It's All About Ensembles

Big Data Analytics Using R

*“Science … resolves **the whole into parts**, the organism into organs, the obscure into the known. … Science gives us knowledge, but only philosophy can give us wisdom … [to] **synthesize knowledge to resolve the obscure into the known.**”*

After the Philosopher Durant.

Graham Williams, Australian Taxation Office

Last Updated 7 October 2014

Ensembles for the Data Scientist

We present an overview of the use of ensembles in Data Mining, particularly in the context of so-called “Big Data”. Starting from the beginning we review how we stumbled on the concept of multiple models, found it useful, and developed it into boosted decision stumps, random forests, and ensembles of nuggets.

Introducing the Concepts of

- Ensembles
- Big Data
- Ensembles for Big Data in R in the ATO

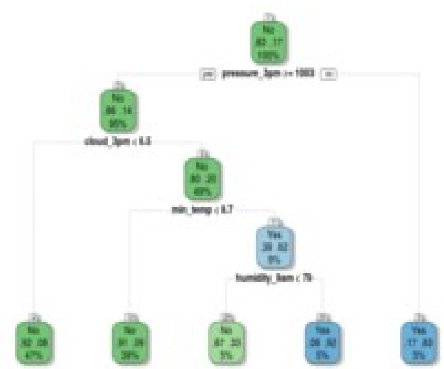
Setting the Scene – 1987

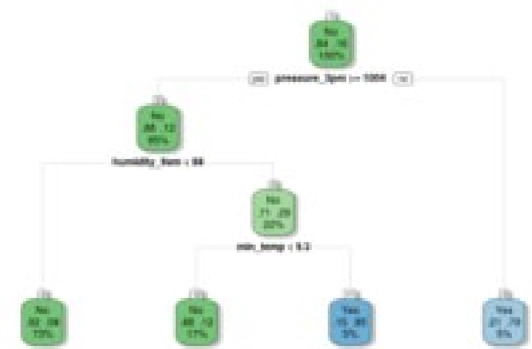
Original brush with ensembles came during PhD research at the ANU in 1987. Implemented a decision tree builder and used it to predict the likelihood of a parcel of land in Australia being suitable for grazing cattle. It used a rather small dataset, allowing calculations to be confirmed by hand – and demonstrating “random” choices for selecting variables.

	date	location	min_temp	max_temp	rainfall	evaporation	sunshine	wind_gust_dir	wind_gus
1	2013-08-31	Canberra	2.0	19.6	0.0	NA	NA	N	39
2	2013-09-01	Canberra	-0.5	21.8	0.0	NA	NA	NNW	22
3	2013-09-02	Canberra	0.9	25.2	0.0	NA	NA	ENE	33
4	2013-09-03	Canberra	2.8	24.1	0.0	NA	NA	WNW	19
5	2013-09-04	Canberra	3.3	21.7	0.0	NA	NA	NNW	35
6	2013-09-05	Canberra	5.1	22.1	0.0	NA	NA	NW	30
7	2013-09-06	Canberra	6.3	22.6	0.0	NA	NA	NNW	41
8	2013-09-07	Canberra	6.5	24.2	0.0	NA	NA	WNW	48
9	2013-09-08	Canberra	3.3	20.4	0.0	NA	NA	NNW	31
10	2013-09-09	Canberra	3.1	23.3	0.0	NA	NA	NNW	39
11	2013-09-10	Canberra	10.4	20.4	0.0	NA	NA	NW	69
12	2013-09-11	Canberra	6.0	16.4	0.0	NA	NA	WNW	52
13	2013-09-12	Canberra	3.2	17.5	0.0	NA	NA	NW	50
14	2013-09-13	Canberra	-3.0	14.9	0.0	NA	NA	ENE	39
15	2013-09-14	Canberra	7.2	18.2	15.2	NA	NA	SSW	50
16	2013-09-15	Canberra	7.3	19.0	0.4	NA	NA	N	26
17	2013-09-16	Canberra	8.3	13.6	0.0	NA	NA	E	33
18	2013-09-17	Canberra	10.9	13.8	57.8	NA	NA	WNW	52
19	2013-09-18	Canberra	10.7	18.3	13.4	NA	NA	WNW	78
20	2013-09-19	Canberra	6.7	13.8	3.2	NA	NA	WNW	54
21	2013-09-20	Canberra	2.0	14.8	0.0	NA	NA	WNW	52

Sample

Sample





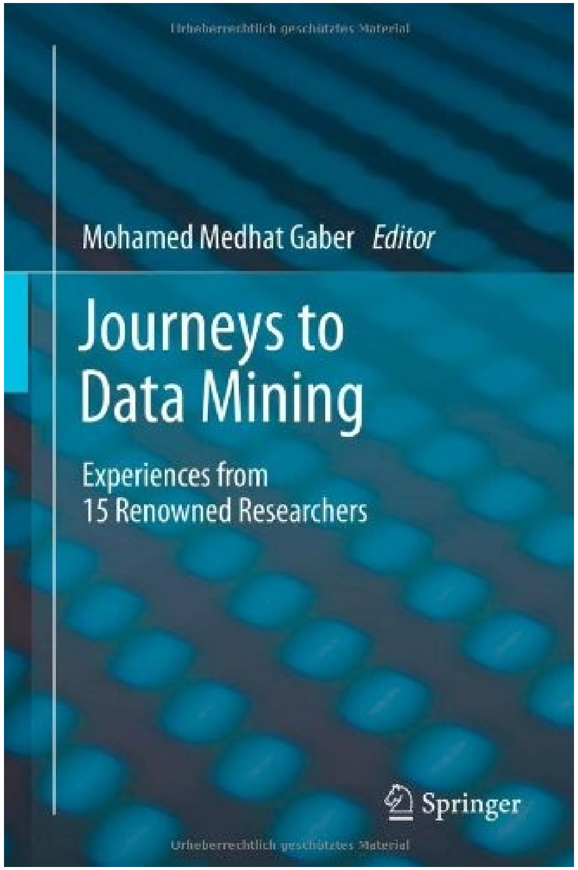
Combining Decision Trees: Initial results from the MIL algorithm, Artificial Intelligence Developments and Applications, edited by J. S. Gero and R. B. Stanton, North-Holland, Elsevier Science Publishers, 1988, Pages 273-289.

Setting the Scene – 1987 – Do Ensembles Make Sense?

The ensemble concept was presented at the first Australian AI Conference in Sydney in November 1987, with Ross Quinlan as the session chair. As recounted in a chapter in Journeys to Data Mining, the idea of not going with the best single model, but combining multiple models was challenged – it is now the approach of choice for many data scientists.

- Concept presented at the first Australian AI Conference in 1987
- Multiple Inductive Learning
- “Why would you build more than one model?” - J. R. Quinlan.

Another chapter as recommended reading on the history of Data Mining is Gregory Piatetski-Shapiro's The Journey of Knowledge Discovery.



Rattle and other Data Mining Tales in Journeys to Data Mining, Experiences from 15 Renowned Researchers, Springer, 2012, 211-230.

Ensembles

Ensembles combine the results from multiple models into a single decision. Over the years ensembles have been demonstrated to produce “better” models than a single model. We might ask the question “Which of several models is actually the best model?” The answer often depends on context. Compare it to a panel of experts.

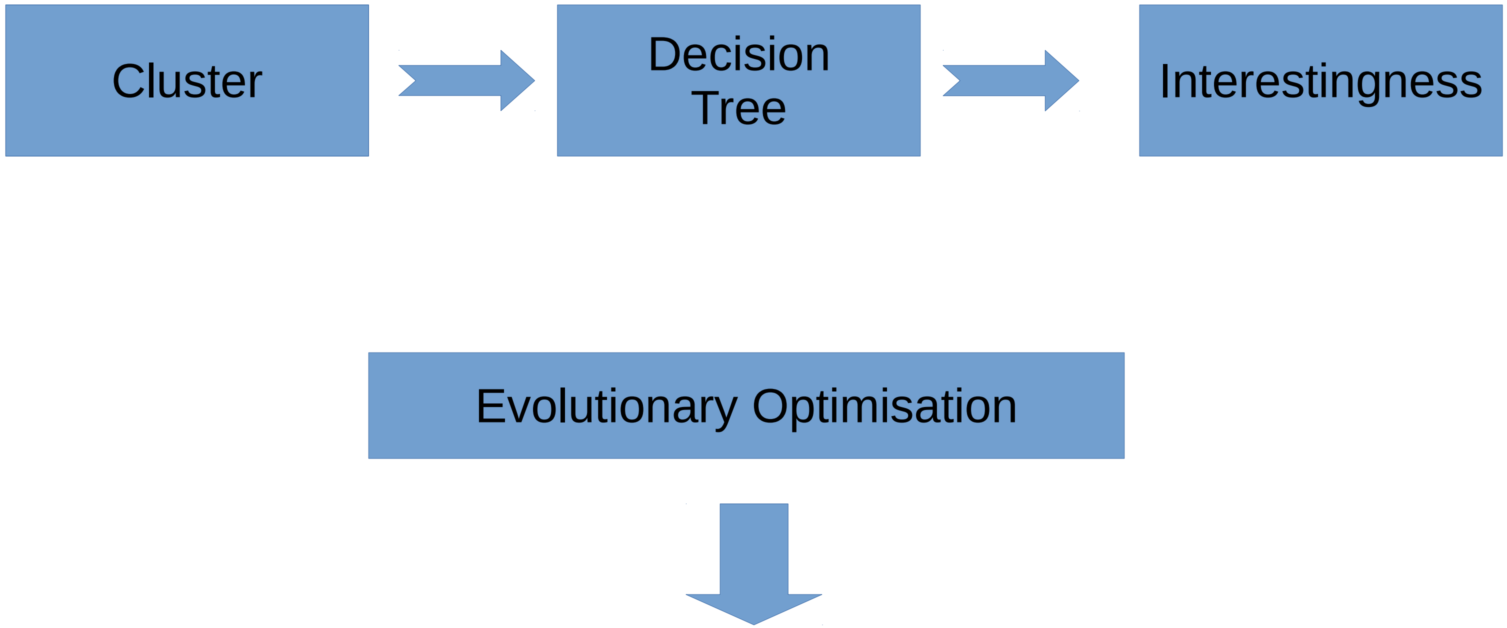
Why not build all of the very similarly good decision trees, and combine them into a single ensemble model?

- Adaptive Boosting
- Bootstrap Aggregation
- Random Forests
- Bucket of Models

By combining multiple models, we can improve accuracy, reduce bias and variance, and provide an overall robust model when applied to new data.

Hot Spots Analysis

Another development from the 1990's is the concept of evolutionary hot spots discovery, working with Professors Zhexue Huang and Xin Yao, applied to health care data. Cluster a large population of M entities into N (<<M) clusters. Describe each cluster by a decision tree, convert to rules, and each rule is then a hot spot measured for interestingness – evolve.



Identify Doctor Shoppers
and Over-Prescribers

Ensemble of Nuggets for Discovering the Unknown

A **new** ensemble approach has been developed to analyse big data using ensembles and hot spots analysis – **nugget discovery**. This experimental approach is being developed through a project which aims to identify compliance issues with Activity Statement refunds.

Annual ATO budget is $3 billion to collect $300 billion revenue for Government

2 million businesses

20 million lodgments

2 million refunds totalling $20 billion

GST Collected each year is approximately $50 billion

Yes

No

$n billion refunds

$m billion protected

Poor targeting leads to significantly “higher touch” than required.

Analytics

m << n

All figures are unofficial approximate/estimates – see, e.g.,

- http://www.igt.gov.au/content/reports/gst_refunds/GST_Refunds-01.asp
- http://budget.gov.au/2014-15/content/bp3/html/bp3_04_part_3.htm

Global Models versus Local Models

Traditional approaches fail for big data as they often attempt to build one model over the whole population. The approach here is to automatically identify previously unknown behavioural groups, and micro model within the groups. An ensemble of local models predict the risk (+ve/-ve) of an entity, which is then aggregated for an overall risk.

- Local populations with markedly different behaviours and properties to those of the global population are key to developing models in the big data world.

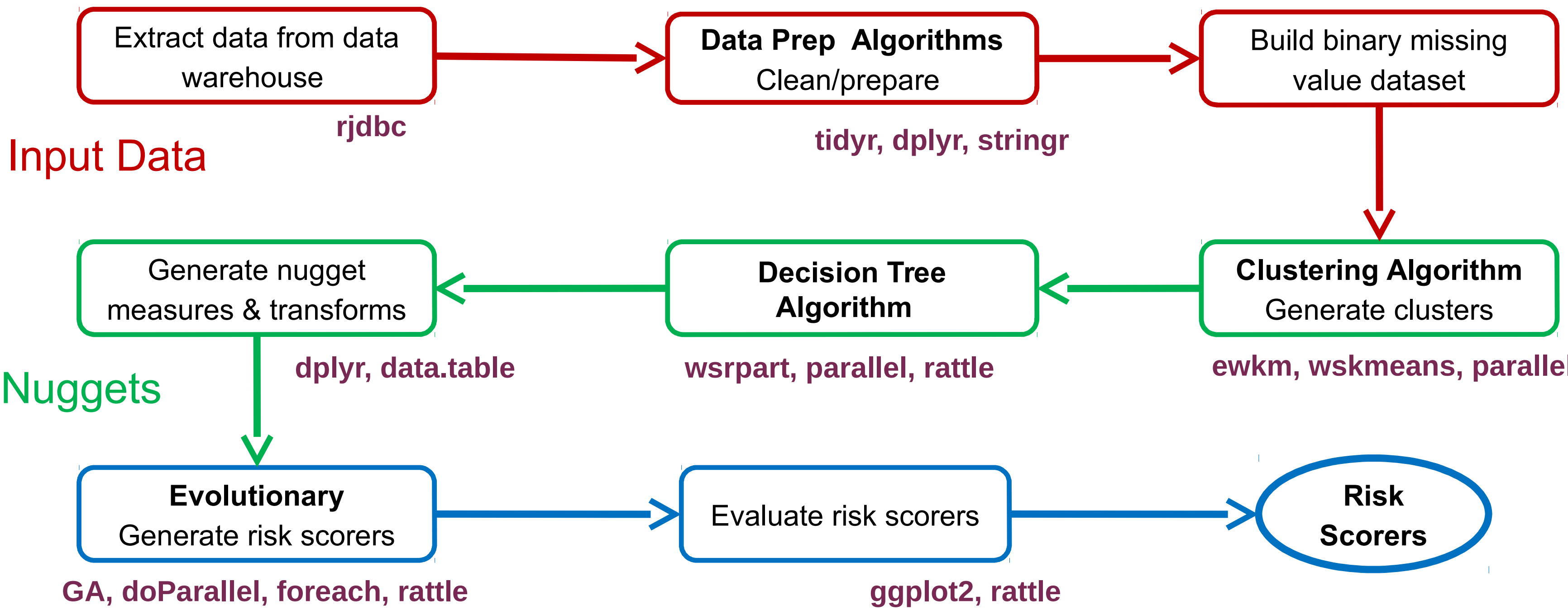
- Our approach here is to develop a massive ensemble of local models that can be aggregated to better reason over the global population.

Acknowledgement ATO Data Scientists: Dr Frank Lu, Dr Lifang Gu, Dr Nandita Sharma, Nicole Wade

Algorithms

The overall algorithm combines **weighted subspace cluster analysis** to identify behaviourally coherent groups, **decision tree induction** to build business interpretable rules, and **evolutionary algorithms** to identify the best risk scorers. An agile approach is used to deliver regular expert feedback into the modelling process.

Training dataset: 2013, 2 million refunds, 1,000 features



Nugget Attributes and Transformation

We might "discover" in a data-driven algorithm 20,000 nuggets covering the 2 million credit Business Activity Statement lodgements over one year. We now need to measure how "interesting" each nugget is. A simple approach is to define a collection of attributes for each nugget, and combine them to measure the nugget.

Population Attributes
- Percentage of GST
- Percentage of BAS
- ...

Demographics Attributes
- Percentage Gov
- Percentage Large
- ...

Training Attributes
- Percentage known fraud
- Percentage known productive cases
- Percentage known revenue protected
- ...

Nugget Attributes
x20

Nugget Transforms
x10

Z-score
Standardise the range of different characteristics.

Shift
Shift the values by the addition on the value on nugget of whole population.

Scale
To normalise the values to 0 to 1 or -1 to 1

Weighting
To make the effective nuggets contribute more and reduce the influence of noisy nuggets

200 Attributes

Risk Scorers – The Genetic Code for Evolutionary Process

Each possible risk scorer is a formula in the language of the measures, transforms, and weights, over the nuggets. There is an infinite number of possible risk scorers for any population of nuggets – and we need to identify good risk scorers across this infinite search space. Heuristic search is required, and evolutionary optimisation is a candidate.

Aggregation function: sum, max, avg, ...

$$r(b) = \sum_{i=1}^{n_c} a_i \times \sum_{j \in \text{ng[b]}} w_{ij} \times v_{ij}$$

$a_i = weight\ for\ attribute\ i$

$n_c = \text{number of chracteristics}$

ng[b] = nuggets that match BAS b

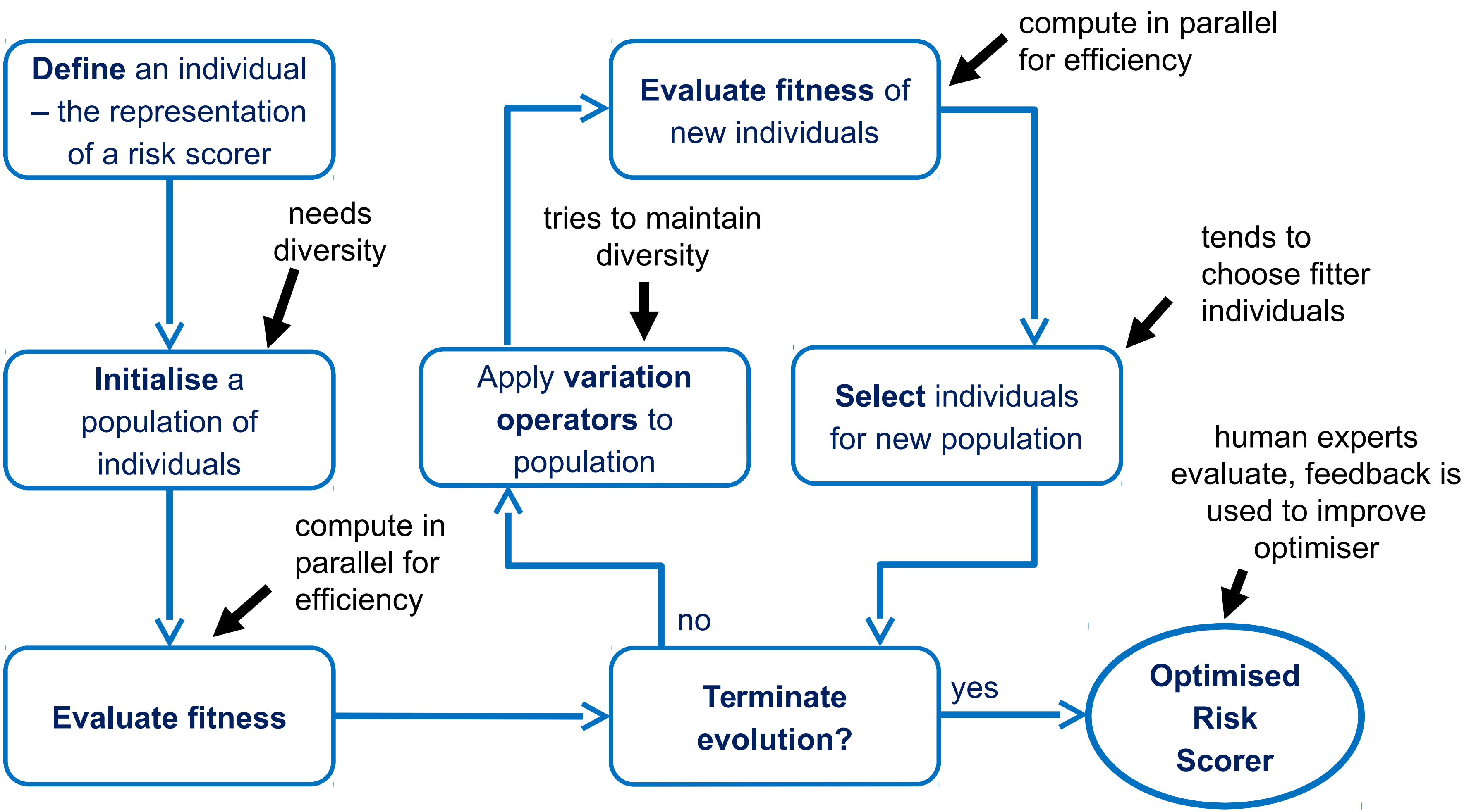
$w_{ij} = \text{weight for attribute i in nugget j}$

$v_{ij} = \text{transformed value of } a_i \text{ in } ng_j$

We need to discover the attribute weights, nugget weights and aggregation function.

Risk Scorers – Evolutionary Optimisation

Evolutionary optimisation is a biologically inspired computational algorithm developed by artificial intelligence researchers. Given a population of genetic material, randomly mutate and cross pollinate the genes, under the guidance of a measure of fitness – Our measure here is based also on feedback of a team of professional auditors.



Infrastructure for Analytics – Can be Cost Effective

The ATO introduced the use of open source tools for data mining over 10 years with the set up of Corporate Analytics. We recognise that we need toolkits with a variety and different tools, changing regularly, open and closed source. As Gartner noted, it is no surprise that the latest technology for data science is coming through the open source route.

- Network of Ubuntu servers: 32 parallel threads, 750GB RAM
- Running open source from the ground up – GNU/Linux OS
- Powerful suite of well established open source Unix tools
  - C, awk, sed, wc, diff, meld, tr, cvs, latex, perl, python, R
  - Concept of many specialised tools working together through a standard interface referred to as "pipes".
  - Pipes now a powerful new concept in R.
- Scaling out rather than scaling up – add new computers to the grid, not necessarily larger computers – R, Spark, Python, ...

… and then there were 7 billion models

The November edition of the IEEE Computational Intelligence Magazine contains an article where I discuss turning ensemble concepts into the extreme, reflecting on the need for the pendulum to swing back toward protecting privacy, and the resulting focus on massively ensembled models, each “model” modelling an individual.

Big Data Opportunities and Challenges – Extreme Data Distribution: Privacy and Ownership.

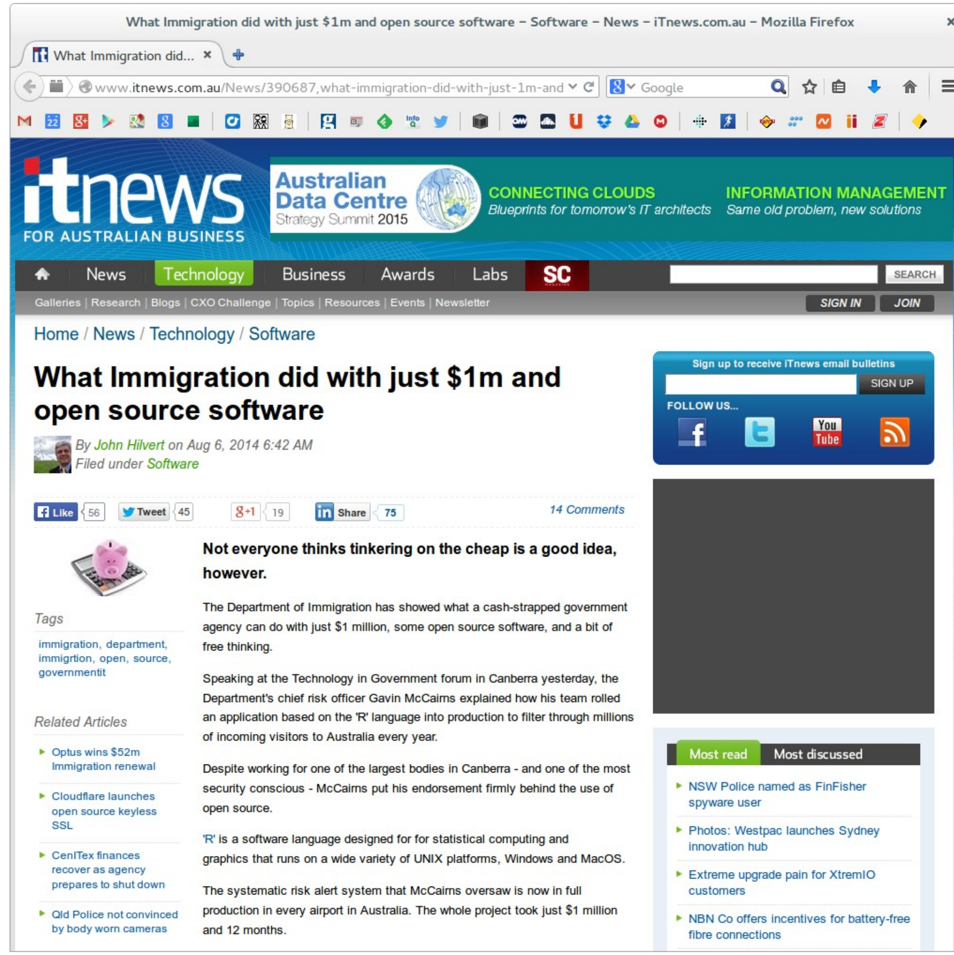
Open Source R as Credible Software

> We continue to be discouraged by the fear, uncertainty and doubt that is offered by many vendors who have a natural concern about loss of business. Instead we need a variety of tools that will make up the most effective toolkits for Data Scientists including the state-of-the-art that only open source can deliver, as identified by Gartner 2014.

Dept Immigration: Data Scientists deliver sophisticated risk models to protect Australia's borders. Gavin McCairns says “the department bought $15 million worth of software---but it's gathering dust."

SAS responded: “... R in a production system, it can be scary …”

Every Australian Tax Return lodged today is risk scored by at least one model developed using open source software (often an R-based model).



What Immigration did with just $1m and open source software – Software – News – iTnews.com.au – Mozilla Firefox

www.itnews.com.au/News/390687,what-immigration-did-with-just-1m-and

itnews FOR AUSTRALIAN BUSINESS

News | Technology | Business | Awards | Labs | SC

Home / News / Technology / Software

What Immigration did with just $1m and open source software

By John Hilvert on Aug 6, 2014 6:42 AM
Filed under Software

Not everyone thinks tinkering on the cheap is a good idea, however.

The Department of Immigration has showed what a cash-strapped government agency can do with just $1 million, some open source software, and a bit of free thinking.

Speaking at the Technology in Government forum in Canberra yesterday, the Department's chief risk officer Gavin McCairns explained how his team rolled an application based on the 'R' language into production to filter through millions of incoming visitors to Australia every year.

Despite working for one of the largest bodies in Canberra - and one of the most security conscious - McCairns put his endorsement firmly behind the use of open source.

'R' is a software language designed for for statistical computing and graphics that runs on a wide variety of UNIX platforms, Windows and MacOS.

The systematic risk alert system that McCairns oversaw is now in full production in every airport in Australia. The whole project took just $1 million and 12 months.

Tags: immigration, department, immigrtion, open, source, governmentit

Related Articles
- Optus wins $52m Immigration renewal
- Cloudflare launches open source keyless SSL
- CenITex finances recover as agency prepares to shut down
- Qld Police not convinced by body worn cameras

Most read | Most discussed
- NSW Police named as FinFisher spyware user
- Photos: Westpac launches Sydney innovation hub
- Extreme upgrade pain for XtremIO customers
- NBN Co offers incentives for battery-free fibre connections

Key Messages – Ensembles

The state-of-the-art Analytic Model developed here introduces a **new approach to big data analytics**. The technology takes us beyond traditional algorithms and prepares us for delivering new capabilities to support the ATO move to providing better interactions with Tax Payers. New ideas undergoing research and refinement to discover the unknowns.

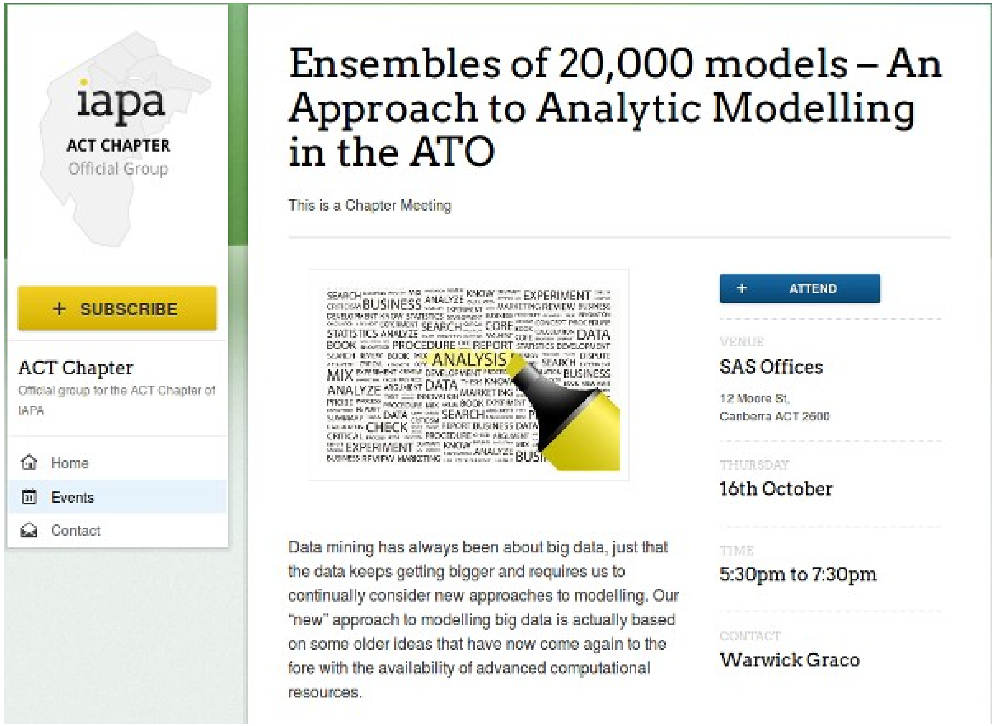
Why have only one model when you can have a population of 20,000 models?

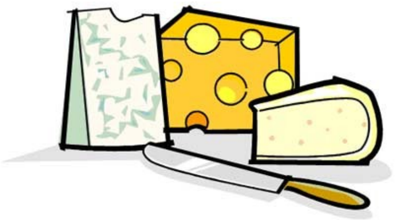
*Data Scientists **synthesise** data into information, information into knowledge, and knowledge into wisdom, to resolve the obscure into the known.*

An ensemble of many models (20,000 or even many more?) delivers the expertise on local understanding to global applicability.

… and now

Now's a great time to grab some snacks and drinks and meet some of our data science colleagues – networking...
I'll also be presenting a broader talk on data science and ensembles at IAPA on Thursday 16th October – next week.
Join us for the next Canberra R User Group and Data Science meeting, first Tuesday in November.





Keep an eye on http://innovationspace.net.au