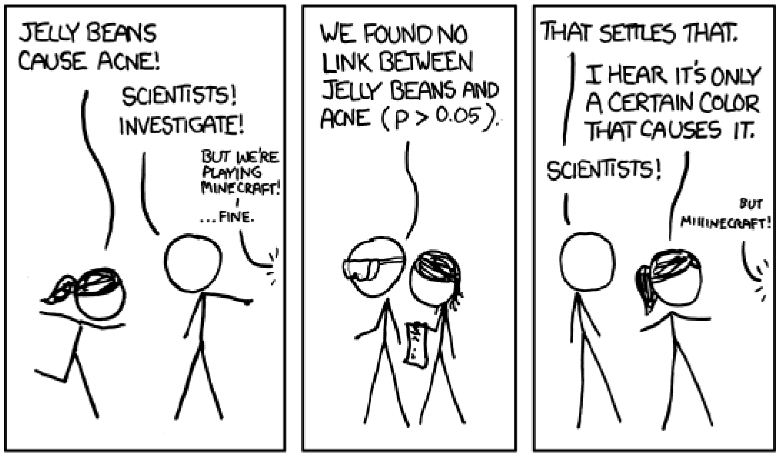
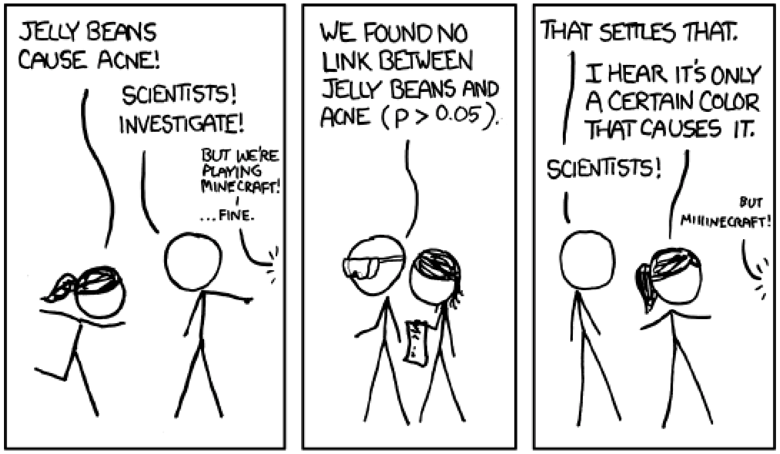
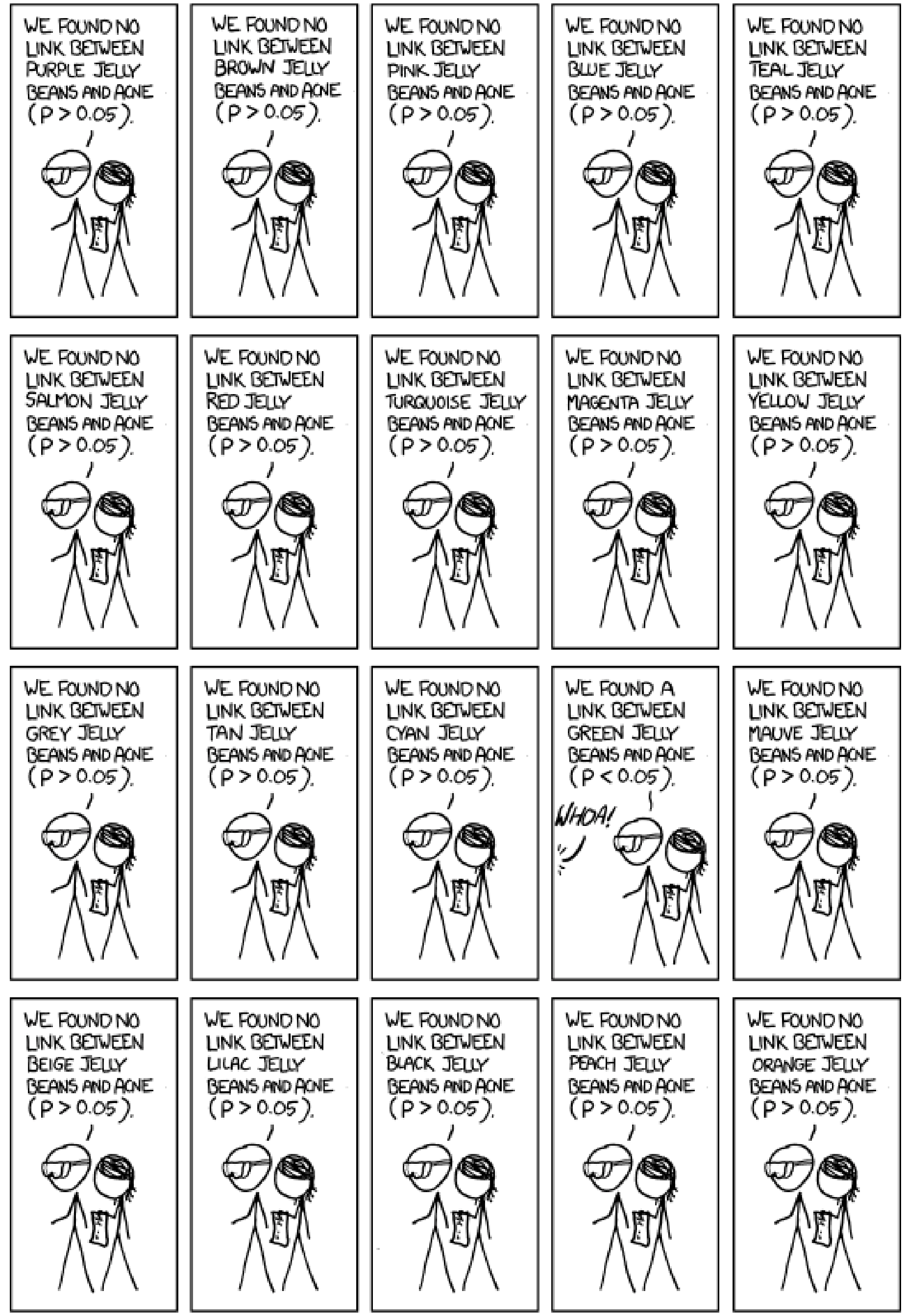
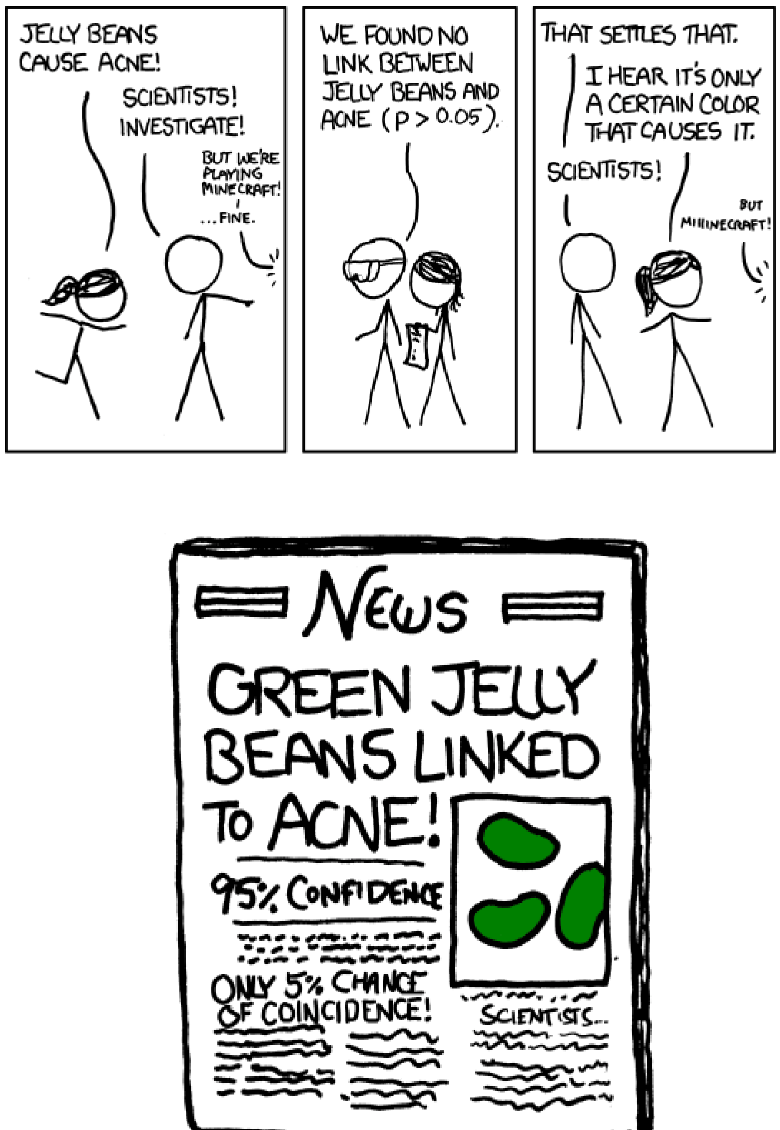
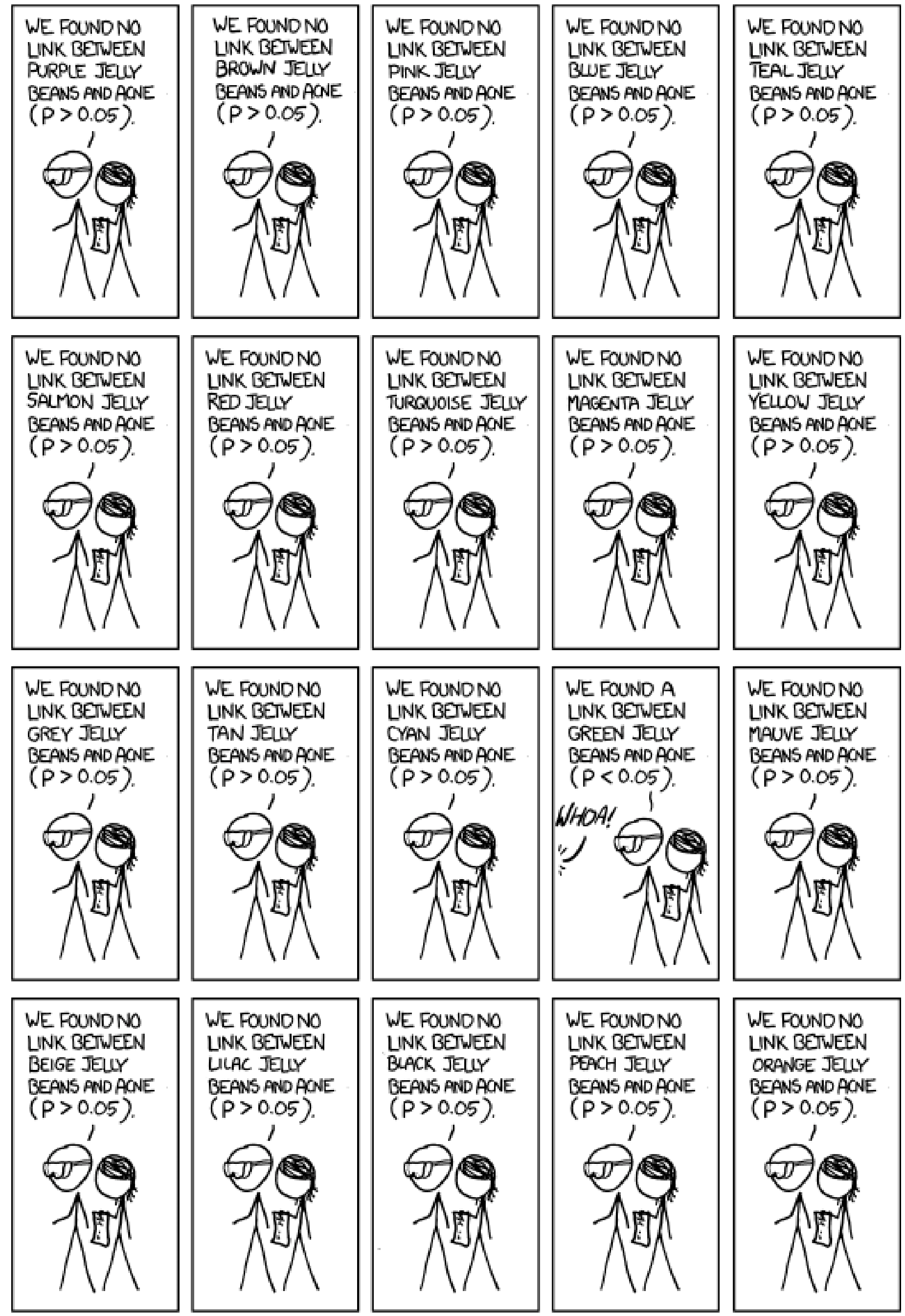

JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
... FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
... FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!


WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
...FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!
NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...


WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).

Xkcd.org

Lucky Ockham

Peter Grünwald

Centrum Wiskunde & Informatica – Amsterdam

Mathematisch Instituut – Universiteit Leiden

Plan of the Talk

- I will describe three settings of inductive inference studied by *many* machine learning theorists and *some* statisticians
  – on-line sequential prediction without stochastics
  – statistical learning with “oracle bounds”
  – statistical learning with “empirical bounds”

Vladimir Vovk

Vladimir Vapnik

Plan of the Talk

- I will describe three settings of inductive inference studied by *many* machine learning theorists and *some* statisticians
  - on-line sequential prediction without stochastics
  - statistical learning with “oracle bounds”
  - statistical learning with “empirical bounds”
- In all three settings a particular form of **Ockham’s razor** plays a crucial role. Goals of the talk are
  1. to introduce these settings to philosophers
  2. to thereby highlight the importance of this form of Ockham’s razor
  3. To argue some specific things about Ockham...

Plan of the Talk

- I will describe three settings of inductive inference studied by *many* machine learning theorists and *some* statisticians
  - on-line sequential prediction without stochastics
  - statistical learning with "oracle bounds"
  - statistical learning with "empirical bounds"

Why these three?

- In all three settings a particular form of **Ockham's razor** plays a crucial role. Goals of the talk are
  1. to introduce these settings to philosophers
  2. to thereby highlight the importance of this form of Ockham's razor
  3. To argue some specific things about Ockham...

Form of Ockham's Razor

- In all three settings, one gets tight bounds on performance of algorithms which involve trade-off between *error term* and *codelength* or *minus log prior* term

$$-\log W(\theta) \qquad \textit{(always > 0)}$$

- can be interpreted as precise form of Occam's razor:
  - if one uses a "complex" model (many bits needed to encode hypothesis) one needs more data before one gets good performance (because one has to counter overfitting)

Three Extreme Positions

- **BAYES:** All these prior-dependent methods are essentially Bayesian, which is as it should be
- **NFL:** These and other description-length/prior-based notions of Ockham's razor are essentially **arbitrary**, because you can make any hypothesis arbitrarily 'simple' or 'complex' by changing the prior
- **MDL/Kolmogorov:** By choosing the "right" priors, these methods can be made "fully objective"

Three Extreme Positions

- **BAYES:** All these prior-dependent methods are essentially Bayesian, which is as it should be
- **NFL:** These and other description-length/prior-based notions of Ockham's razor are essentially **arbitrary**, because you can make any hypothesis arbitrarily 'simple' or 'complex' by changing the prior
- **MDL/Kolmogorov:** By choosing the "right" priors, these methods can be made fully "objective"

I'll argue that, simply and boldly,
all three positions are nonsensical

Three Extreme Positions

Settings are game-theoretic/frequentist

- **BAYES:** All these prior-dependent methods are essentially Bayesian, which is as it should be
- **NFL:** These and other description-length/prior-based notions of Ockham's razor are essentially **arbitrary**, because you can make any hypothesis arbitrarily 'simple' or 'complex' by changing the prior
- **MDL/Kolmogorov:** By choosing the "right" priors, these methods can be made "fully objective"

all three positions are nonsensical

bounds are tight + not nearly everything goes!

Three Extreme Positions

Settings are game-theoretic/frequentist

- **BAYES:** All these prior-dependent methods are essentially Bayesian, which is as it should be
- **NFL:** These and other description-length/prior-based notions of Ockham's razor are essentially **arbitrary**, because you can make any hypothesis arbitrarily 'simple' or 'complex' by changing the prior
- **MDL/Kolmogorov:** By choosing the "right" priors, these methods can be made "fully objective"

all three positions are nonsensical

bounds are tight + not nearly everything goes!

Three Extreme Positions

Settings are game-theoretic/frequentist

- **BAYES:** All these prior-dependent methods are essentially Bayesian, which is as it should be
- **NFL:** These and other description-length/prior-based notions of Ockham's razor are essentially **arbitrary**, because you can make any hypothesis arbitrarily 'simple' or 'complex' by changing the prior
- **MDL/Kolmogorov:** By choosing the "right" priors, these methods can be made "fully objective"

there is a subjective component but it is to be understood as **luckiness** rather than **belief**

all three positions are nonsensical

Menu

1. On-Line Sequential Prediction
   - no stochastic assumptions

- log prior will pop up

2. Statistical Learning
   - i.i.d. assumption (but no "model true")
   - oracle bounds, confidence bounds

- log prior will pop up

- log prior will pop up

Menu

1. On-Line Sequential Prediction
   - no stochastic assumptions
2. Statistical Learning
   - i.i.d. assumption (but no “model true”)
   - oracle bounds, confidence bounds
3. What do priors have to do with Ockham?
   - *not* Bayesian validation of Ockham
4. What is the role of subjectivity? **Luckiness!**

Universal Prediction

- There exist prediction strategies for sequentially predicting data that always work well (in a relative sense), no matter what data are observed

Universal Prediction

- Suppose we have two weather forecasters
  - Marjon de Hond (Dutch public TV)
  - Peter Timofeeff (Dutch commercial TV)
- On each i (day), Marjon and Peter announce the probability that $y_{i+1} = 1$,i.e. that it will rain on day $i+1$

Universal Prediction

- Suppose we have two weather forecasters
  - Marjon de Hond (NOS, public TV)
  - Peter Timofeeff (RTL4, commercial TV)
- On each i (day), Marjon and Peter announce the probability that $y_{i+1} = 1$,i.e. that it will rain on day $i+1$
- We would like to combine their predictions in some way such that for every sequence $y_1, \ldots, y_n \in \{0,1\}^n$ we predict almost as well as whoever turns out to be the best forecaster for that sequence

  (note: we know *nothing* about weather physics ourselves)

Universal Prediction

- Suppose we have two weather forecasters
  - Marjon de Hond (NOS, public TV)
  - Peter Timofeeff (RTL4, commercial TV)
- On each i (day), Marjon and Peter announce the probability that $y_{i+1} = 1$,i.e. that it will rain on day $i+1$
- We would like to combine their predictions in some way such that for every sequence $y_1, \ldots, y_n \in \{0,1\}^n$ we predict almost as well as whoever turns out to be the best forecaster for that sequence
  - If, with hindsight, Marjon was better, we predict as well as Marjon
  - If, with hindsight, Peter was better, we predict as well as Peter

Universal Prediction

- A prediction strategy S is a function that, at each time point i, based on inputs available at time i, outputs a prediction $S(i+1)$ (probability distribution for y_{i+1})
- Marjon and Peter are prediction strategies (using inputs and algorithms that we don't know)
- Our goal: design prediction strategy $\bar{S}$ that, at time i,
  - uses as inputs only past data and past and current predictions of Marjon and Peter, and
  - for every sequence $y_1, \ldots, y_n \in \{0,1\}^n$ predicts almost as well as the best forecaster for that sequence
- Surprisingly, there exist strategies that achieve this.

Logarithmic Loss

- To compare performance of different prediction strategies, we need a measure of prediction quality
- In this talk, prediction quality measured by log loss:

$$\text{loss}(y, P) := -\log_2 P(y)$$

$$\text{loss}(y^n, S) = \text{loss}(y_1, \ldots, y_n, S) := \sum_{i=1}^{n} \text{loss}(y_i, S(i))$$

- corresponds to two important practical settings:
  - data compression, sequential gambling with reinvestiment

Universal prediction with log loss

- We would like to combine predictions such that for every sequence $y_1, \ldots, y_n \in \{0,1\}^n$ we predict almost as well as the best forecaster for that sequence
- It turns out that there exists a universal strategy $\bar{\mathrm{S}}$ such that, for all $n, y^n \in \{0,1\}^n$

$$\text{loss}(y^n, \bar{\mathbf{S}}) \leq \min\{\text{loss}(y^n, S_{\text{Marjon}}), \text{loss}(y^n, S_{\text{Peter}})\} + 1.$$

- Losses increase linearly in n so this is very good!

$$\text{loss}(y^n, S) := \sum_{i=1}^{n} \text{loss}(y_i, S(i))$$

Universal prediction with log loss

- Let Θ be a countable set and let $\{P_\theta \mid \theta \in \Theta\}$ be "probabilistic" predictors, identified with distributions on $\mathcal{Y}^\infty$
- Examples:
  - Θ is a finite set of weather forecasters
  - Θ represents set of all Markov chains of each order with rational-valued parameters
  - Θ represents all polynomials of each degree with rational-valued coefficients, turned into distributions by the Gauss device
- GOAL: given $\{P_\theta \mid \theta \in \Theta\}$, construct a new predictor predicting future data 'essentially as well' as any of the P_θ

A Bayesian Strategy

- One possibility is to act Bayesian:
  1. Put some prior W on Θ
  2. Predict with Bayesian predictive distribution

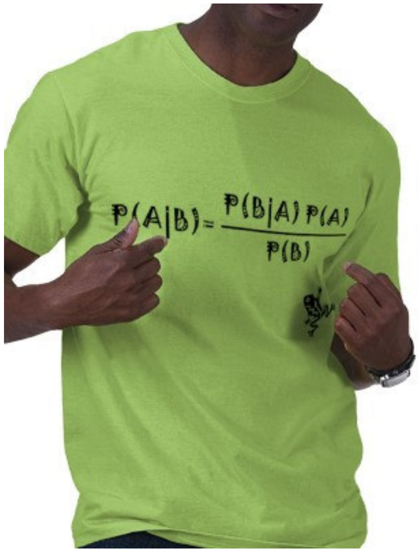
$$P_{\text{Bayes}}(y_{i+1} \mid y_1, \ldots, y_i) = \sum_{\theta} P_\theta(y_{i+1} \mid y^i) W(\theta \mid y^i)$$

A Bayesian Strategy

- One possibility is to act Bayesian:
  1. Put some prior W on Θ
  2. Predict with Bayesian predictive distribution

$$P_{\text{Bayes}}(y_{i+1} \mid y_1, \ldots, y_i) = \sum_{\theta} P_\theta(y_{i+1} \mid y^i) W(\theta \mid y^i)$$

$$W(\theta \mid y^i) = \frac{P_\theta(y^i) \cdot W(\theta)}{\sum_{\theta=1} P_\theta(y^i) W(\theta)}$$ is Bayes posterior!



Bayes is very good universal predictor

- For **all** n, y^n, all θ :

$$\text{log-loss}(y^n, P_{\text{Bayes}}) \leq \text{log-loss}(y^n, P_\theta) - \log W(\theta)$$

- For all sequences of each length n, regret of Bayes bounded by constant depending on θ, not on n

Bayes is very good universal predictor

- For all n, y^n, all θ :

$$\text{log-loss}(y^n, P_{\text{Bayes}}) \leq \text{log-loss}(y^n, P_\theta) - \log W(\theta)$$

- For all sequences of each length n, regret of Bayes bounded by constant depending on θ, not on n
- For “nonmixable” loss functions like 0/1-loss and absolute loss, need to change this a little (Vovk!)
- But first we’ll say something about Luckiness and Ockham

First Glimpse of Luckiness

if nonuniform: "luckiness term"

- For all n, y^n, all θ :

$$\text{log-loss}(y^n, P_{\text{Bayes}}) \leq \text{log-loss}(y^n, P_\theta) - \log W(\theta)$$

First Glimpse of Luckiness

- For all n, y^n, all θ :

$$\text{log-loss}(y^n, P_{\text{Bayes}}) \leq \text{log-loss}(y^n, P_\theta) - \log W(\theta)$$

- If the best θ turns out be one on which you had put high prior, then you are **lucky on the data**
  - good (bound on) performance
- If you had put low prior on all good θ you have bad **luck**
- yet bound holds for **all data**, irrespective of your luck

First Glimpse of Luckiness

- For all n, y^n, all θ :

$$\text{log-loss}(y^n, P_{\text{Bayes}}) \leq \text{log-loss}(y^n, P_\theta) - \log W(\theta)$$

- If the best θ turns out be one on which you had put high prior, then you are **lucky on the data**
  - good (bound on) performance
- If you had put low prior on all good θ you have bad **luck**
- yet bound holds for **all data**, irrespective of your luck

you can put high prior on θ if you believe that it's likely to lead to good predictions (much weaker than: if it is 'true'),
but also if... (see later)

Ockham

Complexity term

- For all n, y^n, all θ :

$$\text{log-loss}(y^n, P_{\text{Bayes}}) \leq \text{log-loss}(y^n, P_\theta) - \log W(\theta)$$

- Term also implies a form of Ockham's Razor:
  entities should not be multiplied beyond necessity

  The more θ I consider, the more data I need before the bound becomes good

  – We are *not* considering complexity of individual θ here – just of the collective!

Ockham

- The more θ I consider, the more data I need
- This is clear for uniform prior on finite Θ

$$|\Theta| = M$$

data compression interpretation easy

$$-\log W(\theta) = \log |M|$$

e.g. $M = 16384 = 2^{14}, -\log W(\theta) = 14$

Ockham

- The more θ I consider, the more data I need
- This is clear for uniform prior on finite Θ

$$|\Theta| = M$$

data compression interpretation easy

$$-\log W(\theta) = \log |M|$$

e.g. $M = 16384 = 2^{14}, -\log W(\theta) = 14$

- But **it still holds for nonuniform prior:**

no matter what prior W you use, at most a fraction of 2^{-K} θ's can be additionally compressed by K bits or more:

$$|\{\theta \in \Theta : -\log W(\theta) \leq M - K\}| \leq 2^{-K}$$

Uncountable "Models"

- If Θ parametric and elements interrelated (say, all Markov chains of order $K = 2^k$) we can discretize in a clever way and put uniform prior on discretized elements, to get, once again, uniform bounds: for all $\theta \in \Theta, n, y^n$,

$$\text{log-loss}(y^n, P_{\text{Bayes}}) \leq \text{log-loss}(y^n, P_\theta) + \frac{k}{2} \log n + \text{const.}$$

can be computed !

Uncountable "Models"

- If Θ parametric and elements interrelated (say, all Markov chains of order $K = 2^k$) we can discretize in a clever way and put uniform prior on discretized elements, to get, once again, uniform bounds: for all $\theta \in \Theta, n, y^n$,

$$\text{log-loss}(y^n, P_{\text{Bayes}}) \leq \text{ log-loss}(y^n, P_\theta) + \frac{k}{2} \log n + \text{const.}$$

- this is worst-case optimal regret
- can in fact get around discretization, though

Uncountable "Models"

- We can now put a meta-prior on k and get, uniformly for all $\theta \in \Theta = \cup_{k=1}^{\infty} \Theta_k, n, y^n$,

$$\text{log-loss}(y^n, P_{\text{Bayes}}) \leq \text{log-loss}(y^n, P_\theta) + \frac{k}{2} \log n + \text{const}_k$$

- Similar things can be done with "nonparametric" models – the regret relative to θ now depends on smoothness properties of θ,e.g. how often is its density differentiable

Ockham+Luckiness

- Ockham+Luckiness Principle: given a large structured 'model' Θ you can (repeatedly!) single out a small, less complex subset Θ_{simple} and construct a meta-prior such that

$$\forall \theta \in \Theta_{\text{simple}}, \ \text{regret}(P'_{\text{Bayes}}, \theta, y^n) \leq \text{regret}(P_{\text{Bayes}} \mid \Theta_{\text{simple}}, \theta, y^n) + 1$$

$$\forall \theta \in \Theta, \ \text{regret}(P'_{\text{Bayes}}, \theta, y^n) \leq \text{regret}(P_{\text{Bayes}} \mid \Theta, \theta, y^n) + 1$$

- Rationale:
  - **If you're lucky, you'll do much better than with the original prior on the large model**
  - **If you're *not* lucky, you will hardly do worse than with the original prior on the large model**

Ockham+Luckiness

$\log n$

- Ockham+Luckiness Principle: given a large structured 'model' Θ you can (repeatedly!) single out a small, less complex subset Θ_{simple} and construct a meta-prior such that

$$\forall \theta \in \Theta_{\text{simple}},\ \text{regret}(P'_{\text{Bayes}}, \theta, y^n) \leq \text{regret}(P_{\text{Bayes}} \mid \Theta_{\text{simple}}, \theta, y^n) + 1$$

$$\forall \theta \in \Theta,\ \text{regret}(P'_{\text{Bayes}}, \theta, y^n) \leq \text{regret}(P_{\text{Bayes}} \mid \Theta, \theta, y^n) + 1$$

$2 \log n$

- Rationale:
  - **If you're lucky, you'll do much better than with the original prior on the large model**
  - **If you're *not* lucky, you will hardly do worse than with the original prior on the large model**

Ockham+Luckiness

Luckiness-Ockham Principle has

- Subjective Component: you can decide on the "simple subset" yourself. Also within the simple subset, you don't have to use a uniform prior
- Objective Component:
  1. Some things simply cannot be arranged by fiddling with the prior (e.g.) "make all second degree polynomials simpler than all first-degree polynomials". The ***set*** of second degree polynomials is inherently more complex!
  2. Some things can be done but are objectively stupid, like discretizing Θ such that $-\log W(\theta)$ regret bound not tight

Ockham+Luckiness

Luckiness-Ockham Principle has

- Subjective Component: you can decide on the "simple subset" yourself. Also within the simple subset, you don't have to use a uniform prior
- Objective Component:
    1. Some things simply cannot be arranged by fiddling with the prior (e.g.) "make all second degree polynomials simpler than all first-degree polynomials". The ***set*** of second degree polynomials is inherently more complex!
    2. Some things can be done but are objectively stupid, like discretizing Θ such that $-\log W(\theta)$ regret bound not tight

Menu

1. On-Line Sequential Prediction
   - no stochastic assumptions
   - still need to go beyond log-loss
2. Statistical Learning
   - i.i.d. assumption (but no “model true”)
   - oracle bounds, confidence bounds

General Loss Functions

- Let $\mathsf{loss} : \mathcal{Y} \times \mathcal{A} \rightarrow [0, \infty]$ be arbitrary loss fn.
- "Expert" (hypothesis) θ is a prediction strategy, i.e. for each i, y^i, it outputs an action $\theta \mid y^i$
- Define

$$\mathsf{loss}(y^n, \theta) := \sum_{i=1}^{n} \mathsf{loss}(y_i, \theta \mid y^{i-1})$$

- Examples:
  - Log-loss: $\mathcal{A}$ is set of distr. on $\mathcal{Y}$, $\mathsf{loss}(y, p_\theta) = -\log p_\theta(y)$
  - 0/1-loss ("rain" or "no rain") $\mathcal{A} = \mathcal{Y} = \{0, 1\}$

$$\mathsf{loss}(y, a) = |y - a|$$

Generalized Bayesian Posterior

Vovk ’90, Audibert ’04, Zhang ’06, Hjorth & Walker,. G. ‘11

$$\text{loss}(y^n, \theta) := \sum_{i=1}^{n} \text{loss}(y_i, \theta \mid y^{i-1})$$

- Define “generalized posterior” as

$$W_{\eta}(\theta \mid y^i) = \frac{W(\theta) e^{-\eta \text{loss}(y^i, \theta)}}{\sum_{\theta' \in \Theta} W(\theta') e^{-\eta \text{loss}(y^i, \theta')}}$$

- With $\eta = 1$ and log-loss, this is just standard posterior

Aggregating Algorithm/Hedge

Vovk '90, Freund & Shapire '98

- Both algorithms work like this:
  1. fix "appropriate" η
  2. For each i, y^i , calculate generalized posterior $W_\eta(\theta \mid y^i)$ and predict y_{i+1} using some fixed function f , $\hat{a}_{i+1} := f(W_\eta(\theta \mid y^i))$

UPSHOT: the algorithm is not Bayes any more, but the bounds still involve priors!

Regret Bounds for AA/Hedge:

- We have for all n, y^n, θ :

regret$(y^n, \mathbf{AA}, \theta) :=$

loss$(y^n, \mathbf{AA}) -$ loss(y^n, θ)

$$\leq \begin{cases} -\log W(\theta) & \text{for log-loss} \\ \frac{1}{\eta} \cdot -\log W(\theta) & \text{for } \eta\text{-mixable loss functions} \\ C \cdot \sqrt{-\log W(\theta)} & \text{for other bounded losses, e.g.} 0/1^* \end{cases}$$

Regret Bounds for AA/Hedge:

Priors remain there even though we have different loss fn!

- We have for all n, y^n, θ :

regret$(y^n, \mathbf{AA}, \theta) :=$

loss$(y^n, \mathbf{AA})$ $-$ loss(y^n, θ)

$$\leq \begin{cases} -\log W(\theta) & \text{for log-loss} \\ \frac{1}{\eta} \cdot -\log W(\theta) & \text{for } \eta\text{-mixable loss functions} \\ C \cdot \sqrt{-\log W(\theta)} & \text{for other bounded losses, e.g.} 0/1^* \end{cases}$$

Regret Bounds for AA/Hedge:

- We have for all n, y^n, θ :

regret$(y^n, \mathbf{AA}, \theta) :=$

loss$(y^n, \mathbf{AA}) -$ loss(y^n, θ)

$$\leq \begin{cases} -\log W(\theta) & \text{for log-loss} \\ \frac{1}{\eta} \cdot -\log W(\theta) & \text{for } \eta\text{-mixable loss functions} \\ C \cdot \sqrt{-\log W(\theta)} & \text{for other bounded losses, e.g.} 0/1^* \end{cases}$$

no stochastic assumptions whatsoever!

Regret Bounds for AA/Hedge:

- We have for all n, y^n, θ :

regret$(y^n, \mathbf{AA}, \theta) :=$

loss$(y^n, \mathbf{AA}) -$ loss(y^n, θ)

$$\leq \begin{cases} -\log W(\theta) & \text{for log-loss} \\ \frac{1}{\eta} \cdot -\log W(\theta) & \text{for } \eta\text{-mixable loss functions} \\ C \cdot \sqrt{-\log W(\theta)} & \text{for other bounded losses, e.g.} 0/1^* \end{cases}$$

These bounds are (in appropriate sense) optimal up to constant factors (Vovk 2001)

Apply to Statistical Learning Theory

Vapnik 1998, many others!

- Let $S_i = (X_i, Y_i)$, $S_1, S_2, \ldots$ i.i.d. $\sim P^*$
- Let Θ be countable set of *predictors* $\Theta : \mathcal{X} \to \mathcal{A}$
  W is prior on Θ
- Example: classification: 0/1 loss, Θ are *classifiers*
  – Spam filtering, object recognition, ...

$$\text{loss}(y, \theta(x)) = |y - \theta(x)|$$

$$\text{risk}(\theta) = \mathbf{E}_{X,Y \sim P^*}[\text{loss}(Y, \theta(X))] = P^*(Y \neq \theta(X))$$

Generalized MAP/2-Part MDL

- The Generalized η-MAP/MDL estimator is defined as

$$\ddot{\theta}_\eta := \arg\min_{\theta \in \Theta} \ \eta \cdot \sum_{i=1}^{n} \underbrace{\mathrm{loss}(y_i, \theta(x_i))}_{\text{error term}} - \underbrace{\log W(\theta)}_{\text{complexity term}}$$

(for log-loss and $\eta = 1$ this is standard MAP)

penalized empirical risk minimization;
ridge regression

Oracle Risk Bounds, 2-Part Estimate

- "risk" is expected loss:

$$\text{risk}(P^*, \theta) = \mathbf{E}_{X,Y \sim P^*}[\text{loss}(Y, \theta(X))]$$

- "excess risk" is concept analogous to "regret"

$$\text{excess-risk}(P^*, \ddot{\theta}_\eta, n) :=$$
$$\mathbf{E}_{S^n \sim P^*}\left[\text{risk}(P^*, \ddot{\theta}_{\eta|S^n}) - \text{risk}(P^*, \theta_{\text{opt}})\right]$$

$$\theta_{\text{opt}} = \arg\min_{\theta \in \Theta} \mathbf{E}_{X,Y \sim P^*}[\text{loss}(Y, \theta(X))]$$

Additional expcted loss incurred by the *learned* predictor compared to the *best* predictor

Oracle Risk Bounds, 2-Part Estimate

- We have for all P^* , for all $0 < \eta < \eta_{\text{crit}}$:

$$\text{excess-risk}(P^*, \ddot{\theta}_\eta, n) \leq \begin{cases} \frac{C}{\eta} \cdot \frac{-\log W(\theta_{\text{opt}})}{n} & \text{for } \eta\text{-mixable loss functions} \\ C \cdot \sqrt{\frac{-\log W(\theta_{\text{opt}})}{n}} & \text{for other bounded losses, e.g. } 0/1^* \end{cases}$$

log-loss $\longrightarrow$ density estimation

- Suppose model correct, i.e. contains "true" P^* , i.e. $P_{\theta_{\mathrm{opt}}}(Y|X) = P^*(Y \mid X)$
- Then log-loss is 1-mixable, and excess-risk is KL divergence

$$\text{excess-risk}(P^*, \ddot{\theta}_\eta, n) \quad \leq \frac{C}{\eta} \cdot \frac{-\log W(\theta_{\mathrm{opt}})}{n} \text{ for } \eta\text{-mixable losses}$$

$$= \mathbf{E}_{S^n \sim P^*}\left[\mathrm{KL}(\theta_{\mathrm{opt}} \| \ddot{\theta}_{\eta|S^n})\right]$$

$$\eta = 1$$

Oracle Risk Bound, Randomized Est.

- Problem: in practice we may have large, nonparametric model, so we cannot assume $W(\theta_{\mathrm{opt}}) > 0$

Oracle Risk Bound, Randomized Est.

- Problem: in practice we may have large, nonparametric model, so we cannot assume $W(\theta_{\text{opt}}) > 0$
- If, instead of doing "generalized MAP", we *randomize* according to the posterior, then we get for all P^*, $\eta < \eta_{\text{crit}}$

$$\text{excess-risk}(P^*, W_\eta \mid Z^n) :=$$

$$\leq \begin{cases} \frac{C}{\eta} \cdot \frac{\text{comp}}{n} & \text{for } \eta\text{-mixable loss functions} \\ \\ C \cdot \sqrt{\frac{\text{comp}}{n}} & \text{for other bounded losses, e.g.} 0/1^* \end{cases}$$

$$\text{comp} = \inf_{\epsilon \geq 0} \{ \, \epsilon - \log W(\theta : \text{excess-risk}(P^*, \theta) \leq \epsilon) \, \}$$

Oracle Risk Bound, Randomized Est.

- Problem: in practice we may have large, nonparametric model, so we cannot assume $W(\theta_{\text{opt}}) > 0$
- If, instead of doing "generalized MAP", we *randomize* according to the posterior, then we get for all P^*, $\eta < \eta_{\text{crit}}$

$$\text{excess-risk}(P^*, W_\eta \mid Z^n) :=$$

$$\leq \begin{cases} \frac{C}{\eta} \cdot \frac{\text{comp}}{n} & \text{for } \eta\text{-mixable loss functions} \\ C \cdot \sqrt{\frac{\text{comp}}{n}} & \text{for other bounded losses, e.g.} 0/1^* \end{cases}$$

These bounds are often* minimax optimal
(Barron '98, Audibert/Tsybakov '04, Zhang '06)

Confidence Risk Bound

- Problem: previous bounds say that generalized Bayes method learns 'as fast as possible', but involve an unknown quantity (P^*)
  - We would like to have a confidence bound for our predictions for actual, given data that does not depend on unknown quantities

Confidence Risk Bound

- Problem: previous bounds say that generalized Bayes method learns 'as fast as possible', but involve an unknown quantity (P^*)
  - We would like to have a confidence bound for our predictions for actual, given data that does not depend on unknown quantities
- Provided by McAllester's PAC-Bayes generalization bounds: for all $P^*, K > 0, \eta > 0$, with prob. at least $1 - e^{-K}$:

$$\text{risk}(P^*, \ddot{\theta}_\eta) \leq \frac{1}{n} \sum_{i=1}^{n} \text{loss}(Y_i, \ddot{\theta}_\eta(X_i)) + \sqrt{\frac{-\log W(\ddot{\theta}_\eta) + K}{n}}$$

Taking Stock

- Complexity-Regularizing Priors appear in
  - nonstochastic worst-case regret bounds (game-theoretic analysis)
  - oracle risk bounds w.r.t. general loss functions (frequentist analysis)
  - oracle confidence bounds wrt general loss fns (frequentist analysis)
- So “priors” may be pretty fundamental!
  - analysis was never Bayesian though (cf. Complete Class Thm.)

Did I deliver?
Three Extreme Positions, Revisited

- **BAYES:** All these prior-dependent methods are essentially Bayesian, which is as it should be
- **NFL:** These and other description-length/prior-based notions of Ockham's razor are essentially **arbitrary**, because you can make any hypothesis arbitrarily 'simple' or 'complex' by changing the prior
- **MDL/Kolmogorov:** By choosing the "right" priors, these methods can be made "fully objective"

"all three positions are nonsensical"

Did I deliver?
Three Extreme Positions, Revisited

- **BAYES:** "All these prior-dependent methods are essentially Bayesian, which is as it should be"
  - no: algorithms were not Bayesian (yet similar)

    purely Bayesian algorithms may fail dramatically in such cases (G. and Langford, 2007)
  - You may assign small prior to certain θ because you think they are not likely to predict well...
  - But also because **they may not be useful**!
  - **bounds hold *irrespective of prior assumptions***
    - If you're lucky, prior is well aligned with data, and bound is strong. But bound holds whether you are lucky or not! There's no such thing in Bayesian inference

Did I deliver?
Three Extreme Positions, Revisited

Note though that I'm certainly not anti-Bayes.

It's just that I think that there exist interesting settings of inductive inference in which Bayes is not the whole story

Similarly I'm not strictly instrumentalist – sometimes one wants to be realist, and it is also interesting to study Occam in that setting

Did I deliver?
Three Extreme Positions, Revisited

- **NFL:** These and other description-length/prior-based notions of Ockham's razor are essentially **arbitrary**, because you can make any hypothesis arbitrarily 'simple' or 'complex' by changing the prior
  - NO NO NO . You cannot make the **set** of second-degree polynomials simpler than the set of first-degree polynomials by fiddling with the prior, unless you use a prior which can be "uniformly beaten" by another prior
  - **And relatedly, nowhere do we make the (false) assumption that "the truth is likely to have a short description"**

Did I deliver?
Three Extreme Positions, Revisited

- **MDL/Kolmogorov:** By choosing the "right" priors, these methods can be made "fully objective"
  - No: a subjective element is inherent. Which "simple" subset do you prefer? There are many
  - For many parametric models "minimax optimal priors" (eg Jeffreys' prior) for a given loss function do not exist
    - You are *forced* to give a preference to a subset of the parameters

I didn’t tell you about...

- **Nonparametric Bayes** inconsistency and Ockham (rel. to Diaconis-Freedman results)
- Ockham in cross-validation (really: prequential validation)

Luckiness

- Idea of combining luckiness with complexity is all over the place in modern statistics, though not always (I admit) with complexity determined in terms of priors
- Prime Example: Adaptive Estimation
- Difference between luckiness and belief-priors... **where are the philosophers???**
- One of the first mentions on a related idea was by Kiefer, in the context of ‘conditionalist frequentist inference’

Some Lucky References

Explicit Luckiness in Statistics and Machine Learning:

- J. Kiefer, Conditional Confidence Statements and Confidence Estimators, JASA, 72(360), 1977. First occurrence (?) of "lucky"
- J. Shawe-Taylor, P. Bartlett, R. Williamson, and M. Anthony. Structural risk minimization over data-dependent hierarchies. IEEE Transacticions on Information Theory 44(5),1998
- R. Herbrich and R. Williamson. Algorithmic Luckiness, *Journal of Machine Learning Research* 3 (2002)

Luckiness + Ockham:

- Ch. 17 of my book, “The Minimum Description Length Principle”
- S. de Rooij and G. . Luckiness and Regret in Minimum Description Length Inference. *Handbook of the Philosophy of Science, Vol. 7: Philosophy of Statistics* (eds. P.Bandyopadhyay and M. Forster). Elsevier 2011

- "Statistics is too complex to be codified in terms of a simple prescription that is a panacea for all settings"

  Jack Kiefer (father of "luckiness" ideas) in:

  The Foundations of Statistics: Are There Any? (Synthese, 1977)
- That still holds today. Nevertheless I firmly believe, and hope to have shown, that some useful unifications are possible based on bits and priors
- **Thank you!**