robustness: "shake your model"

QUT

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

What is the role of *Bayesian* Statistics in the Big Data Era?

"A computational revolution"

1. Merger of computational maths, stats, machine learning
2. Expectation, not optimisation
3. Computational platforms:
   Hadoop, Spark
4. Software:
   R, Python, Spark R, Tessera, H2O
5. Algorithms:
   MCMC, approximations (INLA, VB, ABC, GP, etc)

QUT

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

Example (David Dunson)

Hybrid high-dimensional density estimation

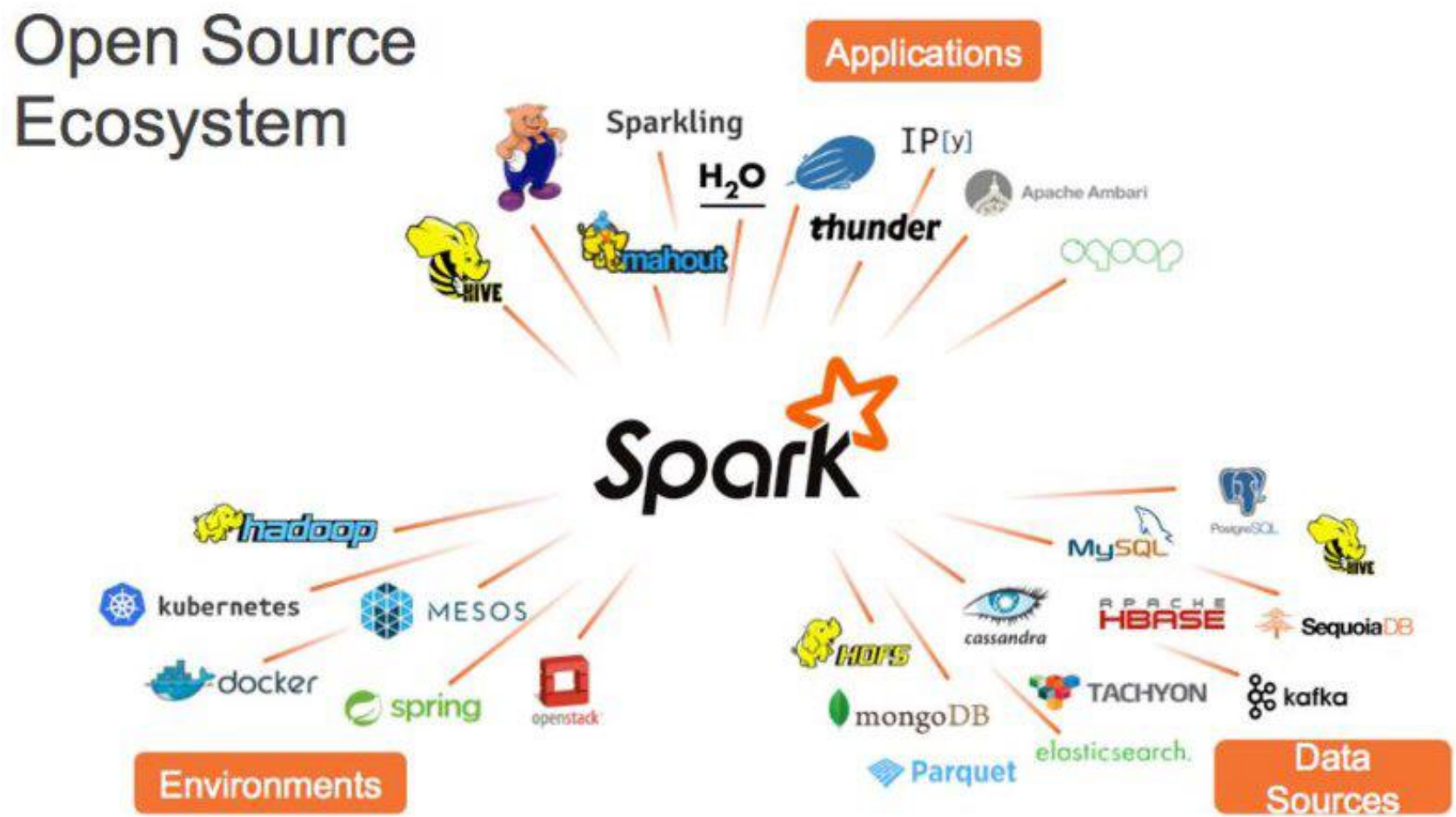
Ye, Canale & Dunson (2016, AISTATS)

- $y_i = (y_{i1}, \ldots, y_{ip})^T \sim f$ with p large & f an unknown density
- Potentially use Dirichlet process mixtures of factor models
- Approach doesn't scale well at all with p
- Instead use hybrid of Gibbs sampling & fast multiscale SVD
- Scalable, excellent mixing & empirical/predictive performance

QUT

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

Big Data Analytics Ecosystem



QUT

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

Technologies

OpenSceneGraph

Neo4j

tableau SOFTWARE

HTML

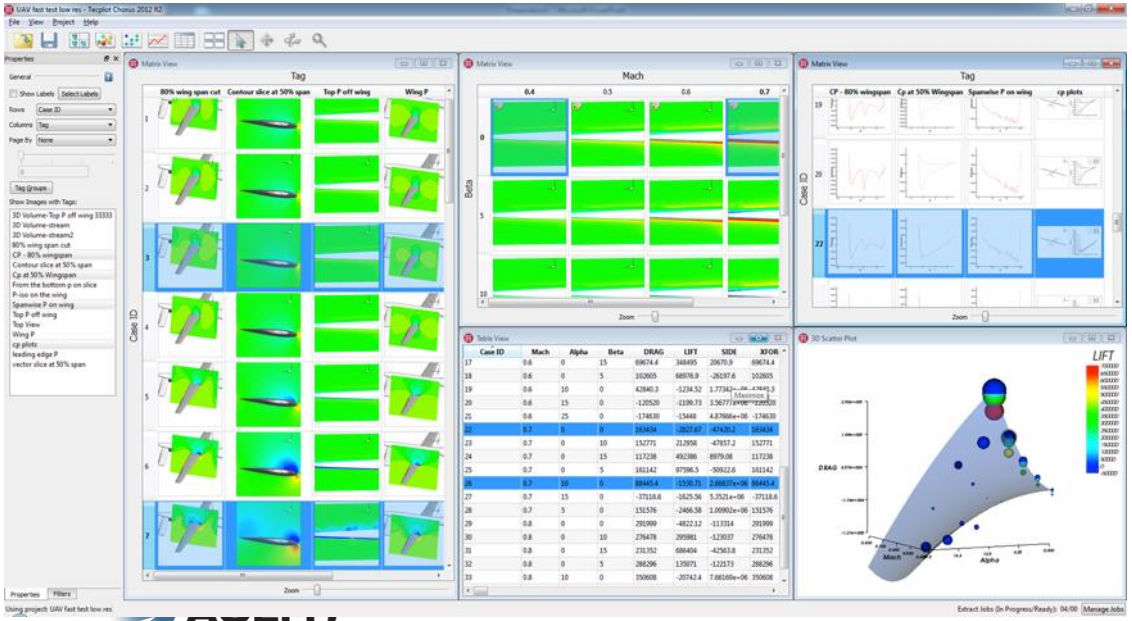
WebGL

tecplot Master the View

SVG HTML CSS JavaScript

D3.JS - Data Driven Documents

SAS® VISUAL ANALYTICS

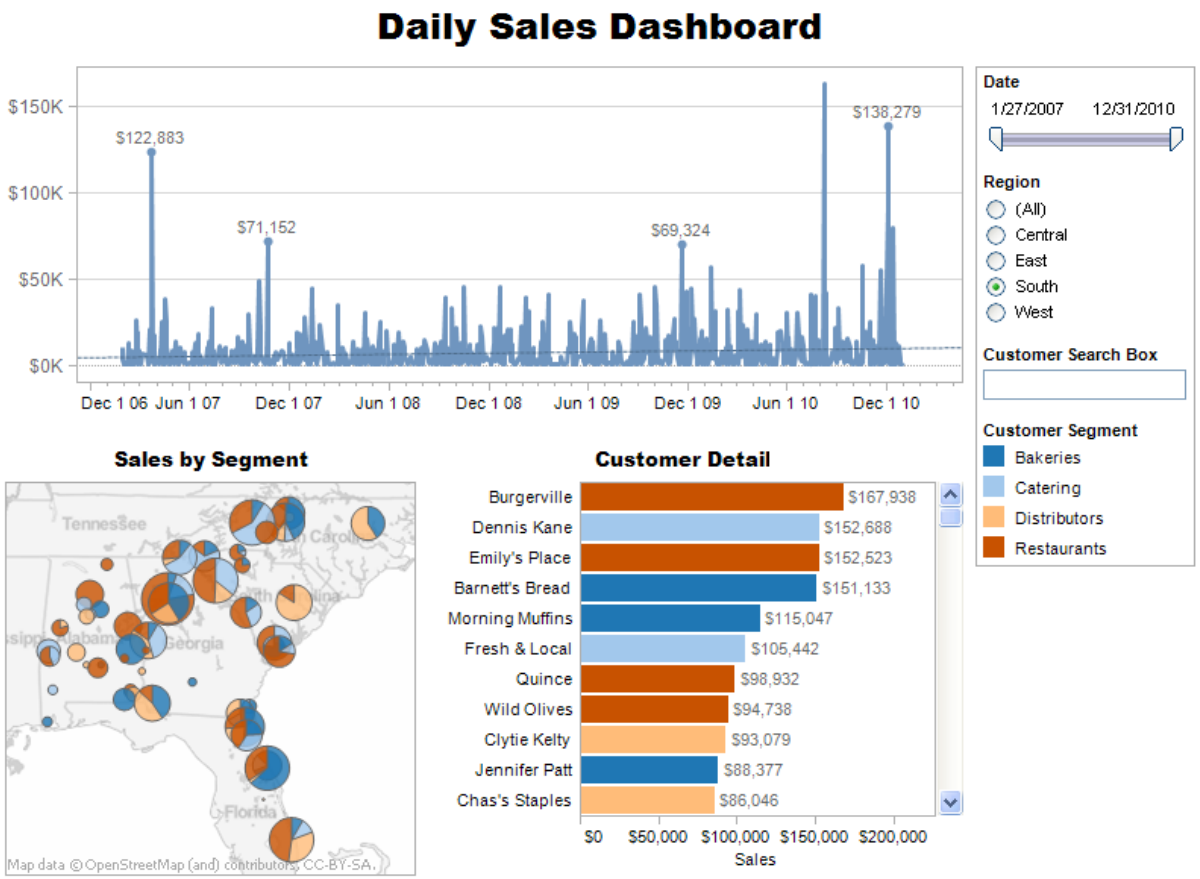


QUT

AUSTRALIAN RESEARCH COUNCIL CENTRE OF EXCELLENCE FOR MATHEMATICAL AND STATISTICAL FRONTIERS

Example: Tableau and Big Data

- Hadoop
- Cloudera Impala
- Cassandra
- HortonWorks
- Karmasphere
- Google Analytics
- Google BigQuery
- Teradata
- Amazon RedShift
- SAP HANA
- MySQL, PostgreSQL
- Many other data source

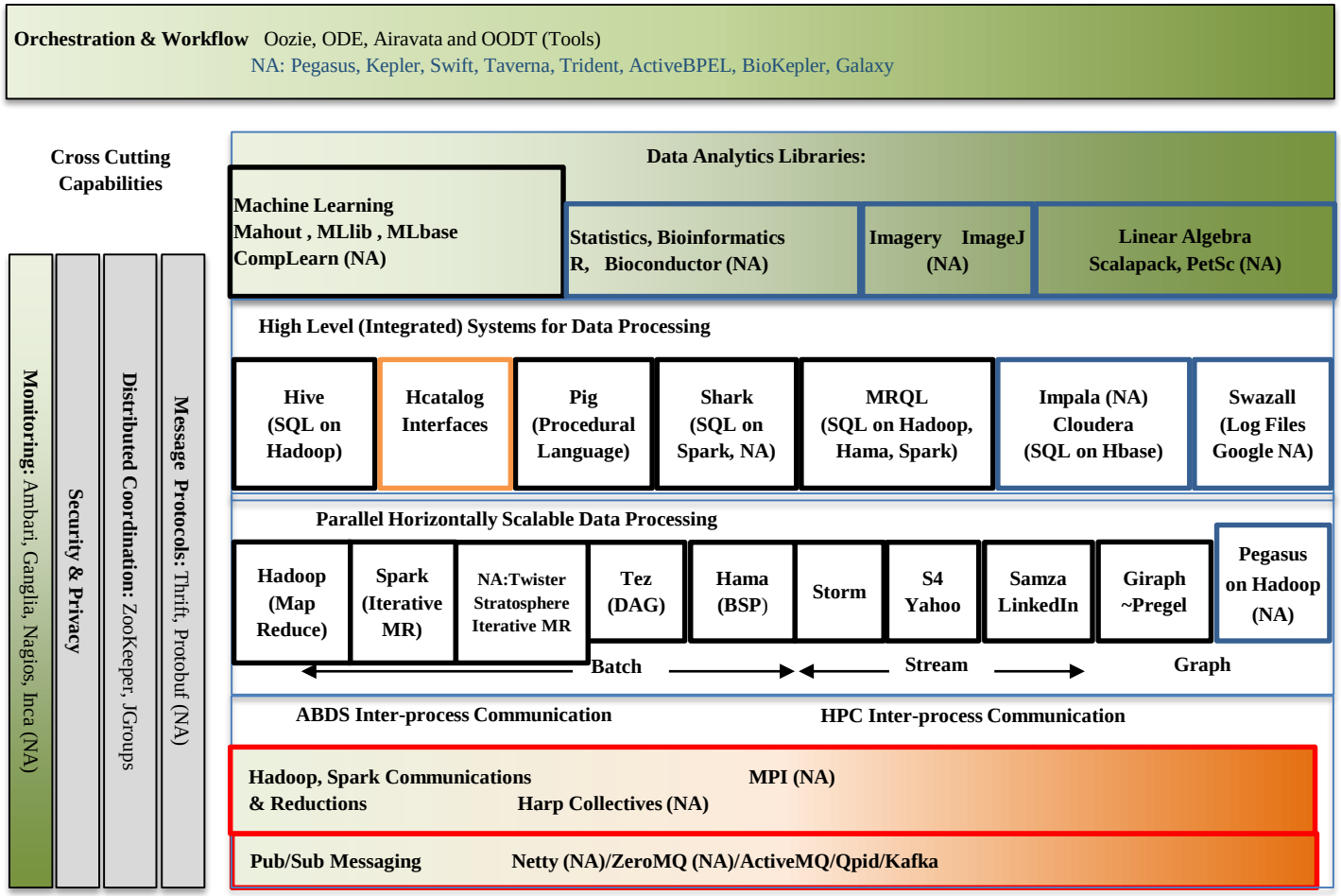


QUT

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

Big Data Technologies – Big Picture

- NA – Non Apache projects
- Green layers are Apache/Commercial Cloud (light) to HPC (darker) integration layers



The figure of layered architecture courtesy of Prof. Geoffrey Fox

Algorithms for Bayesian Big Data Analytics

Markov chain Monte Carlo

Approximate the posterior distribution by Monte Carlo simulation using a Markov chain algorithm

- *Pros*: very easy, effective, flexible; many variations
- *Cons*: computational time

MCMC and computational Bayes Expert:
Professor Antonietta Mira

QUT

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

Algorithms for Bayesian Big Data Analytics

Divide and Conquer

Analyse subsets of the data in parallel by different processors, via either multi-core CPU or massively parallel GPU, and combine results

- *Pros*: very effective
- *Cons*: (i) can't alleviate bottlenecks related to memory or disk; (ii) requires sophisticated programming; (iii) different communication costs, so need different algorithms

QUT

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

Algorithms for Bayesian Big Data Analytics

Consensus Monte Carlo

Run a separate Monte Carlo algorithm on each machine, then average individual Monte Carlo draws across machines

- *Pros:* embarrassingly parallel, so can be run on virtually any system for parallel computing, multi-core systems, or networks of workstations
- *Cons:* need rules for combining posterior samples; still requires careful programming

QUT

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

Algorithms for Bayesian Big Data Analytics

Resampling algorithms

- Bag of little bootstraps (Kleiner et al., 2014)
- Resampling-based approximations (Liang et al. 2013)
- Properties of MCMC resampling and subsampling algorithms (Bardenet et al., 2014, 2015)

QUT

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

Algorithms for Bayesian Big Data Analytics

Approximation algorithms

- INLA (Rue et al. 2009):
  - *Pros:* fast, wide variety of models
  - *Cons:* computational cost of INLA is exponential in the dimension of the parameter space
- Variational Bayes (Hinton and Van Camp, 1993):
  - *Pros:* approximate posteriors give good approximations to individual marginal distributions
  - *Cons*: not such good approximations to the joint distribution as a whole
- Approximate Bayesian Computation (Mira):

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)

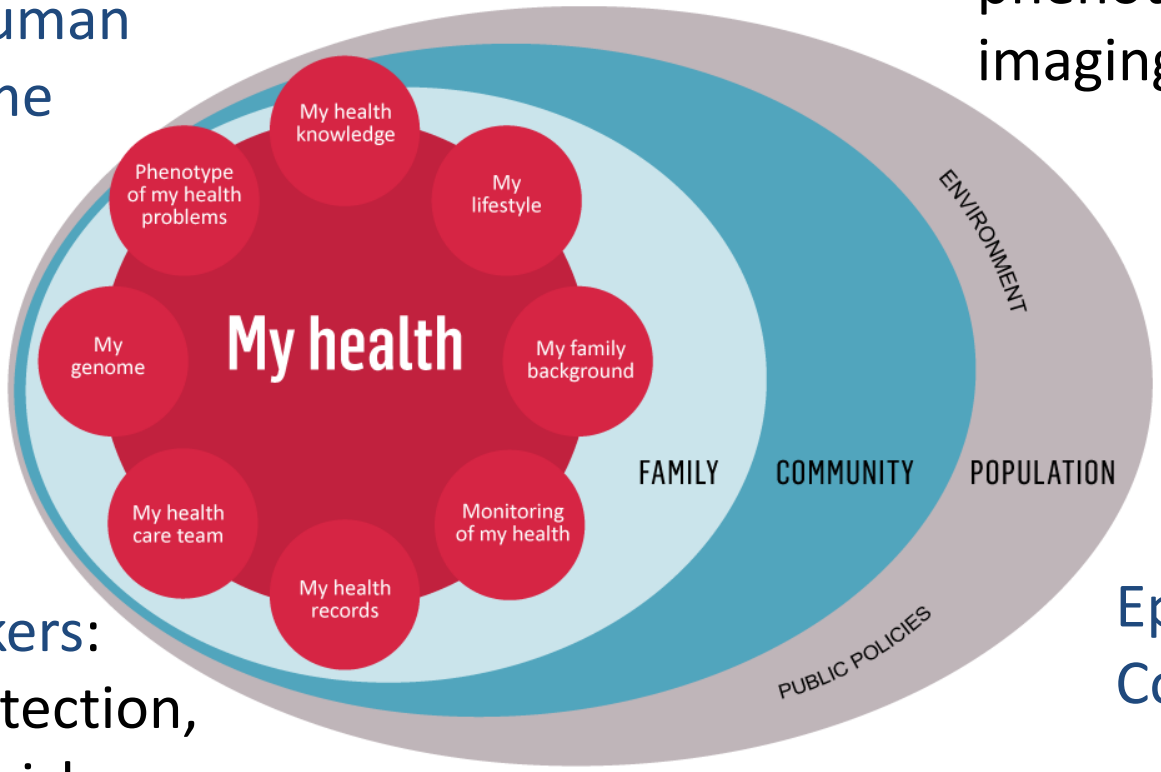
+

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

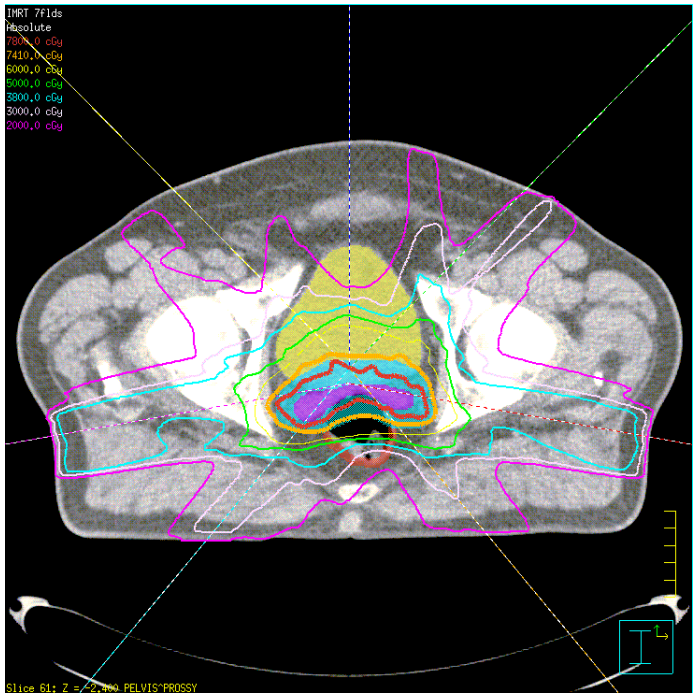
ACEMS
AUSTRALIAN RESEARCH COUNCIL CENTRE OF EXCELLENCE FOR
MATHEMATICAL AND STATISTICAL FRONTIERS

Bayesian statistics?

Imaging

Personalised Medicine

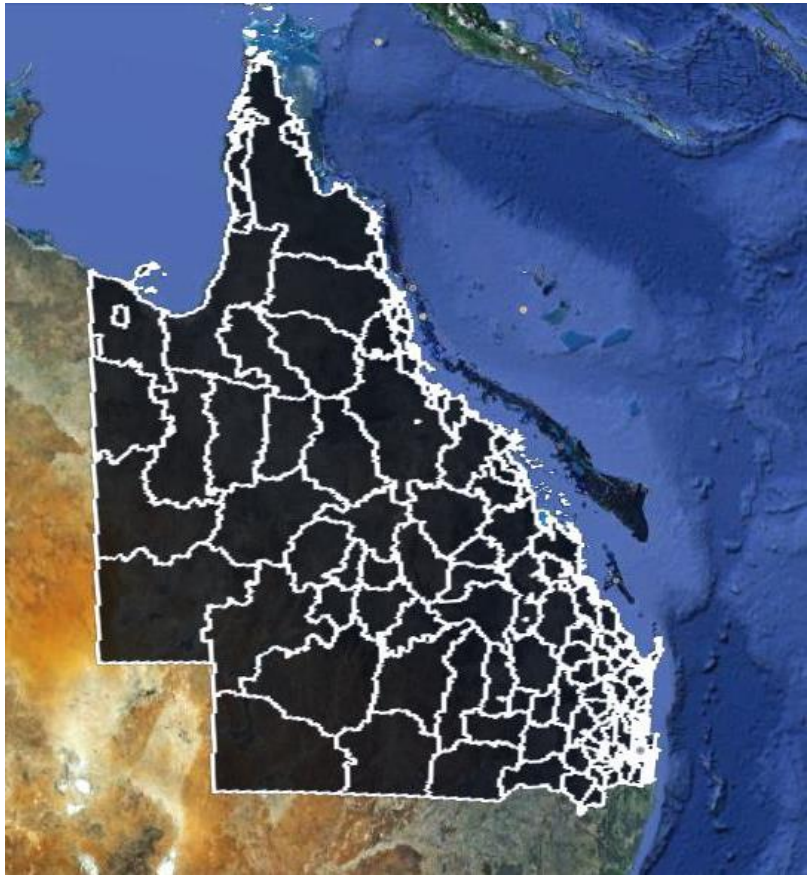
Deep Brain Stimulation for Parkinson's Disease



QUT

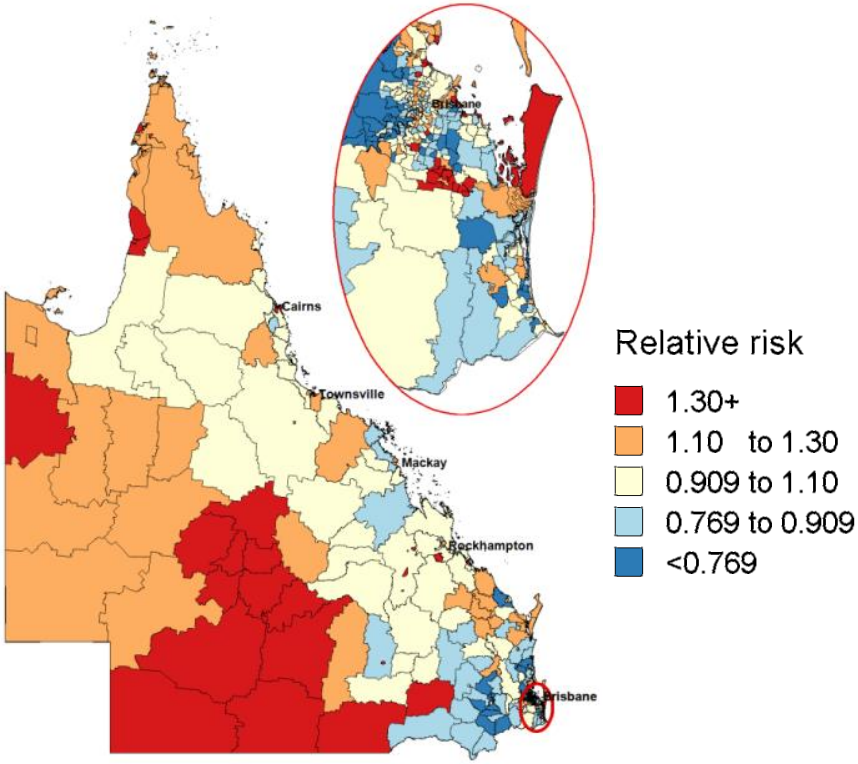
ACEMS
AUSTRALIAN RESEARCH COUNCIL CENTRE OF EXCELLENCE FOR
MATHEMATICAL AND STATISTICAL FRONTIERS

Does “place” impact on cancer survival?



QUT

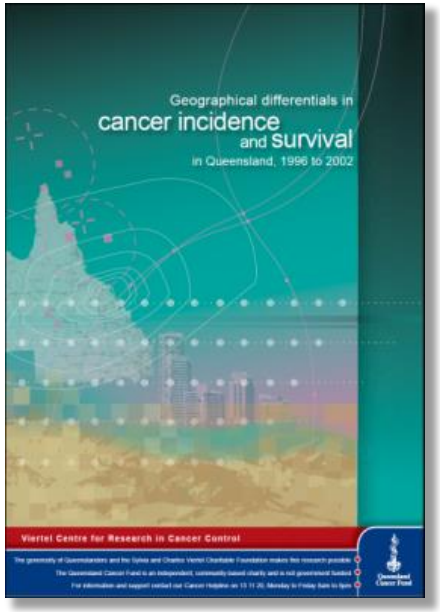
ACEMS
AUSTRALIAN RESEARCH COUNCIL CENTRE OF EXCELLENCE FOR
MATHEMATICAL AND STATISTICAL FRONTIERS



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

Supercomputing + Modelling = Long-term vision

The Australian experience

- New supercomputer planned in 2013
- Announced in 2014
- Awarded in mid-2015
- Release in mid-2016
- Planned upgrade for 2018 (ensure the supercomputer is meeting needs)

  Will bring the new supercomputer up from 1.6 to 5 petaflops

http://www.itnews.com.au/news/cray-to-build-australias-biggest-supercomputer-406827#ixzz4BLkUv2P2

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/

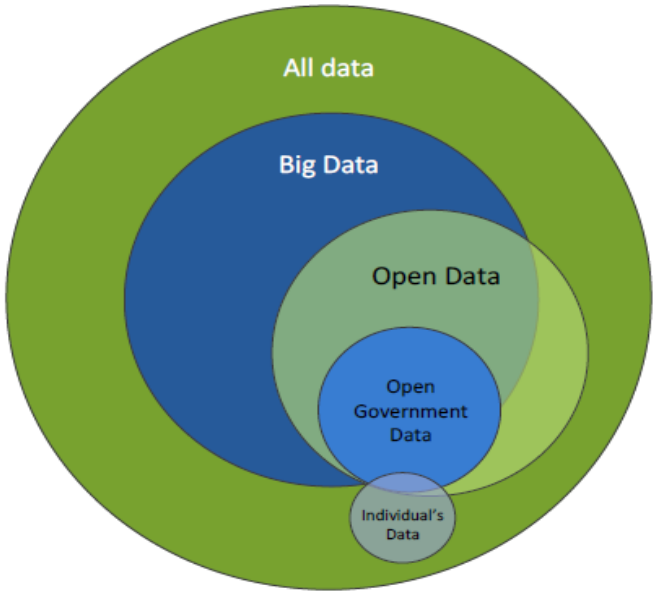
QUT

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

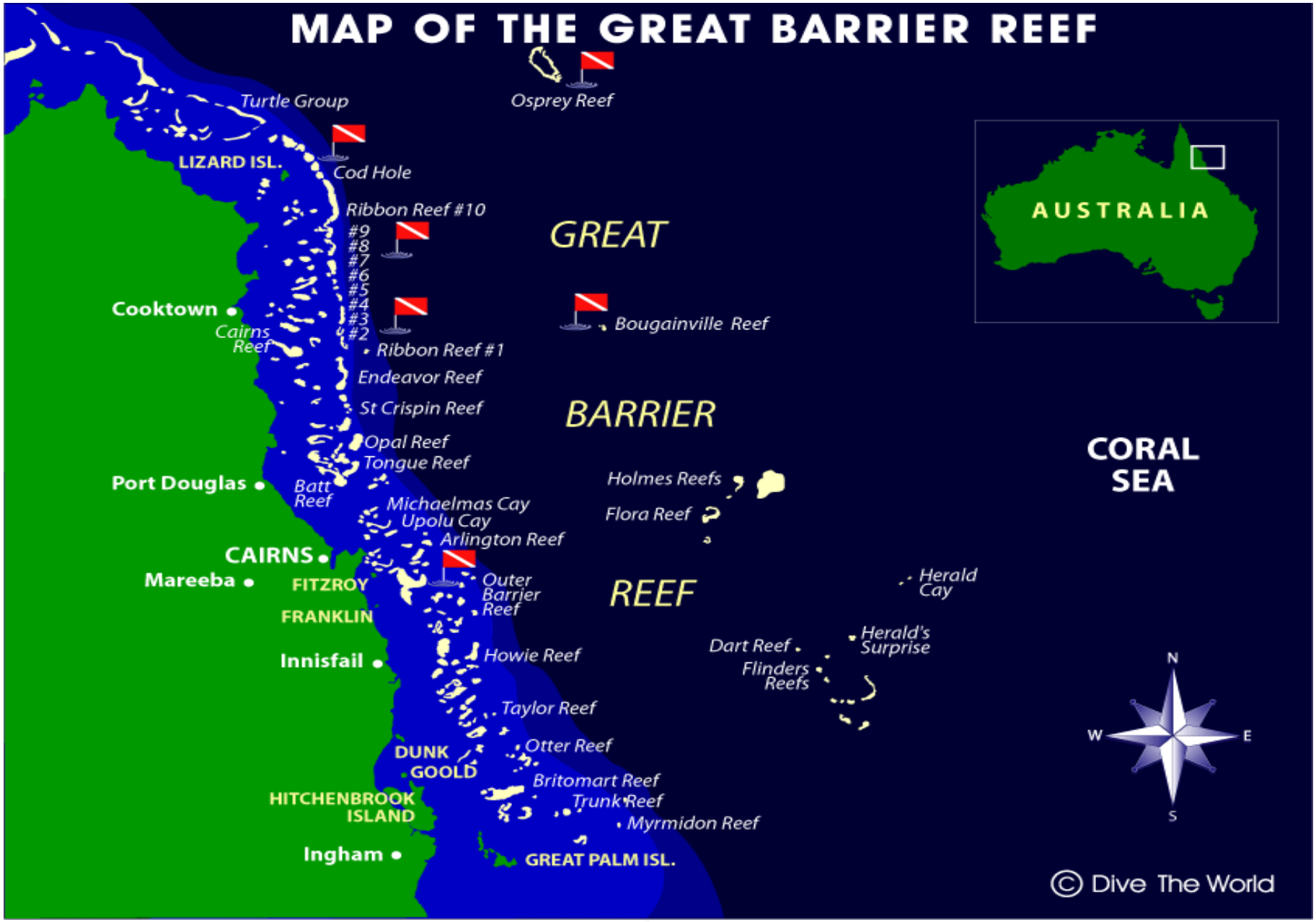
Cyclone damage devastating to the Great Barrier Reef

Sep 2 2013
Thousands gather for Rally for the Reef

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

QUT

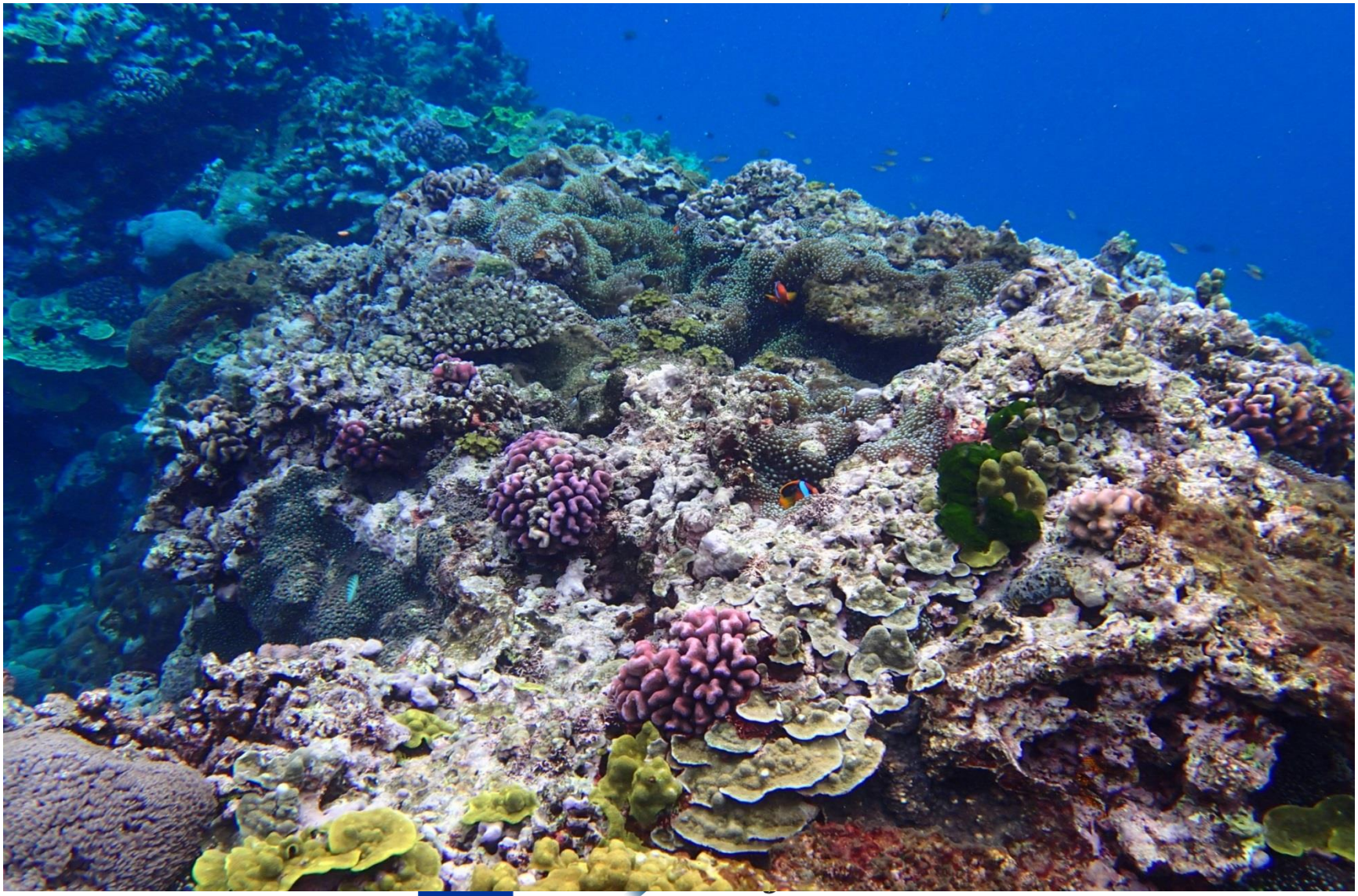
ACEMS
AUSTRALIAN RESEARCH COUNCIL CENTRE OF EXCELLENCE FOR MATHEMATICAL AND STATISTICAL FRONTIERS



QUT

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

Data source 1: observations

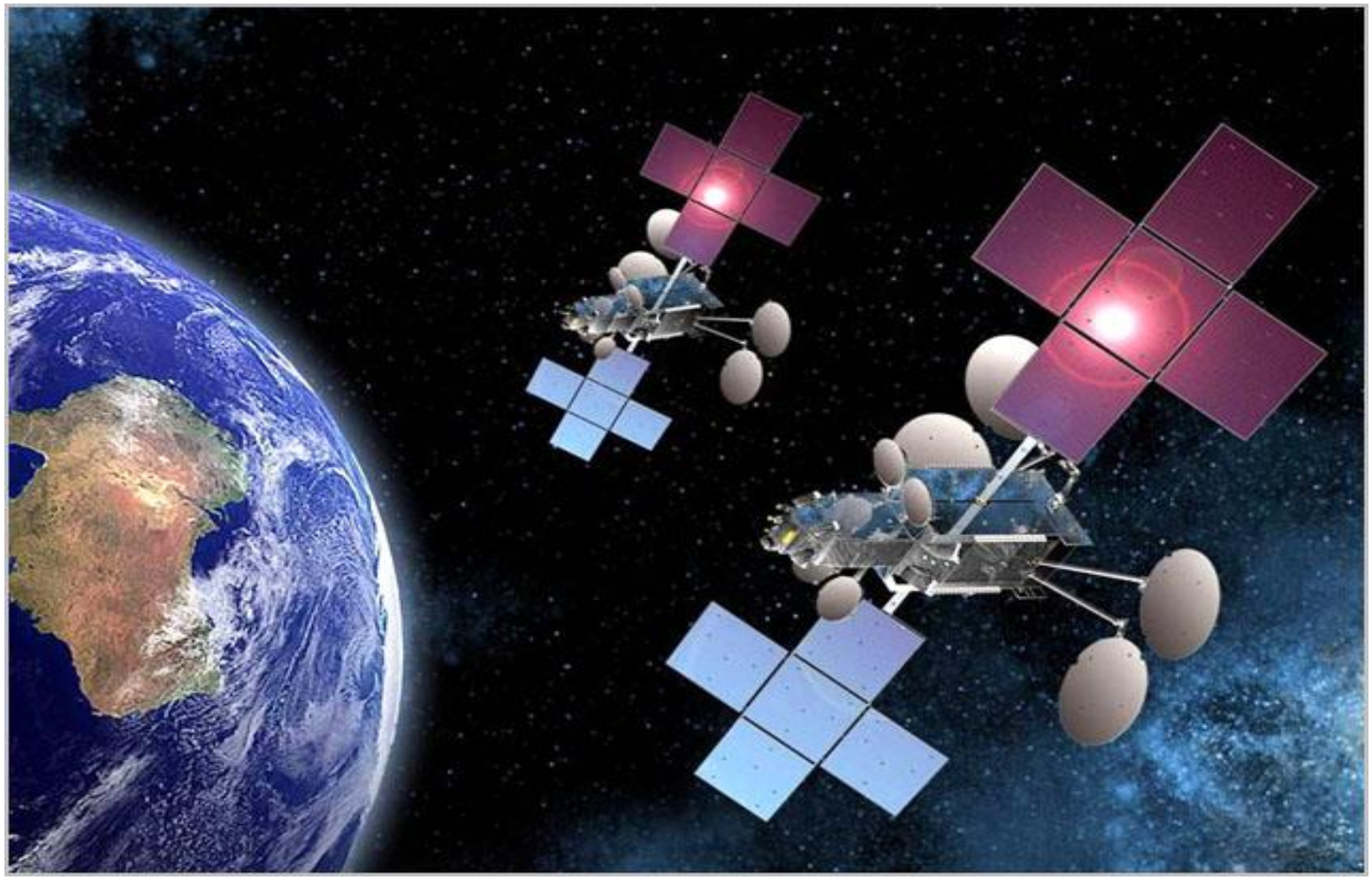


Data source 2: sensors

QUT

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

Data source 3: images



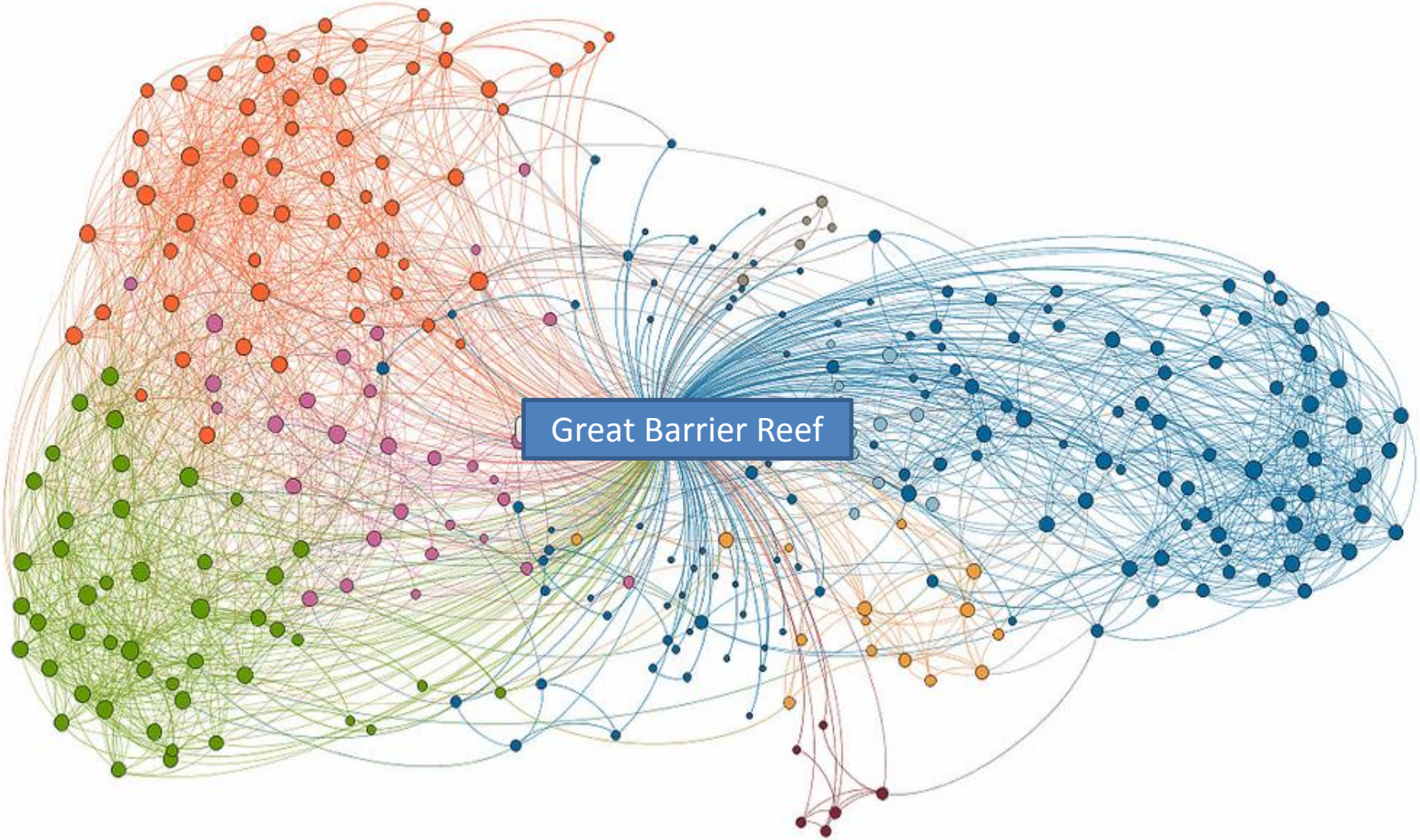
delimiter.com.au

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

Bayesian statistics and big data

Clusters

Patterns

MODEL

Drivers

Predictions

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)*

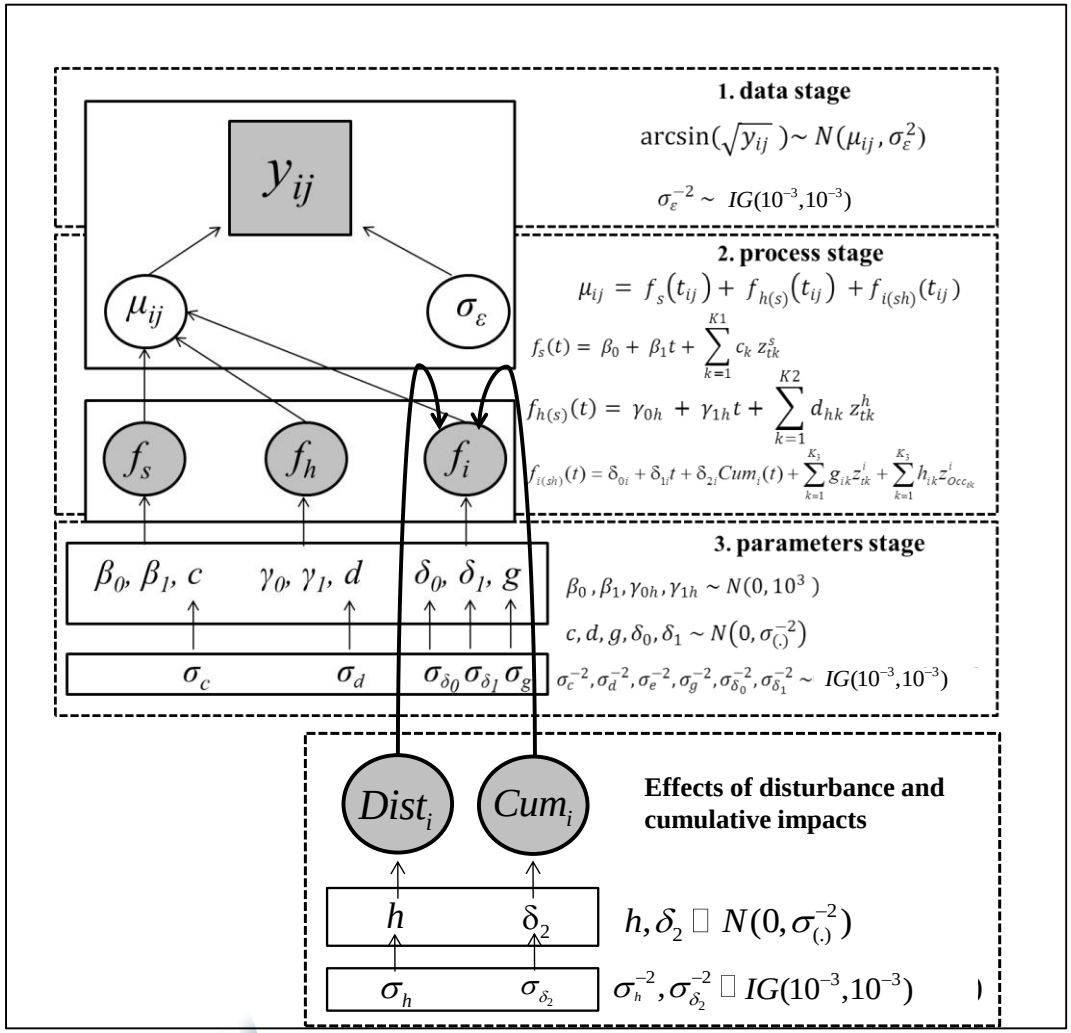
Hierarchical modelling

Data

Process

Parameters

Drivers



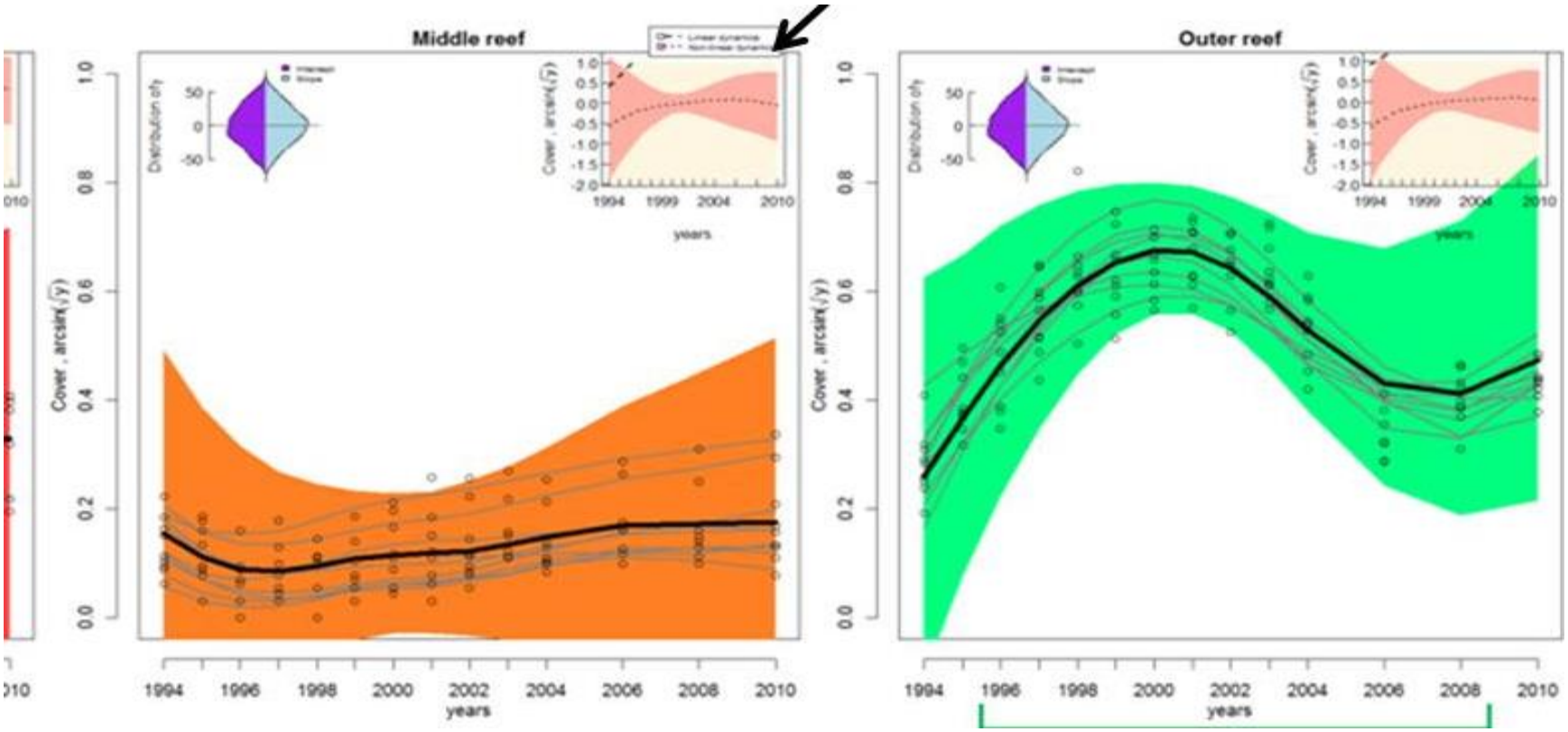
QUT

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

Meeting tomorrow's demands today

Virtual reality and Immersive environments have arrived!

https://www.youtube.com/watch?v=8FggsGUK5iA

VR/IE + Bayesian Statistics in Conservation:

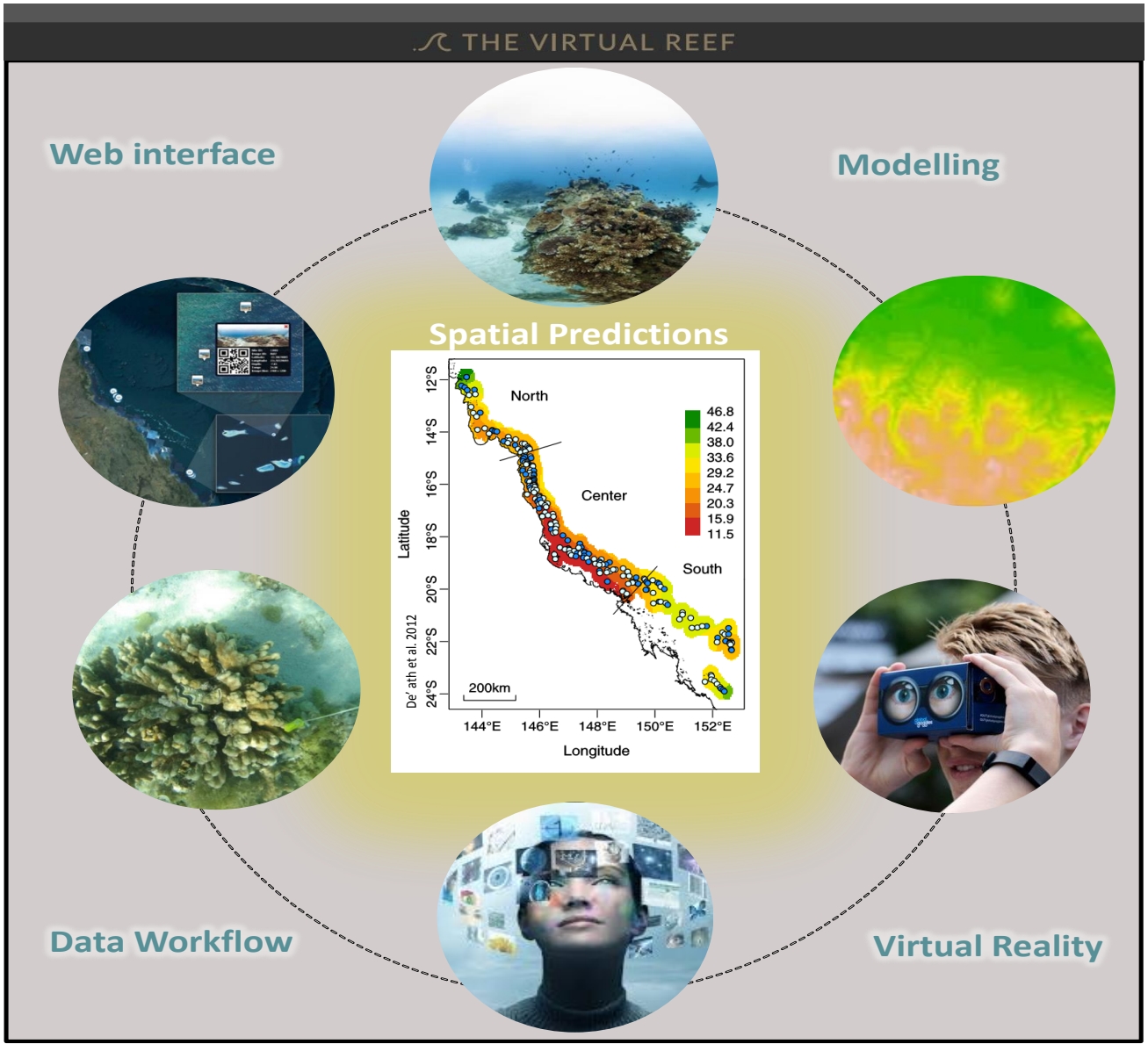
"Eyes on the Reef" https://youtu.be/1VU-q-uQPXU

"Eyes on the Wild" vis.stats.technology

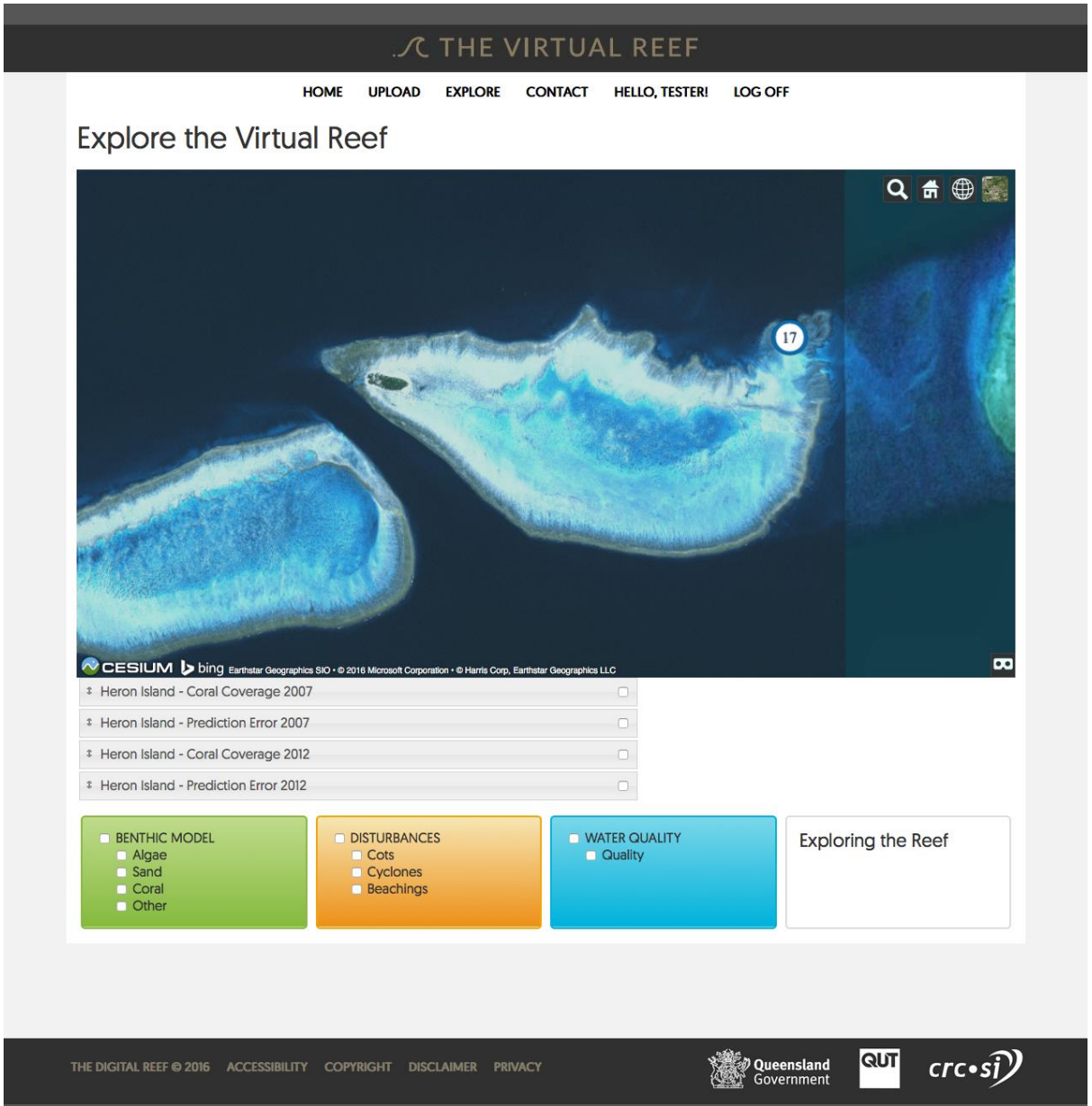
QUT

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

Monitoring the Barrier Reef



(Citizen science) user interface

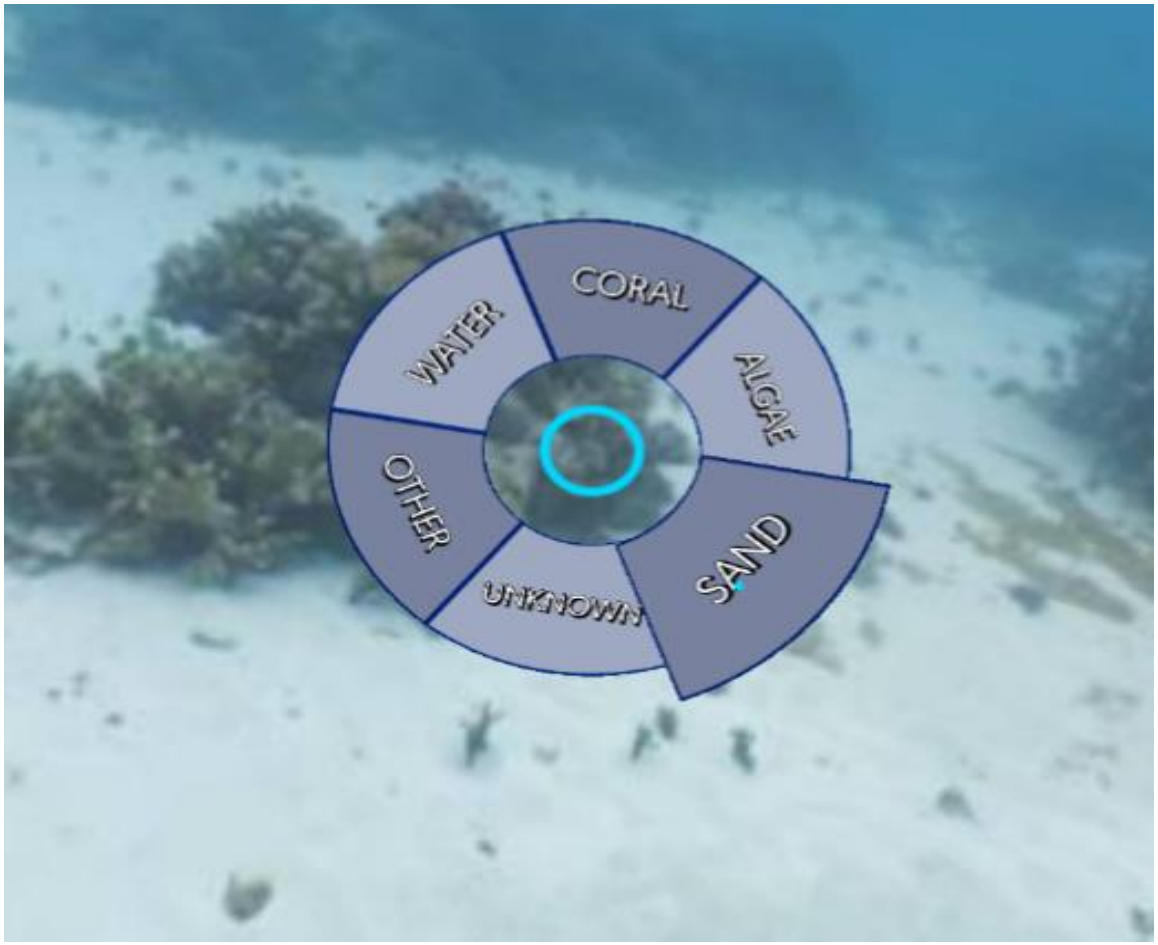


Elicitation of information from photos/film

QUT

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

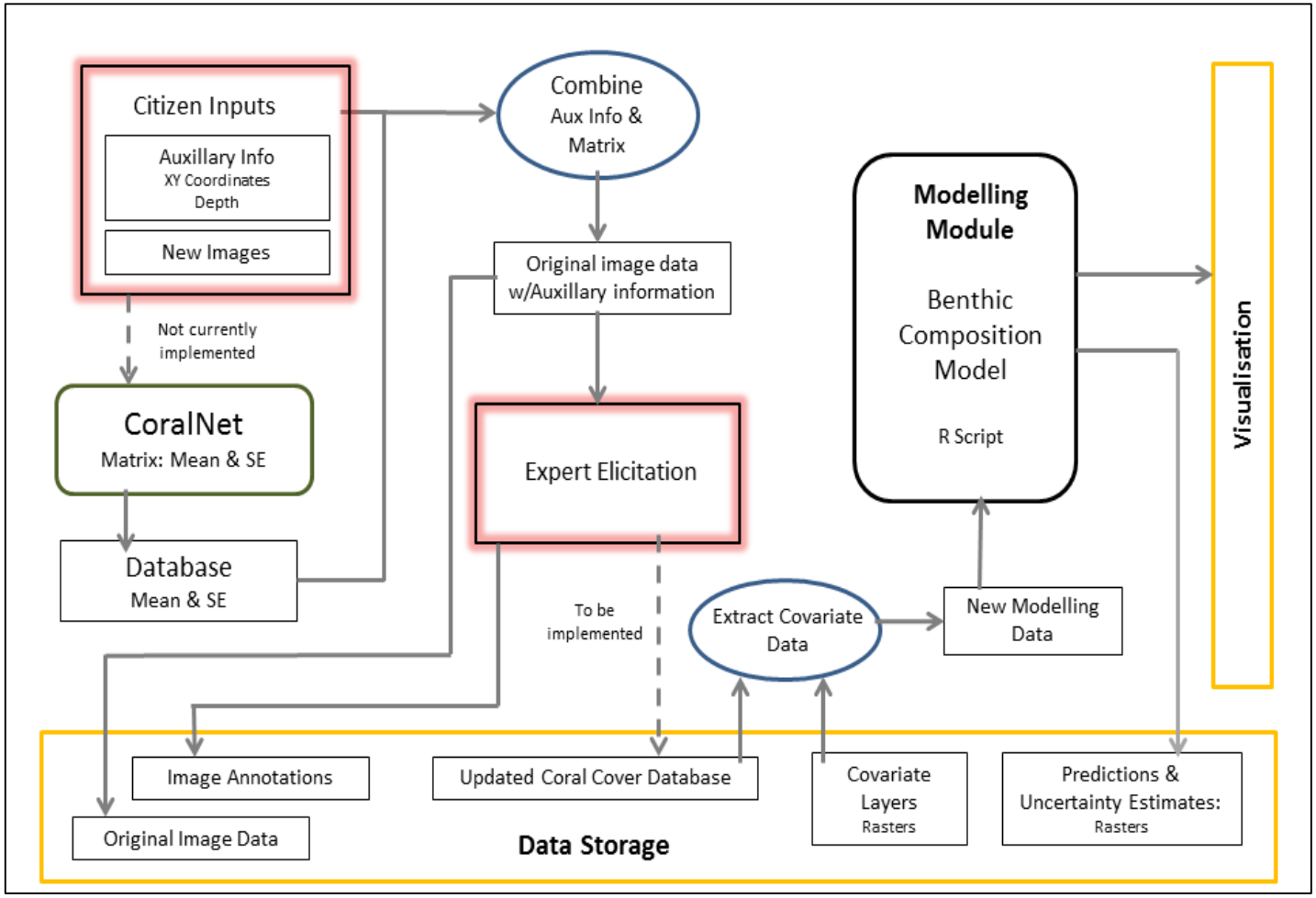
Elicitation of information from photos/film



QUT

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

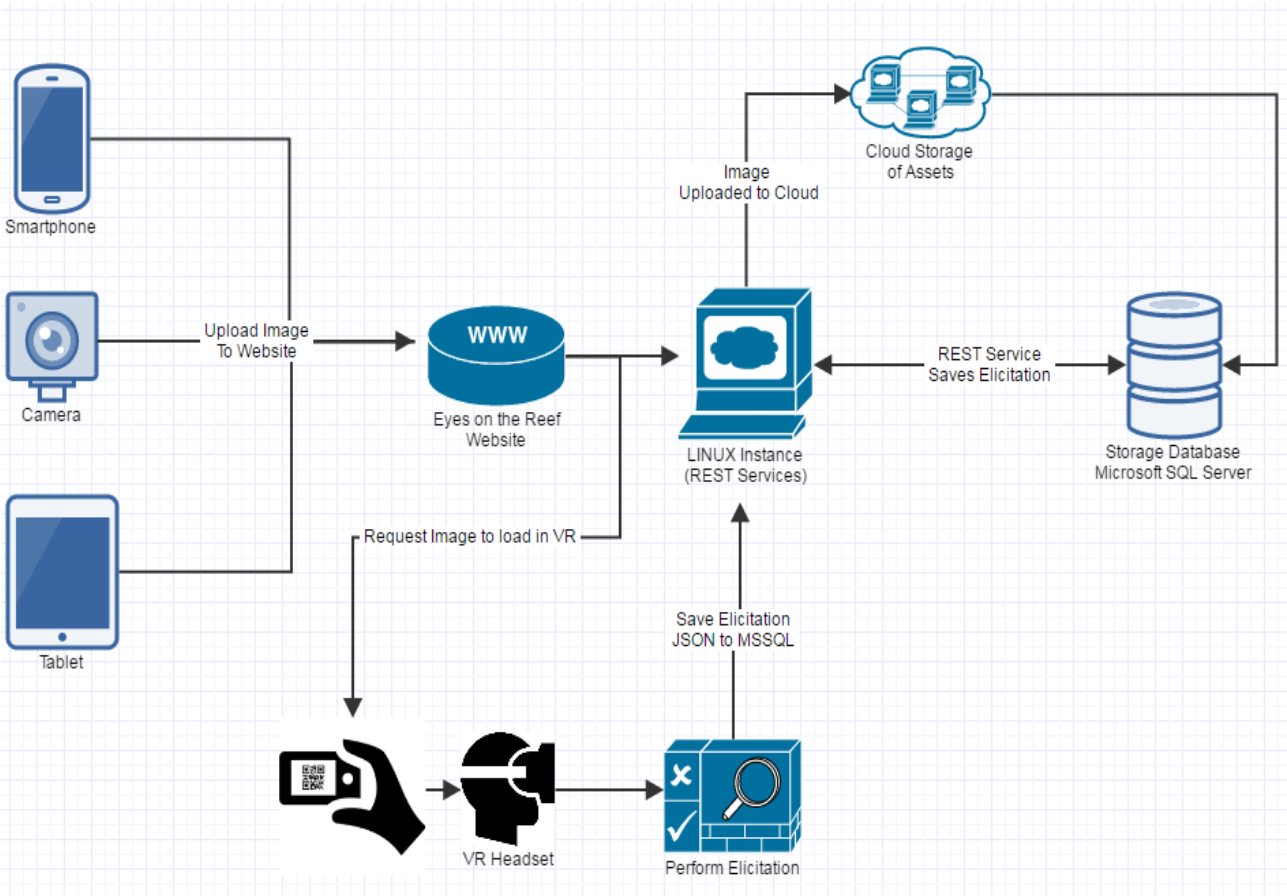
Workflow



QUT

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

Workflow for elicitation upload



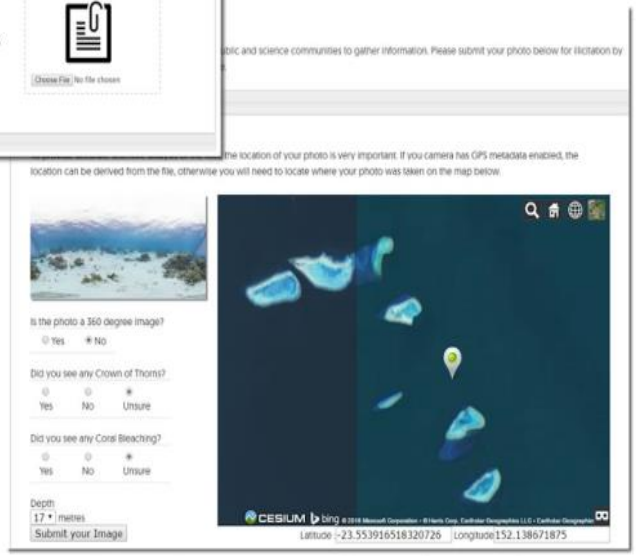
QUT

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

Workflow for elicitation upload

User contribution Photo Upload Process

1. Image Upload



2. Image Location and Metadata gathered

3. QR Code generated

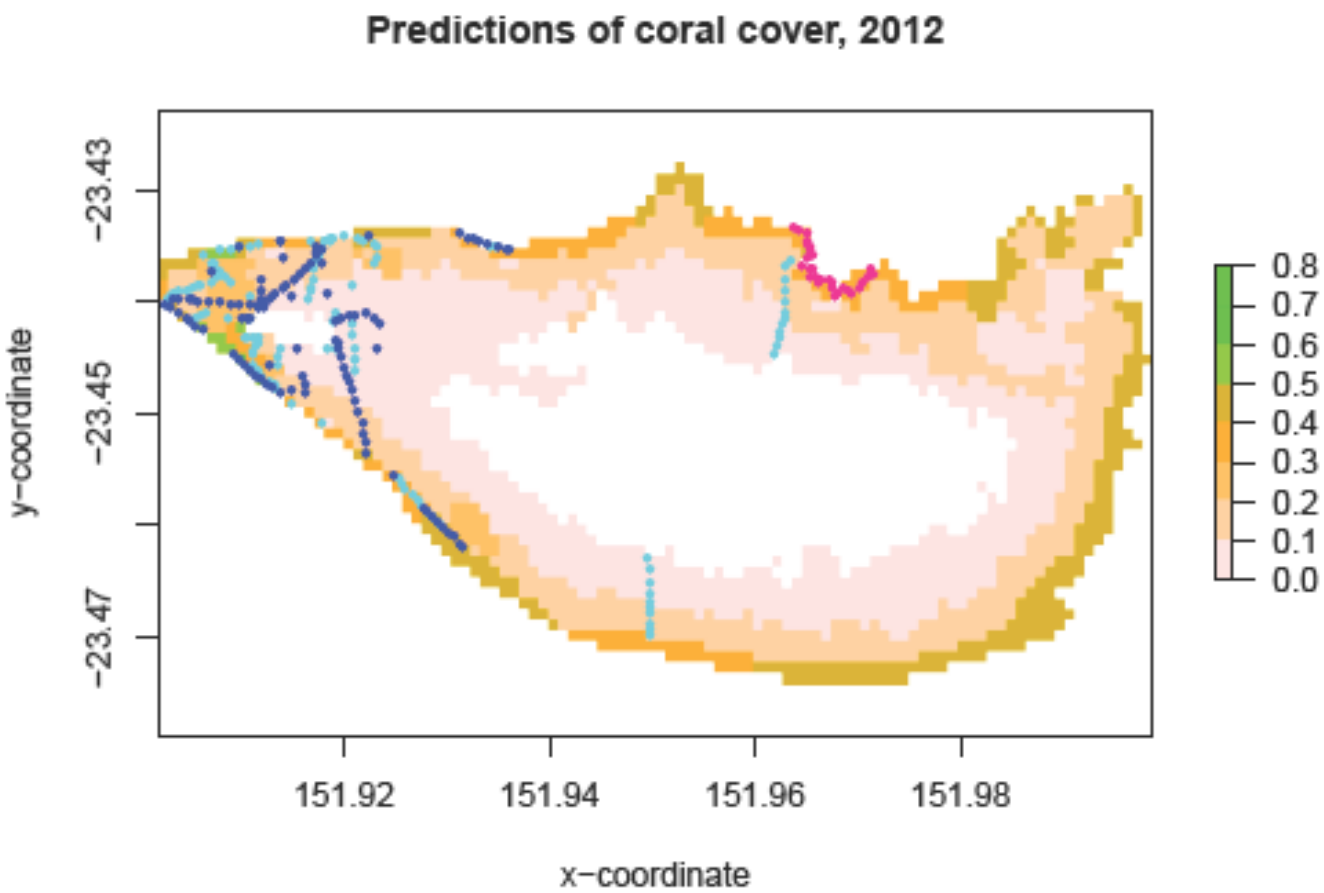
A user uploads an image to the website which enables the creation of a QR code associated with the image.
When scanned with the VR application, it launches a specific eilicitation session with the related image.

Upload : Select Image / Gather Metadata / Generate QR code for VR Elicitation

QUT

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

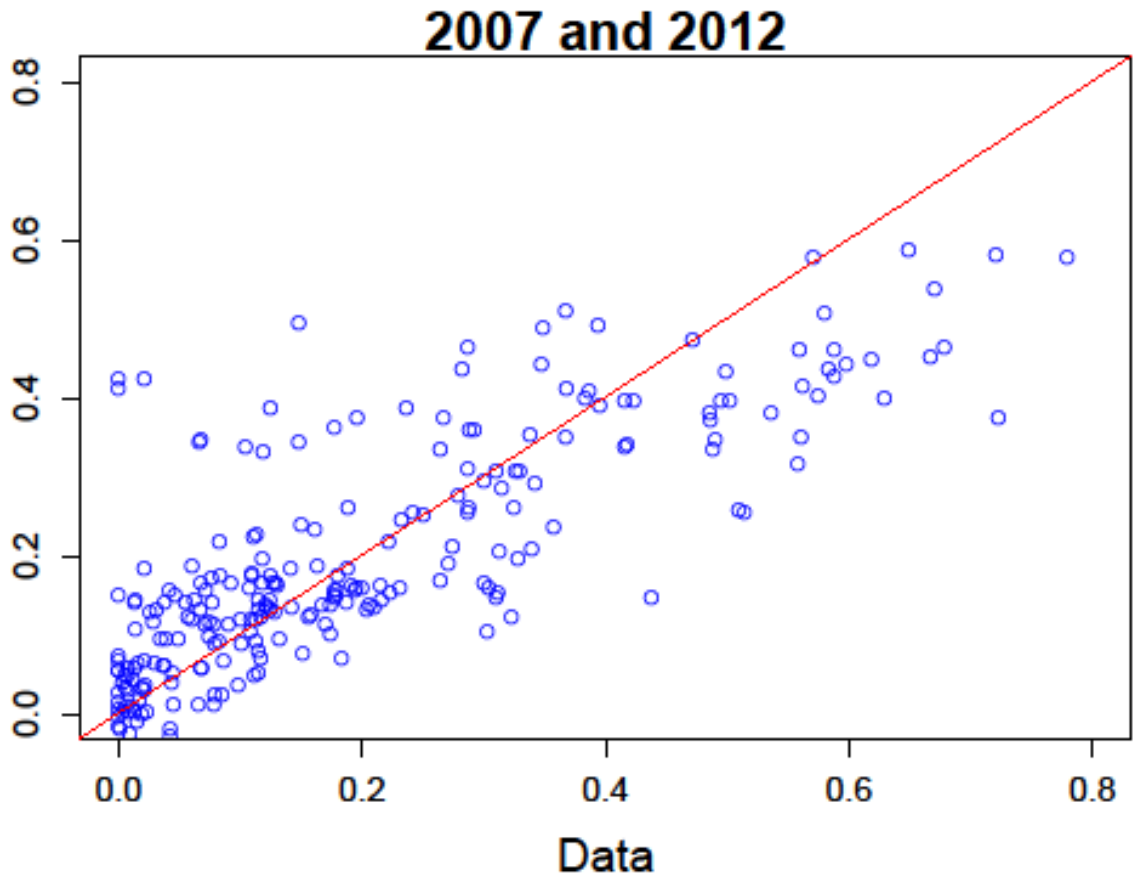
(Bayesian) spatial outputs
Posterior predictions based on data + photos/films



QUT

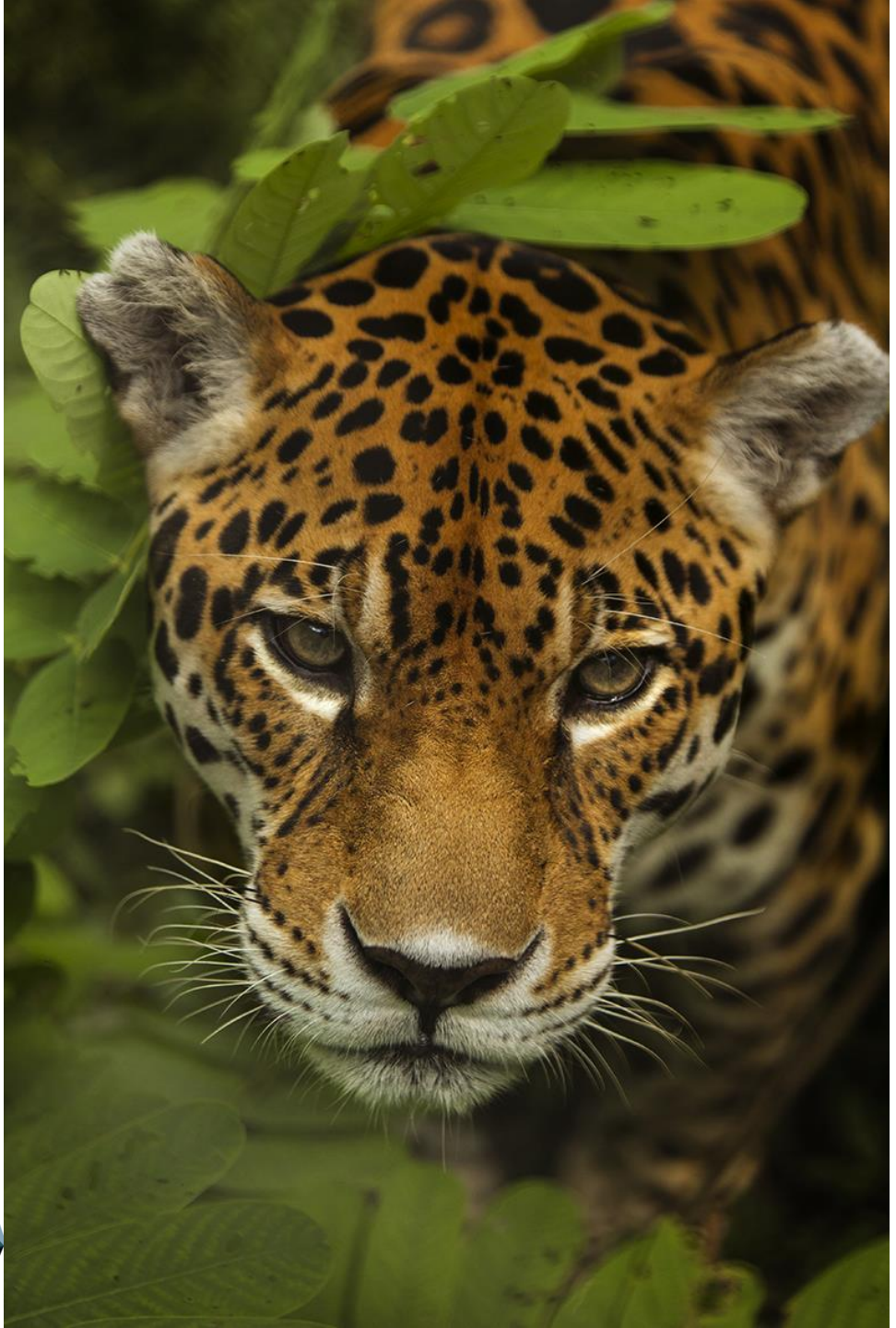
ACEMS
AUSTRALIAN RESEARCH COUNCIL CENTRE OF EXCELLENCE FOR
MATHEMATICAL AND STATISTICAL FRONTIERS

(Bayesian) spatial outputs
Predictive accuracy



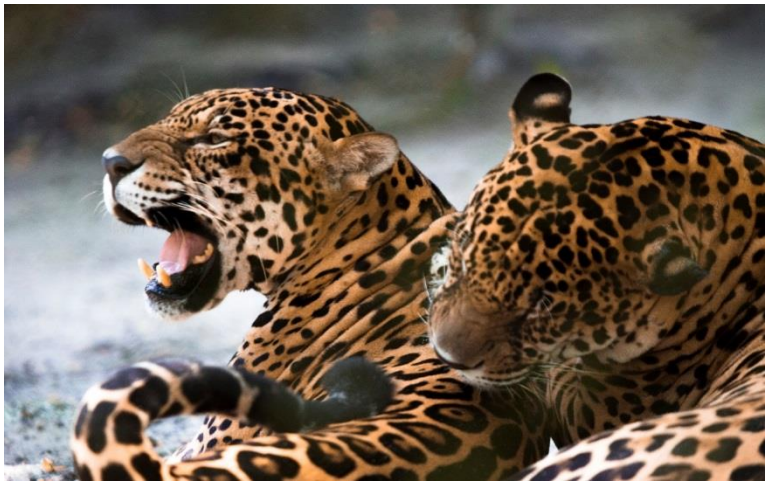
Hunting jaguars in the Peruvian Amazon

Aim: to create a jaguar corridor across Peru



“Eyes on the Wild”

What’s (Bayesian) statistics got to do with it?



1. Identify prime jaguar habitat
2. Estimate jaguar populations, their spatial distributions, and trends in jaguar numbers over time
3. Collate information and develop models about key threats and impacts

LUPUNALUZ Foundation

ACEMS

The expedition

Two locations:

1.Imiria Reserve

2. Pacaya Samiria Reserve

Imiria Reserve

Pacaya Samiria Reserve

Primary information sources

Data

Local Knowledge

Experts

QUT

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

Local knowledge

"Tell us your stories"

“Stars on a map”

Collecting other information

"Where would jaguars live?"

Asking the experts

"Instead of bringing experts to the Amazon, how can we take the Amazon to the experts?"

Create virtual worlds and immersive environments of habitat

Virtual environments

“Putting the expert into the environment”

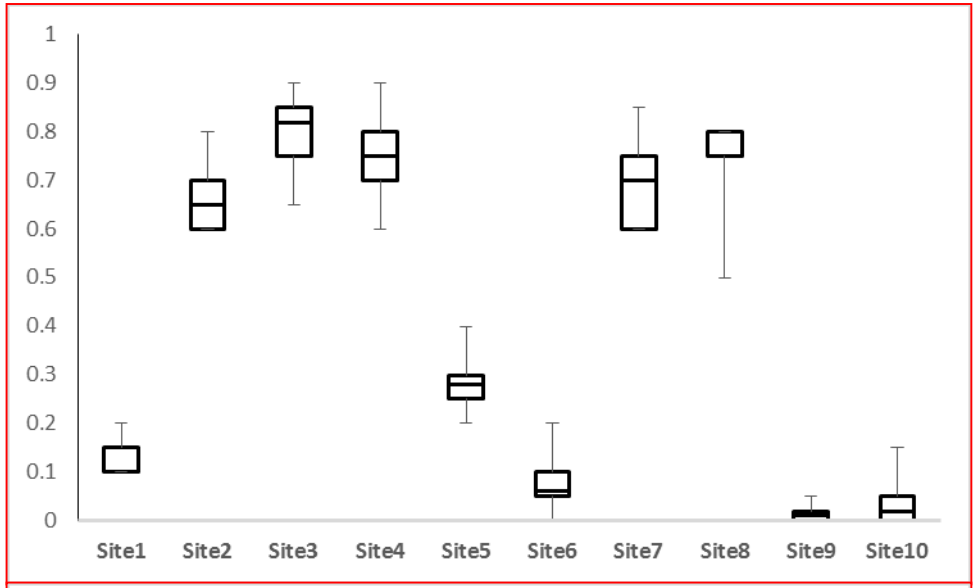
QUT

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

Results: pilot study

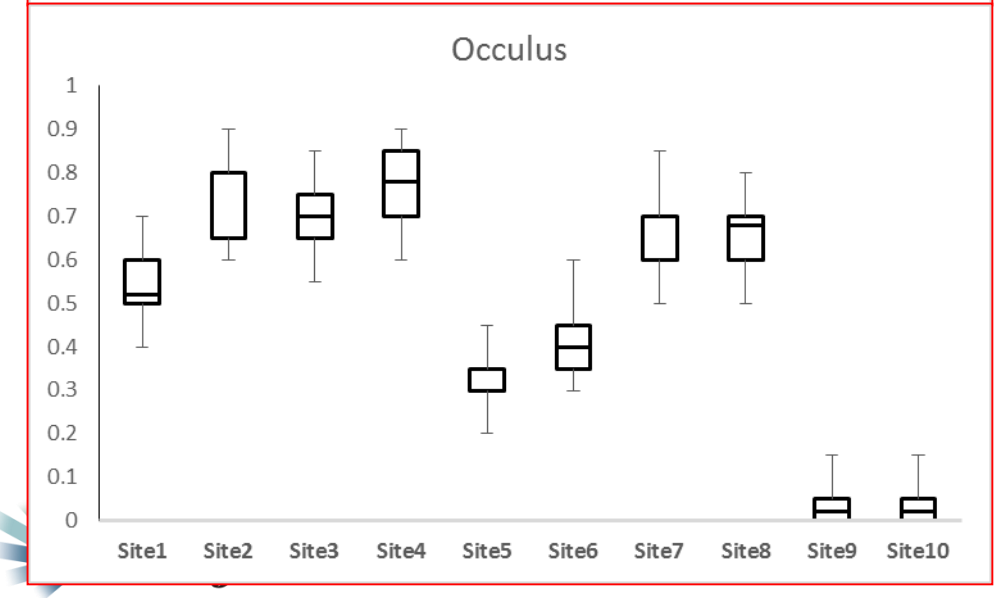
Normal screen

Normal Screen			
	Site1	Site2	Site3
Min	0.1	0.6	0.65
Q_1	0.1	0.6	0.75
Best estimate	0.1	0.65	0.82
Q_3	0.15	0.7	0.85
Max	0.2	0.8	0.9



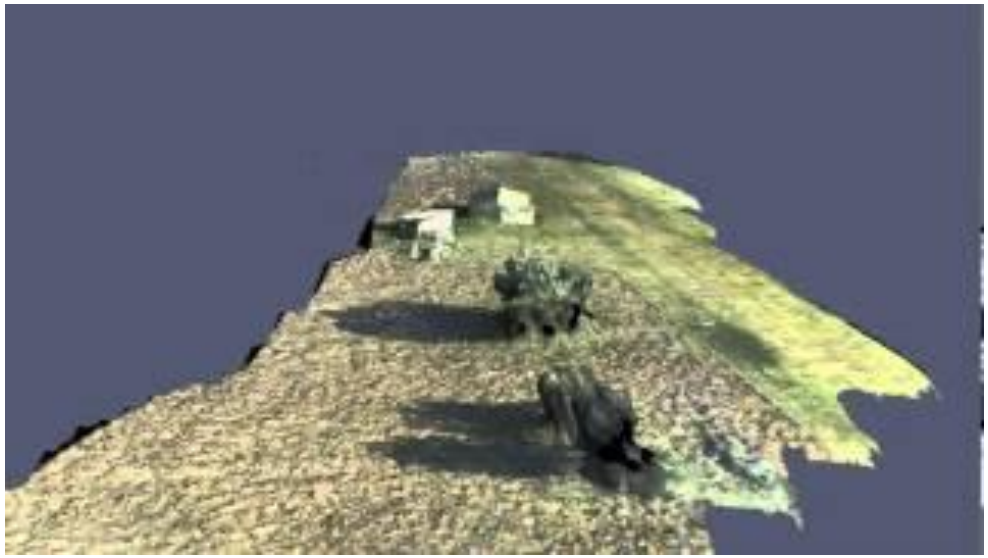
Occulus

Occulus			
	Site1	Site2	Site3
Min	0.4	0.6	0.55
Q_1	0.5	0.65	0.65
Best estimate	0.52	0.8	0.7
Q_3	0.6	0.8	0.75
Max	0.7	0.9	0.85



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

Drones & UAVs



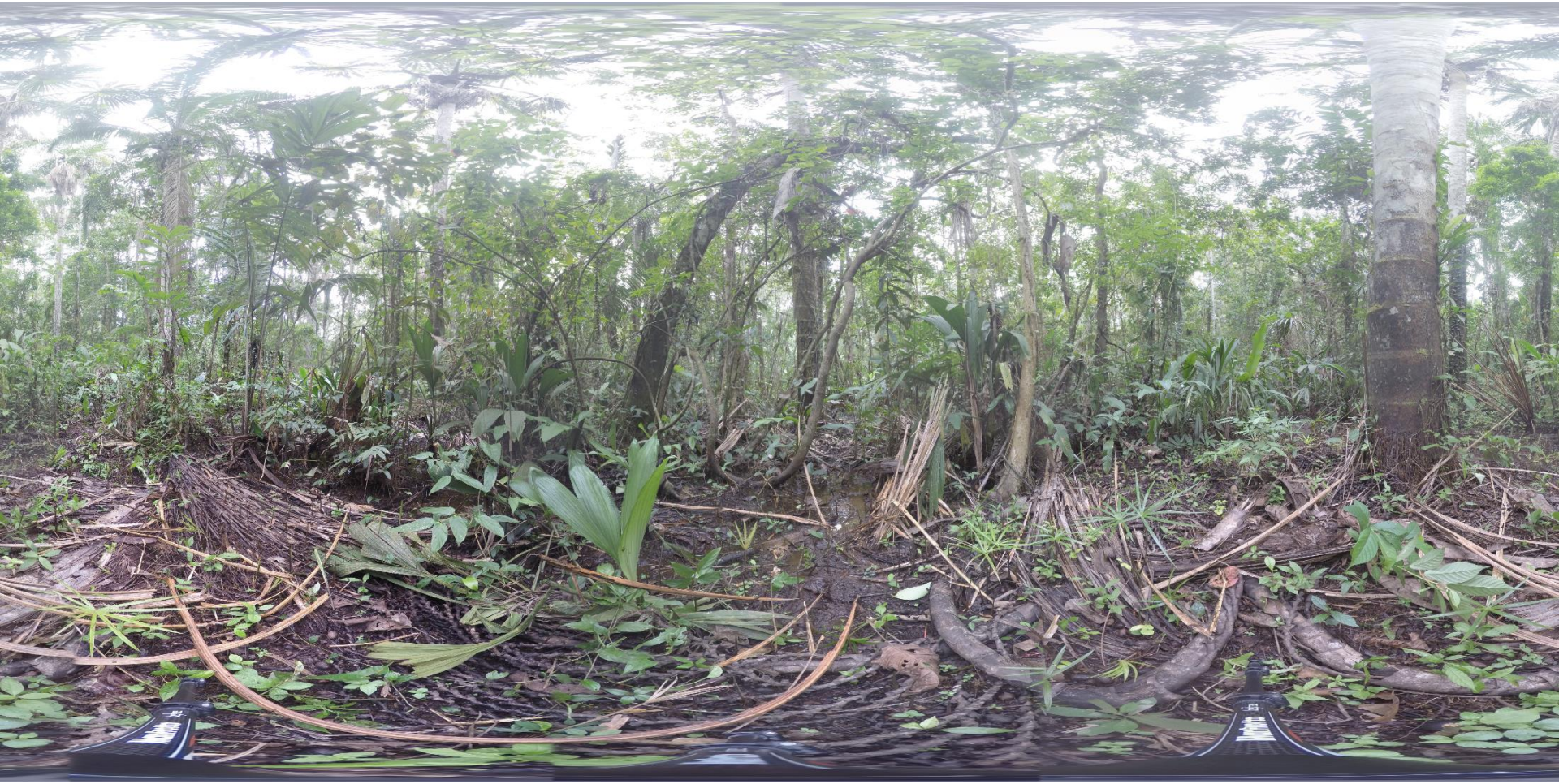
Australian News: "Aerial robots, artificial intelligence and statistics revolutionise wildlife tracking and research"

New Reality

Go-pros
360 cameras
3D cameras
VR headset
+ GPS

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

vis.stats.technology



QUT

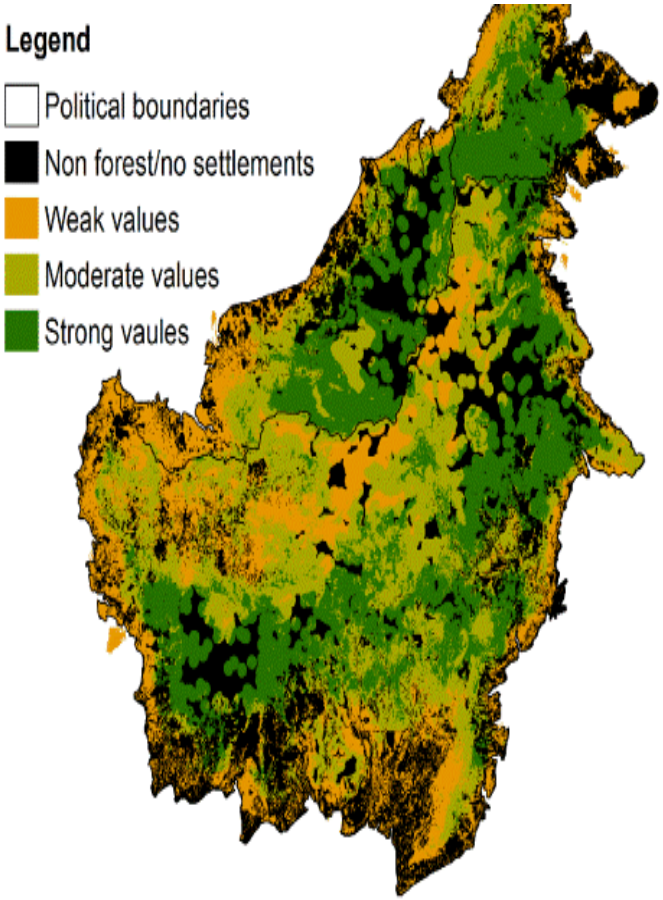
ACEMS
AUSTRALIAN RESEARCH COUNCIL CENTRE OF EXCELLENCE FOR
MATHEMATICAL AND STATISTICAL FRONTIERS

So what about the maths and stats?

Create maths/stats models

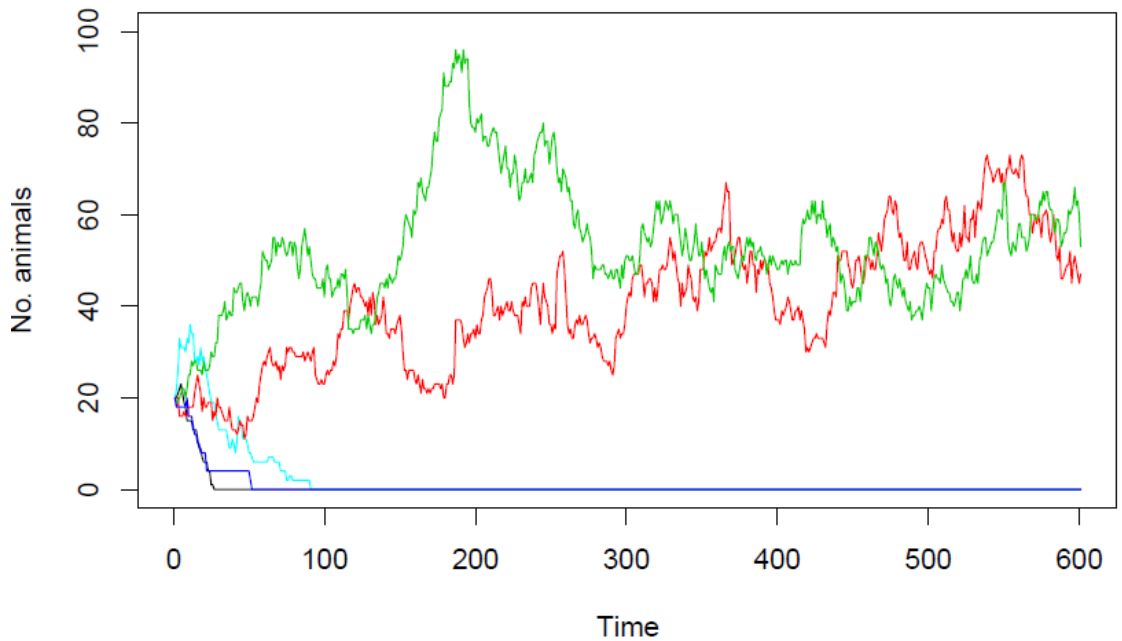
- Spatial model of jaguar encounters
- Markov model of jaguar movements
- Bayesian network of key factors and impacts
- Bayesian models using data

Spatial model of encounters

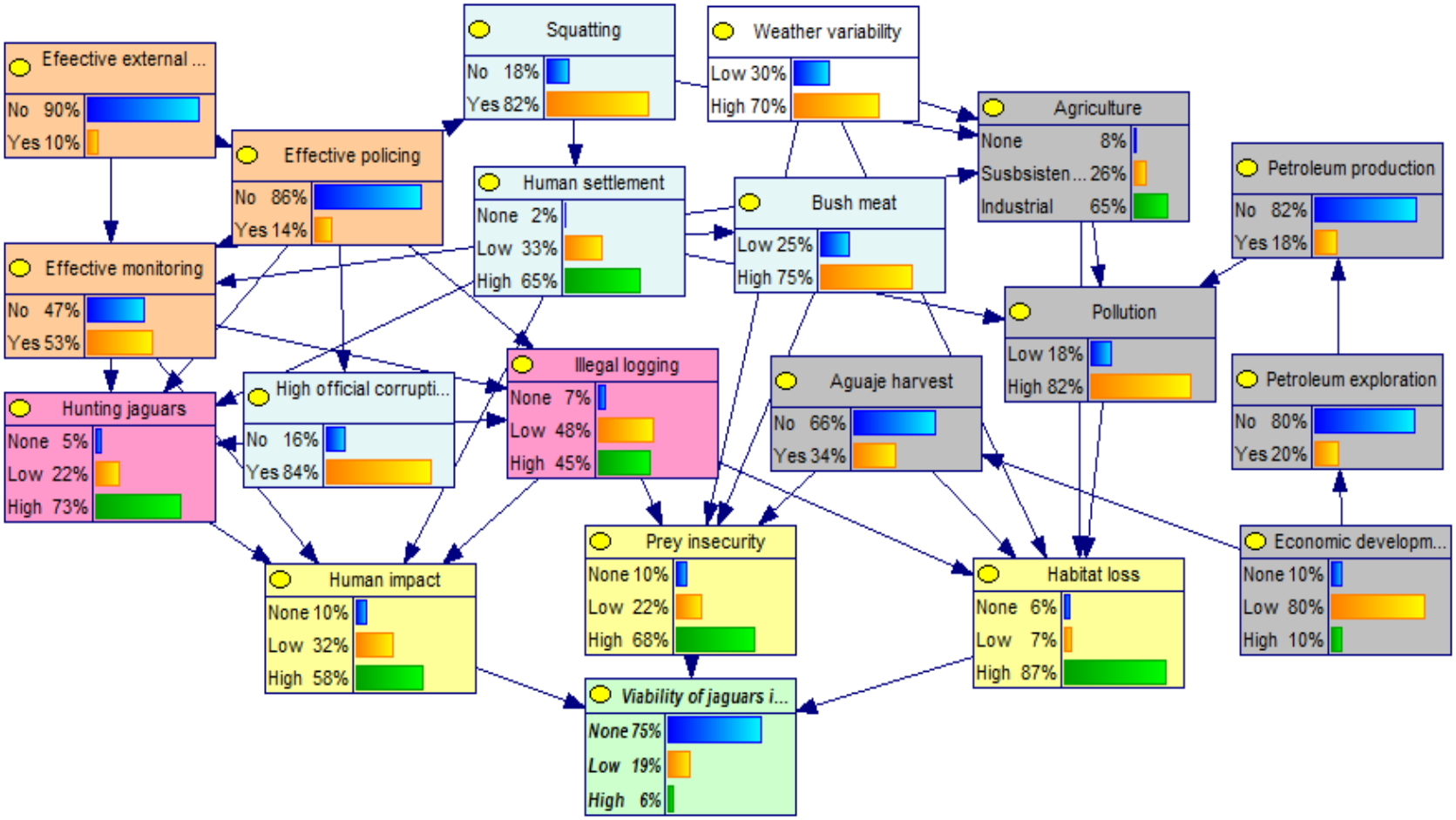


Population dynamics model

Use elicited info about jaguar behaviour, types of prey, birth/death rates, territories, impact of hunting/habitat loss



Network model of key factors



“So what?”

- *For Lupunaluz:* New project to estimate jaguar abundance

- *For us:* New ways to combine new technology and maths/stats

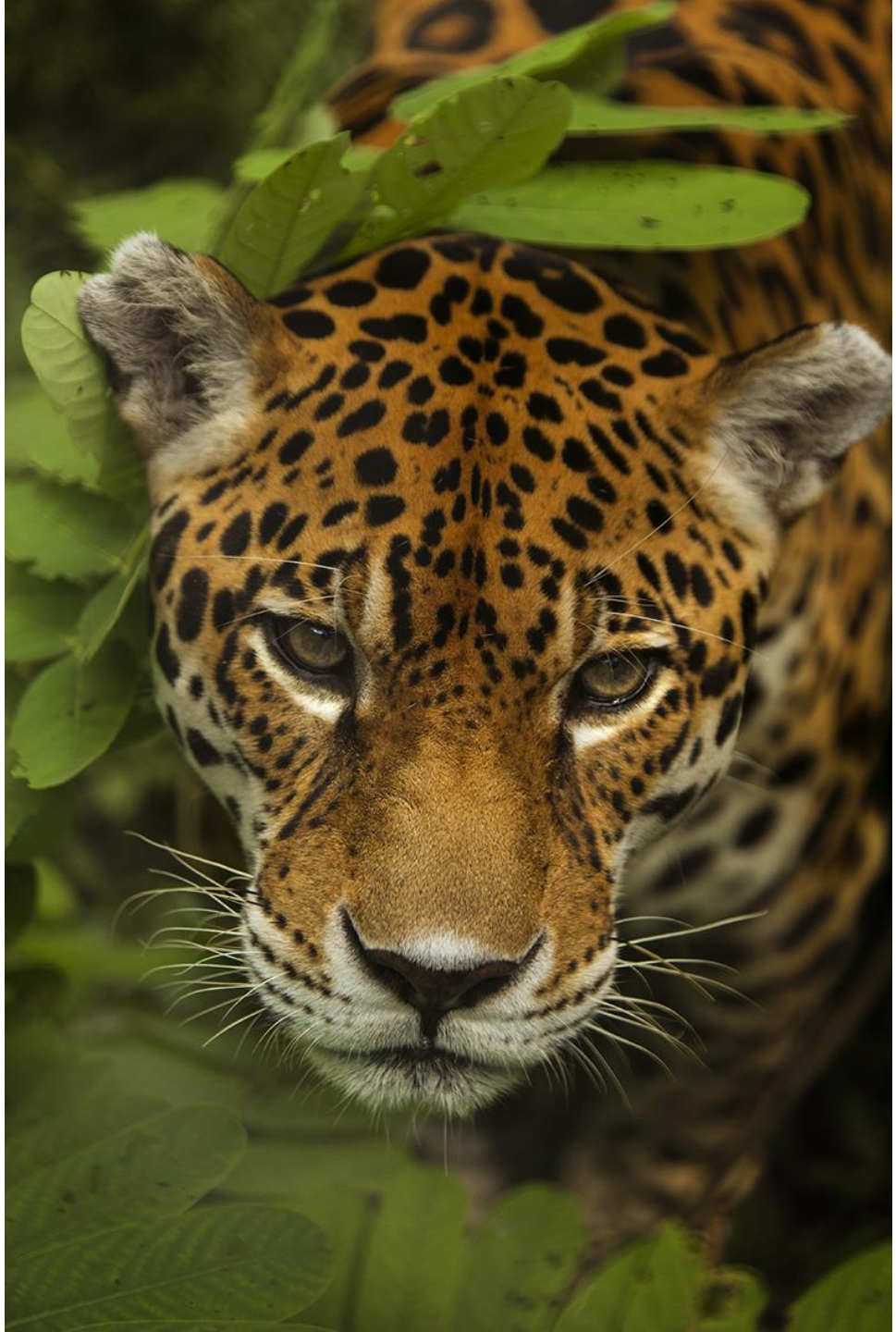
- *For the world!* Input to the Global Climate Change conference in Paris

Final question…
Did we see a jaguar?

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

Final answer:

no!



But, with supercomputers, big data
and Bayesian statistics, there will
hopefully be a next time….

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