ThinkBIG

Patterns in Big Data: Methods, Applications and Implications

ERC Advanced Grant

Nello Cristianini

Professor of Artificial Intelligence

University of Bristol

Outline

Part 1 - Objectives and Results of ThinkBIG (2014-2019)

Big Data: Analysis of media content + Social consequences

Part 2 - Future Opportunities and Challenges in AI

Priority: learning how to coexist

Part 3 - Links between AI and other disciplines

Key action in interface with social science, biology, humanities

Part 4 - The Future of this sub-area

Relation with individuals and society

AI and Data…

… there is a close relation between modern AI and Big-Data.

Modern AI is data-driven, it is based on the application of Machine Learning Algorithms to vast masses of data, to discover statistical signals that can be useful.

This leads to both benefits and risks ...

ThinkBIG: "Patterns in Big Data: Methods, Applications and Implications"

Stated goals were: “Understanding, exploiting and managing the paradigm shift to Data-Driven approaches (2014-2019). “

1) developing **new types of algorithms** and methods to take full advantage of this opportunity. ✔
2) exploring **new areas of opportunity** for big-data to make an impact, with particular attention to the growing field of computational social sciences. ✔

3) developing a set of cultural, legal and technical tools to **reduce the risks** associated with the application of these technologies to science and society. (investigating the ethical and epistemological challenges that arise from the transition towards a data-driven way of running society, business and science.) ✔

Some Results (we finish in 4 months)

New **Methods**:

- Hypothesis testing in massive multiple-testing setting
- Hashing methods for large data streams
- Sentiment through translation
- Removal of bias from deep NN
- Removal of bias from word embeddings

Applications:

Humanities: vast scale analysis of historical newspapers

- **Gender bias over time**
- Italian newspapers - full stack
- Narrative networks in Britain
- Periodic structures in US and UK news

- Social sciences: vast scale analysis of social media
  - **Gender bias in modern news**
  - Gender bias in embeddings
  - **Mood in twitter** for brexit
- Psychology and Biology also touched:
  - **Diurnal and Seasonal structures** in psychometric indicators
  - Also in wikipedia, and OTC medication, and google queries,
  - Recently sunlight, google queries, prescriptions ?

Collaborations with philosophers, lawyers, biologists, historians, economists, ..

Results (we finish in 4 months) -

Some highlights 1

- **Humanities**:
  - We can use big-data and AI to analyse the contents of historical newspapers
  - We can go from a box of microfilms to a map
  - We need historians to be part of the journey
  - We can see changes in social structure and even values
  - A new set of signals for historians
- **Social Science**
  - There is gender bias in the content of newspapers (there has been for past 200 years at least)
  - It can be absorbed into language models
  - We can probably remove it
  - A new frontier for social sciences

Psychology

- 73 psychometric indicators follow a 24-cycle in twitter content
- 5 mood indicators follow a seasonal cycle
- Wikipedia access, over-the-counter medication, google searches all follow similar patterns
- A new source of signals for medical and psychological research

-

Results

Some highlights 2

Implications:

- Machine decisions and human consequences
- Machine persuasion
- Machine psychometrics
- Social machines and Algorithmic Regulation

Intelligent Machines can influence user behaviour, and this has real ethical and social implications

PNAS

Content analysis of 150 years of British periodicals

Thomas Lansdall-Welfare[a], Saatviga Sudhahar[a], James Thompson[b], Justin Lewis[c], FindMyPast Newspaper Team[d,1], and Nello Cristianini[a,2]

[a]Intelligent Systems Laboratory, University of Bristol, Bristol BS8 1UB, United Kingdom; [b]Department of History, University of Bristol, Bristol BS8 1TB, United Kingdom; [c]School of Journalism, Media and Cultural Studies, University of Cardiff, Cardiff CF10 3NB, United Kingdom; and [d]FindMyPast Newspaper Archive Limited (www.britishnewspaperarchive.co.uk), Dundee DD2 1TP, Scotland

Edited by Kenneth W. Wachter, University of California, Berkeley, CA, and approved November 30, 2016 (received for review April 21, 2016)

Previous studies have shown that it is possible to detect macroscopic patterns of cultural change over periods of centuries by analyzing large textual time series, specifically digitized books. This method promises to empower scholars with a quantitative and data-driven tool to study culture and society, but its power querying of individual words. It concluded by advocating for the use of big data methods for newspaper analysis and proposing specific criteria for the design of such experiments.

Although the "Culturomics" study (1) was based on the idea of introducing quantitative and measurable aspects to the study

PLOS ONE

RESEARCH ARTICLE

Diurnal variations of psychometric indicators in Twitter content

Fabon Dzogang[1], Stafford Lightman[2], Nello Cristianini[1]*

1 Intelligent Systems Laboratory, University of Bristol, Bristol, United Kingdom, **2** Henry Wellcome Laboratories for Integrative Neuroscience and Endocrinology, University of Bristol, Bristol, United Kingdom

Research Paper

Circadian mood variations in Twitter content

Fabon Dzogang[1], Stafford Lightman[2] and Nello Cristianini[1]

Brain and Neuroscience Advances

Brain and Neuroscience Advances
Volume XX: 1–14

DOI: 10.1177/2398212817744501
journals.sagepub.com/home/bna
SAGE

Large-scale content analysis of historical newspapers in the town of Gorizia 1873–1914

Nello Cristianini, Thomas Lansdall-Welfare, and Gaetano Dato

Department of Engineering Mathematics, University of Bristol

ABSTRACT

We have digitised a corpus of Italian newspapers published in 1873–1914 in Gorizia, the county town of an area in the North Adriatic at the crossroad of the Latin, Slavic and Germanic civilizations, then part of the Habsburg Empire and now divided between Italy and Slovenia. This new corpus (of 47,466 pages) is analysed along with a comparable set of local Slovenian newspapers, already digitised by the Slovenian National Library. This large and multilingual effort in digital humanities reveals the statistical traces of events and ideas that shaped a remarkable place and period. The emerging picture is one of rapid cultural, social and technological transformation, and of rising national awareness, combining the larger European pattern with uniquely local aspects.

KEYWORDS

Austro-Hungarian empire; digital humanities; digital newspaper archives; Gorizia

Part 2 - AI: Opportunities and Concerns

ThinkBIG was all about identifying opportunities and concerns

- They are BOTH at the interface between AI and other disciplines
- We benefit from automation, in terms of accuracy and efficiency;
- We lose from wild-data leaks, reckless-shortcuts, ethical-debt, and hyper-personalisation business models…
- (Employment MIGHT be a consideration too, but not as big as media think…)

Can we access information about society and psychology without violating individual rights? Can we provide personalised services without violating user autonomy?

Part 3 -Multidisciplinary Applications of AI

ThinkBIG was all about crossing the border with:

- social sciences,
- psychology,
- humanities,
- law,
- ethics,
- philosophy
- Sociology

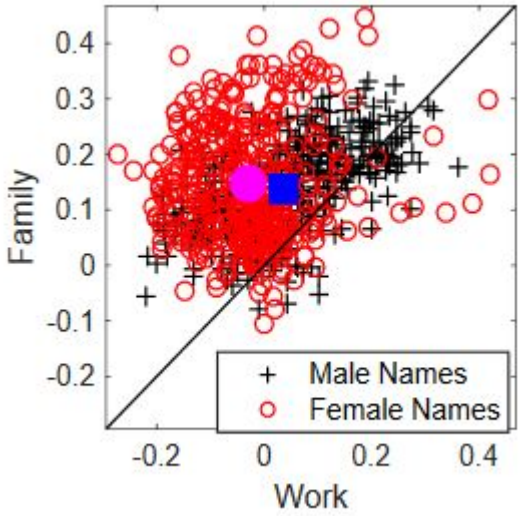
That interface is very active, it is AI's future natural habitat.

The future of AI as a new medium.

One Example - on fairness…

Why would an AI algorithm behave like that?

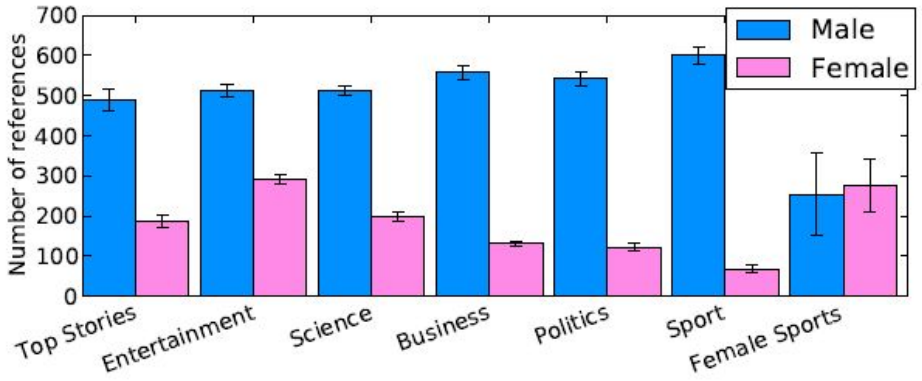
In the models within AI agents …



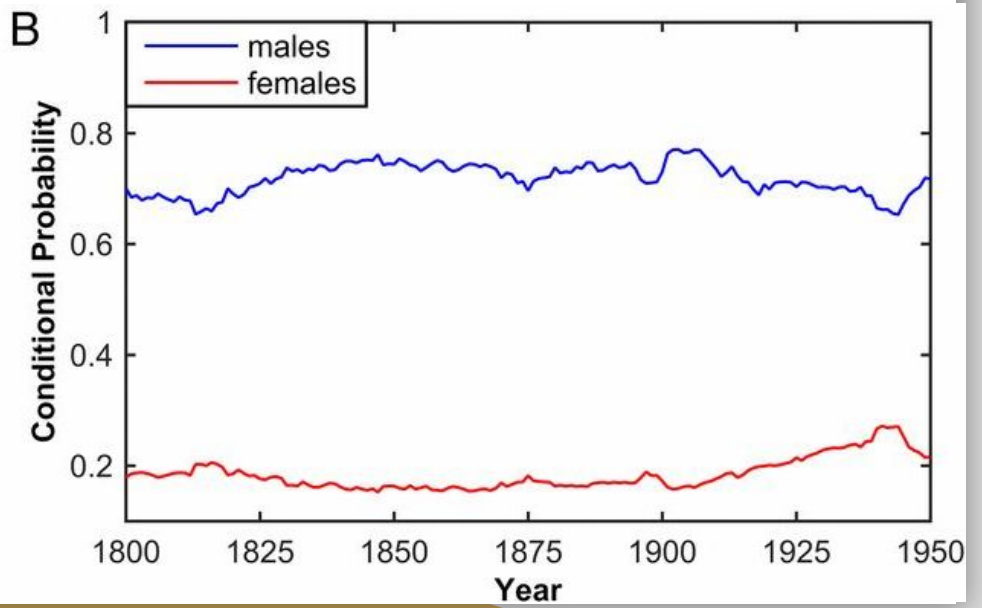
Our recent study in Bristol ..

TABLE I: List of the top 10 occupations per gender by their association with gender.

Gender	Occupations most associated with a gender
Male	Manager, Engineer, Coach, Executive, Surveyor, Secretary, Architect, Driver, Police, Caretaker, Director
Female	Housekeeper, Nurse, Therapist, Bartender, Psychologist, Designer, Pharmacist, Supervisor, Radiographer, Underwriter



In the data
(often used to train them)



To be systematic…

We are planning to go through all protected categories (gender, race, age, ability, etc) … and protected domains (jobs, housing, justice, …) and develop tests and checks on data and models …

We are designing new and more robust algorithms ….

New ways to measure fairness (and various other things too)

… not an easy task

JUST AN EXAMPLE of what it will mean to CO-EXIST with AI.

Part 4- Outlook: which way is AI heading?

- data-driven AI is only one way of doing things, we need to keep an eye also on other techniques (eg reasoning)
- we will need to become good at **co-existing** with this technology;
- will need to become good at thinking new business models (and more assertive about our rights)
- Technically and socially: we will be digesting the implications of this turn for a while, I hope, before the next major change …
  we need time to digest this ….

The future of the field - benefits and problems?

- Machines that make decisions **about people** need to be transparent, accountable, private, fair, ...
- Risks come from putting data-driven AI machines in the position to make decisions that affect humans, based on **biased data from the wild**
- we know about bias in the wild, and we know about how this leaks into our AIs

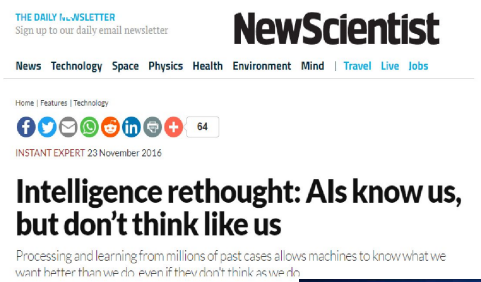
We need urgent work to understand this LEAK

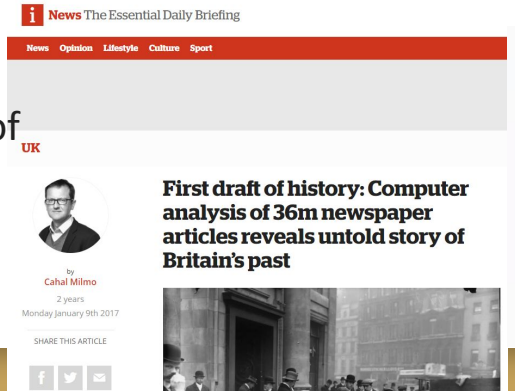
- We will not be able to go all the way in AI unless we have strong guarantees against abuses
- There MUST be **a European path to AI** - we will be part of finding it
- **BENEFITS:** benefits of automation can be manifold, mostly they will be in the two dimensions of
  - Increased efficiency
  - Increased accuracy
- But also remember: automation EQUALS removing humans

Reaching out…

Crucial duty at this time

- Articles on New Scientist
- Tv interviews
- National newspapers
- Keynotes
- EU Parliament
- Council of Europe
- Meetings at JRC
- Relation with UK Department of Media, Culture, Digital





Two Remarks…

- We need to chart a European path to AI (not dominated by State nor by Industry, but defined by citizen / user values)
- ERC has already been funding research in this area for several years…

ThinkBIG Team