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

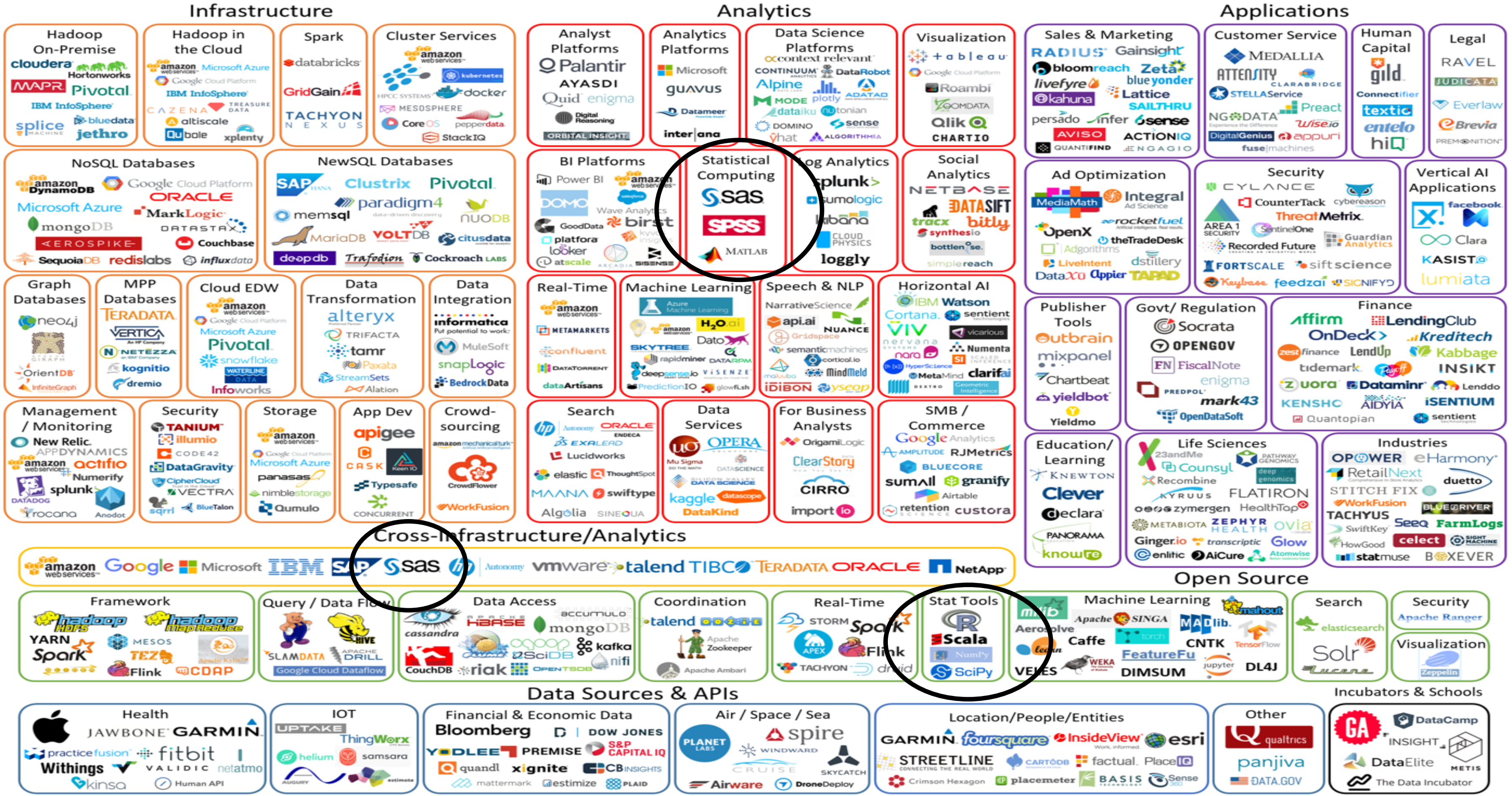
Agenda

- **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

Big Data Landscape 2016



Infrastructure

Hadoop On-Premise
cloudera, Hortonworks, MAPR, Pivotal, IBM InfoSphere, splice MACHINE, bluedata, jethro

Hadoop in the Cloud
amazon web services, Microsoft Azure, Google Cloud Platform, IBM InfoSphere, CAZENA, TREASURE DATA, altiscale, Qubole, xplenty

Spark
databricks, GridGain, TACHYON NEXUS

Cluster Services
amazon web services, kubernetes, HPCC SYSTEMS, docker, MESOSPHERE, CoreOS, pepperdata, StackIQ

NoSQL Databases
amazon DynamoDB, Google Cloud Platform, ORACLE, Microsoft Azure, MarkLogic, mongoDB, DATASTAX, AEROSPIKE, Couchbase, SequoiaDB, redislabs, influxdata

NewSQL Databases
SAP HANA, Clustrix, Pivotal, paradigm4 data-driven discovery, memsql, NUODB, MariaDB, VOLTDB, citusdata, deepdb, Trafodion, Cockroach LABS

Graph Databases
neo4j, APACHE GIRAPH, OrientDB, InfiniteGraph

MPP Databases
TERADATA, VERTICA An HP Company, NETEZZA an IBM Company, kognitio, dremio

Cloud EDW
amazon web services, Google Cloud Platform, Microsoft Azure, Pivotal, snowflake, WATERLINE DATA, Infoworks

Data Transformation
alteryx, TRIFACTA, tamr, Paxata, StreamSets, Alation

Data Integration
informatica Put potential to work, MuleSoft, snapLogic, BedrockData

Management / Monitoring
New Relic, APPDYNAMICS, amazon web services, actifio, Numerify, DATADOG, splunk, rocana, Anodot

Security
TANIUM, illumio, CODE42, DataGravity, CipherCloud, VECTRA, sqrrl, BlueTalon

Storage
amazon web services, Google Cloud Platform, Microsoft Azure, panasas, nimblestorage, Qumulo

App Dev
apigee, CASK, Keen IO, Typesafe, CONCURRENT

Crowd-sourcing
amazon mechanicalturk, CrowdFlower, WorkFusion

Analytics

Analyst Platforms
Palantir, AYASDI, Quid, enigma, Digital Reasoning, ORBITAL INSIGHT

Analytics Platforms
Microsoft, guavus, Datameer, inter|ana

Data Science Platforms
context relevant, CONTINUUM ANALYTICS, DataRobot, Alpine, plotly, ADATAO, MODE, dataiku, nutonian, DOMINO, sense, yhat, ALGORITHMIA

Visualization
tableau, Google Cloud Platform, Roambi, ZOOMDATA, Qlik, CHARTIO

BI Platforms
Power BI, amazon web services, DOMO, salesforce Wave Analytics, GoodData, birst, platfora, kyvos insights, looker, atscale, ARCADIA, SISENSE

Statistical Computing
sas, SPSS, MATLAB

Log Analytics
splunk>, sumologic, kibana, CLOUD PHYSICS, loggly

Social Analytics
NETBASE, DATASIFT, tracx, bitly, synthesio, bottlenose, simplereach

Real-Time
amazon web services, METAMARKETS, confluent, DATATORRENT, dataArtisans

Machine Learning
Azure Machine Learning, amazon web services, H2O.ai, Dato, SKYTREE, rapidminer, DATARPM, deepsense.io, VISENZE, PredictionIO, glowfi.sh

Speech & NLP
NarrativeScience, api.ai, NUANCE, Gridspace, semanticmachines, cortical.io, maluuba, MindMeld, IDIBON, yseop

Horizontal AI
IBM Watson, Cortana, sentient technologies, VIV, vicarious, nervana systems, nara, Numenta, SCALED INFERENCE, HyperScience, MetaMind, clarifai, DEXTRO, Geometric Intelligence

Search
hp Autonomy, ORACLE ENDECA, EXALEAD, Lucidworks, elastic, ThoughtSpot, MAANA, swiftype, Algolia, SINEQUA

Data Services
Mu Sigma DO THE MATH, OPERA SOLUTIONS, DATASCIENCE, SILICON VALLEY DATA SCIENCE, kaggle, datascope, DataKind

For Business Analysts
OrigamiLogic, ClearStory DATA, CIRRO, import io

SMB / Commerce
Google Analytics, AMPLITUDE, RJMetrics, BLUECORE, sumAll, granify, Airtable, retention science, custora

Applications

Sales & Marketing
RADIUS, Gainsight, bloomreach, Zeta, livefyre, blue yonder, kahuna, Lattice, SAILTHRU, persado, infer, 6sense, AVISO, ACTIONIQ, QUANTIFIND, ENGAGIO

Customer Service
MEDALLIA, ATTENSITY, CLARABRIDGE, STELLAService, Preact, NGDATA Experience the Difference, Wise.io, DigitalGenius, appuri, fuse|machines

Human Capital
gild, Connectifier, textio, entelo, hiQ

Legal
RAVEL, JUDICATA, Everlaw, Brevia, PREMONITION

Ad Optimization
MediaMath, Integral Ad Science, rocketfuel Artificial intelligence. Real results., OpenX, theTradeDesk, Adgorithms, LiveIntent, dstillery, DataXu, Appier, TAPAD

Security
CYLANCE, cybereason, CounterTack, ThreatMetrix, AREA 1 SECURITY, SentinelOne, Recorded Future, Guardian Analytics, FORTSCALE, sift science, Keybase, feedzai, SIGNIFYD

Vertical AI Applications
X., facebook, Clara, KASISTO, lumiata

Publisher Tools
outbrain, mixpanel, Chartbeat, yieldbot, Yieldmo

Govt/ Regulation
Socrata, OPENGOV, FiscalNote, enigma, PREDPOL, mark43, OpenDataSoft

Finance
affirm, LendingClub, OnDeck, Kreditech, zestfinance, LendUp, Kabbage, tidemark, Payoff, INSIKT, zuora, Dataminr, Lenddo, KENSHO, AIDYIA, iSENTIUM, Quantopian, sentient technologies

Education/ Learning
KNEWTON, Clever, declara, PANORAMA EDUCATION, knowre

Life Sciences
23andMe, Counsyl, PATHWAY GENOMICS, deep genomics, Recombine, KYRUUS, FLATIRON, zymergen, HealthTap, METABIOTA, ZEPHYR HEALTH, ovia, Ginger.io, transcriptic, Glow, enlitic, AiCure, Atomwise

Industries
OPOWER, eHarmony, RetailNext, duetto, STITCH FIX, WorkFusion, BLUE RIVER, TACHYUS, Seeq, FarmLogs, SwiftKey, HowGood, celect, SIGHT MACHINE, statmuse, BOXEVER

Cross-Infrastructure/Analytics

amazon web services, Google, Microsoft, IBM, SAP, sas, hp Autonomy, vmware, talend, TIBCO, TERADATA, ORACLE, NetApp

Open Source

Framework
hadoop HDFS, hadoop MapReduce, YARN, Spark, MESOS, TEZ, Apache Kylin, Flink, CDAP

Query / Data Flow
HIVE, SLAMDATA, APACHE DRILL, Google Cloud Dataflow

Data Access
cassandra, APACHE HBASE, accumulo, mongoDB, SciDB, kafka, nifi, CouchDB, riak, OPENTSDB

Coordination
talend, OOZIE, Apache Zookeeper, Apache Ambari

Real-Time
STORM, Spark, APEX, Flink, TACHYON, druid

Stat Tools
R, Scala, NumPy, SciPy

Machine Learning
mllib, Apache SINGA, MADlib., mahout, Aerosolve, Caffe, torch, CNTK, TensorFlow, learn, FeatureFu, jupyter, VELES, WEKA, DIMSUM, DL4J

Search
elasticsearch, Solr, Lucene

Security
Apache Ranger

Visualization
Zeppelin

Data Sources & APIs

Health
JAWBONE, GARMIN, practice fusion, fitbit, Withings, VALIDIC, netatmo, kinsa, Human API

IOT
UPTAKE, ThingWorx, helium, samsara, AUGURY, estimote

Financial & Economic Data
Bloomberg, DOW JONES, YODLEE, PREMISE, S&P CAPITAL IQ, quandl, xignite, CB INSIGHTS, mattermark, estimize, PLAID

Air / Space / Sea
PLANET LABS, spire, WINDWARD, CRUISE, SKYCATCH, Airware, DroneDeploy

Location/People/Entities
GARMIN, foursquare, InsideView Work. informed., esri, STREETLINE CONNECTING THE REAL WORLD, CARTODB, factual., PlaceIQ, Crimson Hexagon, placemeter, BASIS, Sense360

Other
qualtrics, panjiva, DATA.GOV

Incubators & Schools

GA, DataCamp, INSIGHT, METIS, DataElite, The Data Incubator

© Matt Turck (@mattturck), Jim Hao (@jimrhao), & FirstMark Capital (@firstmarkcap)

FIRSTMARK

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

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/11500048/

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

Please sign up!

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

Evidence
Disease
Decision
Diversified
Hypotheses
Failed
Address
Scientific
High
Continued
Answer
Expertise
Technical Visualisation
Structured
Ir
Knowledge
Far
Out
Patients
Rate
Drive
Hta
Technology
Proper
Kinds
Generate
Insights
Includes
Mass
Explore
Combination
Abuse
Area
Statistics
Within
Company
Educated
Makers
Analyse
Support
Applied
More
Sigma
Understand
Extract
Skills
Tackle
Failure
Deriving
Needs
Six
Survey
Business
Order
Field
Geeks
Interpreting
Gain
Programming
Inform
Forms
Verify
Setup
Phase
Math
See
Sets
Process
Large
Big
Ai
Stat
Sources
House
Already
Quite
Computer
Specific
Today
Prediction
Design
Collection
Wiser
Looking
Ways
People
Access
Such
Benefit
Various
Acumen
Driven
Interdisciplinary
Being
Enriched
Systems
Domain
Science
Dm
Crf
Analytics
Use
Extra
Based
Concepts
Charge
Scientists
Simulation
Further
Unstructured
Events
Data

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

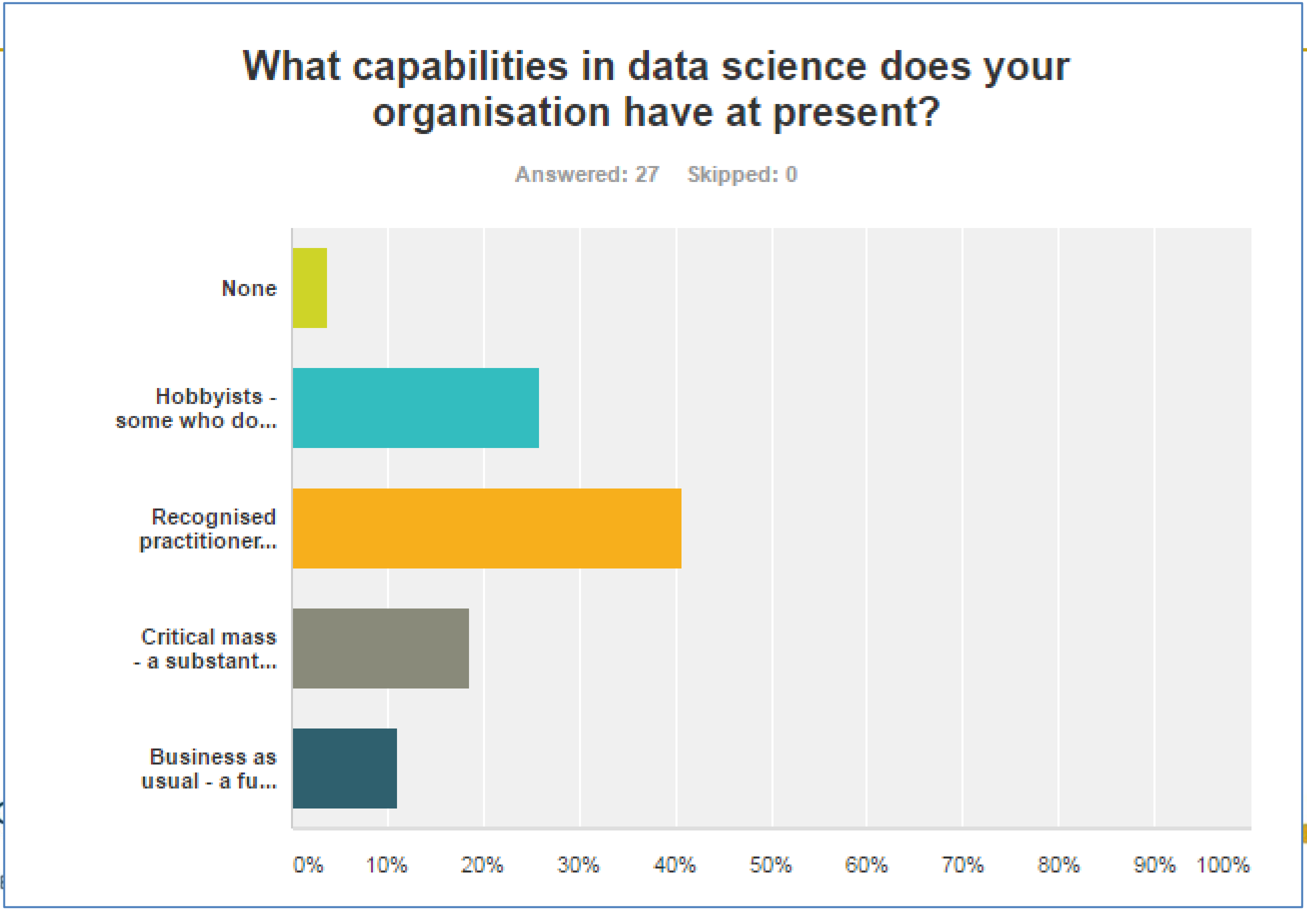
Evidence that we have failed as a statistical discipline

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

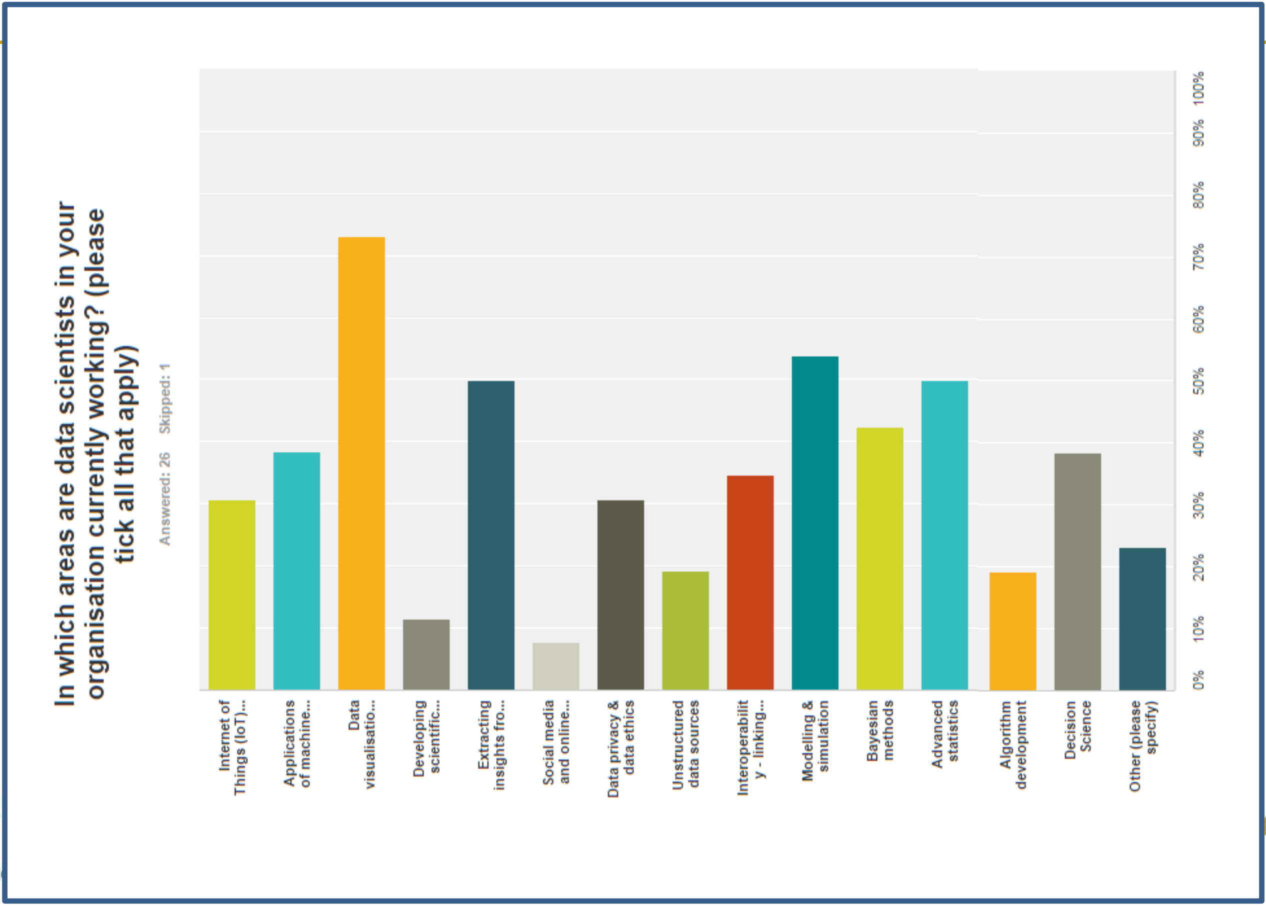
Data science is recognised in most organisations



What capabilities in data science does your organisation have at present?

Answered: 27 Skipped: 0

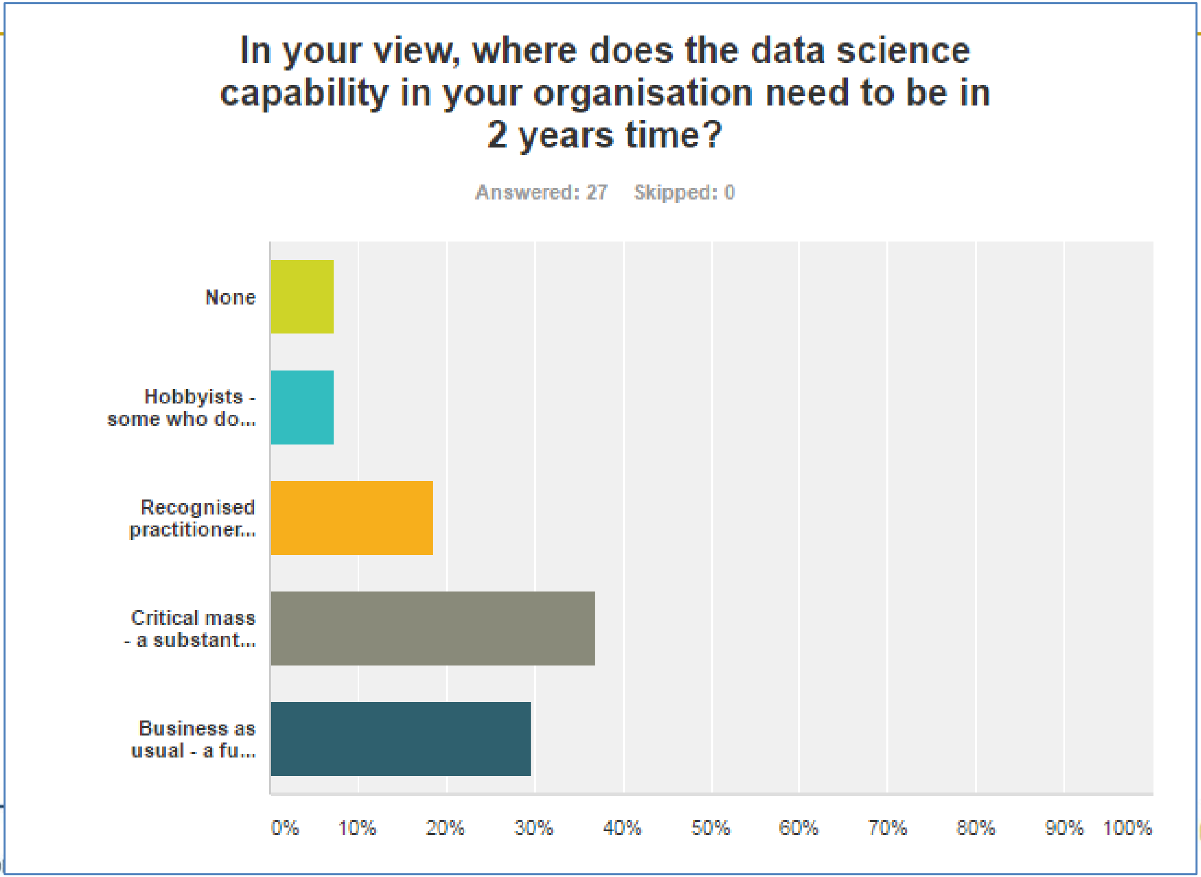
Broad range of contributions from data scientists



In which areas are data scientists in your organisation currently working? (please tick all that apply)

Answered: 26 Skipped: 1

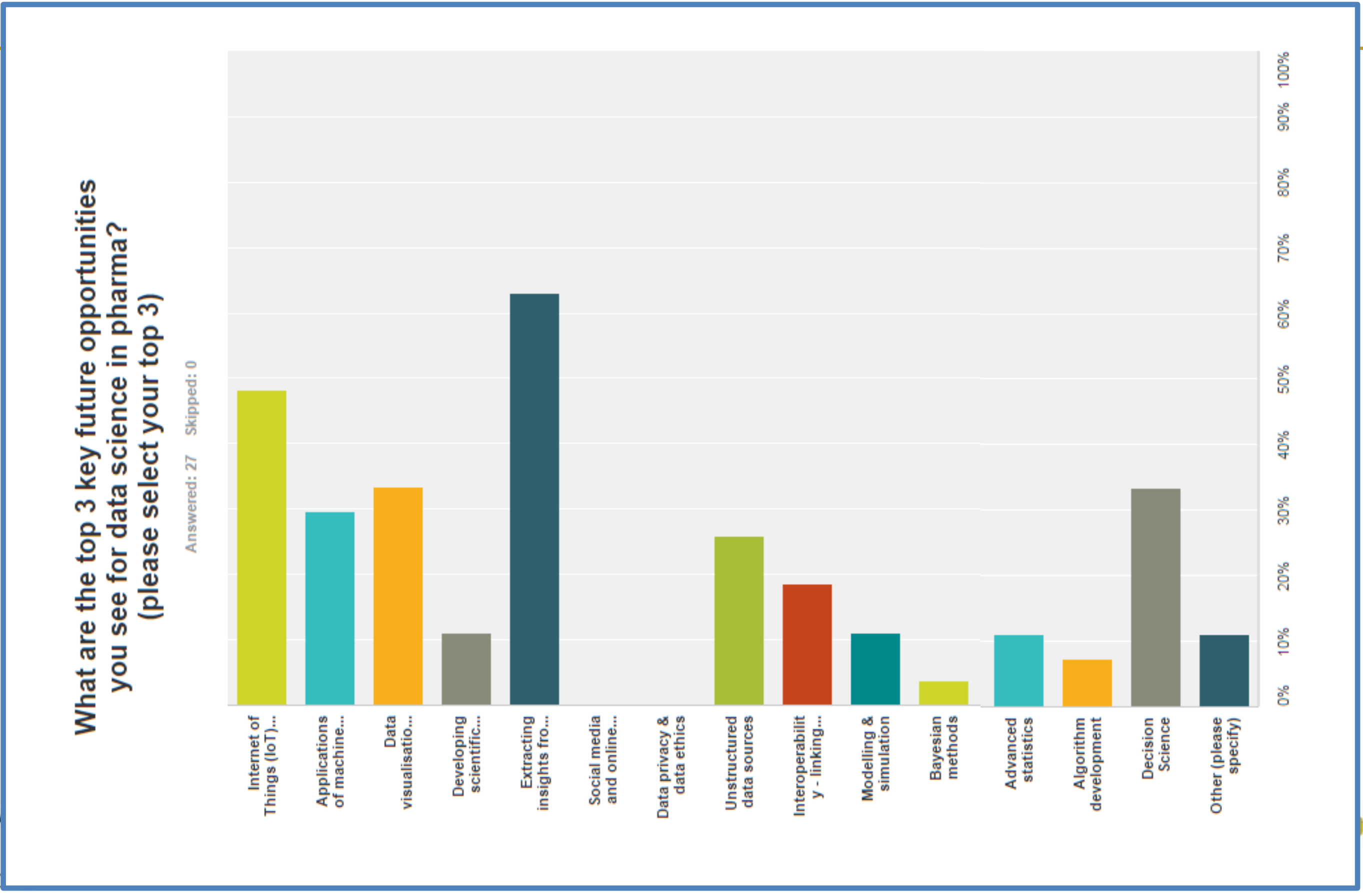
Most believe a more mature data science capability is needed



In your view, where does the data science capability in your organisation need to be in 2 years time?

Answered: 27 Skipped: 0

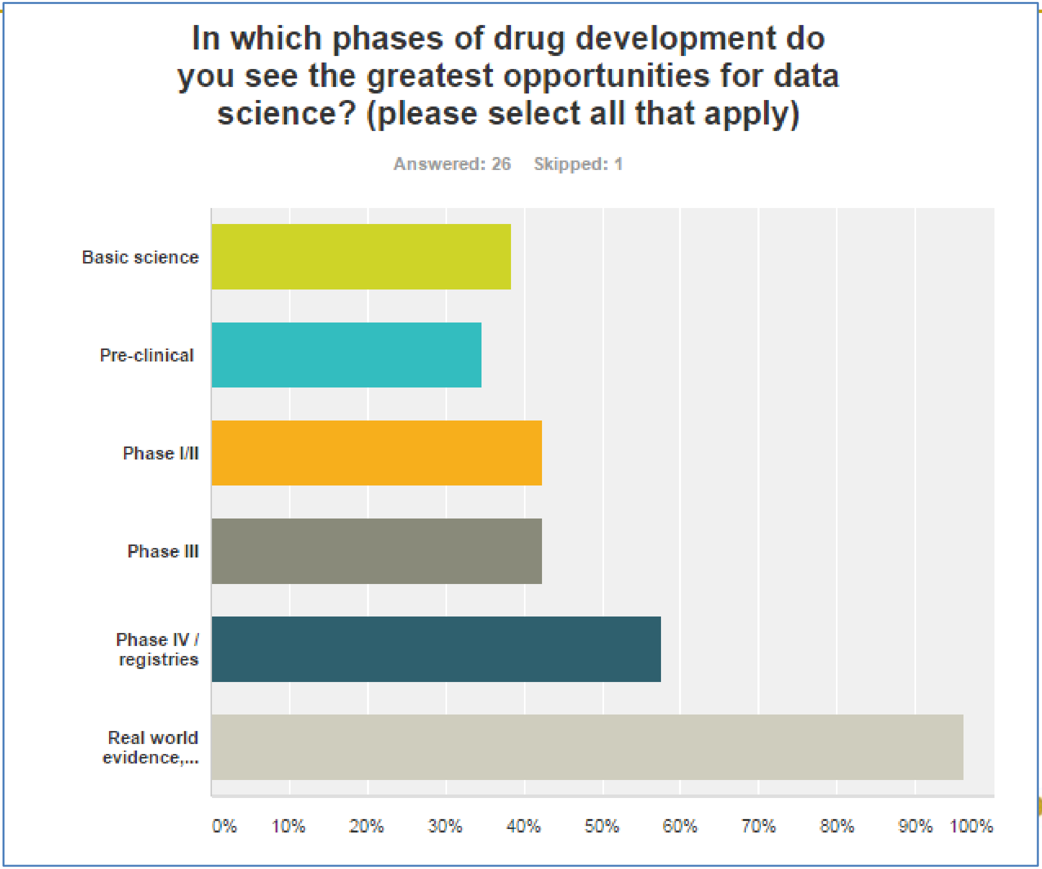
Future look: Insights, IoT, visualisation & decision science



What are the top 3 key future opportunities you see for data science in pharma? (please select your top 3)

Answered: 27 Skipped: 0

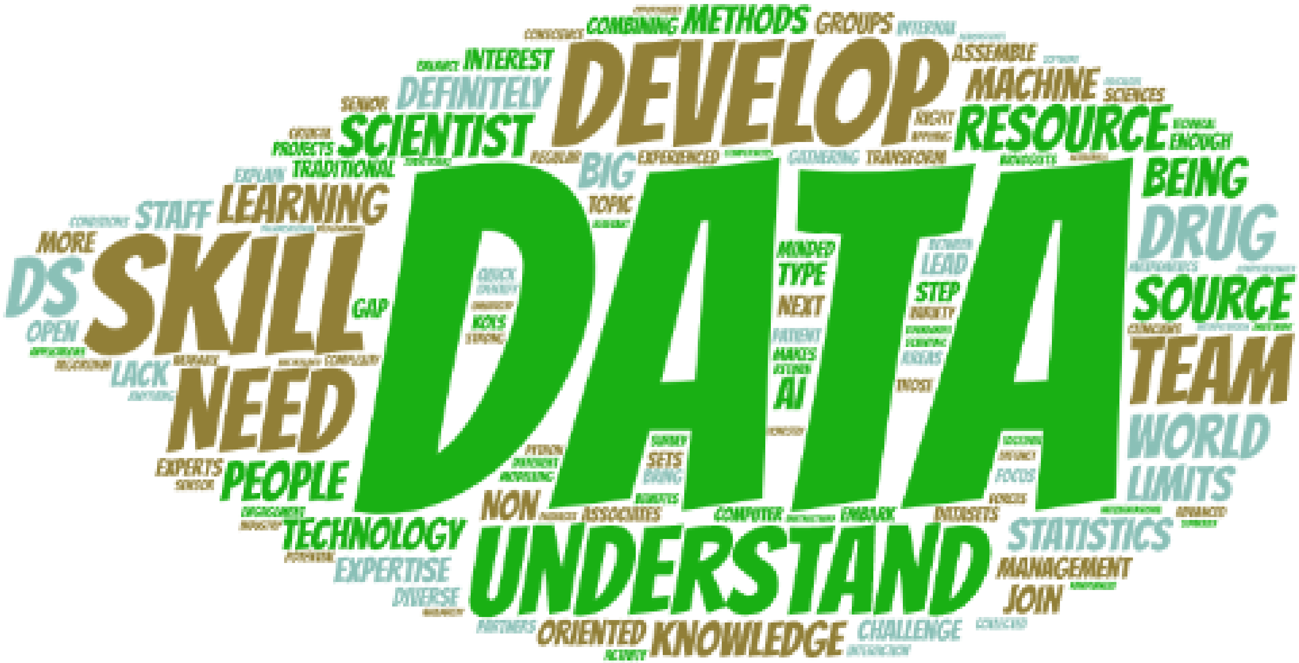
Opportunities for data science throughout development



In which phases of drug development do you see the greatest opportunities for data science? (please select all that apply)

Answered: 26 Skipped: 1

Where is the gap?



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

Strong programming skills

Unstructured data

Agenda

- Introduction – 15 mins

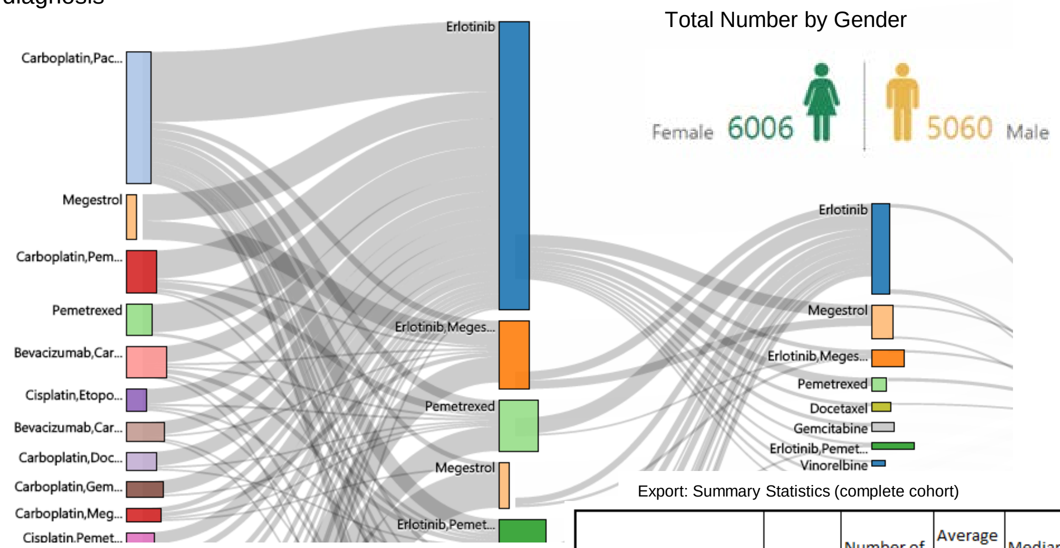
LUNCH

- 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

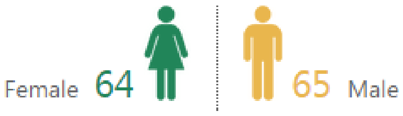
Patient Flows in EHR data

OncologyFlo

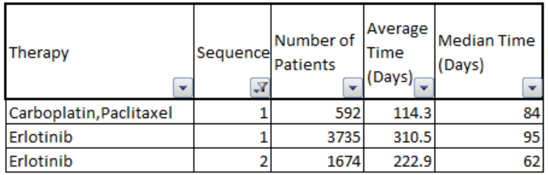
Result: Example showing treatment pathways of Lung cancer patients treated with erlotinib after diagnosis


Total Number by Gender

Median Age at First Diagnosis

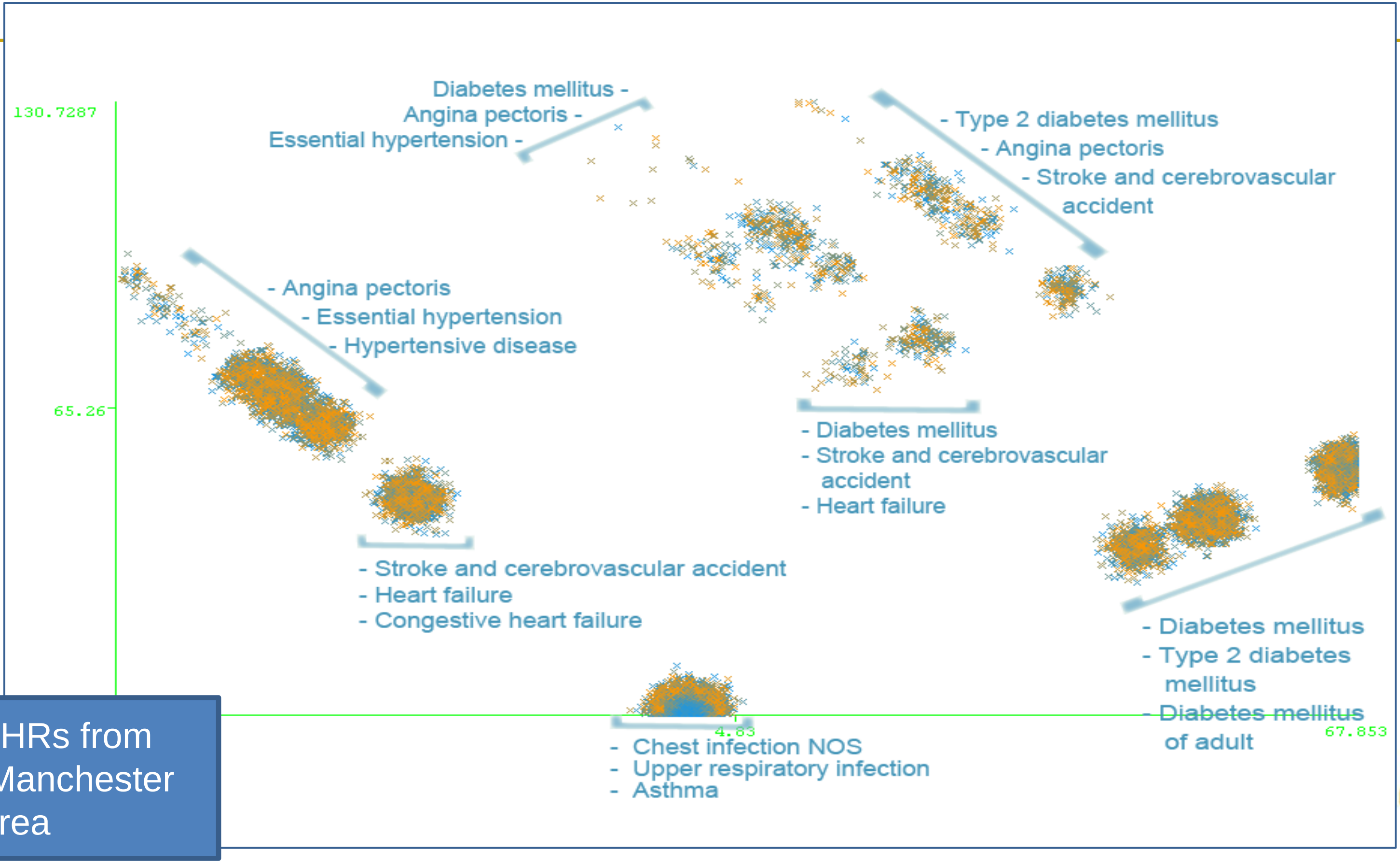


Export: Summary Statistics (complete cohort)



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

Unsupervised machine learning – Insights into healthcare



Diabetes mellitus -
Angina pectoris -
Essential hypertension -

- Type 2 diabetes mellitus
- Angina pectoris
- Stroke and cerebrovascular accident

- Angina pectoris
- Essential hypertension
- Hypertensive disease

- Diabetes mellitus
- Stroke and cerebrovascular accident
- Heart failure

- Stroke and cerebrovascular accident
- Heart failure
- Congestive heart failure

- Diabetes mellitus
- Type 2 diabetes mellitus
- Diabetes mellitus of adult

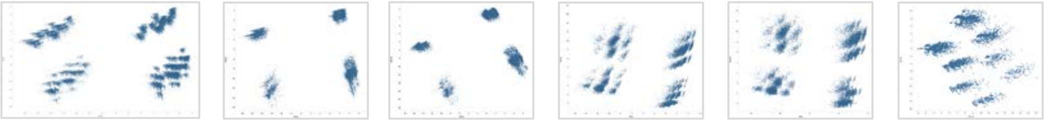
- Chest infection NOS
- Upper respiratory infection
- Asthma

Data: EHRs from Greater Manchester area

“Seven Ages of Man” healthcare clustering

Infant & schoolboy	Lover	Solider	Justice	Old age	Incapacity
Age 0-17	Age 18-29	Age 30-39	Age 40-59	Age 60-79	Age > 80

PCA figures:
(1PC vs. 2PC)



PCA analysis:

- Rashes (e.g nappy rashes)
- Acne
- Eczema

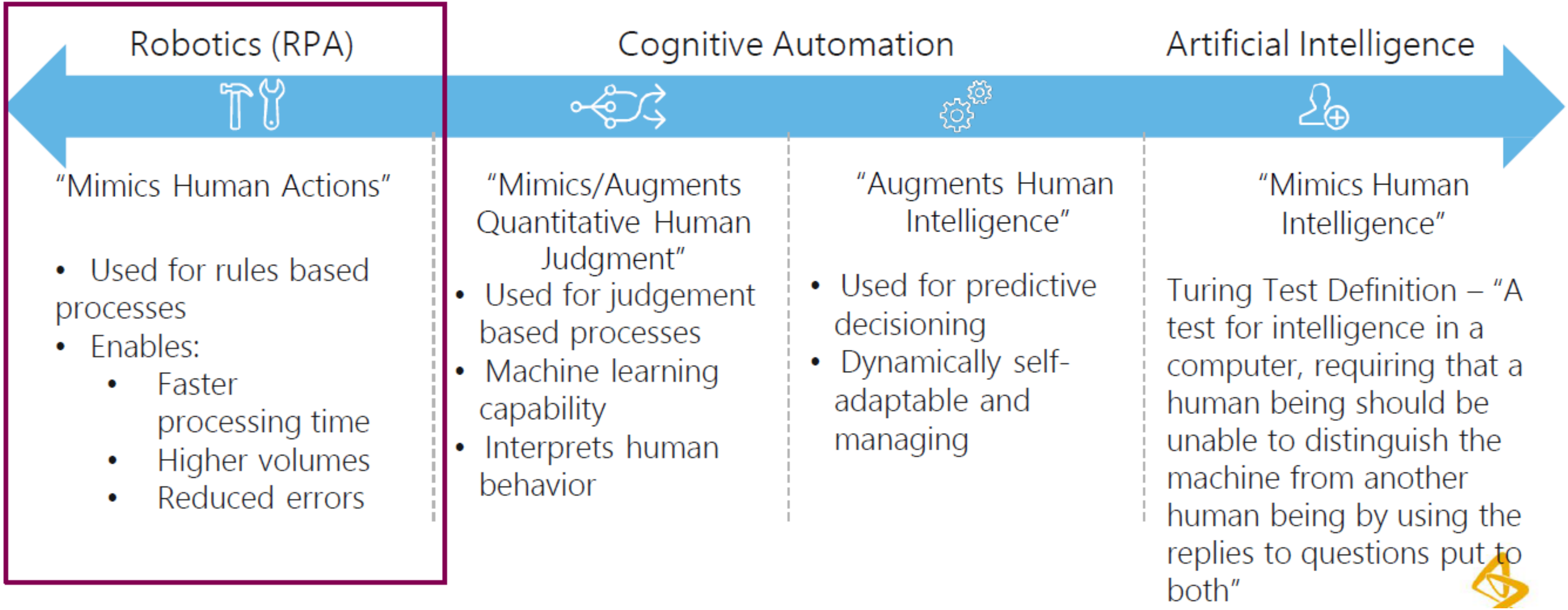
- Injuries (e.g sports)
- Pains (knee pain, ankle pain, etc…)
- skin and subcutaneous tissue disease

- 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)

What is robotic process automation (RPA)

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



Robotics (RPA)

"Mimics Human Actions"

- Used for rules based processes
- Enables:
  - Faster processing time
  - Higher volumes
  - Reduced errors

Cognitive Automation

"Mimics/Augments Quantitative Human Judgment"

- Used for judgement based processes
- Machine learning capability
- Interprets human behavior

"Augments Human Intelligence"

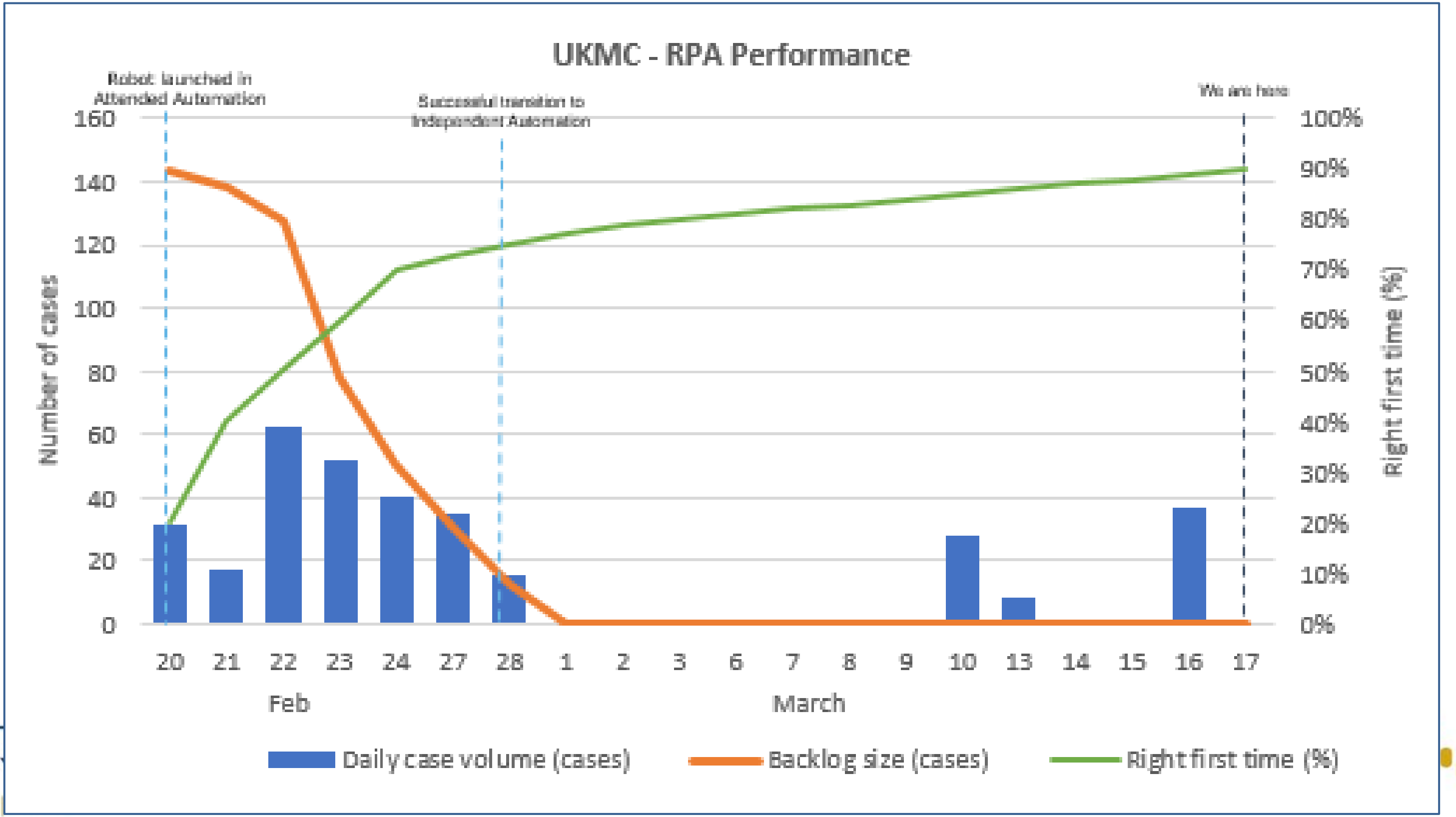
- Used for predictive decisioning
- Dynamically self-adaptable and managing

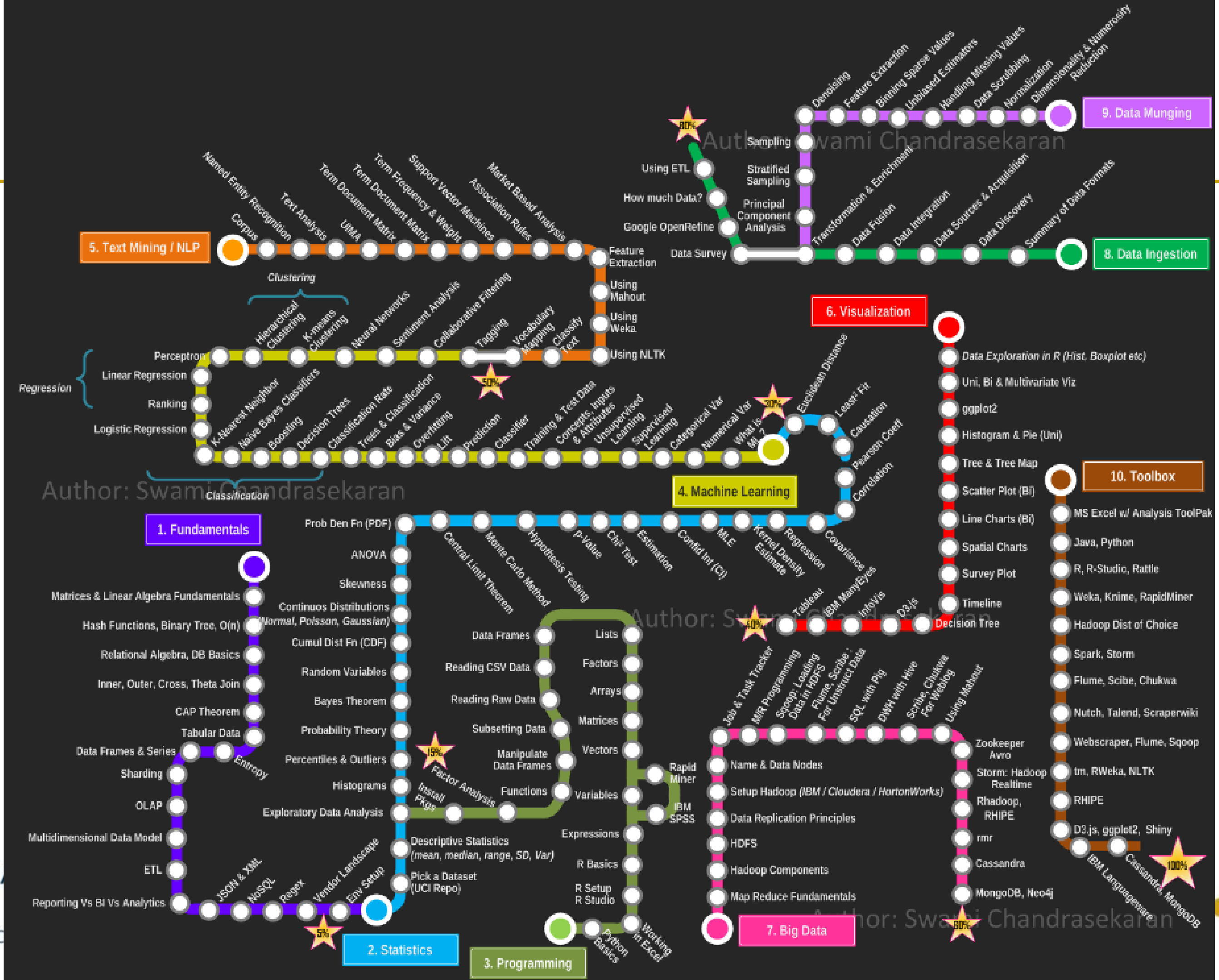
Artificial Intelligence

"Mimics Human Intelligence"

Turing Test Definition – "A test for intelligence in a computer, requiring that a human being should be unable to distinguish the machine from another human being by using the replies to questions put to both"

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

Entropy

Data Frames & Series

Sharding

OLAP

Multidimensional Data Model

ETL

Reporting Vs BI Vs Analytics

JSON & XML

NoSQL

Regex

Vendor Landscape

Env Setup

5%

2. Statistics

Pick a Dataset (UCI Repo)

Descriptive Statistics (*mean, median, range, SD, Var)*

Exploratory Data Analysis

Histograms

Percentiles & Outliers

Probability Theory

Bayes Theorem

Random Variables

Cumul Dist Fn (CDF)

Continuos Distributions (*Normal, Poisson, Gaussian)*

Skewness

ANOVA

Prob Den Fn (PDF)

Central Limit Theorem

Monte Carlo Method

Hypothesis Testing

p-Value

Chi² Test

Estimation

Confid Int (CI)

MLE

Kernel Density Estimate

Regression

Covariance

Correlation

Pearson Coeff

Causation

Least² Fit

Euclidean Distance

3. Programming

Python Basics

Working in Excel

R Setup R Studio

R Basics

Expressions

Variables

IBM SPSS

Rapid Miner

Vectors

Matrices

Arrays

Factors

Lists

Data Frames

Reading CSV Data

Reading Raw Data

Subsetting Data

Manipulate Data Frames

Functions

Factor Analysis

Install Pkgs

15%

4. Machine Learning

What is ML ?

30%

Numerical Var

Categorical Var

Supervised Learning

Unsupervised Learning

Concepts, Inputs & Attributes

Training & Test Data

Classifier

Prediction

Lift

Overfitting

Bias & Variance

Trees & Classification

Classification Rate

Decision Trees

Boosting

Naive Bayes Classifiers

K-Nearest Neighbor

Classification

Logistic Regression

Ranking

Linear Regression

Regression

Perceptron

Hierarchical Clustering

K-means Clustering

Clustering

Neural Networks

Sentiment Analysis

Collaborative Filtering

Tagging

50%

5. Text Mining / NLP

Corpus

Named Entity Recognition

Text Analysis

UIMA

Term Document Matrix

Term Frequency & Weight

Support Vector Machines

Association Rules

Market Based Analysis

Feature Extraction

Using Mahout

Using Weka

Using NLTK

Classify Text

Vocabulary Mapping

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

D3.js

InfoVis

IBM ManyEyes

Tableau

40%

7. Big Data

Map Reduce Fundamentals

Hadoop Components

HDFS

Data Replication Principles

Setup Hadoop (*IBM / Cloudera / HortonWorks)*

Name & Data Nodes

Job & Task Tracker

M/R Programming

Sqoop: Loading Data in HDFS

Flume, Scribe: For Unstruct Data

SQL with Pig

DWH with Hive

Scribe, Chukwa For Weblog

Using Mahout

Zookeeper Avro

Storm: Hadoop Realtime

Rhadoop, RHIPE

rmr

Cassandra

MongoDB, Neo4j

60%

8. Data Ingestion

Summary of Data Formats

Data Discovery

Data Sources & Acquisition

Data Integration

Data Fusion

Transformation & Enrichment

Data Survey

Google OpenRefine

How much Data?

Using ETL

80%

9. Data Munging

Dimensionality & Numerosity Reduction

Normalization

Data Scrubbing

Handling Missing Values

Unbiased Estimators

Binning Sparse Values

Feature Extraction

Denoising

Sampling

Stratified Sampling

Principal Component Analysis

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

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?

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

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

The End

Thank you!