Society of Actuaries in Ireland

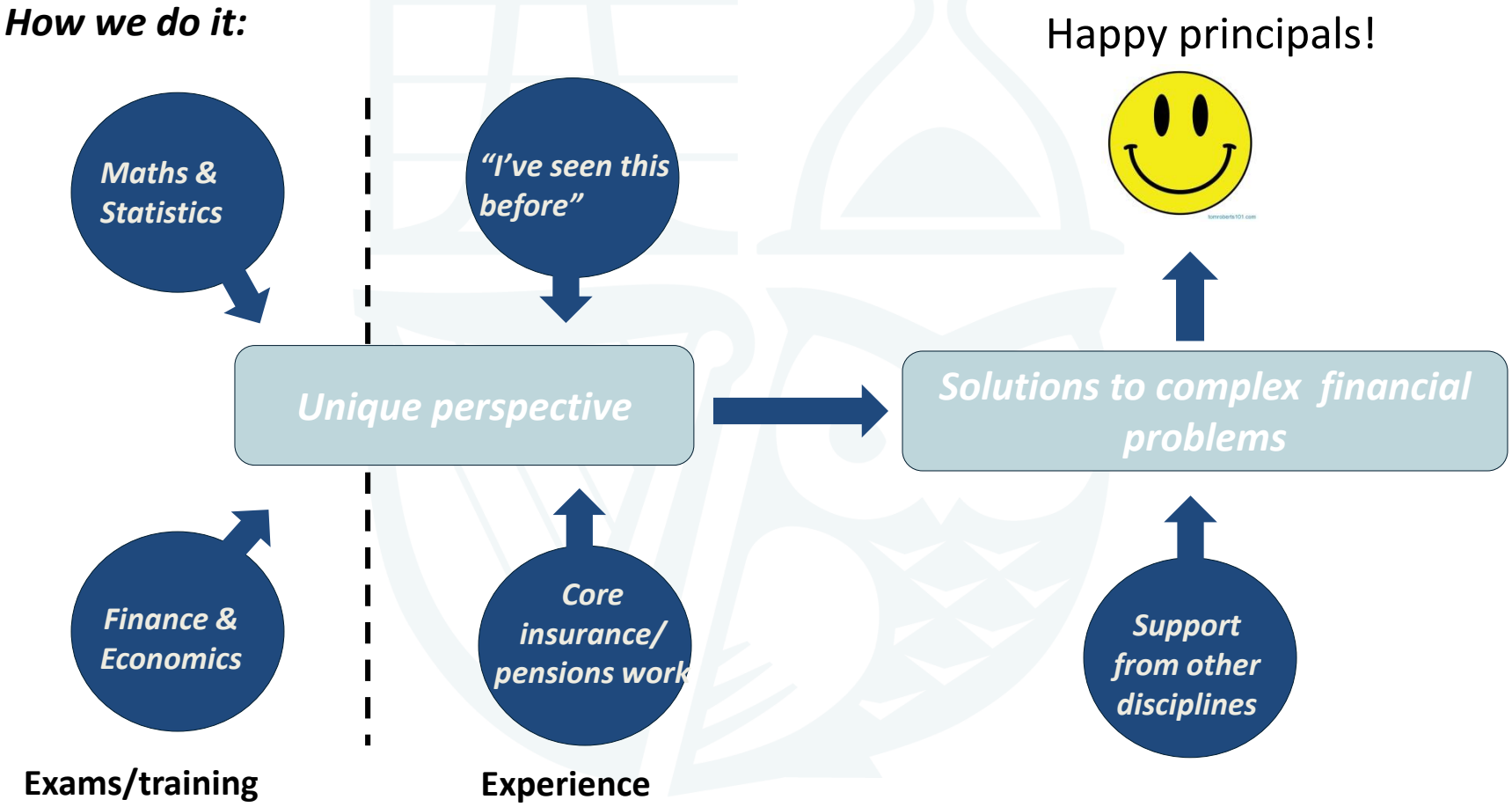
Who is the driver? - Applying data analytics to actuarial and insurance challenges

Thursday 25 June

Actuaries' core competence

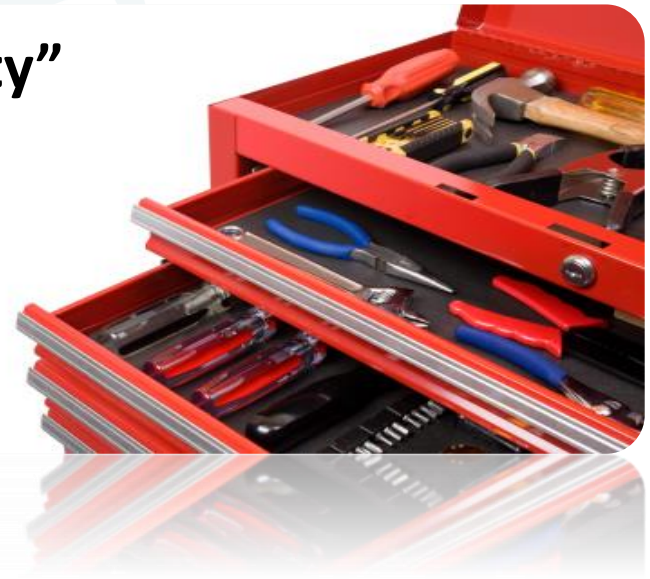
Analyse the financial consequences of uncertain future events

How we do it:



Common problems solved

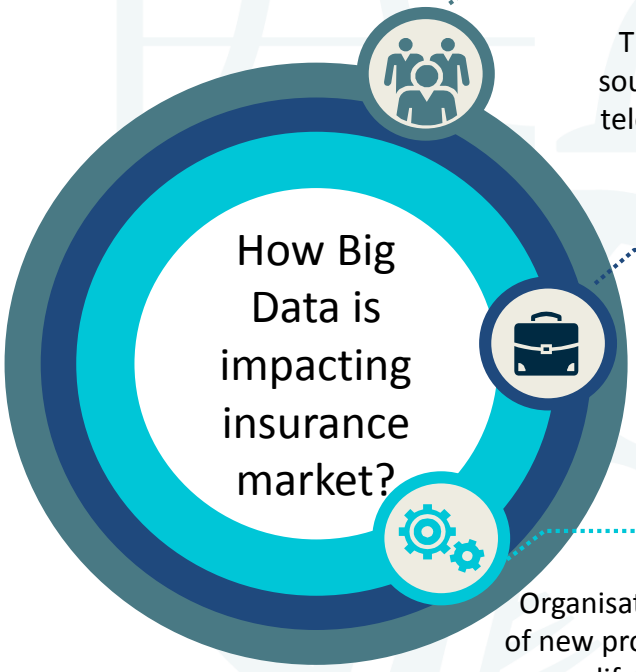
- **“Model uncertainty”** – we develop models to help prioritise and make business choices
- **“Make data talk”** – we use data to derive business insights and drive evidence-based change
- **“Quantify complex risks and uncertainty”** – we help quantify and manage volatile business risks and uncertainty



Big data and analytics are impacting on insurance

The insurance sector has a long history of creating and leveraging mostly domain-specific, data-oriented models

"Insurance is the most obvious" industry "about to explode" with uses for big data (Eric Schmidt, Google Chairman)



Shaping the competition landscape

The data explosion has created a new ecosystem of data sources incorporating a diverse range of new players from telecommunications, retail and automotive companies, to domain-specific data providers and aggregators

Opening new market opportunities

The application of analytical techniques to available data gives visibility on more forward-looking business opportunities. Examples include capital modelling and economic scenario and catastrophe simulations.

Rethinking the products

Organisations are looking at how big data can drive development of new products, such as those based on longer term policies (e.g. life assurance, consumer healthcare) and smart technology provision (e.g. car insurance)

Actuaries and the data science community

- There is a vibrant data science/analytics community
- The data science community thrives on exchange and challenge, e.g. Kaggle as main platform of data science competitions posted by Fortune 500 companies
- Actuaries are still focused on our own community
- Recent actuarial graduates are flexible and better trained to actively participate and benefit
- How can we trigger an interest and get our members interested and involved in these communities?

The Titanic Competition

Knowledge • 2,937 teams

Titanic: Machine Learning from Disaster

Fri 28 Sep 2012 — Thu 31 Dec 2015 (6 months to go)

- An ideal starting point for those interested in data analytics.
- *"In this challenge, we ask you to complete the analysis of what sorts of people were likely to survive. In particular, we ask you to apply the tools of machine learning to predict which passengers survived the tragedy."*
- No cash prize but knowledge.
- Tutorials available in data analytics, R and Python.
- The competition runs from 1st July to 31st December.
- Contact Emily O'Gara at emily.ogara@actuaries.ie if you would like to join.

Next Steps

- Join the competition!
- September Titanic workshop
- End of competition event – share your experience!

- *"R is a language and environment for statistical computing and graphics."* (www.r-project.org)
- Why R? R is free, portable, flexible, widely used and fit for purpose. There are many libraries and tutorials freely available online
- What are the alternatives?
  - Free: RapidMiner, SAS University Edition, Julia...
  - Not so free: SAS, SPSS, Mathematica, STATA...

Who is the driver?

Applying data analytics to actuarial and insurance challenges

Gábor Stikkel

Who am I?

Az 1994/95. évi matematika Országos Középiskolai Tanulmányi Verseny eredményei

- I. díj: Németh Sándor Géza, IV. o., Vác, Boronkay György Műszaki Szakközépiskola, felkészítő tanár: Benedek Ilona
- II. díj: Zaupper Bence, III. o., Győr, Krúdy Gyula Gimnázium és Vendéglátóipari Szki., felkészítő tanár: Babarczi Imréné
- III. díj: Bányai Attila, IV. o., Kaposvár, Eötvös Loránd Műszaki Középiskola, felkészítő tanár: Demeter László
- 4. Kiss Béla, II. o., Vác, Boronkay György Műszaki Szakközépiskola, felkészítő tanár: Újvári István
- 5. Kiss Olivér, IV. o., Debrecen, Mechwart András Gépipari Műszaki Szki., felkészítő tanár: dr. Rutovszky Ede
- 6. Stikkel Gábor, III. o., Eger, Neumann János Közgazdasági Szki. és Gimnázium, felkészítő tanár: Máté Mihályné

Agenda

- AXA Driver Telematics competition on Kaggle
  - Insurance related data science competitions and use cases
- Importance of data analytics in other industries
- Basic skills required to be a data scientist
  - Profile of a 'typical' data analyst
- Software tools for applying data science
  - R Studio
  - A basic solution for the Titanic competition
- Summary

Data science competitions

- Netflix prize: 1 000 000 USD to improve movie recommendations by 10%, 2009
- Foundation of Kaggle, 2010
  - Crowdsourcing predictive modeling
  - Has 326 000+ members
  - Selling data analytics service of the best perform
  - Hosted 179 competitions

AXA Driver Telematics Analysis

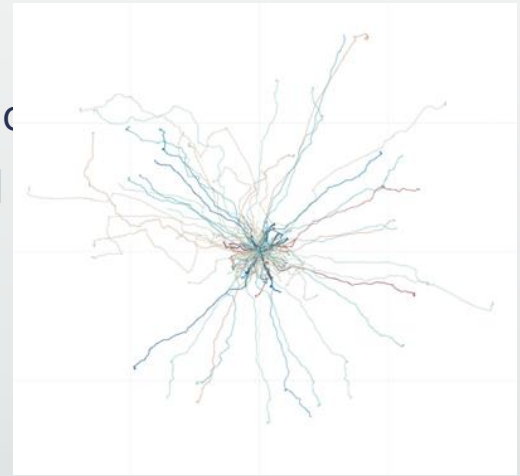
- The intent of this competition is to develop an algorithmic signature of driving type
- Come up with a "telematic fingerprint" capable of distinguishing when a trip was driven by a given driver
- Data available:
  - 2736 folder with 200 trips in each
  - the majority of the trips in one folder belongs to d
  - trip (all centered to origin, randomly rotated and

x,y

0.0,0.0

18.6,-11.1

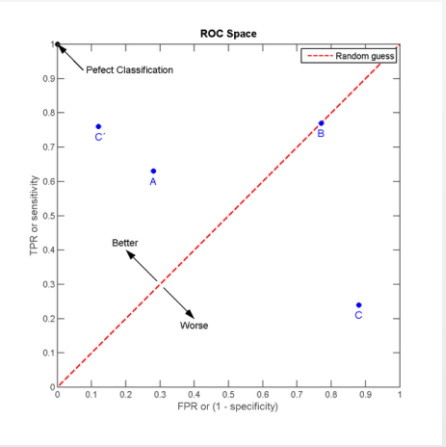
36.1,-21.9



Evaluation

- Submissions (probabilities of each trip belonging to a particular driver) are judged on area under the receiver operating curve (AUC)
- The ROC area is calculated in a global manner (all predictions together)
- You should therefore aim to submit calibrated probabilities between the drivers
- Interpretation of global AUC:

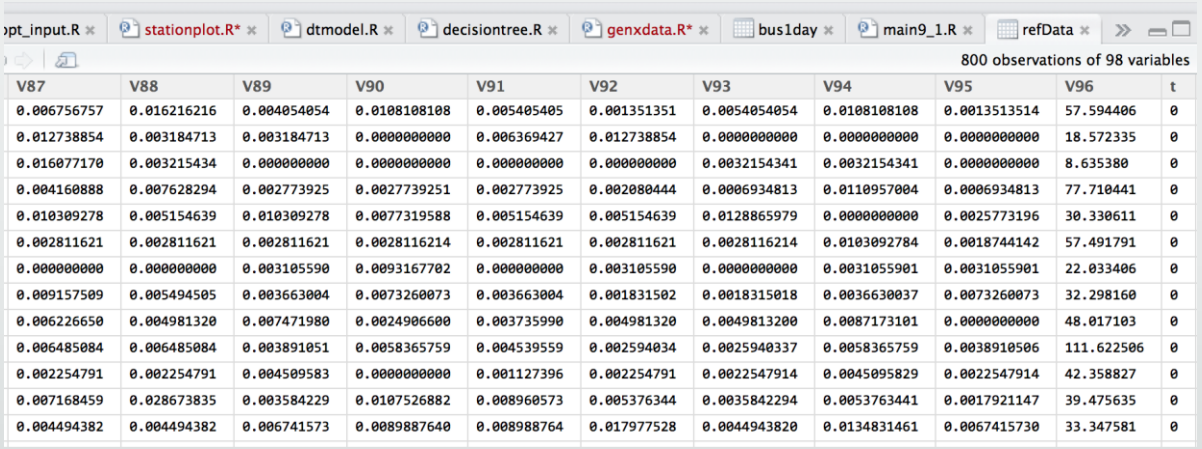
```
count = 0
   for i=1 to m
        for j=1 to n
            if( p[i] > p[j] )
                  count = count + 1;
             else if( p[i] < p[j] )
                  count = count - 1;
AUC = 0.5 + (count/(2*m*n))
```



Here i runs over all m data points with true label 1, and j runs over all n data points with true label 0; p[i] and p[j] denote the probability score assigned by the classifier to data point i and j, respectively.

Main strategy

- Assume that all 200 trips were driven by the same driver, label them as 1
- Select trips from other drivers and label them as 0
  - Strategy of how to do it can vary
- Run supervised learning on this dataset and gather probabilities for the first 200 trips



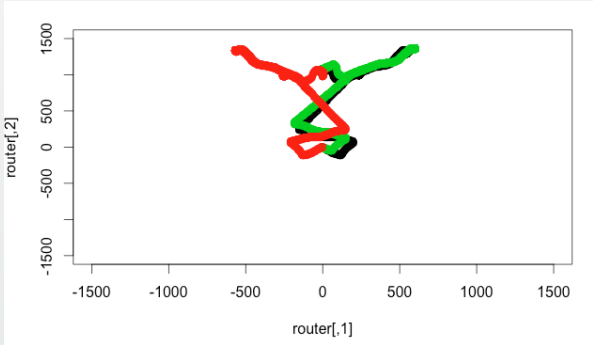
opt_input.R | stationplot.R* | dtmodel.R | decisiontree.R | genxdata.R* | bus1day | main9_1.R | refData

800 observations of 98 variables

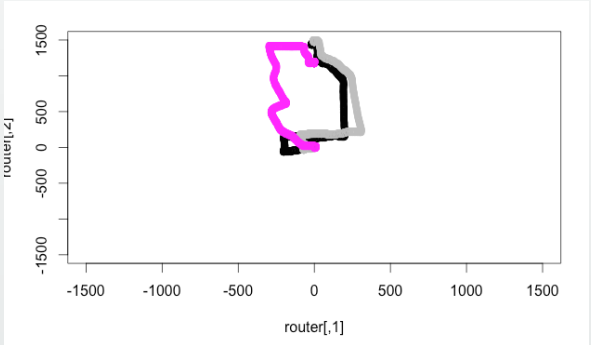
V87	V88	V89	V90	V91	V92	V93	V94	V95	V96	t
0.006756757	0.016216216	0.004054054	0.0108108108	0.005405405	0.001351351	0.0054054054	0.0108108108	0.0013513514	57.594406	0
0.012738854	0.003184713	0.003184713	0.0000000000	0.006369427	0.012738854	0.0000000000	0.0000000000	0.0000000000	18.572335	0
0.016077170	0.003215434	0.000000000	0.0000000000	0.000000000	0.000000000	0.0032154341	0.0032154341	0.0000000000	8.635380	0
0.004160888	0.007628294	0.002773925	0.0027739251	0.002773925	0.002080444	0.0006934813	0.0110957004	0.0006934813	77.710441	0
0.010309278	0.005154639	0.010309278	0.0077319588	0.005154639	0.005154639	0.0128865979	0.0000000000	0.0025773196	30.330611	0
0.002811621	0.002811621	0.002811621	0.0028116214	0.002811621	0.002811621	0.0028116214	0.0103092784	0.0018744142	57.491791	0
0.000000000	0.000000000	0.003105590	0.0093167702	0.000000000	0.003105590	0.0000000000	0.0031055901	0.0031055901	22.033406	0
0.009157509	0.005494505	0.003663004	0.0073260073	0.003663004	0.001831502	0.0018315018	0.0036630037	0.0073260073	32.298160	0
0.006226650	0.004981320	0.007471980	0.0024906600	0.003735990	0.004981320	0.0049813200	0.0087173101	0.0000000000	48.017103	0
0.006485084	0.006485084	0.003891051	0.0058365759	0.004539559	0.002594034	0.0025940337	0.0058365759	0.0038910506	111.622506	0
0.002254791	0.002254791	0.004509583	0.0000000000	0.001127396	0.002254791	0.0022547914	0.0045095829	0.0022547914	42.358827	0
0.007168459	0.028673835	0.003584229	0.0107526882	0.008960573	0.005376344	0.0035842294	0.0053763441	0.0017921147	39.475635	0
0.004494382	0.004494382	0.006741573	0.0089887640	0.008988764	0.017977528	0.0044943820	0.0134831461	0.0067415730	33.347581	0

Features – trip matching

Heuristics work



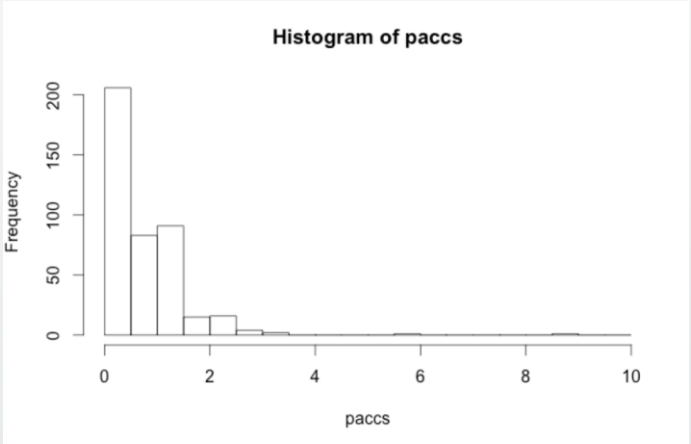
Heuristics does not work



Heuristics: rotate the trip to have the endpoint on y-axis and check which rectangle it can fit

Features - cont.

- Distribution of
  - acceleration (discretized: 0, 0.5, …, 10 m/s^2)
  - deceleration (-10, -9.5, …, 0 m/s^2)
  - speed (0, 2, …, 50 m/s)
  - angles (0, 0.025, …, 0.5 radian)
- Length of trip in seconds
- Distance of trip in meter
- Quantiles of speed (0, 10%, …, 100%)
- Horizontal extension of rotated trip
- Vertical extension of rotated trip
- Sum of angle changes



Major milestones

In terms of AUC score:

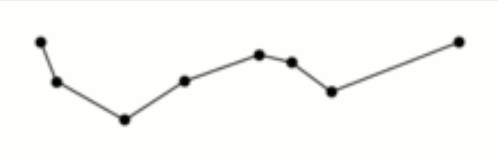
- 0.50961: 1st submission, heuristics of trip length
- 0.62222: still heuristics of trip length and horizontal extension
- 0.69965: zscore on heuristics of trip shape, speed, acceleration
- 0.70150: Random Forest on acceleration and deceleration histograms
- 0.81130: RF on acc, dec, speed and trip length (2x200 random trips)
- 0.90307: RF on full feature set (20x40 random trips)
- 0.91237: blending 10 different RF on full feature set (20x40 random trips)

Winning features

- 0.93398 AUC (7th place) can be achieved by using
  - The derivative of acceleration: jerk
  - Centripetal acceleration
  - Tangential acceleration
  - Heading change distribution
  - Stopping time
  - Number of stops
- Ensembling with Lasso-regression instead of just averaging

Winning features and ideas

- Trip matching does count!
- Ramer-Douglas-Peucker algorithm:



```
function DouglasPeucker(PointList[], epsilon)
    // Find the point with the maximum distance
    dmax = 0
    index = 0
    end = length(PointList)
    for i = 2 to ( end - 1) {
        d = shortestDistanceToSegment(PointList[i], Line(PointList[1], PointList[end]))
        if ( d > dmax ) {
            index = i
            dmax = d
        }
    }
    // If max distance is greater than epsilon, recursively simplify
    if ( dmax > epsilon ) {
        // Recursive call
        recResults1[] = DouglasPeucker(PointList[1...index], epsilon)
        recResults2[] = DouglasPeucker(PointList[index...end], epsilon)

        // Build the result list
        ResultList[] = {recResults1[1...length(recResults1)-1] recResults2[1...length(recResults2)]}
    } else {
        ResultList[] = {PointList[1], PointList[end]}
    }
    // Return the result
    return ResultList[]
end
```

Insurance related competitions

Competition Name	Reward	Teams	Deadline
 **Heritage Health Prize** Identify patients who will be admitted to a hospital within the next year using historical claims data. (Enter by 06:59:59 UTC Oct 4 2012)	$500,000	1353	2 years ago
Liberty Mutual Group - Fire Peril Loss Cost Predict expected fire losses for insurance policies	$25,000	634	9 months ago
 **Claim Prediction Challenge (Allstate)** A key part of insurance is charging each customer the appropriate price for the risk	$10,000	102	3 years ago
limited **As the World Churns** Predict which customers will leave an insurance company in the next 12 months.	Private	37	18 months ago

Insurance related big data use cases

- PWC's approach to cure churn:

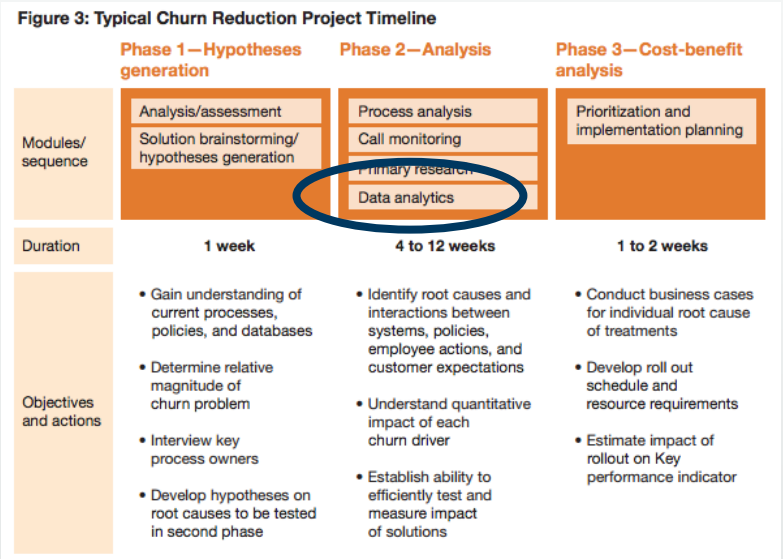


Figure 3: Typical Churn Reduction Project Timeline

	Phase 1—Hypotheses generation	Phase 2—Analysis	Phase 3—Cost-benefit analysis
Modules/ sequence	Analysis/assessment Solution brainstorming/ hypotheses generation	Process analysis Call monitoring Primary research Data analytics	Prioritization and implementation planning
Duration	1 week	4 to 12 weeks	1 to 2 weeks
Objectives and actions	• Gain understanding of current processes, policies, and databases • Determine relative magnitude of churn problem • Interview key process owners • Develop hypotheses on root causes to be tested in second phase	• Identify root causes and interactions between systems, policies, employee actions, and customer expectations • Understand quantitative impact of each churn driver • Establish ability to efficiently test and measure impact of solutions	• Conduct business cases for individual root cause of treatments • Develop roll out schedule and resource requirements • Estimate impact of rollout on Key performance indicator

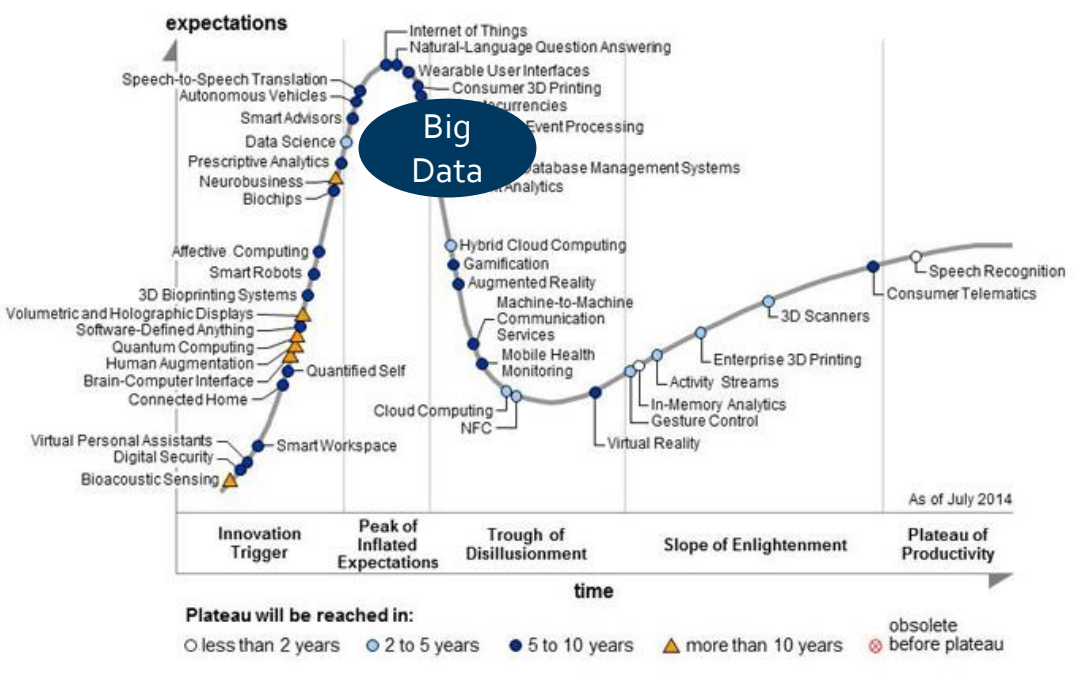
- Smartcasco: cellphone assisted car insurance (30% cheaper than normal)
- Liberty Mutual Insurance and Nest Partner to Reward Customers For Protecting Their Homes With Innovative Technology

A possible application of driver's signature

Importance of data analytics in other industries

- Data is the new oil
- Data is the new gold
- Report from McKinsey Global Institute (2011) on Big Data claims it is the next frontier for innovation, competition, and productivity
- The trend is visible everywhere: from telecom through agriculture to psychology
  - Connected vehicle cloud platform for pay-as-you-go car insurance policies
- Even following blogs on big data is time consuming ☺
  - Paco Nathan
  - Bernard Marr

Data is the new gold

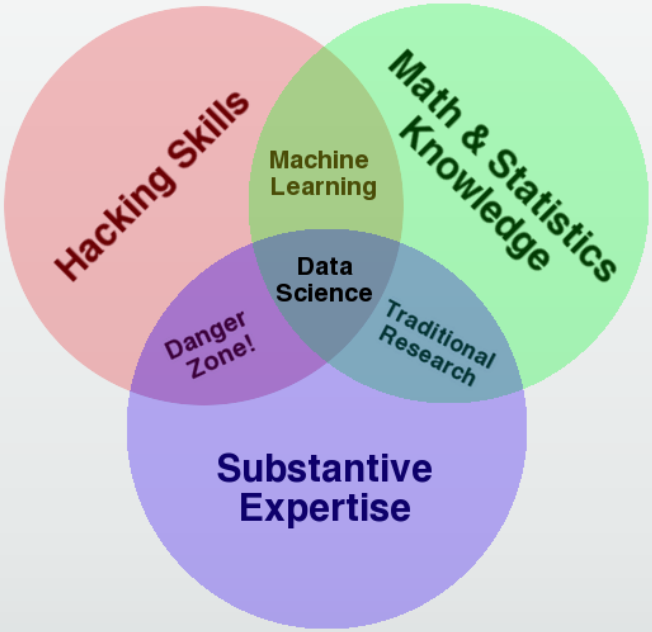


Gartner, August 2014

Agenda

- AXA Driver Telematics competition on Kaggle
  - Insurance related data science competitions
- Importance of data analytics in other industries
- **Basic skills required to be a data scientist**
  - Profile of a 'typical' data analyst
- Software tools for applying data science
  - R Studio
  - A basic solution for the Titanic competition
- Summary

Data science Venn diagram

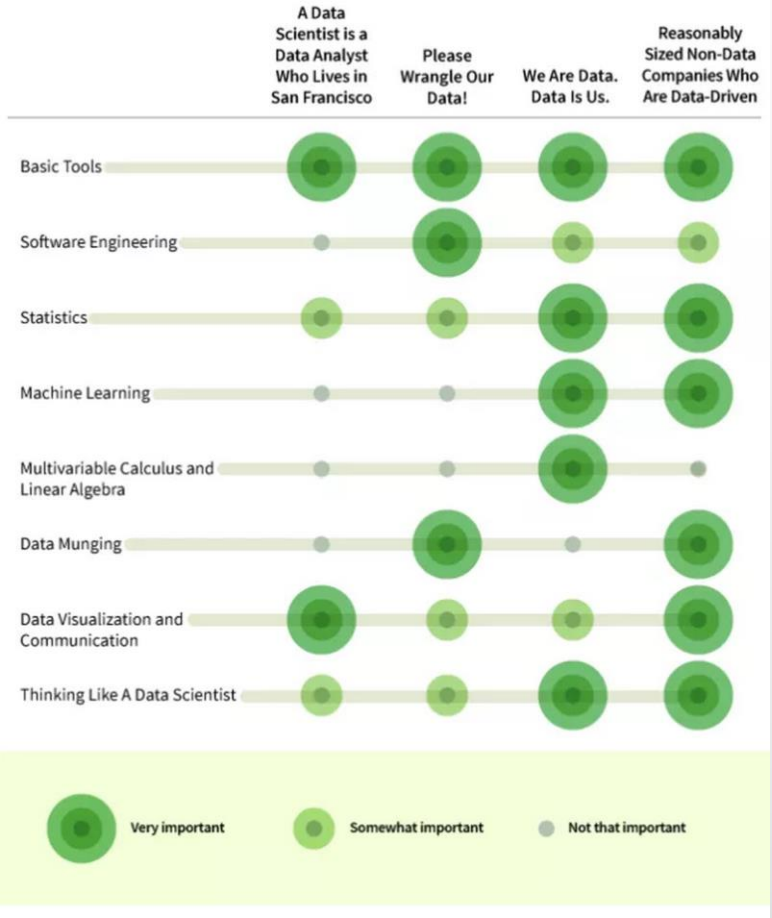


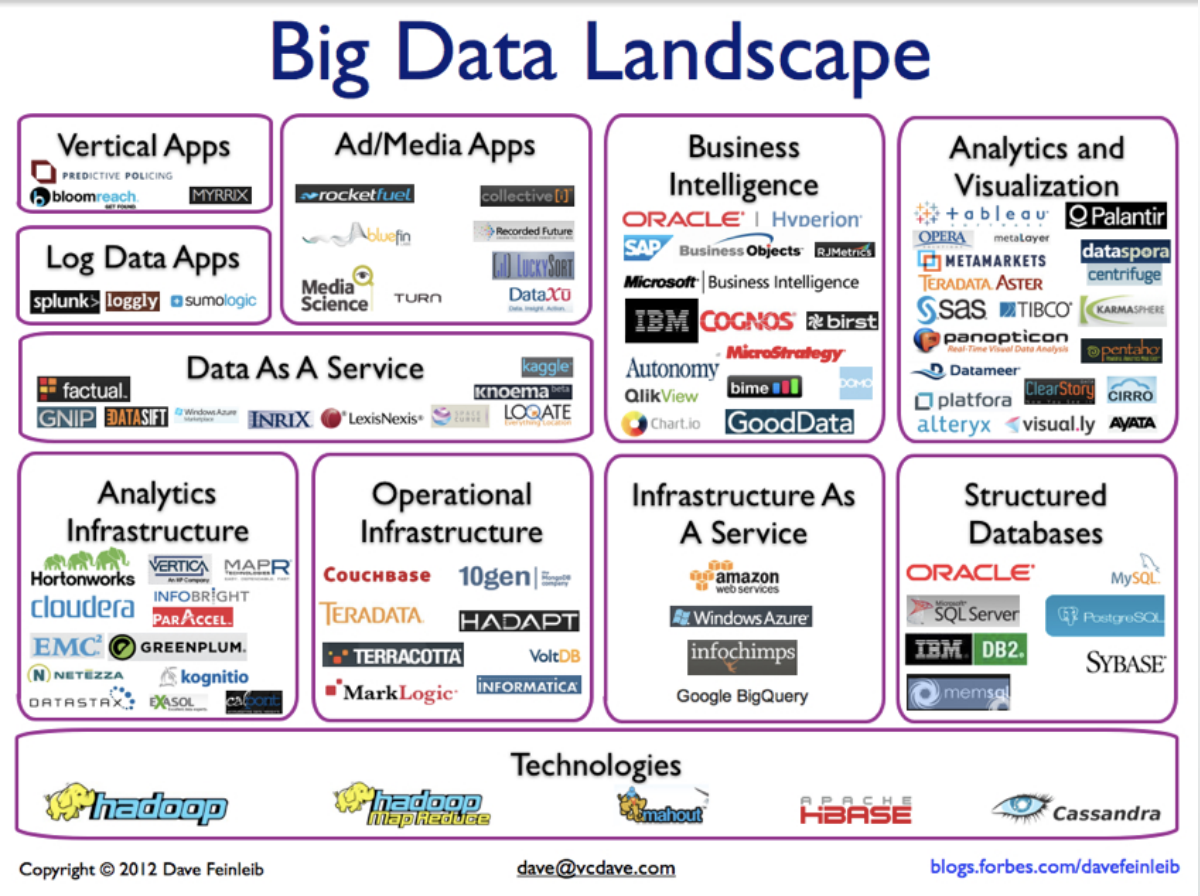
Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician.

http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

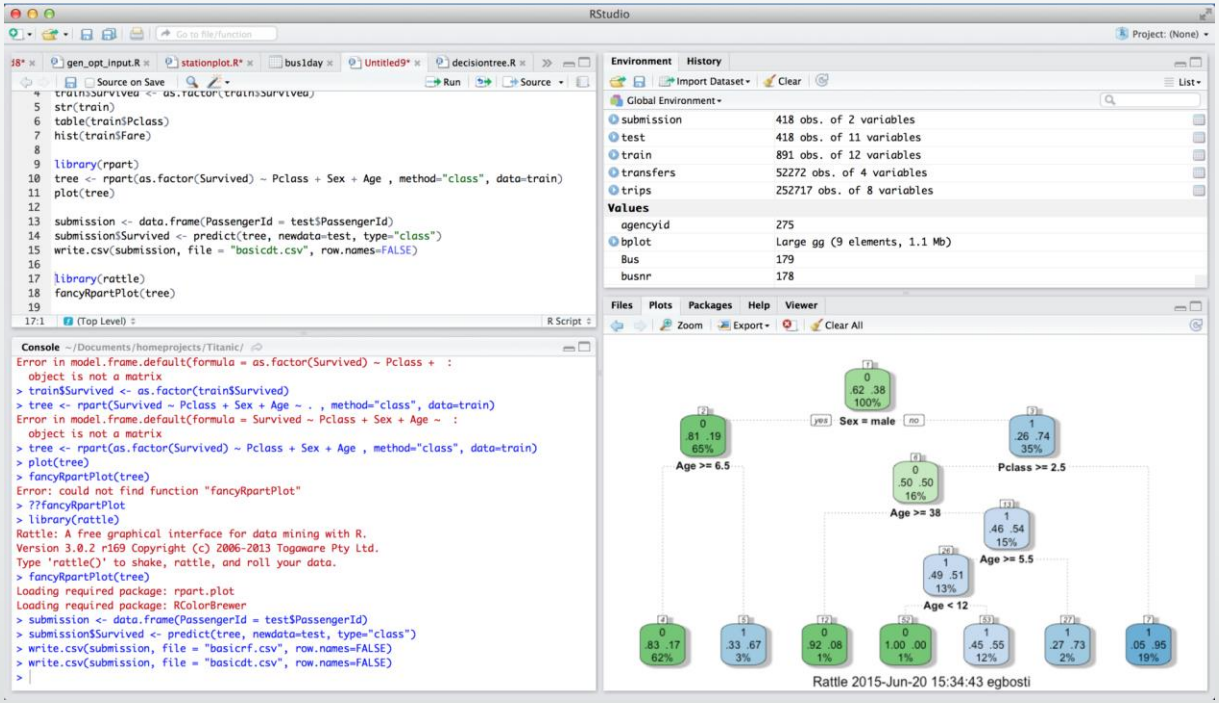
Skills required for data science

- 4 types of data science jobs with a breakdown of the 8 skills you need
- The sexiest job of the 21st century
- Coursera learnings
- MIT edX course on data analytics with Spark




Big Data Landscape
Vertical Apps
PREDICTIVE POLICING
bloomreach
MYRRIX
Log Data Apps
splunk
loggly
sumologic
Ad/Media Apps
rocketfuel
collective[i]
bluefin
Recorded Future
Media Science
TURN
LuckySort
DataXu
Business Intelligence
ORACLE | Hyperion
SAP
Business Objects
RJMetrics
Microsoft | Business Intelligence
IBM
COGNOS
birst
MicroStrategy
Autonomy
QlikView
bime
Chart.io
GoodData
Analytics and Visualization
+ableau
Palantir
OPERA
metaLayer
METAMARKETS
dataspora
centrifuge
TERADATA ASTER
SAS
TIBCO
KARMASPHERE
panopticon
pentaho
Datameer
platfora
ClearStory
CIRRO
alteryx
visual.ly
AYATA
Data As A Service
kaggle
factual
knoema
GNIP
DATASIFT
INRIX
LexisNexis
LOQATE
Analytics Infrastructure
Hortonworks
VERTICA
MAPR
cloudera
INFOBRIGHT
ParAccel
EMC²
GREENPLUM
NETEZZA
kognitio
DATASTAX
EXASOL
calpont
Operational Infrastructure
Couchbase
10gen
TERADATA
HADAPT
TERRACOTTA
VoltDB
MarkLogic
INFORMATICA
Infrastructure As A Service
amazon web services
Windows Azure
infochimps
Google BigQuery
Structured Databases
ORACLE
MySQL
SQL Server
PostgreSQL
IBM DB2
SYBASE
memsql
Technologies
hadoop
hadoop MapReduce
mahout
APACHE HBASE
Cassandra
Copyright © 2012 Dave Feinleib
dave@vcdave.com
blogs.forbes.com/davefeinleib

Framework for data science



Summary

- Fun to compete especially when discussing ideas with your colleagues at the coffee machine
- One will always find something new to learn in R
  - Running the AXA models on multicore, link to my code in Github
- Feature engineering tests creativity
- Data is the new gold

References

- The competition's webpage: https://www.kaggle.com/c/axa-driver-telematics-analysis
- My blog post about the AXA telematics competition: http://www.ericsson.com/research-blog/data-knowledge/adventures-in-the-data-mine/
- Andrei Olariu's blog: http://webmining.olariu.org/kaggle-driver-telematics/
- Connected vehicle cloud concept of Ericsson: https://www.youtube.com/watch?v=NK12lhW_Siw
- R Studio: http://www.rstudio.com/

„Prediction is very difficult, especially about the future“

Niels Bohr