“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.
Outline

AI and ML - Trees and Ensembles
Open Source ML - R and Rattle
Elastic Data Science - Azure
Linux Data Science VM
What is Machine Learning, Actually?

How to identify patterns in the data – associated with the outcome of interest – Rain Tomorrow?

Recursive Partitioning
... aka Divide and Conquer
... aka Map and Reduce

One of the earliest algorithms and still going very strong!

Even though deep learning with massive data and massive compute characterises our current AI/ML surge.
Ensembles as Foundation for all Learning

- Introduced the concept of an ensemble of decision trees (1987 Australian AI Conference)

- Idea of multiple models was challenging at the time. Why not just build the single best model?

- Ensembles are now the approach of choice (Journeys to Data Mining)

Everything is an ensemble.....
(hammers and nails?)
What is R?

**Language Platform**
- The most popular statistical programming language
- A data visualisation tool
- Open source

**Community**
- 3 million users?
- Taught in most universities
- New and recent graduates use it
- Thriving user groups worldwide

**Ecosystem**
- 10,000+ contributed packages
- Rich application & platform integration
#1 Software for Advanced Analytics

## R Usage Growth
**Rexer Data Miner Survey, 2007-2015**

- 76% of data scientists report using R
- 36% select R as their primary tool

## Language Popularity
**IEEE Spectrum Top Programming Languages**

<table>
<thead>
<tr>
<th>Language Rank</th>
<th>Types</th>
<th>Spectrum Ranking</th>
<th>Spectrum Ranking</th>
<th>Spectrum Ranking</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. C</td>
<td>🎨💡📱</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
</tr>
<tr>
<td>2. Java</td>
<td>🎨💡📱</td>
<td>98.1</td>
<td>99.9</td>
<td>99.3</td>
</tr>
<tr>
<td>3. Python</td>
<td>🎨💡📱</td>
<td>98.0</td>
<td>99.4</td>
<td>95.5</td>
</tr>
<tr>
<td>4. C++</td>
<td>🎨💡📱</td>
<td>95.9</td>
<td>96.5</td>
<td>93.5</td>
</tr>
<tr>
<td>5. R</td>
<td>🎨💡📱</td>
<td>87.9 2016</td>
<td>93.3</td>
<td>92.4</td>
</tr>
<tr>
<td>6. C#</td>
<td>🎨💡📱</td>
<td>86.7</td>
<td>84.8 2015</td>
<td>84.8</td>
</tr>
<tr>
<td>7. PHP</td>
<td>🎨💡📱</td>
<td>82.8</td>
<td>84.5</td>
<td>84.5</td>
</tr>
<tr>
<td>8. JavaScript</td>
<td>🎨💡📱</td>
<td>82.2</td>
<td>83.0</td>
<td>83.0</td>
</tr>
<tr>
<td>9. Ruby</td>
<td>🎨💡📱</td>
<td>74.5</td>
<td>76.2</td>
<td>78.9</td>
</tr>
<tr>
<td>10. Go</td>
<td>🎨💡📱</td>
<td>71.9</td>
<td>72.4</td>
<td>72.8</td>
</tr>
</tbody>
</table>


“C is No. 1, but big data is still the big winner”
Rattle – GUI for Data Mining

Using

- Glade point and click GUI builder (XML)
- RGtk2 bindings for the cross platform GUI
- R to implement all the callbacks

- >20,000 downloads per month

Log tab collects documented, formatted, R scripts as a starting point for real work in R
Demo: Rattle Data Mining

First Model in 4 Clicks
Open Source R but ...

- In-Memory Operation
- Data Movement & Duplication
- Lack of Parallelism
Alpha Version: Rattle with MRS

No limit on the dataset sizes

Parallel data processing and model building

Now supports Microsoft R Server – Big Data
Linux Data Science Virtual Machine - Azure

Vowpal Wabbit

Anaconda

Jupyter

R

Hadoop

TensorFlow

git

#!/bin/bash

R

Eclipse

Julia

Python

WEKA

R Studio

dm1c

XGBoost
Resources

- Overview of the Linux Data Science Virtual Machine
  https://aka.ms/linuxdsvm

- Essentials Guide to Setting Up a Linux DSVM and R and Rstudio and Rattle
  https://aka.ms/lds vm

- Rattle Home Page
  https://rattle.togaware.com