Course description:

We are living in the age of big data. The abundance of big data can be utilized by economists and policy makers to address key socio-economic problems related to education, health, climate change etc. However, in recent times, we have often seen big data to be misrepresented and misdirected not only by the media and politicians, but also in academic and policy circles. This course introduces students to the basics of big data and how to analyze it correctly.

This course is a hybrid of theory needed for analysis of big data and its practical application. The first half of the semester would be devoted to learning how big data is represented by economists. For this, the basics of mathematical statistics would be discussed. The second half would be based on application and case studies exploring these theoretical aspects from a practical viewpoint. Special topics include the spread of misinformation via misuse of big data, fake news and political bias, data manipulation, p-hacking, data privacy and the economic impacts of incorrectly using big data. One of the emerging issues in contemporary politics is the role played by the media in shaping public opinion. This course helps students to understand how big data is used/misused by the media to manipulate public opinion. This course would enable students to understand how big data can be used correctly to address some of the major socio-economic problems we are facing today.

Course work:

There is no particular textbook that students are required to follow per se for this course. Workload would include reading materials from a combination of textbooks, research papers and newspaper articles, respond to videos/documentaries and podcasts. Students are encouraged to focus on the main ideas, rather than the technical and econometric details of many of these readings. All materials would be provided and regularly updated on the moodle website.

Additionally, you may find these books to be useful references for this course and in future while dealing with big data:
- Mostly harmless econometrics: an empiricist’s companion by Joshua D. Angrist and Jörn-Steffen Pischke (A & P)
- Mastering ‘metrics’: The path from cause to effect by Joshua D. Angrist and Jörn-Steffen Pischke
- A field guide to lies by Daniel Levitin
- Spurious correlations by Tyler Vigen

Requirements and grading:

The requirements for this course would be based on overall class participation, one assignment, a midterm assessment, one class presentation and response papers.

Response papers and class participation (35%):
Students are supposed to read the required readings and write a response not exceeding 500 words. There would be 8 response papers spread due mostly on Sundays 11pm EST. Unless specified, they will be due on Sundays prior to Monday’s lecture. Students are allowed to miss one response paper without any penalty. They will submit their response papers on the moodle course website. These will be critical summaries/ reactions to a set of readings. Each of these responses should be in the range of 500 -1000 words. It is crucial for students to remember that this is not a summary of the readings, it should be their own reflection on the readings. They are expected to raise at least one critical question in each of these response papers pertaining to the topic, which would be discussed in class by other students and the instructor.
Since this is a mix of lecture and discussion, class participation is crucial. Participation includes not only attendance but engagement during the lecture. Students are expected to discuss, share ideas, ask questions, and critically engage in the class.

Assignment (15%):
There would be an analytical assignment which students are expected to submit by February 28th 11pm EST. This would be based on the course materials of the third and fourth week.

Mid-term assessment (25%): There would be an open book 2-hour long midterm assessment on Week 6. It would broadly be based on the concepts learnt in the class till March 10th. Students would be given a choice to pick a 2-hour slot on either March 11th or March 12th to take the assessment. Students with proper justification should inform me before the deadline if they need to reschedule it to any other date.

Class presentation (25%):
There would be class presentations on selected readings by group of students. Each group would consist of 2 students and would be needed to critically present at least one of the readings listed for the presentation to the class. They are expected to facilitate and lead a discussion rather than present a summary of the readings to the class. Doodle poll would be conducted before Week 6 to fix time slots for presentation.
A note on late submission:
Late submissions would not be accepted without a University sanctioned excused absence or an extension granted by me. To obtain an extension of deadline, students must contact me before the deadline. For more on class absence policies, please check
https://www.umass.edu/registrar/students/policies-and-practices/class-absence-policy

Academic honesty:
Please note that academic dishonesty is strictly prohibited and will be met with appropriate measures. Academic dishonesty includes but is not limited to cheating, plagiarism, citing or quoting without proper references, fabrication and facilitating dishonesty. More information on the university’s academic honesty can be found on:
http://www.umass.edu/dean_students/codeofconduct/acadhonesty/#policy

Grading:
Grading would be based on the criterion discussed above. The following schedule would be maintained for calculating the final grades:

<table>
<thead>
<tr>
<th>Cut offs</th>
<th>Grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>93</td>
<td>A</td>
</tr>
<tr>
<td>90</td>
<td>A-</td>
</tr>
<tr>
<td>87</td>
<td>B+</td>
</tr>
<tr>
<td>83</td>
<td>B</td>
</tr>
<tr>
<td>80</td>
<td>B-</td>
</tr>
<tr>
<td>77</td>
<td>C+</td>
</tr>
<tr>
<td>73</td>
<td>C</td>
</tr>
<tr>
<td>70</td>
<td>C-</td>
</tr>
<tr>
<td>60</td>
<td>D</td>
</tr>
<tr>
<td>-</td>
<td>F</td>
</tr>
</tbody>
</table>

Syllabus¹:

Week 1 (Feb 1, 3)
Introduction

Week 2 (Feb 8, 10)
Big data and the art of sifting
The art of sifting among seasoned adults: can economists be trusted?
https://scholar.princeton.edu/sites/default/files/reinhardt/files/ec_100_siffing.pdf

¹ This is a comprehensive list of all the materials that would be useful for this course, it will be updated regularly.
Watch: Big data is better data  
https://www.youtube.com/watch?v=8pHzROP1D-w

Watch: How algorithms shape our world?  
https://www.youtube.com/watch?v=TDaFwnOiKVe

Watch: Why fascism is so tempting and how your data could power it  
https://www.youtube.com/watch?v=xHHb7R3kx40

Watch: What do we do with all this big data?  
https://www.youtube.com/watch?v=AWPrOvzzqZk

Response paper 1 on The Art of Sifting and any one of the videos due on Feb 7th, 11pm EST

Week 3 (Feb 15, 17)²
Representing big data
Types of charts and diagrams:
- Line chart, Bar chart, Pie chart, Pictogram, Histogram, Frequency polygon, Ogive, Frequency distributions
- Case studies on incorrect representation of data

Reference: W & W (Chapter 2)

Week 4 (Feb 22, 24)
Use of mathematical statistics in the analysis of big data
Measures of central tendency:
- Arithmetic mean, Geometric Mean, Harmonic Mean, Median, Mode, Quartiles, Deciles, Percentiles
Measures of dispersion, moments, skewness and kurtosis:
- Range, Quartile deviation, Mean deviation, Standard Deviation, Lorenz curve, Moments, Skewness, Kurtosis

Reference: W & W (Chapters 3, 4)

Assignment based on Week 3 and 4 due on Feb 28th, 11pm EST

Week 5 (Mar 1, 3)
Association versus Causality
Basics of Correlation, Covariance, Regression, scatter plots, Interpretation and use of r in economics,
Basic rules to remember in Impact evaluation study, Validity issues

Reference: W & W (Chapters 5, 6, 7)

Response paper 2 based on Week 5 due on March 7th, 11pm EST

² Lecture notes will be provided for Week 3, 4 and 5
Week 6 (Mar 8, 10)
“Correlation is not causation”

Case studies from Spurious correlations by Tyler Vigen
Case study:

*Midterm assessment on March 11th - 12th.*

Week 7 (Mar 15, 17)
Big data in post-truth world

Case study:

https://www.washingtonpost.com/opinions/2021/01/27/republicans-have-strategy-take-back-power-heres-why-it-could-fail/


*Response paper 3 on Tandoc et al., (2018) due on March 14th 11pm EST.*

Reading for presentation:

Week 8 (Mar 22, 24)
Detection of rumor and misinformation

*Response paper 4 due on March 21st 11pm EST*

Readings for presentation:


**Week 9 (Mar 29, 31)**

**Media and spread of misinformation via misuse of big data:**
https://www.buzzfeednews.com/article/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook#.oeZ50QAke
Watch What Facebook and Google are hiding from the world
https://www.youtube.com/watch?v=p6vM4dhI9I8

*Response paper 5 due on March 28th, 11pm EST*

**Readings for presentation:**


**Week 10 (Apr 5, 7)**

**Big data privacy**
Guest lecture details TBD

**Week 11 (Apr 12, 14)**

**Special topics: p-hacking**
Watch p-hacking and power calculations
https://www.youtube.com/watch?v=UFhJefdVCjE

Watch Life after p-hacking
https://www.youtube.com/watch?v=8wDwcp1EwNM

Watch Scientific Studies: Last week with John Oliver
https://www.youtube.com/watch?v=0Rnq1NpHd㎡

Response paper 6 due on April 11th, 11 pm EST

Reading for presentation:

Week 12 (Apr 19, 21)
An economist’s guide to use and interpret big data

Case study:
1. A. Reinhart Carmen and Rogoff Kenneth: Growth in a time of debt
   https://pubs.aeaweb.org/doi/pdf/10.1257/aer.100.2.573


2. The challenges of charting regional inequality

Response paper 7 due on April 18th 11pm EST

Reading for presentation


Week 13 (Apr 26, 28)
Advantages of using big data


How one university used big data to boost graduation rates

Listen: [https://freakonomics.com/podcast/american-dream-really-dead/](https://freakonomics.com/podcast/american-dream-really-dead/)

Response paper 8 due on April 25th, 11pm EST

Readings for Presentation:


Week 14 (May 3)
Wrapping up