Northeastern University

SANTA FE INSTITUTE

The Complex Systems View of AI Ethics

Tina Eliassi-Rad

RADLAB

Trustworthy Network Science

- Adversarial Robustness
- Sensitivity & Stability
- Explainability
- Fairness & Equity
- Reliability
- Topological interpretations
- Geometric interpretations

Just Machine Learning

- Data issues
- Model issues
- Task issues
- Community issues
- Democratic backsliding
- Impacts on high stakes situations
- Representation issues

RADLAB

Trustworthy Network Science

- Adversarial Robustness
- Sensitivity & Stability
- Explainability
- Fairness & Equity
- Reliability
- Topological interpretations
- Geometric interpretations

Just Machine Learning

- Data issues
- Model issues
- Task issues
- Community issues
- Democratic backsliding
- Impacts on high stakes situations
- Representation issues

Machine learning (ML) systems are not islands.

They are part of broader complex systems.

To understand and mitigate the risks and harms of using ML, we must remove our optimization blinders & study the broader complex systems in which ML systems operate.

Machine learning life-cycle

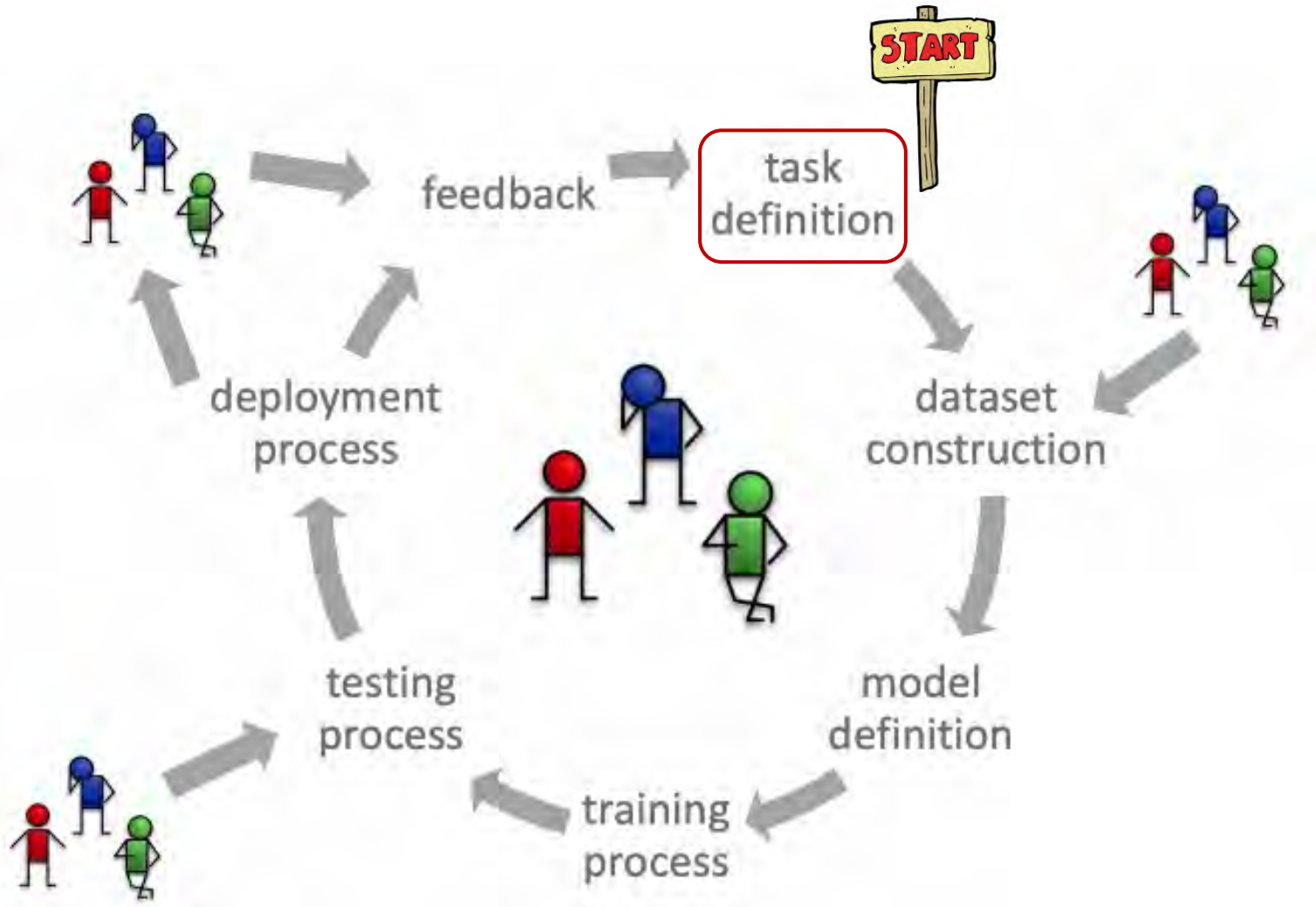


Image courtesy of [Jennifer Wortman Vaughan]

Age of Prediction

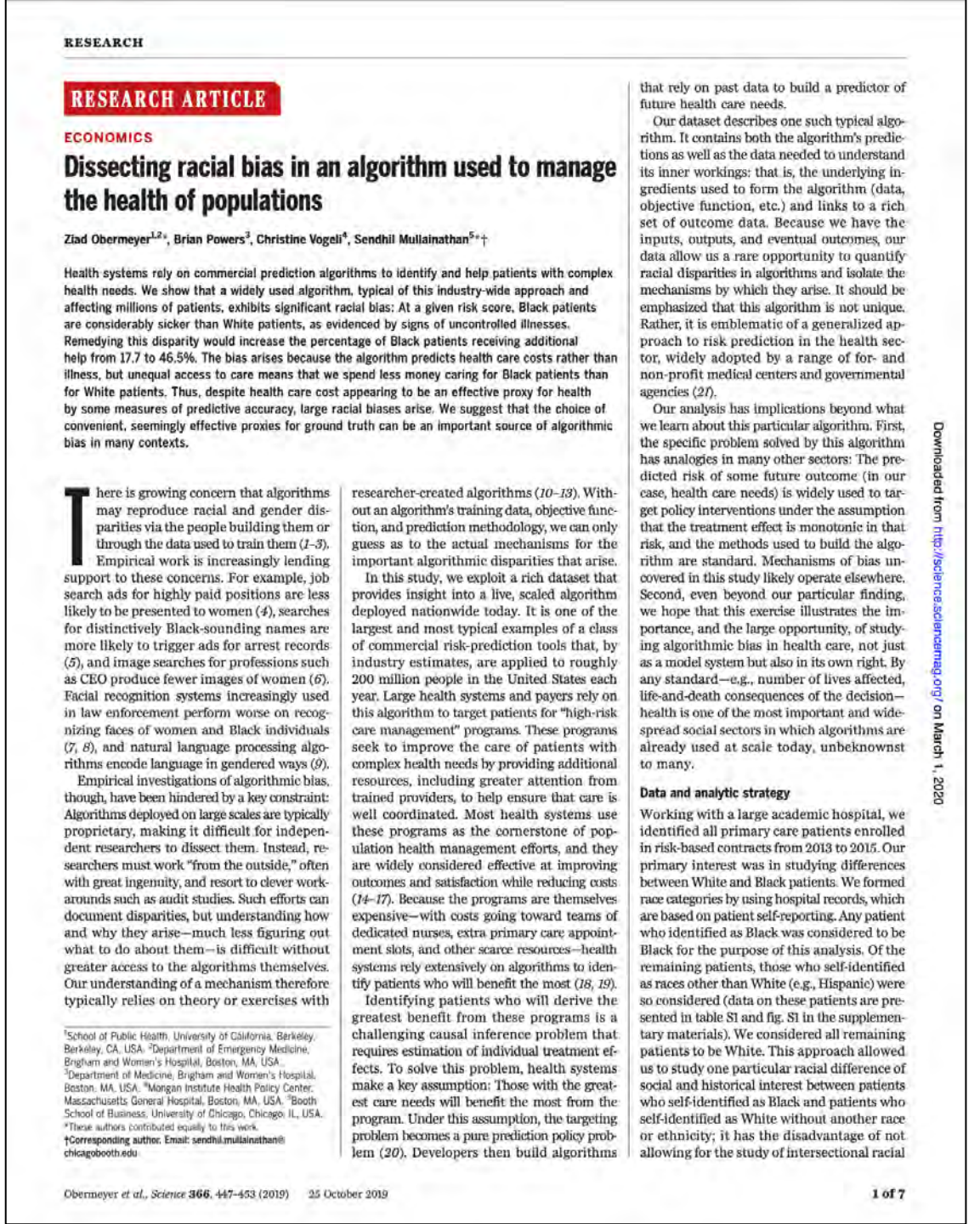
A prediction task

"At a given risk score, Black patients are considerably sicker than White patients, … The bias arises because the algorithm predicts health care costs rather than illness, but unequal access to care means that we spend less money caring for Black patients than for White patients."

what they were measuring

≠

what they thought they were measuring



RESEARCH ARTICLE

ECONOMICS

Dissecting racial bias in an algorithm used to manage the health of populations

Ziad Obermeyer[1,2]*, Brian Powers[3], Christine Vogeli[4], Sendhil Mullainathan[5]*†

Health systems rely on commercial prediction algorithms to identify and help patients with complex health needs. We show that a widely used algorithm, typical of this industry-wide approach and affecting millions of patients, exhibits significant racial bias: At a given risk score, Black patients are considerably sicker than White patients, as evidenced by signs of uncontrolled illnesses. Remedying this disparity would increase the percentage of Black patients receiving additional help from 17.7 to 46.5%. The bias arises because the algorithm predicts health care costs rather than illness, but unequal access to care means that we spend less money caring for Black patients than for White patients. Thus, despite health care cost appearing to be an effective proxy for health by some measures of predictive accuracy, large racial biases arise. We suggest that the choice of convenient, seemingly effective proxies for ground truth can be an important source of algorithmic bias in many contexts.

There is growing concern that algorithms may reproduce racial and gender disparities via the people building them or through the data used to train them (*1*–*3*). Empirical work is increasingly lending support to these concerns. For example, job search ads for highly paid positions are less likely to be presented to women (*4*), searches for distinctively Black-sounding names are more likely to trigger ads for arrest records (*5*), and image searches for professions such as CEO produce fewer images of women (*6*). Facial recognition systems increasingly used in law enforcement perform worse on recognizing faces of women and Black individuals (*7*, *8*), and natural language processing algorithms encode language in gendered ways (*9*).

Empirical investigations of algorithmic bias, though, have been hindered by a key constraint: Algorithms deployed on large scales are typically proprietary, making it difficult for independent researchers to dissect them. Instead, researchers must work "from the outside," often with great ingenuity, and resort to clever workarounds such as audit studies. Such efforts can document disparities, but understanding how and why they arise—much less figuring out what to do about them—is difficult without greater access to the algorithms themselves. Our understanding of a mechanism therefore typically relies on theory or exercises with researcher-created algorithms (*10*–*13*). Without an algorithm's training data, objective function, and prediction methodology, we can only guess as to the actual mechanisms for the important algorithmic disparities that arise.

In this study, we exploit a rich dataset that provides insight into a live, scaled algorithm deployed nationwide today. It is one of the largest and most typical examples of a class of commercial risk-prediction tools that, by industry estimates, are applied to roughly 200 million people in the United States each year. Large health systems and payers rely on this algorithm to target patients for "high-risk care management" programs. These programs seek to improve the care of patients with complex health needs by providing additional resources, including greater attention from trained providers, to help ensure that care is well coordinated. Most health systems use these programs as the cornerstone of population health management efforts, and they are widely considered effective at improving outcomes and satisfaction while reducing costs (*14*–*17*). Because the programs are themselves expensive—with costs going toward teams of dedicated nurses, extra primary care appointment slots, and other scarce resources—health systems rely extensively on algorithms to identify patients who will benefit the most (*18*, *19*).

Identifying patients who will derive the greatest benefit from these programs is a challenging causal inference problem that requires estimation of individual treatment effects. To solve this problem, health systems make a key assumption: Those with the greatest care needs will benefit the most from the program. Under this assumption, the targeting problem becomes a pure prediction policy problem (*20*). Developers then build algorithms that rely on past data to build a predictor of future health care needs.

Our dataset describes one such typical algorithm. It contains both the algorithm's predictions as well as the data needed to understand its inner workings: that is, the underlying ingredients used to form the algorithm (data, objective function, etc.) and links to a rich set of outcome data. Because we have the inputs, outputs, and eventual outcomes, our data allow us a rare opportunity to quantify racial disparities in algorithms and isolate the mechanisms by which they arise. It should be emphasized that this algorithm is not unique. Rather, it is emblematic of a generalized approach to risk prediction in the health sector, widely adopted by a range of for- and non-profit medical centers and governmental agencies (*21*).

Our analysis has implications beyond what we learn about this particular algorithm. First, the specific problem solved by this algorithm has analogies in many other sectors: The predicted risk of some future outcome (in our case, health care needs) is widely used to target policy interventions under the assumption that the treatment effect is monotonic in that risk, and the methods used to build the algorithm are standard. Mechanisms of bias uncovered in this study likely operate elsewhere. Second, even beyond our particular finding, we hope that this exercise illustrates the importance, and the large opportunity, of studying algorithmic bias in health care, not just as a model system but also in its own right. By any standard—e.g., number of lives affected, life-and-death consequences of the decision—health is one of the most important and widespread social sectors in which algorithms are already used at scale today, unbeknownst to many.

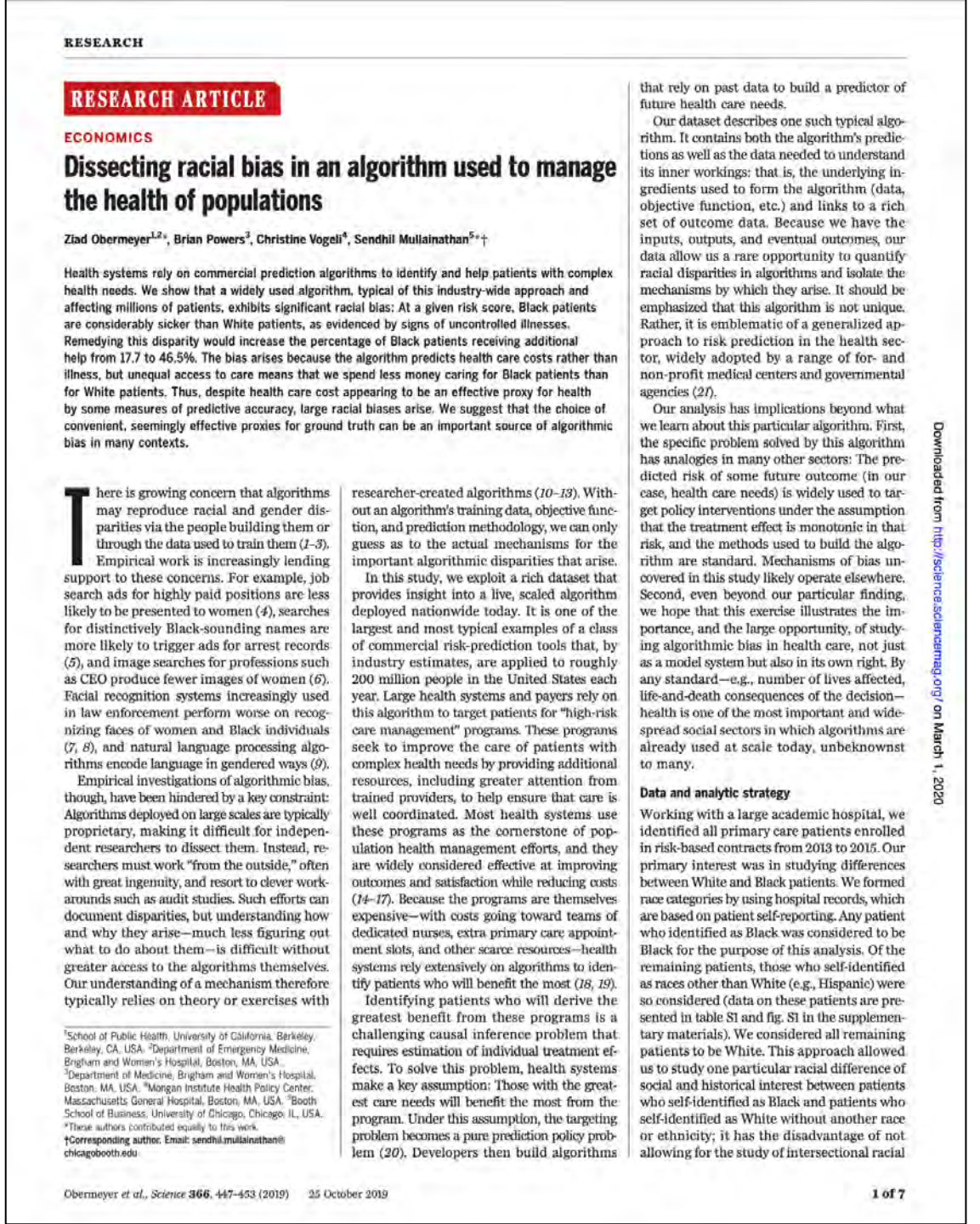
Data and analytic strategy

Working with a large academic hospital, we identified all primary care patients enrolled in risk-based contracts from 2013 to 2015. Our primary interest was in studying differences between White and Black patients. We formed race categories by using hospital records, which are based on patient self-reporting. Any patient who identified as Black was considered to be Black for the purpose of this analysis. Of the remaining patients, those who self-identified as races other than White (e.g., Hispanic) were so considered (data on these patients are presented in table S1 and fig. S1 in the supplementary materials). We considered all remaining patients to be White. This approach allowed us to study one particular racial difference of social and historical interest between patients who self-identified as Black and patients who self-identified as White without another race or ethnicity; it has the disadvantage of not allowing for the study of intersectional racial

[1]School of Public Health, University of California, Berkeley, Berkeley, CA, USA. [2]Department of Emergency Medicine, Brigham and Women's Hospital, Boston, MA, USA. [3]Department of Medicine, Brigham and Women's Hospital, Boston, MA, USA. [4]Morgan Institute Health Policy Center, Massachusetts General Hospital, Boston, MA, USA. [5]Booth School of Business, University of Chicago, Chicago, IL, USA.
*These authors contributed equally to this work.
†Corresponding author. Email: sendhil.mullainathan@chicagobooth.edu

The prediction task

"At a given risk score, Black patients are considerably sicker than White patients, … The bias arises because the algorithm predicts health care costs rather than illness, but unequal access to care means that we spend less money caring for Black patients than for White patients."



RESEARCH ARTICLE

ECONOMICS

Dissecting racial bias in an algorithm used to manage the health of populations

Ziad Obermeyer[1,2]*, Brian Powers[3], Christine Vogeli[4], Sendhil Mullainathan[5]*†

Health systems rely on commercial prediction algorithms to identify and help patients with complex health needs. We show that a widely used algorithm, typical of this industry-wide approach and affecting millions of patients, exhibits significant racial bias: At a given risk score, Black patients are considerably sicker than White patients, as evidenced by signs of uncontrolled illnesses. Remedying this disparity would increase the percentage of Black patients receiving additional help from 17.7 to 46.5%. The bias arises because the algorithm predicts health care costs rather than illness, but unequal access to care means that we spend less money caring for Black patients than for White patients. Thus, despite health care cost appearing to be an effective proxy for health by some measures of predictive accuracy, large racial biases arise. We suggest that the choice of convenient, seemingly effective proxies for ground truth can be an important source of algorithmic bias in many contexts.

There is growing concern that algorithms may reproduce racial and gender disparities via the people building them or through the data used to train them (*1*–*3*). Empirical work is increasingly lending support to these concerns. For example, job search ads for highly paid positions are less likely to be presented to women (*4*), searches for distinctively Black-sounding names are more likely to trigger ads for arrest records (*5*), and image searches for professions such as CEO produce fewer images of women (*6*). Facial recognition systems increasingly used in law enforcement perform worse on recognizing faces of women and Black individuals (*7*, *8*), and natural language processing algorithms encode language in gendered ways (*9*).

Empirical investigations of algorithmic bias, though, have been hindered by a key constraint: Algorithms deployed on large scales are typically proprietary, making it difficult for independent researchers to dissect them. Instead, researchers must work "from the outside," often with great ingenuity, and resort to clever workarounds such as audit studies. Such efforts can document disparities, but understanding how and why they arise—much less figuring out what to do about them—is difficult without greater access to the algorithms themselves. Our understanding of a mechanism therefore typically relies on theory or exercises with researcher-created algorithms (*10*–*13*). Without an algorithm's training data, objective function, and prediction methodology, we can only guess as to the actual mechanisms for the important algorithmic disparities that arise.

In this study, we exploit a rich dataset that provides insight into a live, scaled algorithm deployed nationwide today. It is one of the largest and most typical examples of a class of commercial risk-prediction tools that, by industry estimates, are applied to roughly 200 million people in the United States each year. Large health systems and payers rely on this algorithm to target patients for "high-risk care management" programs. These programs seek to improve the care of patients with complex health needs by providing additional resources, including greater attention from trained providers, to help ensure that care is well coordinated. Most health systems use these programs as the cornerstone of population health management efforts, and they are widely considered effective at improving outcomes and satisfaction while reducing costs (*14*–*17*). Because the programs are themselves expensive—with costs going toward teams of dedicated nurses, extra primary care appointment slots, and other scarce resources—health systems rely extensively on algorithms to identify patients who will benefit the most (*18*, *19*).

Identifying patients who will derive the greatest benefit from these programs is a challenging causal inference problem that requires estimation of individual treatment effects. To solve this problem, health systems make a key assumption: Those with the greatest care needs will benefit the most from the program. Under this assumption, the targeting problem becomes a pure prediction policy problem (*20*). Developers then build algorithms that rely on past data to build a predictor of future health care needs.

Our dataset describes one such typical algorithm. It contains both the algorithm's predictions as well as the data needed to understand its inner workings: that is, the underlying ingredients used to form the algorithm (data, objective function, etc.) and links to a rich set of outcome data. Because we have the inputs, outputs, and eventual outcomes, our data allow us a rare opportunity to quantify racial disparities in algorithms and isolate the mechanisms by which they arise. It should be emphasized that this algorithm is not unique. Rather, it is emblematic of a generalized approach to risk prediction in the health sector, widely adopted by a range of for- and non-profit medical centers and governmental agencies (*21*).

Our analysis has implications beyond what we learn about this particular algorithm. First, the specific problem solved by this algorithm has analogies in many other sectors: The predicted risk of some future outcome (in our case, health care needs) is widely used to target policy interventions under the assumption that the treatment effect is monotonic in that risk, and the methods used to build the algorithm are standard. Mechanisms of bias uncovered in this study likely operate elsewhere. Second, even beyond our particular finding, we hope that this exercise illustrates the importance, and the large opportunity, of studying algorithmic bias in health care, not just as a model system but also in its own right. By any standard—e.g., number of lives affected, life-and-death consequences of the decision—health is one of the most important and widespread social sectors in which algorithms are already used at scale today, unbeknownst to many.

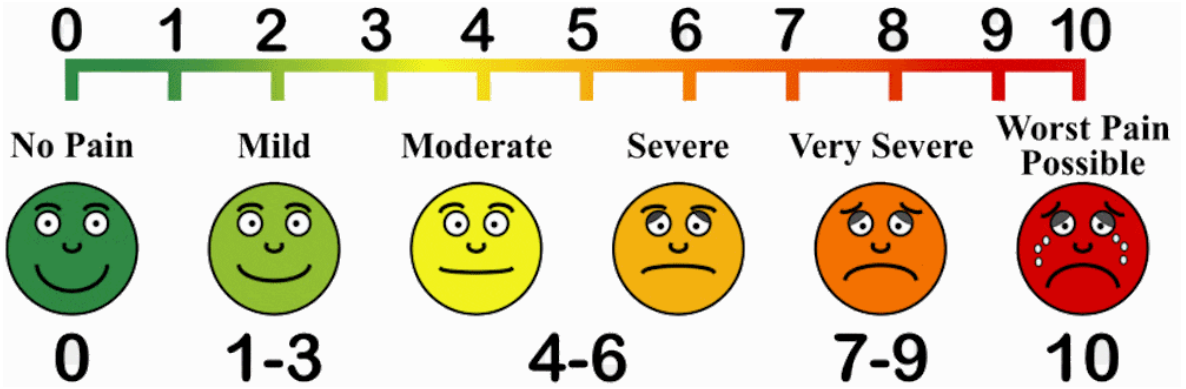
Data and analytic strategy

Working with a large academic hospital, we identified all primary care patients enrolled in risk-based contracts from 2013 to 2015. Our primary interest was in studying differences between White and Black patients. We formed race categories by using hospital records, which are based on patient self-reporting. Any patient who identified as Black was considered to be Black for the purpose of this analysis. Of the remaining patients, those who self-identified as races other than White (e.g., Hispanic) were so considered (data on these patients are presented in table S1 and fig. S1 in the supplementary materials). We considered all remaining patients to be White. This approach allowed us to study one particular racial difference of social and historical interest between patients who self-identified as Black and patients who self-identified as White without another race or ethnicity; it has the disadvantage of not allowing for the study of intersectional racial

[1]School of Public Health, University of California, Berkeley, Berkeley, CA, USA. [2]Department of Emergency Medicine, Brigham and Women's Hospital, Boston, MA, USA. [3]Department of Medicine, Brigham and Women's Hospital, Boston, MA, USA. [4]Morgan Institute Health Policy Center, Massachusetts General Hospital, Boston, MA, USA. [5]Booth School of Business, University of Chicago, Chicago, IL, USA.
*These authors contributed equally to this work.
†Corresponding author. Email: sendhil.mullainathan@chicagobooth.edu

Risk assessment: A popular prediction task

- Why is it popular?
    1. Because machine learning people know a lot about it – i.e., it is a "low-hanging fruit".
    2. Because human decision-makers like the output of these tasks; they are easily understandable.



Issues with risk assessment


MISSION IMPOSSIBLE

Parity (a.k.a. fairness) measures come from the confusion (a.k.a. error) matrix

		True condition	
	Total population	Condition positive	Condition negative
Predicted condition	Predicted condition positive	**True positive**	**False positive**, Type I error
	Predicted condition negative	**False negative**, Type II error	**True negative**

Image from https://en.wikipedia.org/wiki/Confusion_matrix

Parity (a.k.a. fairness) measures come from the confusion (a.k.a. error) matrix

		True condition				
	Total population	Condition positive	Condition negative	Prevalence $= \frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) $= \frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	
Predicted condition	Predicted condition positive	**True positive**	**False positive**, Type I error	Positive predictive value (PPV), Precision $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$	
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$	
		True positive rate (TPR), Recall, Sensitivity, probability of detection, Power $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR−}}$	F_1 score $= 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$
		False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) $= \frac{\text{FNR}}{\text{TNR}}$		

Image from https://en.wikipedia.org/wiki/Confusion_matrix

Impossibility result for risk assessment

Condition	Chouldechova (2016)
Mutual exclusivity between groups a and b	✓
Unequal base rates: $Pr_a(Y = 1) \neq Pr_b(Y = 1)$	✓

Impossibility result for risk assessment

Condition	Chouldechova (2016)
Mutual exclusivity between groups a and b	✓
Unequal base rates: $Pr_a(Y = 1) \neq Pr_b(Y = 1)$	✓
Imperfect classifier: $Pr_a(C = 1 \mid Y = 0) \neq 0$ and $Pr_b(C = 1 \mid Y = 0) \neq 0$ and $Pr_a(C = 1 \mid Y = 1) \neq 1$ and $Pr_b(C = 1 \mid Y = 1) \neq 1$	
Non-zero precision: $Pr_a(Y = 1 \mid C = 1) \neq 0$ or $Pr_b(Y = 1 \mid C = 1) \neq 0$	
Statistical parity: $Pr_a(C = 1) = Pr_b(C = 1)$	✓

Impossibility result for risk assessment

Condition	Chouldechova (2016)
Mutual exclusivity between groups a and b	✓
Unequal base rates: $Pr_a(Y = 1) \neq Pr_b(Y = 1)$	✓
Imperfect classifier: $Pr_a(C = 1 \mid Y = 0) \neq 0$ and $Pr_b(C = 1 \mid Y = 0) \neq 0$ and $Pr_a(C = 1 \mid Y = 1) \neq 1$ and $Pr_b(C = 1 \mid Y = 1) \neq 1$	
Non-zero precision: $Pr_a(Y = 1 \mid C = 1) \neq 0$ or $Pr_b(Y = 1 \mid C = 1) \neq 0$	
Statistical parity: $Pr_a(C = 1) = Pr_b(C = 1)$	✓
Predictive parity: $Pr_a(Y = 1 \mid C = 1) = Pr_b(Y = 1 \mid C = 1)$	✗
True positive parity: $Pr_a(C = 1 \mid Y = 1) = Pr_b(C = 1 \mid Y = 1)$	✗
False positive parity: $Pr_a(C = 1 \mid Y = 0) = Pr_b(C = 1 \mid Y = 0)$	✗

Impossibility result for risk assessment

Condition	Chouldechova (2016)	Kleinberg et al. (2016)
Mutual exclusivity between groups a and b	✓	✓
Unequal base rates: $Pr_a(Y = 1) \neq Pr_b(Y = 1)$	✓	✓
Imperfect classifier: $Pr_a(C = 1 \mid Y = 0) \neq 0$ and $Pr_b(C = 1 \mid Y = 0) \neq 0$ and $Pr_a(C = 1 \mid Y = 1) \neq 1$ and $Pr_b(C = 1 \mid Y = 1) \neq 1$		
Non-zero precision: $Pr_a(Y = 1 \mid C = 1) \neq 0$ or $Pr_b(Y = 1 \mid C = 1) \neq 0$		
Statistical parity: $Pr_a(C = 1) = Pr_b(C = 1)$	✓	
Predictive parity: $Pr_a(Y = 1 \mid C = 1) = Pr_b(Y = 1 \mid C = 1)$	✗	
True positive parity: $Pr_a(C = 1 \mid Y = 1) = Pr_b(C = 1 \mid Y = 1)$	✗	
False positive parity: $Pr_a(C = 1 \mid Y = 0) = Pr_b(C = 1 \mid Y = 0)$	✗	

Impossibility result for risk assessment

Condition	Chouldechova (2016)	Kleinberg et al. (2016)
Mutual exclusivity between groups a and b	✓	✓
Unequal base rates: $Pr_a(Y = 1) \neq Pr_b(Y = 1)$	✓	✓
Imperfect classifier: $Pr_a(C = 1 \mid Y = 0) \neq 0$ and $Pr_b(C = 1 \mid Y = 0) \neq 0$ and $Pr_a(C = 1 \mid Y = 1) \neq 1$ and $Pr_b(C = 1 \mid Y = 1) \neq 1$		✓
Non-zero precision: $Pr_a(Y = 1 \mid C = 1) \neq 0$ or $Pr_b(Y = 1 \mid C = 1) \neq 0$		✓
Statistical parity: $Pr_a(C = 1) = Pr_b(C = 1)$	✓	
Predictive parity: $Pr_a(Y = 1 \mid C = 1) = Pr_b(Y = 1 \mid C = 1)$	✗	
True positive parity: $Pr_a(C = 1 \mid Y = 1) = Pr_b(C = 1 \mid Y = 1)$	✗	
False positive parity: $Pr_a(C = 1 \mid Y = 0) = Pr_b(C = 1 \mid Y = 0)$	✗	

Impossibility result for risk assessment

Condition	Chouldechova (2016)	Kleinberg et al. (2016)
Mutual exclusivity between groups a and b	✓	✓
Unequal base rates: $Pr_a(Y = 1) \neq Pr_b(Y = 1)$	✓	✓
Imperfect classifier: $Pr_a(C = 1 \mid Y = 0) \neq 0$ and $Pr_b(C = 1 \mid Y = 0) \neq 0$ and $Pr_a(C = 1 \mid Y = 1) \neq 1$ and $Pr_b(C = 1 \mid Y = 1) \neq 1$		✓
Non-zero precision: $Pr_a(Y = 1 \mid C = 1) \neq 0$ or $Pr_b(Y = 1 \mid C = 1) \neq 0$		✓
Statistical parity: $Pr_a(C = 1) = Pr_b(C = 1)$	✓	
Predictive parity: $Pr_a(Y = 1 \mid C = 1) = Pr_b(Y = 1 \mid C = 1)$	✗	✗
True positive parity: $Pr_a(C = 1 \mid Y = 1) = Pr_b(C = 1 \mid Y = 1)$	✗	✗
False positive parity: $Pr_a(C = 1 \mid Y = 0) = Pr_b(C = 1 \mid Y = 0)$	✗	✗

Impossibility result for risk assessment

Condition	Chouldechova (2016)	Kleinberg et al. (2016)	Eliassi-Rad & Fitelson (2021)
Mutual exclusivity between groups a and b	✓	✓	
Unequal base rates: $Pr_a(Y = 1) \neq Pr_b(Y = 1)$	✓	✓	✓
Imperfect classifier: $Pr_a(C = 1 \mid Y = 0) \neq 0$ and $Pr_b(C = 1 \mid Y = 0) \neq 0$ and $Pr_a(C = 1 \mid Y = 1) \neq 1$ and $Pr_b(C = 1 \mid Y = 1) \neq 1$		✓	
Non-zero precision: $Pr_a(Y = 1 \mid C = 1) \neq 0$ or $Pr_b(Y = 1 \mid C = 1) \neq 0$		✓	
Statistical parity: $Pr_a(C = 1) = Pr_b(C = 1)$	✓		
Predictive parity: $Pr_a(Y = 1 \mid C = 1) = Pr_b(Y = 1 \mid C = 1)$	✗	✗	
True positive parity: $Pr_a(C = 1 \mid Y = 1) = Pr_b(C = 1 \mid Y = 1)$	✗	✗	
False positive parity: $Pr_a(C = 1 \mid Y = 0) = Pr_b(C = 1 \mid Y = 0)$	✗	✗	

Impossibility result for risk assessment

Condition	Chouldechova (2016)	Kleinberg et al. (2016)	Eliassi-Rad & Fitelson (2021)
Mutual exclusivity between groups a and b	✓	✓	
Unequal base rates: $Pr_a(Y = 1) \neq Pr_b(Y = 1)$	✓	✓	✓
Imperfect classifier: $Pr_a(C = 1 \mid Y = 0) \neq 0$ and $Pr_b(C = 1 \mid Y = 0) \neq 0$ and $Pr_a(C = 1 \mid Y = 1) \neq 1$ and $Pr_b(C = 1 \mid Y = 1) \neq 1$		✓	✓
Non-zero precision: $Pr_a(Y = 1 \mid C = 1) \neq 0$ or $Pr_b(Y = 1 \mid C = 1) \neq 0$		✓	✓
Statistical parity: $Pr_a(C = 1) = Pr_b(C = 1)$	✓		
Predictive parity: $Pr_a(Y = 1 \mid C = 1) = Pr_b(Y = 1 \mid C = 1)$	✗	✗	
True positive parity: $Pr_a(C = 1 \mid Y = 1) = Pr_b(C = 1 \mid Y = 1)$	✗	✗	
False positive parity: $Pr_a(C = 1 \mid Y = 0) = Pr_b(C = 1 \mid Y = 0)$	✗	✗	

Impossibility result for risk assessment

Condition	Chouldechova (2016)	Kleinberg et al. (2016)	Eliassi-Rad & Fitelson (2021)
Mutual exclusivity between groups a and b	✓	✓	
Unequal base rates: $Pr_a(Y = 1) \neq Pr_b(Y = 1)$	✓	✓	✓
Imperfect classifier: $Pr_a(C = 1 \mid Y = 0) \neq 0$ and $Pr_b(C = 1 \mid Y = 0) \neq 0$ and $Pr_a(C = 1 \mid Y = 1) \neq 1$ and $Pr_b(C = 1 \mid Y = 1) \neq 1$		✓	✓
Non-zero precision: $Pr_a(Y = 1 \mid C = 1) \neq 0$ or $Pr_b(Y = 1 \mid C = 1) \neq 0$		✓	✓
Statistical parity: $Pr_a(C = 1) = Pr_b(C = 1)$	✓		
Predictive parity: $Pr_a(Y = 1 \mid C = 1) = Pr_b(Y = 1 \mid C = 1)$	✗	✗	✗
True positive parity: $Pr_a(C = 1 \mid Y = 1) = Pr_b(C = 1 \mid Y = 1)$	✗	✗	✗
False positive parity: $Pr_a(C = 1 \mid Y = 0) = Pr_b(C = 1 \mid Y = 0)$	✗	✗	✗

Impossibility result for risk assessment

Condition	Kleinberg et al. (2016)	Eliassi-Rad & Fitelson (2021)	Eliassi-Rad & Fitelson (2021)
Mutual exclusivity between groups a and b	✓		
Unequal base rates: $Pr_a(Y = 1) \neq Pr_b(Y = 1)$	✓	✓	✓
Imperfect classifier: $Pr_a(C = 1 \mid Y = 0) \neq 0$ and $Pr_b(C = 1 \mid Y = 0) \neq 0$ and $Pr_a(C = 1 \mid Y = 1) \neq 1$ and $Pr_b(C = 1 \mid Y = 1) \neq 1$	✓	✓	
Non-zero precision: $Pr_a(Y = 1 \mid C = 1) \neq 0$ or $Pr_b(Y = 1 \mid C = 1) \neq 0$	✓	✓	
Regular priors: Prior Pr(•) only assigns extremal probability to non-contingent propositions			✓
Statistical parity: $Pr_a(C = 1) = Pr_b(C = 1)$			
Predictive parity: $Pr_a(Y = 1 \mid C = 1) = Pr_b(Y = 1 \mid C = 1)$	×	×	×
True positive parity: $Pr_a(C = 1 \mid Y = 1) = Pr_b(C = 1 \mid Y = 1)$	×	×	×
False positive parity: $Pr_a(C = 1 \mid Y = 0) = Pr_b(C = 1 \mid Y = 0)$	×	×	×

There are more impossibility results for risk assessment and group fairness

Branden Fitelson

- To read about more impossibilities, download http://fitelson.org/exploring_impossibility.pdf
- To find new ones, download the Mathematica notebook at http://fitelson.org/exploring_impossibility.nb

Fallout from the impossibility theorems

Predictive parity: $Pr_a(Y = 1 \mid C = 1) = Pr_b(Y = 1 \mid C = 1)$
True positive parity: $Pr_a(C = 1 \mid Y = 1) = Pr_b(C = 1 \mid Y = 1)$
False positive parity: $Pr_a(C = 1 \mid Y = 0) = Pr_b(C = 1 \mid Y = 0)$

- Get rid of one of the parities
- Put bounds on the parities
- Deborah Hellman (University of Virginia Law School)
  - Predictive parity captures "what you ought to believe"
  - True positive and false positive parities capture "what you ought to do"
  - The algorithm ought not be thinking about the right-making properties when deliberating in many cases

    ➔ If you are going to drop a parity, drop predictive parity

https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3418528

What about individual fairness?

Will Fleisher

- Dwork et al. [2012]: *Similar individuals should be treated similarly*.
- Will Fleisher. What's fair about individual fairness? *AIES* 2021: 480-490.
- Four problems for individual fairness as a definition and as a method for ensuring fairness:
    1. Insufficiency of similar treatment
    2. Systematic bias and arbiters (implicit human biases)
    3. Prior moral judgments about task relevance and moral values
    4. Incommensurability
- Takeaways:
    - Individual fairness is inadequate as a definition of fairness.
    - Individual fairness should not be used as a sole means for determining fairness (or detecting bias).

Machine learning life-cycle

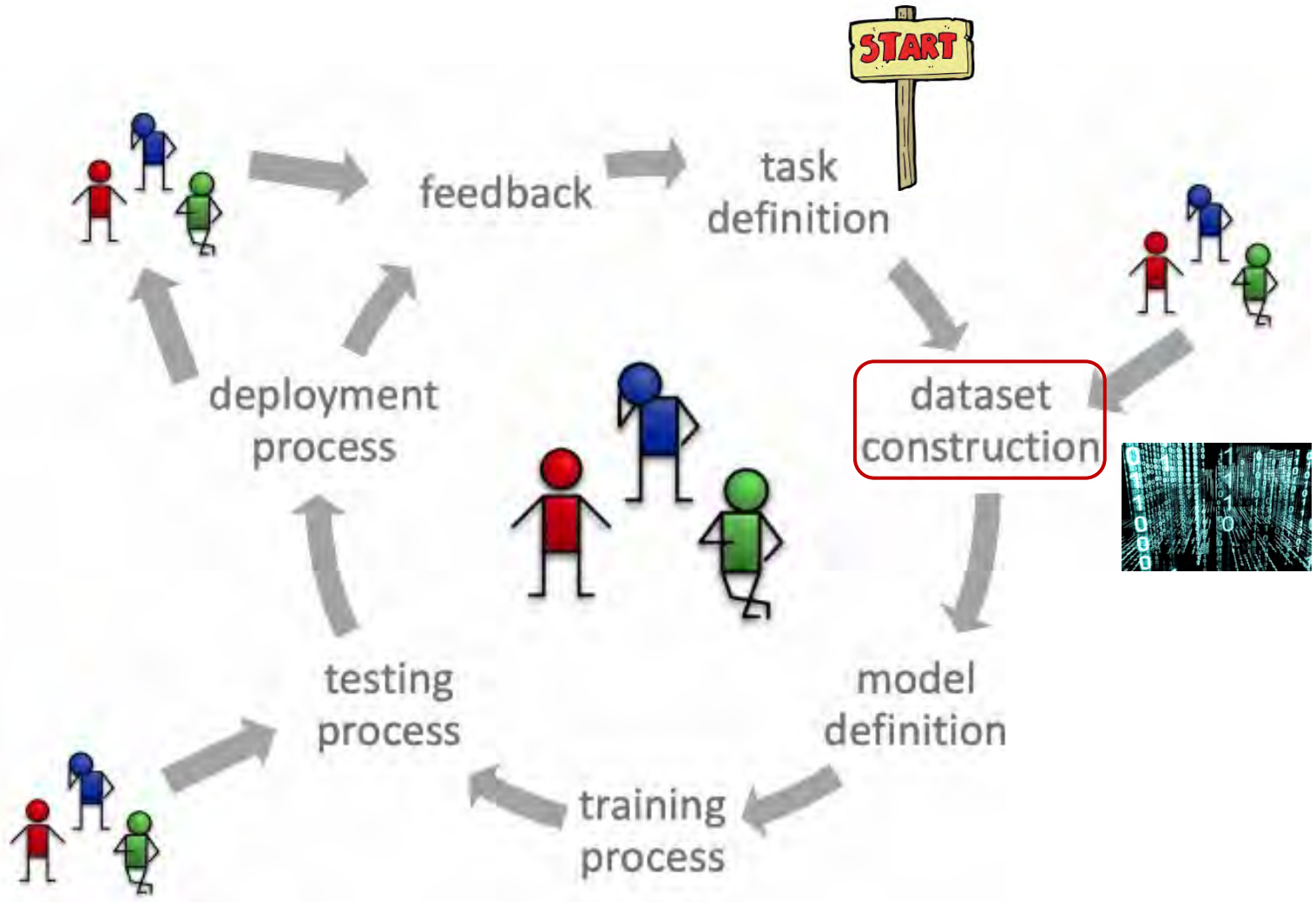


Image courtesy of [Jennifer Wortman Vaughan]

Oximeter Data

CORONAVIRUS GUIDE

Devices Used In COVID-19 Treatment Can Give Errors For Patients With Dark Skin

December 16, 2020 · 5:01 PM ET

RICHARD HARRIS

A paramedic uses a pulse oximeter to check a patient's vital signs during an August home visit in the Bronx borough of New York.

Angus Mordant/Bloomberg via Getty Images

https://n.pr/3vRUjbf

https://www.nejm.org/doi/full/10.1056/NEJMc2029240
(Dec 2020)

The NEW ENGLAND JOURNAL of MEDICINE

CORRESPONDENCE

Racial Bias in Pulse Oximetry Measurement

TO THE EDITOR: Oxygen is among the most frequently administered medical therapies, with a level that is commonly adjusted according to the reading on a pulse oximeter that measures patients' oxygen saturation. Questions about pulse oximeter technology have been raised, given its original development in populations that were not racially diverse.[1,2] The clinical significance of potential racial bias in pulse oximetry measurement is unknown.

Our study involved adult inpatients who were receiving supplemental oxygen at the University of Michigan Hospital (from January through July 2020) and patients in intensive care units at 178 hospitals (from 2014 through 2015).[3] We analyzed paired pulse oximetry measures of oxygen saturation and measures of arterial oxygen saturation in arterial blood gas, with all evaluations performed within 10 minutes of each other. To ensure that the arterial oxygen saturation was directly measured by co-oximetry, we limited analyses to measures of arterial blood gas that included carboxyhemoglobin and methemoglobin saturations.

We tested for occult hypoxemia (i.e., an arterial oxygen saturation of <88% despite an oxygen saturation of 92 to 96% on pulse oximetry) among patients who identified their race as Black or White. Since a low level of peripheral perfusion could lower the accuracy of oxygen saturation values,[4] we also estimated the percentage of patients with occult hypoxemia after adjusting for age, sex, and cardiovascular score on the Sequential Organ Failure Assessment (SOFA) in the University of Michigan cohort. Additional details regarding the methods that were used in the study are provided in the Supplementary Appendix, available with the full text of this letter at NEJM.org.

We analyzed 10,789 pairs of measures of oxygen saturation by pulse oximetry and arterial oxygen saturation in arterial blood gas obtained from 1333 White patients and 276 Black patients in the University of Michigan cohort and 37,308 pairs obtained from 7342 White patients and 1050 Black patients in the multicenter cohort. In the University of Michigan cohort, among the patients who had an oxygen saturation of 92 to 96% on pulse oximetry, an arterial oxygen saturation of less than 88% was found in 88 of 749 arterial blood gas measurements in Black patients (11.7%; 95% confidence interval [CI], 8.5 to 16.0) and in 99 of 2778 measurements in White patients (3.6%; 95% CI, 2.7 to 4.7) (Fig. 1). The findings in the adjusted analyses were similar to those in the unadjusted analyses, with an arterial blood gas oxygen saturation of less than 88% in 11.4% (95% CI, 7.6 to 15.2) of the measurements in Black patients and in 3.6% (95% CI, 2.5 to 4.6) of those in White patients. Results were

https://emcrit.org/pulmcrit/racism-oximetry/ (Dec 2020)

PulmCrit – Dismantling the systemic racism of pulse oximetry

December 21, 2020 by **Josh Farkas** — **9 Comments**

Oximetry is fundamental to critical care. Consequently, even small biases in pulse oximetry measurements could have real clinical impact (especially when leveraged across innumerable measurements among thousands of patients).

Racial bias in pulse oximetry was the subject of two studies in 2005 and 2007. The topic was then largely ignored over the past 13 years. I was vaguely aware of this issue, but given a lack of modern research on it I had assumed that it was a technical glitch restricted to older pulse oximeters. A fresh publication in the New England Journal of Medicine shows that I was wrong. Let's start by reviewing the evidence.

https://doi.org/10.1097/00000542-200504000-00004 (April 2005)

Anesthesiology 2005; 102:715–9

Effects of Skin Pigmentation on Pulse Oximeter Accuracy at Low Saturation

Philip E. Bickler, M.D., Ph.D., John R. Feiner, M.D.,† John W. Severinghaus, M.D.‡*

Background: **It is uncertain whether skin pigmentation affects pulse oximeter accuracy at low HbO_2 saturation.**

Methods: **The accuracy of finger pulse oximeters during stable, plateau levels of arterial oxygen saturation (Sao_2) between 60 and 100% were evaluated in 11 subjects with darkly pigmented skin and in 10 with light skin pigmentation. Oximeters tested were the Nellcor N-595 with the OxiMax-A probe (Nellcor Inc., Pleasanton, CA), the Novametrix 513 (Novametrix Inc., Wallingford, CT), and the Nonin Onyx (Nonin Inc., Plymouth, MN). Semisupine subjects breathed air–nitrogen–carbon dioxide mixtures through a mouthpiece. A computer used end-tidal oxygen and carbon dioxide concentrations determined by mass spectrometry to estimate breath-by-breath Sao_2, from which an operator adjusted inspired gas to rapidly achieve 2- to 3-min stable plateaus of desaturation. Comparisons of oxygen saturation measured by pulse oximetry (Spo_2) with Sao_2 (by Radiometer OSM3) were used in a multivariate model to determine the interrelation between saturation, skin pigmentation, and oximeter bias ($Spo_2 - Sao_2$).**

Results: **At 60–70% Sao_2, Spo_2 (mean of three oximeters) overestimated Sao_2 (bias ± SD) by 3.56 ± 2.45% (n = 29) in darkly pigmented subjects, compared with 0.37 ± 3.20% (n = 58) in lightly pigmented subjects ($P < 0.0001$). The SD of bias was not greater with dark than light skin. The dark–light skin differences at 60–70% Sao_2 were 2.35% (Nonin), 3.38% (Novametrix), and 4.30% (Nellcor). Skin pigment–related differences were significant with Nonin below 70% Sao_2, with Novametrix below 90%, and with Nellcor at all ranges. Pigment-related bias increased approximately in proportion to desaturation.**

Conclusions: **The three tested pulse oximeters overestimated arterial oxygen saturation during hypoxia in dark-skinned individuals.**

PULSE oximetry theoretically can compute arterial hemoglobin oxygen saturation from the ratio of the pulsatile to the total transmitted red light divided by the same ratio for infrared light transilluminating a finger, ear, or other tissue. The derived saturation should be independent of skin pigmentation, and many other variables, such as hemoglobin concentration, nail polish, dirt, and jaundice. Several large controlled studies comparing black and white patients (380 subjects)[1,2] reported no significant pigment-related errors in pulse oximeters at normal saturation.

However, Severinghaus and Kelleher[3] reviewed data from several investigators who had reported anecdotal errors (+3 to +5%) in black patients.[4–7] Model simulations of errors due to various pigments were reviewed by Ralston *et al.*[8] Cote *et al.*[9] reported that nail polish and ink on skin surface can cause errors, a finding confirmed anecdotally by others from fingerprinting ink,[10] henna,[11] and meconium.[12] Intravenously injected dyes cause transient errors.[13] Lee *et al.*[14] found overestimation of saturation, especially at low saturation in pigmented patients (Indian, Malay *vs.* Chinese). The Technology Subcommittee of the Working Group on Critical Care, Ontario Ministry of Health,[15] reported unacceptable errors in pulse oximetry at low saturation in pigmented subjects. Zeballos and Weisman[16] compared the accuracy of the Hewlett-Packard (Sunnyvale, CA) ear oximeter and the Biox II pulse oximeter (Ohmeda, Andover, MA) in 33 young black men exercising at three different simulated altitudes. At an altitude of 4,000 m, where arterial oxygen saturation (Sao_2) ranged from 75 to 84%, the Hewlett-Packard underestimated Sao_2 by 4.8 ± 1.6%, whereas the Biox overestimated Sao_2 by 9.8 ± 1.8% (n = 22). It was stated that these errors, previously reported in whites, were both exaggerated in blacks.

During our many years of testing pulse oximeter accuracy at oxygen saturations as low as 50%, we have occasionally noted unusually high positive bias, particularly at very low saturation levels, in some but not in other deeply pigmented subjects. This investigation was therefore specifically designed to determine whether errors at low Sao_2 correlate with skin color.

All pulse oximeters marketed in the United States are required by the US Food and Drug Administration to have been tested and to be certified as accurate to less than ±3% root mean square error at Sao_2 values between 70 and 100%. The great majority of calibration and confirmation tests have been conducted in volunteer subjects with light skin pigmentation.

The Food and Drug Administration has recently suggested that studies of pulse oximeter accuracy submitted for Food and Drug Administration device approval include subjects with a range of skin pigmentations, although no quantitative requirement has been distributed. We are aware of no data that support this action.

If there is a significant and reproducible positive bias at low saturation in dark-skinned subjects, inclusion of

* Professor, † Associate Professor, ‡ Professor Emeritus.

Received from the Department of Anesthesia and Perioperative Care, University of California at San Francisco, San Francisco, California. Submitted for publication July 6, 2004. Accepted for publication November 10, 2004. Supported by a UCSF Department of Anesthesia research fund generated from the clinical testing of pulse oximeters. No manufacturers were involved in any part of the concept, design or analysis of this study.

Address reprint requests to Dr. Bickler: Sciences 255, Box 0542, University of California Medical Center, 513 Parnassus Avenue, San Francisco, California 94143-0542. Address electronic mail to: bicklerp@anesthesia.ucsf.edu. Individual article reprints may be purchased through the Journal Web site, www.anesthesiology.org.

https://doi.org/10.1097/00000542-199206000-00024 (June 1992)

REVIEW ARTICLE
Julien F. Biebuyck, M.B., D.Phil., Editor

Anesthesiology
76:1018–1038, 1992

Recent Developments in Pulse Oximetry

John W. Severinghaus, M.D., Joseph F. Kelleher, M.D., M.A.†*

CONTENTS

Pulse oximeters now occupy most critical care arenas and virtually every operating room in the United States. They are manufactured by more than 35 firms, with 1989 annual world wide sales estimated at 65,000 units valued at $200 million.

In January 1989, two comprehensive reviews of pulse oximetry were published. One gave relative emphasis to theory of operation and other technical aspects,[1] while the other focused primarily on clinical issues.[2] Since the reviews of 1989 were completed, more than 500 additional publications have described methods, uses, problems, progress, and effects of pulse oximetry—135 of them in the 6-month period prior to the end date of this review (October 1, 1991). The past 3 yr have seen a variety of other reviews concerning some recent developments[3–12] as well as the history of pulse oximetry.[13,14]

The purpose of this article is to summarize the literature on pulse oximetry that has appeared since the major reviews of early 1989. Expressions such as "before 1988" and "since 1988," unless otherwise indicated, refer herein to the mid-1988 cutoff date for the references appearing in those reviews.

Methodological Developments

There is relatively little to report as to methodological advance in pulse oximetry since 1988. A potential exception is surface reflectance ("surface") oximetry, which has received significant recent experimental attention but does not appear ready for widespread clinical use. In 1988, a review of pulse oximetry[2] could dismiss the topic of reflectance oximetry by observing that it was "one desirable possibility for the future of pulse oximetry . . . in which measurements of reflected light would allow monitoring at nontransilluminable sites, such as the fetal presenting part during labor." With the possible exception of its use in labor, reflectance oximetry is still largely investigational. Some models now couple the oximeter to the ECG (see "Limitations of Pulse Oximetry: Motion

* Department of Anesthesia and Cardiovascular Research Institute, University of California, San Francisco, California.
† Anesthesia Service Medical Group, Inc., San Diego, California.
Received from the Department of Anesthesia and Cardiovascular Research Institute, University of California, San Francisco and the Anesthesia Service Medical Group, Inc., San Diego, California. Accepted for publication January 21, 1992.
This review contains portions of Severinghaus JW: Oximetry: Uses and limitations. ASA Refresher Courses in Anesthesiology 19:139–152, 1991.

Address reprint requests to Dr. Severinghaus: 1386HSE, UCSF, San Francisco, California 94143.
Key words: Blood: oxygen measurement; oxygen saturation. Complications: hypoxia; postoperative hypoxia; respiratory problems. Equipment: optical; oximeters. Monitoring: oximetry.

https://doi.org/10.1378/chest.97.6.1420 (June 1990)

Reliability of Pulse Oximetry in Titrating Supplemental Oxygen Therapy in Ventilator-Dependent Patients*

Amal Jubran, M.D.;† and Martin J. Tobin, M.D., F.C.C.P.‡

Pulse oximetry is widely used in the critical care setting, but few studies have examined its usefulness in clinical decision making. One area where pulse oximetry might be useful is in the titration of fractional inspired O_2 concentration (FIO_2) in ventilator-dependent patients. Unfortunately, documented guidelines for this use do not exist, and in a survey of directors of intensive care units, we found that they employed a wide range of target O_2 saturation (SpO_2) values. Consequently, we undertook a study to determine if SpO_2 could be reliably substituted for measurements of arterial O_2 tension (PaO_2) when adjusting FIO_2 in ventilator-dependent patients. We examined a number of SpO_2 target values in 54 critically ill patients aiming for a PaO_2 of ≥60 mm Hg, while minimizing the risk of O_2 toxicity. In white patients, we found that a SpO_2 target of 92 percent was reliable in predicting a satisfactory level of oxygenation. However, in black patients, such a SpO_2 reading was commonly associated with significant hypoxemia (PaO_2 as low as 49 mm Hg), and a higher SpO_2 target, 95 percent, was required. In addition, inaccurate oximetry readings (*ie*, >4 percent difference between SpO_2 and direct SaO_2 measurements) were more common in black (27 percent) than in white patients (11 percent, $p<0.05$). In conclusion, a SpO_2 target of 92 percent was reliable when titrating supplemental O_2 in white patients receiving mechanical ventilation; however, in black patients, such a SpO_2 reading was commonly associated with significant hypoxemia, and a higher SpO_2 target, 95 percent, was required to ensure a satisfactory level of oxygenation. *(Chest 1990; 97:1420-25)*

SpO_2 = pulse oximetry based oxygen saturation

Pulse oximetry has gained enormous popularity since its introduction into clinical medicine. In 1988, there were 23 manufacturers of pulse oximeters, 45 different models available, and approximately 45,000 units in use.[1] Indeed, Severinghaus and Astrup[2] recently considered pulse oximetry to be "the most significant technologic advance ever made in monitoring the well-being and safety of patients during anesthesia, recovery, and critical care." Although pulse oximeters have become ubiquitous in intensive care units, their role in clinical decision making is not well defined because most studies have focused on their accuracy rather than on their clinical usefulness.[3]

One area where oximetry might be useful is in the titration of optimal levels of supplemental oxygen (O_2). Recently, we had to totally depend on pulse oximetry to adjust O_2 therapy in a patient with severe adult respiratory distress syndrome who had significant hypoxemia (arterial O_2 tension of 55 mm Hg) while receiving a fractional inspired O_2 concentration (FIO_2) of 0.80. This young woman refused permission for the insertion of an arterial catheter, and also requested that no further arterial blood gas samples would be drawn. While she had a pulse oximeter probe attached to her finger, we did not know what optimal O_2 saturation target was to aim for (*ie*, ensure adequate oxygenation while minimizing the risk of O_2 toxicity), because published guidelines do not exist. Consequently, we undertook a systematic study of this problem consisting of two parts. First, we conducted a telephone survey of directors of intensive care units (ICUs) throughout the United States to determine if pulse oximetry is being employed to titrate FIO_2 in ventilator-dependent patients and the manner in which this is done. Secondly, we prospectively examined a number of algorithms to determine if pulse oximetry is reliable in assessing alterations in oxygenation that result from titration of FIO_2 in ventilator-dependent patients.

*From the Division of Pulmonary Medicine and Critical Care, University of Texas Health Science Center, Houston.
†Research Fellow.
‡Associate Professor.
Manuscript received September 21; revision accepted December 18.
Reprint requests: Dr. Jubran, University of Texas at Houston, 6431 Fannin, Houston 77030

Methods

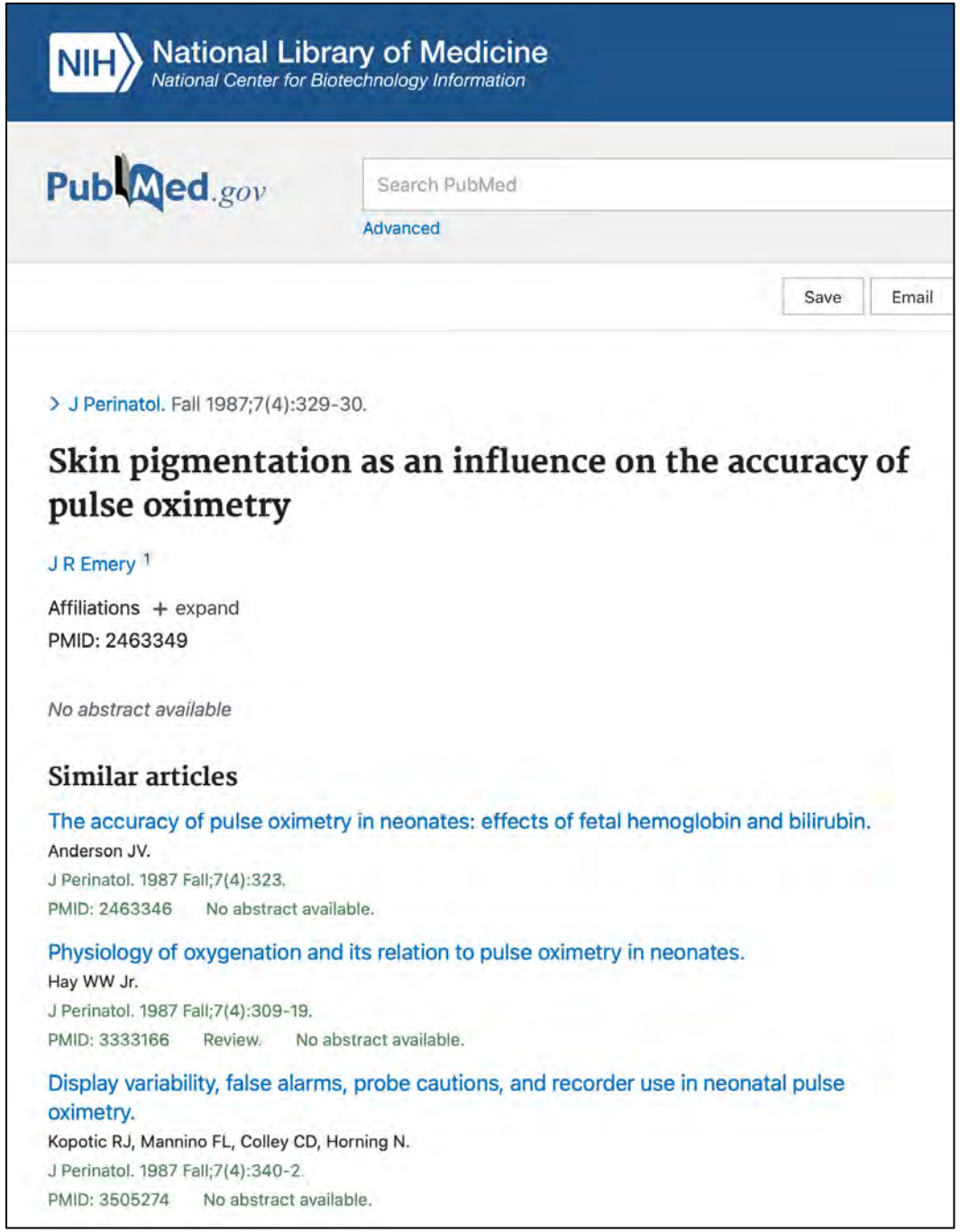
Hospital Survey

A telephone survey was conducted to determine if pulse oximetry is being used to evaluate the response to titrations in FIO_2. The medical directors of 25 hospitals (both university affiliated and private community hospitals) from various geographic locations throughout the United States were contacted. The medical director was asked whether he/she used pulse oximetry to titrate FIO_2 in ventilator-dependent patients, and, if so, what target SpO_2 he/she employed.

Patient Study

Patients: Fifty-four patients admitted to the medical intensive care units of Hermann Hospital and M. D. Anderson Hospital in

https://pubmed.ncbi.nlm.nih.gov/2463349/ (Fall 1987)



National Library of Medicine
National Center for Biotechnology Information

PubMed.gov

Search PubMed

Advanced

Save Email

> J Perinatol. Fall 1987;7(4):329-30.

Skin pigmentation as an influence on the accuracy of pulse oximetry

J R Emery [1]

Affiliations + expand

PMID: 2463349

No abstract available

Similar articles

The accuracy of pulse oximetry in neonates: effects of fetal hemoglobin and bilirubin.
Anderson JV.
J Perinatol. 1987 Fall;7(4):323.
PMID: 2463346 No abstract available.

Physiology of oxygenation and its relation to pulse oximetry in neonates.
Hay WW Jr.
J Perinatol. 1987 Fall;7(4):309-19.
PMID: 3333166 Review. No abstract available.

Display variability, false alarms, probe cautions, and recorder use in neonatal pulse oximetry.
Kopotic RJ, Mannino FL, Colley CD, Horning N.
J Perinatol. 1987 Fall;7(4):340-2.
PMID: 3505274 No abstract available.

Datasheets for Datasets by Timnit Gebru et al.

Example datasheet for Pang and Lee's polarity dataset [ACL 2004]

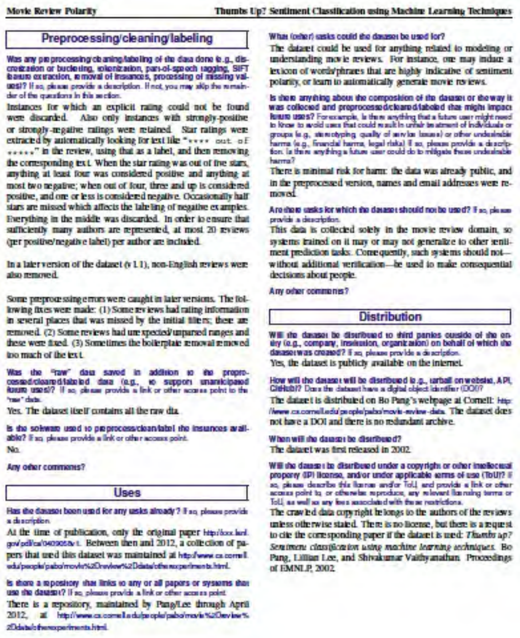
Movie Review Polarity — Thumbs Up? Sentiment Classification using Machine Learning Techniques

Motivation

Composition

Movie Review Polarity — Thumbs Up? Sentiment Classification using Machine Learning Techniques

Collection Process

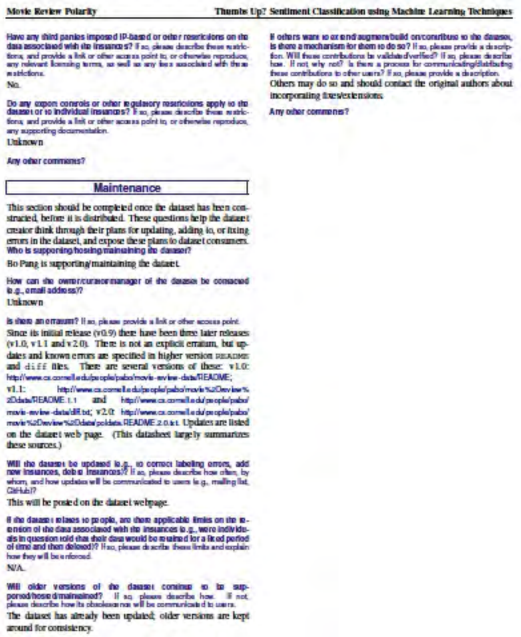


Movie Review Polarity — Thumbs Up? Sentiment Classification using Machine Learning Techniques

Preprocessing/cleaning/labeling

Uses

Distribution



Movie Review Polarity — Thumbs Up? Sentiment Classification using Machine Learning Techniques

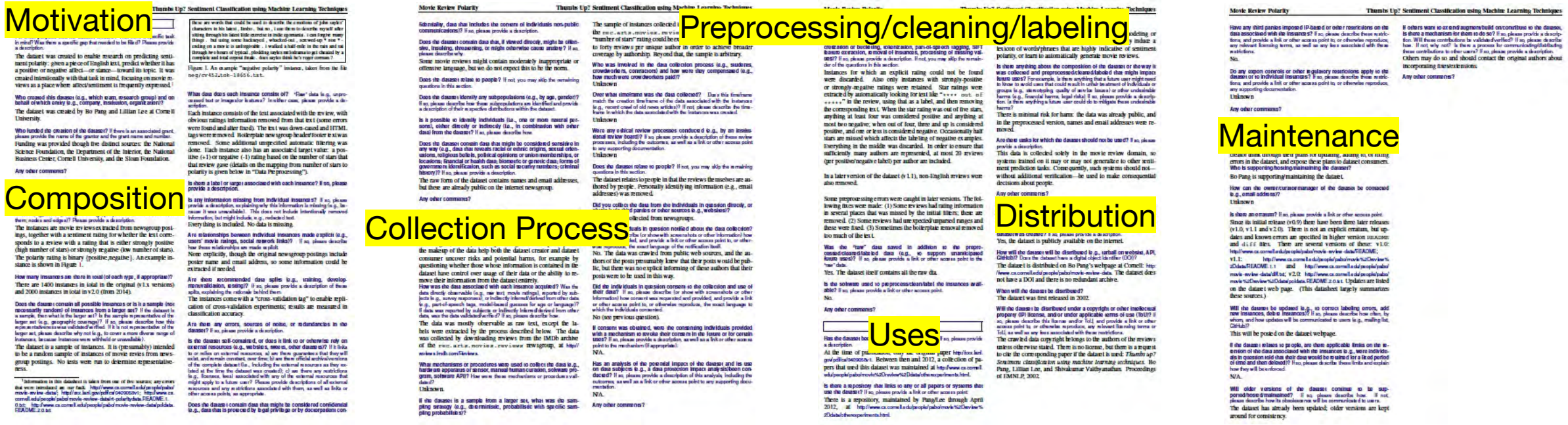
Maintenance

https://arxiv.org/abs/1803.09010

https://cacm.acm.org/magazines/2021/12/256932-datasheets-for-datasets/ (Dec 2021)

Datasheets for Datasets by Timnit Gebru et al.

Example datasheet for Pang and Lee's polarity dataset [ACL 2004]



https://arxiv.org/abs/1803.09010
https://cacm.acm.org/magazines/2021/12/256932-datasheets-for-datasets/ (Dec 2021)

Can we generate “aspirational” data?

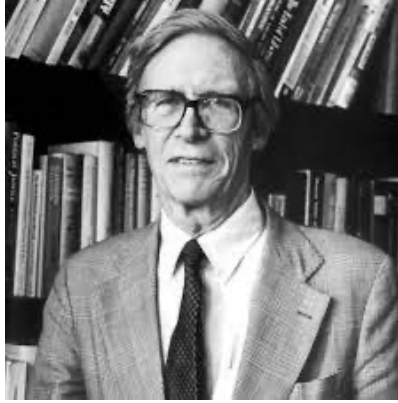
Deborah Ramirez

Can we generate “aspirational” data?

Aspirational data: synthetic data from ideally fair circumstances

Deborah Ramirez

Rawls' well-ordered society



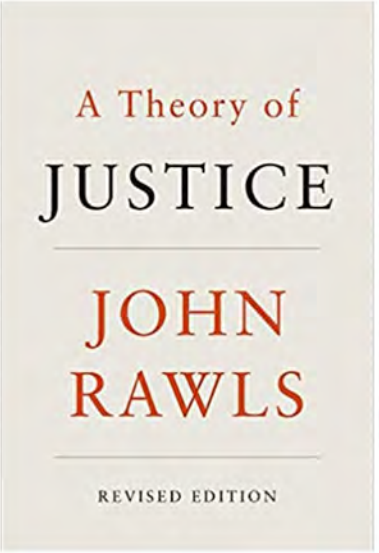
John Rawls

A society is well-ordered when it …

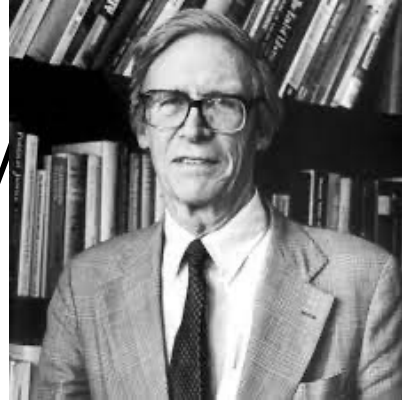
"advances the good of its members"

and

"is effectively regulated by a public conception of justice"



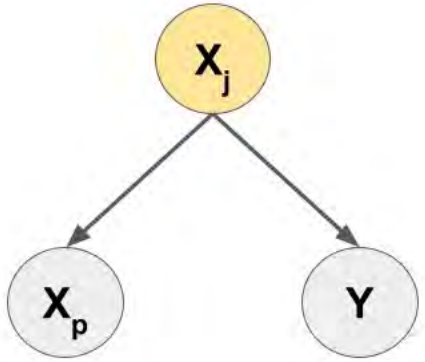
Rawls' Principle of Fair Equality of Opportunity (FEO)



John Rawls

- FEO governs the fair allocation of advantageous positions (e.g., high-paying jobs) in society.
- FEO can be formalized as a conditional independence (CI)
  - The probability of securing an advantageous position ought to be independent of protected variables (e.g., race) given justified variables (e.g., talent):

$$Y \perp\!\!\!\perp X_p | X_j$$

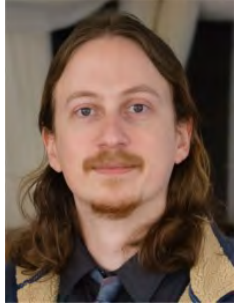


RAWLSNET


David Liu


Zohair Shafi


Will Fleisher


Scott Alfeld

Question: Given an unfair (in the Rawlsian sense) outcome and the capability to alter some (but not all) decision-making processes, how can one satisfy FEO?

Answer: **RAWLSNET**, a system for altering the parameters of Bayesian Network (BN) models to satisfy FEO

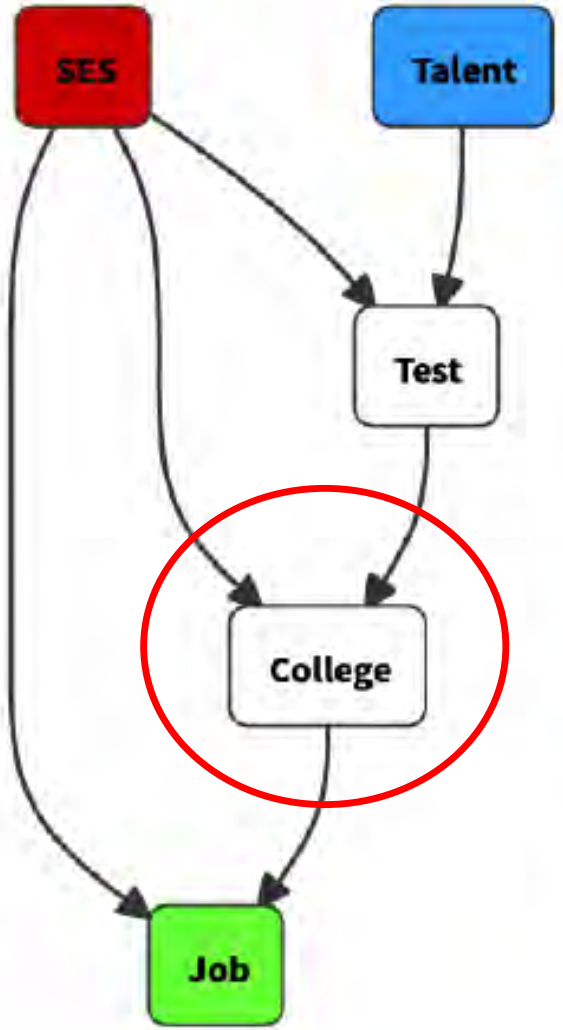
David Liu, Zohair Shafi, Will Fleisher, et al. RAWLSNET: Altering Bayesian Networks to Encode Rawlsian Fair Equality of Opportunity. *AIES* 2021: 745-755.

RAWLSNET

- RAWLSNET has three components:
  1. Learn a BN.
  2. Determine relevance to FEO.
  3. Update parameters of the BN to satisfy FEO if possible.
     Otherwise, update the parameters to approximately satisfy FEO.

College Admissions

SES	P
Low	0.8
High	0.2



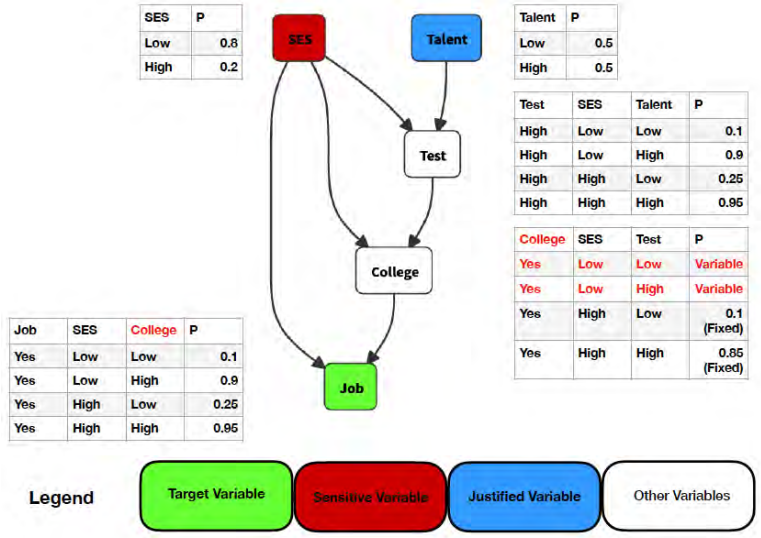
Talent	P
Low	0.5
High	0.5

Test	SES	Talent	P
High	Low	Low	0.1
High	Low	High	0.9
High	High	Low	0.25
High	High	High	0.95

College	SES	Test	P
Yes	Low	Low	Variable
Yes	Low	High	Variable
Yes	High	Low	0.1 (Fixed)
Yes	High	High	0.85 (Fixed)

Job	SES	College	P
Yes	Low	Low	0.1
Yes	Low	High	0.9
Yes	High	Low	0.25
Yes	High	High	0.95

Optimize for FEO



SES	P
Low	0.8
High	0.2

Talent	P
Low	0.5
High	0.5

Test	SES	Talent	P
High	Low	Low	0.1
High	Low	High	0.9
High	High	Low	0.25
High	High	High	0.95

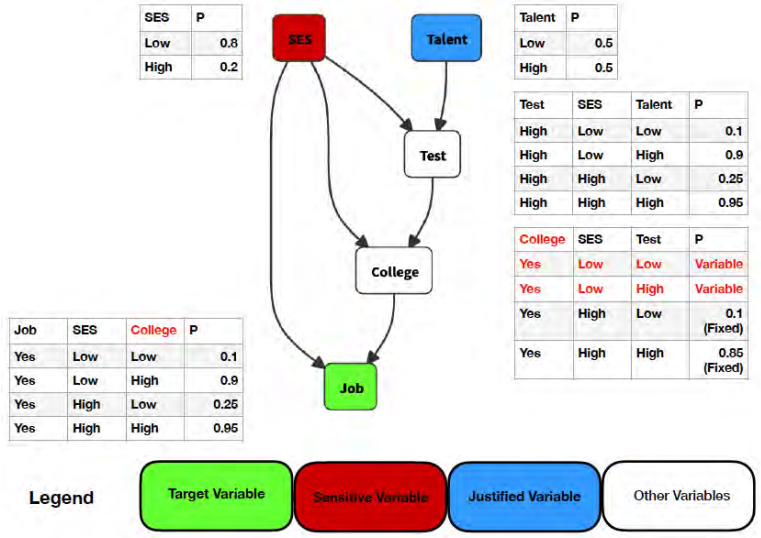
College	SES	Test	P
Yes	Low	Low	Variable
Yes	Low	High	Variable
Yes	High	Low	0.1 (Fixed)
Yes	High	High	0.85 (Fixed)

Job	SES	College	P
Yes	Low	Low	0.1
Yes	Low	High	0.9
Yes	High	Low	0.25
Yes	High	High	0.95

- Admissions policy is θ_{College}

$$\theta_{\text{College}} = \left\{\theta_{ij} = P(\text{College}|\text{Test} = i, \text{SES} = j)\right\} \forall i, j \in \{0, 1\}$$

Optimize for FEO



SES	P
Low	0.8
High	0.2

Talent	P
Low	0.5
High	0.5

Test	SES	Talent	P
High	Low	Low	0.1
High	Low	High	0.9
High	High	Low	0.25
High	High	High	0.95

College	SES	Test	P
Yes	Low	Low	Variable
Yes	Low	High	Variable
Yes	High	Low	0.1 (Fixed)
Yes	High	High	0.85 (Fixed)

Job	SES	College	P
Yes	Low	Low	0.1
Yes	Low	High	0.9
Yes	High	Low	0.25
Yes	High	High	0.95

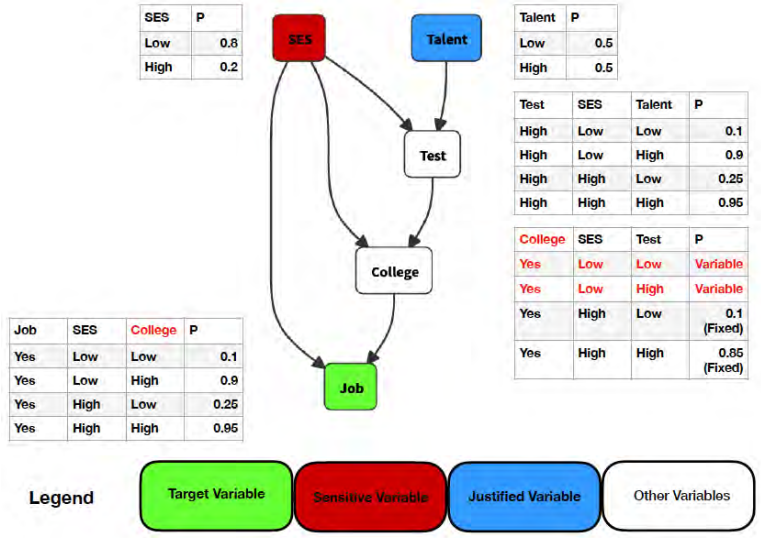
- Admissions policy is θ_{College}

$$\theta_{\text{College}} = \{\theta_{ij} = P(\text{College}|\text{Test} = i, \text{SES} = j)\}\ \forall i, j \in \{0, 1\}$$

- Optimization problem:

$$\underset{\theta_{\text{College}}}{\text{argmin}}\ [P(\text{Job}|\text{Talent} = 0, \text{SES} = 0) - P(\text{Job}|\text{Talent} = 0, \text{SES} = 1)]^2 +$$
$$[P(\text{Job}|\text{Talent} = 1, \text{SES} = 0) - P(\text{Job}|\text{Talent} = 1, \text{SES} = 1)]^2$$
$$\text{s.t. } \theta_{i,j} \in \{0, 1\}\ \forall i, j \text{ and } f_{\text{feasibility}}(\theta_{\text{College}}) \leq c_{\text{feasibility}}$$

Optimize for FEO



- Admissions policy is θ_{College}

$$\theta_{\text{College}} = \{\theta_{ij} = P(\text{College}|\text{Test} = i, \text{SES} = j)\}\ \forall i, j \in \{0, 1\}$$

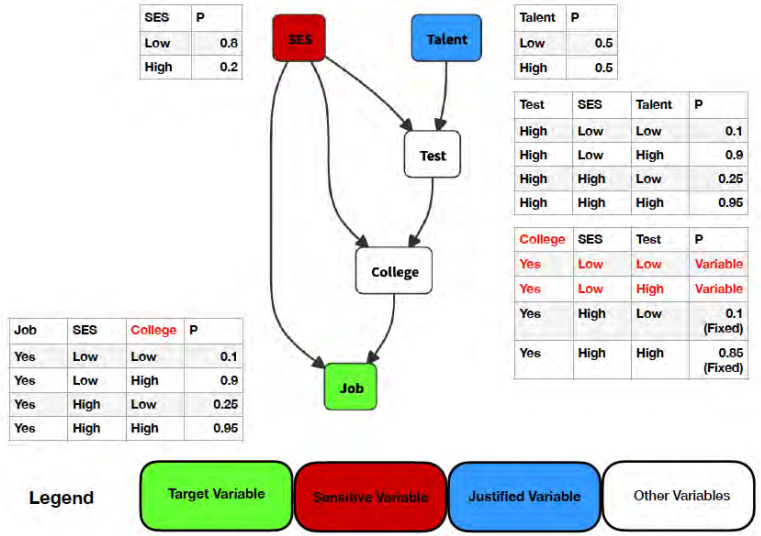
- Optimization problem:

$$\underset{\theta_{\text{College}}}{\text{argmin}} \left[P(\text{Job}|\text{Talent} = 0, \text{SES} = 0) - P(\text{Job}|\text{Talent} = 0, \text{SES} = 1)\right]^2 +$$

$$\left[P(\text{Job}|\text{Talent} = 1, \text{SES} = 0) - P(\text{Job}|\text{Talent} = 1, \text{SES} = 1)\right]^2$$

$$\text{s.t. } \theta_{i,j} \in \{0, 1\}\ \forall i, j \text{ and } f_{\text{feasibility}}(\theta_{\text{College}}) \leq c_{\text{feasibility}}$$

Optimize for FEO



SES	P
Low	0.8
High	0.2

Talent	P
Low	0.5
High	0.5

Test	SES	Talent	P
High	Low	Low	0.1
High	Low	High	0.9
High	High	Low	0.25
High	High	High	0.95

College	SES	Test	P
Yes	Low	Low	Variable
Yes	Low	High	Variable
Yes	High	Low	0.1 (Fixed)
Yes	High	High	0.85 (Fixed)

Job	SES	College	P
Yes	Low	Low	0.1
Yes	Low	High	0.9
Yes	High	Low	0.25
Yes	High	High	0.95

- Admissions policy is θ_{College}

$$\theta_{\text{College}} = \{\theta_{ij} = P(\text{College}|\text{Test} = i, \text{SES} = j)\}\ \forall i, j \in \{0, 1\}$$

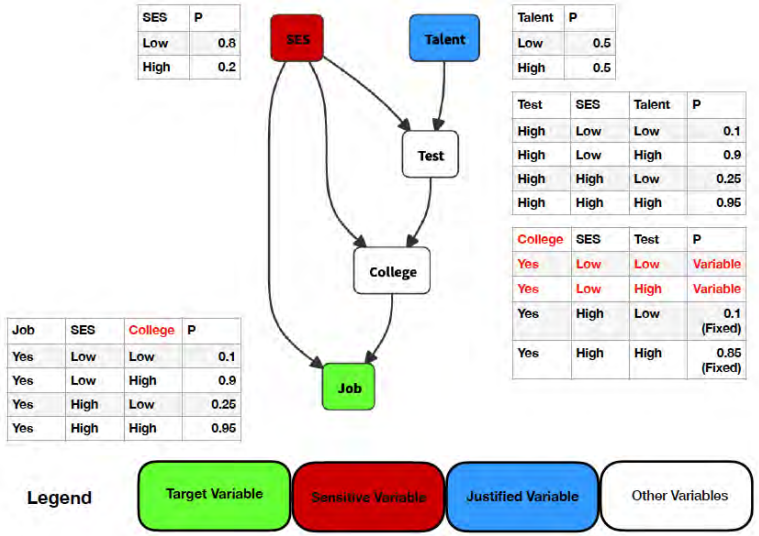
- Optimization problem:

$$\underset{\theta_{\text{College}}}{\text{argmin}}\ [P(\text{Job}|\text{Talent} = 0, \text{SES} = 0) - P(\text{Job}|\text{Talent} = 0, \text{SES} = 1)]^2 +$$

$$[P(\text{Job}|\text{Talent} = 1, \text{SES} = 0) - P(\text{Job}|\text{Talent} = 1, \text{SES} = 1)]^2$$

$$\text{s.t. } \theta_{i,j} \in \{0, 1\}\ \forall i, j \text{ and } f_{\text{feasibility}}(\theta_{\text{College}}) \leq c_{\text{feasibility}}$$

Optimize for FEO



SES	P
Low	0.8
High	0.2

Talent	P
Low	0.5
High	0.5

Test	SES	Talent	P
High	Low	Low	0.1
High	Low	High	0.9
High	High	Low	0.25
High	High	High	0.95

College	SES	Test	P
Yes	Low	Low	Variable
Yes	Low	High	Variable
Yes	High	Low	0.1 (Fixed)
Yes	High	High	0.85 (Fixed)

Job	SES	College	P
Yes	Low	Low	0.1
Yes	Low	High	0.9
Yes	High	Low	0.25
Yes	High	High	0.95

- Admissions policy is θ_{College}

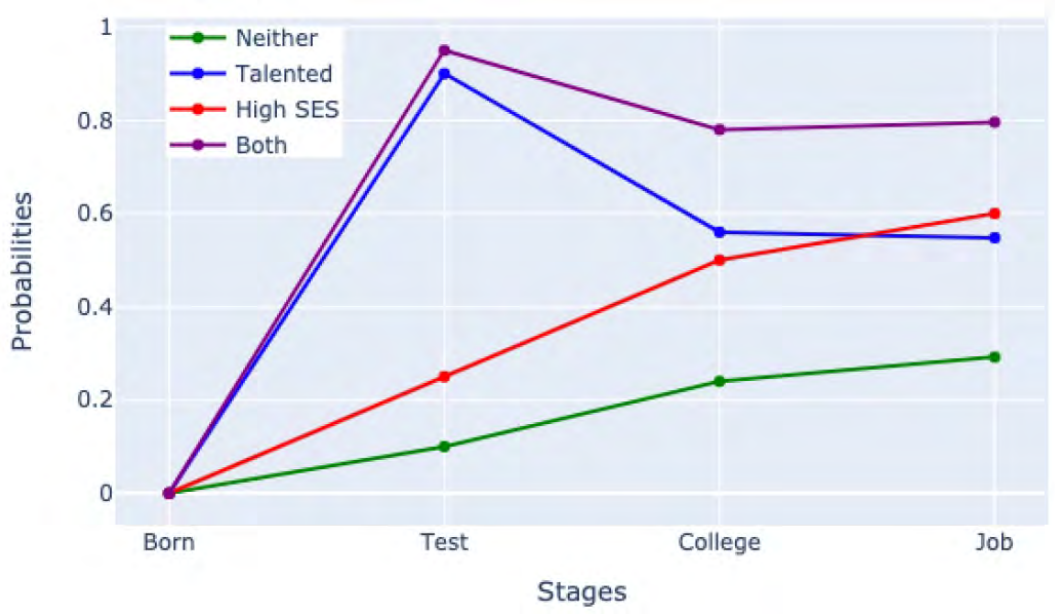
$$\theta_{\text{College}} = \{\theta_{ij} = P(\text{College}|\text{Test} = i, \text{SES} = j)\}\ \forall i, j \in \{0, 1\}$$

- Optimization problem:

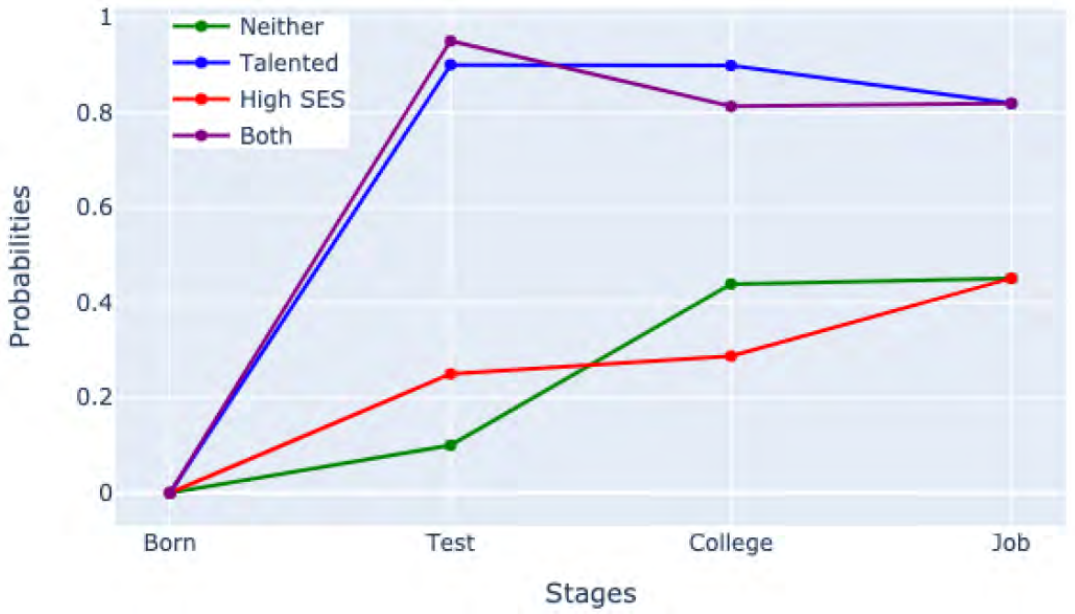
$$\underset{\theta_{\text{College}}}{\text{argmin}}\ [P(\text{Job}|\text{Talent} = 0, \text{SES} = 0) - P(\text{Job}|\text{Talent} = 0, \text{SES} = 1)]^2 +$$
$$[P(\text{Job}|\text{Talent} = 1, \text{SES} = 0) - P(\text{Job}|\text{Talent} = 1, \text{SES} = 1)]^2$$

s.t. $\theta_{i,j} \in \{0, 1\}\ \forall i, j$ and $f_{\text{feasibility}}(\theta_{\text{College}}) \leq c_{\text{feasibility}}$

Before and After RAWLSNET

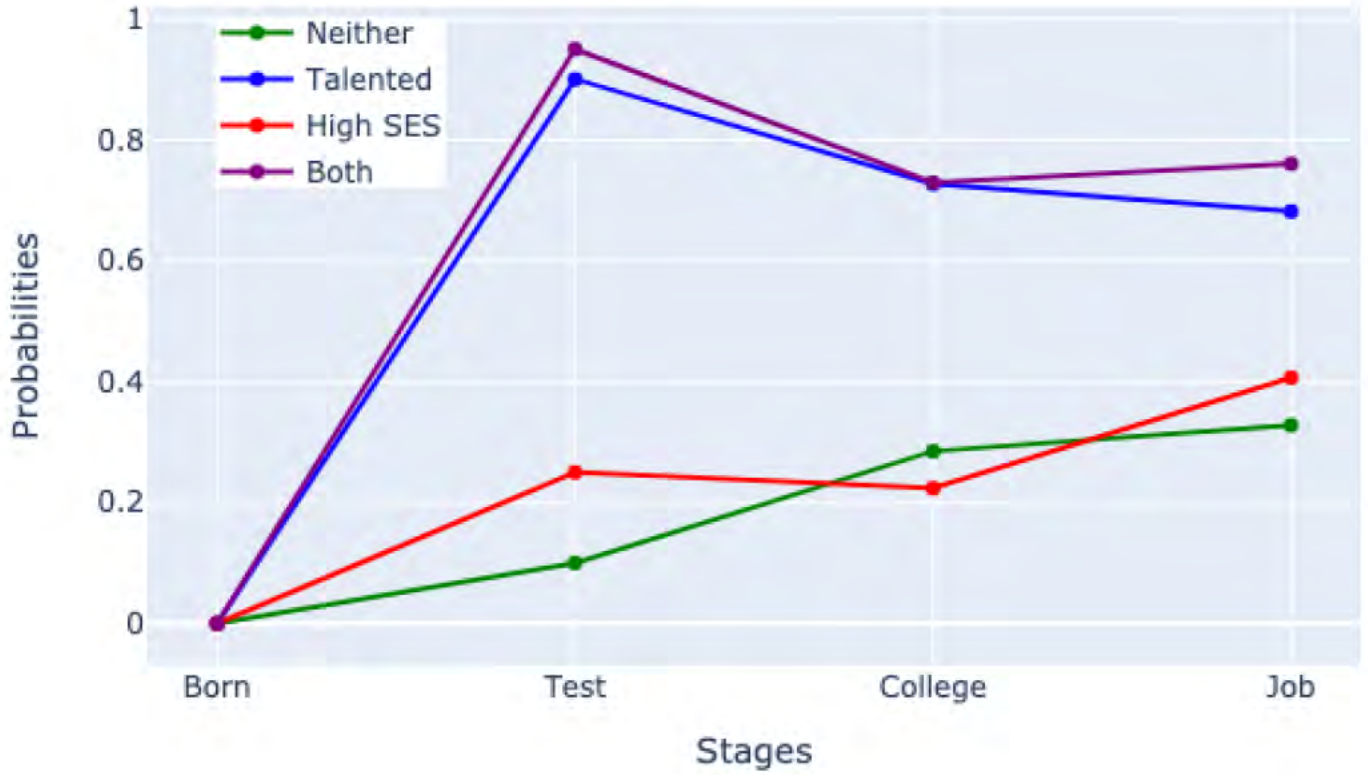


Probability of desired outcome **before** applying RAWLSNET



Probability of desired outcome **after** applying RAWLSNET

RAWLSNET with feasibility constraints



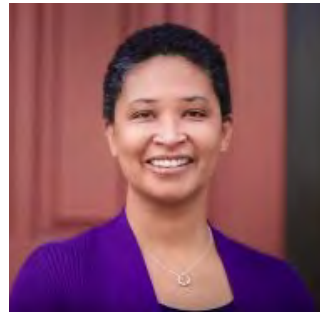
Probability of desired outcome
after applying RAWLSNET with **feasibility constraints**

Software available at https://github.com/dliu18/rawlsnet

Uses of RAWLSNET

1. It can be used to generate “aspirational” data: synthetic data from ideally fair circumstances.
2. It can aid policy makers in decision-making.
   - Example: RAWLSNET might be used to advise acceptance decisions for a college admissions committee.
     - Assuming distribution of talent is available in one of these ways, RAWLSNET will calculate the acceptance rates for applicants from different groups needed to satisfy FEO.

A Proposal for Decreasing Geographical Inequality in College Admissions



Danielle Allen

Branden Fitelson

- Talent is everywhere
- Can we use zip codes and merit to enhance diversity?
- We proposed an algorithm in 2014
  - http://fitelson.org/tie.pdf
- In 2021, it was adapted by Boston Latin School for admissions decisions

Algorithm 1 Calculating weights w_z for the weighted average of the $\{R^z_{ys}\}$.

```
c := 10^-6                                          Set exponentially decaying constant.
for z = 1 to n do
    w_z := 0                                        Initialize popularity weight w_z of zip+4 code z.
end for
for y = 1 to m do                                   Consider data from the past m years.
    for z = 1 to n do                               Iterate over all the zip+4 codes.
        if (w_z = 0) and (R^z_ys > 0) then          R^z_ys > 0 presently but not previously.
            w_z := 1
        else if (w_z > 0) and (R^z_ys = 0) then     R^z_ys > 0 previously but not presently.
            w_z := ((1 − c) × w_z)
        else if (w_z > 0) and (R^z_ys > 0) then     R^z_ys > 0 previously and presently.
            w_z := ((1 − c) × w_z) + 1
        end if
    end for
end for
for z = 1 to n do
    w_z := w_z / max({w_z})                         Normalize weights such that w_z ∈ [0, 1].
end for
```

T. Eliassi-Rad, B. Fitelson. **A Proposal for Decreasing Geographical Inequality in College Admissions**. Appendix of Chapter 12 (Talent is Everywhere by Danielle Allen) in *The Future of Affirmative Action*, Eds: J. Renker and J. Miller, The Century Foundation Press, 2014.

Machine learning life-cycle

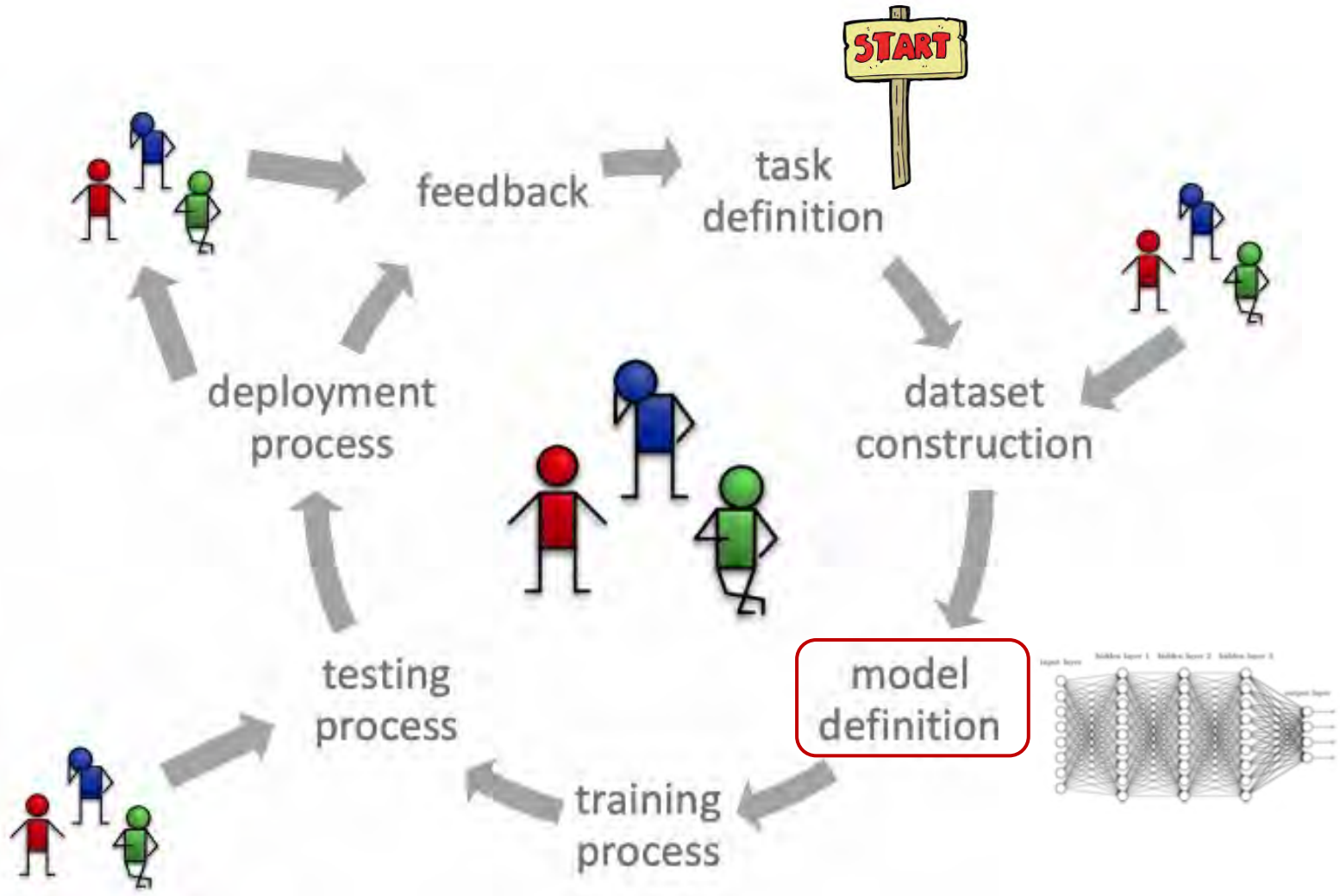


Image courtesy of [Jennifer Wortman Vaughan]

Ongoing work

When data are fair, but the model isn’t

Question: Given a fair data distribution and the structure of a BN, does maximum likelihood estimation learn parameters that produce a fair posterior distribution (i.e., one which preserves the fairness in the data distribution)?

Ongoing work

When data are fair, but the model isn't

Question: Given a fair data distribution and the structure of a BN, does maximum likelihood estimation learn parameters that produce a fair posterior distribution (i.e., one which preserves the fairness in the data distribution)?

Answer: **Not necessarily**. It depends on (1) the correctness of the BN's structure, (2) the faithfulness of the learned joint distribution to the BN's structure, (3) the correctness of the learned joint distribution, and (4) whether one is only interested in asymptotic behavior.

Model Cards for Model Reporting

by Margaret Mitchell et al.

https://arxiv.org/abs/1810.03993

Model Card - Smiling Detection in Images

Model Details

- Developed by researchers at Google and the University of Toronto, 2018, v1.
- Convolutional Neural Net.
- Pretrained for face recognition then fine-tuned with cross-entropy loss for binary smiling classification.

Intended Use

- Intended to be used for fun applications, such as creating cartoon smiles on real images; augmentative applications, such as providing details for people who are blind; or assisting applications such as automatically finding smiling photos.
- Particularly intended for younger audiences.
- Not suitable for emotion detection or determining affect; smiles were annotated based on physical appearance, and not underlying emotions.

Factors

- Based on known problems with computer vision face technology, potential relevant factors include groups for gender, age, race, and Fitzpatrick skin type; hardware factors of camera type and lens type; and environmental factors of lighting and humidity.
- Evaluation factors are gender and age group, as annotated in the publicly available dataset CelebA [36]. Further possible factors not currently available in a public smiling dataset. Gender and age determined by third-party annotators based on visual presentation, following a set of examples of male/female gender and young/old age. Further details available in [36].

Metrics

- Evaluation metrics include **False Positive Rate** and **False Negative Rate** to measure disproportionate model performance errors across subgroups. **False Discovery Rate** and **False Omission Rate**, which measure the fraction of negative (not smiling) and positive (smiling) predictions that are incorrectly predicted to be positive and negative, respectively, are also reported. [48]
- Together, these four metrics provide values for different errors that can be calculated from the confusion matrix for binary classification systems.
- These also correspond to metrics in recent definitions of "fairness" in machine learning (cf. [6, 26]), where parity across subgroups for different metrics correspond to different fairness criteria.
- 95% confidence intervals calculated with bootstrap resampling.
- All metrics reported at the .5 decision threshold, where all error types (FPR, FNR, FDR, FOR) are within the same range (0.04 - 0.14).

Training Data

- CelebA [36], training data split.

Evaluation Data

- CelebA [36], test data split.
- Chosen as a basic proof-of-concept.

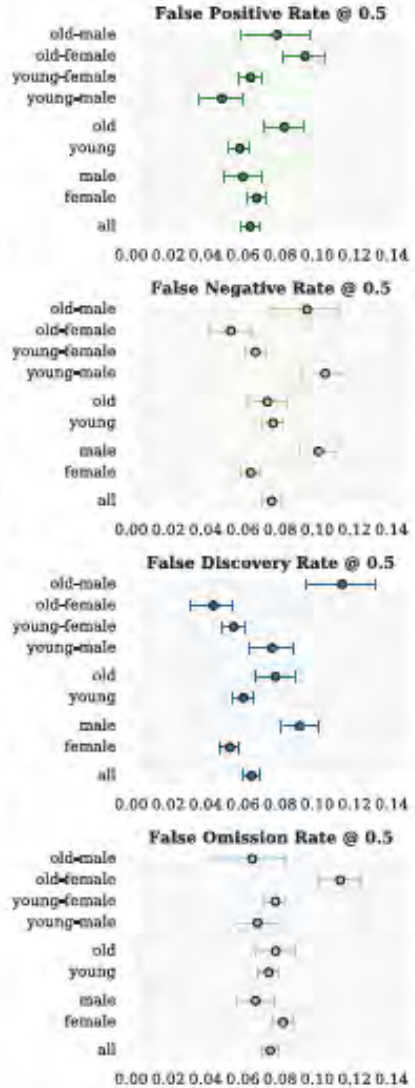
Ethical Considerations

- Faces and annotations based on public figures (celebrities). No new information is inferred or annotated.

Caveats and Recommendations

- Does not capture race or skin type, which has been reported as a source of disproportionate errors [5].
- Given gender classes are binary (male/not male), which we include as male/female. Further work needed to evaluate across a spectrum of genders.
- An ideal evaluation dataset would additionally include annotations for Fitzpatrick skin type, camera details, and environment (lighting/humidity) details.

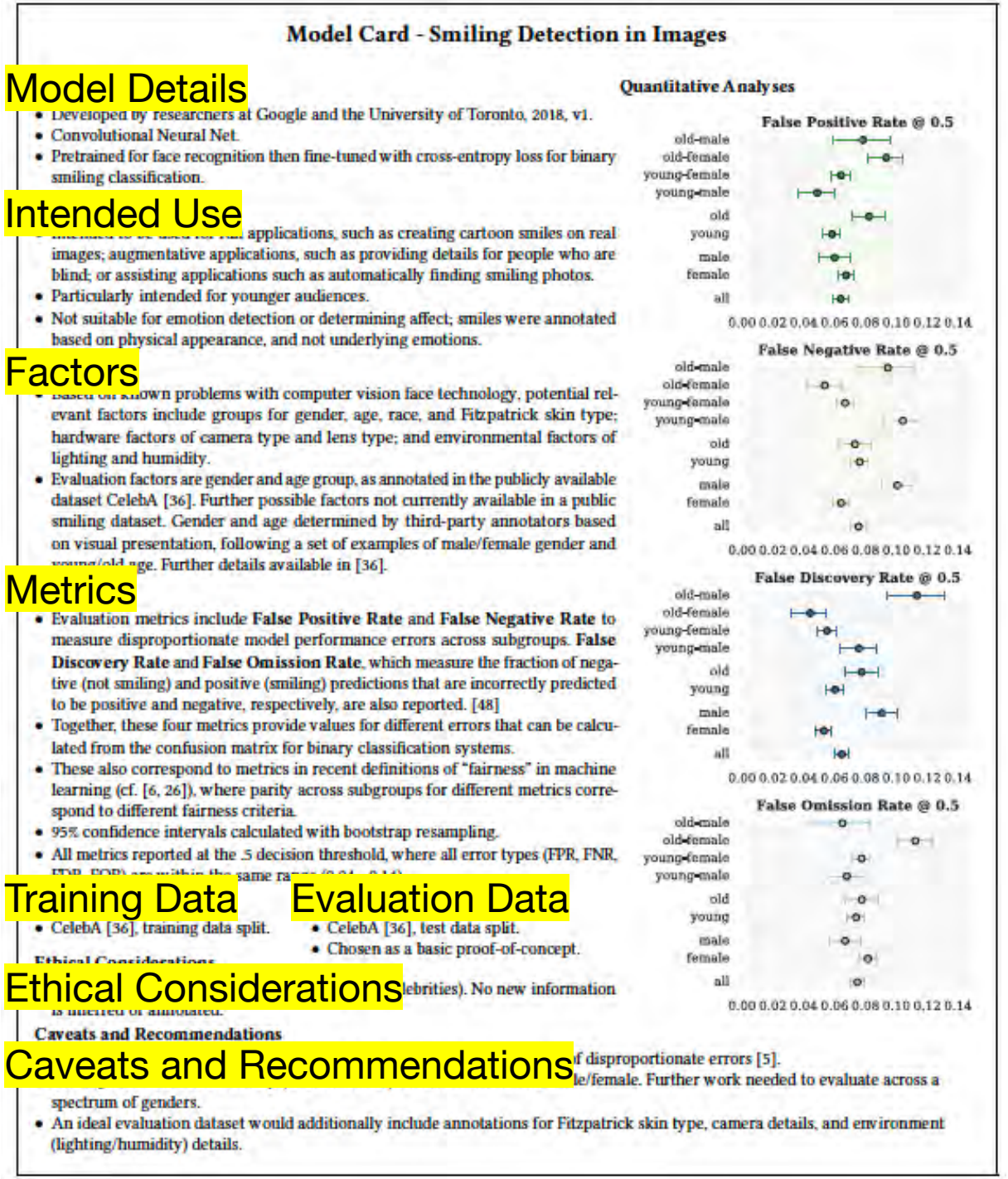
Quantitative Analyses



Model Cards for Model Reporting

by Margaret Mitchell et al.

https://arxiv.org/abs/1810.03993



Model Card - Smiling Detection in Images

Model Details

- Developed by researchers at Google and the University of Toronto, 2018, v1.
- Convolutional Neural Net.
- Pretrained for face recognition then fine-tuned with cross-entropy loss for binary smiling classification.

Intended Use

- applications, such as creating cartoon smiles on real images; augmentative applications, such as providing details for people who are blind; or assisting applications such as automatically finding smiling photos.
- Particularly intended for younger audiences.
- Not suitable for emotion detection or determining affect; smiles were annotated based on physical appearance, and not underlying emotions.

Factors

- known problems with computer vision face technology, potential relevant factors include groups for gender, age, race, and Fitzpatrick skin type; hardware factors of camera type and lens type; and environmental factors of lighting and humidity.
- Evaluation factors are gender and age group, as annotated in the publicly available dataset CelebA [36]. Further possible factors not currently available in a public smiling dataset. Gender and age determined by third-party annotators based on visual presentation, following a set of examples of male/female gender and young/old age. Further details available in [36].

Metrics

- Evaluation metrics include **False Positive Rate** and **False Negative Rate** to measure disproportionate model performance errors across subgroups. **False Discovery Rate** and **False Omission Rate**, which measure the fraction of negative (not smiling) and positive (smiling) predictions that are incorrectly predicted to be positive and negative, respectively, are also reported. [48]
- Together, these four metrics provide values for different errors that can be calculated from the confusion matrix for binary classification systems.
- These also correspond to metrics in recent definitions of "fairness" in machine learning (cf. [6, 26]), where parity across subgroups for different metrics correspond to different fairness criteria.
- 95% confidence intervals calculated with bootstrap resampling.
- All metrics reported at the .5 decision threshold, where all error types (FPR, FNR, FDR, FOR) are within the same ra

Training Data

- CelebA [36], training data split.

Evaluation Data

- CelebA [36], test data split.
- Chosen as a basic proof-of-concept.

Ethical Considerations

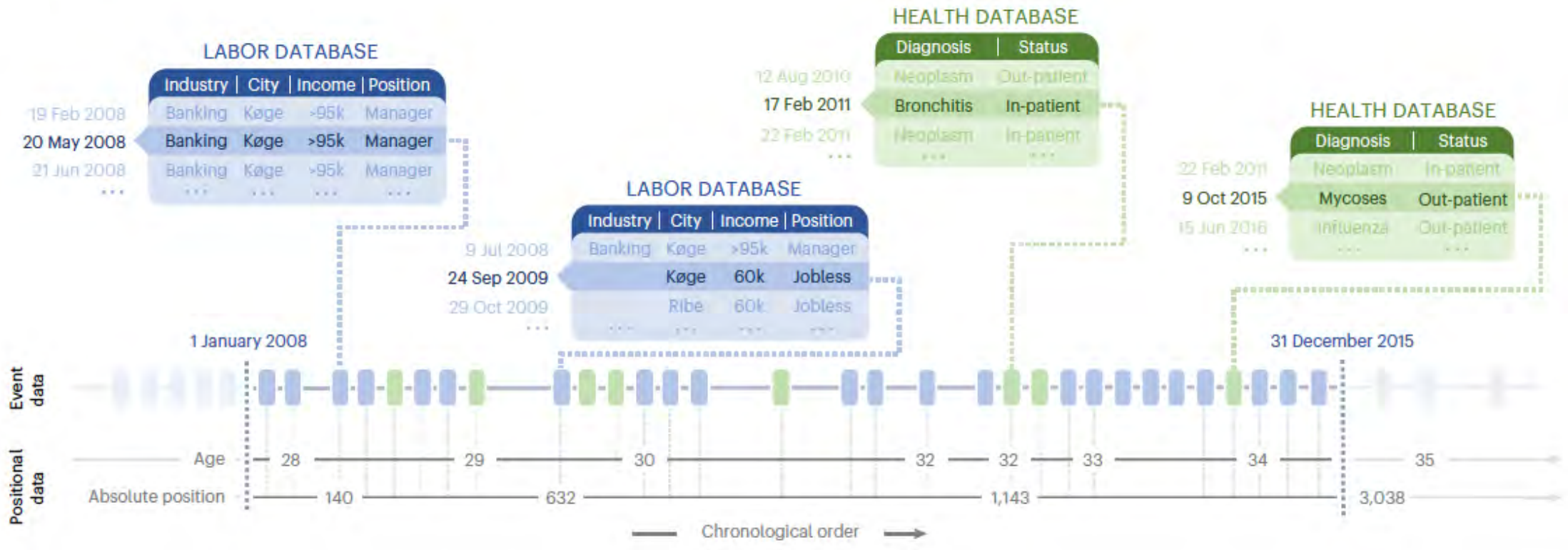
lebrities). No new information is inferred or annotated.

Caveats and Recommendations

f disproportionate errors [5].
e/female. Further work needed to evaluate across a spectrum of genders.
- An ideal evaluation dataset would additionally include annotations for Fitzpatrick skin type, camera details, and environment (lighting/humidity) details.

Quantitative Analyses

life2vec: vector representations of human lives



Germans Savcisens, et al. **Using sequences of life-events to predict human lives.**
Nature Computational Science, 2023. https://doi.org/10.1038/s43588-023-00573-5

Machine learning life-cycle

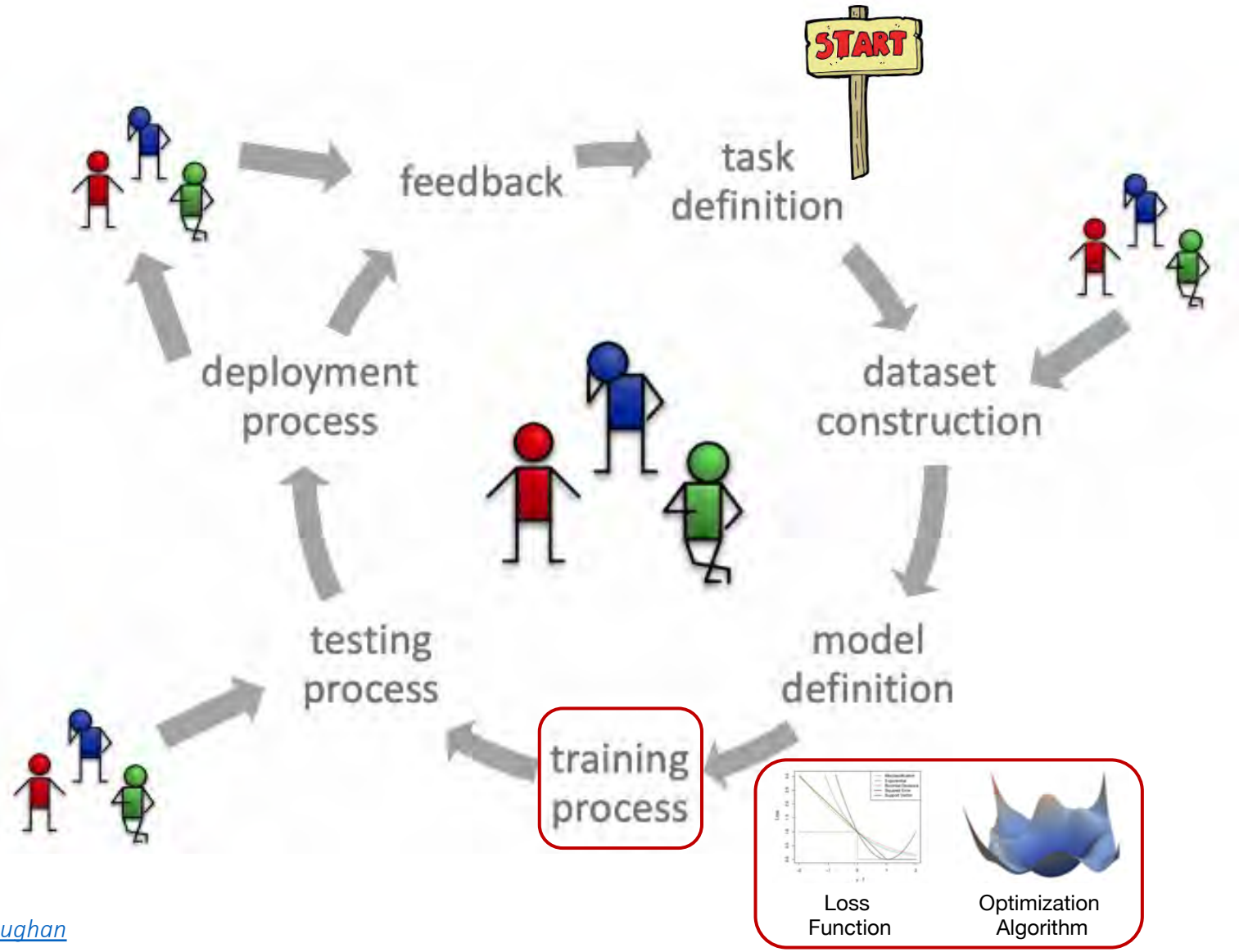


Image courtesy of [Jennifer Wortman Vaughan]

Machine learning life-cycle

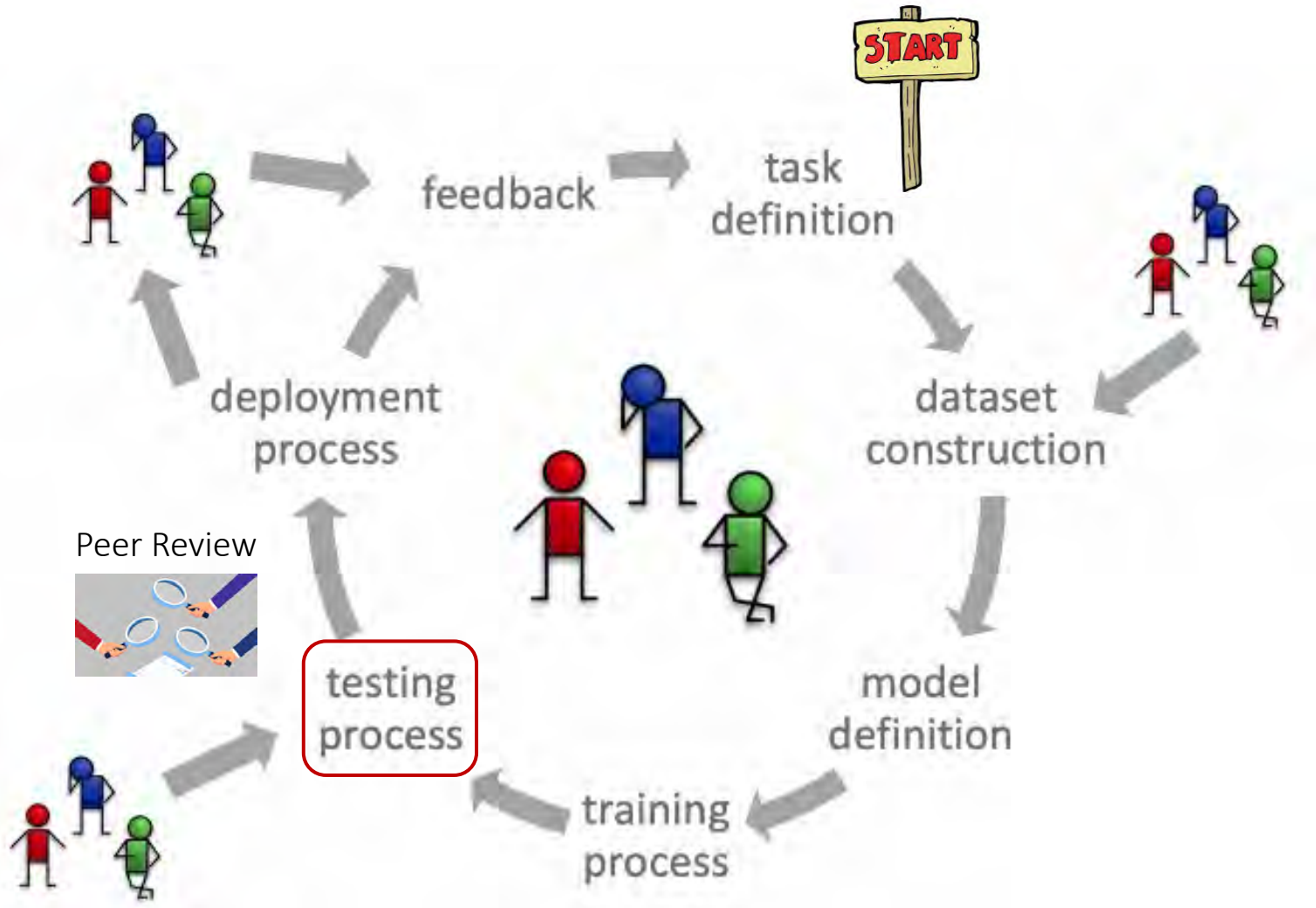


Image courtesy of [Jennifer Wortman Vaughan]

Machine learning life-cycle

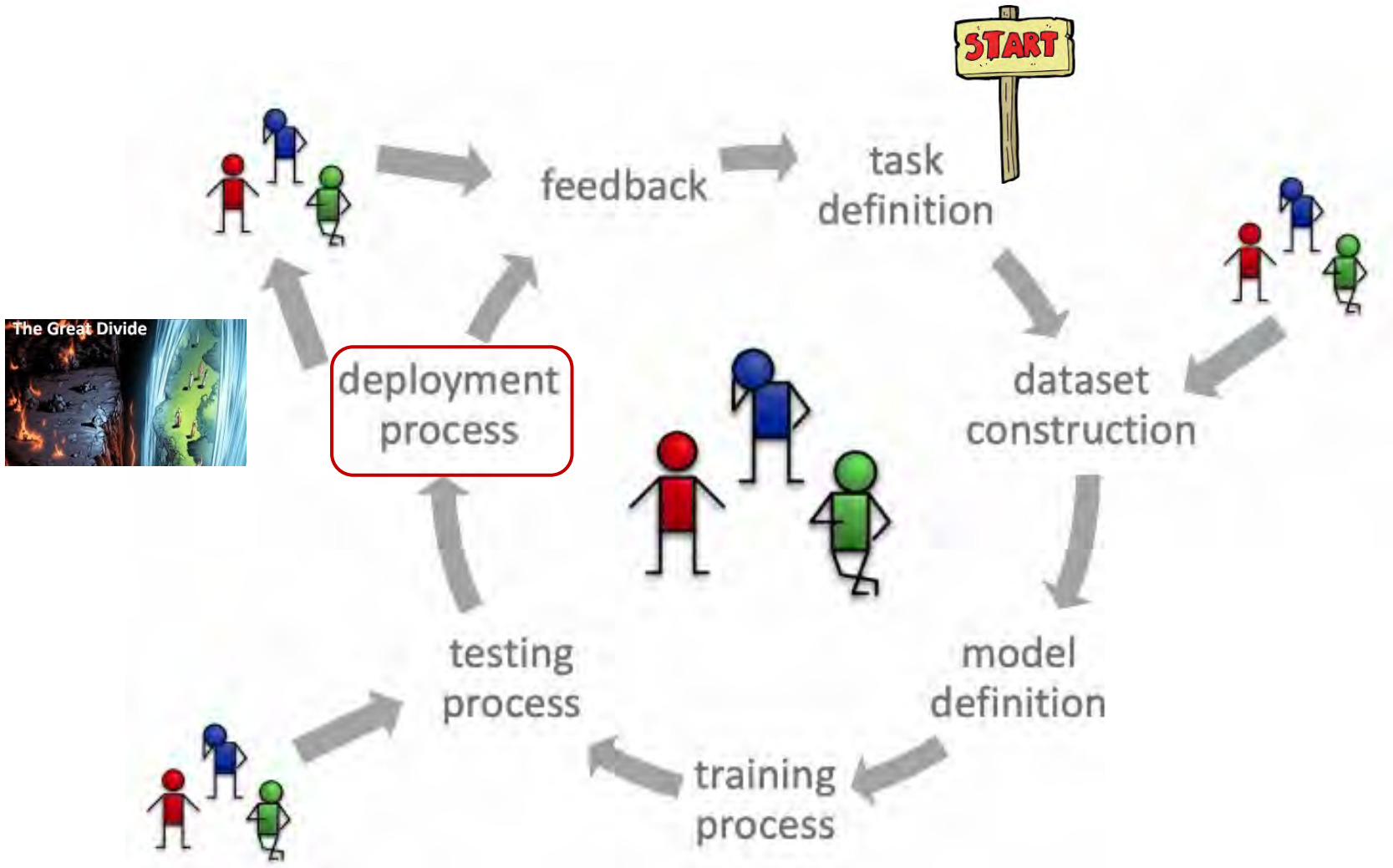


Image courtesy of [Jennifer Wortman Vaughan]

David Liu

Priyanka Nanayakkara

How do AI researchers describe and respond to the negative impact of their work on society?

D. Liu, P. Nanayakkara et al. Examining Responsibility and Deliberation in AI Impact Statements and Ethics Reviews. In *Proceedings of the 2022 AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society* (AIES), August 2022.

Examining Responsibility and Deliberation in AI Impact Statements and Ethics Reviews in NeurIPS 2021

Impact Statements	Ethics Reviews
• **Lack of agency** • Adversarial users • Improper input or application • **Denying responsibility** • Explicitly by minimizing negative societal impact • Implicitly by not proposing mitigation strategies • **Assigning responsibility** • To practitioners • To the subfield • To future work	• **Issues** • Policy vs. non-policy • Scope • **Recommendations** • Identification vs. mitigation • **Interaction: Justification mechanisms** • Citing • Retracting

David Liu

Priyanka Nanayakkara

How do AI researchers describe and respond to the negative impact of their work on society? Badly.

D. Liu, P. Nanayakkara et al. Examining Responsibility and Deliberation in AI Impact Statements and Ethics Reviews. In *Proceedings of the 2022 AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society* (AIES), August 2022.

Machine learning life-cycle

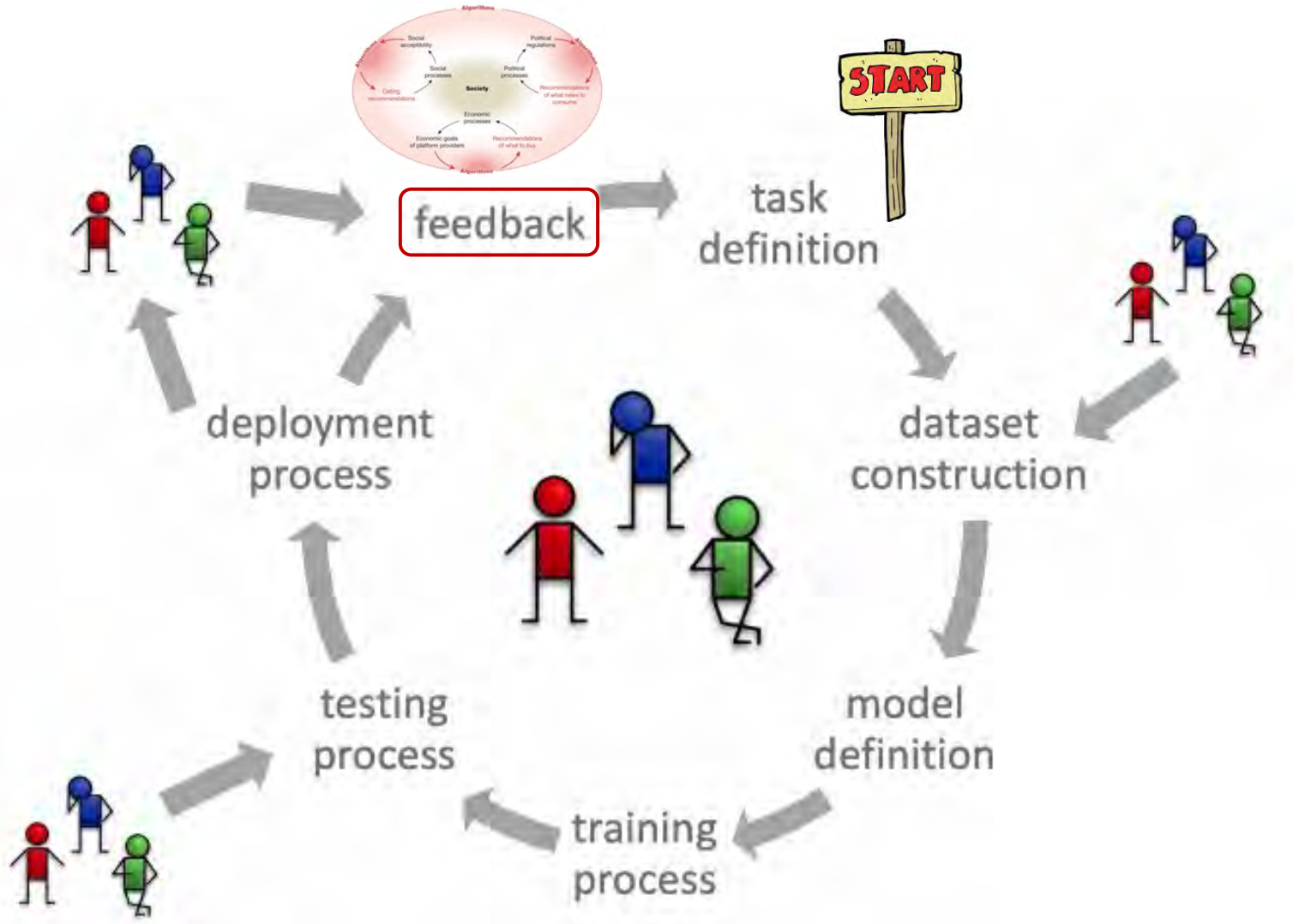
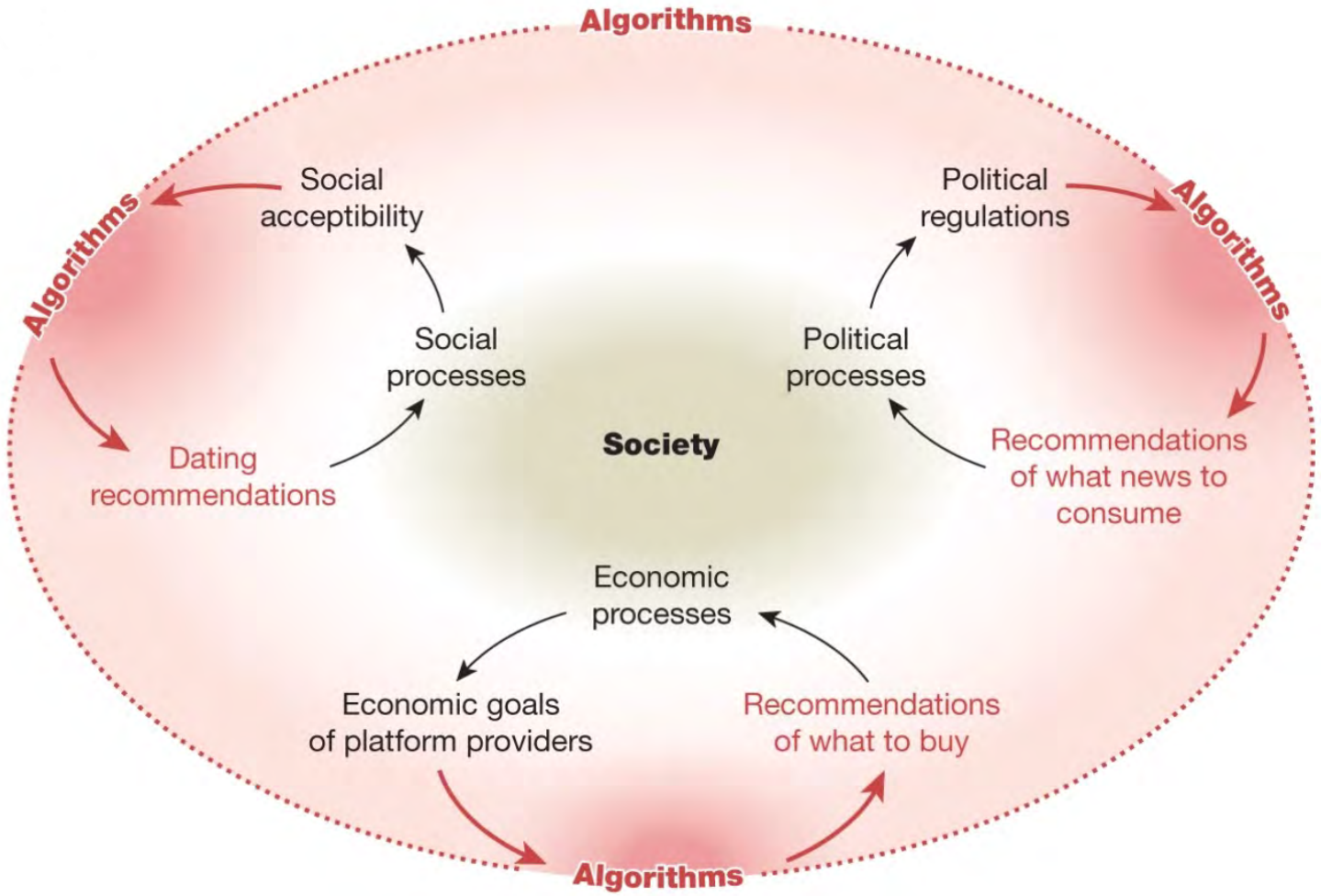


Image courtesy of [Jennifer Wortman Vaughan]

Measuring Algorithmically Infused Societies




Claudia Wagner


Markus Strohmaier


Alexandra Olteanu


Emre Kiciman


Noshir Contractor

C. Wagner, M. Strohmaier, A. Olteanu, E. Kiciman, N. Contractor, T. Eliassi-Rad. **Measuring Algorithmically Infused Societies**. *Nature*, 595: 197-204, 2021. https://doi.org/10.1038/s41586-021-03666-1

Three challenges

- Insufficient quality of measurements
- Complex consequences of (mis)measurements
- Limits of existing social theories

How to address these challenges?

- Insufficient quality of measurements
  - Triangulate data to examine the measurement quality
  - Develop guidelines and best practices
- Complex consequences of (mis)measurements
  - Reflect on what to measure and what not to measure
  - Develop professional norms.
- Limits of existing social theories
  - Integrate data and measurements into the theory construction process
  - Establish transparent, participatory processes for examining algorithmically infused societies

ML Life-cycle

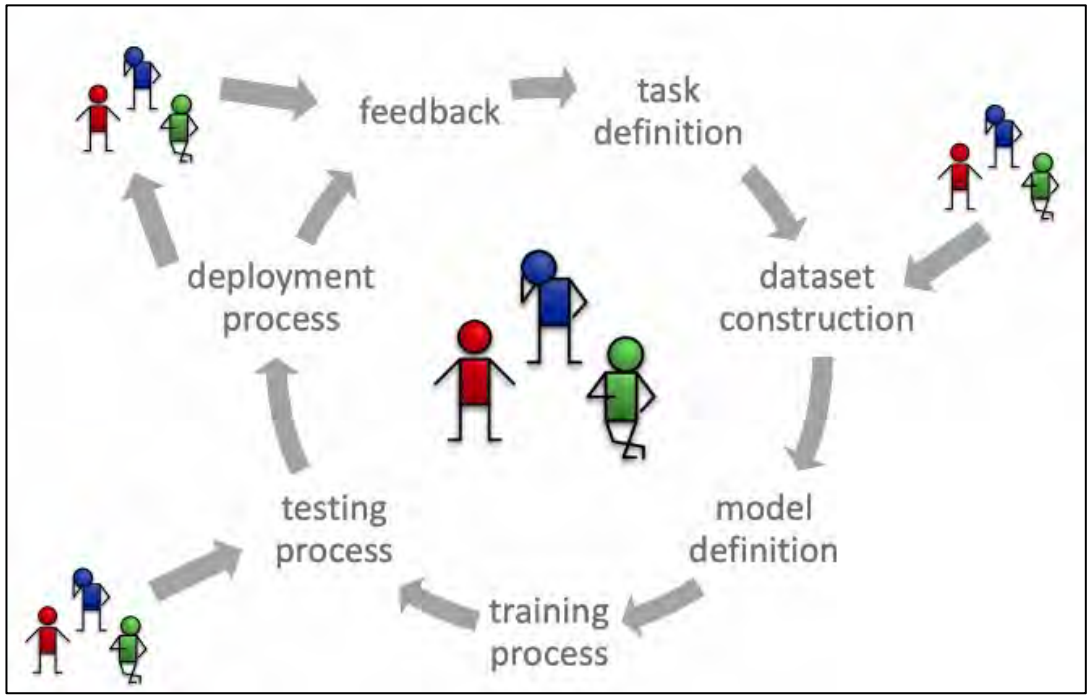


Image by Jennifer Wortman Vaughan

Complex System

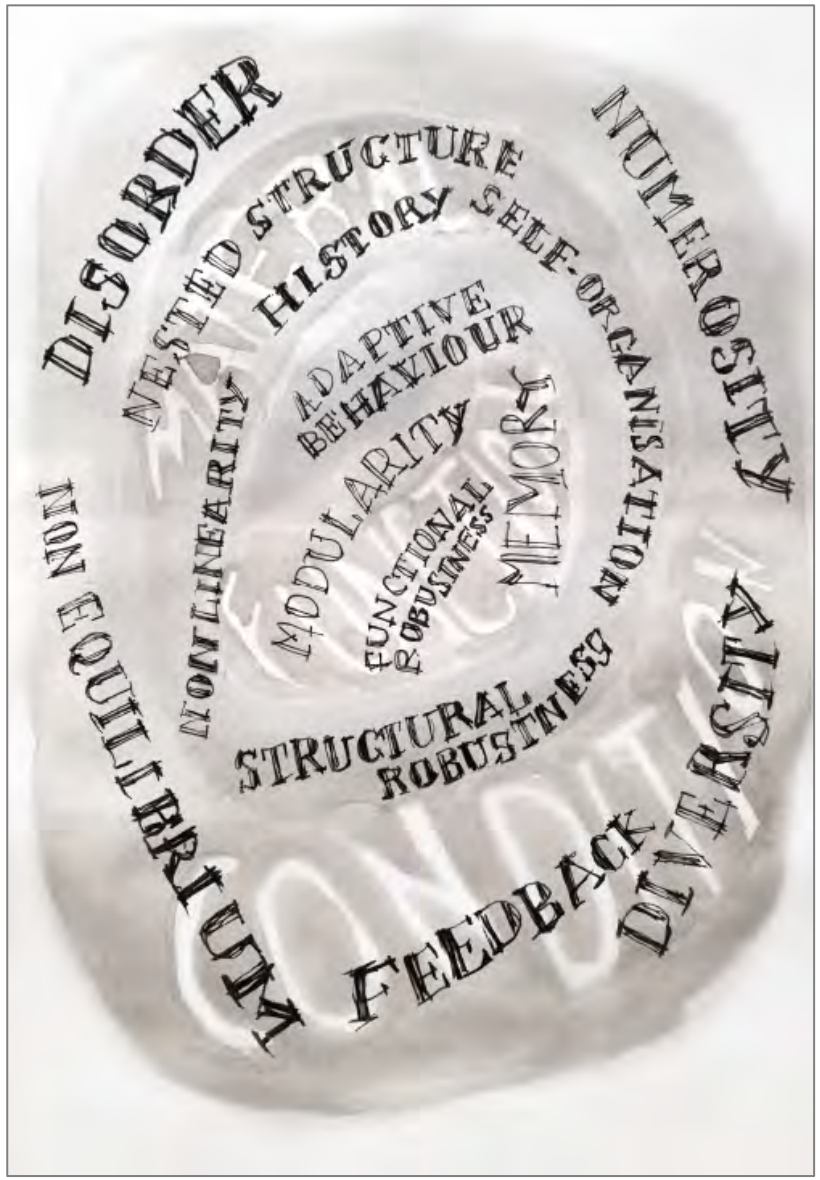


Image by E Bozzarelli

ML Life-cycle

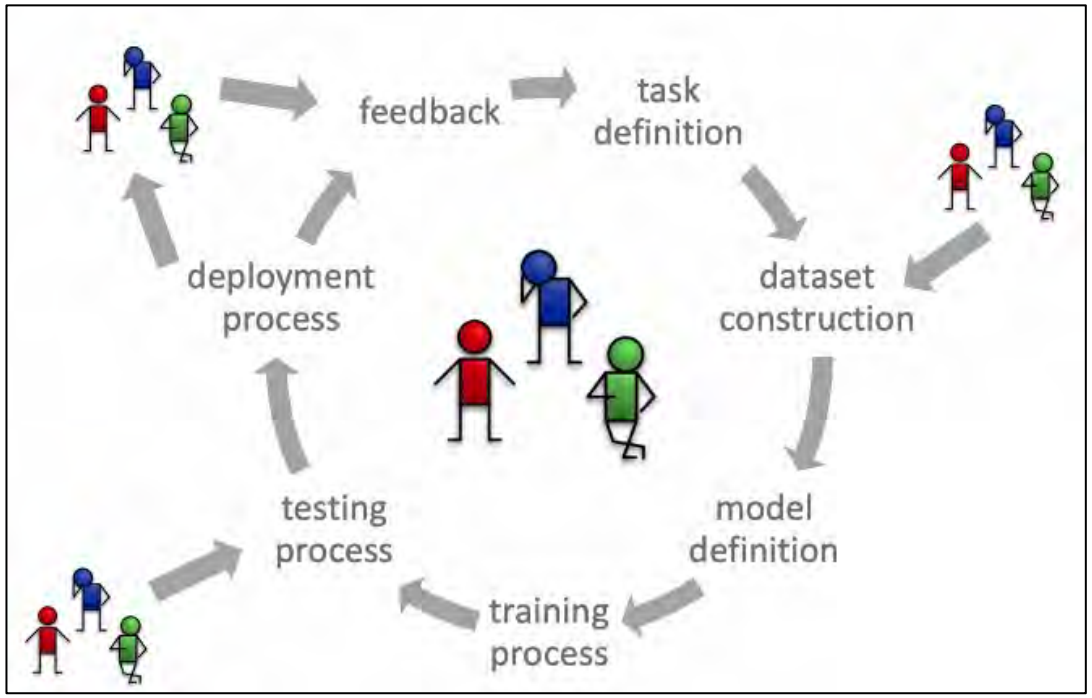
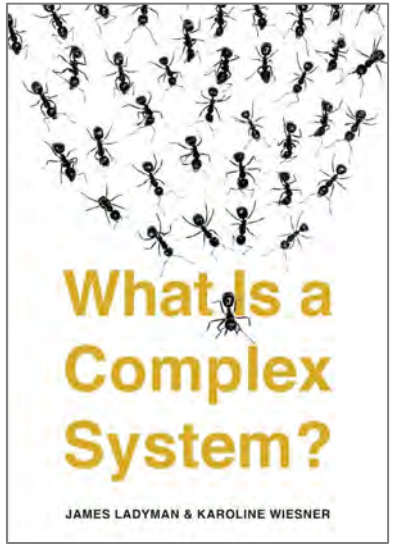


Image by Jennifer Wortman Vaughan



Complex System

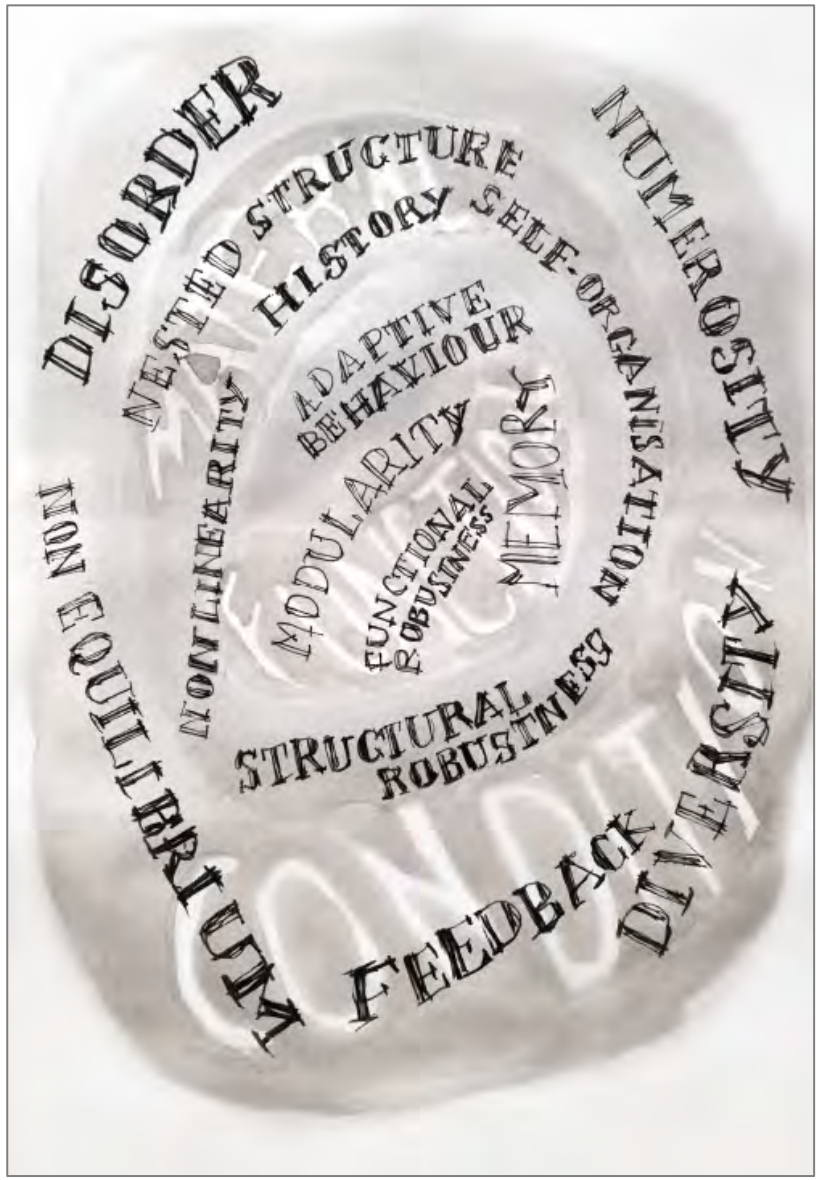
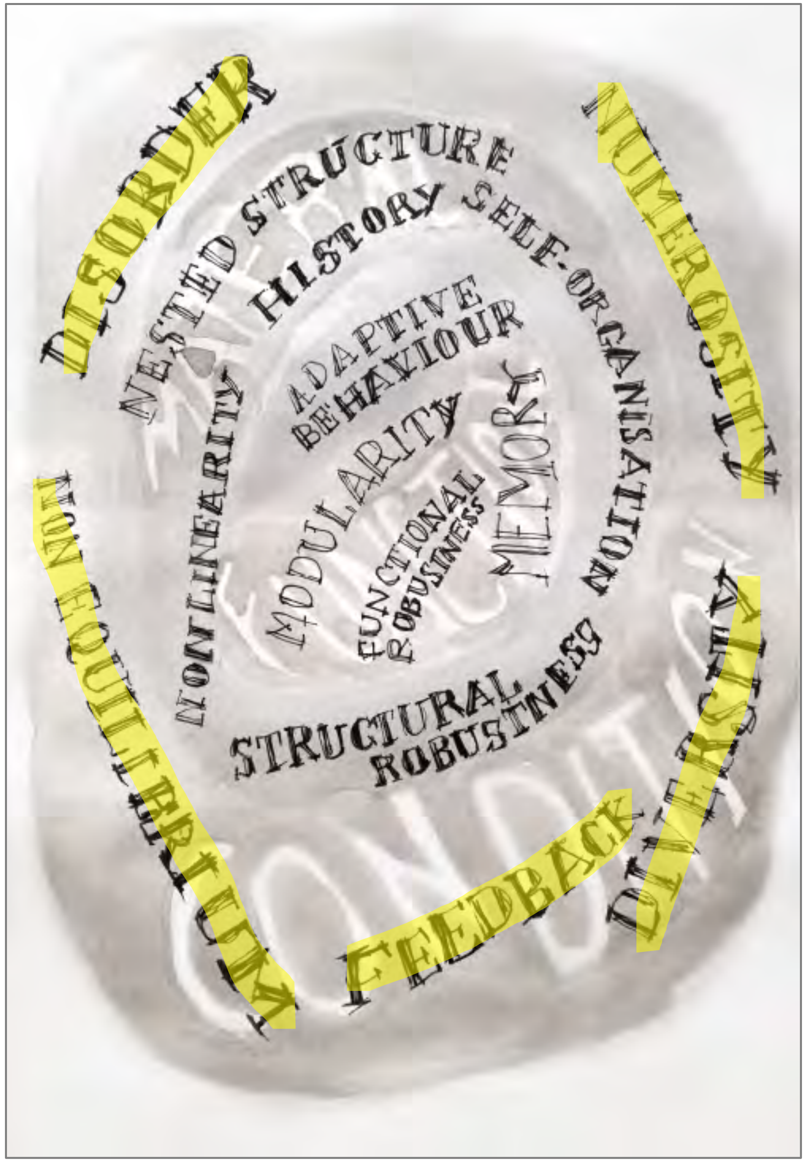


Image by E Bozzarelli

Necessary Conditions

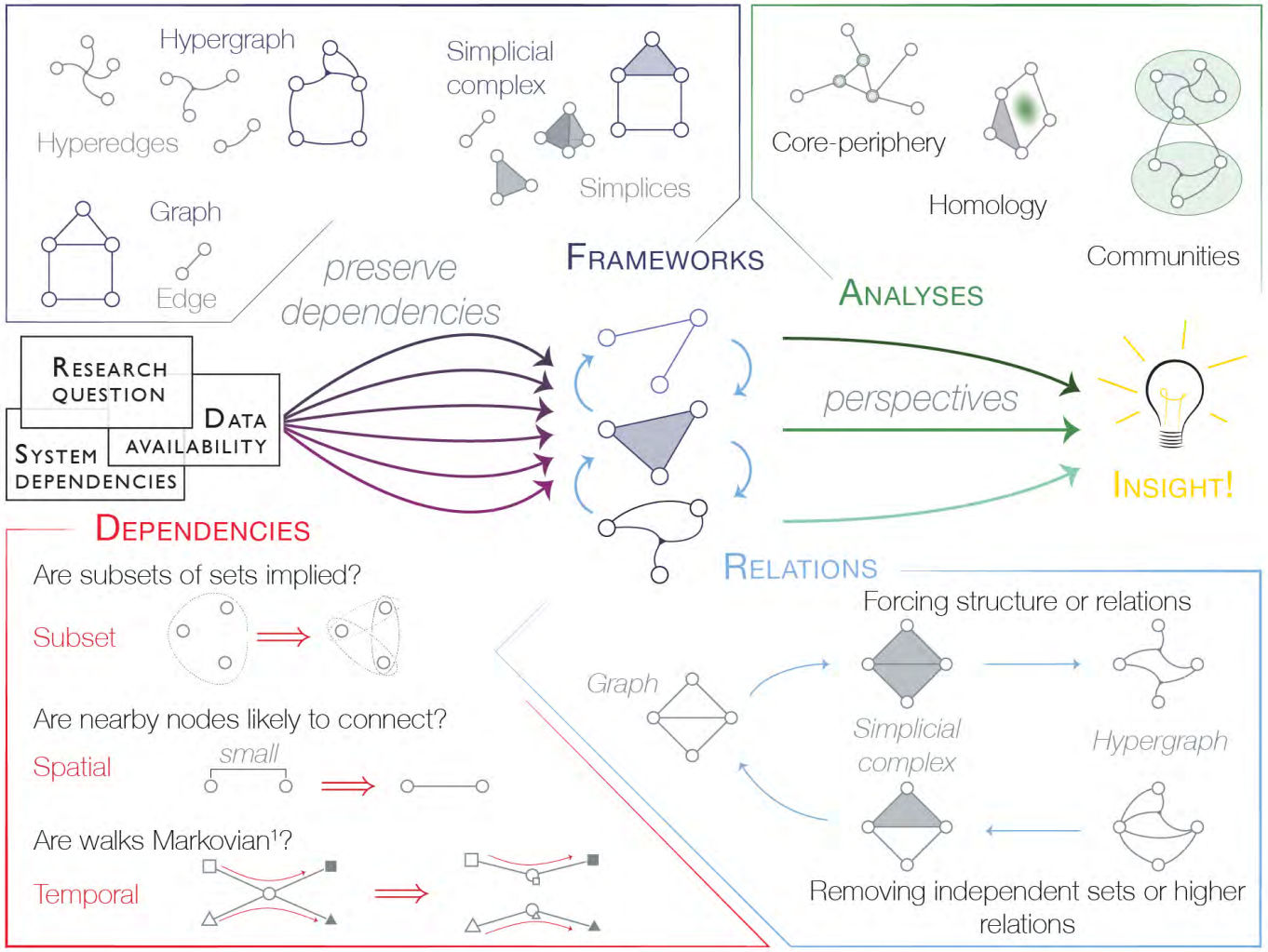


Emergent Features

Functional Features

Image by E Bozzarelli

Representations of Complex Systems



Leo Torres

Ann Sizemore Blevins

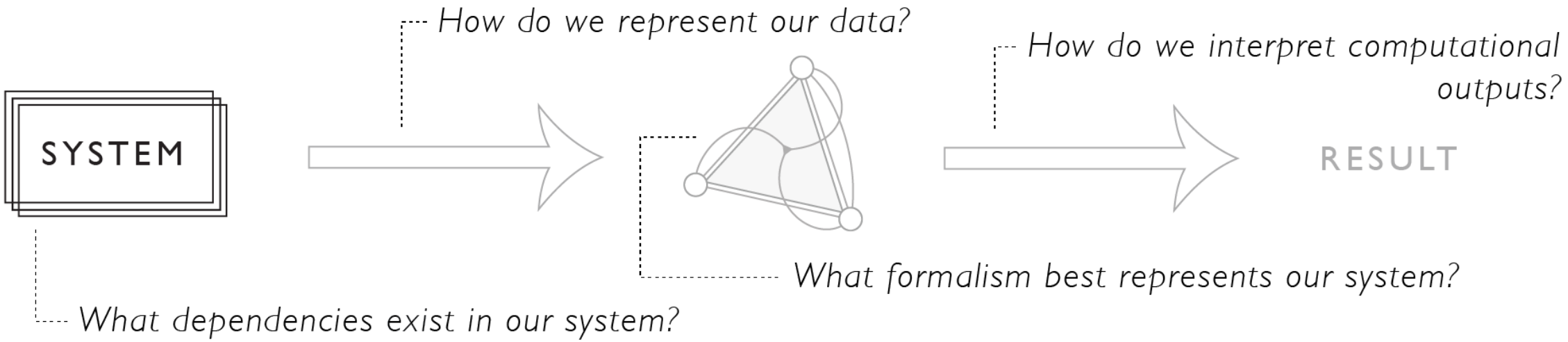
Danielle Bassett

L. Torres, A. Sizemore Blevins, D. S. Bassett, T. Eliassi-Rad: **The Why, How, and When of Representations for Complex Systems**. *SIAM Review* 63(3): 435-485 (2021). https://doi.org/10.1137/20M1355896

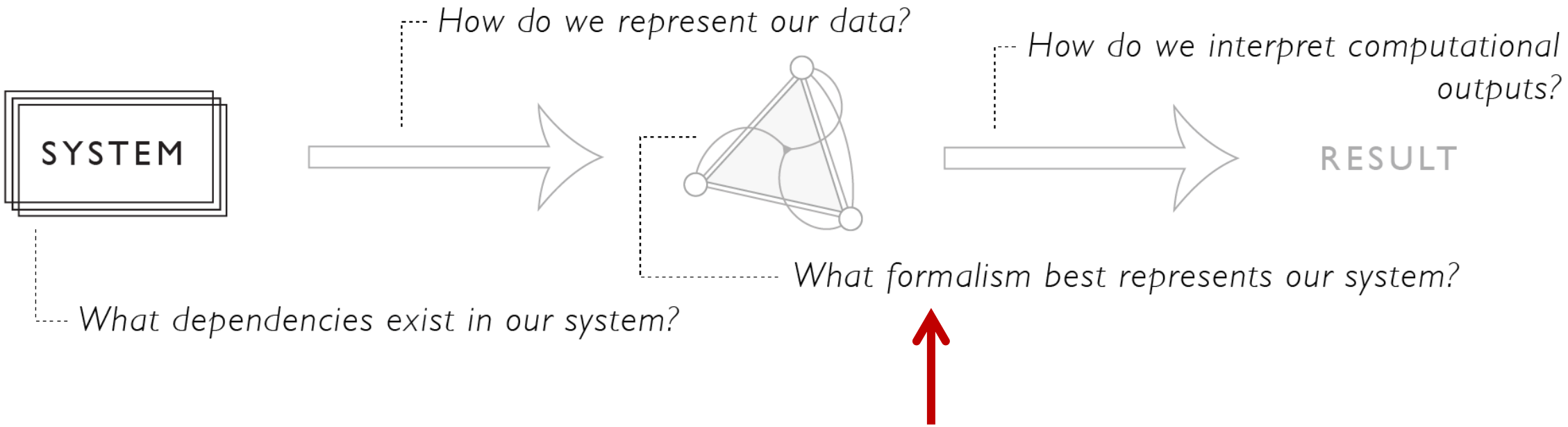
Spoiler ⚠️

- There is no perfect way to analyze a complex system.
- Modeling decisions made when examining a data set from one system are not necessarily …
  - transferable to another system,
  - or even to another data set from the same system.

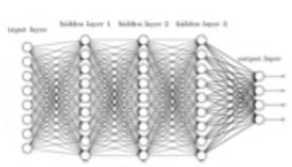
Prototypical analysis pipeline for complex systems



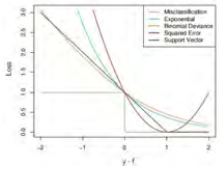
Prototypical analysis pipeline for complex systems



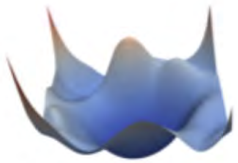
ML enters the pipeline here.



Model



Loss Function

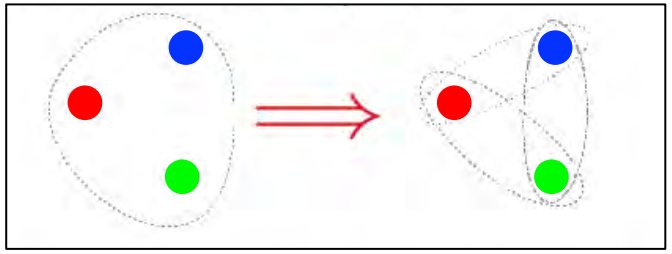


Optimization Algorithm

What dependencies exist in our system?

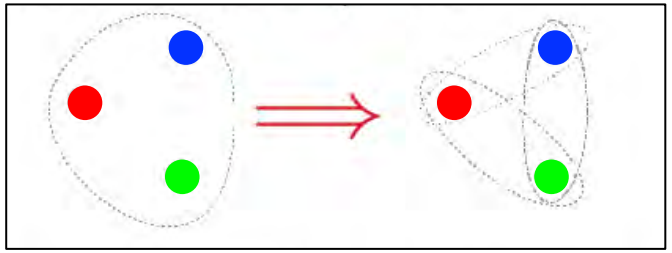
What dependencies exist in our system?

- Subset: Are subsets of sets implied?

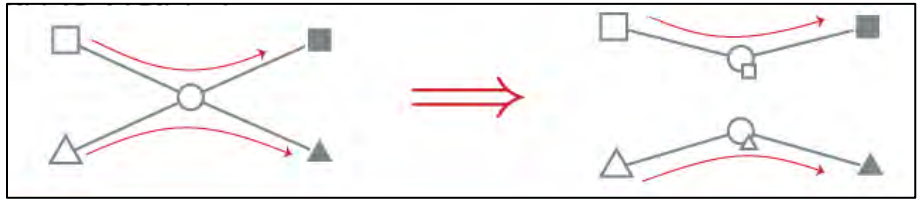


What dependencies exist in our system?

- Subset: Are subsets of sets implied?

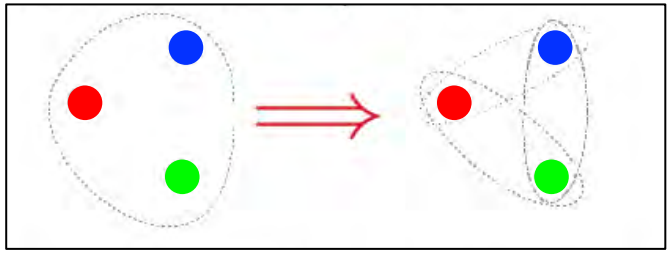


- Temporal: Are walks Markovian?

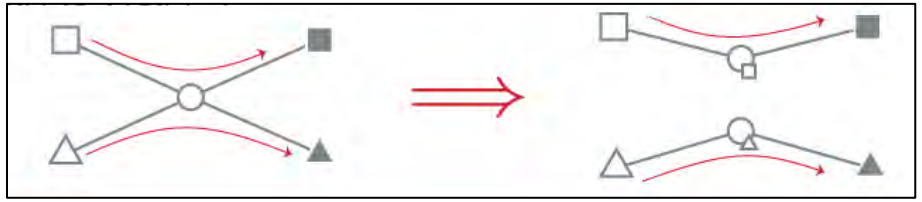


What dependencies exist in our system?

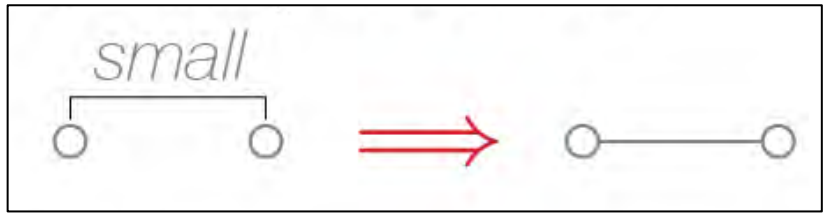
- Subset: Are subsets of sets implied?



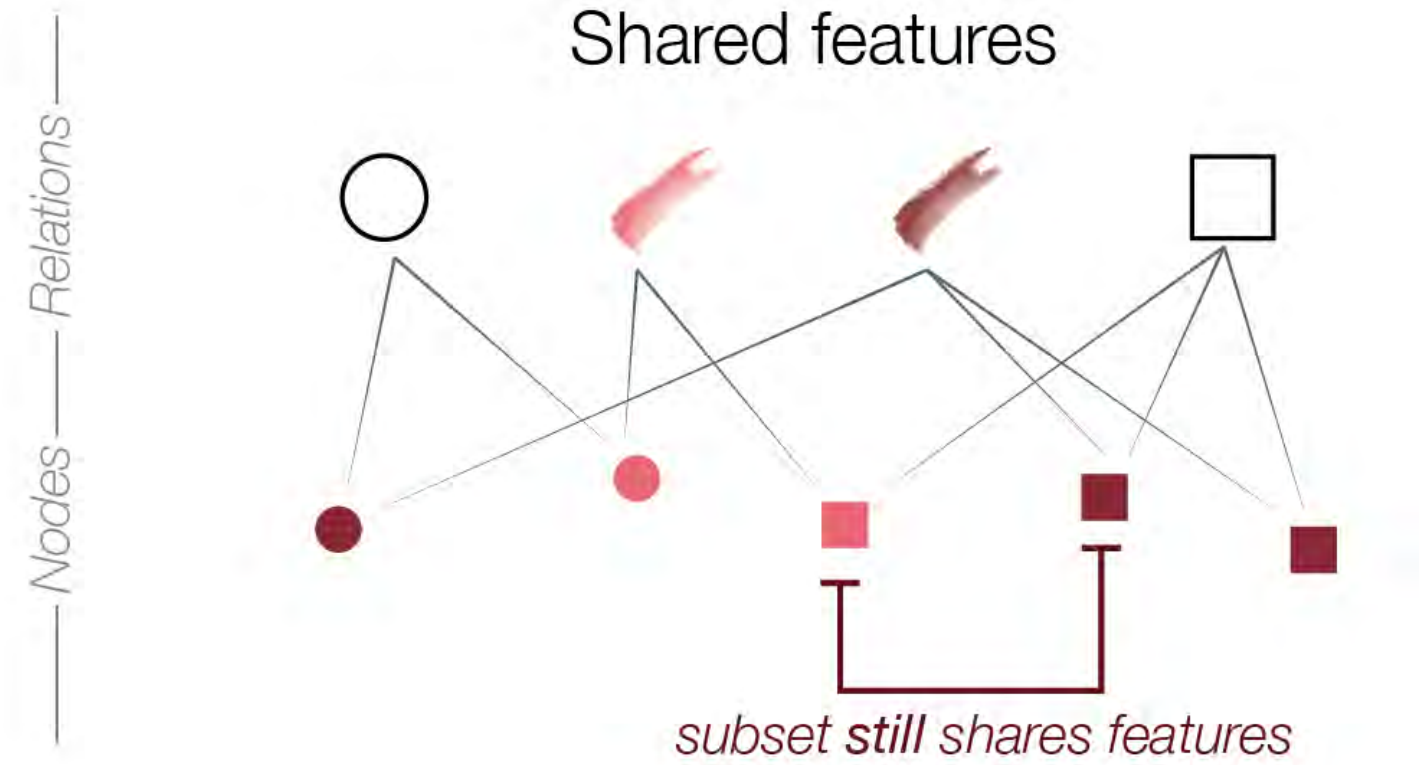
- Temporal: Are walks Markovian?



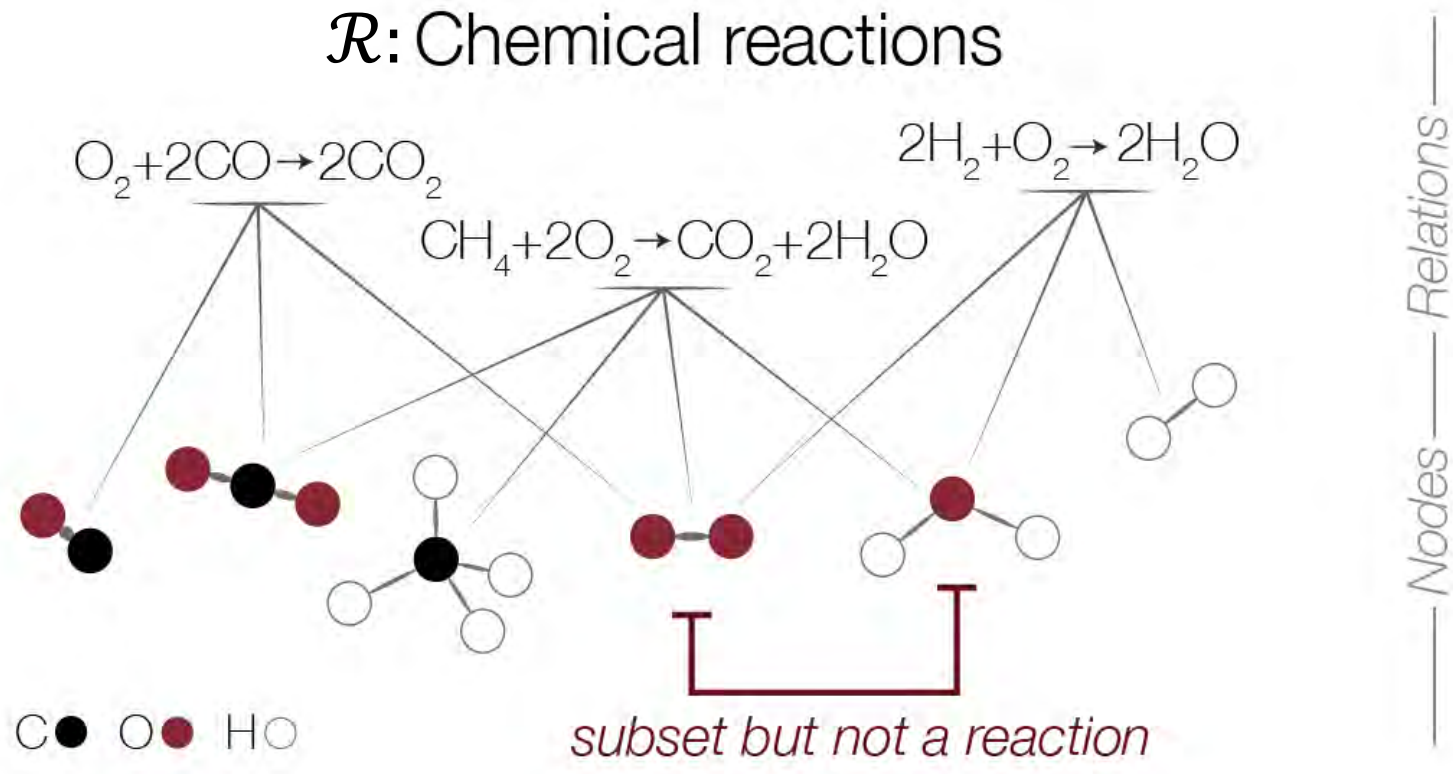
- Spatial: Are nearby nodes likely to connect?



Are subsets of related nodes necessarily related?

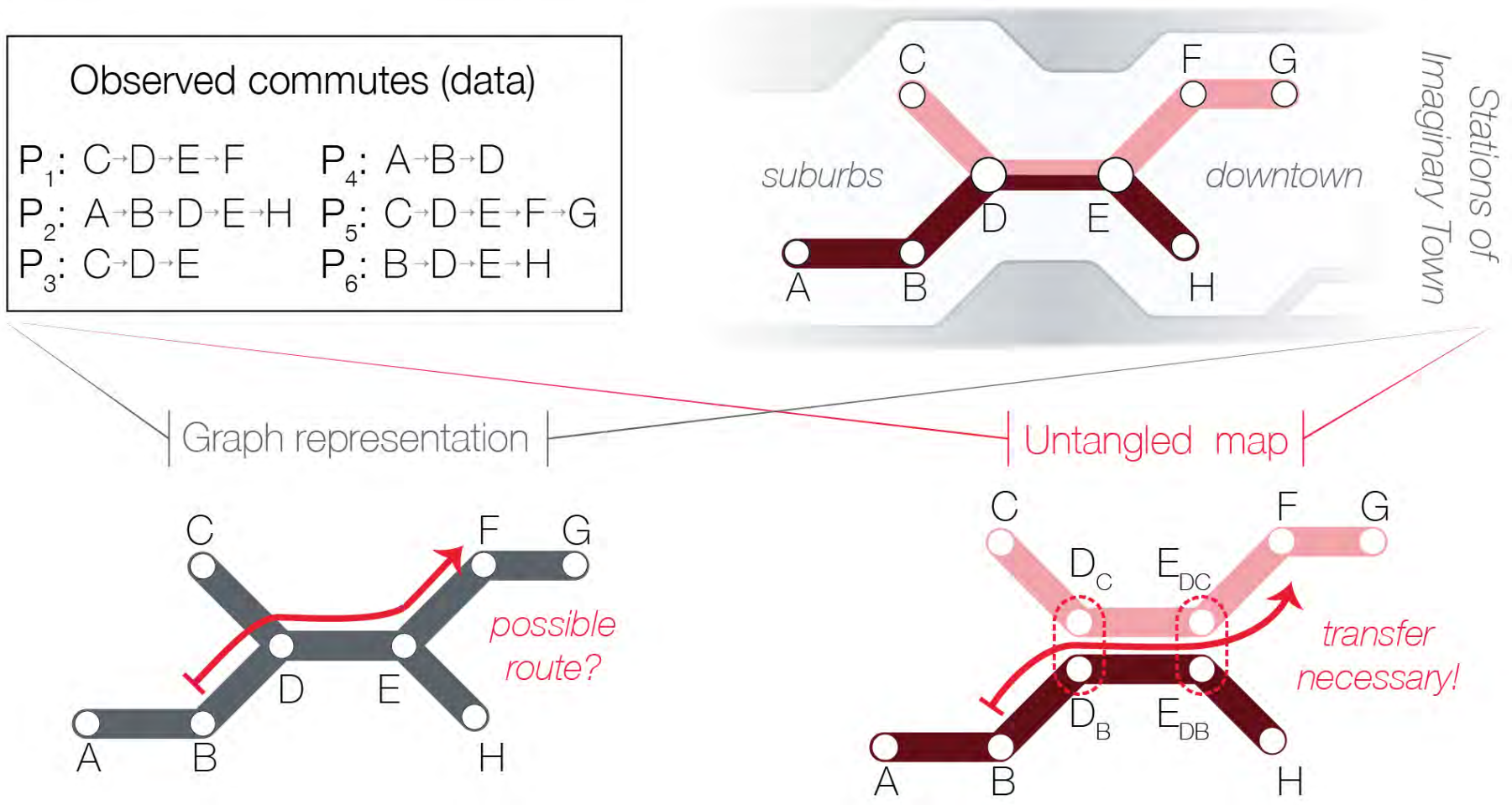


Are subsets of related nodes necessarily related?

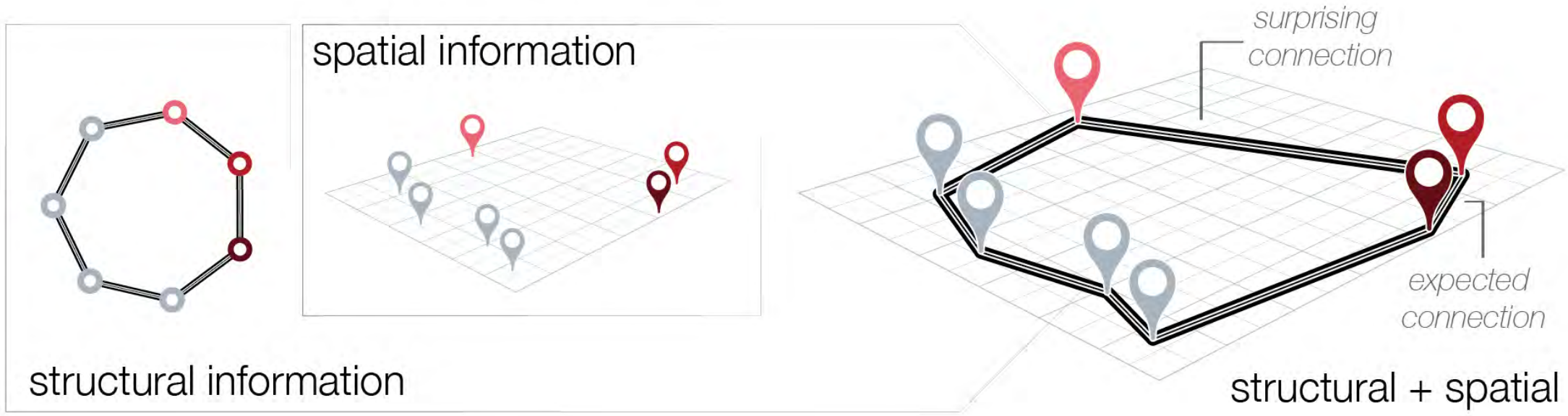


- $\{O_2; H_2O\} \subset \{H_2; O_2; H_2O\}$ **and**
- $\{H_2; O_2; H_2O\} \in \mathcal{R}$ **but**
- $\{O_2; H_2O\} \notin \mathcal{R}$

Will incorporating temporal dependencies yield more accurate representations?



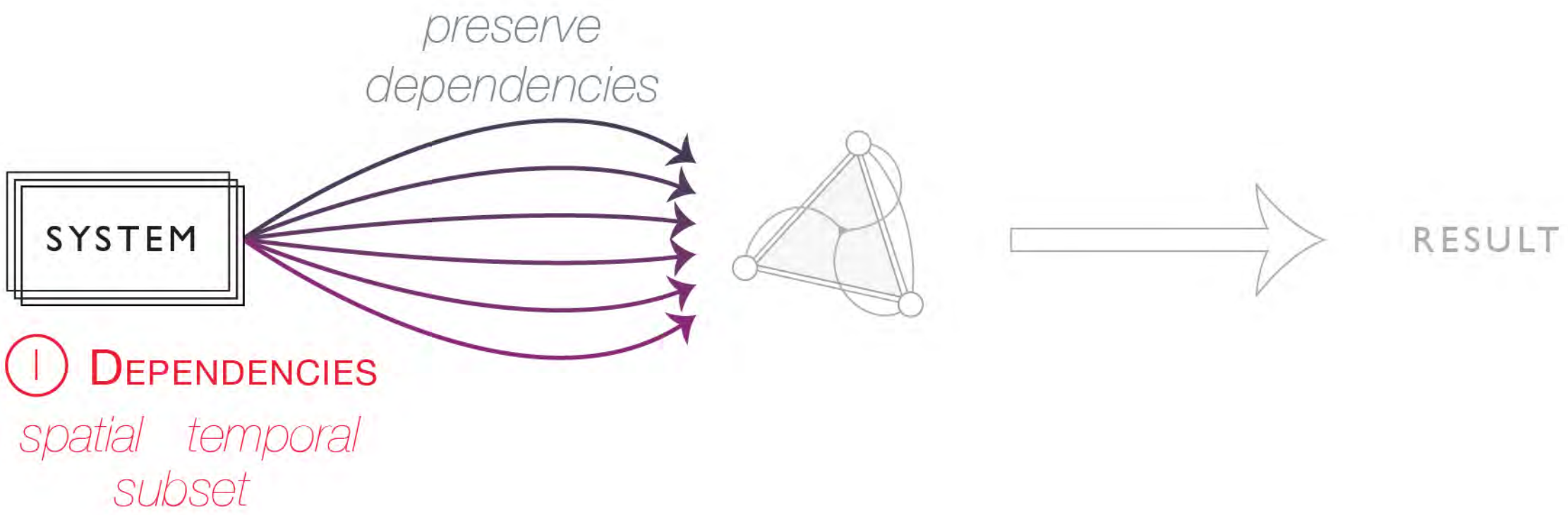
Spatial dependencies within a system can complicate our representations of the data



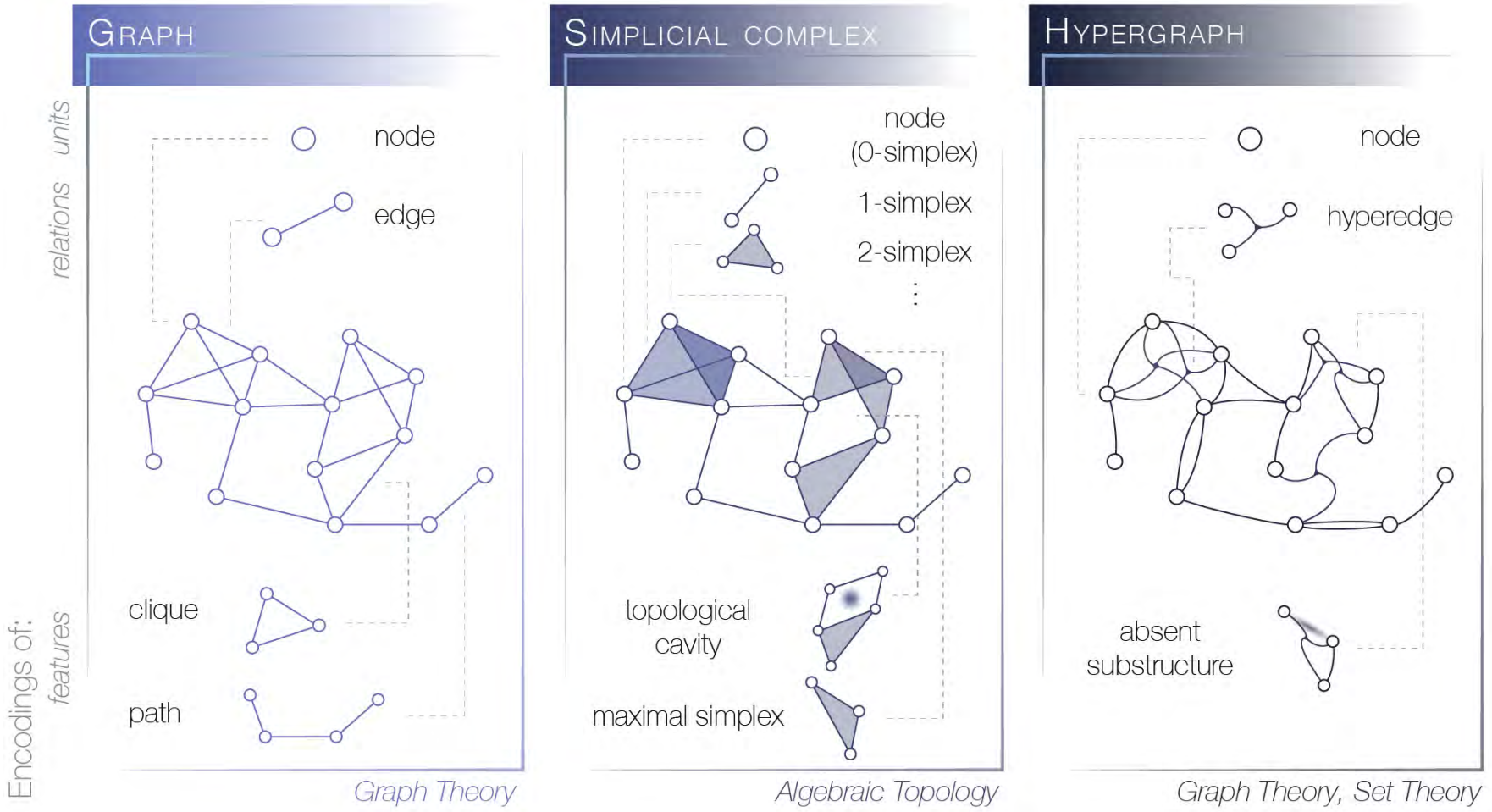
External sources of dependencies

- Data availability
  - However big, your data is incomplete.
  - KDD'19 tutorial on network discovery: https://eliassi.org/kdd19tut.html
- Data acquisition and processing
  - Transitivity dependency in projected bipartite graphs does not show up in unipartite graphs.
  - [Gupte & Eliassi-Rad, WebSci'12]
- Research question
  - The question should influence the definition of a relation.
  - Could a common food have caused a disease outbreak?

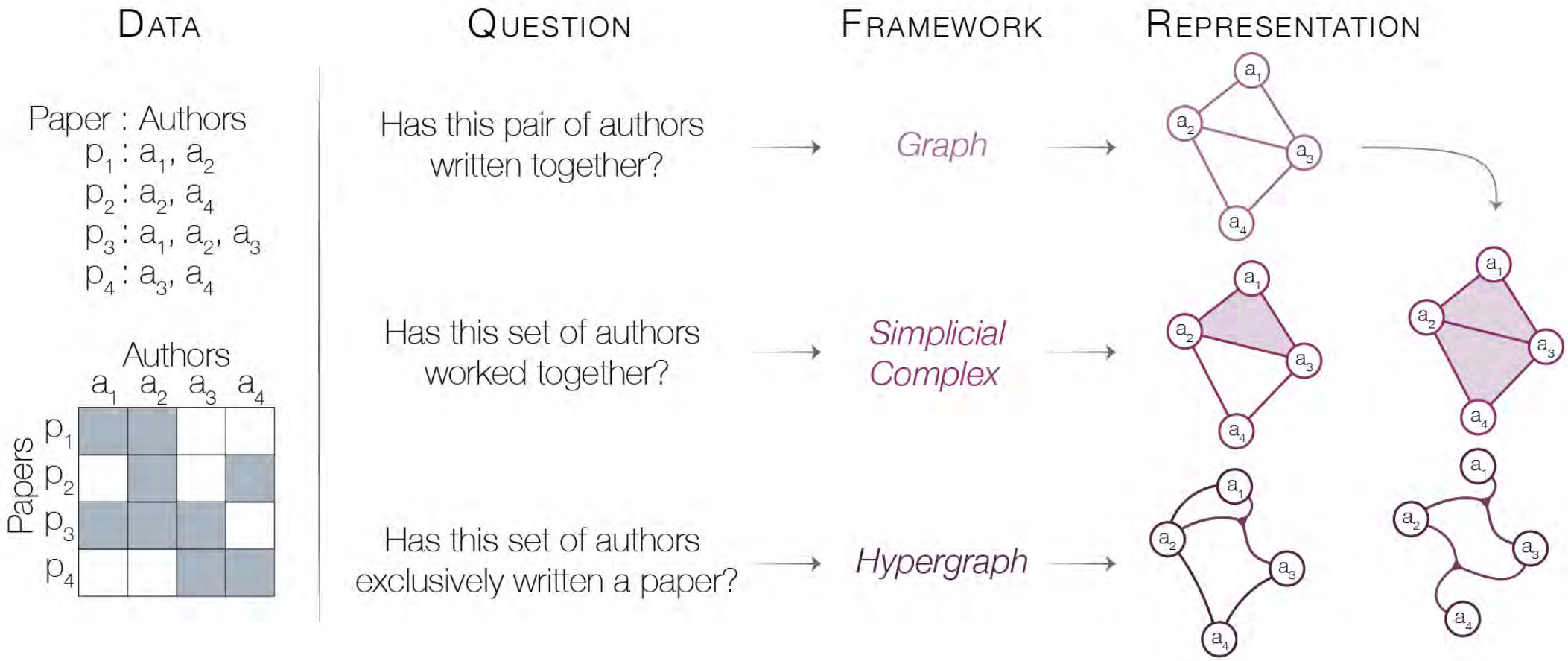
Understanding system dependencies is a first step in the complex system analysis pipeline



Three types of frameworks composed from nodes and relations



Co-authorship data: Each framework offers its own perspective



Tenure-track positions are advantageous positions in society

Samantha Dies

- Getting a tenure-track position depends on many factors
  - School/department: ranking and reputation
  - Advisor(s): professional standing
  - Publications: where and how many
  - Letter writers: professional standing
  - …
- Letter writers (at least in STEM fields) are often co-authors
- We study the ‘who you know’ effect
- **Assumption**: *Researchers who collaborate with prominent faculty may receive unfair advantages when applying for faculty positions*

Dies et al. *Conference on Complex Systems* 2023

‘Who you know’ effect

Samantha Dies

- We study CS faculty placement using graph and hypergraph representations of the temporal co-authorship networks.
- We use GNNs to capture the signals in these networks and predict whether a researcher is hired at a high, medium, or low ranked university based on their co-author network leading up to their hire.
- Preliminary results:
  - There is more signal in the exclusivity that hypergraphs capture than pairwise interactions of a simple graph.
  - Publishing within a tightly-knit community of prominent and productive researchers result in a hiring advantage.

Dies et al. *Conference on Complex Systems* 2023

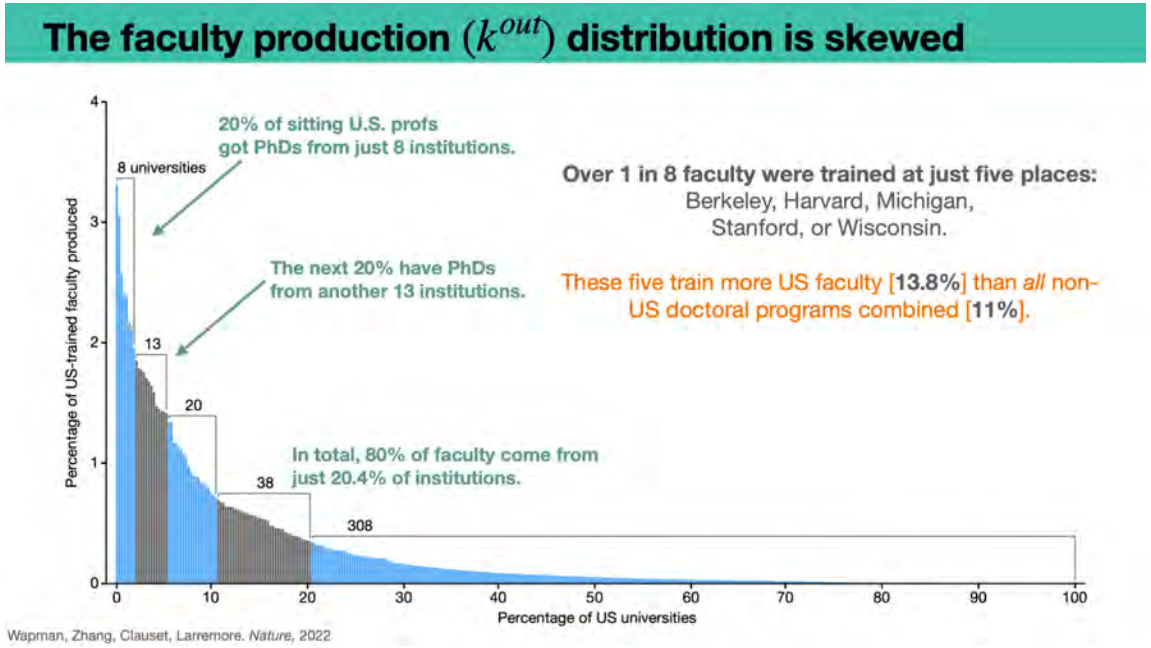
Ongoing work

Interventions

Samantha Dies

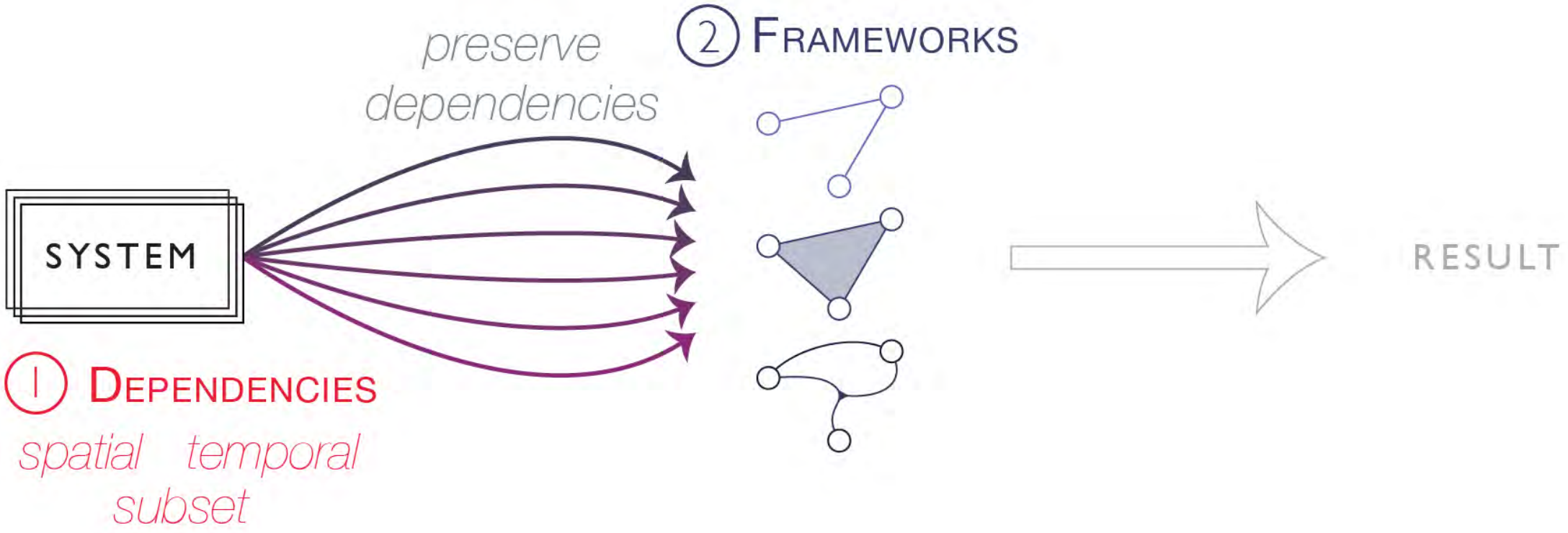
- So what? Tell me something I don't know.

- What are the number and types of additional collaborations which will improve a hiring candidate's likelihood of being hired at a high-ranked university?

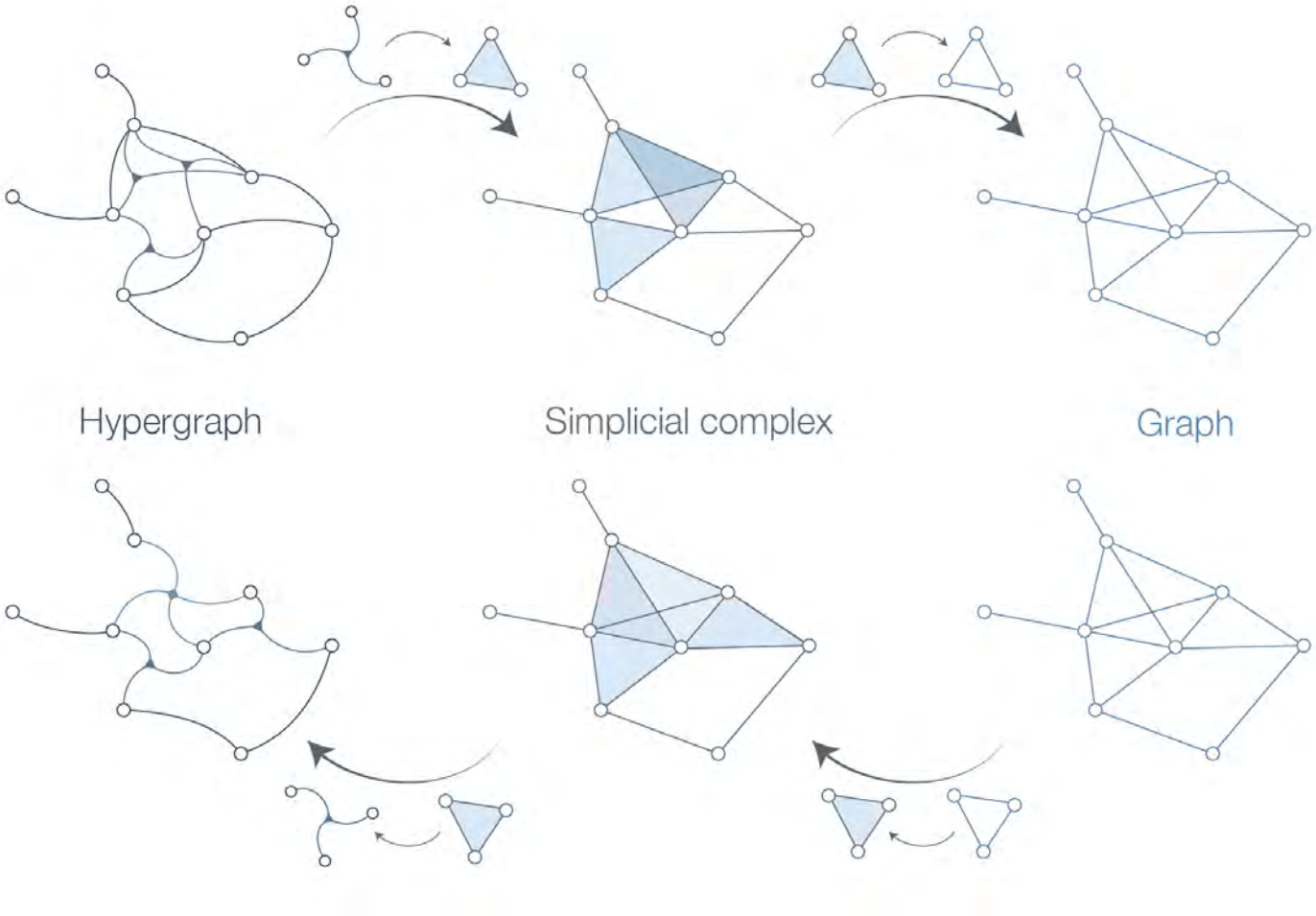


Dies et al. *Conference on Complex Systems* 2023

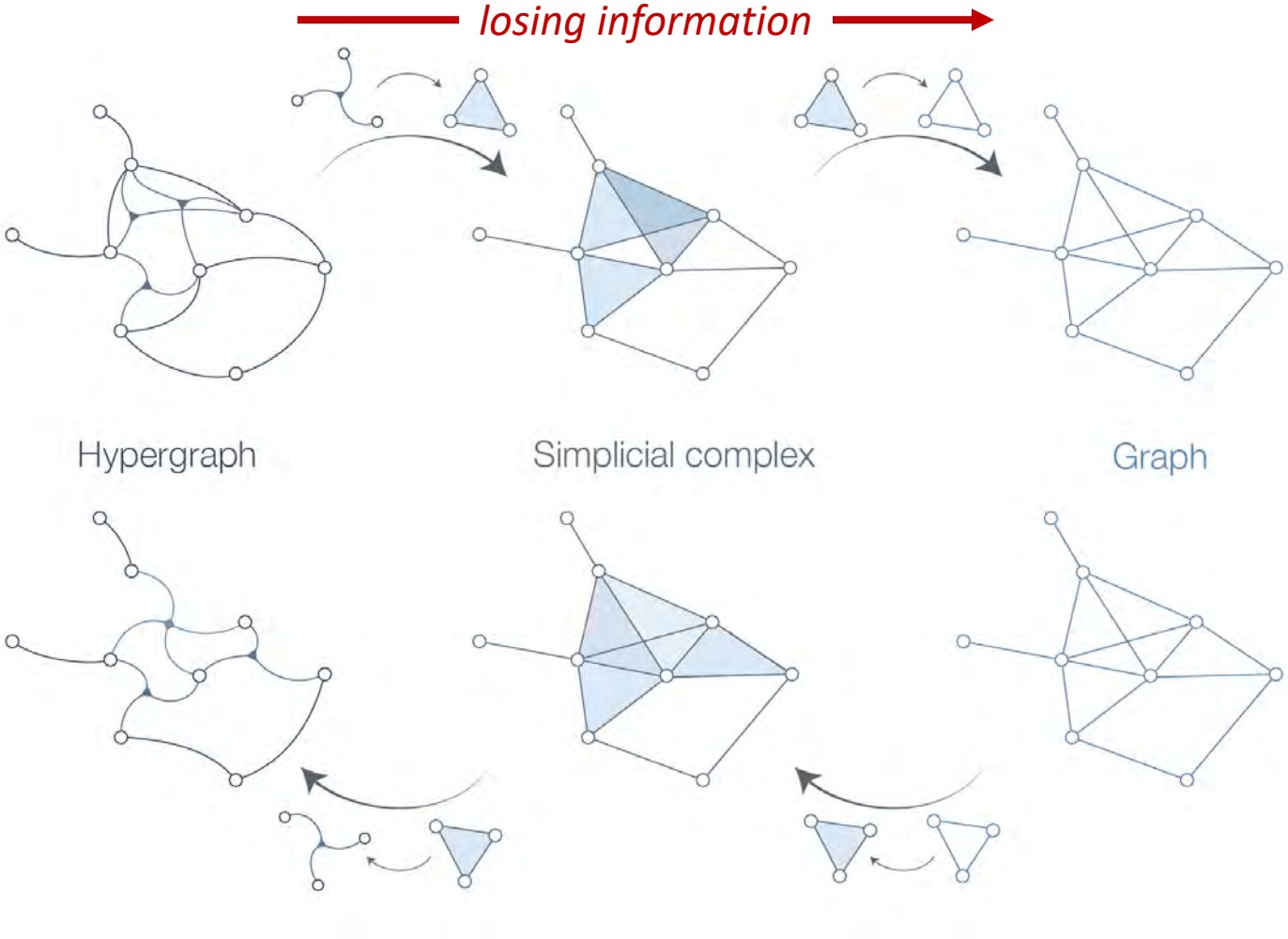
Choosing a framework marks the second step of our analysis pipeline



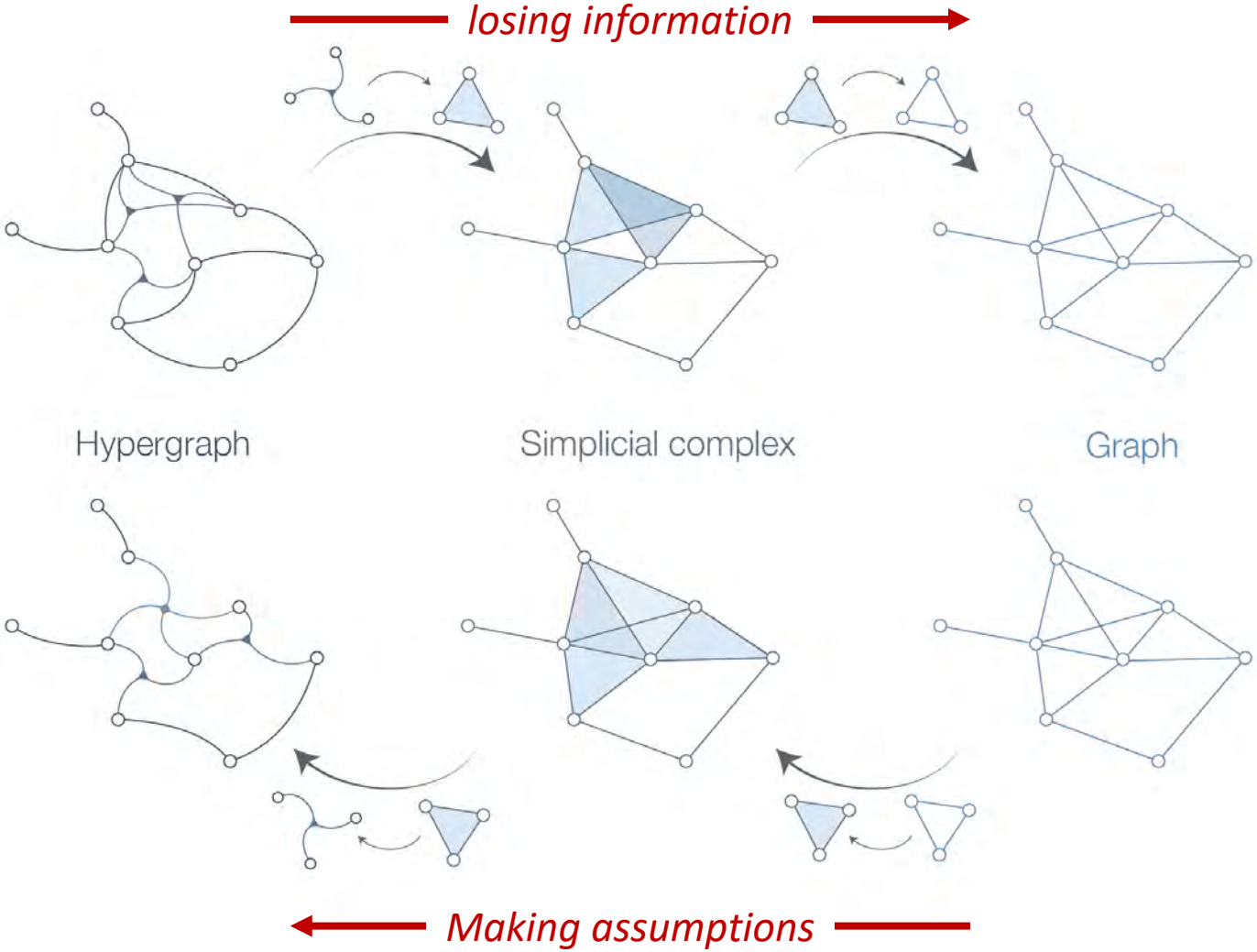
Transitioning between frameworks requires added assumptions or engenders forgetting information



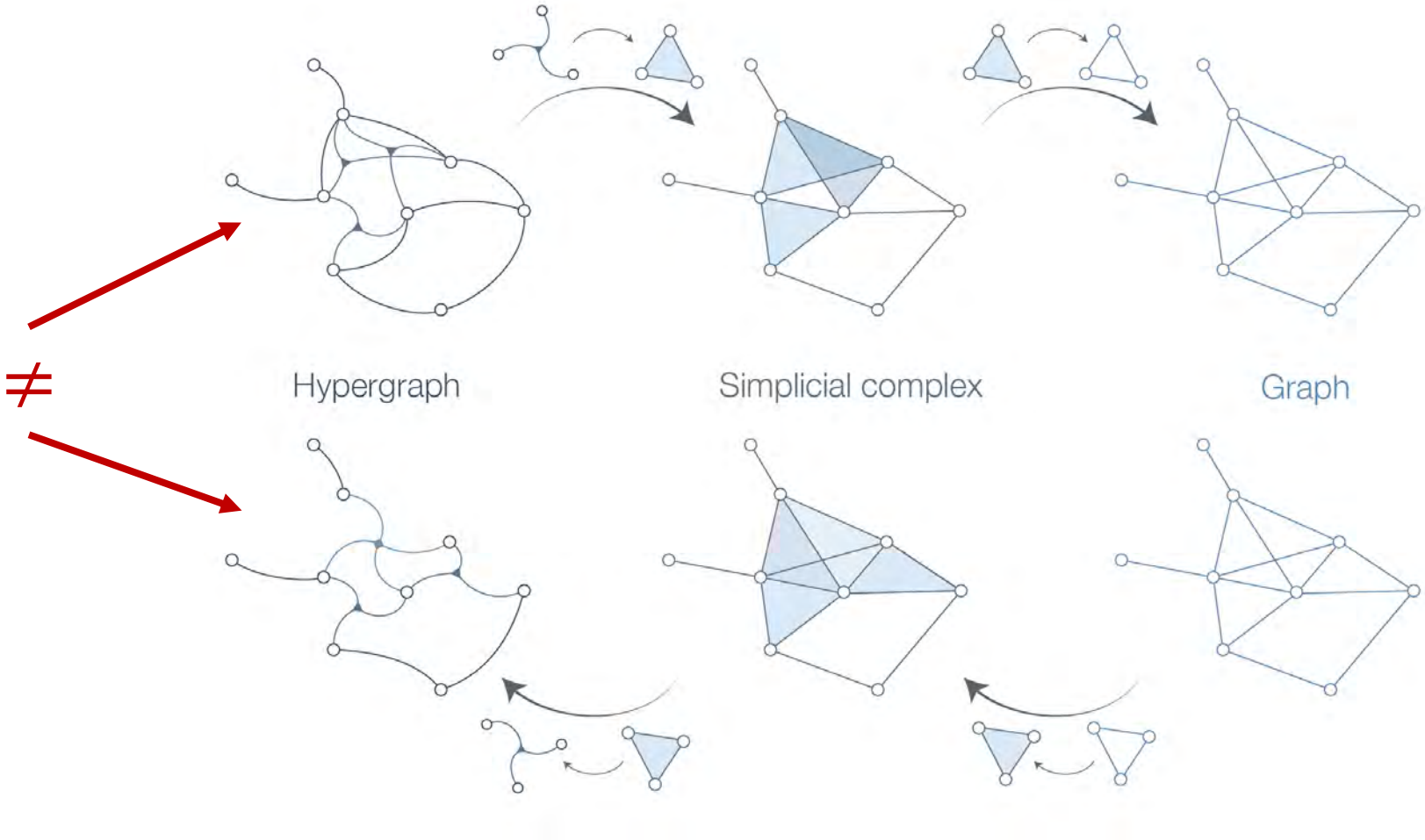
Transitioning between frameworks requires added assumptions or engenders forgetting information



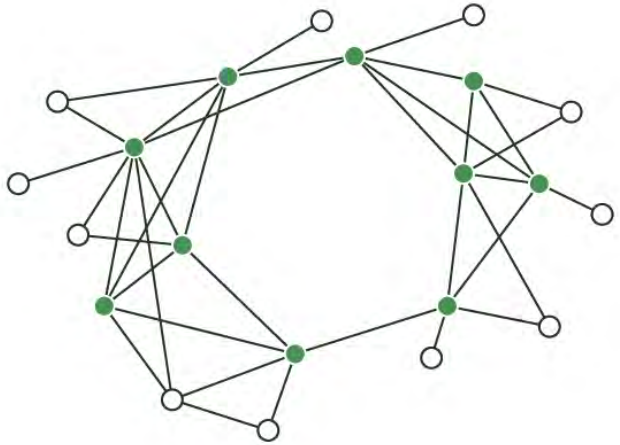
Transitioning between frameworks requires added assumptions or engenders forgetting information



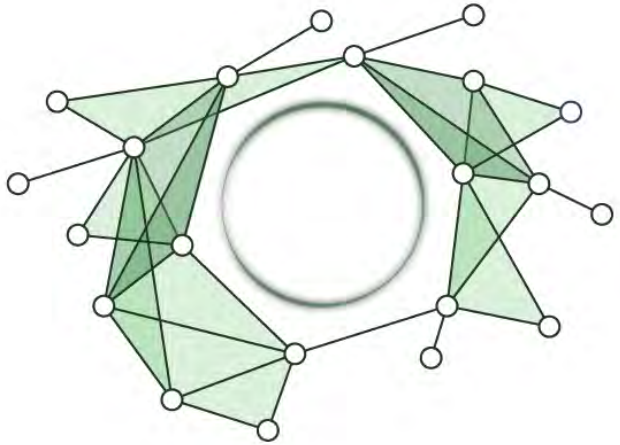
Transitioning between frameworks requires added assumptions or engenders forgetting information



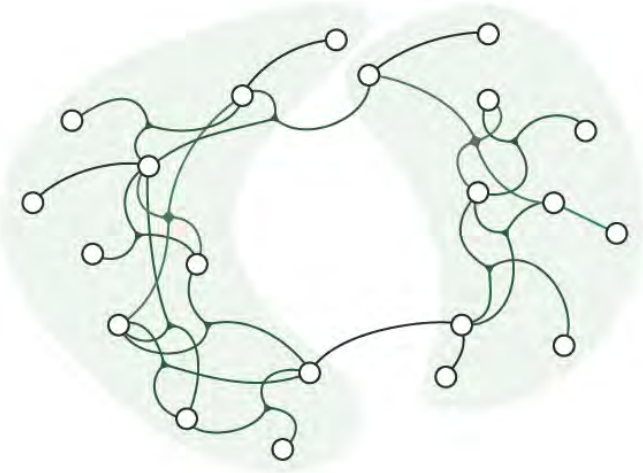
The three frameworks and their analyses can offer different perspectives on a complex system



Core-periphery



Globally circular



Two communities

Variations

- Directed
- Weighted
- Dynamic
- Multiplex
- Multilayer
- Higher-order

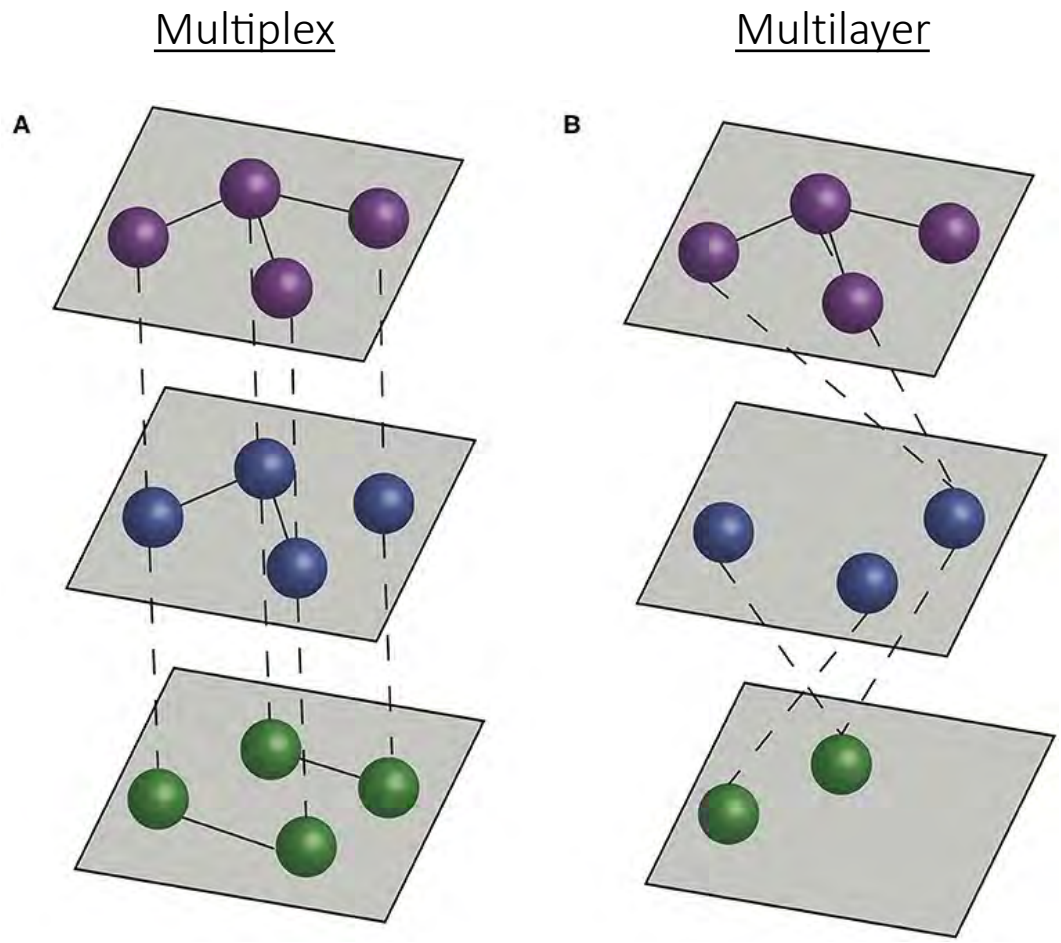


Image from Kinsley AC, Rossi G, Silk MJ and VanderWaal K (2020) Multilayer and Multiplex Networks: An Introduction to Their Use in Veterinary Epidemiology. *Frontiers in Veterinary Science* 7:596. doi: 10.3389/fvets.2020.00596

Some of our efforts

- **Talent is everywhere**
  - Allen, Eliassi-Rad, Fitelson. *The Future of Affirmative Action*, 2014
  - Liu et al. *AI, Ethics, and Society* 2021
- **Democratic backsliding**
  - Wiesner et al. *European J. of Physics* 2019
  - Eliassi-Rad et al. *Humanities & Social Sciences Communications* 2020
- **Information access equality**
  - Wang, Varol, Eliassi-Rad. *Applied Network Science* 2022
- **COVID-19 & the US criminal justice systems**
  - Klein et al. *Nature* 2023
- **Interventions on academic hiring**
  - Dies et al. *Conference on Complex Systems* 2023
- **Using sequences of life-events to predict human lives**
  - Germans Savcisens, et al. *Nature Computational Science* 2023

Some of our efforts

- Talent is everywhere
  - Allen, Eliassi-Rad, Fitelson. *The Future of Affirmative Action*, 2014
  - Liu et al. *AI, Ethics, and Society* 2021
- Democratic backsliding
  - Wiesner et al. *European J. of Physics* 2019
  - Eliassi-Rad et al. *Humanities & Social Sciences Communications* 2020
- Information access equality
  - Wang, Varol, Eliassi-Rad. *Applied Network Science* 2022
- COVID-19 & the US criminal justice systems
  - Klein et al. *Nature* 2023
- Interventions on academic hiring
  - Dies et al. *Conference on Complex Systems* 2023
- Using sequences of life-events to predict human lives
  - Germans Savcisens, et al. *Nature Computational Science* 2023

Some of our efforts

- Talent is everywhere
  - Allen, Eliassi-Rad, Fitelson. *The Future of Affirmative Action*, 2014
  - Liu et al. *AI, Ethics, and Society* 2021
- Democratic backsliding
  - Wiesner et al. *European J. of Physics* 2019
  - Eliassi-Rad et al. *Humanities & Social Sciences Communications* 2020
- Information access equality
  - Wang, Varol, Eliassi-Rad. *Applied Network Science* 2022
- COVID-19 & the US criminal justice systems
  - Klein et al. *Nature* 2023
- Interventions on academic hiring
  - Dies et al. *Conference on Complex Systems* 2023
- Using sequences of life-events to predict human lives
  - Germans Savcisens, et al. *Nature Computational Science* 2023

Democracy in the balance

September 2020

- According to The Economist (13 November 2023), 76 countries will hold elections in 2024.
- That is equivalent to more than half of the world's population going to the polls in 2024.
- EIU has a Democracy Index for 71 of the 76 countries.
  - Of these, 43 countries (more than 60%) have free and fair elections.
  - In the other 28 countries, this is not the case.

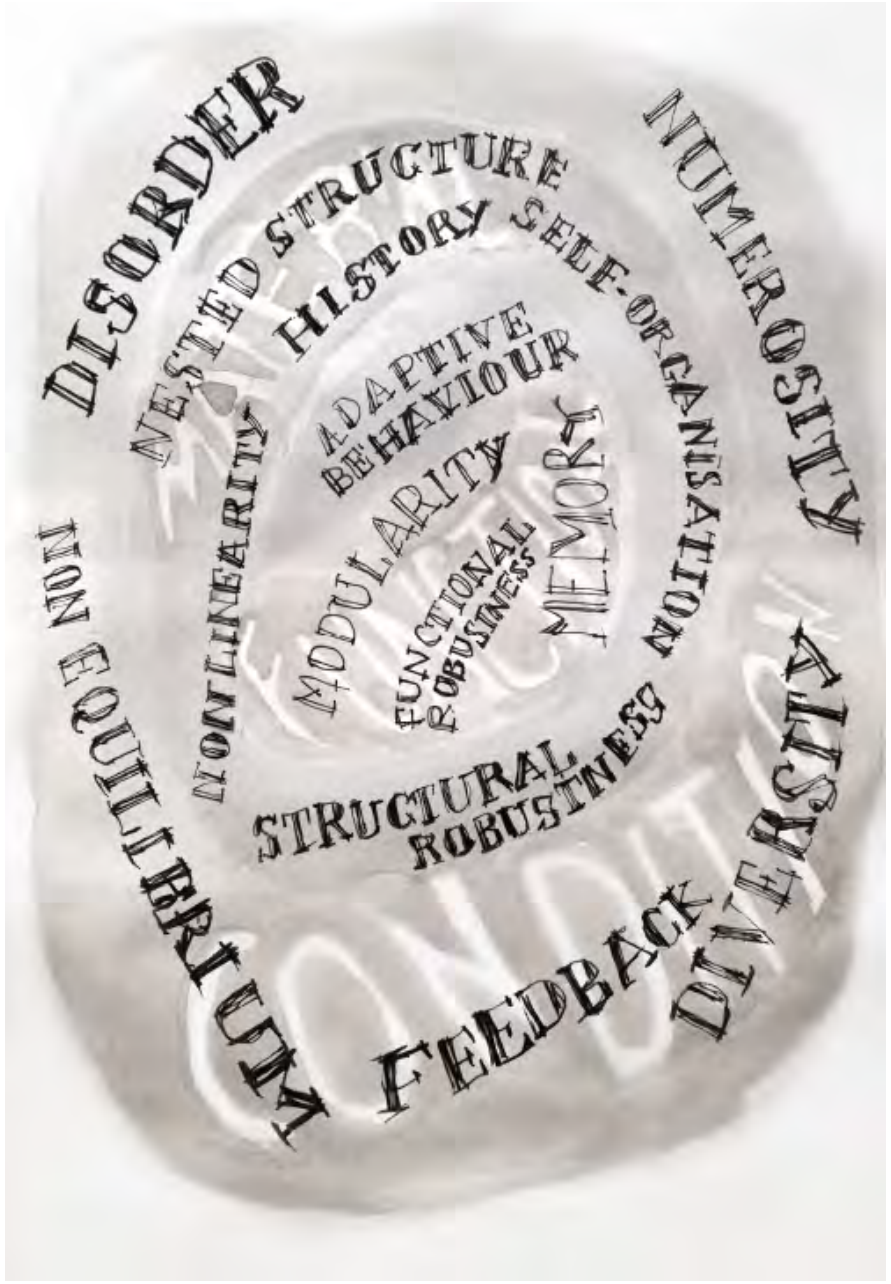


Image by E Bozzarelli

Democracy has all the features of a complex system.

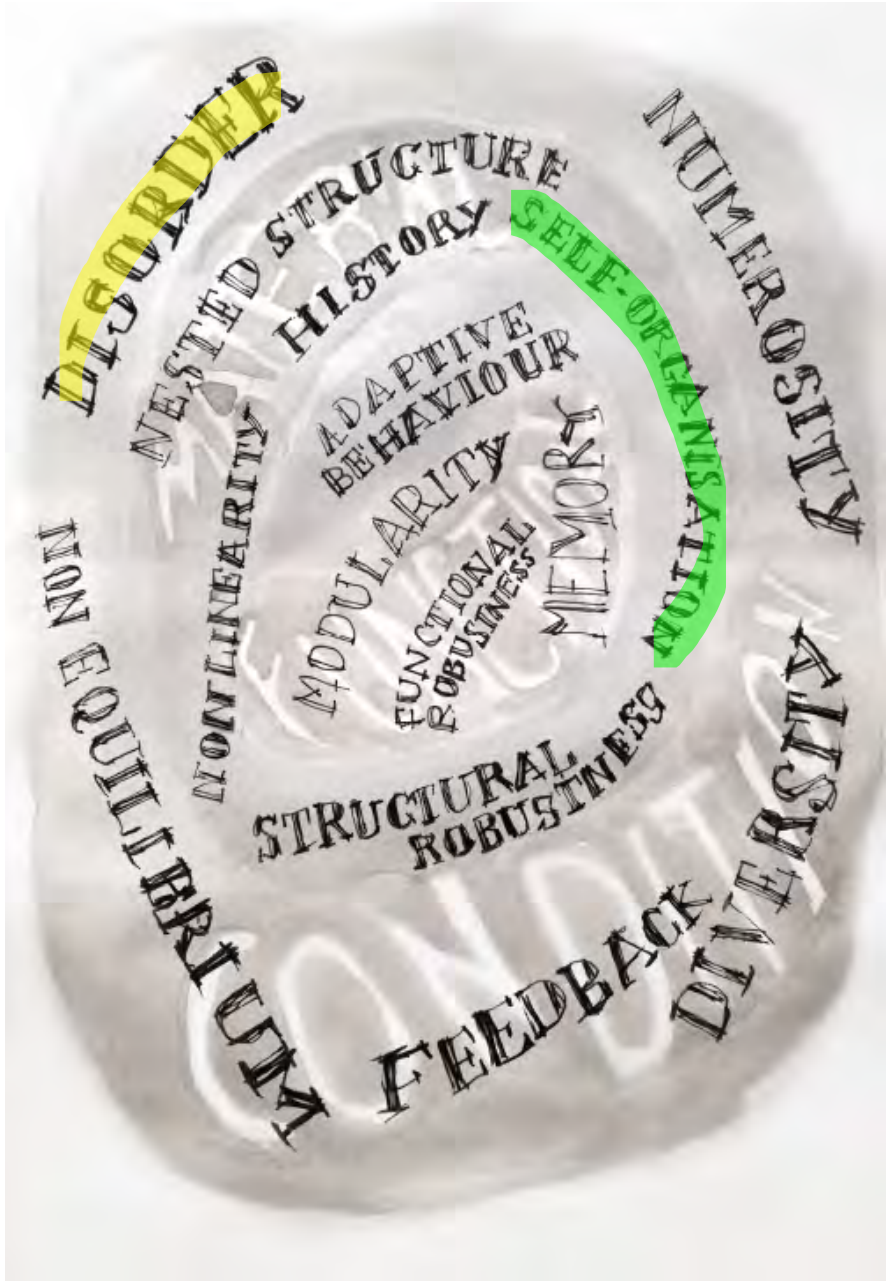


Image by E Bozzarelli

- Randomness of interactions is important to self-organization
- Democracy requires an unstructured exchange of opinions and ideas between citizens

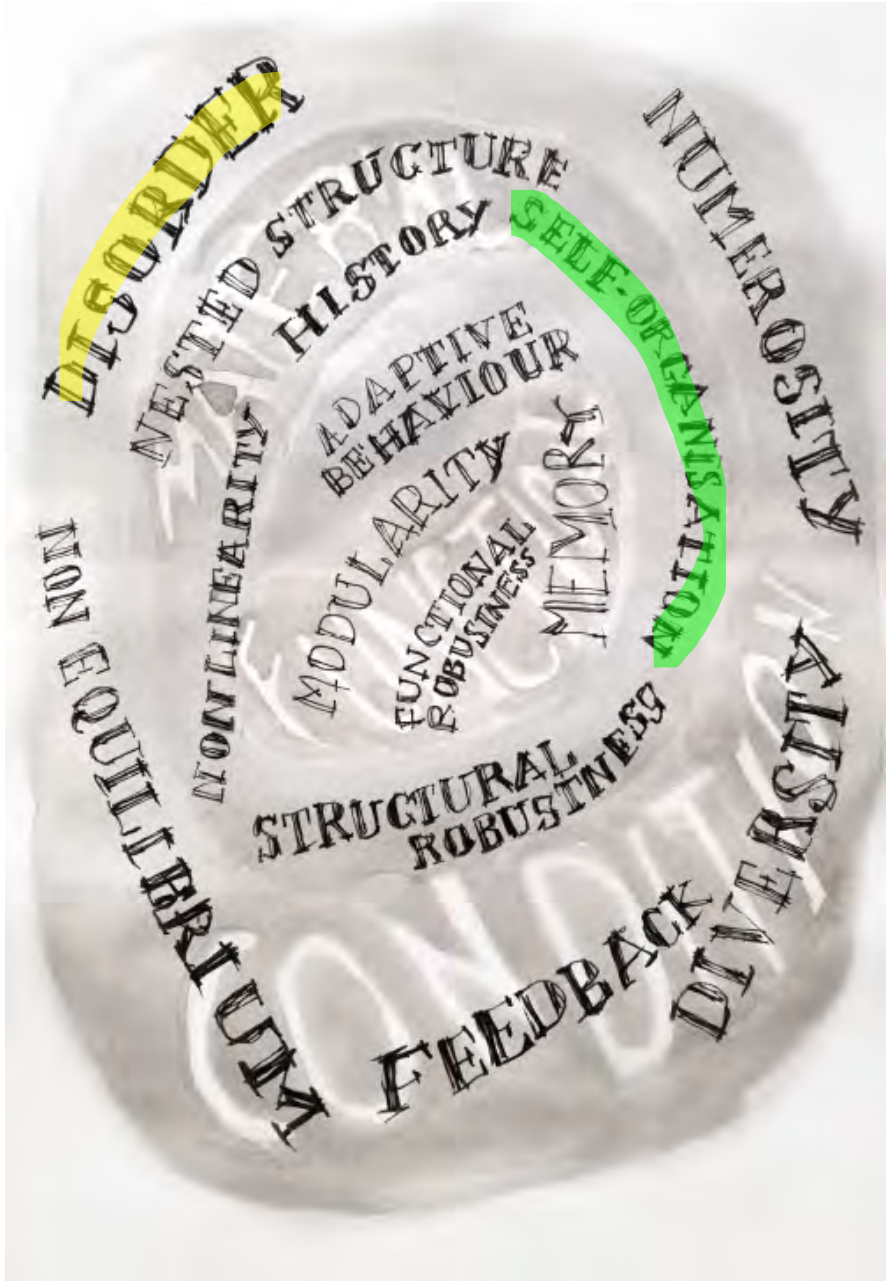


Image by E Bozzarelli

- Randomness of interactions is important to self-organization
- Democracy requires an unstructured exchange of opinions and ideas between citizens
- Caveat: Instability arises when randomness increases beyond a critical level
- Instability: chaos and collapse in terms of consensual norms

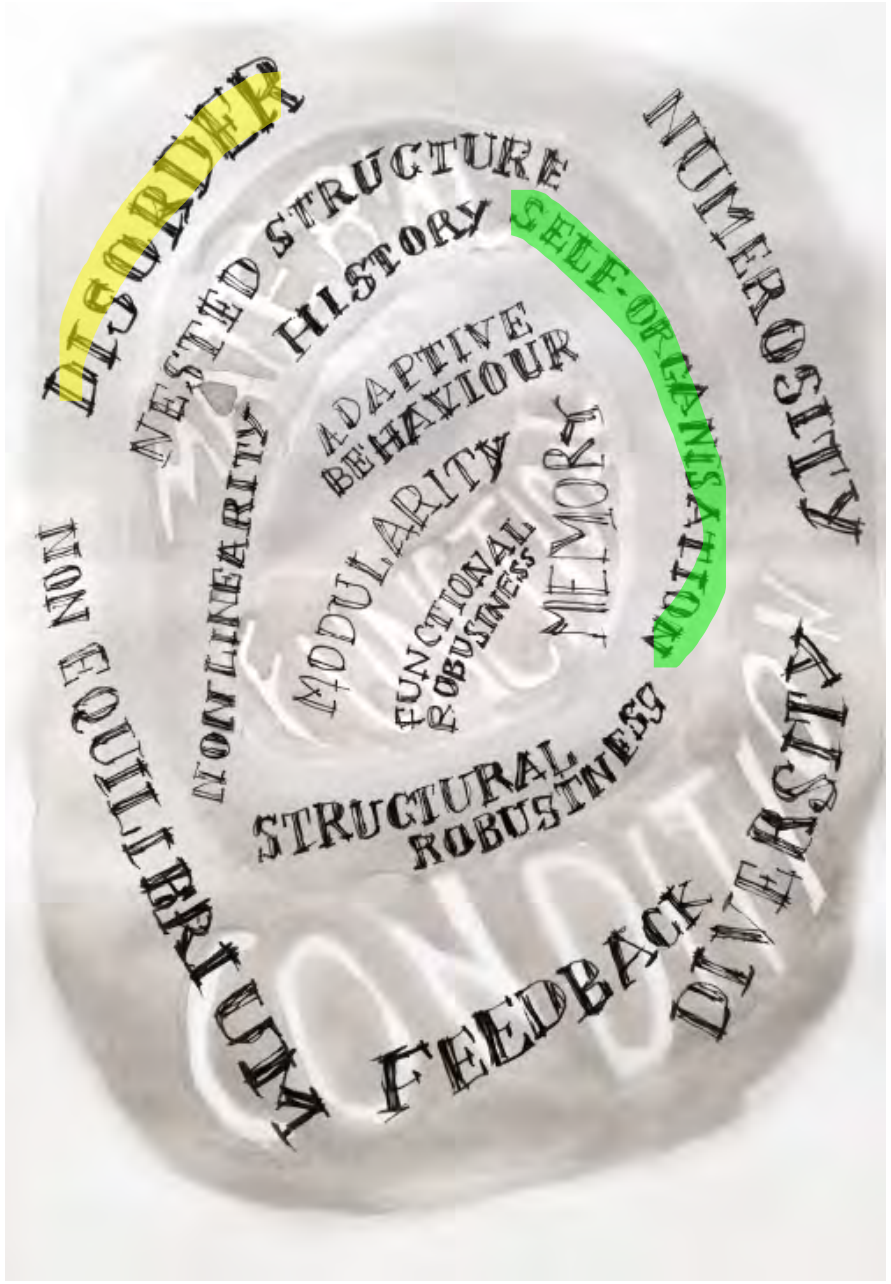
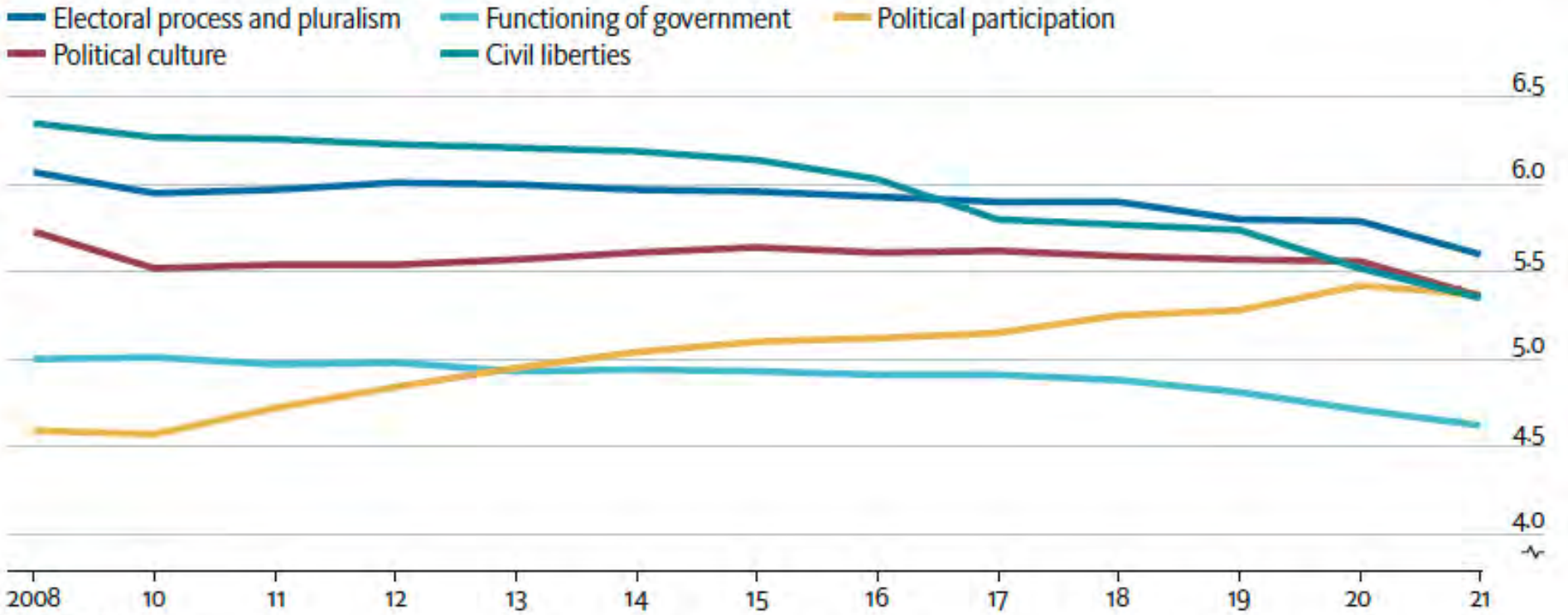


Image by E Bozzarelli

- Instability: chaos and collapse in terms of consensual norms
- "… propels people into accommodating "lying demagogues" because their brazen lies signal opposition to the disdained establishment."
  Hahl et al. (2018)

Democratic backsliding

Evolution of democracy by category, 2008-21
(Index score out of 10, 10 being best)



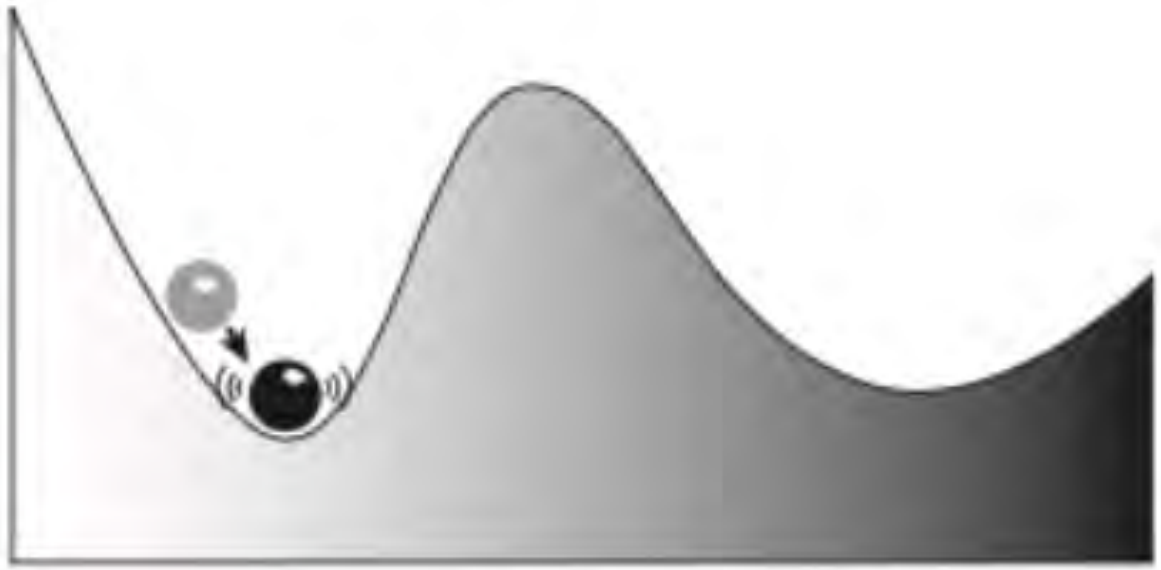
Source: EIU.

Democratic backsliding is, in part, the result of instability in the democratic system.

AI technology that amplifies misinformation and disinformation increases instability in democracy.

Stability

(a)



🙂

(b)

😕

For a detailed discussion of stability, see https://bit.ly/3gFfRzI

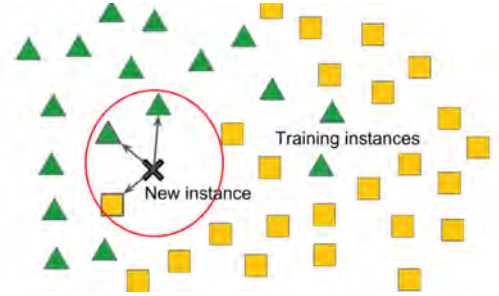
(Temporal) Stability vs. (Modal) Robustness

Stability

- Persistence over time
- Resistance to perturbation or change
- Example: a rock on top of a peak is not stable

Robustness

- Insensitivity or independence of the behavior of the system to changes in possible microscopic realization
- Example: *k*NN is not robust



For a detailed discussion of stability, see https://bit.ly/3gFfRzI

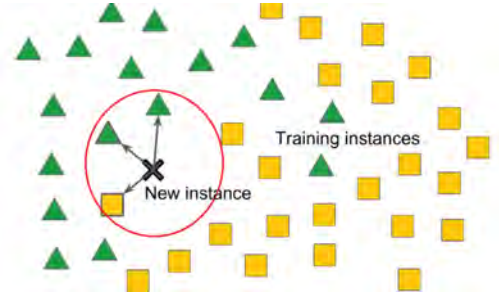
(Temporal) Stability vs. (Modal) Robustness

Stability

- Persistence over time
- Resistance to perturbation or change
- Example: a rock on top of a peak is not stable

Robustness

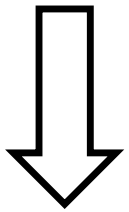
- Insensitivity or independence of the behavior of the system to changes in possible microscopic realization
- Example: *k*NN is not robust



A stable system need not be robust nor a robust system stable.

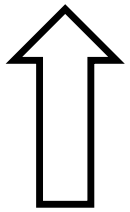
Understanding democracy as a complex system enables us to make policy recommendations to counteract democratic backsliding (including ones caused by AI technology).

Policy recommendations



Top-down

1. Entrench diversity by regulation
2. Monitor feedback
3. Ensure connectivity



Bottom-up

1. Recruit credible communicators in estranged regions
2. Recognize limits to message control
3. Emphasize persistence and limits to forecasting

https://rdcu.be/b5yne *(February 2020)*

Disclaimer on policy recommendations

- **Our recommendations do not form a hierarchy**
- **Their effectiveness is context dependent**
  - The way in which these recommendations can be put into practice will differ from country to country
- **Our recommendations are also not independent of each other**
  - Example: failure in recruiting credible communicators increases the likelihood of failure in recognizing limits of message control
- **Democracy is an evolving project**

https://rdcu.be/b5yne *(February 2020)*

OPEN ACCESS
IOP Publishing
European Journal of Physics
Eur. J. Phys. 40 (2019) 014002 (17pp)
https://doi.org/10.1088/1361-6404/aaeb4d

Stability of democracies: a complex systems perspective

K Wiesner[1,2], A Birdi[3], T Eliassi-Rad[4], H Farrell[5], D Garcia[2,6], S Lewandowsky[7], P Palacios[8], D Ross[9,10,11], D Sornette[12] and K Thébault[13]

[1] School of Mathematics, University of Bristol, United Kingdom
[2] Complexity Science Hub Vienna, Austria
[3] Department of Economics, University of Bristol, United Kingdom
[4] Network Science Institute, College of Computer and Information Science, Northeastern University, Boston, United States of America
[5] Department of Political Science, George Washington University, Washington, D.C., United States of America
[6] Section for Science of Complex Systems, Center for Medical Statistics, Informatics, and Intelligent Systems, Medical University of Vienna, Austria
[7] School of Psychological Science, University of Bristol, United Kingdom
[8] Department of Philosophy, University of Salzburg, Austria
[9] School of Sociology, Philosophy, Criminology, Government, and Politics, University College Cork, Ireland
[10] School of Economics, University of Cape Town, South Africa
[11] Center for Economic Analysis of Risk, Georgia State University, United States of America
[12] Department of Management, Technology and Economics, ETH Zurich, Switzerland
[13] Department of Philosophy, University of Bristol, United Kingdom

E-mail: k.wiesner@bristol.ac.uk

Received 25 June 2018, revised 21 September 2018
Accepted for publication 25 October 2018
Published 27 November 2018

Abstract

The idea that democracy is under threat, after being largely dormant for at least 40 years, is looming increasingly large in public discourse. Complex systems theory offers a range of powerful new tools to analyse the stability of social institutions in general, and democracy in particular. What makes a democracy stable? And which processes potentially lead to instability of a democratic system? This paper offers a complex systems perspective on this question, informed by areas of the mathematical, natural, and social sciences. We

Humanities & Social Sciences Communications

COMMENT

https://doi.org/10.1057/s41599-020-0518-0 OPEN

What science can do for democracy: a complexity science approach

Tina Eliassi-Rad[1], Henry Farrell[2], David Garcia[3,4], Stephan Lewandowsky[5], Patricia Palacios[6], Don Ross[7,8,9], Didier Sornette[10], Karim Thébault[11] & Karoline Wiesner[12]✉

Political scientists have conventionally assumed that achieving democracy is a one-way ratchet. Only very recently has the question of "democratic backsliding" attracted any research attention. We argue that democratic instability is best understood with tools from complexity science. The explanatory power of complexity science arises from several features of complex systems. Their relevance in the context of democracy is discussed. Several policy recommendations are offered to help (re)stabilize current systems of representative democracy.

[1] Network Science Institute, College of Computer and Information Science, Northeastern University, Boston, MA, USA. [2] Department of Political Science, George Washington University, Washington, DC, USA. [3] Section for Science of Complex Systems, Center for Medical Statistics, Informatics, and Intelligent Systems, Medical University of Vienna, Vienna, Austria. [4] Complexity Science Hub Vienna, Vienna, Austria. [5] School of Psychological Science, University of Bristol, Bristol, UK. [6] Department of Philosophy, University of Salzburg, Salzburg, Austria. [7] School of Sociology, Philosophy, Criminology, Government, and Politics, University College Cork, Ireland, UK. [8] School of Economics, University of Cape Town, Cape Town, South Africa. [9] Center for Economic Analysis of Risk, Georgia State University, Atlanta, GA, USA. [10] Department of Management, Technology and Economics, ETH Zurich, Zürich, Switzerland. [11] Department of Philosophy, University of Bristol, Bristol, UK. [12] School of Mathematics, University of Bristol, Bristol, UK. ✉email: k.wiesner@bristol.ac.uk

Downloadable from http://eliassi.org/pubs.html

Machine learning (ML) systems are not islands.

They are part of broader complex systems.

To understand and mitigate the risks and harms of using ML, we must remove our optimization blinders & study the broader complex systems in which ML systems operate.

Thank you!

Any questions?

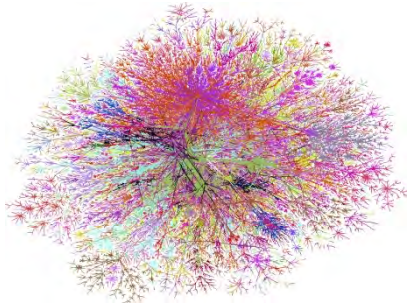
Some of our efforts

- Talent is everywhere
  - Allen, Eliassi-Rad, Fitelson. *The Future of Affirmative Action*, 2014
  - Liu et al. *AI, Ethics, and Society* 2021
- Democratic backsliding
  - Wiesner et al. *European J. of Physics* 2019
  - Eliassi-Rad et al. *Humanities & Social Sciences Communications* 2020
- Information access equality
  - Wang, Varol, Eliassi-Rad. *Applied Network Science* 2022
- COVID-19 & the US criminal justice systems
  - Klein et al. *Nature* 2023
- Interventions on academic hiring
  - Dies et al. *Conference on Complex Systems* 2023
- Using sequences of life-events to predict human lives
  - Germans Savcisens, et al. *Nature Computational Science* 2023

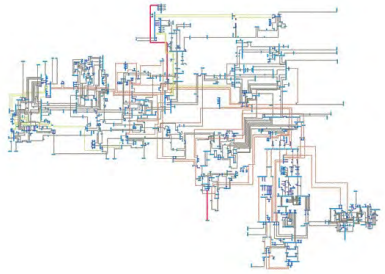
Complex networks are ubiquitous

Technological Networks

Internet

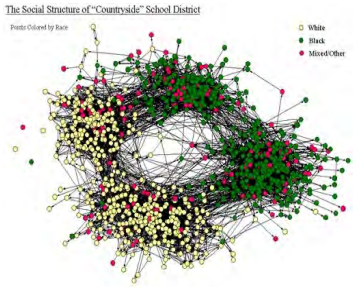


NY State Power Grid



Social Networks

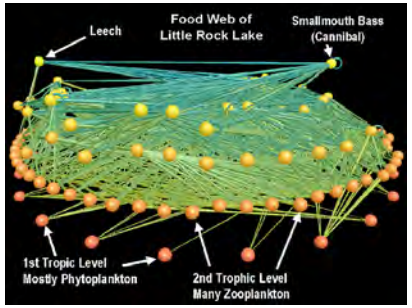
Friendship



HP Emails

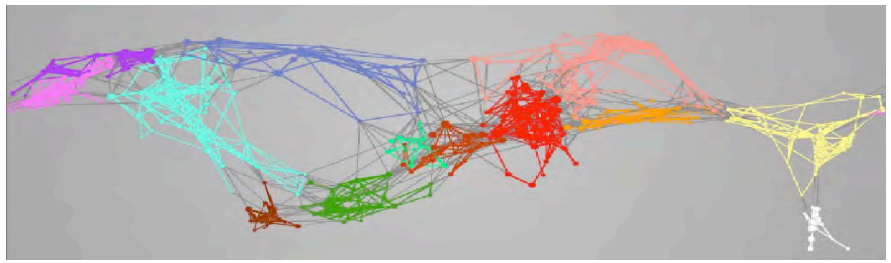
Biological networks

Food Web

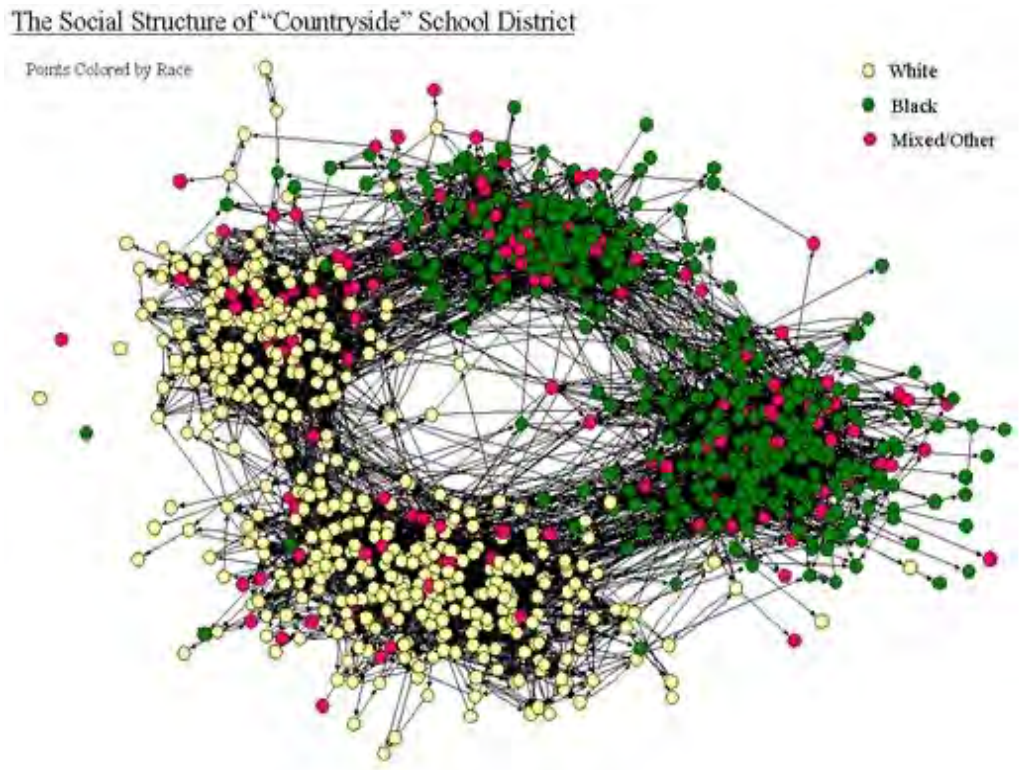


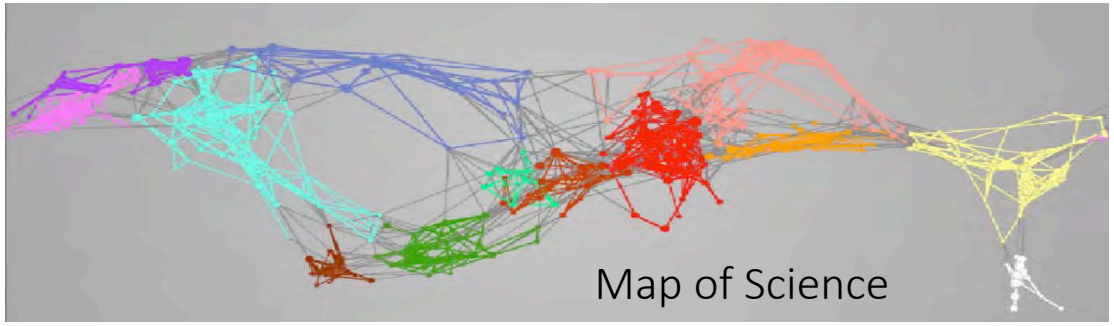
Contagion of TB

Information Networks



Social and information networks





Common mechanisms and properties limiting minority nodes' access to information

Preferential attachment

Homophily

Node activity

Edge density

L. Espín-Noboa, C. Wagner, M. Strohmaier, F. Karimi. **Inequality and Inequity in Network-based Ranking and Recommendation Algorithms**. *Scientific Reports* 12 (1): 1-14 (2022)

Xindi Wang

Onur Varol

What is the interplay between network structure and the spreading process for information access equality?

X. Wang, Onur Varol, T. Eliassi-Rad. **Information Access Equality on Generative Models of Complex Networks**. *Applied Network Science*, Volume 7, Article 54, 2022. https://doi.org/10.1007/s41109-022-00494-8

Growth mechanisms

- Majority/minority dichotomy
- Homophily
- Preferential attachment
- Diversity

Growth mechanisms

- Majority/minority dichotomy
- Homophily
- Preferential attachment
- Diversity

Spreading processes

- Simple vs. complex contagions
- Symmetric vs. asymmetric transmission rates
- Various seeding conditions

Information access equality

- Spreading processes can take different times in different complex networks

Information access equality

- Spreading processes can take different times in different complex networks
- Measure information access equality at different stages of the spreading process

Information access equality

- Spreading processes can take different times in different complex networks
- Measure information access equality at different stages of the spreading process
- Normalize fraction of nodes in state I (Infected) at each time step t by the length of the spreading process T to obtain $I(t/T)$

$$\Delta I(t/T) = \frac{I_{maj}(t/T) - I_{min}(t/T)}{I_{maj}(t/T) + I_{min}(t/T)} \in [-1, 1]$$

Information access equality

- Spreading processes can take different times in different complex networks
- Measure information access equality at different stages of the spreading process
- Normalize fraction of nodes in state I (Infected) at each time step t by the length of the spreading process T to obtain $I(t/T)$

$$\Delta I(t/T) = \frac{I_{maj}(t/T) - I_{min}(t/T)}{I_{maj}(t/T) + I_{min}(t/T)} \in [-1, 1]$$

$\Delta I(t/T)$	Meaning
< 0	Minority group is at an advantage
= 0	Equality
> 0	Majority group is at an advantage

What did we find?

1. Information access equality is a complex interplay between network structures and the spreading processes.

What did we find?

1. Information access equality is a complex interplay between network structures and the spreading processes.
2. There is a trade-off between equality and efficiency of information access under certain circumstances (e.g., low inter-group edges and asymmetric transmission).

What did we find?

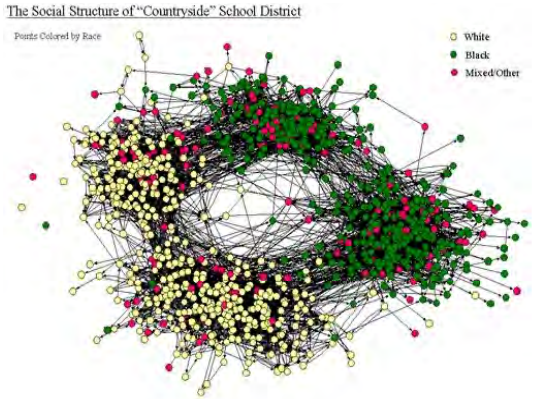
1. Information access equality is a complex interplay between network structures and the spreading processes.
2. There is a trade-off between equality and efficiency of information access under certain circumstances (e.g., low inter-group edges and asymmetric transmission).
3. Spreading process features are statistically significant (p-value ≤ 0.05) when it comes to information access equality.

What did we find?

1. Information access equality is a complex interplay between network structures and the spreading processes.
2. There is a trade-off between equality and efficiency of information access under certain circumstances (e.g., low inter-group edges and asymmetric transmission).
3. Spreading process features are statistically significant (p-value ≤ 0.05) when it comes to information access equality.
4. Network features are not always statistically significant. But two network features stand out w.r.t. information access equality: degree inequality and network distance.

Downstream impact

- Our findings can be used to recommend connections that steer an online social network toward information access equality for a given topic.
- Ideally, such recommendation systems will first classify the spreading process, then based on the classification recommend new connections to their users.

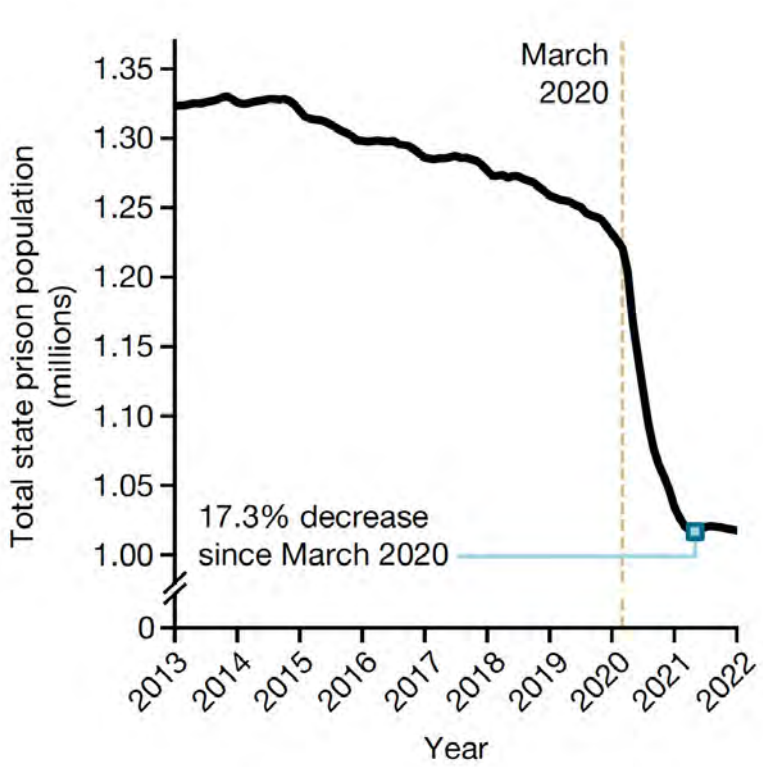


Some of our efforts

- **Talent is everywhere**
  - Allen, Eliassi-Rad, Fitelson. *The Future of Affirmative Action*, 2014
  - Liu et al. *AI, Ethics, and Society* 2021
- **Democratic backsliding**
  - Wiesner et al. *European J. of Physics* 2019
  - Eliassi-Rad et al. *Humanities & Social Sciences Communications* 2020
- **Information access equality**
  - Wang, Varol, Eliassi-Rad. *Applied Network Science* 2022
- **COVID-19 & the US criminal justice systems**
  - Klein et al. *Nature* 2023
- **Interventions on academic hiring**
  - Dies et al. *Conference on Complex Systems* 2023
- **Using sequences of life-events to predict human lives**
  - Germans Savcisens, et al. *Nature Computational Science* 2023

Prison population fell during COVID-19

1. Almost all courts shut down, reducing the admission rate by 70 percent.
2. Prisoners were released in response to the pandemic.



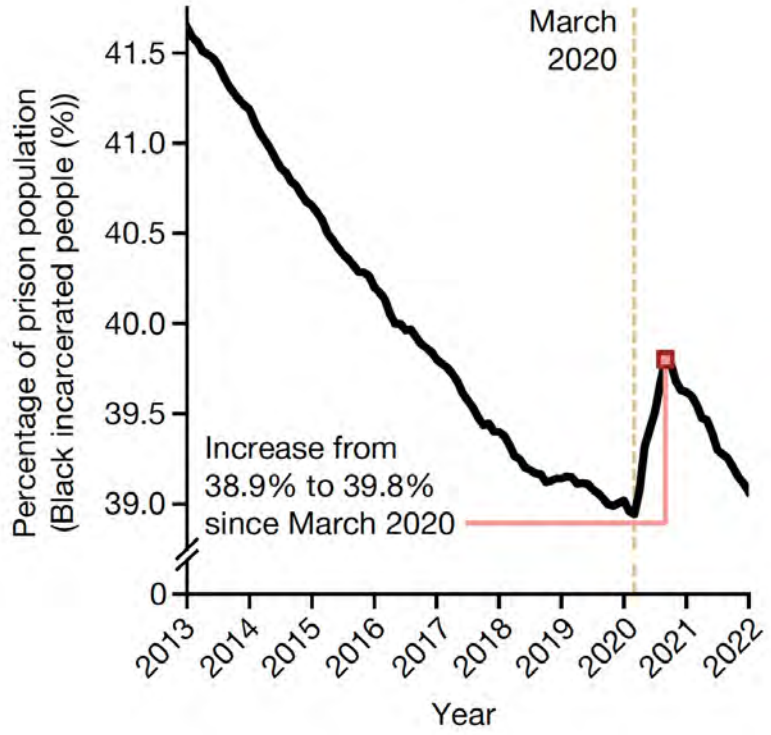
Brennan Klein

B. Klein et al. [The COVID-19 Pandemic Amplified long-standing Racial Disparities in the United States Criminal Justice System](). ***Nature***, 2023.

Prison population fell during COVID-19

- The fraction of incarcerated Black and Latino people increased.
- But in most states, the released prisoners were not disproportionately white.

🤔



Dynamics of the US prison population

1. **Non-white people tend to get longer sentences than white people for the same crimes.**
   - With fewer prisoners coming in over the course of the pandemic and with white prisoners disproportionately serving shorter sentences, the population skewed Black and Latino.
2. **During the pandemic, prosecutors pushed hard for pre-trial plea deals to complete cases.**
   - Plea deals result in a disproportionate number of Black defendants spending time in prison.
3. **Decreasing the flow of new admissions increased the non-white population.**
   - The Black-white disparity in new prison admissions is typically a ratio of 2:1, whereas it is closer to 6:1 for the total incarcerated.

Why should you care?

- There is still **a large backlog of cases** in the criminal legal system stemming from delays in the early stages of the pandemic

- The **structural problem** in the U.S. criminal legal system **will continue to worsen unless we address the sentencing inequities** in the legal system and work toward reforms that will produce a more equitable and just system