Introduction to Machine Learning and Data Science

Joshua Blumenstock

Harnessing Big Data for National Statistics and Evidence-Based Policymaking

Blantyre, August 2025

This talk: Outline

Introduction to data science and machine learning

- Traditional statistics vs. data science and machine learning
- What is Machine Learning (and why is it relevant to public policy)?
  - Examples of ML in policymaking
  - Supervised vs. unsupervised learning
  - Key concepts in machine learning
  - Important tradeoffs and considerations
- Concluding remarks

Key Concepts (today's lecture)

By the end of the lecture, my hope is that you'll have some understanding of:

- What is "Data Science"?
- Machine learning vs. traditional programming vs. traditional statistics
- Supervised learning vs. unsupervised Learning
- Key concepts: representation, evaluation, generalization, overfitting
- Cross-validation
- Why is machine learning increasingly relevant to policy?

From traditional statistics to data science

We stand at an **inflection point**

- "…we are on the cusp of a tremendous wave of innovation, productivity, and growth… all driven by big data…" (McKinsey Global)

Driven by "Big Data" (Volume, Velocity, Variety)

Remote sensing: High-resolution imagery available daily

Mobile phones: 96% mobile phone penetration globally

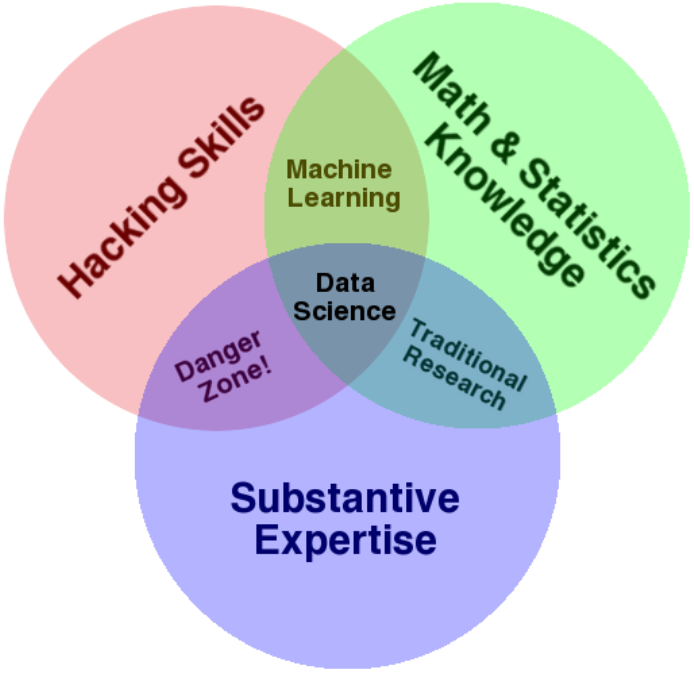
Digital traces: 3.2 billion active Facebook users

Other sources: …

What is Data Science?

New (and old) tools for working with big data

- "A set of fundamental principles that support and guide the principled extraction of information and knowledge from data"



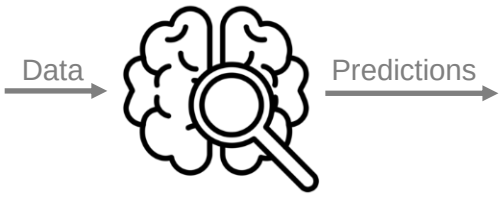
Key data science skills

- **Domain expert:** Understanding the nuances of the problem
- **Computer scientist:** Algorithms, data structures, programming
- **Analyst:** Statistics, machine learning, econometrics

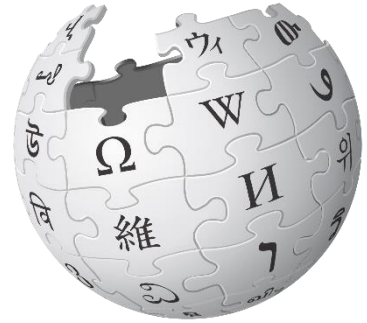
What is Machine Learning?

What is machine learning?

- **Informal definition**: Algorithms that learn from data to make predictions



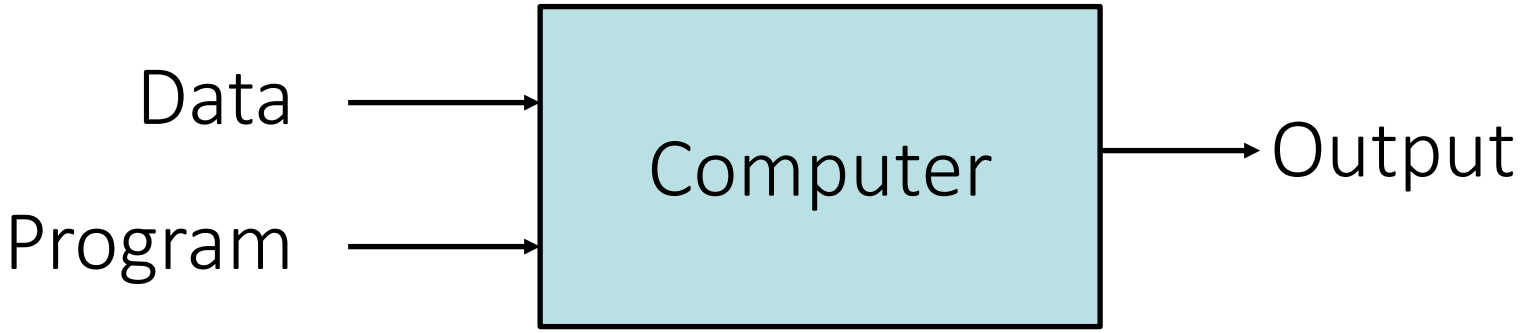
- **Wikipedia**: A field of study in artificial intelligence concerned with the development and study of statistical algorithms that can learn from data and generalize to unseen data, and thus perform tasks without explicit instructions



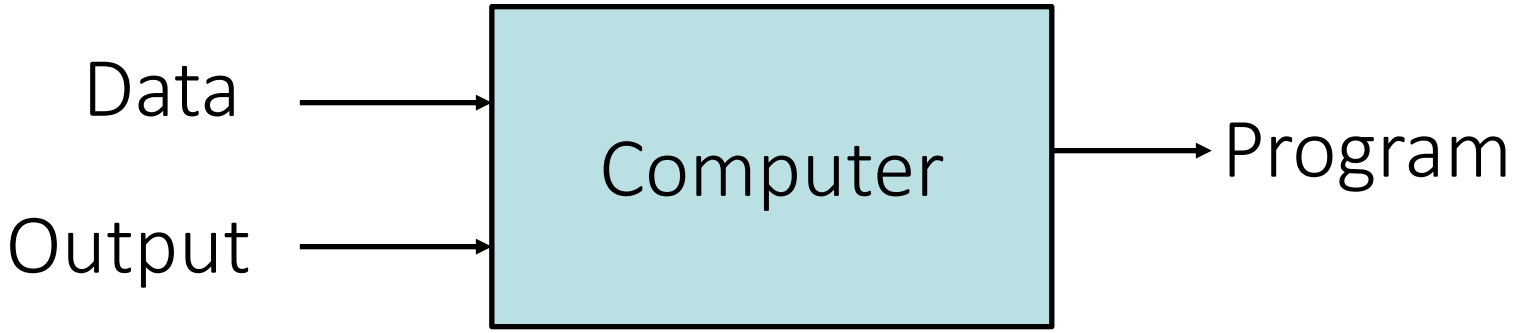
- **Analogy**: Like teaching a child by showing examples rather than giving step-by-step instructions

ML vs programming

Traditional Programming



Machine Learning



"Learning from examples vs. hard-coded rules"

ML vs traditional statistics

Traditional statistics

- We start with a model $f(\cdot)$, that relates inputs X to outputs Y
  - For instance, linear regression (OLS): $Y = \alpha + \beta X + \epsilon$
- We often care about causal relationships between X and Y (i.e., $\hat{\beta}$)
  - 💡 How we specify $f(\cdot)$ is critical – the validity of inferences depends on it

Machine learning:

- We start with a model $f(\cdot)$, that relates inputs X to outputs Y
- Now, our focus is on *accurate predictions of* $\hat{Y}$
  - 💡 Now, we can use much more creative approaches to $f(\cdot)$
  - In other words, ML prioritizes **predictive accuracy** over **interpretability**

This talk: Outline

Introduction to data science and machine mearning

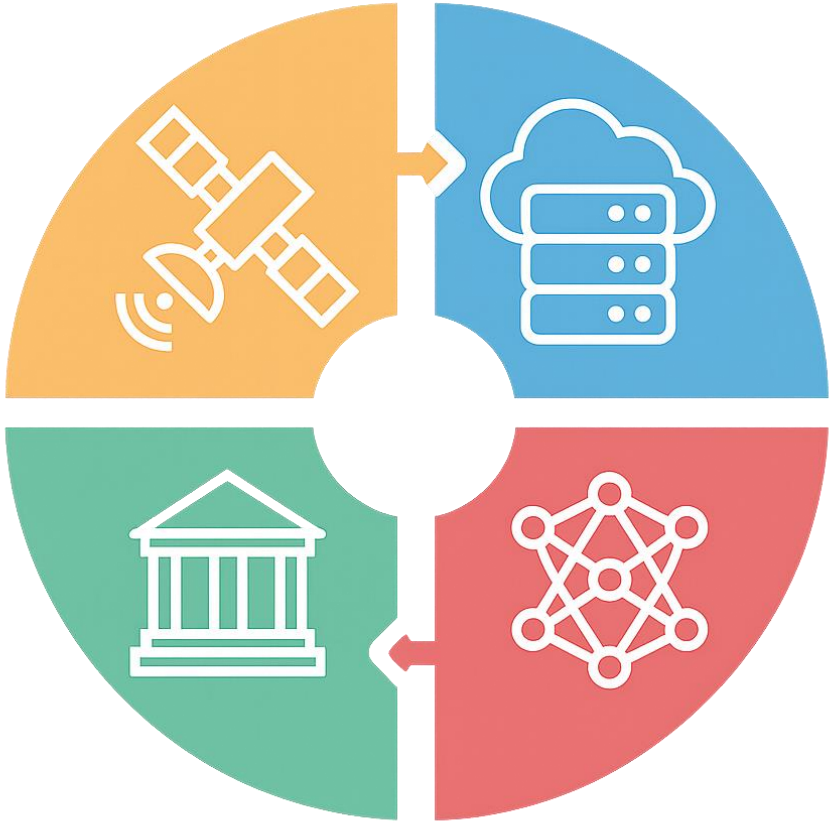
- Traditional statistics vs. machine learning
- **What is Machine Learning (and why is it relevant to public policy)?**
  - Examples of ML in policymaking
  - Supervised vs. unsupervised learning
  - Key concepts in machine learning
  - Important tradeoffs and considerations
- Concluding remarks

ML and policymaking

Why is machine learning "suddenly" relevant to policymaking?

Explosion in data availability

Satellites, remote sensing, mobile phones, digital government records, admin data, networked sensors



Advances in computing and cloud infrastructure

High-speed computing and affordable cloud storage

Policy demands for timely, evidence-based decisions

Ability to systematically maximize impact, real-time insights to inform rapid interventions

Improved algorithms and open-source tools

Modern machine learning tools integrated into common frameworks, point-and-click dashboards

ML and policymaking

How can machine learning be **useful** in policymaking?

Filling data gaps

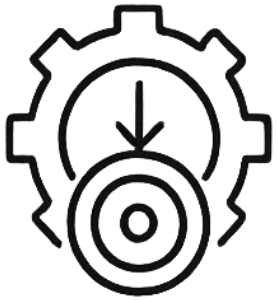
ML-based predictions can help fill gaps in traditional data collection infrastructure

Predictive modeling

Can detect patterns too complex for humans, even in massive and heterogeneous datasets

National-scale, real-time analysis

Enables program monitoring and evaluation at a national-scale in real-time



Rapid and effective interventions

Limited resources can be targeted efficiently to maximize impact

Examples of ML in policymaking

Common applications of ML (have been around for decades!)

Proxy means tests: Observable proxies (e.g., roof type) are used to estimate (*predict*) the living standards of a household

Credit scoring: Financial transactions are used to *predict* whether someone will pay back a loan

Weather forecasts: Current (and historical) meteorological data are used to *predict* local weather conditions in the future

- And much more!
  - Public health surveillance and risk prediction
  - Tax evasion and fraud detection
  - Agricultural monitoring

Examples of ML in policymaking

Modern applications of ML (last few years)

Satellite-based poverty maps: Observable proxies (e.g., satellite pixels) are used to *predict* the living standards of a household

Digital credit: Phone and app data are used to *predict* whether someone will pay back a loan

Hunger forecasts: Current (and historical) meteorological data are used to *predict* acute malnutrition

- And much, much more!
  - Medical diagnosis, prognosis, treatment; Hiring, firing, recruiting, admissions
  - Parole, bail, sentencing, policing, military, security decisions, screening
  - e-Government, criminal/tax audits, content recommendations

This talk: Outline

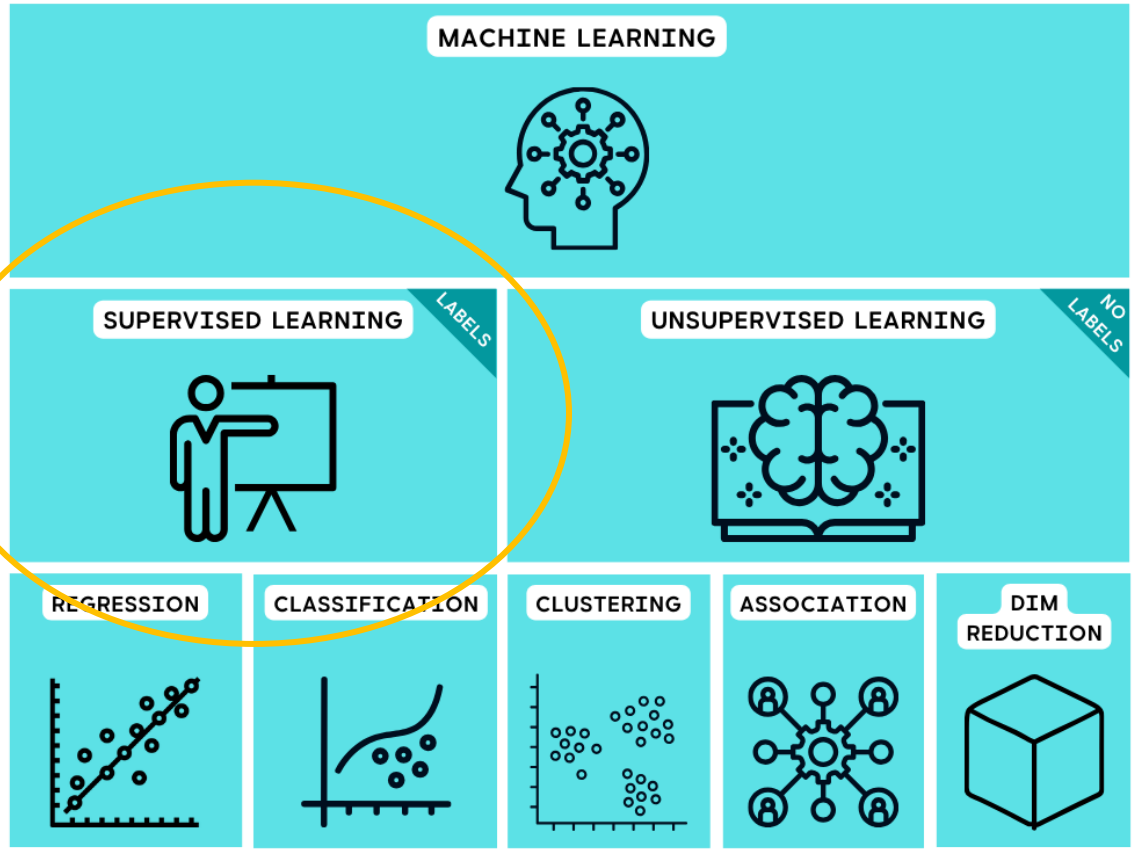
Introduction to data science and machine mearning

- Traditional statistics vs. data science
- What is Machine Learning (and why is it relevant to public policy)?
  - Examples of ML in policymaking
  - **Supervised vs. unsupervised learning**
  - Key concepts in machine learning
  - Important tradeoffs and considerations
- Concluding remarks

Supervised vs. Unsupervised learning

Most of these are examples of **supervised learning**

- This means we have labels; we know the "right answer" for some data



Supervised vs. Unsupervised learning

Key differences between supervised and unsupervised learning

	Supervised learning	Unsupervised learning
Input data	Data have labels	No labels
When to use	We have inputs X and labels Y and want to model the X-Y relationship	We just have inputs X and want to better understand the structure of X
Goal	Predict Y for new X	Discover patterns in X
Common methods	• Linear models (linear/logistic regression) • Decision trees, random forests • Neural networks, ensemble methods • Deep learning	• K-means and hierarchical clustering • Principal Component analysis • SVD, NMA, LDA

Other approaches to ML

Semi-supervised learning

- Combines a small amount of labeled data with a large amount of unlabeled data to improve learning efficiency and accuracy.

Reinforcement learning

- An agent learns to make decisions by interacting with an environment and receiving rewards or penalties, optimizing a long-term objective.

Transfer learning

- Leverages knowledge from one task or domain (where data is plentiful) to improve performance in another task (where data is scarce).ties, optimizing a long-term objective.

And more!

- Adversarial learning
- Online learning
- FATE: Fair, Accountable, Transparent, and Ethical ML

This talk: Outline

Introduction to data science and machine mearning

- Traditional statistics vs. data science
- What is Machine Learning (and why is it relevant to public policy)?
  - Examples of ML in policymaking
  - Supervised vs. unsupervised learning
  - **Key concepts in machine learning**
  - Important tradeoffs and considerations
- Concluding remarks

Key concepts in machine learning

Three key principles of modern machine learning

1. **Representation**: How do we model relationships in the data?
   - Examples: Linear regression, random forest, gradient boosting, deep neural nets, …
   - Insight 💡 : To enable *accurate prediction,* we often explore many representations

2. **Evaluation**: How do we know if our model is effective?
   - Accuracy, error rates, precision/recall, R^2
   - Insight 💡 : Traditional ML focuses on stats; policies may define other objectives

3. **Generalization:** The fundamental goal of machine learning is to generalize beyond data that has already been seen

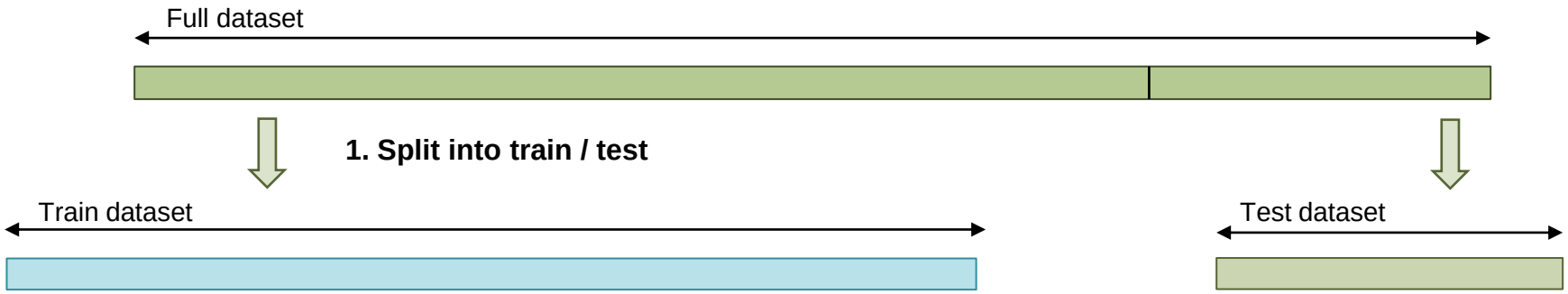
Generalization via Training and Testing

How can we ensure that a model will generalize?

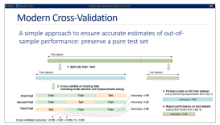
- **"Generalize"** means he model will perform well on new data

A simple – and effective – trick: Split data into a training set and a test set

- **Training set**: used to fit the model
- **Test set**: used to evaluate performance on new data
- This is a simple way to see how model generalizes (from train to test)



Generalization via Cross-Validation

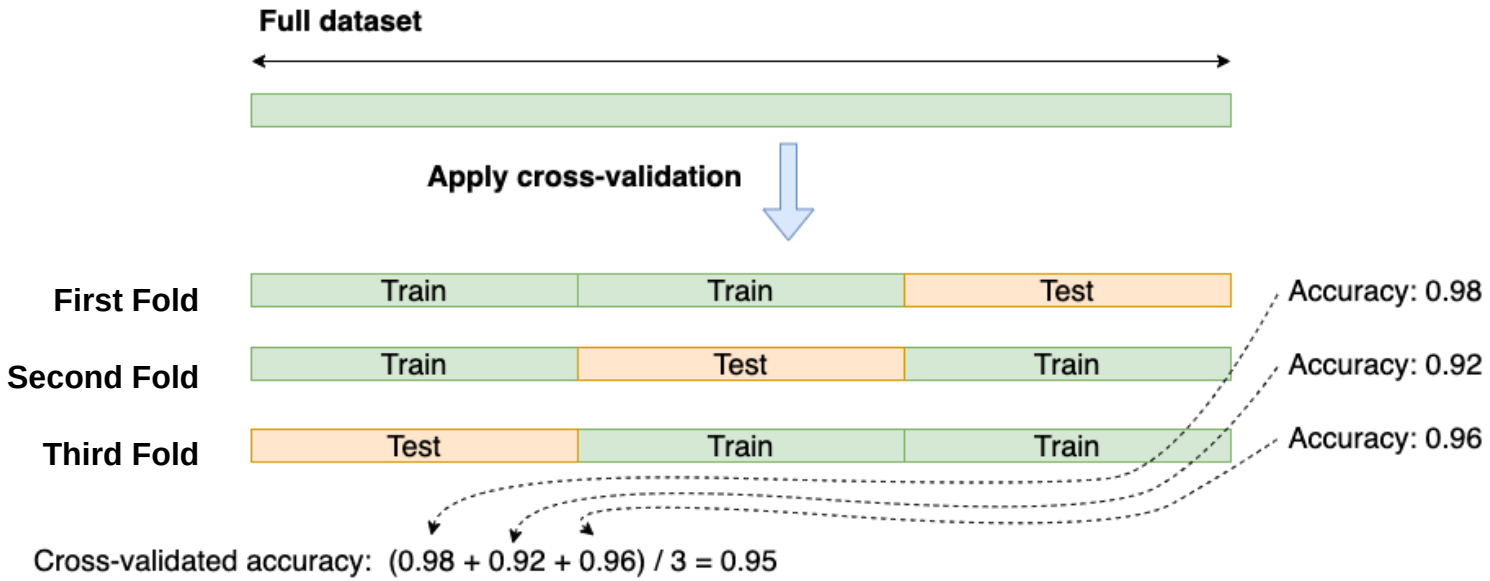


A single test-train split is good, but not great

- With a single split, we might get unlucky
- We might also accidentally "overfit" to our test data

A better idea: **Cross-validation**

- Data are split into *k* random "folds"; each fold is used as test set once



Model complexity and regularization

💡 A really, really important point 💡

- Careful cross-validation establishes a framework for evaluating many models, to rigorously determine what works best

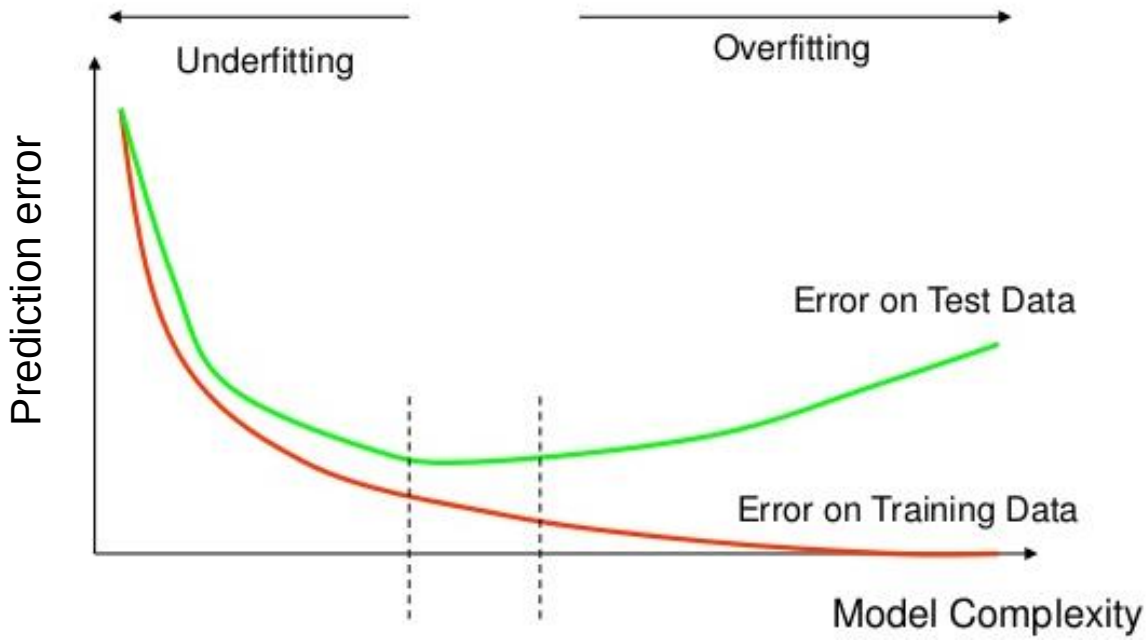
We can determine (for instance)

- Which is more accurate: linear regression or random forests?
- What features (predictors) should be included in the model?
- Should our random forest have 100 trees or 1000 trees?

Model complexity and regularization

Generalization vs. Overfitting

- A model is "overfit" if it is accurate on the training data well, but is not accurate on the test data
- This often happens when the model is too flexible/complex:



This talk: Outline

Introduction to data science and machine mearning

- Traditional statistics vs. data science
- What is Machine Learning (and why is it relevant to public policy)?
  - Examples of ML in policymaking
  - Supervised vs. unsupervised learning
  - Key concepts in machine learning
  - **Important tradeoffs and considerations**
- Concluding remarks

Tradeoffs

Machine learning algorithms optimize for *predictive accuracy*

- By default, algorithms are designed to minimize prediction error

But: many policy settings require balancing multiple objectives

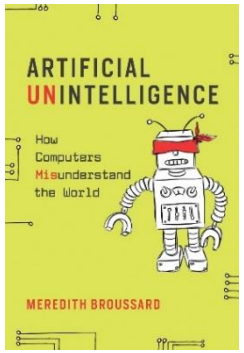
- Accuracy
- Interpretability / explainability
- Resource requirements
- Fairness
- Auditability

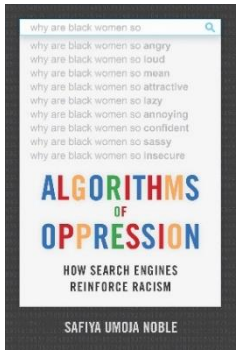
Achieving this balance is possible, but requires attention & care

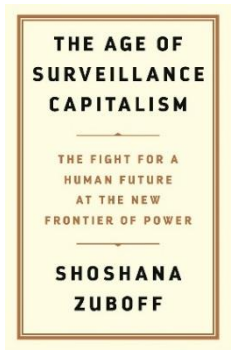
Other considerations

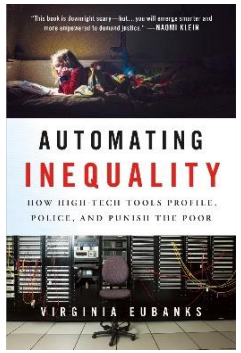
Machine learning is far from a silver bullet!

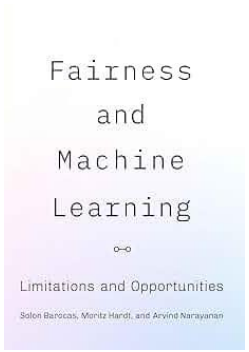
- Many potential sources of bias ("garbage in, garbage out")
  - Historical bias, Representation bias, Measurement bias, Aggregation Bias, Learning Bias, Evaluation bias, Deployment bias
- Opacity and interpretability
- Localization / adaptation / switching costs
- Hype and Hubris

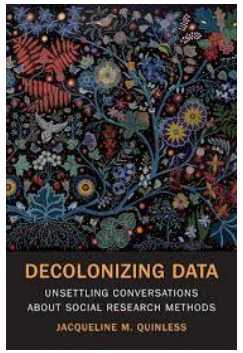


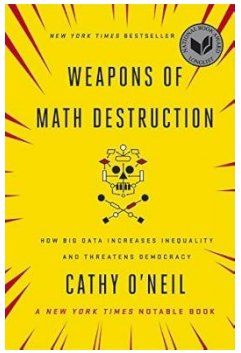


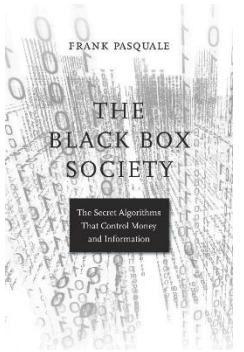


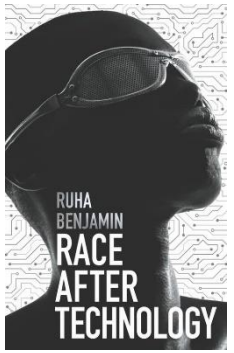


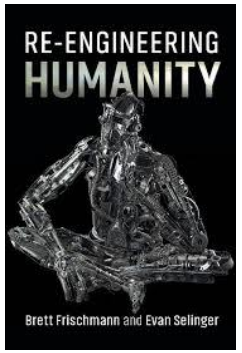












This talk: Outline

Introduction to data science and machine mearning

- Traditional statistics vs. data science
- What is Machine Learning (and why is it relevant to public policy)?
  - Examples of ML in policymaking
  - Supervised vs. unsupervised learning
  - Key concepts in machine learning
  - Important tradeoffs and considerations
- **Concluding remarks**

ML and data science: Key takeaways

1. **What is ML**? Machine learning enables systems to learn from data and make predictions — without being explicitly programmed
2. **Why care?** Advances in data availability, computing power, and algorithms make ML newly relevant for government use
3. **How to use?** Effective use of ML requires careful adaptation; algorithms rarely work immediately out-of-the-box

Take a 5 minute break!