



Machine Learning: The Probabilistic Perspective

Konstantin Tretyakov

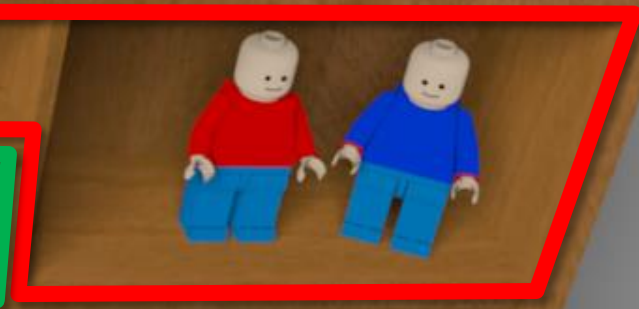
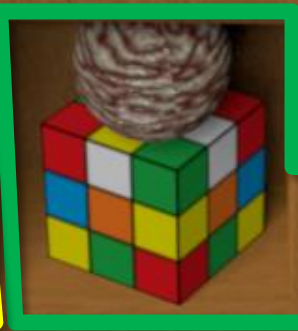
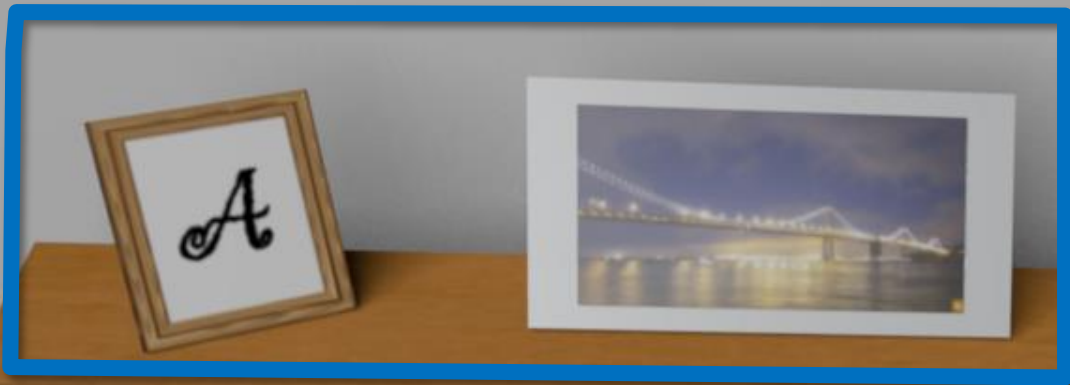
<http://kt.era.ee>

**AACIMP Summer
School 2015**

STACC

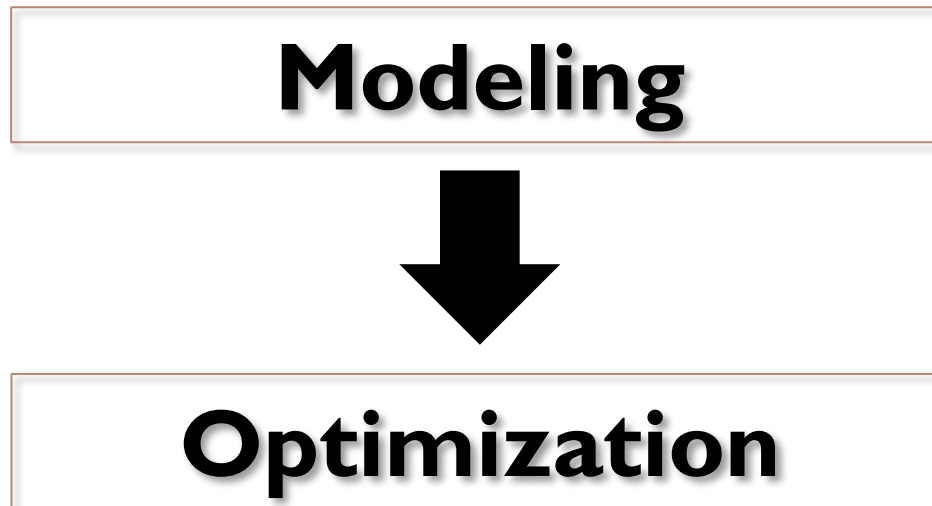
Software Technology and
Applications Competence Center





So far...

- ▶ Machine learning is important and interesting
- ▶ The general concept:



So far...

- ▶ Machine learning is important and interesting
- ▶ The general concept:

**Probability
Theory**

Optimization

The Land of Machine Learning

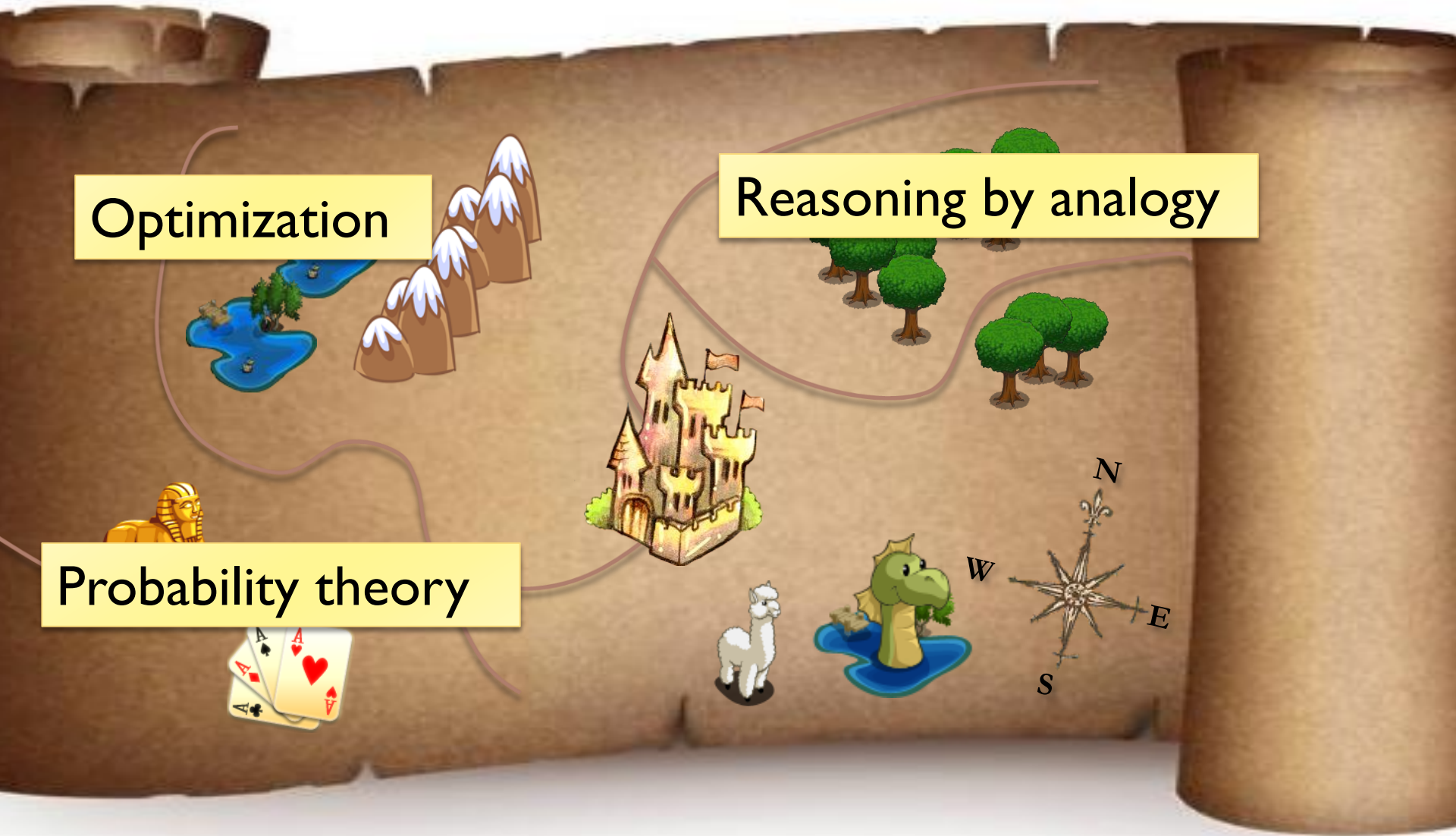


The Land of Machine Learning

Optimization

Reasoning by analogy

Probability theory



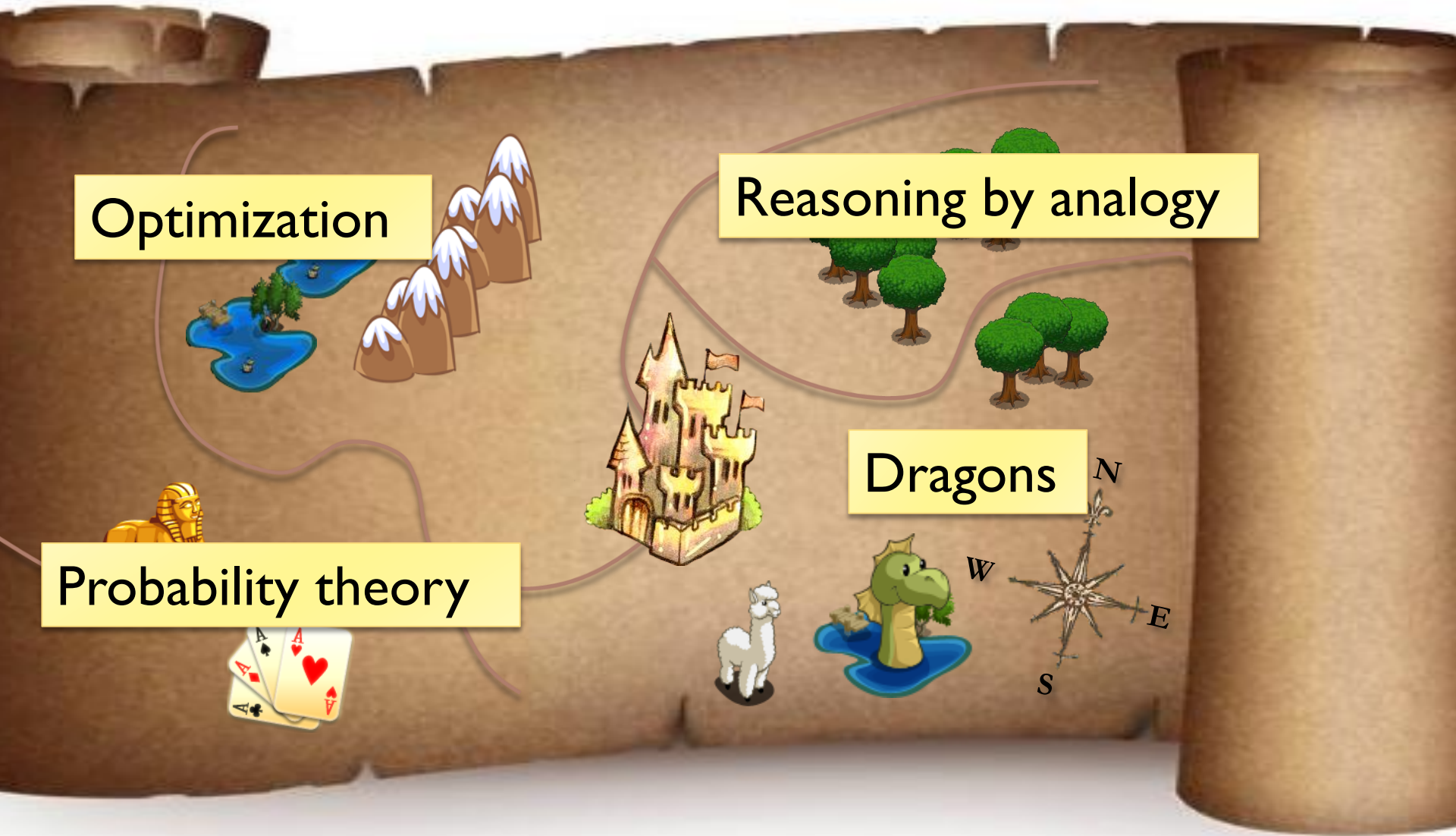
The Land of Machine Learning

Optimization

Reasoning by analogy

Probability theory

Dragons



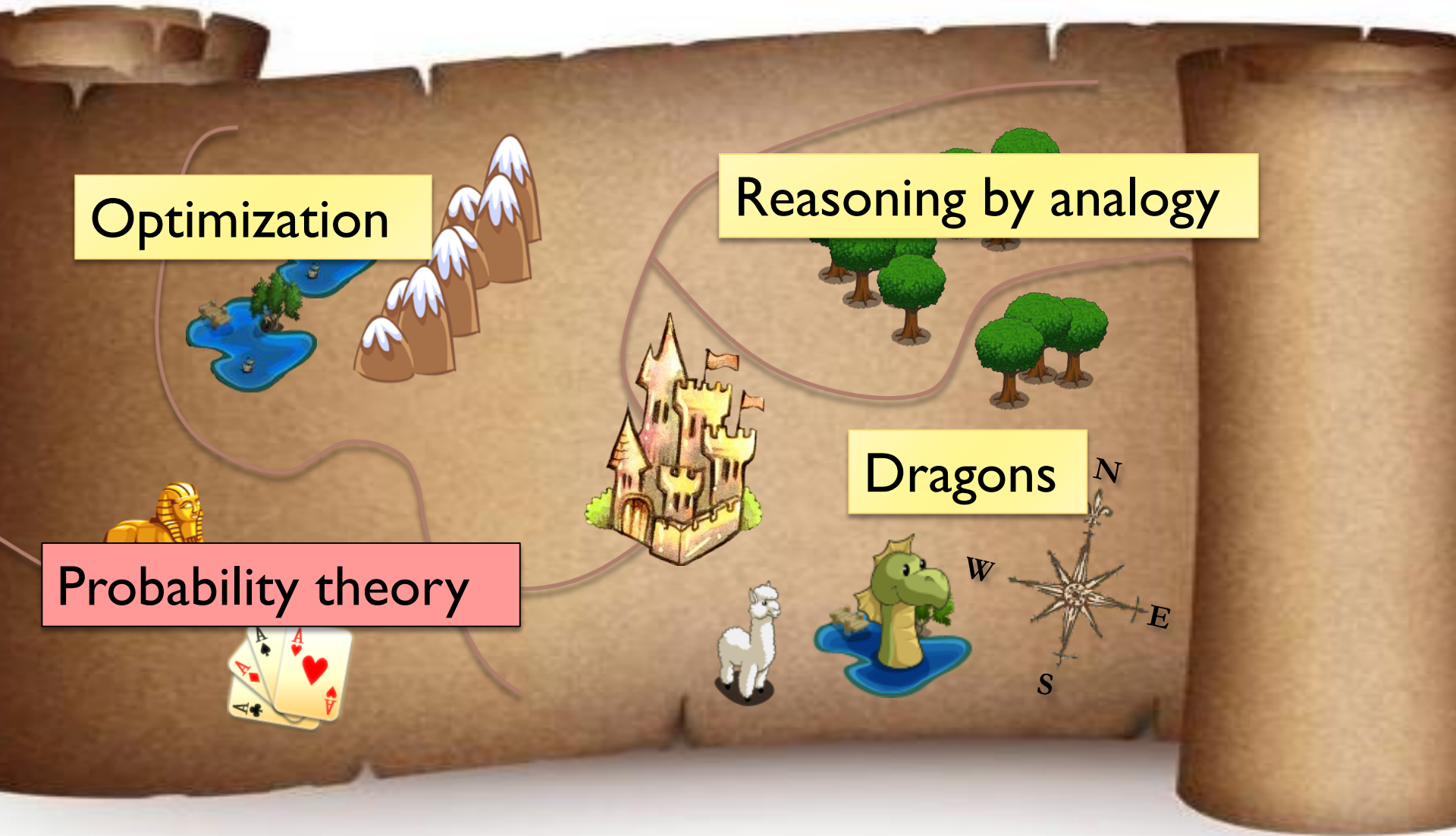
The Land of Machine Learning

Optimization

Reasoning by analogy

Probability theory

Dragons



Next



Why should the model,
tuned on the **training set**,
generalize to the test set?



The “No Free Lunch” Principle

Learning **purely from data** is, in general, impossible

| X | Y | Output |
|---|---|--------|
| 0 | 0 | False |
| 0 | 1 | True |
| 1 | 0 | True |
| 1 | 1 | ? |



The “No Free Lunch” Principle

Learning **purely from data** is, in general, impossible

- ▶ Is it good or bad?



- ▶ What should we do to enable learning?



The “No Free Lunch” Principle

Learning **purely from data** is, in general, impossible

- ▶ Is it good or bad?
 - ▶ Good for cryptographers, bad for data miners
- ▶ What should we do to enable learning?
 - ▶ Introduce **assumptions about data** (“inductive bias”):
 1. **How does existing data relate to the future data?**
 2. **What is the system we are learning?**

The “No Free Lunch” Principle

Learning **purely from data** is, in general, impossible

► Is it good or bad?

► Good for cryptographers, bad for data miners

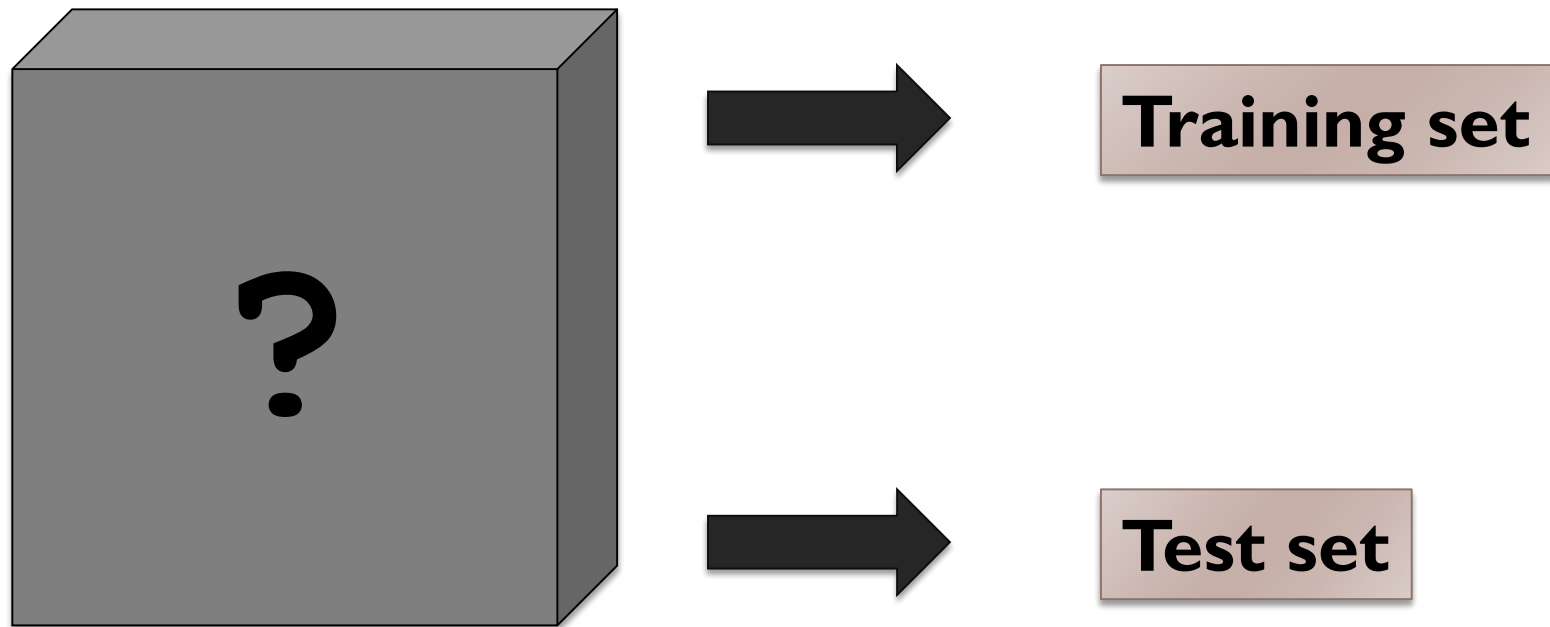
► What should we do to enable learning?

► Introduce **assumptions about data** (“inductive bias”).

1. **How does existing data relate to the future data?**

2. **What is the system we are learning?**

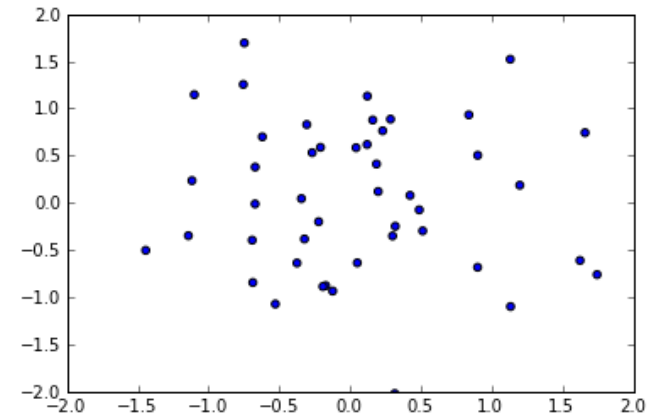
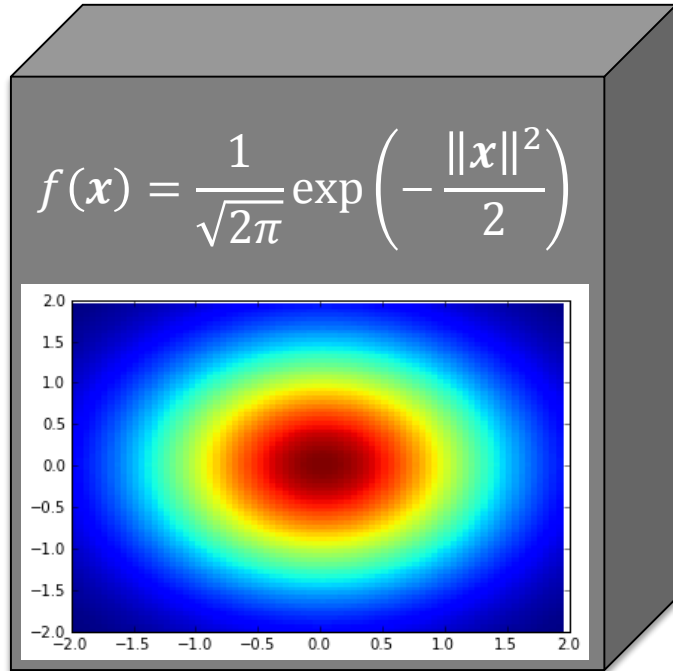
How does existing data relate to future data?



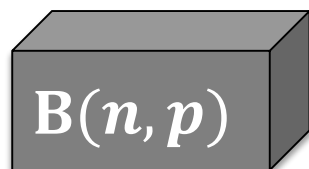
| X | P(X) |
|-------|------|
| heads | 0.5 |
| tails | 0.5 |

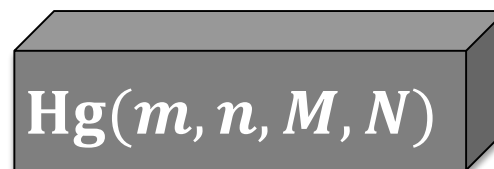


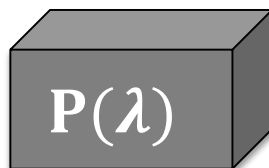
**heads,
heads,
tails,
heads,
tails,
...**

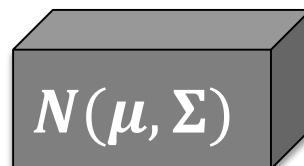


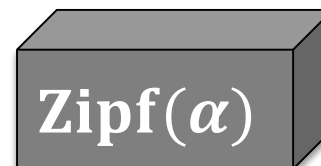
Probability theory

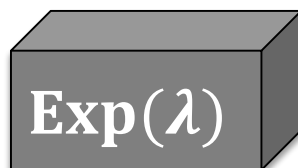

$$B(n, p)$$

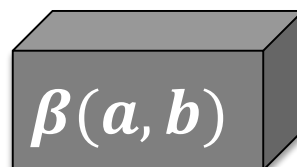

$$Hg(m, n, M, N)$$

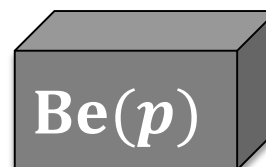

$$P(\lambda)$$

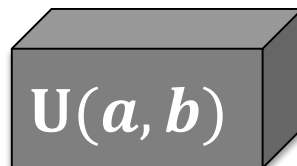

$$N(\mu, \Sigma)$$


$$\text{Zipf}(\alpha)$$


$$\text{Exp}(\lambda)$$


$$\beta(a, b)$$


$$\text{Be}(p)$$


$$U(a, b)$$


$$\dots$$

| | | |
|-----------|---|--------|
| V • T • E | Probability distributions | [hide] |
| | Discrete univariate with finite support | [hide] |
| | Benford • Bernoulli • Beta-binomial • binomial • categorical • hypergeometric • Poisson binomial • Rademacher • discrete uniform • Zipf • Zipf-Mandelbrot | |
| | Discrete univariate with infinite support | [hide] |
| | beta negative binomial • Boltzmann • Conway–Maxwell–Poisson • discrete phase-type • Delaporte • extended negative binomial • Gauss–Kuzmin • geometric • logarithmic • negative binomial • parabolic fractal • Poisson • Skellam • Yule–Simon • zeta | |
| | Continuous univariate supported on a bounded interval, e.g. [0,1] | [hide] |
| | Arcsine • ARGUS • Balding–Nichols • Bates • Beta • Beta rectangular • Irwin–Hall • Kumaraswamy • logit-normal • Noncentral beta • raised cosine • triangular • U-quadratic • uniform • Wigner semicircle | |
| | Continuous univariate supported on a semi-infinite interval, usually [0,∞) | [hide] |
| | Benini • Benktander 1st kind • Benktander 2nd kind • Beta prime • Bose–Einstein • Burr • chi-squared • chi • Coxian • Dagum • Davis • Erlang • exponential • F • Fermi–Dirac • folded normal • Fréchet • Gamma • generalized inverse Gaussian • half-logistic • half-normal • Hotelling's T-squared • hyper-exponential • hypoexponential • inverse chi-squared (scaled-inverse-chi-squared) • inverse Gaussian • inverse gamma • Kolmogorov • Lévy • log-Cauchy • log-Laplace • log-logistic • log-normal • Maxwell–Boltzmann • Maxwell speed • Mittag–Leffler • Nakagami • noncentral chi-squared • Pareto • phase-type • Rayleigh • relativistic Breit–Wigner • Rice • Rosin–Rammler • shifted Gompertz • truncated normal • type-2 Gumbel • Weibull • Wilks' lambda | |
| | Continuous univariate supported on the whole real line (−∞, ∞) | [hide] |
| | Cauchy • exponential power • Fisher's z • generalized normal • generalized hyperbolic • geometric stable • Gumbel • Holtzmark • hyperbolic secant • Landau • Laplace • Linnik • logistic • noncentral t • normal (Gaussian) • normal-inverse Gaussian • skew normal • slash • stable • Student's t • type-1 Gumbel • variance-gamma • Voigt | |
| | Continuous univariate with support whose type varies | [hide] |
| | generalized extreme value • generalized Pareto • Tukey lambda • q-Gaussian • q-exponential • shifted log-logistic | |
| | Mixed continuous-discrete univariate distributions | [hide] |
| | rectified Gaussian | |
| | Multivariate (joint) | [hide] |
| | <i>Discrete:</i> Ewens • multinomial • Dirichlet-multinomial • negative multinomial | |
| | <i>Continuous:</i> Dirichlet • Generalized Dirichlet • multivariate normal • Multivariate stable • multivariate Student • normal-scaled inverse gamma • normal-gamma | |
| | <i>Matrix-valued:</i> inverse matrix gamma • inverse-Wishart • matrix normal • matrix t • matrix gamma • normal-inverse-Wishart • normal-Wishart • Wishart | |
| | Directional | [hide] |
| | <i>Univariate (circular) directional:</i> Circular uniform • univariate von Mises • wrapped normal • wrapped Cauchy • wrapped exponential • wrapped Lévy | |
| | <i>Bivariate (spherical):</i> Kent • <i>Bivariate (toroidal):</i> bivariate von Mises | |
| | <i>Multivariate:</i> von Mises–Fisher • Bingham | |
| | Degenerate and singular | [hide] |
| | <i>Degenerate:</i> discrete degenerate • Dirac delta function | |
| | <i>Singular:</i> Cantor | |
| | Families | [hide] |

Probability theory

```
from numpy.random import beta, binomial,  
chisquare, dirichlet, exponential, f, gamma,  
geometric, gumbel, hypergeometric, ...
```

```
>>> numpy.random.seed(1)
```

```
>>> binomial(10, 0.2)
```

```
::: 2
```

Probability theory

```
from scipy.stats.distributions import beta,  
binom, chisquare, ...
```

```
>>> numpy.random.seed(1)
```

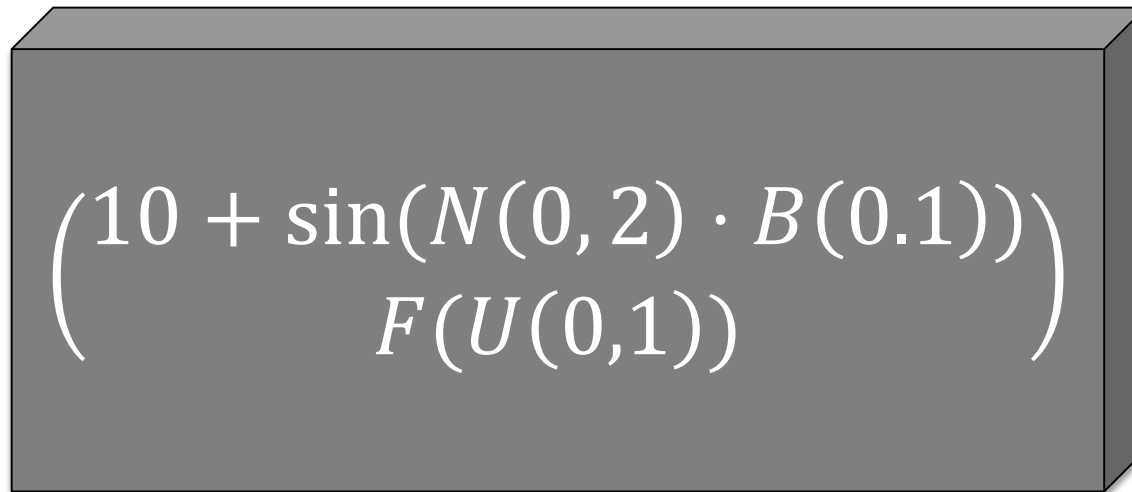
```
>>> X = binom(10, 0.2)
```

```
>>> X.rvs()
```

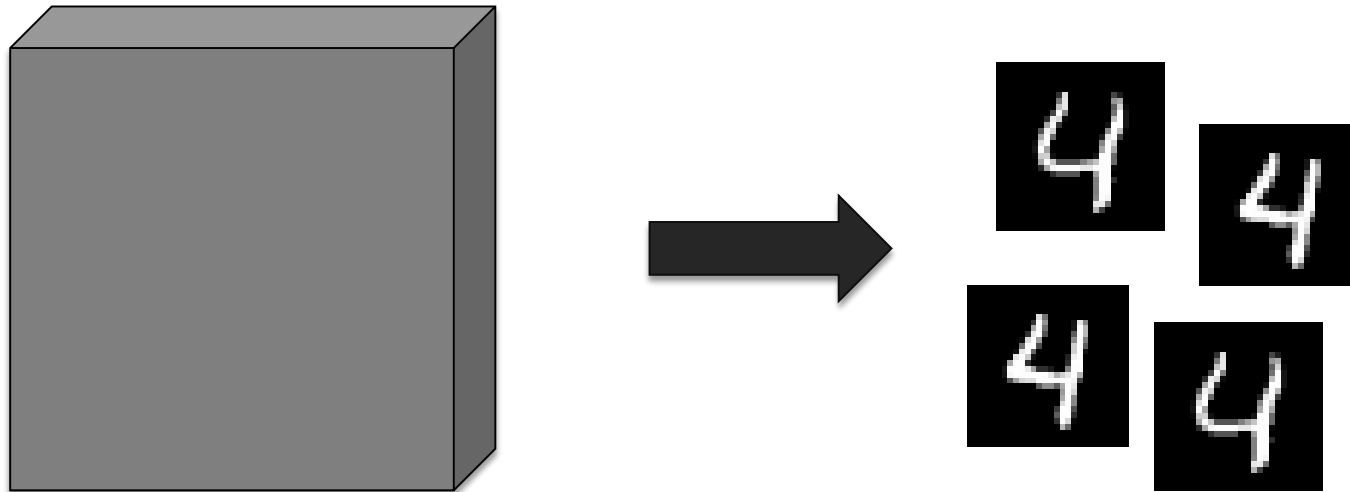
```
::: 2
```

```
>>> X.pmf(2), X.cdf(2), X.mean(), X.std(), ...
```


Probability theory

A gray 3D rectangular box with a black outline, containing a mathematical expression in white text.
$$\left(\begin{array}{c} 10 + \sin(N(0, 2) \cdot B(0.1)) \\ F(U(0,1)) \end{array} \right)$$

Probability theory



Everything is Probabilistic?

What is your height?



Everything is Probabilistic?

What is your height?

Is it a fixed number?



Everything is Probabilistic?

What is your height?

Is it a fixed number?

- ▶ Frequentist: **Yes, it is**, we just don't know it precisely.
- ▶ Bayesian: **No, it is not.** It is a **distribution**.



Everything is Probabilistic?

What is your height?

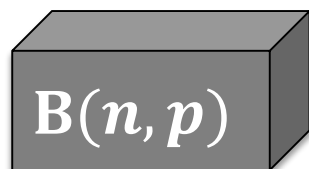
Is it a fixed number?

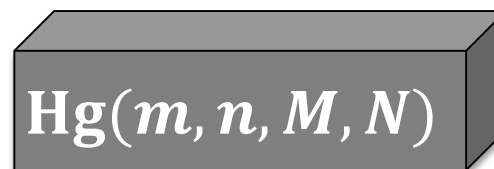
- ▶ Frequentist: **Yes, it is**, we just don't know it precisely.
- ▶ Bayesian: **No, it is not**. It is a **distribution**.

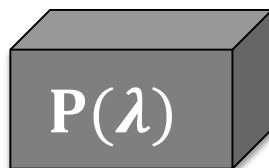
In any case, we need probabilistic reasoning.

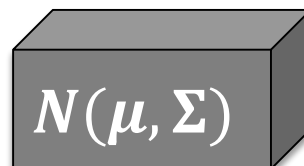


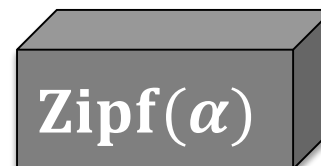
Probability theory

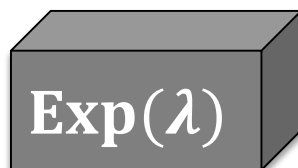

$$B(n, p)$$

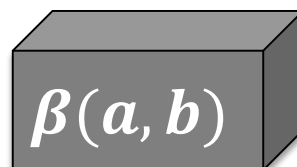

$$Hg(m, n, M, N)$$

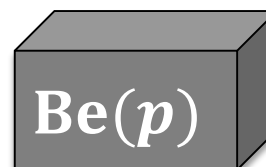

$$P(\lambda)$$

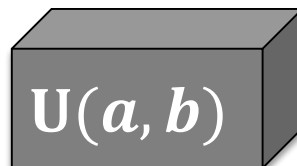

$$N(\mu, \Sigma)$$


$$\text{Zipf}(\alpha)$$


$$\text{Exp}(\lambda)$$


$$\beta(a, b)$$

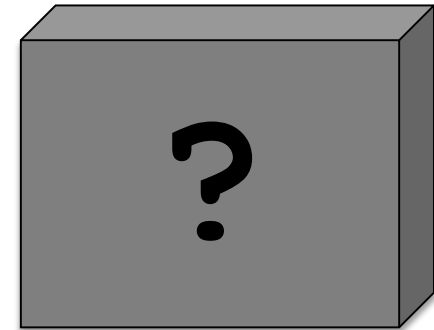
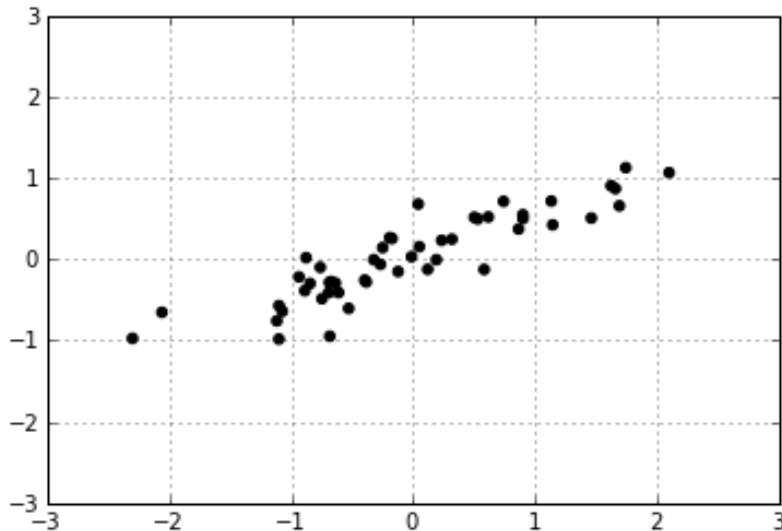

$$\text{Be}(p)$$


$$U(a, b)$$


$$\dots$$

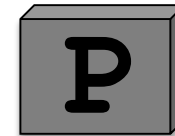
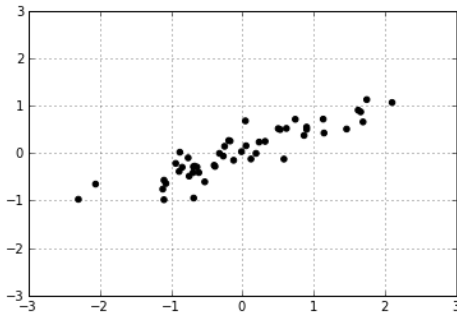
Statistics

- ▶ How do we **infer** a probabilistic **model** based on data?



Statistics

- ▶ How do we **infer** a probabilistic **model** based on data?

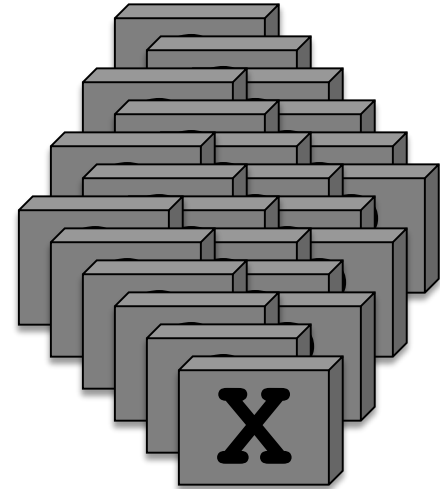
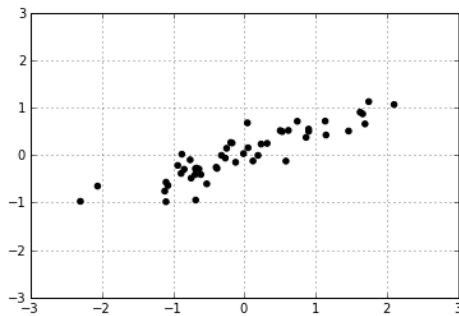


or not?

Hypothesis testing

Statistics

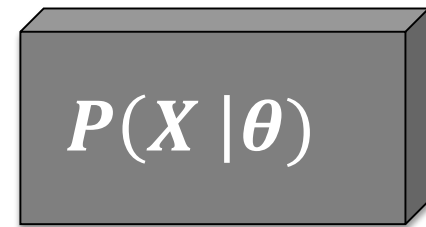
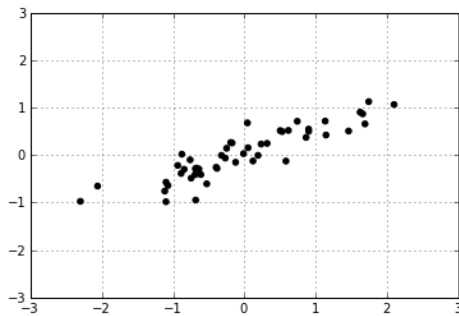
- ▶ How do we **infer** a probabilistic **model** based on data?



Model selection

Statistics

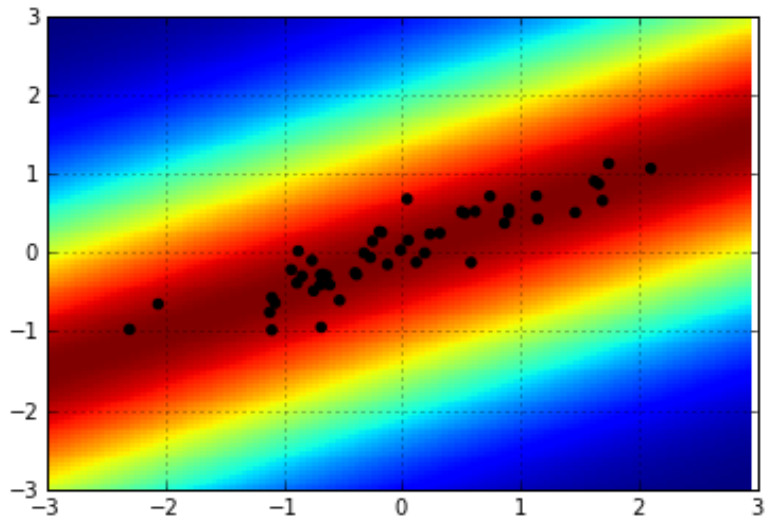
- ▶ How do we **infer** a probabilistic **model** based on data?



Parameter inference

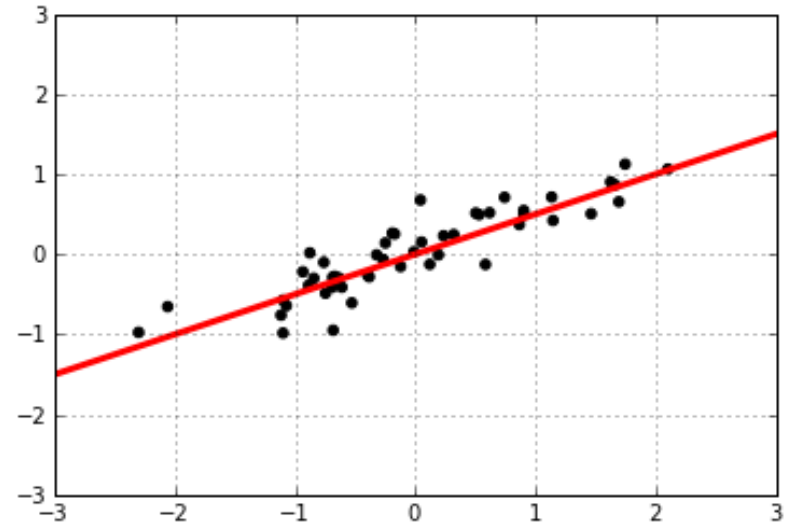
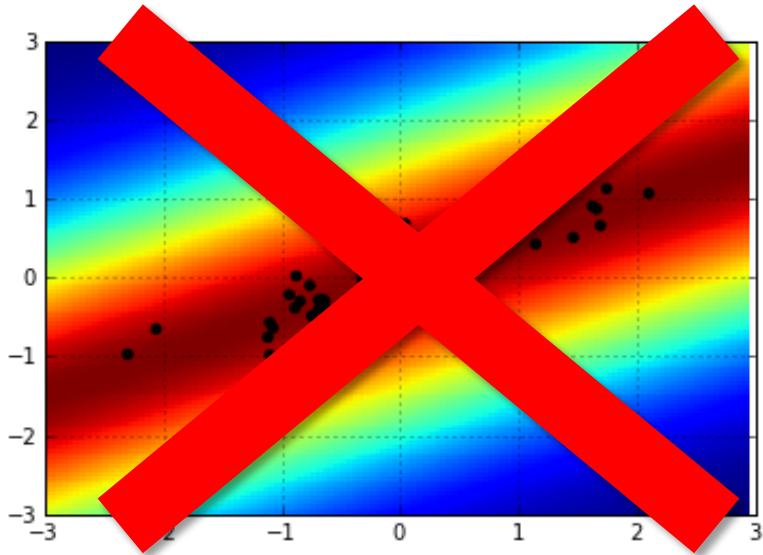
Decision Theory

- ▶ How do we **use** a probabilistic model **to act**?



Decision Theory

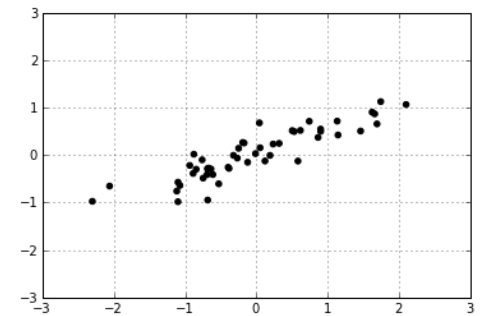
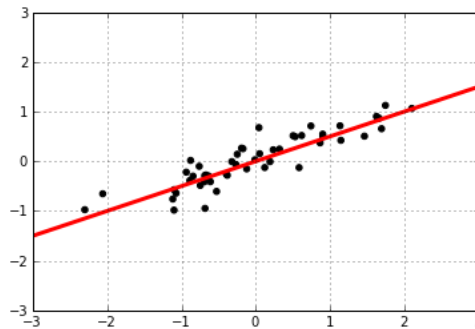
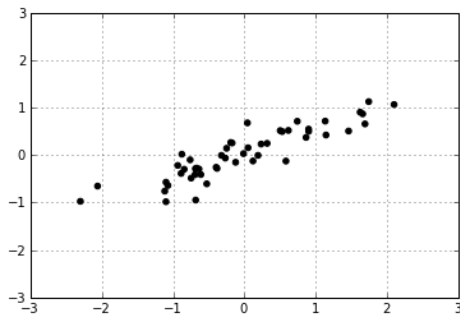
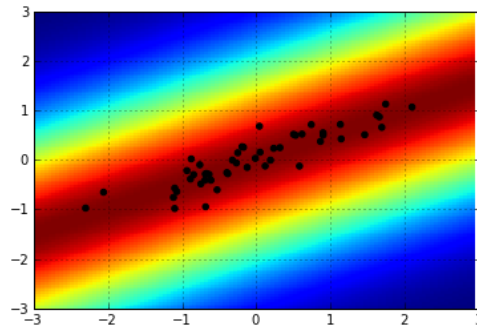
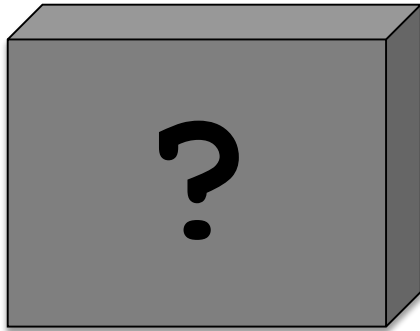
- ▶ How do we **use** a probabilistic model **to act**?



Quiz

- ▶ Model, trained on the training set might work well on the test set because:
 - ▶ Because we **assume** a single underlying mechanism.
 - ▶ Because we **use statistical inference** to infer the mechanism.
 - ▶ Because we **use decision theory** to produce optimal decisions.

Quiz



Example: Biased coin

What is the next output?



1,1,0,1,1,?



Example: Biased coin

Step 1: Modeling

$Be(p)$

| X | $P(X)$ |
|-----|--------|
| 1 | p |
| 0 | $1-p$ |

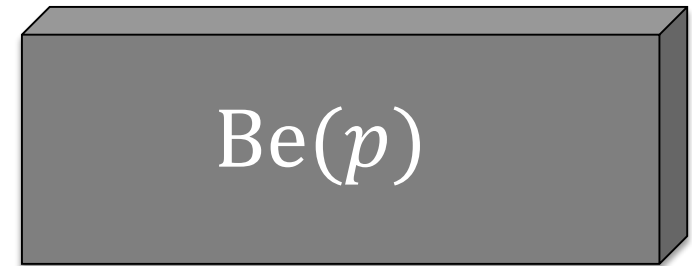


1,1,0,1,1

Example: Biased coin

Step 2: Parameter inference

1,1,0,1,1



$p = ?$

Maximum Likelihood Estimation

- ▶ Data Likelihood:

$$\Pr[\text{Data} \mid \text{Model}]$$

- ▶ Example:

- ▶ Model: $\text{Be}(0.5)$
- ▶ Data: 1,1,0,1,1
- ▶ Likelihood: ?

Maximum Likelihood Estimation

- ▶ Data Likelihood:

$$\Pr[\text{Data} \mid \text{Model}]$$

- ▶ Example:

- ▶ Model: $\text{Be}(0.5)$

- ▶ Data: 1,1,0,1,1

- ▶ Likelihood: $0.5 \cdot 0.5 \cdot 0.5 \cdot 0.5 \cdot 0.5 = 2^{-5}$

0.03125

Maximum Likelihood Estimation

- ▶ **Data Likelihood:**

$$\Pr[\text{Data} \mid \text{Model}]$$

- ▶ **Example:**

- ▶ Model: $\text{Be}(0.2)$
- ▶ Data: 1,1,0,1,1
- ▶ Likelihood: ?

Maximum Likelihood Estimation

- ▶ Data Likelihood:

$$\Pr[\text{Data} \mid \text{Model}]$$

- ▶ Example:

- ▶ Model: $\text{Be}(0.2)$

- ▶ Data: 1,1,0,1,1

- ▶ Likelihood: $0.2 \cdot 0.2 \cdot 0.8 \cdot 0.2 \cdot 0.2 = 0.2^4 \cdot 0.8$

0.00128

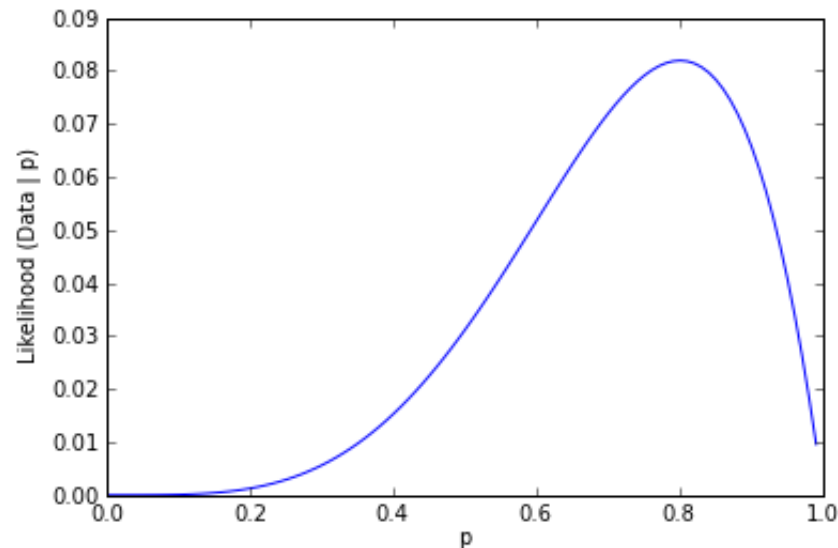
Maximum Likelihood Estimation

► Example:

► Model: $\text{Be}(p)$

► Data: 1,1,0,1,1

► Likelihood: $p \cdot p \cdot (1 - p) \cdot p \cdot p = p^{n_1} (1 - p)^{n_0}$



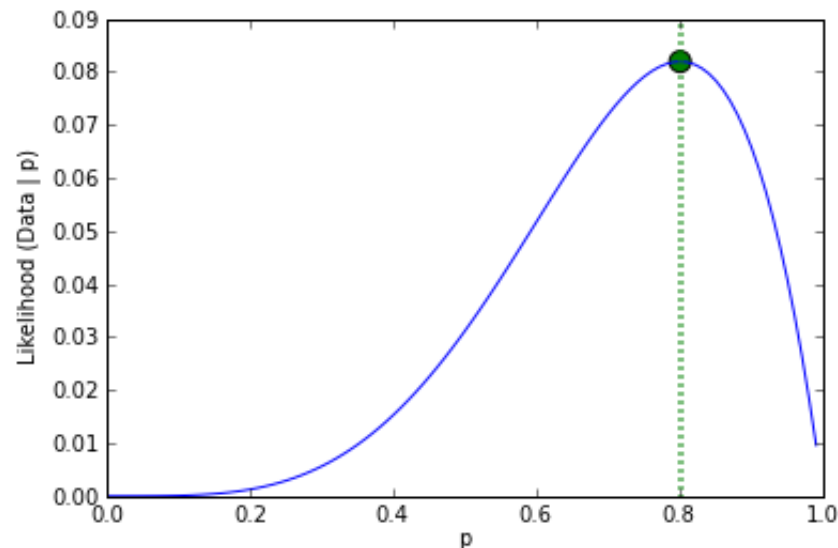
Maximum Likelihood Estimation

► Example:

► Model: $\text{Be}(p)$

► Data: 1,1,0,1,1

► Likelihood: $p \cdot p \cdot (1 - p) \cdot p \cdot p = p^{n_1} (1 - p)^{n_0}$

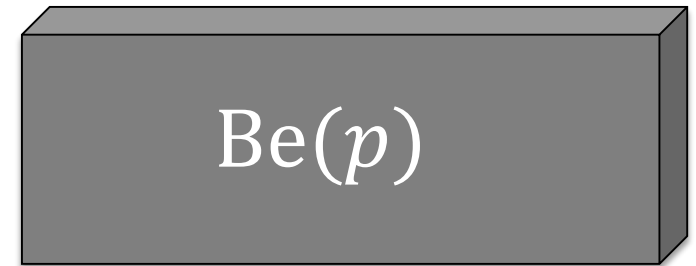


$$\hat{p} = \frac{n_1}{n_0 + n_1}$$

Example: Biased coin

Step 2: Parameter inference

1,1,0,1,1



$$p = 0.8$$

Maximum Likelihood Estimation

► Maximum Likelihood Estimation:

$$\operatorname{argmax}_{\text{Model}} \Pr(\text{Data} | \text{Model})$$

Problems of MLE

- ▶ You are on a trip in an exotic country and you meet a person who happens to be from Ukraine.
- ▶ Is he a member of the Rada (Ukrainian Parliament)?

Problems of MLE

- ▶ Data: “X is from Ukraine”
- ▶ Models:
 - ▶ “X is a member of the Rada”,
 - ▶ “X is not a member of the Rada”

Problems of MLE

- ▶ Data: “X is from Ukraine”
- ▶ Models:
 - ▶ “X is a member of the Rada”,
 - ▶ “X is not a member of the Rada”
- ▶ Likelihoods:
 - ▶ $P(\text{X is from Ukraine} \mid \text{X is a member of the Rada}) =$
 - ▶ $P(\text{X is from Ukraine} \mid \text{X is **not** a member the Rada}) =$

Problems of MLE

- ▶ Data: “X is from Ukraine”
- ▶ Models:
 - ▶ “X is a member of the Rada”,
 - ▶ “X is not a member of the Rada”
- ▶ Likelihoods:
 - ▶ $P(X \text{ is from Ukraine} \mid X \text{ is a member of the Rada}) = 1$
 - ▶ $P(X \text{ is from Ukraine} \mid X \text{ is **not** a member the Rada}) = \frac{45}{7000}$

Problems of MLE

- ▶ Data: “X is from Ukraine”

- ▶ Models:

- ▶ “X is a member of the Rada”,

- ▶ MLE treats all candidate models as equal and can thus **overfit**

- ▶ $P(X \text{ is from Ukraine} \mid X \text{ is a member of the Rada}) = 1$

- ▶ $P(X \text{ is from Ukraine} \mid X \text{ is **not** a member the Rada}) = \frac{45}{7000}$

Maximum A-posteriori Estimation

- ▶ Maximum Likelihood Estimate (MLE):

$$\operatorname{argmax}_{\text{Model}} \Pr(\text{Data} | \text{Model})$$

- ▶ Maximum A-posteriori Estimate (MAP):

$$\operatorname{argmax}_{\text{Model}} \Pr(\text{Model} | \text{Data})$$

MAP Estimation

$$\operatorname{argmax}_{\text{Model}} \Pr(\text{Model}|\text{Data})$$

MAP Estimation

$$\operatorname{argmax}_{\text{Model}} \Pr(\text{Model}|\text{Data})$$

$$\operatorname{argmax}_{\text{Model}} \frac{\Pr(\text{Model}, \text{Data})}{\Pr(\text{Data})}$$

$$\operatorname{argmax}_{\text{Model}} \Pr(\text{Model}, \text{Data})$$

MAP Estimation

$$\operatorname{argmax}_{\text{Model}} \Pr(\text{Model}|\text{Data})$$

$$\operatorname{argmax}_{\text{Model}} \frac{\Pr(\text{Model}, \text{Data})}{\Pr(\text{Data})}$$

$$\operatorname{argmax}_{\text{Model}} \Pr(\text{Model}, \text{Data})$$

$$\operatorname{argmax}_{\text{Model}} \Pr(\text{Data} | \text{Model}) \cdot \Pr(\text{Model})$$

MAP Estimation

$$\operatorname{argmax}_{\text{Model}} \Pr(\text{Model} | \text{Data})$$

$$\operatorname{argmax}_{\text{Model}} \frac{\Pr(\text{Model}, \text{Data})}{\Pr(\text{Data})}$$

Model posterior

$$\operatorname{argmax}_{\text{Model}} \Pr(\text{Model}, \text{Data})$$

$$\operatorname{argmax}_{\text{Model}} \Pr(\text{Data} | \text{Model}) \Pr(\text{Model})$$

Likelihood

Model prior

Summary

- ▶ Maximum Likelihood Estimate (MLE):

$$\operatorname{argmax}_{\text{Model}} \Pr(\text{Data} \mid \text{Model})$$

- ▶ Maximum A-posteriori Estimate (MAP):

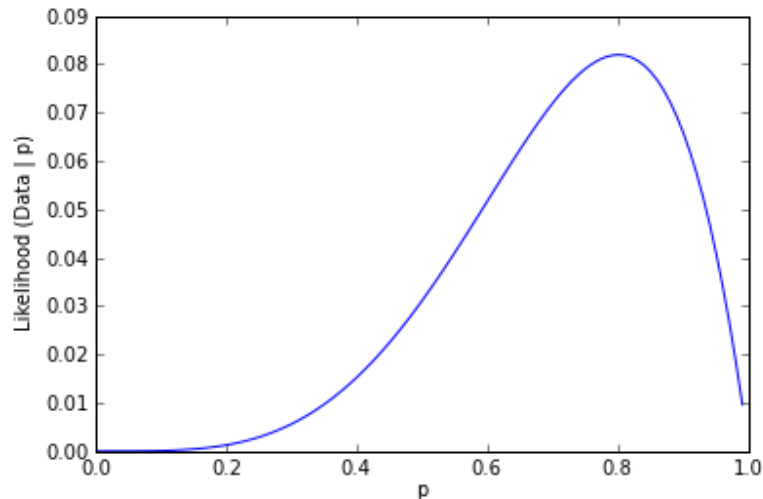
$$\operatorname{argmax}_{\text{Model}} \Pr(\text{Data} \mid \text{Model}) \Pr(\text{Model})$$

MAP Estimation

► Model: $\text{Be}(p)$

Data: 1,1,0,1,1

Likelihood: $p^4(1 - p)$



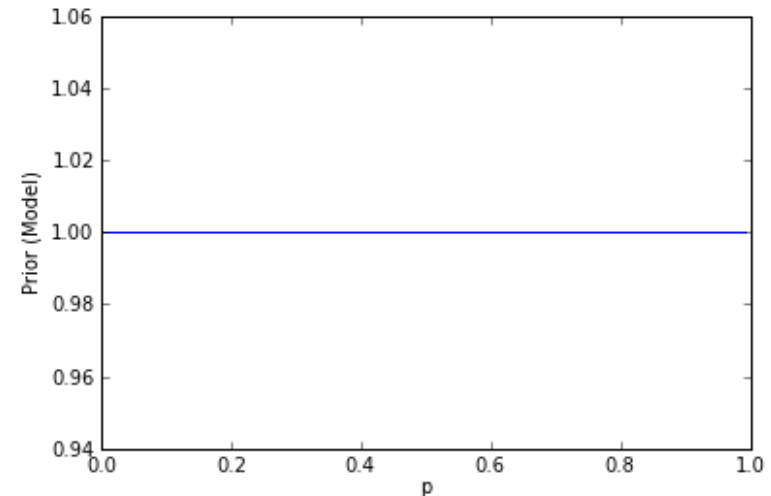
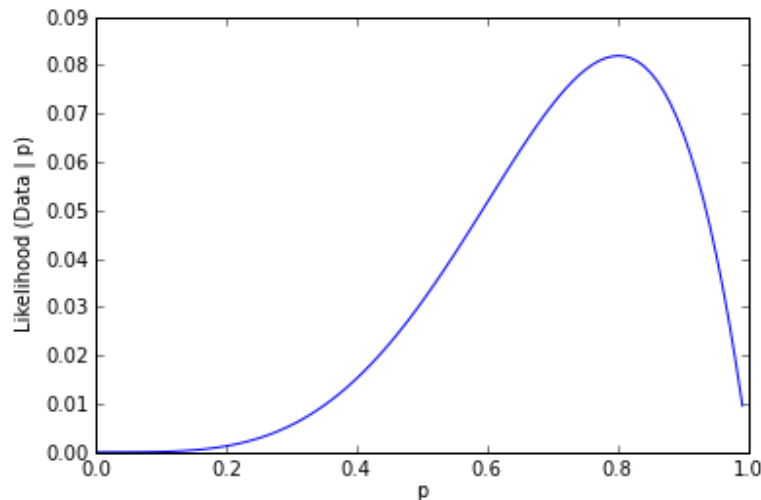
MAP Estimation

► Model: $\text{Be}(p)$

Data: 1,1,0,1,1

Likelihood: $p^4(1 - p)$

Prior: $U(0,1)$



$$\hat{p}_{MAP} = \hat{p}_{MLE} = \frac{n_1}{n_0 + n_1}$$

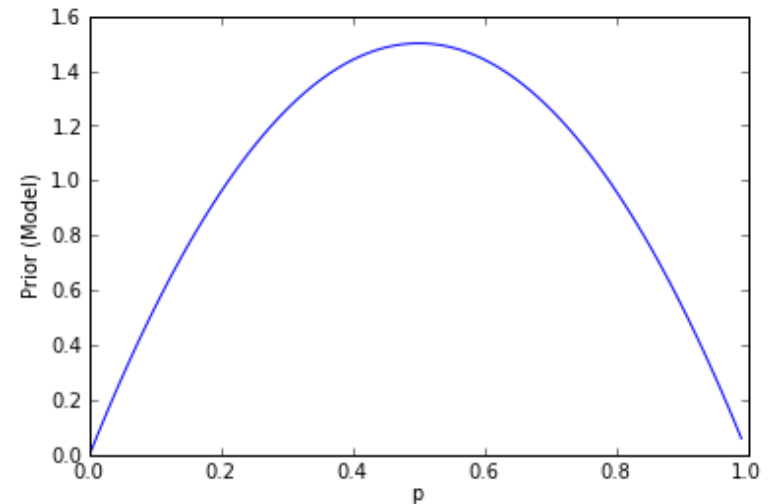
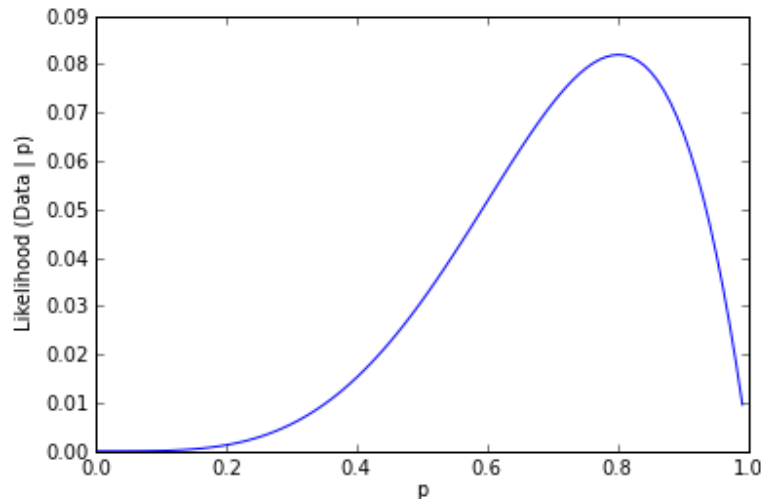
MAP Estimation

► Model: $\text{Be}(p)$

Data: 1,1,0,1,1

Likelihood: $p^4(1 - p)$

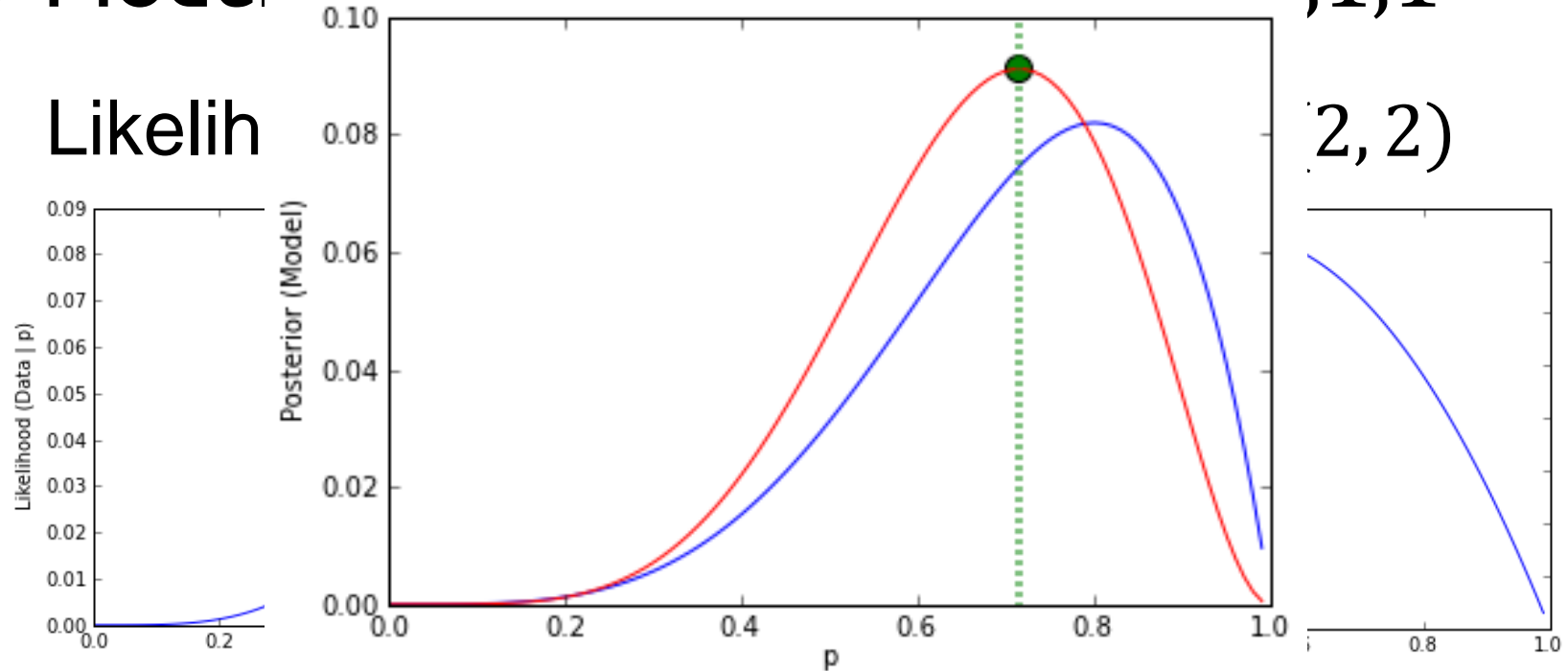
Prior: $\text{Beta}(2, 2)$



MAP Estimation

► Model: $B_\theta(p)$

Data: 1 1 0,1,1

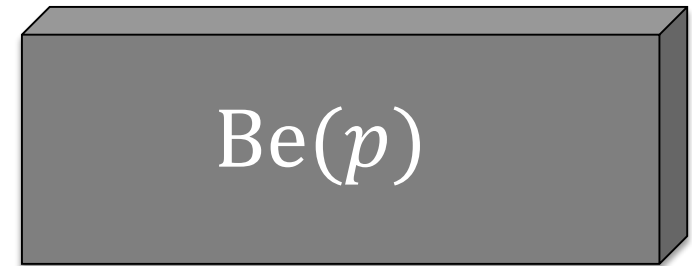


$$\hat{p}_{MAP} = \frac{n_1 + 1}{n_0 + n_1 + 2}$$

Example: Biased coin

Step 2: Parameter inference

1,1,0,1,1



$$p = 0.7 ?$$

MAP Estimation

$$\operatorname{argmax}_{\text{Model}} \Pr(\text{Data} \mid \text{Model}) \cdot \Pr(\text{Model})$$

MAP Estimation

$$\operatorname{argmax}_{\text{Model}} \Pr(\text{Data} \mid \text{Model}) \cdot \Pr(\text{Model})$$

$$\operatorname{argmax}_{\text{Model}} \log (\Pr(\text{Data} \mid \text{Model}) \cdot \Pr(\text{Model}))$$

MAP Estimation

$$\operatorname{argmax}_{\text{Model}} \Pr(\text{Data} \mid \text{Model}) \cdot \Pr(\text{Model})$$

$$\operatorname{argmax}_{\text{Model}} \log (\Pr(\text{Data} \mid \text{Model}) \cdot \Pr(\text{Model}))$$

$$\operatorname{argmax}_{\text{Model}} \log \Pr(\text{Data} \mid \text{Model}) + \log \Pr(\text{Model})$$

MAP Estimation

$$\operatorname{argmax}_{\text{Model}} \Pr(\text{Data} \mid \text{Model}) \cdot \Pr(\text{Model})$$

$$\operatorname{argmax}_{\text{Model}} \log (\Pr(\text{Data} \mid \text{Model}) \cdot \Pr(\text{Model}))$$

$$\operatorname{argmax}_{\text{Model}} \log \Pr(\text{Data} \mid \text{Model}) + \log \Pr(\text{Model})$$

MAP Estimation

$$\operatorname{argmax}_{\text{Model}} \Pr(\text{Data} \mid \text{Model}) \cdot \Pr(\text{Model})$$

$$\operatorname{argmax}_{\text{Model}} \log (\Pr(\text{Data} \mid \text{Model}) \cdot \Pr(\text{Model}))$$

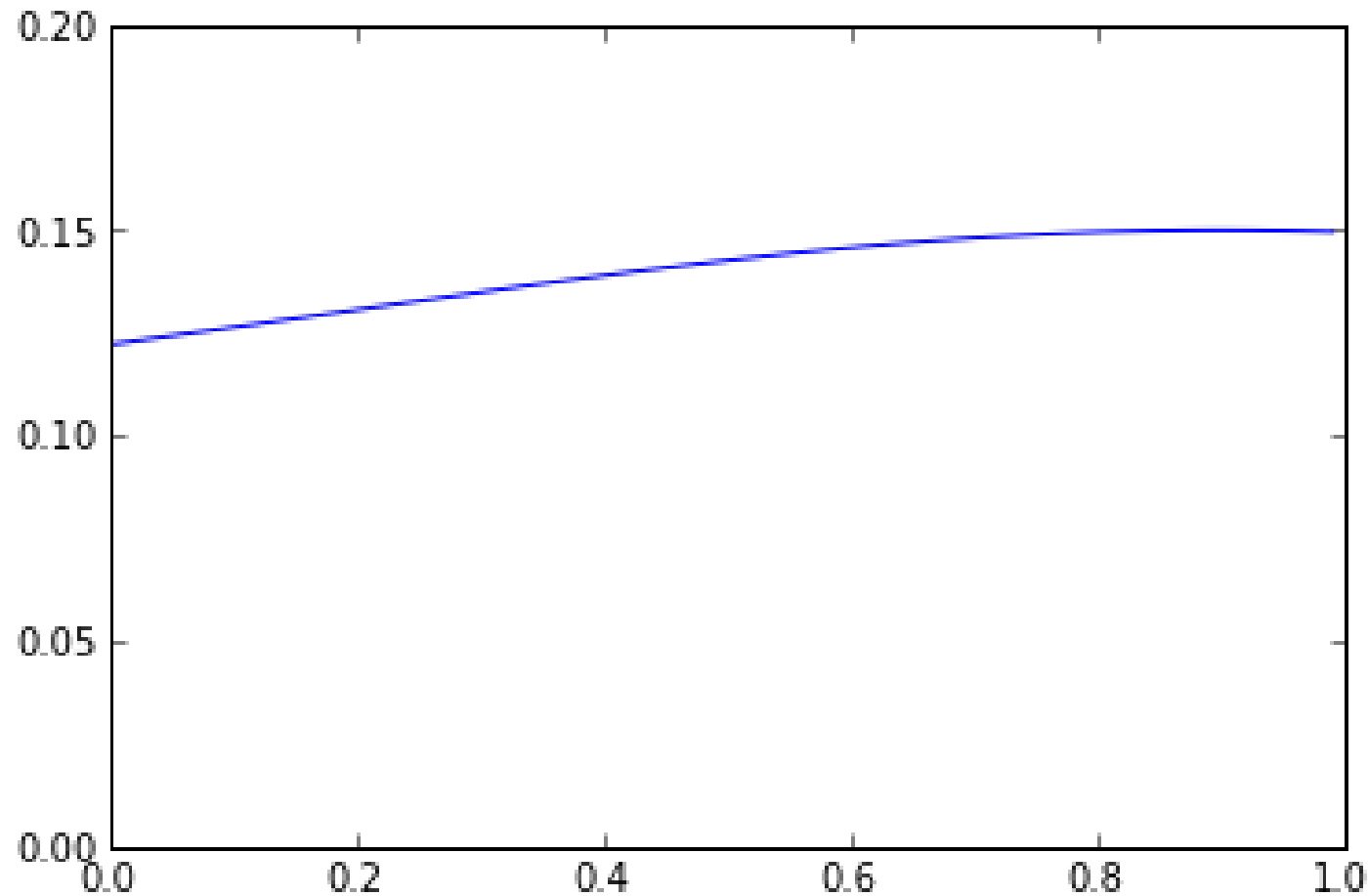
$$\operatorname{argmax}_{\text{Model}} \log \Pr(\text{Data} \mid \text{Model}) + \log \Pr(\text{Model})$$

$$\operatorname{argmin}_{\mathbf{w}} \text{Error}(\text{Data}, \mathbf{w}) + \text{Complexity}(\mathbf{w})$$

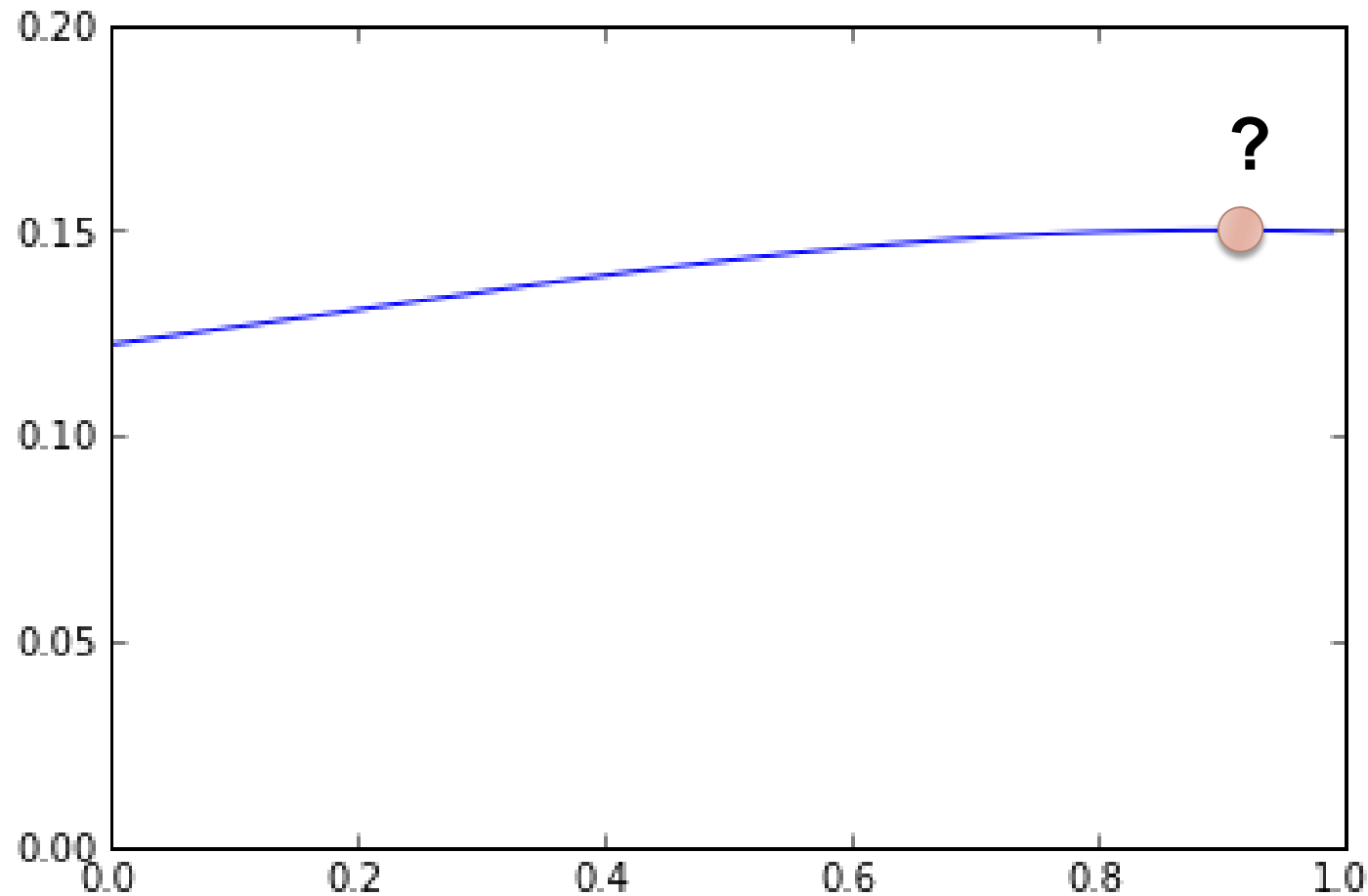
Problems of MAP estimation



Problems of MAP estimation

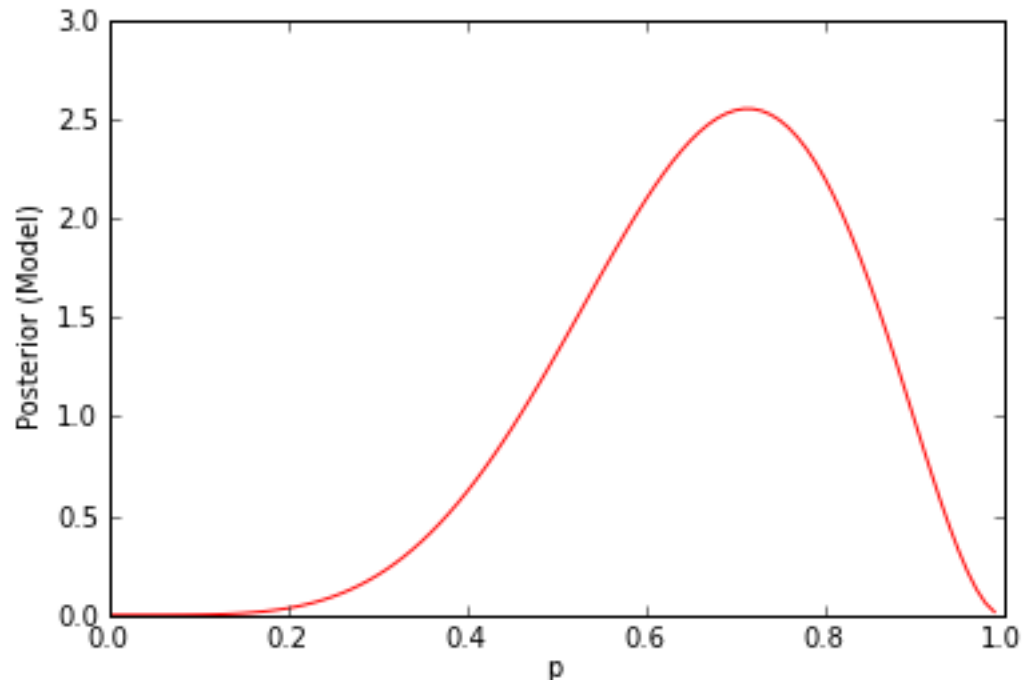


Problems of MAP estimation



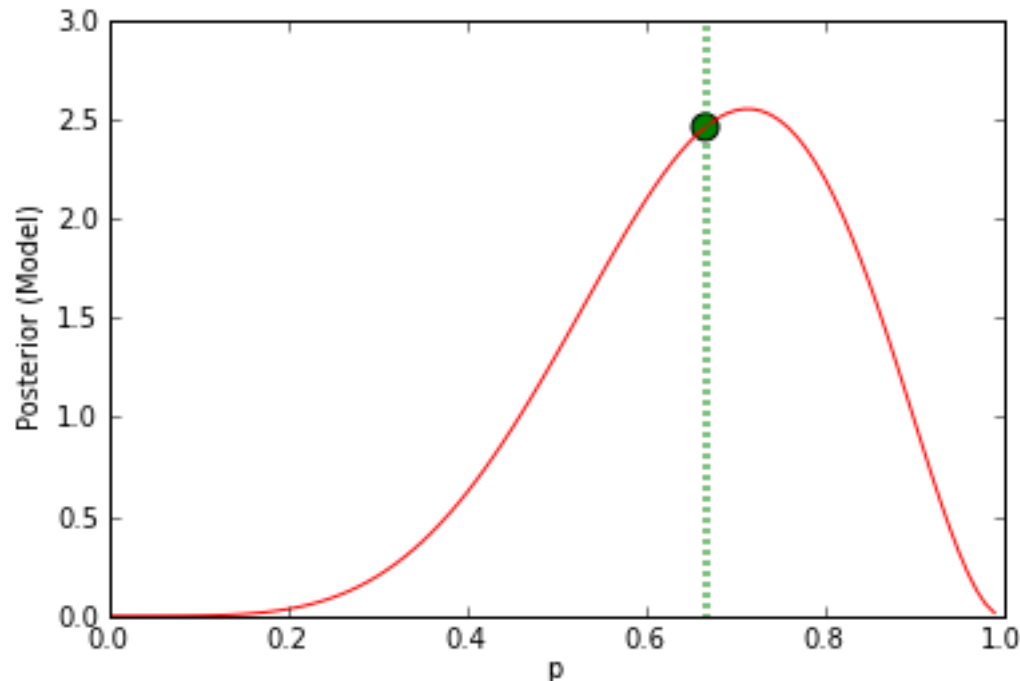
Bayesian estimation

- Pick the model with minimal *expected risk*
 $E(\text{Model} | \text{Data})$



Bayesian estimation

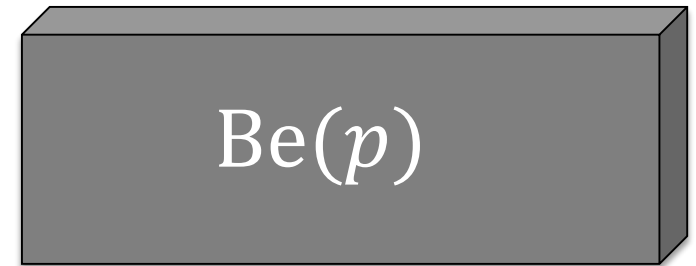
- Pick the model with minimal *expected risk*
 $E(\text{Model} | \text{Data})$



Example: Biased coin

Step 2: Parameter inference

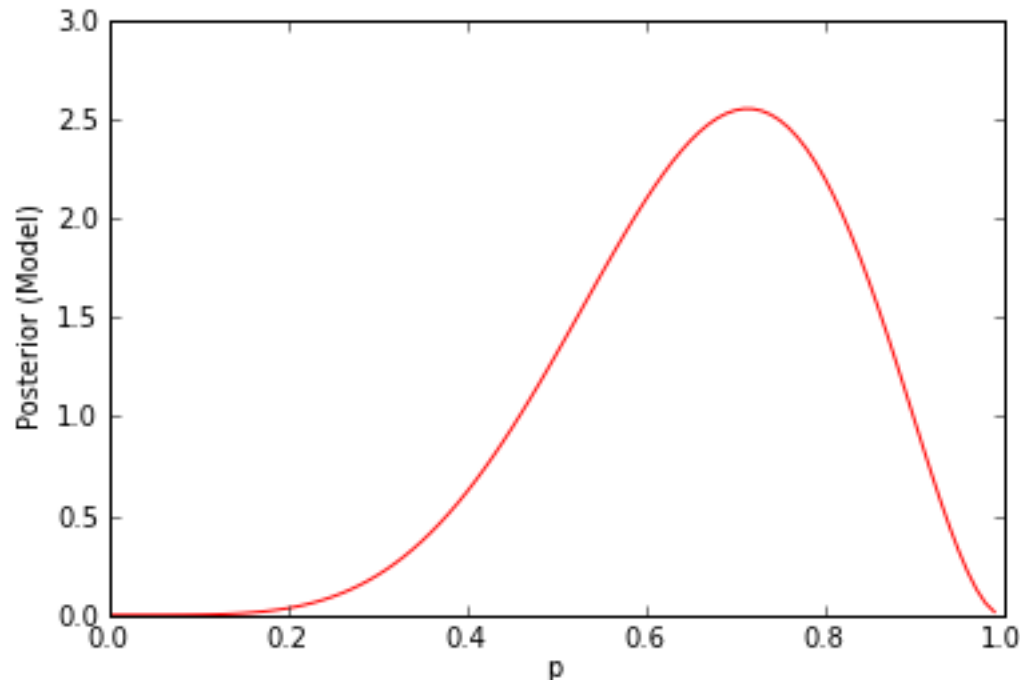
1,1,0,1,1



$$p = 0.65 ?$$

Bayesian estimation +

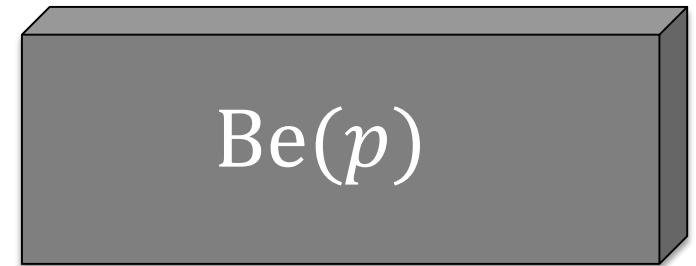
- Use the full posterior distribution
 $\Pr(\text{Model} | \text{Data})$



Example: Biased coin

Step 2: Parameter inference

1,1,0,1,1



$p \sim \text{Beta}(5, 2) ?$

Quiz

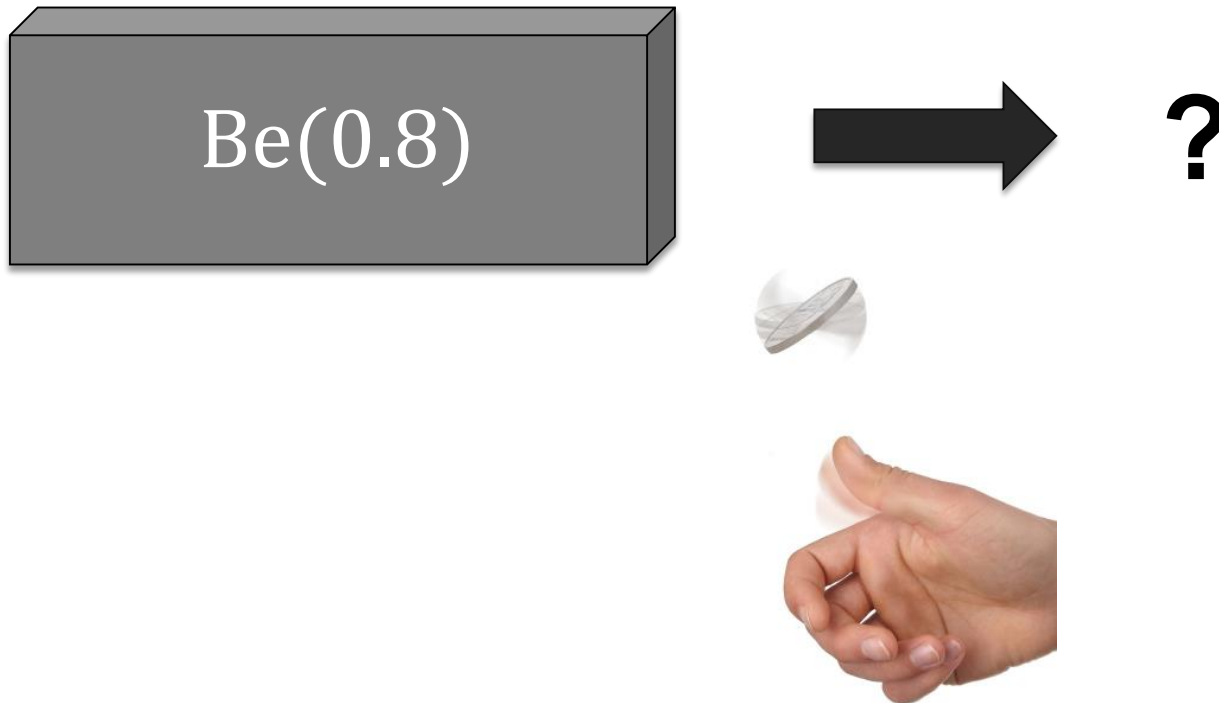


- ▶ Three major model inference methods are:



Example: Biased coin

Step 3: Decision making



Decision Theory

Model



| X | P(X) |
|----------|-------------|
| 1 | 0.8 |
| 0 | 0.2 |

Decision Theory

Model



| X | P(X) |
|----------|-------------|
| 1 | 0.8 |
| 0 | 0.2 |



Decision

| X | P(X) | “1” | “0” |
|----------|-------------|------------|------------|
| 1 | 0.8 | | |
| 0 | 0.2 | | |

Decision Theory

Model



| X | P(X) |
|----------|-------------|
| 1 | 0.8 |
| 0 | 0.2 |



Decision

| X | P(X) | “1” | “0” |
|----------|-------------|------------|------------|
| 1 | 0.8 | 0 | 1 |
| 0 | 0.2 | 5 | 0 |

Decision Theory

Model



| X | P(X) |
|----------|-------------|
| 1 | 0.8 |
| 0 | 0.2 |



Decision

| X | P(X) | “1” | “0” |
|---------------|-------------|------------|------------|
| 1 | 0.8 | 0 | 1 |
| 0 | 0.2 | 5 | 0 |
| Expected Risk | | | |

Decision Theory

Model



| X | P(X) |
|----------|-------------|
| 1 | 0.8 |
| 0 | 0.2 |



Decision

| X | P(X) | “1” | “0” |
|---------------|-------------|------------|------------|
| 1 | 0.8 | 0 | 1 |
| 0 | 0.2 | 5 | 0 |
| Expected Risk | | 1 | 0.8 |

Decision Theory

Model



| X | P(X) |
|----------|-------------|
| 1 | 0.8 |
| 0 | 0.2 |

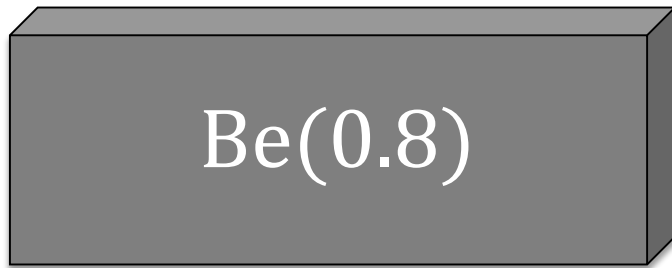


Decision

| X | P(X) | “1” | “0” |
|---------------|-------------|------------|------------|
| 1 | 0.8 | 0 | 1 |
| 0 | 0.2 | 5 | 0 |
| Expected Risk | | 1 | 0.8 |

Example: Biased coin

Step 3: Decision making

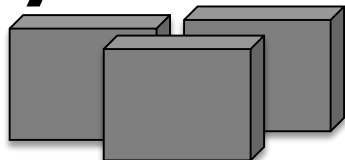


predict “0”



Summary

- ▶ **Probability** for modeling



- ▶ **Statistics** for estimation

MLE

MAP

- ▶ **Decision theory** for prediction



The Tennis Dataset

| Day | Outlook | Temp | Humidity | Wind | <i>PlayTennis</i> |
|-----|-----------------|-------------|---------------|---------------|-------------------|
| D1 | <i>Sunny</i> | <i>Hot</i> | <i>High</i> | <i>Weak</i> | <i>No</i> |
| D2 | <i>Sunny</i> | <i>Hot</i> | <i>High</i> | <i>Strong</i> | <i>No</i> |
| D3 | <i>Overcast</i> | <i>Hot</i> | <i>High</i> | <i>Weak</i> | <i>Yes</i> |
| D4 | <i>Rain</i> | <i>Mild</i> | <i>High</i> | <i>Weak</i> | <i>Yes</i> |
| D5 | <i>Rain</i> | <i>Cool</i> | <i>Normal</i> | <i>Weak</i> | <i>Yes</i> |
| D6 | <i>Rain</i> | <i>Cool</i> | <i>Normal</i> | <i>Strong</i> | <i>No</i> |
| D7 | <i>Overcast</i> | <i>Cool</i> | <i>Normal</i> | <i>Strong</i> | <i>Yes</i> |
| D8 | <i>Sunny</i> | <i>Mild</i> | <i>High</i> | <i>Weak</i> | <i>No</i> |
| D9 | <i>Sunny</i> | <i>Cool</i> | <i>Normal</i> | <i>Weak</i> | <i>Yes</i> |
| D10 | <i>Rain</i> | <i>Mild</i> | <i>Normal</i> | <i>Weak</i> | <i>Yes</i> |
| D11 | <i>Sunny</i> | <i>Mild</i> | <i>Normal</i> | <i>Strong</i> | <i>Yes</i> |
| D12 | <i>Overcast</i> | <i>Mild</i> | <i>High</i> | <i>Strong</i> | <i>Yes</i> |
| D13 | <i>Overcast</i> | <i>Hot</i> | <i>Normal</i> | <i>Weak</i> | <i>Yes</i> |
| D14 | <i>Rain</i> | <i>Mild</i> | <i>High</i> | <i>Strong</i> | <i>No</i> |

Shall we play tennis today?

| <i>PlayTennis</i> |
|-------------------|
| <i>No</i> |
| <i>No</i> |
| <i>Yes</i> |
| <i>Yes</i> |
| <i>Yes</i> |
| <i>No</i> |
| <i>Yes</i> |
| <i>No</i> |
| <i>Yes</i> |
| <i>Yes</i> |
| <i>Yes</i> |
| <i>Yes</i> |
| <i>Yes</i> |
| <i>No</i> |

Shall we play tennis today?

| <i>PlayTennis</i> |
|-------------------|
| <i>No</i> |
| <i>No</i> |
| <i>Yes</i> |
| <i>Yes</i> |
| <i>Yes</i> |
| <i>No</i> |
| <i>Yes</i> |
| <i>No</i> |
| <i>Yes</i> |
| <i>Yes</i> |
| <i>Yes</i> |
| <i>Yes</i> |
| <i>Yes</i> |
| <i>No</i> |

Estimate a
probabilistic model
and predict:

$$\Pr(\text{Yes}) = 9/14 = 0.64$$

$$\Pr(\text{No}) = 5/14 = 0.36$$

➔ Yes

It's windy today. Tennis, anyone?

| Wind | <i>PlayTennis</i> |
|---------------|-------------------|
| <i>Weak</i> | <i>No</i> |
| <i>Strong</i> | <i>No</i> |
| <i>Weak</i> | <i>Yes</i> |
| <i>Weak</i> | <i>Yes</i> |
| <i>Weak</i> | <i>Yes</i> |
| <i>Strong</i> | <i>No</i> |
| <i>Strong</i> | <i>Yes</i> |
| <i>Weak</i> | <i>No</i> |
| <i>Weak</i> | <i>Yes</i> |
| <i>Weak</i> | <i>Yes</i> |
| <i>Strong</i> | <i>Yes</i> |
| <i>Strong</i> | <i>Yes</i> |
| <i>Weak</i> | <i>Yes</i> |
| <i>Strong</i> | <i>No</i> |

It's windy today. Tennis, anyone?

| Wind | <i>PlayTennis</i> |
|---------------|-------------------|
| <i>Weak</i> | <i>No</i> |
| <i>Strong</i> | <i>No</i> |
| <i>Weak</i> | <i>Yes</i> |
| <i>Weak</i> | <i>Yes</i> |
| <i>Weak</i> | <i>Yes</i> |
| <i>Strong</i> | <i>No</i> |
| <i>Strong</i> | <i>Yes</i> |
| <i>Weak</i> | <i>No</i> |
| <i>Weak</i> | <i>Yes</i> |
| <i>Weak</i> | <i>Yes</i> |
| <i>Strong</i> | <i>Yes</i> |
| <i>Strong</i> | <i>Yes</i> |
| <i>Weak</i> | <i>Yes</i> |
| <i>Strong</i> | <i>No</i> |

$$\text{Pr}(\text{Yes} \mid \text{Weak}) = 6/8$$

$$\text{Pr}(\text{No} \mid \text{Weak}) = 2/8$$

$$\text{Pr}(\text{Yes} \mid \text{Strong}) = 3/6$$

$$\text{Pr}(\text{No} \mid \text{Strong}) = 3/6$$

More attributes

| Humidity | Wind | PlayTennis |
|----------|--------|------------|
| High | Weak | No |
| High | Strong | No |
| High | Weak | Yes |
| High | Weak | Yes |
| Normal | Weak | Yes |
| Normal | Strong | No |
| Normal | Strong | Yes |
| High | Weak | No |
| Normal | Weak | Yes |
| Normal | Weak | Yes |
| Normal | Strong | Yes |
| High | Strong | Yes |
| Normal | Weak | Yes |
| High | Strong | No |

$$\Pr(\text{Yes} \mid \text{High, Weak}) = 2/4$$

$$\Pr(\text{No} \mid \text{High, Weak}) = 2/4$$

$$\Pr(\text{Yes} \mid \text{High, Strong}) = 1/3$$

$$\Pr(\text{No} \mid \text{High, Strong}) = 2/3$$

...

The Bayesian Classifier

In general:

1. **Estimate from data:**

$$\Pr(\text{Class} | x_1, x_2, x_3, \dots)$$

2. **For a given instance** (x_1, x_2, x_3, \dots)
predict class whose conditional
probability is greater*:

$$\Pr(C_1 | x_1, x_2, x_3, \dots) > \Pr(C_2 | x_1, x_2, x_3, \dots)$$

→ predict C_1

Problem

► We need exponential amount of data

| Humidity | Wind | <i>PlayTennis</i> |
|---------------|---------------|-------------------|
| <i>High</i> | <i>Weak</i> | <i>No</i> |
| <i>High</i> | <i>Strong</i> | <i>No</i> |
| <i>High</i> | <i>Weak</i> | <i>Yes</i> |
| <i>High</i> | <i>Weak</i> | <i>Yes</i> |
| <i>Normal</i> | <i>Weak</i> | <i>Yes</i> |
| <i>Normal</i> | <i>Strong</i> | <i>No</i> |
| <i>Normal</i> | <i>Strong</i> | <i>Yes</i> |
| <i>High</i> | <i>Weak</i> | <i>No</i> |
| <i>Normal</i> | <i>Weak</i> | <i>Yes</i> |
| <i>Normal</i> | <i>Weak</i> | <i>Yes</i> |
| <i>Normal</i> | <i>Strong</i> | <i>Yes</i> |
| <i>High</i> | <i>Strong</i> | <i>Yes</i> |
| <i>Normal</i> | <i>Weak</i> | <i>Yes</i> |
| <i>High</i> | <i>Strong</i> | <i>No</i> |

$$\Pr(\text{Yes} \mid \text{High, Weak}) = 2/4$$

$$\Pr(\text{No} \mid \text{High, Weak}) = 2/4$$

$$\Pr(\text{Yes} \mid \text{High, Strong}) = 1/3$$

$$\Pr(\text{No} \mid \text{High, Strong}) = 2/3$$

...

Naïve Bayes Classifier

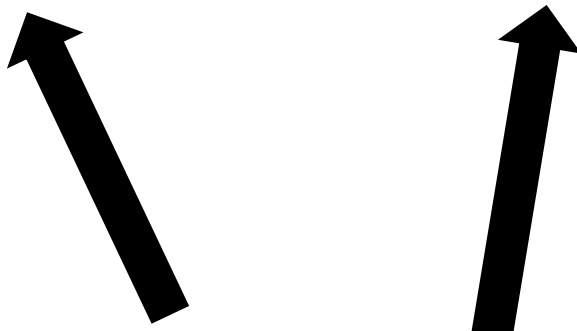
To scale beyond 2-3 attributes, use a hack:

Assume that attributes are independent within each class:

$$\begin{aligned} \Pr(x_1, x_2, x_3 \mid \text{Class}) \\ = \Pr(x_1 \mid \text{Class}) \Pr(x_2 \mid \text{Class}) \Pr(x_3 \mid \text{Class}) \dots \end{aligned}$$

Naïve Bayes Classifier

1. $\Pr(C_1 | \mathbf{x}) > \Pr(C_2 | \mathbf{x})$
 \rightarrow predict C_1


$$P(A|B) = \frac{P(B|A) P(A)}{P(B)}$$

Naïve Bayes Classifier

1. $\Pr(C_1 | \mathbf{x}) > \Pr(C_2 | \mathbf{x})$
 \rightarrow predict C_1

2. $\frac{\Pr(C_1) \Pr(\mathbf{x} | C_1)}{\Pr(\mathbf{x})} > \frac{\Pr(C_2) \Pr(\mathbf{x} | C_2)}{\Pr(\mathbf{x})}$
 \rightarrow predict C_1

Naïve Bayes Classifier

1. $\Pr(C_1 | \mathbf{x}) > \Pr(C_2 | \mathbf{x})$

→ predict C_1

2. $\frac{\Pr(C_1) \Pr(\mathbf{x} | C_1)}{\Pr(\mathbf{x})} > \frac{\Pr(C_2) \Pr(\mathbf{x} | C_2)}{\Pr(\mathbf{x})}$

→ predict C_1

3. $\Pr(C_1) \Pr(\mathbf{x} | C_1) > \Pr(C_2) \Pr(\mathbf{x} | C_2)$

→ predict C_1

Naïve Bayes Classifier

1. $\Pr(C_1 | \mathbf{x}) > \Pr(C_2 | \mathbf{x})$

→ predict C_1

2. $\frac{\Pr(C_1) \Pr(\mathbf{x} | C_1)}{\Pr(\mathbf{x})} > \frac{\Pr(C_2) \Pr(\mathbf{x} | C_2)}{\Pr(\mathbf{x})}$

→ predict C_1

3. $\Pr(C_1) \Pr(\mathbf{x} | C_1) > \Pr(C_2) \Pr(\mathbf{x} | C_2)$

→ predict C_1

4. $\Pr(C_1) \cdot \Pr(x_1 | C_1) \Pr(x_2 | C_1) \dots \Pr(x_m | C_1) >$
 $\Pr(C_2) \cdot \Pr(x_1 | C_2) \Pr(x_2 | C_2) \dots \Pr(x_m | C_2)$

→ predict C_1

Naïve Bayes Classifier

1. $\Pr(C_1 | \mathbf{x}) > \Pr(C_2 | \mathbf{x})$
 \rightarrow predict C_1

2. $\frac{\Pr(C_1) \Pr(\mathbf{x} | C_1)}{\Pr(\mathbf{x})} > \frac{\Pr(C_2) \Pr(\mathbf{x} | C_2)}{\Pr(\mathbf{x})}$
 \rightarrow predict C_1

3. $\Pr(C_1) \Pr(\mathbf{x} | C_1) > \Pr(C_2) \Pr(\mathbf{x} | C_2)$
 \rightarrow predict C_1

4. $\frac{\Pr(C_1) \cdot \Pr(x_1 | C_1) \Pr(x_2 | C_1) \dots \Pr(x_m | C_1)}{\Pr(C_2) \cdot \Pr(x_1 | C_2) \Pr(x_2 | C_2) \dots \Pr(x_m | C_2)} >$
 \rightarrow predict C_1

Naïve Bayes Classifier

- ▶ Works for both discrete and continuous attributes.

- ▶ The goods:
 - ▶ Easy to implement, efficient
 - ▶ Won't overfit, interpretable
 - ▶ Works better than you would expect (e.g. spam filtering)

- ▶ The bads
 - ▶ “Naïve”, linear
 - ▶ Usually won't work well for too many classes
 - ▶ Not a good probability estimator

Naïve Bayes Classifier

```
from sklearn.naive_bayes import  
    BernoulliNB,  
    MultinomialNB,  
    GaussianNB
```

Quiz

► MLE:

$\operatorname{argmax}_{\text{Model}} \underline{\hspace{10em}}$

► MAP:

$\operatorname{argmax}_{\text{Model}} \underline{\hspace{10em}}$

Quiz



▶ Naïve Bayesian classifier assumption:

- ▶ $\Pr(C|x_1, x_2) = \Pr(C|x_1) \Pr(C|x_2)$
- ▶ $\Pr(x_1, x_2|C) = \Pr(x_1|C) \Pr(x_2|C)$
- ▶ $\Pr(C_1, C_2|x) = \Pr(C_1|x) \Pr(C_2|x)$
- ▶ $\Pr(x|C_1, C_2) = \Pr(x|C_1) \Pr(x|C_2)$



-
- ▶ All machine learning methods we have mentioned so far rely on MLE or MAP
 - ▶ Yes
 - ▶ No

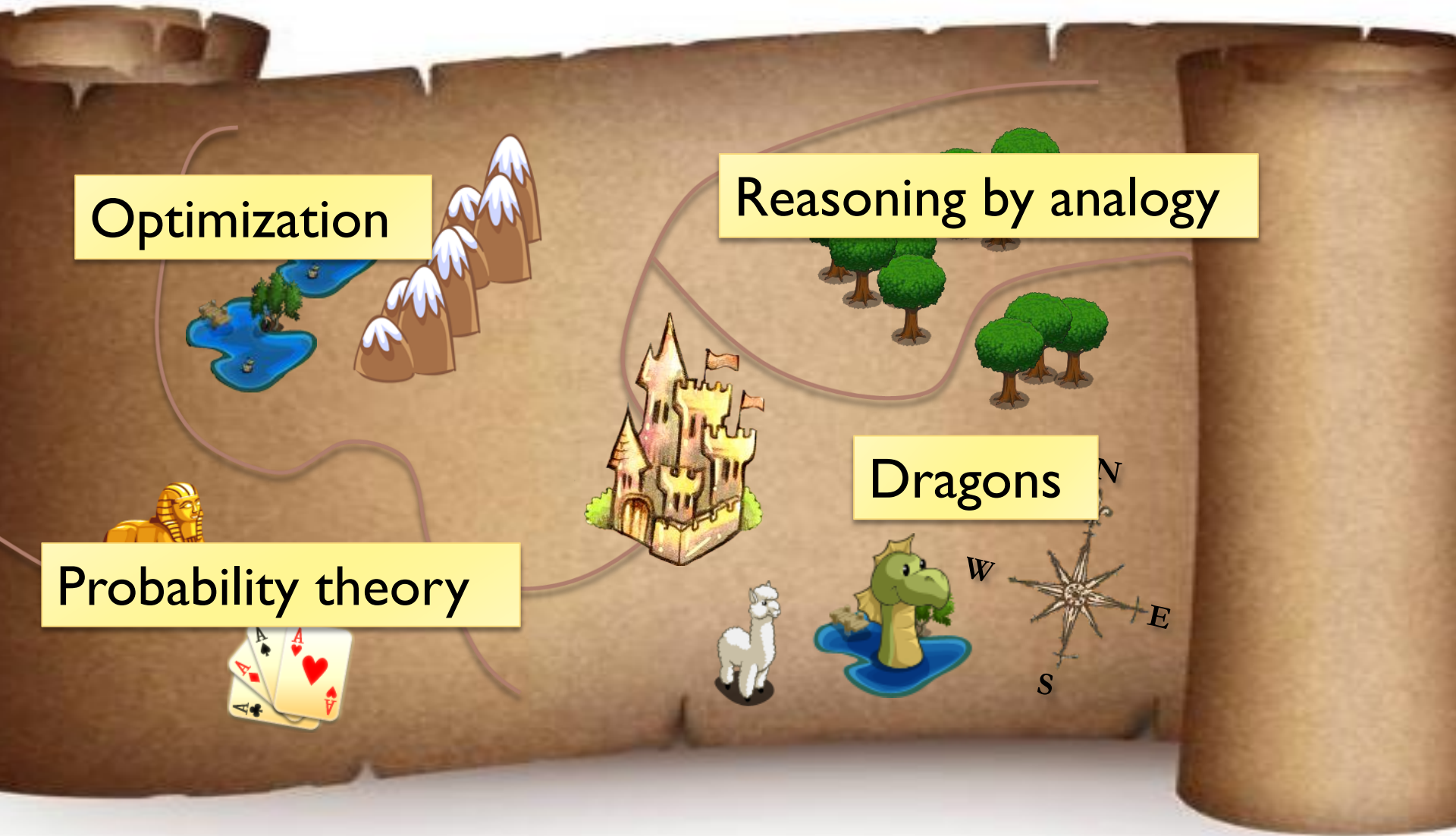
The Land of Machine Learning

Optimization

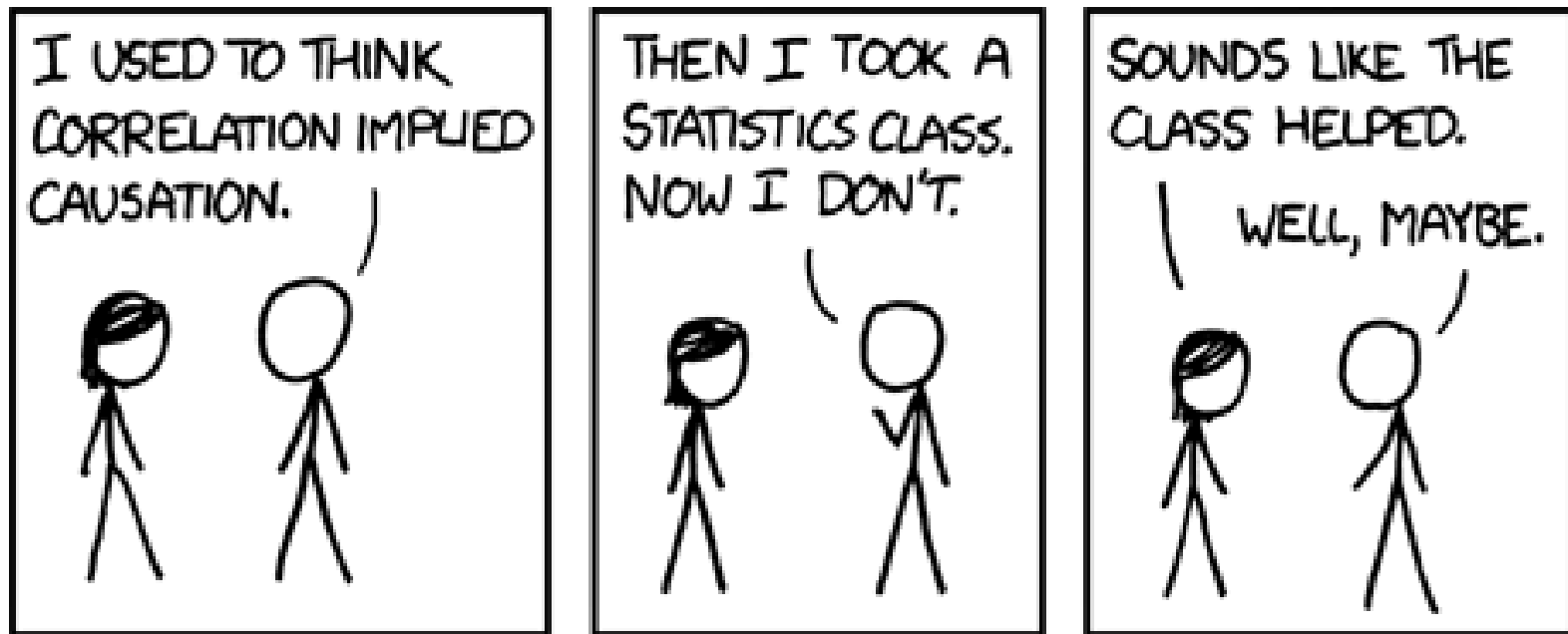
Reasoning by analogy

Probability theory

Dragons



Questions?



<http://xkcd.com/552/>