# Discussion of “A hierarchy of limitations in machine learning”

**Momin M. Malik**

Senior Data Science Analyst – AI Ethics, Center for Digital Health, Mayo Clinic
School of Social Policy & Practice, University of Pennsylvania
Institute in Critical Quantitative, Computational, & Mixed Methodologies

Philosophy of Science & Technology of Computer Simulation
Höchstleistungsrechenzentrum Stuttgart, University of Stuttgart
[online], 2024 June 06. **Slides: https://MominMalik.com/hlrs2022.pdf**

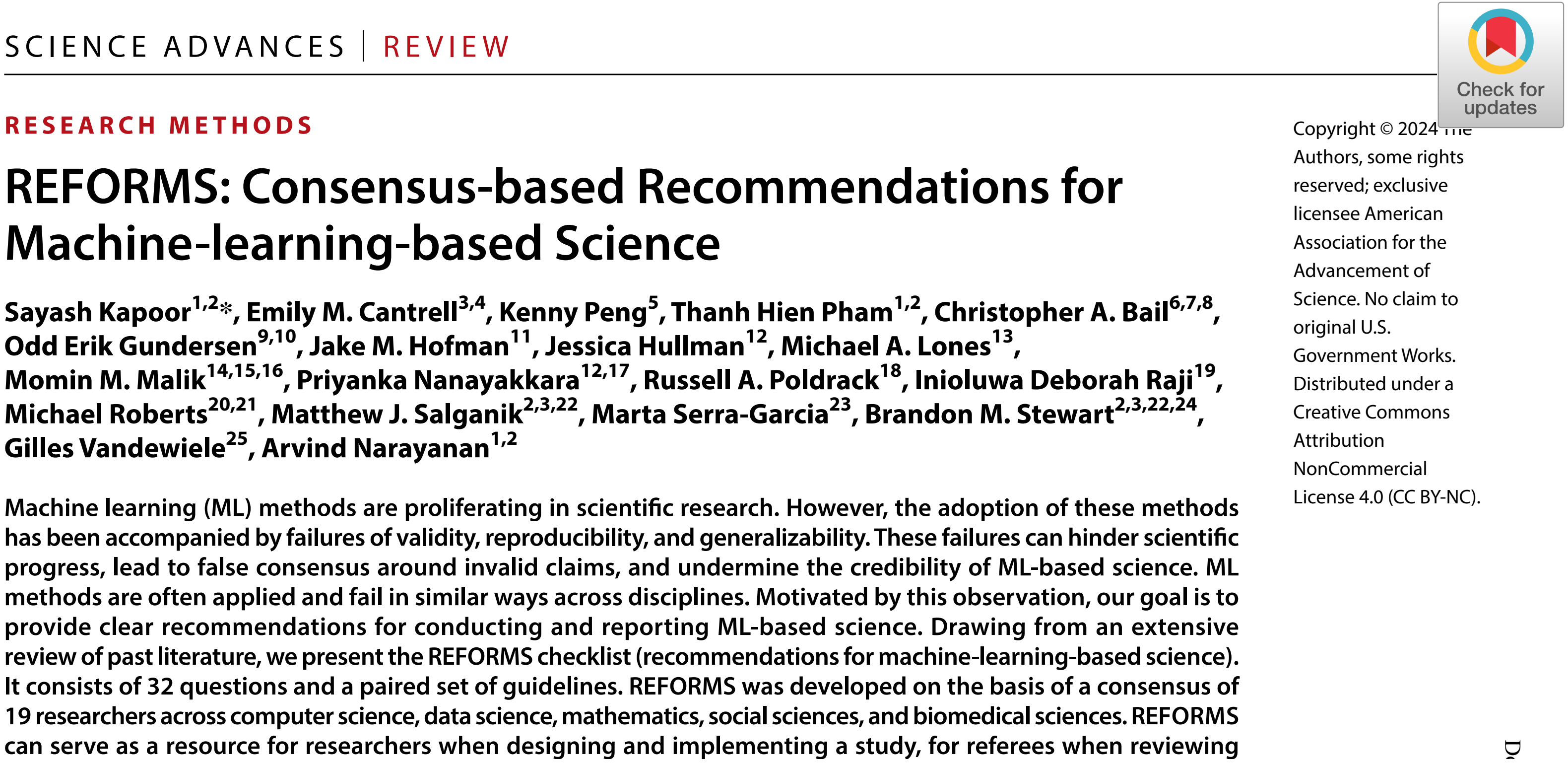
# Update: reproducibility in ML out!

**Paper out!**

The effect of dependencies on machine learning

Other concerns

References



SCIENCE ADVANCES | REVIEW

RESEARCH METHODS

## REFORMS: Consensus-based Recommendations for Machine-learning-based Science

**Sayash Kapoor[1,2]*, Emily M. Cantrell[3,4], Kenny Peng[5], Thanh Hien Pham[1,2], Christopher A. Bail[6,7,8], Odd Erik Gundersen[9,10], Jake M. Hofman[11], Jessica Hullman[12], Michael A. Lones[13], Momin M. Malik[14,15,16], Priyanka Nanayakkara[12,17], Russell A. Poldrack[18], Inioluwa Deborah Raji[19], Michael Roberts[20,21], Matthew J. Salganik[2,3,22], Marta Serra-Garcia[23], Brandon M. Stewart[2,3,22,24], Gilles Vandewiele[25], Arvind Narayanan[1,2]**

Copyright © 2024 The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original U.S. Government Works. Distributed under a Creative Commons Attribution NonCommercial License 4.0 (CC BY-NC).

Machine learning (ML) methods are proliferating in scientific research. However, the adoption of these methods has been accompanied by failures of validity, reproducibility, and generalizability. These failures can hinder scientific progress, lead to false consensus around invalid claims, and undermine the credibility of ML-based science. ML methods are often applied and fail in similar ways across disciplines. Motivated by this observation, our goal is to provide clear recommendations for conducting and reporting ML-based science. Drawing from an extensive review of past literature, we present the REFORMS checklist (recommendations for machine-learning-based science). It consists of 32 questions and a paired set of guidelines. REFORMS was developed on the basis of a consensus of 19 researchers across computer science, data science, mathematics, social sciences, and biomedical sciences. REFORMS can serve as a resource for researchers when designing and implementing a study, for referees when reviewing

# The effect of dependencies in machine learning

# Summary

- Machine learning generally ignores dependencies between observations (assumes iid)
- This is usually justified for model *fitting*; and the major impact of dependencies is on *inference*.
- The problem is in our ability to estimate model *performance*; we think we are doing better than we actually are

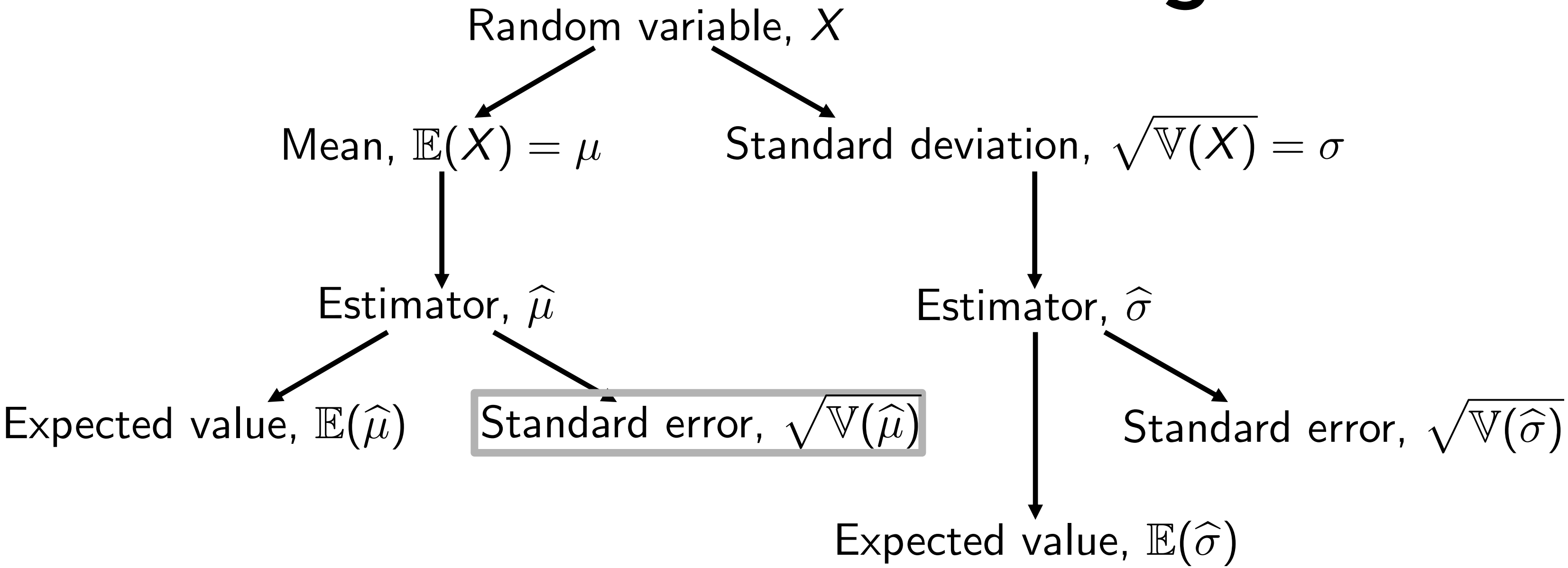
# Without (conditionally) iid, nonparametric models are unidentifiable

"A number of problems, some quite fundamental, occur when nonparametric regression is attempted in the presence of correlated errors. Indeed, **in the most general setting where no parametric shape is assumed for the mean nor the correlation function, the model is essentially unidentifiable**, so that it is theoretically impossible to estimate either function separately." (Opsomer et al. 2001)

# Estimator properties of estimates of model performance

- “Metaprediction—the prediction about predictions—is… an integral component of the predictive enterprise itself… Indeed, to characterize someone as a reliable predictor… is in effect to predict on one’s own account that this agent’s predictions will generally come true—and is thereby to make a metaprediction of sorts.” (Rescher 1998)
- Metaprediction to at least third order is worthwhile
- First-order prediction: the prediction itself, $\hat{y}$
- Second-order prediction: $\mathbb{E}(\hat{y})$, estimate via CV
- Third-order prediction: $\mathbb{E}(\hat{\mathbb{E}}(\hat{y}))$, look at properties of CV

# Inference (in statistics): If uncertainty of an estimator is less than the “signal”



Random variable, $X$

Mean, $\mathbb{E}(X) = \mu$

Standard deviation, $\sqrt{\mathbb{V}(X)} = \sigma$

Estimator, $\widehat{\mu}$

Estimator, $\widehat{\sigma}$

Expected value, $\mathbb{E}(\widehat{\mu})$

Standard error, $\sqrt{\mathbb{V}(\widehat{\mu})}$

Standard error, $\sqrt{\mathbb{V}(\widehat{\sigma})}$

Expected value, $\mathbb{E}(\widehat{\sigma})$

The *variance* of the *estimator* of the *mean* gives us the uncertainty of the estimate, and is given the special name of the *standard error*. If the uncertainty is small enough, we say we have made an *inference* to the underlying data-generating process.

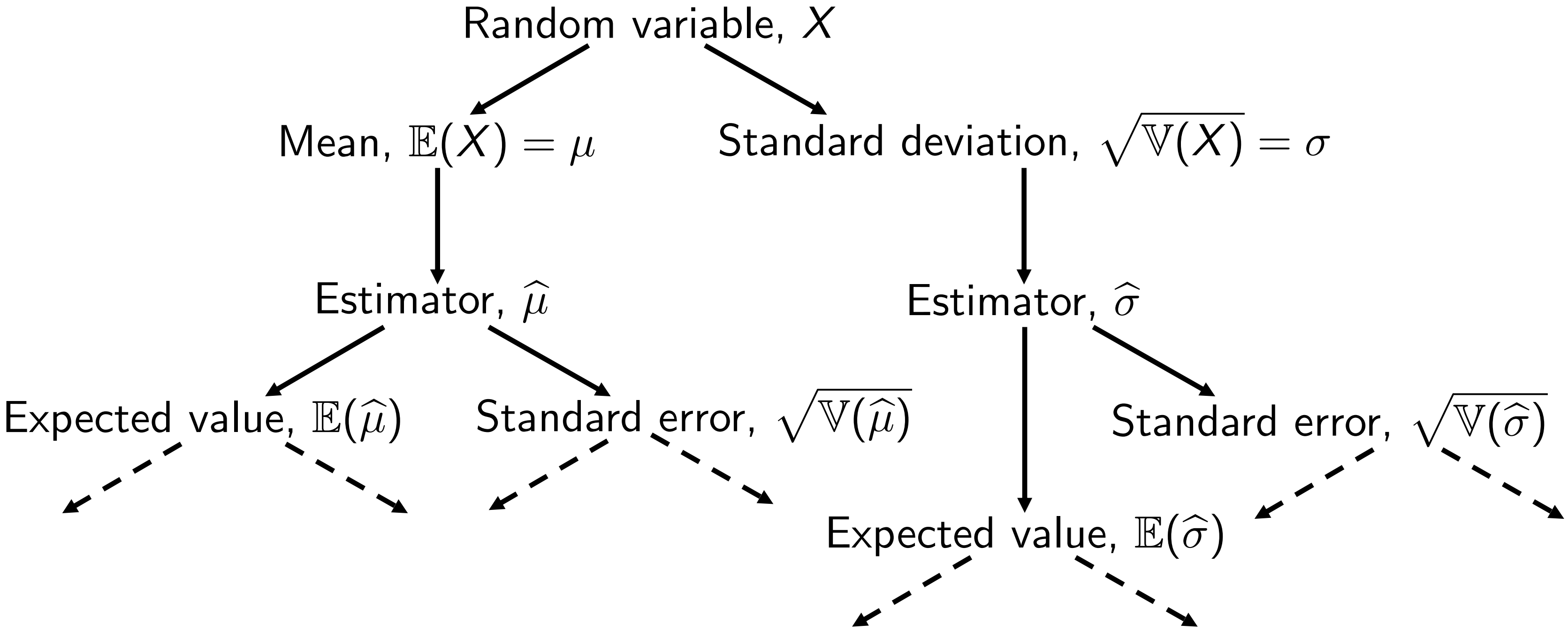
Paper out!

**The effect of dependencies on machine learning**

Other concerns

References

# Going on *ad infinitum*…



Random variable, $X$

Mean, $\mathbb{E}(X) = \mu$

Standard deviation, $\sqrt{\mathbb{V}(X)} = \sigma$

Estimator, $\widehat{\mu}$

Estimator, $\widehat{\sigma}$

Expected value, $\mathbb{E}(\widehat{\mu})$

Standard error, $\sqrt{\mathbb{V}(\widehat{\mu})}$

Standard error, $\sqrt{\mathbb{V}(\widehat{\sigma})}$

Expected value, $\mathbb{E}(\widehat{\sigma})$

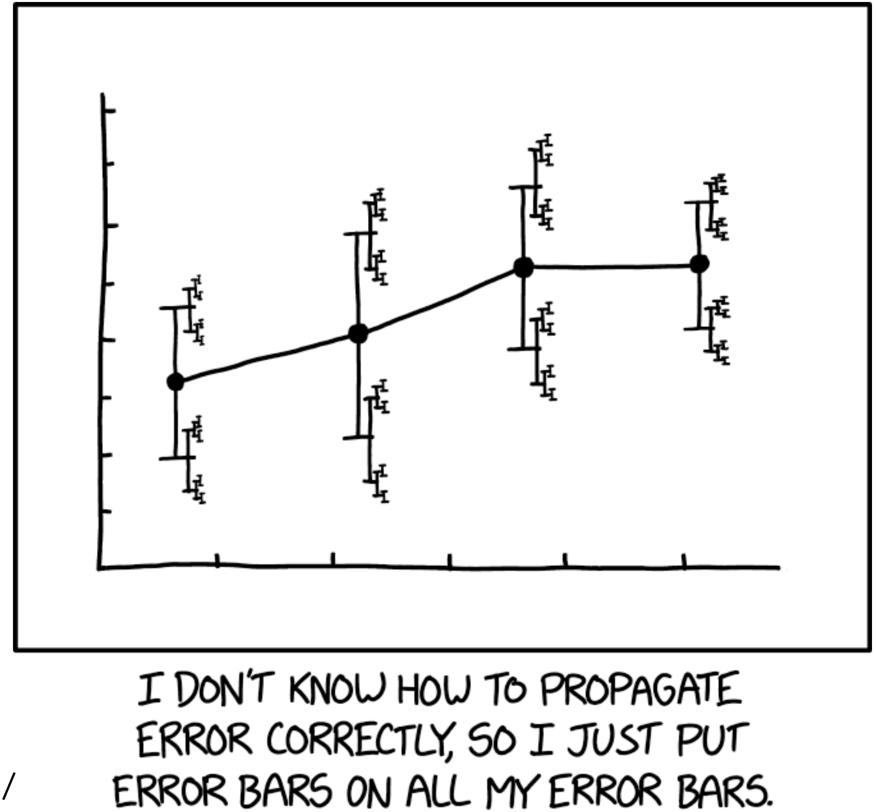
Paper out!

**The effect of dependencies on machine learning**

Other concerns

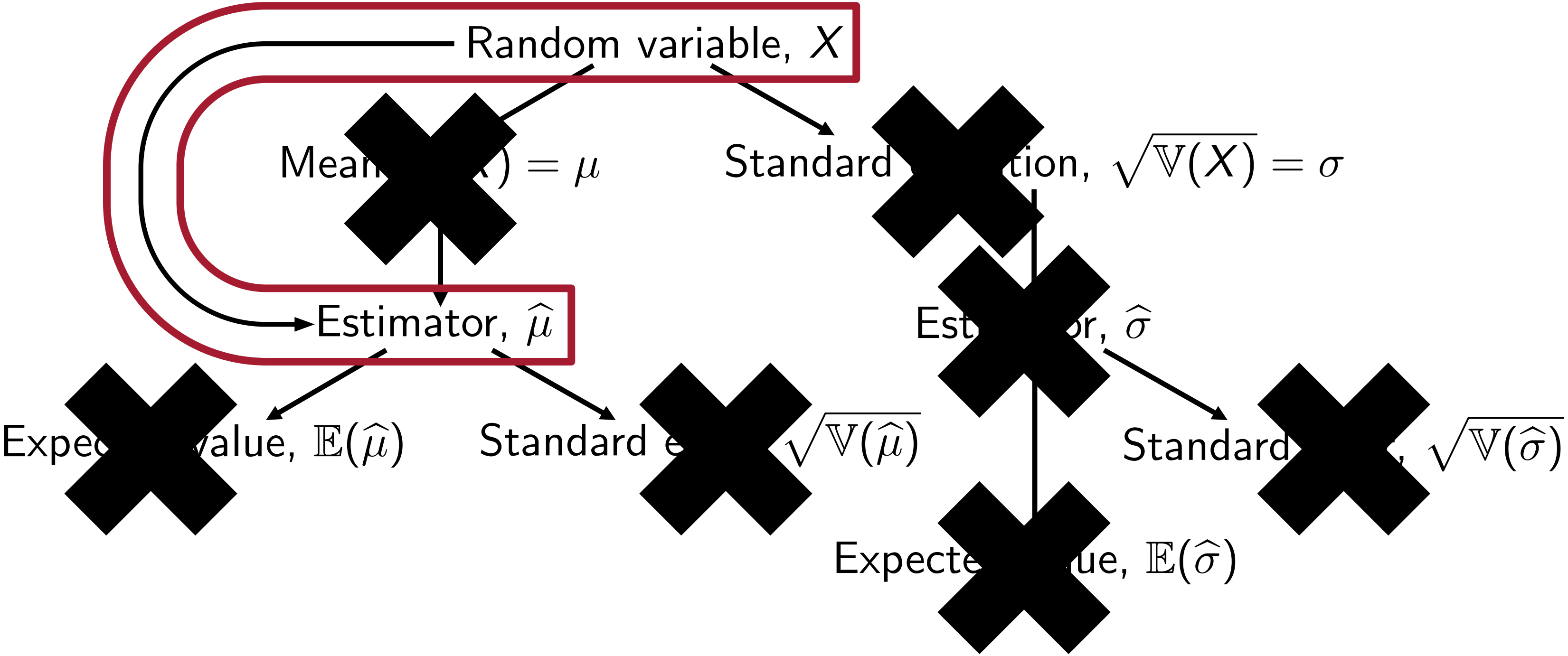
References

# Going on *ad infinitum*…


I DON'T KNOW HOW TO PROPAGATE ERROR CORRECTLY, SO I JUST PUT ERROR BARS ON ALL MY ERROR BARS.

https://xkcd.com/2110/

# Machine learning: Instrumentalist


Random variable, $X$
Mea ) $= \mu$
Standard tion, $\sqrt{\mathbb{V}(X)} = \sigma$
Estimator, $\widehat{\mu}$
Est r, $\widehat{\sigma}$
Expe value, $\mathbb{E}(\widehat{\mu})$
Standard e $\sqrt{\mathbb{V}(\widehat{\mu})}$
Standard , $\sqrt{\mathbb{V}(\widehat{\sigma})}$
Expecte ue, $\mathbb{E}(\widehat{\sigma})$

ML skips over the entire machinery of inference, creating estimators only to recover some aspect of held-out data. (*Statistical machine learning* brings theory back in, but for the purpose of seeing what best predicts, not what recovers information.) Part of what we argue in "REFORMS": must bring back in examination of properties of estimators of estimators (like held-out data)

# Matrix bias-variance decomposition

$$
\begin{aligned}
\operatorname{err}(\hat{\mu}) &= \tfrac{1}{n}\mathbb{E}_f \| Y - \widehat{Y} \|_2^2 \\
&= \frac{1}{n}\left[\mathbb{E}_f \| Y \|_2^2 + \mathbb{E}_f \| \widehat{Y} \|_2^2 - 2\mathbb{E}_f(Y^T \widehat{Y})\right] \\
&= \frac{1}{n}\left[\mathbb{E}_f \| Y \|_2^2 + \mathbb{E}_f \| \widehat{Y} \|_2^2 - 2\operatorname{tr}\mathbb{E}_f(Y \widehat{Y}^T)\right] \\
&\quad + \frac{1}{n}\left[\mu^T \mu + \mathbb{E}_f(\widehat{Y})^T \mathbb{E}_f(\widehat{Y}) + 2\operatorname{tr}\mu\mathbb{E}_f(\widehat{Y})^T\right] \\
&\quad + \frac{1}{n}\left[-\mu^T \mu - \mathbb{E}_f(\widehat{Y}) \mathbb{E}_f(\widehat{Y})^T - 2\mu^T\mathbb{E}_f(\widehat{Y})\right] \\
&= \frac{1}{n}\left[\operatorname{tr}\Sigma + \| \mu - \mathbb{E}(\widehat{Y}) \|_2^2 + \operatorname{tr}\operatorname{Var}_f(\widehat{Y}) - 2\operatorname{tr}\operatorname{Cov}_f(Y, \widehat{Y})\right]
\end{aligned}
$$

irreducible (“Bayes”) error

bias squared

variance of the estimator

“optimism”

# Classic argument for CV

Training:

$$\begin{aligned} \mathrm{err}(\hat{\mu}) &= \tfrac{1}{n}\mathbb{E}_f\|Y - \widehat{Y}\|_2^2 \\ &= \frac{1}{n}\left[\mathrm{tr}\,\Sigma + \|\mu - \mathbb{E}(\widehat{Y})\|_2^2 + \mathrm{tr}\,\mathrm{Var}_f(\widehat{Y}) - 2\,\mathrm{tr}\,\mathrm{Cov}_f(Y, \widehat{Y})\right] \end{aligned}$$

Testing:

$$\begin{aligned} \mathrm{Err}(\hat{\mu}) &= \tfrac{1}{n}\mathbb{E}_f\|Y^* - \widehat{Y}\|_2^2 \\ &= \frac{1}{n}\left[\mathrm{tr}\,\Sigma + \|\mu - \mathbb{E}(\widehat{Y})\|_2^2 + \mathrm{tr}\,\mathrm{Var}_f(\widehat{Y}) - \cancel{2\,\mathrm{tr}\,\mathrm{Cov}_f(Y^*, \widehat{Y})}\right] \end{aligned}$$

The difference is the *optimism* (Efron 2004; Rosset and Tibshirani 2020):

$$\mathrm{Opt}(\hat{\mu}) = \mathrm{Err}(\hat{\mu}) - \mathrm{err}(\hat{\mu}) = \frac{2}{n}\,\mathrm{tr}\,\mathrm{Cov}_f(Y, \widehat{Y})$$

# Apply this to non-iid data

- Imagine we have, for $\mathbf{\Sigma}_{ii} = \sigma^2$ and $\mathbf{\Sigma}_{ij} = \rho\sigma^2, \quad i \neq j$

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} \sim \mathcal{N}\left( \begin{bmatrix} \mathbf{X} \\ \mathbf{X} \end{bmatrix} \boldsymbol{\beta}, \begin{bmatrix} \mathbf{\Sigma} & \rho\sigma^2 \mathbf{1}\mathbf{1}^T \\ \rho\sigma^2 \mathbf{1}\mathbf{1}^T & \mathbf{\Sigma} \end{bmatrix} \right)$$

- Then, optimism in the training set is:

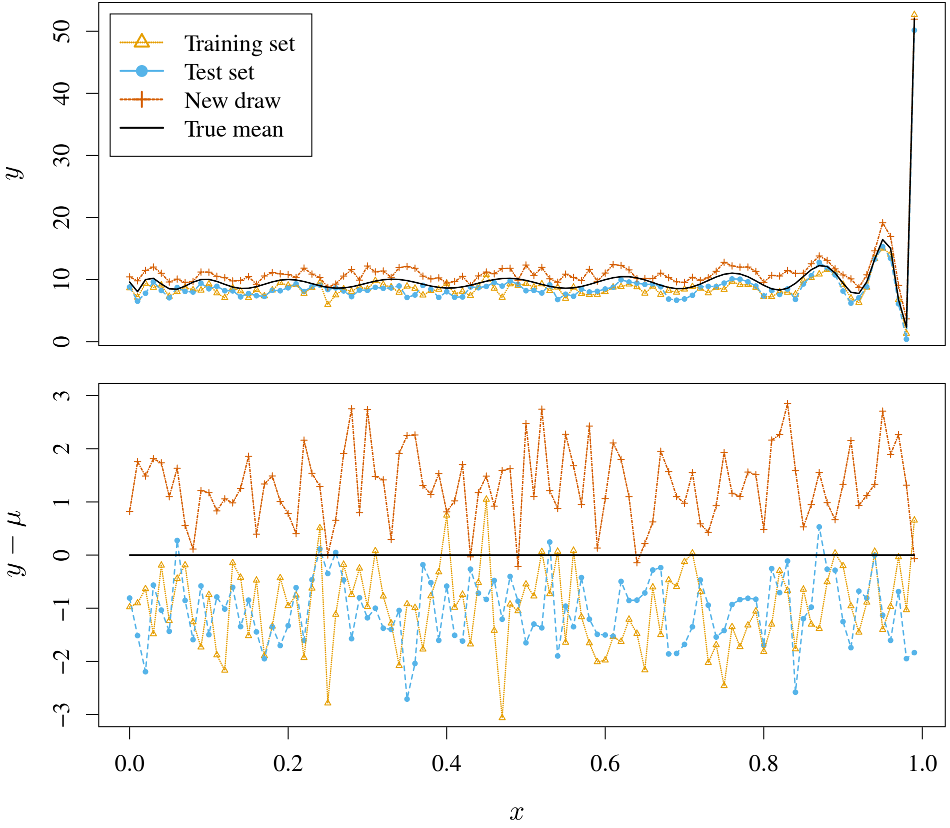
$$\frac{2}{n} \operatorname{tr} \operatorname{Cov}_f(Y_1, \widehat{Y}_1) = \frac{2}{n} \operatorname{tr} \operatorname{Cov}_f(Y_1, \mathbf{H} Y_1) = \frac{2}{n} \operatorname{tr} \mathbf{H} \operatorname{Var}_f(Y_1) = \frac{2}{n} \operatorname{tr} \mathbf{H\Sigma}$$

- But test set also has nonzero optimism!

$$\frac{2}{n} \operatorname{tr} \operatorname{Cov}_f(Y_2, \widehat{Y}_1) = \frac{2}{n} \operatorname{tr} \operatorname{Cov}_f(Y_2, \mathbf{H} Y_1) = \frac{2\rho\sigma^2}{n} \operatorname{tr} \mathbf{H}\mathbf{1}\mathbf{1}^T = 2\rho\sigma^2$$

# One draw as an example

Correlation between observations can pull training and test observations close to one another, but potentially far from an independent draw



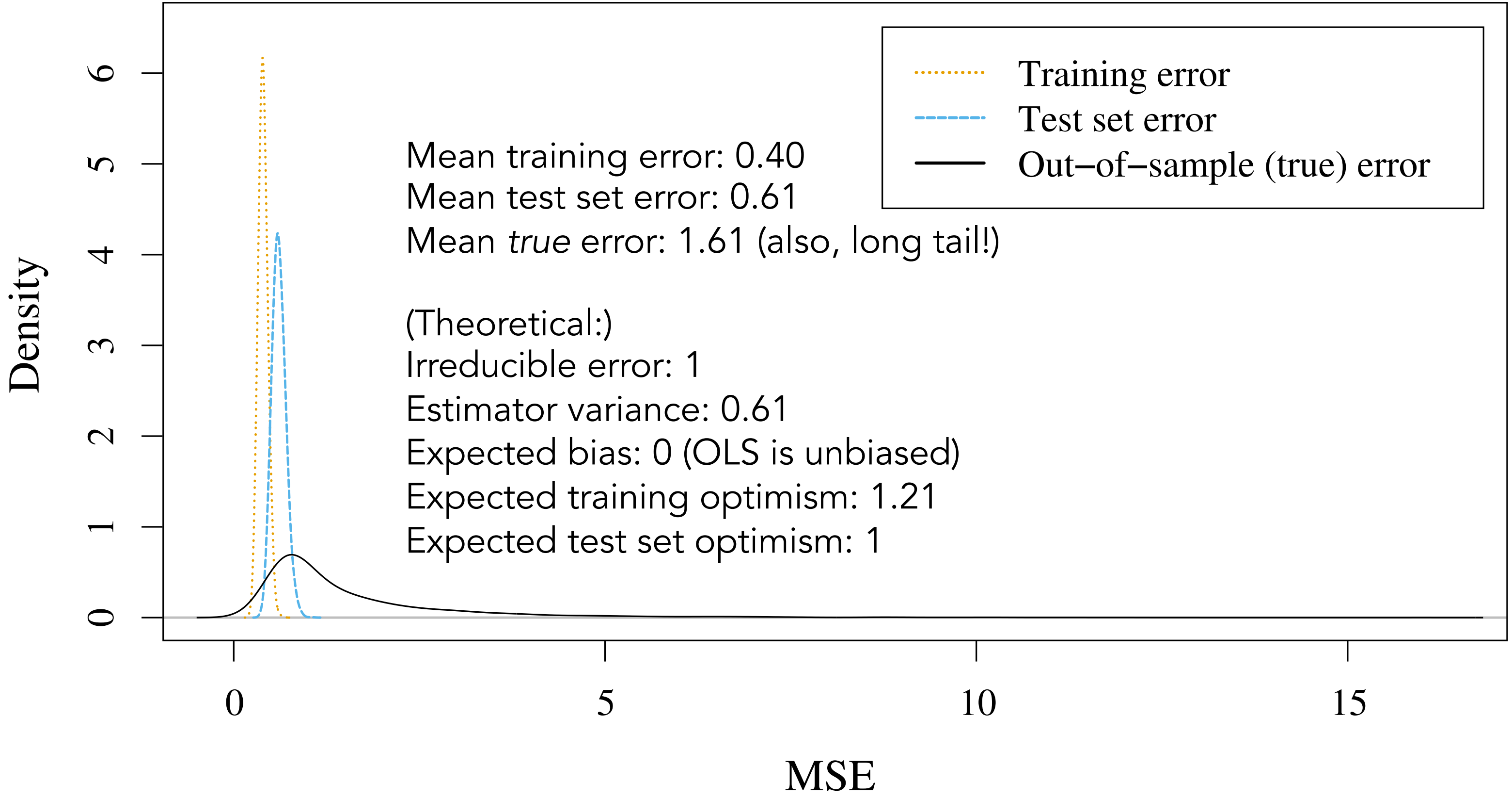
Paper out!

**The effect of dependencies on machine learning**

Other concerns

References

# Simulated MSE

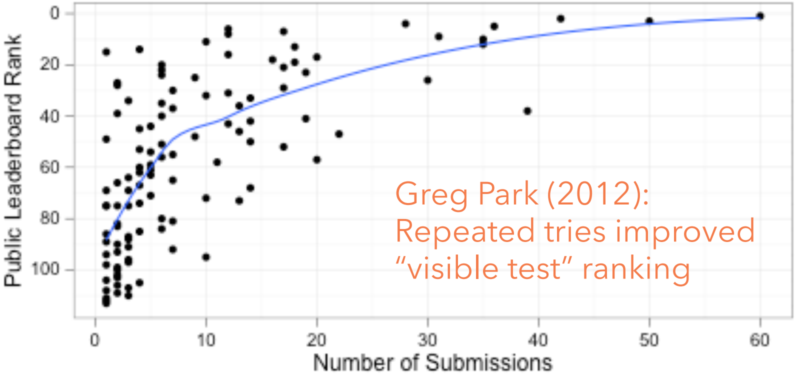


Mean training error: 0.40
Mean test set error: 0.61
Mean *true* error: 1.61 (also, long tail!)

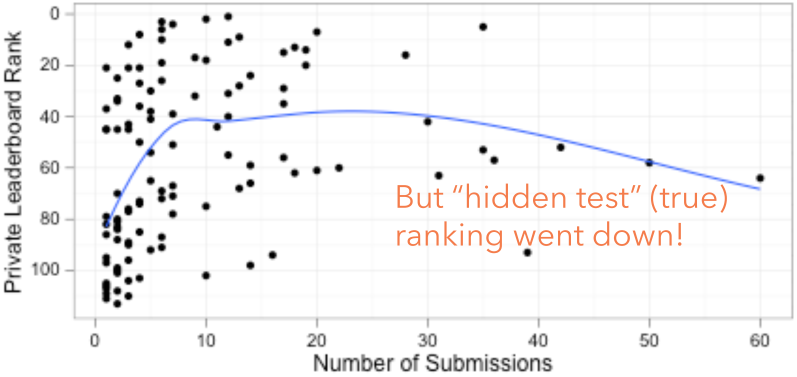
(Theoretical:)
Irreducible error: 1
Estimator variance: 0.61
Expected bias: 0 (OLS is unbiased)
Expected training optimism: 1.21
Expected test set optimism: 1

# Dependencies and CV: examples

- Highly-cited "Twitter mood predicts the stock market" trains on future values, tests on past values: that is "time-traveling"! (see critique by Lachanski and Pav 2017)
- A colleague of mine trained a model to recognize birds on his windowsill in webcam images, splitting frames randomly…
- Park (2012) has a great example of overfitting to the test set in Kaggle. Having a "private leaderboard" helps catch overfitting in Kaggle (see also Dwork et al. 2015)
  - I agree with Wagstaff (2012) that in research, it's probably not worth having a test set we only use once (do we give up if performance is bad?). But we *should* temper our claims, and do out-of-sample testing



Greg Park (2012): Repeated tries improved "visible test" ranking



But "hidden test" (true) ranking went down!

# Applying to networks

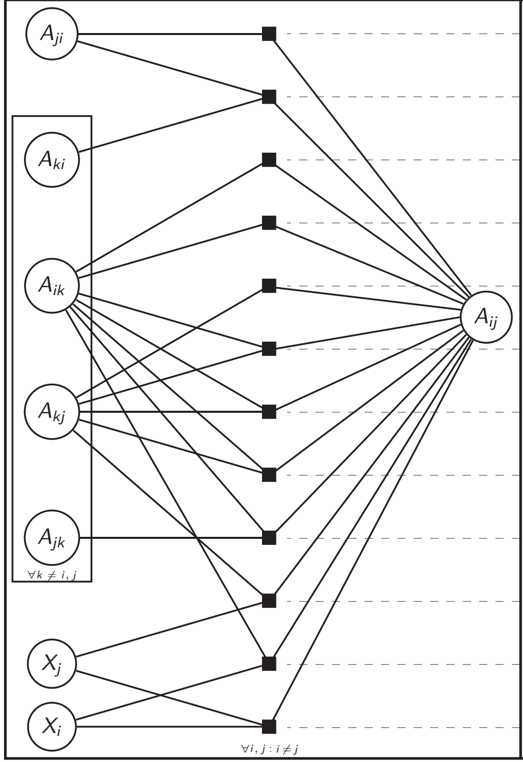
- This formulation would apply to a network autocorrelation model, where network is nuisance parameter
- But what if we are modeling the *edges*, which represent dependencies between observations?

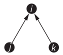
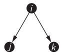
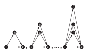
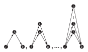
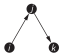
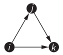
# Modeling the *edges*

| | $Y$ | $X_1$ | $X_2$ | $\cdots$ | $X_d$ |
|---|---|---|---|---|---|
| 1 | $y_1$ | $x_{11}$ | $x_{12}$ | $\cdots$ | $x_{1d}$ |
| 2 | $y_2$ | $x_{21}$ | $x_{22}$ | $\cdots$ | $x_{2d}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ |
| $n$ | $y_n$ | $x_{n1}$ | $x_{n2}$ | $\cdots$ | $x_{nd}$ |

| *index* | *from* | *to* | $Y$ | $W_1$ | $W_2$ | $W_3$ | $\cdots$ |
|---|---|---|---|---|---|---|---|
| $e_1$ | 1 | 2 | $y_{12}$ | $\mathbf{1}(x_{11} = x_{21})$ | $x_{12} - x_{22}$ | $x_{13}$ | $\cdots$ |
| $e_2$ | 2 | 3 | $y_{23}$ | $\mathbf{1}(x_{11} = x_{31})$ | $x_{12} - x_{32}$ | $x_{13}$ | $\cdots$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | |
| $e_{n+1}$ | 2 | 1 | $y_{21}$ | $\mathbf{1}(x_{21} = x_{11})$ | $x_{22} - x_{12}$ | $x_{23}$ | $\cdots$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | |
| $e_{2\binom{n}{2}}$ | $n-1$ | $n$ | $y_{(n-1)n}$ | $\mathbf{1}(x_{(n-1)1} = x_{n1})$ | $x_{(n-1)2} - x_{n2}$ | $x_{(n-1)3}$ | $\cdots$ |

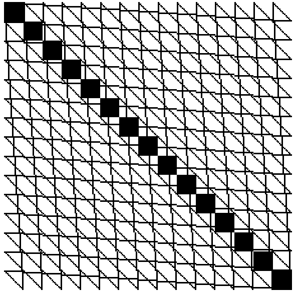
# But dyads are dependent too!



| Factor graph | Parameter name | Network Motif | Parameterization | Matrix notation |
|---|---|---|---|---|
| | mutual dyads | | $\sum_{i<j} A_{ij} A_{ji}$ | $\frac{1}{2}\,\mathrm{tr}\left(\mathbf{A}\mathbf{A}^T\right)$ |
| | in-two-stars |  | $\sum_{(i,j,k)} A_{ji} A_{ki}$ | $\mathrm{sum}\left(\mathbf{A}\mathbf{A}^T\right) - \mathrm{tr}\left(\mathbf{A}\mathbf{A}^T\right)$ |
| | out-two-stars |  | $\sum_{(i,j,k)} A_{ij} A_{ik}$ | $\mathrm{sum}\left(\mathbf{A}^T\mathbf{A}\right) - \mathrm{tr}\left(\mathbf{A}^T\mathbf{A}\right)$ |
| | geom. weighted out-degrees | — | $\sum_i \exp\left\{-\alpha \sum_k A_{ik}\right\}$ | $\mathrm{sum}\left(\exp\{-\alpha\ \mathrm{rowsum}\left(\mathbf{A}\right)\}\right)$ |
| | geom. weighted in-degrees | — | $\sum_j \exp\left\{-\alpha \sum_k A_{kj}\right\}$ | $\mathrm{sum}\left(\exp\{-\alpha\ \mathrm{colsum}\left(\mathbf{A}\right)\}\right)$ |
| | alternating transitive $k$-triplets |  | $\lambda \sum_{i,j} A_{ij} \left\{1 - \left(1 - \frac{1}{\lambda}\right)^{\sum_{k \neq i,j} A_{ik} A_{kj}}\right\}$ | $\lambda\ \mathrm{sum}\left(\mathbf{A} \odot \left(1 - \left(1 - \frac{1}{\lambda}\right)^{\mathbf{AA} - \mathrm{diag}(\mathbf{AA})}\right)\right)$ |
| | alternating indep. two-paths |  | $\lambda \sum_{i,j} \left\{1 - \left(1 - \frac{1}{\lambda}\right)^{\sum_{k \neq i,j} A_{ik} A_{kj}}\right\}$ | $\lambda\ \mathrm{sum}\left(1 - \left(1 - \frac{1}{\lambda}\right)^{\mathbf{AA} - \mathrm{diag}(\mathbf{AA})}\right)$ |
| | two-paths (mixed two-stars) |  | $\sum_{(i,k,j)} A_{ik} A_{kj}$ | $\mathrm{sum}\left(\mathbf{AA}\right) - \mathrm{tr}\left(\mathbf{AA}\right)$ |
| | transitive triads |  | $\sum_{(i,j,k)} A_{ij} A_{jk} A_{ik}$ | $\mathrm{tr}\left(\mathbf{AAA}^T\right)$ |
| | activity effect | | $\sum_i X_i \sum_j A_{ij}$ | $\mathrm{sum}\left(\mathbf{X} \odot \mathrm{rowsum}\left(\mathbf{A}\right)\right)$ |
| | popularity effect | | $\sum_j X_j \sum_i A_{ij}$ | $\mathrm{sum}\left(\mathbf{X} \odot \mathrm{colsum}\left(\mathbf{A}\right)\right)$ |
| | similarity effect | | $\sum_{i,j} A_{ij} \left(1 - \frac{\lvert X_i - X_j \rvert}{\max_{k,l} \lvert X_k - X_l \rvert}\right)$ | $\mathrm{sum}\left(\mathbf{A} \odot \mathbf{S}\right)$ |

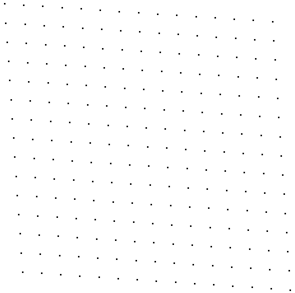
Graphical model and matrix notations for ERGM specification terms given in: Snijders et al. 2006. Joint work with Antonis Manousis and Naji Shajarisales, 2018.

# Covariance structure of *edges* (*n* = 15)

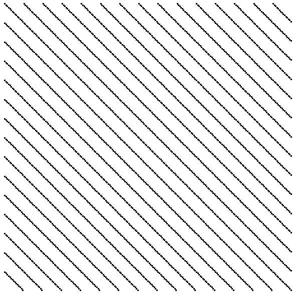


Total covariance between dyads

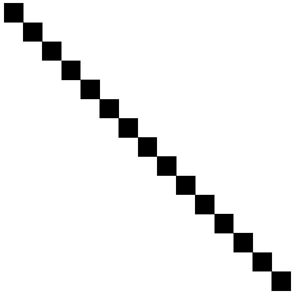
- The pairs of edges that are present together, or aren't present together
- Note: A theoretical construct, since we only see edges once (or once per time slice)



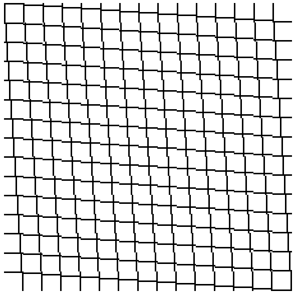
Mutual dyads



In-2-stars

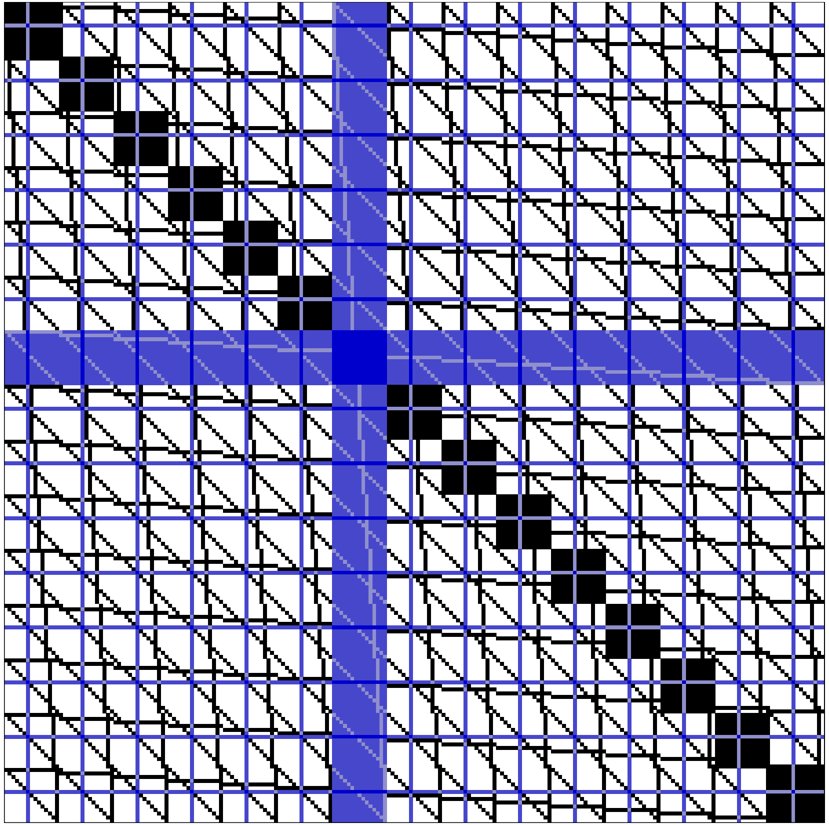


Out-2-stars



2-paths

# No data split would allow generalizable estimates



- Partition nodes into training and test sets?
  - Breaks up triads; omitted edges “share” information across training and test (diagram: blue are edges that include node 7)
- Partition dyads?
  - Breaks up nodes; even worse
- Can’t *eliminate*, but can *minimize* optimism by careful data splitting

# Other concerns

Quantification

Bias-variance tradeoff implies a “false” model can ‘predict’ better than the “true” model

Explanation (correlation) vs. prediction (causation)

Levels of prediction

“Prediction” and other language

Guarantees of what?

# Quantification and “ways to understand a person” (see Kiviat 2023)

| | **As a case (quant)** | **In narrative (qual)** |
|---|---|---|
| Context/circumstance | Stripped away | Key |
| Mental states | Absent (for the most part) | Crucial; constitutive |
| Relevant features | Determined in advance | Emergent |
| Orientation to time | Atemporal | Chronological |
| Ordering of features | Unimportant | Meaningful |
| Other actors | Invisible | Often present |
| Causal logic | Mathematical | Theoretical |
| Boost predictive validity | Add cases | Know person better |

Slide from Barbara Kiviat, based on “Bowker and Star 2000; Bruner 1986; Desrosières 1998; Espeland 1998; Espeland and Stevens 1998, 2008; Fourcade and Healy 2017; Hacking 1990; Porter 1994, 1995; Ricouer 1998; White 1980, 1984”. I would add: Abbott 1988

# Unbiased vs. minimizing loss: “True” model can “predict” worse!

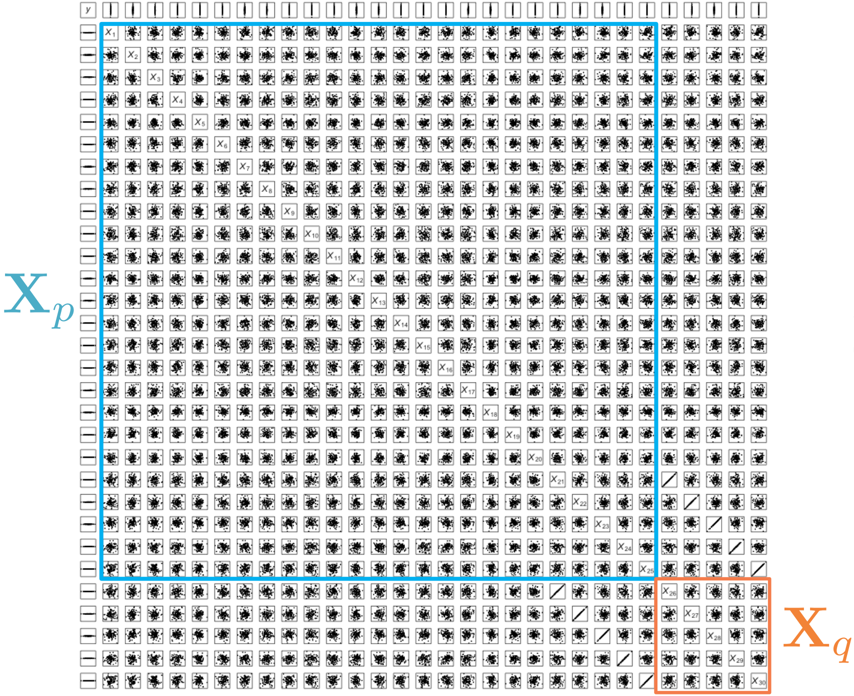
- A linear data-generating process.

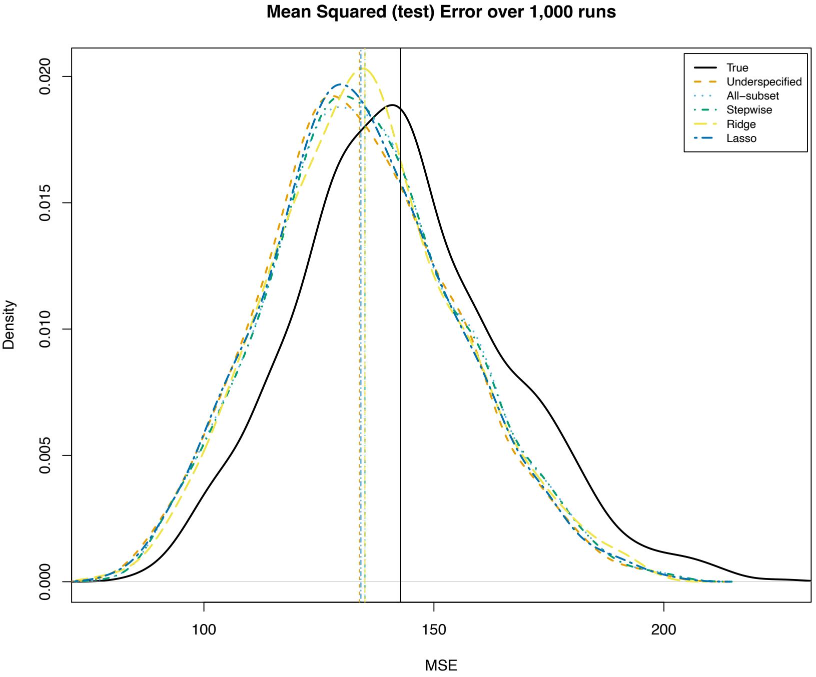
$$\mathbf{y} \sim \mathcal{N}\left(\beta_{\boldsymbol{p}} \mathbf{X}_{\boldsymbol{p}} + \beta_{\boldsymbol{q}} \mathbf{X}_{\boldsymbol{q}}, \sigma^2 \mathbf{I}\right)$$

- Wu et al. (2007): Fitting only $\mathbf{X}_p$ has lower expected MSE than fitting the model that generated the data if and only if:

$$\beta_q^T \mathbf{X}_q^T (\mathbf{I}_n - \mathbf{H}_p) \mathbf{X}_q \beta_p < q\sigma^2$$

# Simulation: 5 weak covariates, each highly correlated with a strong covariate





Simulation of Wu et al. (2007)

# *How* the underspecified model, and regularized models, do better

Paper out!

The effect of dependencies on machine learning

**Other concerns**

Quantification

**Bias-variance tradeoff**

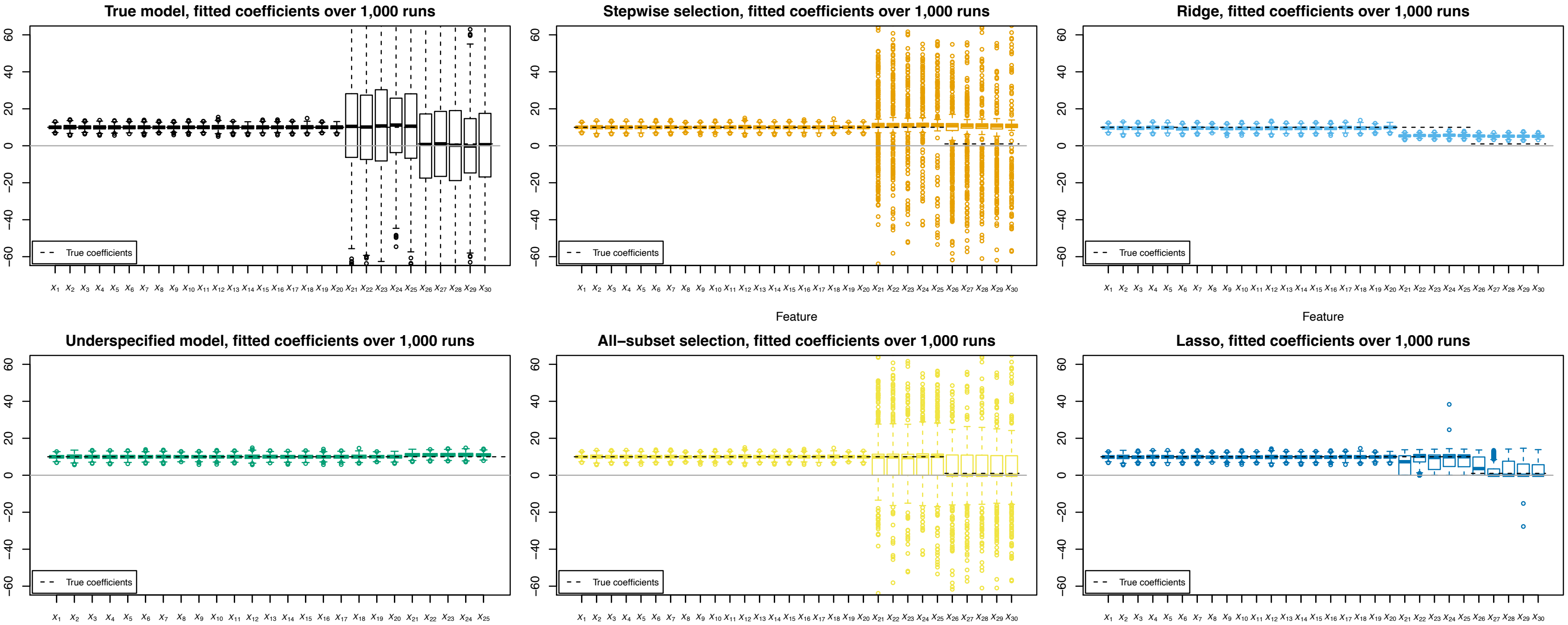
Explanation vs prediction

Levels of prediction

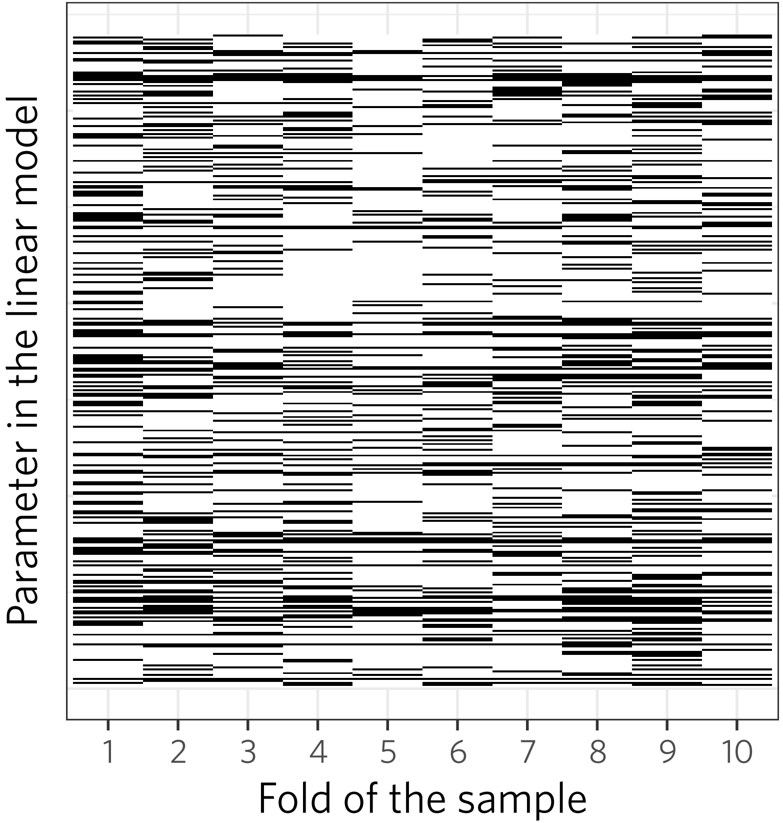
“Prediction”

Guarantees of what?

References



# Explanation (causation) vs. prediction (correlation)



- Very different sets of correlations can “predict” equally well (Mullainathan and Spiess 2017); Breiman (2001) called this the “Rashomon effect” and saw it as a point in favor of prediction over trying to get at causation
- But if we want to intervene, we need causation

Paper out!

The effect of dependencies on machine learning

**Other concerns**

Quantific-ation

Bias-variance tradeoff

Explanation vs prediction

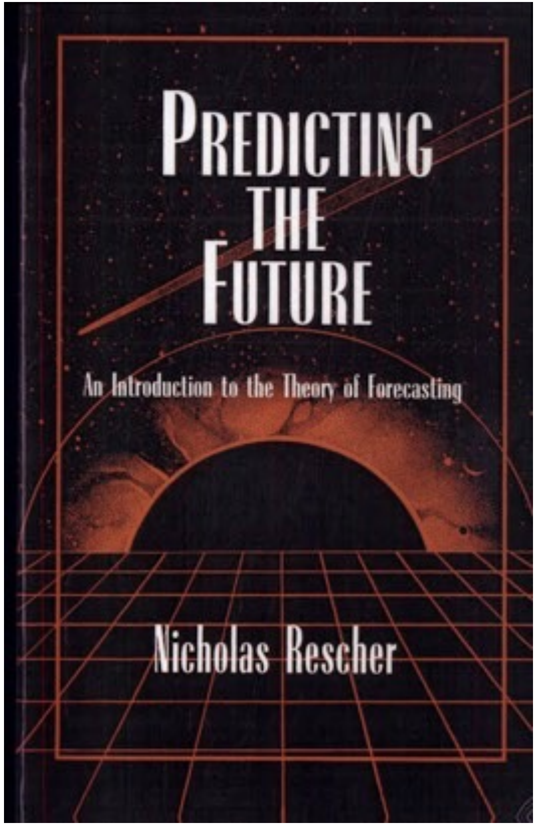
**Levels of prediction**

"Prediction"

Guarantees of what?

References

# Levels of prediction (Rescher 1998)



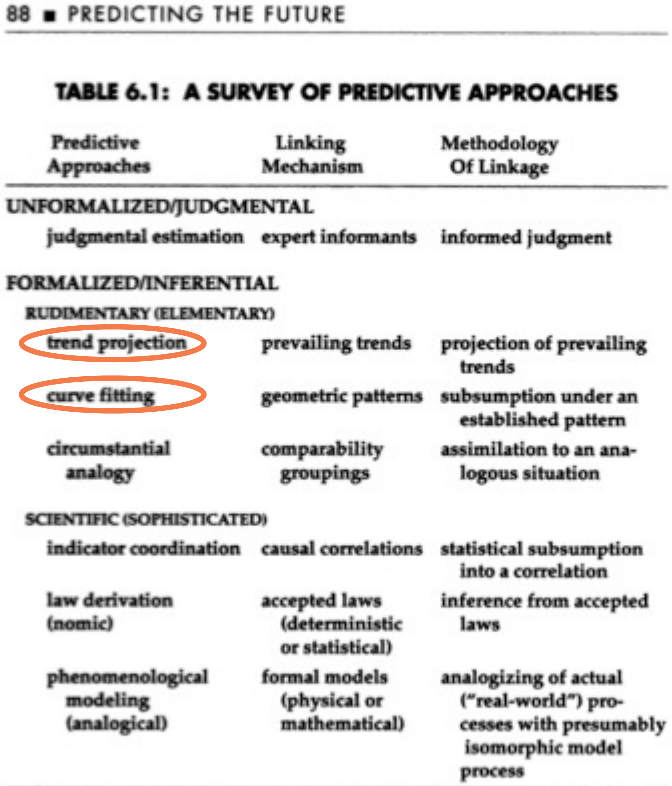


**TABLE 6.1: A SURVEY OF PREDICTIVE APPROACHES**

| Predictive Approaches | Linking Mechanism | Methodology Of Linkage |
|---|---|---|
| UNFORMALIZED/JUDGMENTAL | | |
| judgmental estimation | expert informants | informed judgment |
| FORMALIZED/INFERENTIAL | | |
| RUDIMENTARY (ELEMENTARY) | | |
| trend projection | prevailing trends | projection of prevailing trends |
| curve fitting | geometric patterns | subsumption under an established pattern |
| circumstantial analogy | comparability groupings | assimilation to an analogous situation |
| SCIENTIFIC (SOPHISTICATED) | | |
| indicator coordination | causal correlations | statistical subsumption into a correlation |
| law derivation (nomic) | accepted laws (deterministic or statistical) | inference from accepted laws |
| phenomenological modeling (analogical) | formal models (physical or mathematical) | analogizing of actual ("real-world") processes with presumably isomorphic model process |

Paper out!

The effect of dependencies on machine learning

**Other concerns**

Quantific-ation

Bias-variance tradeoff

Explanation vs prediction

Levels of prediction

**“Prediction”**

Guarantees of what?

References

# “Prediction” and other language

- Communication: **stop saying “prediction” if it is really “correlation”**
  - **The use of ‘prediction’ leads to false, inflated expectations.** Instead of saying “prediction” for post-hoc demonstrations (Gayo-Avello 2012), use “retrodiction”: it is awkward, but that’s what we need. For time series: nowcasting, back-testing (although better language is not enough:
  - Partial correlation (i.e., for “ceteris paribus” interpretations) can be described with “association”
- “Prediction” is overused as it is
  - Statements like ”predict the probability of risk”, or “calculate the probability of a likelihood” exist and are redundant if not nonsensical (akin to, “a probability of a probability [of a probability]”).
    - Probabilities and risks are always latent (and indeed, are hypothetical and metaphysical), so how can we “predict” them? We should say that we *estimate* probabilities and risk (say *estimated probabilities*, etc.), and not overload on synonyms for probability
  - Use “detection” or “classification” if labels are manifest but unknown. E.g., we don’t “predict” race; “detecting” and “predicting” cancer imply two very different tasks; etc.
- **Models, not algorithms** (unless you really do mean an optimization algorithm). Why? Specificity: logistic regression is a *model*, IRLS is an algorithm. Random forests are a *model*, CART is an algorithm. And: we already know “all models are wrong” (Box 1979)

# Guarantees of what?

- The Frequentist-Bayesian issues come back again
- To get information about the world, we want our models to give us $\mathbb{P}(H \mid \mathcal{D})$
- But we want to use methods with frequentist guarantees (e.g., a 95% credible interval, if repeated, will *not* necessarily contain the true value 95% of the time)
- There’s no way to get $\mathbb{P}(H \mid \mathcal{D})$ without a prior, and with priors, we don’t get frequency guarantees
- No frequency analysis is about the specific situation; it’s a property of the *procedure* (including what I did here)

# Generalize to what?

- If by “generalizability,” we mean that a fitted model will apply to very different contexts, probably very few ML models will generalize (at least for the social world)–or at least our theory gives us no guarantees that this will happen
- Our interest is in the quality of predictions that we can make *with a specific model*, but all our analysis refers to is if the ML *procedure* will generalize.
- Note that, despite many in ML claiming that it is Bayesian (e.g., Kevin Murphy’s textbook), data splitting is a deeply frequentist procedure and so is mainstream ML overall

# References

Abbott, Andrew. 1988. “Transcending General Linear Reality.” *Sociological Theory* 6 (2): 169–186. https://doi.org/10.2307/202114

Box, George E. P. 1979. “Robustness in the Strategy of Scientific Model Building.” Technical Report #1954. Mathematics Research Center, University of Wisconsin-Madison.

Breiman, Leo. 2001. “Statistical Modeling: The Two Cultures.” *Statistical Science* 16 (3): 199–231. https://doi.org/10.1214/ss/1009213726

Dwork, Cynthia, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. 2015. “The Reusable Holdout: Preserving Validity in Adaptive Data Analysis.” *Science* 349 (6248): 636–638. https://doi.org/10.1126/science.aaa9375

Efron, Bradley. 2004. “The Estimation of Prediction Error: Covariance Penalties and Cross-Validation.” *Journal of the American Statistical Association* 99 (467): 619–632. https://doi.org/10.1198/016214504000000692.

Gayo-Avello, Daniel. 2012. “‘I Wanted to Predict Elections with Twitter and all I got was this Lousy Paper’: A Balanced Survey on Election Prediction using Twitter Data.” https://arxiv.org/abs/1204.6441.

Kapoor, Sayash, Emily M. Cantrell, Kenny Peng, Thanh Hien Pham, Christopher A. Bail, Odd Erik Gundersen, Jake M. Hofman, Jessica Hullman, Michael A. Lones, **Momin M. Malik**, et al. 2024. “REFORMS: Consensus-Based Recommendations for Machine-Learning-Based Science.” *Science Advances* 10 (18): eadk3452. https://doi.org/10.1126/sciadv.adk3452.

Kiviat, Barbara. 2023. “The Moral Affordances of Construing People as Cases: How Algorithms and the Data They Depend on Obscure Narrative and Noncomparative Justice.” *Sociological Theory* 41 (3). https://doi.org/10.1177/07352751231186797.

Lachanski, Michael and Steven Pav. 2017. “Shy of the Character Limit: ‘Twitter Mood Predicts the Stock Market’ Revisited.” *Econ Journal Watch* 14 (3): 302–345. https://econjwatch.org/articles/shy-of-the-character-limit-twitter-mood-predicts-the-stock-market-revisited

**Malik, Momin M.** 2020. “A Hierarchy of Limitations in Machine Learning.” https://www.arxiv.org/abs/2002.05193.

Mullainathan, Sendhil and Jann Spiess. 2017. “Machine Learning: An Applied Econometric Approach.” *Journal of Economic Perspectives* 31 (2): 87–106. https://doi.org/10.1257/jep.31.2.87

Opsomer, Jean, Yuedong Wang, and Yuhong Yang. 2001. “Nonparametric Regression with Correlated Errors.” *Statistical Science* 16 (2): 134–153. https://doi.org/10.1214/ss/1009213287

Park, Greg. 2012. “The Dangers of Overfitting: A Kaggle Postmortem.” https://web.archive.org/web/20231204120801/http://gregpark.io/blog/Kaggle-Psychopathy-Postmortem/

Rescher, Nicholas. 1998. *Predicting the Future: An Introduction to the Theory of Forecasting*. State University of New York Press.

Rosset, Saharon, and Ryan J. Tibshirani. 2020. “From Fixed-X to Random-X Regression: Bias-Variance Decompositions, Covariance Penalties, and Prediction Error Estimation.” Journal of the American Statistical Association 115 (529): 138–151. https://doi.org/10.1080/01621459.2018.1424632.

Shmueli, Galit. 2010. “To Explain or to Predict?” *Statistical Science* 25 (3): 289–310. https://doi.org/10.1214/10-STS330.

Snijders, Tom A. B. , Philippa E. Pattison, Garry L. Robins, and Mark S. Handcock. 2006. “New Specifications for Exponential Random Graph Models.” Sociological Methodology 36 (1): 99–153. https://doi.org/10.1111/j.1467-9531.2006.00176.x

Wagstaff, Kiri L. 2012. “Machine Learning that Matters.” In *Proceedings of the 29th International Conference on International Conference on Machine Learning*, ICML’12, pages 1851–1856. https://icml.cc/2012/papers/298.pdf.

Wu, Shaohua, T. J. Harris, and K. B. McAuley. 2007. “The Use of Simplified or Misspecified Models: Linear Case.” *The Canadian Journal of Chemical Engineering* 85 (4): 386–398. https://doi.org/10.1002/cjce.5450850401.