

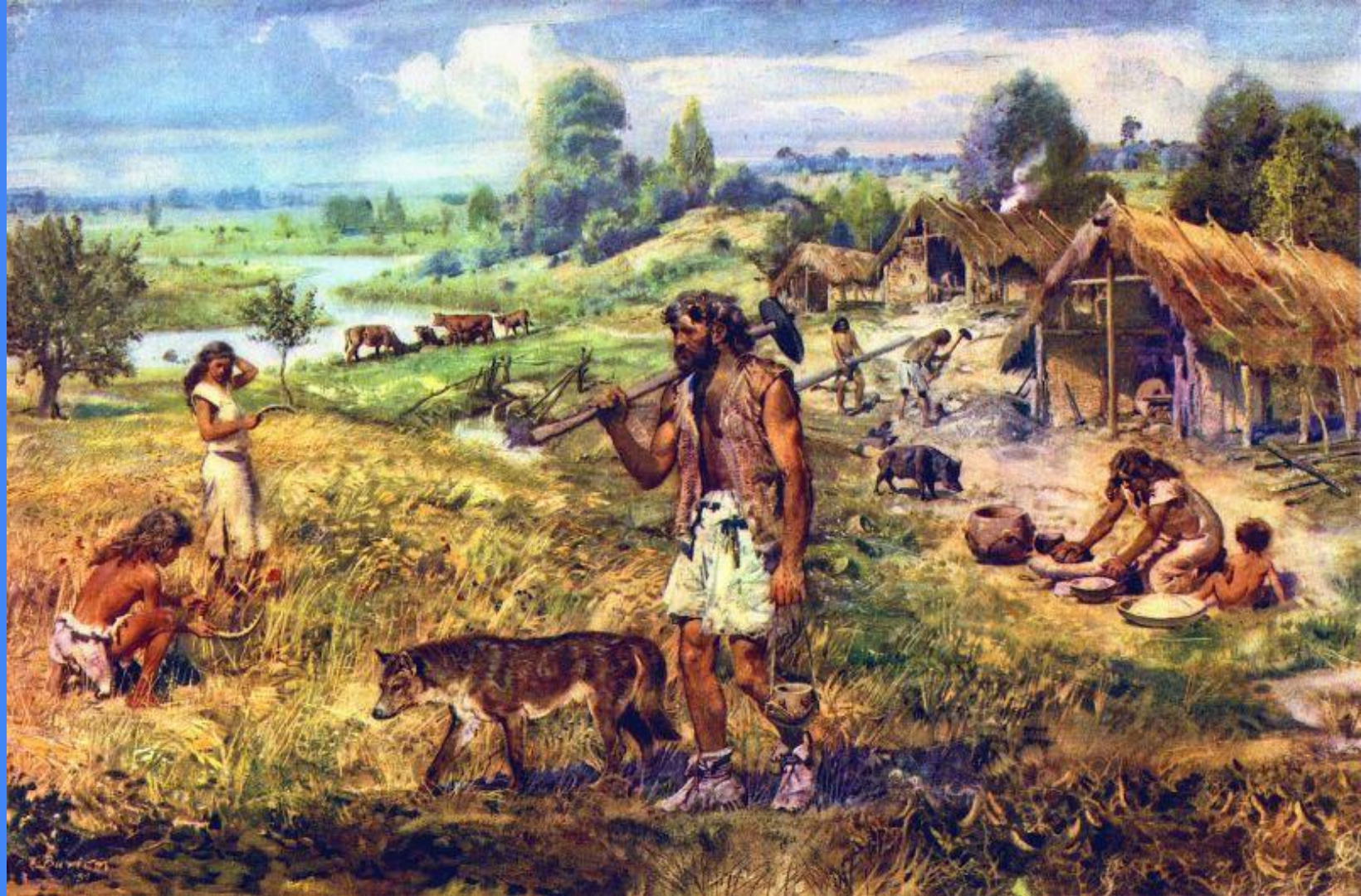
# Prediction and explanation in social systems

Jake M. Hofman, Amit Sharma, Duncan J. Watts

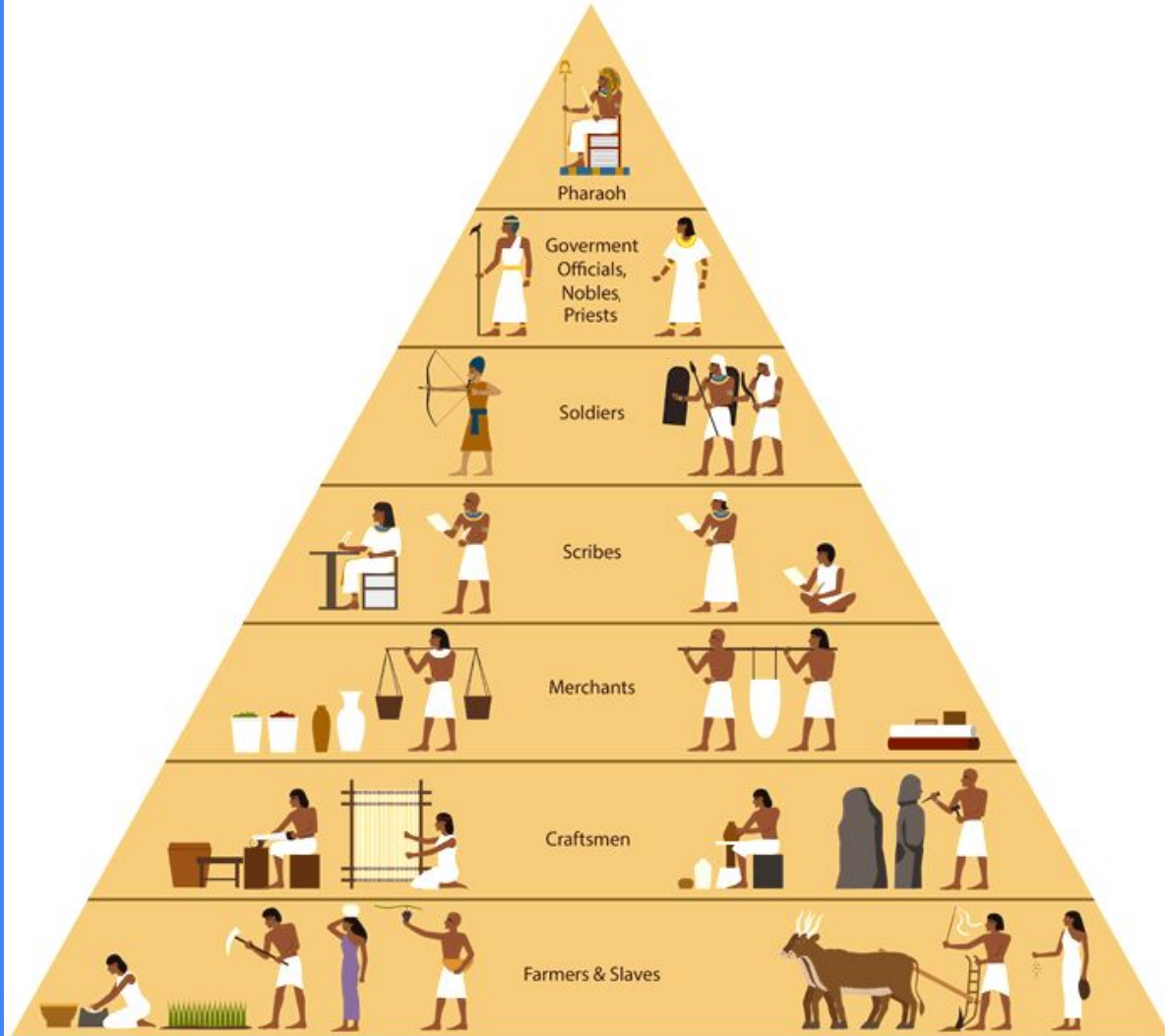


Let's predict large scale evolution  
in social systems!

However, it's not that easy...







Subscribe for the  
**INDUSTRIAL WORKER**  
Foremost Exposure of Revolutionary  
Industrial Unionism  
Published in the English Language  
One Dollar per Year  
Three Months 25c  
P. O. Box 3129 Spokane, Washington

**CAPITALISM**

**WE RULK YOU**

**WE FOOL YOU**

**WE SHOOT AT YOU**

**WE LAY FOR YOU**

**WE WORK FOR ALL**

**WE FEED ALL**

**PYRAMID OF CAPITALIST SYSTEM**

Problem



# Problem

- Social scientists emphasis on providing interpretable causal mechanisms, not on prediction
  - Social scientists are often not concerned about  $R^2$  (explained variation/total variation)

An Example from sociology

# Class Advantage, Commitment Penalty: The Gendered Effect of Social Class Signals in an Elite Labor Market

Lauren A. Rivera and András Tilcsik

American Sociological Review 2016,  
Vol. 81(6) 1097–1131

**Table 6.** OLS Models Predicting the Strength of Interview Recommendations in the Survey Experiment

	Model 12	Model 13	Model 14
	None	Respondents Who Have Ever Worked at a Law Firm	Respondents Currently at a Law Firm
<b>Applicant Characteristics</b>			
Male applicant	-.176 (.228)	-.561* (.269)	-.532 (.297)
Higher-class signals	-.027 (.225)	-.280 (.305)	-.236 (.329)
Male applicant × higher-class signals	.627* (.318)	.983* (.401)	1.113* (.443)
<b>Survey Respondent Characteristics</b>			
Respondent age	.008 (.008)	-.004 (.009)	-.002 (.010)
Male respondent	.095 (.173)	.177 (.216)	.189 (.229)
Respondent has worked at law firm	.191 (.289)		
Race dummies	Yes	Yes	Yes
Current position dummies	Yes	Yes	Yes
Current organization dummies	Yes	Yes	No
Constant	6.064*** (1.215)	6.947*** (1.372)	4.875*** (1.046)
<i>R</i> -squared	.20	.27	.29
Observations	210	122	97

\* $p < .05$ ; \*\* $p < .01$ ; \*\*\* $p < .001$  (two-tailed tests).

## What Social Scientists care about

**Table 6.** OLS Models Predicting the Strength of Interview Recommendations in the Survey Experiment

	Model 12	Model 13	Model 14
Sample Restriction:	None	Respondents Who Have Ever Worked at a Law Firm	Respondents Currently at a Law Firm
<b>Applicant Characteristics</b>			
Male applicant	-.176 (.228)	-.561* (.269)	-.532 (.297)
Higher-class signals	-.027 (.225)	-.280 (.305)	-.236 (.329)
Male applicant × higher-class signals	.627* (.318)	.983* (.401)	1.113* (.443)
<b>Survey Respondent Characteristics</b>			
Respondent age	.008 (.008)	-.004 (.009)	-.002 (.010)
Male respondent	.095 (.173)	.177 (.216)	.189 (.229)
Respondent has worked at law firm	.191 (.289)		
Race dummies	Yes	Yes	Yes
Current position dummies	Yes	Yes	Yes
Current organization dummies	Yes	Yes	No
Constant	6.064*** (1.215)	6.947*** (1.372)	4.875*** (1.046)
<i>R</i> -squared	.20	.27	.29
Observations	210	122	97

\* $p < .05$ ; \*\* $p < .01$ ; \*\*\* $p < .001$  (two-tailed tests).

What people in predictive modeling care about

**Table 6.** OLS Models Predicting the Strength of Interview Recommendations in the Survey Experiment

	Model 12	Model 13	Model 14
Sample Restriction:	None	Respondents Who Have Ever Worked at a Law Firm	Respondents Currently at a Law Firm
<b>Applicant Characteristics</b>			
Male applicant	-.176 (.228)	-.561* (.269)	-.532 (.297)
Higher-class signals	-.027 (.225)	-.280 (.305)	-.236 (.329)
Male applicant × higher-class signals	.627* (.318)	.983* (.401)	1.113* (.443)
<b>Survey Respondent Characteristics</b>			
Respondent age	.008 (.008)	-.004 (.009)	-.002 (.010)
Male respondent	.095 (.173)	.177 (.216)	.189 (.229)
Respondent has worked at law firm	.191 (.289)		
Race dummies	Yes	Yes	Yes
Current position dummies	Yes	Yes	Yes
Current organization dummies	Yes	Yes	No
Constant	6.064*** (1.215)	6.947*** (1.372)	4.875*** (1.313)
<i>R</i> -squared	.20	.27	.29
Observations	210	122	97

\* $p < .05$ ; \*\* $p < .01$ ; \*\*\* $p < .001$  (two-tailed tests).

# Problem

- Social scientists emphasis on providing interpretable causal mechanisms, not on prediction
    - Social scientists are often not concerned about  $R^2$  (explained variation/total variation)
- Lack of predictive accuracy
- Unbiased estimates more important than prediction
- Search for statistical significance

# Benefit

- Social systems have meaning → Theory has to account for that
- Unbiased estimates can help to identify causal mechanisms → Substantive theory (beware: fallacy of common sense)

# Problems in predictive modeling

- Problem has to be translated into a computational task
- Models, data sets, predictive measures are subject to individual choice → Researcher degrees of freedom
- Predictions vary based on researcher decisions



Data source

Regression

Prediction task

Classification

$R^2$

RMSE

MAE

Evaluation metric

Accuracy

AUC

F1

0 1 10 100

0 1 10 100

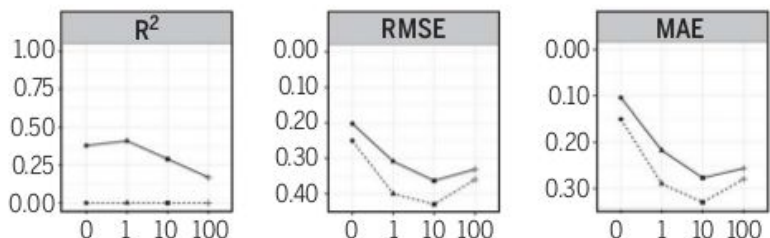
0 1 10 100

1 10 100

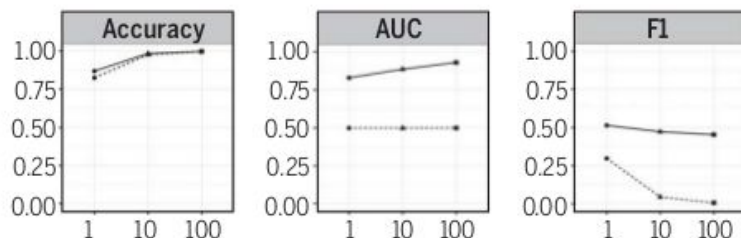
1 10 100

1 10 100

Data processing



Threshold for filtering unpopular content



Threshold for defining successful outcomes

Solution

# 1. Standardization of current practices

1. Finding substantive problems that are regarded as important
2. Testing these algorithmic findings on different data sets based on commonly agreed standards
3. Differentiate between confirmatory and explanatory research

## 2. Theoretical limits of predictability

1. Human behavior ranges from very predictable (phone user example) to almost unpredictable (impact of the web)
2. Finding a balance between model performance (theoretical maximum) and meaningful explanations of phenomena (identifying what limits the prediction)

# 3. Predictive accuracy and interpretability as complementary practices

- Often the claim that predictive models are too complex to be generalized → spurious claims
  1. Simple models do not necessarily generalize better
  2. Trade-off between prediction and interpretability smaller than expected
  3. Subjective understanding should not be equated with true understanding of the model

## The proposed solution

**Table 6.** OLS Models Predicting the Strength of Interview Recommendations in the Survey Experiment

	Model 12	Model 13	Model 14
Sample Restriction:	None	Respondents Who Have Ever Worked at a Law Firm	Respondents Currently at a Law Firm
<b>Applicant Characteristics</b>			
Male applicant	-.176 (.228)	-.561* (.269)	-.532 (.297)
Higher-class signals	-.027 (.225)	-.280 (.305)	-.236 (.329)
Male applicant × higher-class signals	.627* (.318)	.983* (.401)	1.113* (.443)
<b>Survey Respondent Characteristics</b>			
Respondent age	.008 (.008)	-.004 (.009)	-.002 (.010)
Male respondent	.095 (.173)	.177 (.216)	.189 (.229)
Respondent has worked at law firm	.191 (.289)		
Race dummies	Yes	Yes	Yes
Current position dummies	Yes	Yes	Yes
Current organization dummies	Yes	Yes	No
Constant	6.064*** (1.215)	6.947*** (1.372)	4.875*** (1.046)
<i>R</i> -squared	.20	.27	.29
Observations	210	122	97

\* $p < .05$ ; \*\* $p < .01$ ; \*\*\* $p < .001$  (two-tailed tests).

Prediction and explanation should  
be complementary in the analysis  
of social systems

# Questions

- What are potential limits of predictability?
- Thinking about predictive policing: Are there ethical pitfalls of prediction
- Which type of data is suitable for prediction?
- How can we overcome disciplinary boundaries?
- How will theory change?