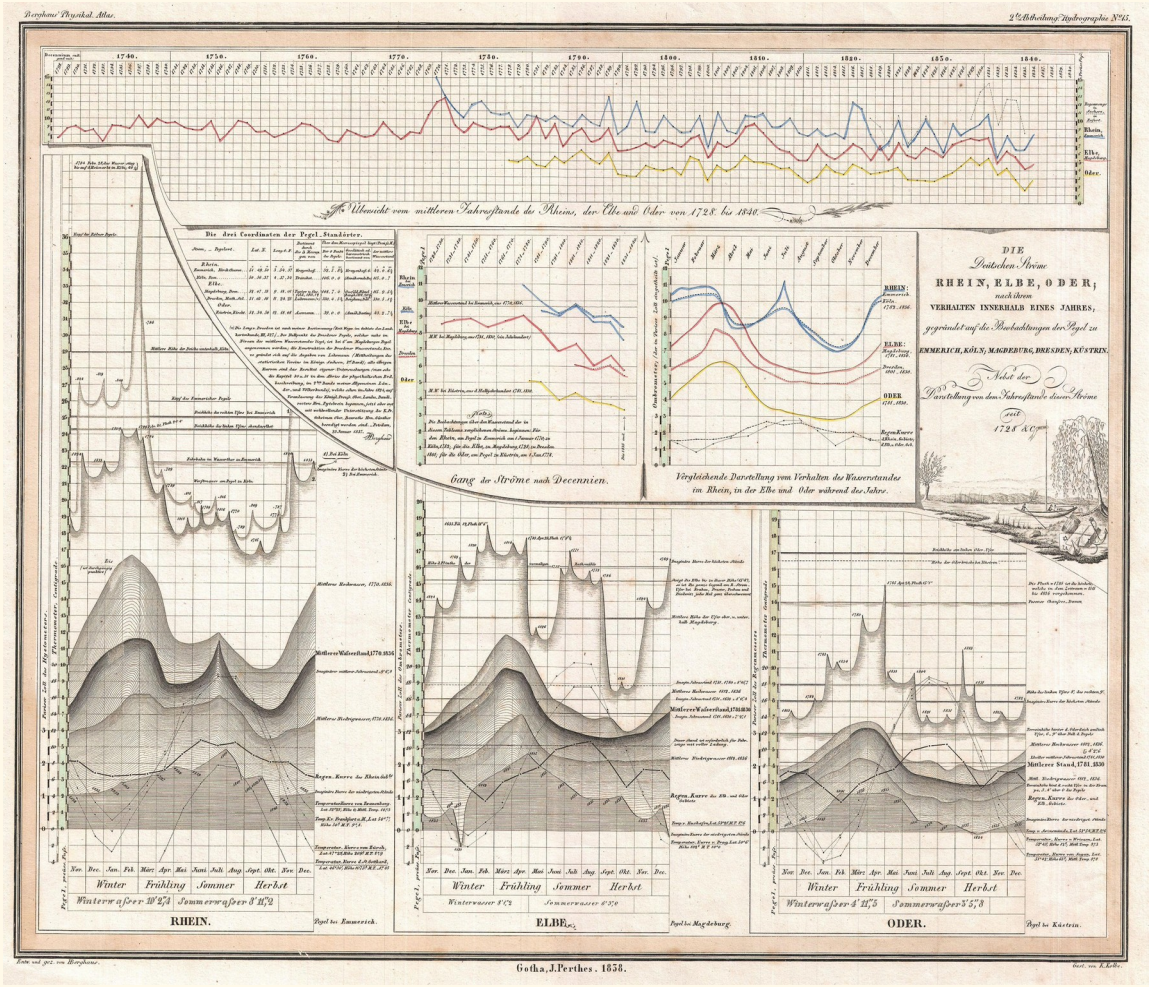
Quantifying Uncertainty

DS3 Göttingen 2019

August 7th 2019

The summer school is funded by the DAAD
with funds of the Federal Foreign Office

Data Science until 80 years ago



River charts by Dr. Heinrich Berghaus from 1838 (public domain)
https://commons.wikimedia.org/wiki/File:1838_Perthes_Chart_of_the_Rhine,_Elbe,_and_Order_Rivers_-_Geographicus_-_RheinElbeOder-perthes-1838.jpg

The summer school is funded by the DAAD with funds of the Federal Foreign Office
/eResearch Alliance Göttingen/

Data Science Today (?)

Tianhe-2 super computer
(creator: user O01326; license: https://creativecommons.org/licenses/by-sa/4.0/deed.en;
source: https://commons.wikimedia.org/w/index.php?title=File:Tianhe-2.jpg&oldid=222583306)

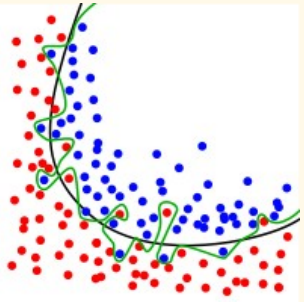
The summer school is funded by the DAAD
with funds of the Federal Foreign Office

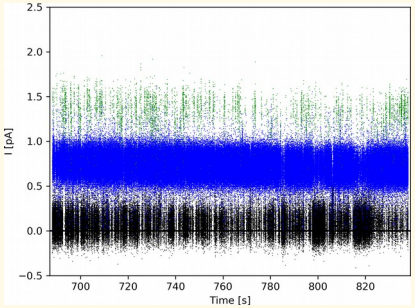
This is more like it…

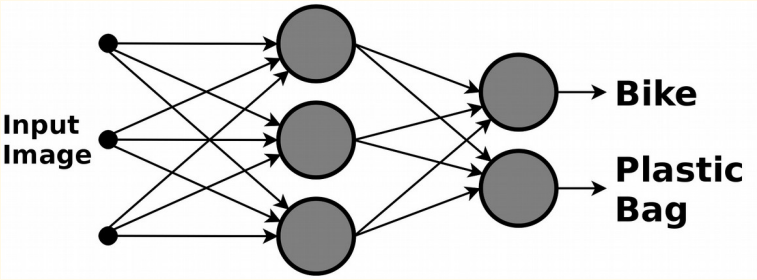
Data Science – How it‘s done.

(creator: user O01326; license: https://creativecommons.org/licenses/by/2.0/deed.en;
source: https://commons.wikimedia.org/w/index.php?title=File:TechCrunch_Disrupt_NY_2015_(17206523159).jpg&oldid=351399991)

A Tale of Two Cities







Statistics

- For small and large data sets
- For low and high dimension
- Classification, Modelling, Summary
- **Quantifies Uncertainty**
- Based on mathematical ingenuity

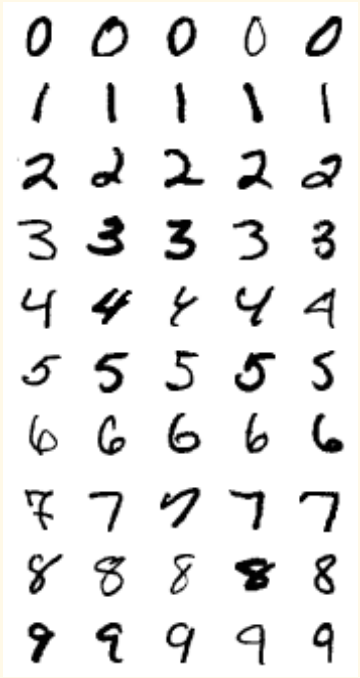
Neural Networks

- Needs large data sets
- Superior in high dimension
- Mostly classification
- Uncertainty empirically checked
- Based on technical ingenuity

The summer school is funded by the DAAD with funds of the Federal Foreign Office

How superior are Neural Networks?

MNIST data set of handwritten digits. Task: identify digits correctly!
large training data set (60,000 images)**, high dimension** (28x28=784)



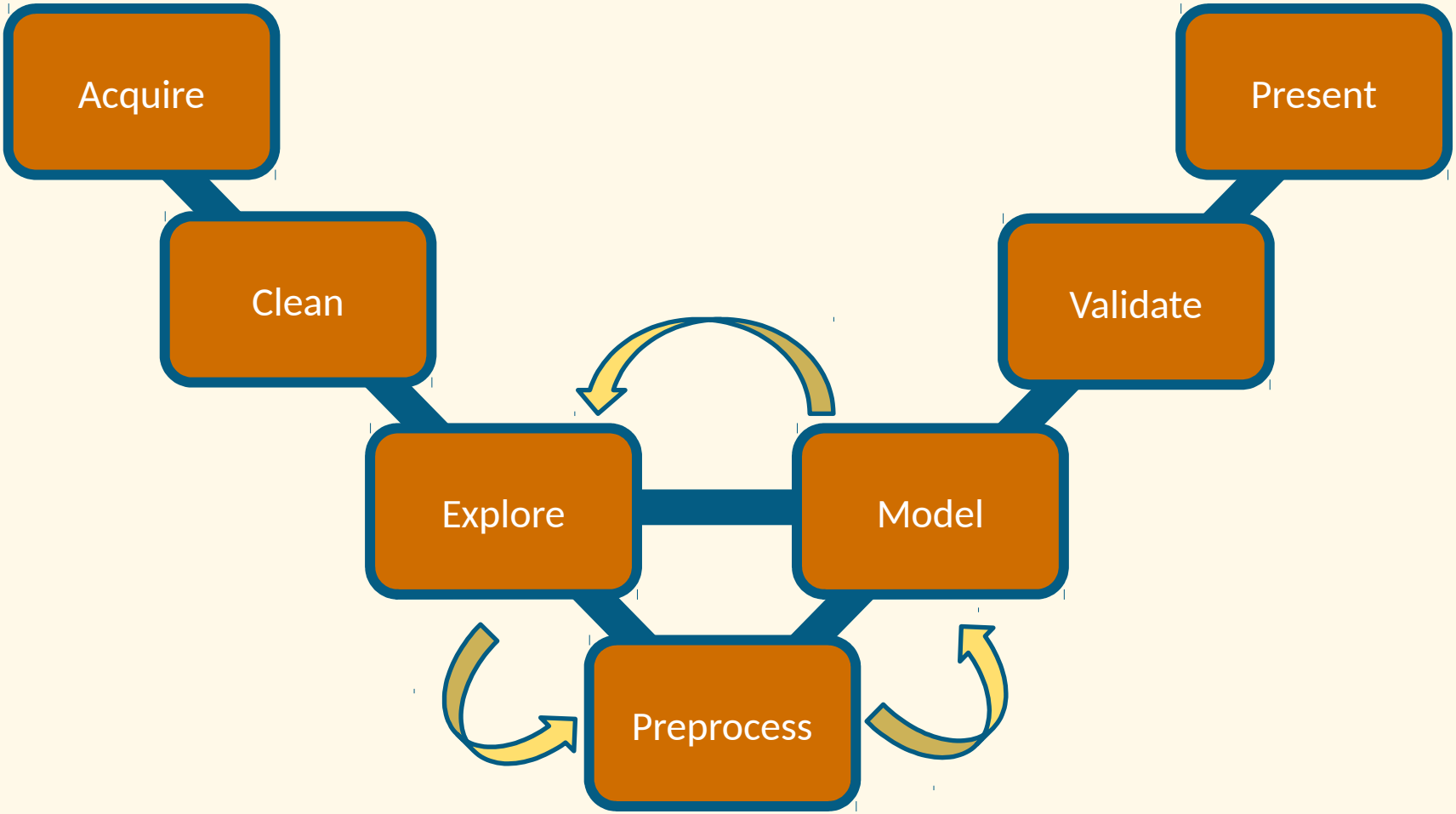
Some example digits[1]

Method	Error Rate	Publication
Gaussian kernel support vector machine	1.4%	
K-nearest-neighbors non-linear deformation (P2DHMDM)	0.52%	Keysers et al. IEEE PAMI 2007
committee of 35 convolutional neural networks	0.23%	Ciresan et al. CVPR 2012

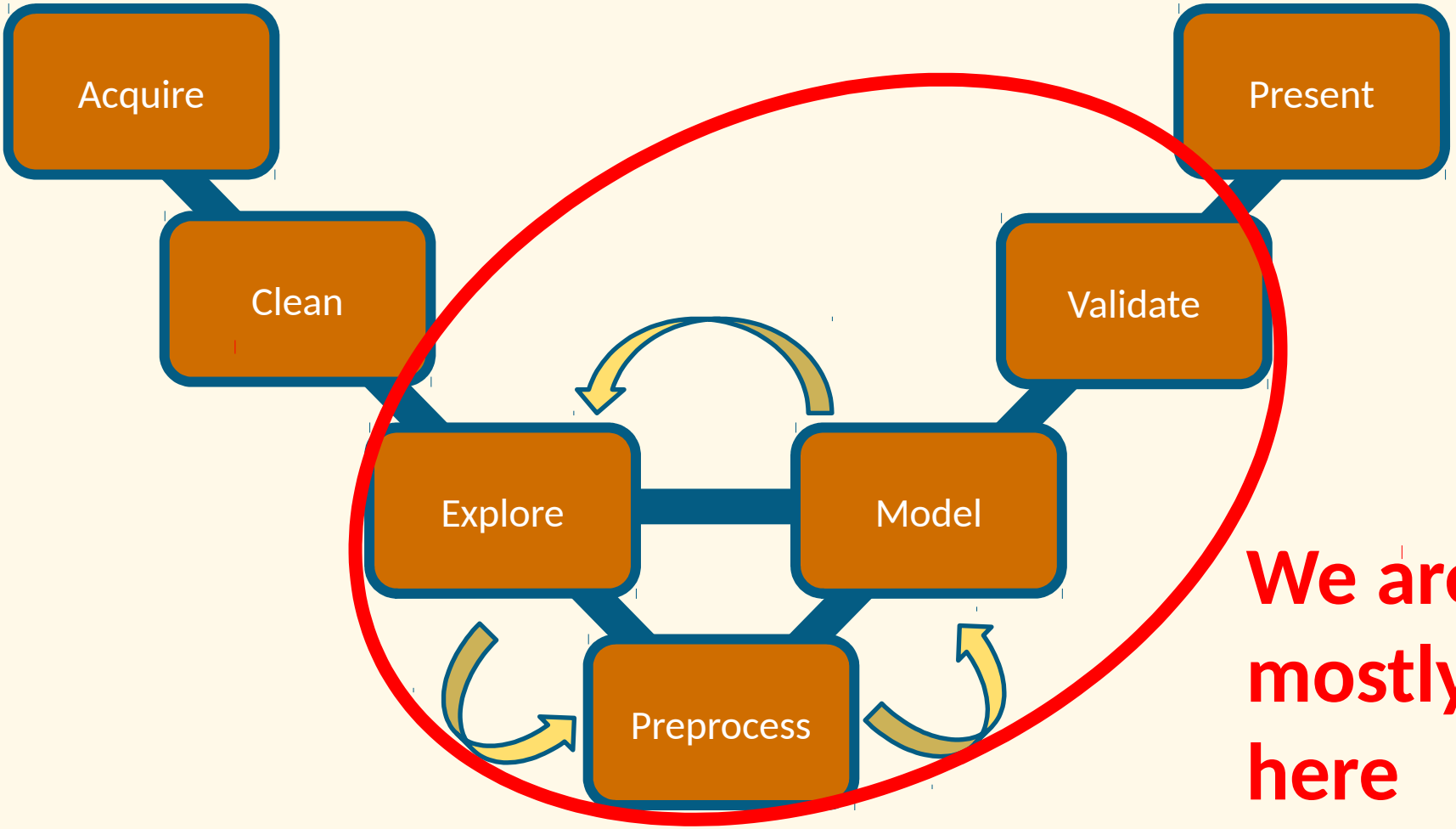
Information from http://yann.lecun.com/exdb/mnist/

[1](original creator: user Jost swd15; license: https://creativecommons.org/licenses/by-sa/4.0/deed.en; source: https://commons.wikimedia.org/w/index.php?title=File:MnistExamples.png&oldid=276154723)

Data Science Building Blocks



Data Science Building Blocks



The summer school is funded by the DAAD with funds of the Federal Foreign Office

Intro: Statistical Experiment

- One system or exact copies, on which one can measure arbitrarily often.
- Measurement results “x” are random; one function f(x) determines the probabilities for all measurements.
- Every measurement result is independent of all other measurements.
- Example: rolling a die, f(x) = 1/6 for x = 1, …, 6.

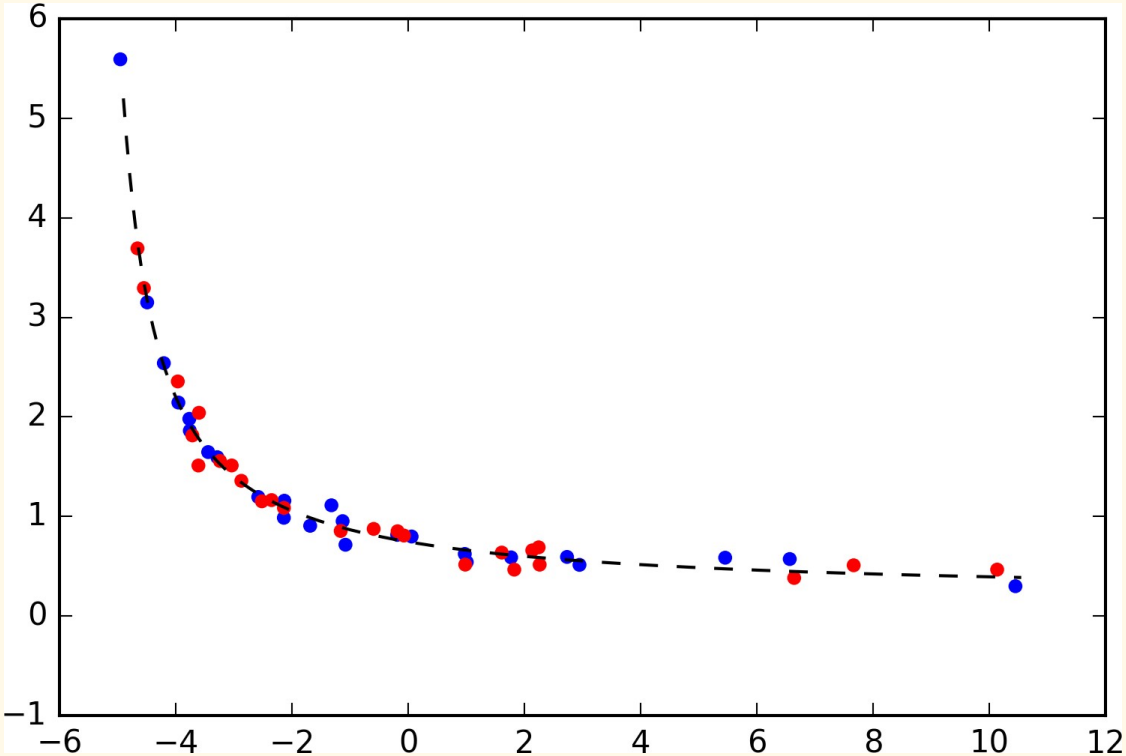
The summer school is funded by the DAAD with funds of the Federal Foreign Office

Regression

- Consider an Experiment, where $x_k = g(t_k) + \varepsilon(t_k)$,
- g(t) is a fixed function,
- values of ε(t) are random (with mean 0) and probabilities are determined by functions $f_t(\varepsilon(t))$.
- Trying to determine g(t) is called regression.

The summer school is funded by the DAAD with funds of the Federal Foreign Office

Overfitting: A Regression Example



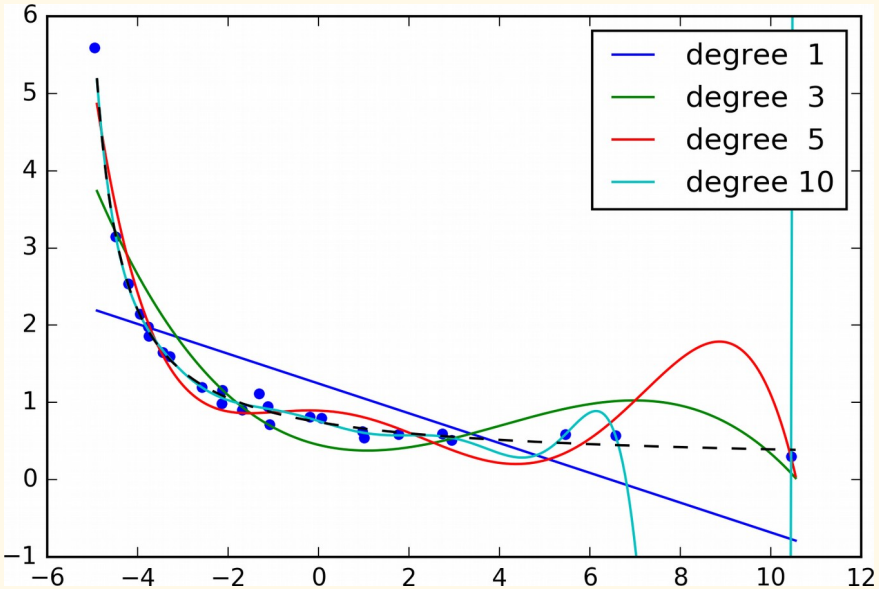
25 Training data

25 Test data

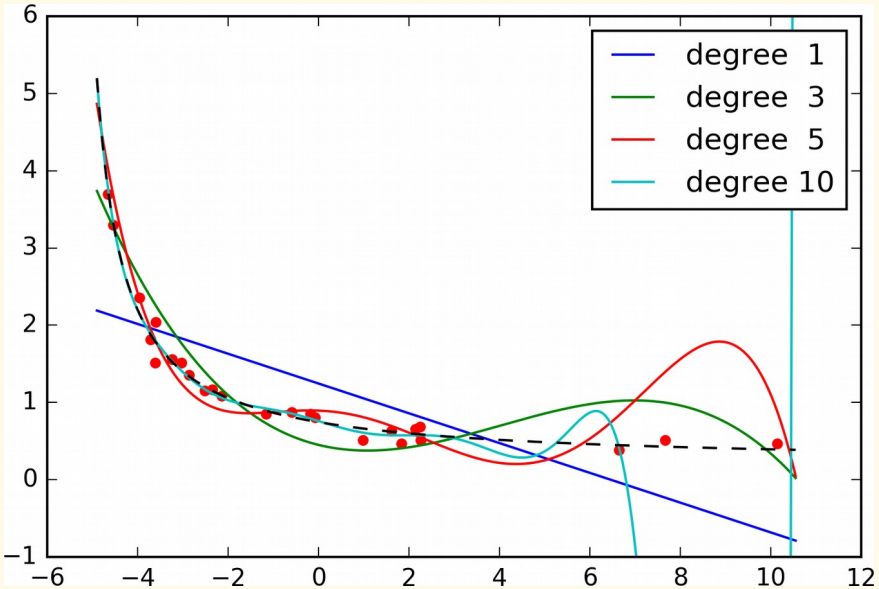
Dashed: g(t)

Polynomial Fits

$$g_d(t) = a_d\, t^d + \ldots + a_1\, t^1 + a_0$$



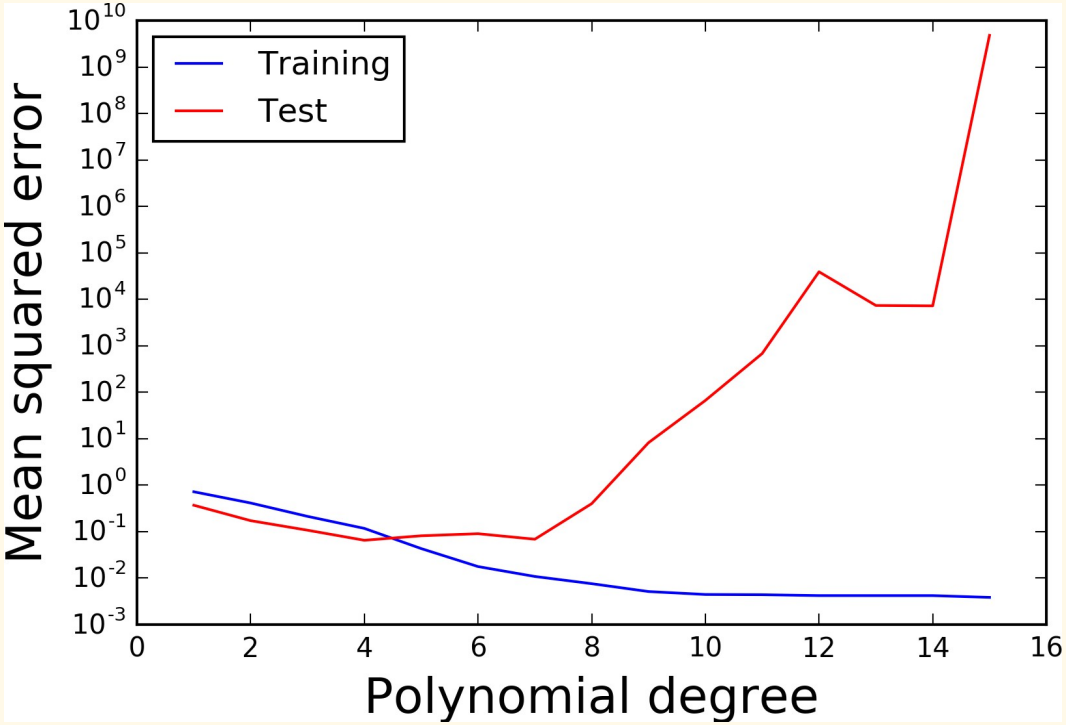
Training data: Good fits



Test data: Bad fits

How to Quantify the Error?

$$\mathrm{MSE} = \frac{1}{n}\left(\left(x_1 - g\left(t_1\right)\right)^2 + \ldots + \left(x_n - g\left(t_n\right)\right)^2\right)$$

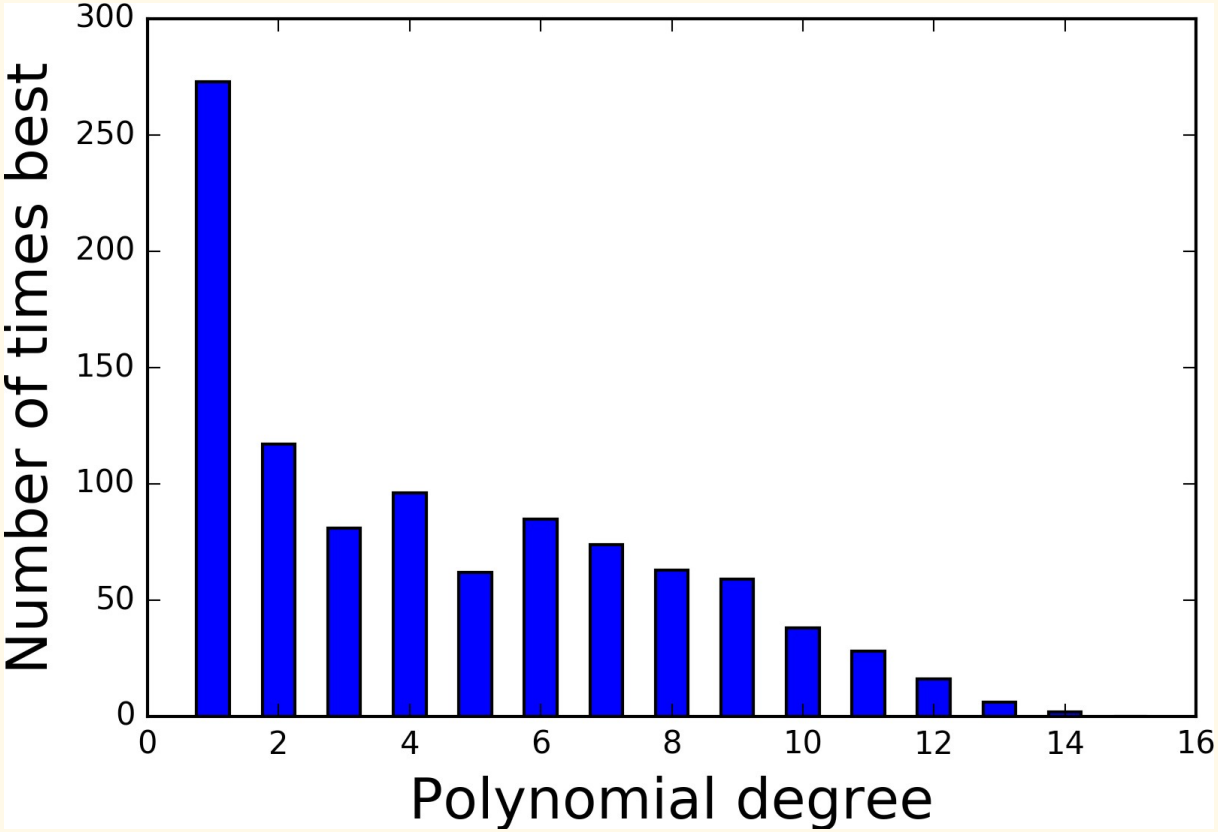


Higher polynomial degree:

better performance on **training** data,

worse performance on **test** data!

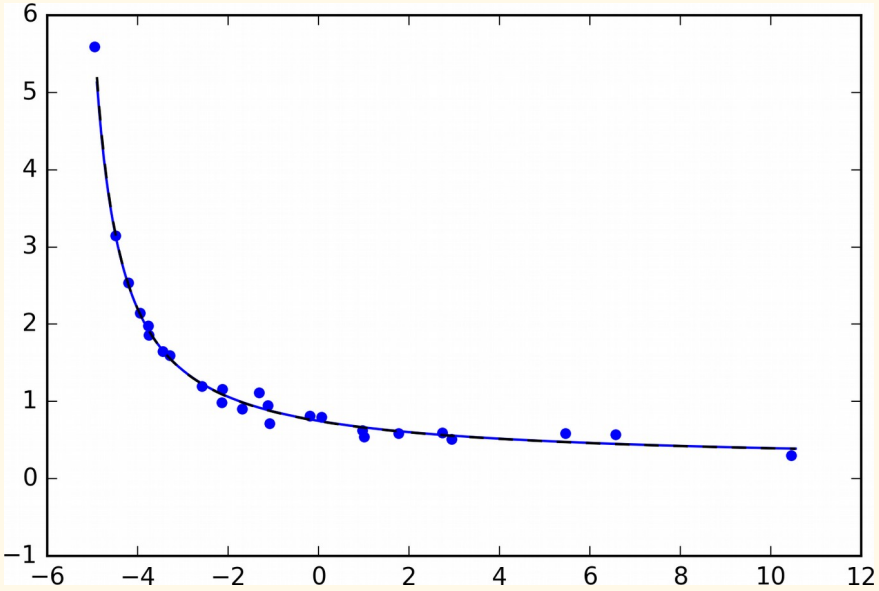
Which Degree is Best?



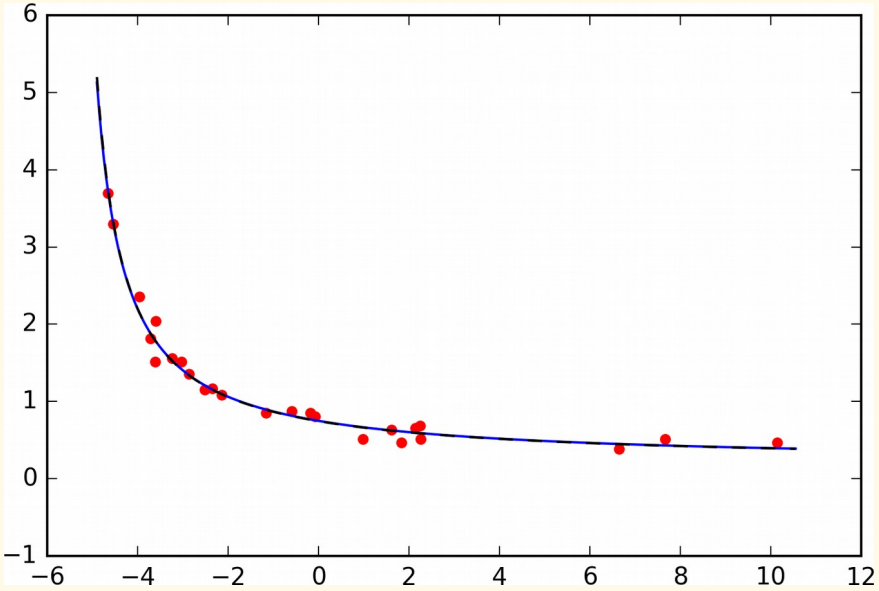
Which polynomial degree has the smallest MSE on test data, 1000 simulated data sets.

A Better Model

$$g(t) = \frac{a_1}{t + a_2} + a_3$$

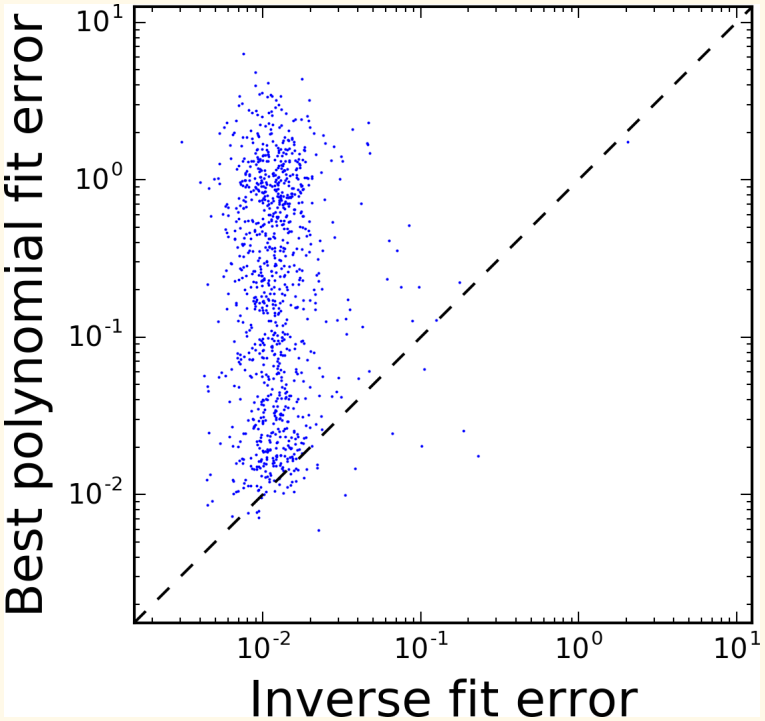


Training data: Good fits



Test data: Good fits

How Much Better is it?



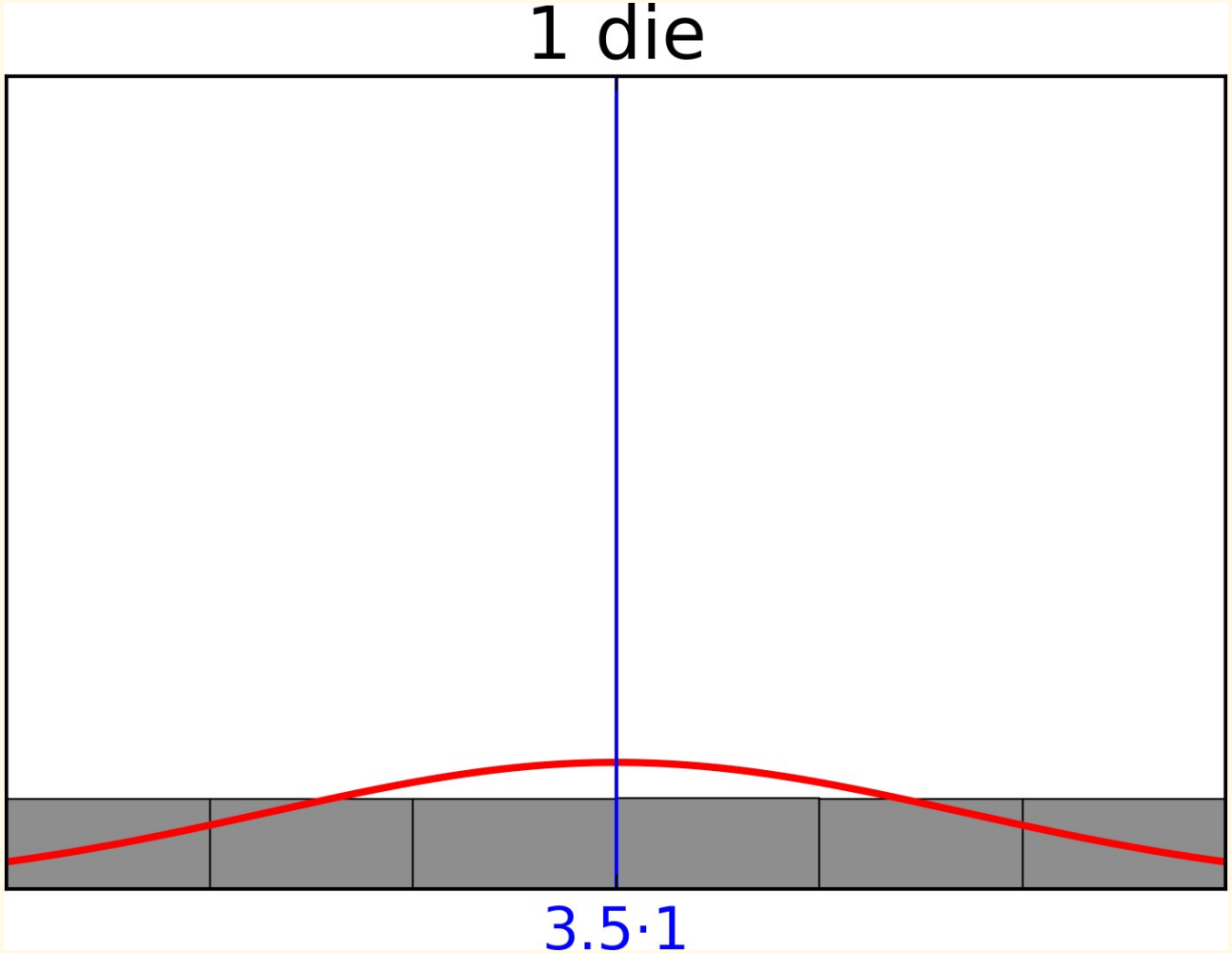
Comparison of MSE on test data for 1000 simulated data sets: 1/x model does much better! (Log scale!)

Observations

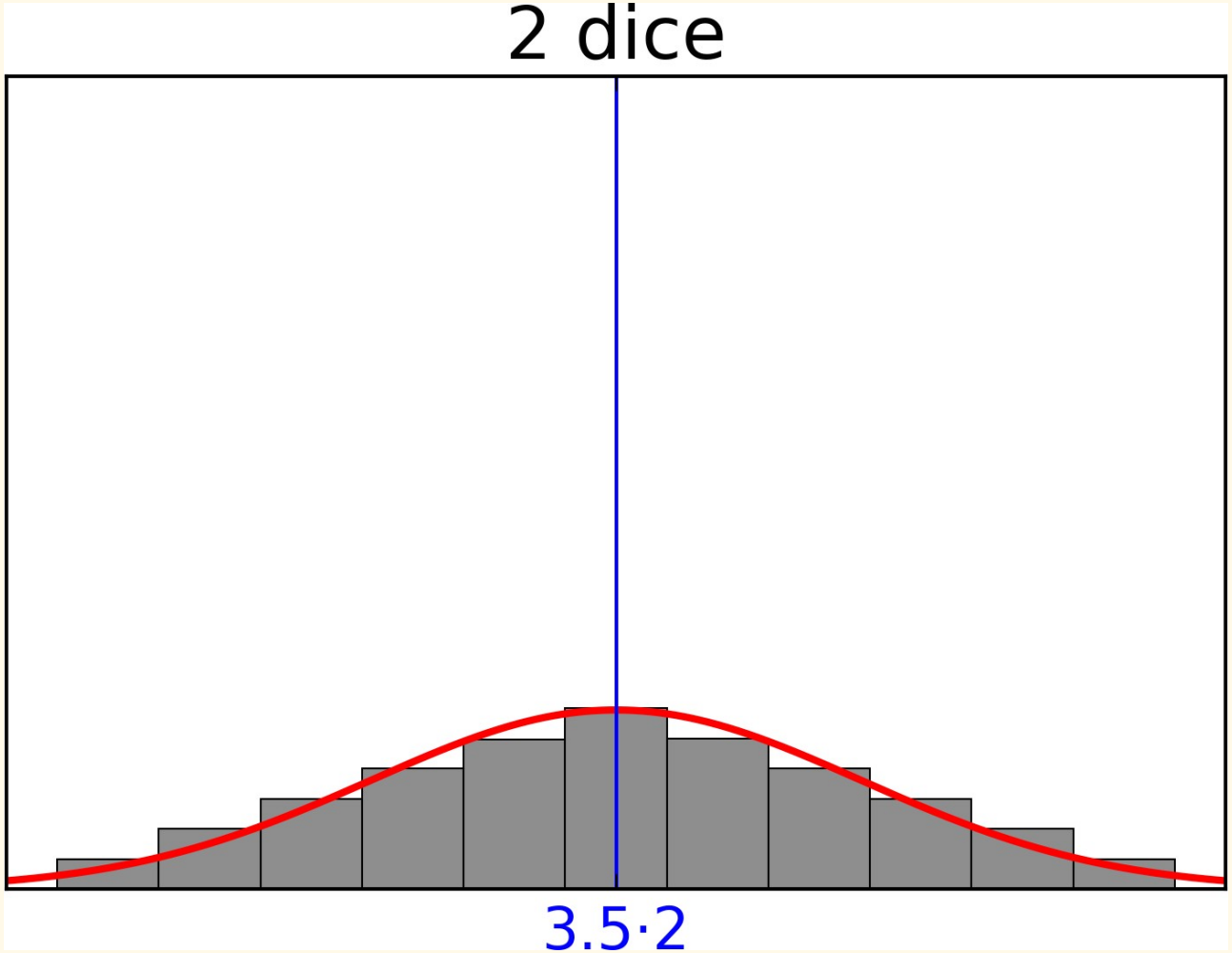
- In general, more model parameters improve the fit for training data, but *too many parameters spoil the fit for test data*!
- Why does this happen? Can one *diagnose this for real data* before looking at test data?
- A *better model* (for the given data) is much more useful than more parameters.

The summer school is funded by the DAAD with funds of the Federal Foreign Office

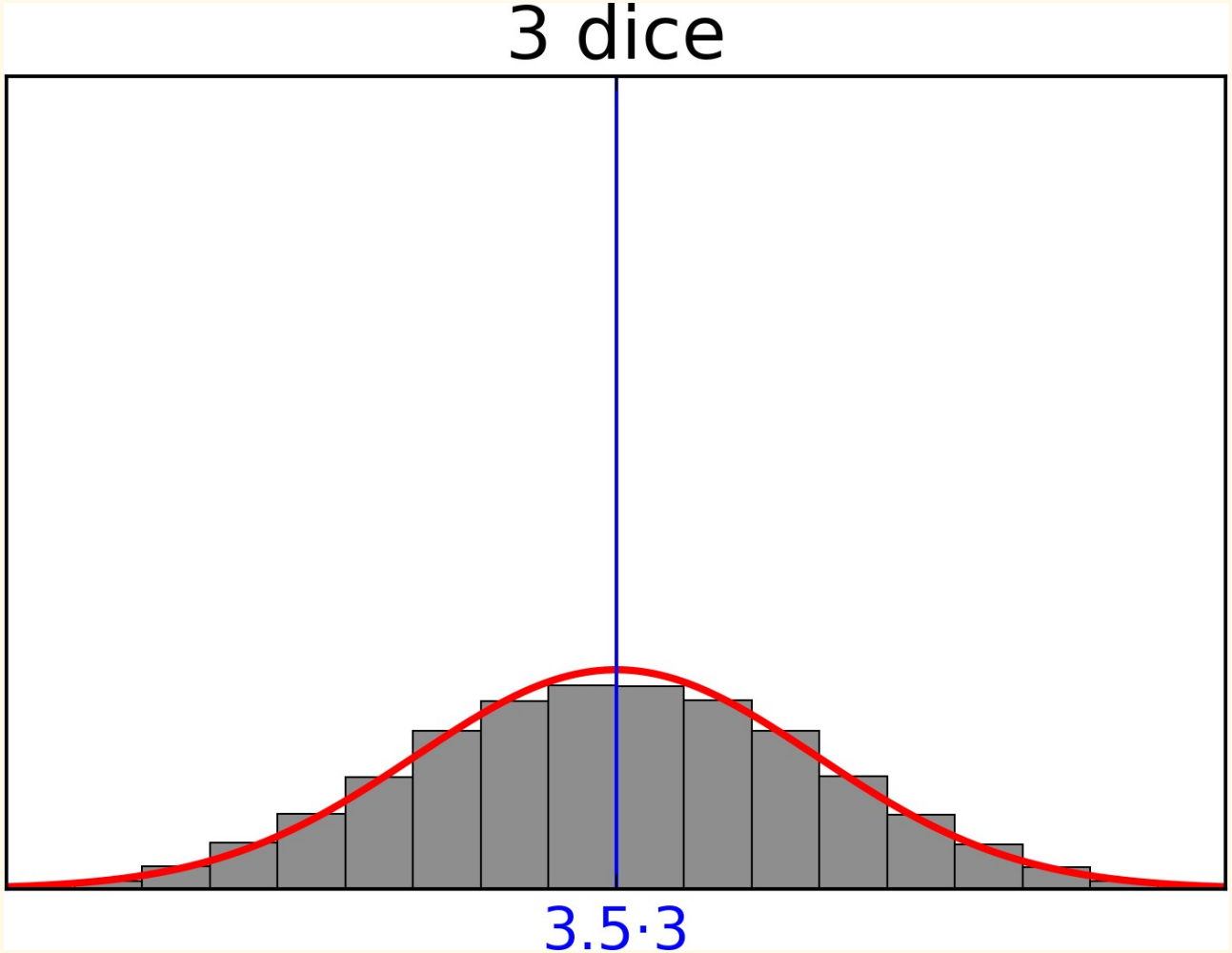
Throwing Dice



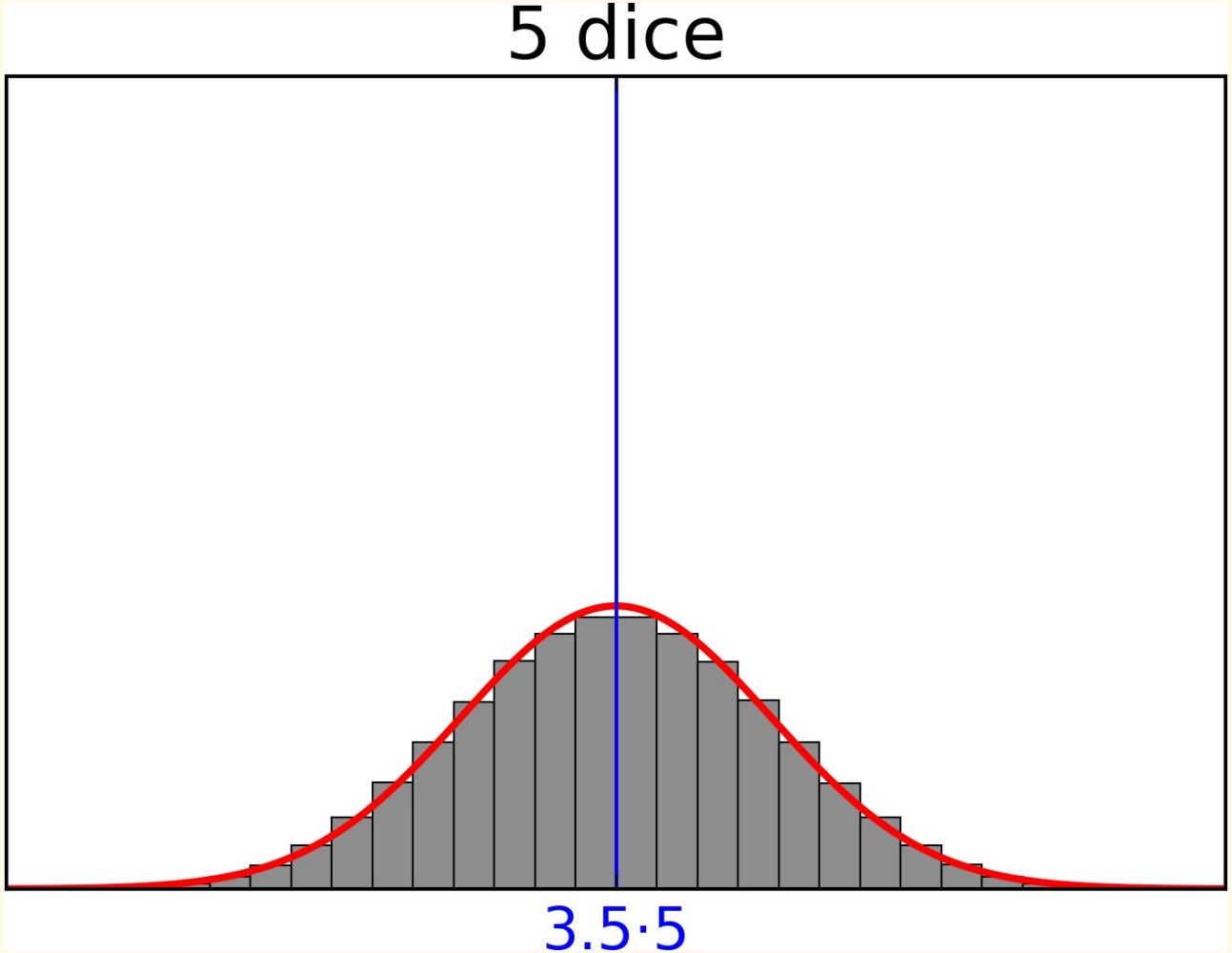
Throwing Dice



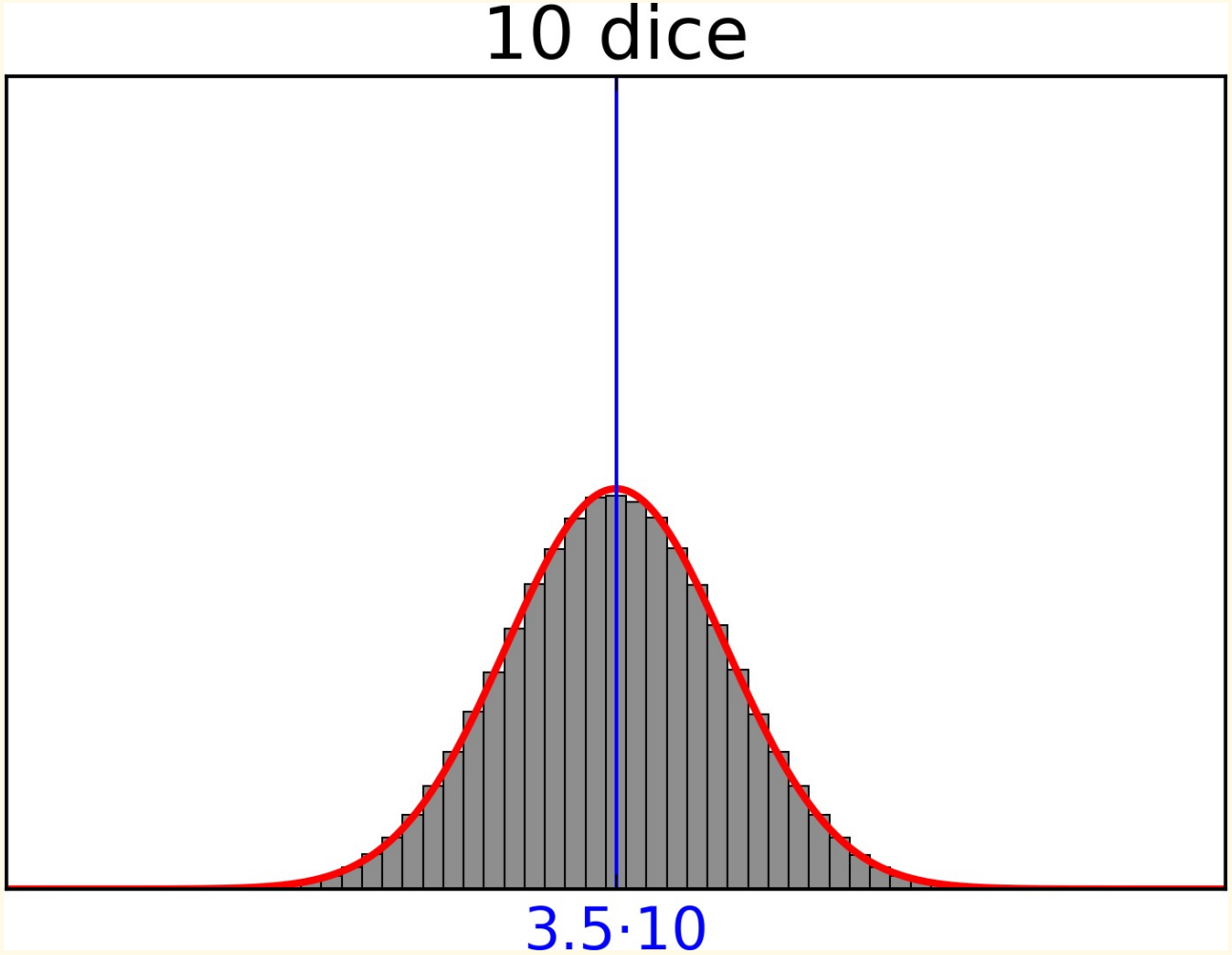
Throwing Dice



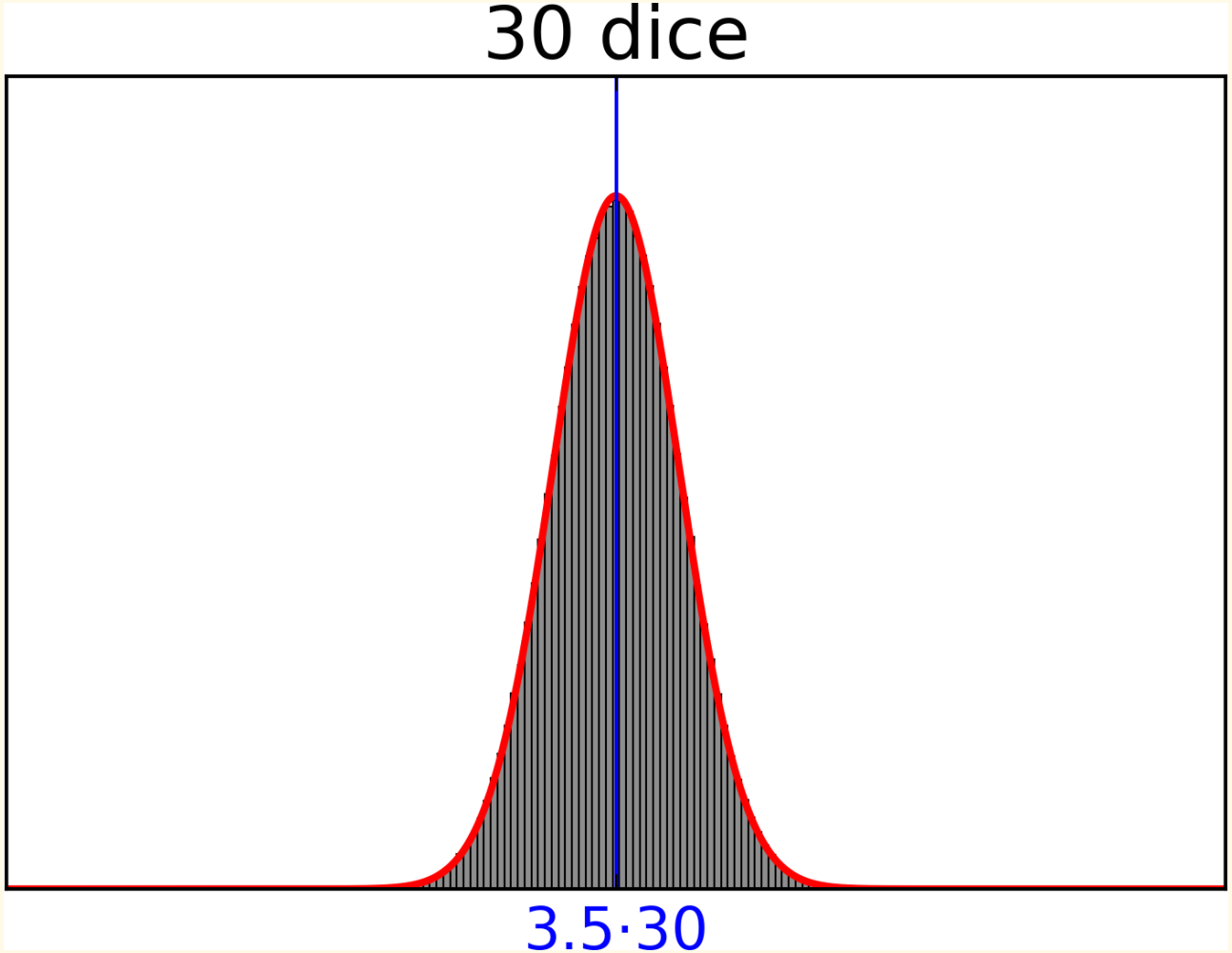
Throwing Dice



Throwing Dice



Throwing Dice

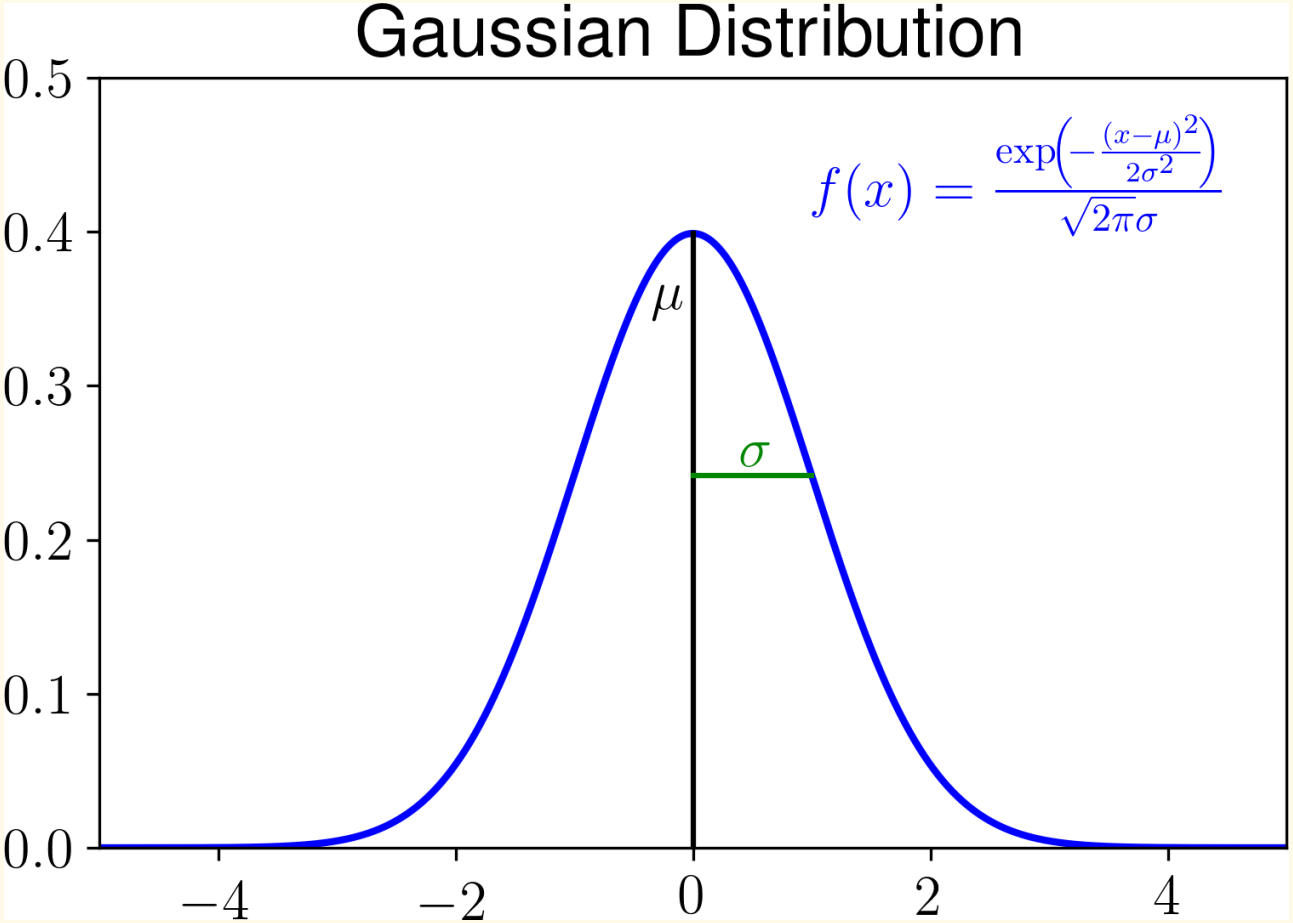


What Happened Here?

- The mean of n dice $\mu_n := (d_1+d_2+\ldots+d_n)/n$ is distributed symmetrically around $\mu_0=3.5$.
- The distribution of μ_n becomes *sharper with increasing n*.
- The distribution approaches a *Gaussian bell shape*.
- Is there some *general principle* at work here?

The summer school is funded by the DAAD with funds of the Federal Foreign Office

Mean μ and Standard Deviation σ



The Central Limit Theorem

Assume the following:

- $x_1, x_2, \ldots, x_n$ are measurements from a statistical experiment and call $\mu_n := (x_1 + x_2 + \ldots + x_n)/n$.
- For the density function f(x) of measurement results, $\mu_0 = \int_{-\infty}^{\infty} x f(x) dx$ and $\int_{-\infty}^{\infty} x^2 f(x) dx$ exist.

Then the distribution of $\sqrt{n}(\mu_n - \mu_0)$ approaches a Gaussian bell shape centered at 0 for $n \to \infty$.

Caveat: In reality, we do not know f(x), so we do not always know, if the Central Limit Theorem holds!

The summer school is funded by the DAAD with funds of the Federal Foreign Office

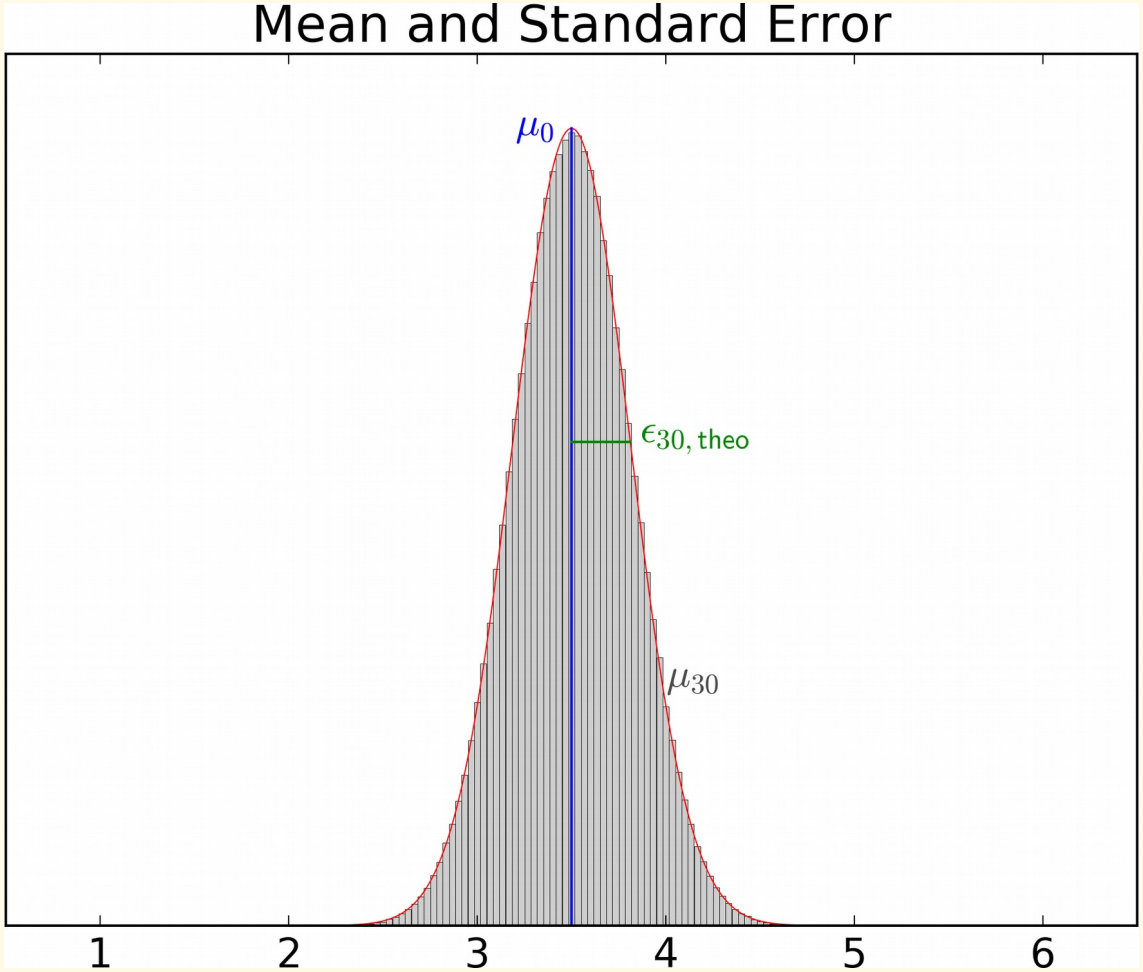
What does this mean?

- $\sigma_n := \sqrt{\frac{1}{n}\left(\left(x_1-\mu_n\right)^2+\left(x_2-\mu_n\right)^2+\ldots+\left(x_n-\mu_n\right)^2\right)}$ is called the standard deviation and estimates the *spread of measured values*.

- The Central Limit Theorem yields the *standard error of the mean* $\epsilon_n := \frac{\sigma_n}{\sqrt{n}} = \frac{1}{n}\sqrt{\left(x_1-\mu_n\right)^2+\left(x_2-\mu_n\right)^2+\ldots+\left(x_n-\mu_n\right)^2}$ which estimates the spread of means of sets of n measured values.

The summer school is funded by the DAAD with funds of the Federal Foreign Office

Behold the Standard Error



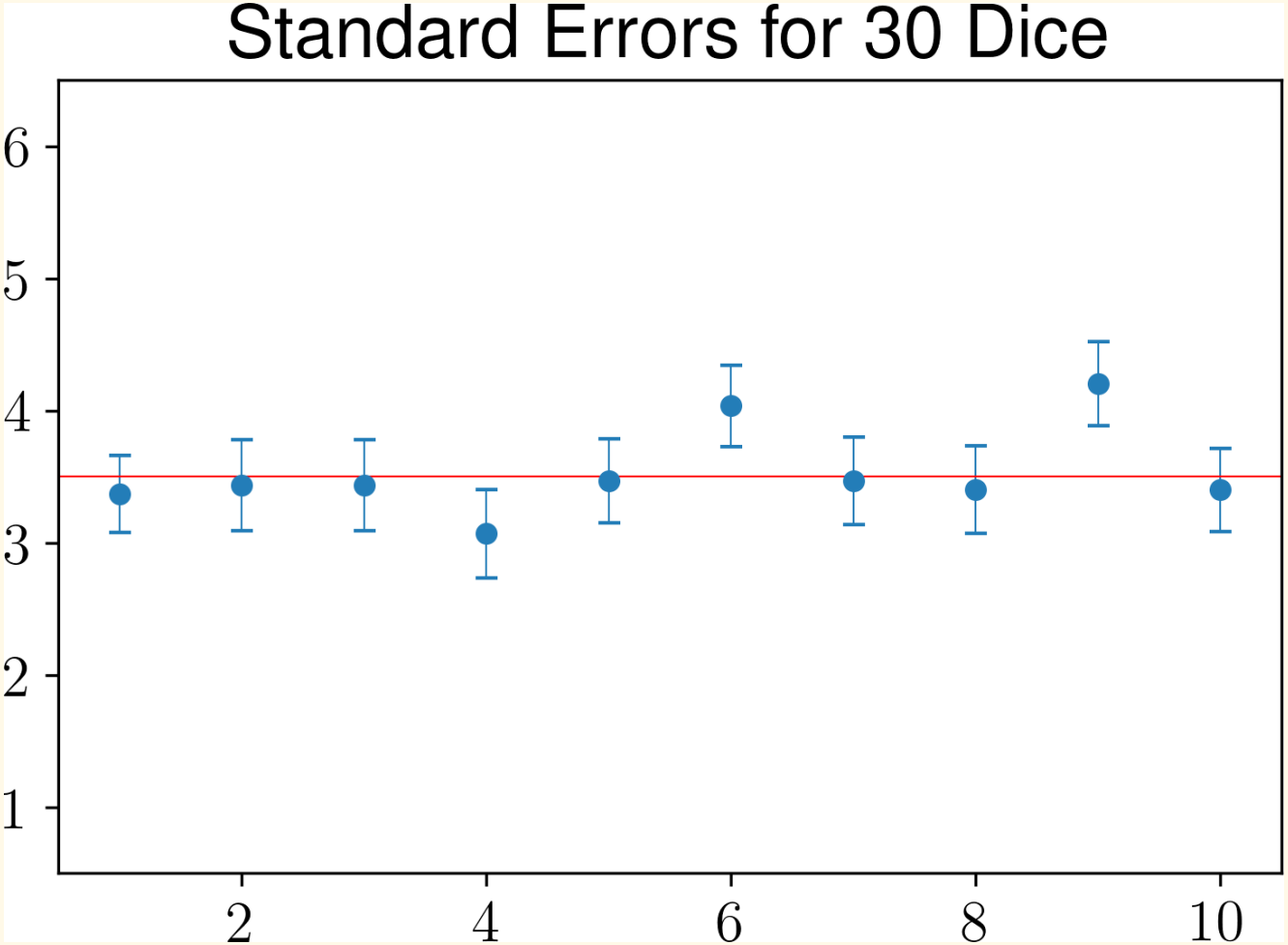
Quantifying Uncertainty

- The standard error ε_n is an estimator for the standard deviation of the distribution of μ_n.

- If the distribution of μ_n is the Gaussian distribution,
  the probability that $\mu_n - \mu_0 < \varepsilon_n$ is ~68%,
  the probability that $\mu_n - \mu_0 < 2\varepsilon_n$ is ~95%,
  the probability that $\mu_n - \mu_0 < 3\varepsilon_n$ is ~99,9%.

The summer school is funded by the DAAD with funds of the Federal Foreign Office

And this is how it looks...



The Root of Overfitting

- Central limit theorem for regression yields uncertainties of regression parameters.
- More model parameters *lead to higher parameter uncertainty*! For $g_d(t) = a_d\, t^d + \ldots + a_1\, t^1 + a_0$ the uncertainty of a_d is often larger than a_d itself!
- Therefore, the models are *not reliable for test data*.
- *Reliability can be quantified* in terms of parameter uncertainties!

The summer school is funded by the DAAD with funds of the Federal Foreign Office

Artificial Neural Networks

- One layer is a map: $L_j : \mathbb{R}^{p_j} \times \mathbb{R}^{n_{j-1}} \to \mathbb{R}^{n_j}$
  p_j : number of parameters
  n_{j-1} : number of input values (nodes of previous layer)
  n_j : number of output values (nodes of this layer)

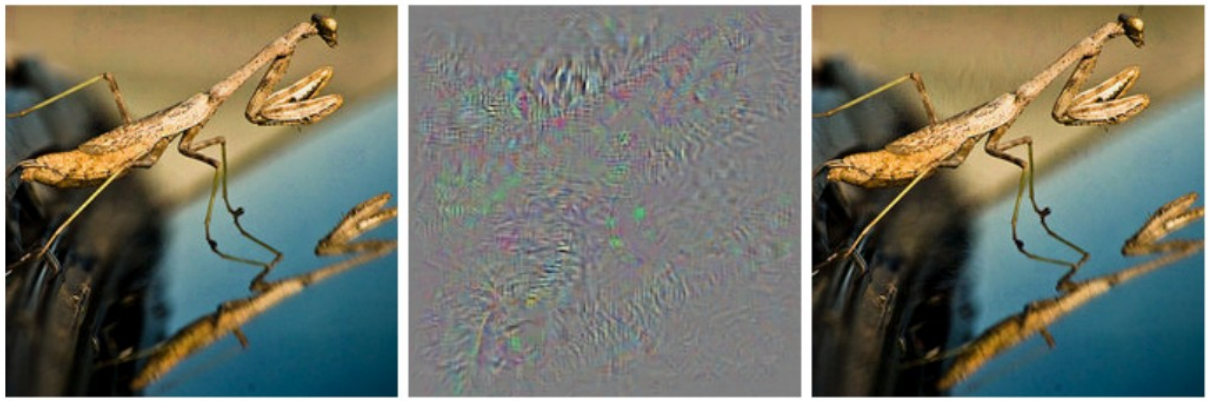
- A network of several layers is a map:
  $L_k \circ \ldots \circ L_2 \circ L_1 : \mathbb{R}^{p_k} \times \ldots \times \mathbb{R}^{p_2} \times \mathbb{R}^{p_1} \times \mathbb{R}^{n_0} \to \mathbb{R}^{n_k}$

- *Many fitted parameters*: one per network node!
  Neural networks are prone to overfitting!

The summer school is funded by the DAAD with funds of the Federal Foreign Office

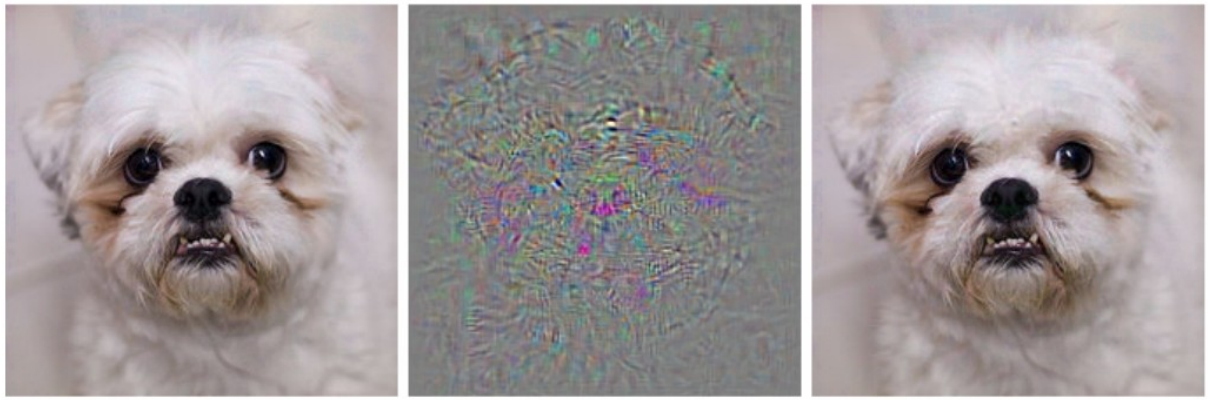
Adversarial Data

Mantis



Ostrich

Dog



Ostrich

- C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, “Intriguing properties of neural networks,” arXiv preprint arXiv:1312.6199, 2013. **Source of above image!**
- A. Nguyen, J. Yosinski, and J. Clune, “Deep neural networks are easily fooled: High confidence predictions for unrecognizable images,” arXiv preprint arXiv:1412.1897, 2014.

Adversarial Data – Definition

- Consider training data $x_1, x_2, \ldots, x_n$ in a high dimensional data space $\mathbb{R}^d$ and a neural network with p parameters which returns one of L labels: $NN : \mathbb{R}^p \times \mathbb{R}^d \to \{1, 2, \ldots, L\}$

- By training on $x_1, x_2, \ldots, x_n$ the parameter values P are determined.

- An *adversarial data point* for a given x_j of the training data is a point $y_j \in \mathbb{R}^d$ such that $NN(P, x_j) \neq NN(P, y_j)$ but the distance $|x_j - y_j|$ is very small.

The summer school is funded by the DAAD with funds of the Federal Foreign Office

Mocking the Flawed Technology

Stephan Huckemann

FOLGEN

Professor for Non-Euclidean Statistics, University of Goettingen
Bestätigte E-Mail-Adresse bei math.uni-goettingen.de - Startseite

Statistical Shape Analysis Pattern Recognition Fingerprint Recognition Biomechanics Biomedical Imaging

TITEL	ZITIERT VON	JAHR
Intrinsic shape analysis: Geodesic PCA for Riemannian manifolds modulo isometric Lie group actions S Huckemann, T Hotz, A Munk Statistica Sinica, 1-58	~~161~~	2010
Intrinsic shape analysis: Geodesic PCA for Riemannian manifolds modulo isometric Lie group actions S Huckemann, T Hotz, A Munk Statistica Sinica, 1-58	161	2010
Adversarial spheres J Gilmer, L Metz, F Faghri, SS Schoenholz, M Raghu, M Wattenberg, ... arXiv preprint arXiv:1801.02774	88	2018
Global models for the orientation field of fingerprints: an approach based on quadratic differentials S Huckemann, T Hotz, A Munk IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (9), 1507-1519	82	2008
Principal component analysis for Riemannian manifolds, with an application to triangular shape spaces S Huckemann, H Ziezold Advances in Applied Probability 38 (2), 299-319	82	2006

How does Adversarial Data work?

J. Gilmer, L. Metz, F. Faghri, S. S. Schoenholz, M. Raghu, M. Wattenberg, and I. Goodfellow. “Adversarial spheres,” arXiv preprint arXiv:1801.02774, 2018.

Consider two data sets of random vectors:

1) $x_j \in \mathbb{R}^{d+1}, \quad |x_j| = 1.0$ uniformly distributed

2) $y_j \in \mathbb{R}^{d+1}, \quad |y_j| = 1.3$ uniformly distributed

For d >> 100 and training data sets of 1,000,000 points, even though *99% of the sphere surfaces* are correctly classified, the *average distance* of a wrongly classified *adversarial point on either sphere* from any training data point can be smaller than 0.1.

The maximal average distance d depending on error surface μ can be bounded using the central limit theorem.

The summer school is funded by the DAAD with funds of the Federal Foreign Office

The Curse of Dimensionality

Theorem

Consider a random vector **X** = $(X_1, X_2, \ldots, X_{d+1})$ of length one, uniformly distributed on the Sphere S^d. Then

$$\text{for any } \epsilon > 0 : \lim_{d \to \infty} P\left(|X_1| < \epsilon\right) = 1$$

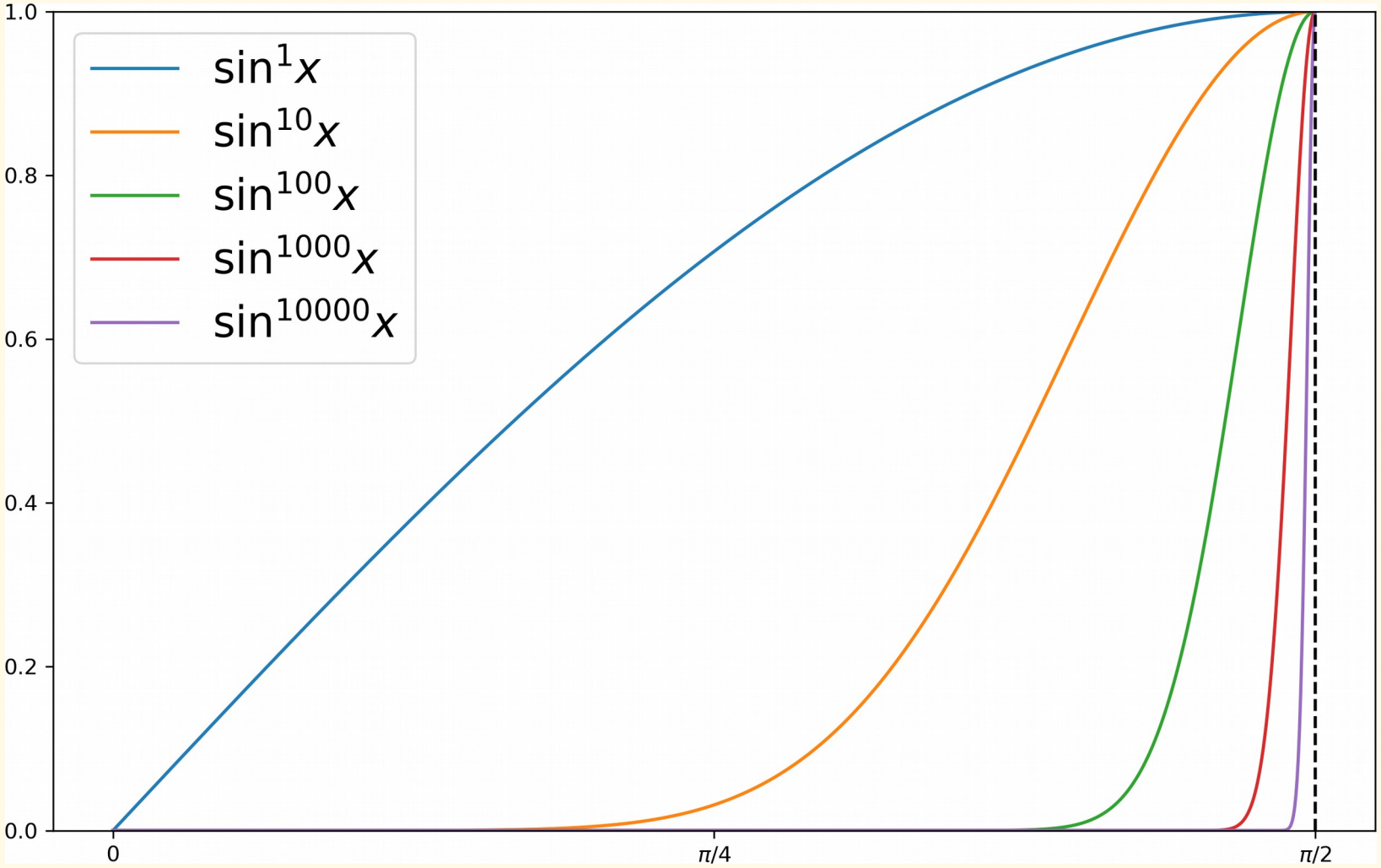
Proof

For $\alpha := \arccos(\epsilon)$ and $f_d(a,b) := \int_a^b \sin^{d-2} x \, dx$ one can write

$$P\left(|X_1| < \epsilon\right) = \frac{f_d(\alpha, \pi/2)}{f_d(0, \pi/2)} = 1 - \frac{f_d(0, \alpha)}{f_d(0, \pi/2)}$$

and note that $\lim_{d \to \infty} f_d(0, a) = 0$. This proves the theorem.

Illustration of the Curse



The summer school is funded by the DAAD
with funds of the Federal Foreign Office

Summary

- Neural Networks have two problems:

  1) Overfitting (many parameters)

  2) Curse of Dimensionality (high data dimension)

- Statistical analysis can help identify problems.

- Solution approach so far: **MOAR DATA!**

- "Neural Networks are terrible at *ignoring useless information*." *Sparsity* could help.

E. R. Balda, A. Behboodi, N. Koep and R. Mathar. "Adversarial Risk Bounds for Neural Networks through Sparsity based Compression," arXiv preprint arXiv:1906.00698, 2019.

And now to something completely different…

The summer school is funded by the DAAD
with funds of the Federal Foreign Office

Revisiting the Mean

The distribution and sample means

$$\mu_0 = \int_{-\infty}^{\infty} x f(x)\, dx \qquad \mu_n = \frac{1}{n}\sum_{k=1}^{n} x_k = \frac{1}{n}\left(x_1 + x_2 + \ldots + x_n\right)$$

cannot be generalized to other spaces. So consider

$$\mu_0 = \operatorname*{argmin}_{\mu \in \mathbb{R}} \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx$$

$$\mu_n = \operatorname*{argmin}_{\mu \in \mathbb{R}} \frac{1}{n}\sum_{k=1}^{n} (x_k - \mu)^2$$

The summer school is funded by the DAAD with funds of the Federal Foreign Office

The Fréchet Mean

To see that the new definitions of the means are equivalent, calculate the minima. The new forms of the means are called population and sample *Fréchet mean*. For a metric space M with metric d:

$$\frac{d}{d\mu}\left(\frac{1}{n}\sum_{k=1}^{n}\left(x_k-\mu\right)^2\right)=0$$

$$\Leftrightarrow \frac{d}{d\mu}\left(\mu^2-2\mu\frac{1}{n}\sum_{k=1}^{n}x_k+\frac{1}{n}\sum_{k=1}^{n}x_k^2\right)=0$$

$$\Leftrightarrow \left(2\mu-\frac{2}{n}\sum_{k=1}^{n}x_k\right)=0$$

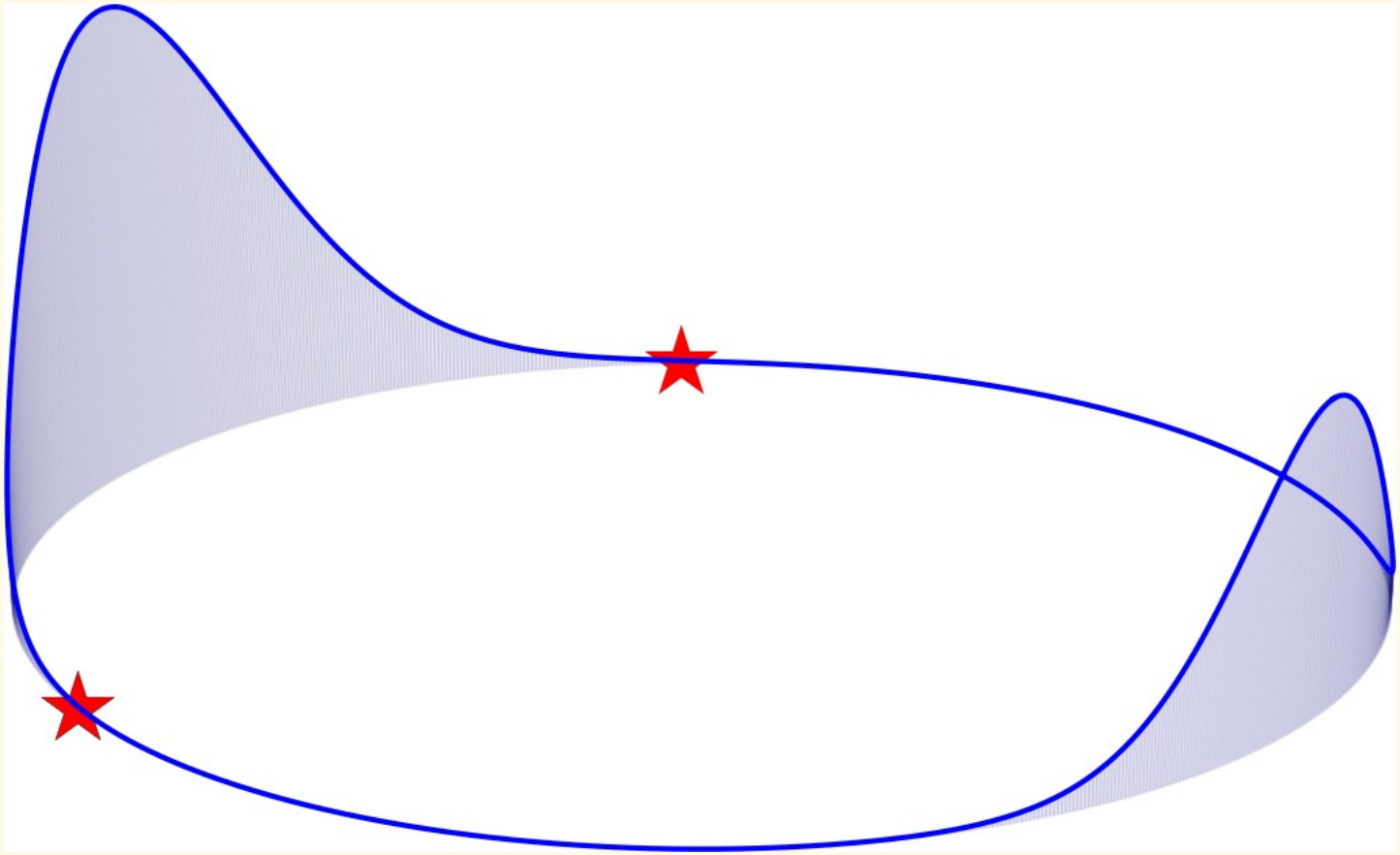
$$\Leftrightarrow \mu=\frac{1}{n}\sum_{k=1}^{n}x_k$$

$$\mu_0=\underset{\mu\in M}{\operatorname{argmin}}\int_M d\left(x,\mu\right)^2 f\left(x\right)dx$$

$$\mu_n=\underset{\mu\in M}{\operatorname{argmin}}\frac{1}{n}\sum_{k=1}^{n}d\left(x_k,\mu\right)^2$$

The Mean on Curved Spaces

On curved spaces, the mean can be ambiguous:



The Central Limit Theorem (CLT)

The Central Limit Theorem can be generalized to the Fréchet mean if

1) μ_0 is unique.

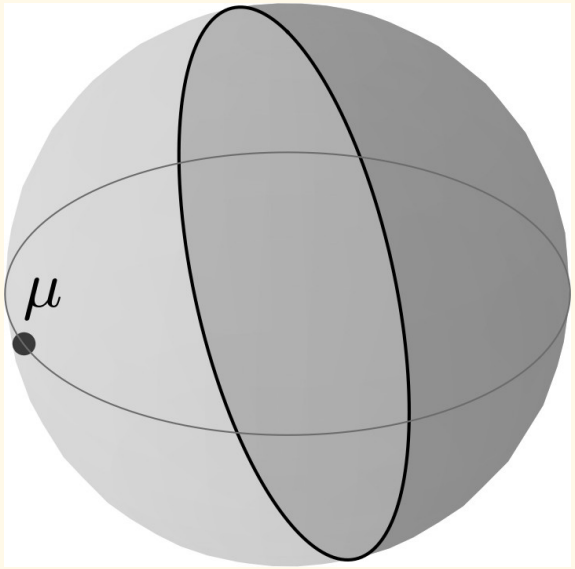
2) The population Fréchet function $F_0(\mu) := \int_M d(x,\mu)^2 f(x)\,dx$ has a second order Taylor expansion

$$F_0(\mu) = F_0(\mu_0) + \frac{d^2 F_0}{d\mu^2}(\mu_0)\,\mu^2 + O(\mu^3)$$

(This can be made precise if M is a *manifold*.)

What if…

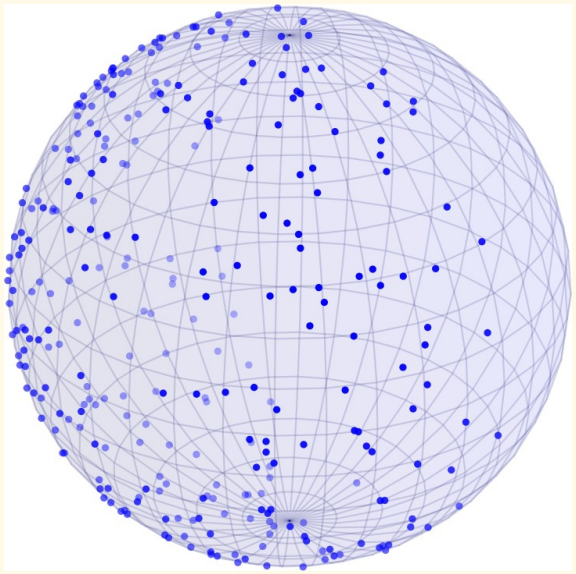
- If the mean is not unique, the Central Limit Theorem can be adjusted.
- If the Taylor expansion does not work, things are more complicated…



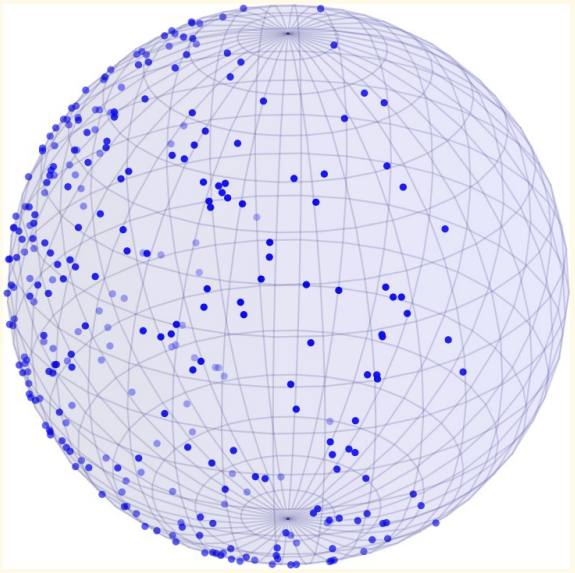
A point mass on the north pole and a uniform distribution on the southern hemisphere: Quite troublesome…

The summer school is funded by the DAAD with funds of the Federal Foreign Office

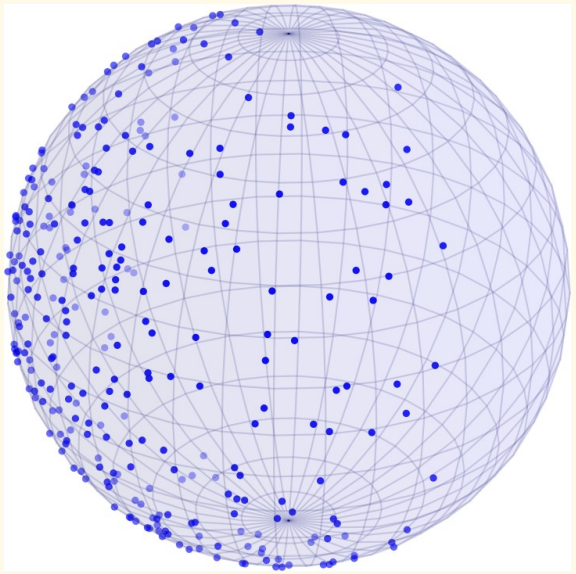
Means for n = 100



Small point mass



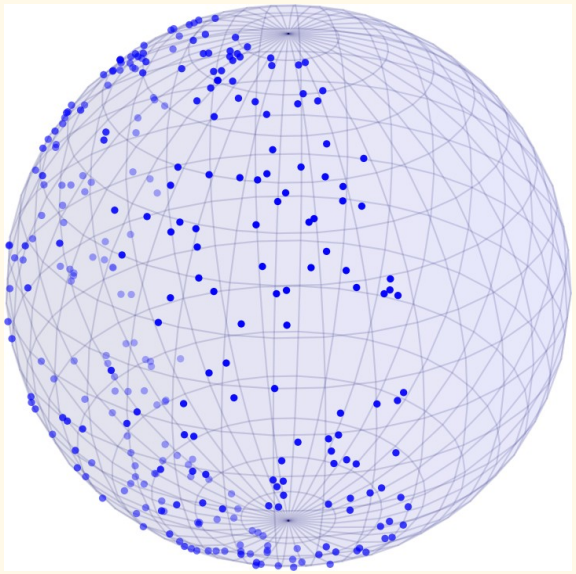
Medium point mass



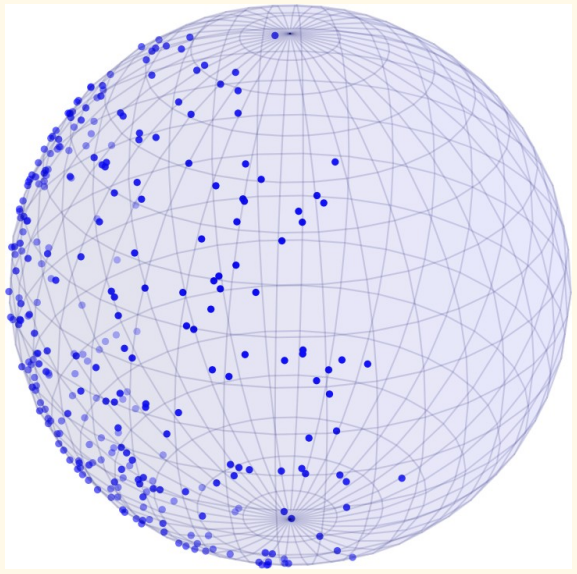
Large point mass

The summer school is funded by the DAAD with funds of the Federal Foreign Office

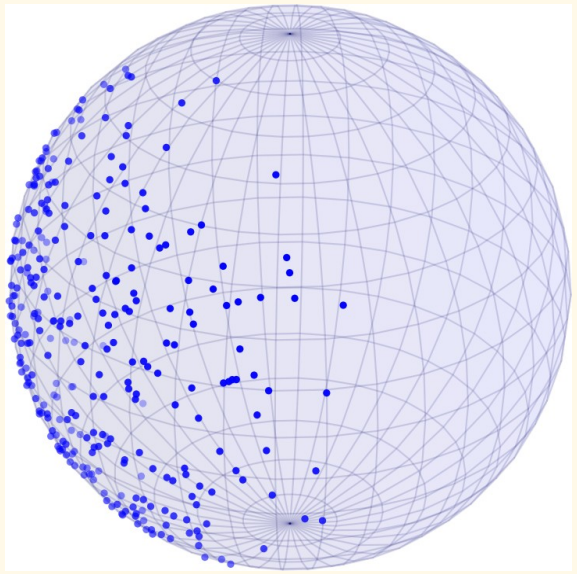
Means for n = 1,000



Small point mass



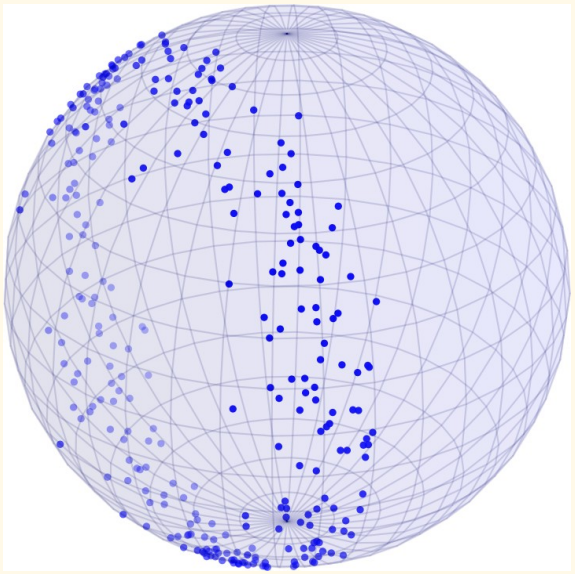
Medium point mass



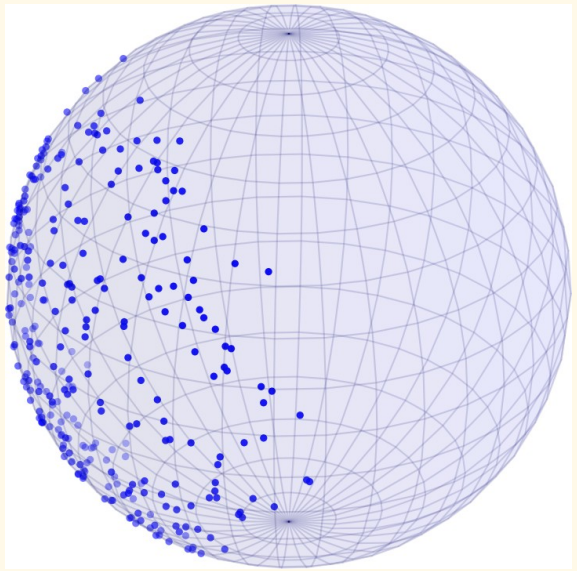
Large point mass

The summer school is funded by the DAAD with funds of the Federal Foreign Office

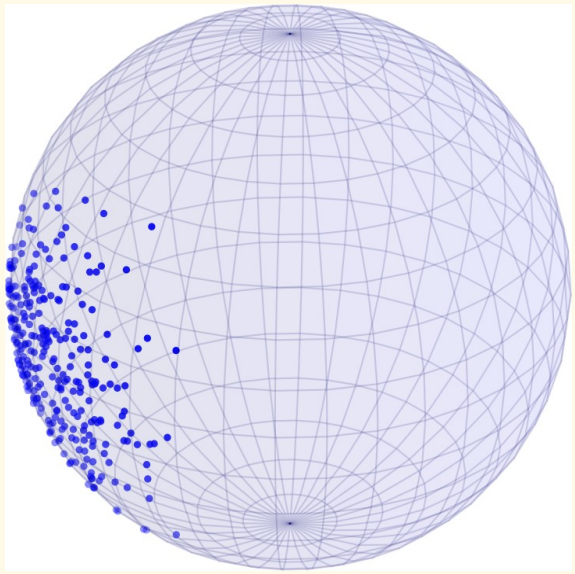
Means for n = 10,000



Small point mass



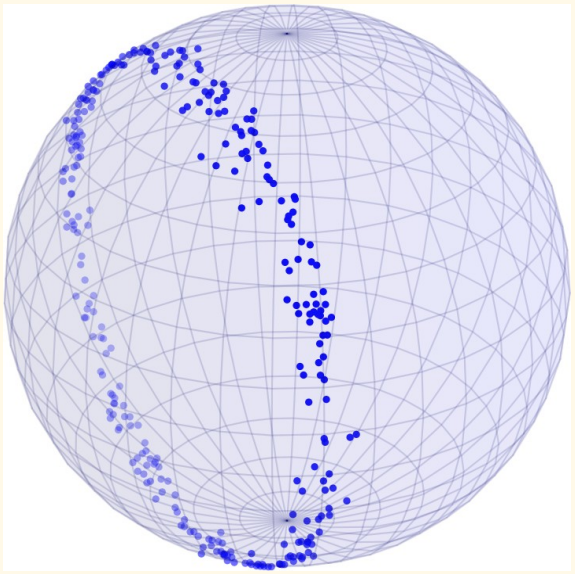
Medium point mass



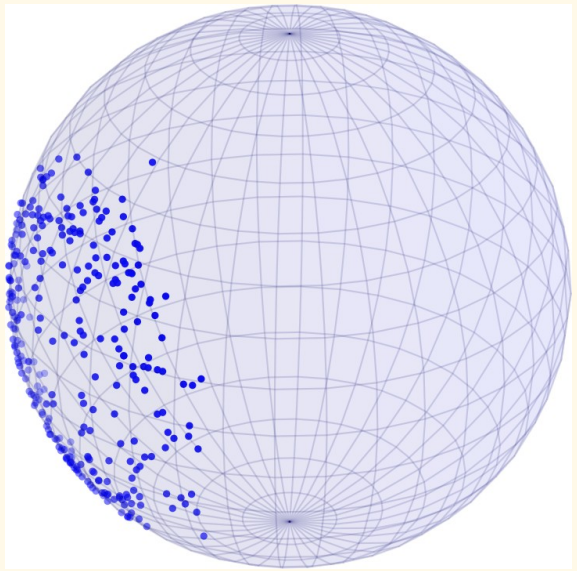
Large point mass

The summer school is funded by the DAAD with funds of the Federal Foreign Office

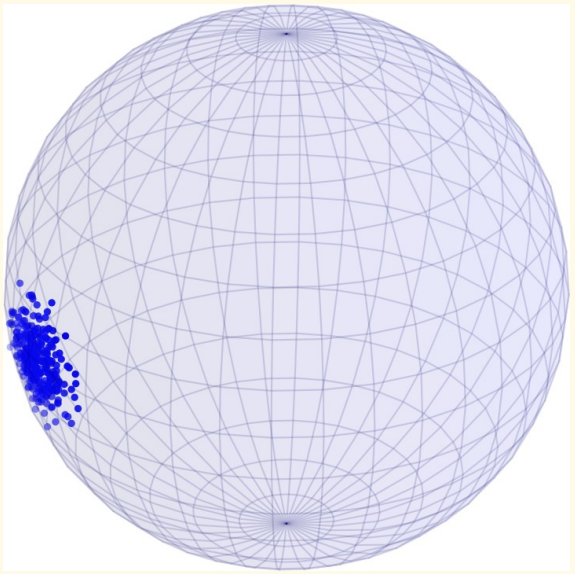
Means for n = 100,000



Small point mass



Medium point mass



Large point mass

The summer school is funded by the DAAD with funds of the Federal Foreign Office

Observations

- For *small point mass*, the *mean is not unique*, but on a circle around the north pole.
- For *large point mass*, the *mean is unique* on the north pole.
- For *medium point mass*, the mean is unique on the north pole, but *convergence is very slow*!

The summer school is funded by the DAAD with funds of the Federal Foreign Office

A Generalized CLT

Assume

1) μ_0 is unique.

2) The population Fréchet function has an expansion

$$F_0(x) - F_0(\mu_0) = \sum_{j=1}^{d} a_j \left| x_j \right|^r + O\left(x^{r+1}\right)$$

Then $\sqrt{n}\left(\mu_n - \mu_0\right)_j \left|\left(\mu_n - \mu_0\right)_j\right|^{r-2}$ approaches a Gaussian distribution centered at 0 for $n \to \infty$.

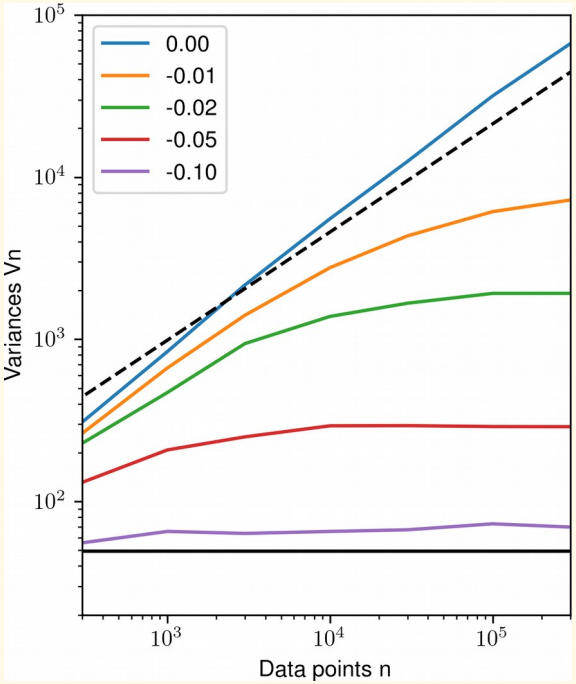
The summer school is funded by the DAAD with funds of the Federal Foreign Office

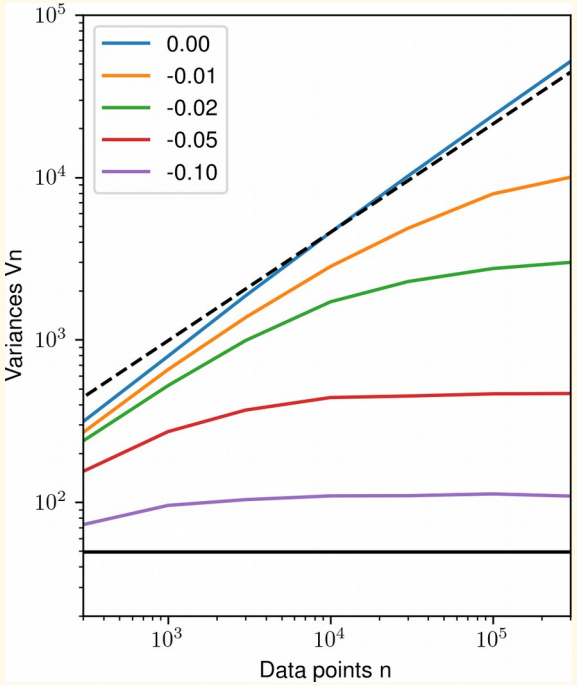
What does this mean?

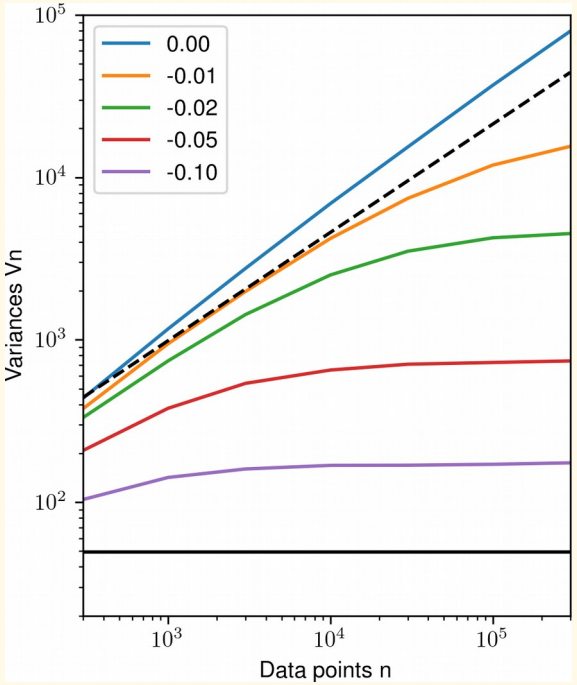
- If the Fréchet function is not a quadratic function at its minimum μ_0, the standard CLT does not hold.
- If the order of the Fréchet function is r, $\left(\mu_n - \mu_0\right)^{r-1}$ converges to 0 as $\frac{1}{\sqrt{n}} = n^{-1/2}$.
- This means that $\mu_n - \mu_0$ converges to 0 as $n^{-1/(2r-2)}$.
- *Larger data sets do not help so much!*

The summer school is funded by the DAAD with funds of the Federal Foreign Office

Only One Pathological Example?





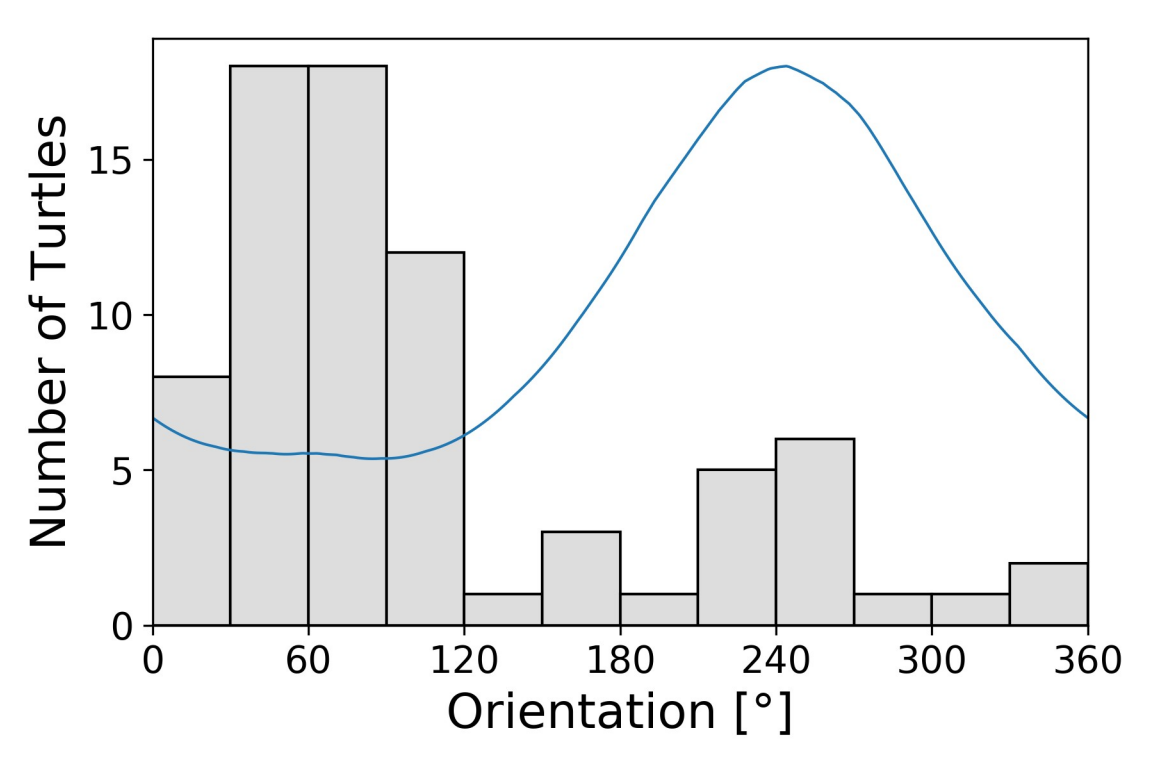


d=2 d=10 d=100

Distributions with larger point mass also *need very large data sets* for normal convergence! *Worse for higher dimension*!

The summer school is funded by the DAAD with funds of the Federal Foreign Office

A Real World™ Example



Movement direction of female turtles after laying eggs on the beach.

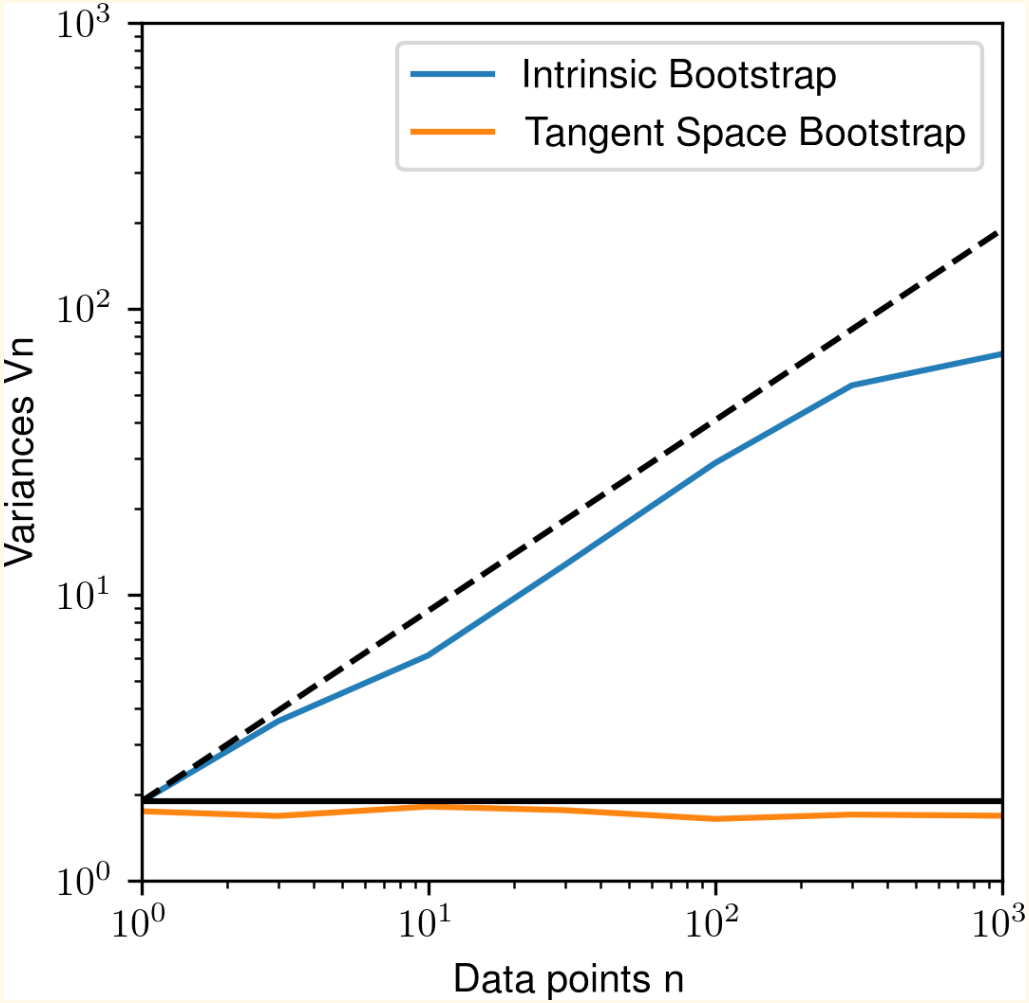
The summer school is funded by the DAAD
with funds of the Federal Foreign Office

How to Check Convergence Rate?

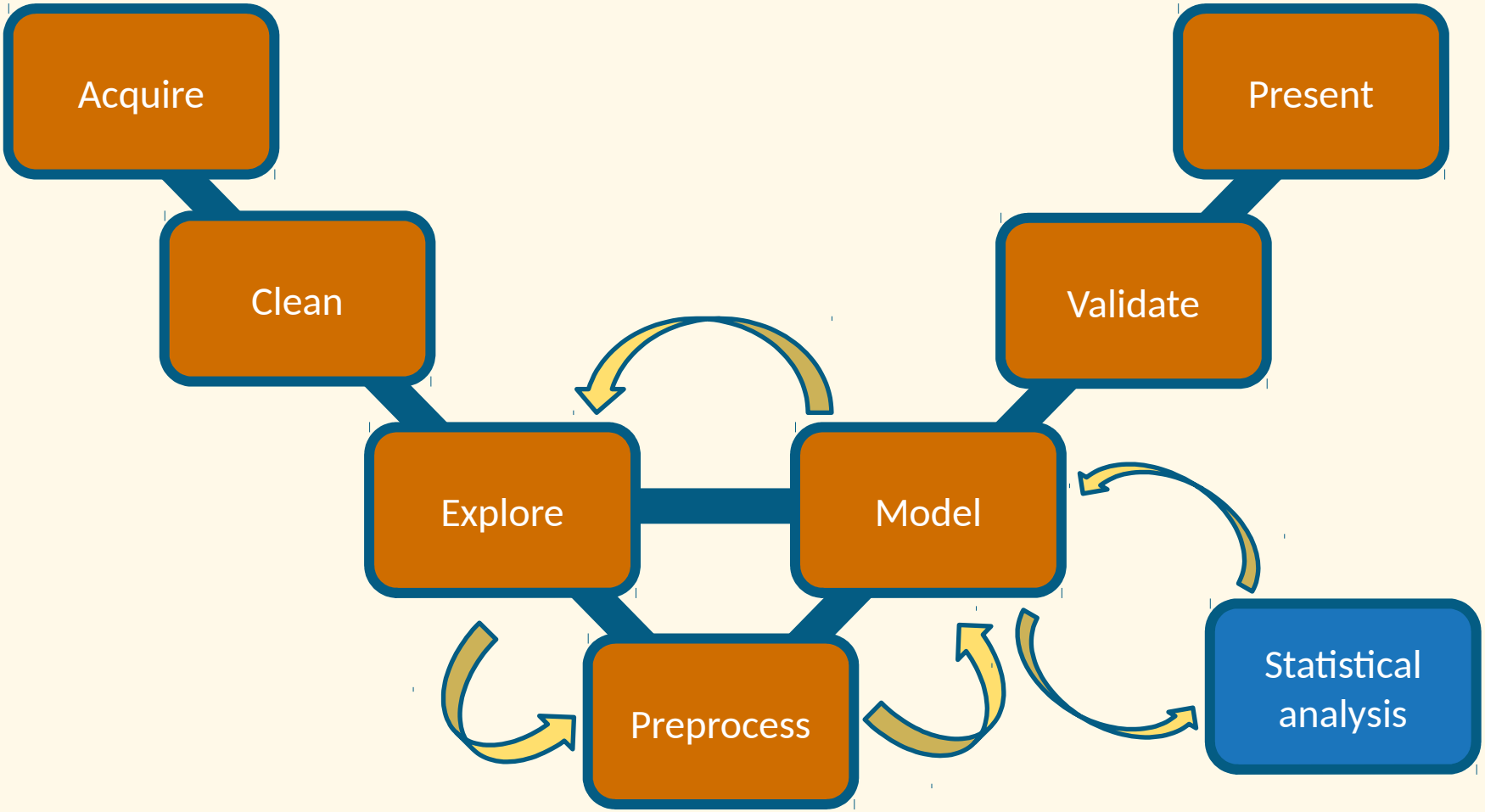
- Consider a data set $x_1, x_2, \ldots, x_n$.
- Draw $y_1, y_2, \ldots, y_m$ with replacement from the data and calculate the mean $\mu_m^* = \operatorname{argmin}_{\mu \in M} \frac{1}{m} \sum_{k=1}^{m} d(y_k, \mu)^2$.
- Do this B times to get $\mu_m^{*1}, \mu_m^{*2}, \ldots, \mu_m^{*B}$. This is called *m-out-of-n bootstrap with B replicates*.
- Then calculate $V_m^* = \frac{1}{B}\left(d(\mu_m^{*1}, \mu_n)^2 + d(\mu_m^{*2}, \mu_n)^2 + \ldots + d(\mu_m^{*B}, \mu_n)^2\right)$.

The summer school is funded by the DAAD with funds of the Federal Foreign Office

Slow Rate of Convergence



How to avoid ruining the V?



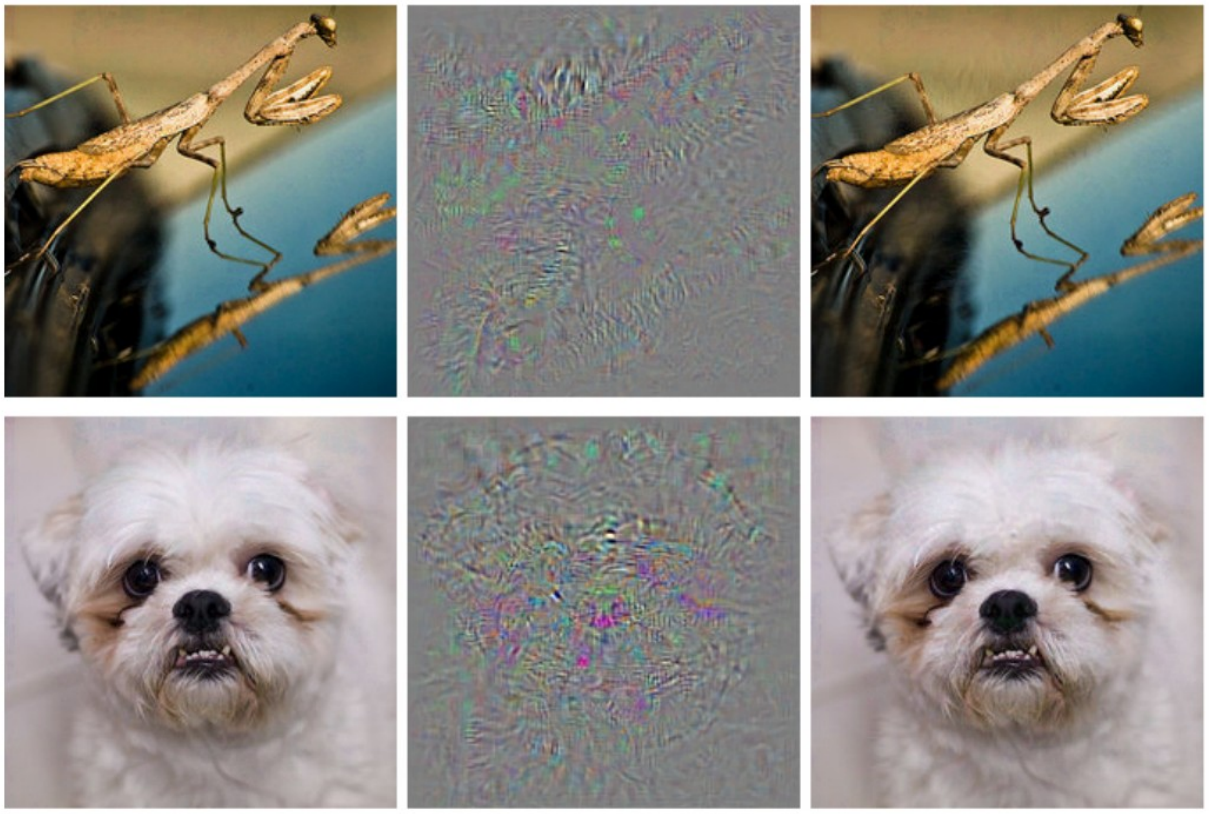
Here's the Upshot

- Terrible things lurk in the shadows: *Overfitting*, *Curse of Dimensionality*, *slow Asymptotics*.
- Statistics knows many of these abominations.
- Statistics can *identify* and *quantify problems*.
- Statistics has learned some *coping strategies*.
- Statistics has the tools to *interpret wreckages*.

The summer school is funded by the DAAD with funds of the Federal Foreign Office

Neural Network without Statistics

Mantis



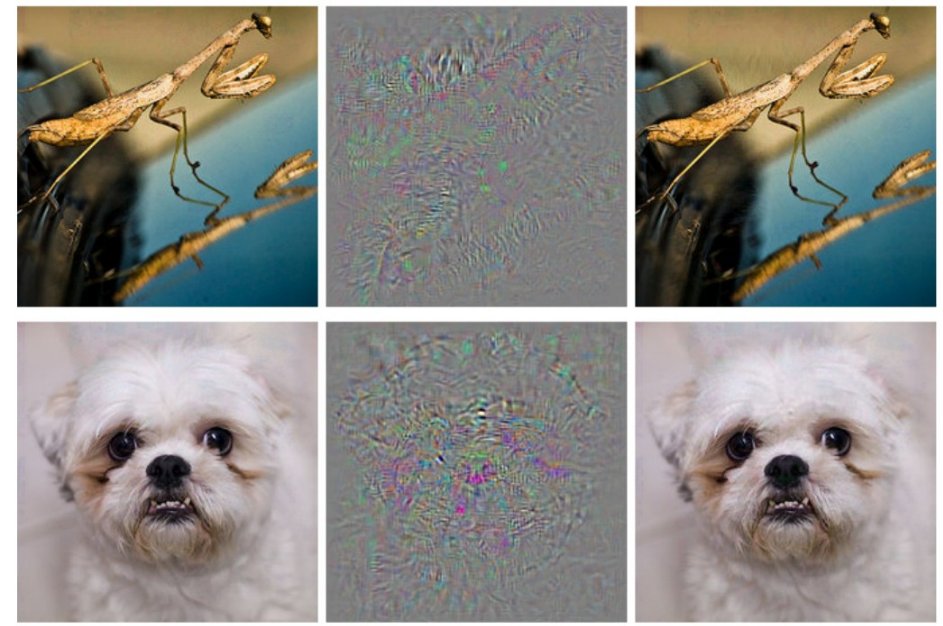
Ostrich

Dog

Ostrich

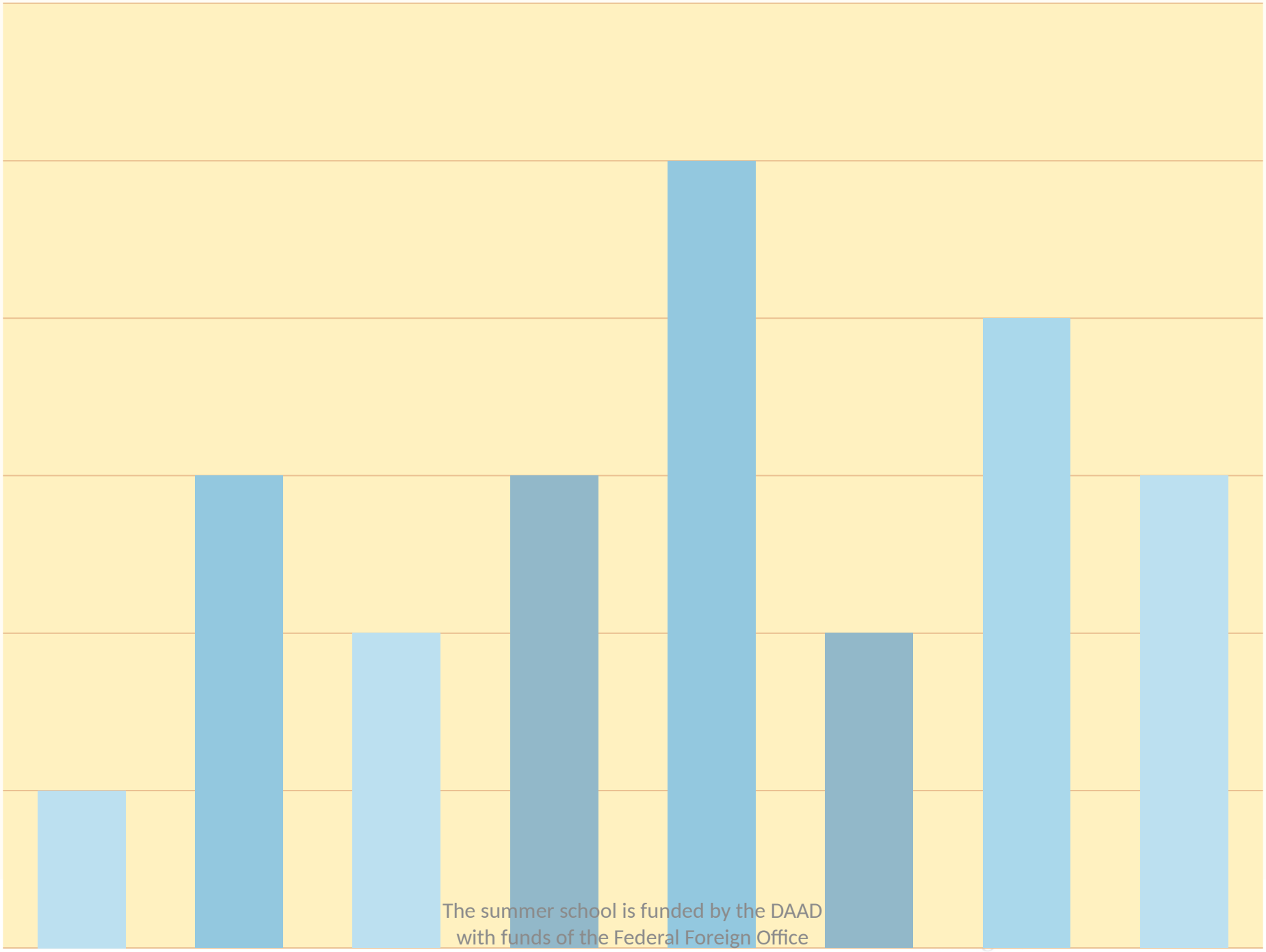
Image: C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, “Intriguing properties of neural networks,” arXiv preprint arXiv:1312.6199, 2013.

Neural Network on Statistics



"With the high dimension and the overfitting, I suppose it could be a mantis, a dog, an ostrich or a submarine. That's assuming that the CLT holds for these data."

Image: C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," arXiv preprint arXiv:1312.6199, 2013.



The summer school is funded by the DAAD
with funds of the Federal Foreign Office