Education for a fast-changing world: Conceptions of Statistical Literacy and Data Science

It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity, it was the season of light, it was the season of darkness, it was the spring of hope, it was the winter of despair.

Charles Dickens (1859)

jim.ridgway@durham.ac.uk

Structure

- Living in interesting times
- Literacy for the data and toys we encounter
- The evidence ecosystem
- Data Science for interesting times
- The needs of citizens
- Reflections on contemporary science
- On Models and Modelling
- Curriculum implications
- Conclusions

Education for a fast-changing world:

- *Government, without popular information, or the means of acquiring it, is but a Prologue to a Farce or a Tragedy; or, perhaps, both* and *…the advancement and diffusion of knowledge* is *the only Guardian of true liberty*
- James Madison (1825)

- *Data are the lifeblood of decision-making and the raw material for accountability. Without high-quality data providing the right information on the right things at the right time; designing, monitoring and evaluating effective policies becomes almost impossible*
- A World that Counts – mobilizing the data revolution for sustainable development (2014, p2) United Nations

jim.ridgway@durham.ac.uk

Data Citizens Encounter vs School Curriculum experiences

Data include images, locations, actions…

Dynamic data, spanning times and locations are common

Phenomena are often multivariate

Non-linear relationships are common

Decisions have been made about measures and operationalization

Novel data sources, data collection methods, and analysis techniques are common

Innovative visualisations have been invented

jim.ridgway@durham.ac.uk

Data Citizens Encounter

Statistics are often embedded in rich texts

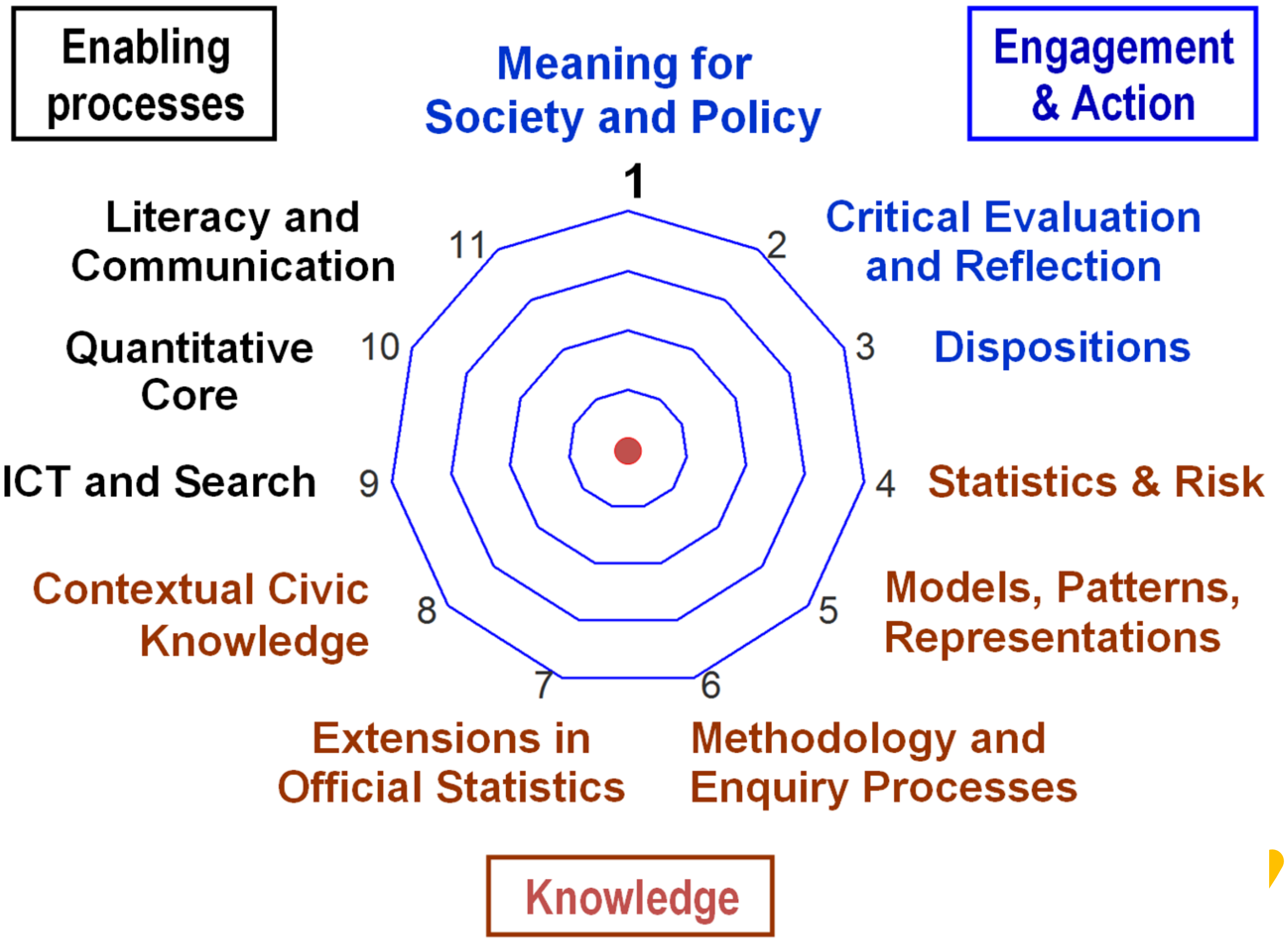
Indicator systems are common

Data are often aggregated

What are we doing it all FOR?

- to establish Causality?
- Conclusions are? Implications are? Consequences are?

Targets for Statistical Literacy



jim.ridgway@durham.ac.uk

What IS Data Science?

- Anything that involves using evidence to inform decisions and actions

Beyond turf wars…

- American Statistical Association *About* page reads *BIGTENT statistics+data science*

- Intellectually disruptive; it raises questions about discovery methods, explanation, and ways to make decisions

The Evidence Ecosystem

-features of ecosystems

Ecosystem Feature	Natural Ecosystem Examples	Evidence Ecosystem Examples
sub-systems need to be understood in their own right	coral reefs, prairies	political systems, the internet
sub-systems are connected	deforestation influences global warming which influences coral reefs	criminal activity can create wealth which can be used for political influence
food chains	rabbits eat grass, foxes eat rabbits	National Statistics Offices gather data, governments use NSO
symbiotic relationships	cattle and egrets	politicians and media
subject to rapid dramatic change	introduction of rats to islands	the internet
pollution	plastics at sea	conspiracy theories
concentration levels can be critical	animal waste as fertiliser or toxin	medical advice can be precise or an 'infodemic'

jim.ridgway@durham.ac.uk

Elements in the evidence ecosystem

Government

Politicians

Video and Print Media

Social Media

Cyber Warriors

Military; private companies; people's intelligence agencies (*Bellingcat*)

Survey Data Collectors and Distributors

NSOs; NGOs – *OECD etc*

Sensor Data Collectors and Distributors

NASA

Elements in the evidence ecosystem

Unaccountable Data Collectors and Distributors

Amazon; Facebook…

Knowledge Creators and Curators

Universities; Curators (*Cochrane collaboration*)

Data Scientists

Advocacy Agencies

UN, Amnesty…

Fact Checkers

Honest (*Fullfact, Politifact*) and Dishonest (*factcheckUK*)

Investigative Journalists

Educators

The Toy Shop

Data…
Analysis…
Theory…
Action…
Ethics…

- Google, Amazon, Facebook, Skype
- recognition of individuals via face, fingerprint, voice, gait, patterns of key presses
- tracking (via fitness trackers, credit card use, data from transport networks)
- speech recognition and language translation
- detection of disease outbreaks via analysis of google search data
- the Internet of Things – smart refrigerators, TVs, cars, and domestic robots
- ‘deep fake’ videos (creating and detecting)
- predicting crime and recommending custodial sentences
- satnav; autonomous vehicles and weapons systems
- mapping dwellings from aerial images, in remote settings
- emotion detectors for classrooms and cars
- Virtual, augmented, and mixed reality devices
- Gene editing
- Brain-driven prosthetics and computer-brain interfaces
- Generating images from voice input

Data Science for Interesting Times

- Denser and denser evidence ecosystems
- Disruptive toys/models continue to be invented
- More and different data and access to data (open data, text, images, sensor data…)
- New creators of knowledge (notably technology companies)
- New distributors, consumers and users of knowledge
- More audiences (social media)
- A more connected world
- A more rapidly changing world
- Global willingness to address ‘wicked’ problems - environment, pandemics, migration
    - e.g. UN Sustainable Development Goals

Have we got the ‘science’ we need?

Disruptive Socio-technical Systems

- Technologies are not *inherently* disruptive – disruption depends on how something is *used*
- Impacts can be unpredictable
  - *Agricultural revolution; Printing press; transport systems; Covid-remote working-education;…*
- Technologies are not neutral
  - *Cars; computers*

Data science creating DST – curriculum implications?

Students need to beaware of existing and potential DST, for their personal welfare, as *citizens* who have a voice, and as *users of tools* i.e. as:

- *spectators* – which simply requires an understanding of what is going on around
- *referees* – where students are empowered to make reasoned political judgements about what *should happen* where the products of data science are likely to have social impact
- *players* – where students are empowered to act, e.g. creating and analysing evidence using data science tools

jim.ridgway@durham.ac.uk

Spectators

- Need to know that disruptive technologies will continue to be invented
- Can be impossible to un-invent
- Are often adopted on a very large scale, before human systems have had time to adapt – as in *DST*
- Need to be aware of the range of applications that have been created, and the promises and potential perils of their use
- Need to be able to evaluate the benefits and potential costs of engaging with different sorts of technology
- Need to be aware of the politics of technology: technologies are never neutral (e.g. cars; the internet)
- Need to be able to identify groups who might benefit from, or be disadvantaged by, emerging technologies
- Need to engage with moral issues (e.g. the dangers of the Panopticon)
- Ability to explore possible unintended consequences (e.g. cyberbullying via social media) via 'what if' games

Referees

- Need to consider the intended and potentially unintended consequences of new technologies, and to be able to identify (and make value judgements about) groups that will benefit from, and be disadvantaged by, disruptive technologies
  - E.g. face recognition; track-and-trace; vulnerable GPS; automated decision tools (including multiple regression!)
- Need a critical awareness of developments, and a *disposition* to speculate about the impact that emerging technologies might have on society at local, national and global levels
- Should engage in political processes to shape the ways that new technologies are used (in as far as this is possible)

Referees and story-telling about the world

Need to understand:

- the characteristics of different sorts of models, the limits of modelling, and the principles of model validation
- The principles underpinning different techniques (e.g. neural nets) and a willingness to explore emerging techniques
- That models can be unstable for a variety of reasons (e.g. insufficient data to produce robust parameters; sample bias)
- That systems can be unstable for a variety of reasons (e.g. the world can change)

Players

Players need the skills associated with Box's idea of *being a good scientist*. This requires considerable sophistication in approaches to understanding and interacting with the world

Components include:

- Modelling skills (conceived broadly), and in particular experiences of using a wide variety of tools
- Wrestling with epistemological issues: the nature of knowledge as conceived in different academic disciplines - how it is created, shared, learned, and used (and by whom, and for what purposes)
- An awareness of the nature of DST, and the motivation to maximise the benefits of emerging technologies for humankind
- Engagement with citizen science

Citizen scientists are *players*; student citizen scientists may become future data scientists, as they engage in developing research questions, designing methods, gathering and analysing data, and communicating results

Reflections on Contemporary Science

jim.ridgway@durham.ac.uk

Systemic Problems with Science - AMPLIFIED by Data Science

Flunking Statistics 101

Sample bias 1: men=people

- Car safety dummies are 'male' – women are more likely to die in comparable crashes
- Drug trials often exclude women

Sample bias 2: Caucasians=humans

- Disease-genome matching
- Medical procedures
  - Coronary screening

- **Replications**
  - Open Science Collaboration (2015) failed to replicate 60/100 'well-known' results in psychology
  - Collins (2015) 0/70 replications of drug trials associated with Lou Gehrig's disease
- **Data dredging**
  - Ioannidis (2005) on *why most published research findings are false*

jim.ridgway@durham.ac.uk

Systemic Problems with Science - AMPLIFIED by Data Science

Flunking Statistics 101

Sample bias 1: men=people

- Car safety dummies are 'male' – women are more likely to die in comparable crashes
- Drug tests often exclude women

Sample bias 2: Caucasians=humans

- Disease-genome matching
- Medical procedures
  - Coronary screening

- **Replications**
  - Open Science Collaboration (2015) failed to replicate 60/100 'well-known' results in psychology
  - Collins (2015) 0/70 replications of drug trials associated with Lou Gehrig's disease
- **Data dredging**
  - Ioannidis (2005) on *why most published research findings are false*

jim.ridgway@durham.ac.uk

Systemic Problems with Science - **AMPLIFIED by Data Science**

Flunking Statistics 101

Sample bias 1: men=people

- Car safety dummies are 'male' – women are more likely to die in comparable crashes
- Drug tests often exclude women

Sample bias 2: Caucasians=humans

  - Disease-genome matching
  - Medical procedures
    - Coronary screening

- **Replications**
  - Open Science Collaboration (2015) failed to replicate 60/100 'well-known' results in psychology
  - Collins (2015) 0/70 replications of drug trials associated with Lou Gehrig's disease
- **Data dredging**
  - Ioannidis (2005) on *why most published research findings are false*

jim.ridgway@durham.ac.uk

Systemic Problems with Science - AMPLIFIED by Data Science

Flunking Statistics 101

Sample bias 1: men=people

- Car safety dummies are ‘male’ – women are more likely to die in comparable crashes
- Drug tests often exclude women

Sample bias 2: Caucasians=humans

  - Disease-genome matching
  - Medical procedures
    - Coronary screening

- **Replications**
  - Open Science Collaboration (2015) failed to replicate 60/100 ‘well-known’ results in psychology
  - Collins (2015) 0/70 replications of drug trials associated with Lou Gehrig’s disease
- **Data dredging**
  - Ioannidis (2005) on *why most published research findings are false*

jim.ridgway@durham.ac.uk

Systemic Problems with Science

Misuse of statistical analyses

p-values

jim.ridgway@durham.ac.uk

Systemic Problems with Science

Poor academic practices

- Publish or perish
- Fraud
- Suppressing results
- Hidden funding
- Open access favours rich countries
- Predatory journals

Epistemological Issues

What counts as evidence (even within disciplines)

qual vs quant divides

'Capture and kill' vs 3d photography

Decolonising knowledge and theory (and inclusive research)

Anthropology of global data (e.g. health; SDG)

Analysing observational data

Deconstructing images and text

jim.ridgway@durham.ac.uk

Representing and understanding the world

-explanation - prediction

All models are wrong, but some are useful

(Box & Draper, 1987, p424)

There are no routine statistical questions; only questionable statistical routines

(Cox, quoted in Chatfield, 1991, p240)

Modelling with Arithmetic

30 musicians take 30 minutes to play Beethoven's 3rd Symphony.

How long would 60 musicians take?

Modelling is about – well, modelling

Mapping Models to Situations

- Anderson’s (2008) end of theory
- Linear additive (school physics)
- Systemic (school biology)
- Macrosystemic – stable (butterfly life cycle)
- Macrosystemic – unstable (implications for humans of climate change)

The end of theory: the data deluge makes the scientific method obsolete
Anderson (2008)

Out with every theory of human behavior, from linguistics to sociology

Forget taxonomy, ontology, and psychology

Who knows why people do what they do?

The point is they do it, and we can track and measure it with unprecedented fidelity

With enough data, the numbers speak for themselves

Beyond Nowcasting…

Kodak: 1976 sold 85% of all film cameras and 90% of all film sold in the US

bankrupt in 2012

AND the digital camera was invented by a Kodak engineer

Manufacturers of mainframe computers (e.g. Control Data Corporation, IBM) missed the emergence of personal computers

Modelling global warming

Challenges associated with AI models

- What you put in determines what you get out
  - Picture recognition
    - African-Americans classified as gorillas (Google)
  - Voice control
    - Drive.ai ignored women
  - Predicting recidivism
    - Afro-Americans receive harsher sentences (wrongly)
  - Genome-disease associations (80% of data from Caucasians)
    - Risks of mis-matches with other ethnic groups

Challenges associated with simple Analytic Models

What you put in determines what you get out

Predicting heart health risks

2009 - Framlingham risk equation is the *model of choice*

2010 - *one of the possible equations to use* (BMA 2014)

Obesity - not measured in the original sample

Generalisation of models across ethnic groups?

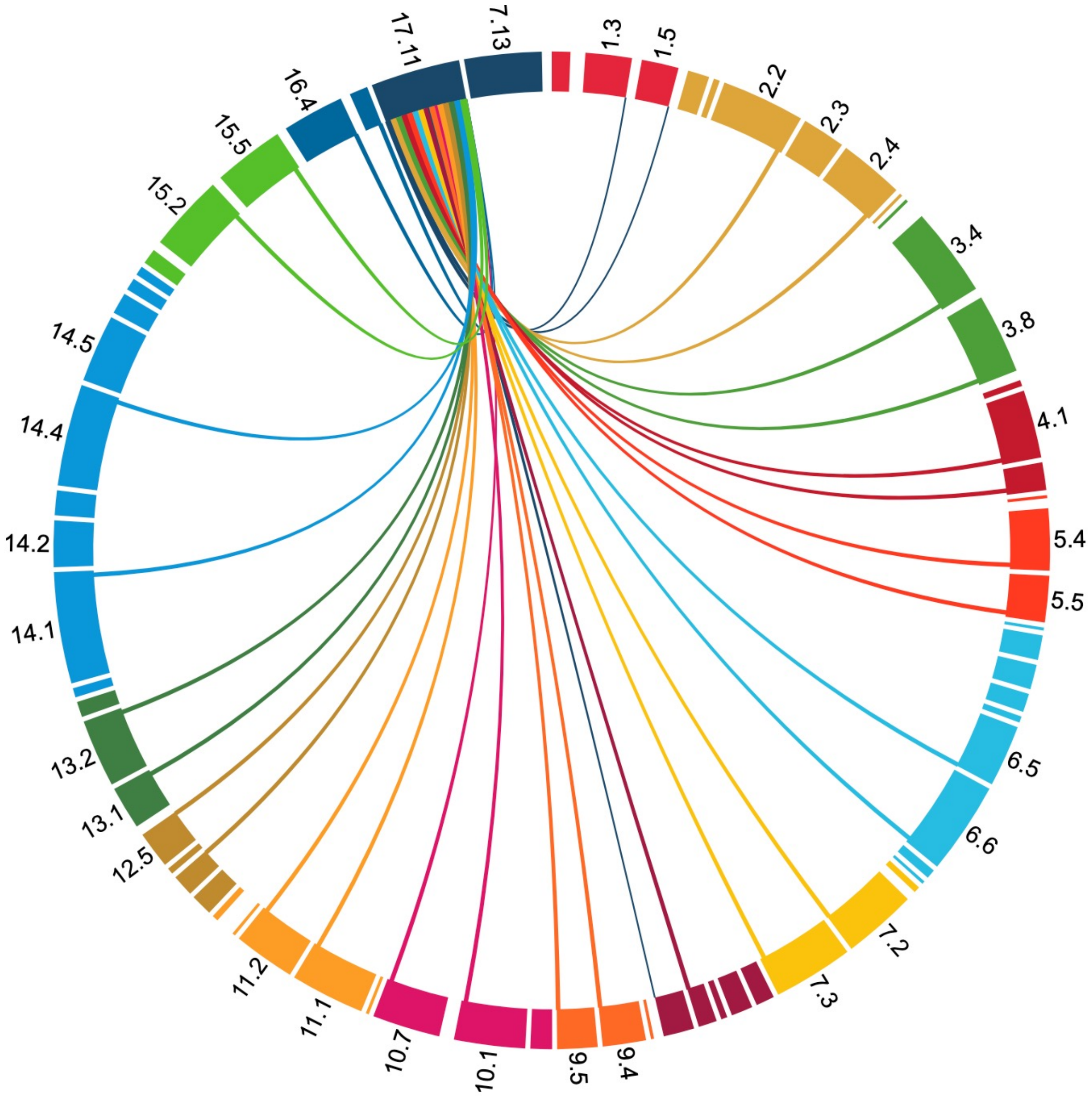
Afro-Americans: mis-diagnosing heart disease; undiagnosed diabetes

jim.ridgway@durham.ac.uk

Problems associated with modelling complex systems SDGs

jim.ridgway@durham.ac.uk

Interlinkages Visualisation: Trade-offs



jim.ridgway@durham.ac.uk

Modelling 101

- Always start off by looking at ‘what is’ – critically
- Always model a situation using at least 2 different representations
- Always model a situation using multiple sources of information
- Explain your methods and assumptions
- Explore the effects of parameter changes
- Calibrate on data sets that are as different as possible
- Review, remodel, recalibrate; go back 3 steps
- Before you advocate a change, do the ‘system-gaming walkthrough’

jim.ridgway@durham.ac.uk

Curriculum Implications Under-represented topics

- Engage with novelty
  - Use Feeds from Scientific American, Nature, NASA, MIT Technology Review, Wired…
- Modelling
  - Exposure to different classes of model
  - *Learning from mistakes*
- Measurement
  - How do you measure anything?
  - Theories implicit in measures (e.g. GDP)
  - Properties of measures (e.g. poverty)
  - Corruptibility of measures
- Data exploration
  - Sampling
  - Data visualisation and caveats (Tufte; FT *Chart Doctor*)

Curriculum Implications Under-represented topics

- Analytic tools
  - Neural nets
  - Classification and Regression Trees
  - Data mining
  - Meta-analysis
- Decision making
  - Argumentation and rhetoric
  - ‘what if’ scenarios

Curriculum Implications Over-represented topics

- Linear models
- The Normal distribution
  - Statistical significance testing (irrelevant for population data)

Conclusions

- We are living in interestingly disruptive times
- Always remember fundamental statistical principles
- Look critically at existing ‘knowledge’
- Engage with novelty
- Understand and explore a variety of models, and be aware of new modellers/styles
- Be eclectic; Synthesise
- Address ETHICS
  - What IS, what COULD BE, what SHOULD BE

jim.ridgway@durham.ac.uk

References

Teaching Data Science and Statistics (eds. MacGillivray, Gould, Ridgway) Special Edition of Teaching Statistics (vol 43, Summer 2021)

Forthcoming: **Statistics for Empowerment and Social Engagement: teaching Civic Statistics for informed citizenship**. (ed. Ridgway) Springer.

2 Back to the Future – Rethinking the Purpose and Nature of Statistics Education

Joachim Engel and Jim Ridgway

3 A Conceptual Framework for Civic Statistics and its Educational Applications

Iddo Gal, James Nicholson, and Jim Ridgway

5 Interactive Data Visualizations for Teaching Civic Statistics

Jim Ridgway, Pedro Campos, James Nicholson, and Sónia Teixeira

22 Data Science, Statistics, and Civic Statistics – education for a fast changing world - Jim Ridgway, Pedro Campos and Rolf Biehler

23 Civic Statistics in Context: Mapping the Global Evidence Ecosystem

Jim Ridgway and Rosie Ridgway

The Epistemological Engine

Jim's on-going research

Use AI (and whatever is to hand) to look at the creation and construction of knowledge

- What are the current problems?
- Can we do better?
- What are the new affordances?
- How do we capitalise on them?

The Epistemological Engine - Under the Hood

Use and develop existing tools

- *Analysing published literature*
- *Automating literature searches, and the conduct of meta-analyses*
- Summarising articles
- Develop tools for critiquing individual papers - use https://casp-uk.net/casp-tools-checklists/
- Extracting 'facts' from text and video
- Semantic analysis of video
- Generating articles

jim.ridgway@durham.ac.uk

The Epistemological Engine - Under the Hood

Develop tools to critique academic practices

- Identify results important for theoretical claims, where the evidence base is weak and should be replicated
- Identify academic areas where replication is not valued
- Identify academic areas where there is insufficient sharing of data, code and workflows
- Identify academic areas that are paradigm-bound (i.e. use few types of model, restricted classes of data, and a small set of analytic tools)
- analyse large corpora of research in to examine the epistemological assumptions made

Critique discovery processes

- *Creating semantic nets of academic papers over time in terms of both content and authorship in order to explore the flow of discovery processes*

Promote discovery processes

- Create analogy generators, to suggest developments in fields other that the one in which a method or tool was developed

Where the Rubber hits the Road – EE by observing People adressing 'Wicked'problems data? analysis? action?

Researching Wicked Problem Research

Analyse people working with multiple sources of messy data, a variety of analytical tools, and a variety of world views

e.g. **Climate Change**

Individual level (e.g. car choice and use)

Local structures (e.g. support for recycling)

National structures (e.g. energy generation policy; policies on house insulation and domestic solar power)

International initiatives (e.g. consensus on restricting carbon emissions)

jim.ridgway@durham.ac.uk

The Epistemological Engine - Under the Hood

Evaluating Rival Models

- Use Common Task Frameworks (Mark Lieberman) to judge success on performance
  - E.g. CTF-SP website
  - Publish training data and have testing data available
  - Define evaluation metrics
  - Compare rival performances
- Creating data repositories
  - e.g. https://registry.opendata.aws/
- Sharing code and workflows to compare different modelling assumptions and analytic techniques
  - e.g. https://github.com/

Notes for Epistemological Engineers

- Be aware of epistemological issues: the nature of knowledge as conceived in different academic disciplines - how it is created, shared, learned, and used (and by whom, and for what purposes)
- Be aware of the epistemological assumptions embedded in Western science – engage with 'Epistemologies of the South'(*apols to Geographers*)
- Be aware of the politics of science/technology: technologies are never neutral
- Explore the reasons for the existence of data sets – all measures are theory-heavy
- Understand modelling and the limits of modelling, and the principles of model validation;
- Create a conceptual web to link between seemingly different methods
- Engage with moral issues (e.g. the dangers of the Panopticon)

Competences for students of Epistemological Engineering

Technical stuff

- Learn to represent the same problem in a variety of ways
- Invent and modify data visualisations
- Understand the principles underpinning different techniques (e.g. neural nets)
- Become fluent in the use of major data repositories
- Share code and workflows

jim.ridgway@durham.ac.uk