6.7900: Machine Learning
Lecture 24

Lecture start: Tues/Thurs 2:35pm
Who's speaking today? Prof. Tamara Broderick
Course website: gradml.mit.edu
Questions? Ask here or on piazza.com/mit/fall2024/67900/
Materials: Slides, video, etc linked from gradml.mit.edu after the lecture (but there is no livestream)

Last Times

I. Lots of methods and algorithms for machine learning
II. Some insight into when they work and when they don't work

Today's Plan

I. What's left in this class
II. A broader perspective on using ML to make decisions
    A. Some (more) ways to do ML wrong

We’re nearing the end!

We’re nearing the end!

- Review problem session from staff
  - Probably Tues Dec 10 in class, but stay tuned for confirmation
  - Dec 11 is the last day of MIT classes for Fall 2024

We’re nearing the end!

- Review problem session from staff
  - Probably Tues Dec 10 in class, but stay tuned for confirmation
  - Dec 11 is the last day of MIT classes for Fall 2024
- Extra office hours on Fri 12/6, 10am–12noon, for Project 2 (due Fri night)

We're nearing the end!

- Review problem session from staff
  - Probably Tues Dec 10 in class, but stay tuned for confirmation
  - Dec 11 is the last day of MIT classes for Fall 2024
- Extra office hours on Fri 12/6, 10am–12noon, for Project 2 (due Fri night)
- Final exam info is up at https://gradml.mit.edu/main/exams/
  - Fri 12/20 9am–12noon
  - Dupont Gym
  - 2 "cheat sheets" (details at link)

We're nearing the end!

- Review problem session from staff
  - Probably Tues Dec 10 in class, but stay tuned for confirmation
  - Dec 11 is the last day of MIT classes for Fall 2024
- Extra office hours on Fri 12/6, 10am–12noon, for Project 2 (due Fri night)
- Final exam info is up at https://gradml.mit.edu/main/exams/
  - Fri 12/20 9am–12noon
  - Dupont Gym
  - 2 "cheat sheets" (details at link)
- Questions? Piazza, office hours, problem session

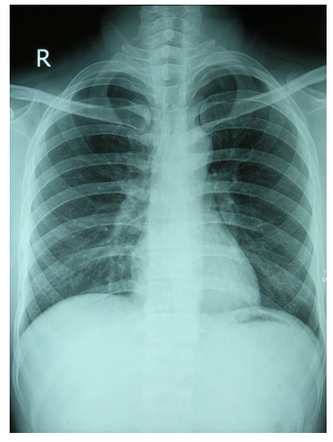
We're nearing the end!

- Review problem session from staff
  - Probably Tues Dec 10 in class, but stay tuned for confirmation
  - Dec 11 is the last day of MIT classes for Fall 2024
- Extra office hours on Fri 12/6, 10am–12noon, for Project 2 (due Fri night)
- Final exam info is up at https://gradml.mit.edu/main/exams/
  - Fri 12/20 9am–12noon
  - Dupont Gym
  - 2 "cheat sheets" (details at link)
- Questions? Piazza, office hours, problem session
- Subject evaluations are open!
  - We take your ideas and opinions very seriously, and they will shape the experience for next year's students.

When can I trust decisions from data?

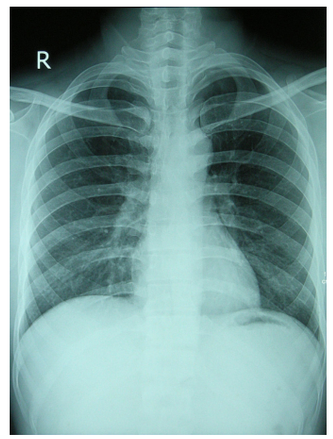
When can I trust decisions from data?

- More data & better computation ➔ data analyses increasingly drive life-changing decisions



When can I trust decisions from data?

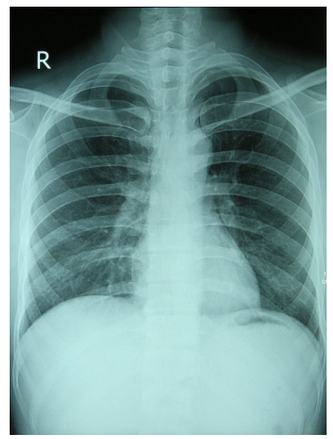
- More data & better computation ➔ data analyses increasingly drive life-changing decisions



- A typical setup: (1) run a data analysis on available data to reach a decision; (2) apply that decision to new data

When can I trust decisions from data?

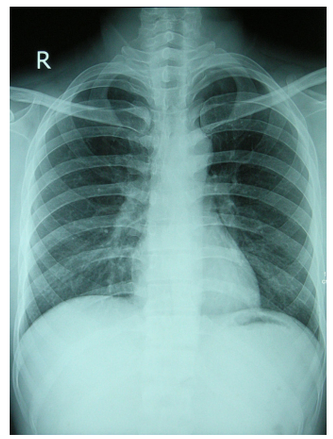
- More data & better computation ➔ data analyses increasingly drive life-changing decisions



- A typical setup: (1) run a data analysis on available data to reach a decision; (2) apply that decision to new data
  - E.g. (1) conclude a medical or economic intervention helps, then (2) distribute it

When can I trust decisions from data?

- More data & better computation ➔ data analyses increasingly drive life-changing decisions

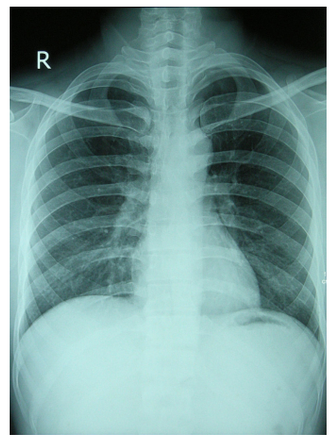


- A typical setup: (1) run a data analysis on available data to reach a decision; (2) apply that decision to new data
  - E.g. (1) conclude a medical or economic intervention helps, then (2) distribute it
- Might worry about *generalization* if replicability fails: E.g. in repeat of 100 psych experiments, <1/2 had same result

[Open Science Collaboration, 2015]

When can I trust decisions from data?

- More data & better computation ➔ data analyses increasingly drive life-changing decisions



- A typical setup: (1) run a data analysis on available data to reach a decision; (2) apply that decision to new data
  - E.g. (1) conclude a medical or economic intervention helps, then (2) distribute it

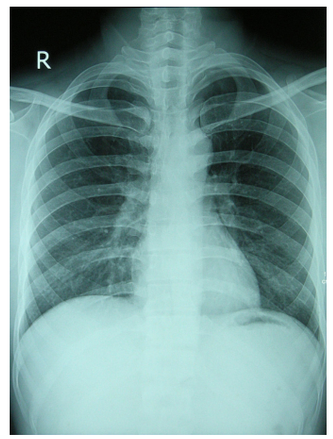
colloquial usage

- Might worry about *generalization* if replicability fails: E.g. in repeat of 100 psych experiments, <1/2 had same result

[Open Science Collaboration, 2015]

When can I trust decisions from data?

- More data & better computation ➔ data analyses increasingly drive life-changing decisions



- A typical setup: (1) run a data analysis on available data to reach a decision; (2) apply that decision to new data
  - E.g. (1) conclude a medical or economic intervention helps, then (2) distribute it

colloquial usage

- Might worry about *generalization* if replicability fails: E.g. in repeat of 100 psych experiments, <1/2 had same result

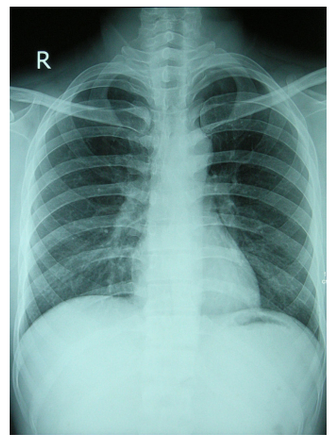
[Open Science Collaboration, 2015]

  - Of 53 hematology/oncology papers, 6 had same result

[Begley, Ellis 2012]

When can I trust decisions from data?

- More data & better computation ➔ data analyses increasingly drive life-changing decisions



- A typical setup: (1) run a data analysis on available data to reach a decision; (2) apply that decision to new data
  - E.g. (1) conclude a medical or economic intervention helps, then (2) distribute it

colloquial usage

- Might worry about *generalization* if replicability fails: E.g. in repeat of 100 psych experiments, <1/2 had same result

[Open Science Collaboration, 2015]

  - Of 53 hematology/oncology papers, 6 had same result

[Begley, Ellis 2012]

- **Today**: how generalization could fail when everyone is well-meaning & mastered all their coursework

Example analysis: microcredit

Example analysis: microcredit

- Making good decisions from data typically requires a lot of work and a lot of hard, necessarily-imperfect choices

Example analysis: microcredit

- Making good decisions from data typically requires a lot of work and a lot of hard, necessarily-imperfect choices

Real-world goal: Want to know if microcredit helps people

Example analysis: microcredit

- Making good decisions from data typically requires a lot of work and a lot of hard, necessarily-imperfect choices

Real-world goal: Want to know if microcredit helps people

1. Have to turn this goal into measurements on actual data

Example analysis: microcredit

- Making good decisions from data typically requires a lot of work and a lot of hard, necessarily-imperfect choices

Real-world goal: Want to know if microcredit helps people

1. Have to turn this goal into measurements on actual data
   - What counts as “help”? E.g. increase business profit

Example analysis: microcredit

- Making good decisions from data typically requires a lot of work and a lot of hard, necessarily-imperfect choices

Real-world goal: Want to know if microcredit helps people

1. Have to turn this goal into measurements on actual data
   - What counts as “help”? E.g. increase business profit
   - Who gets measured? Particular people, location, etc

Example analysis: microcredit

- Making good decisions from data typically requires a lot of work and a lot of hard, necessarily-imperfect choices

Real-world goal: Want to know if microcredit helps people

1. Have to turn this goal into measurements on actual data
   - What counts as “help”? E.g. increase business profit
   - Who gets measured? Particular people, location, etc
   - How is data collected? Survey: In-person, phone, mail

Example analysis: microcredit

- Making good decisions from data typically requires a lot of work and a lot of hard, necessarily-imperfect choices

Real-world goal: Want to know if microcredit helps people

1. Have to turn this goal into measurements on actual data
    - What counts as “help”? E.g. increase business profit
    - Who gets measured? Particular people, location, etc
    - How is data collected? Survey: In-person, phone, mail
    - How is it distributed? Randomized controlled trial

Example analysis: microcredit

- Making good decisions from data typically requires a lot of work and a lot of hard, necessarily-imperfect choices

Real-world goal: Want to know if microcredit helps people

1. Have to turn this goal into measurements on actual data
   - What counts as “help”? E.g. increase business profit
   - Who gets measured? Particular people, location, etc
   - How is data collected? Survey: In-person, phone, mail
   - How is it distributed? Randomized controlled trial
2. Have to figure out how to make a decision from this data

Example analysis: microcredit

- Making good decisions from data typically requires a lot of work and a lot of hard, necessarily-imperfect choices

Real-world goal: Want to know if microcredit helps people

1. Have to turn this goal into measurements on actual data
   - What counts as “help”? E.g. increase business profit
   - Who gets measured? Particular people, location, etc
   - How is data collected? Survey: In-person, phone, mail
   - How is it distributed? Randomized controlled trial
2. Have to figure out how to make a decision from this data
   - What counts as “increasing profit”?

Example analysis: microcredit

- Making good decisions from data typically requires a lot of work and a lot of hard, necessarily-imperfect choices

Real-world goal: Want to know if microcredit helps people

1. Have to turn this goal into measurements on actual data
   - What counts as “help”? E.g. increase business profit
   - Who gets measured? Particular people, location, etc
   - How is data collected? Survey: In-person, phone, mail
   - How is it distributed? Randomized controlled trial
2. Have to figure out how to make a decision from this data
   - What counts as “increasing profit”? E.g. Compare mean profit of groups receiving and not receiving microcredit. Look for a statistically significant positive effect.

Example analysis: microcredit

- Making good decisions from data typically requires a lot of work and a lot of hard, necessarily-imperfect choices

Real-world goal: Want to know if microcredit helps people

1. Have to turn this goal into measurements on actual data
   - What counts as “help”? E.g. increase business profit
   - Who gets measured? Particular people, location, etc
   - How is data collected? Survey: In-person, phone, mail
   - How is it distributed? Randomized controlled trial
2. Have to figure out how to make a decision from this data
   - What counts as “increasing profit”? E.g. Compare mean profit of groups receiving and not receiving microcredit. Look for a statistically significant positive effect. Or: OLS, GPs, deep learning, Bayes, doubly robust learning, etc.

Example analysis: microcredit

- Making good decisions from data typically requires a lot of work and a lot of hard, necessarily-imperfect choices

Real-world goal: Want to know if microcredit helps people

1. Have to turn this goal into measurements on actual data
   - What counts as “help”? E.g. increase business profit
   - Who gets measured? Particular people, location, etc
   - How is data collected? Survey: In-person, phone, mail
   - How is it distributed? Randomized controlled trial
2. Have to figure out how to make a decision from this data
   - What counts as “increasing profit”? E.g. Compare mean profit of groups receiving and not receiving microcredit. Look for a statistically significant positive effect. Or: OLS, GPs, deep learning, Bayes, doubly robust learning, etc.
3. Have to choose an algorithm

Example analysis: microcredit

- Making good decisions from data typically requires a lot of work and a lot of hard, necessarily-imperfect choices

Real-world goal: Want to know if microcredit helps people

1. Have to turn this goal into measurements on actual data
   - What counts as “help”? E.g. increase business profit
   - Who gets measured? Particular people, location, etc
   - How is data collected? Survey: In-person, phone, mail
   - How is it distributed? Randomized controlled trial
2. Have to figure out how to make a decision from this data
   - What counts as “increasing profit”? E.g. Compare mean profit of groups receiving and not receiving microcredit. Look for a statistically significant positive effect. Or: OLS, GPs, deep learning, Bayes, doubly robust learning, etc.
3. Have to choose an algorithm
4. Have to choose code: Packages, but also full pipeline

Example analysis: microcredit

- Making good decisions from data typically requires a lot of work and a lot of hard, necessarily-imperfect choices

Real-world goal: Want to know if microcredit helps people

1. Have to turn this goal into **measurements on actual data**
   - What counts as "help"? E.g. increase business profit
   - Who gets measured? Particular people, location, etc
   - How is data collected? Survey: In-person, phone, mail
   - How is it distributed? Randomized controlled trial
2. Have to figure out how to **make a decision** from this data
   - What counts as "increasing profit"? E.g. Compare mean profit of groups receiving and not receiving microcredit. Look for a statistically significant positive effect. Or: OLS, GPs, deep learning, Bayes, doubly robust learning, etc.
3. Have to choose an **algorithm**
4. Have to choose **code**: Packages, but also full pipeline

Choosing actual measurements

Choosing actual measurements

- It’s hard to measure what I care about. Is a proxy ok?

Choosing actual measurements

- It’s hard to measure what I care about. Is a proxy ok?
  - Hard to avoid some type of proxy, so let’s hope so!

Choosing actual measurements

- It's hard to measure what I care about. Is a proxy ok?
  - Hard to avoid some type of proxy, so let's hope so! But:
  - Some populations face anemia: not enough healthy red blood cells/hemoglobin to carry oxygen to the body

[Charles+ 11; Rappaport+ 17; Wieringa+ 16, www.mayoclinic.org/diseases-conditions/anemia/]

Choosing actual measurements

- It's hard to measure what I care about. Is a proxy ok?
  - Hard to avoid some type of proxy, so let's hope so! But:
  - Some populations face anemia: not enough healthy red blood cells/hemoglobin to carry oxygen to the body

[Charles+ 11; Rappaport+ 17; Wieringa+ 16, www.mayoclinic.org/diseases-conditions/anemia/]

    - 2 (of many) causes: iron deficiency, genetic disorders

Choosing actual measurements

- It's hard to measure what I care about. Is a proxy ok?
  - Hard to avoid some type of proxy, so let's hope so! But:
  - Some populations face anemia: not enough healthy red blood cells/hemoglobin to carry oxygen to the body

[Charles+ 11; Rappaport+ 17; Wieringa+ 16, www.mayoclinic.org/diseases-conditions/anemia/]

    - 2 (of many) causes: iron deficiency, genetic disorders
    - An intervention might increase measures of iron in blood without reducing anemia

Choosing actual measurements

- It's hard to measure what I care about. Is a proxy ok?
  - Hard to avoid some type of proxy, so let's hope so! But:
  - Some populations face anemia: not enough healthy red blood cells/hemoglobin to carry oxygen to the body

[Charles+ 11; Rappaport+ 17; Wieringa+ 16, www.mayoclinic.org/diseases-conditions/anemia/]

    - 2 (of many) causes: iron deficiency, genetic disorders
    - An intervention might increase measures of iron in blood without reducing anemia
- If I run a randomized controlled trial (RCT), doesn't that guarantee any benefit I find (or don't find) will generalize?

Choosing actual measurements

- It's hard to measure what I care about. Is a proxy ok?
  - Hard to avoid some type of proxy, so let's hope so! But:
  - Some populations face anemia: not enough healthy red blood cells/hemoglobin to carry oxygen to the body

[Charles+ 11; Rappaport+ 17; Wieringa+ 16, www.mayoclinic.org/diseases-conditions/anemia/]

    - 2 (of many) causes: iron deficiency, genetic disorders
    - An intervention might increase measures of iron in blood without reducing anemia
- If I run a randomized controlled trial (RCT), doesn't that guarantee any benefit I find (or don't find) will generalize?
  - RCT on the efficacy of umbrellas for keeping people dry

Choosing actual measurements

- It's hard to measure what I care about. Is a proxy ok?
  - Hard to avoid some type of proxy, so let's hope so! But:
  - Some populations face anemia: not enough healthy red blood cells/hemoglobin to carry oxygen to the body

[Charles+ 11; Rappaport+ 17; Wieringa+ 16, www.mayoclinic.org/diseases-conditions/anemia/]

    - 2 (of many) causes: iron deficiency, genetic disorders
    - An intervention might increase measures of iron in blood without reducing anemia
- If I run a randomized controlled trial (RCT), doesn't that guarantee any benefit I find (or don't find) will generalize?
  - RCT on the efficacy of umbrellas for keeping people dry
    - We randomly assign people umbrellas or no umbrellas in a desert. We see no difference in dryness

Choosing actual measurements

- It's hard to measure what I care about. Is a proxy ok?
  - Hard to avoid some type of proxy, so let's hope so! But:
  - Some populations face anemia: not enough healthy red blood cells/hemoglobin to carry oxygen to the body

[Charles+ 11; Rappaport+ 17; Wieringa+ 16, www.mayoclinic.org/diseases-conditions/anemia/]

    - 2 (of many) causes: iron deficiency, genetic disorders
    - An intervention might increase measures of iron in blood without reducing anemia
- If I run a randomized controlled trial (RCT), doesn't that guarantee any benefit I find (or don't find) will generalize?
  - RCT on the efficacy of umbrellas for keeping people dry
    - We randomly assign people umbrellas or no umbrellas in a desert. We see no difference in dryness
    - Conclusion: umbrellas don't work

Choosing actual measurements

- It's hard to measure what I care about. Is a proxy ok?
  - Hard to avoid some type of proxy, so let's hope so! But:
  - Some populations face anemia: not enough healthy red blood cells/hemoglobin to carry oxygen to the body

[Charles+ 11; Rappaport+ 17; Wieringa+ 16, www.mayoclinic.org/diseases-conditions/anemia/]

    - 2 (of many) causes: iron deficiency, genetic disorders
    - An intervention might increase measures of iron in blood without reducing anemia
- If I run a randomized controlled trial (RCT), doesn't that guarantee any benefit I find (or don't find) will generalize?
  - RCT on the efficacy of umbrellas for keeping people dry
    - We randomly assign people umbrellas or no umbrellas in a desert. We see no difference in dryness
    - Conclusion: umbrellas don't work
    - Cf. medical, econ experiments.

Choosing actual measurements

- It's hard to measure what I care about. Is a proxy ok?
  - Hard to avoid some type of proxy, so let's hope so! But:
  - Some populations face anemia: not enough healthy red blood cells/hemoglobin to carry oxygen to the body

[Charles+ 11; Rappaport+ 17; Wieringa+ 16, www.mayoclinic.org/diseases-conditions/anemia/]

    - 2 (of many) causes: iron deficiency, genetic disorders
    - An intervention might increase measures of iron in blood without reducing anemia
- If I run a randomized controlled trial (RCT), doesn't that guarantee any benefit I find (or don't find) will generalize?
  - RCT on the efficacy of umbrellas for keeping people dry
    - We randomly assign people umbrellas or no umbrellas in a desert. We see no difference in dryness
    - Conclusion: umbrellas don't work
    - Cf. medical, econ experiments. Psychology: "WEIRD: Western, Educated, Industrialized, Rich, Democratic"

Choosing measurements: Externalities

Choosing measurements: Externalities

- Job placement assistance (RCT) [Crépon et al 2013]

Choosing measurements: Externalities

- Job placement assistance (RCT) [Crépon et al 2013]
  - “private agencies are contracted to provide intensive placement services to young graduates (with at least a two-year college degree) who have been unemployed for at least six months”
    - In France, so also other standard forms of assistance

Choosing measurements: Externalities

- Job placement assistance (RCT) [Crépon et al 2013]
  - "private agencies are contracted to provide intensive placement services to young graduates (with at least a two-year college degree) who have been unemployed for at least six months"
    - In France, so also other standard forms of assistance
  - Concluded benefits at 8 months, gone by 12 months

Choosing measurements: Externalities

- Job placement assistance (RCT) [Crépon et al 2013]
  - "private agencies are contracted to provide intensive placement services to young graduates (with at least a two-year college degree) who have been unemployed for at least six months"
    - In France, so also other standard forms of assistance
  - Concluded benefits at 8 months, gone by 12 months
  - Benefits at least partly at expense of other workers

Choosing measurements: Externalities

- Job placement assistance (RCT) [Crépon et al 2013]
  - "private agencies are contracted to provide intensive placement services to young graduates (with at least a two-year college degree) who have been unemployed for at least six months"
    - In France, so also other standard forms of assistance
  - Concluded benefits at 8 months, gone by 12 months
  - Benefits at least partly at expense of other workers
  - Length of study; country vs. worker; mechanism

Choosing measurements: Externalities

- Job placement assistance (RCT) [Crépon et al 2013]
  - "private agencies are contracted to provide intensive placement services to young graduates (with at least a two-year college degree) who have been unemployed for at least six months"
    - In France, so also other standard forms of assistance
  - Concluded benefits at 8 months, gone by 12 months
    - Benefits at least partly at expense of other workers
    - Length of study; country vs. worker; mechanism
- A new farming technique could increase crop yield. But if demand is constant, price could drop.

Choosing measurements: Externalities

- Job placement assistance (RCT) [Crépon et al 2013]
  - "private agencies are contracted to provide intensive placement services to young graduates (with at least a two-year college degree) who have been unemployed for at least six months"
    - In France, so also other standard forms of assistance
  - Concluded benefits at 8 months, gone by 12 months
    - Benefits at least partly at expense of other workers
    - Length of study; country vs. worker; mechanism
- A new farming technique could increase crop yield. But if demand is constant, price could drop.
- A medical intervention that reduces disease spread: extrapolation from individual could underestimate benefit (cf. apparent extinction of influenza strain B/Yamagata)

Choosing measurements: Externalities

- Job placement assistance (RCT) [Crépon et al 2013]
  - "private agencies are contracted to provide intensive placement services to young graduates (with at least a two-year college degree) who have been unemployed for at least six months"
    - In France, so also other standard forms of assistance
  - Concluded benefits at 8 months, gone by 12 months
    - Benefits at least partly at expense of other workers
    - Length of study; country vs. worker; mechanism
- A new farming technique could increase crop yield. But if demand is constant, price could drop.
- A medical intervention that reduces disease spread: extrapolation from individual could underestimate benefit (cf. apparent extinction of influenza strain B/Yamagata)
- Can't avoid thinking about where data comes from, what is measured, and what is possible to know from that data

Code: bugs matter

Code: bugs matter

- An econ paper widely used to justify austerity policies omitted data for 5 of 20 countries due to an error

[Reinhart, Rogoff 2010; Herndon et al 2014]

Code: bugs matter

- An econ paper widely used to justify austerity policies omitted data for 5 of 20 countries due to an error
  [Reinhart, Rogoff 2010; Herndon et al 2014]
- Bugs (e.g. off-by-one errors) in data analyses have led to unjustified treatment of patients in oncology clinical trials
  [Baggerly, Coombes 2009]

Code: bugs matter

- An econ paper widely used to justify austerity policies omitted data for 5 of 20 countries due to an error [Reinhart, Rogoff 2010; Herndon et al 2014]
- Bugs (e.g. off-by-one errors) in data analyses have led to unjustified treatment of patients in oncology clinical trials [Baggerly, Coombes 2009]
- British Post Office scandal [Flinders 2022; Computer Weekly 2009–2022]

Code: bugs matter

- An econ paper widely used to justify austerity policies omitted data for 5 of 20 countries due to an error [Reinhart, Rogoff 2010; Herndon et al 2014]
- Bugs (e.g. off-by-one errors) in data analyses have led to unjustified treatment of patients in oncology clinical trials [Baggerly, Coombes 2009]
- British Post Office scandal [Flinders 2022; Computer Weekly 2009–2022]
  - Horizon: accounting software from Fujitsu

Code: bugs matter

- An econ paper widely used to justify austerity policies omitted data for 5 of 20 countries due to an error [Reinhart, Rogoff 2010; Herndon et al 2014]
- Bugs (e.g. off-by-one errors) in data analyses have led to unjustified treatment of patients in oncology clinical trials [Baggerly, Coombes 2009]
- British Post Office scandal [Flinders 2022; Computer Weekly 2009–2022]
  - Horizon: accounting software from Fujitsu
  - Bug sometimes showed two items as sold though only one was sold → left cash in till short

Code: bugs matter

- An econ paper widely used to justify austerity policies omitted data for 5 of 20 countries due to an error [Reinhart, Rogoff 2010; Herndon et al 2014]
- Bugs (e.g. off-by-one errors) in data analyses have led to unjustified treatment of patients in oncology clinical trials [Baggerly, Coombes 2009]
- British Post Office scandal [Flinders 2022; Computer Weekly 2009–2022]
  - Horizon: accounting software from Fujitsu
  - Bug sometimes showed two items as sold though only one was sold → left cash in till short
  - "Between 2000 and 2015, 736 subpostmasters were prosecuted by the Post Office, with many convicted and sent to prison" [Flinders 2022] "Linked to at least four suicides" [Sweney 2024]

Code: bugs matter

- An econ paper widely used to justify austerity policies omitted data for 5 of 20 countries due to an error [Reinhart, Rogoff 2010; Herndon et al 2014]
- Bugs (e.g. off-by-one errors) in data analyses have led to unjustified treatment of patients in oncology clinical trials [Baggerly, Coombes 2009]
- British Post Office scandal [Flinders 2022; Computer Weekly 2009–2022]
  - Horizon: accounting software from Fujitsu
  - Bug sometimes showed two items as sold though only one was sold → left cash in till short
  - "Between 2000 and 2015, 736 subpostmasters were prosecuted by the Post Office, with many convicted and sent to prison" [Flinders 2022] "Linked to at least four suicides" [Sweney 2024]
  - 2006: Computer science graduate Mark Kelly wrote his own code to check, identified the bug, reported the bug

Code: bugs matter

- An econ paper widely used to justify austerity policies omitted data for 5 of 20 countries due to an error [Reinhart, Rogoff 2010; Herndon et al 2014]
- Bugs (e.g. off-by-one errors) in data analyses have led to unjustified treatment of patients in oncology clinical trials [Baggerly, Coombes 2009]
- British Post Office scandal [Flinders 2022; Computer Weekly 2009–2022]
  - Horizon: accounting software from Fujitsu
  - Bug sometimes showed two items as sold though only one was sold → left cash in till short
  - "Between 2000 and 2015, 736 subpostmasters were prosecuted by the Post Office, with many convicted and sent to prison" [Flinders 2022] "Linked to at least four suicides" [Sweney 2024]
  - 2006: Computer science graduate Mark Kelly wrote his own code to check, identified the bug, reported the bug
  - For a decade, Post Office said no errors
  - Finally changed story after High Court litigation in 2019

Code: challenges & mitigations

- OK but am I safe from bugs if I use standard packages and data analyses?

Code: challenges & mitigations

- OK but am I safe from bugs if I use standard packages and data analyses?
  - Natural to outsource code (team science)

Code: challenges & mitigations

- OK but am I safe from bugs if I use standard packages and data analyses?
  - Natural to outsource code (team science)
  - The issues highlighted on the previous slide are arguably in pre-processing

Code: challenges & mitigations

- OK but am I safe from bugs if I use standard packages and data analyses?
  - Natural to outsource code (team science)
  - The issues highlighted on the previous slide are arguably in pre-processing
  - 2016: 20% of studied papers had Excel error that "autocorrects" certain gene names into dates
    - E.g. "MARCH1, SEPT1, Oct-4, jun" [Ziemann, Abeysooriya 2021]

[Zeeberg et al 2004; Ziemann et al 2016]

Code: challenges & mitigations

- OK but am I safe from bugs if I use standard packages and data analyses?
  - Natural to outsource code (team science)
  - The issues highlighted on the previous slide are arguably in pre-processing
  - 2016: 20% of studied papers had Excel error that "autocorrects" certain gene names into dates [Zeeberg et al 2004; Ziemann et al 2016]
  - E.g. "MARCH1, SEPT1, Oct-4, jun" [Ziemann, Abeysooriya 2021]
  - 2021: 30% of studied papers had this issue [Abeysooriya et al 2021; Lewis 2021]

Code: challenges & mitigations

- OK but am I safe from bugs if I use standard packages and data analyses?
  - Natural to outsource code (team science)
  - The issues highlighted on the previous slide are arguably in pre-processing
  - 2016: 20% of studied papers had Excel error that "autocorrects" certain gene names into dates [Zeeberg et al 2004; Ziemann et al 2016]
    - E.g. "MARCH1, SEPT1, Oct-4, jun" [Ziemann, Abeysooriya 2021]
    - 2021: 30% of studied papers had this issue [Abeysooriya et al 2021; Lewis 2021]
- What can we do?

Code: challenges & mitigations

- OK but am I safe from bugs if I use standard packages and data analyses?
  - Natural to outsource code (team science)
  - The issues highlighted on the previous slide are arguably in pre-processing
  - 2016: 20% of studied papers had Excel error that "autocorrects" certain gene names into dates [Zeeberg et al 2004; Ziemann et al 2016]
    - E.g. "MARCH1, SEPT1, Oct-4, jun" [Ziemann, Abeysooriya 2021]
    - 2021: 30% of studied papers had this issue [Abeysooriya et al 2021; Lewis 2021]
- What can we do?
  - Shaming people is often ineffective (or backfires). People worry their code isn't perfect (it's not!) and don't share it.

Code: challenges & mitigations

- OK but am I safe from bugs if I use standard packages and data analyses?
  - Natural to outsource code (team science)
  - The issues highlighted on the previous slide are arguably in pre-processing
  - 2016: 20% of studied papers had Excel error that "autocorrects" certain gene names into dates [Zeeberg et al 2004; Ziemann et al 2016]
    - E.g. "MARCH1, SEPT1, Oct-4, jun" [Ziemann, Abeysooriya 2021]
    - 2021: 30% of studied papers had this issue [Abeysooriya et al 2021; Lewis 2021]
- What can we do?
  - Shaming people is often ineffective (or backfires). People worry their code isn't perfect (it's not!) and don't share it.
  - Bugs are inevitable; how we respond is not.

Code: challenges & mitigations

- OK but am I safe from bugs if I use standard packages and data analyses?
  - Natural to outsource code (team science)
  - The issues highlighted on the previous slide are arguably in pre-processing
  - 2016: 20% of studied papers had Excel error that "autocorrects" certain gene names into dates [Zeeberg et al 2004; Ziemann et al 2016]
    - E.g. "MARCH1, SEPT1, Oct-4, jun" [Ziemann, Abeysooriya 2021]
    - 2021: 30% of studied papers had this issue [Abeysooriya et al 2021; Lewis 2021]
- What can we do?
  - Shaming people is often ineffective (or backfires). People worry their code isn't perfect (it's not!) and don't share it.
  - Bugs are inevitable; how we respond is not.
  - First step: Share code. Rules to share code.

Code: challenges & mitigations

- OK but am I safe from bugs if I use standard packages and data analyses?
  - Natural to outsource code (team science)
  - The issues highlighted on the previous slide are arguably in pre-processing
  - 2016: 20% of studied papers had Excel error that "autocorrects" certain gene names into dates [Zeeberg et al 2004; Ziemann et al 2016]
    - E.g. "MARCH1, SEPT1, Oct-4, jun" [Ziemann, Abeysooriya 2021]
    - 2021: 30% of studied papers had this issue [Abeysooriya et al 2021; Lewis 2021]
- What can we do?
  - Shaming people is often ineffective (or backfires). People worry their code isn't perfect (it's not!) and don't share it.
  - Bugs are inevitable; how we respond is not.
  - First step: Share code. Rules to share code.
  - Tools to support tracking models/packages/pipelines.

[Stodden et al 2014, 2016; Vartak et al 2016; Gebru et al 2021; Heil et al 2021; wandb.ai]

Code: challenges & mitigations

Code: challenges & mitigations

- Recognize biased sampling

Code: challenges & mitigations

- Recognize biased sampling
  - If a scientific area is singled out for a problem with bugs (or replicability, etc), it's because someone checked

Code: challenges & mitigations

- Recognize biased sampling
  - If a scientific area is singled out for a problem with bugs (or replicability, etc), it's because someone checked
  - Areas easier to check might find more issues even if they have fewer overall

Code: challenges & mitigations

- Recognize biased sampling
  - If a scientific area is singled out for a problem with bugs (or replicability, etc), it's because someone checked
  - Areas easier to check might find more issues even if they have fewer overall
  - It's easy to disincentivize checking for problems

Code: challenges & mitigations

- Recognize biased sampling
  - If a scientific area is singled out for a problem with bugs (or replicability, etc), it's because someone checked
  - Areas easier to check might find more issues even if they have fewer overall
  - It's easy to disincentivize checking for problems
- But machine learning/AI code can get very complex. Does it really need to be shared?

Code: challenges & mitigations

- Recognize biased sampling
  - If a scientific area is singled out for a problem with bugs (or replicability, etc), it's because someone checked
  - Areas easier to check might find more issues even if they have fewer overall
  - It's easy to disincentivize checking for problems
- But machine learning/AI code can get very complex. Does it really need to be shared?
- *Nature* states they require authors to share code
  - Gives pass to AI for detecting breast cancer: no code or full detail of the method

[McKinney et al 2020; Haibe-Kains 2021]

Code: challenges & mitigations

- Recognize biased sampling
  - If a scientific area is singled out for a problem with bugs (or replicability, etc), it's because someone checked
  - Areas easier to check might find more issues even if they have fewer overall
  - It's easy to disincentivize checking for problems
- But machine learning/AI code can get very complex. Does it really need to be shared?
- *Nature* states they require authors to share code
  - Gives pass to AI for detecting breast cancer: no code or full detail of the method

[McKinney et al 2020; Haibe-Kains 2021]

- 13 of 62 AI methods for medical diagnosis from images shared code

[Roberts et al 2021]

Code: challenges & mitigations

- Recognize biased sampling
  - If a scientific area is singled out for a problem with bugs (or replicability, etc), it's because someone checked
  - Areas easier to check might find more issues even if they have fewer overall
  - It's easy to disincentivize checking for problems
- But machine learning/AI code can get very complex. Does it really need to be shared?
  - *Nature* states they require authors to share code
    - Gives pass to AI for detecting breast cancer: no code or full detail of the method [McKinney et al 2020; Haibe-Kains 2021]
  - 13 of 62 AI methods for medical diagnosis from images shared code [Roberts et al 2021]
  - When there's no code, other groups can't check the results/algorithm/method (and can't build on the work)

Algorithms

Algorithms

- Didn't past machine learning experts prove this algorithm "works"? If there's an issue, did they lie to me?

Algorithms

- Didn't past machine learning experts prove this algorithm "works"? If there's an issue, did they lie to me?
  - Important to be aware that all algorithms/methods are justified under precise assumptions

Algorithms

- Didn't past machine learning experts prove this algorithm "works"? If there's an issue, did they lie to me?
  - Important to be aware that all algorithms/methods are justified under precise assumptions
    - E.g. optimization guarantees proved under convexity assumptions

Algorithms

- Didn't past machine learning experts prove this algorithm "works"? If there's an issue, did they lie to me?
  - Important to be aware that all algorithms/methods are justified under precise assumptions
    - E.g. optimization guarantees proved under convexity assumptions
    - E.g. covariate shift guarantees proved under bounded density ratios (sufficient data and well-distributed)

Algorithms

- Didn't past machine learning experts prove this algorithm "works"? If there's an issue, did they lie to me?
  - Important to be aware that all algorithms/methods are justified under precise assumptions
    - E.g. optimization guarantees proved under convexity assumptions
    - E.g. covariate shift guarantees proved under bounded density ratios (sufficient data and well-distributed)
  - Terms of art usually ≠ colloquial meaning.

Algorithms

- Didn't past machine learning experts prove this algorithm "works"? If there's an issue, did they lie to me?
  - Important to be aware that all algorithms/methods are justified under precise assumptions
    - E.g. optimization guarantees proved under convexity assumptions
    - E.g. covariate shift guarantees proved under bounded density ratios (sufficient data and well-distributed)
- Terms of art usually ≠ colloquial meaning. E.g.
  - Uncertainty quantification (vs. Bayesian uncertainty under chosen model, frequentist sampling uncert, etc)

Algorithms

- Didn't past machine learning experts prove this algorithm "works"? If there's an issue, did they lie to me?
  - Important to be aware that all algorithms/methods are justified under precise assumptions
    - E.g. optimization guarantees proved under convexity assumptions
    - E.g. covariate shift guarantees proved under bounded density ratios (sufficient data and well-distributed)
- Terms of art usually ≠ colloquial meaning. E.g.
  - Uncertainty quantification (vs. Bayesian uncertainty under chosen model, frequentist sampling uncert, etc)
  - Robustness quantification (Measure/approximate robustness to a particular change. Importance varies by context)

Algorithms

- Didn't past machine learning experts prove this algorithm "works"? If there's an issue, did they lie to me?
  - Important to be aware that all algorithms/methods are justified under precise assumptions
    - E.g. optimization guarantees proved under convexity assumptions
    - E.g. covariate shift guarantees proved under bounded density ratios (sufficient data and well-distributed)
  - Terms of art usually ≠ colloquial meaning. E.g.
    - Uncertainty quantification (vs. Bayesian uncertainty under chosen model, frequentist sampling uncert, etc)
    - Robustness quantification (Measure/approximate robustness to a particular change. Importance varies by context)
  - Still need to run sense checks on what you're trying to do

Turning high-level goals into math

Turning high-level goals into math

- If we're good data analysts with bug-free code and the same data and using reasonable/vetted algorithms, will we always reach the same conclusions?

Turning high-level goals into math

- If we're good data analysts with bug-free code and the same data and using reasonable/vetted algorithms, will we always reach the same conclusions?
  - 29 teams used same data to answer "are soccer referees more likely to give red cards to dark-skin-toned players?" [Silberzahn et al 2018]

Turning high-level goals into math

- If we're good data analysts with bug-free code and the same data and using reasonable/vetted algorithms, will we always reach the same conclusions?
  - 29 teams used same data to answer "are soccer referees more likely to give red cards to dark-skin-toned players?" [Silberzahn et al 2018]
    - 20 teams concluded yes, 9 teams did not
    - Researchers did <u>not</u> find that the variability was due to beliefs, expertise, quality

Turning high-level goals into math

- If we're good data analysts with bug-free code and the same data and using reasonable/vetted algorithms, will we always reach the same conclusions?
  - 29 teams used same data to answer "are soccer referees more likely to give red cards to dark-skin-toned players?" [Silberzahn et al 2018]
    - 20 teams concluded yes, 9 teams did not
    - Researchers did <u>not</u> find that the variability was due to beliefs, expertise, quality
- Engineering, not just math

Turning high-level goals into math

- If we're good data analysts with bug-free code and the same data and using reasonable/vetted algorithms, will we always reach the same conclusions?
  - 29 teams used same data to answer "are soccer referees more likely to give red cards to dark-skin-toned players?" [Silberzahn et al 2018]
    - 20 teams concluded yes, 9 teams did not
    - Researchers did <u>not</u> find that the variability was due to beliefs, expertise, quality
- Engineering, not just math
- In real life, we start with high-level goals (E.g. "does microcredit help people?")

Turning high-level goals into math

- If we're good data analysts with bug-free code and the same data and using reasonable/vetted algorithms, will we always reach the same conclusions?
  - 29 teams used same data to answer "are soccer referees more likely to give red cards to dark-skin-toned players?" [Silberzahn et al 2018]
    - 20 teams concluded yes, 9 teams did not
      - Researchers did <u>not</u> find that the variability was due to beliefs, expertise, quality
- Engineering, not just math
- In real life, we start with high-level goals (E.g. "does microcredit help people?")
  - But we have to make a lot of subjective choices before we can formalize a mathematical problem as a proxy for the high-level goal (e.g. what to measure, what counts as helping, should we control for certain features and how)

Turning high-level goals into math

Turning high-level goals into math

- Subjectivity doesn’t mean there are no wrong answers

Turning high-level goals into math

- Subjectivity doesn’t mean there are no wrong answers
- High-level goal: Diagnose disease

Turning high-level goals into math

- Subjectivity doesn't mean there are no wrong answers
- High-level goal: Diagnose disease
- Formalization: Maximize correct predictions of disease status with AI in a held-out subset of data

Turning high-level goals into math

- Subjectivity doesn't mean there are no wrong answers
- High-level goal: Diagnose disease
- Formalization: Maximize correct predictions of disease status with AI in a held-out subset of data
  - AI for detecting skin cancer can learn to recognize experts' ink marks (study looked at a "CNN approved for use as a medical device in the European market")

[Winkler et al 2019]

Turning high-level goals into math

- Subjectivity doesn't mean there are no wrong answers
- High-level goal: Diagnose disease
- Formalization: Maximize correct predictions of disease status with AI in a held-out subset of data
  - AI for detecting skin cancer can learn to recognize experts' ink marks (study looked at a "CNN approved for use as a medical device in the European market") [Winkler et al 2019]
  - AI using chest scans of children as no-covid controls can learn to identify children rather than no-covid [Roberts et al 2021; Heaven 2021]

Turning high-level goals into math

- Subjectivity doesn't mean there are no wrong answers
- High-level goal: Diagnose disease
- Formalization: Maximize correct predictions of disease status with AI in a held-out subset of data
  - AI for detecting skin cancer can learn to recognize experts' ink marks (study looked at a "CNN approved for use as a medical device in the European market") [Winkler et al 2019]
  - AI using chest scans of children as no-covid controls can learn to identify children rather than no-covid [Roberts et al 2021; Heaven 2021]
    - MIT Technology Review: "Hundreds of AI tools have been built to catch covid. None of them helped. Some have been used in hospitals, despite not being properly tested."

Turning high-level goals into math

- Subjectivity doesn't mean there are no wrong answers
- High-level goal: Diagnose disease
- Formalization: Maximize correct predictions of disease status with AI in a held-out subset of data
  - AI for detecting skin cancer can learn to recognize experts' ink marks (study looked at a "CNN approved for use as a medical device in the European market") [Winkler et al 2019]
  - AI using chest scans of children as no-covid controls can learn to identify children rather than no-covid [Roberts et al 2021; Heaven 2021]
    - MIT Technology Review: "Hundreds of AI tools have been built to catch covid. None of them helped. Some have been used in hospitals, despite not being properly tested."
  - If full dataset has duplicates (e.g. amalgam data), AI can learn to identify a particular patient or scan [Roberts et al 2021; Heaven 2021]

Turning high-level goals into math

- Subjectivity doesn't mean there are no wrong answers
- High-level goal: Diagnose disease
- Formalization: Maximize correct predictions of disease status with AI in a held-out subset of data
  - AI for detecting skin cancer can learn to recognize experts' ink marks (study looked at a "CNN approved for use as a medical device in the European market") [Winkler et al 2019]
  - AI using chest scans of children as no-covid controls can learn to identify children rather than no-covid [Roberts et al 2021; Heaven 2021]
    - MIT Technology Review: "Hundreds of AI tools have been built to catch covid. None of them helped. Some have been used in hospitals, despite not being properly tested."
  - If full dataset has duplicates (e.g. amalgam data), AI can learn to identify a particular patient or scan [Roberts et al 2021; Heaven 2021]
- Optimizing this particular objective is a convenient proxy, but it isn't the same as diagnosing disease well

Turning high-level goals into math

- Subjectivity doesn't mean there are no wrong answers
- High-level goal: Diagnose disease
- Formalization: Maximize correct predictions of disease status with AI in a held-out subset of data
  - AI for detecting skin cancer can learn to recognize experts' ink marks (study looked at a "CNN approved for use as a medical device in the European market") [Winkler et al 2019]
  - AI using chest scans of children as no-covid controls can learn to identify children rather than no-covid [Roberts et al 2021; Heaven 2021]
    - MIT Technology Review: "Hundreds of AI tools have been built to catch covid. None of them helped. Some have been used in hospitals, despite not being properly tested."
  - If full dataset has duplicates (e.g. amalgam data), AI can learn to identify a particular patient or scan [Roberts et al 2021; Heaven 2021]
- Optimizing this particular objective is a convenient proxy, but it isn't the same as diagnosing disease well
- Why cross-validation isn't a cure-all

Turning high-level goals into math

Turning high-level goals into math

- Is the complex AI the problem? When using simpler/vetted methods, we are often still using convenient proxies

Turning high-level goals into math

- Is the complex AI the problem? When using simpler/vetted methods, we are often still using convenient proxies
- High-level goal: Figure out if microcredit is helping people

Turning high-level goals into math

- Is the complex AI the problem? When using simpler/vetted methods, we are often still using convenient proxies
- High-level goal: Figure out if microcredit is helping people
- Probably what we meant: Figure out if microcredit is improving the lives of many in a meaningful way

Turning high-level goals into math

- Is the complex AI the problem? When using simpler/vetted methods, we are often still using convenient proxies
- High-level goal: Figure out if microcredit is helping people
- Probably what we meant: Figure out if microcredit is improving the lives of many in a meaningful way
- Formalization in practice: Decide microcredit is helpful if mean business profit is higher in group with microcredit

Turning high-level goals into math

- Is the complex AI the problem? When using simpler/vetted methods, we are often still using convenient proxies
- High-level goal: Figure out if microcredit is helping people
- Probably what we meant: Figure out if microcredit is improving the lives of many in a meaningful way
- Formalization in practice: Decide microcredit is helpful if mean business profit is higher in group with microcredit
  - Imagine a world where no one benefits from microcredit except for a tiny handful of people → Conclude: it helps

Turning high-level goals into math

- Is the complex AI the problem? When using simpler/vetted methods, we are often still using convenient proxies
- High-level goal: Figure out if microcredit is helping people
- Probably what we meant: Figure out if microcredit is improving the lives of many in a meaningful way
- Formalization in practice: Decide microcredit is helpful if mean business profit is higher in group with microcredit
  - Imagine a world where no one benefits from microcredit except for a tiny handful of people → Conclude: it helps
- In US, household net worth mean ~$1M, median ~$193K [Federal Reserve Board's Division of Research and Statistics, 2023]

Turning high-level goals into math

- Is the complex AI the problem? When using simpler/vetted methods, we are often still using convenient proxies
- High-level goal: Figure out if microcredit is helping people
- Probably what we meant: Figure out if microcredit is improving the lives of many in a meaningful way
- Formalization in practice: Decide microcredit is helpful if mean business profit is higher in group with microcredit
  - Imagine a world where no one benefits from microcredit except for a tiny handful of people → Conclude: it helps
  - In US, household net worth mean ~$1M, median ~$193K [Federal Reserve Board's Division of Research and Statistics, 2023]
  - Can be seen as an issue with squared loss

Turning high-level goals into math

- Is the complex AI the problem? When using simpler/vetted methods, we are often still using convenient proxies
- High-level goal: Figure out if microcredit is helping people
- Probably what we meant: Figure out if microcredit is improving the lives of many in a meaningful way
- Formalization in practice: Decide microcredit is helpful if mean business profit is higher in group with microcredit
  - Imagine a world where no one benefits from microcredit except for a tiny handful of people → Conclude: it helps
  - In US, household net worth mean ~$1M, median ~$193K [Federal Reserve Board's Division of Research and Statistics, 2023]
  - Can be seen as an issue with squared loss
    - Squared loss is sensitive to outliers

Turning high-level goals into math

- Is the complex AI the problem? When using simpler/vetted methods, we are often still using convenient proxies
- High-level goal: Figure out if microcredit is helping people
- Probably what we meant: Figure out if microcredit is improving the lives of many in a meaningful way
- Formalization in practice: Decide microcredit is helpful if mean business profit is higher in group with microcredit
  - Imagine a world where no one benefits from microcredit except for a tiny handful of people → Conclude: it helps
  - In US, household net worth mean ~$1M, median ~$193K [Federal Reserve Board's Division of Research and Statistics, 2023]
  - Can be seen as an issue with squared loss
    - Squared loss is sensitive to outliers
    - Squared loss is extremely widely used, not just in OLS [Castro Torres & Akbaritabar 2024]

Turning high-level goals into math

- Is the complex AI the problem? When using simpler/vetted methods, we are often still using convenient proxies
- High-level goal: Figure out if microcredit is helping people
- Probably what we meant: Figure out if microcredit is improving the lives of many in a meaningful way
- Formalization in practice: Decide microcredit is helpful if mean business profit is higher in group with microcredit
  - Imagine a world where no one benefits from microcredit except for a tiny handful of people → Conclude: it helps
  - In US, household net worth mean ~$1M, median ~$193K [Federal Reserve Board's Division of Research and Statistics, 2023]
  - Can be seen as an issue with squared loss
    - Squared loss is sensitive to outliers
    - Squared loss is extremely widely used, not just in OLS [Castro Torres & Akbaritabar 2024]
    - OLS has many advantages: (typically) closed-form solution. Well-vetted code, theory. Easy to understand

Turning high-level goals into math

Turning high-level goals into math

- Removing outliers isn't a panacea

Turning high-level goals into math

- Removing outliers isn't a panacea
  - E.g. ozone depletion first flagged as outliers to NASA

Turning high-level goals into math

- Removing outliers isn't a panacea
  - E.g. ozone depletion first flagged as outliers to NASA
  - When scientists checked over the outliers, realized there was a problem [Earth Observatory, NASA, 2001; Pukelsheim, 1990]

Turning high-level goals into math

- Removing outliers isn't a panacea
  - E.g. ozone depletion first flagged as outliers to NASA
  - When scientists checked over the outliers, realized there was a problem [Earth Observatory, NASA, 2001; Pukelsheim, 1990]
  - Apparently an apocryphal version of this story says NASA used "robust" data analysis methods and ignored the outliers [Pukelsheim, 1990]

Turning high-level goals into math

- Removing outliers isn't a panacea
  - E.g. ozone depletion first flagged as outliers to NASA
  - When scientists checked over the outliers, realized there was a problem [Earth Observatory, NASA, 2001; Pukelsheim, 1990]
  - Apparently an apocryphal version of this story says NASA used "robust" data analysis methods and ignored the outliers [Pukelsheim, 1990]
- But don't p-values tell me if my result is right/generalizable?

Turning high-level goals into math

- Removing outliers isn't a panacea
  - E.g. ozone depletion first flagged as outliers to NASA
  - When scientists checked over the outliers, realized there was a problem [Earth Observatory, NASA, 2001; Pukelsheim, 1990]
  - Apparently an apocryphal version of this story says NASA used "robust" data analysis methods and ignored the outliers [Pukelsheim, 1990]
- But don't p-values tell me if my result is right/generalizable?
  - Roughly, a p-value tells us how likely our data (summary) is under one particular model you specify in advance

Turning high-level goals into math

- Removing outliers isn't a panacea
  - E.g. ozone depletion first flagged as outliers to NASA
  - When scientists checked over the outliers, realized there was a problem [Earth Observatory, NASA, 2001; Pukelsheim, 1990]
  - Apparently an apocryphal version of this story says NASA used "robust" data analysis methods and ignored the outliers [Pukelsheim, 1990]
- But don't p-values tell me if my result is right/generalizable?
  - Roughly, a p-value tells us how likely our data (summary) is under one particular model you specify in advance
    - If the data is sufficiently unlikely, "reject" that model

Turning high-level goals into math

- Removing outliers isn't a panacea
  - E.g. ozone depletion first flagged as outliers to NASA
  - When scientists checked over the outliers, realized there was a problem [Earth Observatory, NASA, 2001; Pukelsheim, 1990]
  - Apparently an apocryphal version of this story says NASA used "robust" data analysis methods and ignored the outliers [Pukelsheim, 1990]
- But don't p-values tell me if my result is right/generalizable?
  - Roughly, a p-value tells us how likely our data (summary) is under one particular model you specify in advance
    - If the data is sufficiently unlikely, "reject" that model
  - "All models are wrong," so expect p-value to get small with enough data (but that's not necessarily meaningful) [Box, 1976; Wang, Long 2022]

So what can we do?

So what can we do?

- We've learned about some amazing, useful ML tools

So what can we do?

- We've learned about some amazing, useful ML tools
- But we still have to decide which tools to use and how to use them responsibly

So what can we do?

- We've learned about some amazing, useful ML tools
- But we still have to decide which tools to use and how to use them responsibly
- Run unit tests. Run sense checks. Visualize.

So what can we do?

- We've learned about some amazing, useful ML tools
- But we still have to decide which tools to use and how to use them responsibly
- Run unit tests. Run sense checks. Visualize.
- Meaningful results should be stable/robust

[Yu, 2013, 2020; Yu, Kumbier 2020]

So what can we do?

- We've learned about some amazing, useful ML tools
- But we still have to decide which tools to use and how to use them responsibly
- Run unit tests. Run sense checks. Visualize.
- Meaningful results should be stable/robust

[Yu, 2013, 2020; Yu, Kumbier 2020]

- Explanation can be a form of stability

[Arrieta et al 2020; Doshi-Velez et al 2017; Zhang et al 2020; Mittelstadt et al 2019]

So what can we do?

- We've learned about some amazing, useful ML tools
- But we still have to decide which tools to use and how to use them responsibly
- Run unit tests. Run sense checks. Visualize.
- Meaningful results should be stable/robust

[Yu, 2013, 2020; Yu, Kumbier 2020]

- Explanation can be a form of stability

[Arrieta et al 2020; Doshi-Velez et al 2017; Zhang et al 2020; Mittelstadt et al 2019]

- Demand (of yourself) to really understand <u>why</u> the results are the way they are. Don't settle for an absence of observed issues (and the resulting perverse incentives)

So what can we do?

- We've learned about some amazing, useful ML tools
- But we still have to decide which tools to use and how to use them responsibly
- Run unit tests. Run sense checks. Visualize.
- Meaningful results should be stable/robust

[Yu, 2013, 2020; Yu, Kumbier 2020]

- Explanation can be a form of stability

[Arrieta et al 2020; Doshi-Velez et al 2017; Zhang et al 2020; Mittelstadt et al 2019]

  - Demand (of yourself) to really understand <u>why</u> the results are the way they are. Don't settle for an absence of observed issues (and the resulting perverse incentives)
- Sometimes answers aren't clear or what you want

So what can we do?

- We've learned about some amazing, useful ML tools
- But we still have to decide which tools to use and how to use them responsibly
- Run unit tests. Run sense checks. Visualize.
- Meaningful results should be stable/robust

[Yu, 2013, 2020; Yu, Kumbier 2020]

- Explanation can be a form of stability

[Arrieta et al 2020; Doshi-Velez et al 2017; Zhang et al 2020; Mittelstadt et al 2019]

  - Demand (of yourself) to really understand <u>why</u> the results are the way they are. Don't settle for an absence of observed issues (and the resulting perverse incentives)
- Sometimes answers aren't clear or what you want
  - Recognize that classes (including this one) are a biased sample: questions are usually chosen for an achievable solution, and examples are chosen to be illustrative

So what can we do?

- We've learned about some amazing, useful ML tools
- But we still have to decide which tools to use and how to use them responsibly
- Run unit tests. Run sense checks. Visualize.
- Meaningful results should be stable/robust

[Yu, 2013, 2020; Yu, Kumbier 2020]

- Explanation can be a form of stability

[Arrieta et al 2020; Doshi-Velez et al 2017; Zhang et al 2020; Mittelstadt et al 2019]

  - Demand (of yourself) to really understand <u>why</u> the results are the way they are. Don't settle for an absence of observed issues (and the resulting perverse incentives)
- Sometimes answers aren't clear or what you want
  - Recognize that classes (including this one) are a biased sample: questions are usually chosen for an achievable solution, and examples are chosen to be illustrative
- Gaps are opportunities for engineering solutions

6.7900: Machine Learning

Staff

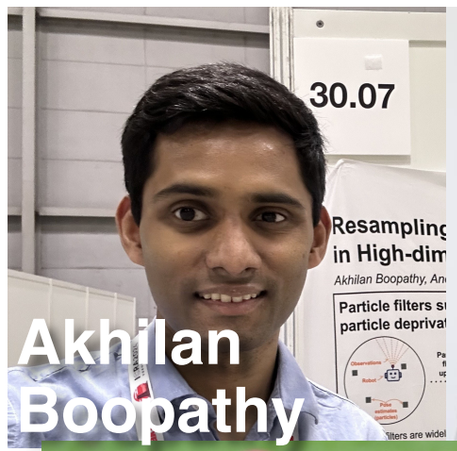
Instructors:

Leslie Kaelbling

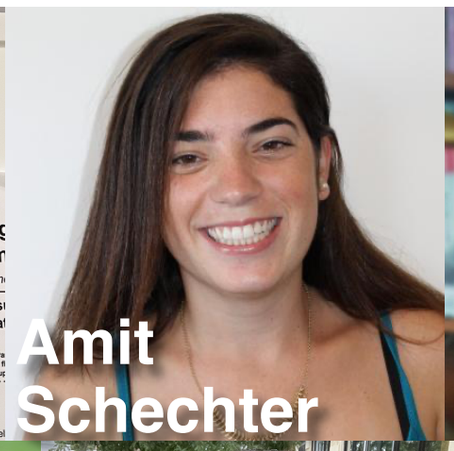
Tamara Broderick

Tommi Jaakkola

Teaching Assistants:

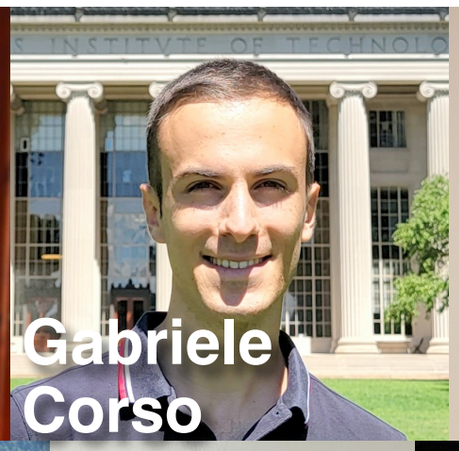


Akhilan Boopathy



Amit Schechter

Feng Zhu



Gabriele Corso

Greg Pylypovych

Jagdeep Bhatia

Morris Yao

Ray Wang

Timmy Qian

Tom Wang

6.7900: Machine Learning

Thank you!

Staff

Instructors:

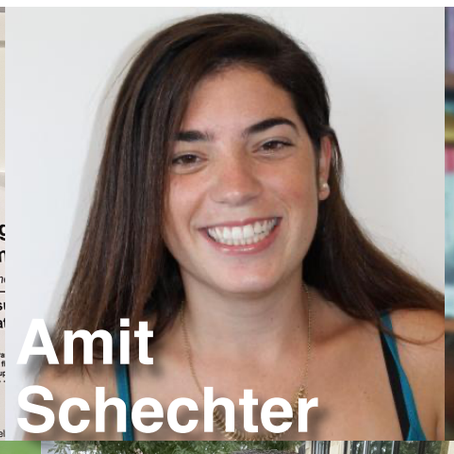
Leslie Kaelbling

Tamara Broderick

Tommi Jaakkola

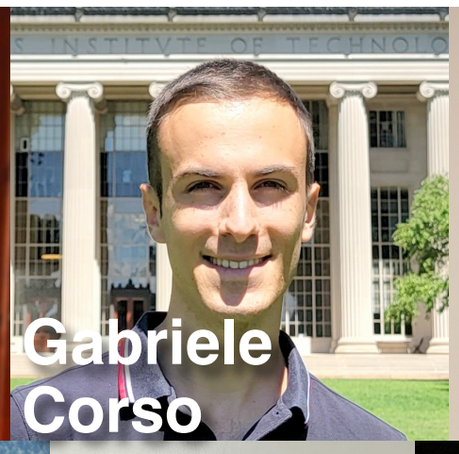
Teaching Assistants:

Akhilan Boopathy



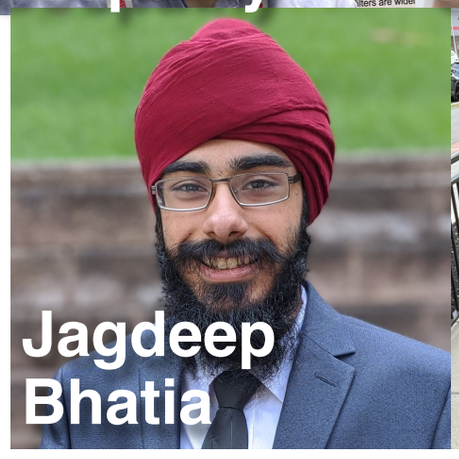
Amit Schechter

Feng Zhu

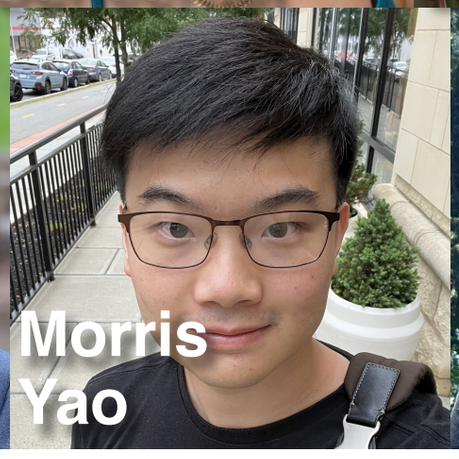


Gabriele Corso

Greg Pylypovych



Jagdeep Bhatia



Morris Yao

Ray Wang

Timmy Qian

Tom Wang

Additional References (1/5)

Abeysooriya et al. Gene name errors: Lessons not learned. *PLOS Computational Biology*, 2021.

Arrieta et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 2020.

Baggerly & Coombes. Deriving chemosensitivity from cell lines: Forensic bioinformatics and reproducible research in high-throughput biology. *Annals of Applied Statistics*, 2009.

Begley & Ellis. Raise standards for preclinical cancer research. *Nature*, 2012.

Box, Science and statistics. *Journal of the American Statistical Association*, 1976.

Charles et al. Iron-deficiency anaemia in rural Cambodia: community trial of a novel iron supplementation technique. *European Journal of Public Health*, 2011.

Crépon et al. Do Labor Market Policies have Displacement Effects? Evidence from a Clustered Randomized Experiment. *The Quarterly Journal of Economics*, 2013.

Doshi-Velez et al. Accountability of AI under the law: The role of explanation. arXiv: 1711.01134, 2017.

Federal Reserve Board's Division of Research and Statistics, Changes in U.S. Family Finances from 2019 to 2022: Evidence from the Survey of Consumer Finances. October 2023. (p. 11)

Additional References (2/5)

Flinders, Post Office scandal victims have criminal convictions overturned in Court of Appeal, ComputerWeekly.com, 2021.

Flinders, Post Office warned of software flaw in 2006, but failed to alert subpostmaster network, ComputerWeekly.com, 2022.

— Describes ComputerWeekly.com coverage from 2009 through 2022.

Gebru et al. Datasheets for datasets. *Communications of the ACM*, 2021.

Haibe-Kains et al. Transparency and reproducibility in artificial intelligence. *Nature*, 2020.

Heaven. Hundreds of AI tools have been built to catch covid. None of them helped. *MIT Technology Review*, 2021.

Heil et al. Reproducibility standards for machine learning in the life sciences. *Nature Methods*, 2021.

Herndon et al. Does High Public Debt Consistently Stifle Economic Growth? A Critique of Reinhart and Rogoff. *Cambridge Journal of Economics*, 2014.

Lewis. Autocorrect errors in Excel still creating genomics headache. *Nature: News*, 2021.

McKinney et al. International evaluation of an AI system for breast cancer screening. *Nature*, 2020.

Additional References (3/5)

Mittelstadt et al. Explaining explanations in AI. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019.

NASA Earth Observatory, Research satellites for atmospheric sciences, 1978–present: Serendipity and stratospheric ozone, 2001. Accessed 2021.

Open Science Collaboration, Estimating the reproducibility of psychological science. *Science*, 2015.

Pukelsheim, Robustness of Statistical Gossip and the Antarctic Ozone Hole. *Letters to the Editor: The IMS Bulletin*, 1990.

Rappaport et al. Randomized controlled trial assessing the efficacy of a reusable fish-shaped iron ingot to increase hemoglobin concentration in anemic, rural Cambodian women. *The American Journal of Clinical Nutrition*, 2017.

Reinhart & Rogoff, Growth in a time of debt. *The American Economic Review*, 2010.

Roberts et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nature Machine Intelligence*, 2021.

Silberzahn et al. Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, 2018.

Additional References (4/5)

Stodden et al. *Implementing Reproducible Research*, CRC Press, 2014.

Stodden et al. Enhancing reproducibility for computational methods. *Science*, 2016.

Sweney, What is the Post Office Horizon IT scandal all about?, *The Guardian*, 2024.

Vartak et al. MODELDB: A system for machine learning model management. *Proceedings of the Workshop on Human-In-the-Loop Data Analytics*, 2016.

Wang & Long. Addressing Common Misuses and Pitfalls of *P* values in Biomedical Research. *Cancer Research*, 2022.

Wieringa et al. Low Prevalence of Iron and Vitamin A Deficiency among Cambodian Women of Reproductive Age. *Nutrients*, 2016.

Winkler et al. Association Between Surgical Skin Markings in Dermoscopic Images and Diagnostic Performance of a Deep Learning Convolutional Neural Network for Melanoma Recognition. *JAMA Dermatology*, 2019.

Yu. Stability. *Bernoulli*, 2013.

Yu. Stability expanded in reality. *Harvard Data Science Review*, 2020.

Yu, Kumbier. Veridical data science. *Proceedings of the National Academy of Sciences*, 2020.

Zeeberg et al. Mistaken Identifiers: Gene name errors can be introduced inadvertently when using Excel in bioinformatics. *BMC Bioinformatics*, 2004.

Additional References (5/5)

Zhang et al. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making, *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020.

Ziemann et al. Gene name errors are widespread in the scientific literature. *Genome Biology*, 2016.

Ziemann, Abeysooriya. Excel autocorrect errors still plague genetic research, raising concerns over scientific rigour. The Conversation, 2021.

Image References

https://commons.wikimedia.org/wiki/File:Money_saving_growth.jpg (Creative Commons CC0 1.0 Universal Public Domain Dedication)

https://commons.wikimedia.org/wiki/File:Chest_X-ray_2346.jpg (Creative Commons CC0 1.0 Universal Public Domain Dedication)

https://commons.wikimedia.org/wiki/File:Wikimedia_in_Education_illustration_books.svg (Creative Commons CC0 1.0 Universal Public Domain Dedication)