University wide course on JustData

Fall, 2017

Co-ordinated by Karoly Boroczky, Mathematics

Number of credits: 2 (4 ECTS credits)

Time period: Fall

Prerequisites: No

Grading: To get credit, a 4-6 pages long case study should be written on a topic chosen by the student. Consultation on Academic Writing and on the Topic is provided.

Learning outcomes

By the end of the course, students will have a clear understanding of the Social Justice aspects of Big Data, including its dark side, like collecting and misusing data, and many good practices.

Short Syllabus

Big Data is all around us – facebook users, records on citizens, the network of neurons in the brain, routes of migrants, impact of publications. The Data itself is neither good or evil, however, it can be used for either purposes. The availability and analysis of big data opens up enormous opportunities for research, but is not without serious dangers. The course explores the amazing potential and the dark side of Big Data.

Topics

1. Data Privacy and Security (co-ordinated by Miklos Koren and Arieda Muco, Economics)
2. Ethics of Big Data (co-ordinated by Chrys Margaritidis, Dean of Student Office)
3. How data helps justice (co-ordinated by Jozsef Martin, Transparency International Hungary)
4. How data helps advancing knowledge (co-ordinated by Roberta Sinatra, Network Science)

1. **Data Privacy and Security** (co-ordinated by Miklos Koren and Arieda Muco, Economics)

This section discuss the technological challenges of dealing with individually identifiable data. Three topics will be covered:

i. Identifying individuals from limited information: challenges and solutions
ii. Protecting your data: Best practices for managing data access
iii. Introduction to cryptography and data encryption

Conceptual lectures by industry experts will be combined by hands-on exercises.

Readings
Books

- An Introduction to Statistical Learning with Applications in R by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani
- Big data : a revolution that will transform how we live, work, and think / Viktor Mayer-Schönberger and Kenneth Cukier
- Big Data at Work: Dispelling the Myths, Uncovering the Opportunities,Thomas H. Davenport
- Data Analytics Made Accessible by Anil Maheshwari
- Elements of statistical Learning: Trevor Hastie, Robert Tibshirani, Jerome Friedman
- Introduction to Machine Learning by Nils J. Nilsson
- Machine Learning in Action: Peter Harrington
- Thinking with Data: How to Turn Information into Insights, Max Shron

Online material

- https://www.coursera.org/learn/machine-learning: Andrew Ng's video
- http://cs229.stanford.edu/materials.html: Andrew Ng's notes
- https://www.quora.com/How-will-machine-learning-impact-economics?redirected_qid=6706789
- http://people.ischool.berkeley.edu/~hal/Papers/2013/ml.pdf Hal R. Varian
- http://datajournalismhandbook.org/1.0/en/getting_data.html

2. **Ethics of Big Data** (co-ordinated by Chrys Margaritidis, Dean of Student Office)

This part of the Just Data university wide seminar will examine some of the ethical questions that Big Data brings to the foreground. We focus on the relationship between Big Data and three issues:

i. Identity and Autonomy,
ii. Privacy and Surveillance,
iii. Freedom of Expression.

We will examine them separately but also to seek the connections between them. In all three discussions, we will first identify the types of changes that Big Data technologies compel in how we think of these concepts and issues. Then, we will look at the ethical consequences of both the technological advances themselves and the reformulation of the concepts. All discussions will be illustrated by case studies.

Readings

- Danah Boyd a & Kate Crawford, (2012), "Critical Questions for Big Data", Information, Communication & Society, Vol. 15, (5), pp.662-679

- Neil M. Richards and Jonathan H. King, (2014), "Big Data Ethics", Wake Forest Law Review, pp.393-432

- Jacob Metcalf, Emily F. Keller and Danah Boyd, (2016), "Perspectives on Big Data, Ethics and Society", The Council for Big Data, Ethics, and Society, pp.1-23

3. **How data helps justice** (co-ordinated by Jozsef Martin, Transparency International Hungary)

- Dr. József Péter Martin (Executive Director, TI Hungary): How to measure corruption?
- Dr. Gabriella Nagy (Head of Public Finance Programs TI Hungary): The role of technology in anti-corruption.
- Dr. Miklós Ligeti (Head of Legal, TI Hungary): Featuring the systemic and centralized corruption in Hungary: what can be quantified, and what can not?

Titles and topics

1. How to measure corruption?
   The relevance and constraints in quantifying corruption exposure. Introduction to most important proxies including TI's Corruption Perception Index and its methodology. The link of corruption with other proxies of democracy and good governance and its impact on economic performance - the case of Hungary in comparative perspective.

2. The role of technology in anti-corruption.
   Innovative and interactive tools to detect corruption risks and promote transparency developed by Transparency International Hungary. Introduction to Redflags.eu, a public procurement corruption risk signaling tool, and to Assist.hu, a data mining tool. The corruption risks in public procurement and in distribution of EU funds - the case of Hungary in comparative perspective.

3. Featuring the systemic and centralized corruption in Hungary: what can be quantified, and what can't?
   Estimating finances of political parties, i.e. an innovative method to calculate campaign spending developed by Transparency international Hungary before the last national election. The use of criminal statistical data on corruption – what do they show, and what do they hide?

Readings

https://www.transparency.org/news/feature/corruption_perceptions_index_2016
https://transparency.hu/en/news/cpi-2016-magyarorszag-tovabbra-is-lejtmenetben/
https://www.transparency.org/research/gcb/gcb_2015_16

http://www.redflags.eu/

https://transparency.hu/en/
https://transparency.hu/en/kozszektor/kozbeszerzes/
https://transparency.hu/en/kozszektor/letelepedesi-magyar-allamkotvenyek/
https://transparency.hu/en/news/integrity-pact-by-ti-to-be-used-on-the-m6-motorway-project/

4. **How data helps advancing knowledge** (co-ordinated by Roberta Sinatra, Network Science)

In this part of the course, we provide an understanding of how we can extract meaning from data and how this can be used for social good. We will do so in a cycle of 2-3 classes, where we will give an overview of recent research, analyzing the goals, challenges and pitfalls that come from advancing knowledge through large datasets. We will cover three overarching topics:

i. use of data to provide actionable predictions (e.g. use of individual mobility data to predict epidemics and pandemics, and the how it is used for policy decision),
ii. use of data to design algorithms that benefit our society and environment (e.g. use of taxi sharing data to design shareability systems that reduce emissions and improve infrastructure efficiency),
iii. use of data to understand the fairness and biases in society, possibly enhanced by algorithms (e.g. how to collect data to test the fairness of ranking algorithms).

Readings

1. use of data to provide actionable predictions (e.g. use of individual mobility data to predict epidemics and pandemics, and the how it is used for policy decision),

Balcan, D., Colizza, V., Gonçalves, B., Hu, H., Ramasco, J. J., & Vespignani, A. (2009). Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences*, *106*(51), 21484-21489.
http://www.pnas.org/content/106/51/21484.full

Science 346 (6213), 1063-1064The parable of Google Flu: traps in big data analysis
D Lazer, R Kennedy, G King, A Vespignani - Science, 2014
https://dash.harvard.edu/bitstream/handle/1/12016836/The%20Parable%20of%20Google%20Flu%20%28WP-Final%29.pdf

Althouse, B. M., Scarpino, S. V., Meyers, L. A., Ayers, J. W., Bargsten, M., Baumbach, J., ... & Del Valle, S. (2015). Enhancing disease surveillance with novel data streams: challenges and opportunities. *EPJ Data Science*, *4*(1), 17.
https://link.springer.com/content/pdf/10.1140%2Fepjds%2Fs13688-015-0054-0.pdf

Gomes, M. F., y Piontti, A. P., Rossi, L., Chao, D., Longini, I., Halloran, M. E., & Vespignani, A. (2014). Assessing the international spreading risk associated with the 2014 West African Ebola outbreak. PLoS currents, 6.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4169359/

Wang, D., Song, C., & Barabási, A. L. (2013). Quantifying long-term scientific impact. *Science*, *342*(6154), 127-132.
https://static1.squarespace.com/static/5877ca6986e6c00f05f58f84/t/58a0acb8bebafb6786e3367b/1486924994671/quantifying-long-term-scientific-impact.pdf

Clauset, A., Larremore, D. B., & Sinatra, R. (2017). Data-driven predictions in the science of science. *Science*, *355*(6324), 477-480
http://www.robertasinatra.com/pdf/SciencePrediction2017_ClausetLarremoreSinatra.pdf

Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D., Marlow, C., Settle, J. E., & Fowler, J. H. (2012). A 61-million-person experiment in social influence and political mobilization. *Nature*, *489*(7415), 295-298.
https://www.nature.com/nature/journal/v489/n7415/pdf/nature11421.pdf

2. use of data to design algorithms that benefit our society and environment (e.g. use of taxi sharing data to design shareability systems that reduce emissions and improve infrastructure efficiency),

P. Santi, G. Resta, M. Szell, S. Sobolevsky, S. Strogatz, C. Ratti. Quantifying the benefits of vehicle pooling with shareability networks. *Proceedings of the National Academy of Sciences 111(37)*, 13290-13294 (2014)
http://www.pnas.org/content/111/37/13290.full.pdf

R. Tachet, O. Sagarra, P. Santi, G. Resta, M. Szell, S. Strogatz, C. Ratti. Scaling law of urban ride sharing. *Scientific Reports 7*, 42868 (2017)
https://www.nature.com/articles/srep42868.pdf

C. Ratti, A, Biederman. From parking lot to paradise. *Scientific American* (2017)
http://senseable.mit.edu/papers/pdf/20170627_RattiBiderman_ParkingParadise_ScientificAmerican.pdf

V. Pandurangan. On Taxis and Rainbows (2014)
https://tech.vijayp.ca/of-taxis-and-rainbows-f6bc289679a1

J.K. Trotter. Public NYC Taxicab Database Lets You See How Celebrities Tip. *Gawker* (2014)
http://gawker.com/the-public-nyc-taxicab-database-that-accidentally-track-1646724546

3. Methods for collecting and using digital trace data; ethical and legal considerations; biases and pitfalls of big data.

Computational social science
D Lazer, AS Pentland, L Adamic, S Aral, AL Barabasi, D Brewer, ...
Science (New York, NY) 323 (5915), 721
https://gking.harvard.edu/files/gking/files/LazPenAda09.pdf

Social media for large studies of behavior
D Ruths, J Pfeffer
https://www.researchgate.net/profile/Juergen_Pfeffer/publication/268879558_Social_Media_for_Large_Studies_of_Behaviour/links/55f87ff508ae07629dd77bbb.pdf

Science 346 (6213), 1063-1064The parable of Google Flu: traps in big data analysis

D Lazer, R Kennedy, G King, A Vespignani - Science, 2014
https://dash.harvard.edu/bitstream/handle/1/12016836/The%20Parable%20of%20Google%20Flu%20%28WP-Final%29.pdf

Social data: Biases, methodological pitfalls, and ethical boundaries
A Olteanu, C Castillo, F Diaz, E Kiciman
https://www.microsoft.com/en-us/research/wp-content/uploads/2017/03/SSRN-id2886526.pdf

Bias in Online Freelance Marketplaces: Evidence from TaskRabbit and Fiverr.
A Hannák, C Wagner, D Garcia, A Mislove, M Strohmaier, C Wilson
http://ancsahannak.me/files/CSCW17.pdf