15.572 ANALYTICS LAB

ACTION LEARNING SEMINAR ON ANALYTICS AND MACHINE LEARNING

Instructors: Professor Erik Brynjolfsson
erikb@mit.edu; URL: http://digital.mit.edu/erik
Professor Abdullah Almaatouq
amaatouq@mit.edu; URL: http://www.amaatouq.com
Office hours by appointment

Class Times: Thursdays 4:00-5:30pm, E62-276
Special Sessions (Pitch Day and Final Presentations):
September 19, 4-8pm, Samberg Conference Center, 6th Floor (E52)
December 13, **2-7pm,** Samberg Conference Center, 7th Floor (E52)

TAs: Jeremy Yang zheny@mit.edu Sebastian Steffen ssteffen@mit.edu

Administrator: Susan Young susany@mit.edu

Summary and Objectives

The growth in big data, analytics and machine learning is transforming management decision-making, operations, marketing, finance, and product innovation. Businesses across the world are wrestling with these challenges and opportunities. We are on the cusp of a second machine age – an era where machines can automate or augment more and more of the mental tasks that previously only humans could do.

The purpose of the Analytics Lab (A-Lab) is to match student teams with leading-edge projects involving analytics and machine learning as they apply to business questions and problems. The primary focus of the projects is on the technical and analytical aspects, but business relevance provides the context and strategy.

Course Principles and Expectations

The primary criterion for projects is to provide a rich learning experience for the students. In addition, the projects should be of high relevance and interest to the supporting organization and senior managers and professionals in it.

Project teams of three to four students are expected to work independent of regular class meetings. Project sponsoring organizations will cover costs of travel and lodging, if any. Each project team will have an MIT-associated mentor to provide guidance and assistance and a link to outside project sponsors on an as-needed basis. The ultimate decision making and responsibility for project direction and completion rests with the team members themselves.

Two special sessions are scheduled: **Pitch Day on September 19 and Final Presentations on December 13. Attendance at both sessions is required.** Please arrange you schedule accordingly.

Notes on Class Activities and Due Dates:

- 9/19: On Pitch Day, we will meet jointly with the representatives from project proposing companies. Each will briefly describe their project as proposed, and students will have an opportunity to meet and informally mix with them and fellow students. The session will be followed by a reception. The chief aim of this session is to help inform student team formation and project selection.

- 9/22, 11:59pm: **DUE: Project Ranks**. Each student should complete the survey separately (link to follow). In the following days, faculty, mentors, and the course support team will work out assignments of projects to students/teams, subject to review by the proposing company.

- 9/26: Final team-project pairings will be communicated to students. MIT and every proposing company have executed a jointly signed NDA. Each student team member will be required to review and sign an acknowledgment stating that all will abide by the terms agreed to in the NDA. Additional information will follow from Ellen Baum.

- 10/10, 11:59pm: **DUE**: **Project plan**. Each team should submit one document to their mentor and Jeremy, the main TA. The project plan is to be developed by the team, reviewed by the team's mentor, and endorsed by the project sponsor before the deadline. It should be thought of as a working document, used by the team and mentor to assess progress and adjust and adapt through the semester. Here is a suggested outline:
    - Purpose and Scope: The project purpose and scope should serve as a compass that guides the team throughout the duration of the project. It should reflect the company proposal, but be more focused. Remember, the project is intended to be a rich learning experience for your team. This project is *not* a consulting engagement with the project sponsor. Bear this in mind when drafting your project purpose and scope. We welcome both creativity and practicality.
    - Objectives: Break the project down into high-level objectives that you intend to achieve. It is wise to define a clear "minimum viable product" as an initial goal, leaving room for expanding on that in a modular fashion as time and resources permit, including a "stretch goal" if all goes well.
    - Tasks: For each objective, list one or more granular tasks. For each task, define the following: a) due date, b) specific deliverables, c) who is responsible, d) current status. Revisit the objectives and tasks at least weekly to see if you're on track. Are there new opportunities or unanticipated barriers? Feel free to use a simple shared spreadsheet or more elaborate project management tool of your choice to track progress.

- 10/31, 10:00am: **DUE: Mid-term presentation slides**. Each team should submit their slides to their mentor and Jeremy.

- 10/31 & 11/7, teams will deliver 3-minute presentations on their project work to date and potential lines of future analysis. The chief aim of these sessions is to help illuminate issues common across teams in order to foster discussion and collaboration. This is a great chance to get feedback and suggestions. Teams should be open and honest about their progress and challenges. Class participants should be supportive, helpful and generous with comments and advice.

- 12/8, 11:59pm: **DUE**:
    - **Final report** (10 pages maximum, 3000 words, not including figures or references). Each team should submit one document to their mentor and one to Jeremy.
    - **Summary of findings** (one-page executive summary); summary should include a high-level statement of the challenge addressed by the project and the key insights the team generated during the semester. This can include a graphic, but should fit on one page. What's interesting, useful and surprising about your work?

- 12/13, 10:00am: **DUE**: **Final presentation slides**. Each team should submit their final slides to their mentor and to Jeremy. Feel free to send them earlier!

- 12/13: During the Final Presentations session, each team will present their project work to an audience of experts, entrepreneurs, and executives, including representatives from project sponsoring organizations, as well as a three outside judges. Teams will have 4 minutes to present their project work, plus 2 minutes for Q&A and judge remarks (6 minutes total per team). See the "Grading" section below for the judges' evaluation criteria.

Please note that teams are required to share your final report, summary of findings, and final presentation slides with project sponsors with enough lead time for them to review for inadvertent disclosure of Confidential Information.

Grading:

- 30% Final presentation content and delivery – team-wide; presentations will be evaluated according to the following criteria:
    - Technical and Analytical: How creative or advanced were the techniques used? Were they appropriate to the task and correctly applied?
    - Effort and Contribution: How much improvement was delivered? What alternative techniques were attempted before the team selected the one(s) that seemed best?
    - Business Impact: Beyond the data analytics described, how clearly did the team convey the bottom-line, real-world impact of their findings? What are the managerial or strategic implications? Can the potential benefits be quantified? Why do these results matter for the business or organization? What more general lessons can be learned? What are the next steps?
    - Presentation: How clear, informative and interesting was the presentation itself and how well was it delivered? Was it fun to see and hear? How did the team handle questions?
- 30% Final report – team-wide, using the same criteria as the presentation. Be sure to carefully and fully document all your references and data sources.
- 15% Summary of Findings – team-wide. Executive summaries are important documents.
- 15% Contribution to class discussions and team project enablement – individual. Students in the class are co-producers of class discussions and collective learning. Your contributions to this learning process all semester long will be appraised in addition to the specific content that you contribute.
    - Independently evaluated by instructor, mentors, and team members.
- 10% Mid-point presentation content and delivery – team-wide.

Required Book:

Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking, Foster Provost and Tom Fawcett. 2013. O'Reilly Media Inc. (Online access available at http://library.mit.edu/item/002221893)

Data Destruction:

The following states the MIT Action Learning Office's policies on data destruction:

Project sponsors share confidential and proprietary information to student teams doing Action Learning projects. MIT Sloan has an obligation to destroy that data at the end of the project so that it does not inadvertently get disclosed to unauthorized people and it is not used for any other purpose than the project.

MIT Sloan depends on the student teams for destroying the data in a timely and appropriate manner. Please note that destruction of data is a requisite step for the completion of course requirements.

What data is required to be destroyed?
Any information supplied by company in any format- emails, notes from a phone meeting, worksheets, records, company documents, any kind of company data. This includes data that is marked confidential and unmarked data. If the company supplied it, it must be destroyed at the end of the project.

What data is NOT required to be destroyed?
Students can keep their final paper and other derivative work that does not include company proprietary or confidential information. If there is any doubt, ask for help to discern what needs to be destroyed.

What are acceptable destruction methods?

- Printed Materials: Documents should be recycled in MIT approved secure recycle bins. Each academic area and many program offices have these bins.
- Digital Data Controlled by Students: If students have the data in Dropbox or on their computer, they must delete the data using appropriate tools.
- Digital Data Controlled by Sloan Technology Services: STS will destroy the data according to MIT Sloan IT policies.

If there are any issues or questions on this issue, please contact Ellen Baum, Contract Administration, at 3-5617 at ebaum@mit.edu.

Class Schedule:

	Date	Time	Session	Lecturer
S1	9/5	4:00-5:30	Welcome – Intro to Analytics and A-lab	Erik Brynjolfsson
S2	9/12	4:00-5:30	Social Analytics	Abdullah Almaatouq
S3	9/19	4:00-8:00	**Pitch Day**	
S4	9/26	4:00-5:30	1. The Emergence of Really Big Data; Universal Connectivity and Unified Network Theory for Humans and Machines 2. Legal considerations and guidelines	Sandy Pentland and Ellen Baum
OS	9/27 **Friday**	4:00-5:00	Optional Skill Seminar: Data Wrangling in R and Python	Sebastian Steffen
S5	10/3	4:00-5:30	Machine Learning/Deep Learning	Jeremy Howard via video
S6	10/10	4:00-5:30	ML Model Survey	Daniel Rock
S7	10/17	4:00-5:30	Causal Inference	Susan Athey
	10/24		No Class - SIP Week	
S8	10/31	4:00-5:30	**Mid-Point Presentations I**	
S9	11/7	4:00-5:30	**Mid-Point Presentations II**	
S10	11/14	4:00-5:30	Skill Seminar: NLP: in pursuit of means and meaning	Jay Alammar
S11	11/21	9:00-5:30	Optional: AI and Future of Work Congress	Multiple speakers at Kresge Auditorium
	11/28		No Class – Thanksgiving	
S12	12/5	4:00-5:30	Building Flexible NLP Pipelines for Sentiment Analysis	Zanele Munyikwa
S13	12/13 **Friday**	2:00-7:00	**Final Presentations Session**	

Reading List:

Session 1: *Welcome – Intro to Analytics* (Erik Brynjolfsson)

1. Review all project proposals and the A-lab syllabus.
2. "*Big Data: The Management Revolution*" Brynjolfsson, E. and McAfee, A. 2012. *Harvard Business Review*, 90(10); October: 60-68; http://libproxy.mit.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=79996279&site=ehost-live.
3. "*The Business of AI*" Brynjolfsson, E. and McAfee, A. 2017. *Harvard Business Review*, July; https://hbr.org/cover-story/2017/07/the-business-of-artificial-intelligence

Optional Reading:

4. "*Chapter 1: Introduction: Data Analytic Thinking*" Provost, F. and Fawcett T. 2013. *Data Science for Business*, O'Reilly Media Inc.; http://library.mit.edu/item/002221893.
5. *"The Unreasonable Effectiveness of Data"* Halevy, A., Norvig, P. and Pereira, F. 2009. *IEEE Intelligent Systems;* http://static.googleusercontent.com/media/research.google.com/en//pubs/archive/35179.pdf
6. *"The Rapid Adoption of Data-Driven Decision-Making."* Brynjolfsson, E. and McElheran, K. 2016. *American Economic Review*, 106(5): 133-39. https://www.aeaweb.org/articles?id=10.1257/aer.p20161016
7. "*Big Data: New Tricks for Econometrics*" Varian, H. 2014. *Journal of Economic Perspectives*, 28(2): 3-28; https://www.aeaweb.org/articles.php?doi=10.1257/jep.28.2.3.
8. "*Lectures on Machine Learning*" Athey, S. and Imbens, G. 2015. NBER; http://conference.nber.org/confer/2015/SI2015/ML/syllabus.pdf.
9. "*The Future of Prediction: How Google Searches Foreshadow Housing Prices and Sales*" Wu, L. and Brynjolfsson, E. 2014. *Economics of Digitization* (A. Goldfarb, S. Greenstein, and C. Tucker, eds.), Univ. of Chicago Press; http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2022293.

Session 2: *Social Analytics* (Abdullah Almaatouq)

1. "*Computational social science.*" Lazer, David, et al. 2009. *Science. 323*(5915), 721-723 https://gking.harvard.edu/files/LazPenAda09.pdf
2. "*Predicting poverty and wealth from mobile phone metadata*." Blumenstock, Joshua, Gabriel Cadamuro, and Robert On. *Science* 350.6264 (2015): 1073-1076. https://www.unhcr.org/innovation/wp-content/uploads/2016/11/blumenstock-science-2015.pdf
3. *"Unique in the shopping mall: On the reidentifiability of credit card metadata."* De Montjoye, Yves-Alexandre, Laura Radaelli, and Vivek Kumar Singh. *Science* 347.6221 (2015): 536-539. https://science.sciencemag.org/content/347/6221/536.full.pdf

Optional Reading:

4. Chapters 1-3 "*Bit by bit: Social research in the digital age*". Salganik, Matthew. Princeton University Press, 2019.

5. *"Mobile communication signatures of unemployment."* Almaatouq, Abdullah, Francisco Prieto-Castrillo, and Alex Pentland. *International conference on social informatics*. Springer, Cham, 2016. https://arxiv.org/pdf/1609.01778.pdf
6. "Unique in the crowd: The privacy bounds of human mobility." De Montjoye, Yves-Alexandre, et al. *Scientific reports* 3 (2013): 1376. https://www.nature.com/articles/srep01376
7. *"The world's technological capacity to store, communicate, and compute information"* Hilbert, Martin, and Priscila López. *Science* (2011). http://www.uvm.edu/pdodds/files/papers/others/2011/hilbert2011a.pdf

Session 3: *Pitch Day*

Optional Reading:

10. "*Chapter 2: Business Problems and Data Science Solutions*" Provost, F. and Fawcett T. 2013. *Data Science for Business*, O'Reilly Media Inc.

Session 4: *Modeling Transaction Behavior* (Sandy Pentland)

11. Dong, Xiaowen, Yoshihiko Suhara, Burçin Bozkaya, Vivek K Singh, Bruno Lepri, and Alex `Sandy' Pentland. 2017. "Social Bridges in Urban Purchase Behavior." *ACM Trans. Intell. Syst. Technol.* 9 (3): 33:1--33:29. http://web.media.mit.edu/~xdong/paper/tist2018.pdf
12. Suhara, Yoshihiko, Mohsen Bahrami, · Burçin, Bozkaya · Alex ', Sandy ' Pentland, Y Suhara, M Bahrami, A Pentland, and B Bozkaya. 2019. "Validating Gravity-Based Market Share Models Using Large-Scale Transactional Data." https://arxiv.org/abs/1902.03488

Optional Skill Seminar: Data Wrangling in R and Python (Sebastian Steffen)

Optional Readings:

13. Grolemund, G., & Wickham, H. (2017). Data transformation. In R for Data Science. Retrieved from https://r4ds.had.co.nz/transform.html
14. Grolemund, G., & Wickham, H. (2017). Tidy data. In R for Data Science. Retrieved from https://r4ds.had.co.nz/tidy-data.html
15. Grolemund, G., & Wickham, H. (2017). Pipes. In R for Data Science. Retrieved from https://r4ds.had.co.nz/pipes.html
16. Grolemund, G., & Wickham, H. (2017). Model building. In R for Data Science. Retrieved from https://r4ds.had.co.nz/model-building.html
17. 10 minutes to pandas — pandas 0.25.1 documentation. (2019). Retrieved September 14, 2019, from https://pandas.pydata.org/pandas-docs/stable/getting_started/10min.html
18. Grolemund, G., & Wickham, H. (2017). Dates and times. Retrieved from https://r4ds.had.co.nz/dates-and-times.html
19. tidyr 1.0.0 - Tidyverse. (2019). Retrieved September 14, 2019, from https://www.tidyverse.org/articles/2019/09/tidyr-1-0-0/
20. Preprocessing data — scikit-learn 0.21.3 documentation. (2019). Retrieved September 14, 2019, from https://scikit-learn.org/stable/modules/preprocessing.html#preprocessing
21. Supervised learning — scikit-learn 0.21.3 documentation. (2019). Retrieved September 14, 2019, from https://scikit-learn.org/stable/supervised_learning.html#supervised-learning

22. Decomposing signals in components (matrix factorization problems) — scikit-learn 0.21.3 documentation. (2019). Retrieved September 14, 2019, from https://scikit-learn.org/stable/modules/decomposition.html#decompositions
23. Model selection and evaluation — scikit-learn 0.21.3 documentation. (2019). Retrieved September 14, 2019, from https://scikit-learn.org/stable/model_selection.html#model-selection

Session 5: ***Machine Learning and Deep Learning*** **(Jeremy Howard)**

24. fast.ai lesson 1. https://course.fast.ai/videos/?lesson=1
25. Algorithmic Bias: https://www.youtube.com/watch?v=pThqge9QDn8&list=PLtmWHNX-gukKocXQOkQjuVxglSDYWsSh9&index=16

Optional Reading:

26. fast.ai lesson 2. https://course.fast.ai/videos/?lesson=2
27. fast.ai lesson 4. https://course.fast.ai/videos/?lesson=4

Session 6: ***Machine Learning Model Survey*** **(Daniel Rock)**

28. The Elements of Statistical Learning chapters 8-10, 14. https://web.stanford.edu/~hastie/Papers/ESLII.pdf
29. SVD Tutorial: https://blog.statsbot.co/singular-value-decomposition-tutorial-52c695315254
30. Scikit-Learn Tutorial: https://www.dataquest.io/blog/sci-kit-learn-tutorial/

Session 7: ***Causal Inference*** **(Susan Athey)**

Required Reading:

31. Athey, Susan. 2017. "Beyond Prediction: Using Big Data for Policy Problems." *Science* 355 (6324): 483 LP – 485. https://doi.org/10.1126/science.aal4321.
32. Uber paper posted on Canvas.
33. Athey, Susan, and Guido W Imbens. 2019. "Machine Learning Methods That Economists Should Know About." *Annual Review of Economics* 11 (1): 685–725. https://doi.org/10.1146/annurev-economics-080217-053433.
34. Athey, Susan, Julie Tibshirani, and Stefan Wager. 2019. "Generalized Random Forests." *Annals of Applied Statistics* 47 (2): 1148–78. https://doi.org/10.1214/18-AOS1709.

Session 10: ***NLP: in pursuit of means and meaning*** **(Jay Alammar)**

Optional Reading:

35. The Illustrated Word2vec https://jalammar.github.io/illustrated-word2vec/

Session 12: ***Building Flexible NLP Pipelines for Sentiment Analysis*** **(Zanele Munyikwa)**

Required Reading:

36. Section 2 of Text Mining with R: Sentiment Analysis with tidy data https://www.tidytextmining.com/sentiment.html
37. Read First Three Sections of Chapter 10: Representing and Mining Text (Introduction, Why Text is Important, Why Text is Difficult) From Provost, F. and Fawcett T. 2013. Data Science for Business, O'Reilly Media Inc.; http://library.mit.edu/item/002221893
38. Golder, Scott A., and Michael W. Macy. "Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures." Science 333, no. 6051 (2011): 1878-1881.

Optional Reading:

39. Hovy, Dirk, and Shannon L. Spruit. "The social impact of natural language processing." In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 591-598. 2016. https://www.aclweb.org/anthology/P16-2096.pdf
40. Kiritchenko, Svetlana, and Saif M. Mohammad. "Examining gender and race bias in two hundred sentiment analysis systems." arXiv preprint arXiv:1805.04508 (2018).