NCRM

COMPUTATIONAL SOCIAL SCIENCE
PROGRAMMING SKILLS FOR SOCIAL SCIENTISTS

Dr. James Allen-Robertson - University of Essex

jallenh@essex.ac.uk
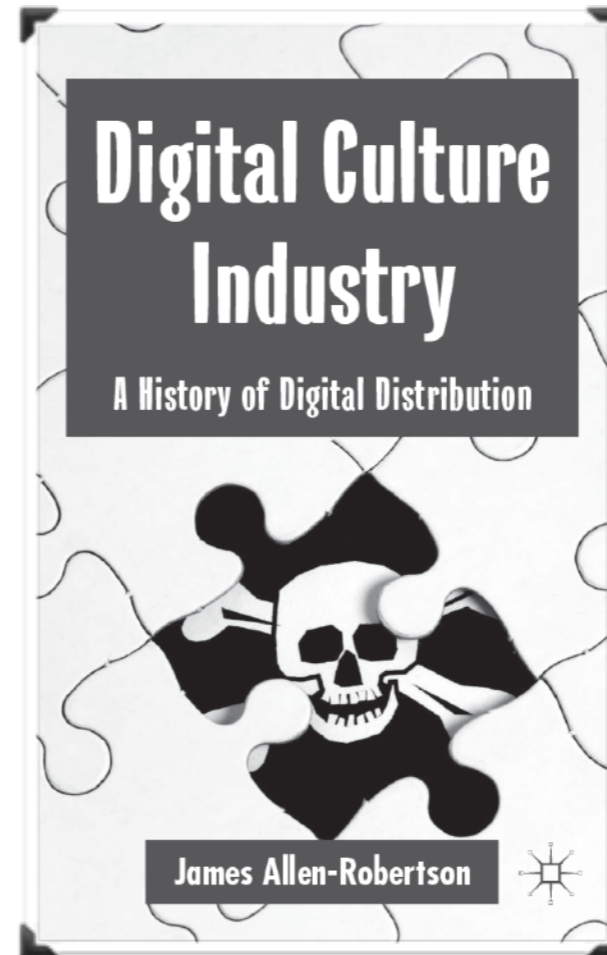
@Minyall   github.com/Minyall

PERSONAL TRAJECTORY

Sociology BA, MA and PhD

▸ Cyberculture

▸ Piracy & Copyright

▸ Utilising digital documents for <u>qualitative</u> research

Digital Culture Industry
A History of Digital Distribution
James Allen-Robertson

Materiality Media Theorist

▸ Design in media systems

▸ User agency

▸ Memory

▸ Algorithmic Power Relations

Allen-Robertson, James. 2017. 'Critically assessing digital documents: materiality and the interpretative role of software', *Information, Communication & Society*: 1-15.

Allen-Robertson, J. 2017. 'The materiality of digital media: The hard disk drive, phonograph, magnetic tape and optical media in technical close-up', *New Media & Society*, 19: 455-70.

PERSONAL TRAJECTORY

Sociology BA, MA and PhD

▸ Cyberculture

▸ Piracy & Copyright

▸ Utilising digital documents for <u>qualitative</u> research

Materiality Media Theorist

▸ Design in media systems

▸ User agency

▸ Memory

▸ Algorithmic Power Relations

Computational Social Science

▸ Building own tools for data analysis.

▸ Automated collection of data at scale.

▸ Focus particularly on text data retaining my qualitative roots.

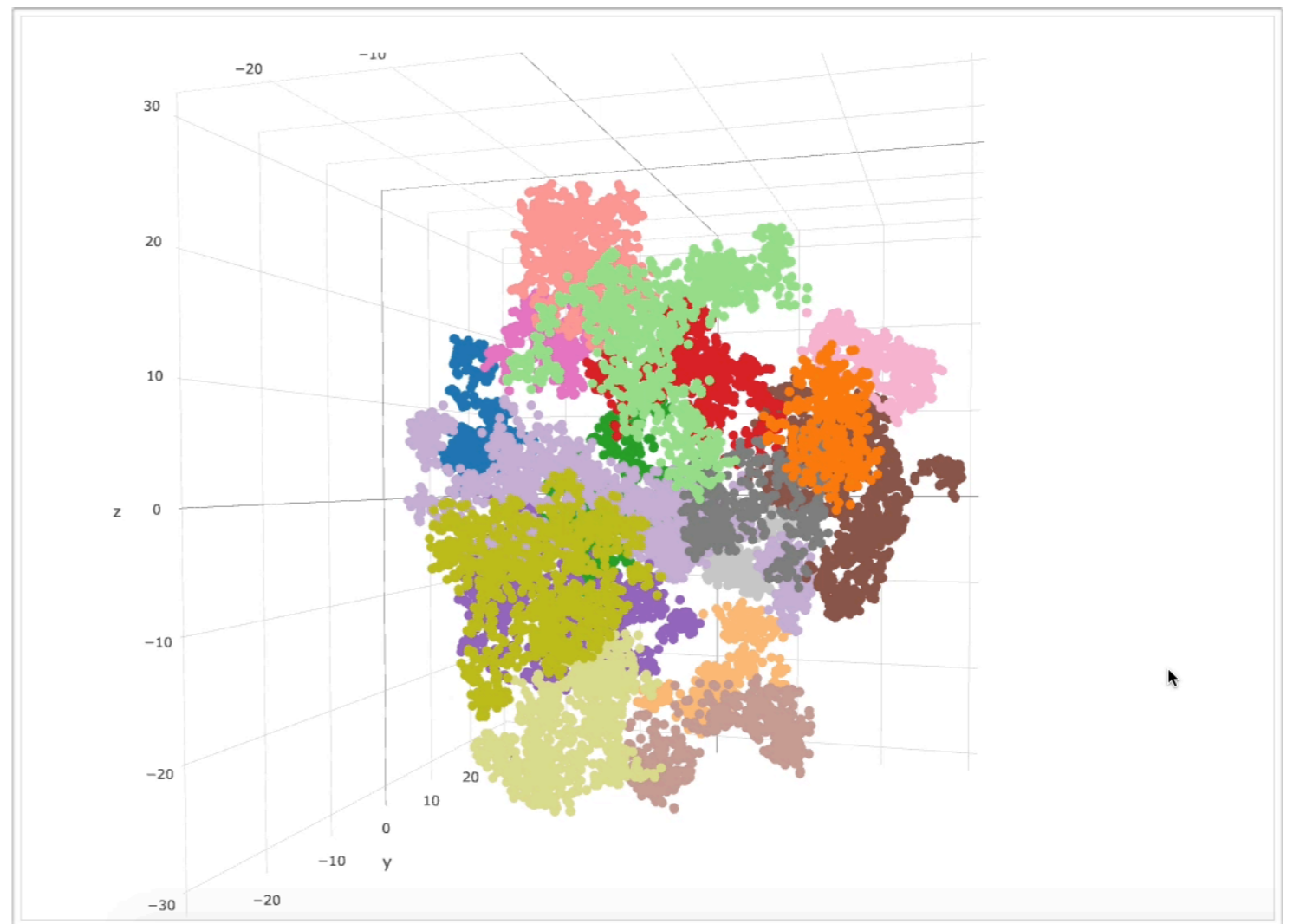▸ Much more collaborative work providing skills and tools to colleagues.

TODAY

COMPUTATIONAL SOCIAL SCIENCE EXAMPLES
LARGE SCALE DOCUMENT CLUSTERING

▸ Uber Driver Discourse

    ▸ 50K threads from Uber driver and Taxi driver forums.

    ▸ Document Similarity + Dimensionality Reduction + Clustering

▸ What are the concerns of workers under algorithmic management?

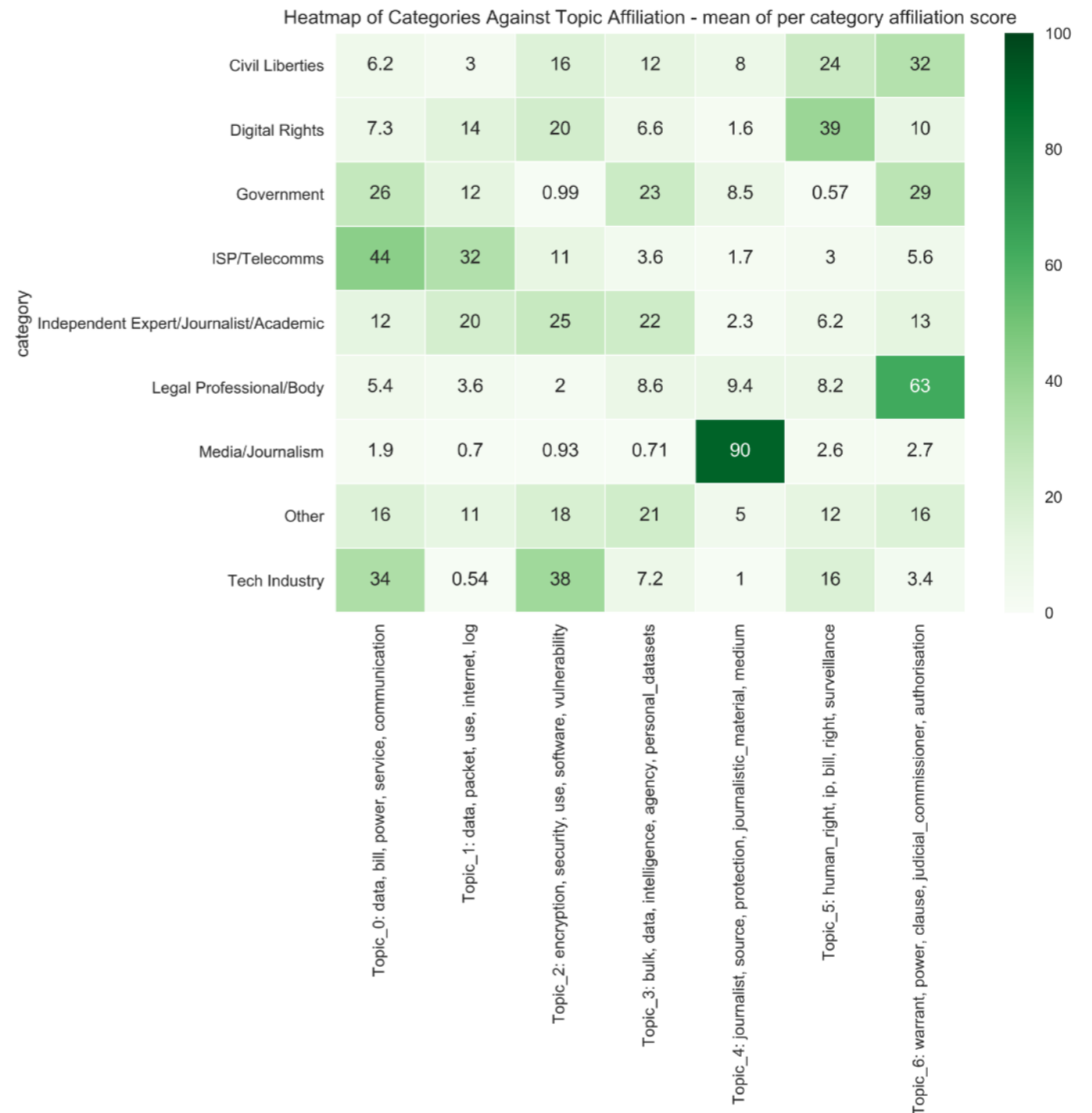▸ Where do Uber Drivers and Taxi Drivers converge and diverge in their concerns?
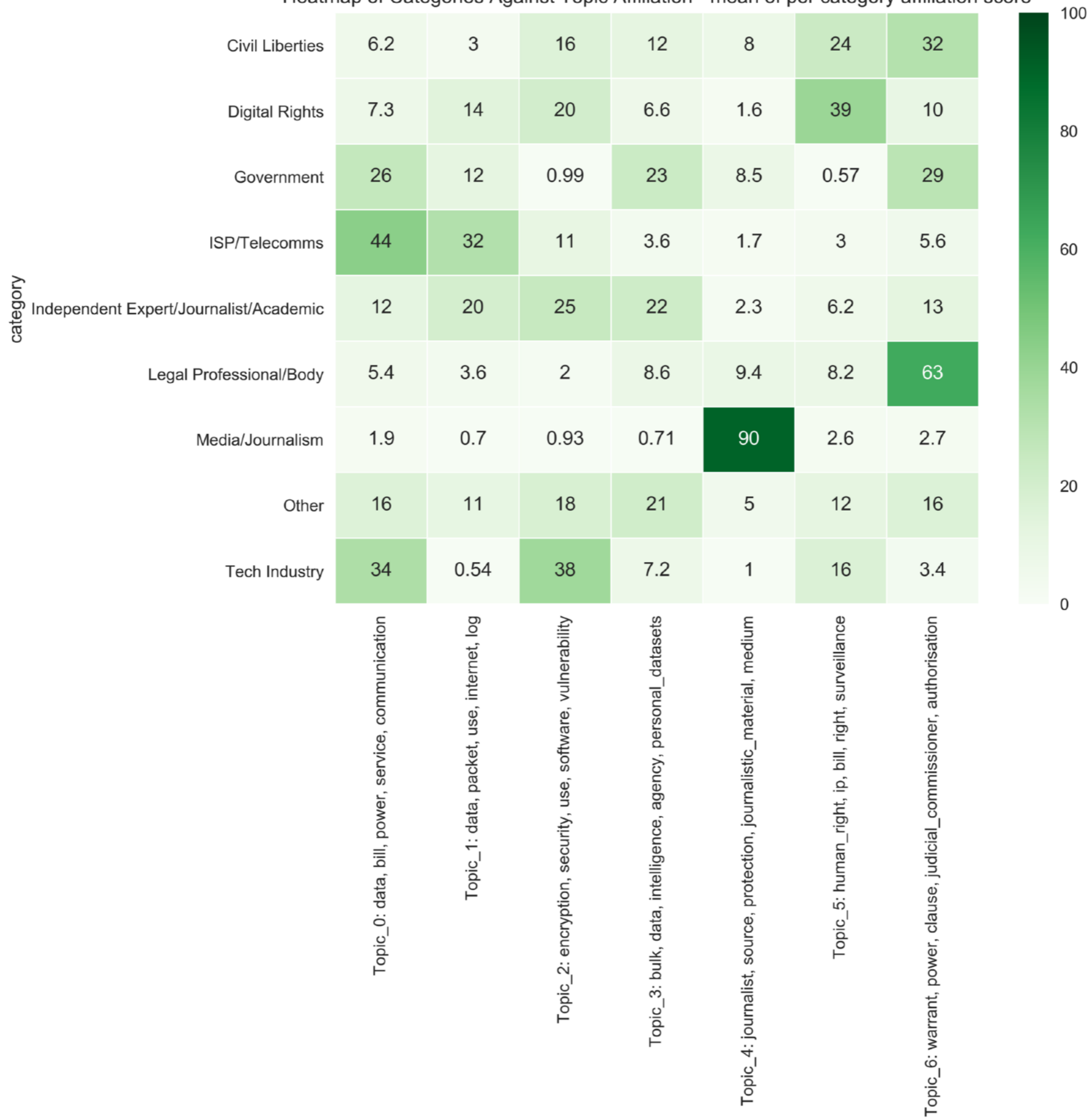
COMPUTATIONAL SOCIAL SCIENCE EXAMPLES
TOPIC MODELLING FOR NUANCED DISCOURSE

▸ Topic Modelling the Investigatory Powers Act (HRBDT)

  ▸ 219 submissions to the consultation categorised by organisation type.

  ▸ NMF Topic Modelling to discover trends in discourse across groups.

  ▸ Focusing in on HR discourse within tech firms. How is human rights being intertwined with tech?

  ▸ In progress

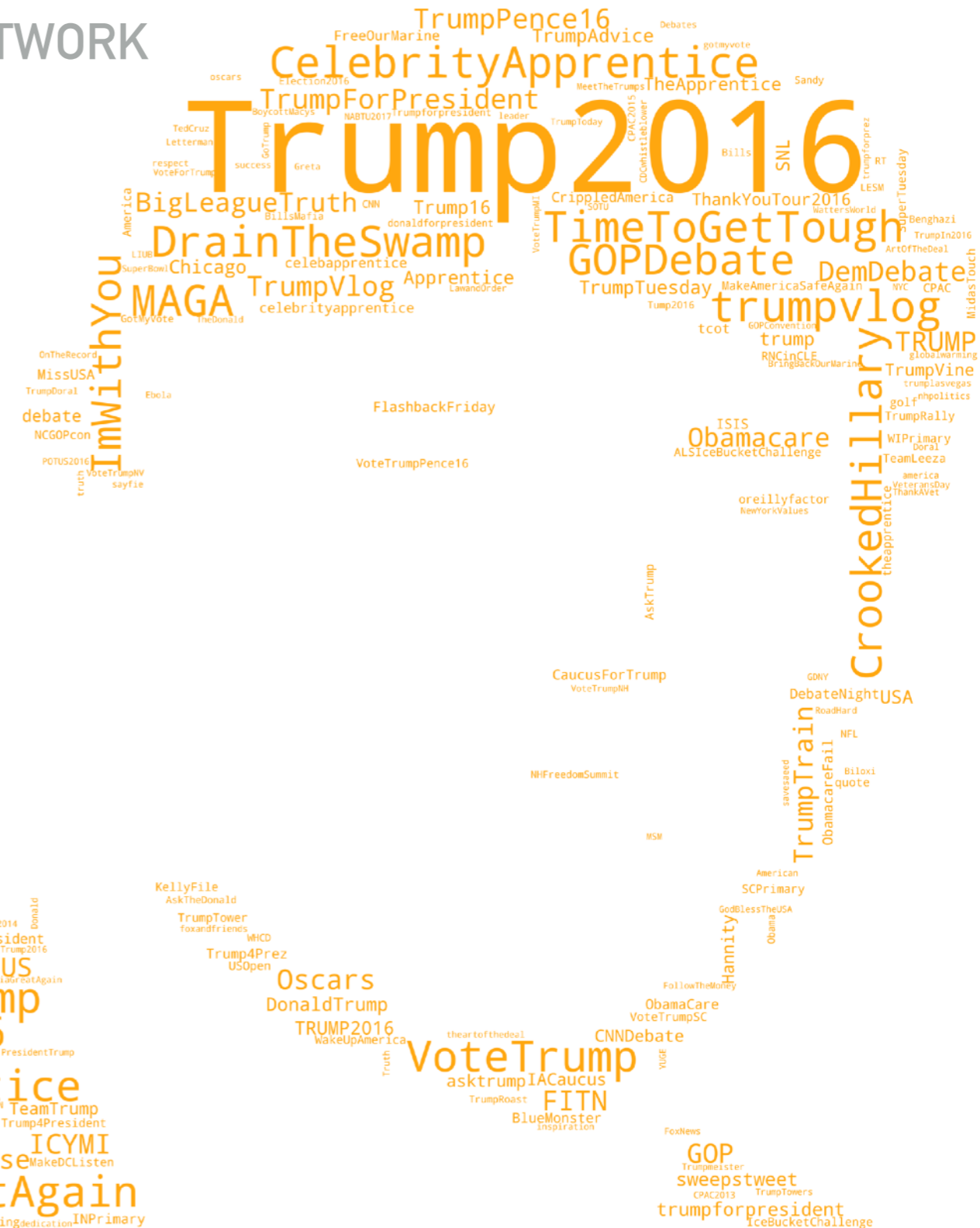Heatmap of Categories Against Topic Affiliation - mean of per category affiliation score

category	Topic_0: data, bill, power, service, communication	Topic_1: data, packet, use, internet, log	Topic_2: encryption, security, use, software, vulnerability	Topic_3: bulk, data, intelligence, agency, personal_datasets	Topic_4: journalist, source, protection, journalistic_material, medium	Topic_5: human_right, ip, bill, right, surveillance	Topic_6: warrant, power, clause, judicial_commissioner, authorisation
Civil Liberties	6.2	3	16	12	8	24	32
Digital Rights	7.3	14	20	6.6	1.6	39	10
Government	26	12	0.99	23	8.5	0.57	29
ISP/Telecomms	44	32	11	3.6	1.7	3	5.6
Independent Expert/Journalist/Academic	12	20	25	22	2.3	6.2	13
Legal Professional/Body	5.4	3.6	2	8.6	9.4	8.2	63
Media/Journalism	1.9	0.7	0.93	0.71	90	2.6	2.7
Other	16	11	18	21	5	12	16
Tech Industry	34	0.54	38	7.2	1	16	3.4

COMPU

HR DISCOU
NUANCES O

▸ Topic M
Powers

   ▸ 219
cons
orga

   ▸ NMI
disc
acro

   ▸ Focu
with
hum
inte

Heatmap of Categories Against Topic Affiliation - mean of per category affiliation score

COMPUTATIONAL SOCIAL SCIENCE EXAMPLES
CLASSIFYING DOCUMENTS WITH A NEURAL NETWORK

▸ Finding Trump Tweets with Neural Networks.

  ▸ Trained a Neural Network to recognise if a Tweet was written by Trump or his Staff.

  ▸ 76% accurate on unseen tweets.

  ▸ 2 day side project.

  ▸ Can AI support us in verifying the source of information?

http://bit.ly/findingtrump

COMPUTATIONAL SOCIAL SCIENCE EXAMPLES
CLASSIFYING DOCUMENTS WITH A NEURAL NETWORK

▸ Finding Trump Tweets with Neural Networks.

  ▸ Trained a Neural Network to recognise if a Tweet was written by Trump or his Staff.

  ▸ 76% accurate on unseen tweets.

  ▸ 2 day side project.

  ▸ Can AI support us in verifying the source of information?

PREDICTED TRUMP

text	prediction_score
...fired. This story is totally made up by the dishonest media.The Chief is doing a FANTASTIC job for me and, more importantly, for the USA!	0.958198
"46% of Americans think the Media is inventing stories about Trump & his Administration." @FoxNews It is actually much worse than this!	0.998146
Mike Flynn should ask for immunity in that this is a witch hunt (excuse for big election loss), by media & Dems, of historic proportion!	0.828915
Big win in Montana for Republicans! We	0.999999
Do you notice we are not having a gun debate right now? That's because they used knives and a truck!	0.999889

PREDICTED STAFF

The failing @NYTimes would do much better if they were honest! https://t.co/ATy8R3knS2	0.908425
Happy Birthday @TheLeeGreenwood! #FlashbackFriday https://t.co/MCrMK6Eurm	0.876031
It was my great honor to celebrate the opening of two extraordinary museums-the Mississippi State History Museum & the Mississippi Civil Rights Museum. We pay solemn tribute to our heroes of the past & dedicate ourselves to building a future of freedom, equality, justice & peace. https://t.co/5AkgVpV8aa	1
It was a great honor to welcome President Petro Poroshenko of Ukraine to the @WhiteHouse today with @VP Pence. ➡... https://t.co/J1ulOd6pYQ	0.99998
THANK YOU @MayorGimenez for following the RULE OF LAW! Sanctuary cities make our country LESS SAFE! Full remarks: https://t.co/Kz2EsQ3f1v https://t.co/vD3BB3hyti	1

http://bit.ly/findingtrump

COMPUTATIONAL SOCIAL SCIENCE EXAMPLES
TWITTER API IMAGE COLLECTION

▸ Collected 3,219 Tweets using #NoTap.

▸ Protest movement against a new pipeline project in Italy.

▸ Utilised a Twitter listener to gather Tweets live, and then extract the images from them for qualitative analysis.

▸ Aim was to study the role of the visual in online protest.

Di Ronco, Anna, James Allen-Robertson, and Nigel South. 'Representing environmental harm and resistance on Twitter: The case of the TAP pipeline in Italy', *Crime, Media, Culture*, 0: 1741659018760106.
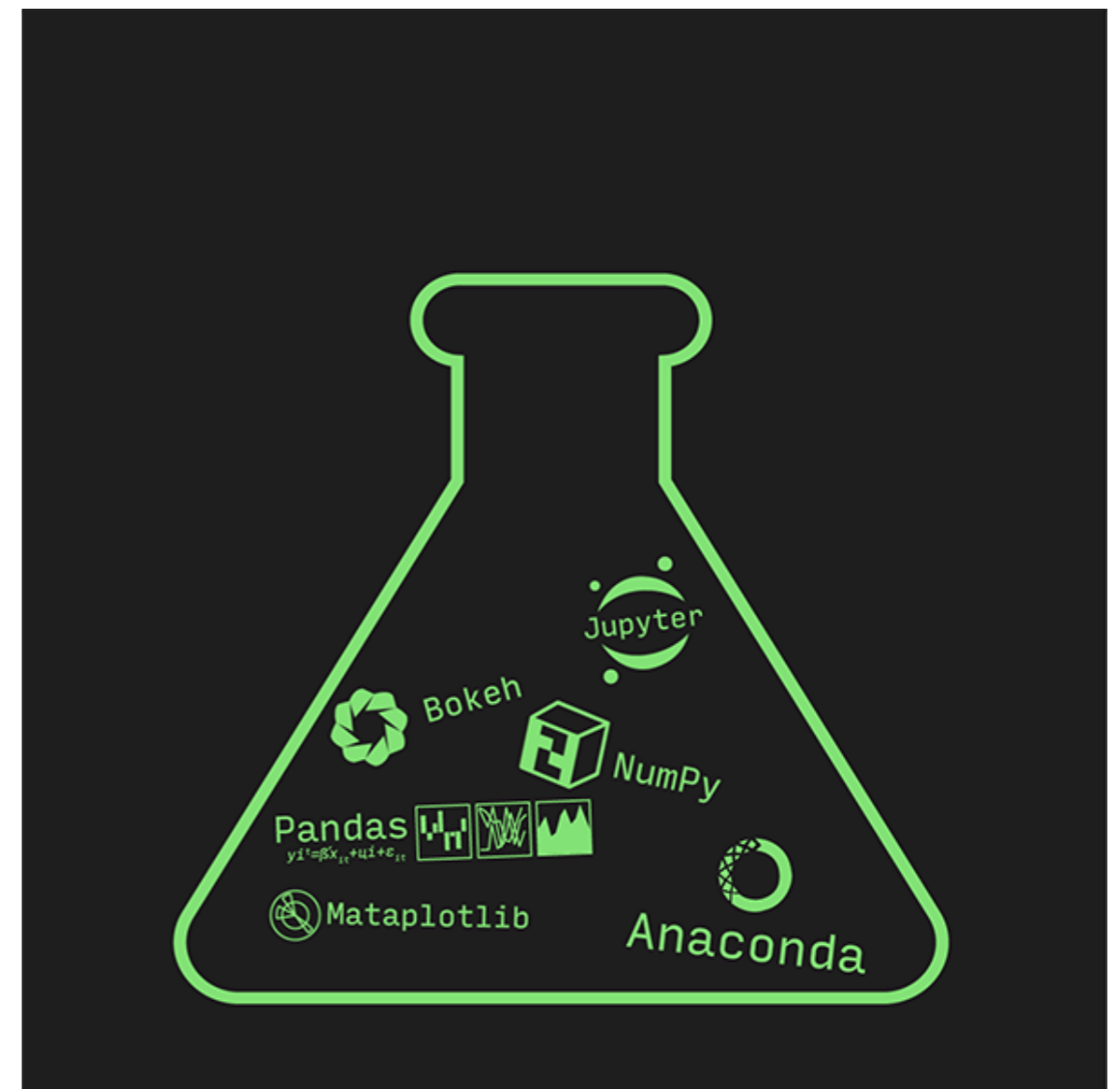
COMPUTATIONAL SOCIAL SCIENCE EXAMPLES
OTHERS

▸ Network relations, bot detection and discourse divergence in the Italian Vaccination Debate.

▸ Automated coding of image content using Neural Network based feature identification > Visual culture in the Catalonian Referendum (4m Tweets).

▸ Mapping the Dark Web - Web Crawler + Trained Classifier > Ethical Issues in good abundance!

Warning: Enthusiasm for computational methods and collaboration may lead to project overload

PROGRAMMING FOR SOCIAL SCIENTISTS
INCENTIVES

▸ Existing software may not scale or have the feature flexibility for your project.

▸ Automated data extraction opens up a range of new sources and questions / possibilities.

▸ Currently very fast development in computer science - new techniques to be repurposed for research.

▸ Allows you to build custom tools specific to your research question.

▸ All components are free!!

GIF by Max Breaux

PROGRAMMING FOR SOCIAL SCIENTISTS
LEARNING

▸ Basics from a 2 hour workshop at a conference…

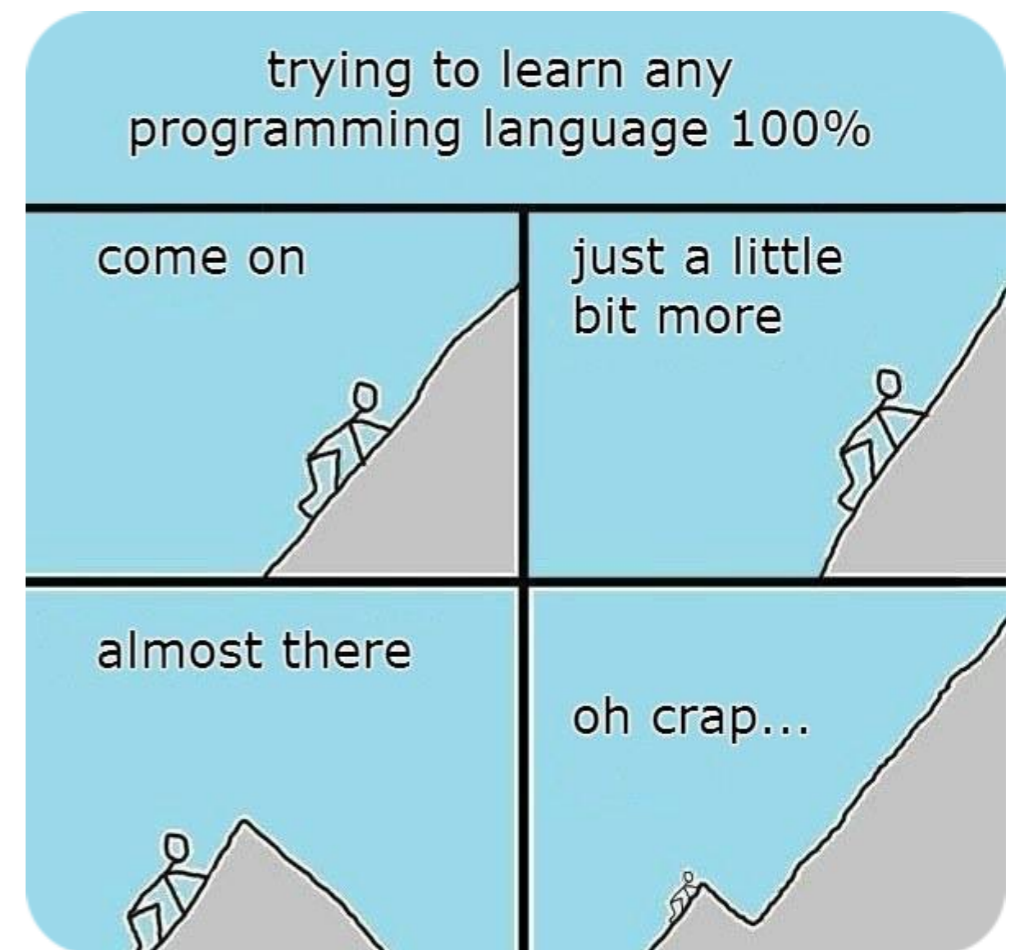▸ Online resources

  ▸ Blogs

  ▸ Tutorials

  ▸ Code Repositories

▸ Online courses

  ▸ Sage Campus (Academia Focus)

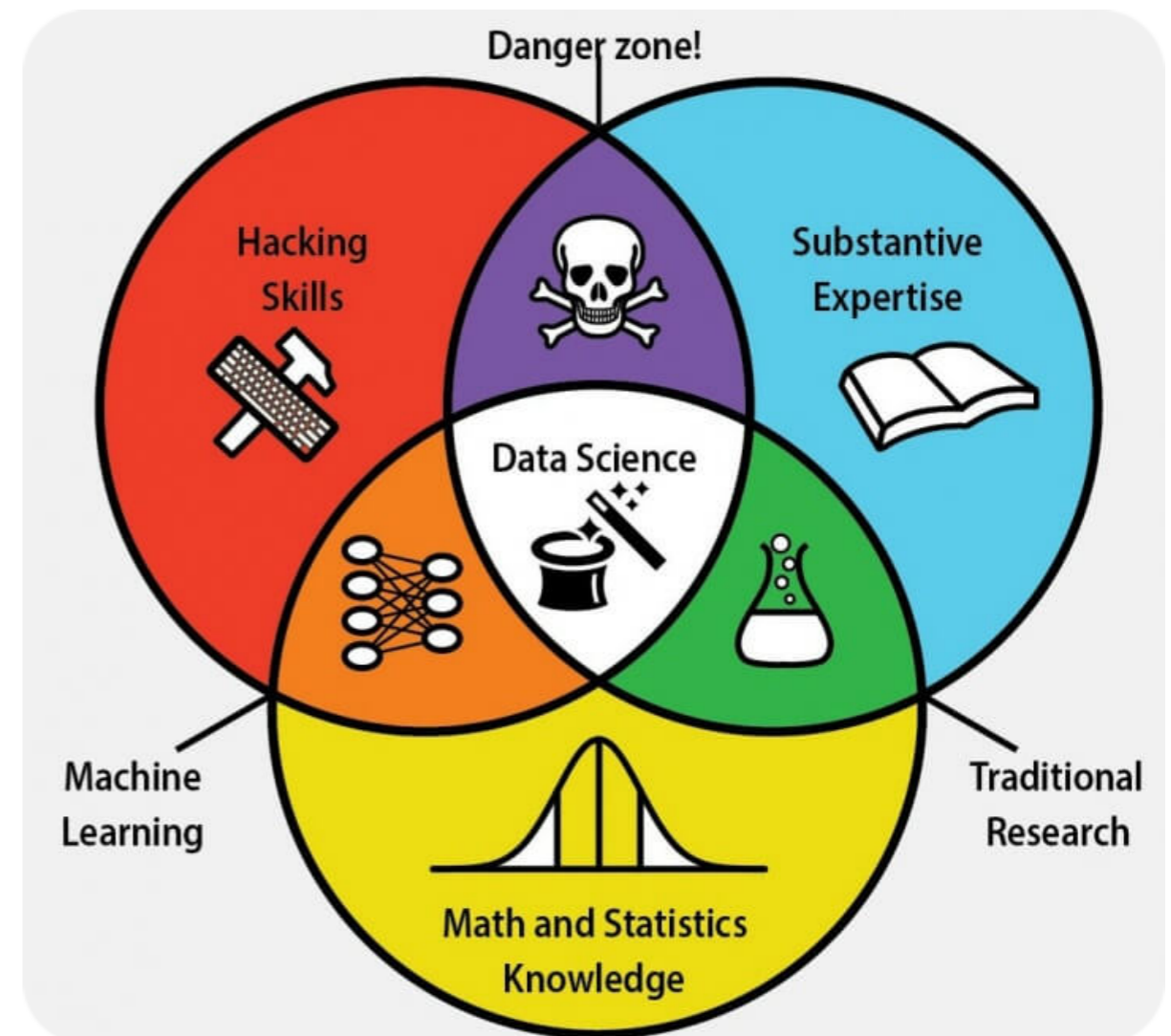 ▸ DataCamp - (Interactive)

  ▸ DataQuest - (Interactive)

  ▸ …

PROGRAMMING FOR SOCIAL SCIENTISTS
CHALLENGES

▸ Translating Techniques to Academia

    ▸ Often tutorials have Industry focus.

    ▸ Sometimes you can utilise tools 'unconventionally'.

▸ Statistics!!

    ▸ Ultimately techniques are reliant on statistical principles.

    ▸ Not statistically minded? > Can make statistics easier(!)

▸ Dirty Data

PROGRAMMING FOR SOCIAL SCIENTISTS
CHALLENGES

▸ Often each stage of a project requires learning something new.

  ▸ Steep initial learning curve.

  ▸ Rabbit hole

▸ Techniques have to change with scale.

  ▸ 100 Docs - Text Documents

  ▸ 1,000 Docs - DataFrames

  ▸ 1M Docs - Databases!

PERSONAL TIMELINE:

2 Weeks > Web Scraper Built

1.5 Year > First Paper

2 Years > Rapid Prototyping and research design.

PROGRAMMING FOR SOCIAL SCIENTISTS
CHALLENGES

▶ Often each stage of a project requires learning something new.

    ▶ Steep initial learning curve.

    ▶ Rabbit hole

▶ Techniques have to change with scale.

    ▶ 100 Docs - Text Documents

    ▶ 1,000 Docs - DataFrames

    ▶ 1M Docs - Databases!

PERSONAL TIMELINE:

2 Weeks > Web Scraper Built

1.5 Year > First Paper

2 Years > Rapid Prototyping and research design.

PROGRAMMING FOR SOCIAL SCIENTISTS
BENEFITS

▸ Collaboration

    ▸ Get involved in a vast range of substantive topics.

    ▸ Talk to colleagues, find out what their approaches are, could they benefit from these techniques? - Not always!

▸ New frontiers == Favourable Editors (Sometimes)

    ▸ Generally a strong positive response from reviewers and editors for methodological innovation.

▸ Currently in vogue - strong skills demand outside academia, research councils very favourable.

Data Science
✔

Data Science
+ Social Science
✔✔✔

PROGRAMMING FOR SOCIAL SCIENTISTS
GETTING STARTED IN PYTHON

▸ Fundamental

 ▸ Pandas - Data storage and structure

 ▸ SciKit Learn -
 Classification / Regression /
 Clustering / Dimensionality Reduction /
 Model Selection / Preprocessing.

 ▸ Seaborn - Data Visualisation

 ▸ Requests - Web Scraping

 ▸ Natural Language Toolkit - Text mining
 and linguistics.

www.anaconda.com/download

Download, install and most of the setup
headache is taken care of. Includes…

Integrates code, output, notes and text
into a single document. Much more
intuitive than a normal IDE.

PROGRAMMING FOR SOCIAL SCIENTISTS
GETTING STARTED IN PYTHON

▸ Advanced

  ▸ Bokeh - Build Interactive Data Visualisations.

  ▸ Scrapy - Powerful high speed web scraping.

  ▸ Keras - Build and train Neural Networks for supervised machine learning classification.

  ▸ Gensim - Text analysis at very large scale.

www.anaconda.com/download

Download, install and most of the setup headache is taken care of. Includes…

Integrates code, output, notes and text into a single document. Much more intuitive than a normal IDE.

PROGRAMMING FOR SOCIAL SCIENTISTS
GETTING STARTED IN PYTHON

HELP!

▸ Stackoverflow.com

▸ Google Groups

▸ Github

▸ Google Search Magic

…or email me

NCRM

...or email me

JALLENH@ESSEX.AC.UK

Dr. James Allen-Robertson - University of Essex

@Minyall   github.com/Minyall

THANKS!