ROYAL STATISTICAL SOCIETY

DATA | EVIDENCE | DECISIONS

Data Science Workshop

Andy Garrett, EVP Global Scientific Operations, ICON

Jim Weatherall, Head Advanced Analytics Centre, AZ

RSS Data Science Section
EFSPI Statistical Leaders' Meeting
4th July 2017

Agenda

- **Introduction – 15 mins**

LUNCH

- Survey analysis – 15 mins
- Case studies – 10 mins
- Group work: four themes – 30 mins
  - The Internet of Things
  - Big Data: EHRs
  - Decision science
  - Automation and artificial intelligence
- Report back – 25 mins
- Discussion – 25 mins

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Big Data Landscape 2016

Data Science Section Remit

To be a professional body that represents data scientists in the UK. The section will organise meetings for a broad range of attendees and generate outputs that are aimed at:

- Promoting good practice by addressing what good Data Science looks like (with exemplars) and what it does not look like.
- Promoting the statistical aspects of Data Science / re-enforcing the statistical framework
- Being a trusted voice on Data Science for employers, including inputting to consultation exercises
- Supporting the Data Science community throughout the UK
- Supporting the pipeline and career development of data scientists and statisticians by elevating skill sets to work in the modern world
- Supporting important emerging topics such as ethics, privacy, algorithmic responsibility and personalization - lifting the quality of the conversation
- Fostering multi-disciplinary connections and the exchanging of ideas

DSS Committee Members

Fran Bennett – Mastodon C

Simon Briscoe (Council representative)

David van Dyk – Imperial / ASA DS Chapter

Andrew Garrett (Chair) - ICON

Martin Goodson – Evolution AI

Mark Girolami – Turing Institute / Imperial

Ioanna Manolopoulou - UCL

Giles Pavey – ex Dunnhumby/Tesco

Harry Powell – Barclays

Richard Pugh (Meetings Secretary) – Mango Solutions

Matthew Upson (Secretary) – Cabinet Office

Leone Wardman - ONS

James Weatherall (Vice Chair) - AZ

ROYAL STATISTICAL SOCIETY

DATA | EVIDENCE | DECISIONS

DSS Launch event

The Industrialisation and Professionalisation of DS (19th June)

- 12 Questions presented, with three formal responses
- An example topic
- President's response
- Q&A

YouTube: https://m.youtube.com/watch?v=5aH3vVvtOfc

DSS Social Media

RSS website: landing page

Twitter: @RSS_DSS

GitHub: https://github.com/rssdatascience

LinkedIn:

https://www.linkedin.com/company-beta/111500048/

Slack: https://rssdatascience.slack.com

Agenda

- Introduction – 15 mins

LUNCH

- Survey analysis – 15 mins
- Case studies – 10 mins
- Group work: four themes - 30 mins
  - **The Internet of Things**
  - **Big Data: EHRs**
  - **Decision science**
  - **Automation and artificial intelligence**
- Report back – 25 mins
- Discussion – 25 mins

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Agenda

- Introduction – 15 mins

LUNCH

- **Survey analysis – 15 mins**
- Case studies – 10 mins
- Group work: four themes - 30 mins
  - The Internet of Things
  - Big Data: EHRs
  - Decision science
  - Automation and artificial intelligence
- Report back – 25 mins
- Discussion – 25 mins

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Personal definitions of data science

ROYAL STATISTICAL SOCIETY

DATA | EVIDENCE | DECISIONS

There are a wide range of perspectives

Gaining Knowledge and Insights from Data

Data Science is an interdisciplinary field of expertise about processes and systems to extract knowledge or insights from data in various forms, either structured or unstructured, in order to address various kinds of technical, scientific and business needs

Data-driven science based on maths, computer science and domain knowledge

Combination of computational and statistical expertise to access and analyse data

Data visualisation, modelling, simulation and AI technologies are applied in Data science

A multidisciplinary field, merging math/stat skills with computer science and

Evidence that we have failed as a statistical discipline

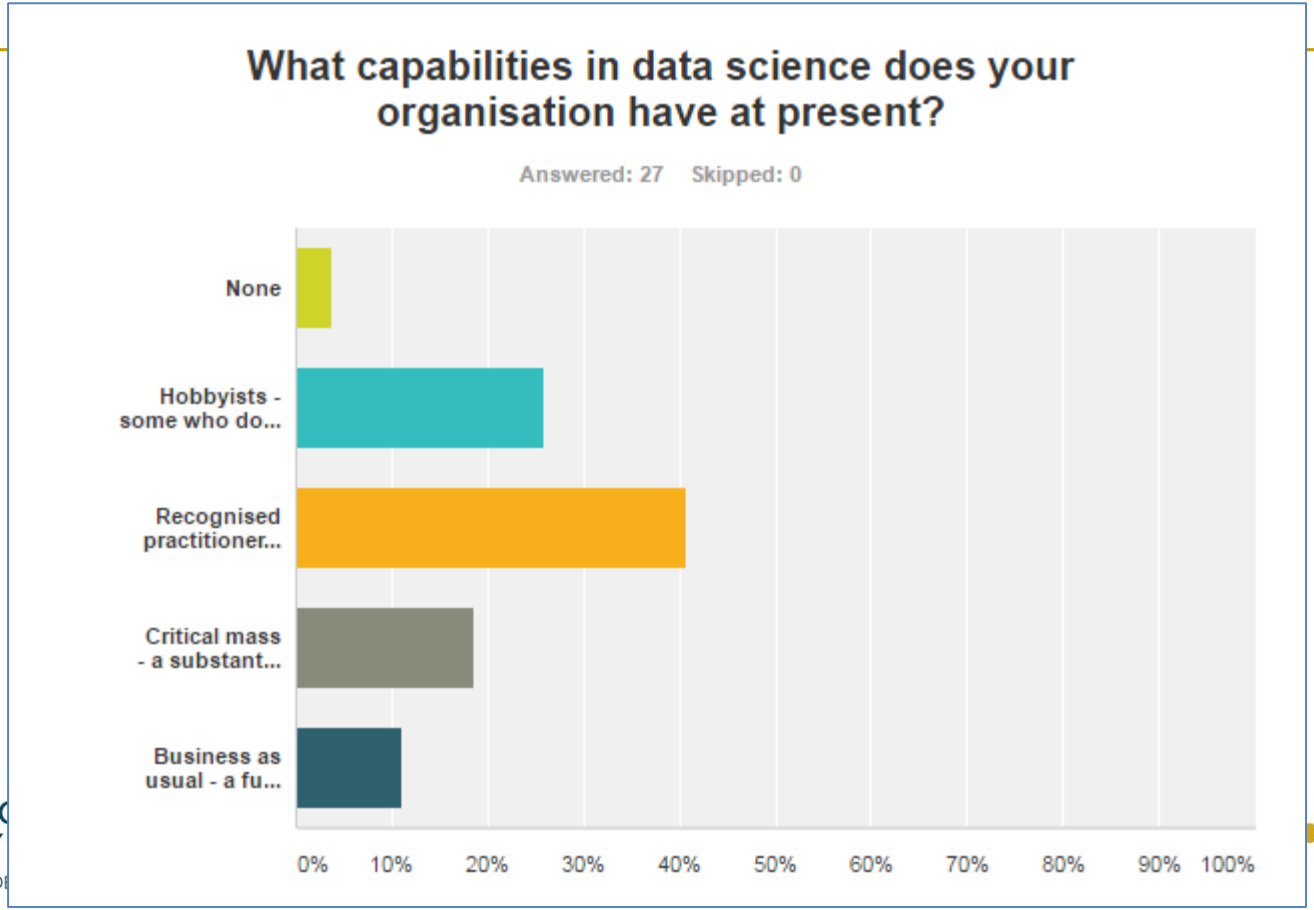
Database setup/programming, CRF design, data management

A blend of statistics, IT and mathematics for big data

Visualisation skills - usually focussed on a specific domain

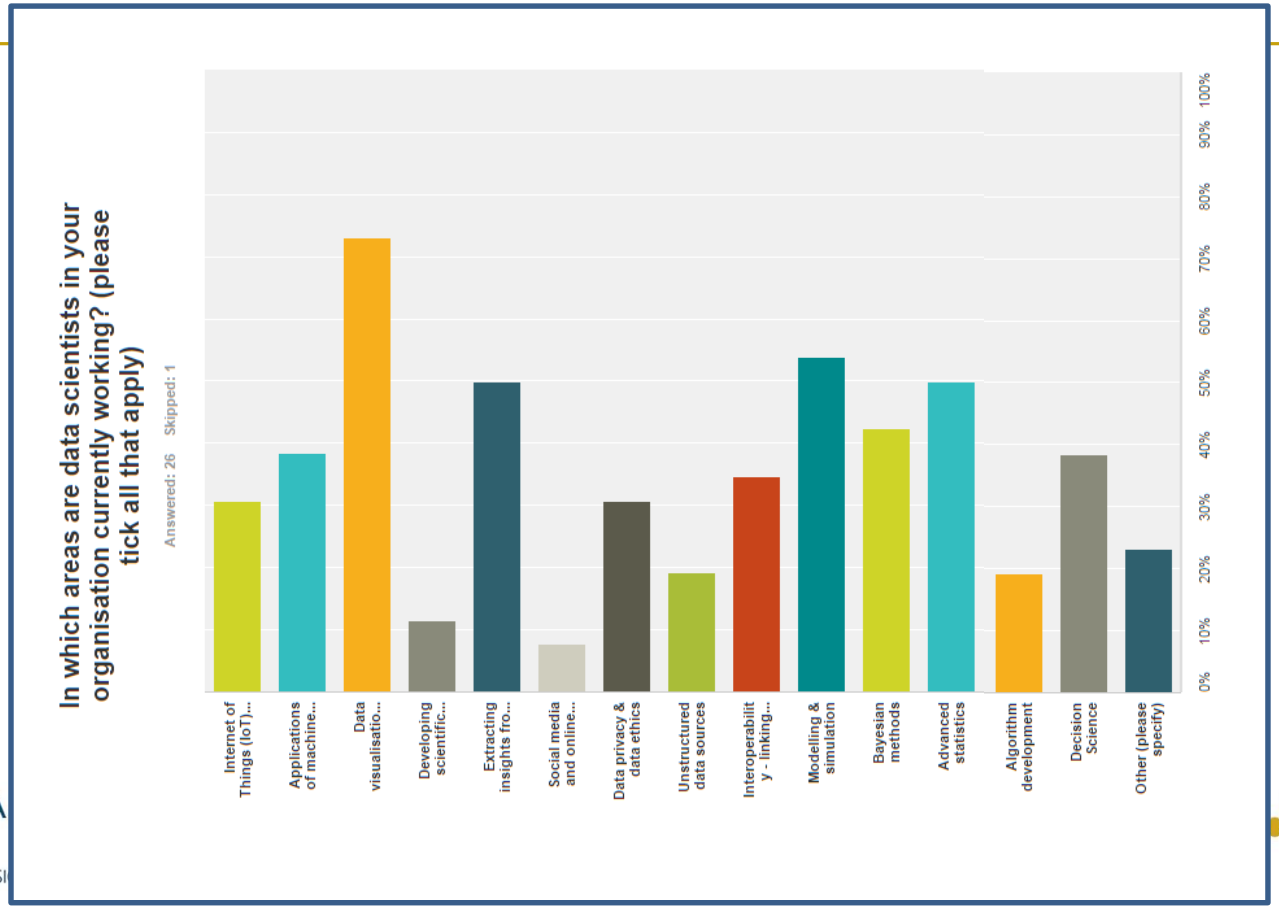
ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Data science is recognised in most organisations



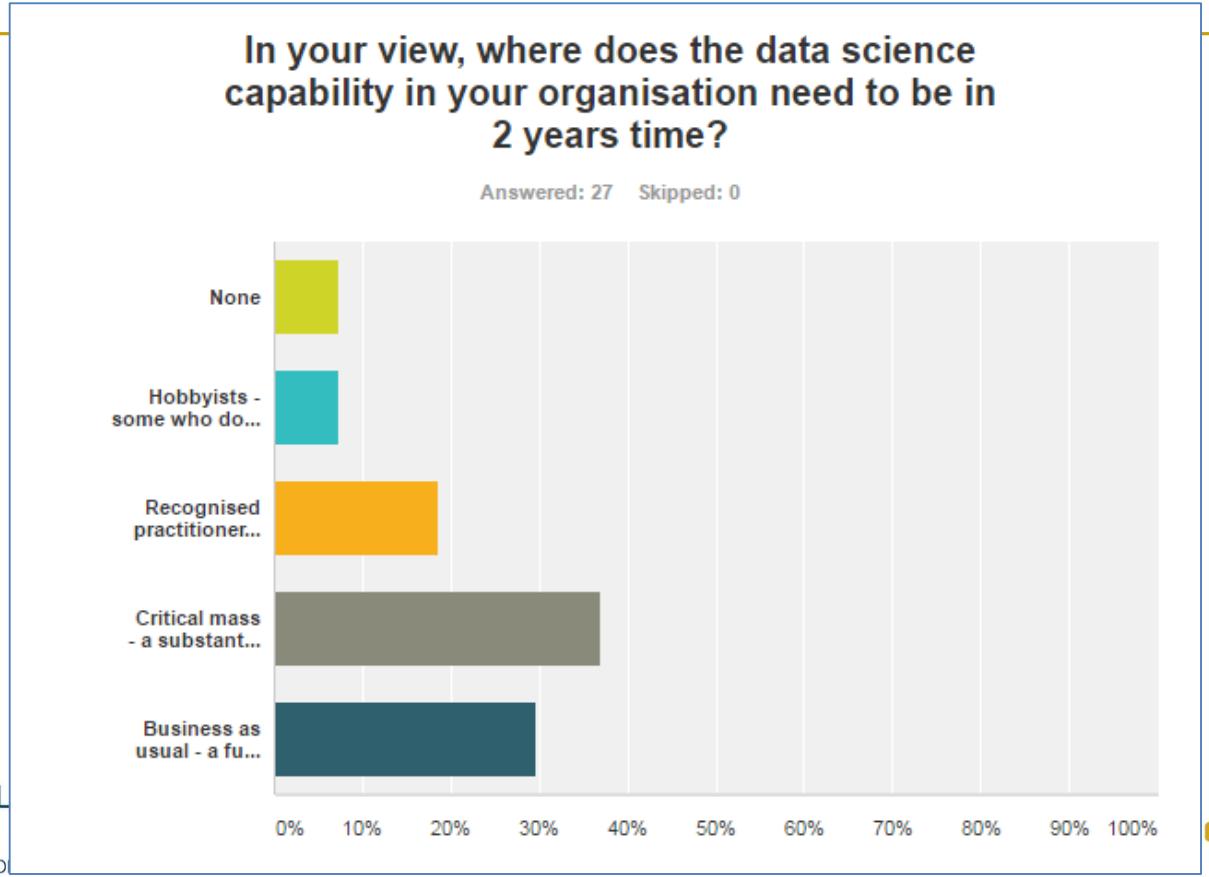
ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DE

Broad range of contributions from data scientists



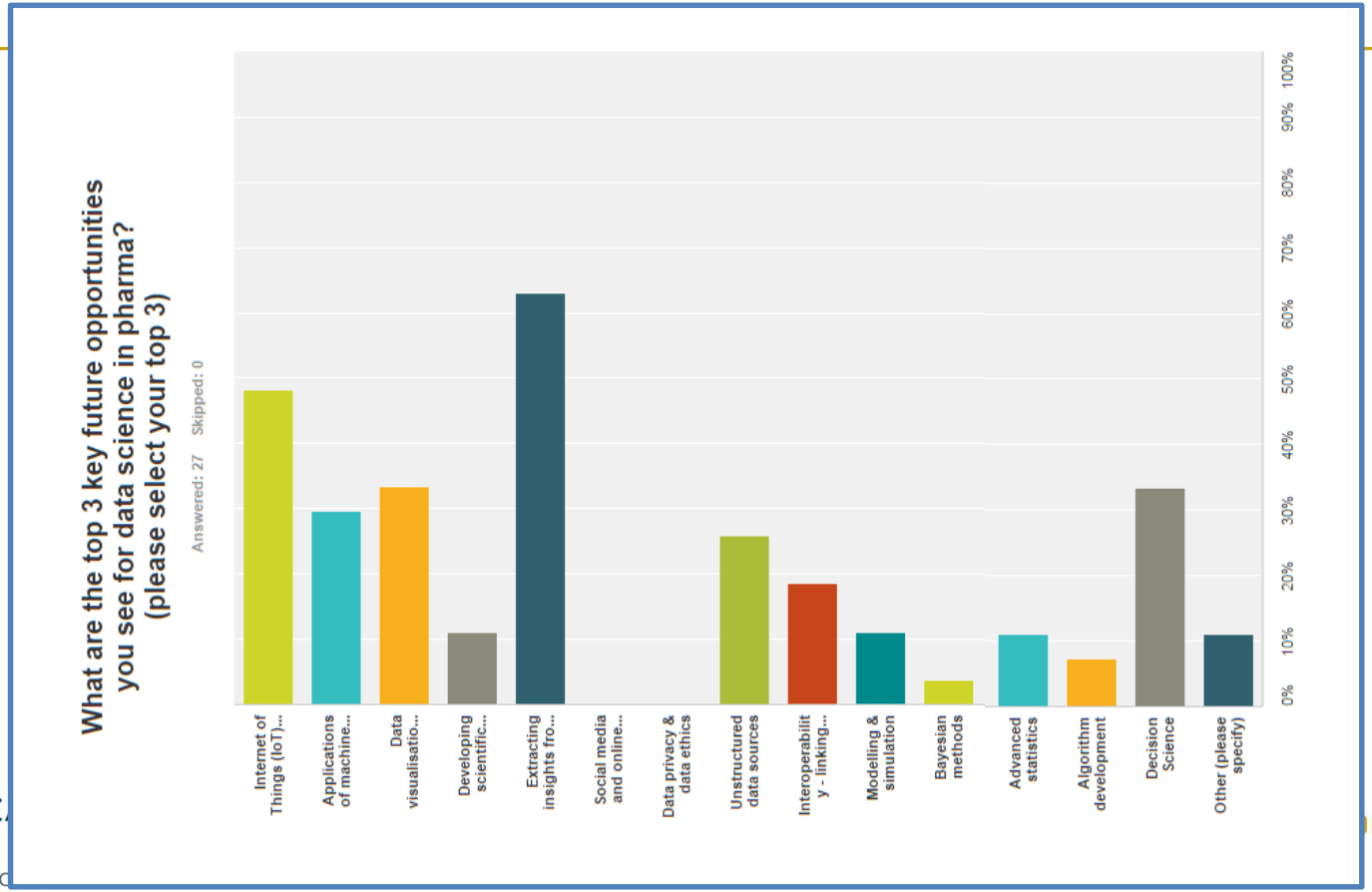
ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISI

Most believe a more mature data science capability is needed



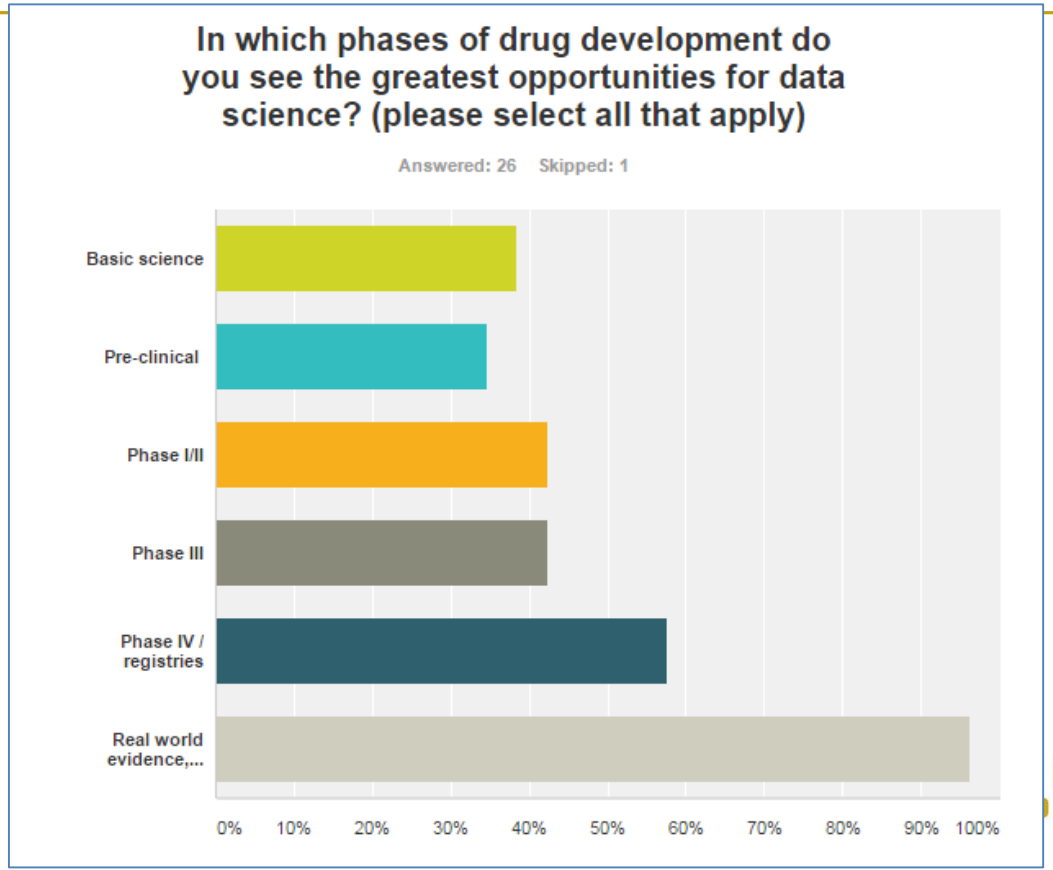
ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIO

Future look: Insights, IoT, visualisation & decision science



ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DEC

Opportunities for data science throughout development



ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Where is the gap?

ROYA STATIS SOCIETY

DATA | EVIDENCE | DECISIONS

There are a wide range of perspectives

we definitely lack people able to assemble or transform the diverse datasets; we also need more associates knowledgeable or experts in Machine learning type of methods

Develop experienced DS teams gathering expertise in technology/mathematics/computer sciences while being open minded and being able to embark and lead DS projects with other scientists (biologists --> clinicians) or internal partners

Organisational boundaries

Complexity of the big data topic and variety of potential applications makes it challenging to focus and join forces between computationally oriented and statistically oriented staff

This is a multidimensional activity needing staff with different skills. Challenge is to have the right balance in the team

Statisticians with an interest in non-traditional data sources people with an interest in non-traditional data sources who understand anything about statistics, uncertainty, randomness

Limited resources/competencies in the critical areas like AI, wearable/sensor technologies

Unstructured data

Strong programming skills

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Agenda

- Introduction – 15 mins

LUNCH

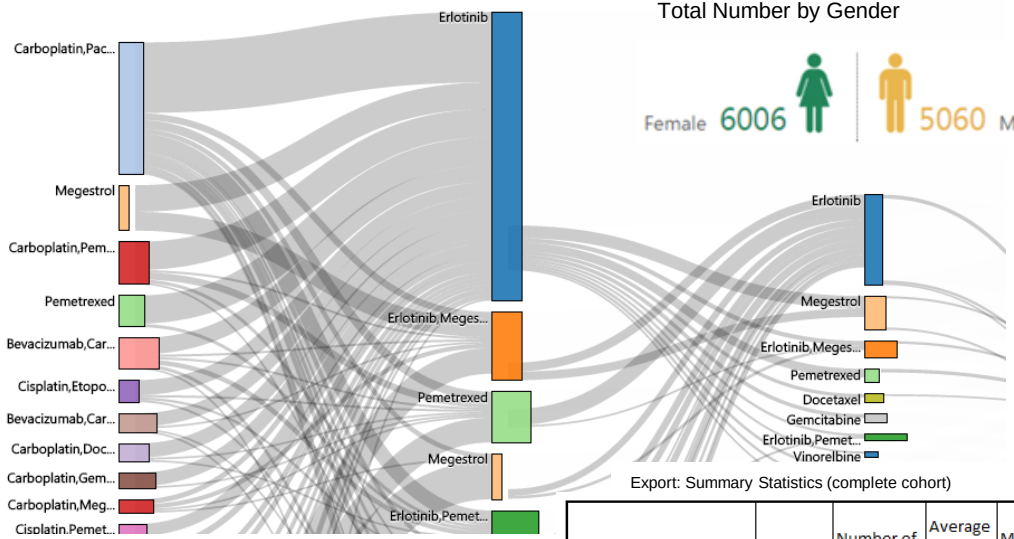
- Survey analysis – 15 mins
- **Case studies – 10 mins**
- Group work: four themes - 30 mins
  - The Internet of Things
  - Big Data: EHRs
  - Decision science
  - Automation and artificial intelligence
- Report back – 25 mins
- Discussion – 25 mins

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Patient Flows in EHR data

OncologyFlo

Result: Example showing treatment pathways of Lung cancer patients treated with erlotinib after diagnosis



Total Number by Gender

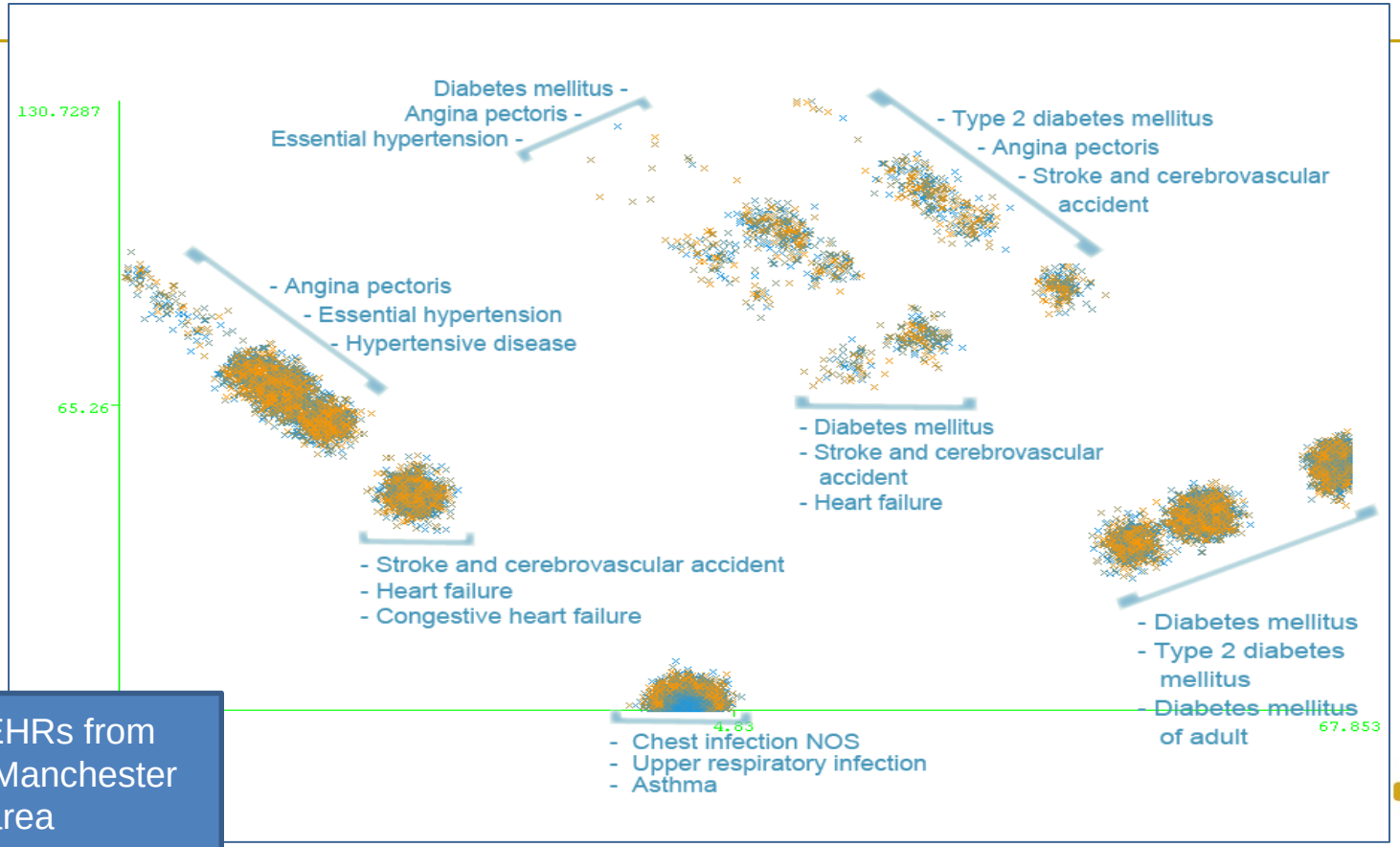
Median Age at First Diagnosis

Export: Summary Statistics (complete cohort)

Therapy	Sequence	Number of Patients	Average Time (Days)	Median Time (Days)
Carboplatin,Paclitaxel	1	592	114.3	84
Erlotinib	1	3735	310.5	95
Erlotinib	2	1674	222.9	62

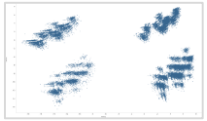
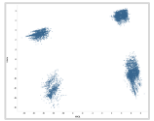
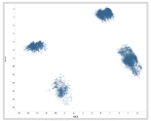
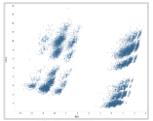
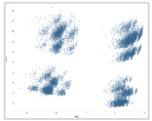
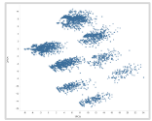
ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Unsupervised machine learning – Insights into healthcare



Data: EHRs from Greater Manchester area

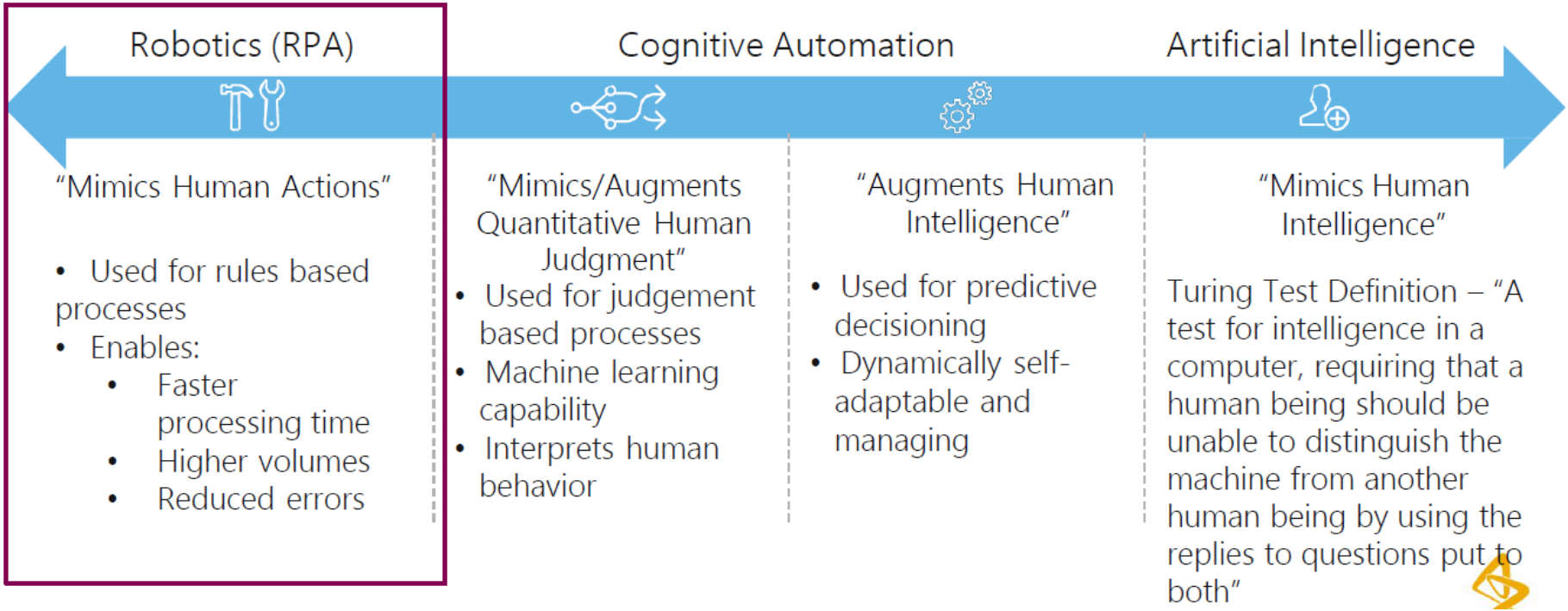
“Seven Ages of Man” healthcare clustering

	Infant & schoolboy Age 0-17	Lover Age 18-29	Solider Age 30-39	Justice Age 40-59	Old age Age 60-79	Incapacity Age > 80
PCA figures: (1PC vs. 2PC)						
						
PCA analysis:	• Rashes (e.g nappy rashes) • Acne • Eczema	• Injuries (e.g sports) • Pains (knee pain, ankle pain, etc…) • skin and subcutaneous tissue disease	• Circulatory system disease (hypertension, atrial fibrillation) • Respiratory system disease (chest infection, throat infection) • Diabetes	• Falls • Pains (back pain, pain in limp) • Urinary system disease (Urinary tract infection)		

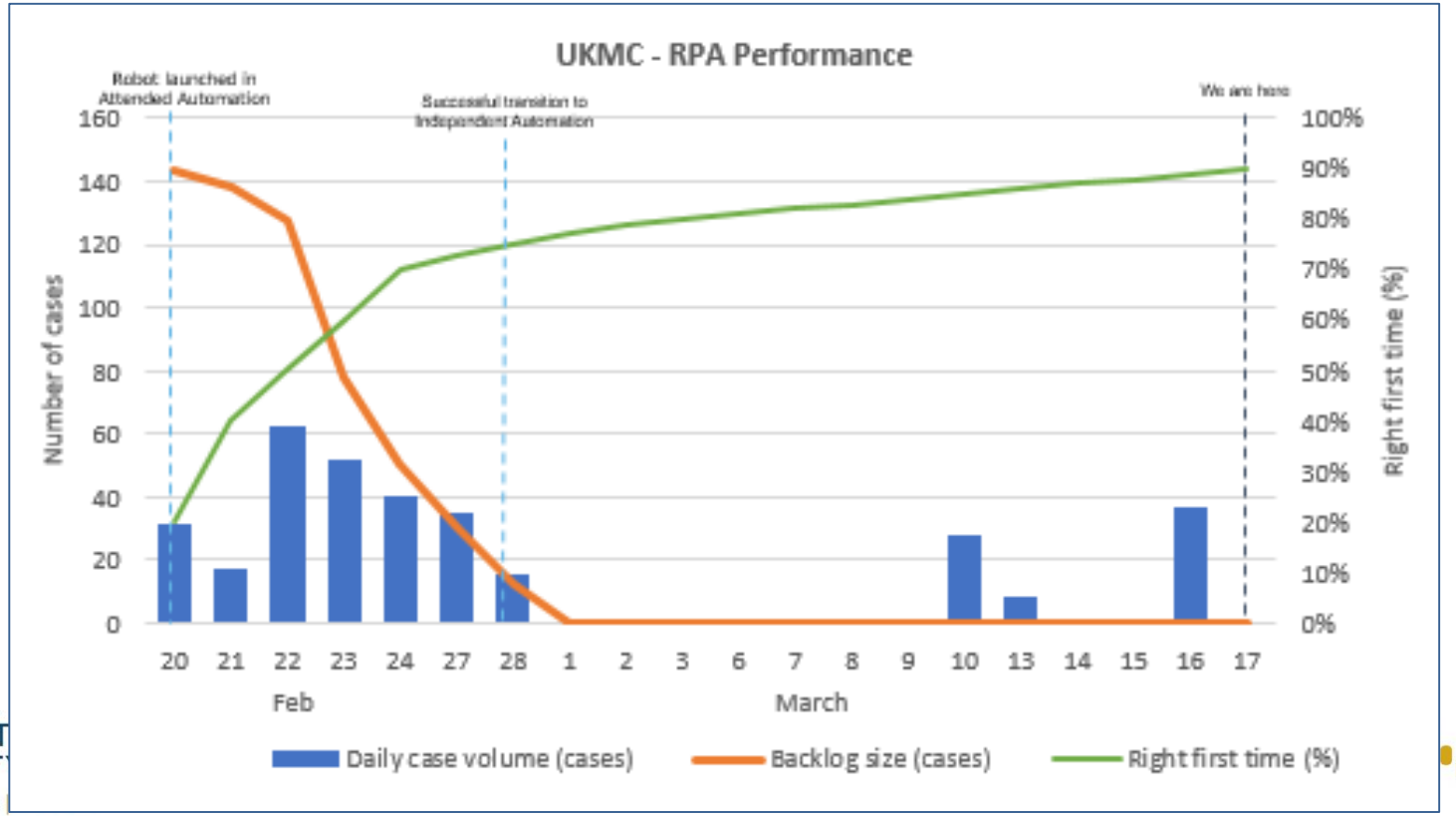
ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

What is robotic process automation (RPA)

- **Software that automates repetitive, rules-based tasks to free up your best people to be your best people**



Safety data collection via Robotic Process Automation




1. Fundamentals
Matrices & Linear Algebra Fundamentals
Hash Functions, Binary Tree, O(n)
Relational Algebra, DB Basics
Inner, Outer, Cross, Theta Join
CAP Theorem
Tabular Data
Data Frames & Series
Sharding
OLAP
Multidimensional Data Model
ETL
Reporting Vs BI Vs Analytics
Entropy
JSON & XML
NoSQL
Regex
Vendor Landscape
Env Setup
5%
2. Statistics
Prob Den Fn (PDF)
ANOVA
Skewness
Continuos Distributions (Normal, Poisson, Gaussian)
Cumul Dist Fn (CDF)
Random Variables
Bayes Theorem
Probability Theory
Percentiles & Outliers
Histograms
Exploratory Data Analysis
Descriptive Statistics (mean, median, range, SD, Var)
Pick a Dataset (UCI Repo)
Central Limit Theorem
Monte Carlo Method
Hypothesis Testing
P-Value
Chi² Test
Estimation
Confid Int (CI)
MLE
Kernel Density Estimate
Regression
Covariance
15%
Factor Analysis
Install Pkgs
3. Programming
Data Frames
Reading CSV Data
Reading Raw Data
Subsetting Data
Manipulate Data Frames
Functions
Lists
Factors
Arrays
Matrices
Vectors
Variables
Expressions
R Basics
R Setup
R Studio
Rapid Miner
IBM SPSS
Python Basics
Working in Excel
4. Machine Learning
Regression
Perceptron
Linear Regression
Ranking
Logistic Regression
Classification
K-Nearest Neighbor
Naive Bayes Classifiers
Boosting
Decision Trees
Classification Rate
Trees & Classification
Bias & Variance
Overfitting
Lift
Prediction
Classifier
Training & Test Data
Concepts, Inputs & Attributes
Unsupervised Learning
Supervised Learning
Categorical Var
Numerical Var
What is ML?
Clustering
Hierarchical Clustering
K-means Clustering
Neural Networks
Sentiment Analysis
Collaborative Filtering
Tagging
Vocabulary Mapping
Classify Text
Using NLTK
Using Weka
Using Mahout
Feature Extraction
5. Text Mining / NLP
Corpus
Named Entity Recognition
Text Analysis
UIMA
Term Document Matrix
Term Document Matrix
Term Frequency & Weight
Support Vector Machines
Association Rules
Market Based Analysis
50%
30%
Euclidean Distance
Least Fit
Causation
Pearson Coeff
Correlation
6. Visualization
Data Exploration in R (Hist, Boxplot etc)
Uni, Bi & Multivariate Viz
ggplot2
Histogram & Pie (Uni)
Tree & Tree Map
Scatter Plot (Bi)
Line Charts (Bi)
Spatial Charts
Survey Plot
Timeline
Decision Tree
10%
Tableau
IBM ManyEyes
Infovis
D3.js
7. Big Data
Job & Task Tracker
MR Programming
Scoop: Loading Data in HDFS
Flume, Scribe : For Unstruct Data
SQL with Pig
DWH with Hive
Scribe, Chukwa For Weblog
Using Mahout
Name & Data Nodes
Setup Hadoop (IBM / Cloudera / HortonWorks)
Data Replication Principles
HDFS
Hadoop Components
Map Reduce Fundamentals
Zookeeper
Avro
Storm: Hadoop Realtime
Rhadoop, RHIPE
rmr
Cassandra
MongoDB, Neo4j
80%
8. Data Ingestion
Data Survey
Transformation & Enrichment
Data Fusion
Data Integration
Data Sources & Acquisition
Data Discovery
Summary of Data Formats
Sampling
Stratified Sampling
Principal Component Analysis
Using ETL
How much Data?
Google OpenRefine
80%
9. Data Munging
Denoising
Feature Extraction
Binning Sparse Values
Unbiased Estimators
Handling Missing Values
Data Scrubbing
Normalization
Dimensionality & Numerosity Reduction
10. Toolbox
MS Excel w/ Analysis ToolPak
Java, Python
R, R-Studio, Rattle
Weka, Knime, RapidMiner
Hadoop Dist of Choice
Spark, Storm
Flume, Scibe, Chukwa
Nutch, Talend, Scraperwiki
Webscraper, Flume, Sqoop
tm, RWeka, NLTK
RHIPE
D3.js, ggplot2, Shiny
IBM Languageware
Cassandra, MongoDB
100%

Agenda

- Introduction – 15 mins

LUNCH

- Survey analysis – 15 mins
- Case studies – 10 mins
- **Group work: four themes – 30 mins**
  - **The Internet of Things**
  - **Big Data: EHRs**
  - **Decision science**
  - **Automation and artificial intelligence**
- Report back – 25 mins
- Discussion – 25 mins

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Group work – three key questions

1. Brainstorm: what are the main opportunities and challenges
2. What are the top 3 areas we should address as statistical leaders
3. What immediate action should we take next?

Agenda

- Introduction – 15 mins

LUNCH

- Survey analysis – 15 mins
- Case studies – 10 mins
- Group work: four themes - 30 mins
    - The Internet of Things
    - Big Data: EHRs
    - Decision science
    - Automation and artificial intelligence
- **Report back – 25 mins**
- Discussion – 25 mins

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Agenda

- Introduction – 15 mins

LUNCH

- Survey analysis – 15 mins
- Case studies – 10 mins
- Group work: four themes - 30 mins
    - The Internet of Things
    - Big Data: EHRs
    - Decision science
    - Automation and artificial intelligence
- Report back – 25 mins
- **Discussion – 25 mins**

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

The End

Thank you!

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS