Purpose

* Results used to determine the proportion of public funding allocated to individual universities for research
* To ensure UK universities are accountable for public investment that support the researches
* Benchmarking university research performance for users such as PhD candidates

Facts

|  |  |
| --- | --- |
| REF 2014 | RAE 2008 |
| * 6 years * 36 units of assessment (subject areas) * 154 universities * 52061 academic staff * 191150 outputs submitted * 76% researches in UK are “internationally excellent” or “world-leading” * 2 billion pounds allocated to universities based on results | * 215507 outputs |

Stages

1. Submission
2. Assessment
3. Publication

REF 2014

Assessment criteria

1. **Output**

* Simplest
* Measure quality of academic work
* Up to 4 research outputs can be nominated for each academic
* Publications: journal articles, book-length studies
* Fruits of research: datasets, new technologies, IP
* Output score for specific uni at specific subject area -
  + how successful its academics are at generating high-quality publications
  + Identify cutting-edge research projects
  + Academics are recognised leaders in the field

1. **Impact**

* New criterion for 2014
* Assess the positive effects of a university’s research beyond the academy
* Assessed using submitted case studies
  + Demonstrate past research effects
  + Strategies for ensuring present and future impact
* Definition:
  + *‘any effect on, change or benefit to the* ***economy, society, culture, public policy or services, health, the environment, or quality of life****, beyond academia’*
* E.g.
  + Medical science research - generate changes to public health policy
  + Arts and humanities - educational outreach, exhibitions in public libraries and galleries
* Impact score for specific uni at specific subject area -
  + Identify high-profile projects and activities outside uni
  + Benefit to society

1. **Environment**

* Most important for prospective research students
* Measure quality of departments, academic units and research groups in universities
  + The environment in which PhD students work
* Assessed based on evidence demonstrating the sustainability and vitality of research environment
  + Continuity of research funding
  + Structures for effective support, supervision and training

1. **Overall**

* Output - 65%
* Impact - 20%
* Environment - 15%

Expert panels

1. Practising Researchers

* Other academics working in a field appropriate to their assigned unit of assessment
* Peer-review

1. Research Users

* Selected from the audience
* Academics using research data
* Representatives of industry, business or policy groups

Data and analysis

**Units of Assessment UOA**

Summary of each UOA: <http://www.ref.ac.uk/media/ref/results/AverageProfile_All%20UOAs.pdf>

**Analysis results by REF official**

<http://www.ref.ac.uk/results/analysis/>

Comparative data

* Average across all universities
  + By UOAs (subject areas)
  + By main panels (faculty)
* Average across the entire UK

**REF impact analysis**

<http://www.hefce.ac.uk/pubs/rereports/Year/2015/analysisREFimpact/>

* 6679 impact case studies submitted to REF 2014
* outline changes and benefits to the economy, society, culture, policy, health, the  environment and quality of life — both within the UK and overseas.
* undertaken by Digital Science, a division of Macmillan Science & Education; working in conjunction with its sister company Nature Publishing Group and the policy institute at King’s College, London
* co-funded by the UK higher education funding bodies, Research Councils UK and Wellcome Trust
* Text mining and qualitative analysis
* Identify general patterns and thematic structures
* Synthetic analysis

Impact case studies data

<http://impact.ref.ac.uk/CaseStudies/>

Pure text documents

Individual pdf links for download

Underlying data for impact topics

* Which case study corresponds to which impact topic
* 0 or 1 matrix

REF impact analysis report available

**Citation data - Scopus**

<http://www.ref.ac.uk/about/guidance/citationdata/>

RAE 2008

Results:

* Outputs less high quality
  + 14% world-leading (vs 22% for REF)
  + 37% internationally excellent (vs 50% for REF)
* Meaning that outputs quality has improved over years
  + 4\* outputs increased by 42%, 3\* by 24%
  + Top 1% world’s most highly cited papers increased by 44%, top 5% by 31%, top 10% by 29%
  + Aligned with independent evidence of the enhanced international standing of UK research

Assessment criteria:

* Impact not assessed
* Significant difference in how environment is assessed - two elements are not comparable

Mapping of UOAs:

<http://www.ref.ac.uk/media/ref/results/Mapping%20of%20UOAs%20across%20RAE%202008%20and%20REF%202014.pdf>

* 67 vs 36 UOAs

Extra sources of data

* Quality of journals
  + Web of science
  + Google Metrics
  + Core conference
* Number of citations
  + Scopus - different papers but similar trajectory of improved research performance

?? need project selection presentation slides

Questions to ask:

**OUTPUTS**

1. What type of outputs have been submitted?
   1. Journal articles; Conference contributions; Books and book chapters; Physical artifacts; Exhibitions and performances; Digital artifacts (including web content)
2. To what extent are the outputs representing **originality, significance and rigour**?
   1. New insights
   2. Contribute significantly to understanding the subject of matter
   3. Using an efficient approach
3. How do outputs represent the above qualities through **methodology, approach, research topics and inter-disciplinarity**?
   1. Research topic, inter-disciplinarity
4. Is there any time delay between the publication of outputs and the academic utilisation of outputs e.g. paper being cited by others, methodology being further developed - contribute to further research advancement? Does the time lags affect the output scores?
   1. For example, publications on tribology are dated back to 10 years ago due to limited numbers of research groups and institutions involved in UK
5. What quantitative data can be extracted from output submissions?
   1. From output summaries
      1. Do output summaries contain quantitative information on **originality, significance and rigour** that can be developed into metrics of research output?
      2. Do output summaries contain information on **methodologies, approaches, research topics and inter-disciplinarity** that can be developed into metrics?
   2. From original publications - not recommended
   3. To quantify **originality, significance and rigour**
6. What quantitative measure can be compared against to evaluate our prediction?
   1. Final REF published score data
7. Does the number/significance of research users - number of citations - affect research output? To what extent?
8. Does research collaboration affect the quality of outputs?
9. Is there any policy/government strategy/research council funding e.g. horizon 2020 encouraging the development in any UOA in particular? To what extent do they affect the output score?
10. Any connection between different UOAs and to what extent interrelated UOAs affect their output scores?

**IMPACT**

Detailed analysis available.

Mainly about text mining and synthetic analysis.

**ENVIRONMENT**

Tbc

PPT - structure

Intro

    Aim of project

        To develop a model to predict REF scores

            Depend on the metrics - toy experiment

            Right way for thinking

            The journey

Literature review

    Impact analysis by KCL

        Upload on github

    Text processing

        tf.idf

    Machine learning

        For the specific purpose

        Prediction model

        Component we need

        Starting model

    Triangular flow?

Methodology

    Identifying parameters that contributes to final results

    Assumptions

        Keywords in collection

        Individual output is not available

    Iterate through the papers for data extraction

        Additional infor

Approach

**Features related to the paper rather than the actual content**

**Questions**

Find the strongest department in the uni

    Find unis with most 4\*

    Find most mentioned words/phrases - tf.idf - additional informations

    Additional infor - related on paper content, author, institution

        Think about dependency

            Machine learning assumes independence but knowing about dependency

        Going down the list, starting from high ranking unis

        But including all papers

        Machine learning hidden variables - latent variables

    Simple statistics on data at the beginning before detailed analysis

        First guess

        Rough properties

        Understanding the data

    Fraction of journal vs conference papers

        Set manual boundaries - binning

        Machine learning approach to choose bins for people

        Mutual validation

    Then identify possible parameters - choose in model that could affect the scores

    Test features

    Have a list of features with good/not-so-good correlations in-between features not always with final scores

        Dependence of variables is given by correlations

Timeline

    Week 4 - 6:

    Data extraction

1. Harvest webpage

        Wget - downloading trees of webpages

        Homebrew - mac

        Cygwin - win

1. Extract infor from webpages

    Python packeges

        One off manual process is possible but documentation

        KEEP RECORD OF DATA PROCESS!

    Data cleaning and management

    Week 7 -11:

    Data analysis

    Model development and testing

    Report writing - conclusion and discussions

        Based on individual work records

Prepare slides

Decide 10 UOAs

    CS must be included

    Choose the ones with more data

    Keep similar structure of information/data

    Different parameters - slightly changes in models

    Similar subjects - less diversity in subject - more homogeneous

    Sample from diff disciplines evenly to learn about diversity

        How does it work across disciplines

    RAE vs REF

        Correlation analysis should be similar

    Learning parameters are only perfect on one dataset

        But **generalisation** required

            Across fields

            Across years

        Models need to be tested on an independent dataset

        Cross validation - rotating testing

            Small element for prediction and many tests

        Overfitting is a big problem

        Model should be able to predict future data

            Time effect/trends

FTE

    Multiplied with quality

    How many full time equivalent staff

Ask github

Ask whether to put more effort in output since it takes up the largest %