04-637 Mobile Big Data Analytics and Management

Preliminary version - subject to change

Instructor: Dr. Emily Aiken
- Postdoctoral scholar at CMU Africa
- Incoming assistant professor at UC San Diego
- Email: eaiken@andrew.cmu.edu
- Office: D206
- Office hours: TBD

Course meetings:
- Mondays and Wednesdays 10:00am - 11:45am, in person in room F205

Prerequisites: There are no formal prerequisites for the course, but **students are expected to be strong programmers**. Students can complete assignments in any programming language they feel comfortable with, but the primary language of instruction will be Python. Background in data science or machine learning will be helpful but not required.

Textbook: There is no textbook for this course. There will be 1-2 readings each week, primarily research papers which will be shared with the class on Canvas. Students will sign up to be responsible for summarizing readings on two classes over the course of the semester, and for critiquing and questioning readings on two classes.

Description and objectives: New sources of digital data from phones, mobile applications, social media companies, financial services providers, satellites, and other digital sensors are opening up new opportunities for societal-scale data analysis and inference. This course will define "mobile big data" broadly to introduce students to methods for storing and analyzing large digital traces, and to explore the potential and limitations of these data to learn about social and environmental phenomena. Much of the course will focus on equipping students with practical skills to work with mobile big data sources, including machine learning and data mining techniques specific to mobile big data's unique characteristics (multi-dimensional, personalized, spatiotemporal, sparse, and real-time). The course will also cover applications of mobile big data, particularly in topics related to sustainable development in Africa, as well as privacy and security challenges to working with mobile big data.

Learning outcomes: At the end of this course, students will be able to...
- Distinguish the characteristics of different types of mobile big data, including mobile phone metadata, internet traces, social media data, satellite data, and in situ sensing
- Apply appropriate data management techniques for mobile big data, including SQL and NoSQL
- Prepare and clean mobile big data for analysis
- Draw on a set of data analysis techniques, including supervised and unsupervised machine learning, natural language processing, geospatial techniques, and time series methods to analyze mobile big data
- Discuss the limitations of mobile big data for societal-scale analysis, and identify use cases where mobile big data may be more or less suited to answering research questions
- Assess the risks of mobile big data analysis, including issues related to privacy, security, fairness, and interpretability, and apply safeguards to mitigate these risks

Assignments and grading: There are no exams in this course. Grading will be based on...

1. **Homework assignments**, worth 8% of the grade each, and 45% in total. There will be five homework assignments for the course, all of which will require coding and data analysis along with critical thinking and writing. You will have two to three weeks to complete each homework assignment (see table below for details on assignment dates).
2. **A final project – conducted individually or in groups** – worth 45% of the grade in total. In addition to the final project paper, there are several interim project milestones, including a proposal, midterm report, and final presentation. See "final project" section below for details.
3. **Participation**, worth 15% of the grade. The participation grade will be based on your class participation (including asking and answering questions, participating class-wide discussions, participating in small group activities, and/or participating in online discussions on Canvas, worth 5% of the overall grade) and completion of in-class labs and assignments (10% of your total grade).

Assignment	Points	Assigned date	Due date
Homework (40% of grade)			
Homework #1: Mobility and mobile phone metadata	8	Wednesday 1/15	Wednesday 2/5
Homework #2: Tracking epidemics with social media data	8	Wednesday 2/5	Monday 2/24
Homework #3: Social media data and conflict	8	Monday 2/24	Monday 3/10
Homework #4: Satellite imagery and poverty prediction	8	Monday 3/10	Wednesday 3/26
Homework #5: Risks and limitations	8	Wednesday 3/26	Monday 4/14
Final project (45% of grade)			
Final project proposal	5	–	Wednesday 2/17
Final project midterm report and presentation	10	–	Monday 3/17
Final project presentation	10	–	Wednesday 4/16
Final research paper	20	–	Wednesday 5/1
Participation (15% of grade)			
In-class participation	5	–	–
Completion of in-class labs and participation assignments	10	–	–

Final project: Students will be expected to produce a high-quality research paper as the final project for this class. The final project should either (a) articulate and answer a novel research question related to mobile big data, including but not limited to any of the data sources covered in class, or (b) extend an existing research paper with novel analysis. **The final output of the project is a publication-quality research paper of 10-20 pages**. **The final project (due Wednesday may 1) can be completed alone or in groups; the standard expected from the final project will be commensurate with the size of the group**. There are several interim milestones for the final project, including a project proposal (due Wednesday February 12) and a midterm report and presentation (due Monday March 17). The teaching team will provide feedback on projects at each of these milestones. Early on during the course there will be in-class forums to discuss project ideas, data access, and form project teams for those who would prefer to conduct the final project in a group. At each project milestone, students will also be asked to submit reflections (written individually) on their work so far and, if conducting the project in a group, the contribution of each of their team members.

Weekly schedule (subject to change):

Day	Topic	Readings	Assignments
M 1/13	Introduction and overview of "mobile big data"	None	
Unit 1: Mobile phone metadata			
W 1/15	Mobile phone metadata overview	• Blondel et al. (2017). A survey of results on mobile phone datasets analysis. *EPJ Data Science.*	**Homework #1 released**
M 1/20	Data storage: SQL	• McGlynn and Santisteban (2007). Introduction to SQL.	
W 1/22	Data storage: Distributed processing Final project brainstorming	• PySpark tutorial: Getting started with PySpark (DataCamp)	
M 1/27	Application: Mobile phone metadata for measuring mobility Methods: Data mining	• Bagrow et al. (2011). Collective response of human populations to large-scale emergencies. *PLOS One.*	
M 1/29	Application: Mobile phone metadata for predicting poverty Methods: Machine learning	• Blumenstock et al. (2015). Predicting poverty and wealth from mobile phone metadata. *Science.*	
M 2/3	Application: Targeting humanitarian aid	• Aiken et al. (2022). Machine learning and mobile phone data can improve the targeting of humanitarian aid. *Nature.*	
Unit 2: Web and social media data			
W 2/5	Web and social media data overview Methods: Sourcing, pre-processing, and cleaning data Final project brainstorming	• Ghani et al. (2019). Social media big data analytics: A survey. *Computers in human behavior.* • Han and Kamber (2006). Data mining: Concepts and techniques. Chapter 3: Data preprocessing.	**Homework #1 due** **Homework #2 released**
M 2/10	Application: Google Search Trends and public health Methods: Time series methods Brief presentations of final project ideas	• Ginsberg et al. (2009). Detecting epidemics using search engine query data. *Nature.* • Lazer et al. (2014). The parable of Google Flu: Traps in big data analysis. *Science*.	
W 2/12	Application: Social media advertising data and digital gender gap	• Kashyap et al. (2020). Monitoring global digital gender inequality using the online populations of Facebook and Google. *Demographic Research.*	
M 2/17	Application: Social media and conflict, part 1 Methods: NLP	• Borge et al. (2015). Content and network dynamics behind Egyption political polarization on Twitter. *CSCW*.	**Final project proposal due**
W 2/19	Application: Social media and conflict, part 2 Methods: Social network analysis	• Schroader et al. (2023). Social media in the global South: A network dataset of the Malian Twittersphere. *Journal of data mining and digital humanities.*	
Unit 3: In situ sensing and the Internet of Things			
M 2/24	In situ sensing overview Application: Smart cities	• Sarker (2022). Smart city data science: Towards data-driven smart cities with	**Homework #2 due** **Homework #3 released**

		open research issues. *Internet of Things.*	
W 2/26	Application: Google Street View and demographic censuses Methods: Computer vision, crowdsourcing	• Gebru et al. (2017). Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the US. *PNAS.*	
M 3/3	Application: Flood prediction Methods: Real time streaming analytics	• Nearing et al. (2024). Global prediction of extreme floods in ungauged watersheds. *Nature.*	
Unit 4: Remote sensing			
W 3/5	Earth observation data overview	• Burke et al. (2021). Using satellite imagery to understand and promote sustainable development. *Science.*	
M 3/10	Application: Earth observation and poverty Methods: Deep learning	• Yeh et al. (2020). Using publicly available satellite imagery and deep learning to understand economic well-being in Africa. *Nature Communications.*	**Homework #3 due** **Homework #4 released**
W 3/12	Application: Earth observation and agriculture Methods: Deep learning	• Burke and Lobell (2017). Satellite-based assessment of yield variation and its determinants in smallholder African systems. *PNAS.*	
M 3/17	Midterm presentations	None	**Final project midterm due**
Unit 5: Risks and limitations			
W 3/19	Privacy and security (focus on mobile phone metadata)	• Taylor (2015). The ethics and analytics of tracking mobility with mobile phone data. *Environment and Planning.* • De Montjoye et al. (2018). On the privacy-conscious use of mobile phone data. *Scientific Data.*	
M 3/24	Fairness, bias, and representativity (focus on web and social media data)	• Blumenstock and Eagle (2012). Divided we call: Disparities in access and use of mobile phones in Rwanda. *ICTD.* • Milusheva et al. (2022). Assessing bias in smartphone mobility estimates in low income countries. *COMPASS.*	
W 3/26	Transparency and interpretability (focus on satellite data)	• Ledesma et al. (2020) Interpretable poverty mapping using social media data, satellite images, and geospatial information.	**Homework #4 due** **Homework #5 released**
M 3/31	Application: Targeting humanitarian aid	• Kahn et al. (2024). Expanding perspectives on data privacy: Insights from rural Togo.	
W 4/2	Topic TBD, catch-up day	None	
M 4/7	Spring break - No class		
W 4/9	Spring break - No class		
Course wrap-up and final project presentations			

M 4/14	Summary of class topics	None	**Homework #5 due**
W 4/16	Final project presentations		
M 4/21	Easter - No class		
W 4/23	Final project presentations		

Late work policy and extensions: To request an extension on an assignment deadline, please email the course instructor and teaching assistant **at least 24 hours before the assignment deadline. No extensions will be granted within 24 hours of the assignment deadline.** Extension requests will be evaluated on a case-by-case basis. For homework assignments, in-class lab assignments, and final project interim milestones (except the final research paper), 10% of the grade will be deducted for each day an assignment is late (without an approved extension). No extensions will be granted to the final research paper deadline (Wednesday May 1).

Generative AI policy: For coding assignments, students may use generative AI to help with coding and syntax. For writing assignments, generative AI may be used for brainstorming, but **no part of the writing assignment may be written by generative AI**. Writing assignments will be screened for AI generation, and assignments that have substantial portions written by generative AI will be given zero credit. If you use generative AI for any of the allowable uses, such as checking code syntax and brainstorming ideas, **you must provide a written statement at the top of your assignment explaining how you used generative AI**.

Academic integrity: Students are encouraged to talk to each other, to the TA, to the instructor, or to anyone else about any of the homework assignments. Any assistance, though, must be limited to discussion of the problem and sketching general approaches to a solution. Each student must write out his or her own homework solutions. Consulting another student's solution is prohibited and submitted solutions may not be copied from any source. Copying from someone else's homework, lab write-up, or exam or allowing another student to copy his/her work, will be considered as cheating. If you have any question about whether some activity would constitute cheating, please feel free to ask.

If you even suspect that you have collaborated with any other person or taken help from online forums, or used material from elsewhere, list their name(s)/sources at the top of the first page of your assignment. Also send an email to the TA/instructor saying that you have used external resources for a particular assignment. This email must be sent BEFORE the assignment submission deadline. Sending this email does not excuse you from charges of plagiarism. It may merely reduce the impact. Instead of getting failed in the course, you may be given a zero on the assignment.

Refer to CMU's Academic Integrity Policy at Academic Integrity - University Policies for further information. You are encouraged to review the three policy violations i.e., cheating, plagiarism, and unauthorized assistance.

You may not have much experience with CMU's Academic Integrity Policy and the issues are sometimes subtle. If you have any question about what the right thing to do, just ask. Any of the faculty or teaching assistants can provide guidance. If you are uncomfortable asking then you probably already know that you are violating the academic integrity policy and should change your actions accordingly.

Accommodations for students with disabilities: Accommodations for students with disabilities will be granted in accordance with CMU's procedure for accommodations. If you have an accommodation from CMU, please contact me at the start of the semester to inform me of your accommodation.

Student well-being: We encourage you to seek a healthy balance during this semester. Universities are in general vibrant communities, places of tremendous vitality and richness that offer abundant opportunities for meaningful work and play. This abundance brings with it the challenge of maintaining a healthy, balanced life – a life characterized by productive tension among such competing needs as work and play, sleep and wakefulness, and solitude and sociability. All members of university communities – students, staff, and faculty – have the responsibility to promote balance in their lives by making thoughtful and balanced choices. I encourage you to make balanced choices.

Balanced choices result from an understanding that human flourishing requires the fulfillment of very real physical, emotional, spiritual, and social needs. Therefore, an understanding that failure is part of the road to success in all endeavors, whether academic, extracurricular, or social is critical to striking a balance between expectations of success and mitigating incidents of failure.

Make sure to move regularly, eat well, and reach out to your support system or me eaiken@andrew.cmu.edu if you need to. We can all benefit from support in times of stress, and this semester is no exception.

Diversity statement: We commit to creating a safer, more inclusive environment that encourages students from diverse backgrounds and perspectives to realize their full potential during this course. This environment should inspire classroom discussion, and respectful and empathetic discourse, where each participant is treated with respect and dignity. I encourage you to abide by this diversity statement for the well-being of all participants.