

Data science and machine learning - a short stroll

Prasanna Bhogale

p.bhogale@kigroup.de

19 May 2017

Disclaimer

The terms *data science*, *machine learning* and *artificial intelligence* are very broad and used in many different and overlapping contexts.

I will attempt to give my own narrow perspective and very likely the next "data scientist" you speak to will disagree with me about everything.

I apologize in advance for any buzzwords you might encounter.

In this talk

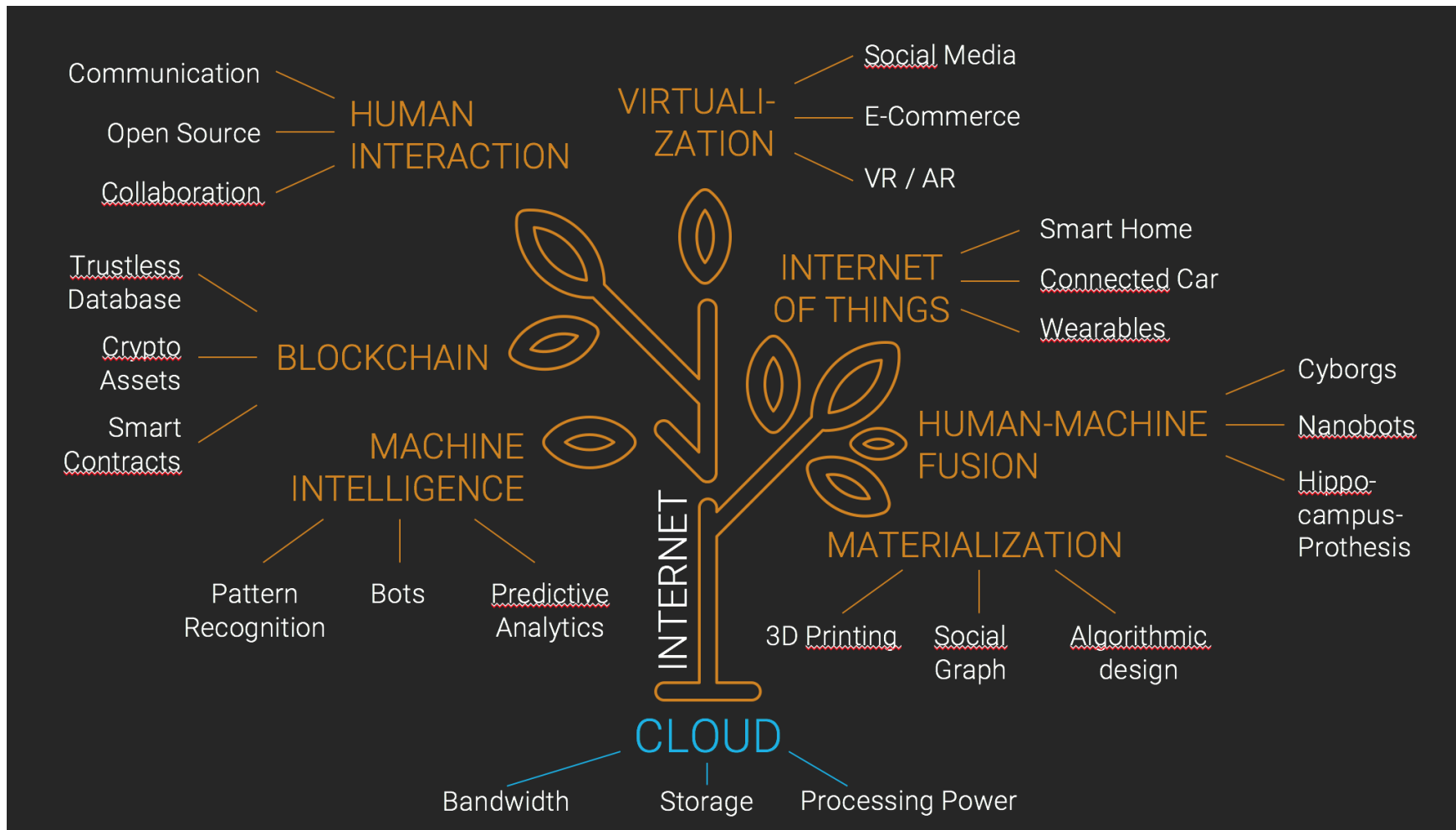
- data science
- machine learning
- lots of links to useful stuff

Why is all this data science happening now ?



-Mark Andersteen

- Moore's law : exponentially growing computing, storage
- Digitization : intensive measurement, storage of ambient information



via [KI-Analytics](#)

More data is produced now than humans can analyse - 2.5
Exabytes/day

Large data sets + clever algorithms \Rightarrow

- replace intuition with statistics
- replace skilled humans with machines
- replace biased humans with rational machines

and other such utopian/dystopian dreams..

Realistically :

- Generate insight
- Improve ability to predict

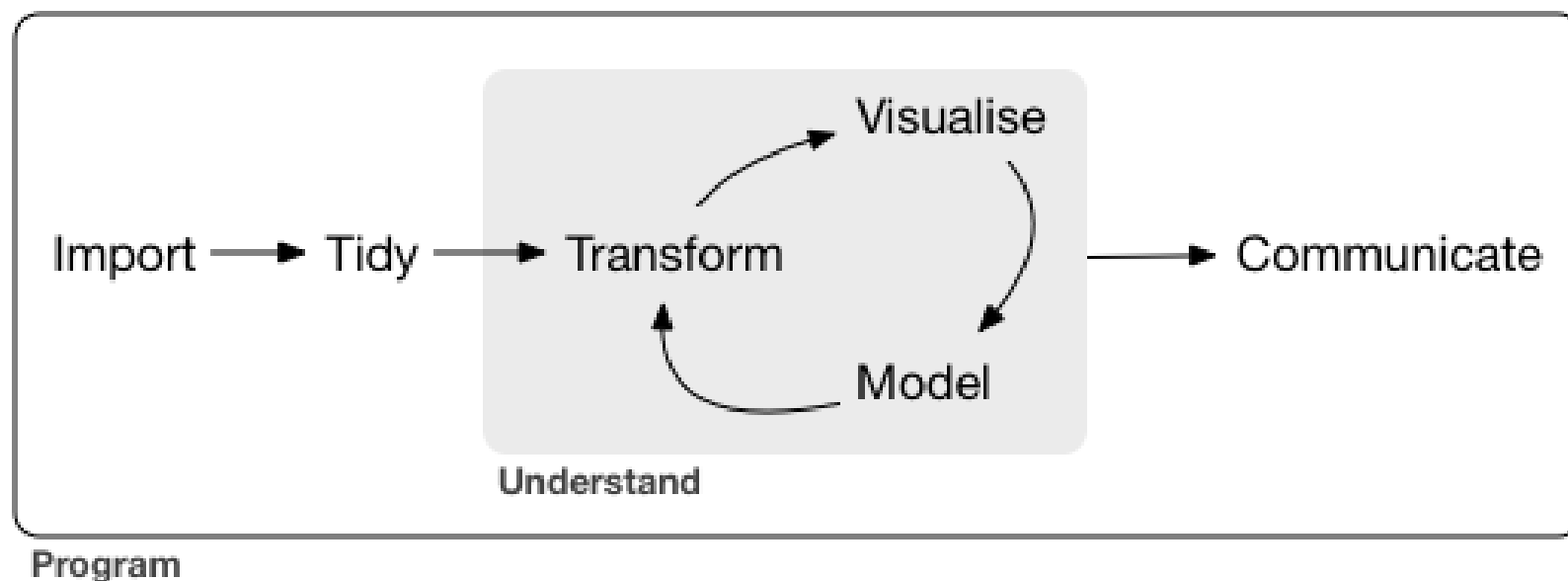
Questions data scientists try to answer

from [KDnuggets](#)

- Is this A or B (or C or D or ...) ? `classification`
- Is this weird ? `anomaly detection`
- How much - or - How many ? `regression`
- How is this organized ? `clustering`
- What should I do next ? `reinforcement learning`

Data scientist is, as data scientist does

and what a data scientist does is programming, programming, programming.



from Wickham's [R for data science](#)

Tools of the trade (minimal)

- Linear algebra, statistics, probability
- R ecosystem
- Python ecosystem
- Efficient visualization – grammar of graphics
- Common machine learning algorithms and libraries like XGboost for gradient boosting

and a MUCH longer list is [here](#)

strongly dependent on context. Most likely, MS Excel, databases, cloud platforms will all be part of the mix.

and now onward to machine learning

First principles - types of reasoning

- Deductive : reasoning from set of premises to reach logically certain conclusion. Eg. Mathematical proof
- Inductive : Data supported probabilistic reasoning. Eg. machine learning
- Abductive : Finding a model which best fits available data. Eg. bayesian inference

Science :

Observations \rightarrow Theory \rightarrow Predictions

Machine learning

Training error = $f(\text{Model}_P, \text{Training set})$

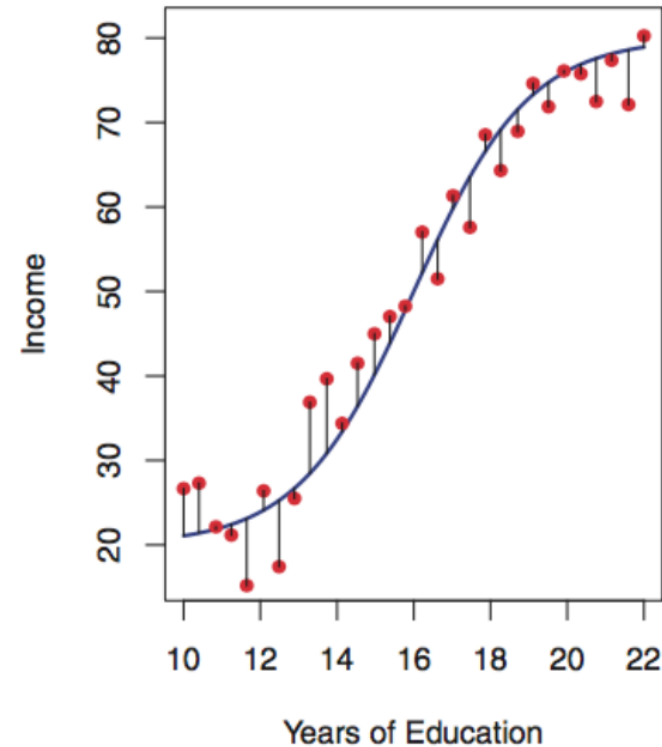
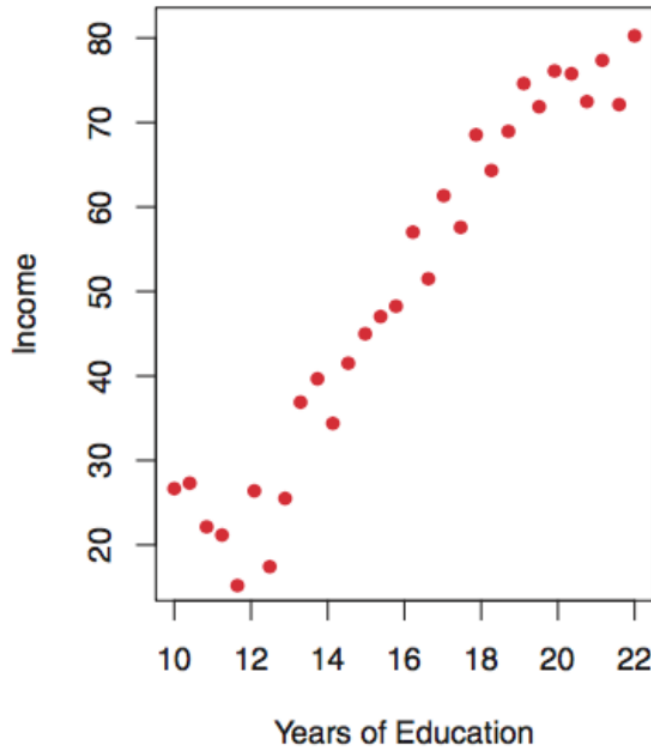
$$\hat{P} = \min_P (f)$$

Test predictions = $\text{Model}_{\hat{P}}(\text{Test set})$

- Test and Training sets should be disjoint.
- ML models are typically trying to capture complex phenomena
- ML models typically have a very large number of parameters
- What is a complex/simple model ? [a good question..](#)

Generalization is everything !

Noise and Data



from [Introduction to Statistical Learning](#)

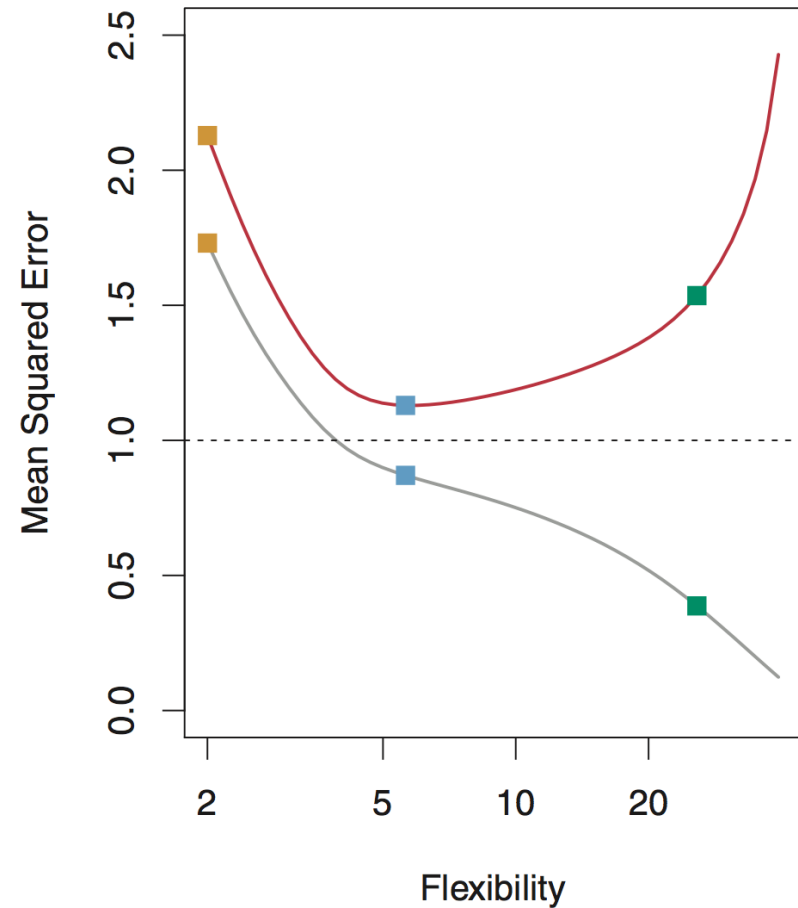
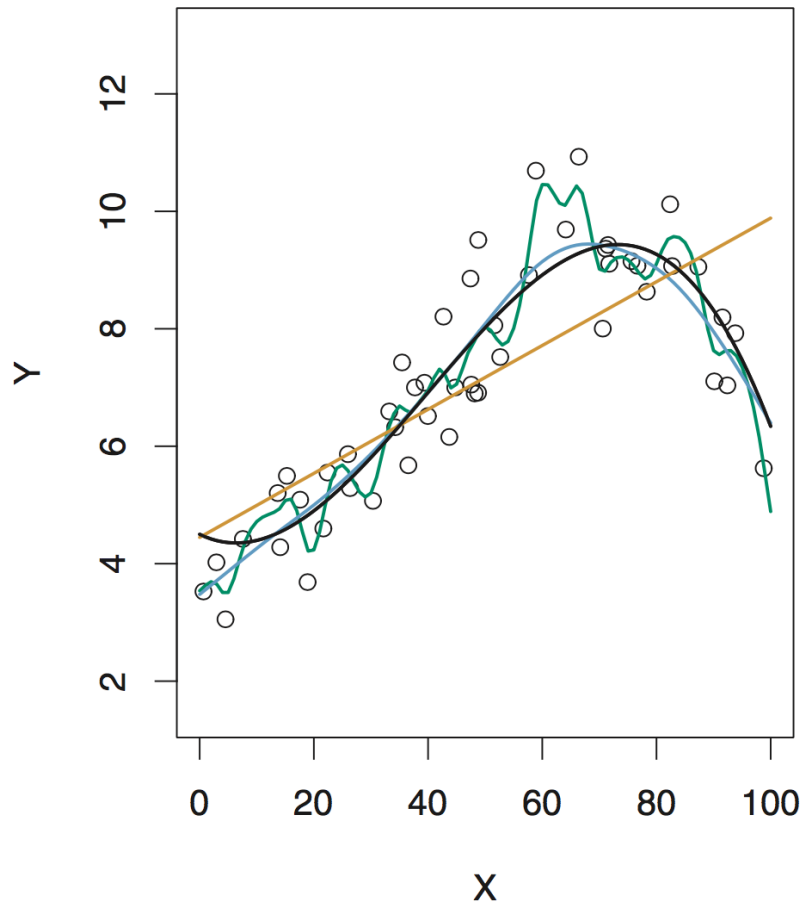
- Reducible error : can be eliminated with more data/better model
- Irreducible error : Inherent randomness and external factors

Bias-variance tradeoff

Properties of a given statistical model -

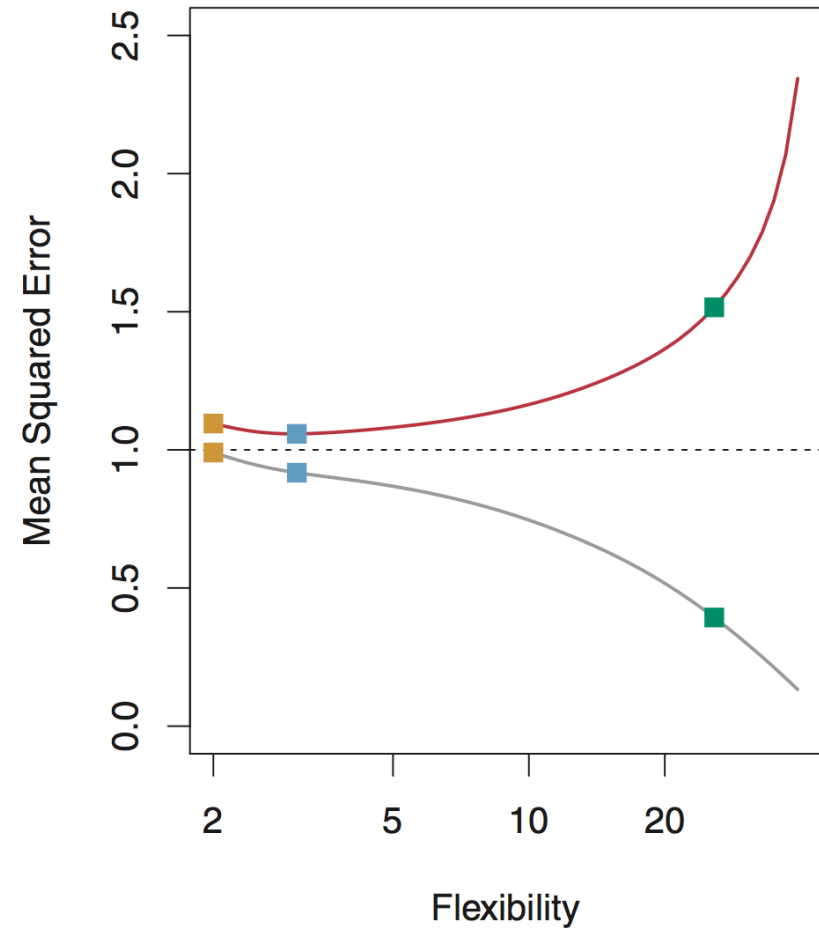
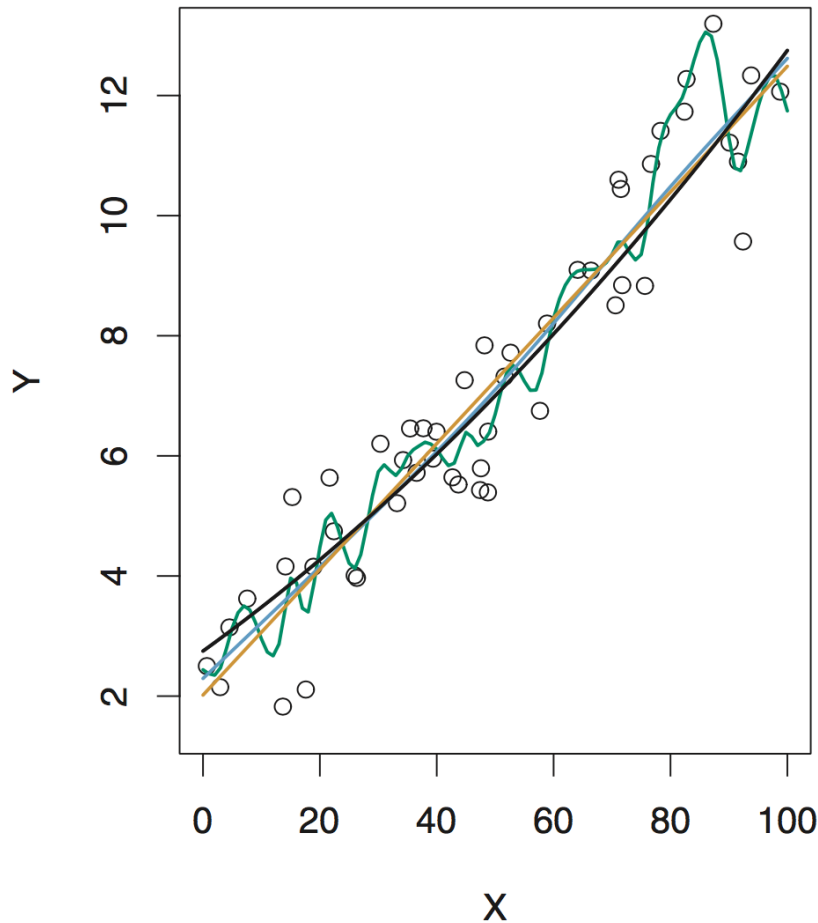
- Bias : Error resulting from approximations in the model. Cannot be improved with more data.
- Variance : Change in prediction due to change in training set.
- Complex (more flexible) models - low bias, high variance
- Simple (less flexible) models - high bias, low variance

Training error always reduces with model complexity



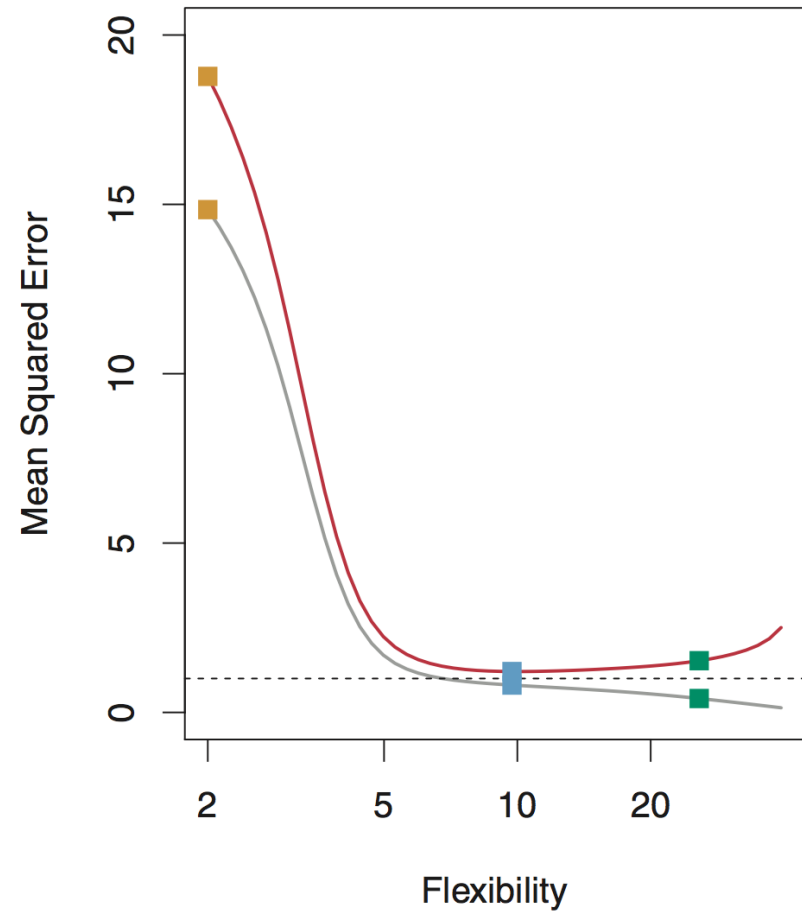
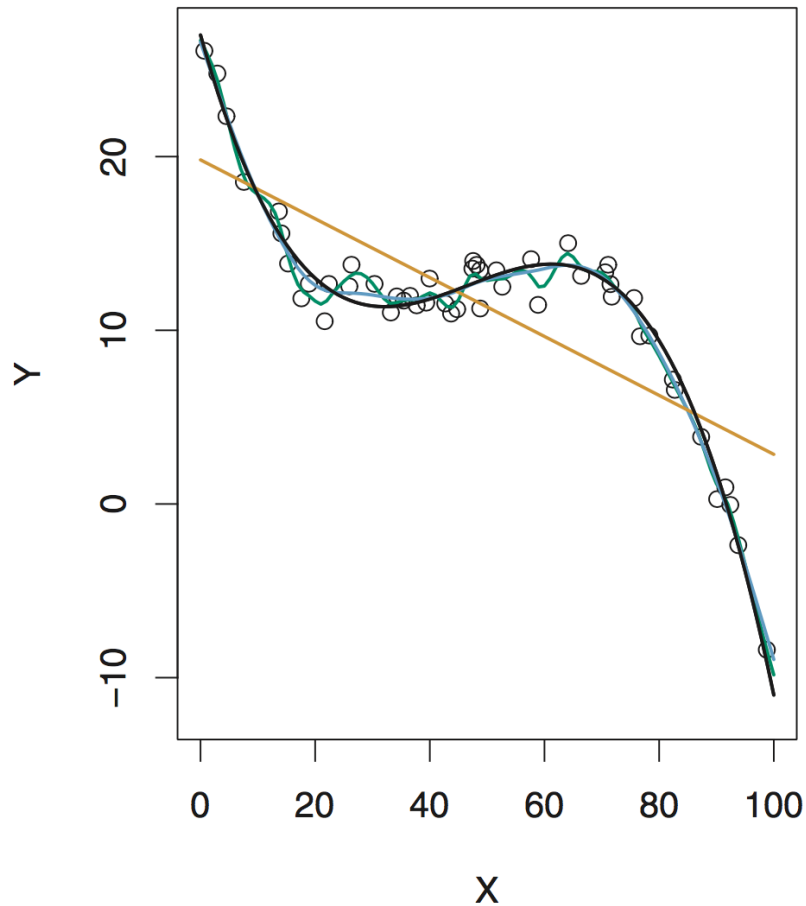
from [Introduction to Statistical Learning](#)

More complex models are not always better



from [Introduction to Statistical Learning](#)

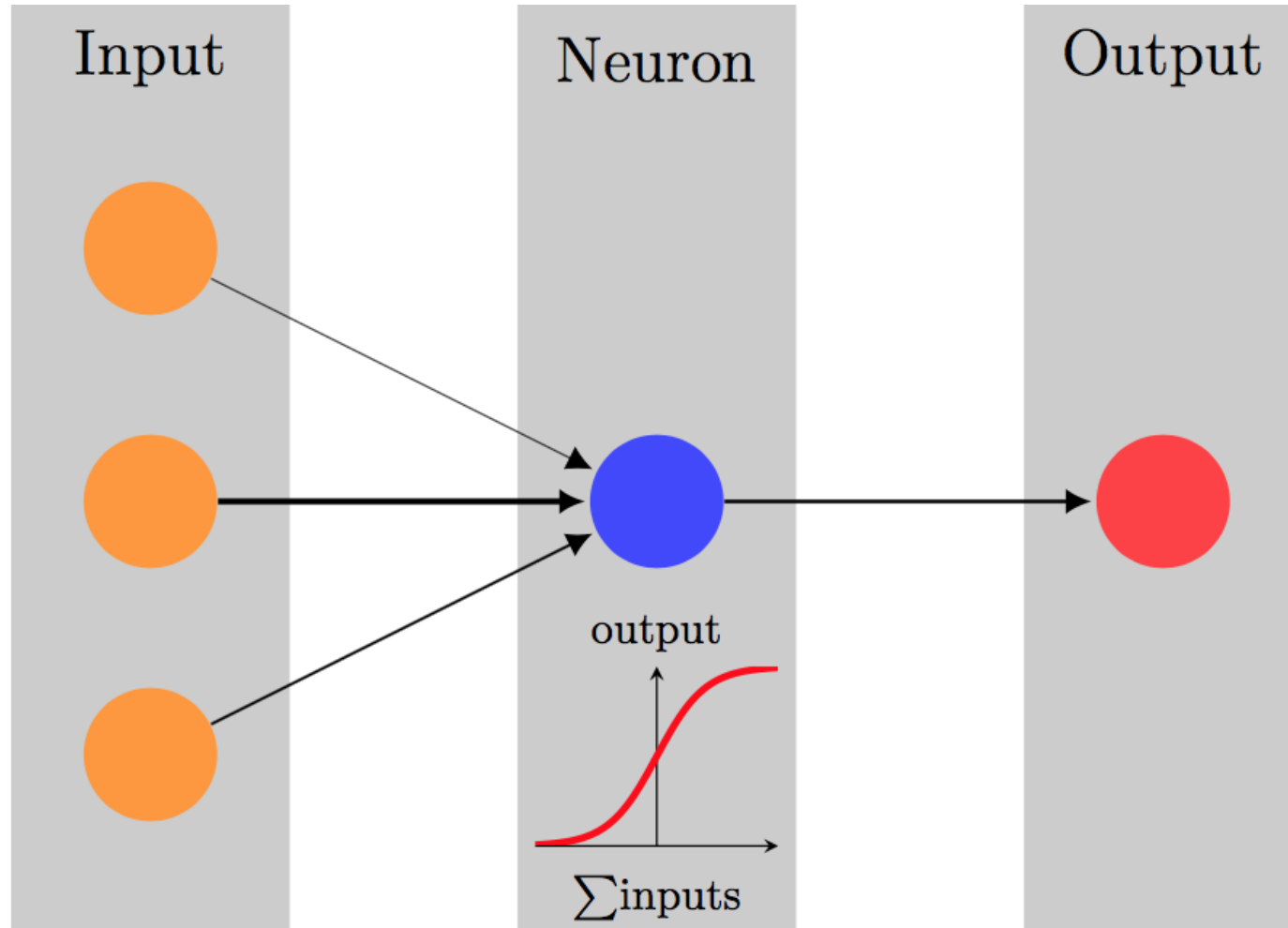
The 'right' model : lowest test error



from [Introduction to Statistical Learning](#)

and we surely must mention that most ubiquitous buzzword...

Differentiable networks



Artificial intelligence : Machine learning that gets spooky

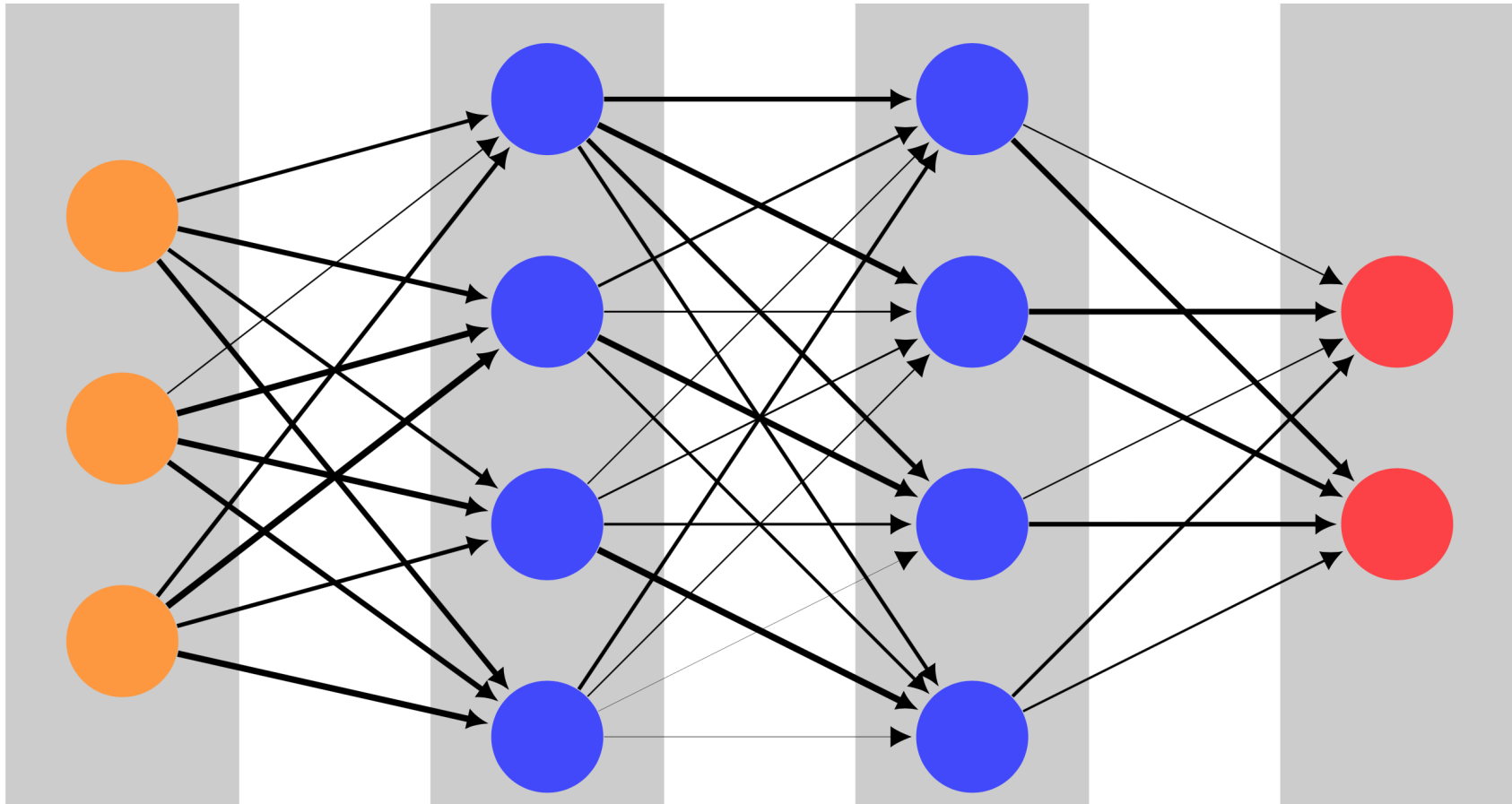
Backpropagation

Input layer

Hidden layer

Hidden layer

Output layer













A wonderful introduction [by Nielsen](#)








kaggle

- Relevant Machine Intelligence Problems
- High quality Data Science Forum
- Data sets, example scripts, kernels

Dashboard ▾ Public Leaderboard - Outbrain Click Prediction

This leaderboard is calculated on approximately 30% of the test data. The final results will be based on the other 70%, so the final standings may be different.

#	Δ1w	Team Name	Score	Entries	Last Submission UTC (Best - Last Submission)
1	—	Three Data Points 	0.69655	102	Mon, 02 Jan 2017 02:39:29 (-2.2d)
2		Neuron 	0.69632	9	Mon, 02 Jan 2017 03:46:46 (-5.1d)
3		CV 	0.69406	20	Fri, 23 Dec 2016 01:34:55 (-41.4h)
4		xiaohaoxx	0.69322	22	Fri, 23 Dec 2016 01:41:41 (-22d)
5		flightrush	0.69274	1	Fri, 23 Dec 2016 01:56:31
6		rokh	0.69264	24	Wed, 21 Dec 2016 19:48:29 (-23.1h)
7		Sangxia	0.69249	3	Mon, 02 Jan 2017 08:59:41
8		Daniel_NTU	0.69093	7	Mon, 26 Dec 2016 10:19:51

	Data Science Bowl 2017 Can you improve lung cancer detection? Featured · 14 days to go	\$1,000,000 1,749 teams
	The Nature Conservancy Fisheries Monitoring Can you detect and classify species of fish? Featured · 14 days to go	\$150,000 2,204 teams
	Intel & MobileODT Cervical Cancer Screening Which cancer treatment will be most effective? Featured · 3 months to go	\$100,000 205 teams
	Google Cloud & YouTube-8M Video Understanding Challenge Can you produce the best video tag predictions? Featured · 2 months to go	\$100,000 364 teams
	NOAA Fisheries Steller Sea Lion Population Count How many sea lions do you see? Featured · 3 months to go	\$25,000 30 teams
	Quora Question Pairs Can you identify question pairs that have the same intent? Featured · 2 months to go	\$25,000 964 teams
	Two Sigma Connect: Rental Listing Inquiries How much interest will a new rental listing on RentHop receive? Recruitment · A month to go	Jobs 1,652 teams

terrific for learning the tools and techniques and a source of many many cool data sets.

Google is your friend, and tutorials on everything are everywhere.

We are hiring.

Prasanna Bhogale

p.bhogale@kigroup.de