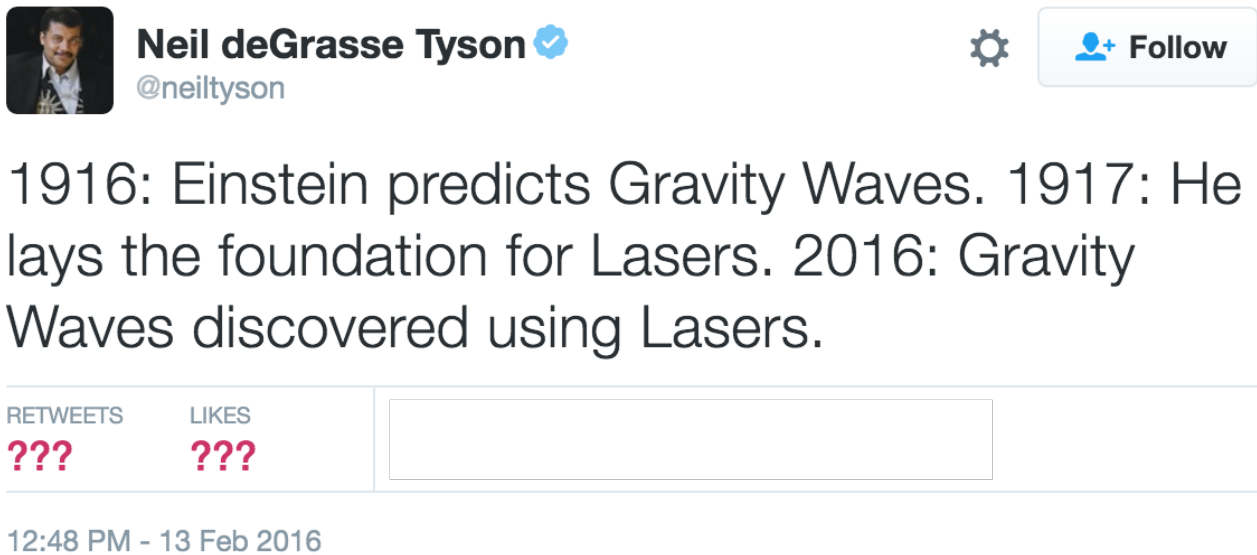
Exploring the limits of prediction

Jake Hofman

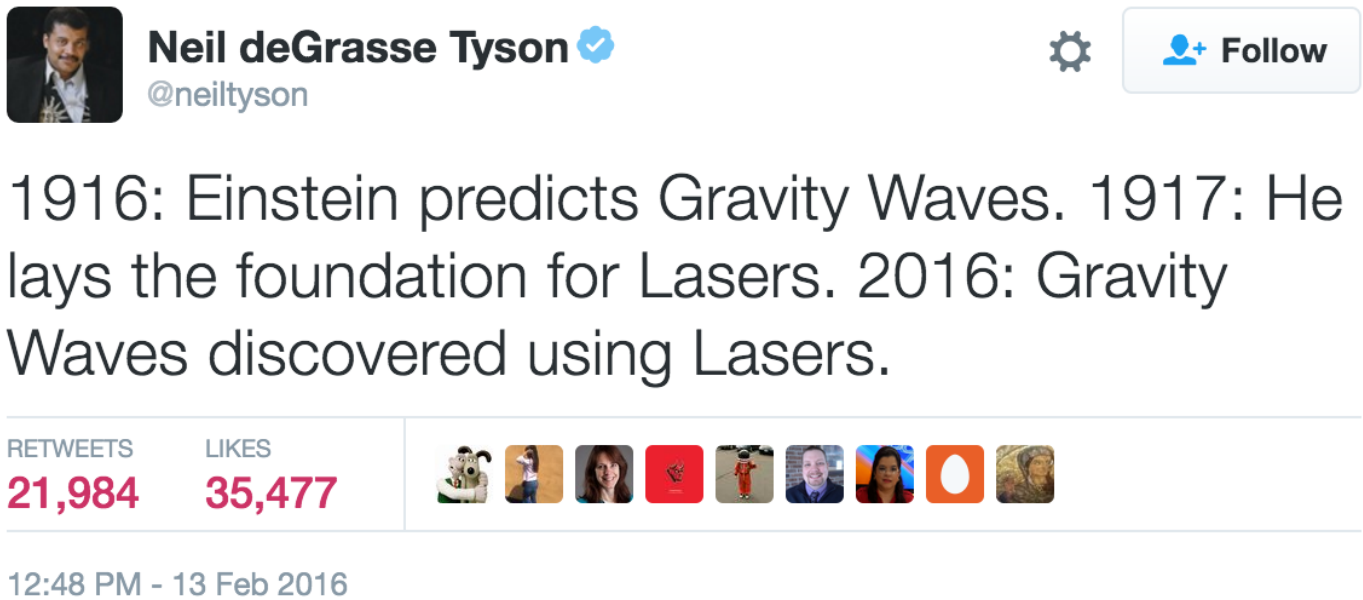
Microsoft Research NYC

January 16, 2018

How far will this spread?

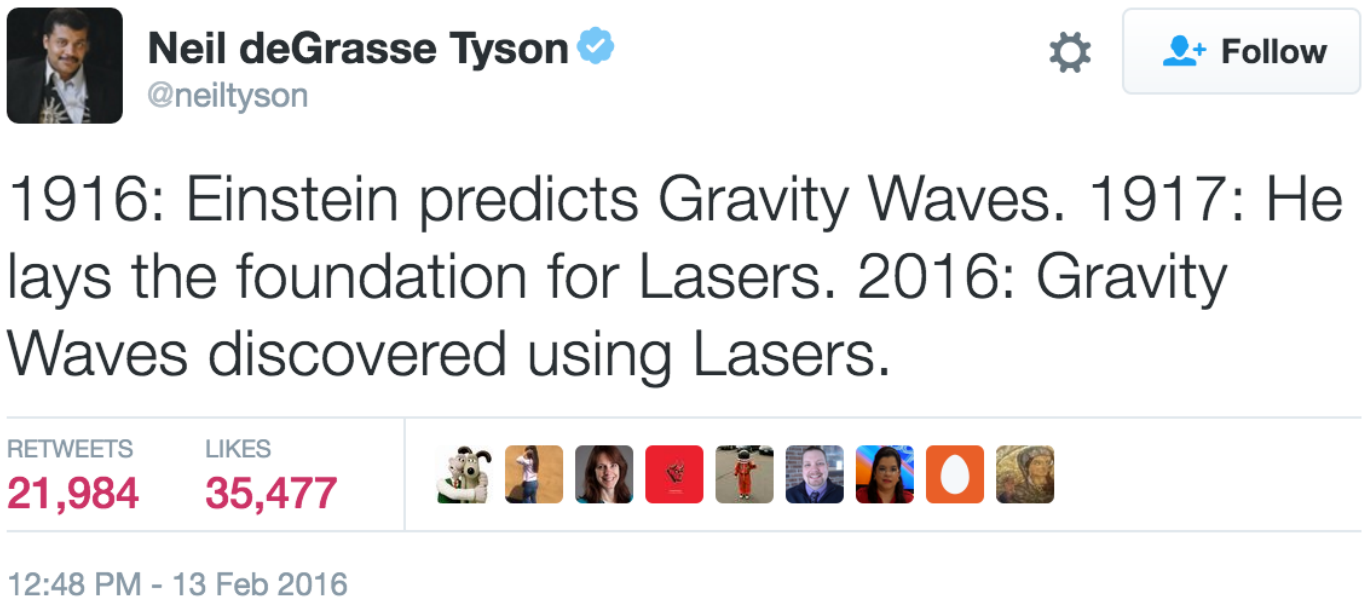


How far will this spread?



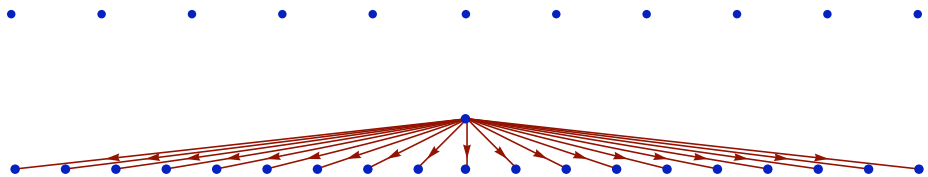
Why is so difficult to predict success?

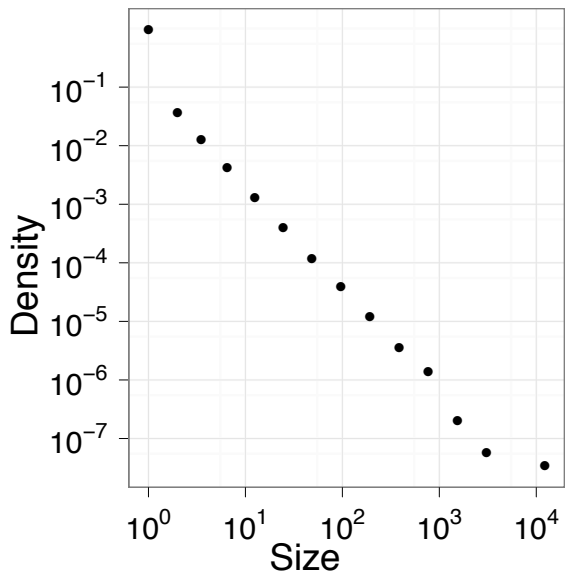
Do we need bigger data and better models?



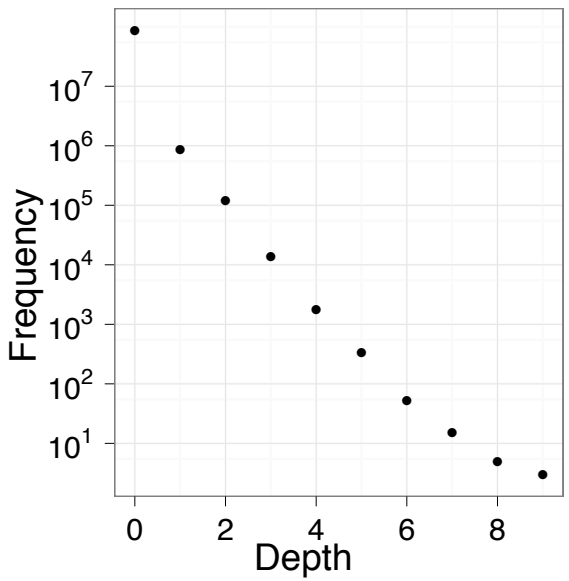
Or is information diffusion inherently unpredictable?

Most things don't spread





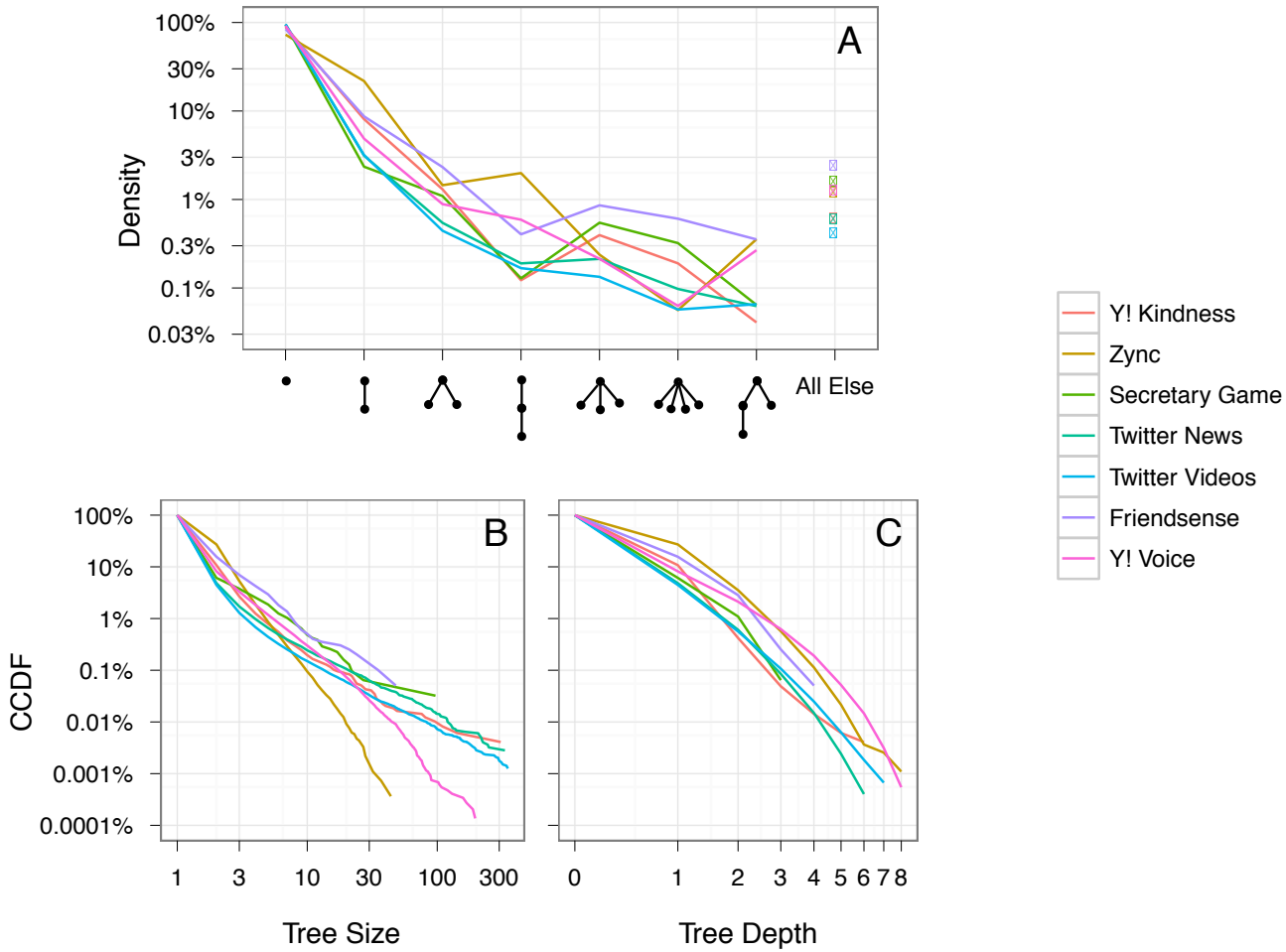
(a) *Cascade Sizes*



(b) *Cascade Depths*

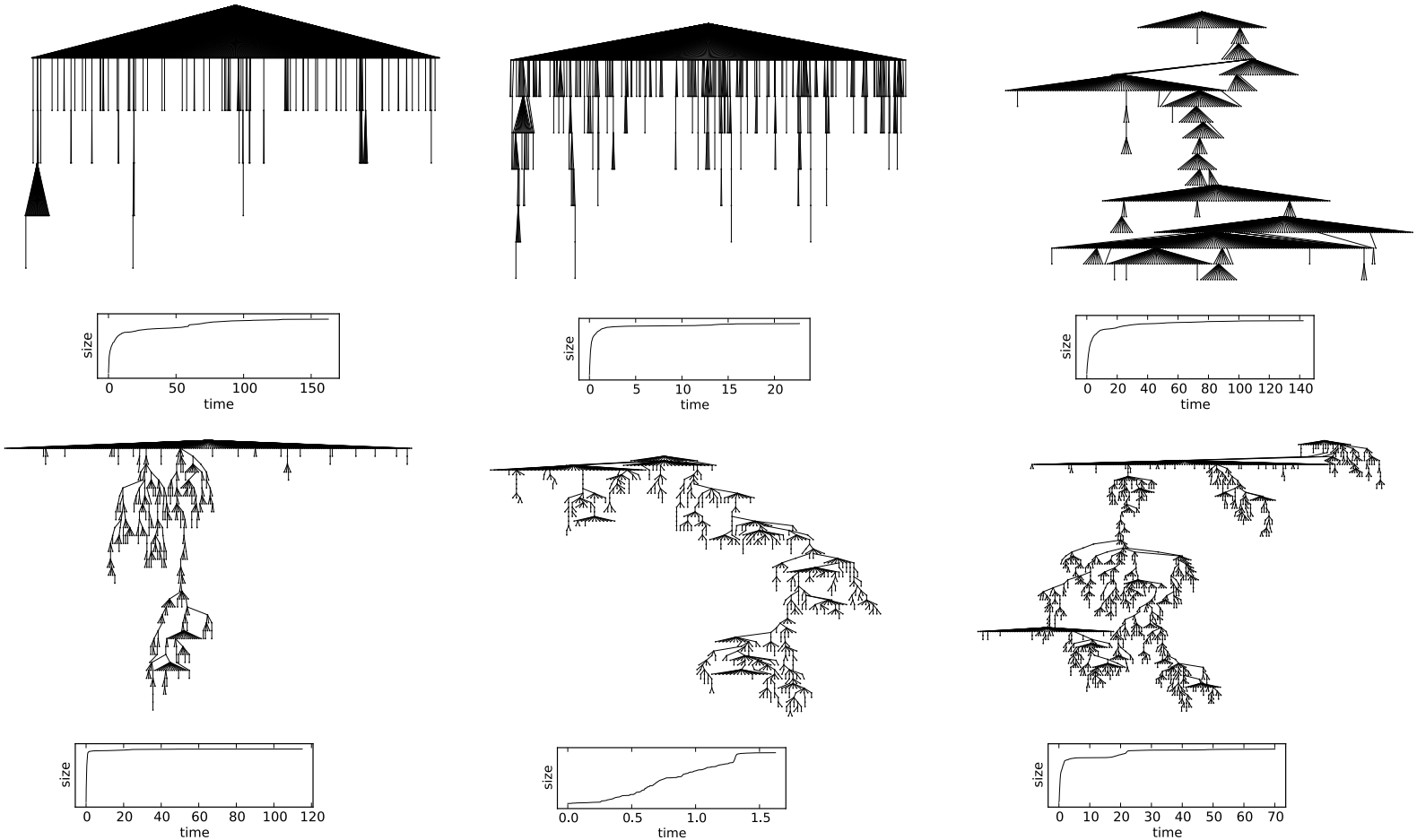
Bakshy, Hofman, Mason, Watts (2011)

Most things don't spread



Goel, Goldstein, Watts (2012)

It's difficult to say how the popular things get popular

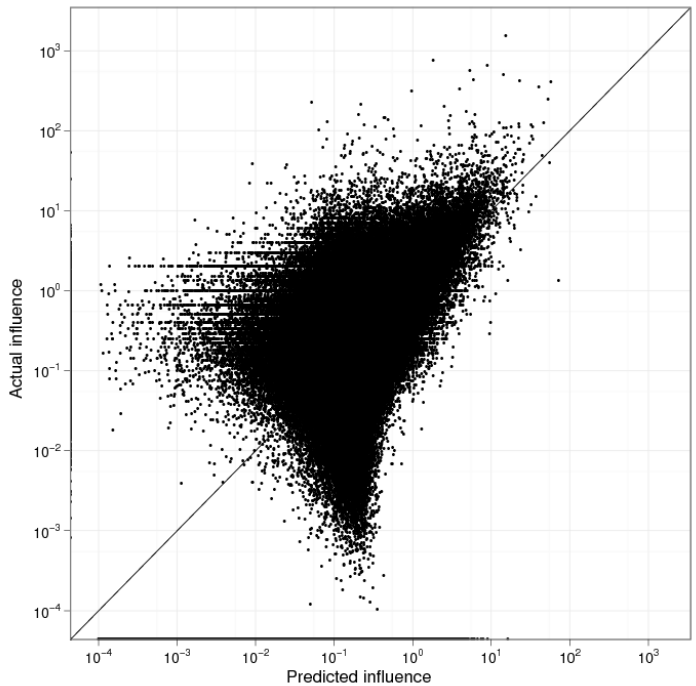


Goel, Anderson, Hofman, Watts (2015)

Predicting success, take 1

Bakshy, Hofman, Mason, Watts (2011)

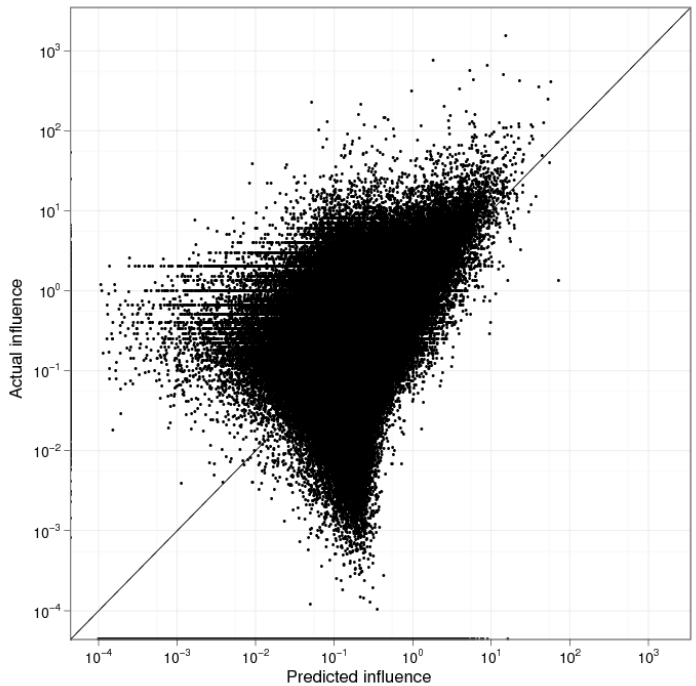
- Looked at 75M diffusion events across 1M users
- Found a relatively low correlation ($R^2 \sim 30\%$) between predicted and actual cascade sizes
- Almost all predictive power comes from examining past performance of a user or piece of content



Predicting success, take 1

Bakshy, Hofman, Mason, Watts (2011)

- Looked at 75M diffusion events across 1M users
- Found a relatively low correlation ($R^2 \sim 30\%$) between predicted and actual cascade sizes
- Almost all predictive power comes from examining past performance of a user or piece of content



How much better can we do?

Related work

- Hong & Davidson (2010): Will a given user be retweeted?
  Topic model features outperform baselines (F1 = 0.47)
- Petrovic et. al. (2011): Will a given tweet be retweeted?
  Social and content features beat humans (F1 = 0.46)
- Jenders et. al. (2013): Will a cascade reach a minimum size?
  Content features lead to good performance (F1 = 0.90)
- Tan et. al. (2014): Which of two tweets will spread further?
  Detailed wording features are informative (Accuracy = 0.65)
- Cheng et. al. (2014): Will a cascade double in size?
  Temporal features provide good performance (AUC = 0.88)

Progress?

All of this work examines a different question with a different measure of success, evaluated on a different subset of data, making it difficult to assess overall progress[1]

[1] http://hunch.net/?p=22

Progress?

Same data, same model ... seemingly different answers

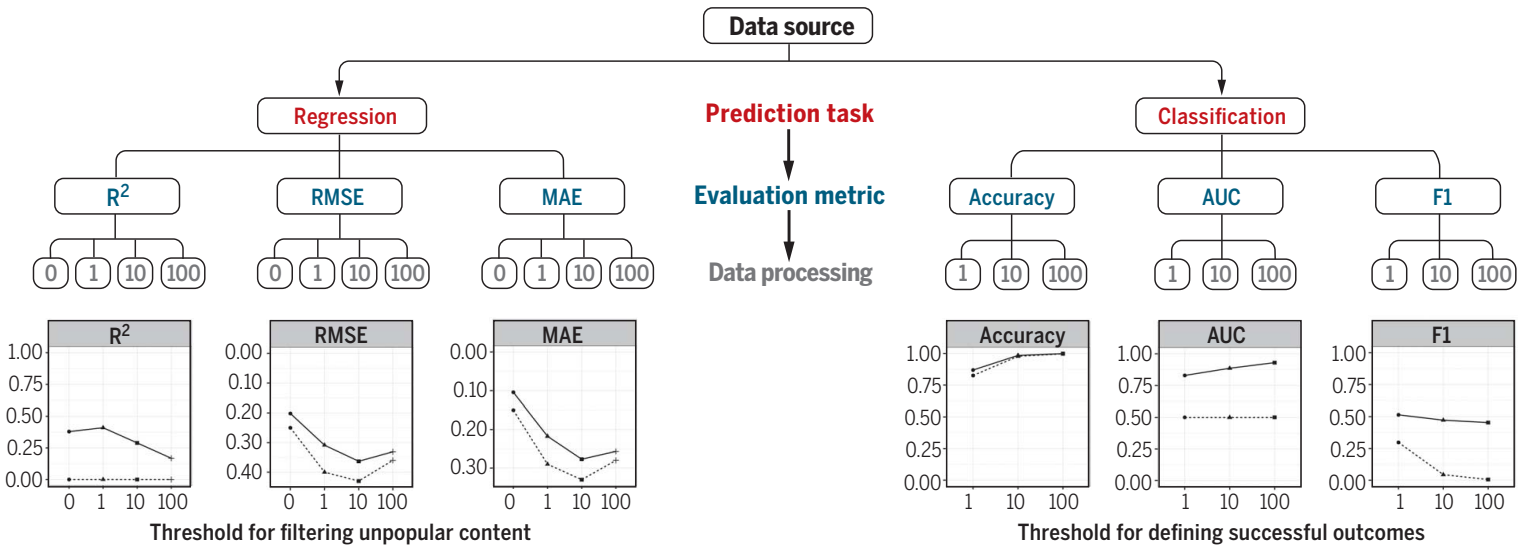


Fig. 1. A single question may correspond to many research designs, each yielding different answers. (**Top**) A depiction of the many choices involved in translating the problem of understanding diffusion cascades into a concrete prediction task, including the choice of data source, task, evaluation metric, and data preprocessing. The preprocessing choices shown at the terminal nodes refer to the threshold used to filter observations for regression or define successful outcomes for classification. Cascade sizes were log-transformed for all of the regression tasks. (**Bottom**) The results of each prediction task, for each metric, as a function of the threshold used in each task. The lower limit of each vertical axis gives the worst possible performance on each metric, and the top gives the best. Dashed lines represent the performance of a naive predictor (always forecasting the global mean for regression or the positive class for classification), and solid lines show the performance of the fitted model. R^2, coefficient of determination; AUC, area under the ROC curve; RMSE, root mean squared error; MAE, mean absolute error; F1 score, the harmonic mean of precision and recall.

Hofman *et al.*, *Science* **355**, 486–488 (2017) 3 February 2017

Data

- Examined all 1.4B tweets containing URLs posted in February 2015

Data

- Examined all 1.4B tweets containing URLs posted in February 2015
- Eliminated spam using internal Microsoft classifier

Data

- Examined all 1.4B tweets containing URLs posted in February 2015
- Eliminated spam using internal Microsoft classifier
- Restricted attention to tweets containing URLs from the top 100 English-speaking domains with the most unique adopters

Data

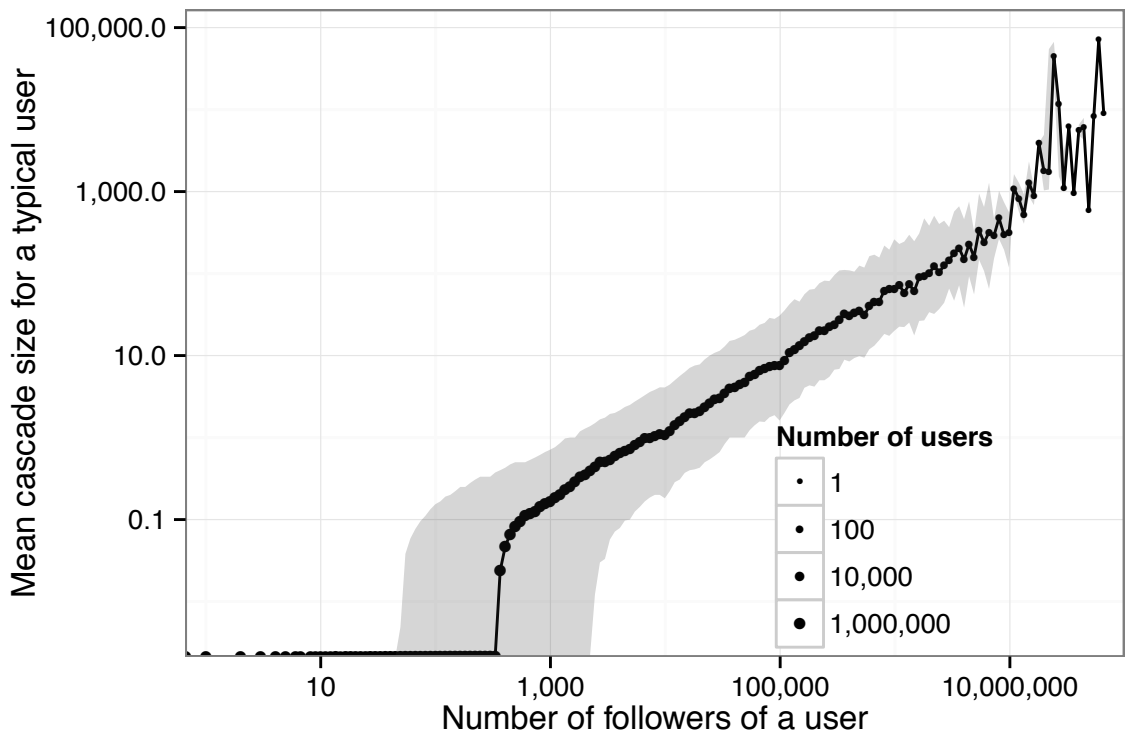
- Examined all 1.4B tweets containing URLs posted in February 2015
- Eliminated spam using internal Microsoft classifier
- Restricted attention to tweets containing URLs from the top 100 English-speaking domains with the most unique adopters
- Resulted in 850M tweets from 50M distinct users covering news, entertainment, videos, images, and products

Data

- Examined all 1.4B tweets containing URLs posted in February 2015
- Eliminated spam using internal Microsoft classifier
- Restricted attention to tweets containing URLs from the top 100 English-speaking domains with the most unique adopters
- Resulted in 850M tweets from 50M distinct users covering news, entertainment, videos, images, and products
- Measured the total cascade size for each seed tweet

Cascade size by degree

There's some signal in the obvious features, but also lots of variance



Predicting success, take 2

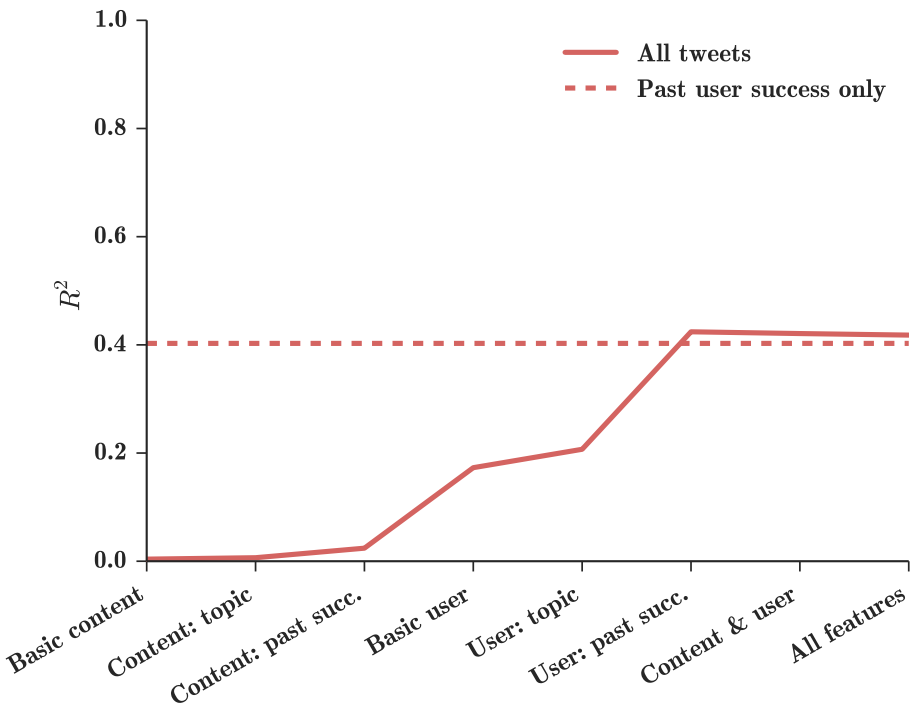
Used a random forest to estimate success (cascade size) given available features

- Basic content features: URL domain, time of tweet, spam score, ODP category
- Basic user features: number of followers, number of friends, number of posts, account creation time
- Topic features: the most probable Latent Dirichlet Allocation topic for each user and tweet, along with an interaction term
- Past success: the average number of retweets received by each URL and user in the past

Predicting success, take 2

Our best model explains roughly 40% of the variance in outcomes

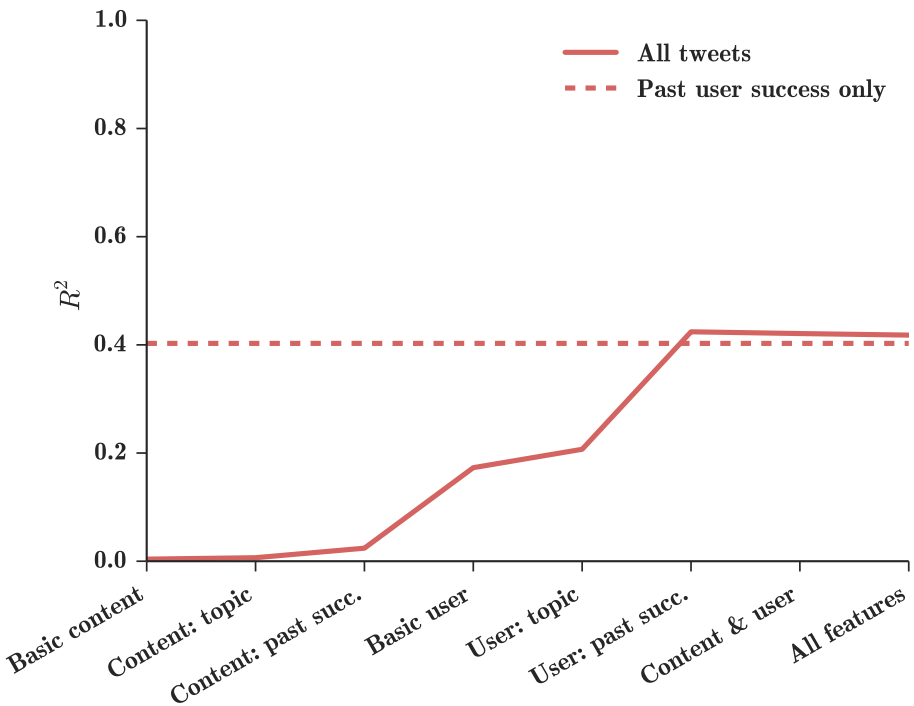
Model	*Tweet time*	*Domain*	*Spam score*	*Category*	*Tweet topic*	*Past url success*	*User time*	*Followers*	*Friends*	*Statuses*	*User topic*	*Past user success*	*Topic interaction*
1. Basic content	✓	✓	✓	✓									
2. Content, topic	✓	✓	✓	✓	✓								
3. Content, past succ.	✓	✓	✓	✓	✓	✓							
4. Basic user							✓	✓	✓	✓			
5. User, topic							✓	✓	✓	✓	✓		
6. User, past succ.							✓	✓	✓	✓	✓	✓	
7. Content, user	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
8. All	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓



Predicting success, take 2

Content features alone perform poorly

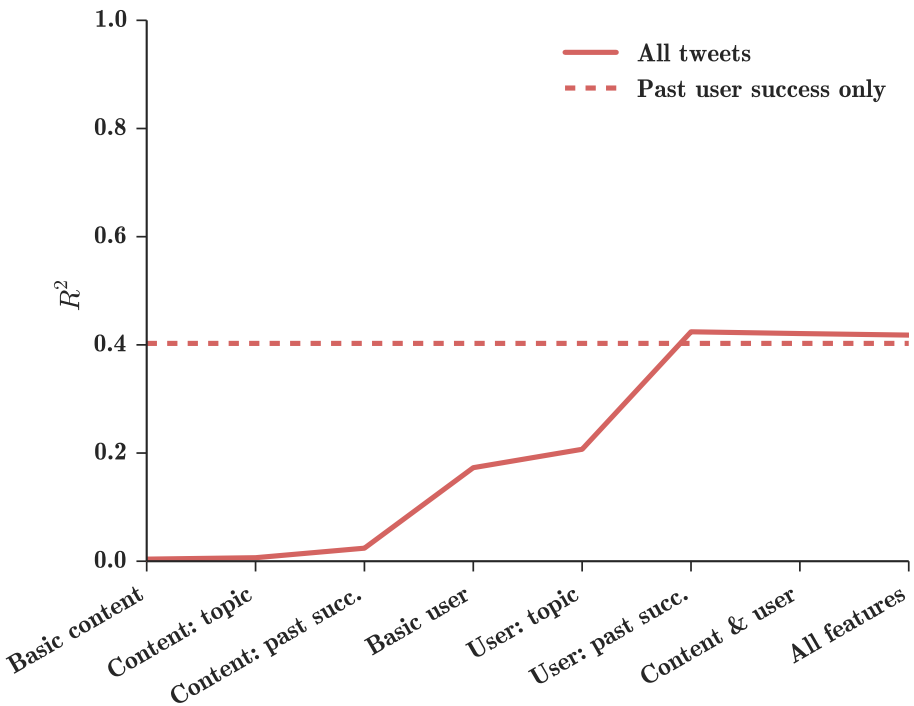
Model	*Tweet time*	*Domain*	*Spam score*	*Category*	*Tweet topic*	*Past url success*	*User time*	*Followers*	*Friends*	*Statuses*	*User topic*	*Past user success*	*Topic interaction*
1. Basic content	✓	✓	✓	✓									
2. Content, topic	✓	✓	✓	✓	✓								
3. Content, past succ.	✓	✓	✓	✓	✓	✓							
4. Basic user							✓	✓	✓	✓			
5. User, topic							✓	✓	✓	✓	✓		
6. User, past succ.							✓	✓	✓	✓	✓	✓	
7. Content, user	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
8. All	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓



Predicting success, take 2

Basic user features provide a reasonable boost in performance

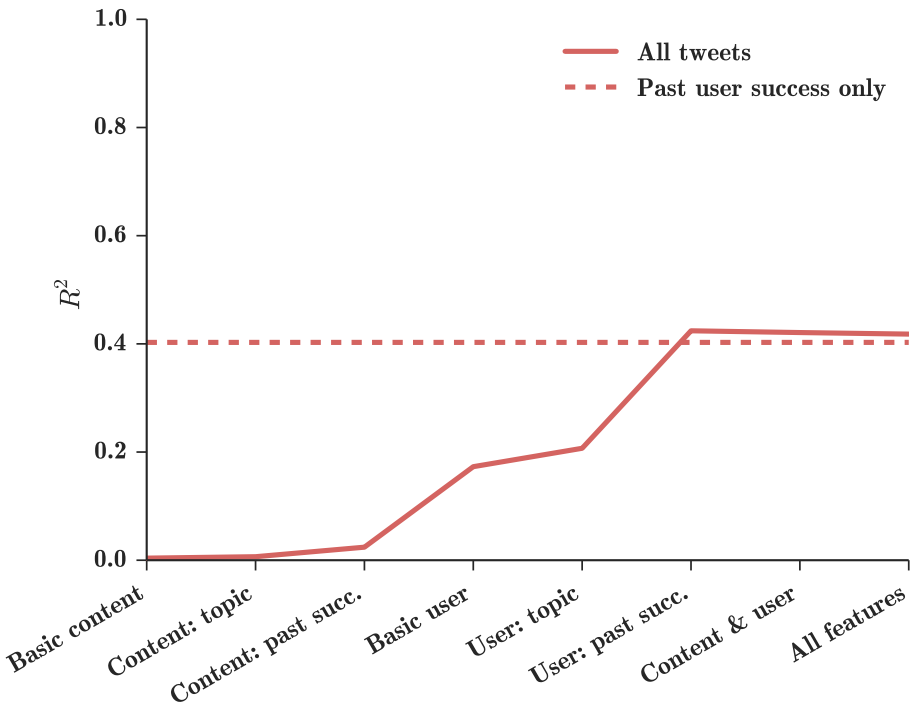
Model	*Tweet time*	*Domain*	*Spam score*	*Category*	*Tweet topic*	*Past url success*	*User time*	*Followers*	*Friends*	*Statuses*	*User topic*	*Past user success*	*Topic interaction*
1. Basic content	✓	✓	✓	✓									
2. Content, topic	✓	✓	✓	✓	✓								
3. Content, past succ.	✓	✓	✓	✓	✓	✓							
4. Basic user							✓	✓	✓	✓			
5. User, topic							✓	✓	✓	✓	✓		
6. User, past succ.							✓	✓	✓	✓	✓	✓	
7. Content, user	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
8. All	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓



Predicting success, take 2

Past user success alone accounts for almost all of predictive power

Model	*Tweet time*	*Domain*	*Spam score*	*Category*	*Tweet topic*	*Past url success*	*User time*	*Followers*	*Friends*	*Statuses*	*User topic*	*Past user success*	*Topic interaction*
1. Basic content	✓	✓	✓	✓									
2. Content, topic	✓	✓	✓	✓	✓								
3. Content, past succ.	✓	✓	✓	✓	✓	✓							
4. Basic user							✓	✓	✓	✓			
5. User, topic							✓	✓	✓	✓	✓		
6. User, past succ.							✓	✓	✓	✓	✓	✓	
7. Content, user	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
8. All	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓



Summary of empirical results

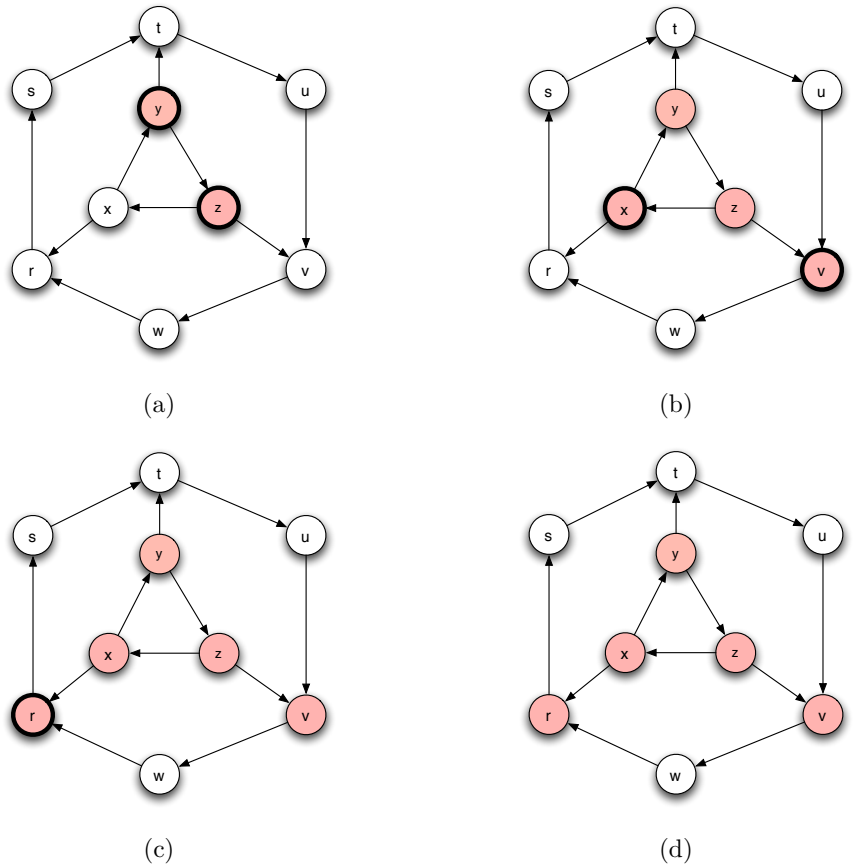
- This is the best known model since Bakshy et. al., boosting performance from $R^2 \sim 30\%$ to $R^2 \sim 40\%$
- Both models derive their predictive power from the same simple feature: a user's past success
- Content features are only weakly informative
- Performance plateaus as we add more features, suggesting a possible limit to the predictability of diffusion outcomes

Simulations

- In practice we can never rule out missing features or superior models, so we turn to numerical simulations where we have full access to and control of all relevant information
- Looked at the variation in outcomes when we repeatedly seed the same user with the same content
- Examined how this varies with content heterogeneity and estimation error

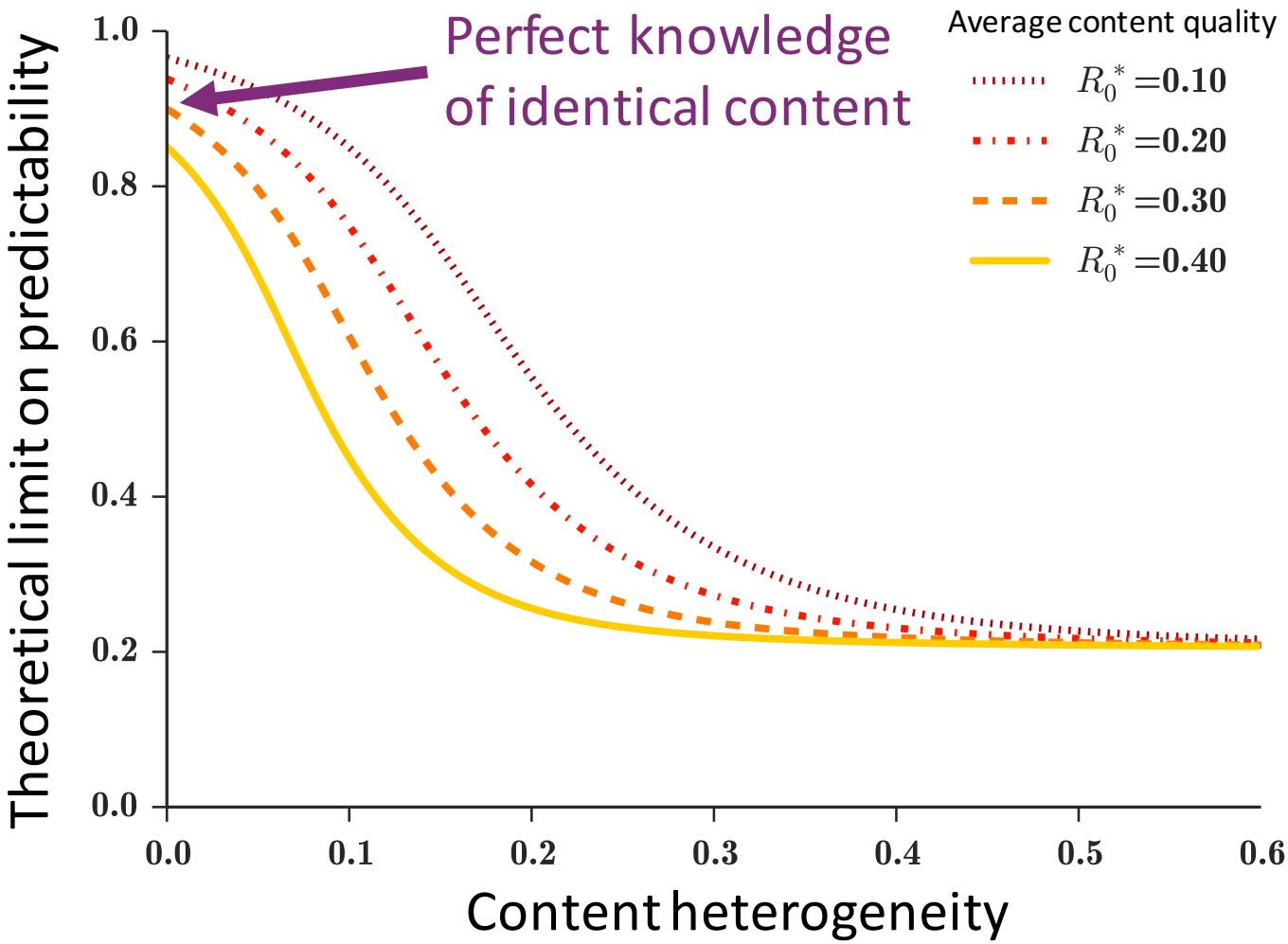
Simulations

- Created a scale-free network similar to Twitter but smaller in size
- Simulated 8B cascades using a standard SI model
- Initiated 1,000 cascades for each combination of 10,000 different seed users and 800 different infectiousness values
- Carefully matched distributions of user activity and cascade size to our empirical data



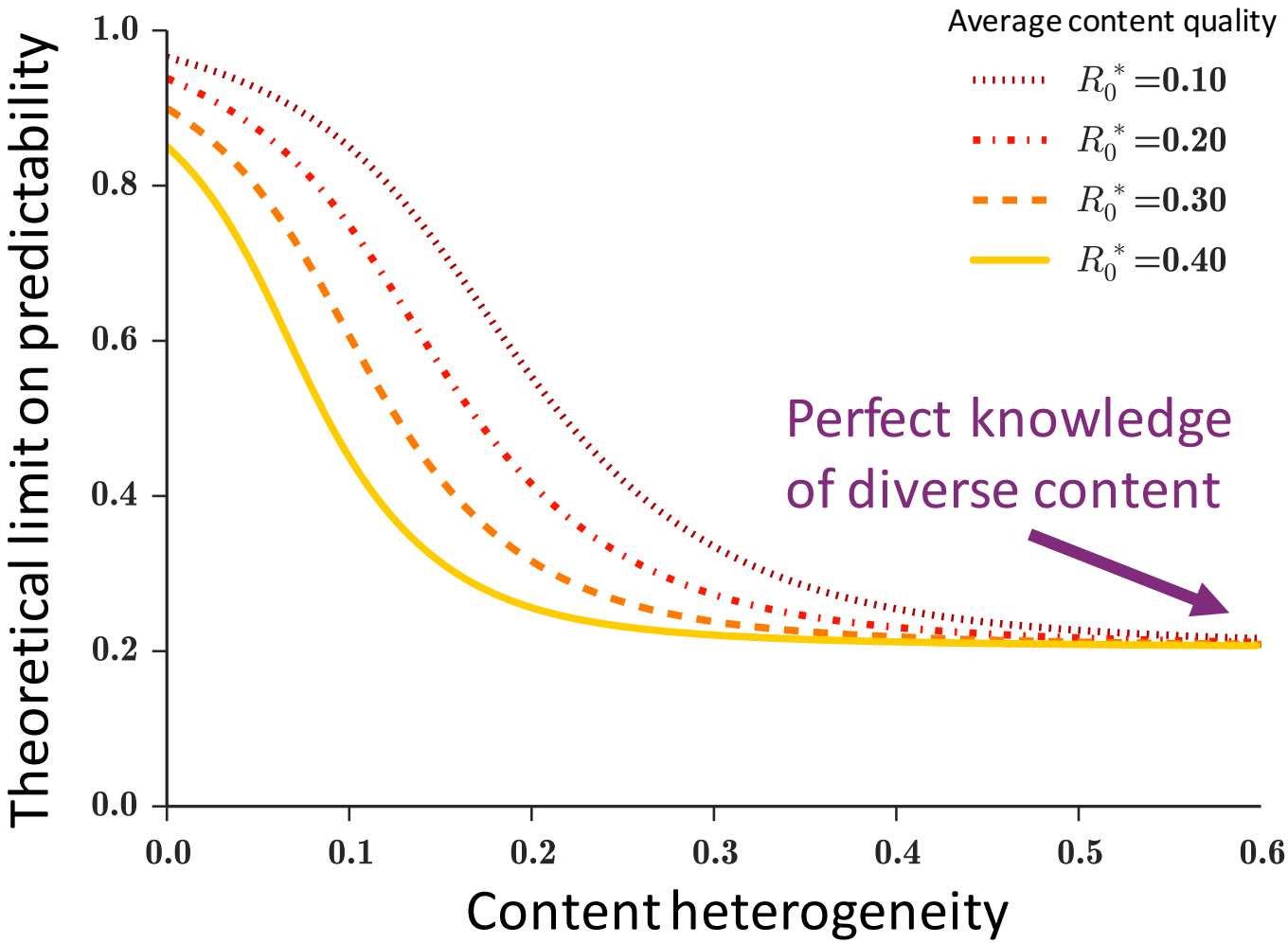
Repeatedly seed the same user with the same content

Outcomes are highly predictable when all content is identical



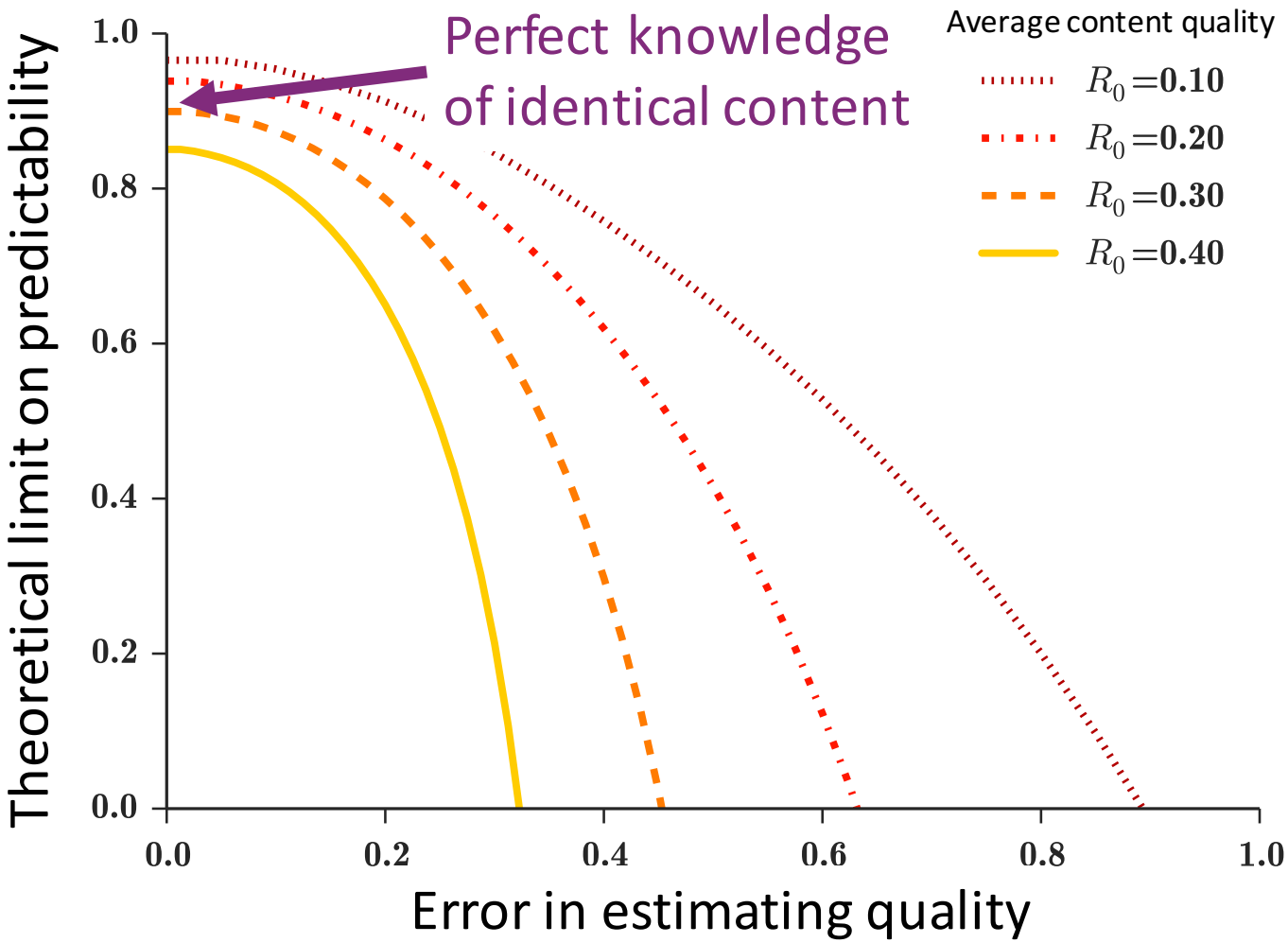
Repeatedly seed the same user with the same content

Predictive performance decreases sharply with content diversity
(e.g., a 15% variation around $R_0^* = 0.2$ gives an R^2 of 60%)



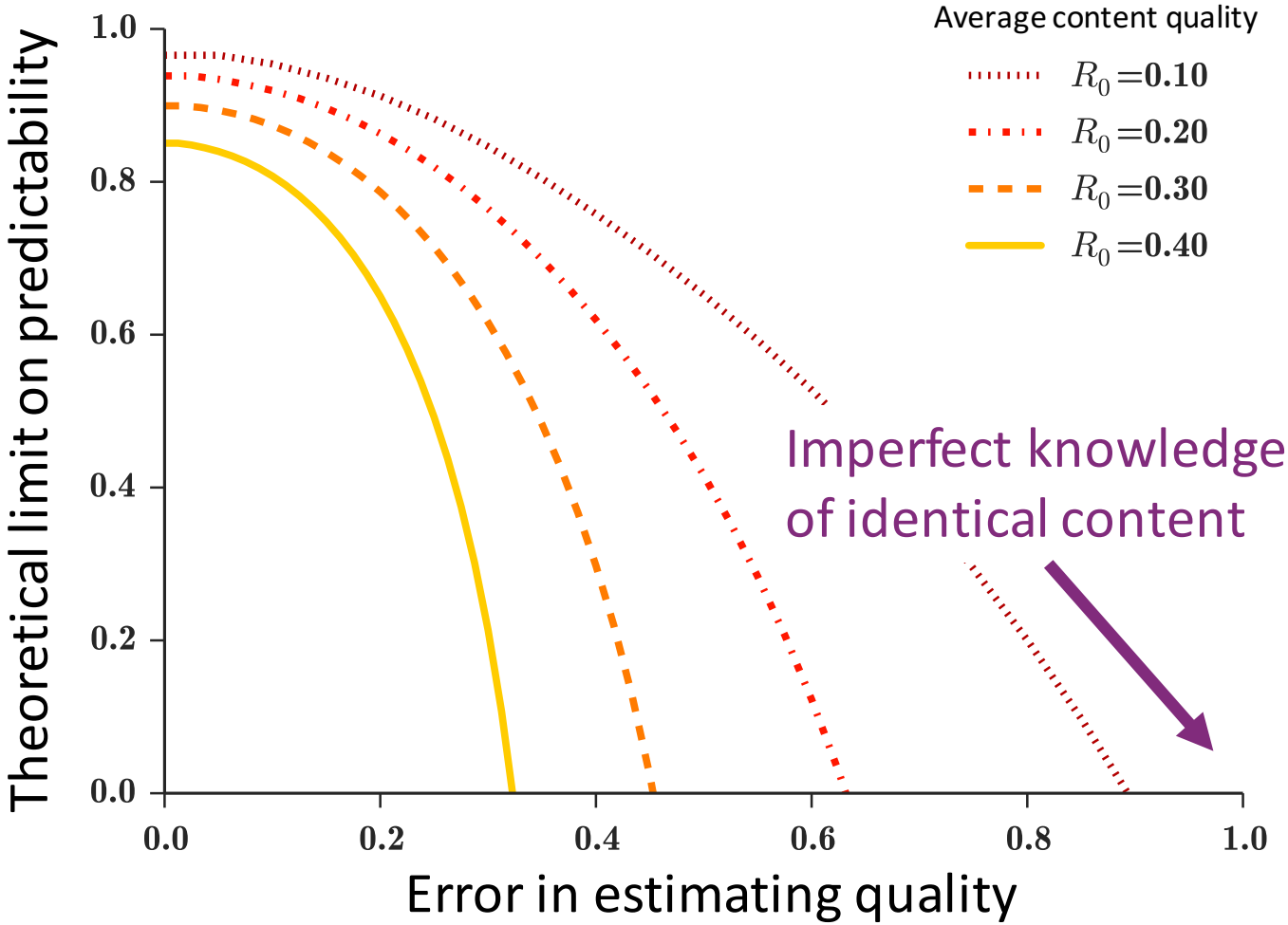
Repeatedly seed the same user with the same content

Outcomes are highly predictable assuming exact quality estimates



Repeatedly seed the same user with the same content

Predictive performance decreases sharply with estimation error
(e.g., $R^2 < 60\%$ with 30% error in estimating $R_0^* = 0.3$)



Conclusions

Conclusions

Most things don't spread, but when they do, it's difficult to predict success

Conclusions

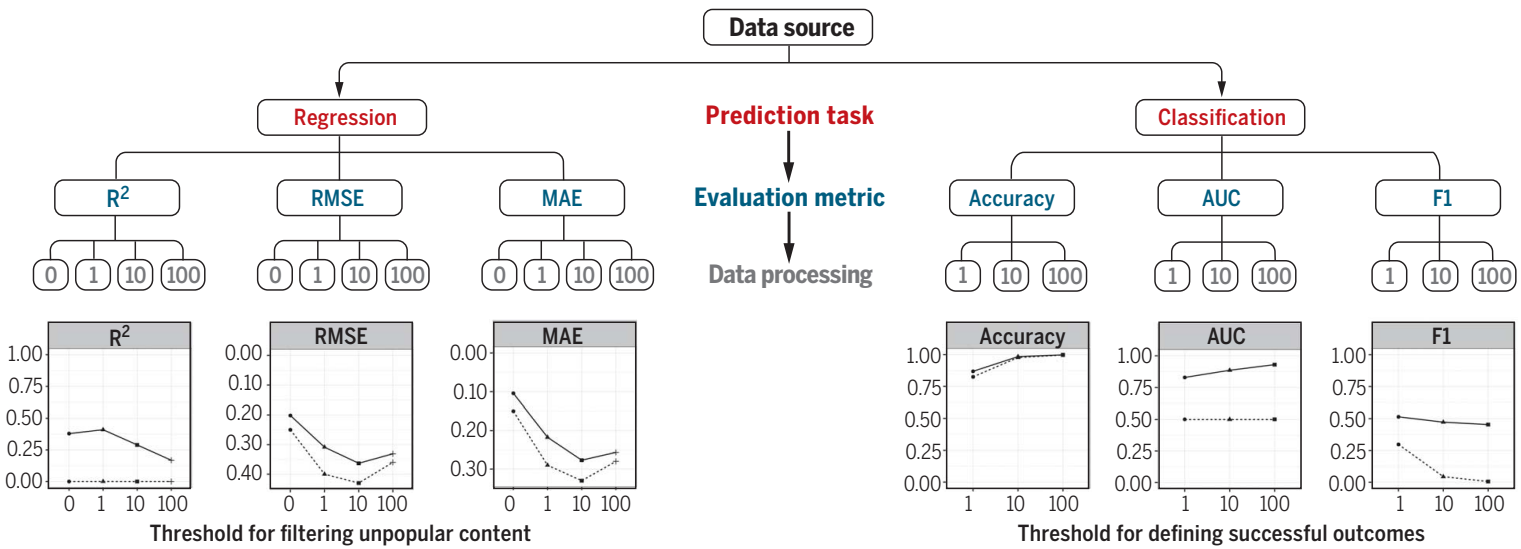
State-of-the-art models explain less than half of the variance in outcomes, based primarily on past success

Conclusions

This performance is likely due to randomness in diffusion process itself, rather than our ability to estimate or model it

Going forward

Focusing solely on prediction can make it difficult to assess long-term progress



Going forward

Important to view prediction and explanation as compliments, not substitutes

Computer science / ML		Social science / Stats
$\hat{y}$	~~vs~~ **and**	$\hat{\beta}$
Predict		Explain

Going forward

Guidelines:

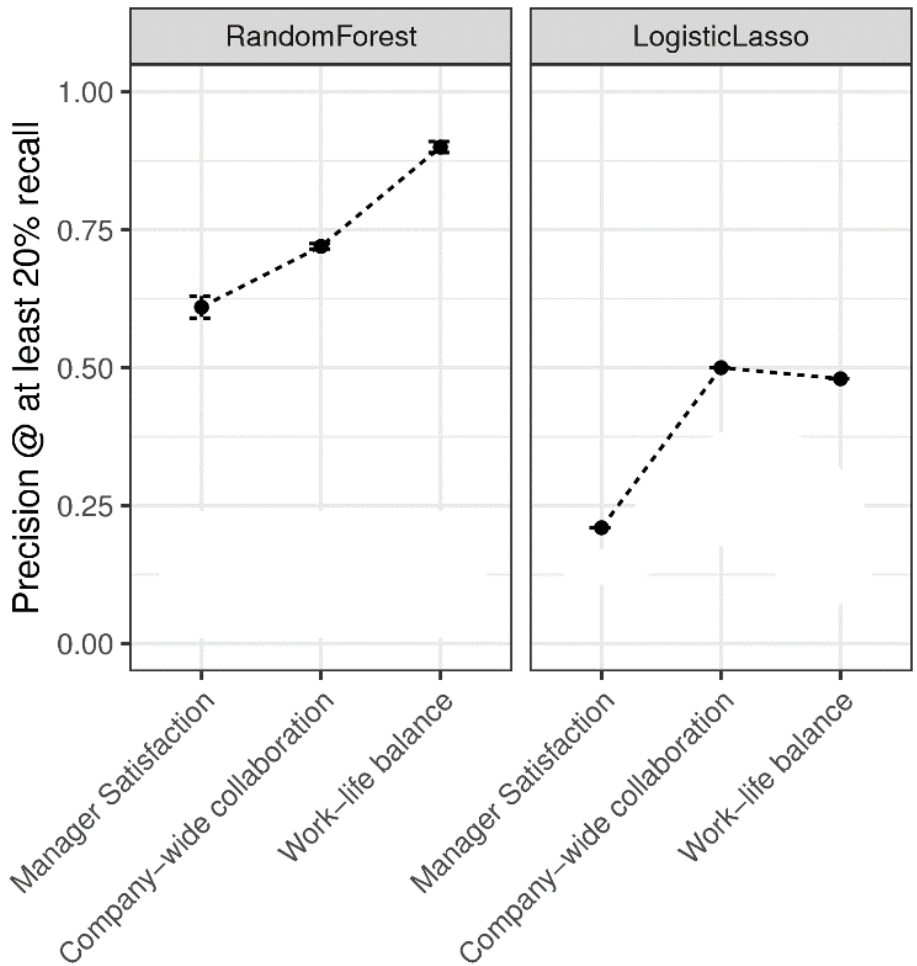
- Improved standards for confirmatory vs. exploratory research (common task framework, pre-registration)
- Focus on models that are both predictive and interpretable
- Entertain complex models for a complex world
- Consider inherent limits to predictability of social phenomena

Example: Predicting employee satisfaction

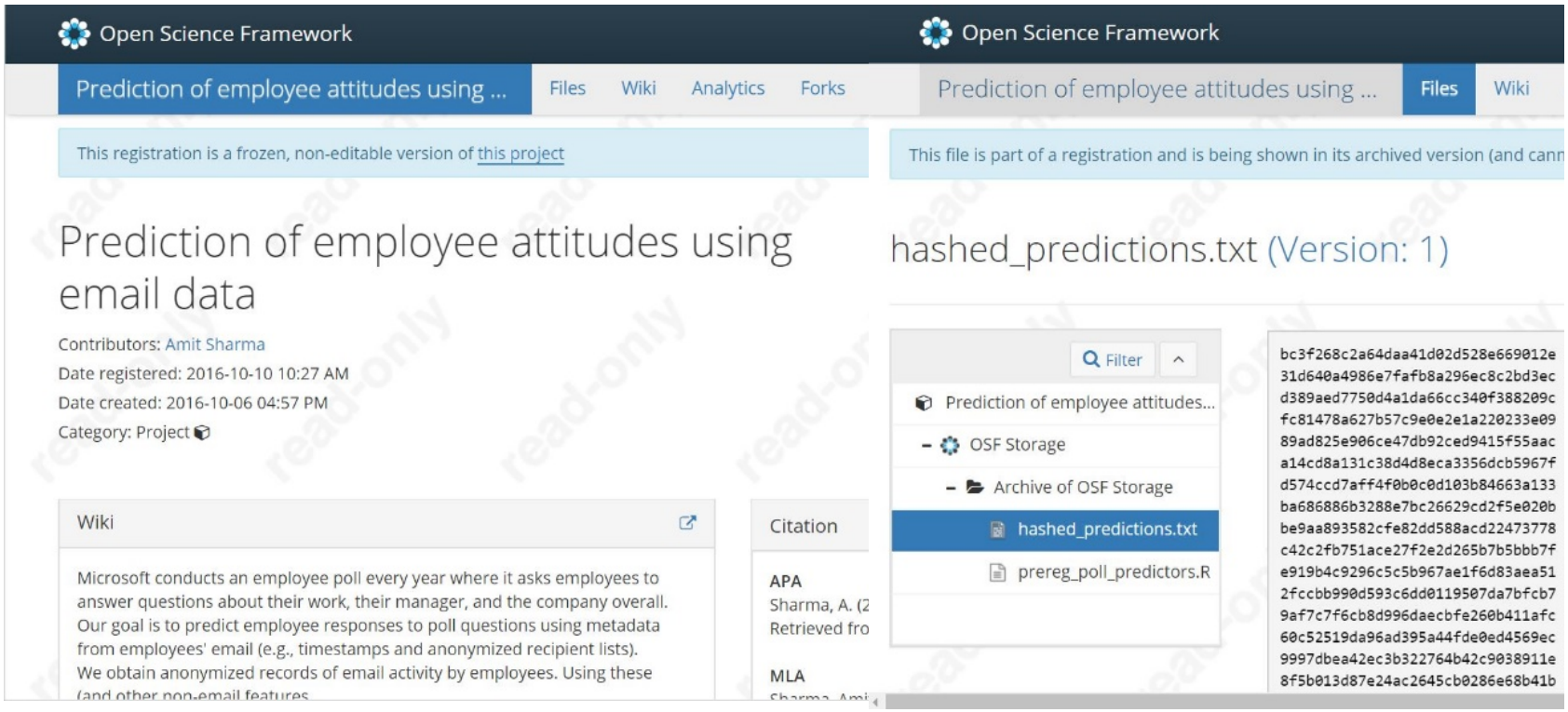
Predicting employee satisfaction on a yearly poll using internal communication data:

- I am satisfied with the level of company-wide collaboration
- I am satisfied with the effectiveness of my manager
- I am satisfied with my work-life balance

Results on the test set in year 1



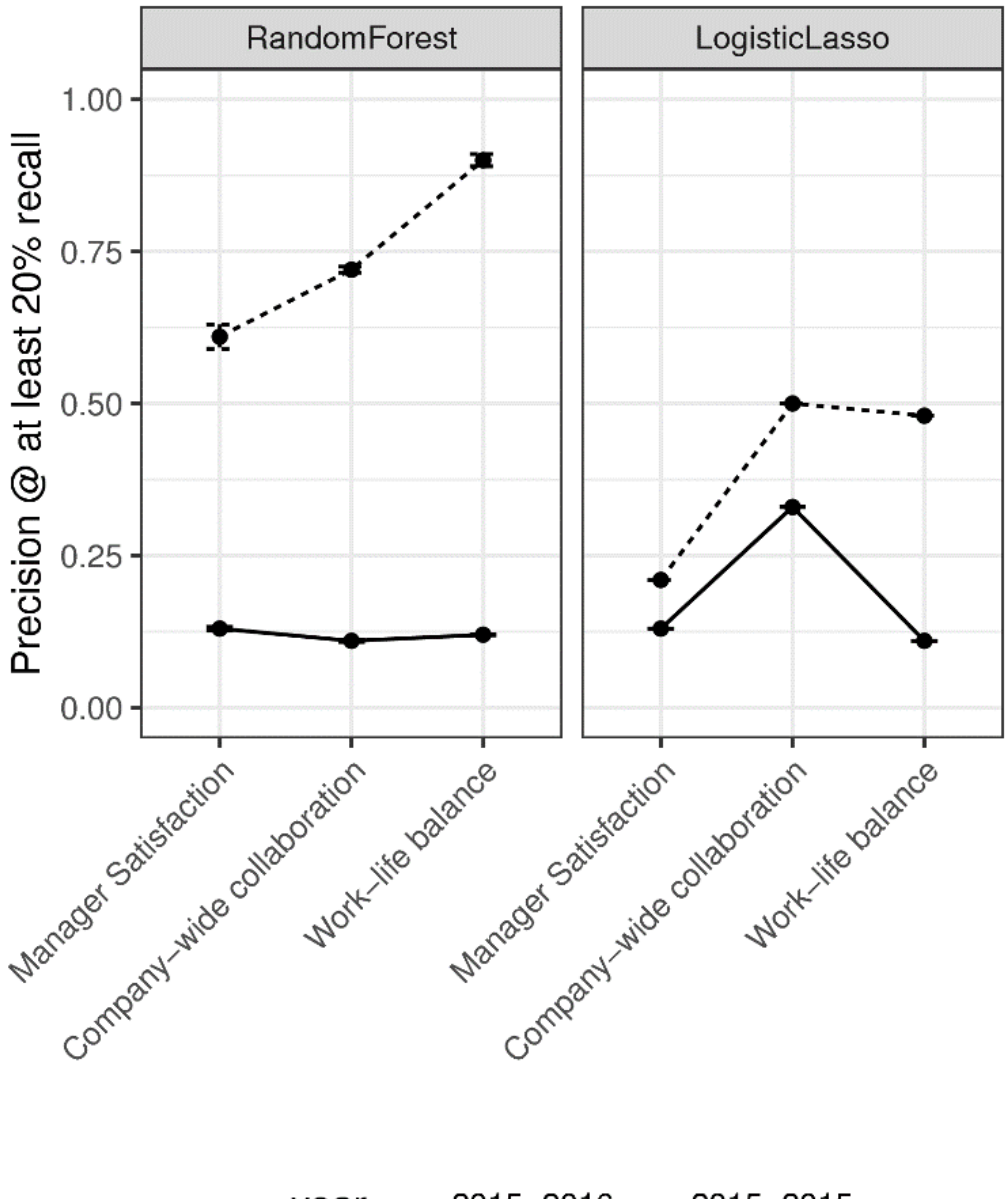
Pre-registering predictive models



Pre-registering predictive models

- Random seed used to initialize pseudo-random generation
- Procedure for dividing data into training, evaluation, test sets
- Outcome variable for prediction
- Features used by the model, along with any transformations
- Model class used with pre-specified hyperparameters
- Metric(s) used for evaluating the trained models predictions

Results on year 2: twice the work, none of the reward



Thanks. Questions?

jmh@microsoft.com

"Exploring limits to prediction in complex social systems"

Travis Martin, Jake Hofman, Amit Sharma, Ashton Anderson, and Duncan Watts

WWW 2016

bit.ly/predlim

"Prediction and explanation in social systems"

Jake Hofman, Amit Sharma, and Duncan Watts

Science 2017

bit.ly/predictexplain