Delivery of an Analytics Capability

Ensembles and Model Delivery for Tax Compliance

Graham Williams

Senior Director and Chief Data Miner, Analytics
Office of the Chief Knowledge Officer
Australian Taxation Office

Adjunct Professor, University of Canberra
Adjunct Professor, Australian National University
Visiting Professor, Chinese Academy of Sciences
Fellow, Institute of Analytics Professionals of Australia

Graham.Williams@togaware.com
http://datamining.togaware.com

Copyright © 2012 Graham.Williams@togaware.com

Overview

Revenue Collection

Establishing Capability

Delivering Outcomes

Ongoing Challenges

- Model Deployment
- Big Data: Many Variables

Copyright © 2012 Graham.Williams@togaware.com

"nothing ... [is] certain, except death and taxes."

Benjamin Franklin, 1789

"The hardest thing in the world to understand is the income tax."

attributed to Albert Einstein

"The avoidance of taxes is the only intellectual pursuit that carries any reward."

John Maynard Keynes

Copyright © 2012 Graham.Williams@togaware.com

"nothing ... [is] certain, except death and taxes."

Benjamin Franklin, 1789

"The hardest thing in the world to understand is the income tax."

attributed to Albert Einstein

"The avoidance of taxes is the only intellectual pursuit that carries any reward."

John Maynard Keynes

Copyright © 2012 Graham.Williams@togaware.com

"nothing ... [is] certain, except death and taxes."

Benjamin Franklin, 1789

"The hardest thing in the world to understand is the income tax."

attributed to Albert Einstein

"The avoidance of taxes is the only intellectual pursuit that carries any reward."

John Maynard Keynes

Copyright © 2012 Graham.Williams@togaware.com

Taxation Records - Historic Data Collection

- **Taxation and Governance**
  Revenue is fundamental to government.
- **Data Collection**
  2500 B.C. Sumerian (Iraq) and Elam (Iran) peoples marked their tax records onto dried mud tablets.
- We still inscribe tax records today — now electronically
- **Data Mining Opportunities**
  Analyse very large collections (20M by 5K) to check tax payer identity, facilitate reporting (lodgement), ensure compliance, pay refunds or collect debts: the four pillars.

Sources `http://www.upenn.edu/almanac/v48/n28/AncientTaxes.html` `http://www.crystalinks.com/cuneiformtablets.html`

Copyright © 2012 Graham.Williams@togaware.com

Taxation Records - Historic Data Collection

- **Taxation and Governance**
  Revenue is fundamental to government.
- **Data Collection**
  2500 B.C. Sumerian (Iraq) and Elam (Iran) peoples marked their tax records onto dried mud tablets.
- We still inscribe tax records today — now electronically
- **Data Mining Opportunities**
  Analyse very large collections (20M by 5K) to check tax payer identity, facilitate reporting (lodgement), ensure compliance, pay refunds or collect debts: the four pillars.

Sources `http://www.upenn.edu/almanac/v48/n28/AncientTaxes.html` `http://www.crystalinks.com/cuneiformtablets.html`

Copyright © 2012 Graham.Williams@togaware.com

Taxation Records - Historic Data Collection

- **Taxation and Governance**
  Revenue is fundamental to government.
- **Data Collection**
  2500 B.C. Sumerian (Iraq) and Elam (Iran) peoples marked their tax records onto dried mud tablets.
- We still inscribe tax records today — now electronically
- **Data Mining Opportunities**
  Analyse very large collections (20M by 5K) to check tax payer identity, facilitate reporting (lodgement), ensure compliance, pay refunds or collect debts: the four pillars.

Sources `http://www.upenn.edu/almanac/v48/n28/AncientTaxes.html` `http://www.crystalinks.com/cuneiformtablets.html`

Copyright © 2012 Graham.Williams@togaware.com

Taxation Records - Historic Data Collection

- **Taxation and Governance**
  Revenue is fundamental to government.
- **Data Collection**
  2500 B.C. Sumerian (Iraq) and Elam (Iran) peoples marked their tax records onto dried mud tablets.
- We still inscribe tax records today — now electronically
- **Data Mining Opportunities**
  Analyse very large collections (20M by 5K) to check tax payer identity, facilitate reporting (lodgement), ensure compliance, pay refunds or collect debts: the four pillars.

Sources `http://www.upenn.edu/almanac/v48/n28/AncientTaxes.html` `http://www.crystalinks.com/cuneiformtablets.html`

Copyright © 2012 Graham.Williams@togaware.com

Australian Taxation Office

. . . continues this "fine tradition"

- Employs 22,000 staff Australia wide
- Role to collect revenue and process refunds

- 12M Individuals, $600B Income, $120B Tax
- 2M Companies..., $2,200B Income, $50B Tax (after expenses)
- GST $46B, Excise $26B, FBT $3B, . . . ≈ $350B total income

- Tax payer's charter:
  Fair but firm; *Assume honesty; Protect privacy*
- Service standards — refunds within days . . . hours . . . seconds?
- Whilst protecting the integrity of the revenue collection

Copyright © 2012 Graham.Williams@togaware.com

Australian Taxation Office

. . . continues this "fine tradition"

- Employs 22,000 staff Australia wide
- Role to collect revenue and process refunds

- 12M Individuals, $600B Income, $120B Tax
- 2M Companies..., $2,200B Income, $50B Tax (after expenses)
- GST $46B, Excise $26B, FBT $3B, . . . ≈ $350B total income

- Tax payer's charter:
  Fair but firm; *Assume honesty; Protect privacy*
- Service standards — refunds within days . . . hours . . . seconds?
- Whilst protecting the integrity of the revenue collection

Copyright © 2012 Graham.Williams@togaware.com

Australian Taxation Office

. . . continues this "fine tradition"

- Employs 22,000 staff Australia wide
- Role to collect revenue and process refunds

- 12M Individuals, $600B Income, $120B Tax
- 2M Companies..., $2,200B Income, $50B Tax (after expenses)
- GST $46B, Excise $26B, FBT $3B, . . . ≈ $350B total income

- Tax payer's charter:
  Fair but firm; *Assume honesty; Protect privacy*
- Service standards — refunds within days . . . hours . . . seconds?
- Whilst protecting the integrity of the revenue collection

Copyright © 2012 Graham.Williams@togaware.com

Australian Taxation Office

. . . continues this "fine tradition"

- Employs 22,000 staff Australia wide
- Role to collect revenue and process refunds

- 12M Individuals, $600B Income, $120B Tax
- 2M Companies..., $2,200B Income, $50B Tax (after expenses)
- GST $46B, Excise $26B, FBT $3B, . . . $\approx$ $350B total income

- Tax payer's charter:
  Fair but firm; *Assume honesty; Protect privacy*
- Service standards — refunds within days . . . hours . . . seconds?
- Whilst protecting the integrity of the revenue collection

Copyright © 2012 Graham.Williams@togaware.com

Overview

Revenue Collection

Establishing Capability

Delivering Outcomes

Ongoing Challenges

- Model Deployment
- Big Data: Many Variables

Copyright © 2012 Graham.Williams@togaware.com

Creating an Analytics Capability

- Established as a corporate capability in 2004
- Strong support by visionary CEO and senior executives
- Core team of 15 data miners then and now 30 data miners
- Wider team of 150 data analysts throughout the organisation
- Shared technology throughout organisation
- Provide framework for whole of ATO risk management

Every tax return lodged in Australia today is risk assessed by at least one data mining model.

- Models delivering benefit:
  - Revenue assurance: impact in $ millions
  - More efficient targeting: resource and tax payer annoyance
  - Better tax payer experience: briefer involvement

Copyright © 2012 Graham.Williams@togaware.com

Creating an Analytics Capability

- Established as a corporate capability in 2004
- Strong support by visionary CEO and senior executives
- Core team of 15 data miners then and now 30 data miners
- Wider team of 150 data analysts throughout the organisation
- Shared technology throughout organisation
- Provide framework for whole of ATO risk management

  Every tax return lodged in Australia today is risk assessed by at least one data mining model.

- Models delivering benefit:
  - Revenue assurance: impact in $ millions
  - More efficient targeting: resource and tax payer annoyance
  - Better tax payer experience: briefer involvement

Copyright © 2012 Graham.Williams@togaware.com

Creating an Analytics Capability

- Established as a corporate capability in 2004
- Strong support by visionary CEO and senior executives
- Core team of 15 data miners then and now 30 data miners
- Wider team of 150 data analysts throughout the organisation
- Shared technology throughout organisation
- Provide framework for whole of ATO risk management

  Every tax return lodged in Australia today is risk assessed by at least one data mining model.

- Models delivering benefit:
  - Revenue assurance: impact in $ millions
  - More efficient targeting: resource and tax payer annoyance
  - Better tax payer experience: briefer involvement

Copyright © 2012 Graham.Williams@togaware.com

Challenge 1: Analytics as IT

- Traditional IT: Buy software first then buy expertise
  c.f. accounting software and accountants

  c.f. data miners as innovative programmers of data
- A culture of multiple tools and dynamic environments
- A culture of sharing algorithms and experiences
- Flexibility in using any technology as and when required
- Not constrained by traditional corporate IT SOE

Copyright © 2012 Graham.Williams@togaware.com

Lesson: Analyst First

"Analytics is a non-repetitive, exploratory and creative process where the outcome is not known at the start, and only a fraction of efforts are expected to result in success."

analystfirst.com

- Analytics is not IT and process.

- Focus is not the tools nor the algorithms!
- Focus needs to be with the skills of the people.
- The people who perform, manage, request and envision analytics.
- Required skill set: Programming, Statistics, Algorithms, Intuition

Copyright © 2012 Graham.Williams@togaware.com

Challenge 2: Missing Technology Ensembles

Ensemble concept developed in the 80's

- Multiple Decision Trees (PhD 1987)
- Remains one of the best off-the-shelf technologies: random forests and boosting.
- Not available in the closed source offerings until recently.
- Yet readily available in the open source community.

Copyright © 2012 Graham.Williams@togaware.com

Lesson: Commodity Platform

A data mining capability need not be expensive.

Build a network of workstations:

- 16 Cores, 64 bit, top CPU speeds
  512GB RAM, 10TB Disk
- Flexible and open source OS
  Ubuntu GNU/Linux
- Open Source data mining tools
  R, Rattle, Weka + SAS, SPSS as rqd
- Open Source does deliver quality

Advert: Strive for open scientific transparency and repeatability – release an R package (or Weka) then it is also available in SAS, SPSS, Netezza, Micro Strategy...

Copyright © 2012 Graham.Williams@togaware.com

Lesson: Commodity Platform

A data mining capability need not be expensive.

Build a network of workstations:

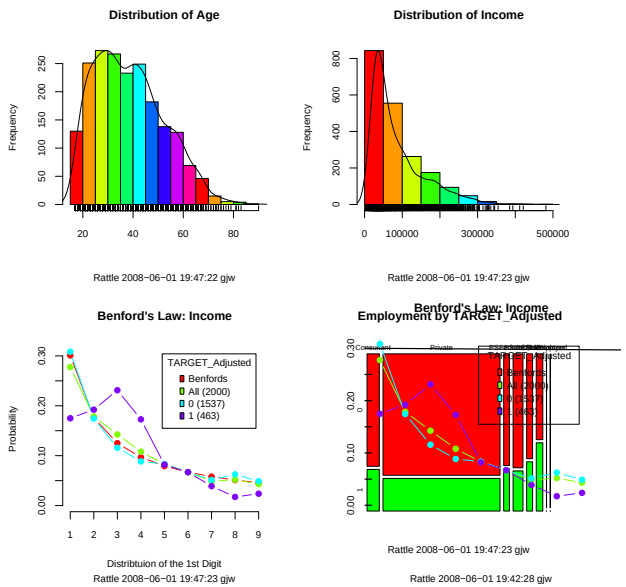
- 16 Cores, 64 bit, top CPU speeds
  512GB RAM, 10TB Disk
- Flexible and open source OS
  Ubuntu GNU/Linux
- Open Source data mining tools
  R, Rattle, Weka + SAS, SPSS as rqd
- Open Source does deliver quality

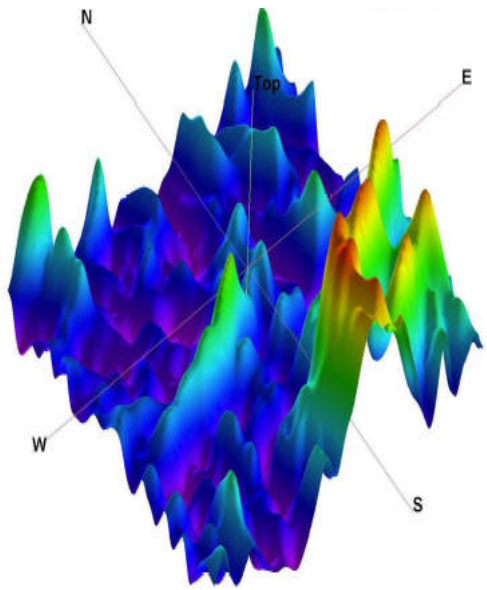
Advert: Strive for open scientific transparency and repeatability – release an R package (or Weka) then it is also available in SAS, SPSS, Netezza, Micro Strategy...

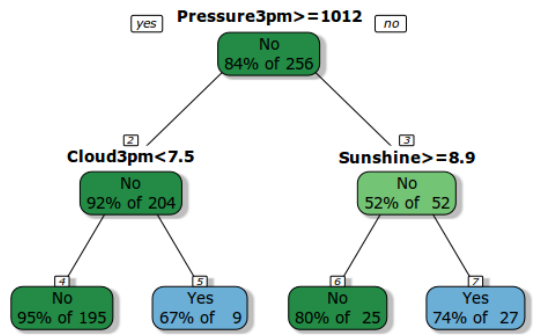
Copyright © 2012 Graham.Williams@togaware.com

Challenge 3: Programming Data Analysis

- Exploratory Data Analysis + Mining: R is second to none
- Analytics is about programming with data, and today's culture is all about GUI - simplified interfaces
- Skills Shortage $\rightarrow$ Train 150 Data Analysts
- Rattle developed as a stepping stone to transparent and repeatable analytics in R.

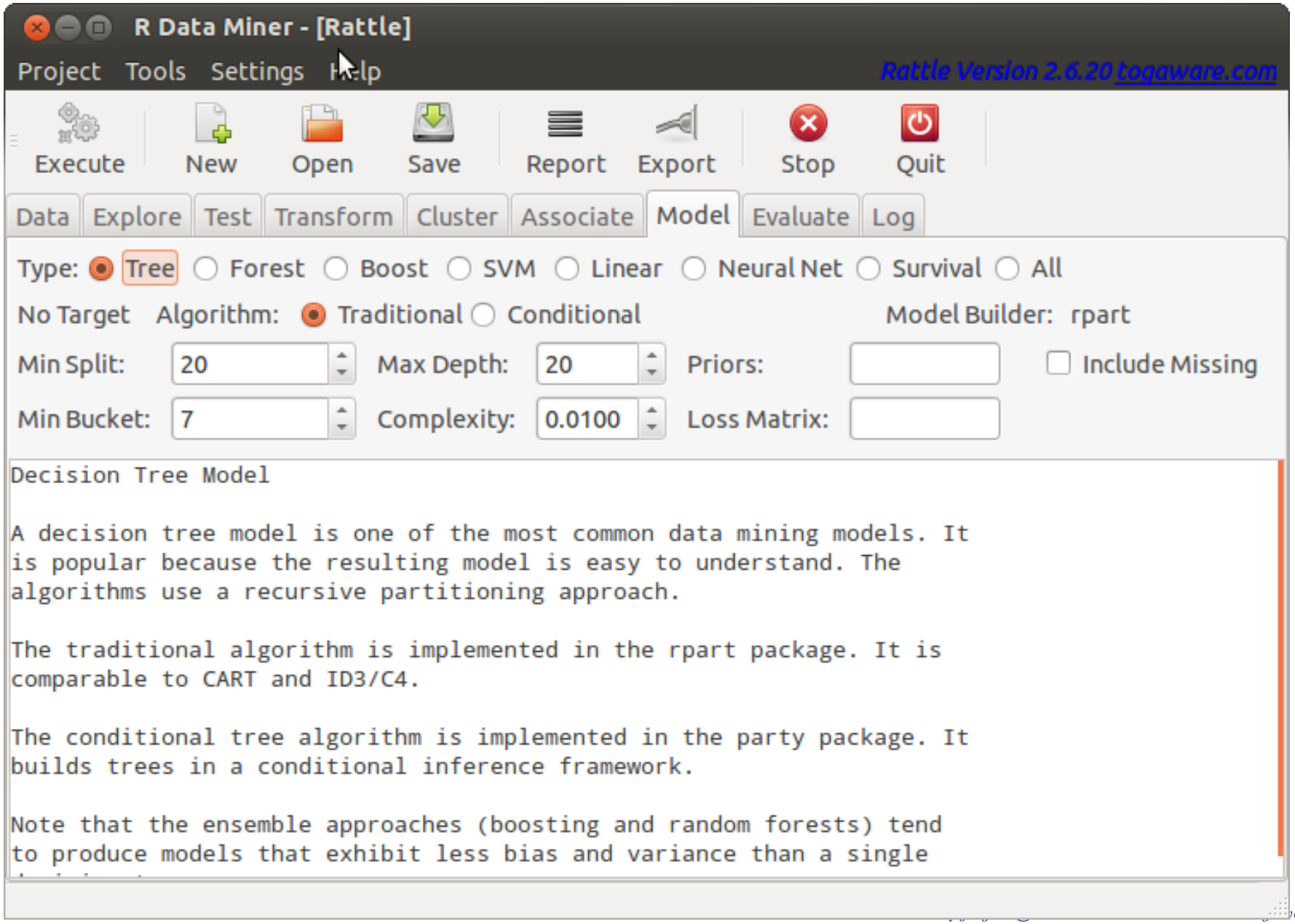






Copyright © 2012 Graham.Williams@togaware.com

Gratuitous Rattle Screenshot



Overview

Revenue Collection

Establishing Capability

Delivering Outcomes

Ongoing Challenges

- Model Deployment
- Big Data: Many Variables

Copyright © 2012 Graham.Williams@togaware.com

Analytics in Action

High Risk Refunds (HRR) identified prior to issuing of refunds.

- Traditional rules identify too many "high risk" refunds.
- Some tests might identify 100,000 cases each year.
- Sometimes as few as 5% are found to require adjustment.
- Revenue at risk can be very significant (from $10m to $1b).

Data Mining modelling for HRR.

- Has identified numerous characteristics to better target risk (5%)
- More effectively deploy resources on productive cases.
- Avoid non-productive audits.
- Uses decision trees and ensembles (random forests).

Copyright © 2012 Graham.Williams@togaware.com

Analytics in Action

High Risk Refunds (HRR) identified prior to issuing of refunds.

- Traditional rules identify too many "high risk" refunds.
- Some tests might identify 100,000 cases each year.
- Sometimes as few as 5% are found to require adjustment.
- Revenue at risk can be very significant (from $10m to $1b).

Data Mining modelling for HRR.

- Has identified numerous characteristics to better target risk (5%)
- More effectively deploy resources on productive cases.
- Avoid non-productive audits.
- Uses decision trees and ensembles (random forests).

Copyright © 2012 Graham.Williams@togaware.com

Other Successful Projects

- High Risk Refunds – ensembles of trees
- Required to Lodge ($110M) – ensembles of trees
- Assessing Levels of Debt – Propensity/Capacity to Pay
- Optimal Treatment Strategies
- Identity Theft – Outliers, Unusual, Out of Pattern
- International and Tax Havens – Text Mining
- Complex Structures – Network and Link Analysis

Copyright © 2012 Graham.Williams@togaware.com

Overview

Revenue Collection

Establishing Capability

Delivering Outcomes

Ongoing Challenges

- Model Deployment
- Big Data: Many Variables

Copyright © 2012 Graham.Williams@togaware.com

Challenge 4: Model Deployment is IT

- Deliver PMML from SAS into Teradata
  open standards and interoperability (KDD98)
- Much data mining is **not** "deployed"
  models are run ad-hoc as required.
- Models "deployed" by conversion
  - SQL: 2 million lines for RF — 20x200x500
  - C: Netezza to score 15M entities in 90 seconds
  - PMML: deployment in e.g. Zementis' Adapa
- Challenge is that many (hundreds of) models have been developed and many need to be "deployed"

Copyright © 2012 Graham.Williams@togaware.com

Model Manager - Models in Production

- Running automatically on demand and event driven.
- Requires a professional production environment:
  - models now have life cycles: traditional dev/test/deploy
  - monitoring of model performance:
    - alerts for out-of-spec models
    - daily dashboards
  - 24/7 support of models

- SAS and SPSS provide model management solutions
- Our developing solution using open source tools:
  - R, Python, Shell, Make (scripting)
  - Bazaar (dev/prod version control)
  - Jenkins (continuous integration framework web interface)
  - ...

Copyright © 2012 Graham.Williams@togaware.com

Challenge 5: Big Data — Many Variables

- Not really too "big" in the ATO
  - 100M transactions
  - 20M entities
  - 5K variables
- Big enough to present some challenges to traditional tools

- The variables are the "big" issue
- Random Forests are good, but could be better!
- Issues when there are many irrelevant variables.

Copyright © 2012 Graham.Williams@togaware.com

Challenge 5: Big Data — Many Variables

- Not really too "big" in the ATO
  - 100M transactions
  - 20M entities
  - 5K variables
- Big enough to present some challenges to traditional tools

- The variables are the "big" issue
- Random Forests are good, but could be better!
- Issues when there are many irrelevant variables.

Copyright © 2012 Graham.Williams@togaware.com

Subspaced Random Forests

Research with Chinese Academy of Sciences
Shenzhen Institutes of Advanced Technology.

- Random forests are a popular classification method building an ensemble of a single type of decision tree.
- Algorithmically intuitive and simple.
- Aim is for an ensemble of very different decision trees, with each decision tree by itself being a good model of the data it is based on.
- How to increase diversity and individual accuracy?

Copyright © 2012 Graham.Williams@togaware.com

Subspaced Random Forests

Research with Chinese Academy of Sciences
Shenzhen Institutes of Advanced Technology.

- Random forests are a popular classification method building an ensemble of a single type of decision tree.
- Algorithmically intuitive and simple.
- Aim is for an ensemble of very different decision trees, with each decision tree by itself being a good model of the data it is based on.
- How to increase diversity and individual accuracy?

Copyright © 2012 Graham.Williams@togaware.com

Hybrid Random Forests

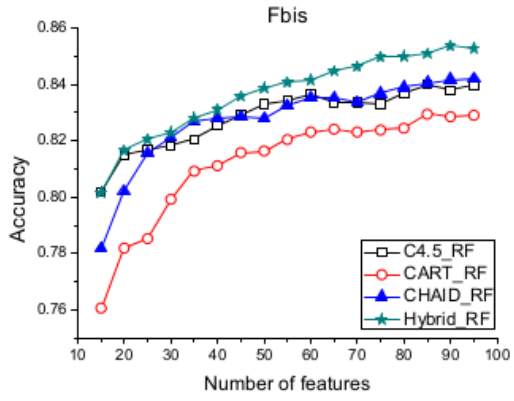
- Simple Idea: *Build different types of decision trees—different algorithms—for each randomly sampled training dataset and choose the best tree model for each.*

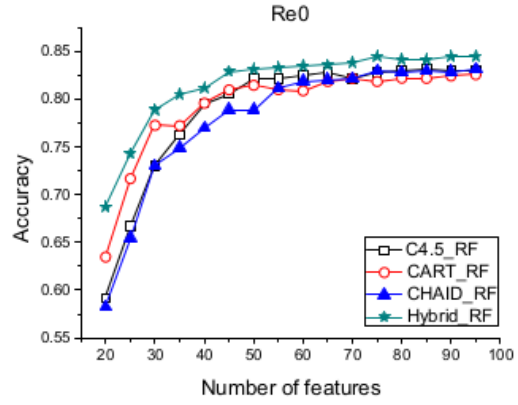
- Tree types considered: C4.5, CART, CHAID
- Further opportunities:
  - HD, EDD, VED (Decision Master)
  - ctree (R/party)

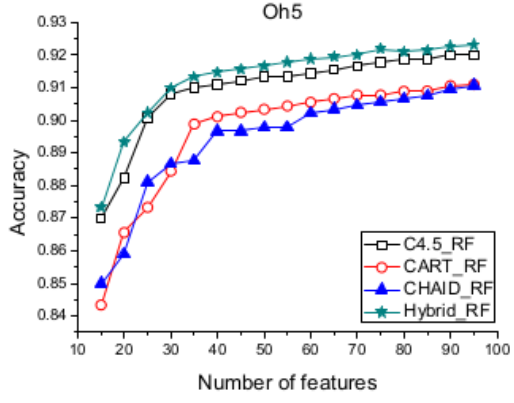
- A series of experiments suggest that the hybrid approach **always** delivers the best model.

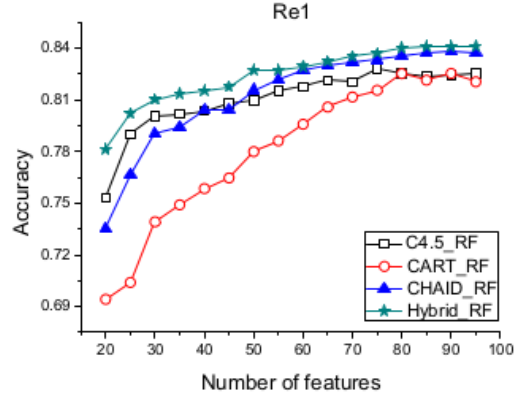
Copyright © 2012 Graham.Williams@togaware.com

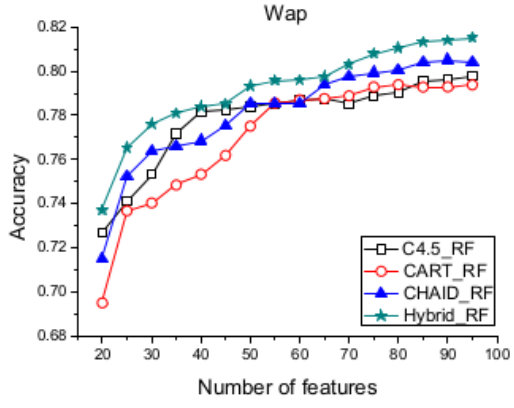
Experimental Results

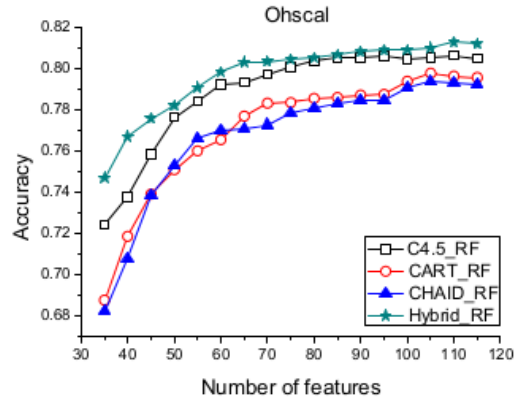












The hybrid random forest always performs best for this collection of datasets!!!

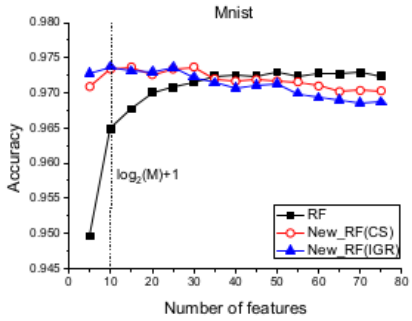
Copyright © 2012 Graham.Williams@togaware.com

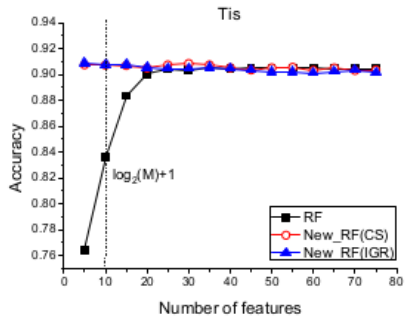
Another Approach: Weighted Subspaces

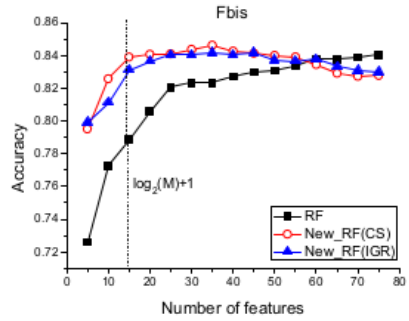
- Performance of a random forest is improved by
  - **Strengthening** each tree
  - Reducing **correlation** between each tree
- Problem of large number of variables:
  - Random selection means too many irrelevant variables
- Introduce the concept of weighted subspace random forests
  - Bias the selection of variables toward most important variables

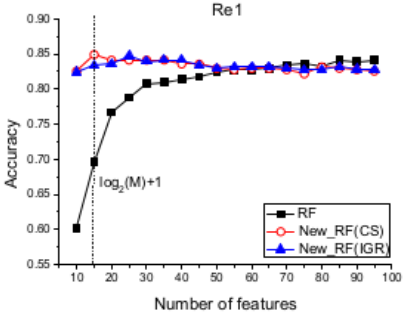
Copyright © 2012 Graham.Williams@togaware.com

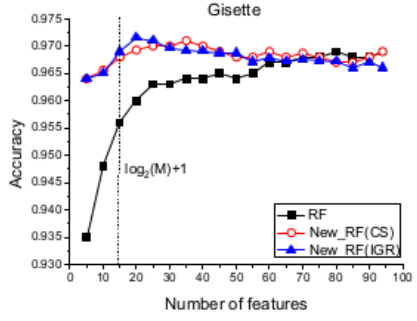
Experimental Results

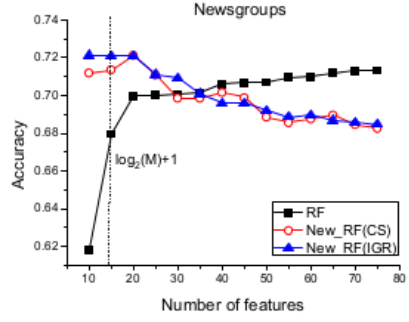


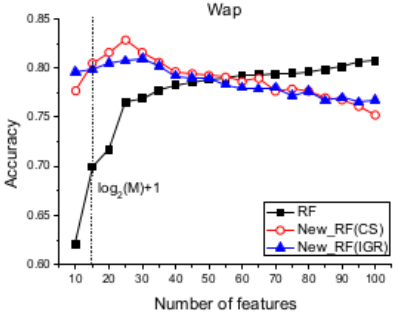


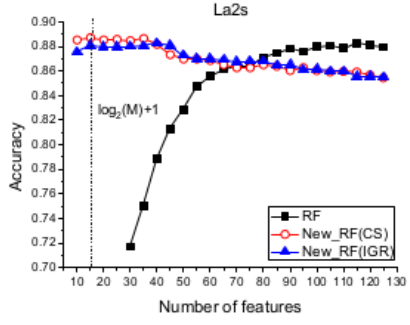


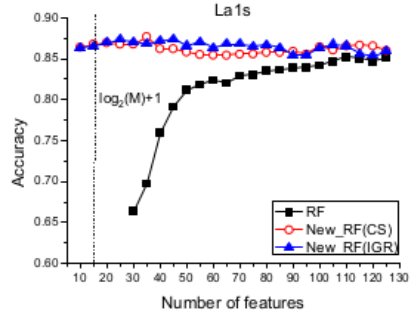












For this collection of experiments the weighted subspace random forest always performs better with many fewer features.

Copyright © 2012 Graham.Williams@togaware.com

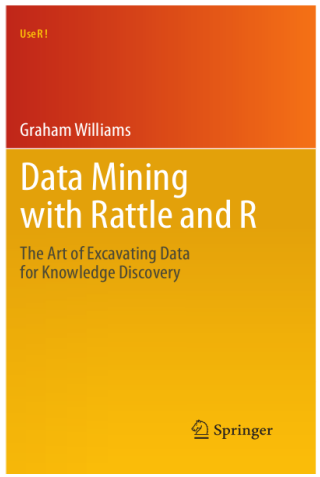
Summary

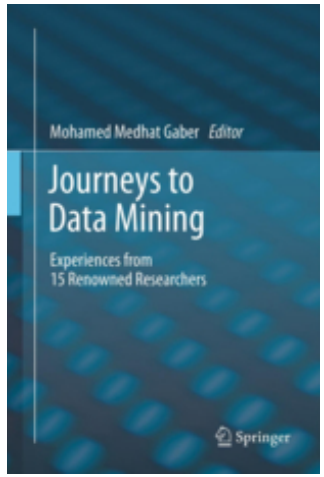
- Focus on the "Analyst First"
- Commodity hardware and software provide excellent capability
- Sharing algorithms through R packages (rattle, wsrpart, wsrf)
- Deployment continues to challenge

Copyright © 2012 Graham.Williams@togaware.com

Resources and References

- Rattle: `rattle.togaware.com`
- Guides: `datamining.togaware.com`
- Practise: `analystfirst.com`
- Book: Data Mining using Rattle/R
- Chapter: Rattle and Other Tales





Copyright © 2012 Graham.Williams@togaware.com

Thank You

Copyright © 2012 Graham.Williams@togaware.com