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.

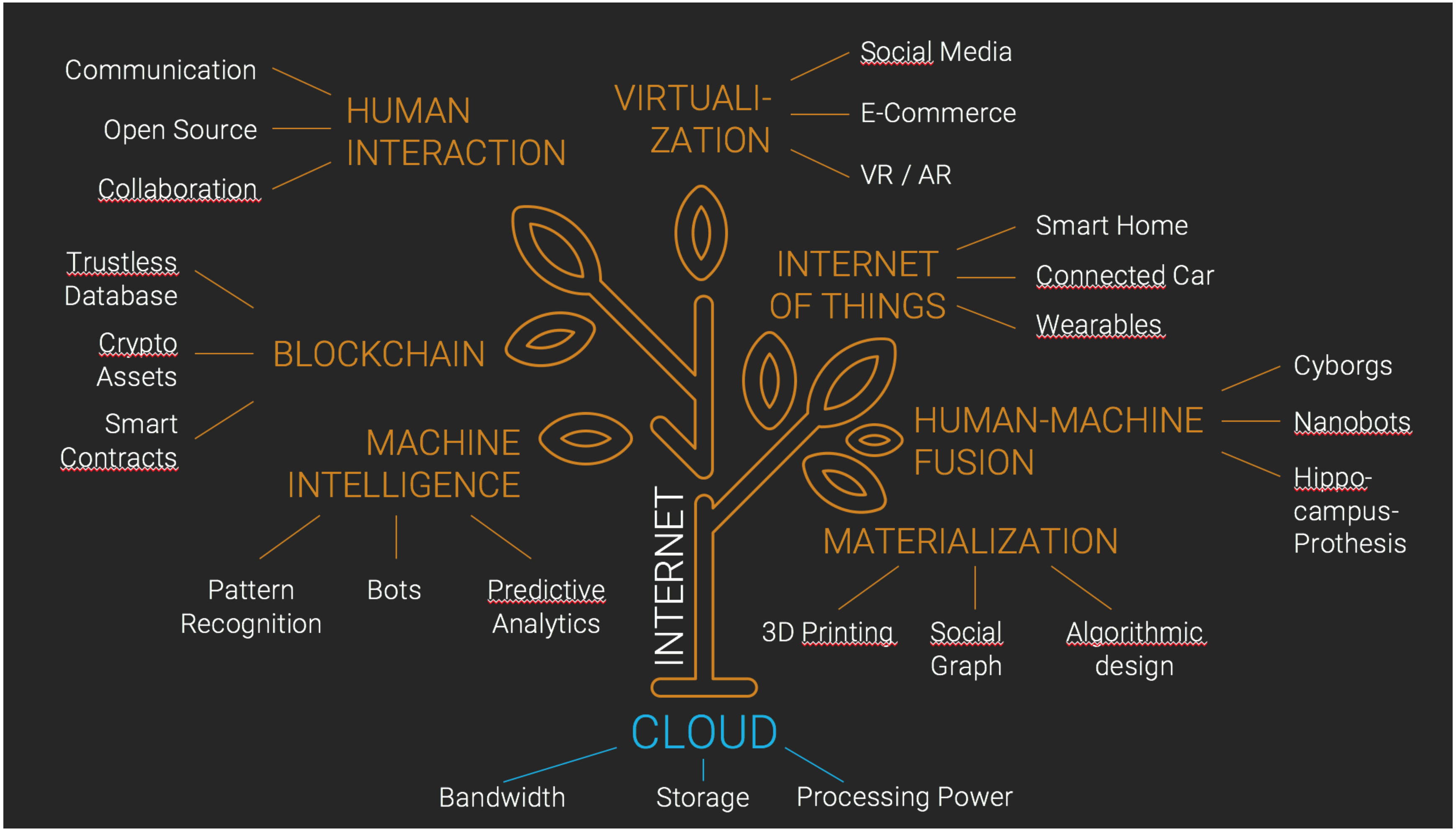
In this talk

- 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 computating, 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

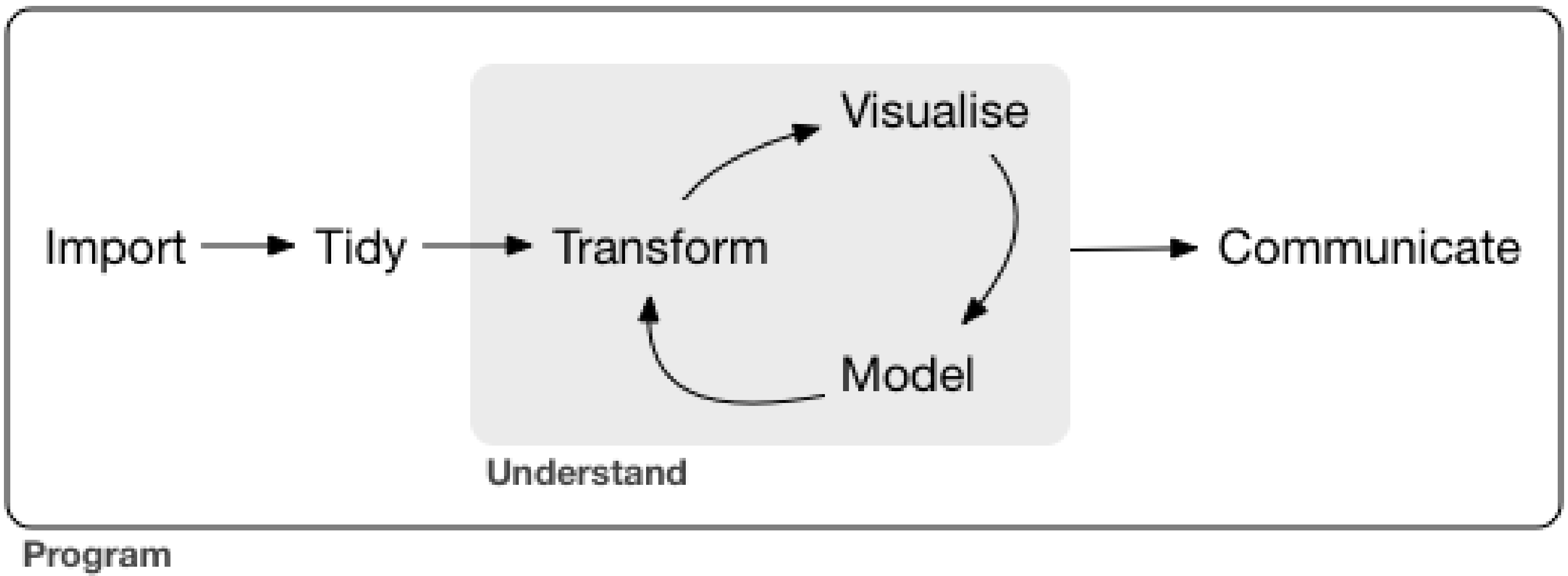
Questions data scientists try to answer

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 :

$$\text{Observations} \rightarrow \text{Theory} \rightarrow \text{Predictions}$$

Machine learning

$$\text{Training error} = f\left(\text{Model}_{\text{P}}, \text{Training set}\right)$$

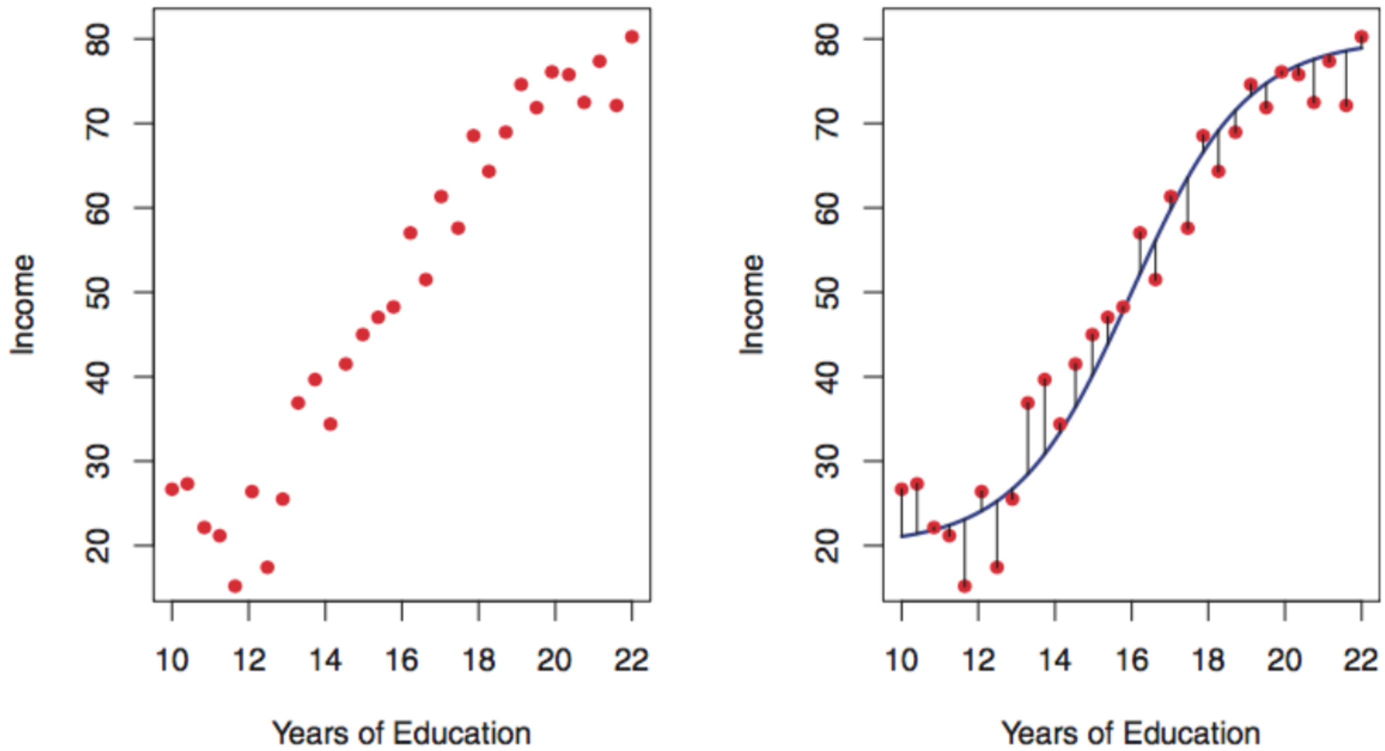
$$\hat{\text{P}} = \min_{\text{P}}\ (f)$$

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

- 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

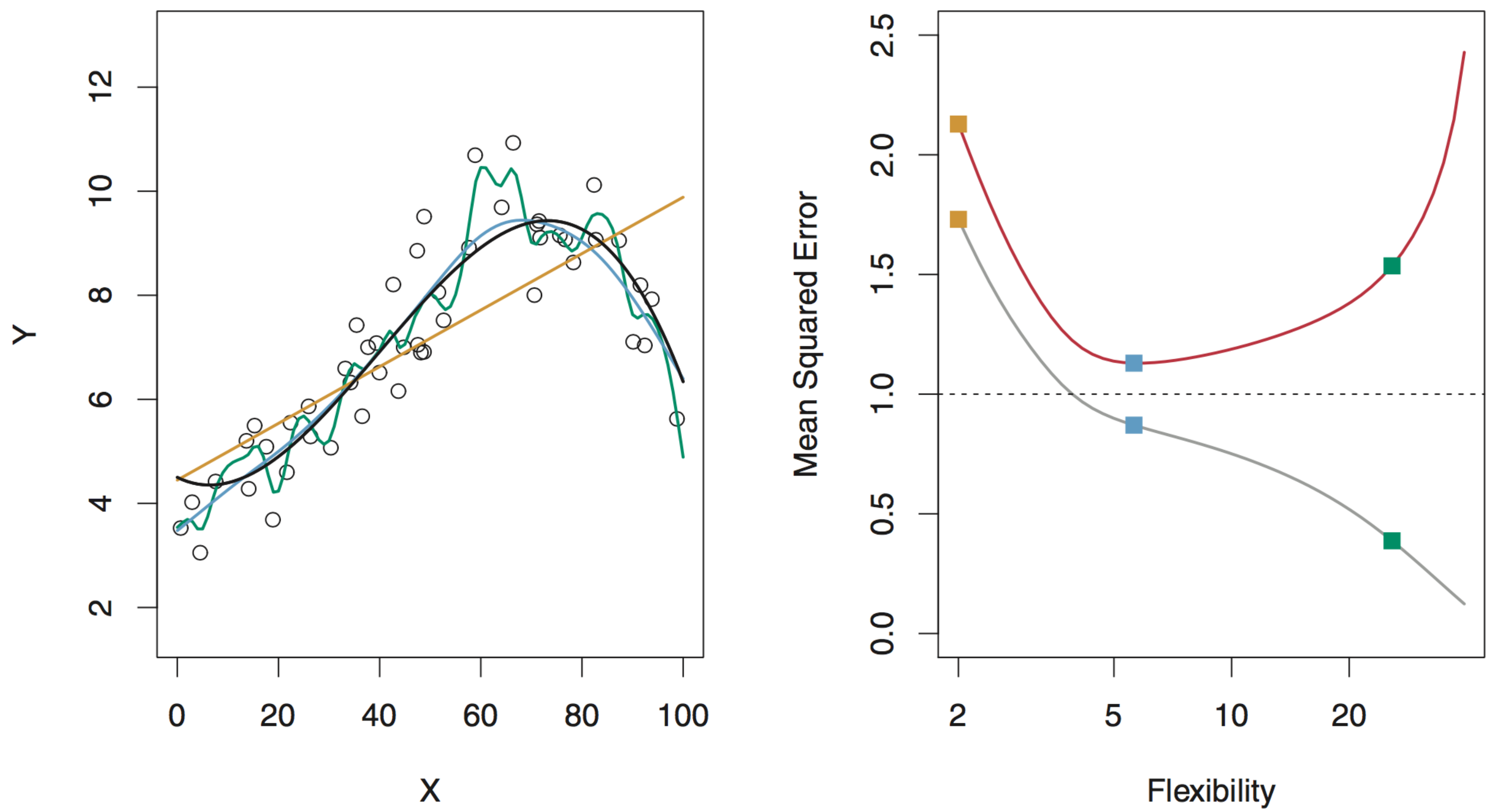
- Reducible error : can be eliminated with more data/better model
- Irreducible error : Inherent randomness and external factors

Bias-variance tradeoff

Properties of a given statistical model -

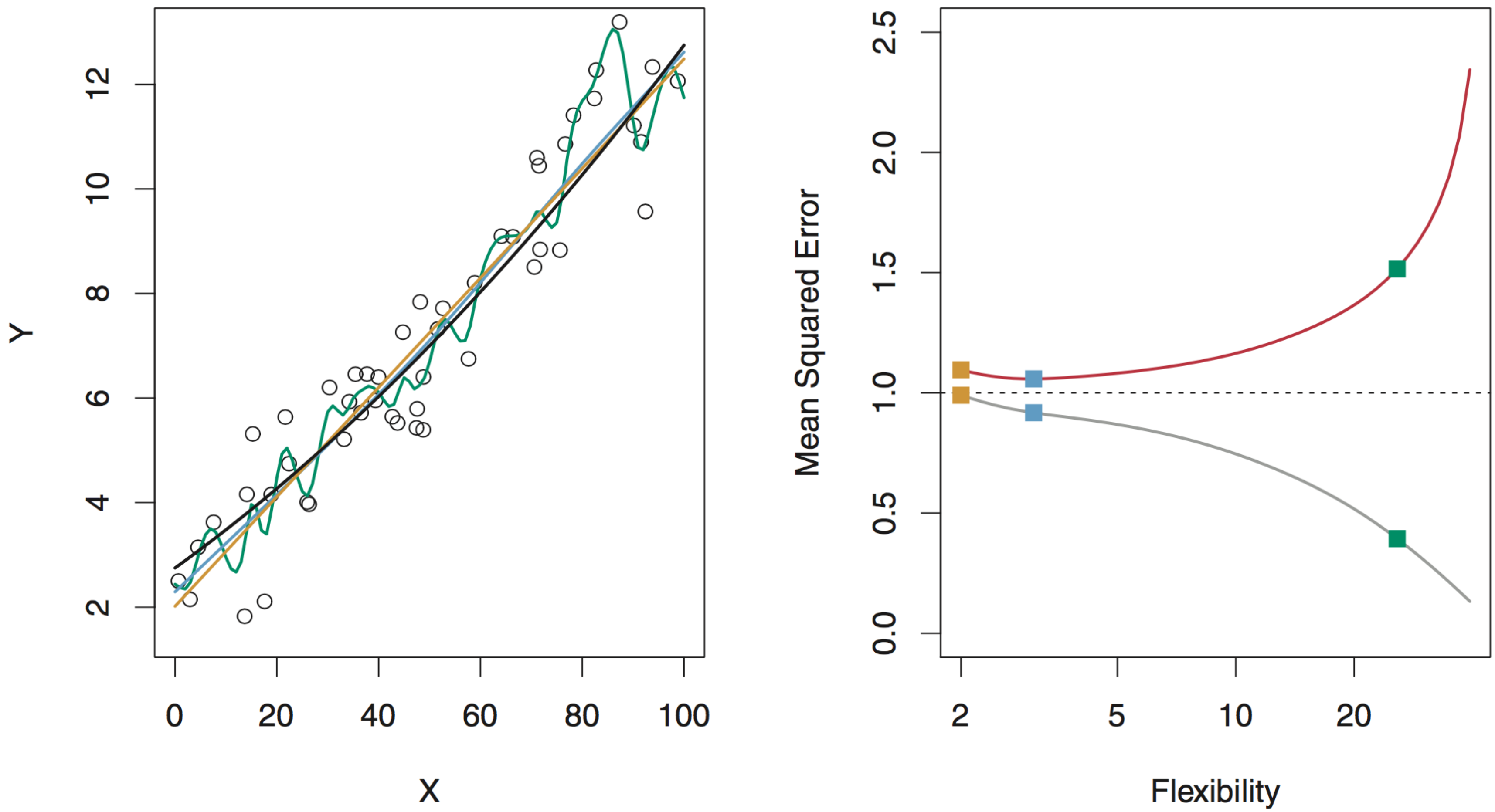
- Bias : Error resulting from approximations in the model. Cannot be improved with more data.
- Variance : Change in prediction due to change in training set.
- Complex (more flexible) models - low bias, high variance
- Simple (less flexible) models - high bias, low variance

Training error always reduces with model complexity



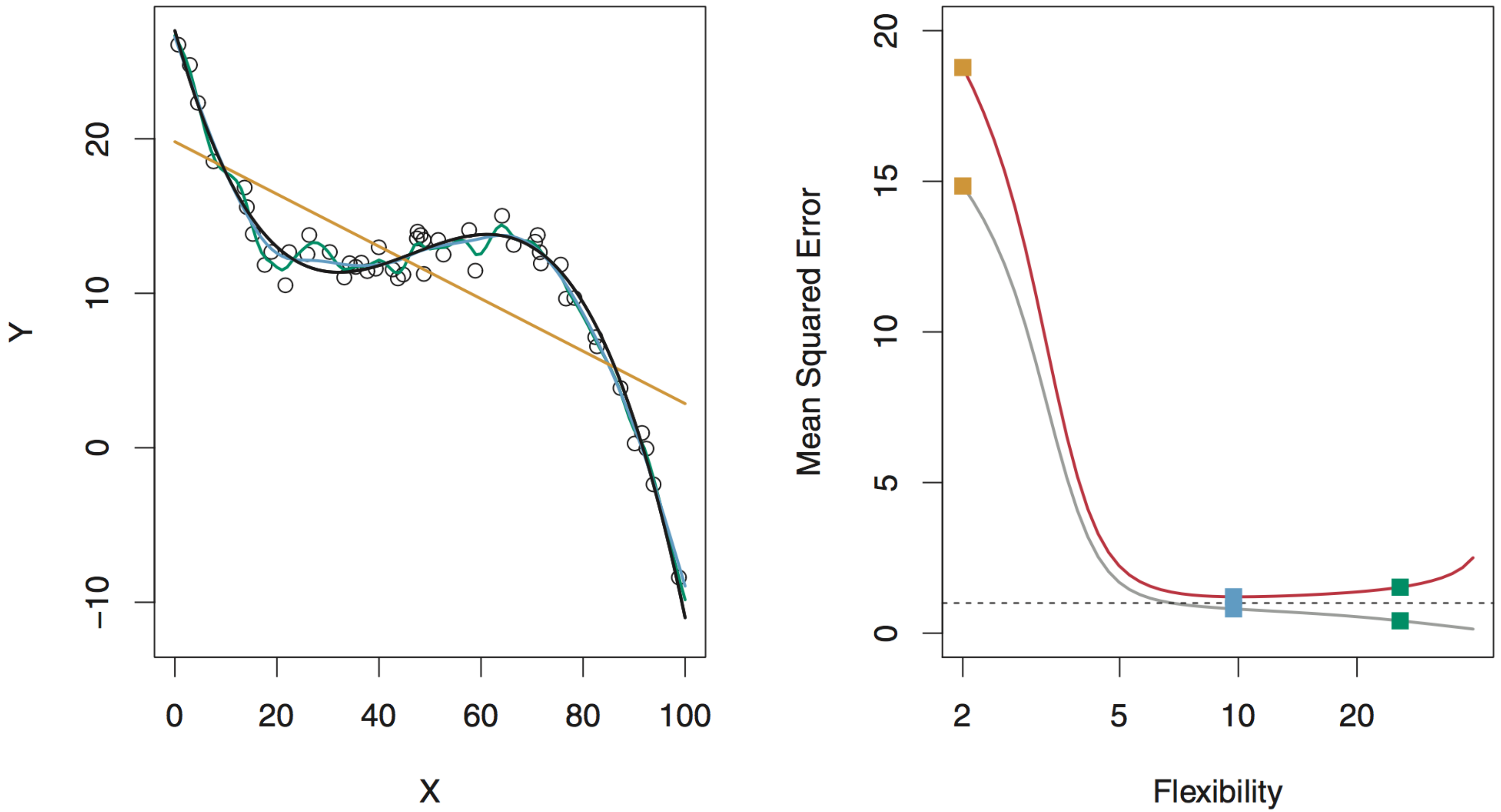
from Introduction to Statistical Learning

More complex models are not always better



from Introduction to Statistical Learning

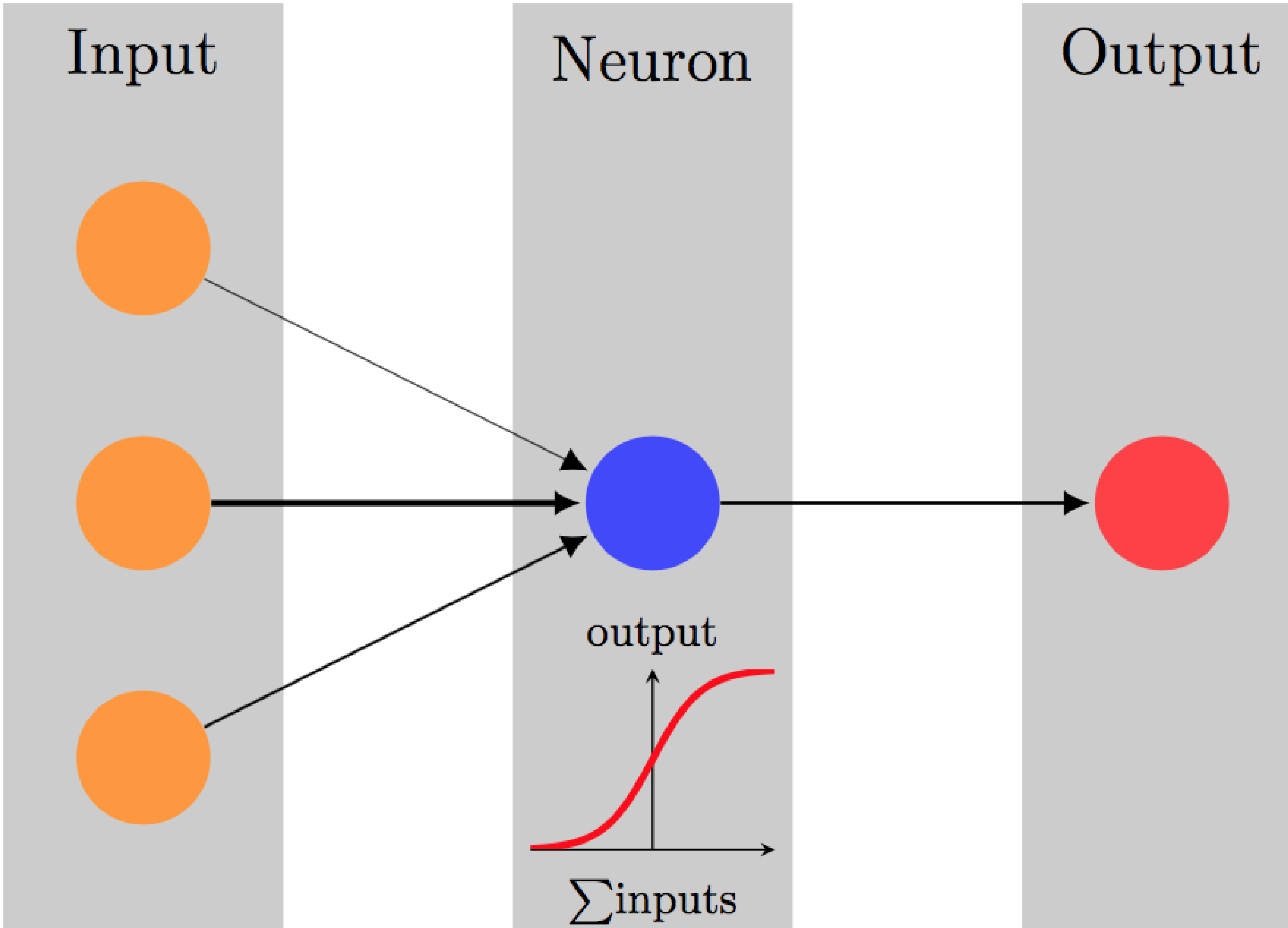
The 'right' model : lowest test error



from Introduction to Statistical Learning

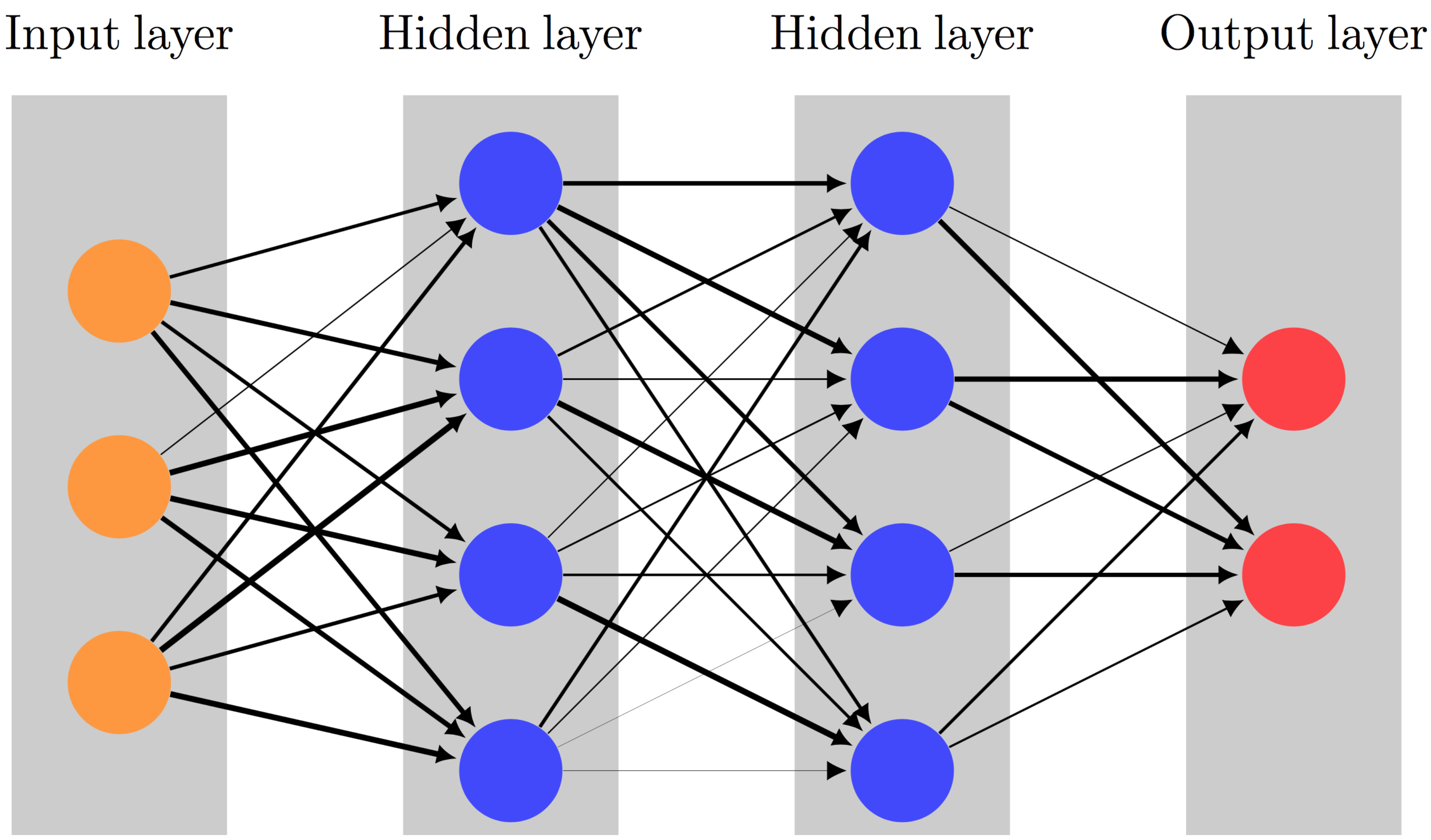
and we surely must mention that most ubiquitous buzzword...

Differentiable networks



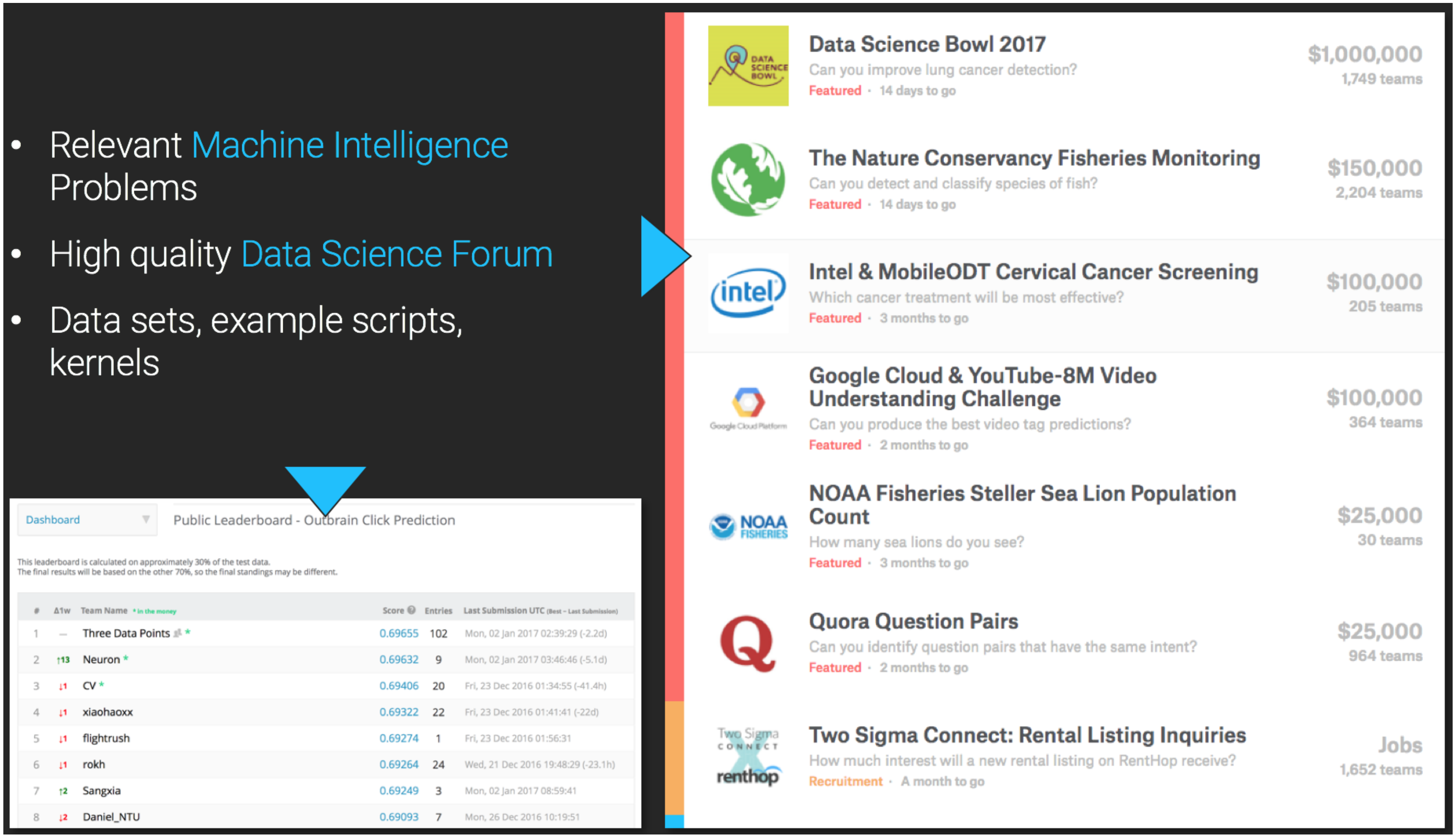
Artificial intelligence : Machine learning that gets spooky

Backpropagation



A wonderful introduction by Nielsen

kaggle



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