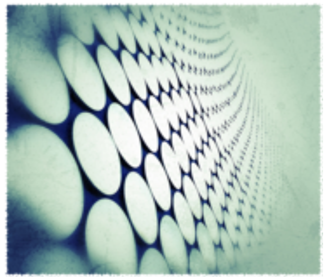



Learning from data: data mining approaches for Energy & Weather/Climate applications

Matteo De Felice, ENEA


http://matteodefelice.name/research
www.utmea.enea.it
matteo.defelice@enea.it

Outline

- Building reliable Climate Services is really challenging
- Cross-disciplinary
- We need to use the latest and most advance research and knowledge
- We need to **use** all available data

It's just a matter of time

* *"Can you tell me how much the climate change will affect the wind power production in Italy for the next two decades?"*
* *"Can you prepare me an early-warning forest fire model?"*

Yes, you can (try)

But how long it will take to elaborate new theories and physical models?

Energy & Climate/ Meteorology

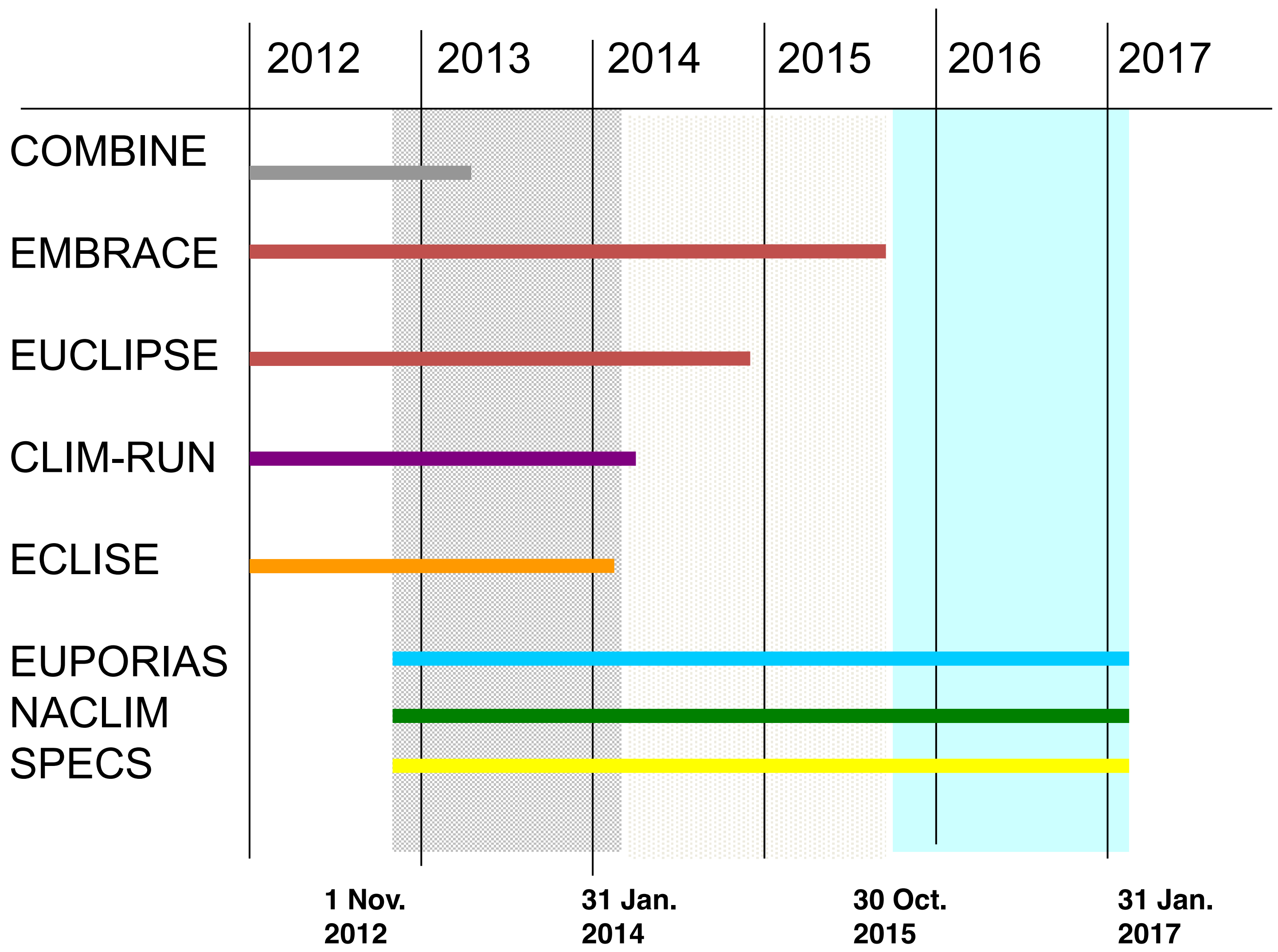
- Link between Energy and Climate/Meteorology is strengthening for several reasons:

1. Diffusion of Renewable Energies
2. Widespread use of air conditioning
3. Necessity of improving efficiency/reliability of power networks (electric utilities)

Energy & Climate/ Meteorology

- Conferences: ICEM
- Projects: CLIM-RUN, EUPORIAS, SPECS
- GFCS Climate Services

EU Projects



from Carlo Buontempo presentation at ICEM 2013

Energy: main challenges

1. Deregulation and competition
2. Climate issue: emissions reduction and higher demands
3. Security and stability: diffusion of non-controllable sources and greater focus on critical infrastructures and dependancies

Renewables Energy

- We need to…
  …find the best place for new plants
  …manage existing plants (efficiently)
  …predict power output (tomorrow, in ten days, in five years)
- We need it **as soon as possible**
- Communicating effectively with stakeholders (and municipalities, regional authorities, etc.) is fundamental

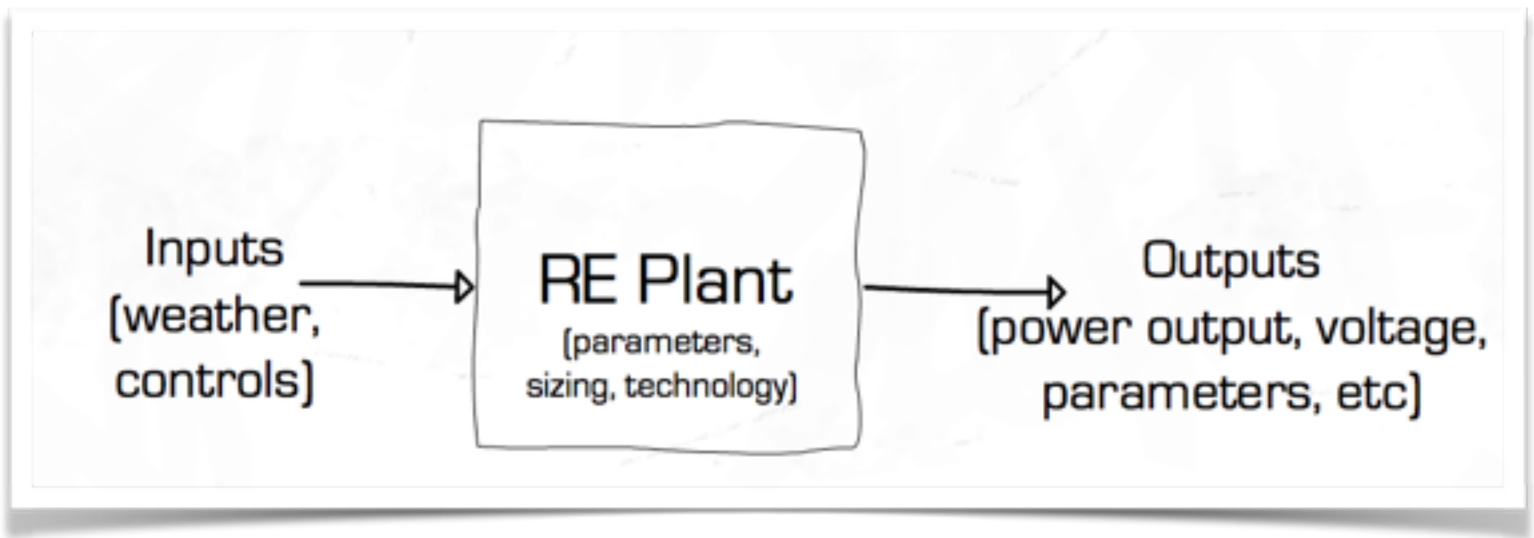
Data

			Type of climatological input data	
Field of application	*Energy element needed*	*Climatological element needed*	*Time scale*	*Space scale*
Power Grid	Daily load and network structure	Air temperature	d, m	g, a
Gas production and distribution	Network structure	Precipitation	h, d	g, a
		Air temperature	d, m	g, a
		Wind speed	d, m	g, a
Wind	Production – high resolution time series	Wind	h, d	s, g
		Pressure	d, m	s, g
Solar	Production data	Radiation	h, d	s, g
		Wind	min, max	s, g
		Temperature	d, m	s, g
		Humidity	d, m	s, g
		Clouds	h, d	s, g
		Aerosols	h, d	s, g
Hydropower and water-based cooling systems	Production data	Temperature	d, m	s, a
	Plants temperature	Rainfall	d, m	s, a
		Runoff	d, m	s, a
		Water level	d, m	s, a
		Snow cover	d, m	s, a
		Snow melt	d, m	s, a
		Soil Moisture	d, m	s, a

from Ruti & De Felice, Climate and Energy Production – A Climate Services Perspective, Climate Vulnerability: Understanding and Addressing Threats to Essential Resources. Elsevier Inc., Academic Press, 2013.

Example

- Goal: “assess the forecast quality for solar and wind energy generation at s2d time scales”



Example

Available RE information

Nothing

Create software models of RE plants based on my assumptions about technology, typology, etc.

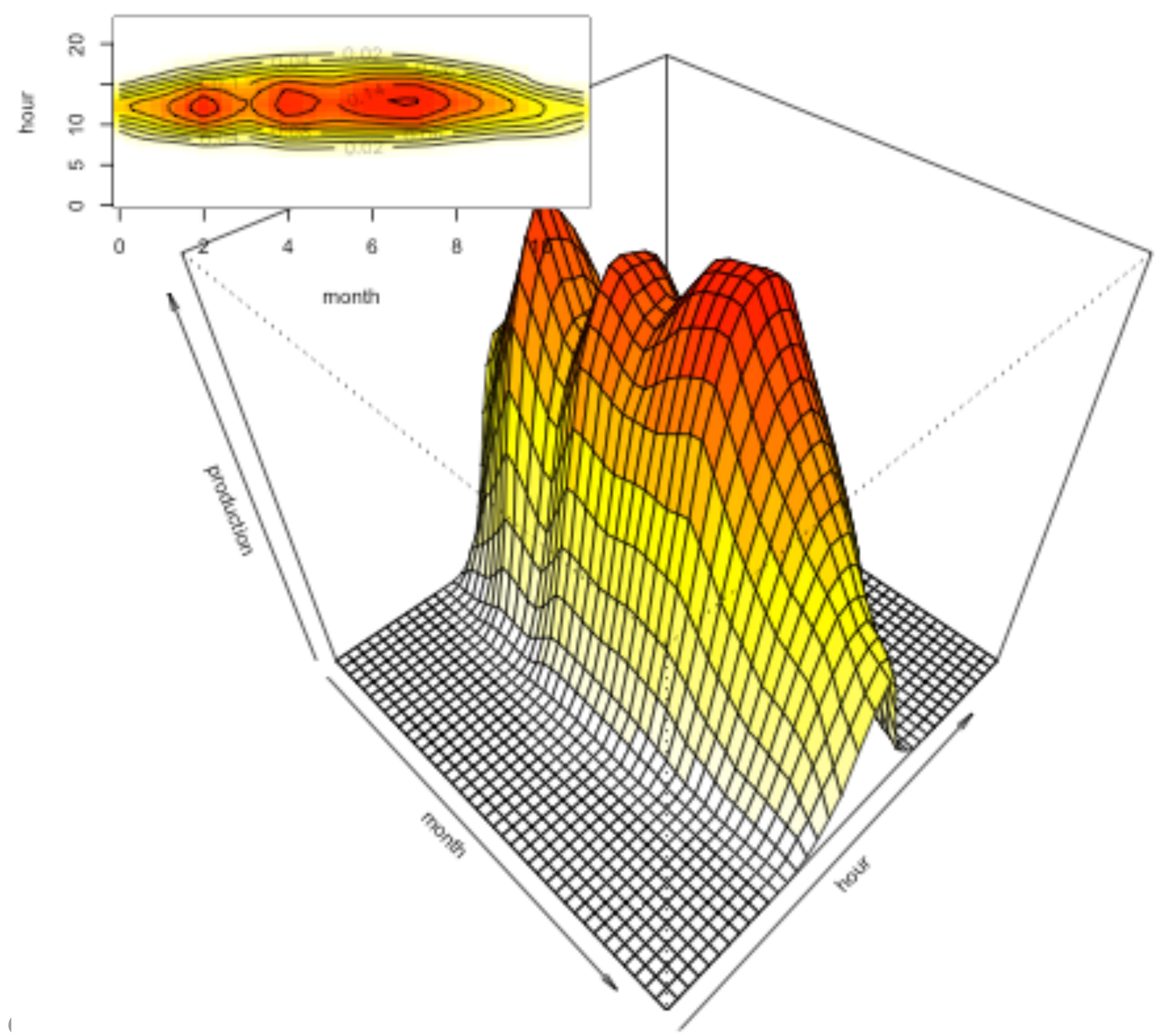
Something

Full

Create models of RE plants using real parameters and options.

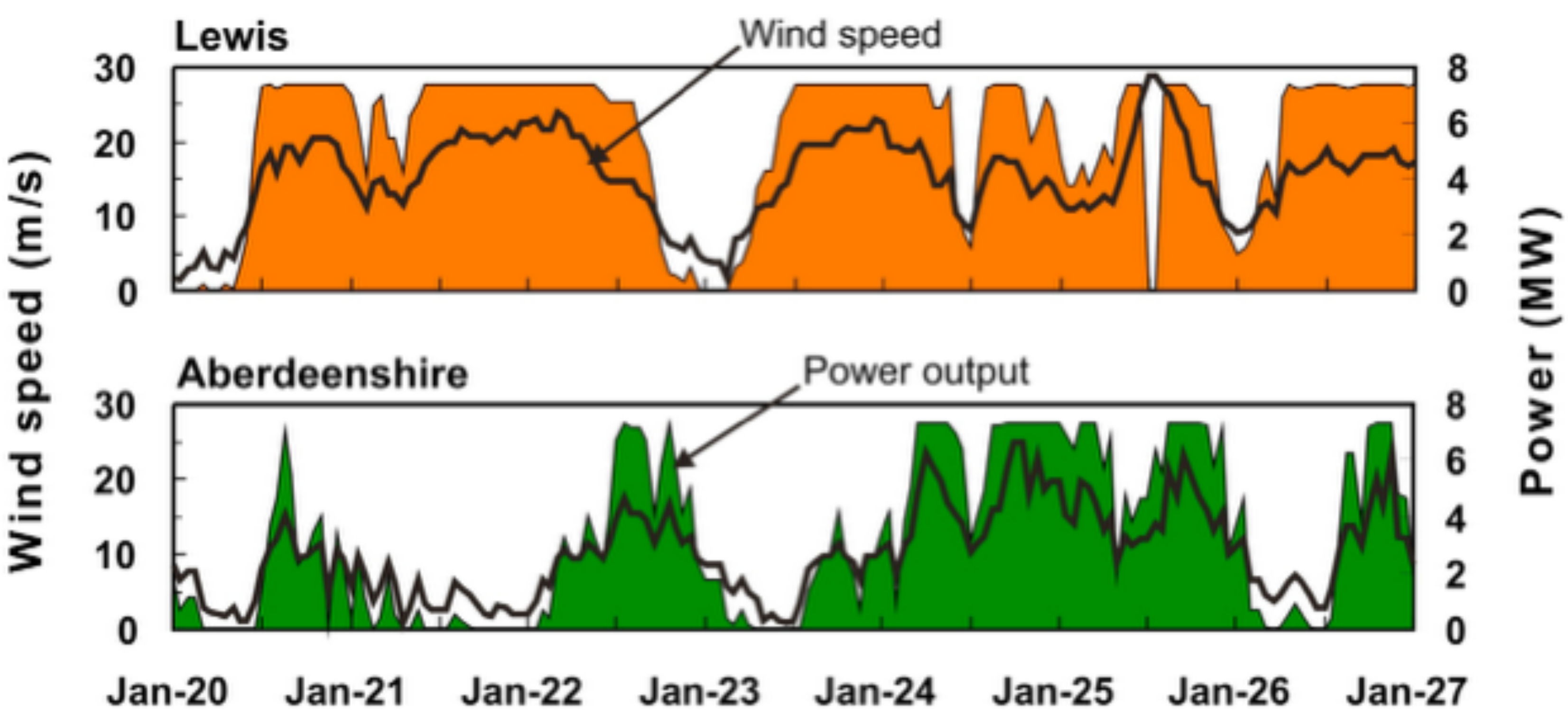
Solar Power

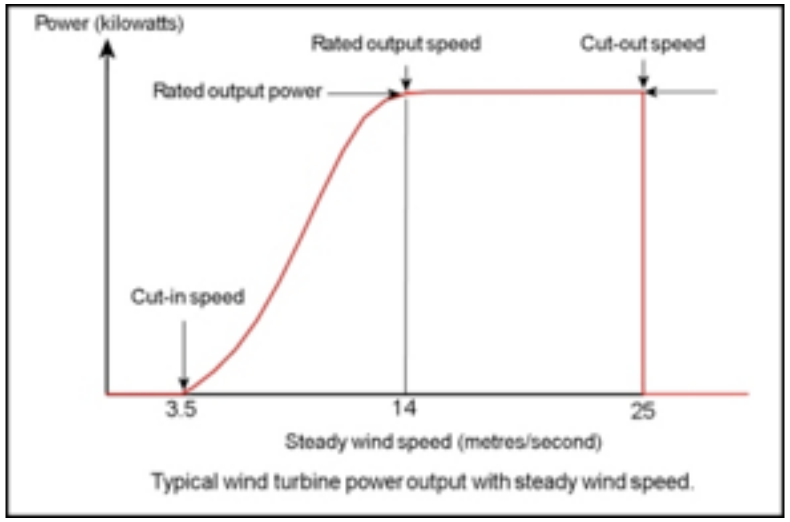
- Power output (photovoltaics) is proportional with solar radiation and affected by with cloud cover
- Rise of temperature leads to reduced efficiency
- Shadowing effects



Wind Power

- Wind speed at specific height is needed (70-100 m)
- Strong non-linear effect (cut-off speed)
- Wind turbines interactions (wake effect)
- Strong intermittence





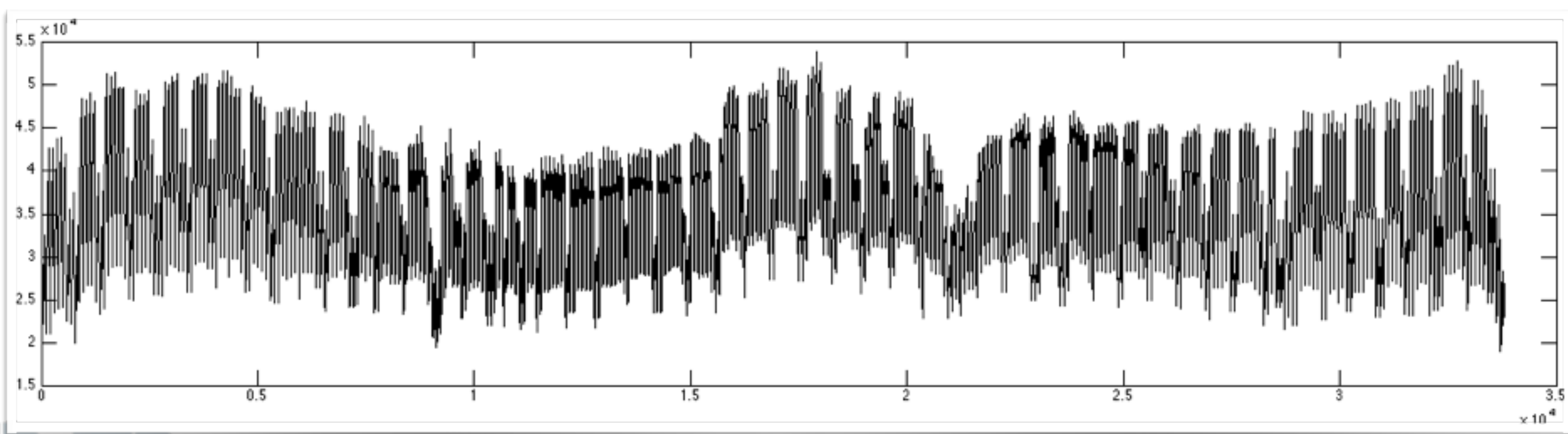
Typical wind turbine power output with steady wind speed.

Hydro-power

- Hydro power is used to store water for electricity generation during the time of year when it is most valuable.
- Precipitation (snow and rain) information needed: when / where / how much (intensity)
- Necessity of hydrological models and in-situ measurements

Electricity Demand

- Electricity demand sensitive to weather conditions
- Currently only climatological data are used for time-scales >14 days
- Demand affected by “human activities” (calendar effects) and economic trends



Energy & Meteorology

- Impact of spatial-temporal variability on energy systems
- Prediction of expected power in high spatial and temporal resolutions
- Operational aspects
- Critical for market operations

How much energy this RE plant will produce in the next three hours?

Energy & Climate

- Longer time-scales -> higher uncertainty
- Climate risk management (e.g. extreme events)
- Use of S2D climate forecasts

How much energy this RE plant will produce every year?

Energy Sector Vulnerability

- Energy Sector is vulnerable to climate change (broadly speaking)
- E.g. During European 2003 heat-wave France reduced electricity export in August of 50% (EDF)

> […] a summer average decrease in capacity of power plants of 6.3–19% in Europe and 4.4–16% in the United States depending on cooling system type and climate scenario for 2031–2060. In addition, probabilities of extreme (>90%) reductions in thermoelectric power production will on average increase by a factor of three.
>
> *(van Vliet et al., Vulnerability of US and European electricity supply to climate change, Nature Climate Change 2(9), 2012)*

Vulnerabilities

- **Hydro-power** depends on hydrological cycle (seasonal pattern). Higher impacts on regions where snowmelt is a relevant factor.
- **Wind power** necessity wind speed data at height >50m. Terrain roughness is a key parameter and it’s affected by vegetation cover.
- **Solar energy** is affected by clouds and water vapour content

Vulnerabilities

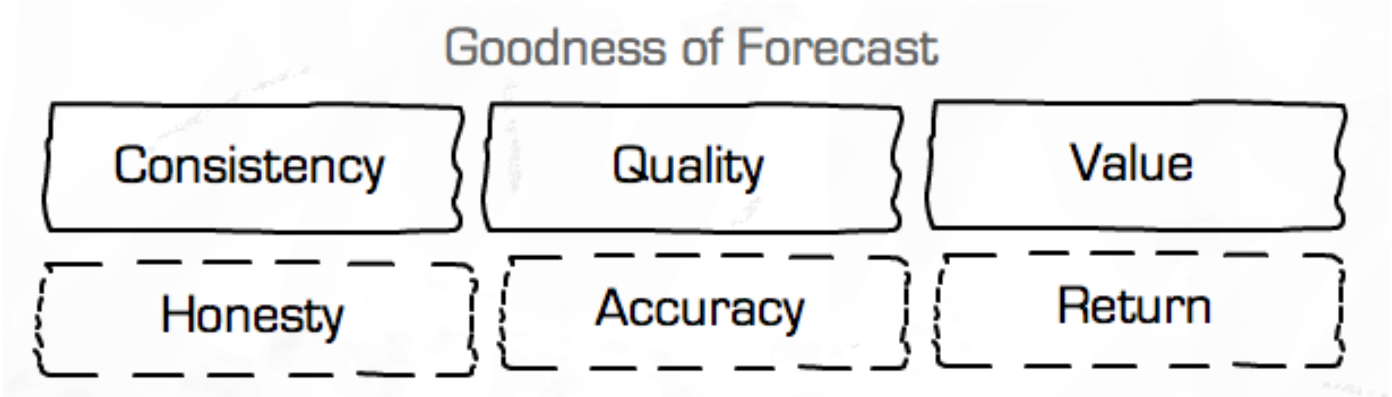
Sector	Variables	Impacts
Gas, coal and nuclear	Air temperature	Cooling water quantity and quality (efficiency)
Oil and gas	Extreme events	Extraction/import distruption
Hydro-power	Precipitation	Water availability
…	…	…

see Schaeffer et al., Energy sector vulnerability to climate change: A review, Energy (38), 2012

Predicting changes

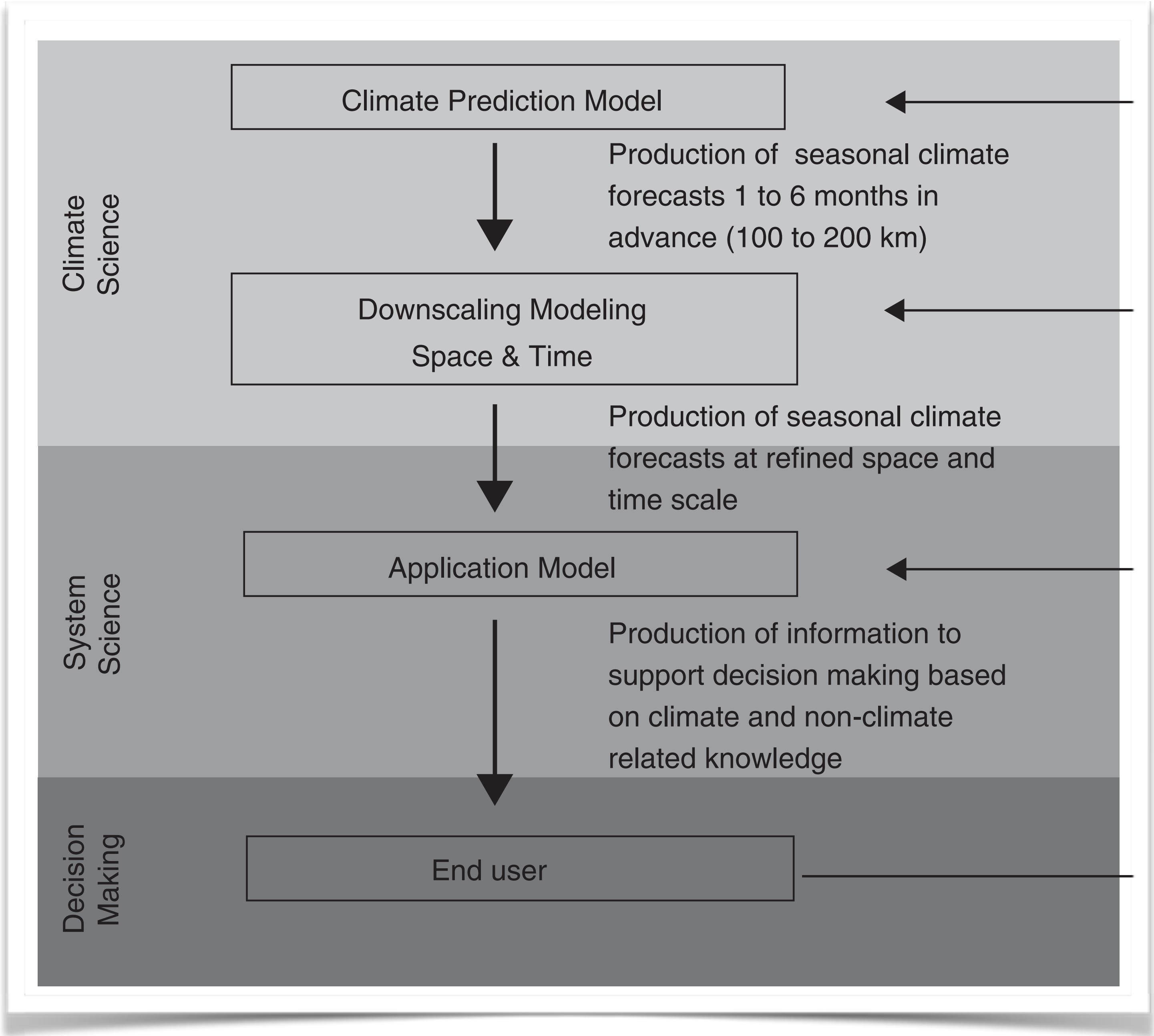
- “Prediction is very difficult, especially about the future”. (Niels Bohr)
- The impact of a predicted change in climate/weather variables on energy production/demand is not straightforward. Highly non-linear.
- How the uncertainty propagates?

Side question



Armour, P. G. (2013). What is a good estimate?: whether forecasting is valuable. Communications of the ACM, 56(6), 31-32.

Forecasting



from Coelho and Costa, Challenges for integrating seasonal climate forecasts in user applications, Current Opinion in Environmental Sustainability (2), 2010

Approaches

- To take decisions we need estimates and predictions
- To generate predictions we need models
- To build models we need data and information

DRIP & Big Data

- DRIP (Data Rich Information Poor) era (Big Data)
- We need tools to deal with high-dimensional and heterogeneous data

Data → ? → Information

"factual information (as measurements or statistics) used as a basis for reasoning, discussion, or calculation"
(Merriam-Webster)

"knowledge obtained from investigation, study, or instruction"
(Merriam-Webster)

Data vs Information

formally-defined

Precise ⟷ Accurate

Data

Information

Useful ⟷ Meaningful

context-related

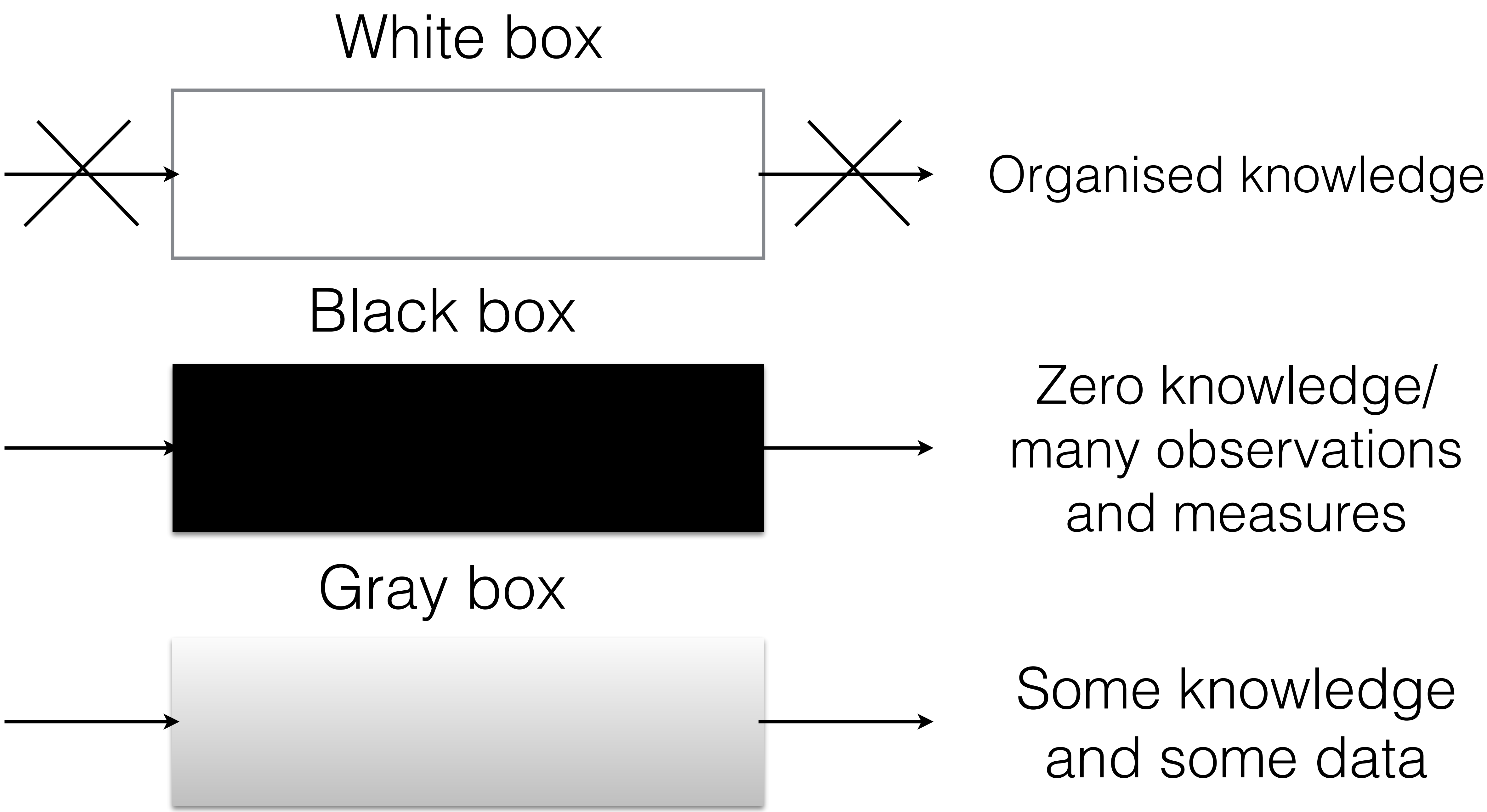
White Box & Black box

- How to use data/information to build models…
  …able to generalise?
  …reliable?
  …consistent?

“Essentially, all models are wrong, but some are useful.”

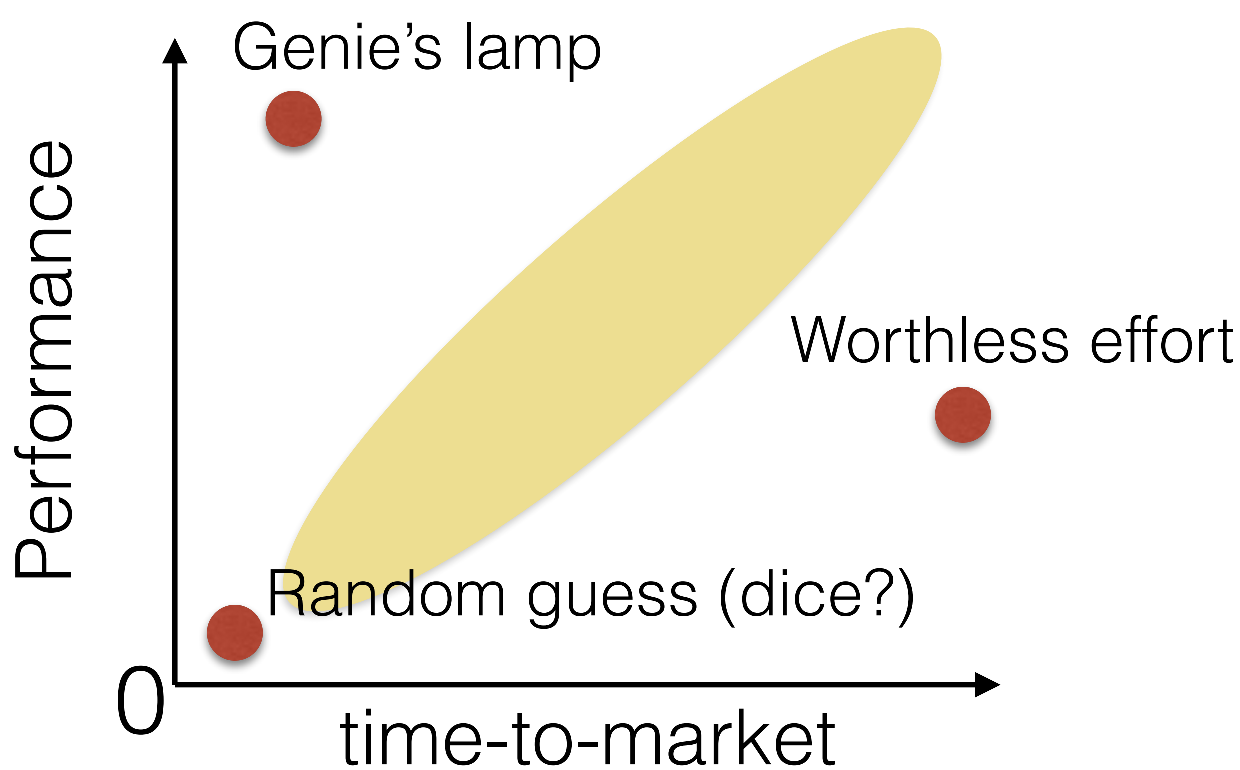
George E. P. Box (English statistician)

Boxes


White box
Organised knowledge
Black box
Zero knowledge/
many observations
and measures
Gray box
Some knowledge
and some data

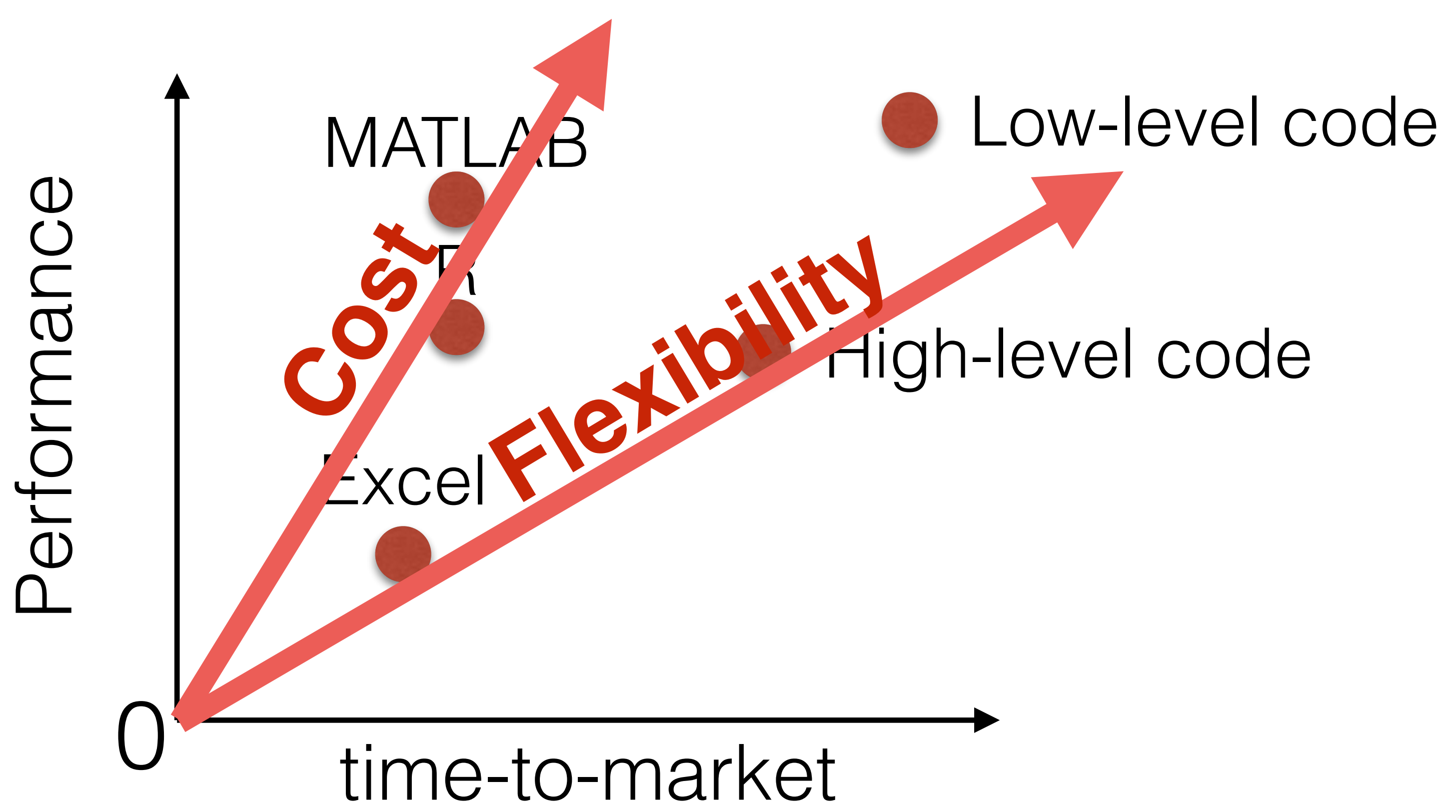
Right tools

- Wolpert "No Free Lunch" theorem: "best model" doesn't exist considering all possible problems
- Exploring the trade-off between time-to-market and efficiency/performances



Modeling software

- Spreadsheet software (e.g. Microsoft Excel)
- Code programming (C/C++, Python)
- Statistical computing environments (e.g. MATLAB, R)

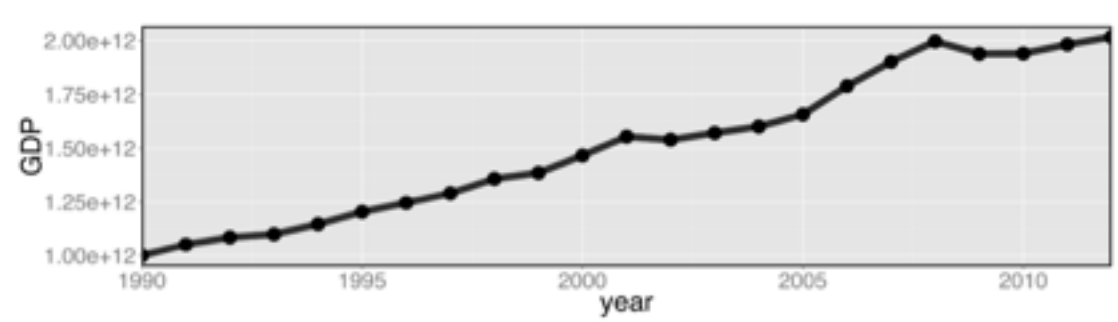


R

- **open source** multi-platform software for statistical computing
- well-established with over **5000** packages

- easy and quick statistic analysis
- availability of free packages developed by research centres worldwide
- quick access to databases and files of any format

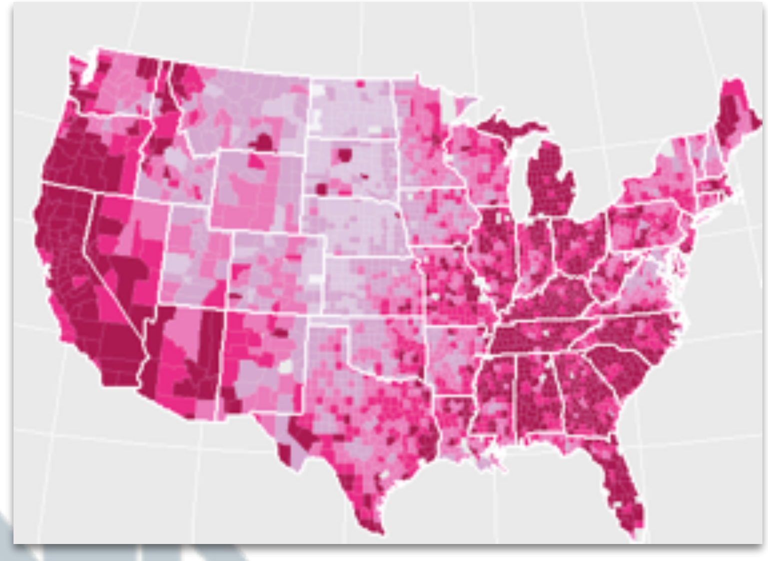
```
> gdp = getWorldBankData(id = 'NY.GDP.MKTP.PP.CD', date = ‘1990:2012')
> plot(gdp)
```

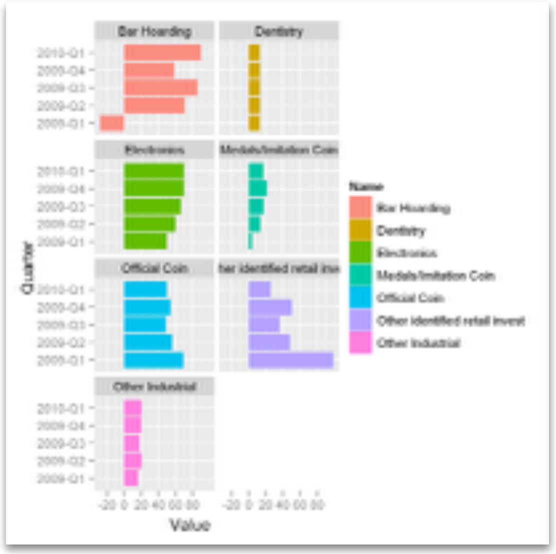


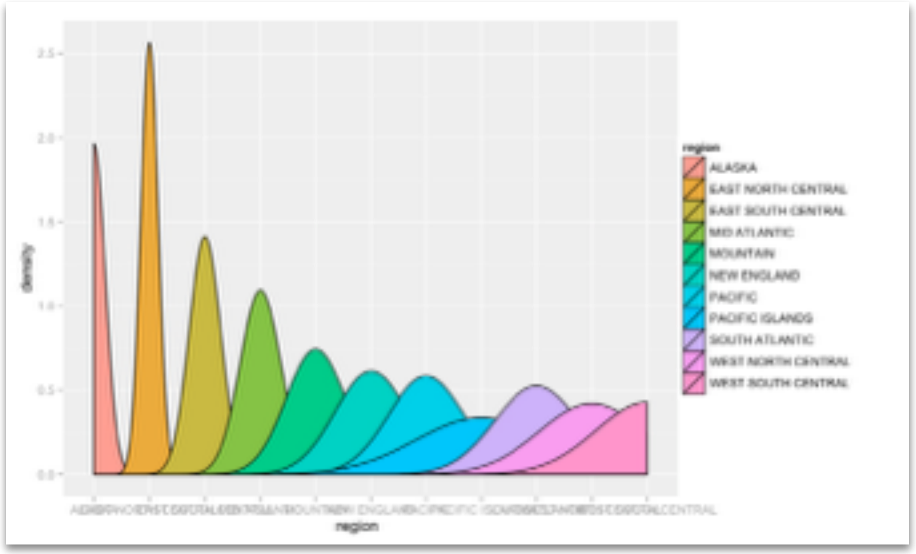
Data Visualisation

1st rule: LOOK AT YOUR DATA
2nd rule: ALWAYS LOOK AT YOUR DATA

- Examine your data to detect evident anomalies
- Nice and clear plots help you to understand your data (and to prepare high-level publications)
- Best tool: ggplot2 (R package)





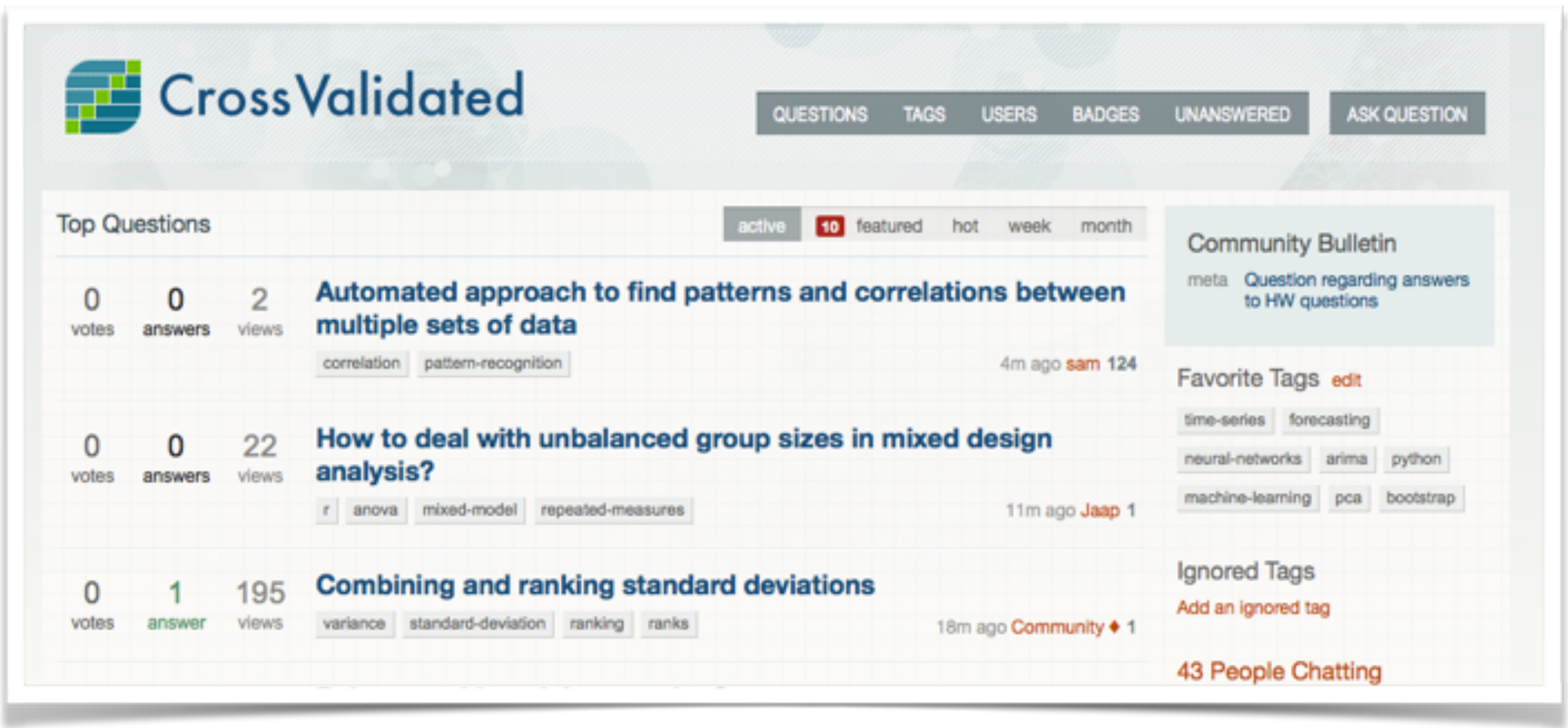


Do you want to learn R?

- Web resources
- Coursera.org courses

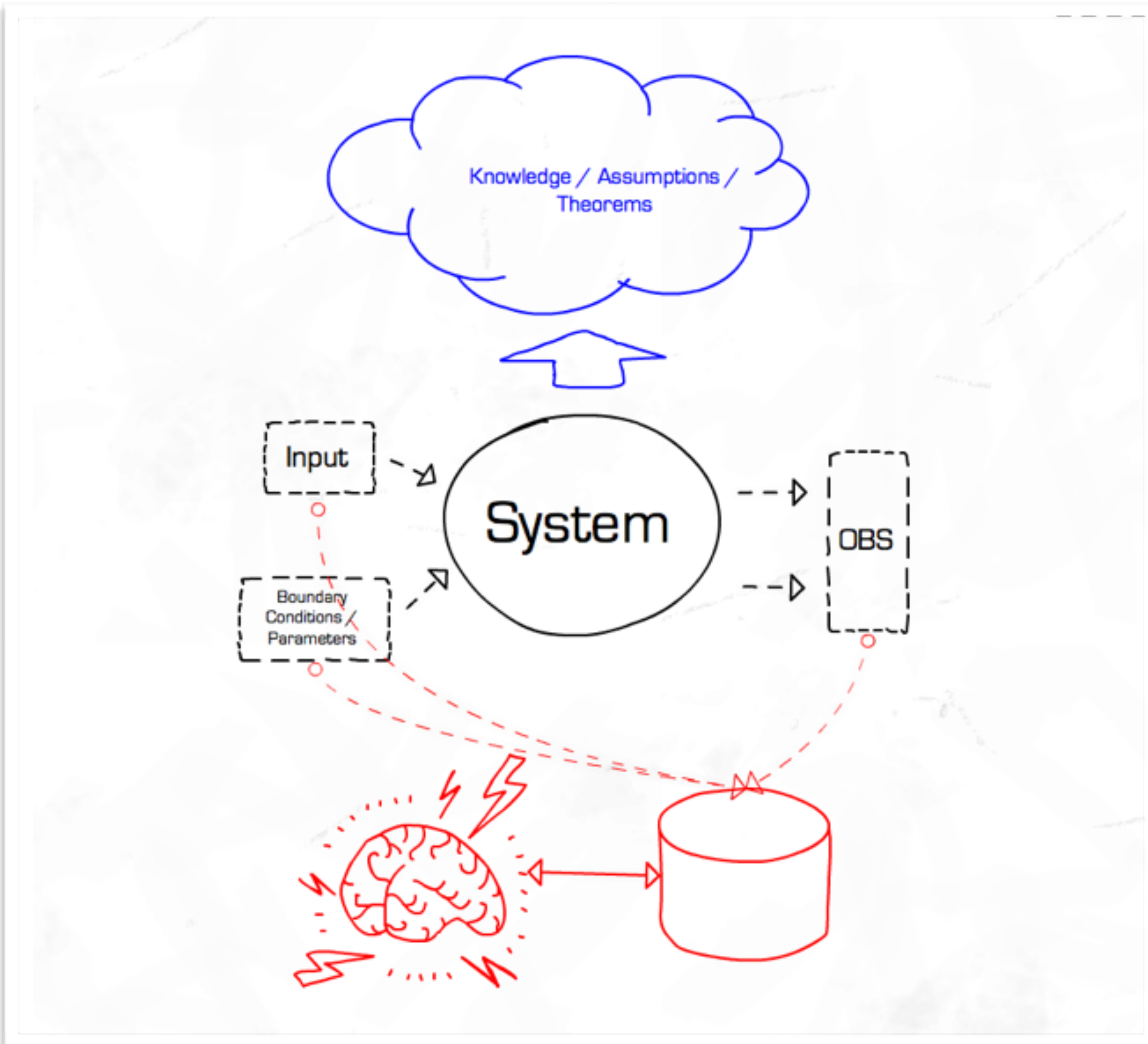
Any question?

- http://stats.stackexchange.com/



http://area51.stackexchange.com/proposals/36296/geoscience

Dealing with Complexity



Data Mining

- We are in the DRIP era (Data Rich Information Poor)
- Big Data
- From data to information

ECMWF stores
50 petabytes of data
($50 \cdot 10^{15}$)

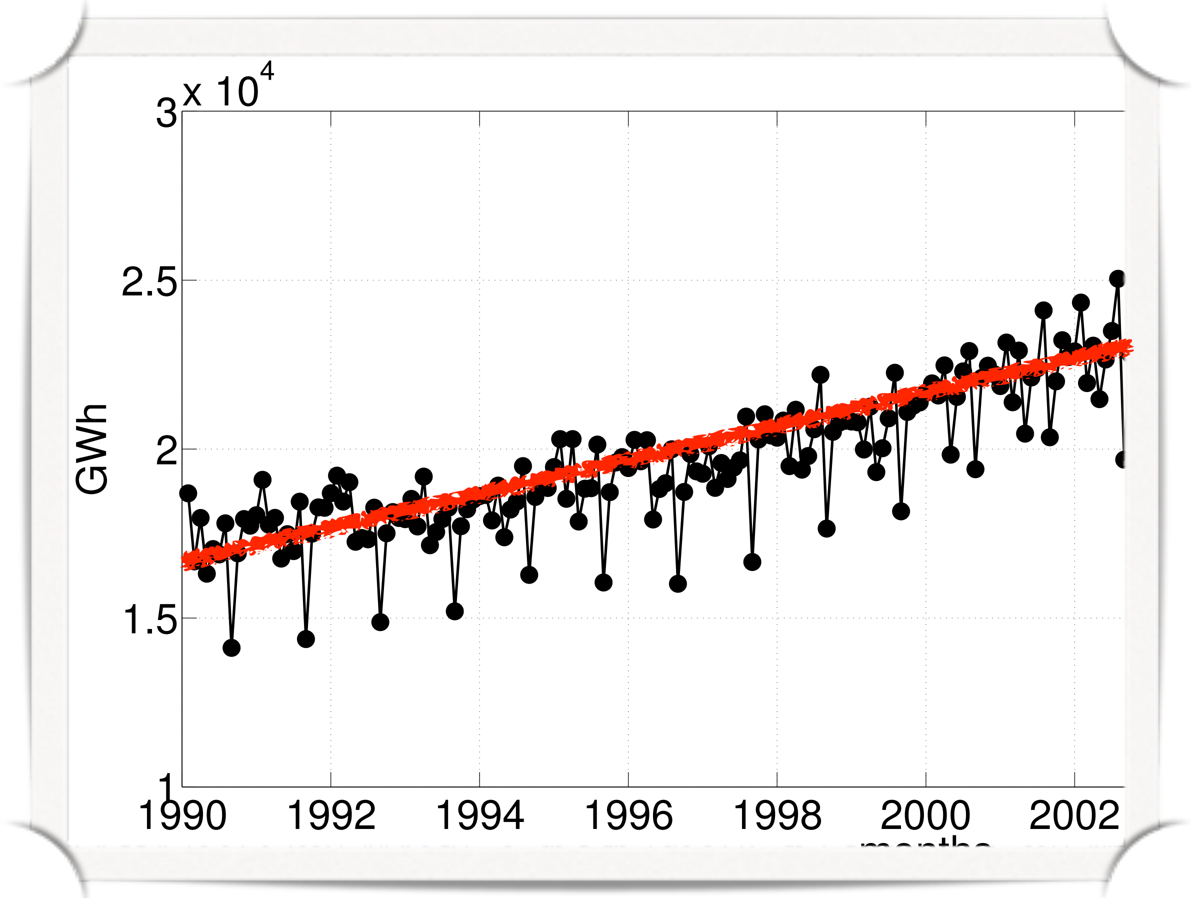
factual information (as measurements or statistics) used as a basis for reasoning, discussion, or calculation

knowledge obtained from investigation, study, or instruction

Data Mining

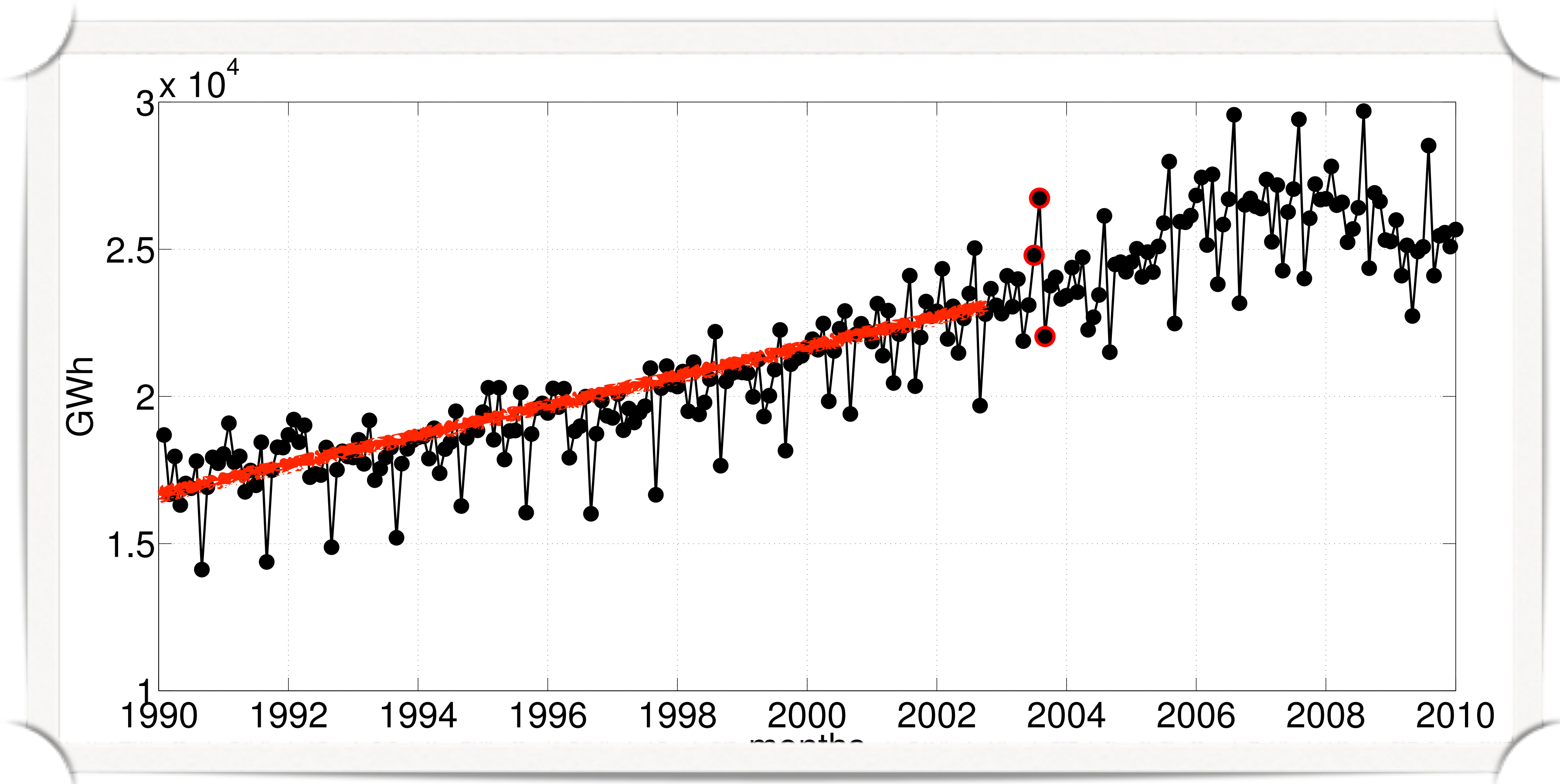
- “Extraction of implicit and **potentially useful** information from data”
- Automated search (software)
- Main difficulties:
  1. Defining “interestingness”
  2. Spurious and accidental coincidences (exceptions)
  3. Missing data and noise

Example

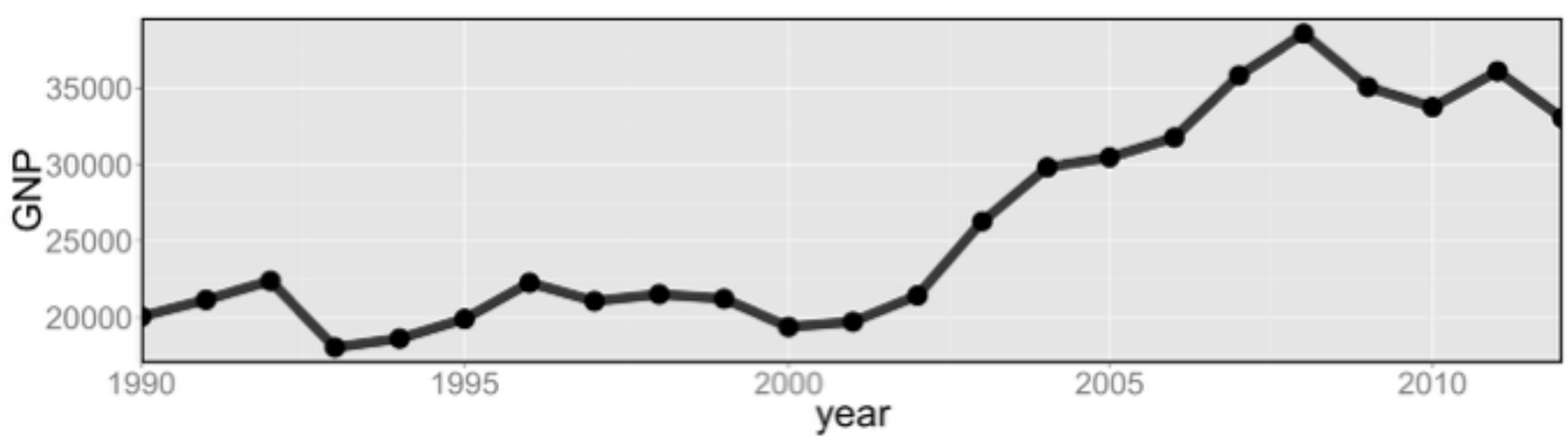


Italian Monthly Electricity Demand

Example



x 10^4
3
2.5
2
1.5
1
GWh
1990 1992 1994 1996 1998 2000 2002 2004 2006 2008 2010



35000
30000
25000
20000
GNP
1990 1995 2000 2005 2010
year

Modelling Methods

- Experience
- Rule-of-thumbs and good practises
- Statistics
- Machine Learning methods

Why?

- Forecasting
- Analysis of past events
- Control
- Anomaly Detection

Example

Outlook	Temperature	Humidity	Windy	Play
Sunny	hot	high	false	no
Sunny	hot	high	true	no
Overcast	hot	high	false	yes
Rainy	mild	high	false	yes
Rainy	cool	normal	false	yes
Rainy	cool	normal	true	no
Overcast	cool	normal	true	yes
Sunny	mild	high	false	no
Sunny	cool	normal	false	yes
Rainy	mild	normal	false	yes
Sunny	mild	normal	true	yes
Overcast	mild	high	true	yes
Overcast	hot	normal	false	yes
Rainy	mild	high	true	no

from Ian Witten, Data Mining: Practical machine learning tools and techniques, Morgan Kaufmann, 2005.

Example

Fire Weather Index system

X	Y	month	day	FFMC	DMC	DC	ISI	Temp	RH	wind	rain	area (ha)
7	5	mar	fri	86.2	26.2	94.3	5.1	8.2	51	6.7	0	0
9	9	jul	tue	85.8	48.3	313.4	3.9	18	42	2.7	0	0.36
3	6	sep	mon	90.9	126.5	686.5	7	15.6	66	3.1	0	0
6	4	aug	thu	95.2	131.7	578.8	10.4	20.3	41	4	0	1.9
6	3	aug	thu	91.6	138.1	621.7	6.3	18.9	41	3.1	0	10.34
7	5	aug	tue	96.1	181.1	671.2	14.3	27.3	63	4.9	6.4	10.82

Burned area of forest fires, in the northeast region of Portugal (517 records)
[Cortez and Morais, 2007]

Example

- Soybean disease database with 307 records
- 19 different diseases
- 35 attributes like month, anomalies in growth, stem, leaves, visible damages, precipitation and temperature, etc.
- Plant pathologist identifies correctly 72%
- Computer-generated rules 97.5%

Methods

- How to represent discovered patterns?

Tables

- Rudimentary and simple
- Look up the appropriate “inputs”

Outlook	Temperature	Humidity	Windy	Play
Sunny	hot	high	false	no
Sunny	hot	high	true	no
Overcast	hot	high	false	yes
Rainy	mild	high	false	yes

- Problem with large (real-world) datasets
- What about continuous variables?

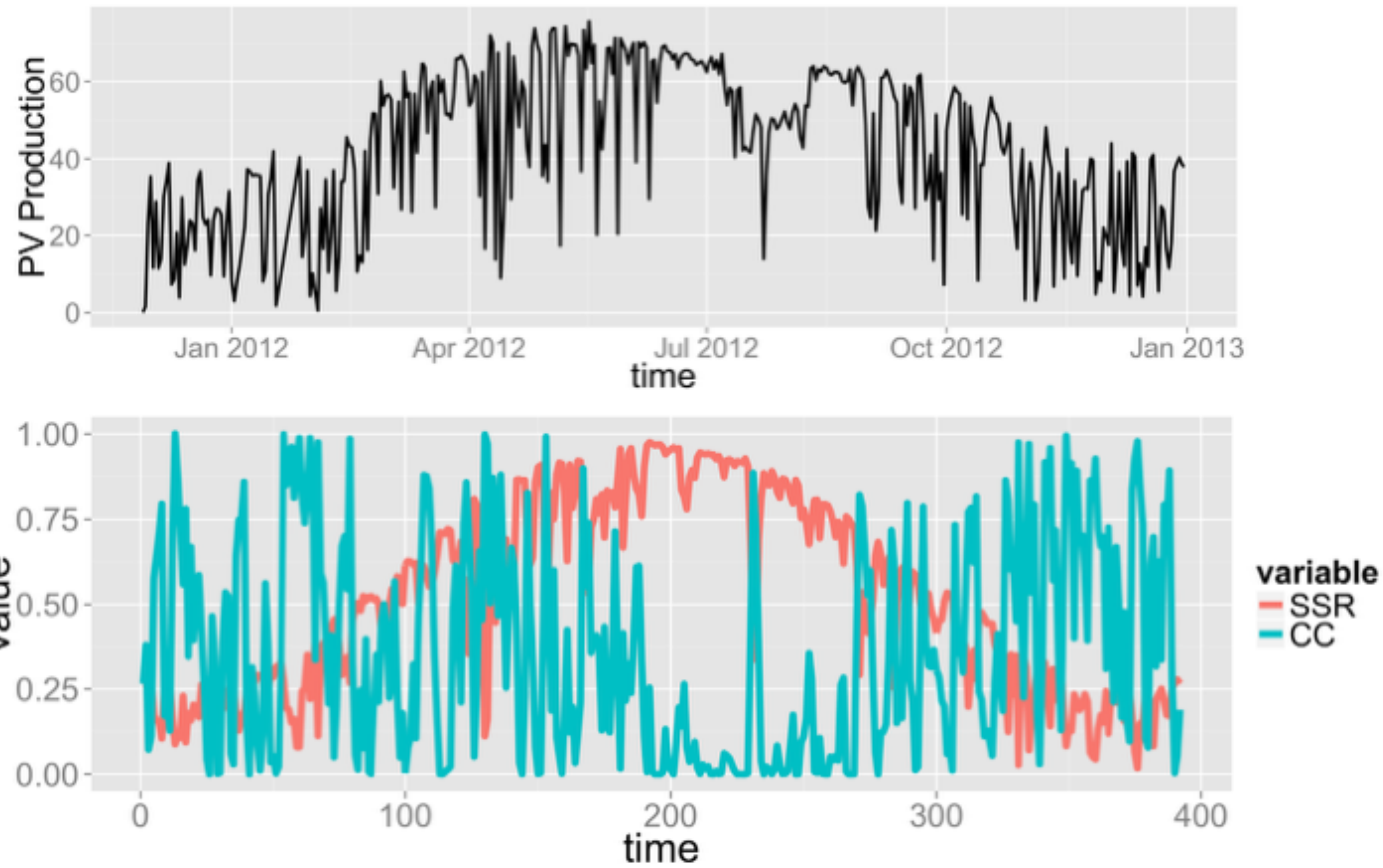
X	Y	month	day	FFMC	DMC	DC	ISI	Temp	RH	wind	rain	area (ha)
6	4	aug	thu	95.2	131.7	578.8	10.4	20.3	41	4	0	1.9

X	Y	month	day	FFMC	DMC	DC	ISI	Temp	RH	wind	rain	area (ha)
6	4	aug	thu	90	135	550	10.4	22	40	4	0	?

Linear models

- Regression (statistics)
- Can be applied to continuous and binary variables

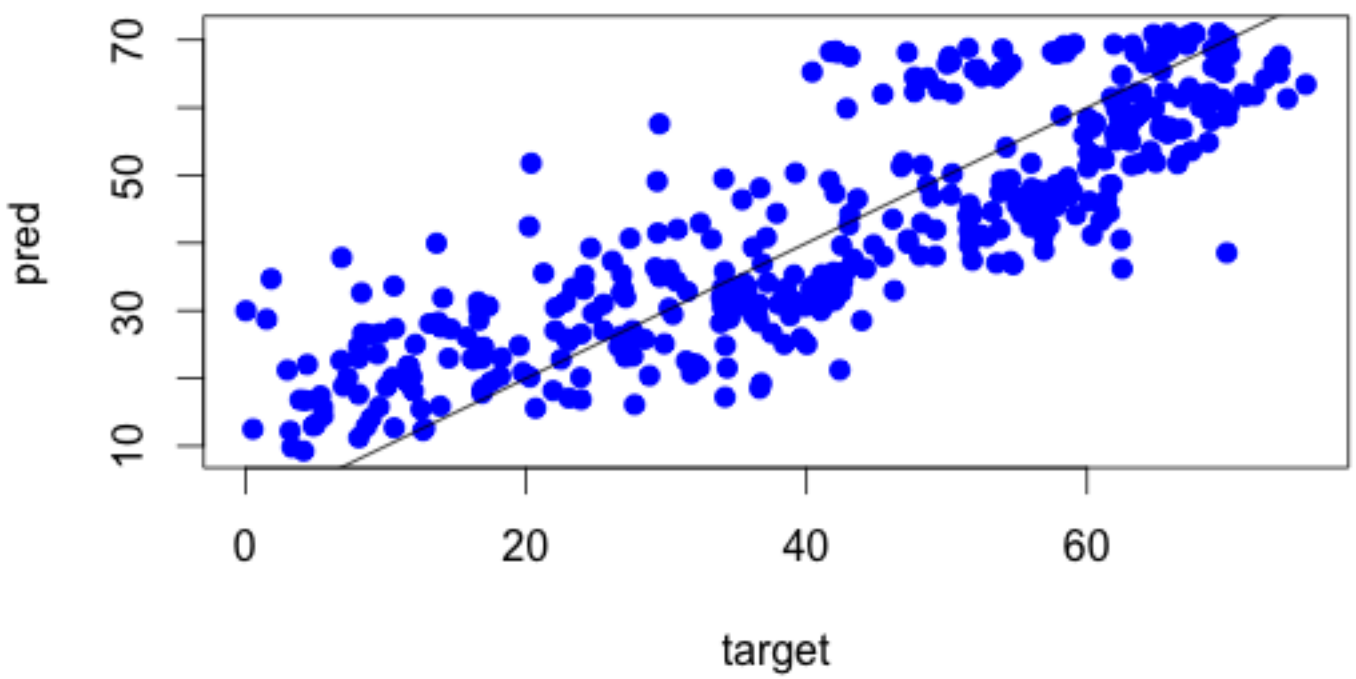
Example



$$y = a_1 \mathrm{SSR} + a_2 \mathrm{CC} + a_3$$

Example #2

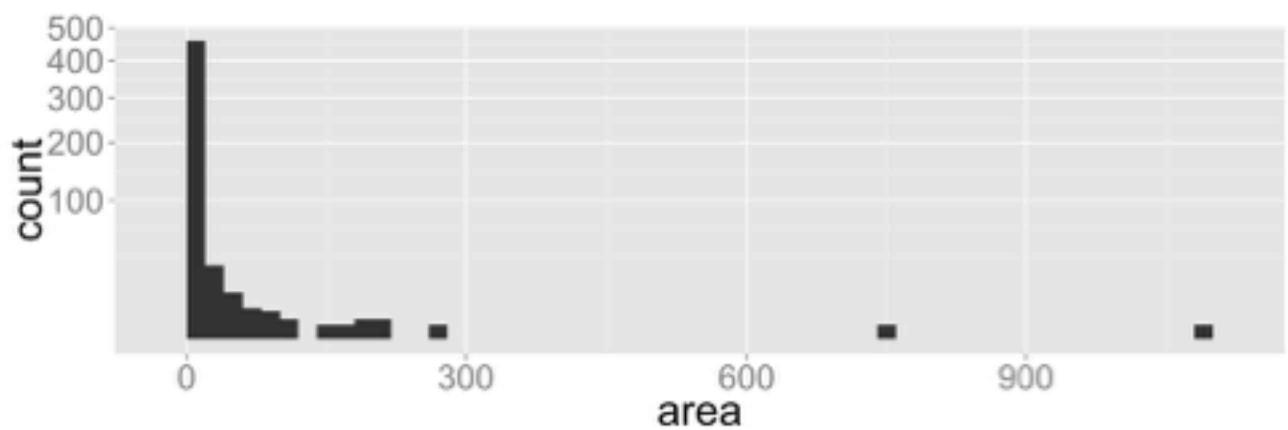
$$y = 54.11\ \mathrm{SSR} - 10.56\ \mathrm{CC} + 18.62$$



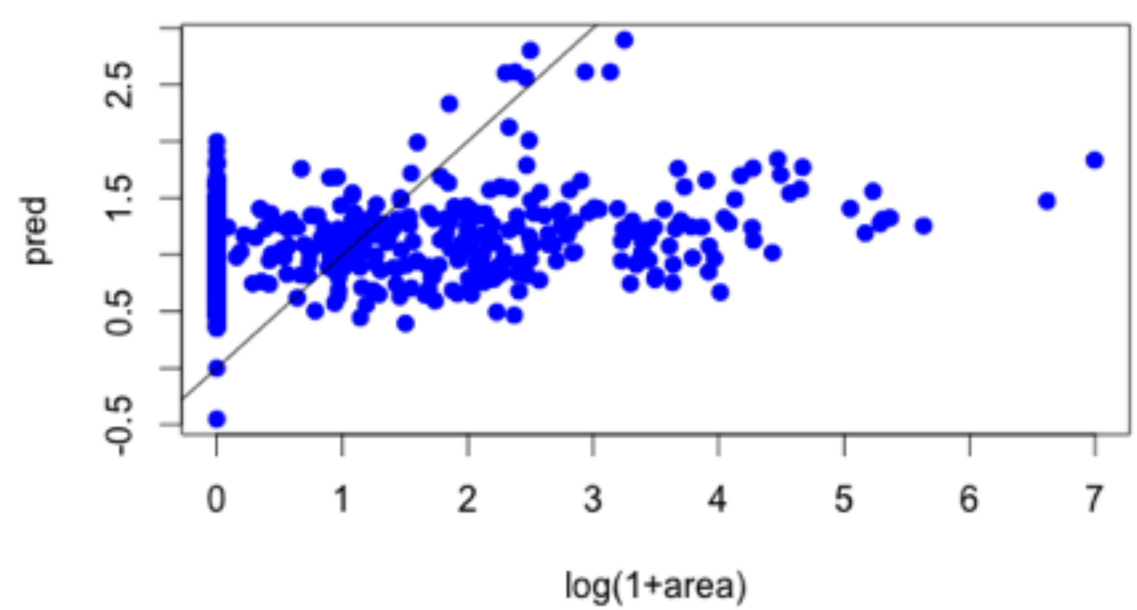
- Correlation of 0.84
- Can we trust this model?

Example

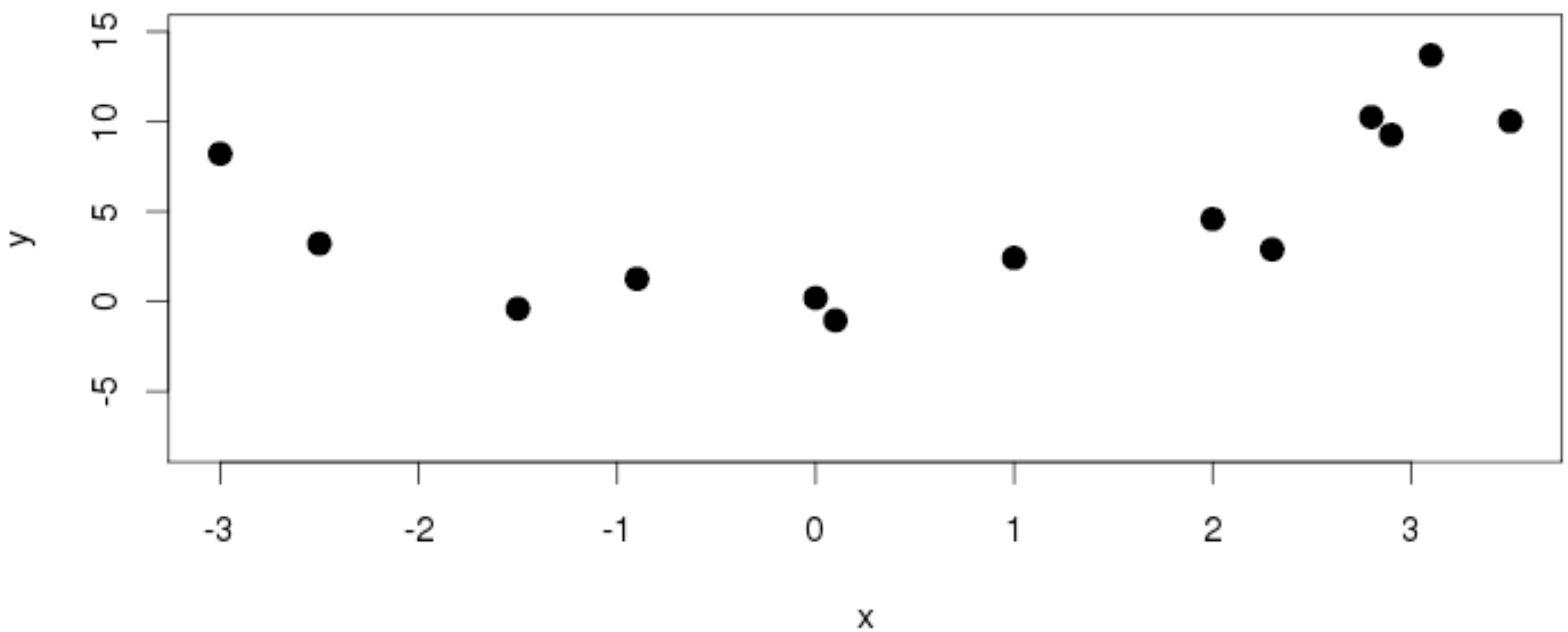
- Forest fires
- Predictand (area) is dramatically skewed (trying with log transformation)

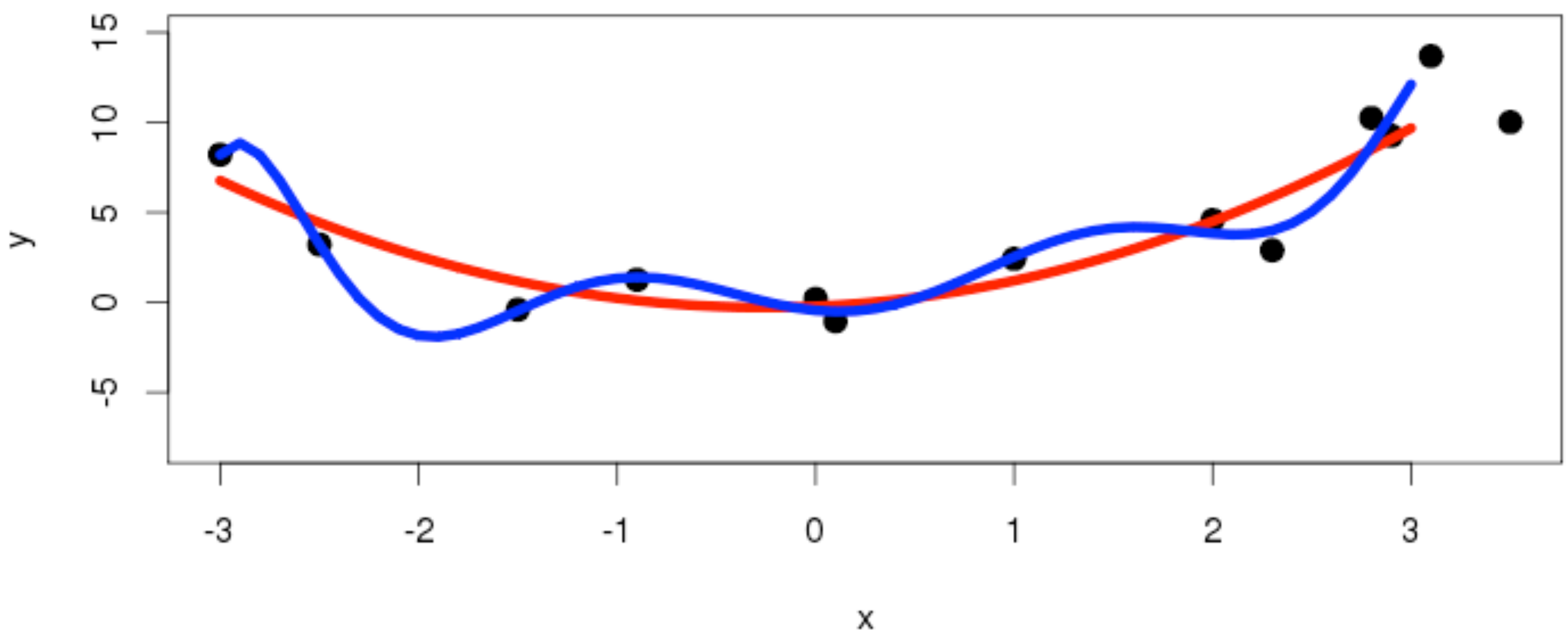


- Poor results with a linear model



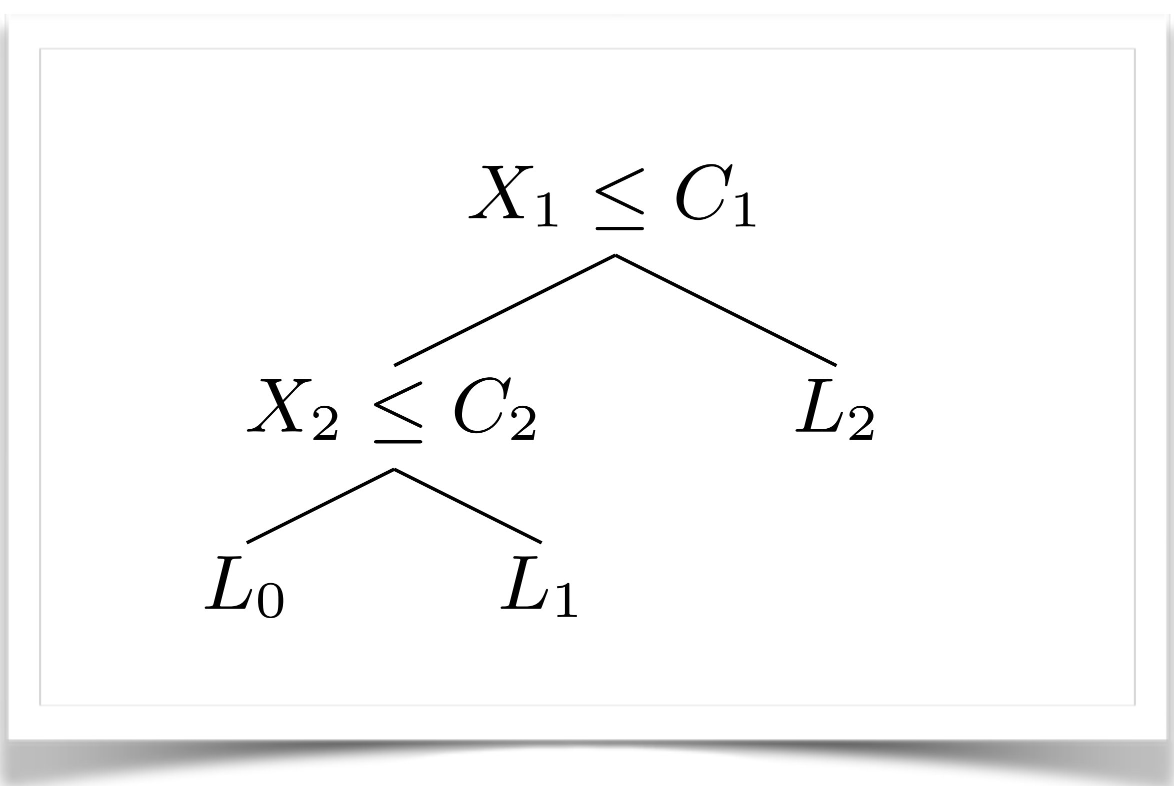
Overfitting

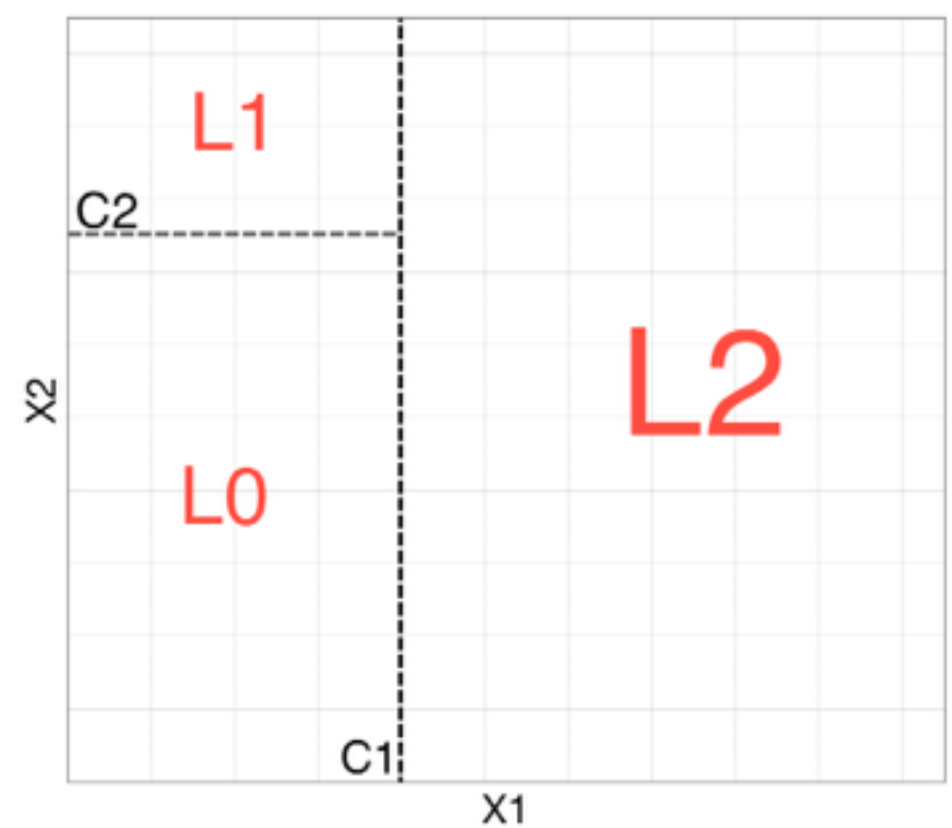




Trees

- Divide-and-conquer approach
- Decision/regression trees





Example

- Again with forest fires
- Converting continuos variable area into factor as:
  - ‘no’: area = 0
  - ‘small’: 0 < area < 5
  - ‘medium’: 5 < area < 50
  - ‘large’: area > 50

```
month = dec: medium (9)
month in {jan,mar,may,nov,oct}:
:...wind <= 8: no (72/23)
:   wind > 8: small (2)
month in {apr,aug,feb,jul,jun,sep}:
:...month = apr:
    :...temp <= 11.8: small (4/2)
    :   temp > 11.8: no (5/1)
    month = jun:
    :...temp > 23.8: medium (4/2)
    :   temp <= 23.8:
    :   :...wind > 4.9: small (3)
    :       wind <= 4.9:
    :       :...temp <= 14.8: small (3/1)
    :           temp > 14.8: no (7)
    month = feb:
    :...temp > 12.4: no (6)
    :   temp <= 12.4:
    :   :...wind > 4.5: medium (5/2)
    :       wind <= 4.5:
    :       :...wind > 2.2: no (3/1)
    :           wind <= 2.2:
[ ... ]
```

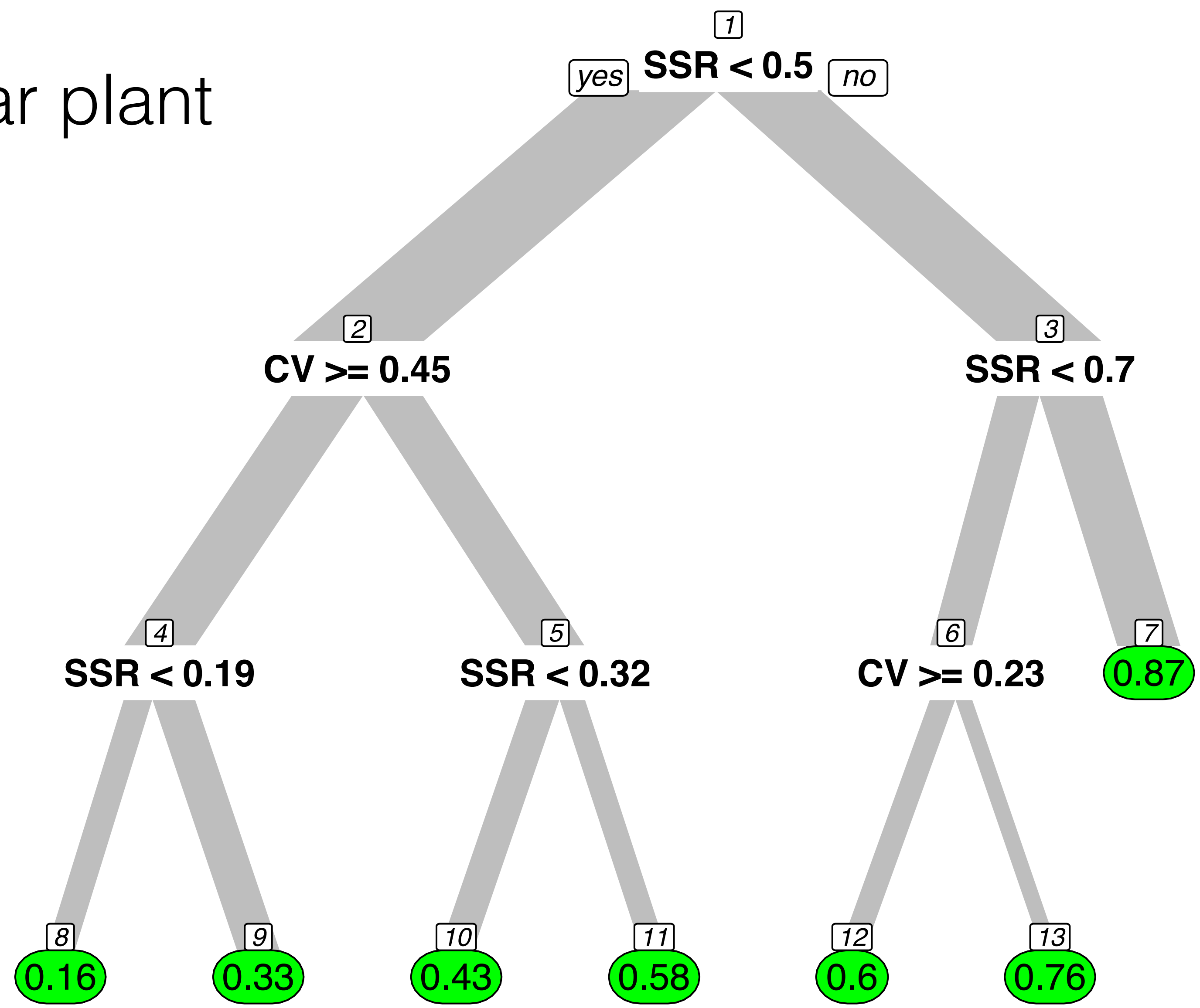
Error: 33.8%

	no	small	med	large
no	214	27	6	0
small	37	75	7	0
med	60	16	51	0
large	13	5	4	2

(Worst results in cross-validation!)

Example

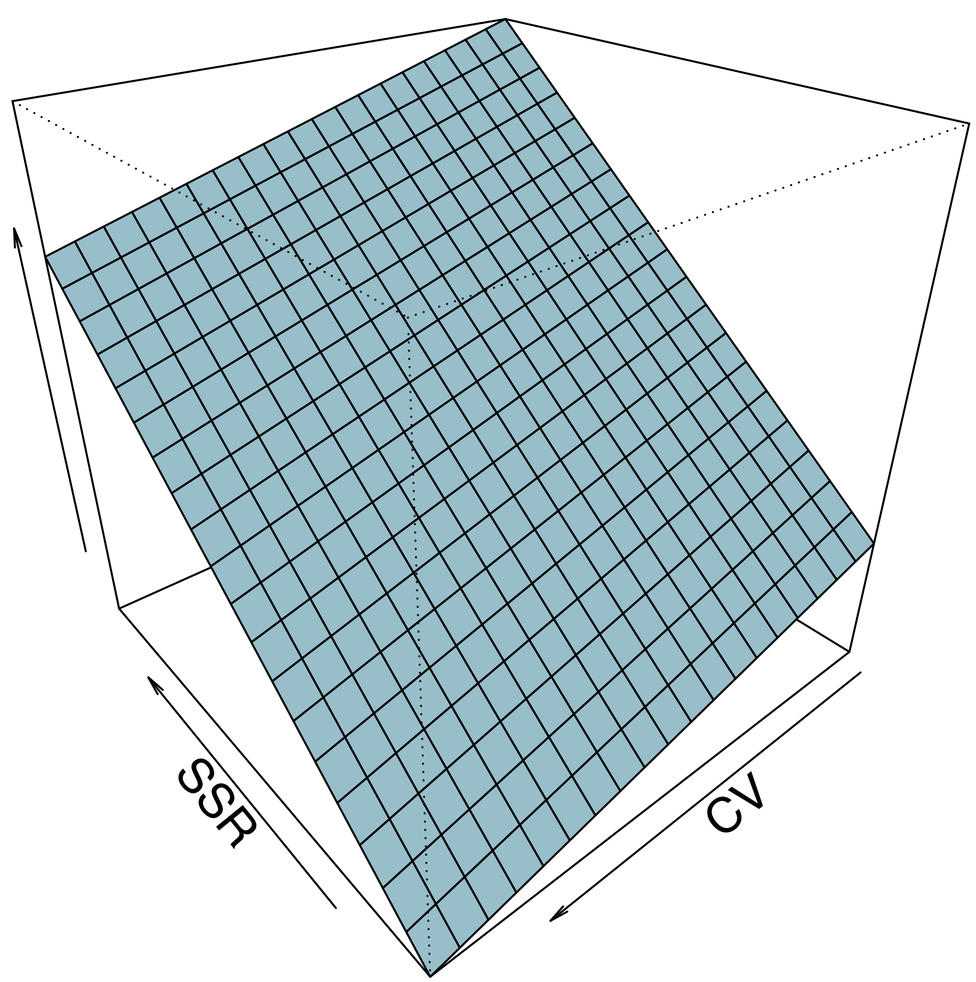
- Data from solar plant



Linear vs Tree

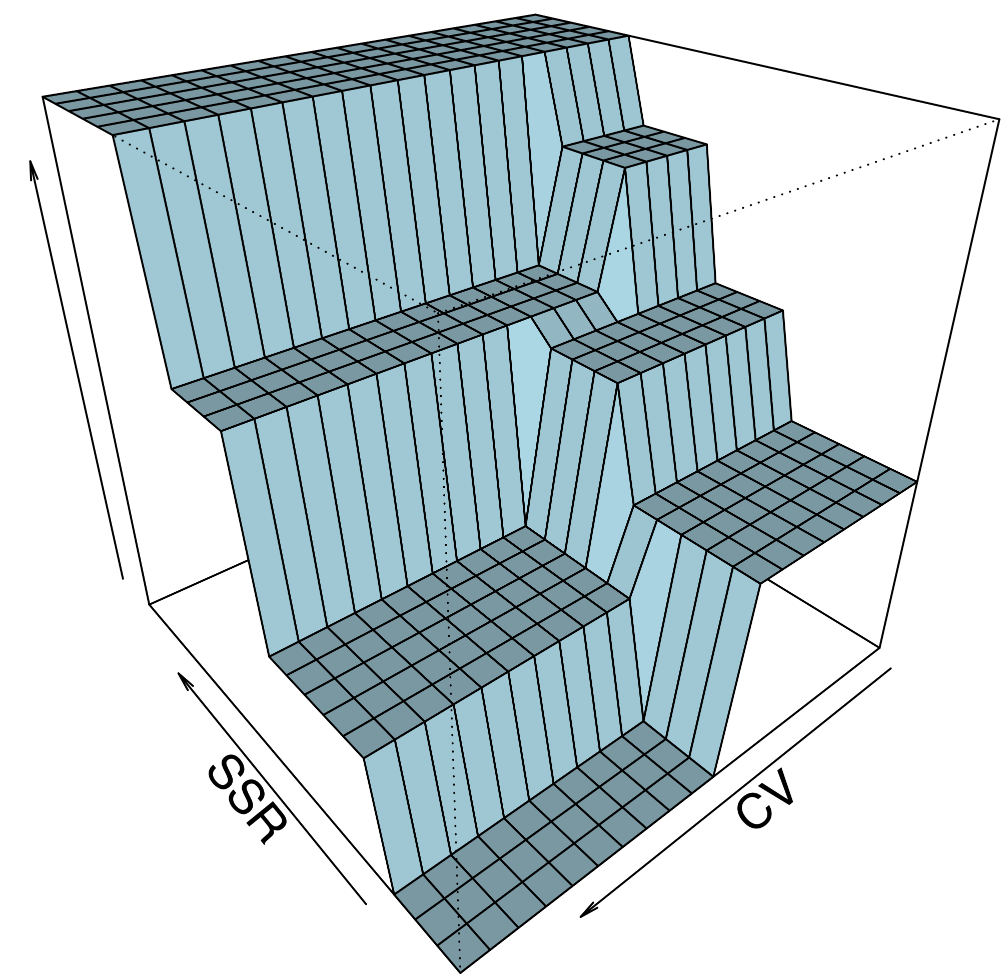
lm(formula=output~SSR+CV,data=df)

SSR: CV

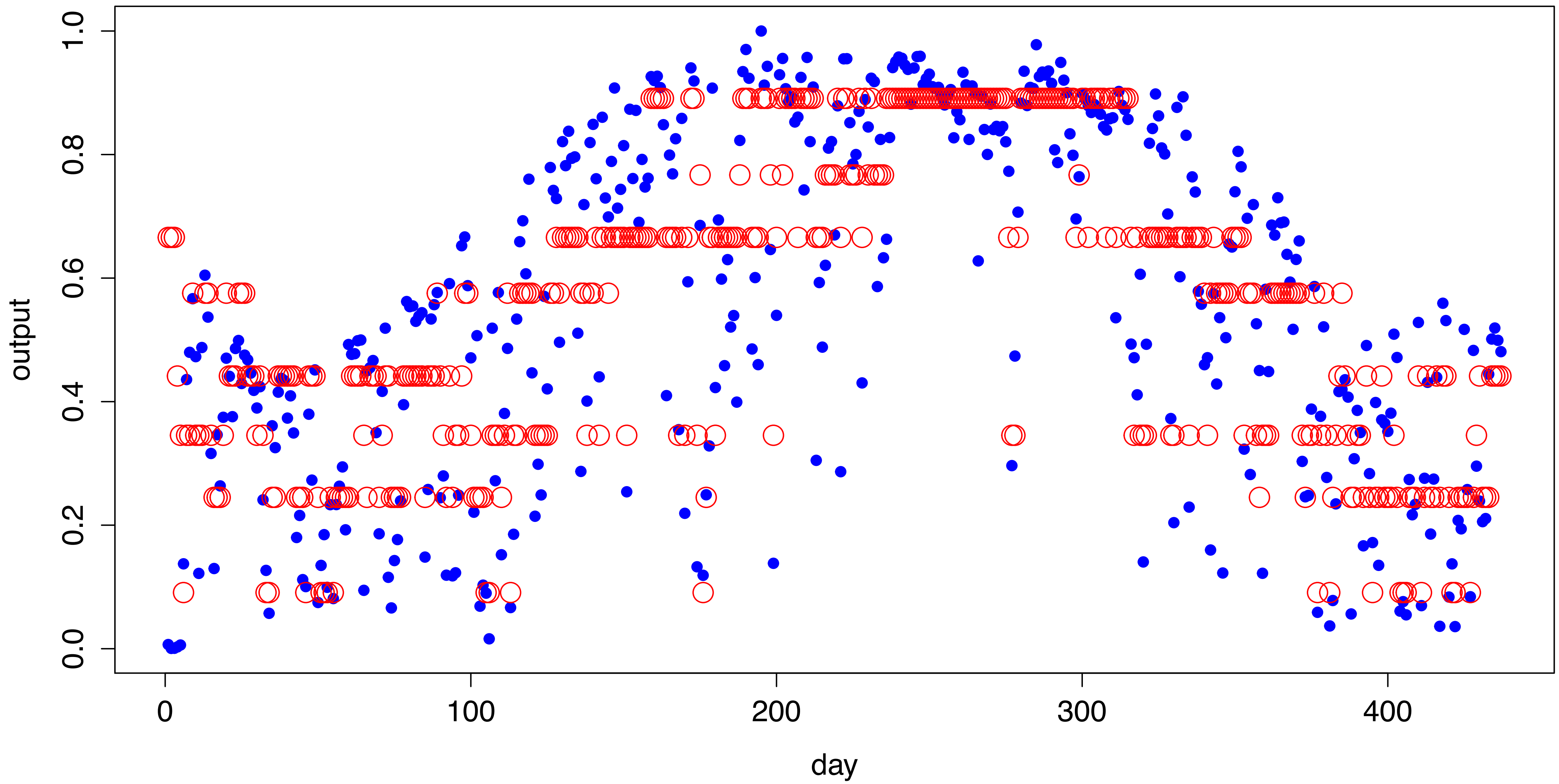


vector rpart(formula=output~SSR+CV,data=df)

SSR: CV

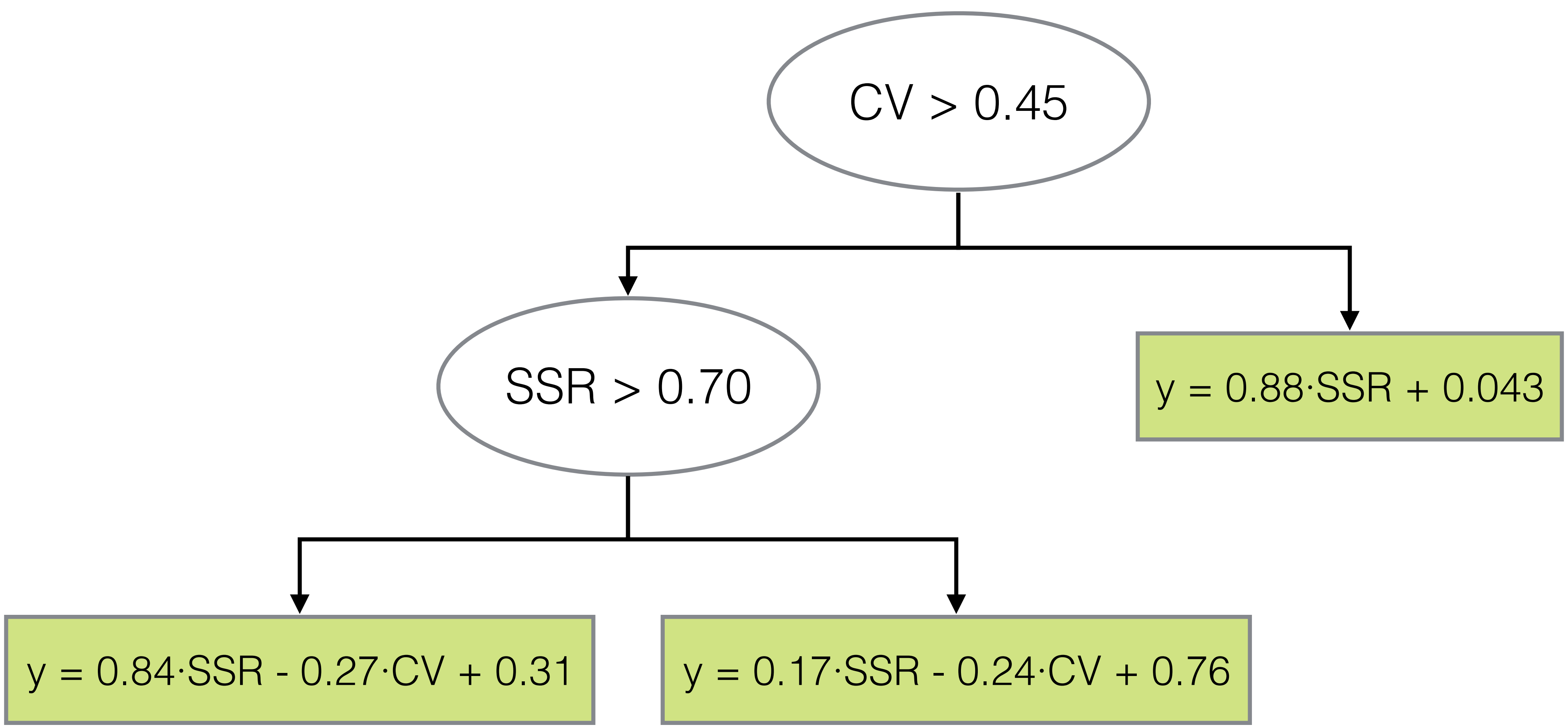


Example

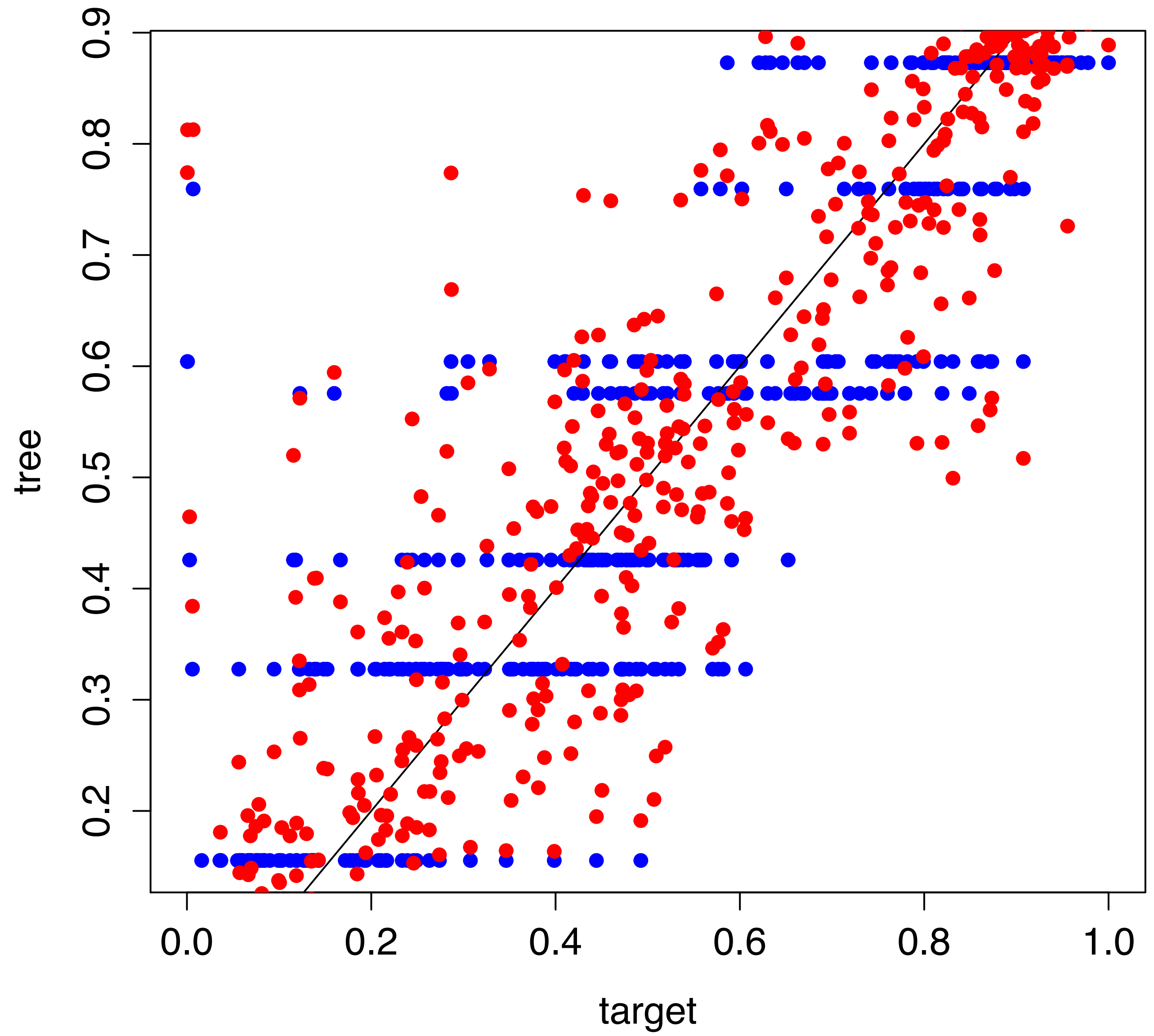


Using model trees

- Linear models as leaves instead of constant values



Using model trees

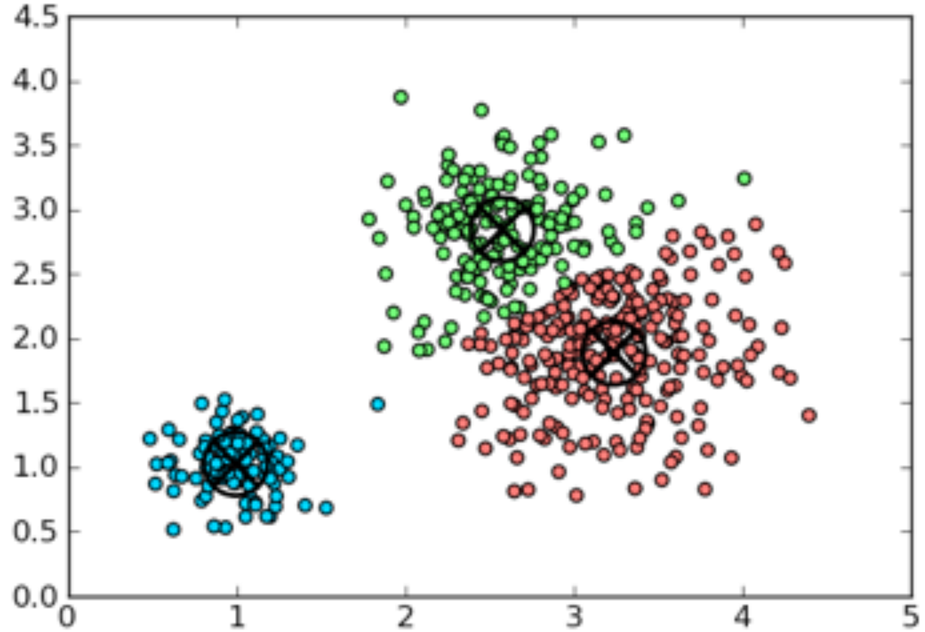


Regression tree and **model tree**

Clustering

- Groups objects by “similarity”
- Many algorithms: most common and simplest centroid-based algorithms like K-Means
- Similarity is often measured with Eucledian distance

$$d = \sqrt{(x_1^1 - x_1^2)^2 + (x_2^1 - x_2^2)^2 + \cdots + (x_k^1 - x_k^2)^2}$$

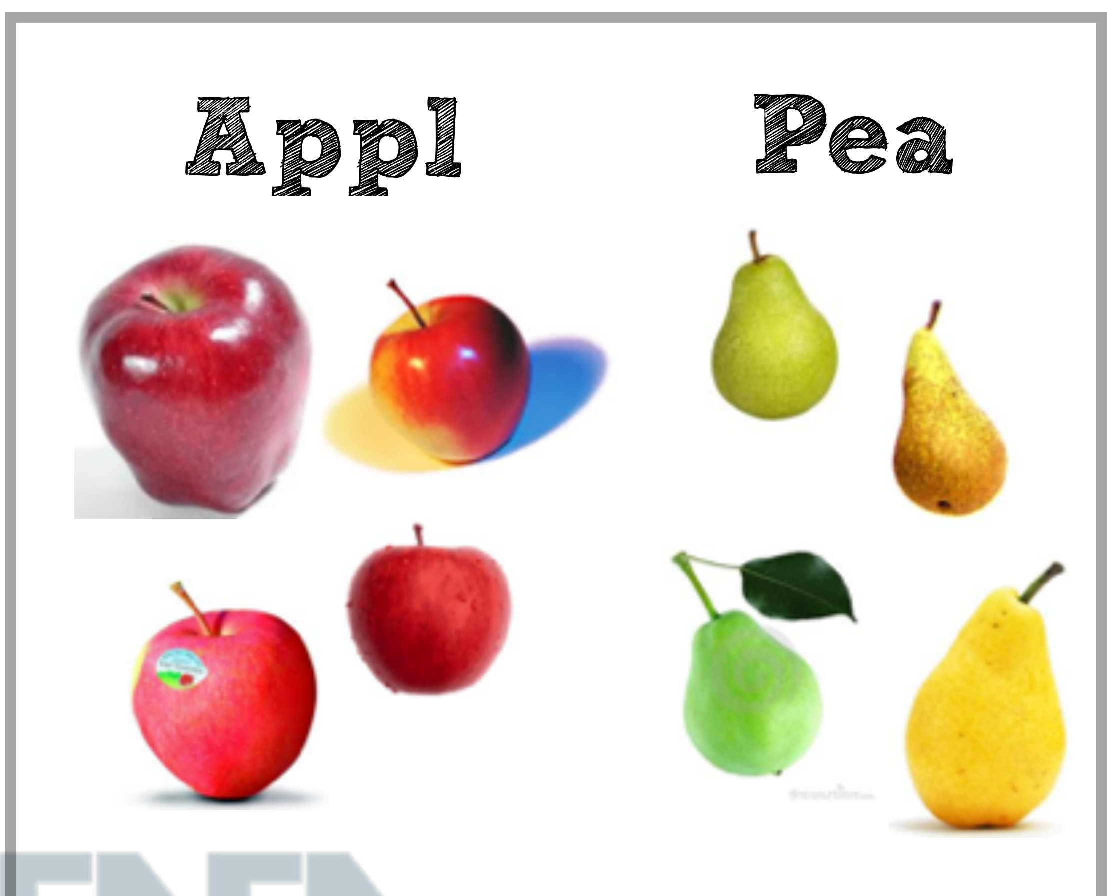


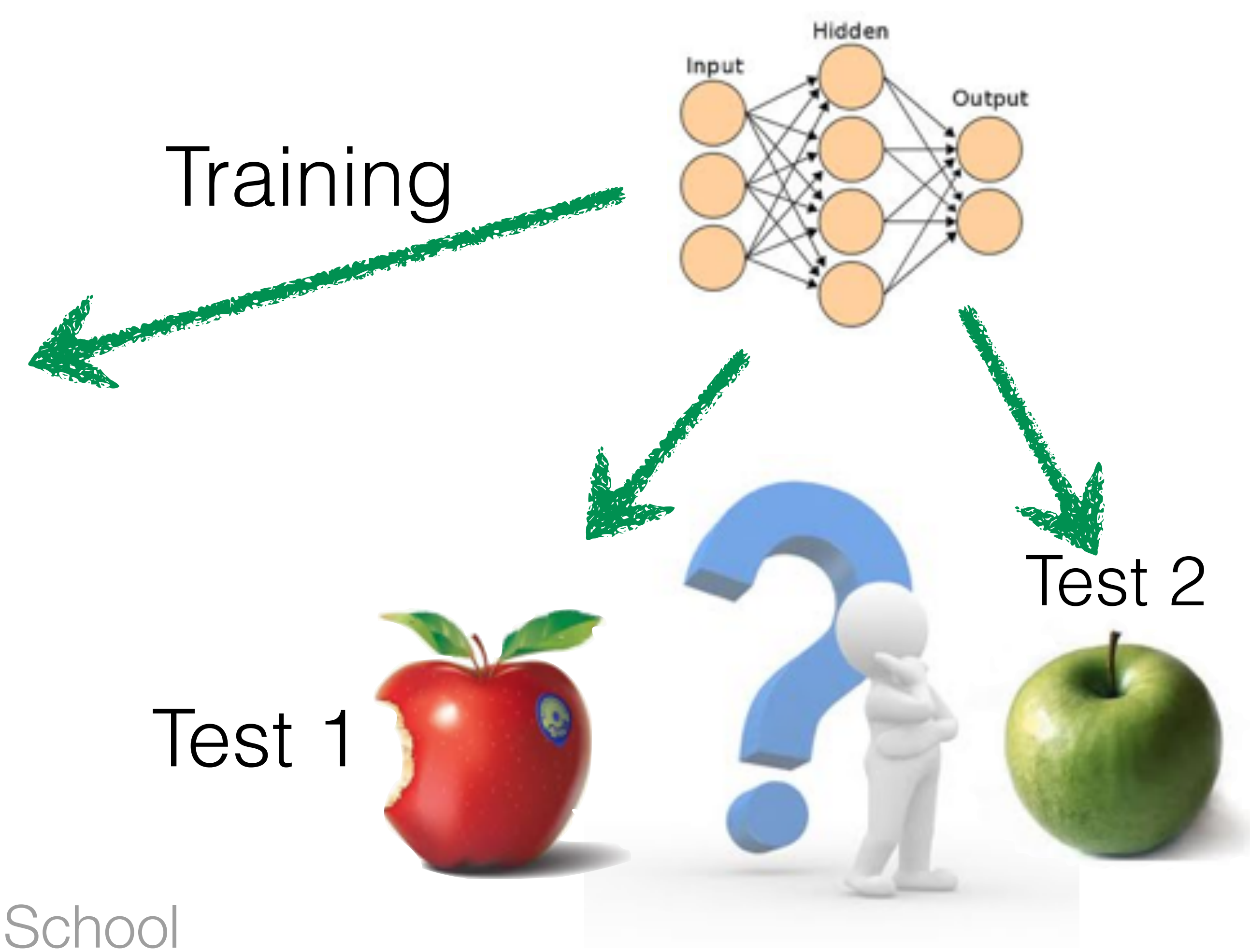
Neural Networks

- Most famous machine learning tool

Pattern recognition

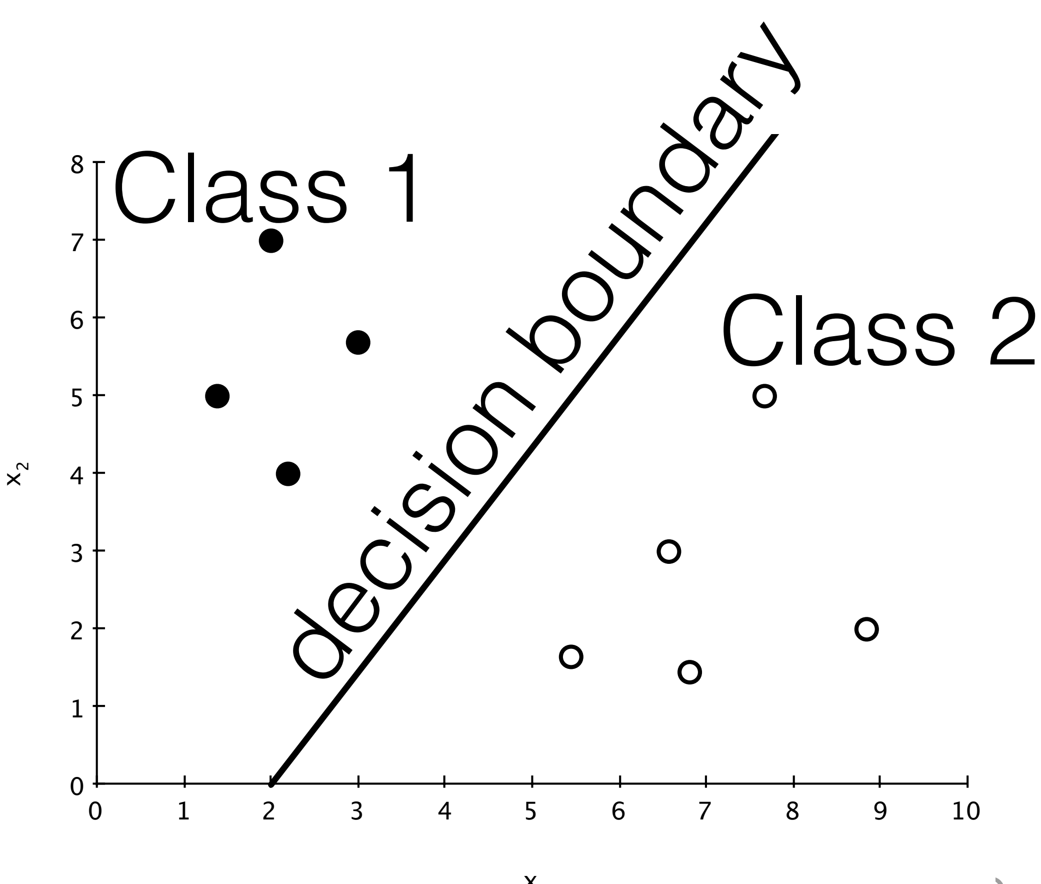
- Traditional machine learning task
- Network ‘learns’ observing input-output pairs
- Generalisation on original (and never observed) data (!)





Perceptron

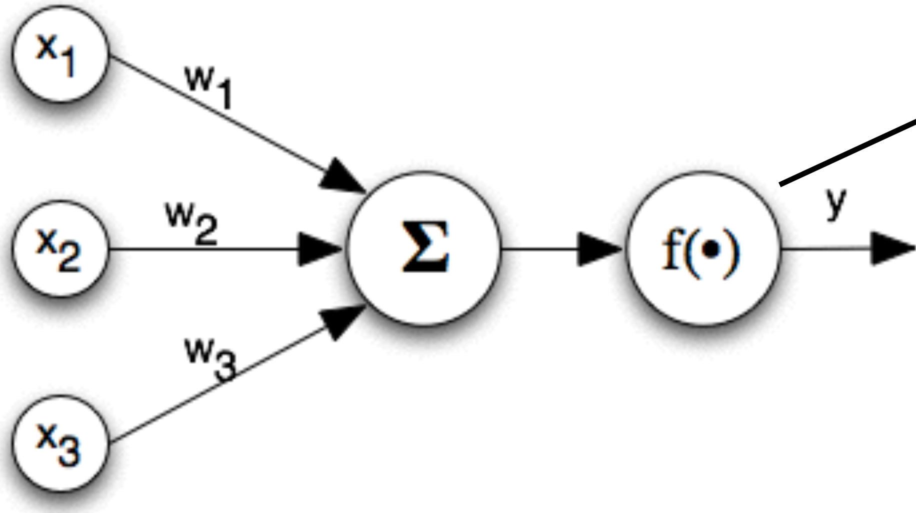
- Simplest neural network
- Used for binary classification of linearly separable data



N-dimensional case: classes separable by an hyperplane

Perceptron

- Binary classifier: ‘sign’ function as output
- **x** input, **w** weights: function y = wx
- How to find optimal parameters?

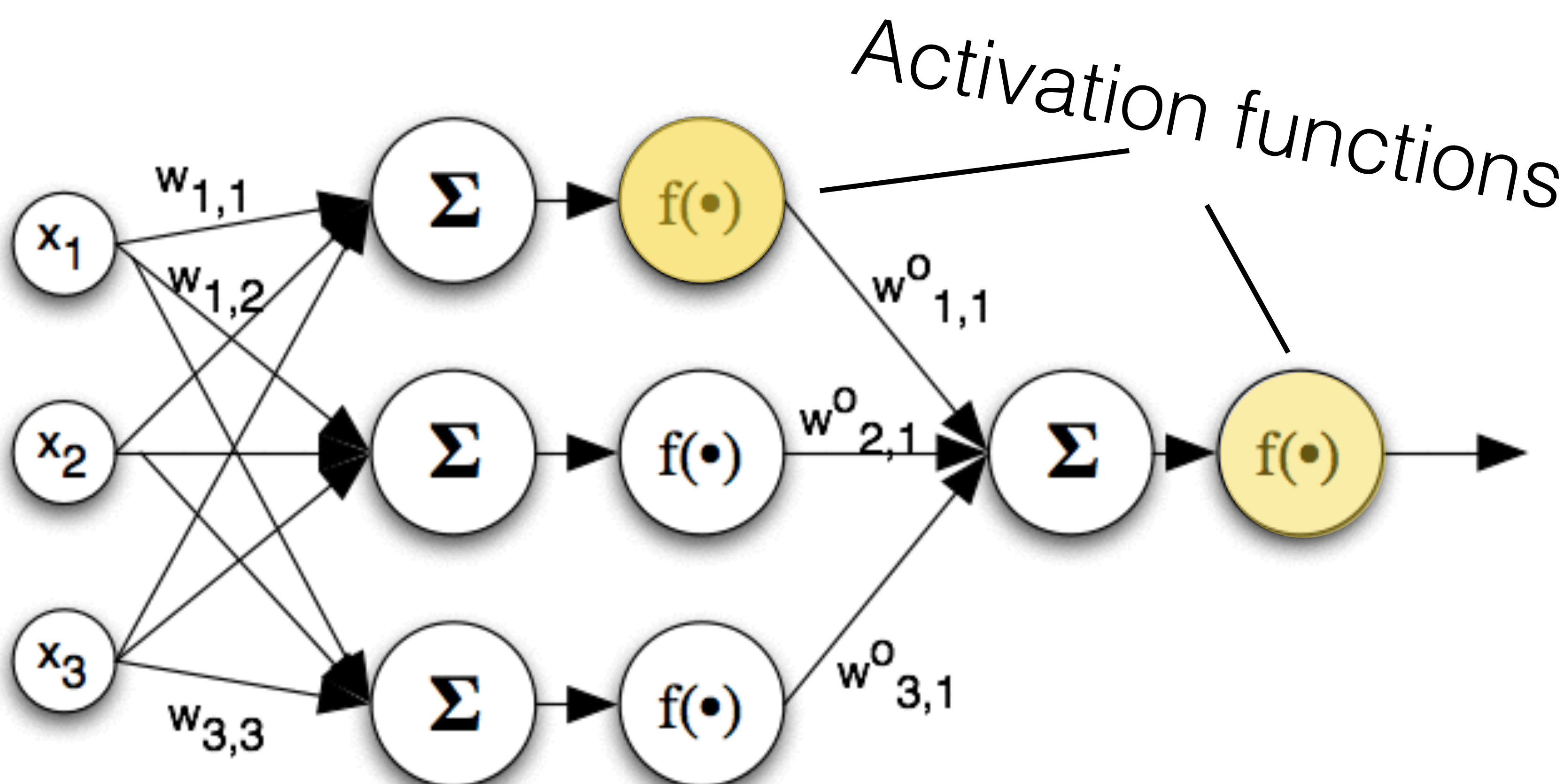


$$f(\mathbf{x}) = \begin{cases} 1 & \text{se } \mathbf{w} \cdot \mathbf{x} > 0 \\ -1 & \text{otherwise} \end{cases}$$

$$y(n) = f(\mathbf{w}^T(n)\,\mathbf{x}(n))$$

Multilayer perceptron (MLP)

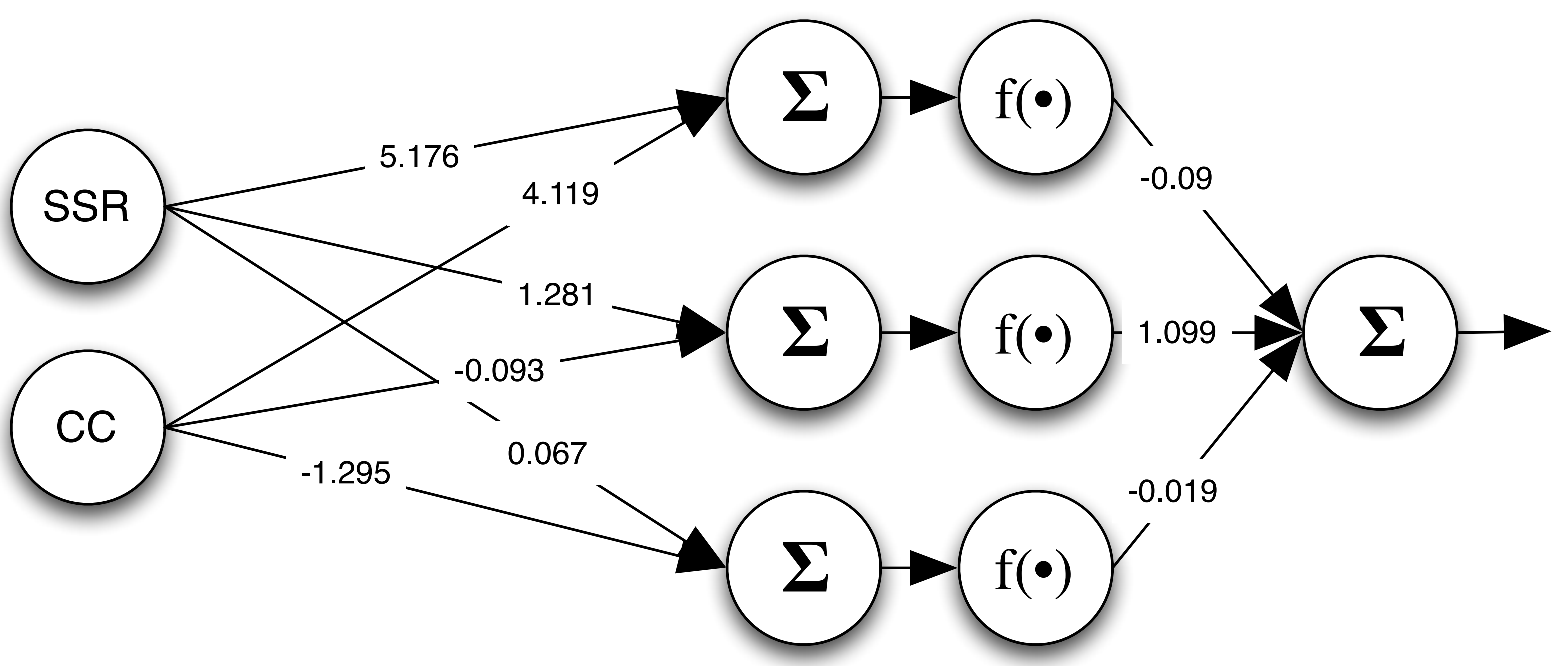
- Neurons with nonlinear differentiable functions
- One or more neurone layers



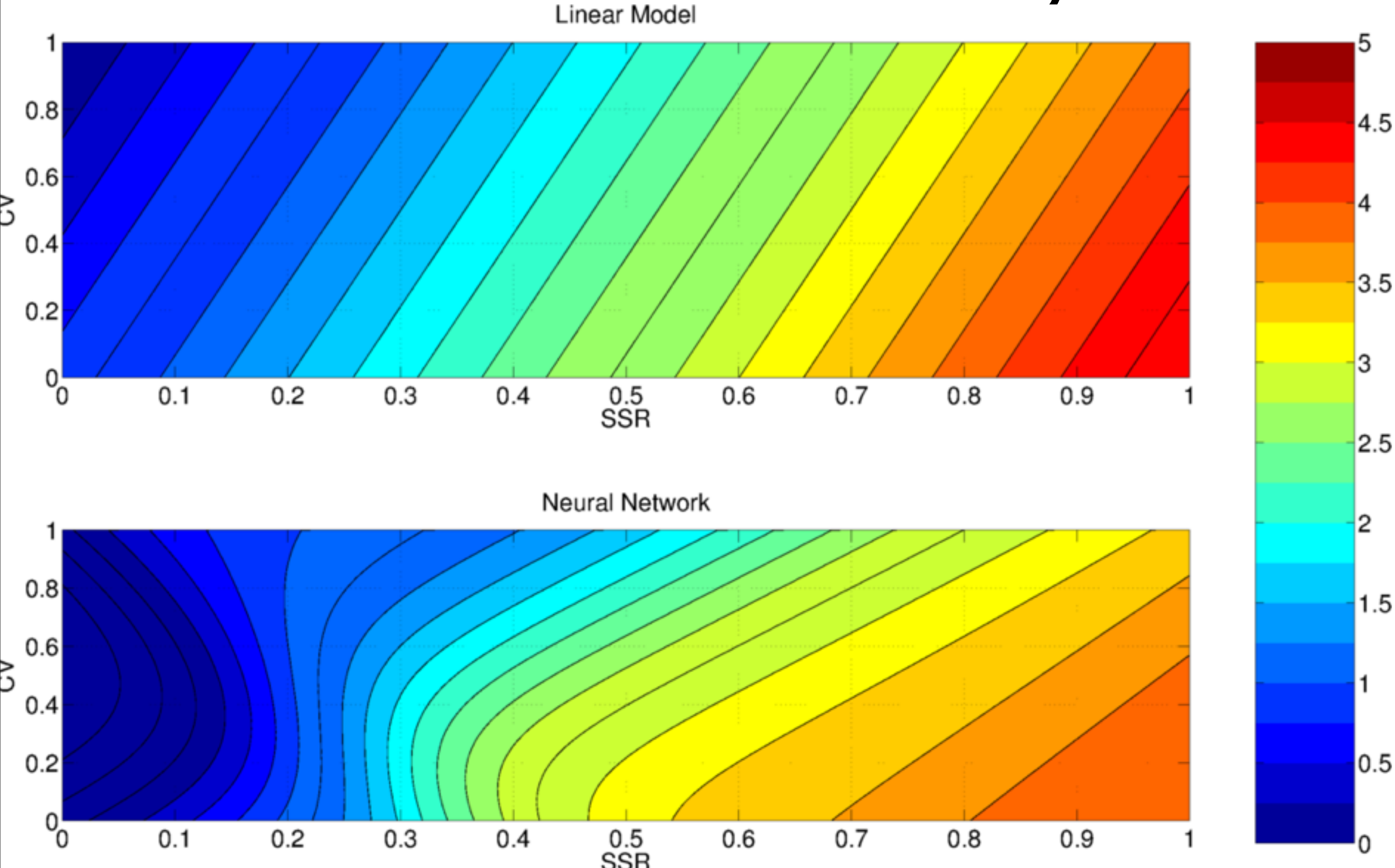
Example

- Neural network vs Linear Regression

$$y = 0.811\ \mathrm{SSR} - 0.156\ \mathrm{CC} + 0.215$$



Non-linearity



Ensemble Learning

- Combining models
- Majority vote is a type of ensemble learning

Simple Methods

- Discrete: Majority vote (or weighted majority vote)
- Continuous: Average (or weighted average)

Pros and Cons

- Easy to implement
- Easy to understand
- Results sometimes hard to analyse

Bagging

"improvements for unstable procedures"

- **B**ootstrap **Agg**regat**ing** [Breiman, 1996]

Training data (size *n*)

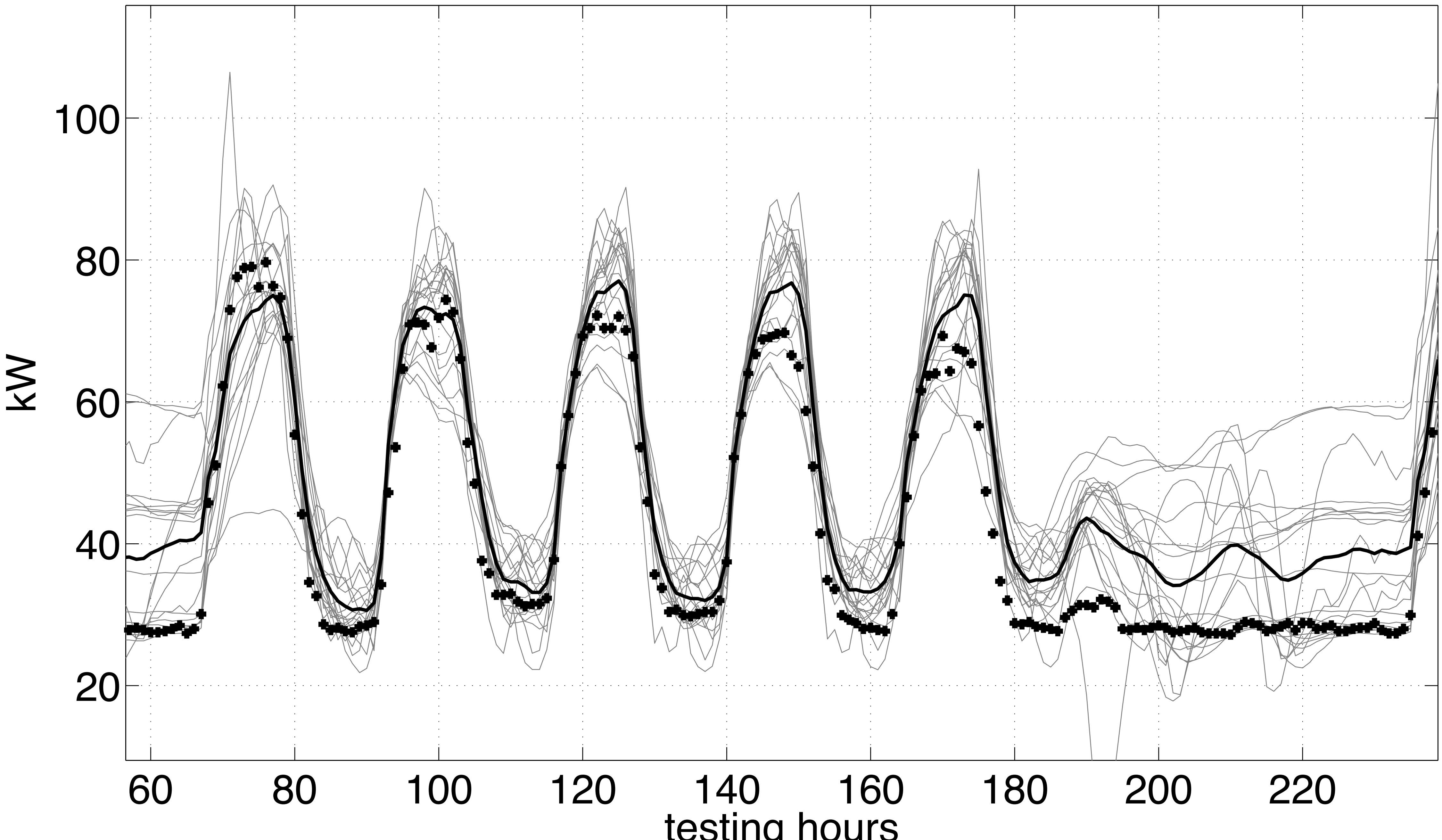
Uniform sampling with replacement

T_1 (size n' < n) | T_2 (size n' < n) | … | T_k (size n' < n)

Average

- Introducing randomisation

Ensemble: example



Error measures

- “Goodness” of a model ➠ Number (error)
- Very importance choice
- Different types:
  1. Absolute errors
  2. Percentage errors
  3. Relative errors

Error measures

Absolute

$$\mathrm{MSE} = \frac{1}{N}\sum_i e_i^2$$

$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}}$$

$$\mathrm{MAE} = \frac{1}{N}\sum_i |e_i|$$

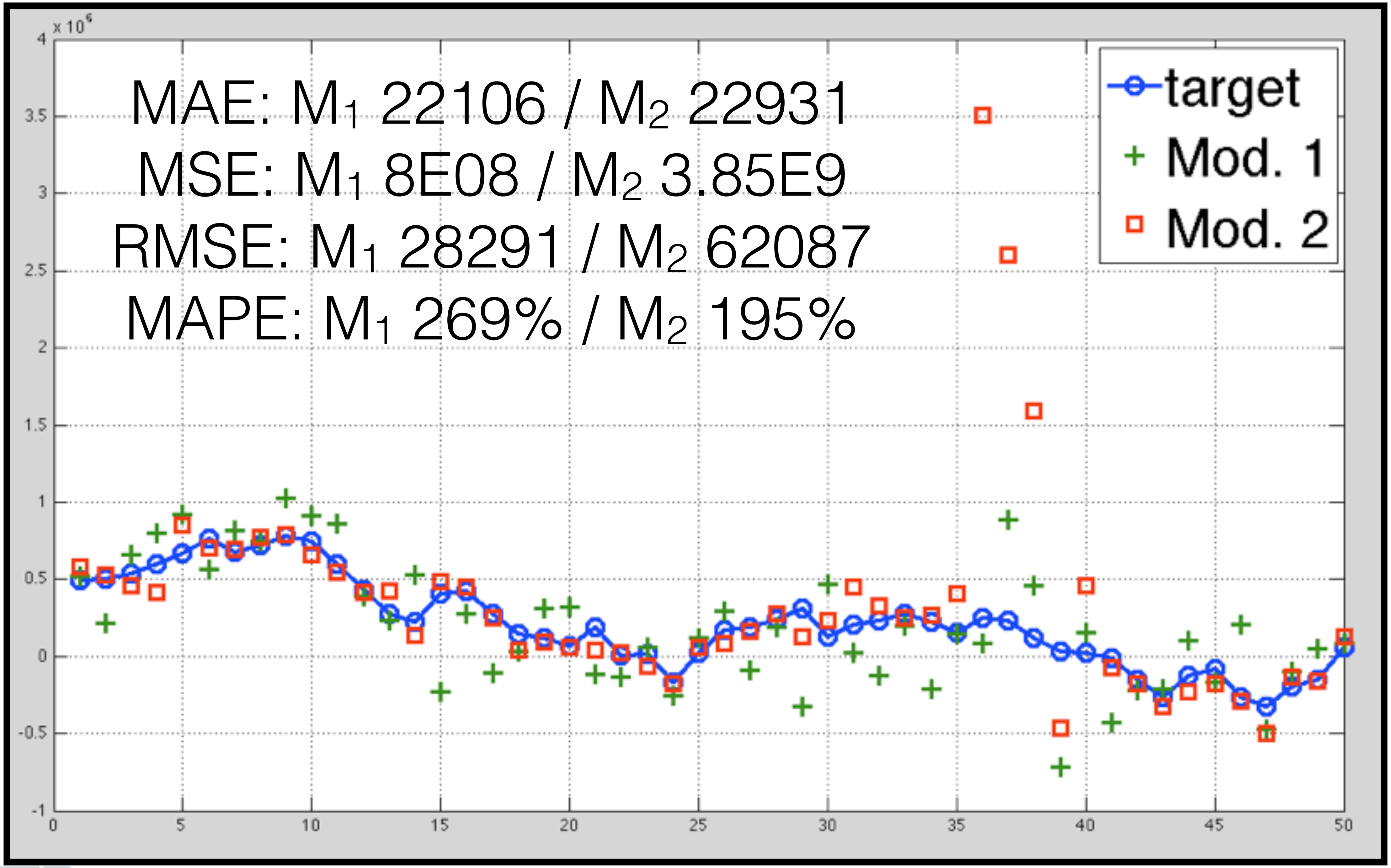
Percentage

$$\mathrm{MAPE} = \frac{1}{N}\sum_i 100\left|\frac{e_i}{y_i}\right|$$

Relative

$$\mathrm{MRAE} = \frac{1}{N}\sum_i \left|\frac{e_i}{e_i^*}\right|$$

Example

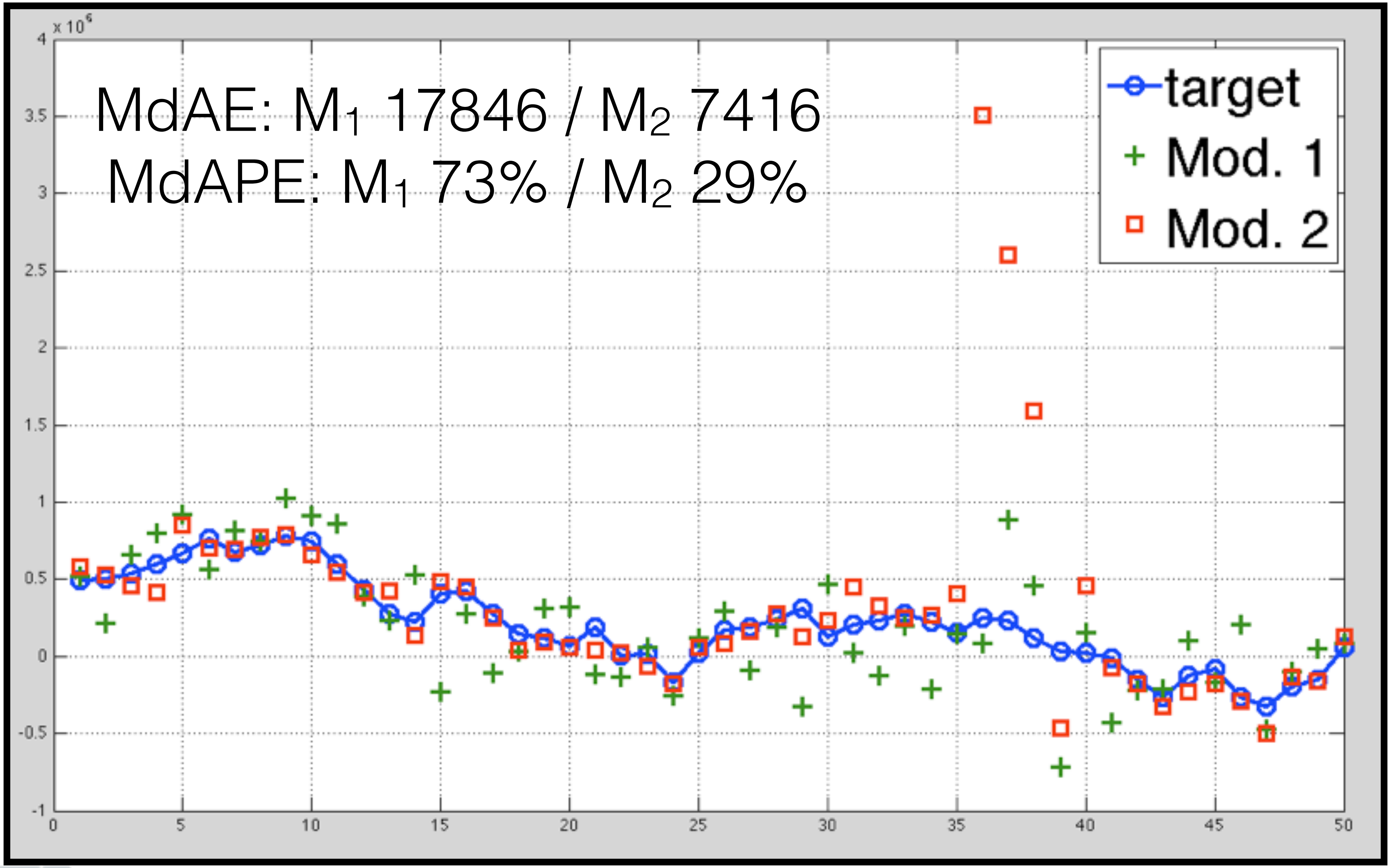


MAE: M_1 22106 / M_2 22931
MSE: M_1 8E08 / M_2 3.85E9
RMSE: M_1 28291 / M_2 62087
MAPE: M_1 269% / M_2 195%
target
Mod. 1
Mod. 2
x 10^6

Good practices

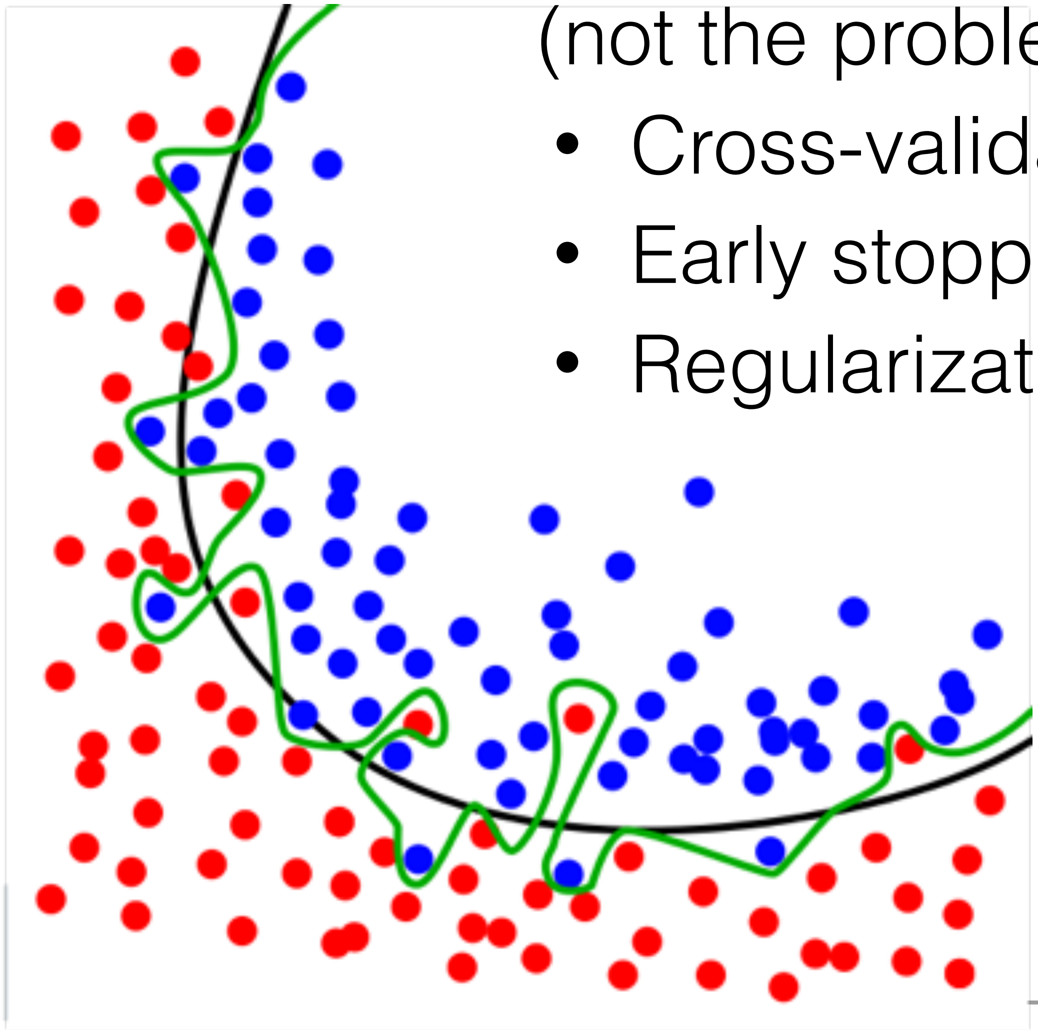
- See your data!
- Look at average AND median (why not quartiles?)
- Ask yoursel “Is it significant?”
- Pay attention to percentage measures mea zero values!

Example



Overfitting

- Machine learning models 'learns' data (not the problem!)
  - Cross-validation
  - Early stopping
  - Regularization



Cross-validation

- Standard method
- Dataset divided in three subsets: training, validation and testing