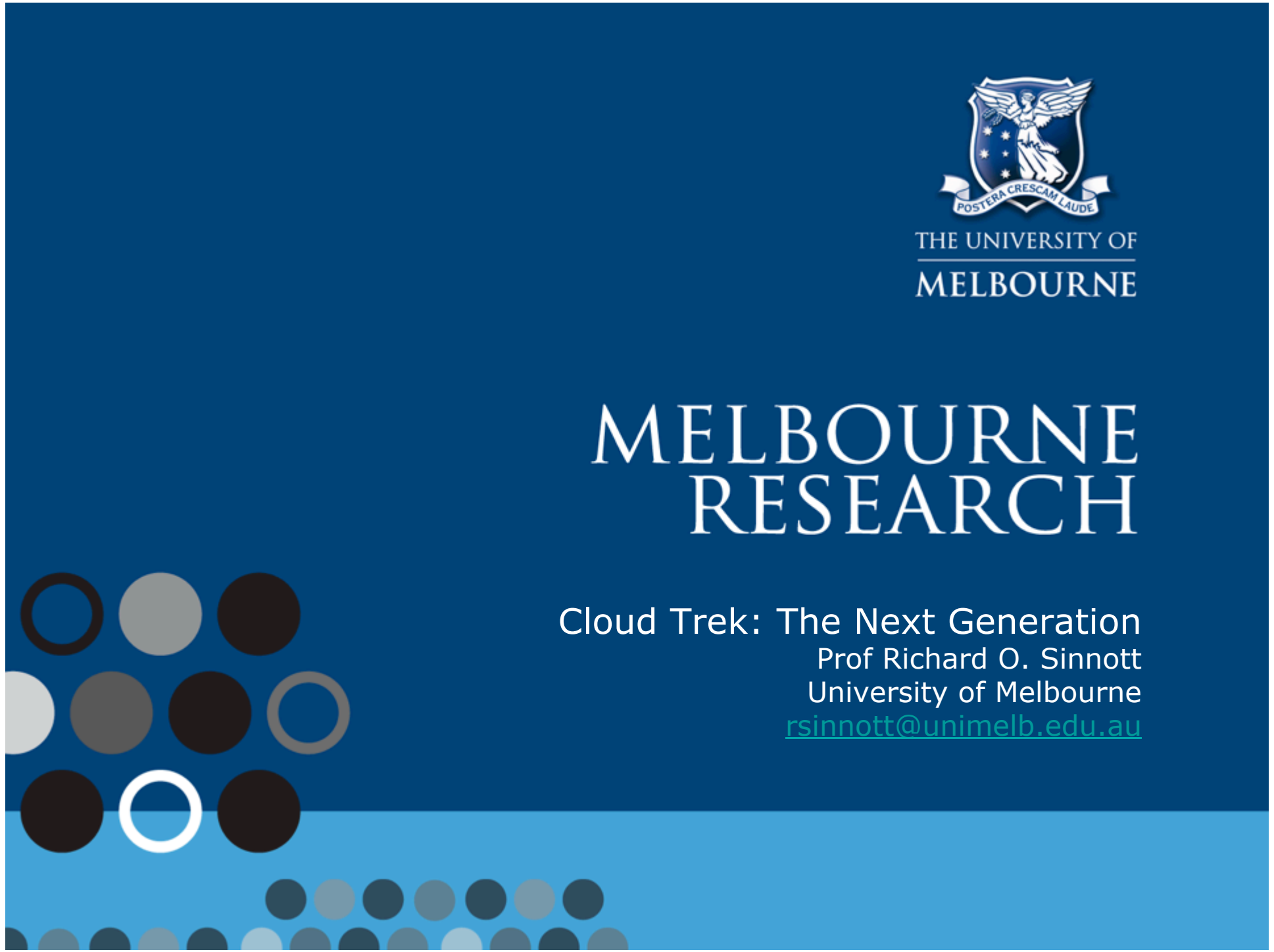


Cloud Trek: The Next Generation

Prof Richard O. Sinnott
University of Melbourne
rsinnott@unimelb.edu.au

Richard

Director, eResearch University of Melbourne

Chair in Applied Computing Systems, University of Melbourne

BSc Theoretical Physics

CEO Own Company (real time systems/telecoms)

Lecturer

Technical Director National e-Science Centre, University of Glasgow

Technical Director, Bioinformatics Research Centre University of Glasgow

PhD Distributed Systems

Post-doc GMD Fokus Berlin

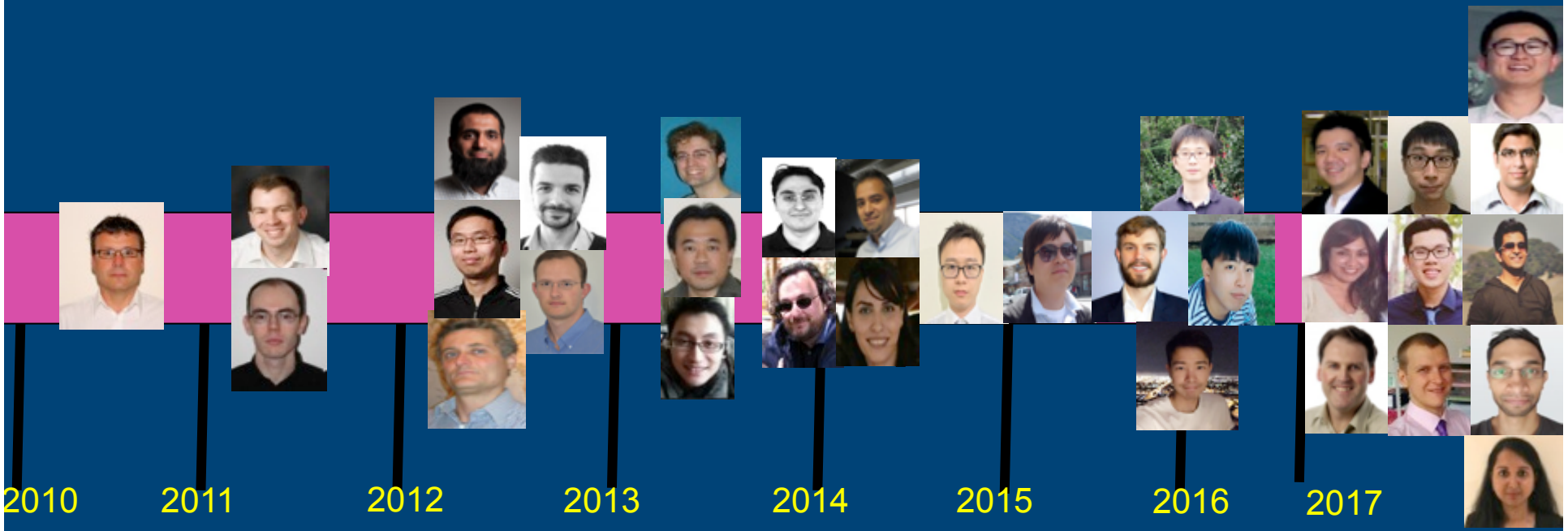
Distributed Systems Standards creator

Educator!!!

MSc Software Engineering

Melbourne eResearch Group

(http://eresearch.unimelb.edu.au)



Over $165m research funding (IARPA, DSTG, EU, NHMRC, ARC, Dept Innovation, Dept Environment, VicHealth, Commercial, …)

15 PhDs + over 250 Masters dissertations

ALL ABOUT APPLIED COMPUTING

Project Portfolio

Completed

- National e-Science Centre (I, II, III)
- Dynamic Virtual Organisations for e-Science Education
- Biomedical Research Informatics Delivered by Grid Enabled Services
- GridNet, GridNet-2
- Grid Enabled Microarray Expression Profile Search
- Glasgow early adoption of Shibboleth
- Joint Data Standards Survey
- ESP-Grid
- HPC Compute cluster award // Sun industrial sponsorship
- OGC Collision
- OMII-Security Portlets // OMII-RAVE
- Integrating VOMS and PERMIS for Superior Grid Authorization
- NCeSS
- CESSDA PPP
- Pharming of Therapeutic RNA
- Grid Enabled Occupational Data Environme
- Towards an e-Infrastructure for e-Science D
- Grid enabled Biochemical Pathway Simulat
- Virtual Organisations for Trials and Epidemi
- A European e-Infrastructure for e-Science F
- Modelling, Inference and Analysis for Biolog
- Drug Discovery Portal
- Parliamentary Discourse
- Scots Words and Placenames
- Qvolution stress management survey syste
- Advanced Grid Authorisation through Sema
- AlstromUK VRE
- Grid-enabled Virtual Safe Settings
- Clinical Streaming Transcription Software
- Enhancing Repositories for Language and
- Proxy Credential Auditing Infrastructure for
- Scottish Bioinformatics Research Network
- Generation Scotland Scottish Family Health
- Breast Cancer Tissue Biobank
- Data Management through e-Social Science (DAMES)
- Meeting the Design Challenges of nanoCMOS Electronics (nanoCMOS)
- EU FW7 AvertIT
- EU FW7 EuroDSD
- NeSC Research Platform (NRP)
- NeSC Information Network (NIN)
- ESF Network for Study of Adrenal Tumors
- Scottish Health Informatics Platform for Research (SHIP)
- National E-Infrastructure for Social Simulation (NeISS)
- EU R4SME Diagnosis of Parkinsons Disease (DiPAR)
- Automating River Pollution Detection (CAPIM)
- Endocrine genomics Virtual Laboratory (endoVL)
- DSDNetwork Australasia

Subset of On-Going

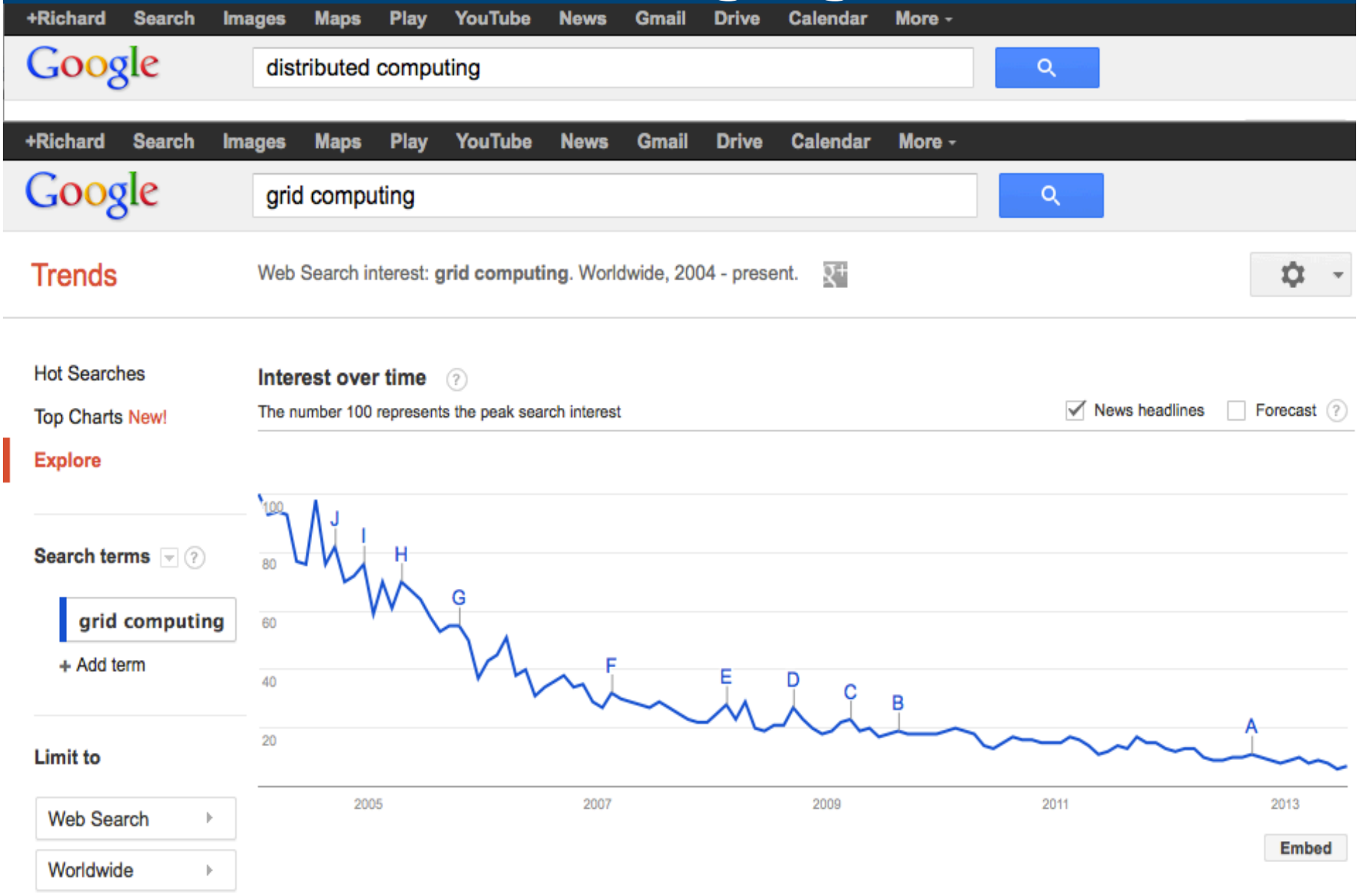
- EU European Platform for Study of Wolfram, Alstrom, Bardet Biedl (EuroWABB)
- Multicenter prospective study of biochemical profiles of monoamine-producing tumors (PMT Study)
- European Society of Hypertension Study on Pheo/PGL
- International DSD
- EU FW7 European Network for Study of Adrenal Tumors Cancer Research Platform (ENSAT-CANCER)
- VicHealth Health Indicators and Spatial Objective Data
- National Spinal Injury Research Platform
- Australian Urban Research Infrastructure Network (AURIN)
- Epilepsy e-Learning portal
- Type-1 Diabetes study of environmental factors on onset of T1D
- Australian Diabetes Data Network (ADDN)
- International Niemann-Pick A, B and C Registry

- Australian Diabetes Data Network – Phase II (ADDN2)
- Helicopter advanced training system, Australian Department of Defence
- Public Records Office Victoria Data Management Solutions
- Complex System Modelling Platform and GPU utilisation
- Spinal Cord Research Hub (SCORH)
- VicHealth 2016 Indicators API
- Helicopter advanced training system Phase II, Australian Department of Defence
- Twitter data analytics for business
- Mobile Applications for Patients with Neuroendocrine Tumours
- Systems Genomics Support Platform
- SWARM: Smartly-aggregated Wiki-style ARgument Marshalling (SWARM)
- ORCA Cognitive Assessment Platform
- VicSpin Victoria-wide Flu Surveillance System
- ElectraNetLIDAR/VectorNZ Lidar

The Buzz


BIG
DATA
INFORMATION
MANAGEMENT
TECHNOLOGIES
ANALYSIS
RESEARCH
SEARCH
INTERNET
PETABYTES
EXABYTES
STORAGE
LARGE
SETS
MILLION
SOFTWARE
VOLUME
PROCESS
GOVERNMENT
SAN
NAS
SOCIAL
SCIENCE
TECHNOLOGY
DEVELOPMENT
FUTURE
NEW
BILLION
PARALLEL
LEARNING
WORLD
BASED
EVERY
AMOUNT
SIZE
DATABASES
PROCESSING
ANALYTICS
SIMULATIONS
FRAMEWORK
BUSINESS
SYSTEMS
FACTOR
ONLINE
DAY
MASSIVE
DEFINITION
BETWEEN
STATISTICS
USING
VISUALIZATION
EFFECTIVE
DEPARTMENT
QUERIES
HIGH
USERS
REAL
STRUCTURE
WORK
RATE
APPLICATIONS
INFRASTRUCTURE
HIGHER
ALGORITHMS
COMPANIES
DETERMINE
WELL
SET
RESULTS
YEARS
DRIVE
REPLICATION
HANDLE
SOURCES
CHALLENGES
LIMITS
COMPLEX
CAPACITY
HAND
ONE
PAST
ANNOUNCED
OUT
INCLUDE
NETWORKS
HYPOTHESIS
END
MOST
NEARLY
MAPREDUCE
CRITIQUES
STATE
CRITIQUE
PEOPLE
TRAFFIC
USE
CURRENTLY
MANY
APPROACH
INITIATIVE
ECONOMIC
TERABYTES
MARKET
VARIETY
RELATED
TIME
SENSOR
WOULD
PROJECT
ANNUAL
NOW
DISTRIBUTED
RELEVANT
USED
PER
SECOND
NATIONAL
FLOW
INTELLIGENCE
PRIVACY
INTERNATIONAL
IMPLEMENTATION
BECOMES
APPROACHES
PROCESSES
PARADIGM
RECORDS
COST
FUNDING
INSIGHT
TOOLS
CENTER
ARCHITECTURE
PRIVATE
TERMS
YEAR
CAPABILITIES
SUCCESSFUL
NEEDED
USES
COMPARED
SHARED
LESS
HUGE
TIMES
VARIOUS

The not so long ago buzz...



The latest buzz...

+Richard Search Images Maps Play YouTube News Gmail Drive Calendar More

Google cloud computing

+Richard Search Images Maps Play YouTube News Gmail Drive Calendar More

Google big data

+Richard Search Images Maps Play YouTube News Gmail Drive Calendar More

Google data analytics

Trends

Web Search interest: **data analytics**. Worldwide, 2004 - present.

Hot Searches

Top Charts New!

Explore

Search terms

data analytics

+ Add term

Limit to

Web Search

Worldwide

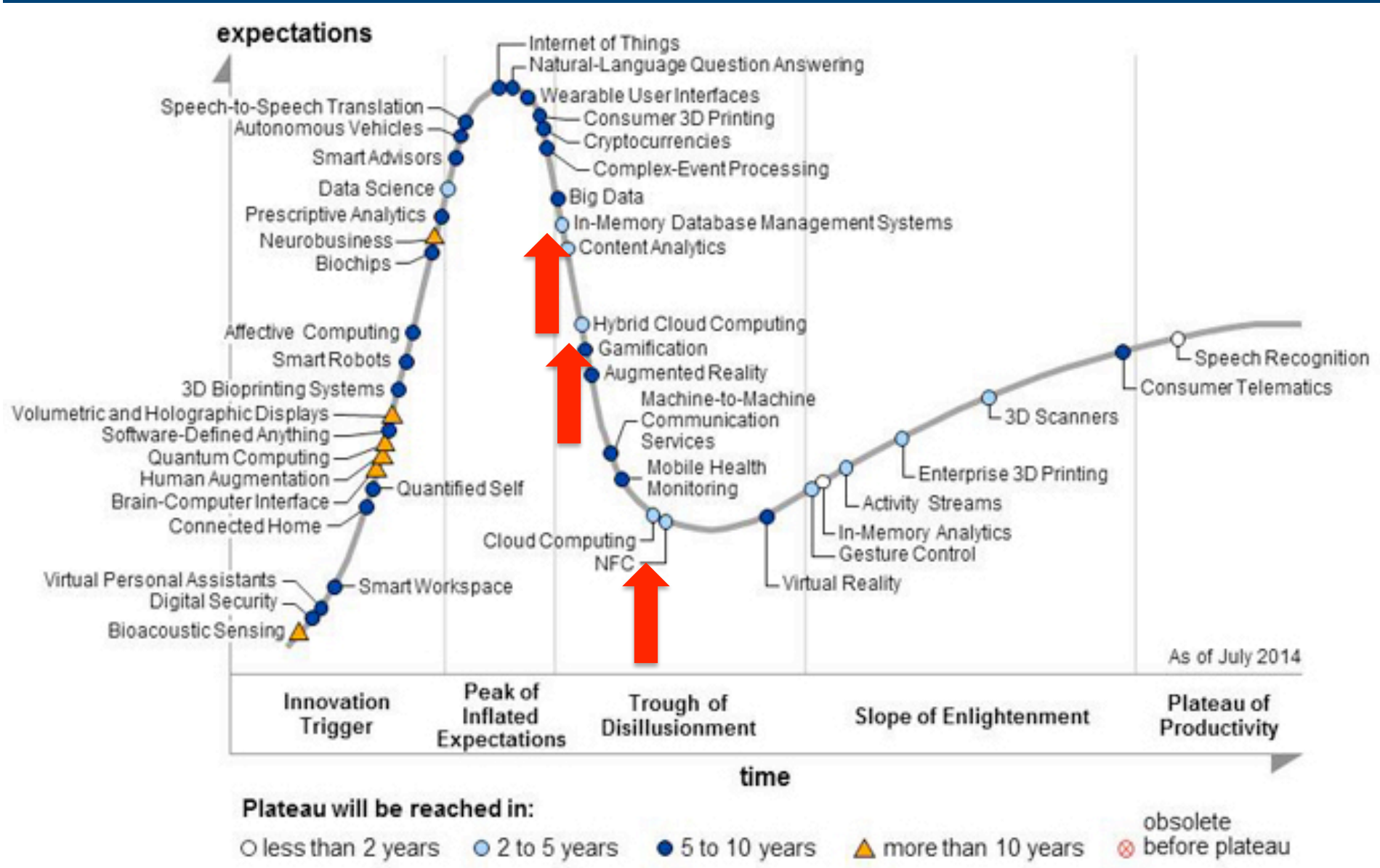
Interest over time

The number 100 represents the peak search interest

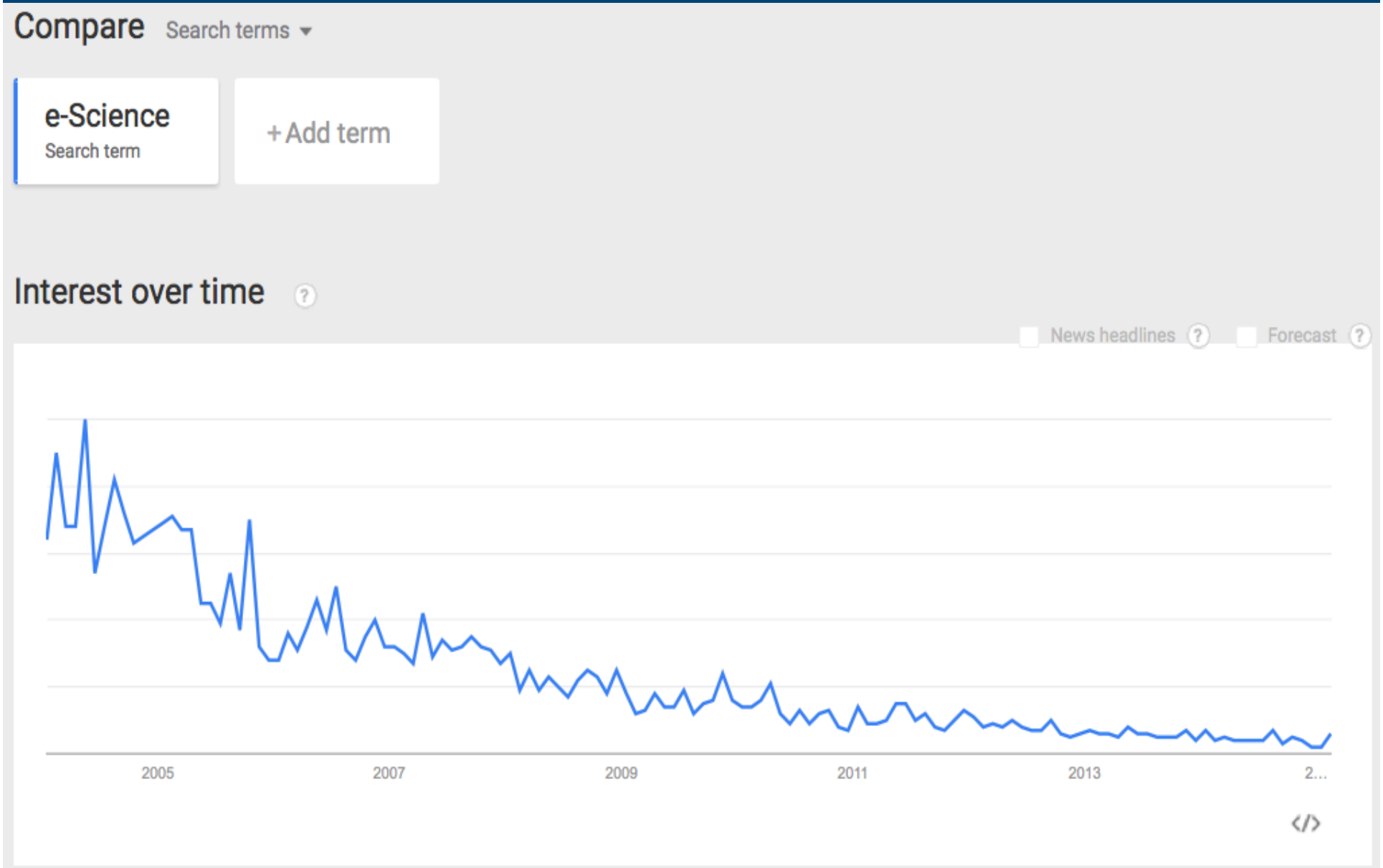
News headlines Forecast

Embed

The Hype Cycle... (Gartner 2015)

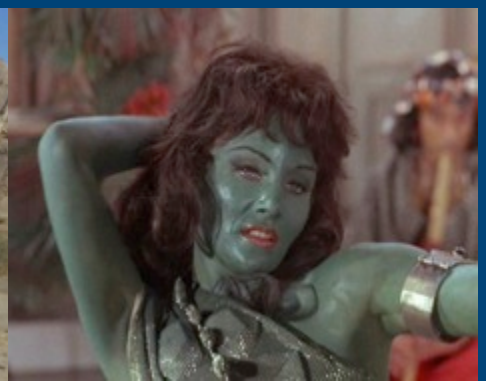


Not just IT trends…



Pedagogy?

- What to teach Computing Science students to support "research" communities?
  - Aka creating e-Scientists, e-Researchers, …. "e-"
  - Ideally, technology future proof'ing



?

Teaching "e-" (Circa 1997)

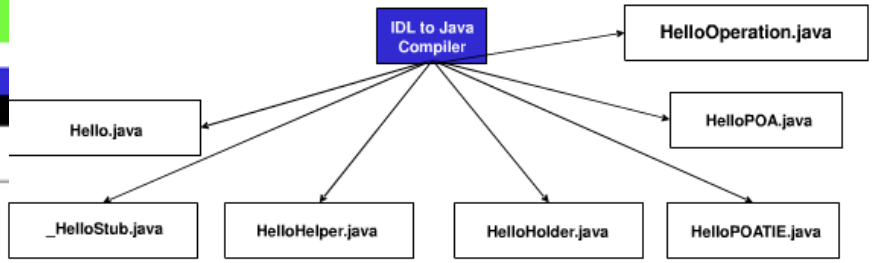
- Distributed systems
  - transparency and heterogeneity of computer-computer interactions
  - finding/discovering resources (trader!)
  - binding to resources in real time
  - run time type checking
  - invoking resources
  - dealing with heterogeneity of systems
    - applications and operating systems

Client.java (Simplified)

Step of Defining IDL

1. Write the IDL as an interface to the object
2. Compiling Hello IDL

```
$ idl -ir -d generated hello.idl
```



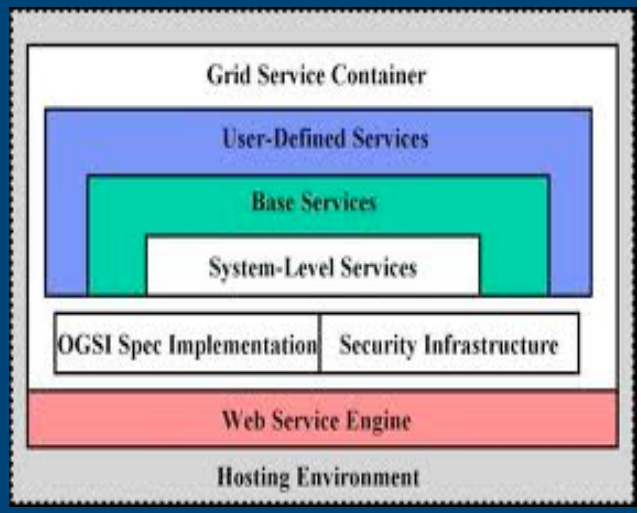
Server.java (Simplified)

Teaching “e-” (circa 2002)

- Globus Toolkit Project – www.globus.org
  - GT2 - Complex software system for large, scale distributed software systems development
  - MANY MB of source code
    - Many software engineers worked in making this
      - and many more in making it work!!!

Teaching "e-" (circa 2004)

- Move to service-based approach
  - Open Grid Services Infrastructure
- GT3 – core technologies re-factored as "Grid Services"
  - stateful Web services
  - extension of Web services interfaces
  - asynchronous notification of state change
  - references to instances of services
  - collections of service instances
  - service state data augmenting constraints of XML Schema definition

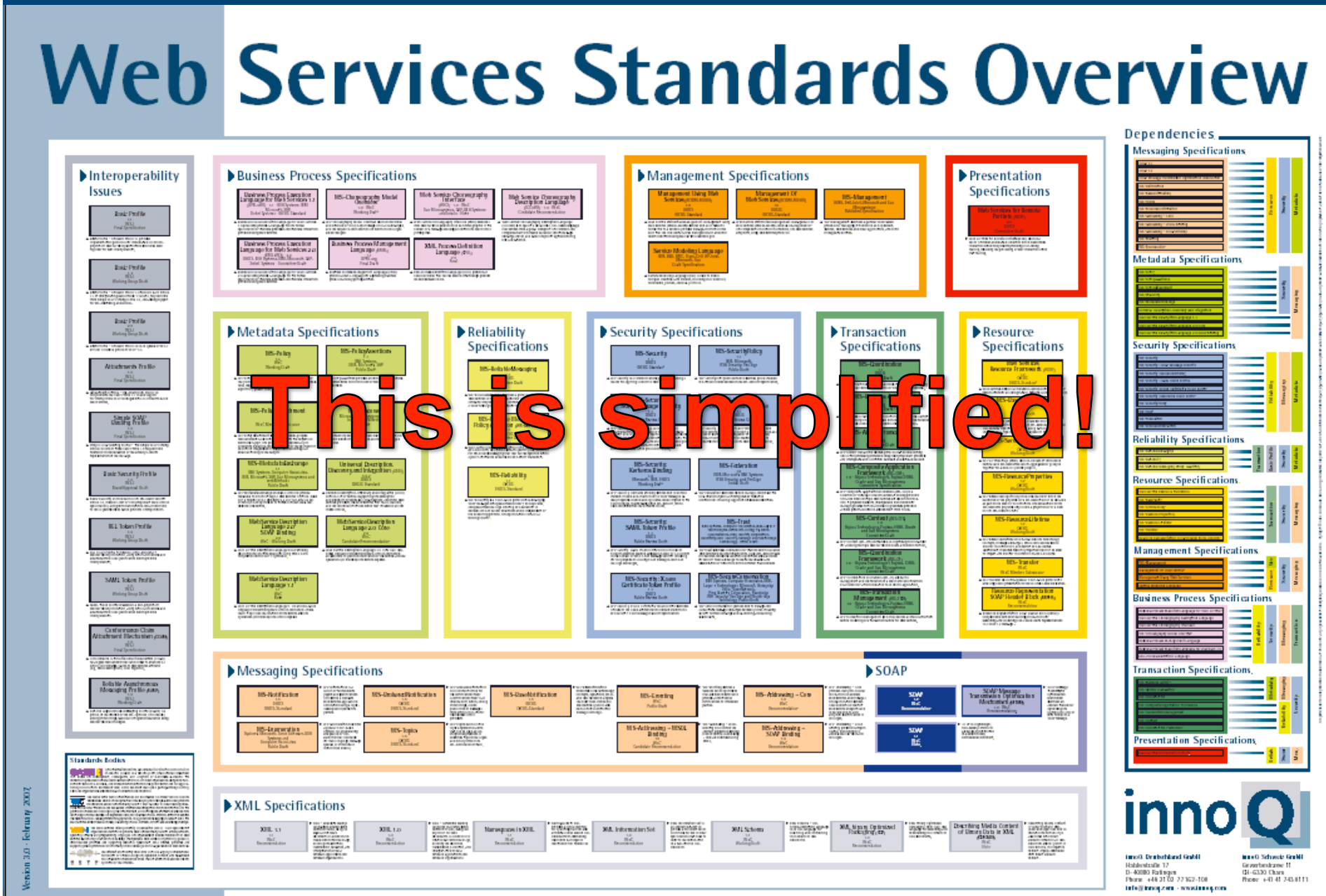


Teaching "e-" (Circa 2009)

- GT4 services
  - Based on Web service resource framework (WSRF)
    - Many of my software engineers hardened their skill sets using this - fair to say that my Glasgow students suffered!!!

- Also MANY other standards and efforts …
  - Business and commercial drivers
    - Vendors shaping standards to their commercial advantage

Flux of Web Service Standards



Common Research Needs -> Teaching "e-"

- It is a fact that:
  - all researchers are now generating more data than ever before
  - all researchers need access to more data than ever before
- They need tools, methodologies to
  - Search for
  - Unlock/Use
  - Analyse
  - Integrate
  - Track
  - Store
  - Destroy
  - Move
  - Check authenticity of
  - Visualise
  - Overcome heterogeneity of
  - Overcome distributed nature of
  - …

  } data

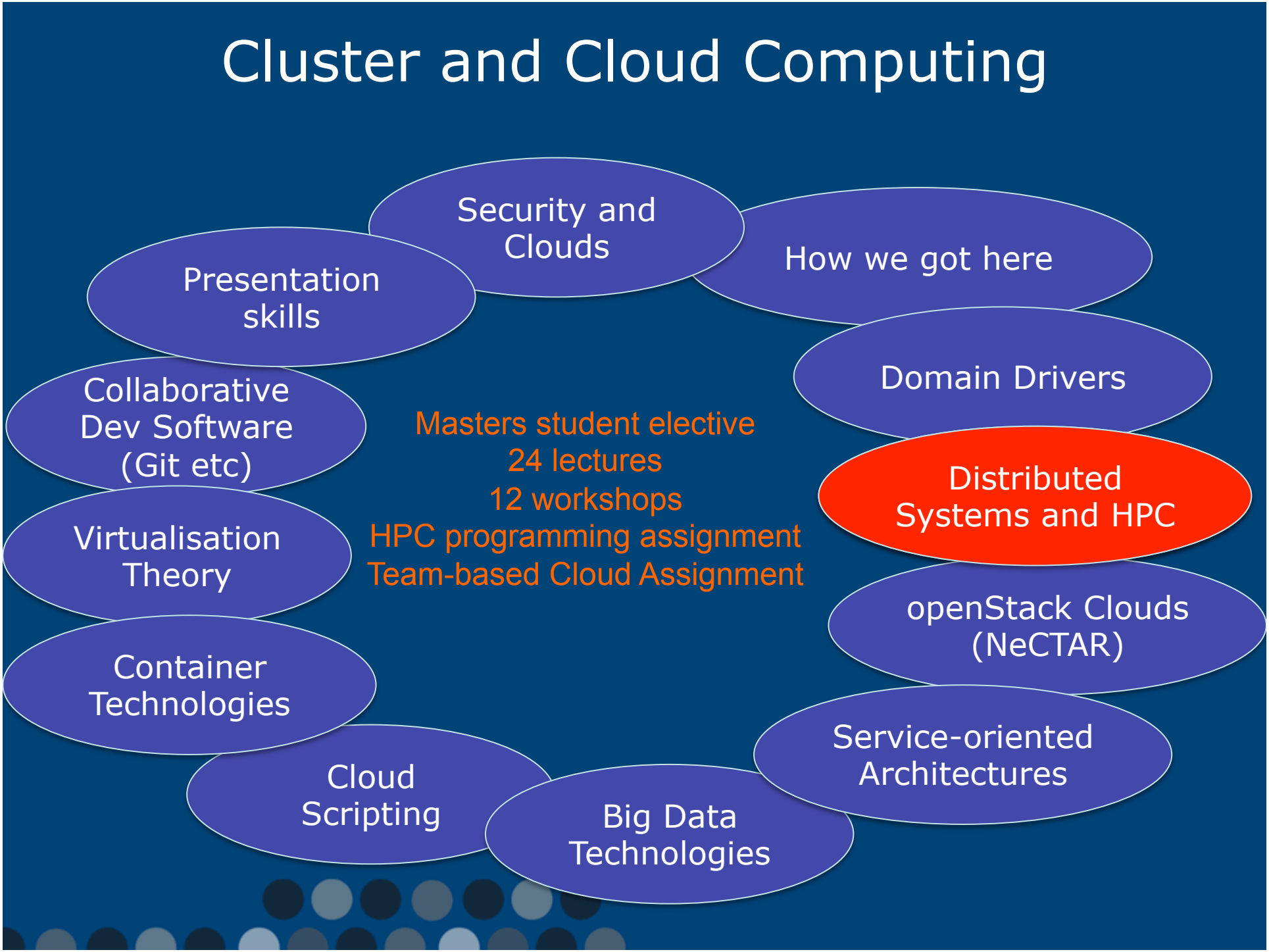
… and this should be tailored to the researcher needs!!!

e-Folk need to be proficient at delivering and supporting these capabilities!!!

Teaching “e-” (2013-17)

- Driven by hands-on experiences derived from many live/production level projects (and lessons learned!)
  - Does the technology work?
    - Is it more pain than gain…?
    - What are it’s limitations?
  - Does it have a lifetime?
- Cluster and Cloud Computing (COMP90024)
  - 55 advanced level computing science students (2013)
  - 71 advanced level computing science students (2014)
  - 139 advanced level computing science students (2015)
  - 147 advanced level computing science students (2016)
  - 197 advanced level computing science students (2017)
  - 200+ (2018)

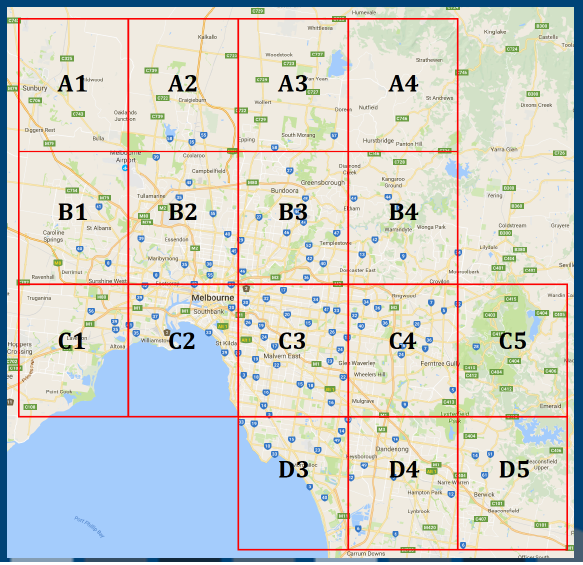
Cluster and Cloud Computing



High Performance Computing

- HPC 3,000 node cluster (UniMelb)
  - Basic HPC/MPI training; Torque/Maui -> Slurm
    - 2014 - Searching/sorting a range of (large) literary works
      - Shakespeare; Tolstoy; Dickens; Verne
    - 2015 – Word concordance (counting)
      - Collection of 100Mb, 1Gb, 5Gb text files analysed
    - 2016 – 10Gb Twitter (JSON) file
      - Most frequent tweeter, trending topic, word search
    - 2017 – 10Gb Twitter (JSON) file
      - Which part of Melbourne tweets the most

1 node/1 core,
1 node/8 cores,
2 nodes/4 cores



(Java/Python/Scripts/C/C++/…)

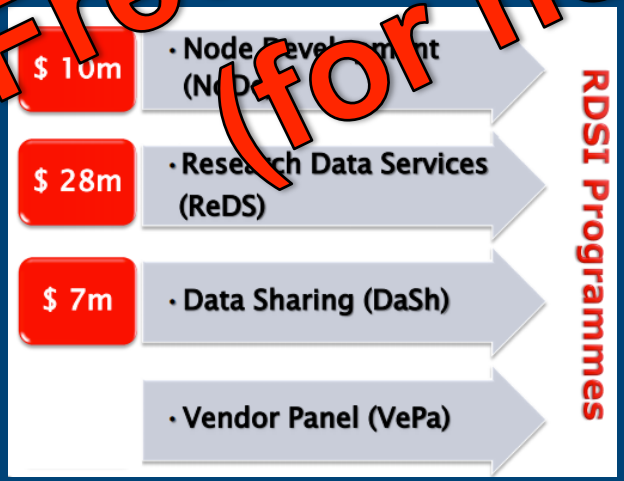
Research Cloud Infrastructure

- National eResearch Collaboration Tools and Resources (NeCTAR – www.nectar.org.au)
  - Federally funded project (~$70m)
  - Lead by University of Melbourne
    - Research Cloud Program
      - OpenStack IaaS
      - 4Gb-64Gb (linux flavours)
      - 30,000 physical servers available across eight availability zones
    - Virtual Laboratories Program
      - Astro,
      - Genomics,
      - Humanities,
      - Climate,
      - Nano-,
      - …endocrine genomics

Free Resource (for now)

Research Data Infrastructure

- Research Data Services (RDS – www.rdsi.edu.au)
- National project to establish data storage resources across Australia
  - ~100 Petabytes national data storage
  - Victoria Node (VicNode)
    - UniMelb, UniMonash for Vic-wide "nationally significant data set"

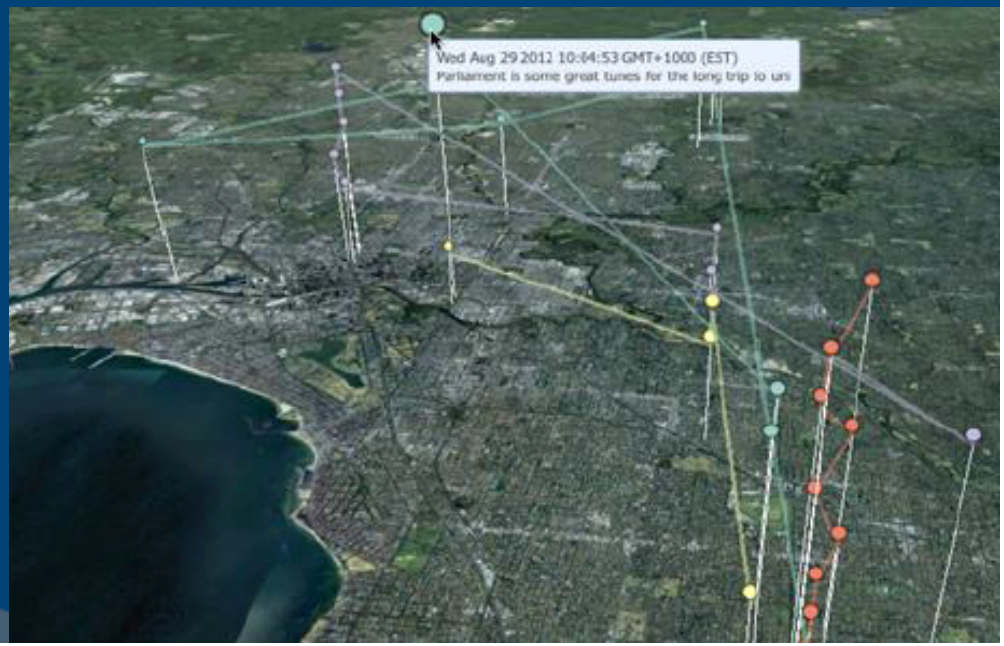


- Used by many diverse communities

Free Resource (for now)

Interesting, Open, Relevant, Creative Data-driven Teaching Scenarios

- Social media, e.g. Twitter
  - 320m+ active users per month
    - 2.9m user accounts in Australia alone
    - (Near) real-time and often geo-located
  - However noisy, erroneous tweets
    - Up to 140 character string -> 9kb data
  - Privacy
    - Data vs metadata
  - Open APIs
    - Search & Streaming
    - Rate limited
  - Big data analytics
    - infrastructure



Cluster and Cloud Computing

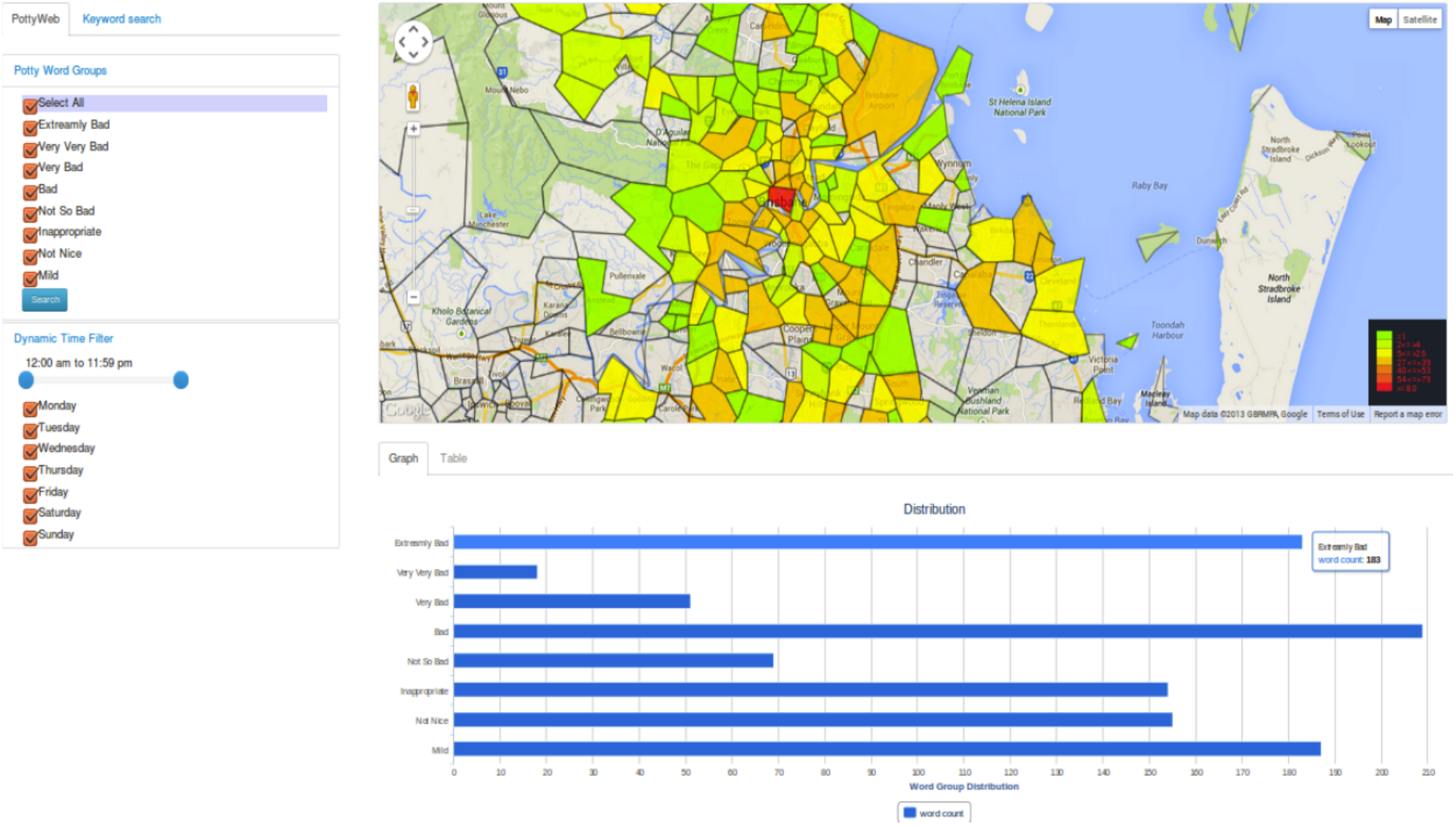
- NeCTAR Research Cloud
  - Assignment 2013
    - Harvest Tweets from Twitter from Australian cities and tell stories
      - Sentiment and topic analysis
        - » 4m -> 150k tweets
      - Analysed with couchDB, MapReduce and using multiple VMs
        - » Each team 8 Medium instances, 100Gb Volume + 100Gb Object storage
      - Application required to scale over the Cloud
        - » dynamic deployment/configuration
        - » Boto, fabric, juju, heat, …

Project: BigTwitter

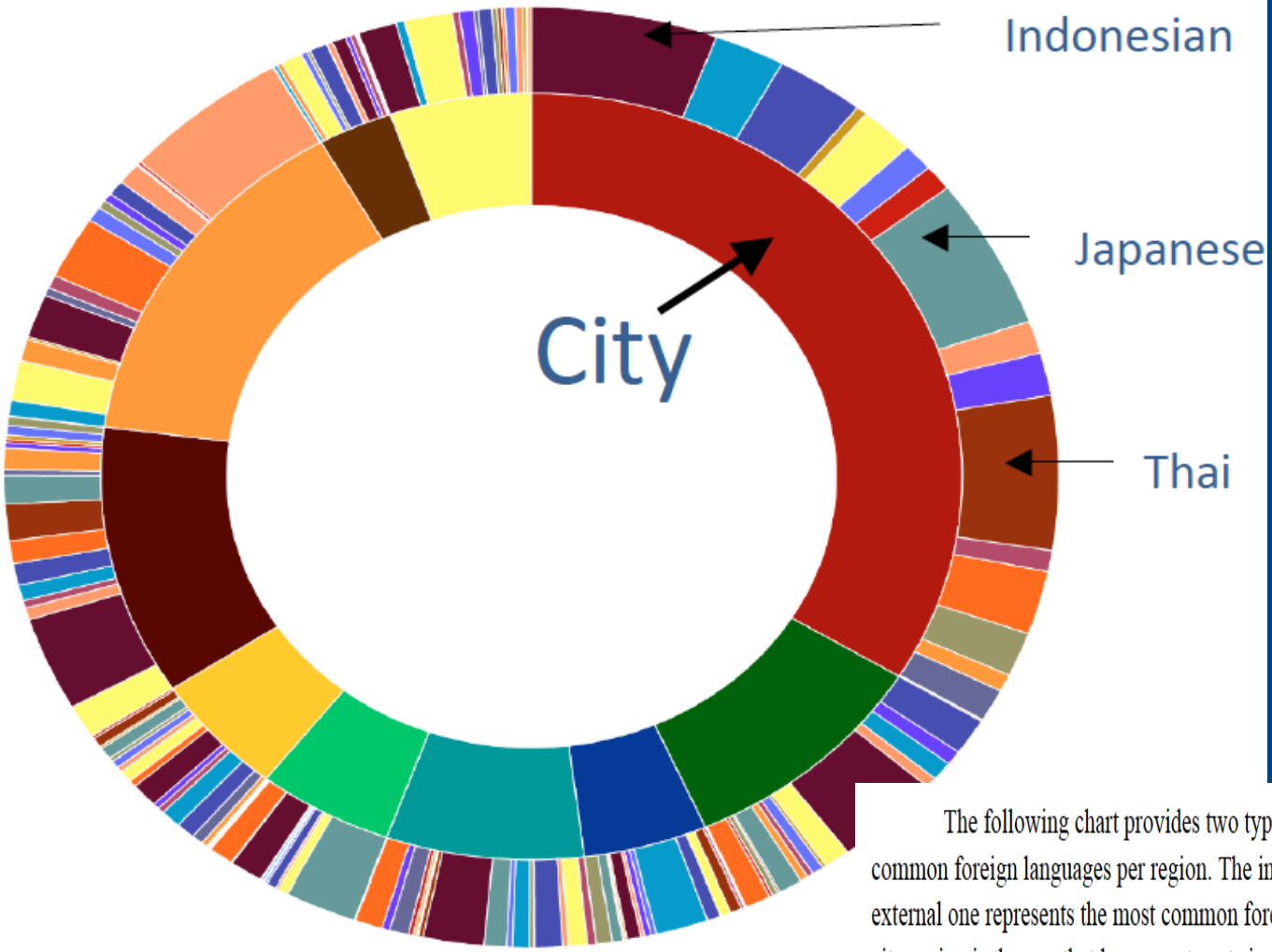
AURIN-workflow
Endo_VL
French_Twitter_Team
InternationalTwitterHarvesting-Team1
InternationalTwitterHarvesting-Team10
InternationalTwitterHarvesting-Team11
InternationalTwitterHarvesting-Team12
InternationalTwitterHarvesting-Team13
InternationalTwitterHarvesting-Team14
InternationalTwitterHarvesting-Team2
InternationalTwitterHarvesting-Team3
InternationalTwitterHarvesting-Team4
InternationalTwitterHarvesting-Team5
InternationalTwitterHarvesting-Team6
InternationalTwitterHarvesting-Team7
InternationalTwitterHarvesting-Team8
InternationalTwitterHarvesting-Team9
TeachingCloudComputing-Team1
TeachingCloudComputing-Team10
TeachingCloudComputing-Team2
TeachingCloudComputing-Team3
TeachingCloudComputing-Team4
TeachingCloudComputing-Team5
TeachingCloudComputing-Team6
TeachingCloudComputing-Team7
TeachingCloudComputing-Team8
TeachingCloudComputing-Team9
Unimelb_National_Twitter_Harvesting
UoM_TeachingCloudComputing
pt-84

Cluster and Cloud Computing

- Examples…

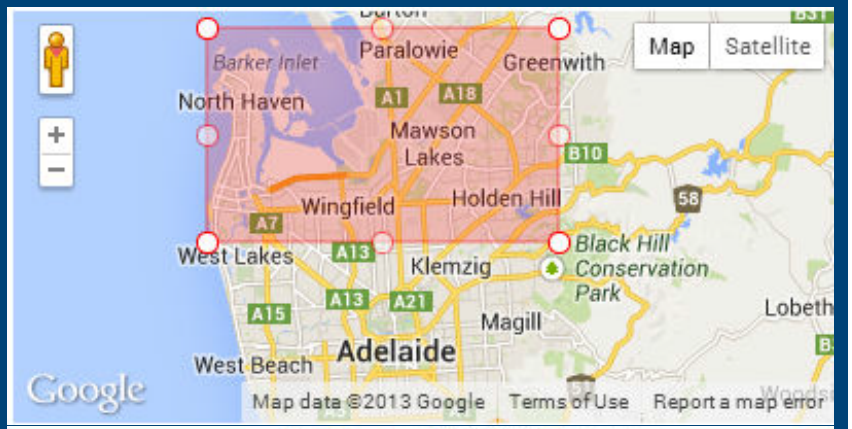


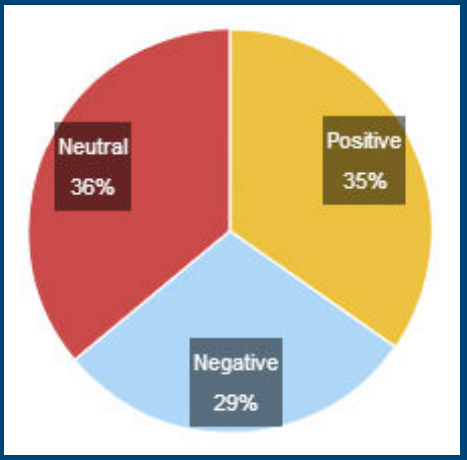
Sydney Language Tweeting

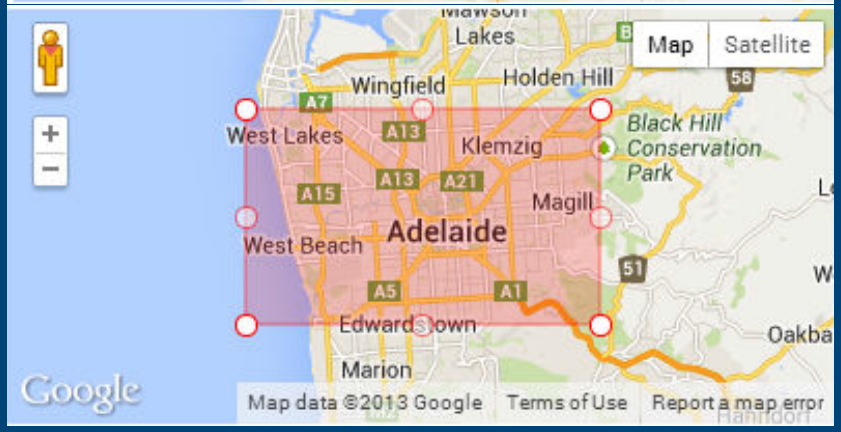


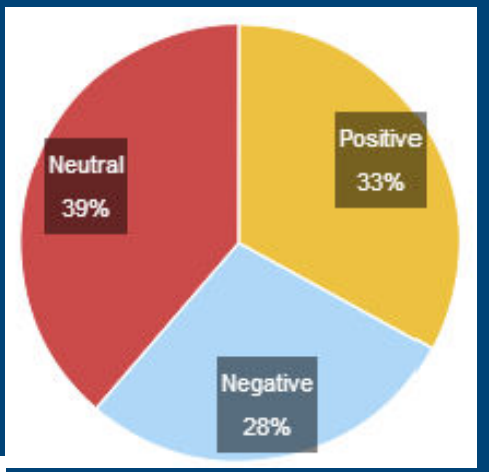
The following chart provides two type of information: The total of tweets and the most common foreign languages per region. The internal ring represents each region of Sydney and the external one represents the most common foreign languages in a particular region. For instance, the city region is the one that has more tweets in Sydney (presented in red colour) and the most common languages there are: Indonesian (violet), Japanese (grey) and Thai (brown). It is interesting to observe that in different regions of Sydney, different foreign languages predominate, for example Spanish (pink) is spoken more in the West (orange).

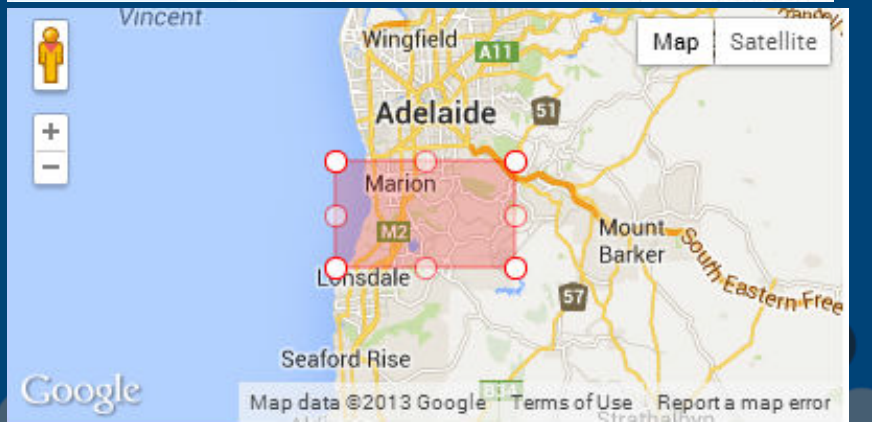
Adelaide Geo-Sentiments Classification

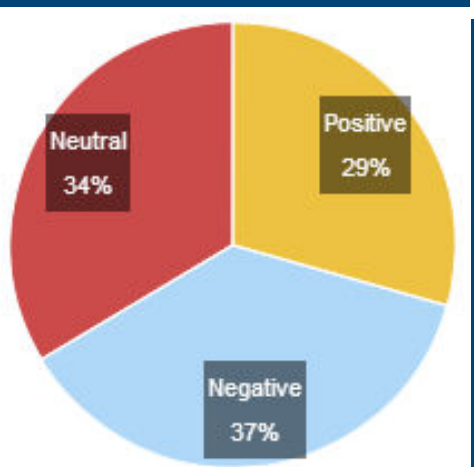




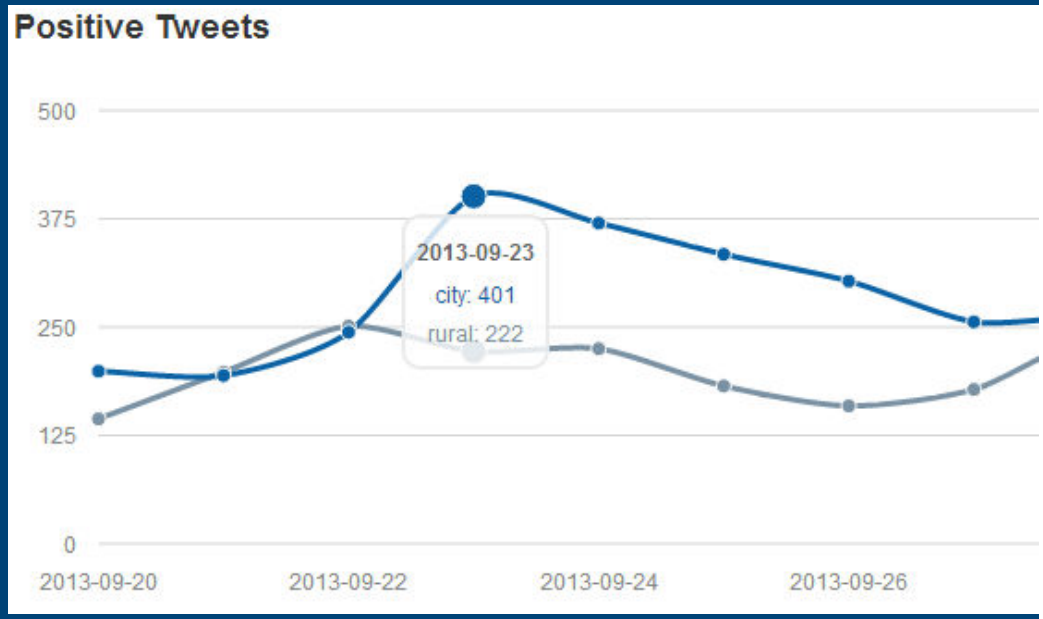








Adelaide One Direction Sentiment Impact



6.3 One Direction Makes Everybody Happy

Kids living in the digital generation tend to control the ebb and flow of trending topics in social media. We looked closely at a date range where there seemed to be an unusual spike of positive tweets. It seems that One Direction's three-day stop in Adelaide would tend to make a lot of their fans elated.

Cluster and Cloud Computing

- NeCTAR Research Cloud
  - Assignment 2014 - BIG(ger) Data for 14 teams
    - Global (English Speaking) Cities
      - *Sydney, New York, Toronto, Philadelphia, Los Angeles, Chicago, Houston, Brisbane, Perth, Boston, Dublin, San Francisco, London, Washington*
      - And compare to Melbourne: is Melbourne the world's most livable city?
        - » (It is!!!)
    - Each team 8 medium VMs, 250Gb volume + 100Gb object storage
      - 2m -> 10m tweets
    - Application required to dynamically scale over the Cloud
      - Ansible, Boto

Challenges when data sets grow!

Cluster and Cloud Computing

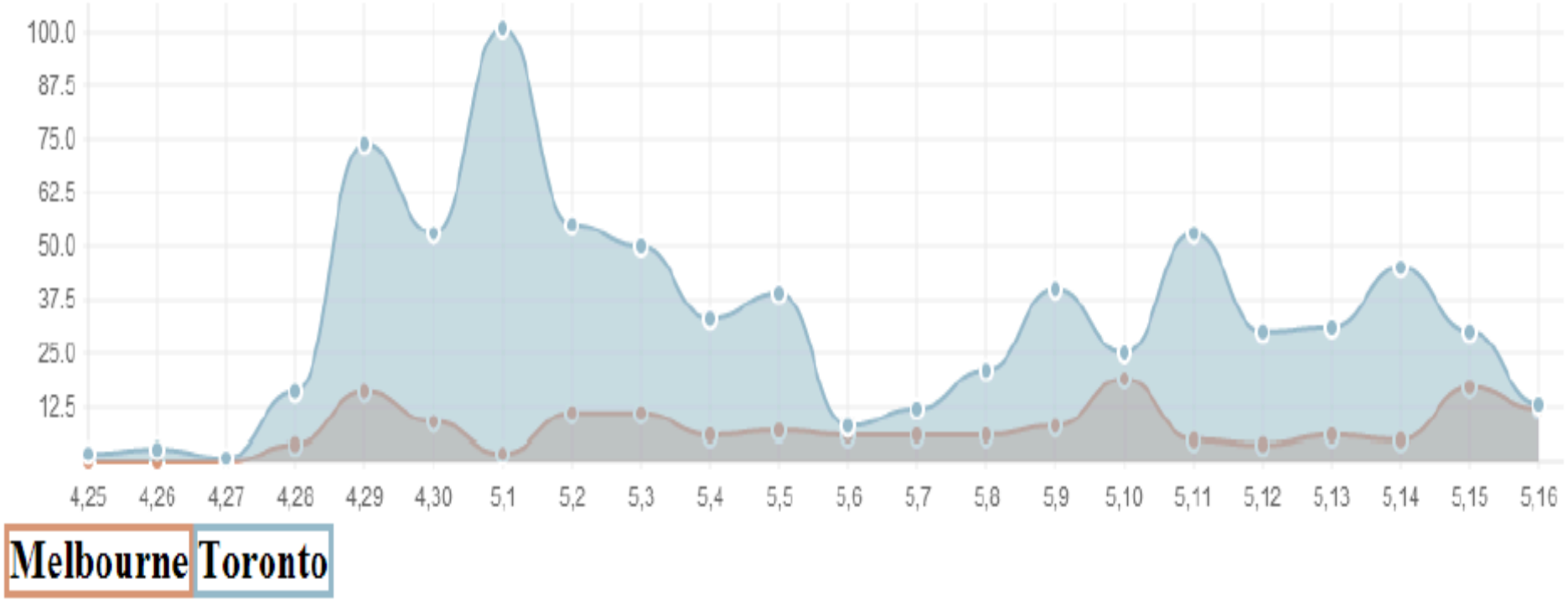
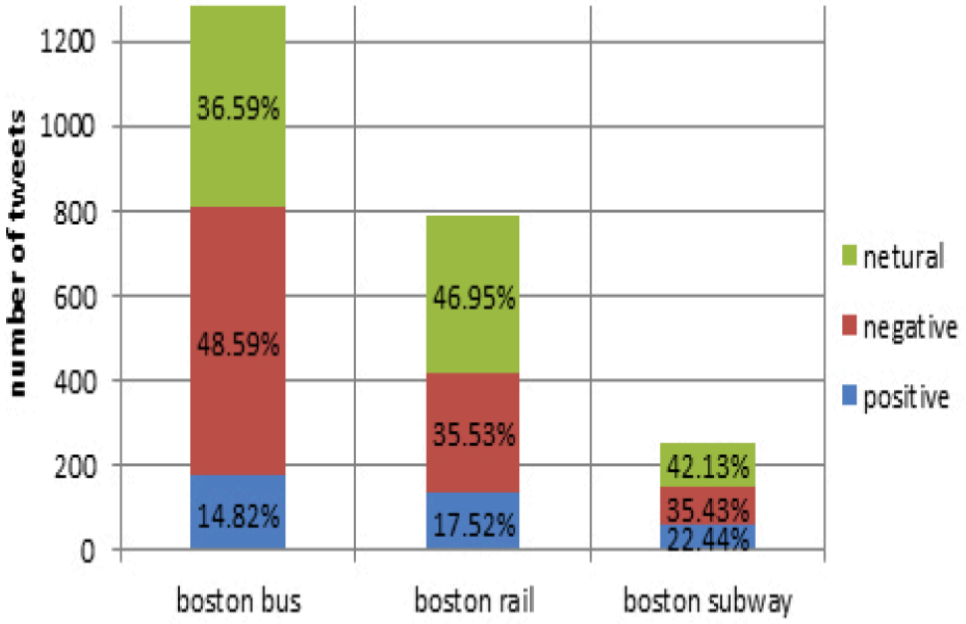
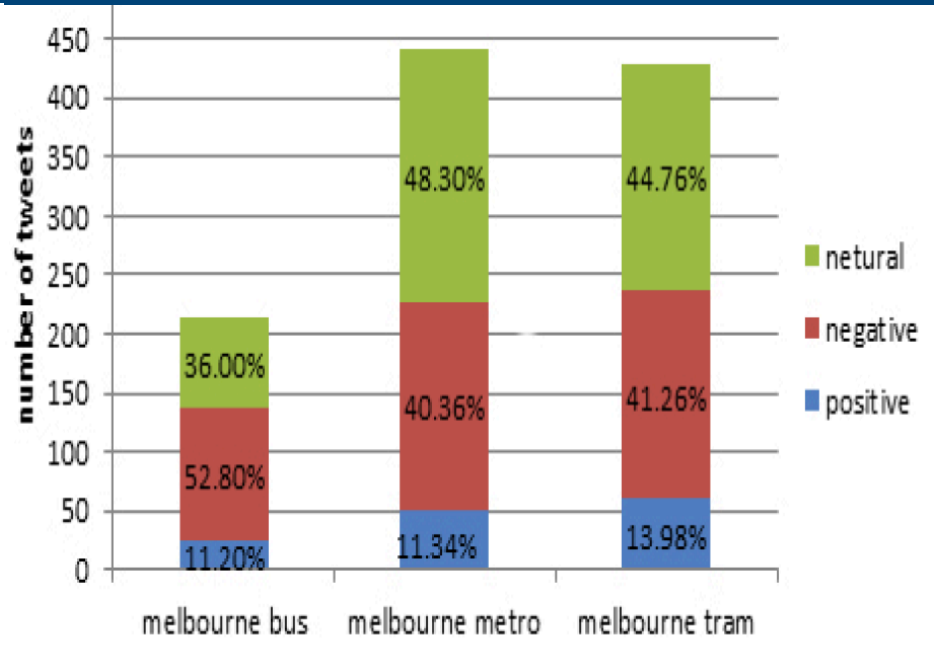


Figure 11 Sentiment and weather correlation analysis

Cluster and Cloud Computing

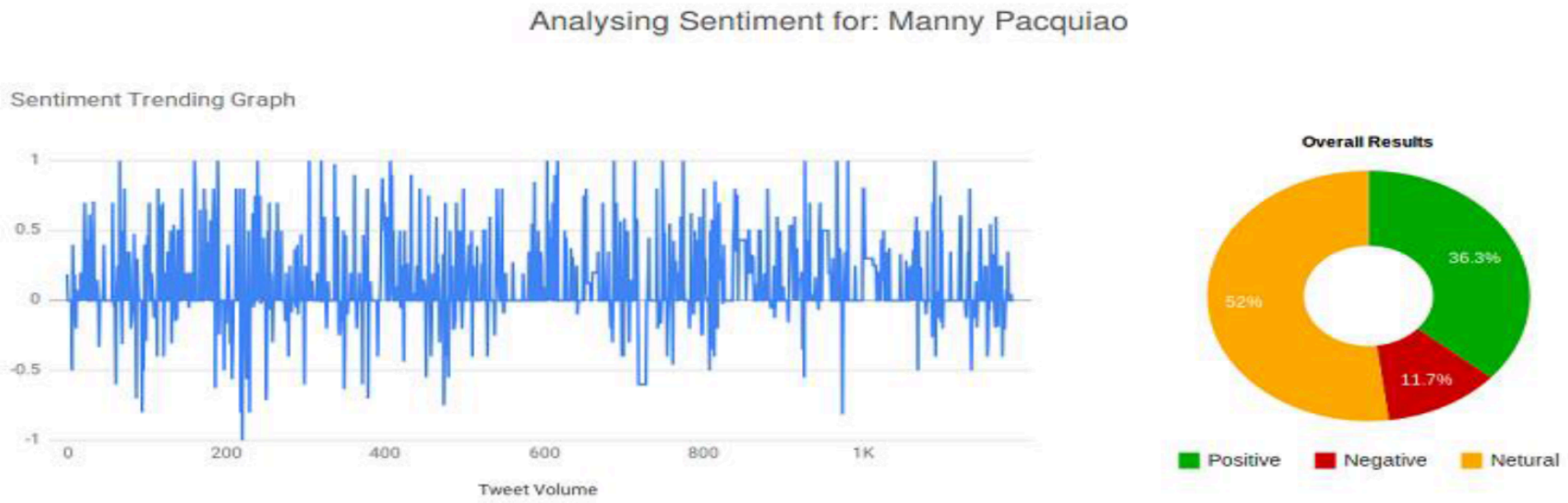


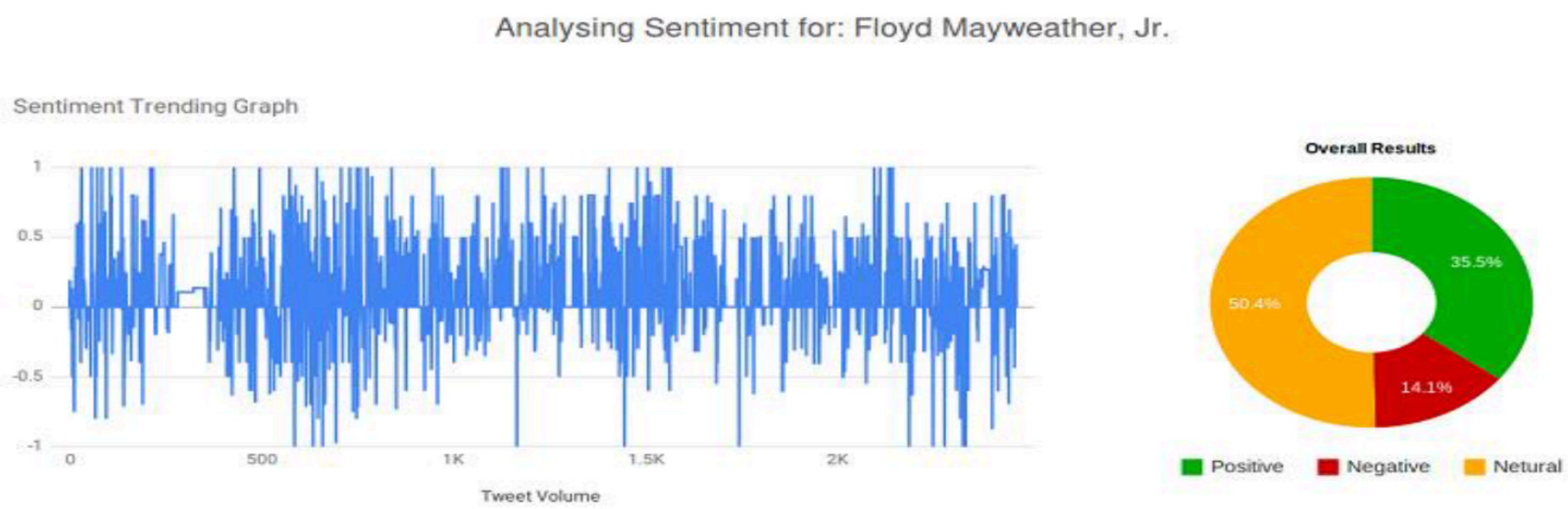


Cluster and Cloud Computing

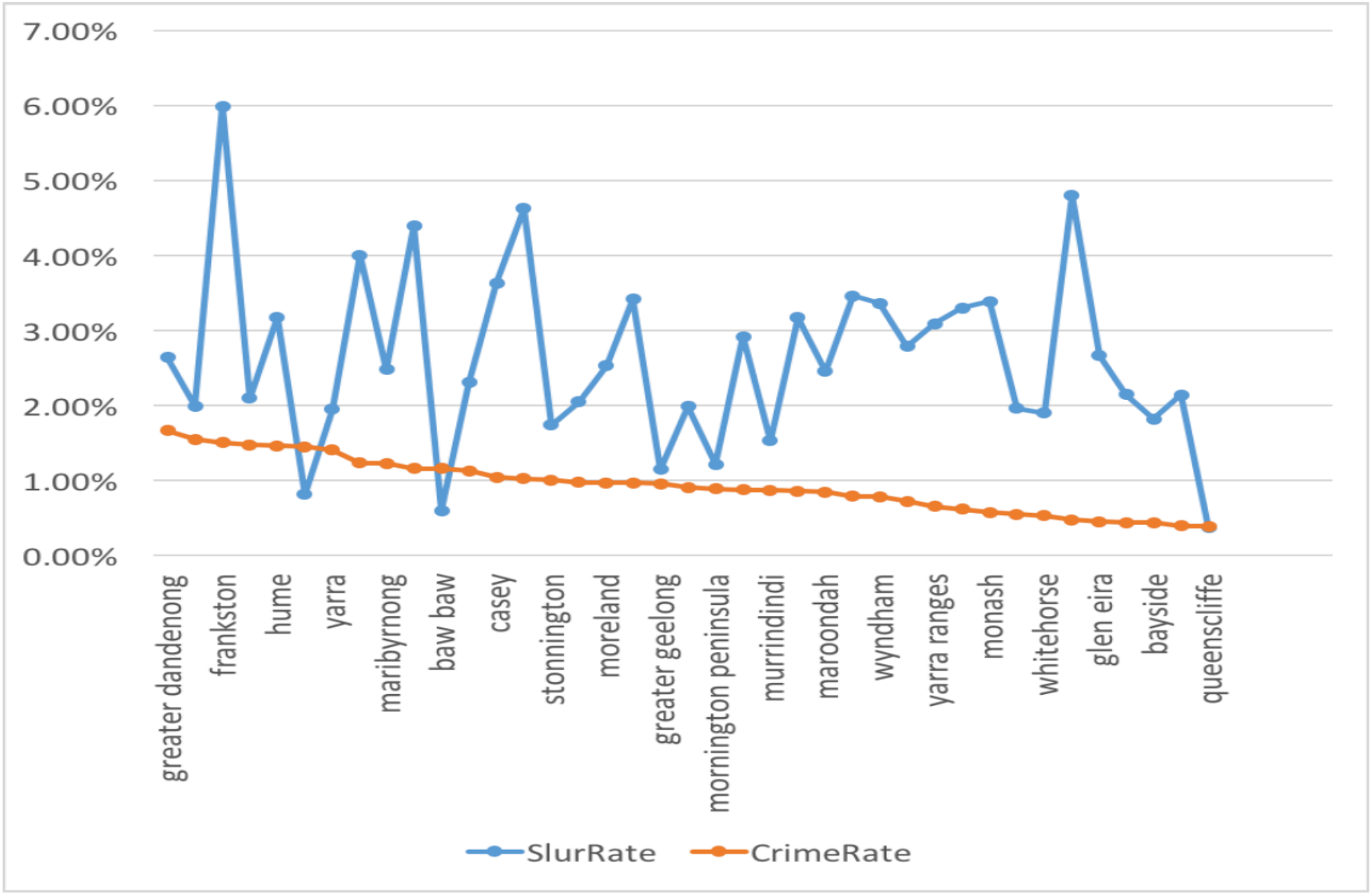
- NeCTAR Research Cloud
  - Assignment 2015 – Twitter analysis
    - 28 Global (English Speaking) Cities
      - *Adelaide, Atlanta, Birmingham, Boston, Brisbane, Chicago, Dallas, Detroit, Dublin, Edinburgh, Glasgow, Houston, London, Los Angeles, Melbourne, Miami, Montreal, New York, Perth, Philadelphia, Phoenix, Singapore, San Antonio, San Diego, San Francisco, Sydney, Toronto, Washington*
    - Each team 8 medium VMs, 250Gb volume + 100Gb object storage
    - Application required to scale over the Cloud (Ansible, boto, chef)
  - 100 million+ Tweets collected from Australian Cities
    - Data challenges driving the technological choices and experiences

Cluster and Cloud Computing

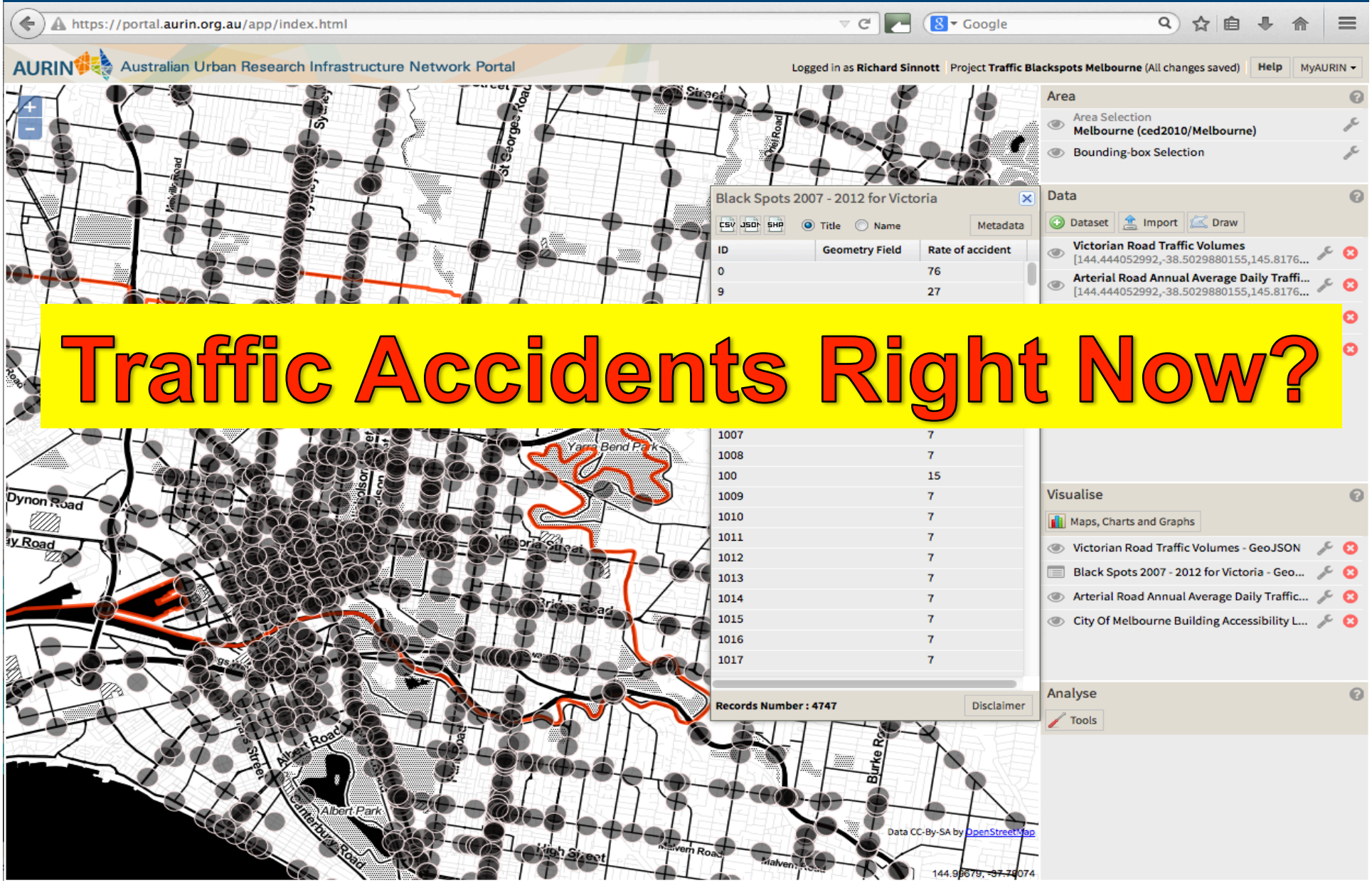




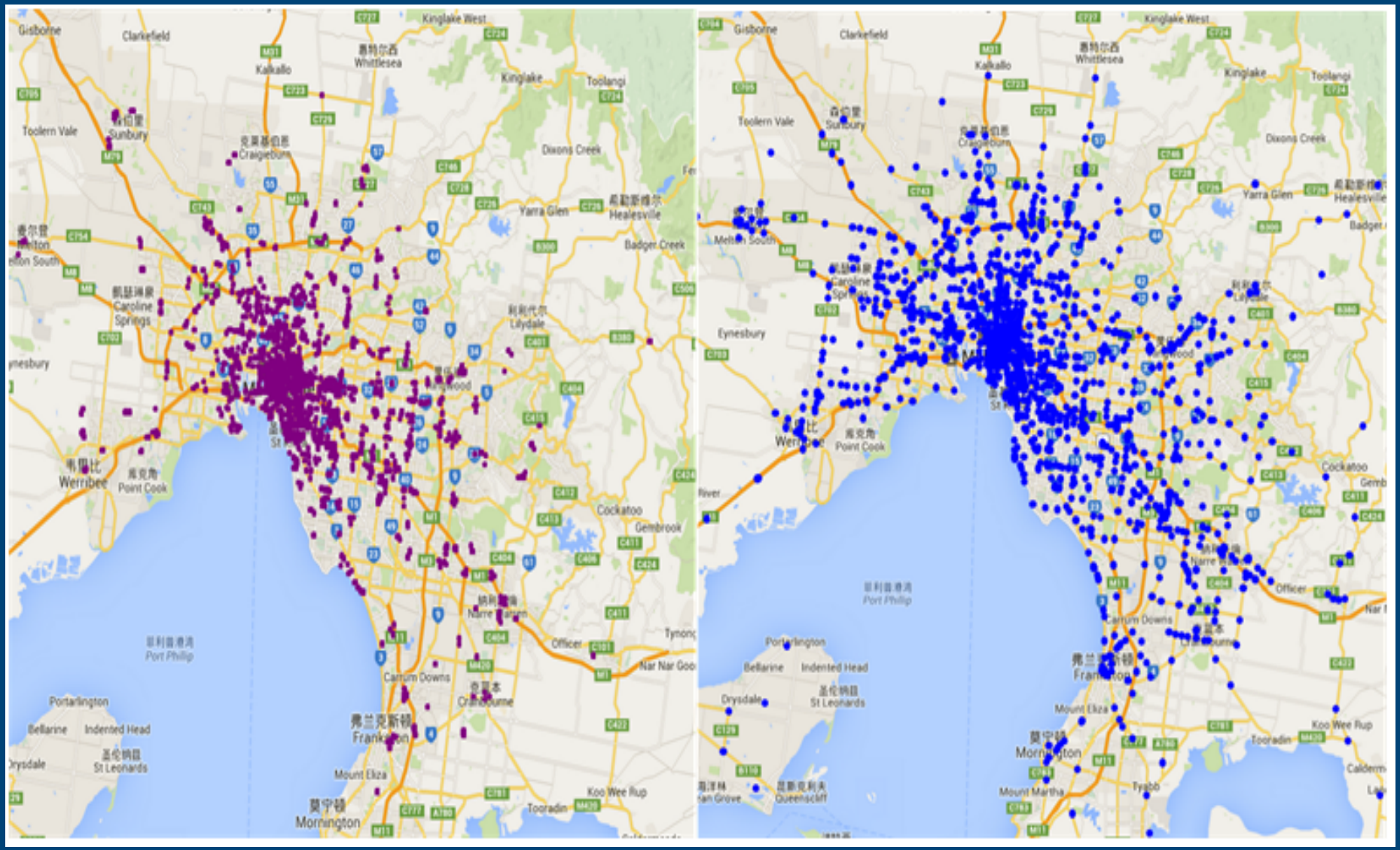
Cluster and Cloud Computing



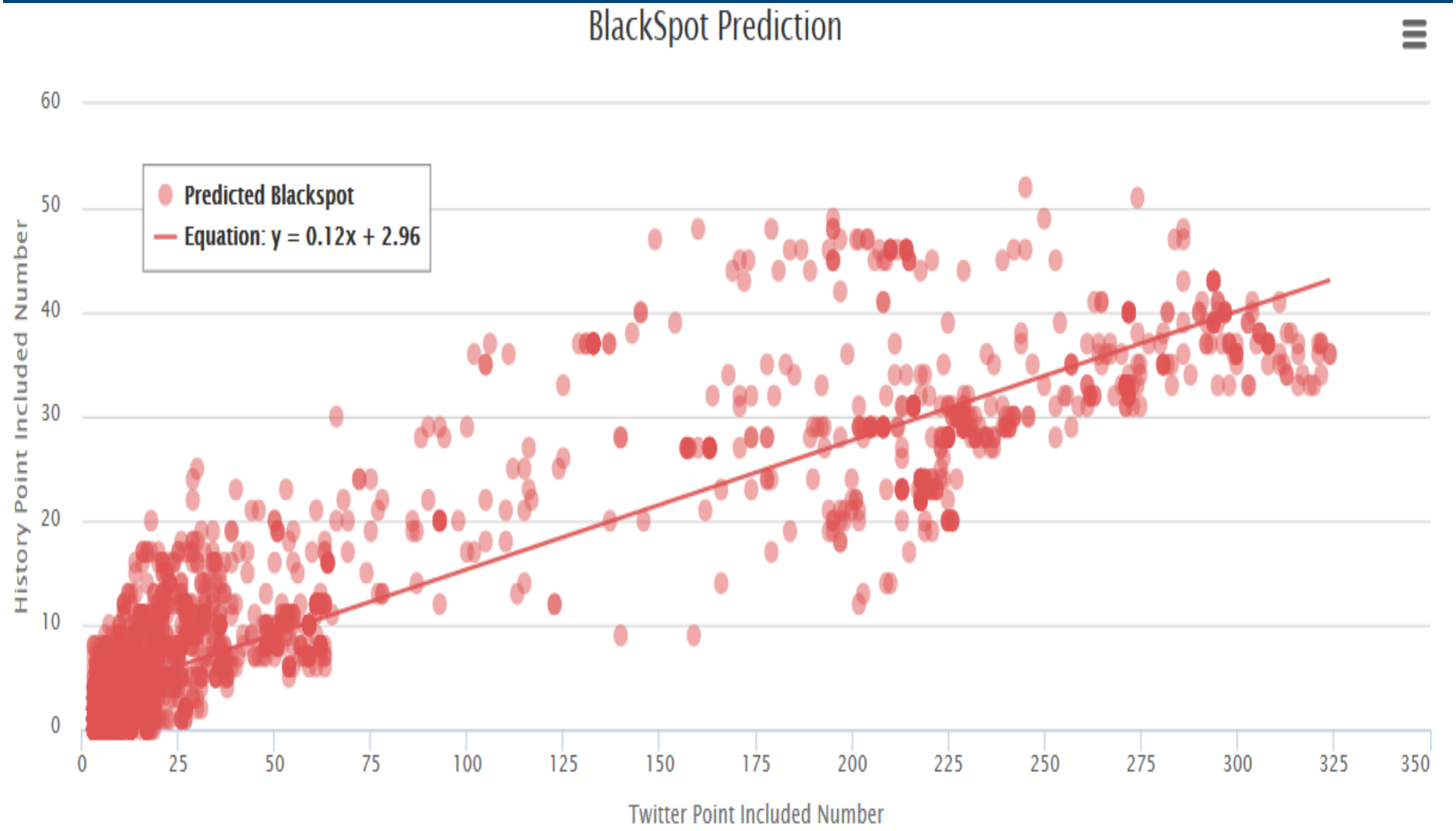
Cluster and Cloud Computing



Predicted Clusters vs Historical Blackspots



Number of Blackspot Predictions

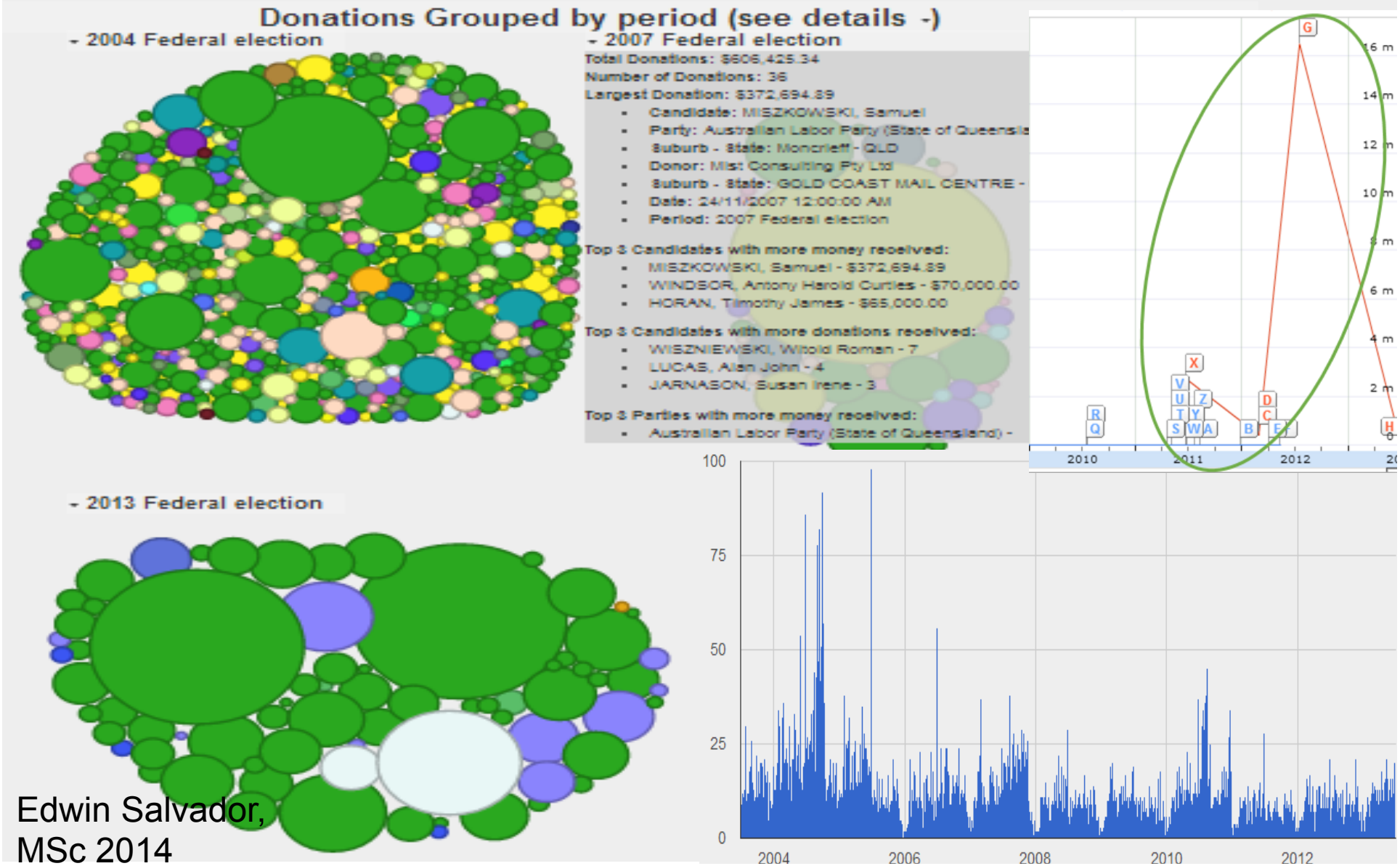


More tweets – better alignment with historical blackspots

Many student dissertations (>250)

- Masters (minor) Dissertation
  - 12 week implementation elective
    - Offers chance for more advanced exploration/analytics
      - Benchmarking technologies
        - » Kubernetes vs Docker Swarm …
      - Exploring algorithms in detail
        - » Scaling machine learning, deep learning over Cloud
      - Challenges of scale of data
      - High velocity data challenges
      - Visual analytics
      - Security and privacy challenges
      - …

Examples



Edwin Salvador, MSc 2014

Crowds…?

How many?

Too many!?

AFL, EPL, Seria A, La Liga

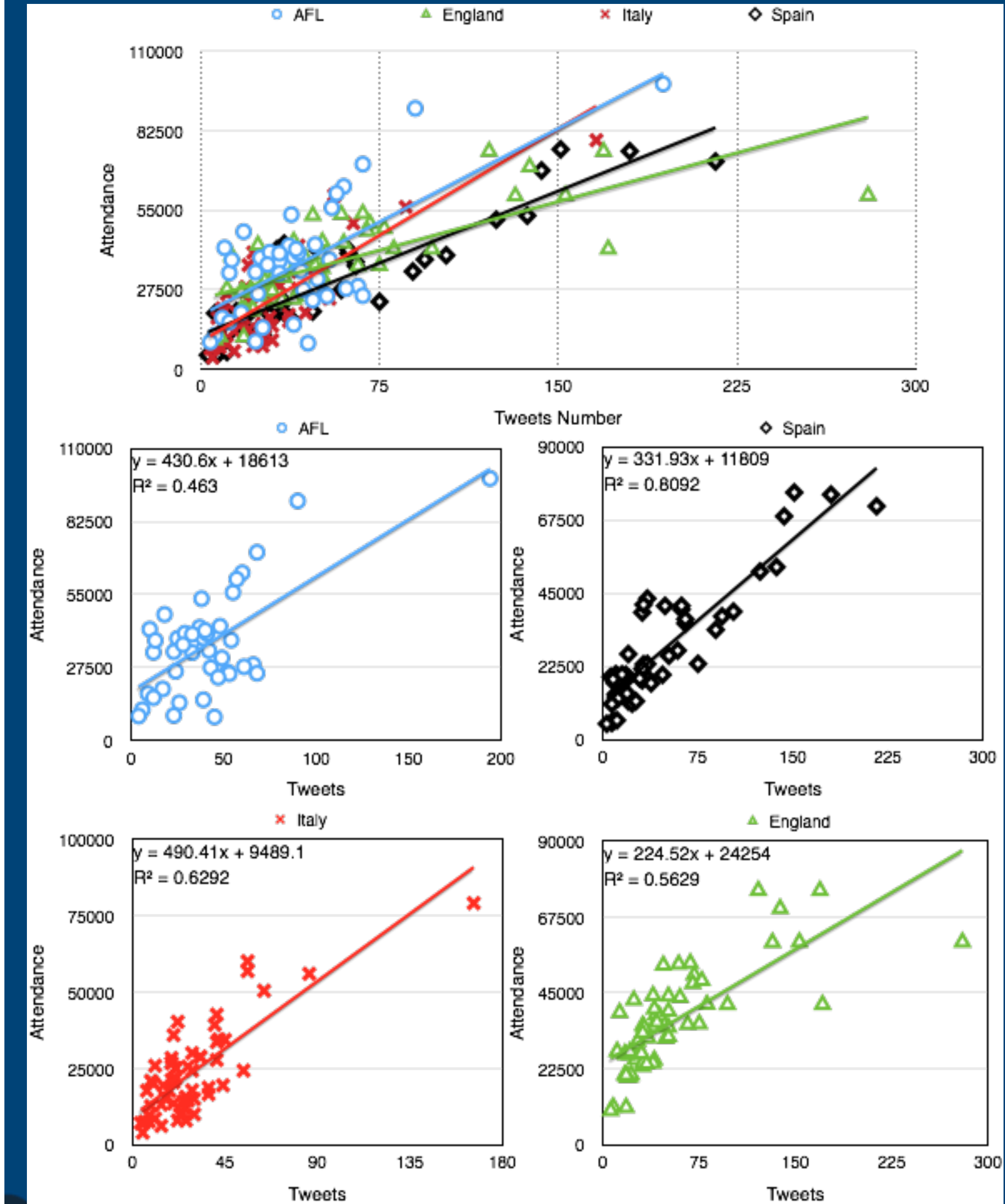


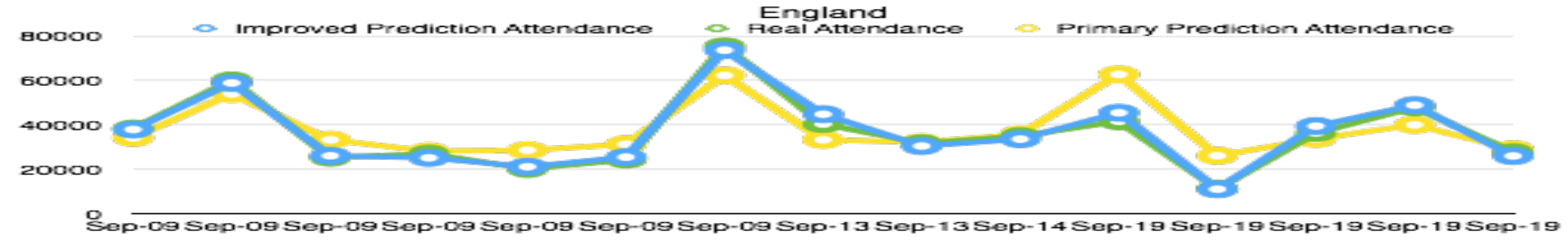
Table 2: Linear Relationship between Tweets and Attendees

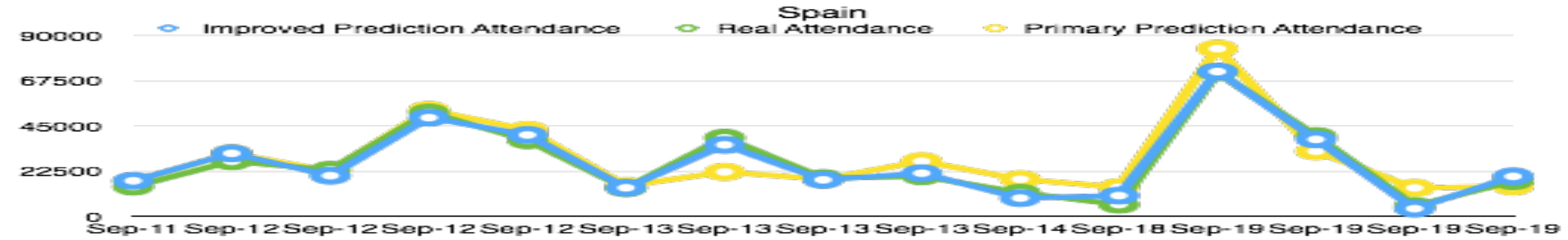
	Equation	R^2
AFL	Attendance = 430.6 * Tweets + 18613	0.463
England	Attendance = 224.52 * Tweets + 24254	0.5629
Spain	Attendance = 331.93 * Tweets + 11809	0.8092
Italy	Attendance = 490.41 * Tweets + 9489.1	0.6292

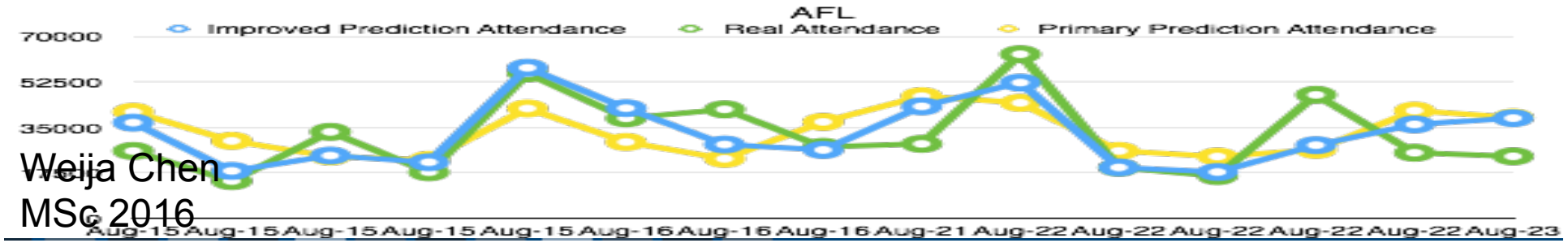
Table 3: Frequent Attendees at Premier League

User Id	User Name	Attendance Frequency
107782177	isalihinabrahim	27
93398111	MrPhilipOldham	8
287183612	cfcdbanks	6
275424947	rifaiabdul4	6
456837227	SlatePaul	6

Tuning the Accuracy

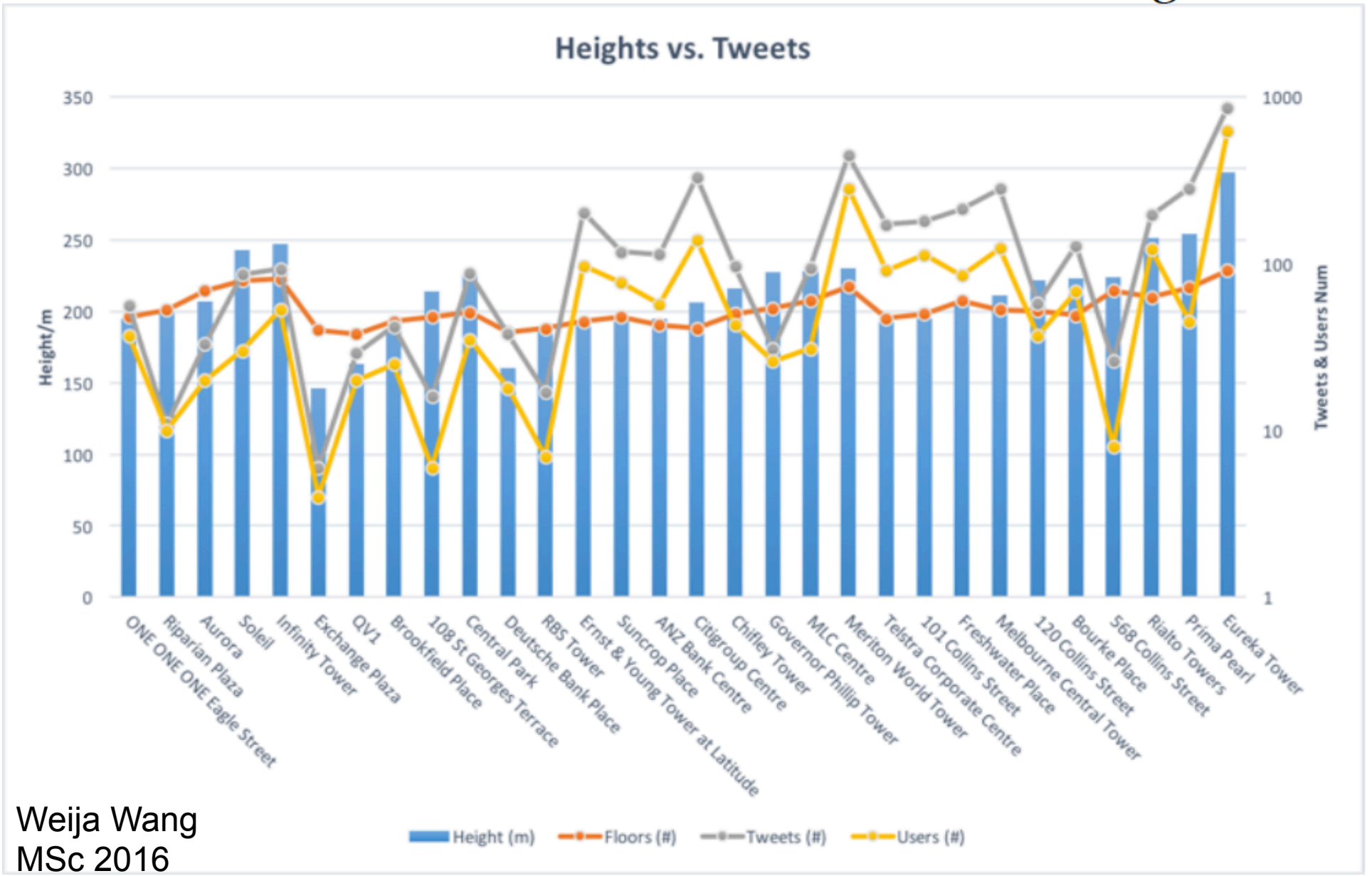






Weija Chen
MSc 2016

Skyscraper (Micro-)Populations



Weija Wang
MSc 2016

Suburb Populations

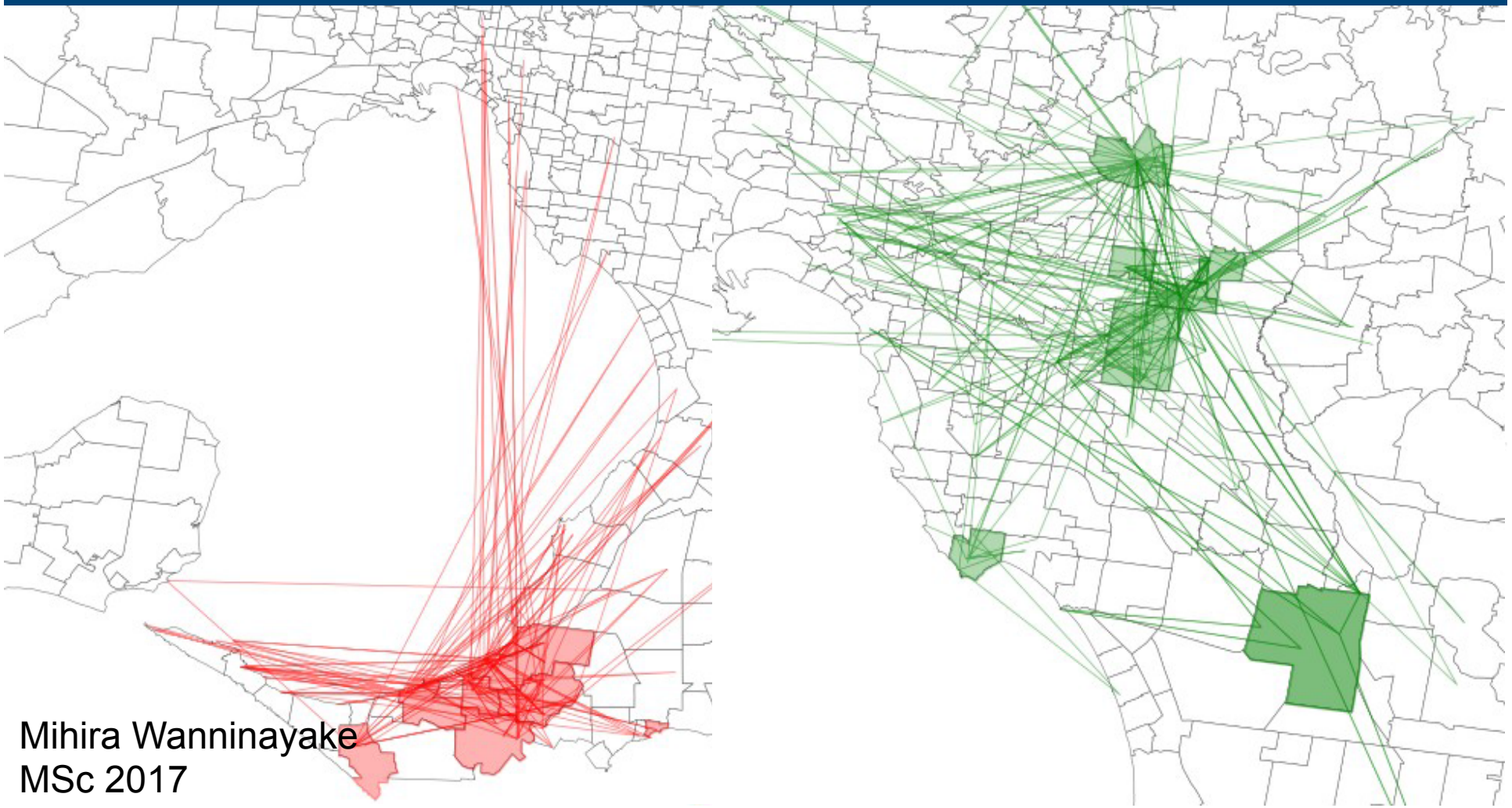
City	Suburb	Pop.	Prediction
Melbourne	Carnegie	17047	16184
	Chelsea Heights	5329	5676
	Flemington	9547	8090
	Malvern-Glen Iris	20279	19654
	Port Melbourne	15413	14213
	Yarra - North	8621	8544
	Bexley	26629	24735
Sydney	Hurstville	21329	20615
	Kogarah Bay - Carlton Allawah	15637	15104
	Sans Souci - Rams-gate	15730	17318
	Morley	21665	21037
Perth	Mosman Park-PeppermintGrove	10828	9234
	Hazelmere - South Guildford	3788	4298
Brisbane	Hamilton (Qld)	4930	5491
	Newmarket	4670	5183
	Windsor	6762	6836

- Factor in:
  - Age
  - Income
  - Ethnicity
  - Geospatial statistical correlations
  - 7% accuracy (according to 2011 census)

Weija Chen
MSc 2016

Commuting Patterns via Social Media

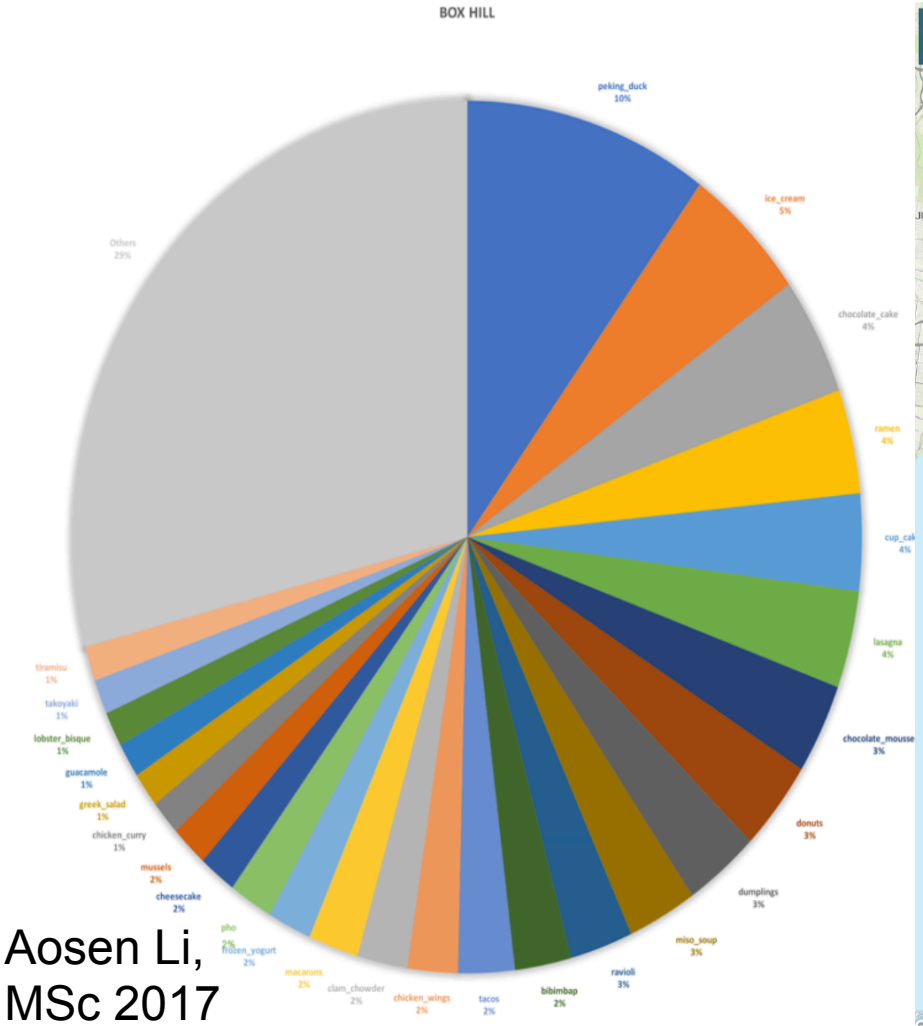
- Twitter, Instagram, Flickr, Foursquare
  - Track usage patterns to identify commuting patterns

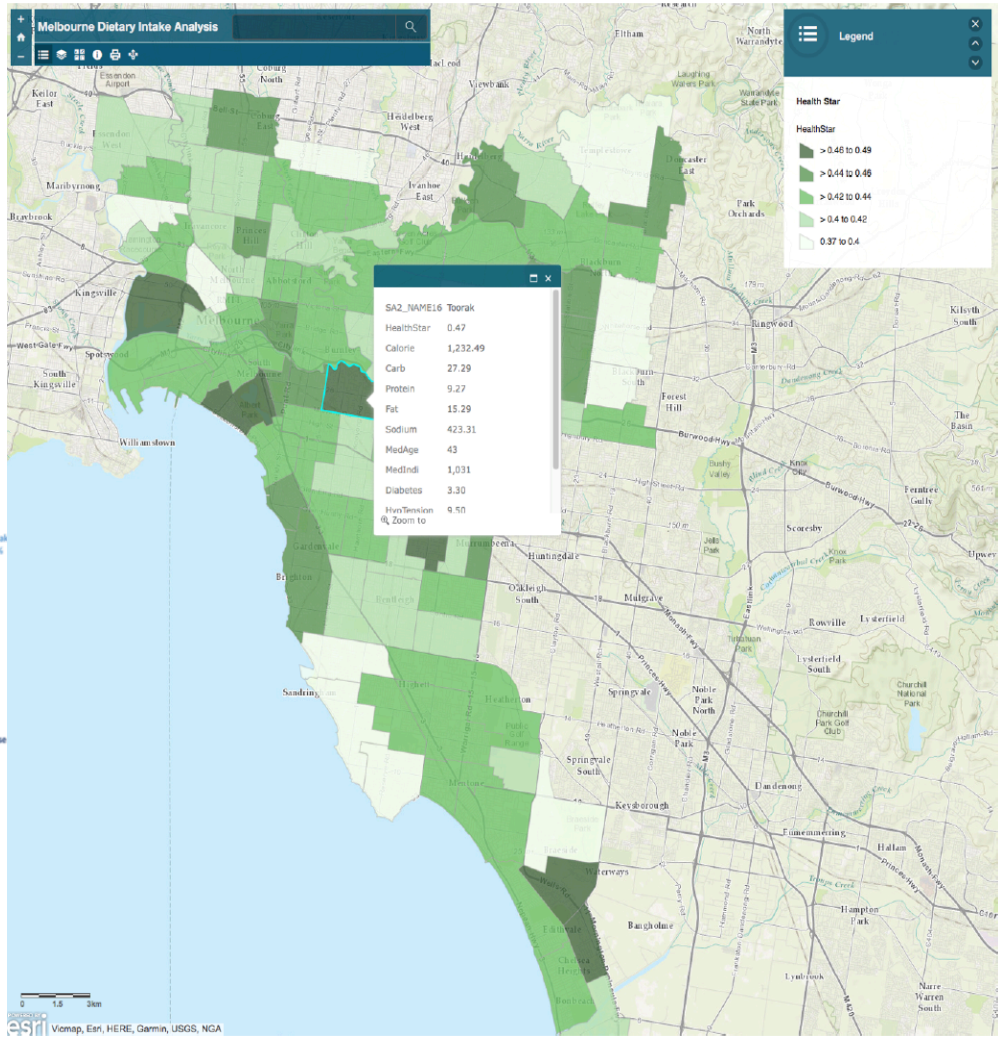


Mihira Wanninayake
MSc 2017

Dietary Analytics

- Instagram image recognition
  - Deep learning (CNN)





Aosen Li,
MSc 2017

Student Workforce

- Technological maturity
  - NeCTAR is not AWS, but hugely improved/stabilised since 2013
- Course is very popular
  - MANY more scenarios, typically for researchers across Uni
    - Gender analytics, type-1 diabetes, openStack mobile apps, Jupyter data analytics, house price bubbles, most popular University, most popular academic, Brexit analytics, Mr TRUMP, …

Questions