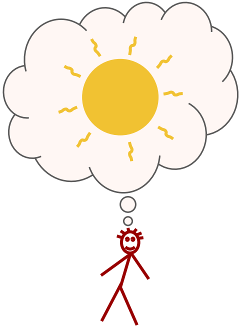
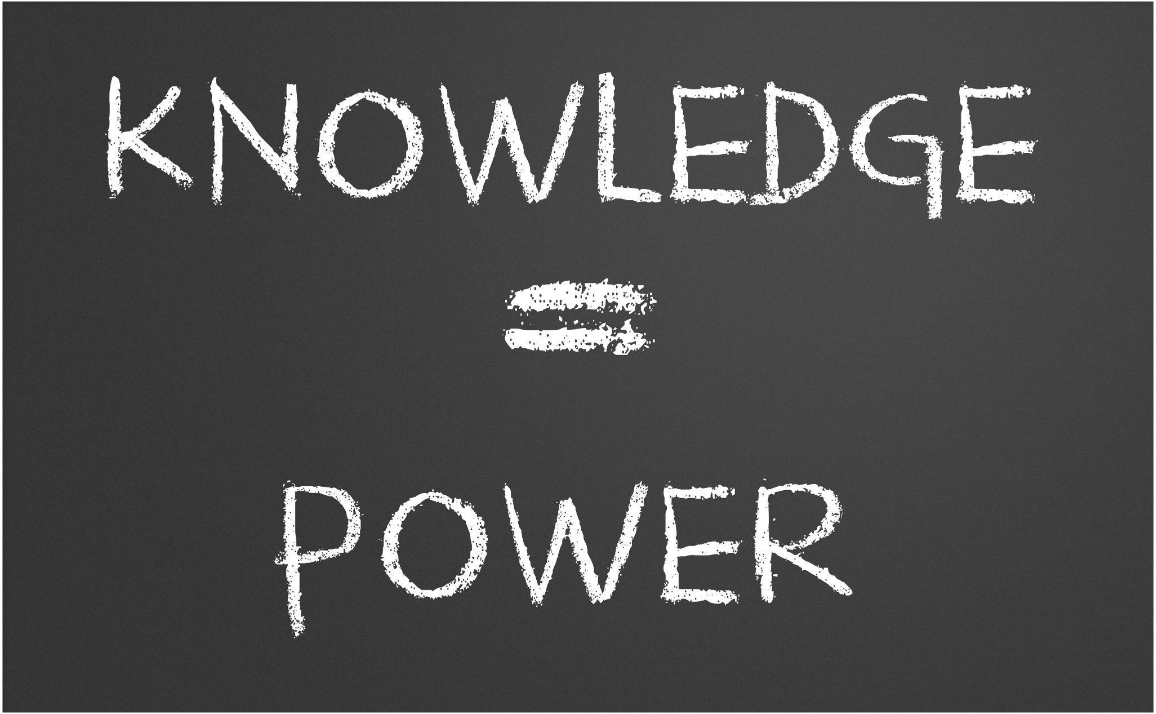
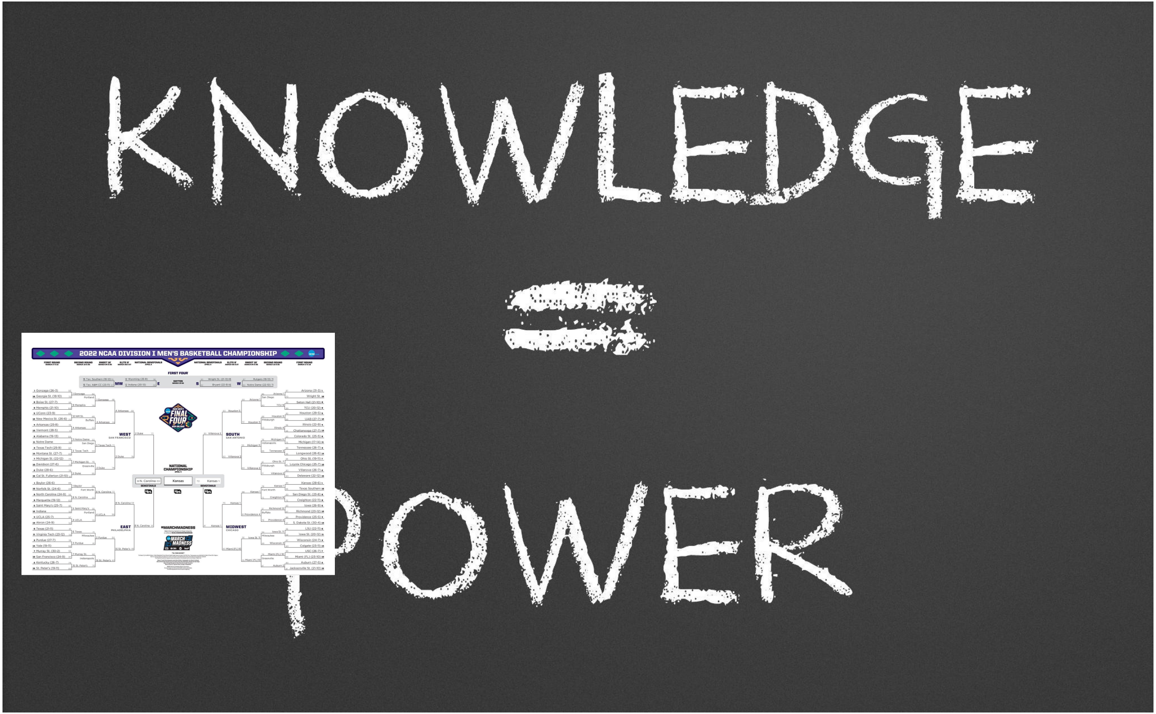
Foundations of Forecasting



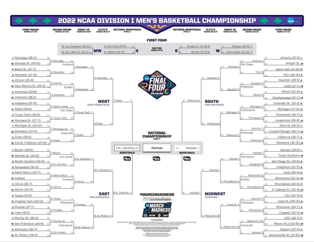
Bo Waggoner
University of Colorado, Boulder

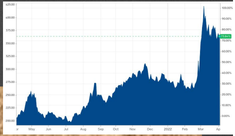
Neumann University
April 6, 2022

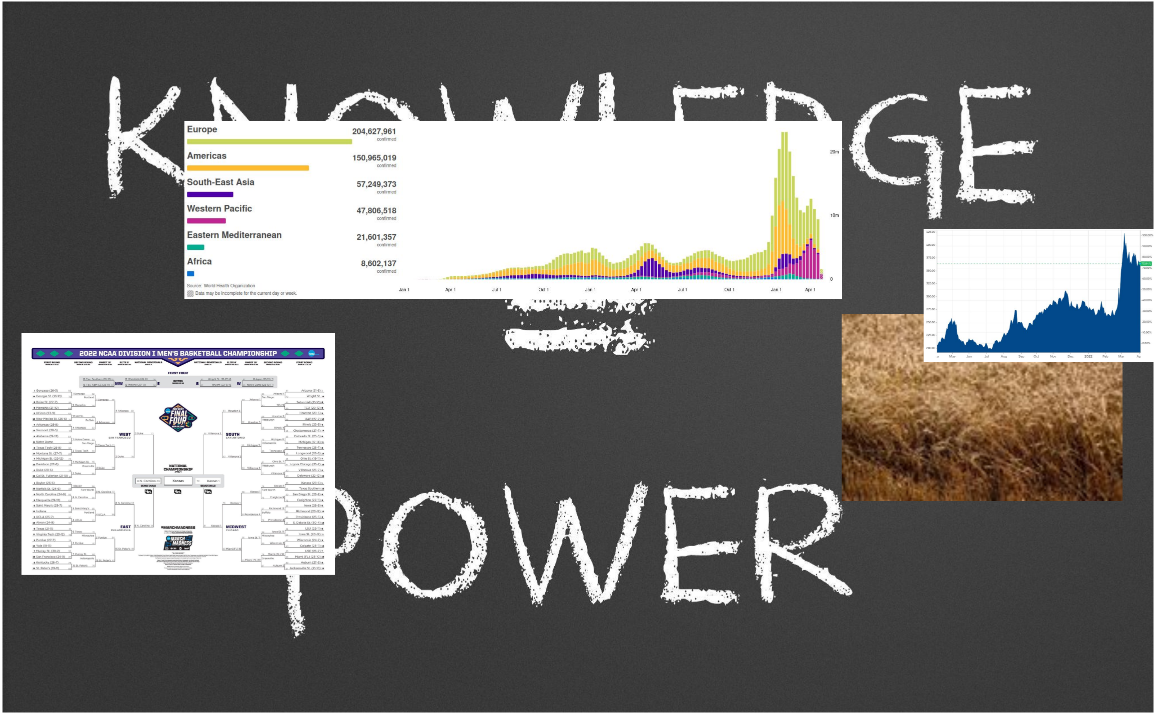

KNOWLEDGE
=
POWER


KNOWLEDGE
=
POWER
2022 NCAA DIVISION I MEN'S BASKETBALL CHAMPIONSHIP

KNOWLEDGE = POWER

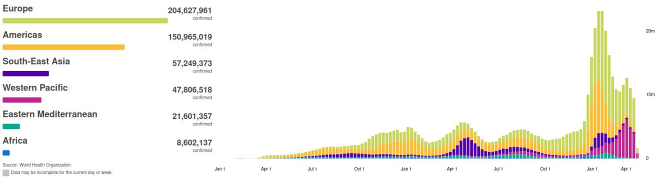


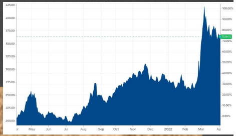


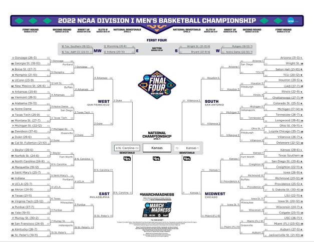


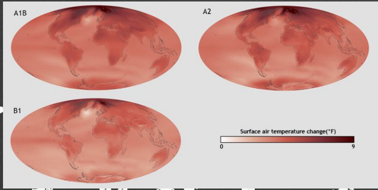
KNOWLEDGE
=
POWER
Europe
204,627,961
confirmed
Americas
150,965,019
confirmed
South-East Asia
57,249,373
confirmed
Western Pacific
47,806,518
confirmed
Eastern Mediterranean
21,601,357
confirmed
Africa
8,602,137
confirmed
Source: World Health Organization
Data may be incomplete for the current day or week.
20m
10m
0
Jan 1
Apr 1
Jul 1
Oct 1
Jan 1
Apr 1
Jul 1
Oct 1
Jan 1
Apr 1
2022 NCAA DIVISION I MEN'S BASKETBALL CHAMPIONSHIP
FIRST FOUR
WEST
SOUTH
EAST
MIDWEST
NATIONAL CHAMPIONSHIP
#MARCHMADNESS

KNOWLEDGE = POWER









Components of **good forecasting:**

Components of **good forecasting:**

- **Evaluation:** checking if a prediction is good

Components of **good forecasting:**

- **Evaluation:** checking if a prediction is good
- **Elicitation:** obtaining good predictions

Components of **good forecasting:**

- **Evaluation:** checking if a prediction is good
- **Elicitation:** obtaining good predictions
- **Aggregation:** combining information into predictions

Components of **good forecasting:**

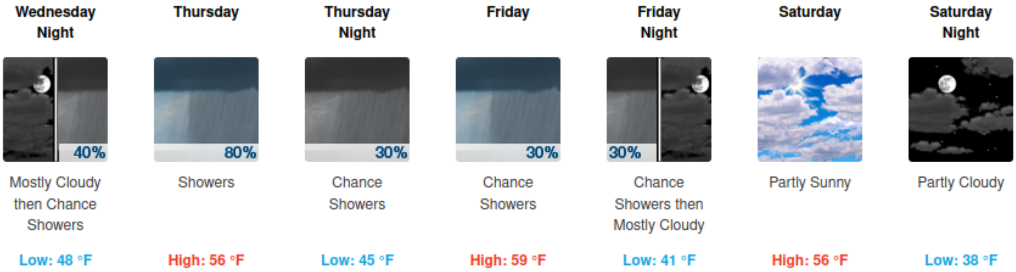
- **Evaluation:** checking if a prediction is good
- **Elicitation:** obtaining good predictions
- **Aggregation:** combining information into predictions
- **Decisionmaking:** making decisions based on predictions

Outline

1. Proper scoring rules
   - Machine learning and loss functions
2. Forecasting in groups
3. Decisionmaking and governance

1. Proper scoring rules

Roots of forecasting



Roots of forecasting

MONTHLY WEATHER REVIEW

EDITOR, JAMES E. CASKEY, JR.

Volume 78
Number 1

JANUARY 1950

Closed March 5, 1950
Issued April 15, 1950

VERIFICATION OF FORECASTS EXPRESSED IN TERMS OF PROBABILITY

GLENN W. BRIER

U. S. Weather Bureau, Washington, D. C.

[Manuscript received February 10, 1950]

INTRODUCTION

Verification of weather forecasts has been a controversial subject for more than a half century. There are a number of reasons why this problem has been so perplexing to meteorologists and others but one of the most important difficulties seems to be in reaching an agreement on the specification of a scale of goodness for weather forecasts. Numerous systems have been proposed but one of the greatest arguments raised against forecast verification is that forecasts which may be the "best" according to the accepted system of arbitrary scores may not be the most useful forecasts. In attempting to resolve this difficulty the forecaster may often find himself in the position of choosing to ignore the verification system or to let it do the forecasting for him by "hedging" or "playing the system." This may lead the forecaster to forecast something other than what he thinks will occur, for it is often easier to analyze the effect of different possible forecasts on the verification score than it is to analyze the weather situation. It is generally agreed that this state of affairs is unsatisfactory, as one essential criterion for satisfactory verification is that the verification scheme should influence numerically have been discussed previously [1, 2, 3, 4] so that the purpose here will not be to emphasize the enhanced usefulness of such forecasts but rather to point out how some aspects of the verification problem are simplified or solved.

VERIFICATION FORMULA

Suppose that on each of n occasions an event can occur in only one of r possible classes or categories and on one such occasion, i, the forecast probabilities are f_{i1}, f_{i2}, . . . f_{ir}, that the event will occur in classes 1, 2, . . . r, respectively. The r classes are chosen to be mutually exclusive and exhaustive so that

$$\sum_{j=1}^{r} f_{ij}=1,\ i=1,2,3,\ldots n \qquad (1)$$

A number of interesting observations can be made about a verification score P defined by

$$P=\frac{1}{n}\sum_{j=1}^{r}\sum_{i=1}^{n}(f_{ij}-E_{ij})^2 \qquad (2)$$

Let’s play a game...

Prediction game: predict a coin toss!

As suggested by Brier: predict *probability* of heads.

Let's play a game...

Prediction game: predict a coin toss!

As suggested by Brier: predict *probability* of heads.

Who had the best prediction?

Proper scoring rules

Brier's solution: a **proper scoring rule**:

Proper scoring rules

Brier's solution: a **proper scoring rule**:

A function $S(p, y)$ where $p =$ prediction and $y =$ observed outcome in $\{0, 1\}$...

Proper scoring rules

Brier's solution: a **proper scoring rule**:

A function $S(p, y)$ where $p =$ prediction and $y =$ observed outcome in $\{0, 1\}$...

...so that the **optimal** prediction is one's true belief.

Proper scoring rules

Brier's solution: a **proper scoring rule**:

A function $S(p, y)$ where $p =$ prediction and $y =$ observed outcome in $\{0, 1\}$...

...so that the **optimal** prediction is one's true belief.

Optimal: maximizes *expected score*.

Proper scoring rules

Brier's solution: a **proper scoring rule**:

A function $S(p, y)$ where $p =$ prediction and $y =$ observed outcome in $\{0, 1\}$...

...so that the **optimal** prediction is one's true belief.

Optimal: maximizes *expected score*.

Example: $S(p, y) = -(y - p)^2$.

From accuracy to game theory (and back)

Squared loss:

- A classic measure of error

From accuracy to game theory (and back)

Squared loss:

- A classic measure of error
- But also **incentivizes truthful forecasts** (is **proper**)

From accuracy to game theory (and back)

Squared loss:

- A classic measure of error
- But also **incentivizes truthful forecasts** (is **proper**)

From accuracy to game theory (and back)

Squared loss:

- A classic measure of error
- But also **incentivizes truthful forecasts** (is **proper**)

Expected score for predicting p when you believe q?
Recall: $S(p, y) = -(y - p)^2$.

From accuracy to game theory (and back)

Squared loss:

- A classic measure of error
- But also **incentivizes truthful forecasts** (is **proper**)

Expected score for predicting p when you believe q?
Recall: $S(p,y) = -(y-p)^2$.

$$S(p;q) \qquad = - \mathop{\mathbb{E}}_{y\sim q} (y-p)^2$$

From accuracy to game theory (and back)

Squared loss:

- A classic measure of error
- But also **incentivizes truthful forecasts** (is **proper**)

Expected score for predicting p when you believe q?
Recall: $S(p, y) = -(y - p)^2$.

$$\begin{aligned} S(p; q) \quad &= -\mathop{\mathbb{E}}_{y \sim q} (y - p)^2 \\ &= -\mathop{\mathbb{E}}_{y \sim q} (y - p + q - q)^2 \end{aligned}$$

From accuracy to game theory (and back)

Squared loss:

- A classic measure of error
- But also **incentivizes truthful forecasts** (is **proper**)

Expected score for predicting p when you believe q?
Recall: $S(p, y) = -(y - p)^2$.

$$
\begin{aligned}
S(p; q) \quad &= - \mathop{\mathbb{E}}_{y \sim q} (y - p)^2 \\
&= - \mathop{\mathbb{E}}_{y \sim q} (y - p + q - q)^2 \\
&= - \mathop{\mathbb{E}}_{y \sim q} \left[(y - q)^2 + (q - p)^2 + 2(y - q)(q - p) \right]
\end{aligned}
$$

From accuracy to game theory (and back)

Squared loss:

- A classic measure of error
- But also **incentivizes truthful forecasts** (is **proper**)

Expected score for predicting p when you believe q?
Recall: $S(p, y) = -(y-p)^2$.

$$
\begin{aligned}
S(p;q) \qquad &= -\mathop{\mathbb{E}}_{y\sim q}(y-p)^2 \\
&= -\mathop{\mathbb{E}}_{y\sim q}(y-p+q-q)^2 \\
&= -\mathop{\mathbb{E}}_{y\sim q}\left[(y-q)^2+(q-p)^2+2(y-q)(q-p)\right] \\
&= -\mathop{\mathbb{E}}_{y\sim q}(y-q)^2 \ - \ (q-p)^2
\end{aligned}
$$

From accuracy to game theory (and back)

Squared loss:

- A classic measure of error
- But also **incentivizes truthful forecasts** (is **proper**)

Expected score for predicting p when you believe q?
Recall: $S(p, y) = -(y-p)^2$.

$$
\begin{aligned}
S(p; q) \quad &= -\mathop{\mathbb{E}}_{y \sim q} (y-p)^2 \\
&= -\mathop{\mathbb{E}}_{y \sim q} (y-p+q-q)^2 \\
&= -\mathop{\mathbb{E}}_{y \sim q} \left[(y-q)^2 + (q-p)^2 + 2(y-q)(q-p) \right] \\
&= -\mathop{\mathbb{E}}_{y \sim q} (y-q)^2 \; - \; (q-p)^2 \\
&= -\mathrm{Var}(q) - (q-p)^2
\end{aligned}
$$

From accuracy to game theory (and back)

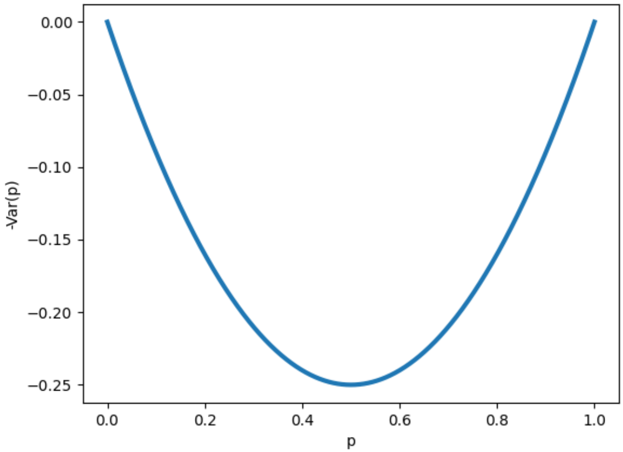
Squared loss:

- A classic measure of error
- But also **incentivizes truthful forecasts** (is **proper**)

Expected score for predicting p when you believe q?
Recall: $S(p,y) = -(y-p)^2$.

$$\begin{aligned}
S(p;q) &= -\mathop{\mathbb{E}}_{y\sim q}(y-p)^2 \\
&= -\mathop{\mathbb{E}}_{y\sim q}(y-p+q-q)^2 \\
&= -\mathop{\mathbb{E}}_{y\sim q}\left[(y-q)^2+(q-p)^2+2(y-q)(q-p)\right] \\
&= -\mathop{\mathbb{E}}_{y\sim q}(y-q)^2 \ - \ (q-p)^2 \\
&= -\mathsf{Var}(q)-(q-p)^2 \\
&\leq -\mathsf{Var}(q).
\end{aligned}$$

Expected score: negative variance



Another proper scoring rule

Good (1952): The scoring rule $S(p, y) = \begin{cases} \log(p) & y = 1 \\ \log(1-p) & y = 0 \end{cases}$.

Another proper scoring rule

Good (1952): The scoring rule $S(p, y) = \begin{cases} \log(p) & y = 1 \\ \log(1-p) & y = 0 \end{cases}$.

Expected score for predicting p when you believe q?

Another proper scoring rule

Good (1952): The scoring rule $S(p, y) = \begin{cases} \log(p) & y = 1 \\ \log(1-p) & y = 0 \end{cases}$.

Expected score for predicting p when you believe q?

$$S(p; q) \qquad = q\log(p) + (1-q)\log(1-p)$$

Another proper scoring rule

Good (1952): The scoring rule $S(p, y) = \begin{cases} \log(p) & y = 1 \\ \log(1-p) & y = 0 \end{cases}$.

Expected score for predicting p when you believe q?

$$\begin{aligned} S(p; q) \qquad &= q \log(p) + (1-q)\log(1-p) \\ &= -\mathsf{H}(q) + q\log(\tfrac{p}{q}) + (1-q)\log(\tfrac{1-p}{1-q}) \end{aligned}$$

Another proper scoring rule

Good (1952): The scoring rule $S(p, y) = \begin{cases} \log(p) & y = 1 \\ \log(1 - p) & y = 0 \end{cases}$.

Expected score for predicting p when you believe q?

$$\begin{aligned} S(p; q) &= q \log(p) + (1 - q) \log(1 - p) \\ &= -\mathsf{H}(q) + q \log(\tfrac{p}{q}) + (1 - q) \log(\tfrac{1-p}{1-q}) \\ &= -\mathsf{H}(q) + \mathsf{KL}(q, p) \end{aligned}$$

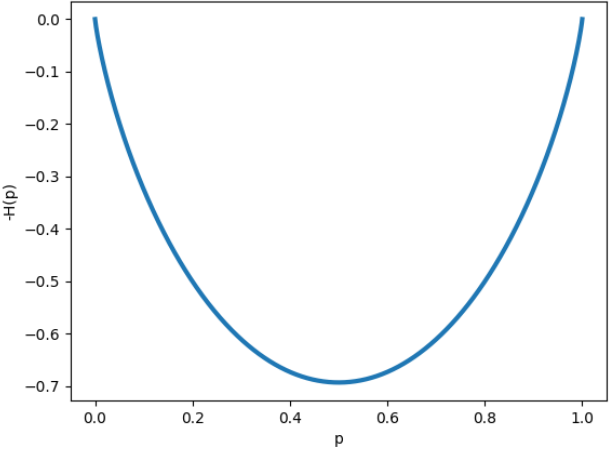
Another proper scoring rule

Good (1952): The scoring rule $S(p, y) = \begin{cases} \log(p) & y = 1 \\ \log(1-p) & y = 0 \end{cases}$.

Expected score for predicting p when you believe q?

$$\begin{aligned} S(p; q) \quad &= q \log(p) + (1-q)\log(1-p) \\ &= -\mathsf{H}(q) + q\log(\tfrac{p}{q}) + (1-q)\log(\tfrac{1-p}{1-q}) \\ &= -\mathsf{H}(q) + \mathsf{KL}(q, p) \\ &\leq -\mathsf{H}(q). \end{aligned}$$

Expected score: negative entropy



Characterization of proper scoring rules

Amazing fact: [McCarthy 1956; Savage 1971; Schervish 1988; Gneiting & Raftery 2007; etc]

Theorem

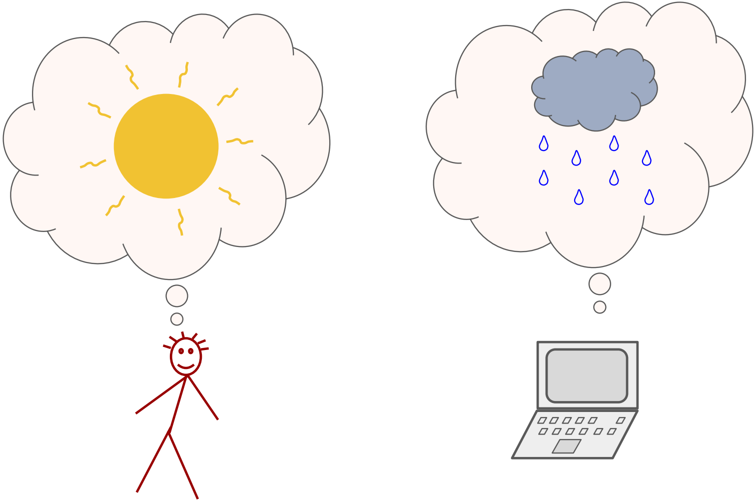
A scoring rule $S(p, y)$ is **proper** *if and only if there exists a convex function G such that*

$$S(p, y) = G(p) + \nabla G(p) \cdot (y - p).$$

1. Proper scoring rules

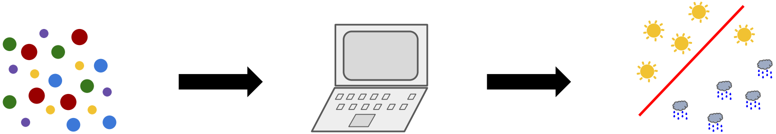
* Machine learning and loss functions

Is an algorithm's prediction different than a human's?



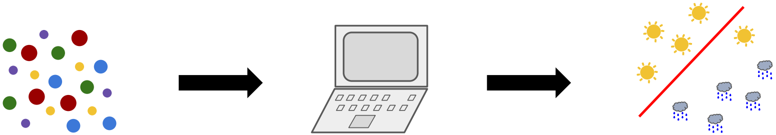
Training algorithms using score (loss)

1. Ask the model to make a prediction p on a data point.



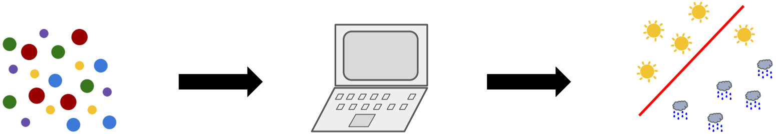
Training algorithms using score (loss)

1. Ask the model to make a prediction p on a data point.
2. Assess its loss $\ell(p, y)$.



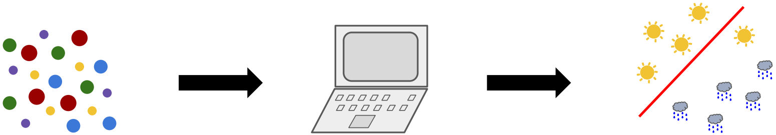
Training algorithms using score (loss)

1. Ask the model to make a prediction p on a data point.
2. Assess its loss $\ell(p, y)$.
3. Adjust the model.



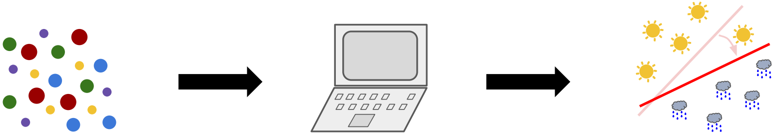
Training algorithms using score (loss)

1. Ask the model to make a prediction p on a data point.
2. Assess its loss $\ell(p, y)$.
3. Adjust the model.
4. Repeat.



Training algorithms using score (loss)

1. Ask the model to make a prediction p on a data point.
2. Assess its loss $\ell(p, y)$.
3. Adjust the model.
4. Repeat.



Choosing a loss

What loss function should we use?

Choosing a loss

What loss function should we use?

$\implies$ As we **train**, the **optimal prediction** should converge to the **truth**.

Choosing a loss

What loss function should we use?

$\implies$ As we **train**, the **optimal prediction** should converge to the **truth**.

What does the **loss** converge to?

Choosing a loss

What loss function should we use?

$\implies$ As we **train**, the **optimal prediction** should converge to the **truth**.

What does the **loss** converge to?

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \ell(p, y_i) \to \mathop{\mathbb{E}}_{y\sim q} \ell(p, y).$$

Choosing a loss

What loss function should we use?

$\implies$ As we **train**, the **optimal prediction** should converge to the **truth**.

What does the **loss** converge to?

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \ell(p, y_i) \to \mathop{\mathbb{E}}_{y\sim q} \ell(p, y).$$

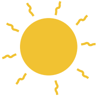
$\implies$ For statistical consistency, we should use a (negated) proper scoring rule!

Predictions other than probabilities

(1) Labels: rain or sun?

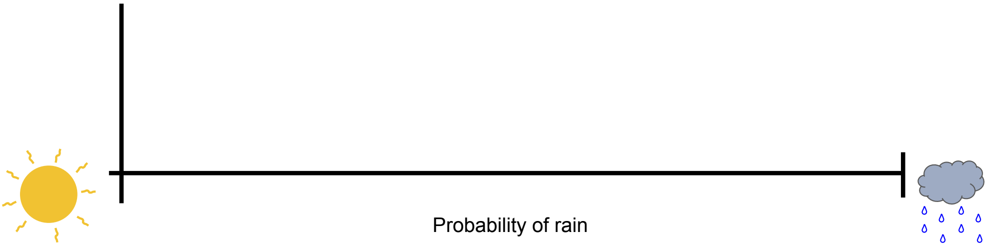
Predictions other than probabilities

(1) Labels: rain or sun?



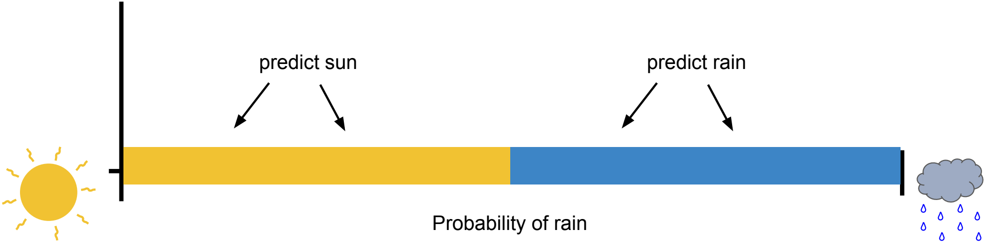
Predictions other than probabilities

(1) Labels: rain or sun?



Predictions other than probabilities

(1) Labels: rain or sun?



Predictions other than probabilities

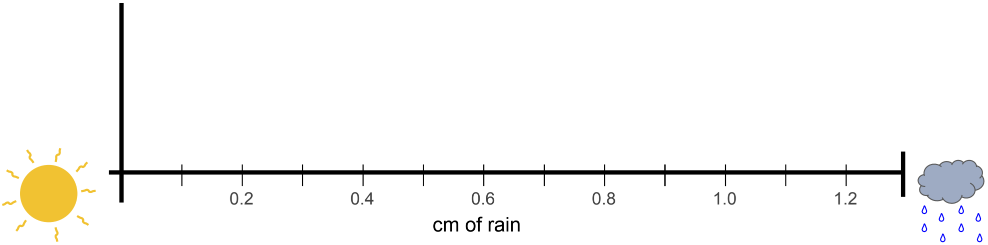
(1) Labels: rain or sun?

(2) Numbers: cm of rain?

Predictions other than probabilities

(1) Labels: rain or sun?

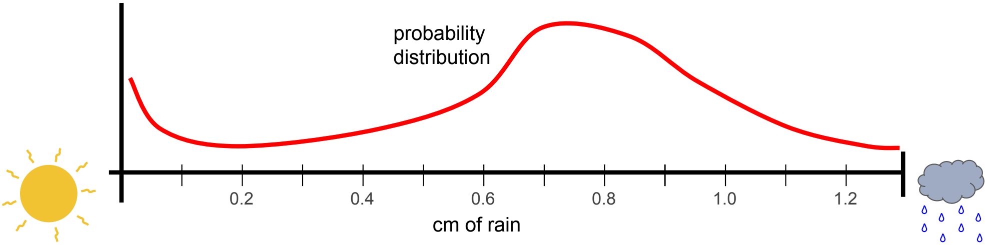
(2) Numbers: cm of rain?



Predictions other than probabilities

(1) Labels: rain or sun?

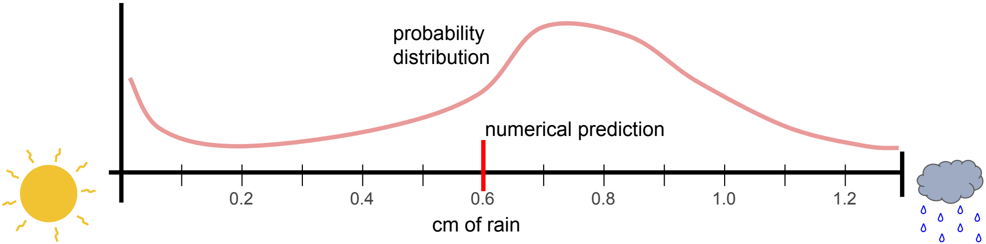
(2) Numbers: cm of rain?



Predictions other than probabilities

(1) Labels: rain or sun?

(2) Numbers: cm of rain?

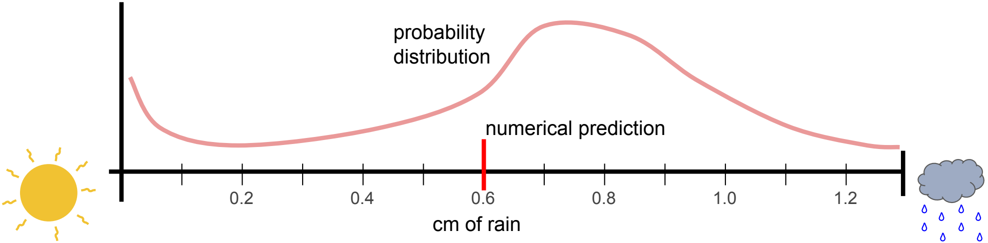


Scoring rules for numerical predictions

How to evaluate a numerical prediction?

a Absolute error, $|p - y|$.

b Squared error, $(p - y)^2$.



Proper scoring rules - conclusion

A. How can we **evaluate** and **elicit** forecasts?

$\Longrightarrow$ **Proper scoring rules** such as the log and Brier score (squared loss).

Proper scoring rules - conclusion

A. How can we **evaluate** and **elicit** forecasts?

$\Longrightarrow$ **Proper scoring rules** such as the log and Brier score (squared loss).

B. What scoring rules are **proper**?

$\Longrightarrow$ Derived from **convex functions**, which represent entropy/uncertainty of the forecast.

Proper scoring rules - conclusion

A. How can we **evaluate** and **elicit** forecasts?

$\Longrightarrow$ **Proper scoring rules** such as the log and Brier score (squared loss).

B. What scoring rules are **proper**?

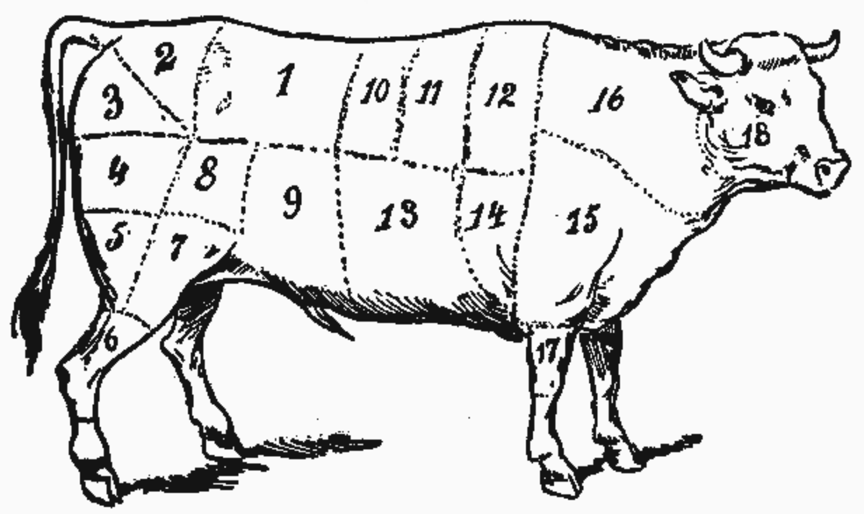
$\Longrightarrow$ Derived from **convex functions**, which represent entropy/uncertainty of the forecast.

C. How should we **train algorithms**?

$\Longrightarrow$ also **proper scoring rules**!

2. Forecasting in groups

Wisdom of the crowd



Beyond wisdom of the crowd

How to aggregate information?

Goals:

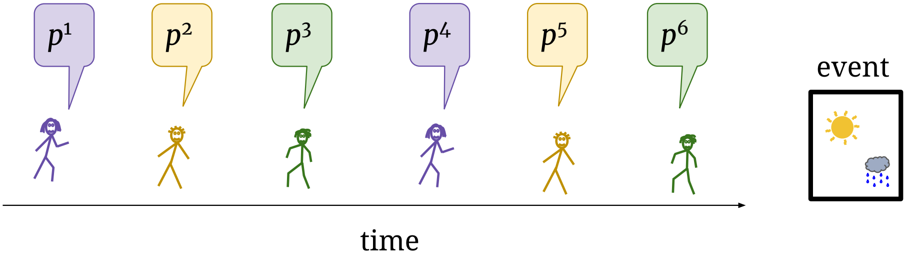
How to aggregate information?

Goals:

- Incentivize **each participant** to provide information
- Handle different **types** of information
- Handle different **strengths** of beliefs

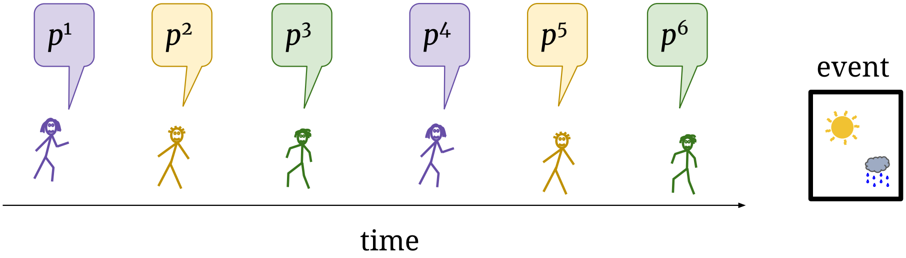
Scoring rule prediction market (Hanson 2003)

- Participants **take turns predicting**.



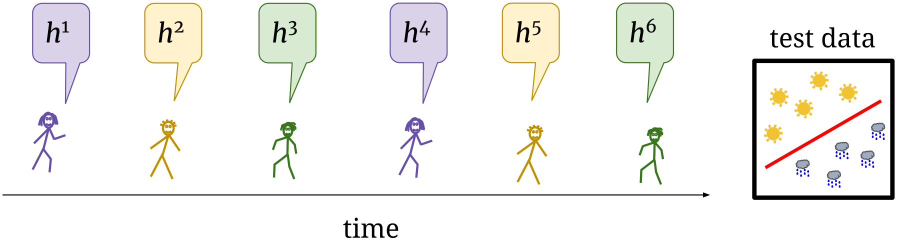
Scoring rule prediction market (Hanson 2003)

- Participants **take turns predicting**.
- After the event, reward is **improvement in score**.
  $S(p^t, y) - S(p^{t-1}, y)$.



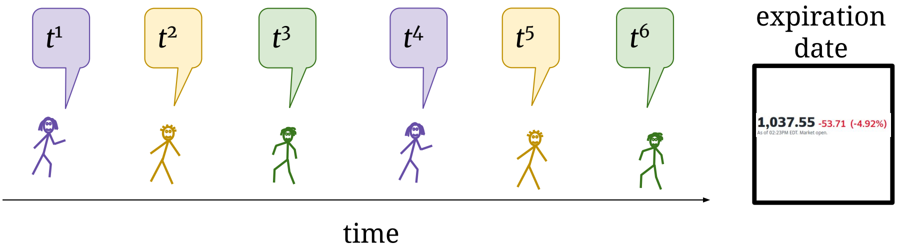
ML collaborative contests (Abernethy, Frongillo 2011)

- Participants **take turns providing models**.
- After the event, reward is **improvement in test score**.



Similarity to financial markets

- Participants **take turns trading**.
- After the event, reward is **net payment**.



Design questions for scoring rule markets

- Can we use SRMs for **label** predictions?

Design questions for scoring rule markets

- Can we use SRMs for **label** predictions?

Design questions for scoring rule markets

- Can we use SRMs for **label** predictions? *Not really!*
- Can we use SRMs for **numerical (mean)** predictions?

Design questions for scoring rule markets

- Can we use SRMs for **label** predictions? *Not really!*
- Can we use SRMs for **numerical (mean)** predictions?

Design questions for scoring rule markets

- Can we use SRMs for **label** predictions? *Not really!*
- Can we use SRMs for **numerical (mean)** predictions? *Yes!*
- Can we use SRMs for **numerical (median)** predictions?

Design questions for scoring rule markets

- Can we use SRMs for **label** predictions? *Not really!*
- Can we use SRMs for **numerical (mean)** predictions? *Yes!*
- Can we use SRMs for **numerical (median)** predictions?

Design questions for scoring rule markets

- Can we use SRMs for **label** predictions? *Not really!*
- Can we use SRMs for **numerical (mean)** predictions? *Yes!*
- Can we use SRMs for **numerical (median)** predictions? *Sort of!*

Group forecasting - conclusion

A. How can we **evaluate** and **elicit** forecasts **from groups**?

$\Longrightarrow$ Design **prediction markets** based on **proper scoring rules**.

Group forecasting - conclusion

A. How can we **evaluate** and **elicit** forecasts **from groups**?

$\Longrightarrow$ Design **prediction markets** based on **proper scoring rules**.

B. What encourages **good group forecasting**?

$\Longrightarrow$ **Sharing** and **iterativel updating** information and predictions.

Group forecasting - conclusion

A. How can we **evaluate** and **elicit** forecasts **from groups**?

$\Longrightarrow$ Design **prediction markets** based on **proper scoring rules**.

B. What encourages **good group forecasting**?

$\Longrightarrow$ **Sharing** and **iterativel updating** information and predictions.

C. What **else** can prediction market designs be used for?

$\Longrightarrow$ Understanding financial markets, designing collaborative contests.

3. Decisionmaking

Decisionmaking in groups

There are **many** paradigms for **decisionmaking in groups**.

Decisionmaking in groups

There are **many** paradigms for **decisionmaking in groups**.

- Direct democracy - voting

Decisionmaking in groups

There are **many** paradigms for **decisionmaking in groups**.

- Direct democracy - voting
- Representative democracy

Decisionmaking in groups

There are **many** paradigms for **decisionmaking in groups**.

- Direct democracy - voting
- Representative democracy
- Corporate structure - delegating authority

Decisionmaking in groups

There are **many** paradigms for **decisionmaking in groups**.

- Direct democracy - voting
- Representative democracy
- Corporate structure - delegating authority
- ...

Decisionmaking in groups

There are **many** paradigms for **decisionmaking in groups**.

- Direct democracy - voting
- Representative democracy
- Corporate structure - delegating authority
- ...

Decisionmaking in groups

There are **many** paradigms for **decisionmaking in groups**.

- Direct democracy - voting
- Representative democracy
- Corporate structure - delegating authority
- ...

Decisions need two inputs:

1. Preferences

Decisionmaking in groups

There are **many** paradigms for **decisionmaking in groups**.

- Direct democracy - voting
- Representative democracy
- Corporate structure - delegating authority
- ...

Decisions need two inputs:

1. Preferences
2. Information

Preferences and information

Using predictions for decisionmaking

Can we incorporate **forecasting** in group decisionmaking?

Using predictions for decisionmaking

Can we incorporate **forecasting** in group decisionmaking?

Challenges:

- Gathering the information

Using predictions for decisionmaking

Can we incorporate **forecasting** in group decisionmaking?

Challenges:

- Gathering the information
- Aggregating it into forecasts

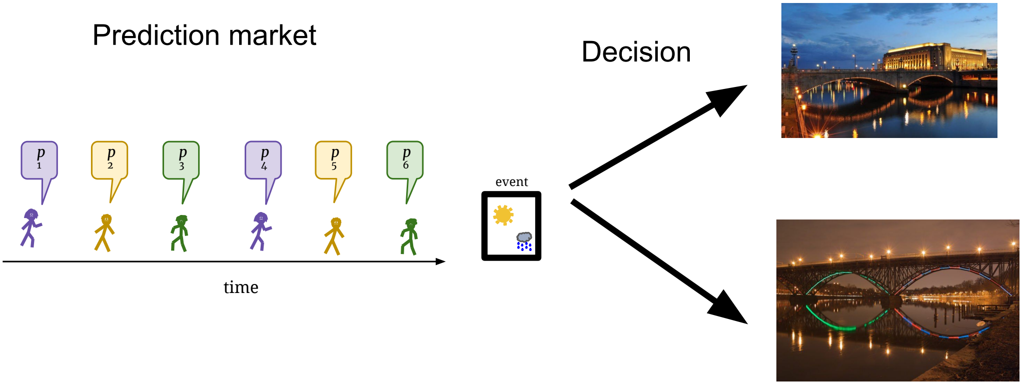
Using predictions for decisionmaking

Can we incorporate **forecasting** in group decisionmaking?

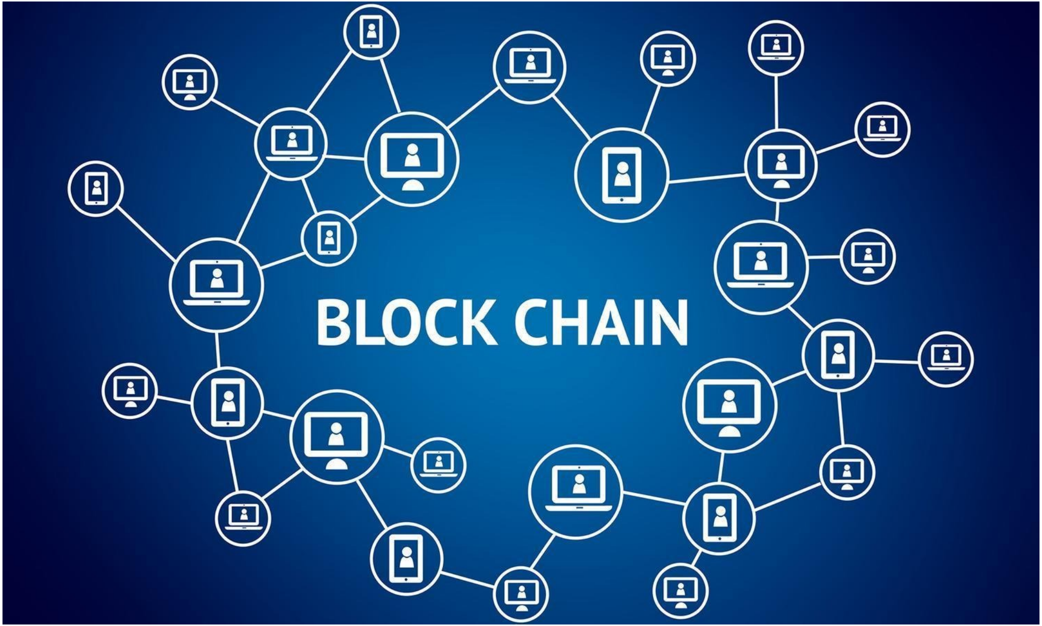
Challenges:

- Gathering the information
- Aggregating it into forecasts
- Incorporating forecasts **and** preferences

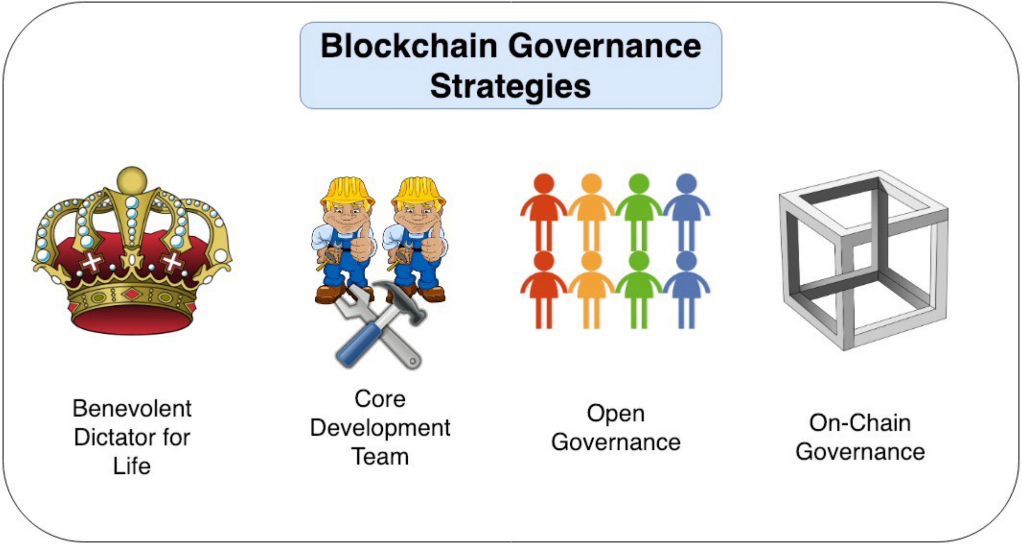
Decision markets



Blockchain applications



Blockchain applications



Conclusions

Conclusions

- It is important to society to **evaluate** forecasts

Conclusions

- It is important to society to **evaluate** forecasts
- How? **proper scoring rules**

Conclusions

- It is important to society to **evaluate** forecasts
- How? **proper scoring rules**
- Applications in **machine learning** and connections to **game theory**

Conclusions

- It is important to society to **evaluate** forecasts
- How? **proper scoring rules**
- Applications in **machine learning** and connections to **game theory**
- Building blocks for **group forecasting** (prediction markets) ...

Conclusions

- It is important to society to **evaluate** forecasts
- How? **proper scoring rules**
- Applications in **machine learning** and connections to **game theory**
- Building blocks for **group forecasting** (prediction markets) ...
- ... and decisionmaking / **governance** proposals.

Conclusions

- It is important to society to **evaluate** forecasts
- How? **proper scoring rules**
- Applications in **machine learning** and connections to **game theory**
- Building blocks for **group forecasting** (prediction markets) ...
- ... and decisionmaking / **governance** proposals.

Conclusions

- It is important to society to **evaluate** forecasts
- How? **proper scoring rules**
- Applications in **machine learning** and connections to **game theory**
- Building blocks for **group forecasting** (prediction markets) ...
- ... and decisionmaking / **governance** proposals.

Thanks to mentors and collaborators, esp. Yiling Chen and Raf Frongillo.

Thanks!