

# Foundations of Data Science syllabus

Wesley Tansey

August 1, 2023

This course will cover the foundations of modern data science from a probabilistic modeling perspective. We will cover the basics of statistical modeling: likelihoods, priors, and posteriors. We will compare and contrast different ways to fit these models, focusing on the trade-offs made between computation and objectives like uncertainty quantification or accuracy. There will be 3 homeworks, each having both problem sets and a programming task. Some lectures will have assigned readings that must be completed before the lecture begins; readings will be sent out the week before the lecture.

**Final project.** There is a final project for the course that accounts for the bulk of the overall grade. A mid-semester report is due with the expectation of substantial progress made. A report on the project is due at the end of the course along with a short presentation. Students should budget substantial time for the final project throughout the entire course.

**Meeting times.** The course will meet Mondays and Wednesdays, 10am-11:30m, in the lower level conference room of the Joy building (321 E 61st St.). All lectures are entirely in-person.

**Prerequisites.** Students are expected to have a working knowledge of linear algebra, calculus, and basic probability. It is also expected that they will have working knowledge of python (numpy, scipy, matplotlib).

**Asking for help.** A Slack channel is available for discussions and questions. All students enrolled will receive an invite automatically. Any auditing students that wish to be invited to the channel should email me.

**Topics.** Below are topics covered in the class. The precise ordering may change through the semester.

- **Background / refresher.** Algebra, calculus, linear algebra, probability, python programming.
- **Intro to statistical inference.** Likelihoods, maximum likelihood estimation, bias of an estimator, variance of an estimator, bias-variance decomposition of MSE, bias-variance trade-off.

- **Convex optimization.** Convexity, logistic regression, gradient descent, Newton's method, coordinate descent.
- **Stochastic optimization.** Stochastic gradient descent, automatic differentiation, pytorch.
- **Deep learning.** Neural networks, convolutional nets, transformers, residual nets, batch norm, U-nets.
- **Bayesian modeling.** Priors, Bayes' rule, MAP inference, prior/penalty duality, L1 and L2 penalties.
- **Bayesian analysis.** Posterior distributions, uncertainty intervals, posterior predictive, posterior predictive checks, model criticism, Box's loop.
- **Graphical models.** Exponential families, conjugate priors, hierarchical models, plate diagrams.
- **Bayesian posterior inference.** Benefits of Bayesian inference in non-convex problems, MCMC, Gibbs sampling, Metropolis-Hastings, computational issues for MCMC.
- **Scalable approximate Bayesian inference.** Empirical Bayes, variational inference, Hamiltonian Monte Carlo, trade-offs of hybrid methods.
- **Deep generative models.** Amortized inference, variational autoencoders, deep empirical Bayes.
- **Causal inference.** Confounders, mediators, colliders, D-separation, structure learning, causal estimation, potential outcomes, estimands (ATE, ITE), double robustness, Bayesian causal inference, missing data perspective, sensitivity analysis.
- **Applications.** Illustrative examples of common types of modeling problems:
  - **Matrix factorization.** Alternating minimization, identifiability, interpretability, cross-validation, strong generalization, tensors, multi-view factorization.
  - **Mixture models.** Latent variable models, expectation maximization, data augmentation.
  - **Spatiotemporal data.** Time series, spatial data, graph data, trend filtering, Gaussian processes.