IE 529
Statistics of Big Data and Clustering
MWF 1-1:50pm, 2200 Lu MEB

Instructor: Prof. Carolyn Beck, office 1270A DCL, *beck3@illinois.edu*
TA: Vincent Leon, office hours TBA, *leon18@illinois.edu*

Reading material: Journal articles, lectures notes, slides and other readings will be posted to the course website; students are expected to check the website regularly and read the posted material (on Canvas).

Prerequisites: A basic course in probability and statistics (e.g., at or above the level of IE 300); an introductory course in linear algebra (e.g., at the level of MATH 415/416); the standard engineering calculus sequence; a bit of real analysis helps.

Tentative List of Course Topics:

Weeks 1-5 **Introduction and Foundations:** (1) brief introduction to big data and clustering; (2) overview of mathematical statistical concepts and principles, including notions of convergence; (3) review of hypothesis testing and basic statistical inference as needed. **See Readings 1-3**

Weeks 5-7 **More Foundations:** (1) review of linear algebra concepts: vector and matrix norms; (2) decompositions, including the singular value decomposition (SVD). **See Reading 4**

Weeks 7-9 **Regression Methods:** (1) linear regression (model and interpretation; model fitting; model diagnosis); (2) logistic regression; (3) local polynomial regression; (4) model evaluations using AIC and BIC. **See Readings 5, 7**

Weeks 9-11 **Dimension Reduction:** (1) principal component analysis (PCA), i.e., the SVD revisited; (2) canonical analysis **See Readings 6, 9**

Weeks 11-14 **Clustering and Classification:** (1) k-means, k-centers, k-medians, k-means++; (2) spectral clustering; (3) expectation-maximization; (4) deterministic annealing. **See Readings 8, 10-13**

Weeks 14-15 **Advanced Topics:** TBD (possibilities include (1) networks and graphs; (2) recommender systems and matrix completion problems; (3) neural networks). *Readings to be posted.*

Assignments:

- There will be 3-4 homework assignments ($\approx 30-35\%$ of grade, analytical/mathematical in style, may include questions based on the readings).
- There will be 2-3 computational (i.e., programming) assignments ($\approx 30-35\%$ of grade, application-oriented in style with given data sets).
- There will be 2 one-hour "in-class" quizzes (20% of grade): these will be **in class/timed-online**; dates to be announced on Canvas (analytical/mathematical material).
- There will be 1 team project based on a journal paper ($\approx 10\%$ of grade), which will include coding and implementing examples using test case data, and writing a summary report of the article (analytical and application components).
- Attendance/participation may be included at 5% - to be discussed.

Programming languages:
There are no strictly required coding languages for the assignments. If you use Matlab, Prof. Beck can help you. If you use Python, TA Leon can help you. Any other coding languages we will try to help.