ROYAL
STATISTICAL
SOCIETY

DATA | EVIDENCE | DECISIONS

Rigour not Rigidity

The Industrialisation and Professionalisation of Data Science

Dr Andrew Garrett

IFoA Data Science Summit

13th September 2017

(All views are my own)

Background

Administrative Data Research Network

Board member / Approvals Panel

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Chair, Data Science Section (previously VP Professional Affairs)

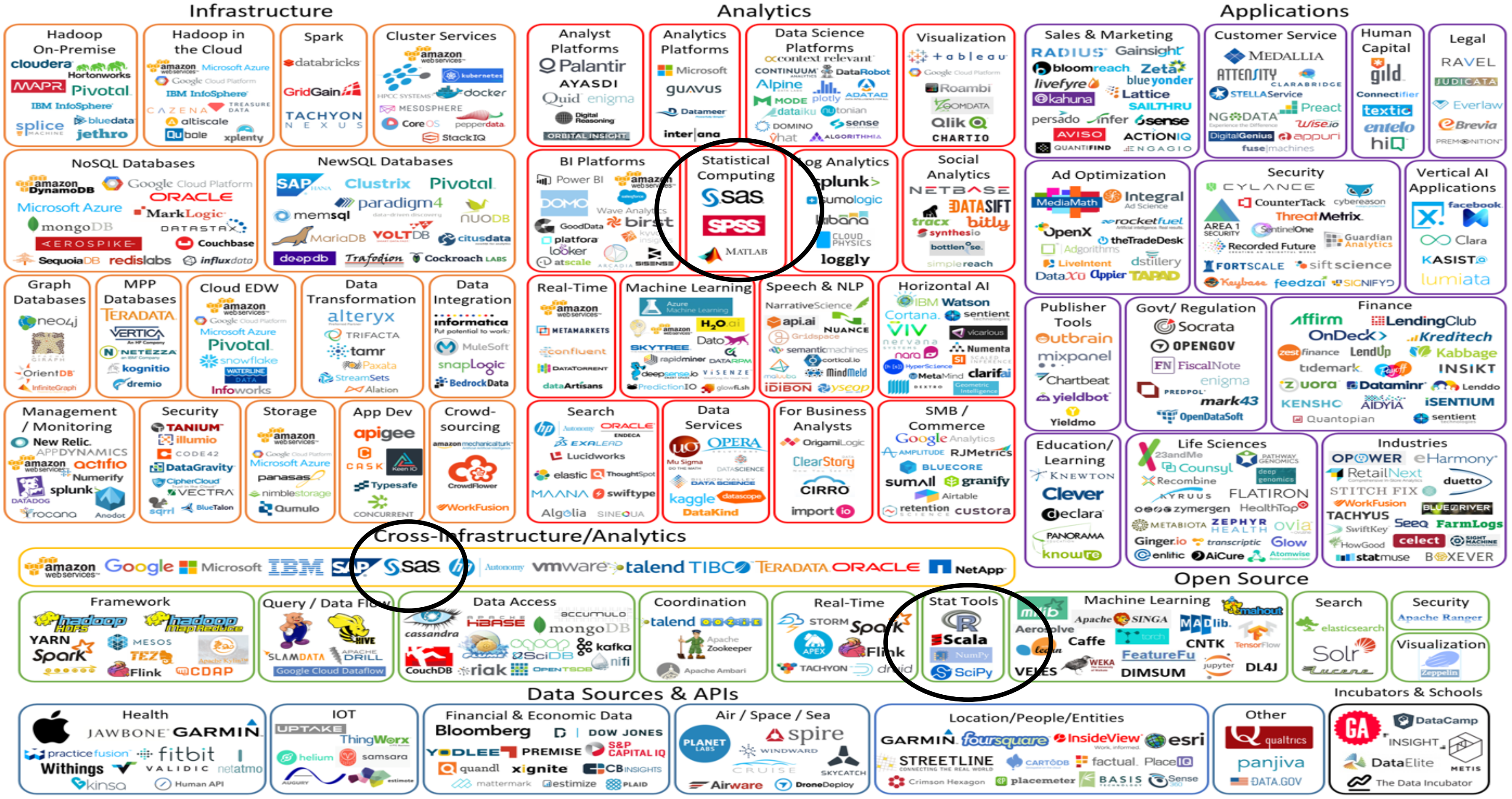
Office for National Statistics

Independent review of methodology function

ICON

Drug development – manage Biostatistics, Imaging etc.

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, memsql, NuoDB, MariaDB, VOLTDB, citusdata, deepdb, Trafodion, Cockroach Labs

Graph Databases: neo4j, Giraph, OrientDB, InfiniteGraph

MPP Databases: TERADATA, VERTICA, NETEZZA, 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, 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 mechanical turk, 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, H2O.ai, amazon web services, Dato, SKYTREE, rapidminer, DATARPM, deepsense.io, VISENZE, PredictionIO, glowfish

Speech & NLP: NarrativeScience, api.ai, NUANCE, Gridspace, semanticmachines, corticol.io, maluuba, MindMeld, iDIBON, yseop

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

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

Data Services: Mu Sigma, OPERA, DATASCIENCE, SILICON VALLEY DATA SCIENCE, kaggle, datascope, DataKind

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

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

Applications

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

Customer Service: MEDALLIA, ATTENSITY, CLARABRIDGE, STELLAService, NGDATA, Preact, DigitalGenius, Wise.io, appuri, fuse machines

Human Capital: gild, Connectifier, textio, entelo, hiQ

Legal: RAVEL, JUDICATA, Everlaw, Brevia, PREMONITION

Ad Optimization: MediaMath, Integral Ad Science, rocketfuel, OpenX, theTradeDesk, Adgorithms, Liveintent, dstillery, DataXu, Appier, TAPAD

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

Vertical AI Applications: facebook, x.ai, 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

Education/ Learning: KNEWTON, Clever, declara, PANORAMA, 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, TACHYUS, WorkFusion, BLUE RIVER, SwiftKey, Seeq, FarmLogs, HowGood, select, 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, MESOS, Spark, TEZ, Apache Kylin, Flink, CDAP

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

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

Coordination: talend, OOZIE, Apache Zookeeper, Apache Ambari

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

Stat Tools: R, Scala, NumPy, SciPy

Machine Learning: mllib, mahout, Apache SINGA, MADlib, Aerosolve, torch, Caffe, TensorFlow, CNTK, learn, WEKA, FeatureFu, jupyter, VELES, 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, esri, STREETLINE, CARTODB, factual, PlaceIQ, Crimson Hexagon, placemeter, BASIS, Sense360

Other: qualtrics, panjiva, DATA.GOV

Incubators & Schools

DataCamp, INSIGHT, DataElite, METIS, The Data Incubator

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

FIRSTMARK

Content

- RSS DSS Remit
- Traditional Industrialised Statistics
- Big Data Challenges
- Big Data Opportunities
- Professionalisation (and Regulation)
- The 12 Questions
- Our Challenge

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:

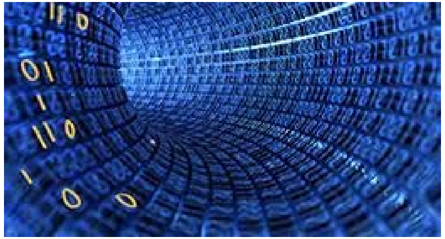
- Promoting good practice by addressing what good Data Science looks like (with exemplars) and what it does not look like.
- Promoting the statistical aspects of Data Science / re-enforcing the statistical framework
- Being a trusted voice on Data Science for employers, including inputting to consultation 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

Traditional industrialised statistics

- Data collected through a controlled process (experiment or survey)
- Drug development – Randomised Clinical Trial
  - e.g. Placebo Controlled Parallel Group Design
- Government – Surveys using Random sampling
  - e.g. British Crime Survey, International Passenger Survey
- Study population or sampling frame well defined
- Properties of “statistics” known
  - specifically in relation to bias and variability (precision)
- Focus on estimation, causation (?), decision-making (policy not patient)

Big data challenges - Bias

- Data collected as part of service provision, not to produce “statistics”
  - Electronic Health (Medical) Records
  - Government Administrative Data



- Data collected passively / automatically (IoT)
- Representation (e.g. social media)?
- Meta data are key
  - How (order) data are collected / how they are checked (consistency, missing)
  - Challenge of adding variables to legacy systems
  - Merging and appending data sets
    - Consistency across data sets / providers (incl. within Government)
    - Identifiers (de-identifiers), temporal consistency

Big data challenges – Precision

- 1/√n
- Precise in what context?
- Estimating a “constant”?
  - Data=ALL as one realisation in time
    - Super-population?
    - Prediction?

Big data challenges - Ethics

- New ethical challenges due to:
  - Digitalisation of images/text/voice
  - Multiple data sources
  - Admin data = ALL

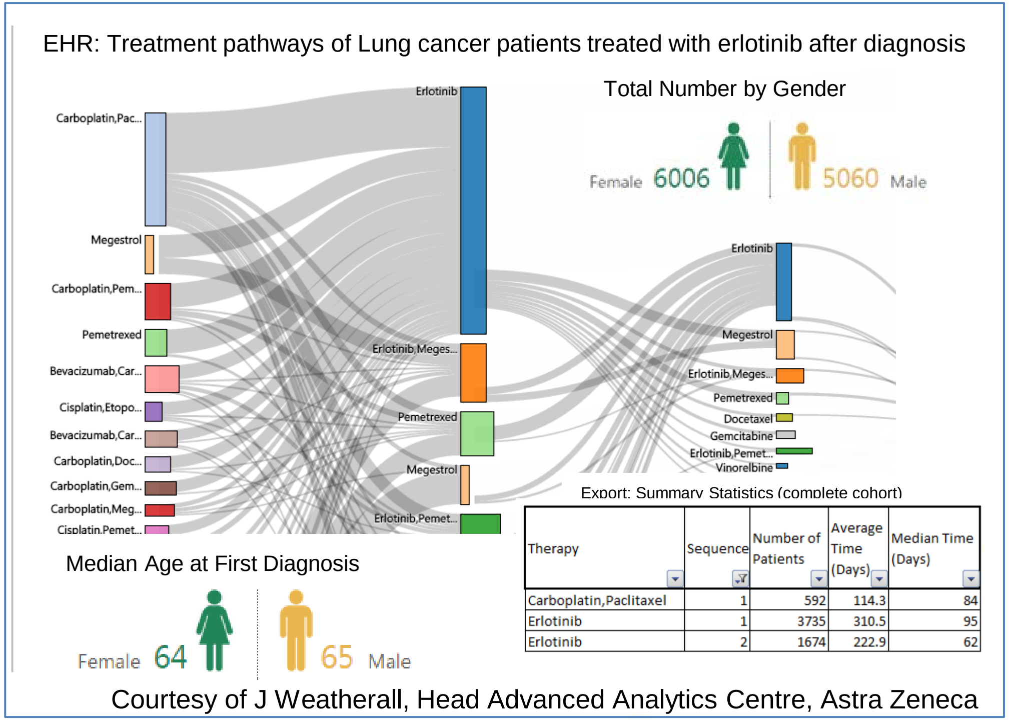
- Algorithmic transparency
  - Fairness
  - Understanding
  - Hacking and malevolence

ROYAL STATISTICAL SOCIETY

DATA | EVIDENCE | DECISIONS

Big Data Opportunities

- Disease trajectories
- Investigating co-morbidities
- Internet of Things



EHR: Treatment pathways of Lung cancer patients treated with erlotinib after 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

Courtesy of J Weatherall, Head Advanced Analytics Centre, Astra Zeneca

Big Data Opportunities

- Early estimates of economic statistics
- Lead indicators
- Private big data – useful for thinking about the consequences of new things (Bean review)
- Granularity – e.g. local inflation rates, sector specific data
- Supplementary and experimental statistics
- How to measure the digital economy

Professionalisation

- Actuaries (IFoA)
  - Certification (to practice/sign), pass exams/ gain exemptions /experience /CPD
- Statisticians (RSS)
  - Chartered Statistician – degree level (must have breadth/depth), 5 years experience, annual re-certification (previously award for life, if paid fee)
  - Code of Conduct
  - Not required to practice
  - Little traction in drug development (global), and government statisticians struggle to meet the “qualification” criterion
- Data Scientists – is something required here?

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Professionalised?

- Regulated (Drug development)
  - ICH E9 Section 1.2: actual responsibility for all statistical work… will lie with an appropriately qualified and experienced statistician… ensure that statistical principles are applied appropriately…… statistician should have a combination of education/training and experience sufficient to implement the principles….
  - In practice most will be Masters level or above
  - Many will be CStat (in UK)
- Regulated (Government)
  - National statistics are a subset of official statistics certified by the UKSA as compliant with its Code of Practice for Official Statistics

Impact of regulation

- Quality and Reliability
- Consistency
- Trust
- Processes / procedures
- Documentation
- Audit
- Pre-specification?
- Qualified people?

- Slow
- Cumbersome
- Expensive
- Blinkered
- Risk-averse / conservative
- Always done it this way
- Uncritical
- Specialised

Data Science – plus ca change?

Data Ingestion = data capture and storage

Data Wrangling = data management / preparation

- DS complain they spend too much time on it

Curation = managing data through its lifecycle

Munging = mapping data from one format to another

Parsing = processing text

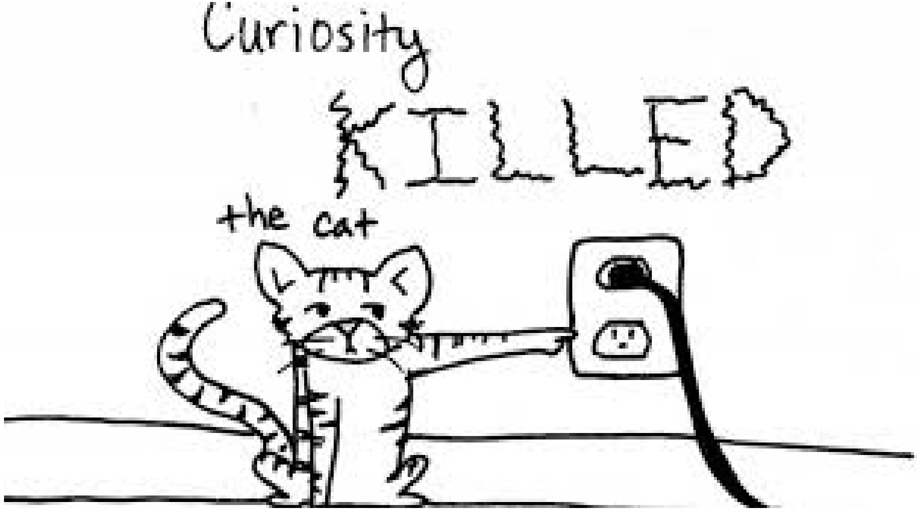
Scraping = getting unstructured data from the Web

Data Lakes = Storing data in its native format

Didn’t Biostatistics start with unicorns?

But let’s not forget…

Regulation is here for a reason – and it evolves in response to events



What might the future look like?

- Processes and procedures introduced in the face of material errors
- Data privacy and fairness concerns bring guidelines and regulation
- Data Scientists specialise and more routine work is done by sub-specialties
- Good Data Science Practice emerges

The Industrialisation and Professionalisation of DS: 12 Questions

1. **What does great DS look like?**
2. **(*) What does a good DS workflow look like?**
3. **What kind of problems can be addressed by DS?**
4. **What are the characteristics of the ideal Data Scientist?**
5. **How should an organisation start a DS function**
6. **(*) How should DS fit into the structure of an organisation**
7. **How should DS business practices change to make a success of DS**
8. **(*) What do executives and managers need to know about DS**
9. **How can an organisation build a coherent DS capability from a collection of DS projects**
10. **What career paths are available to a Data Scientist?**
11. **How can Data Scientists measure the value they create?**
12. **(*) What is a Data Scientist’s responsibility to wider society?**

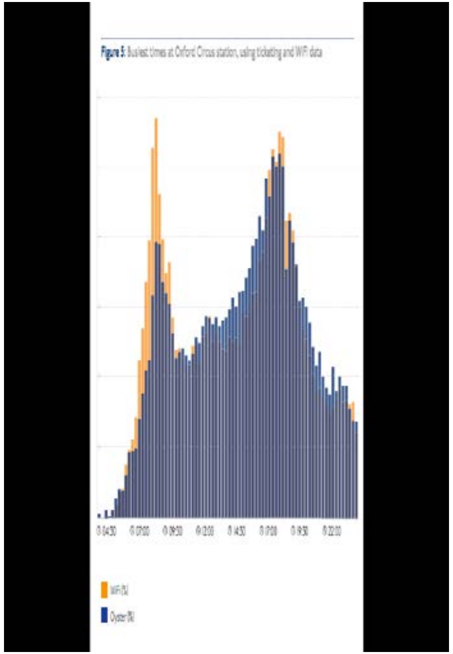
DS is great, DS is rubbish (this week)

Neural network to generate Pub names

- 1053 NE England Pub names fed in, network creates its own rules on how to make names, through iteration
    - *Mingside Arms, Castle Stan, Burn Horse Hotel…..*

TfL WiFi tracking trial (Oyster cards don’t give route)

- King’s Cross → Waterloo (32% via Oxford Circus) – investigate over-crowding in stations



Data Science – Our Challenge

- Engage and be open – DS is a different mind set to traditional “industrial” statistics
- Call out bad practices and encourage good practices – there are some really neat things being done
- Be enablers, not gatekeepers – we should be pleased with the raised profile, and we don’t have all the answers
- Bring rigour, not rigidity – statistical under-pinning is key
- You don’t have to call yourself a data scientist to do data science – but it might help