

Big Data Analytics and its Applications

Steve King, Rolls-Royce Engineering Associate Fellow & EHM Specialist - Analytical methods

IMA Lecture – 16th May 2018

© 2018 Rolls-Royce plc and/or its subsidiaries

The information in this document is the property of Rolls-Royce plc and/or its subsidiaries and may not be copied or communicated to a third party, or used for any purpose other than that for which it is supplied without the express written consent of Rolls-Royce plc and/or its subsidiaries.

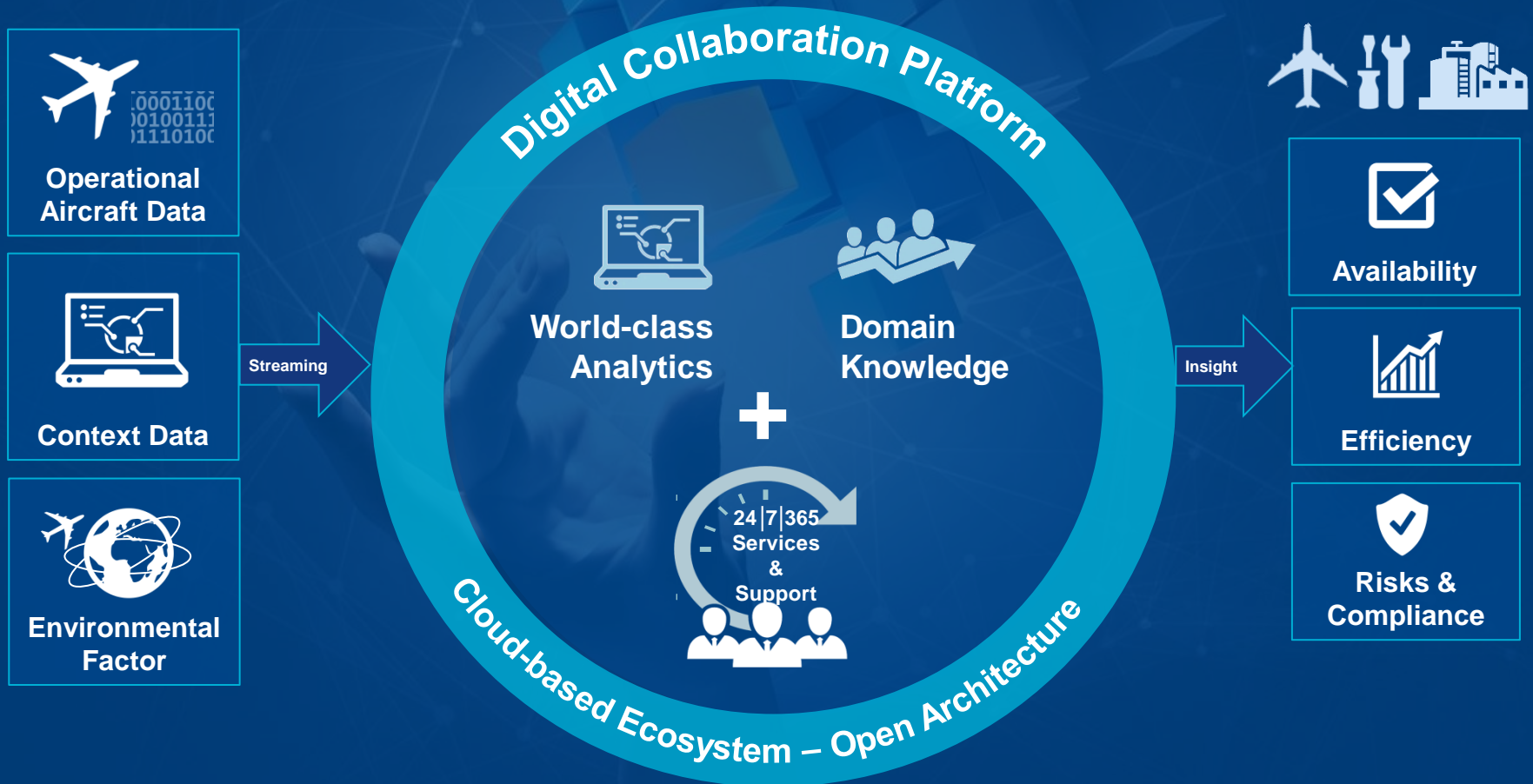
This information is given in good faith based upon the latest information available to Rolls-Royce plc and/or its subsidiaries, no warranty or representation is given concerning such information, which must not be taken as establishing any contractual or other commitment binding upon Rolls-Royce plc and/or its subsidiaries.

Trusted to deliver excellence



Rolls-Royce

Leading provider of actionable insights to the civil aerospace industry



Decades of experience turning data into insight

Close to
30 Years
of experience
in Data Integration
and Analytics

1,300
Aerospace Customers
of our Data Services

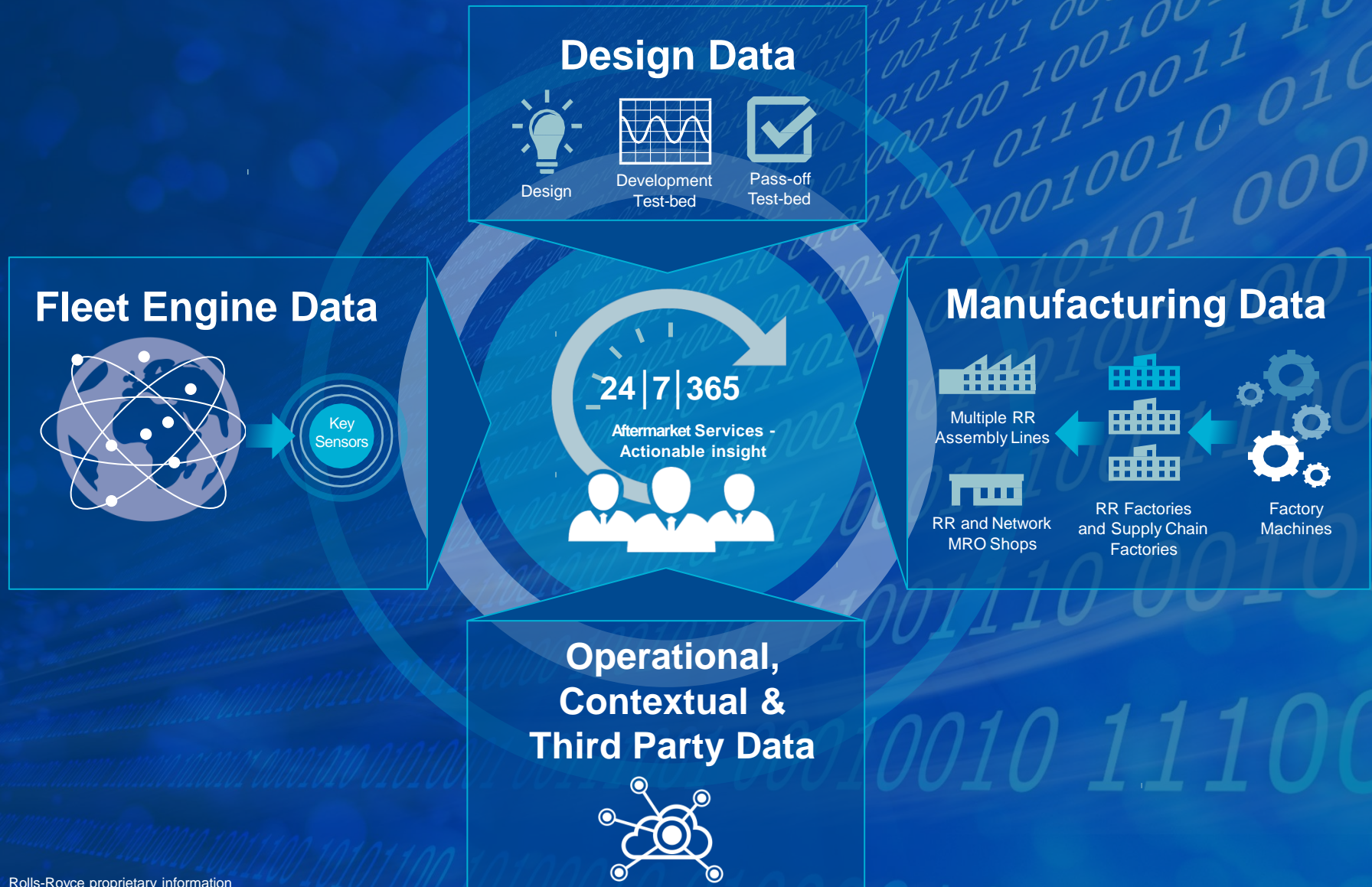
13,000
Engines,
and
6000-8000
commercial flights
monitored per day in
civil aerospace

Billions
of data points
analysed on-board
per flight



EHM

Connecting the end to end data system



Airlines demand predictable operation



- **Customers want a dependable service**
- **Cost of disruption is significant, for example, an in-flight shut-down can lead to:**
 - **diversion to remote site**
 - **overnight for passengers**
 - **replacement aircraft**
 - **supply spare engine**
 - **disrupt follow-on flights**



Rolls-Royce

Engines are the most complex systems on the aircraft...

Ultra efficient swept fan for reduced operational noise & optimum core protection.

High pressure turbine generates over 50,000hp. Each blade generates 800 hp \approx formula 1 racing car

Compressor exit conditions $\sim 700^{\circ}\text{C}$, 50bar pressure

Force on each fan blade at take-off is ~ 100 tonnes



Delivers 75,000-97,000 lb_f thrust with a 10% fuel burn advantage over legacy engines

Combustor incorporates advanced ceramic coatings to allow mix of air and fuel to burn at temps exceeding $2,000^{\circ}\text{C}$

Trent XWB Engine



Rolls-Royce



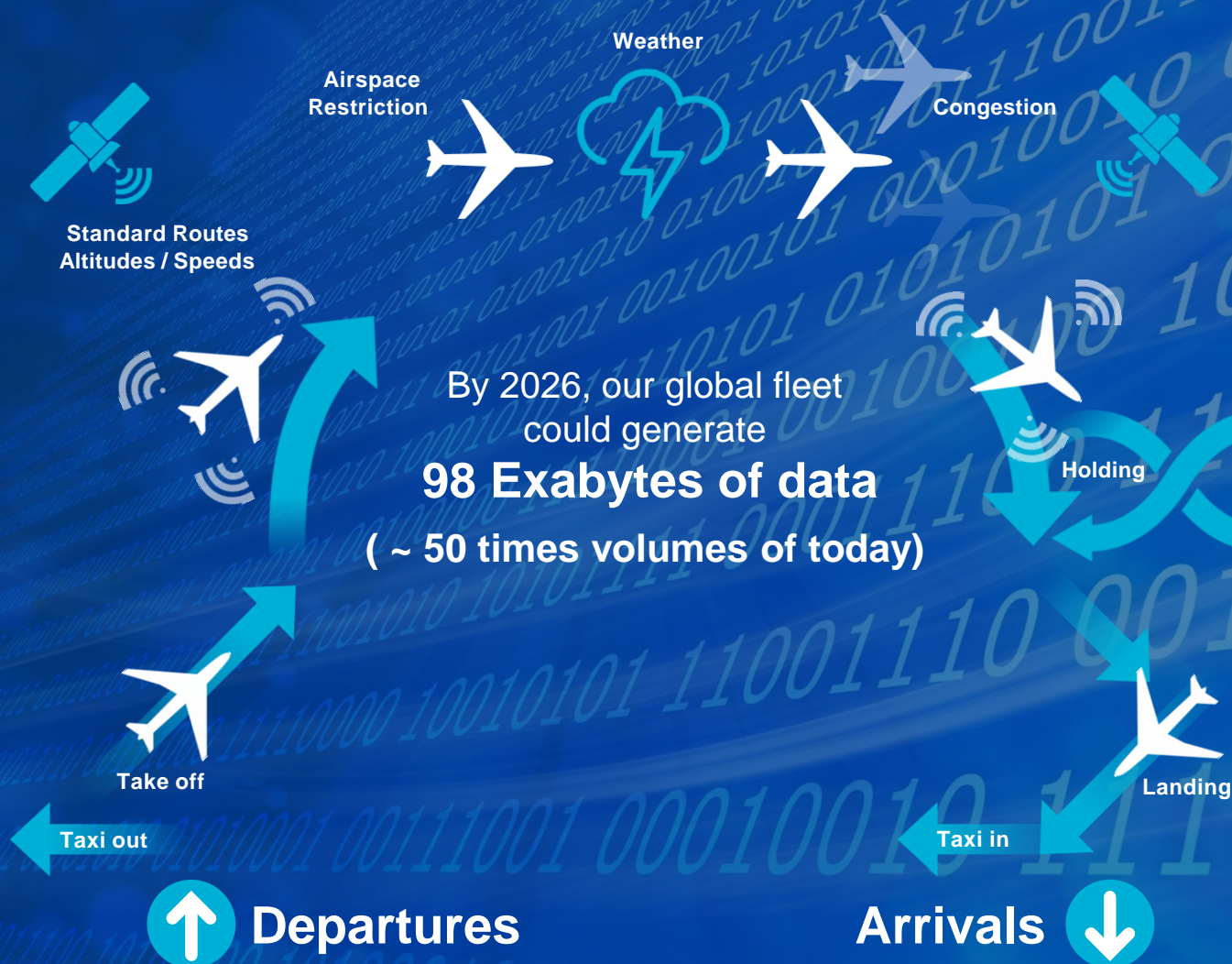
Digital

An exciting technological future
Private - Rolls-Royce Data

An exciting technological future
Private - Rolls-Royce Data

Data growth & complexity

Operational Data volumes are doubling every 2 years



What is Big Data?

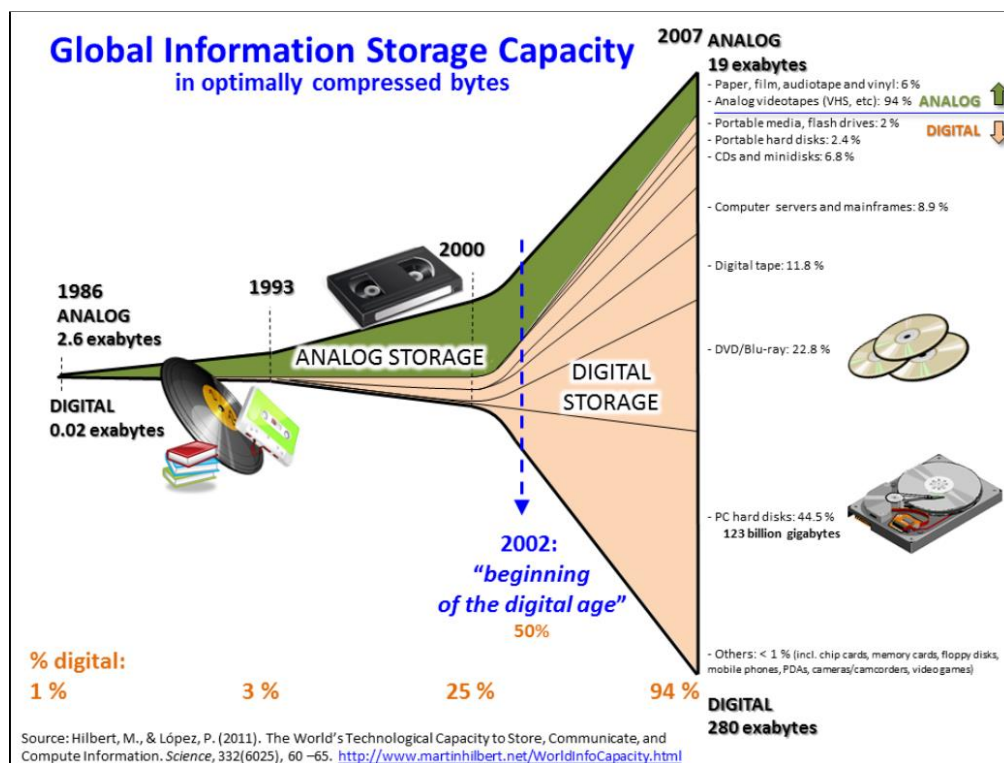
- Term used for data-sets that are so large or complex that traditional data processing apps are inadequate to deal with them:
 - Analysis, capture, curation, searching, storage, transfer, etc.
- It is estimated that our technological capacity to store information has approximately doubled every 40 months since 1980 (source Wikipedia):
 - Higher resolution cameras, wireless sensor networks, mobile devices, increasing storage density (doubles every 12 months) at lower cost



Rolls-Royce

What is Big Data

- Definition of big data at this [link](#) includes the cartoon below.
- This implies digital age commenced in ~2002.
- Coincides with initiative by UK government to fund various e-Science projects related to Grid computing



The term is actually relative and constantly a moving target.

Gigabytes (10^9 bytes)



Terabytes (10^{12} bytes)



Petabytes (10^{15} bytes)



Exabyte (10^{18} bytes)



Rolls-Royce

What is the Grid?

The term is inspired by the electric power grid, which implements standards for electrical power transmission that:

- allows for the decoupling of consumer and provider
- provides a mechanism to link diverse providers into a managed utility

The Grid is a software infrastructure that enables flexible, secure, co-ordinated resource sharing among dynamic collections of individuals, institutions and resources includes computational systems and data storage resources and specialized facilities.

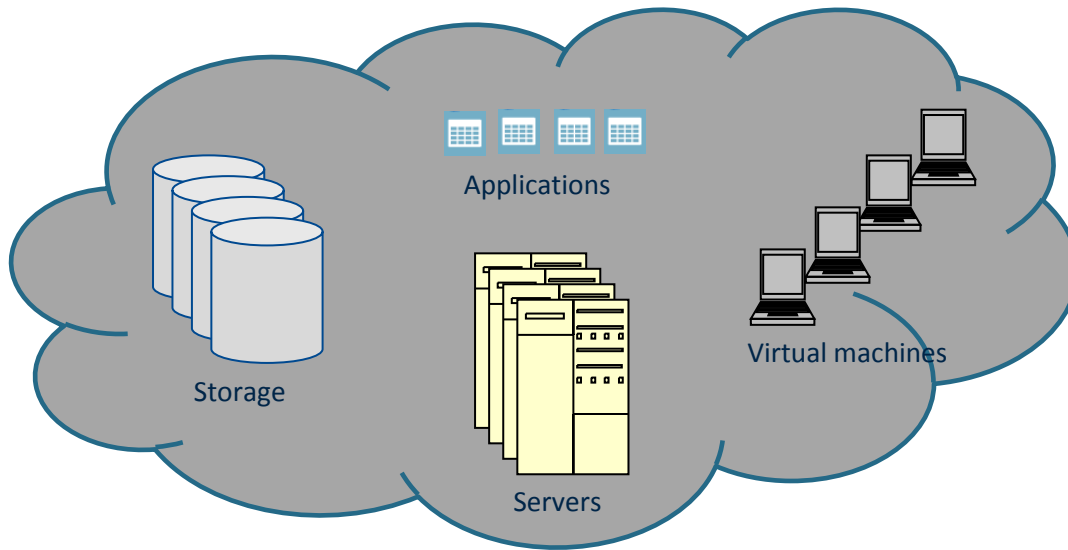
It is also an enabler for transient 'virtual organisations'

Much of the technology developed at that time now forms key components of what is often referred to as cloud computing with many elements available as *open-source* solutions



Rolls-Royce

Cloud Computing



- Network based computing often utilising internet connectivity
- Provides on demand services anywhere, anytime
- Pay for use as when needed with elastic capability
 - Flexibly scale up and down in capacity & functionality
- Hardware and software services available to general public, corporate enterprise & commerce
- Framework to enable data streaming from remote connected devices (e.g. Internet Of Things)

Consider 'Raspberry Pi' - Model 3B

- Quad-Core ARM Cortex-A53
- 1.2GHz
- 1 GB RAM
- Built-in Wi-Fi and 4-port USB hub
- ~£30
- Very compact
- Not many years ago this would have been considered a decent desk-top computer spec
- Compute capability in small foot-print allows easier connectivity of devices/instrumentation for wide range of IoT applications (remote sensing)



Rolls-Royce

Big Data Analytics

Big data analytics involves the process of examining large data volumes to identify hidden patterns, unknown correlations, trends and other descriptive statistics that can provide useful insight from which business value can be derived.

Example Applications include:

- Decision Support
- Financial Analysis
- Social Media Monitoring
- Watson
- Data Exploration & Visualisation
- Anomaly Detection (Fraud, Espionage, Security, Surveillance, Health monitoring – Medical and Industrial)
- Sensor Monitoring
- Telecoms network Monitoring
- Traffic Management
- Text Mining & Language Translation
- Automatic Image & Video Scene recognition/tagging



Rolls-Royce

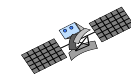
Consider a typical Service Application (Equipment Health Monitoring)

Sense



Acquire

Engine Monitoring Unit



ACMS Reports
via ACARS

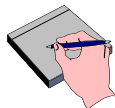


Ground-based
information,
e.g. oil uplift

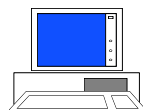
QAR, DFDR



Flight Log
Sheets



Ground
Station



Transfer

Global Network
eg: SITA



24x7 Engine Health Center



Condition monitoring,
Data processing & storage,
Data access & reports,
Forecasting services

Analyse

Maintenance Centre



Customer



Internet, e-mail, pager

OEM



Service Rep



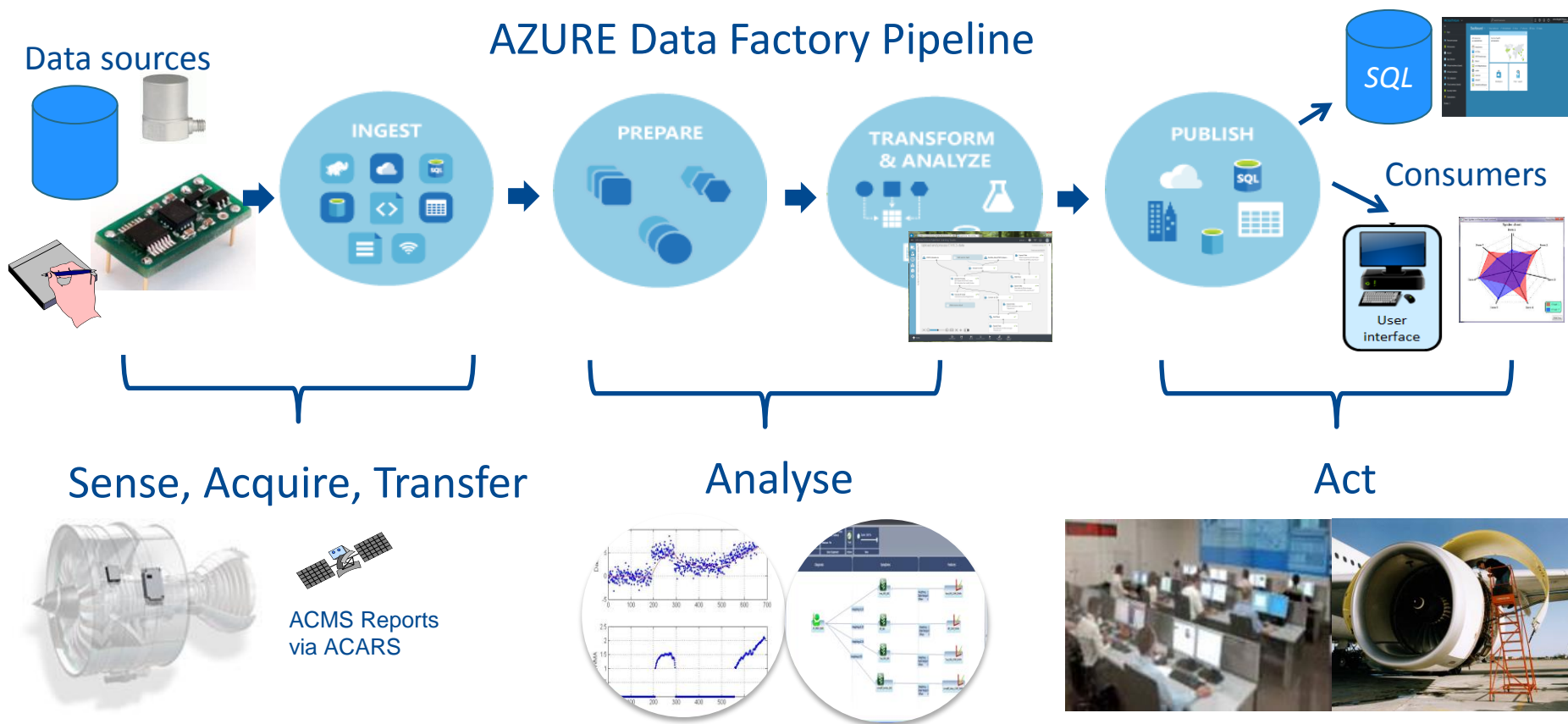
Act



Rolls-Royce

Equipment Health Monitoring

Data volumes to manage fleets continuously increasing, therefore growth in Big-data analytics, IoT and cloud capabilities



Rolls-Royce

Big Data/Cloud computing

- Big data is therefore:
 - high volume,
 - often high velocity
 - consisting of different data types (variety)
- The above often referred to the three key V's of big data
- veracity is another important factor and relates to the quality of captured data, and its variation between sources, which can impact any subsequent analysis.
- Although considered disruptive technology, much of the analysis methods underpinning Big data capability are based on established techniques (e.g. signal processing, machine learning/data mining, statistical methods).
- Scalability of compute resources, provided by cloud technologies, and the mechanics of plumbing these tools together are key enablers for handling *Big data*.



Rolls-Royce

What is Machine Learning?

Some claims:

- A technique to derive knowledge from data (e.g. “if-then” style rules).
- A technique to establish how parameters in a process are related.
- A method to assist in the understanding and visualisation of complex data.
- Considered to be a branch of AI. Essentially, its the mechanisms that enable computers to learn without being explicitly programmed (e.g. a regression model)



Rolls-Royce

Data Analytics & Data Mining?

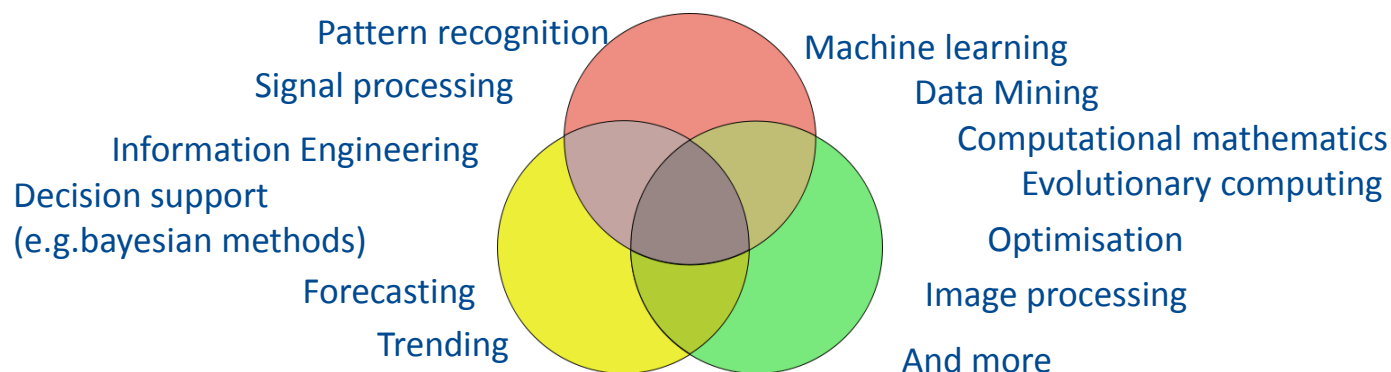
- The science of discovery from data and processing data to extract evidence so that we can make informed decisions
- Traditional exploratory data analysis, based on statistical methods, tend to give emphasis on *inference* through hypothesis testing – e.g. ‘*Is there evidence in the data suggesting x is the same as y to a satisfactory level of significance*’
- In *DM* the emphasis is more on *description*. Hence, focus is more on producing a solution that can generate useful predictions – asking questions like ‘*what relationships/patterns are in my database?*’



Rolls-Royce

Data Analytics & Data Mining?

- Therefore, *Data Mining* accepts among others a "black-box" approach to data exploration or knowledge discovery and uses not only the traditional statistical techniques, but also other techniques such as neural networks, clustering and so on.
- *DM* practitioners may be seen as rushing in where Statistician's fear to tread
 - *They possibly achieve more, but need to take care to avoid mistakes in interpretation.*
- These same techniques can be found in a number of other related topics areas



Rolls-Royce

Some past observations:

- All Models are wrong, some are useful (George Box)
- But then Chris Anderson states: *Today companies like Google, which have grown up in an era of massively abundant data, don't have to settle for wrong models. **Indeed, they don't have to settle for models at all.** Google's founding philosophy is that we don't know why this page is better than that one: If the statistics of incoming links say it is, that's good enough (The End of Theory 2008)*
 - May have a point, but need to consider wider view



Rolls-Royce

Advantages of Data Mining Methods

- Data driven
- Learn relationships from the data
- Do not require explicit models
- Succeed where theoretical models are unknown
- Even when a theoretical model exists they can offer advantages in computational speed or reduced complexity



Rolls-Royce

Disadvantage of Data Mining Methods

- Data driven (need lots of it)
- Learn relationships from the data
 - In many applications a good data analyst needs to understand the data -
- May offer its structure and origin/context
 - functional form may be obscure, or no evidence in sampled data

Hence, there needs to be a balance between the use of these data driven methods and appropriate domain knowledge, particularly when engineering judgments will be made on the results of our models.



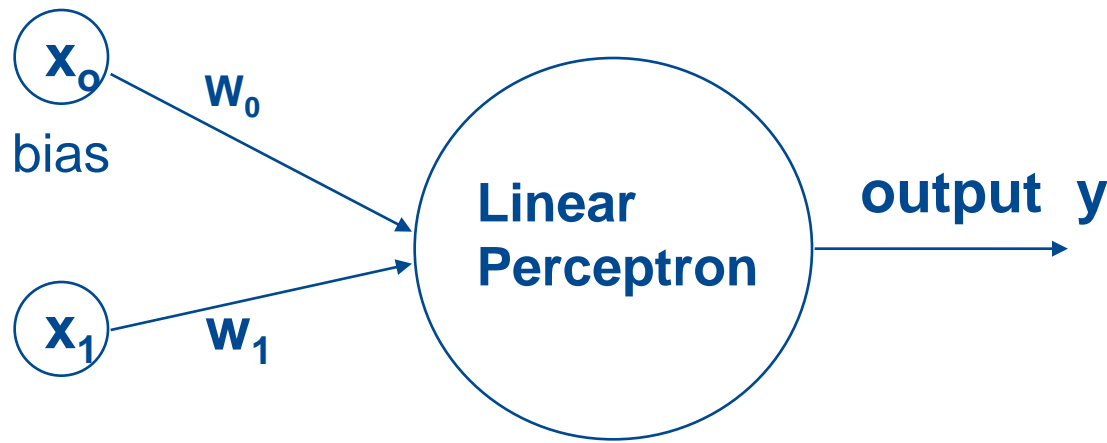
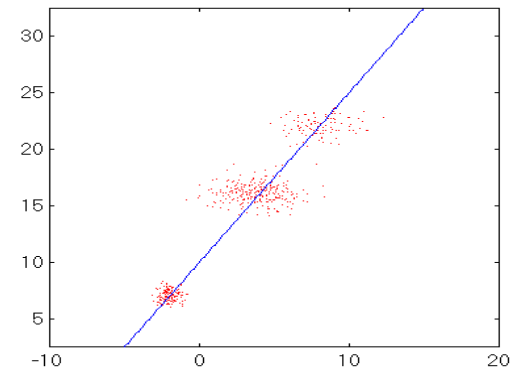
Rolls-Royce

So what's special about these black-box methods – Neural networks ?

Consider a simple problem of curve fitting.

**We wish to approximate some function $f(x)$ by a straight line
such that $y = w_0 + w_1x$**

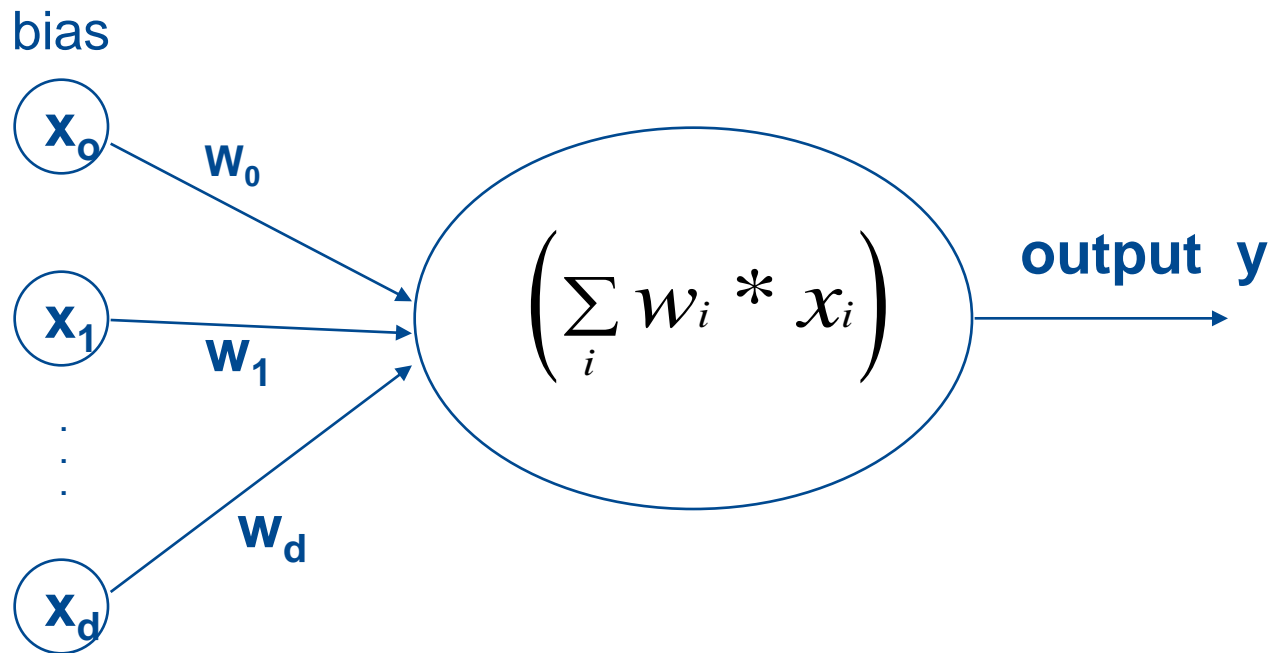
**This is solved by minimising a cost
function $E = (f(x) - y)^2$
with respect to the w_s**



Rolls-Royce

Neural networks

Real problems often deal with multi parameters (multivariate regression) and so the linear perceptron model takes the more general form for the case of d input parameters:



Rolls-Royce

Neural networks

In some cases a linear model may not be adequate for the accuracy required and it may be necessary to consider higher order polynomials.

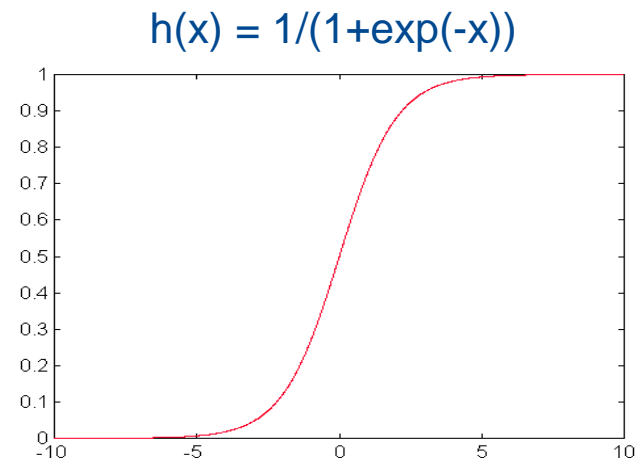
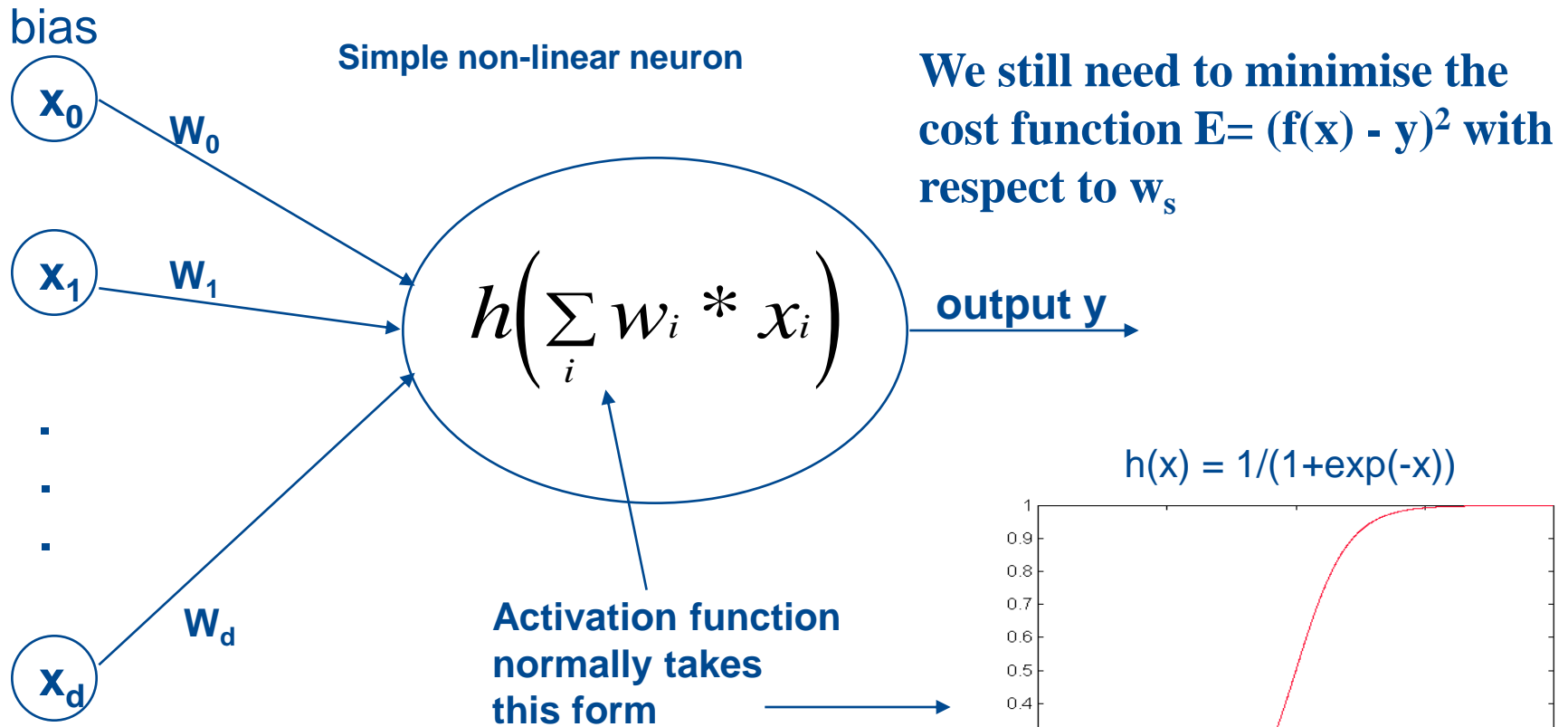
However, the representation of high dimensional problems with high order polynomials, of degree m , generates a model with a large number of ‘free parameters’ (i.e. the w_s) which can grow at the rate d^m - requires lots of data to train.

This problem can be overcome by using linear combinations of non-linear functions of the type shown in the next slide.



Rolls-Royce

Neural networks



Key aspect of this function is it can be differentiated, key to enable training of model

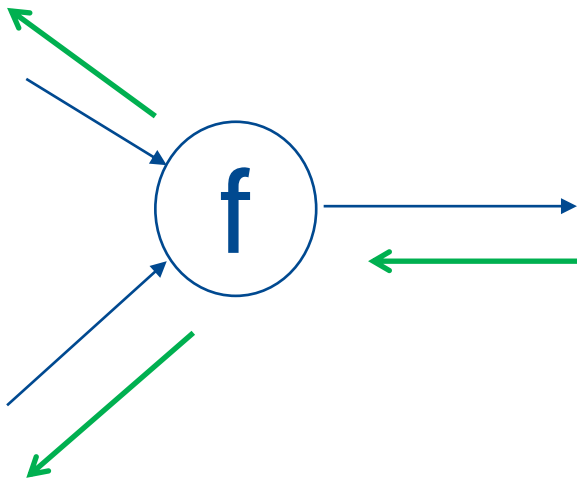


Rolls-Royce

Neural networks

A node/neuron can be **anything** as long as it's differentiable (i.e. you can input value and get output and gradient)

It can be addition, a wavelet transform, $\exp(\sin(\log(x^2)))$, Gaussian kernel smoothing - anything that is differentiable!

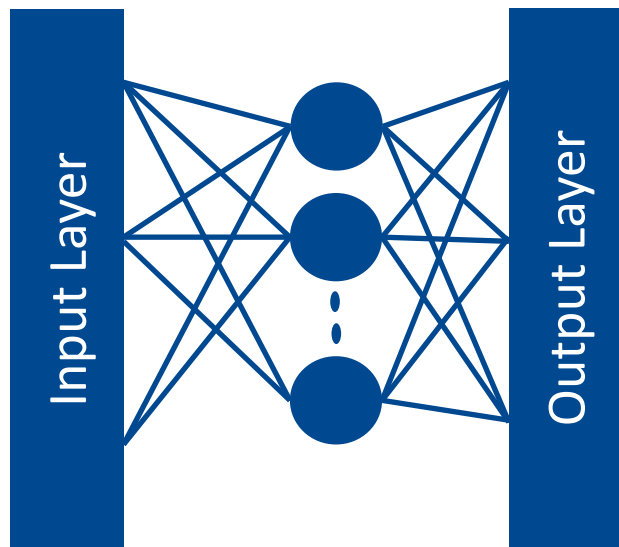


- Having an activation function that's differentiable means it's easy to propagate errors back through the network during training and hence deciding on the level of adjustments required on each input weight.
- Also means we can construct different connected network structures to perform different operations – see next slide

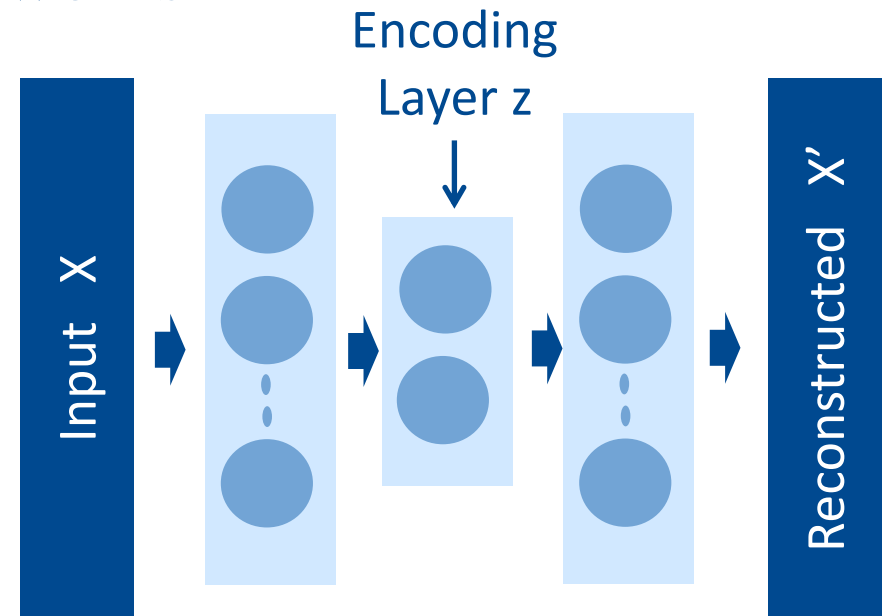


Rolls-Royce

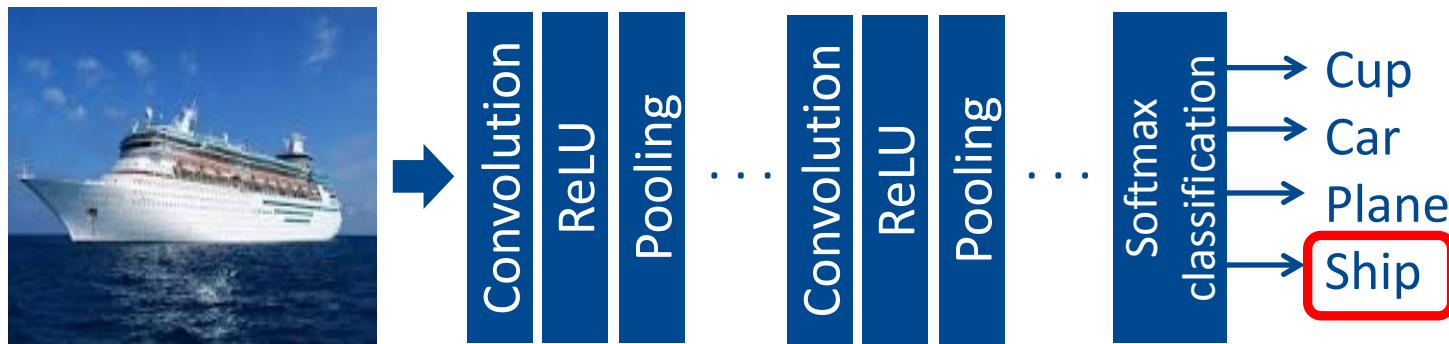
Neural networks



Three-layer network



Autoencoder network



Deep Learning network for image Classification



Rolls-Royce

But, beware of some deep learning solutions:



+



=



Panda (57.7% conf)

Nematode (8.2% conf)

Gibbon (99.3% conf)

Source: Goodfellow, Shlens & Szegedy 2015)



Rolls-Royce

Another example



Model interpretability **to understand your data**: husky vs wolf



(a) Husky classified as wolf



(b) Explanation

Ribeiro et al. 2016, LIME: Why should I trust you? Explaining the predictions of any classifier



Rolls-Royce

Role of Data Mining Techniques

common techniques:

- Data driven modelling/machine learning (e.g. regression)
 - Use of neural networks for regression (prediction error can be used as a measure of novelty)
 - Rule extraction
- Pattern recognition
 - Visualisation of complex data using dimensional reduction techniques to help understand data structure(PCA, Sammon mapping, etc)
 - Cluster analysis as a mechanism to model data and potentially as a method of classification (k-means clustering)
- Knowledge representation (Case-base reasoning)



Rolls-Royce

Data Mining (Information Engineering)

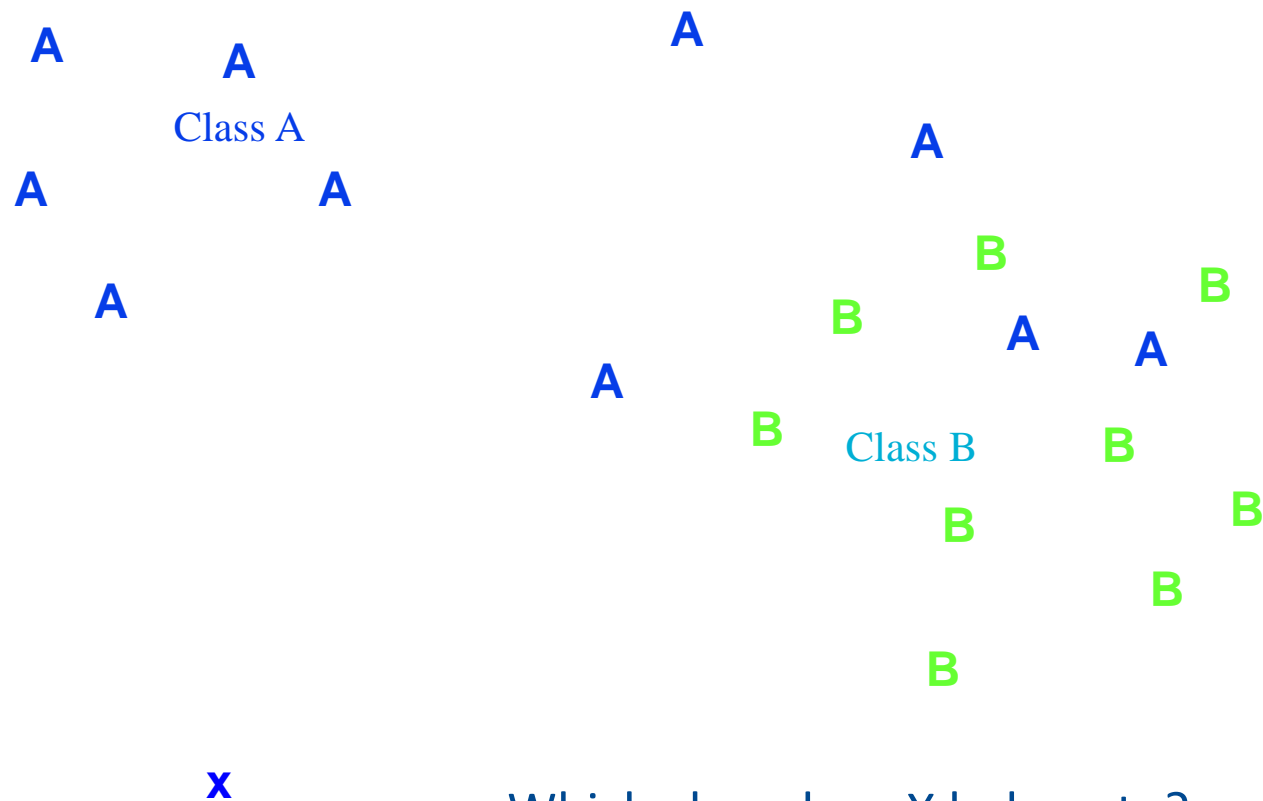
Visualisation of high-dimensional data

- Principal Component Analysis provides a mechanism to identify key directions of variance in data and select the optimum set of vectors that preserves the data variability when data is plotted wrt to these vectors
- Other dimensional reduction techniques rely on preserving geometric structure of the data.



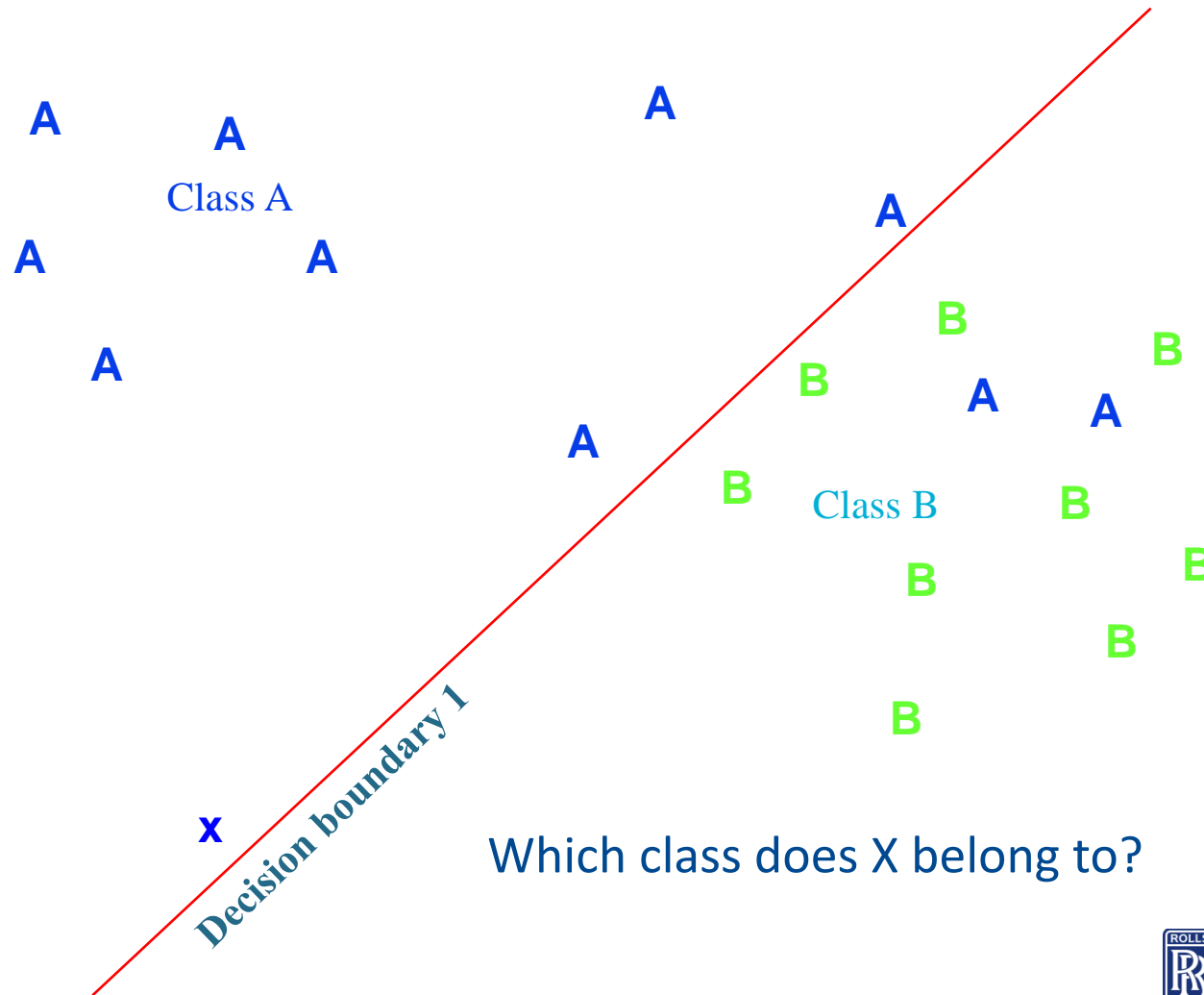
Rolls-Royce

Why visualise the data?

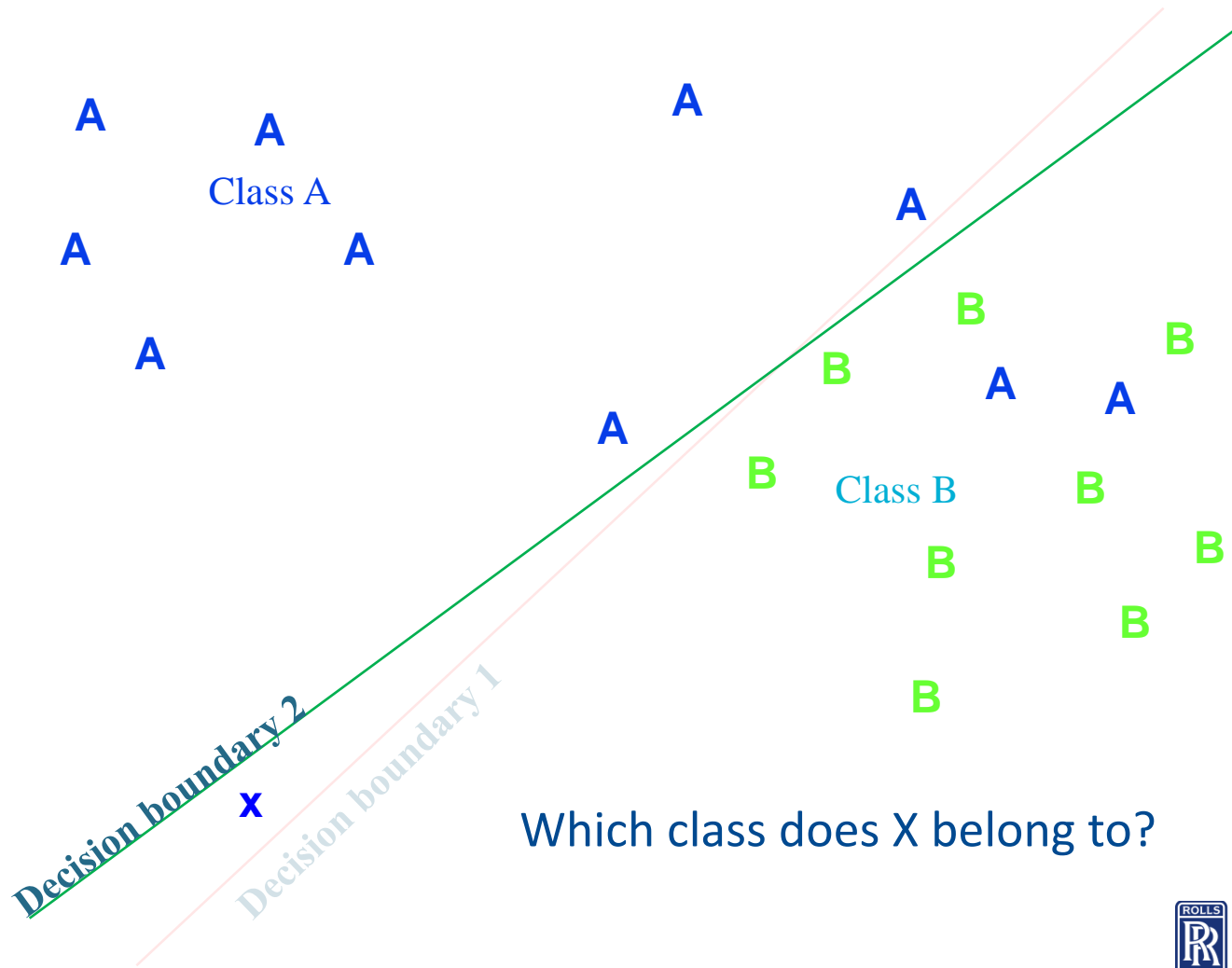


Rolls-Royce

Why visualise the data?

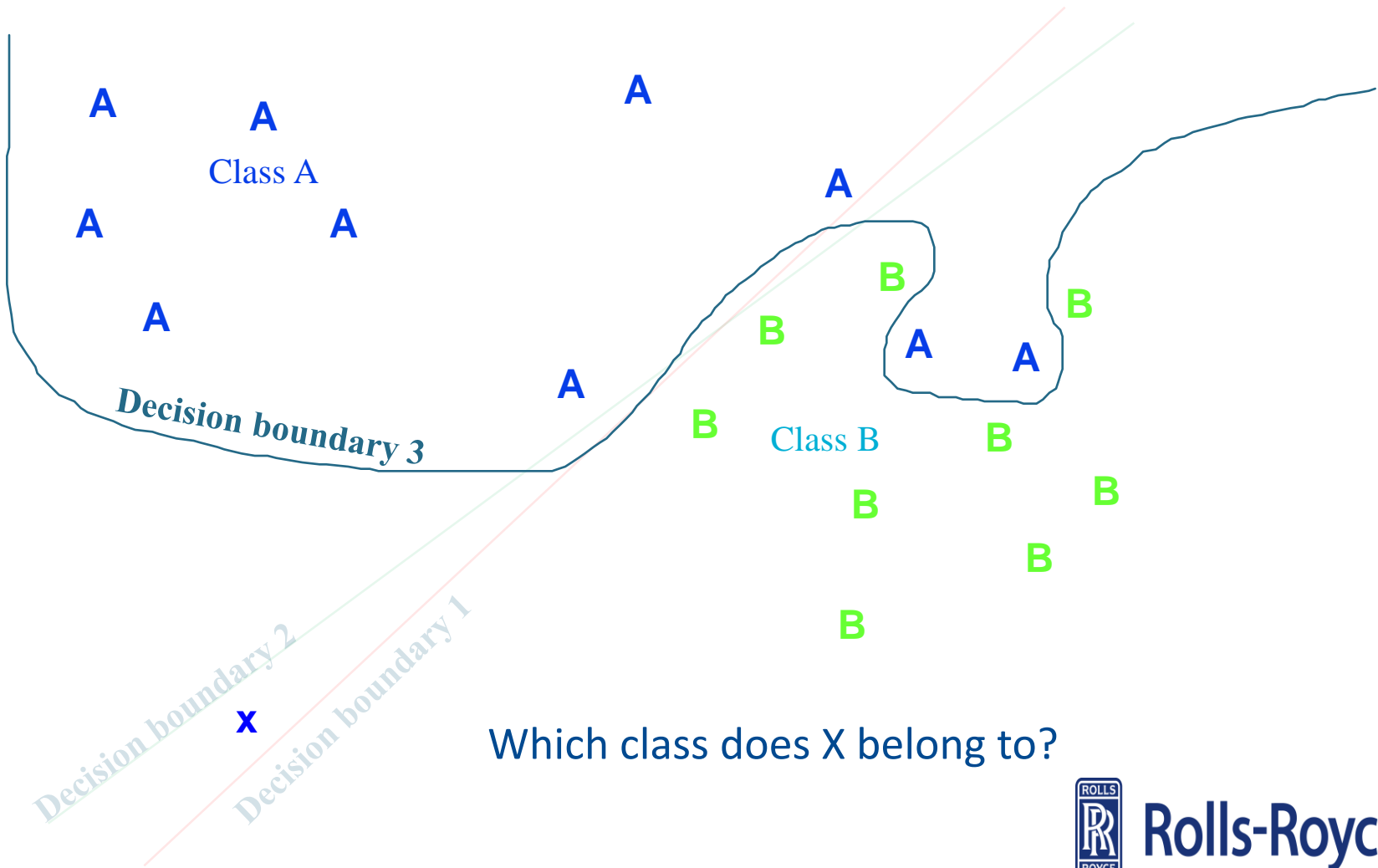


Why visualise the data?



Rolls-Royce

Why visualise the data?

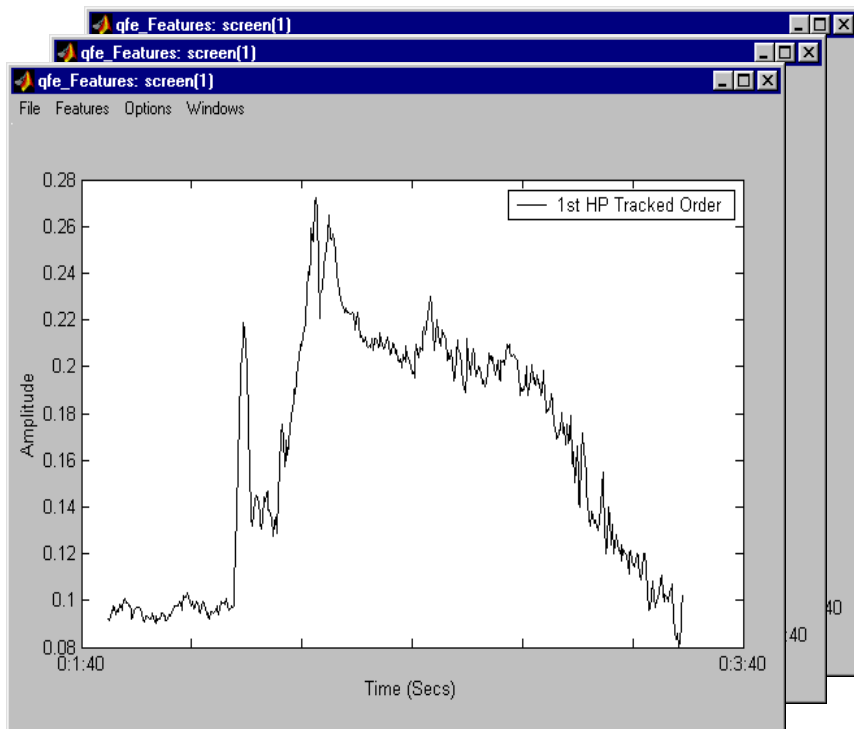


Which class does X belong to?

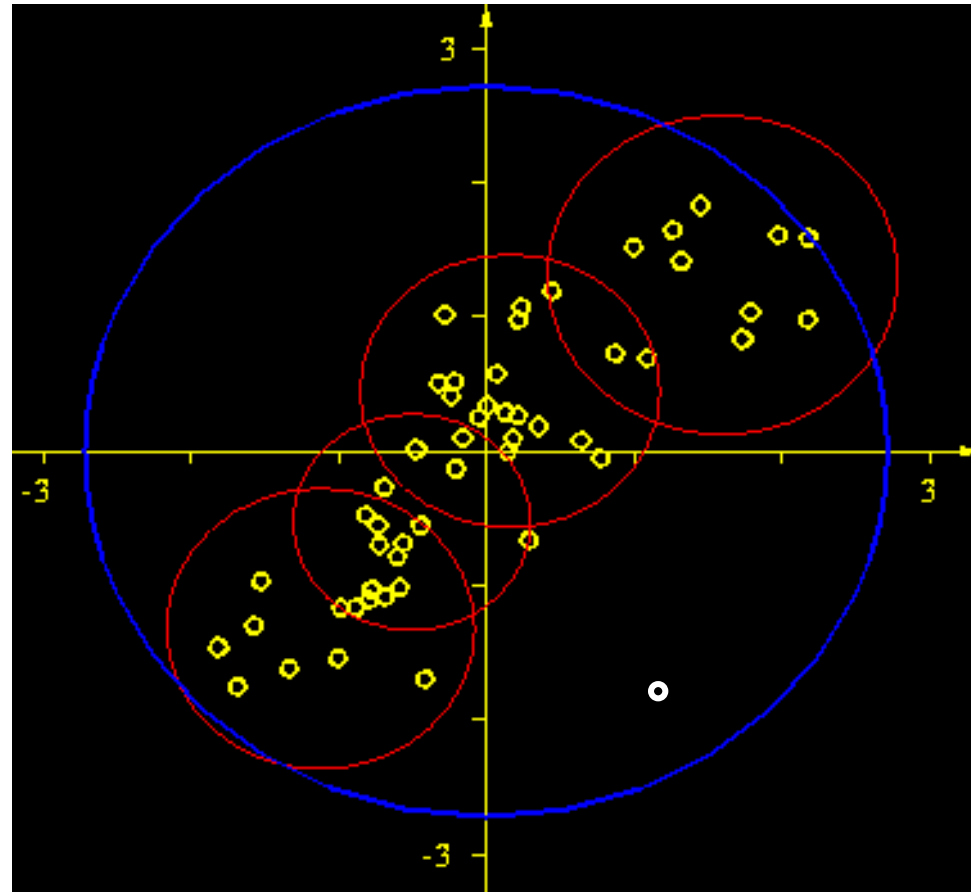


Rolls-Royce

Why Visualise data? - Novelty detection



- Each curve represents the vibration profile of an engine accel (~500 points) during a pass-off test
- We would like to construct a model of normality using clustering. But how many clusters are required?



Rolls-Royce

Case Studies:

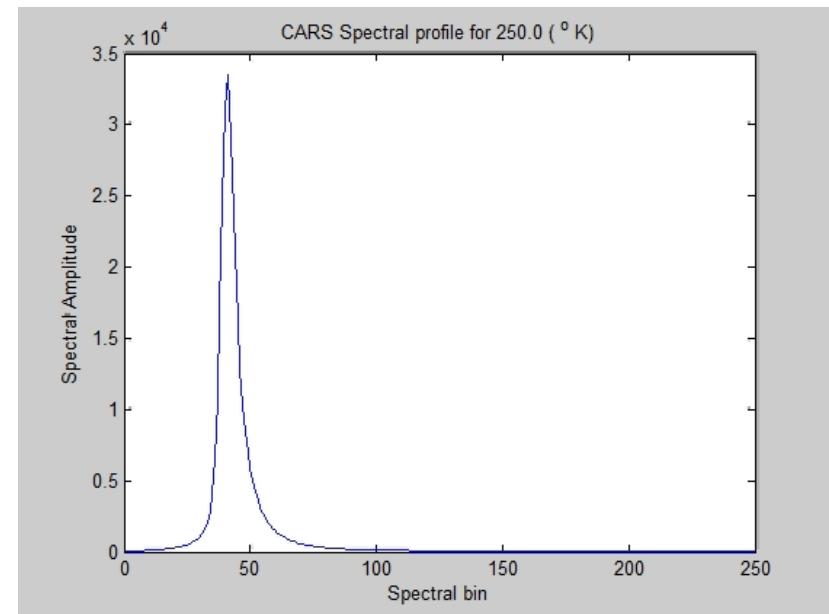
- Estimation of temperature from CARS spectral data (Regression)
- Deriving Rules from Manufacturing data (Rule Induction)
- Skip



Rolls-Royce

Case study – Estimation of temperature from CARS spectral data

- CARS is a non-intrusive thermometry technique used for temperature measurements in combustion environments.
- Data collected from thermal readings each represented by 250 spectral lines (56 examples).
- Each spectrum corresponds to a given temperature value in the range 0-3000 °K.
- Objective - derive a functional model that maps spectral content to a predicted temperature.
 - Need significant reduction of input dimension to derive compact model with available training data due to limited available training data.



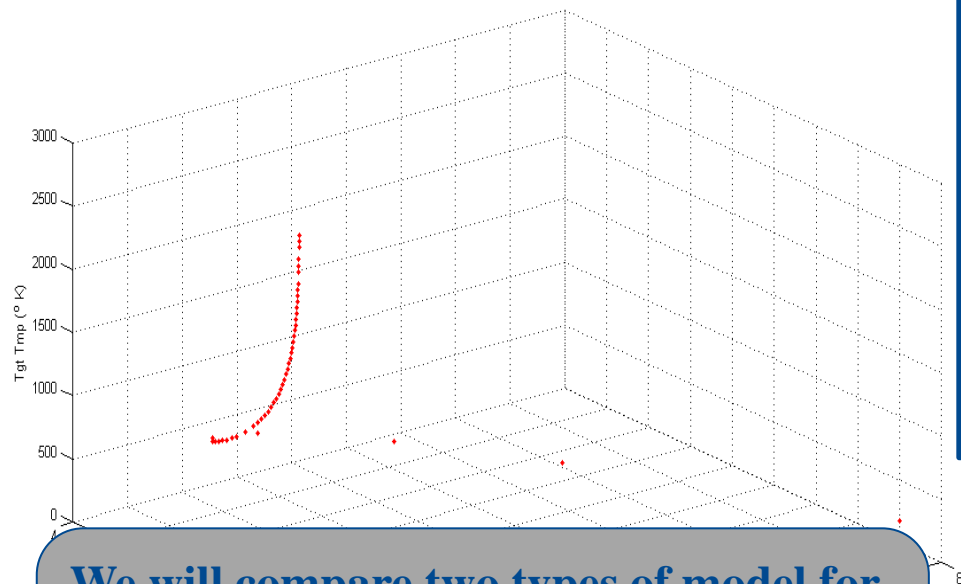
Rolls-Royce

Estimation of temperature from spectral data

Input: Temperature Spectrum
(250D vector)

Dimensional
Reduction
(~2D vector)

Functional
Mapping
To provide Temperature
Estimate



PCA Scores:

PC	variability retained
1	0.9917
2	0.9992
3	0.9997
4	1.0000
5	1.0000
.	.
.	.
.	.
250	1.0

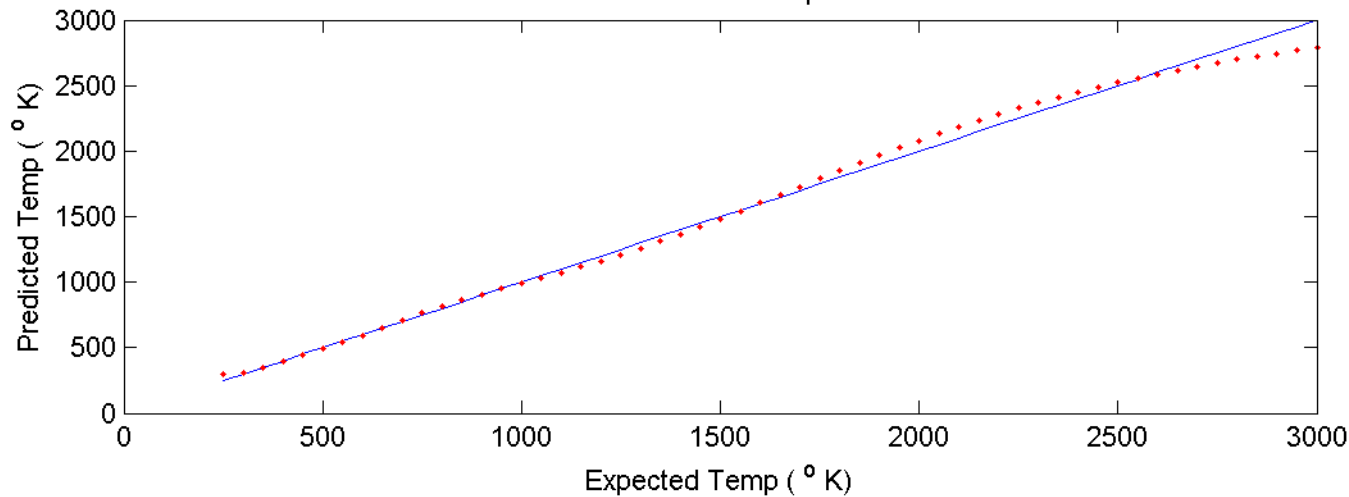
We will compare two types of model for this problem:
A three layer MLP, &
Radial basis function network (RBF)



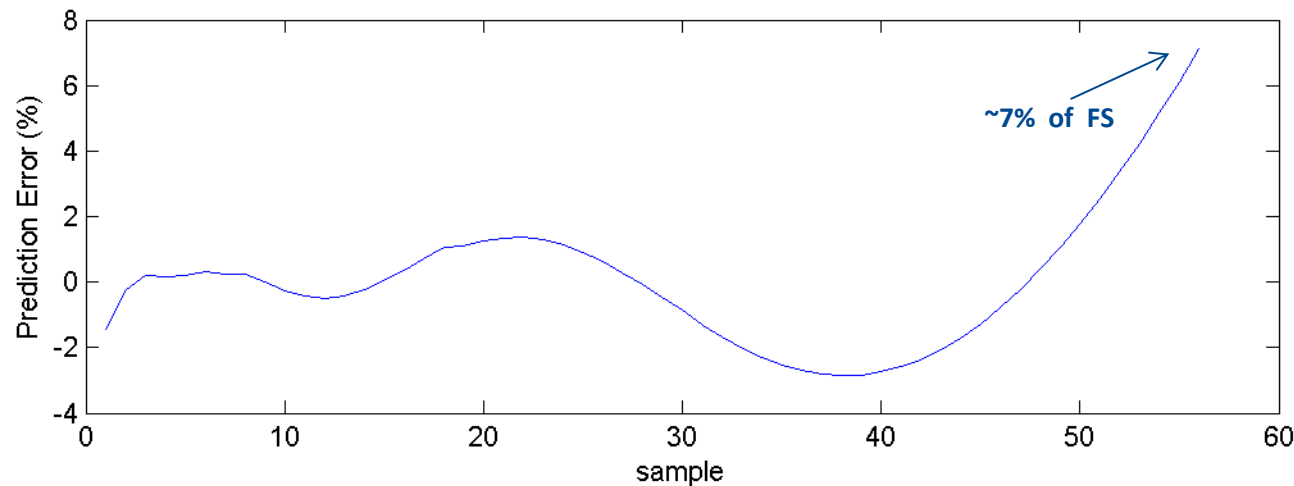
Rolls-Royce

Estimation of temperature from spectral data

MLP Estimate of Temperature



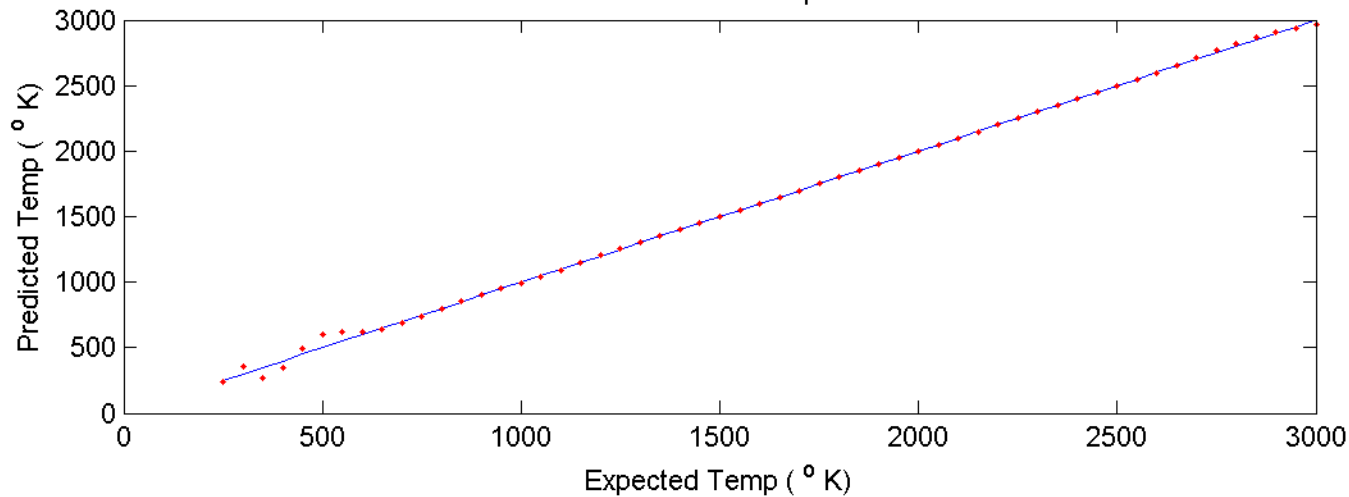
Model based on a three-layer MLP with two inputs, 3 hidden nodes, and one target output.



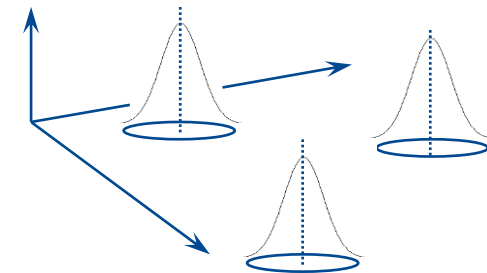
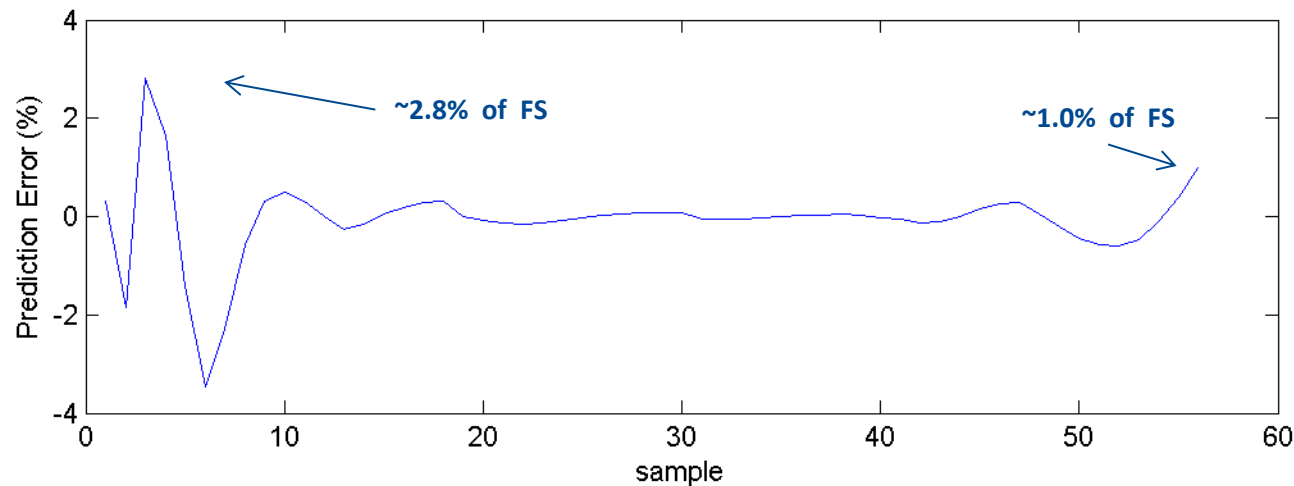
Rolls-Royce

Estimation of temperature from spectral data

RBF Estimate of Temperature

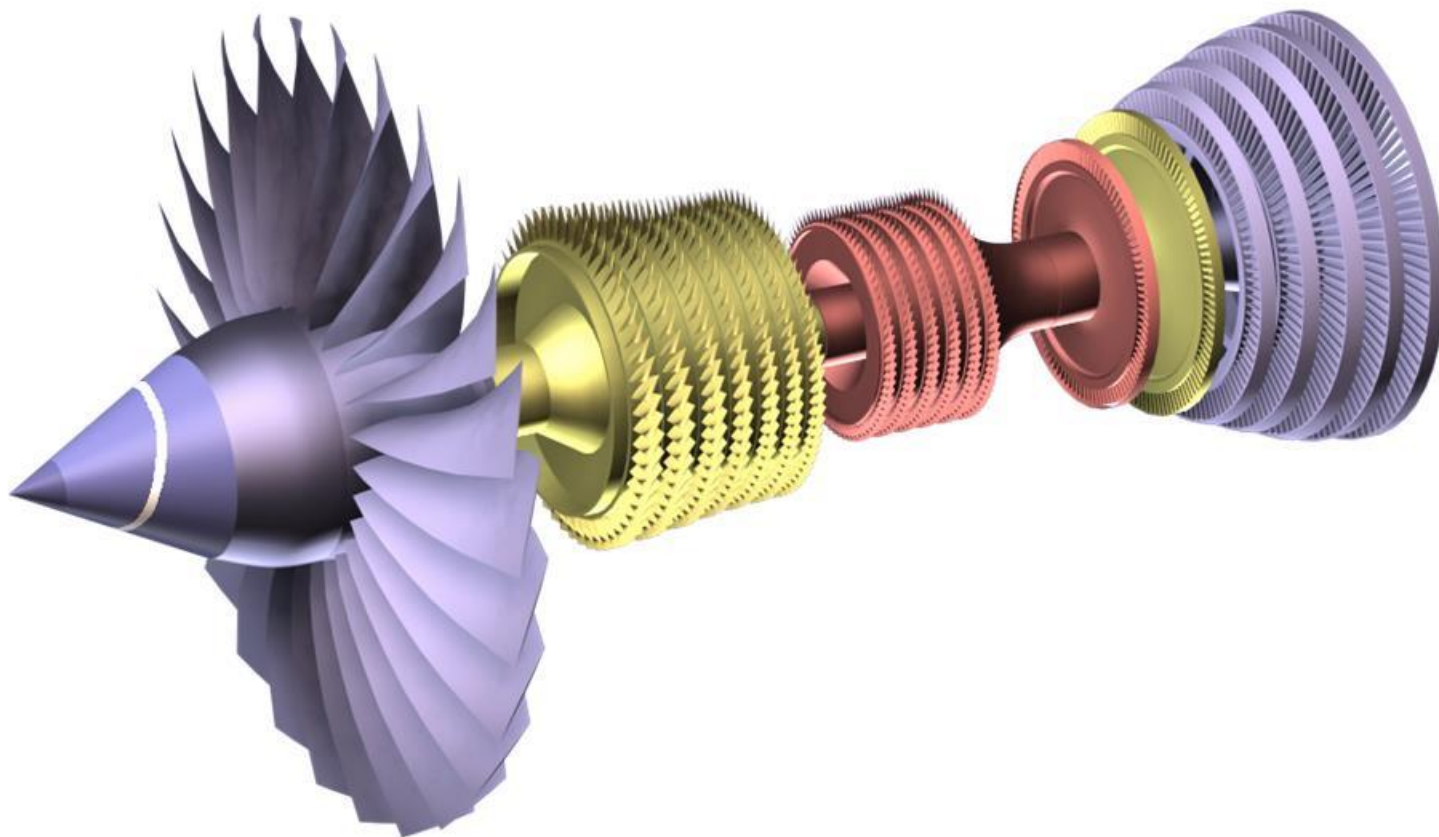


Model based on a Radial basis network (ie mixture of Gaussians in 2D space) using 10 centres.

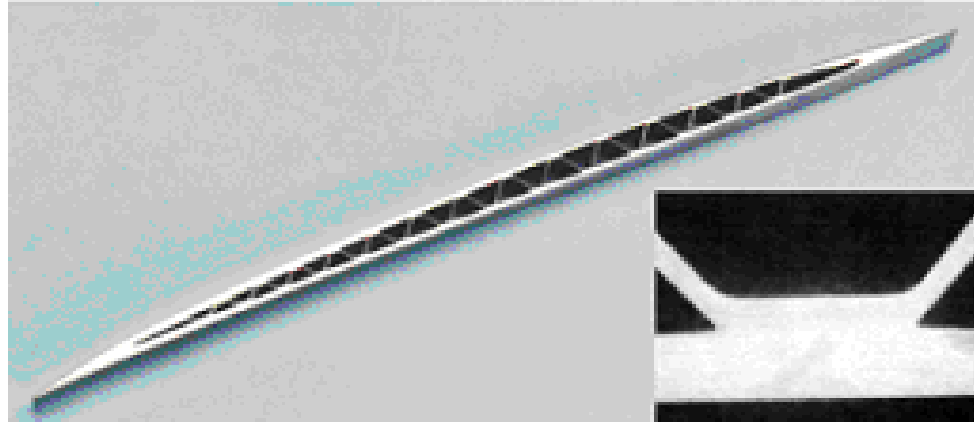
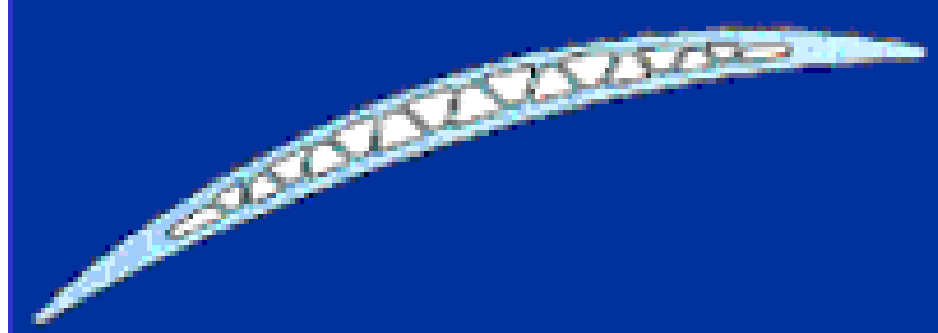


Rolls-Royce

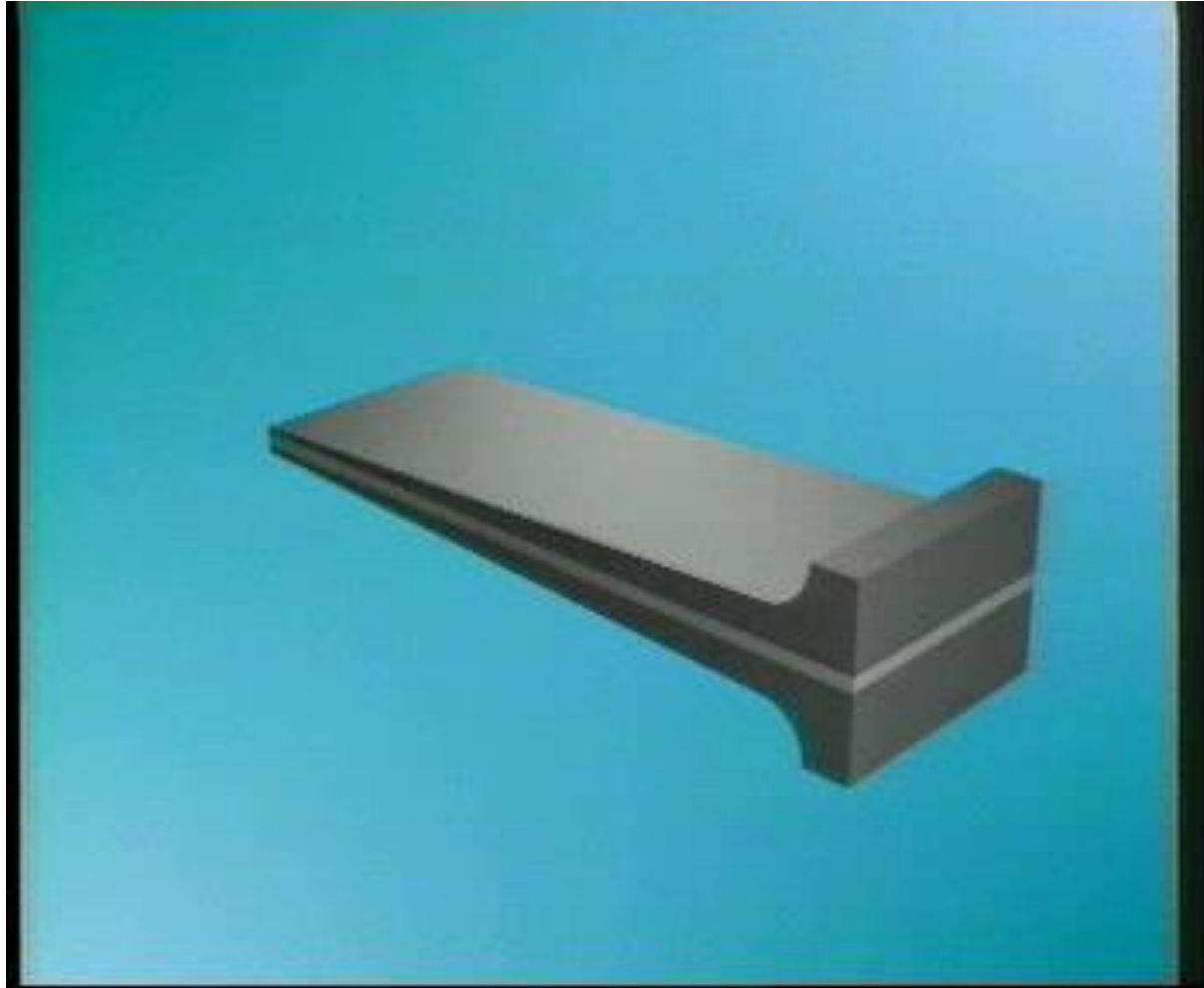
Case Study – Extracting Rules from Manufacturing Data



Extracting Rules from Manufacturing Data



Extracting Rules from Manufacturing Data

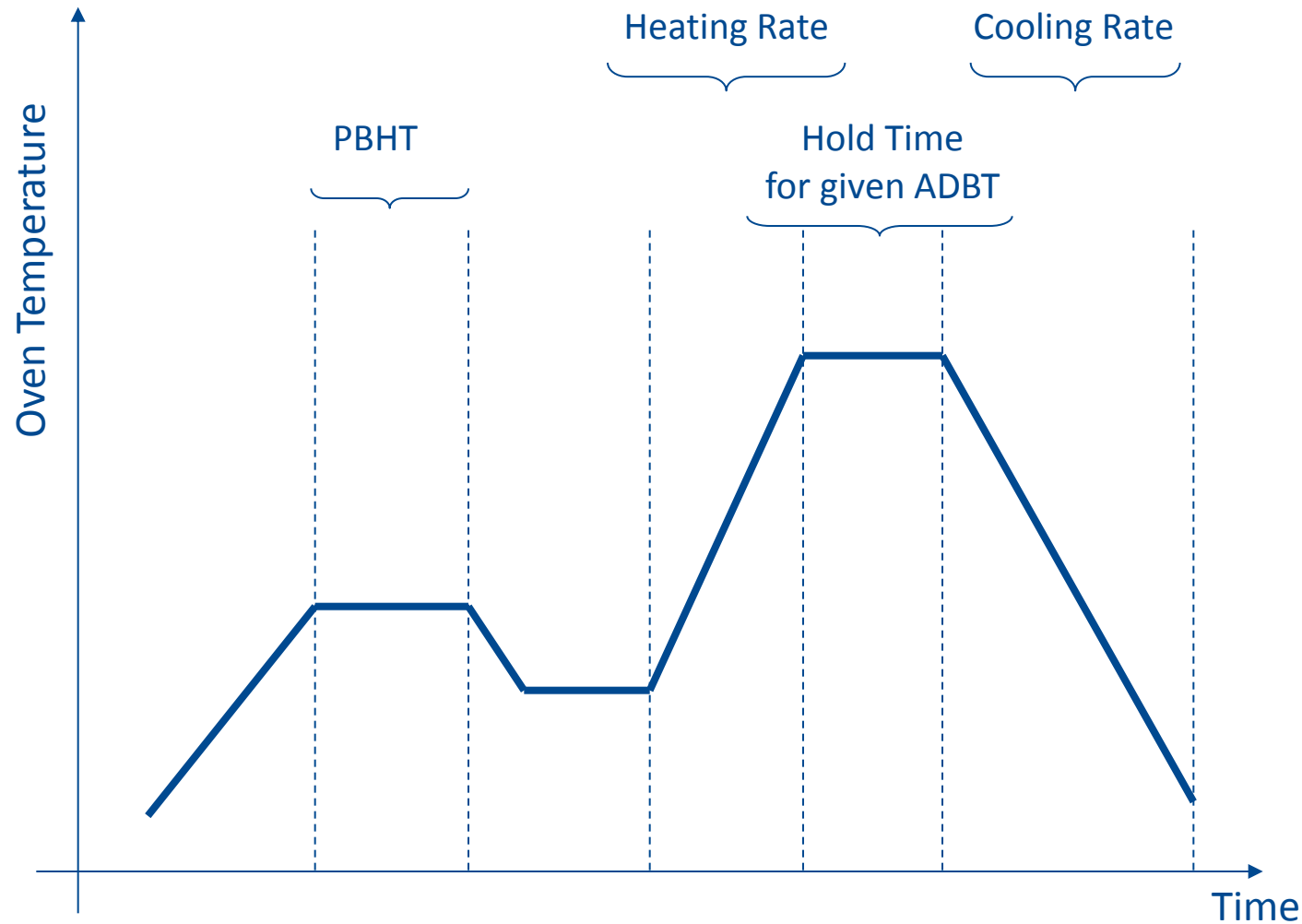


Extracting Rules from Manufacturing Data

- Modified surface layer (MSL) is a thin layer (around 10 microns) formed at the blade's surface where it comes into contact with the honeycomb structure during the manufacturing process.
- Given the random nature of its micro-structure, its existence is very beneficial in preventing crack growth propagation during operation.
- The presence of MSL can only be identified by cutting open a blade and performing visual inspection. Also the extent of its coverage is influenced by the manufacturing process (e.g. variation in stages of heat treatment).

Extracting Rules from Manufacturing Data

Heat Treatment Process of Honeycomb Fan-blade



Extracting Rules from Manufacturing Data

Experiment No	PBHT (mins)	Heating Rate	Hold Time (mins)	ABD Temp (°C)	Cooling Rate	MSL Coverage
1	60	Slow	60	960	Slow	Isolated MSL
2	None	Fast	45	940	Fast	Consistent MSL
3	60	Fast	45	940	Fast	Consistent MSL
4	30	Fast	45	940	Fast	Consistent MSL
5	30	Fast	60	960	Fast	Consistent MSL
6	30	Slow	60	960	Fast	Consistent MSL
7	30	Slow	45	960	Fast	Consistent MSL
8	30	Slow	60	940	Fast	Intermittent MSL
9	30	Fast	45	960	Slow	NONE
10	30	Fast	60	940	Slow	NONE
11	60	Typical	60	960	Typical	NONE
12	None	Slow	45	940	Fast	Intermittent MSL
13	60	Slow	45	940	Slow	Isolated MSL
14	None	Fast	45	940	Slow	NONE
15	None	Slow	60	940	Slow	Isolated MSL
16	60	Slow	60	940	Fast	Consistent MSL
17	None	Fast	60	940	Fast	Intermittent MSL
18	60	Fast	60	940	Slow	Isolated MSL
19	None	Slow	45	960	Slow	Isolated MSL
20	60	Slow	45	960	Fast	Consistent MSL
21	None	Fast	45	960	Fast	Consistent MSL
22	60	Fast	45	960	Slow	Isolated MSL
23	None	Slow	60	960	Fast	Consistent MSL
24	None	Fast	60	960	Slow	Isolated MSL
25	60	Fast	60	960	Fast	Intermittent MSL
26	30	Fast	45	960	Fast	Intermittent MSL
27	None	Fast	90	960	Fast	Consistent MSL
28	30	Slow	60	975	Fast	NONE



Rolls-Royce

Extracting Rules from Manufacturing Data

Experiment No	PBHT (mins)	Heating Rate	Hold Time (mins)	ABD Temp (°C)	Cooling Rate	MSL Coverage
1	60	Slow	60	960	Slow	Isolated MSL
2	None	Fast	45	940	Fast	Consistent MSL
3	60	Fast	45	940	Fast	Consistent MSL
4	30	Fast	45	940	Fast	Consistent MSL
5	30	Fast	60	960	Fast	Consistent MSL
6	30	Slow	60	960	Fast	Consistent MSL
7	30	Slow	45	960	Fast	Consistent MSL
8	30	Slow	60	940	Fast	Intermittent MSL
9	30	Fast	45	960	Slow	NONE
10	30	Fast	60	940	Slow	NONE
11	60	Typical	60	960	Typical	NONE
12	None	Slow	45	940	Fast	Intermittent MSL
13	60	Slow	45	940	Slow	Isolated MSL
14	None	Fast	45	940	Slow	NONE
15	None	Slow	60	940	Slow	Isolated MSL
16	60	Slow	60	940	Fast	Consistent MSL
17	None	Fast	60	940	Fast	Intermittent MSL
18	60	Fast	60	940	Slow	Isolated MSL
19	None	Slow	45	960	Slow	Isolated MSL
20	60	Slow	45	960	Fast	Consistent MSL
21	None	Fast	45	960	Fast	Consistent MSL
22	60	Fast	45	960	Slow	Isolated MSL
23	None	Slow	60	960	Fast	Consistent MSL
24	None	Fast	60	960	Slow	Isolated MSL
25	60	Fast	60	960	Fast	Intermittent MSL
26	30	Fast	45	960	Fast	Intermittent MSL
27	None	Fast	90	960	Fast	Consistent MSL
28	30	Slow	60	975	Fast	NONE

Can we extract a set of rules from this data that help determine what parts of the process can be controlled to ensure MSL coverage is ‘Consistent‘?

Construct decision tree by determining which combination of attributes provides best measure of classification. This is achieved by a measure of an attribute’s Information Gain.

In order to calculate this value we also need a measure of Entropy which characterises the level of impurity in the data (or measure of homogeneity of examples).

Extracting Rules from Data

- Entropy can be calculated as follows:

$$Entropy(S) = \sum_{i=1}^{numberclasses} -p_i \log_2(p_i)$$

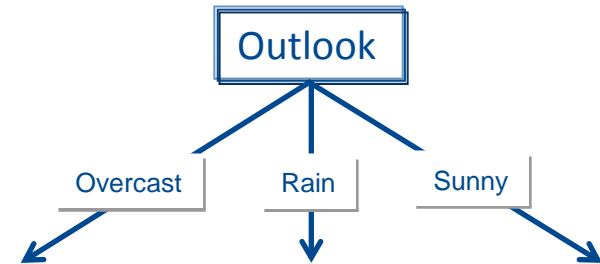
- Information Gain is the expected reduction in entropy caused by partitioning the data according to a selected attribute.

$$Gain(S, A) = Entropy(S) - \sum_{v \in values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

Extracting Rules from Data

Simple worked example

Outlook	Temperature	Humidity	Wind	Play Tennis
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rain	Mild	High	Weak	Yes
Rain	Cool	Normal	Weak	Yes
Rain	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rain	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rain	Mild	High	Strong	No



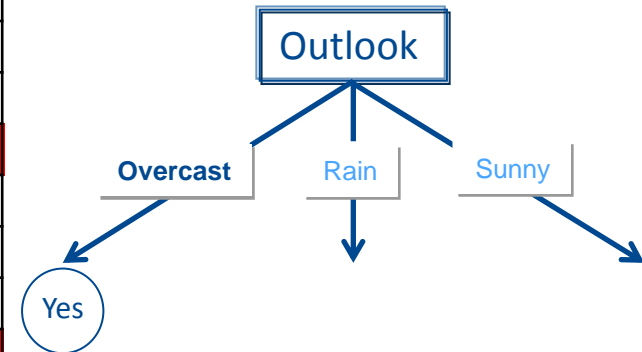
0.2467	0.0292	0.1518	0.0481
--------	--------	--------	--------

Information Gain

Extracting Rules from Data

Simple worked example

Outlook	Temperature	Humidity	Wind	Play Tennis
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rain	Mild	High	Weak	Yes
Rain	Cool	Normal	Weak	Yes
Rain	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rain	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rain	Mild	High	Strong	No



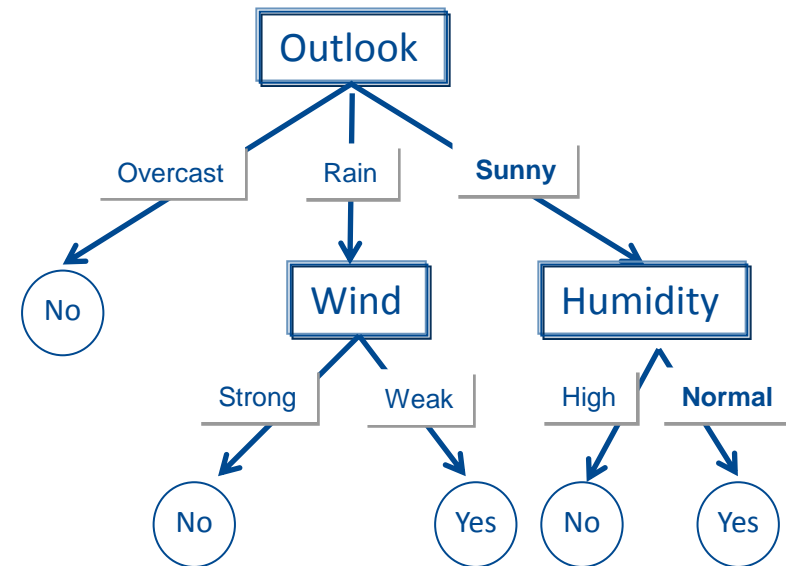
0	0	0	0
---	---	---	---

Information Gain

Extracting Rules from Data

Simple worked example

Outlook	Temperature	Humidity	Wind	Play Tennis
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rain	Mild	High	Weak	Yes
Rain	Cool	Normal	Weak	Yes
Rain	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rain	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rain	Mild	High	Strong	No



Extracting Rules from Manufacturing Data

Experiment No	PBHT (mins)	Heating Rate	Hold Time (mins)	ABD Temp (°C)	Cooling Rate	MSL Coverage
1	60	Slow	60	960	Slow	Isolated MSL
2	None	Fast	45	940	Fast	Consistent MSL
3	60	Fast	45	940	Fast	Consistent MSL
4	30	Fast	45	940	Fast	Consistent MSL
5	30	Fast	60	960	Fast	Consistent MSL
6	30	Slow	60	960	Fast	Consistent MSL
7	30	Slow	45	960	Fast	Consistent MSL
8	30	Slow	60	940	Fast	Intermittent MSL
9	30	Fast	45	960	Slow	NONE
10	30	Fast	60	940	Slow	NONE
11	60	Typical	60	960	Typical	NONE
12	None	Slow	45	940	Fast	Intermittent MSL
13	60	Slow	45	940	Slow	Isolated MSL
14	None	Fast	45	940	Slow	NONE
15	None	Slow	60	940	Slow	Isolated MSL
16	60	Slow	60	940	Fast	Consistent MSL
17	None	Fast	60	940	Fast	Intermittent MSL
18	60	Fast	60	940	Slow	Isolated MSL
19	None	Slow	45	960	Slow	Isolated MSL
20	60	Slow	45	960	Fast	Consistent MSL
21	None	Fast	45	960	Fast	Consistent MSL
22	60	Fast	45	960	Slow	Isolated MSL
23	None	Slow	60	960	Fast	Consistent MSL
24	None	Fast	60	960	Slow	Isolated MSL
25	60	Fast	60	960	Fast	Intermittent MSL
26	30	Fast	45	960	Fast	Intermittent MSL
27	None	Fast	90	960	Fast	Consistent MSL
28	30	Slow	60	975	Fast	NONE



Rolls-Royce

Extracting Rules from Manufacturing Data

Weka Explorer | MathUtils | VMware Remote Console | Devices | Select attributes | Visualize

Classifier: Choose **J48 -C 0.5 -M 2**

Test options:

- ☒ Use training set
- ☐ Supplied test set (Set...)
- ☐ Cross-validation (Folds: 10)
- ☐ Percentage split (%: 66)
- More options...

(Nom) mslcoverage

Start Stop

Result list (right-click for options):

- 10:04:52 - trees.J48

Classifier output:

```

coolingrate = slow
| heatingrate = slow: isolated (4.0)
| heatingrate = fast
| | pbht <= 30: none (4.0/1.0)
| | pbht > 30: isolated (2.0)
| heatingrate = typical: isolated (0.0)
coolingrate = fast
| activationbondtemp <= 940
| | pbht <= 30
| | | heatingrate = slow: intermittent (2.0)
| | | heatingrate = fast: consistent (3.0/1.0)
| | | heatingrate = typical: intermittent (0.0)
| | pbht > 30: consistent (2.0)
| activationbondtemp > 940: consistent (10.0/3.0)
coolingrate = typical: none (1.0)

Number of Leaves :    10
Size of the tree :    16

Time taken to build model: 0 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances      23           82.1429 %
Incorrectly Classified Instances     5           17.8571 %
Kappa statistic                    0.7417
Mean absolute error                 0.1327
Root mean squared error             0.2576
Relative absolute error             36.7078 %
Root relative squared error         60.7277 %
Total Number of Instances          28

=== Detailed Accuracy By Class ===
  
```

Status: OK

Log x 0

Extracting Rules from Manufacturing Data

J48 pruned tree from Weka utility

```

coolingrate = slow
| heatingrate = slow: isolated (4.0)
| heatingrate = fast
| | pbht <= 30: none (4.0/1.0)
| | pbht > 30: isolated (2.0)
| heatingrate = typical: isolated (0.0)
coolingrate = fast
| activationbondtemp <= 940
| | pbht <= 30
| | | heatingrate = slow: intermittent (2.0)
| | | heatingrate = fast: consistent (3.0/1.0)
| | | heatingrate = typical: intermittent (0.0)
| | pbht > 30: consistent (2.0)
| activationbondtemp > 940: consistent (10.0/3.0)
coolingrate = typical: none (1.0)
  
```

Conclusion:

Ensuring *Cooling rate is fast* and *activation bond temperature > 940° C* is likely to result in consistent MSL.

=== Confusion Matrix ===

a	b	c	d	<-- classified as
6	0	0	1	a = isolated
0	11	0	0	b = consistent
0	3	2	0	c = intermittent
0	1	0	4	d = none

23 correct classifications
5 incorrect classifications

Main points to consider for ML approaches

- Big Data Analytics: examination of large data-sets to reveal hidden patterns, correlations and other insights that may provide useful knowledge – but beware of correlations, they are not the same as causal relationships
- Ensure sufficient understanding of data structure & context
- Be willing to follow up false positives
- Be clear on the question you're trying to ask of the data
 - *An approximate answer to the right question is worth a great deal more than a precise answer to the wrong question*
- Think very hard about the complexity of the model you think you need (over-fitting is your enemy, yet so easy to achieve)
- Feature engineering likely to consume most model development time
 - *Make maximum use of available domain knowledge in selection and aggregation/enrichment of features*



Rolls-Royce