# TTK4260 - introduction to multivariate data analysis

## Lesson L1

## January 6, 2020

Big Data Cybernetics gang

Welcome to TTK4260 !

# Bigdatacybernetics

- Instructors
  - Harald MARTENS
  - Øivind RIIS
  - Kristin TØNDEL
  - Damiano VARAGNOLO
  - Frank Ove WESTAD
  - Adil RASHEED

Homework: Who is who ?

# Some suggestions, just to start

- Sit together and at the ‘center’ of the first rows
  *(this to interact better)*

- Write your names on some ‘nametags’
  *(this to learn your names – please write them sufficiently big!)*

# Today's lesson

- Practical information about the course
- Why multivariate data analysis
- Overview of the data analysis methods types and traditions

# Practical information about the course

# What is multivariate data analysis?

= data analysis of more than one statistical outcome variable at a time

# What is multivariate data analysis?

= data analysis of more than one statistical outcome variable at a time

Aim: take advantage of the multitude of information sources available now

# Intended learning outcomes *(in brief)*

- Get an in-depth understanding of commonly used Multivariate Data Analysis tools
- Make an informed choice of tools to solve a particular problem
- Understand their relationship to more advanced but black box machine learning methods
- Apply the knowledge gained from the course to solve real life problems *(on realistic data sets)*

# What are our expectations from you?

- Do the flipped classrooms, and thus
  - Register on www.scalable-learning.com
  - Watch the videos before the lessons
  - Check scalable learning before the lessons

# What are our expectations from you?

- Do the flipped classrooms, and thus
  - Register on www.scalable-learning.com
  - Watch the videos before the lessons
  - Check scalable learning before the lessons
- Do the assignments, and thus
  - Make mistakes to learn from it
  - Try different algorithms on the same datasets to appreciate the pros and const of using different algorithms

# What are our expectations from you?

- Do the flipped classrooms, and thus
  - Register on www.scalable-learning.com
  - Watch the videos before the lessons
  - Check scalable learning before the lessons
- Do the assignments, and thus
  - Make mistakes to learn from it
  - Try different algorithms on the same datasets to appreciate the pros and const of using different algorithms
- Prepare for the lectures and test the algorithms by yourself (70% of the work is on these two points)

# What are our expectations from you?

- Do the flipped classrooms, and thus
  - Register on www.scalable-learning.com
  - Watch the videos before the lessons
  - Check scalable learning before the lessons
- Do the assignments, and thus
  - Make mistakes to learn from it
  - Try different algorithms on the same datasets to appreciate the pros and const of using different algorithms
- Prepare for the lectures and test the algorithms by yourself (70% of the work is on these two points)
- Reserve some time for interacting with your peers

# What are our expectations from you?

- Do the flipped classrooms, and thus
  - Register on www.scalable-learning.com
  - Watch the videos before the lessons
  - Check scalable learning before the lessons
- Do the assignments, and thus
  - Make mistakes to learn from it
  - Try different algorithms on the same datasets to appreciate the pros and const of using different algorithms
- Prepare for the lectures and test the algorithms by yourself (70% of the work is on these two points)
- Reserve some time for interacting with your peers
- Make sure to understand everything

# What do we expect you to do?

- Provide feedback in real time (preferably directly but in a nice way without hampering the progress in the class)
- Send the feedback directly to the lecturer (preferably as soon as possible so that the following lectures can be adapted on the fly)

# What do you expect from us?

(think and discuss this 1 minute also with your peers)

# Schedule of the lectures

https://www.ntnu.no/studier/emner/TTK4260#tab=timeplan

# Overarching division of the course

- Part 1: Refreshing the background knowledge
- Part 2: Detailed analysis of some core algorithms / problems
- Part 3: Overview and demonstration of the capabilities / working strategies of other relevant algorithms

# Part 1: refreshing the background knowledge

- L1: Motivations, underlying points of view, overview of the course, and notation
- L2: Use cases through analysing datasets and problems relative to the marine and medical domains
- L3: Use cases through analysing datasets and problems relative to the process industry and robotics domains
- L4: Least squares, in all its interpretations and flavors
- L5: Maximum Likelihood and Maximum A Posteriori
- L6: Statistical performance indexes
- L7: Bias vs variance tradeoff, and model validation algorithms
- L8: Assessment

# Part 2: Detailed analysis of some core algorithms / problems

L9: Design of experiments

L10-11: Principal components analysis

L12: Model order selection

L13: Independent component analysis

L14-16: Principal component regression and partial least squares

L17: Metamodelling

L18: Summary of the various types of plots and their meanings

L19: Multiblock algorithms

L20: PARAFAC algorithms

L21-22: Time series modelling, identification, and prediction

L23: Assessment

# Part 3: overview and demonstration of the capabilities / working strategies of other relevant algorithms

L24: Change detection and subspace identification

L25: IDLE and informative converse methods

L26: Neural networks and random forests

L27: Support vector machines and t-SNE

L28: Other clustering and classification methods

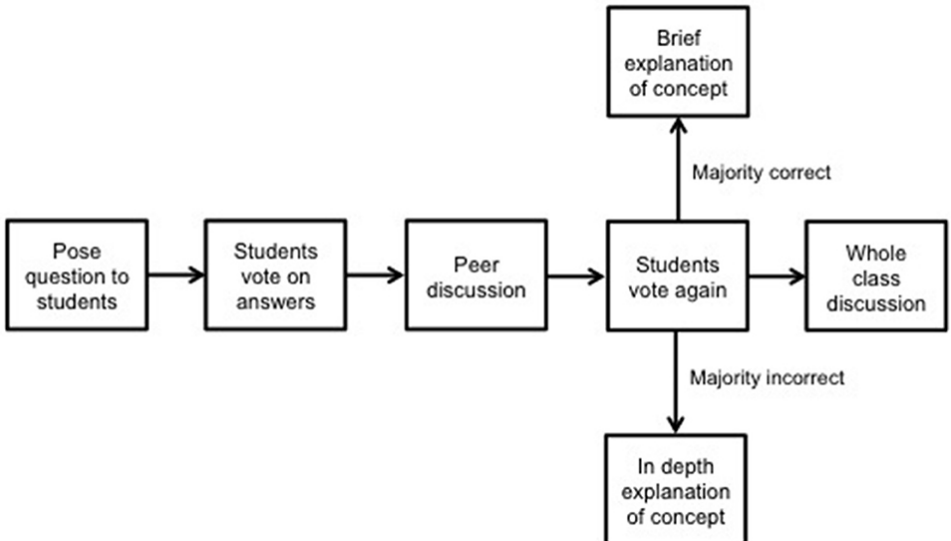
# Peer instructions - example

# Peer instructions

**Why?**

- discussions among and with students
- sharpen focus and understanding

# Peer instructions: working



example: which of these are reasons for the seasons?

go to www.kahoot.it

# Scalable Learning

# Introduction to Scalable-Learning

- Register on `www.scalable-learning.com`
- Watch the videos before the lessons
- Check scalable learning before the lessons

Any data put in is fully annonymized

# Scalable learning - how to enroll in the portal

1. Register an account in `https://www.scalable-learning.com/#/users/login` using your institutional email (i.e., `xxx@ntnu.no`)
2. Go to `https://www.scalable-learning.com/#/courses/enroll?id=GUKXE-22315` and follow the instructions
3. In case the portal asks, use `GUKXE-22315` as the course enrollment key

direct link to the course:
`https://www.scalable-learning.com/#/courses/4363/course_information`

# Exams and evaluation

Written exams: tentative date, assignments

# Data to work with

- A selected set of data will be shared in the Lecture 2
- Students are expected to search for more relevant dataset on websites like Kaggle
- Bring a laptop with Matlab, Python and Unscrambler installed to preprocess the data for the subsequent lectures

# To do list (before Lecture 2)

- Install Matlab (Damiano)
  `https://innsida.ntnu.no/wiki/-/wiki/English/Matlab+for+students`
- Install Python (Adil)
  `https://www.anaconda.com/distribution/`
- Install Unscrambler (Frank)

# Evaluation of the prerequisites

*Today @ 14.15 in R9, realfag*

What are you supposed to do?
→ do module *"assessment of the prerequisites"* in Scalable-Learning

direct link: `https://www.scalable-learning.com/#/courses/4363/modules/17002/courseware/lectures/53828`

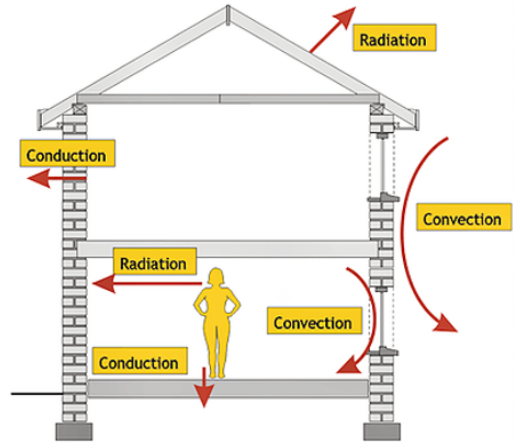
# Expectation from today's lecture

- Why data driven models when first principle mechanistic models are available in abundance?
- Why to take a course in Multivariate Data Analysis when so many courses on ML are available?
- What makes this course different from other courses on similar topics?
- Some taste of Big Data Cybernetics
- Get to know each other and encourage to prepare our own data
- Install the required softwares
- Get familiar with the content and schedule of the course

# Overview

- Mechanistic Modeling
- Data-driven modeling (Deep Learning)
- Big Data Cybernetics
- Preparing the grounds for Multivariate Data Analysis

# Mechanistic modeling approach

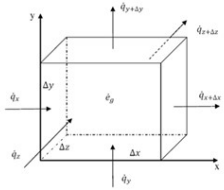


# Modeling

$$\dot{q}_{x+dx} = \dot{q}_x + \frac{\partial \dot{q}_x}{\partial x}\,dx$$

or,

$$\dot{q}_x - \dot{q}_{x+dx} = -\frac{\partial \dot{q}_x}{\partial x}\,dx$$




## Fourier's law of heat conduction

$$\dot{q}_x = -k\,dy\,dz\,\frac{\partial T}{\partial x}$$

$$\dot{q}_x - \dot{q}_{x+dx} = \frac{\partial}{\partial x}\left(k\,\frac{\partial T}{\partial x}\right)dx\,dy\,dz$$

$$\dot{q}_y - \dot{q}_{y+dy} = \frac{\partial}{\partial y}\left(k\,\frac{\partial T}{\partial y}\right)dx\,dy\,dz$$

and,

$$\dot{q}_z - \dot{q}_{z+dz} = \frac{\partial}{\partial z}\left(k\,\frac{\partial T}{\partial z}\right)dx\,dy\,dz$$

*Rate of heat input + Rate of heat generation*

*= Rate of heat output*

*+ Rate of change of heat energy within the body*

$$\dot{u} = (\rho dx\,dy\,dz)c_p\left(\frac{\partial T}{\partial t}\right)$$

$$(\dot{q}_x + \dot{q}_y + \dot{q}_z) + (\dot{e}_g\,dx\,dy\,dz) = (\dot{q}_{x+dx} + \dot{q}_{y+dy} + \dot{q}_{z+dz}) + (\rho dx\,dy\,dz)c_p\left(\frac{\partial T}{\partial t}\right)$$

$$(\dot{q}_x - \dot{q}_{x+dx}) + (\dot{q}_y - \dot{q}_{y+dy}) + (\dot{q}_z - \dot{q}_{z+dz}) + (\dot{e}_g dx\,dy\,dz) = (\rho dx\,dy\,dz)c_p\left(\frac{\partial T}{\partial t}\right)$$

or,

$$\left\{\frac{\partial}{\partial x}\left(k\,\frac{\partial T}{\partial x}\right) + \frac{\partial}{\partial y}\left(k\,\frac{\partial T}{\partial y}\right) + \frac{\partial}{\partial z}\left(k\,\frac{\partial T}{\partial z}\right)\right\}(dx\,dy\,dz) + (\dot{e}_g\,dx\,dy\,dz) = (\rho dx\,dy\,dz)c_p\left(\frac{\partial T}{\partial t}\right)$$

or,

$$\left\{\frac{\partial}{\partial x}\left(k\,\frac{\partial T}{\partial x}\right) + \frac{\partial}{\partial y}\left(k\,\frac{\partial T}{\partial y}\right) + \frac{\partial}{\partial z}\left(k\,\frac{\partial T}{\partial z}\right)\right\} + \dot{e}_g = \rho c_p\left(\frac{\partial T}{\partial t}\right)$$

# Model simplification

- Reducing dimension (eg.using symmetry)

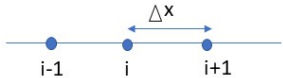
$$\frac{\partial T}{\partial t} = \nu \frac{\partial^2 T}{\partial x^2}$$

- Approximating thermophysical properties
- Approximating shapes
- ....
- ....

# Numerical Discretization

1D diffusion equation

$$\frac{\partial u}{\partial t} = \nu \frac{\partial^2 u}{\partial x^2}$$




$$u_{i+1} = u_i + \Delta x \frac{\partial u}{\partial x}\bigg|_i + \frac{\Delta x^2}{2}\frac{\partial^2 u}{\partial x^2}\bigg|_i + \frac{\Delta x^3}{3!}\frac{\partial^3 u}{\partial x^3}\bigg|_i + O(\Delta x^4)$$

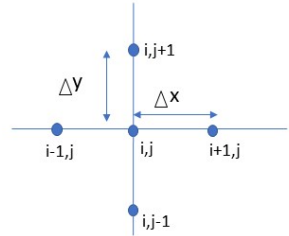
$$u_{i-1} = u_i - \Delta x \frac{\partial u}{\partial x}\bigg|_i + \frac{\Delta x^2}{2}\frac{\partial^2 u}{\partial x^2}\bigg|_i - \frac{\Delta x^3}{3!}\frac{\partial^3 u}{\partial x^3}\bigg|_i + O(\Delta x^4)$$

$$\frac{\partial^2 u}{\partial x^2} = \frac{u_{i+1} - 2u_i + u_{i-1}}{\Delta x^2} + O(\Delta x^2)$$

$$\frac{u_i^{n+1} - u_i^n}{\Delta t} = \nu \frac{u_{i+1}^n - 2u_i^n + u_{i-1}^n}{\Delta x^2}$$

$$u_i^{n+1} = u_i^n + \frac{\nu \Delta t}{\Delta x^2}(u_{i+1}^n - 2u_i^n + u_{i-1}^n)$$

2D diffusion equation

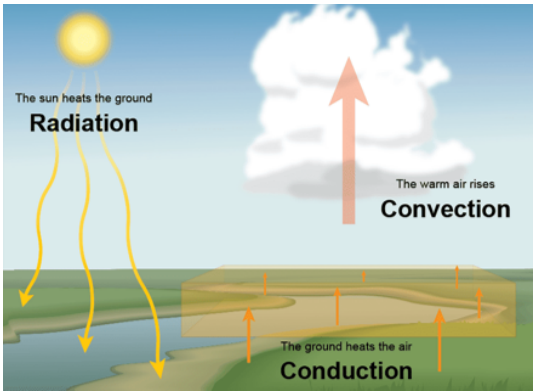


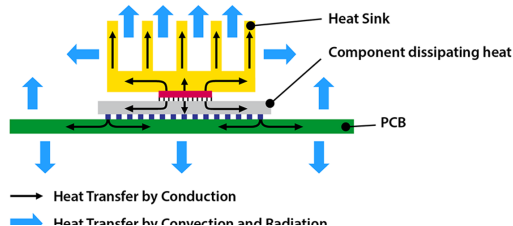

$$\frac{\partial u}{\partial t} = \nu \frac{\partial^2 u}{\partial x^2} + \nu \frac{\partial^2 u}{\partial y^2}$$

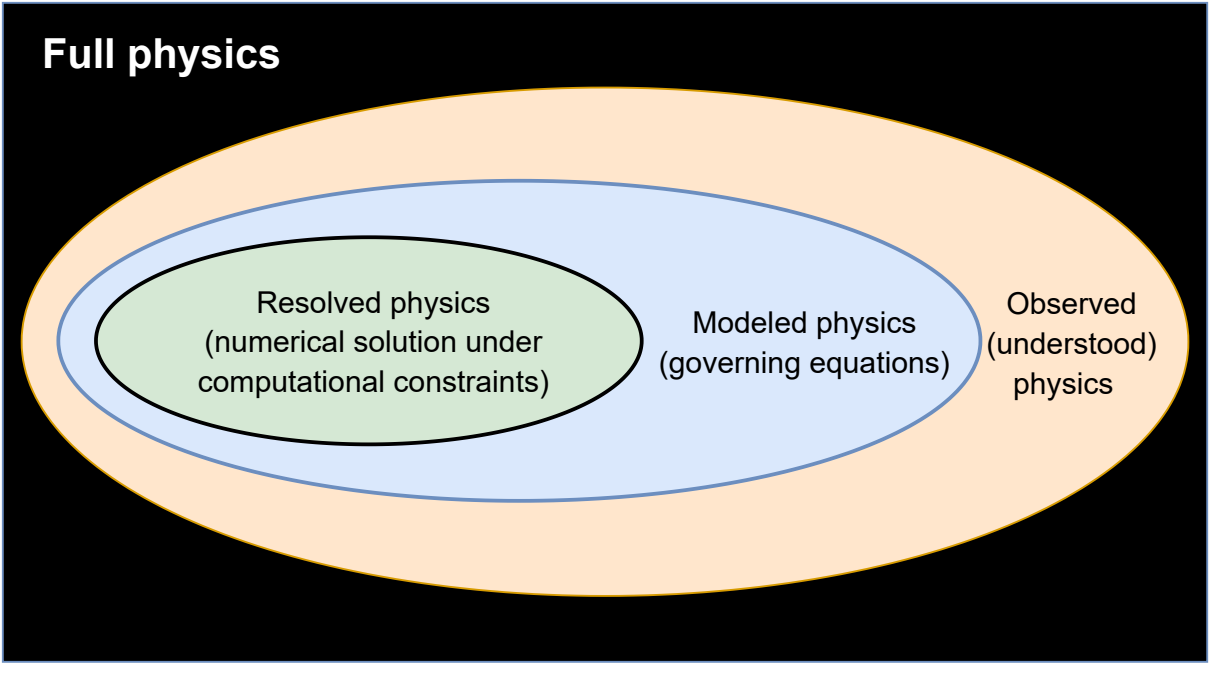
$$\frac{u_{i,j}^{n+1} - u_{i,j}^n}{\Delta t} = \nu \frac{u_{i+1,j}^n - 2u_{i,j}^n + u_{i-1,j}^n}{\Delta x^2} + \nu \frac{u_{i,j+1}^n - 2u_{i,j}^n + u_{i,j-1}^n}{\Delta y^2}$$

$$u_{i,j}^{n+1} = u_{i,j}^n + \frac{\nu \Delta t}{\Delta x^2}(u_{i+1,j}^n - 2u_{i,j}^n + u_{i-1,j}^n) + \frac{\nu \Delta t}{\Delta y^2}(u_{i,j+1}^n - 2u_{i,j}^n + u_{i,j-1}^n)$$

# Generalization to other problems



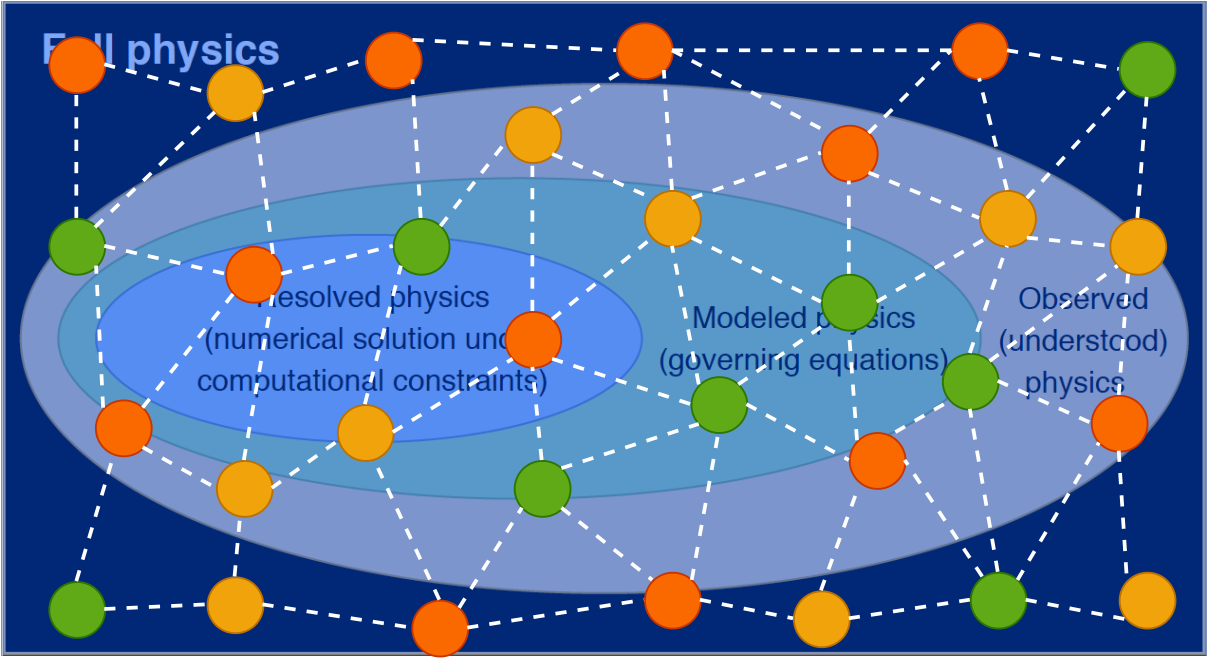




Full physics
Resolved physics
(numerical solution under
computational constraints)
Modeled physics
(governing equations)
Observed
(understood)
physics

# Assignment 1

- Figure out a phenomena governed by the 1D, 2D
- Decide on the initial condition and boundary condition
- Solve it using the numerical methods
- Convince yourself about the generalizability of the mechanistic models

https://github.com/adil-rasheed/TK8117/blob/master/Lecture1/Physicsbased.ipynb


Full physics
Observed
(understood)
physics

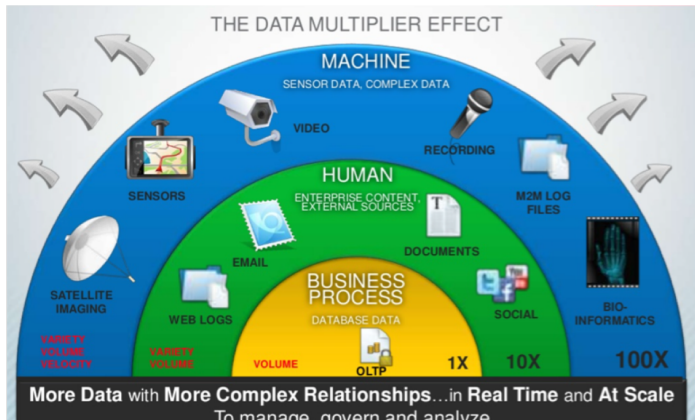
# Technology push and Market pull

- New and existing algorithms
- Digitalization wave
- Cheap sensors and observation data (mobile data, cams)
- Easy-to-use software libraries
- Cheap GPU's and TPUs

- Digital Twin
- Internet of Things
- Smart Cities
- Autonomy
- Precision in predictions
- Real time control systems

# Sources of big data

- Multichannel sensor data
- Web log data/log files
- Computer simulation data
- 4Vs



# Supervised Learning

Classification

- Binary
- Multiclass
- Multilabel

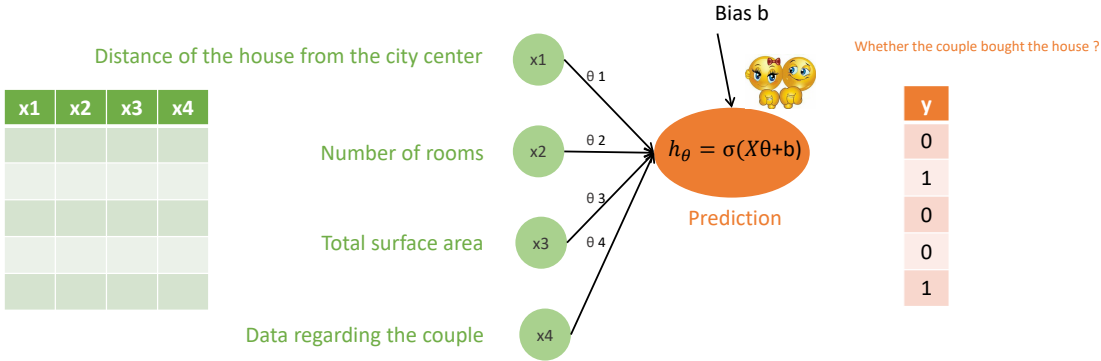
Regression

- Continuous output
- Curve fitting

# Unsupervised Learning

1. Given an input $X$, find interesting structure in the data
2. More general than a supervised approach
3. Compares to human and animal learning

# Logistic Regression

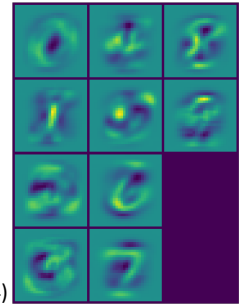



Categorically cross entropy loss

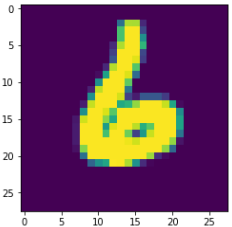
$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}[y^{(i)}log(h_\theta(x^{(i)})) + (1-y^{(i)})log(1-h_\theta(x^{(i)}))]$$

https://github.com/adil-rasheed/TK8117/blob/master/Lecture1/BinaryClassLogisticRegression.ipynb

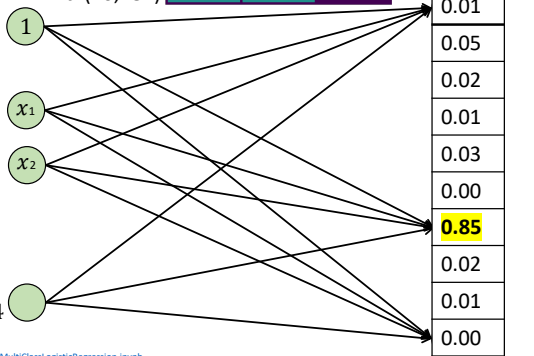
# Logistic Regression



$\theta$ (10,784)




X(1,784)



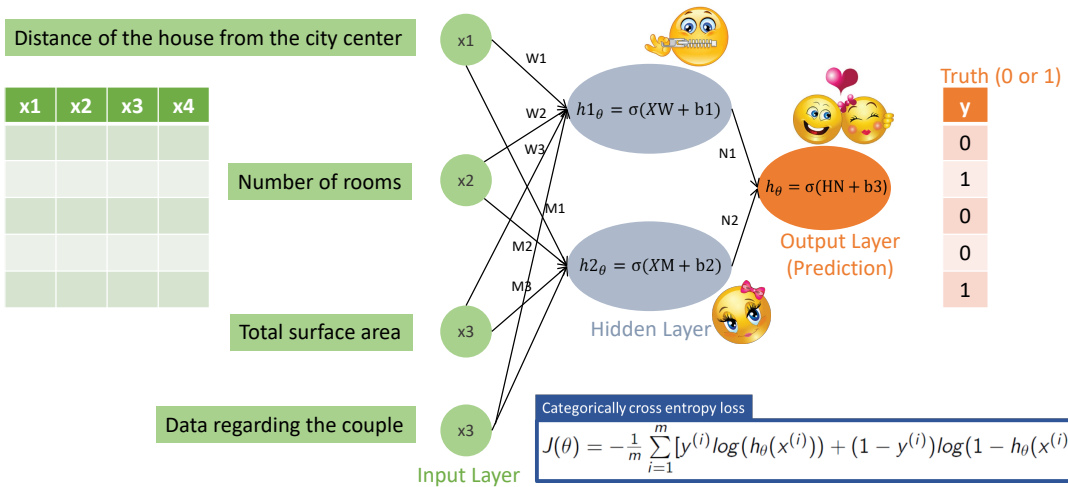

$\sigma(\theta X + b)$

| |
|---|
| 0.01 |
| 0.05 |
| 0.02 |
| 0.01 |
| 0.03 |
| 0.00 |
| **0.85** |
| 0.02 |
| 0.01 |
| 0.00 |

92% accuracy

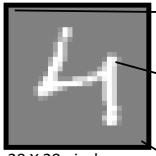
https://github.com/adil-rasheed/TK8117/blob/master/Lecture1/MultiClassLogisticRegression.ipynb

# Artificial Neural Networks

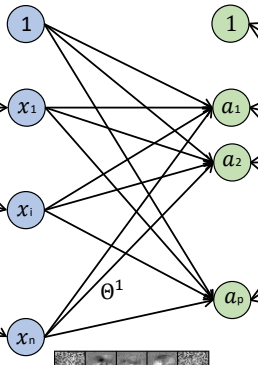



https://github.com/adil-rasheed/TK8117/blob/master/Lecture1/Aritifial_Neural_Network.ipynb

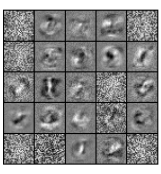
# Shallow neural network

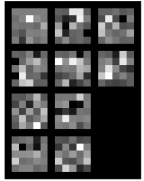


28 X 28 pixels




| Prediction | 1-Hot-Encode | Label |
|---|---|---|
| 0.05 | 0 | 0 |
| 0.01 | 0 | 1 |
| 0.02 | 0 | 2 |
| 0.05 | 0 | 3 |
| **0.8** | **1** | **4** |
| 0.1 | 0 | 5 |
| 0.04 | 0 | 6 |
| 0.01 | 0 | 7 |
| 0.04 | 0 | 8 |
| 0 | 0 | 9 |





96% accuracy

https://github.com/adil-rasheed/TK8117/blob/master/Lecture1/Aritifial_Neural_Network.ipynb

# Deep Neural Networks

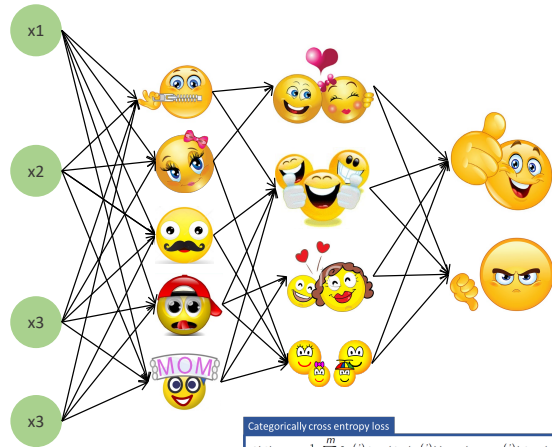
Distance of the house from the city center

| x1 | x2 | x3 | x4 |
|---|---|---|---|
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |

Number of rooms

Total surface area

Data regarding the couple

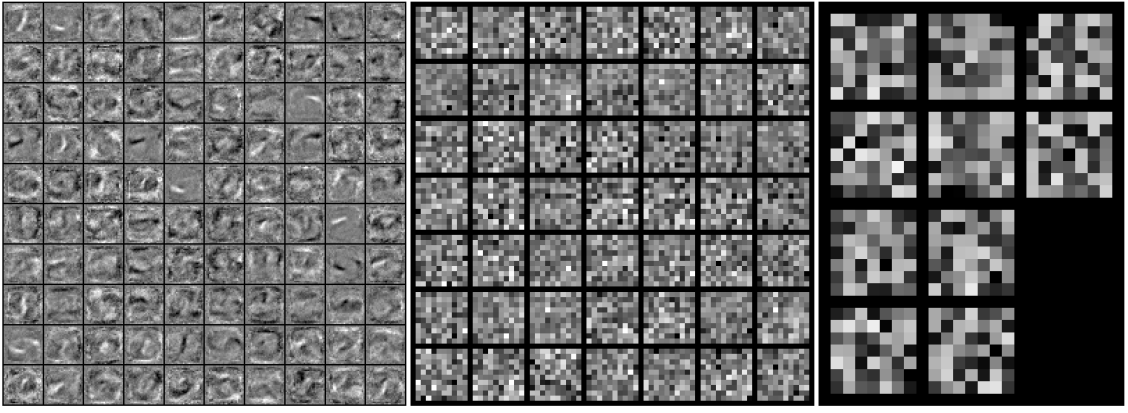



| Truth |
|---|
| y |
| 0 |
| 1 |
| 0 |
| 0 |
| 1 |

Categorically cross entropy loss

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}[y^{(i)}log(h_\theta(x^{(i)})) + (1-y^{(i)})log(1-h_\theta(x^{(i)}))]$$

# How the weights look



98% accuracy on the validation data

https://github.com/adil-rasheed/TK8117/blob/master/Lecture1/Aritifial_Neural_Network.ipynb

Trained a Conditional adversarial nets on NVIDIA GEFORCE 1080 Ti for 10 hours on 2000 images to come up with a model that can convert any grayscale image to its colored version.

Some other applications

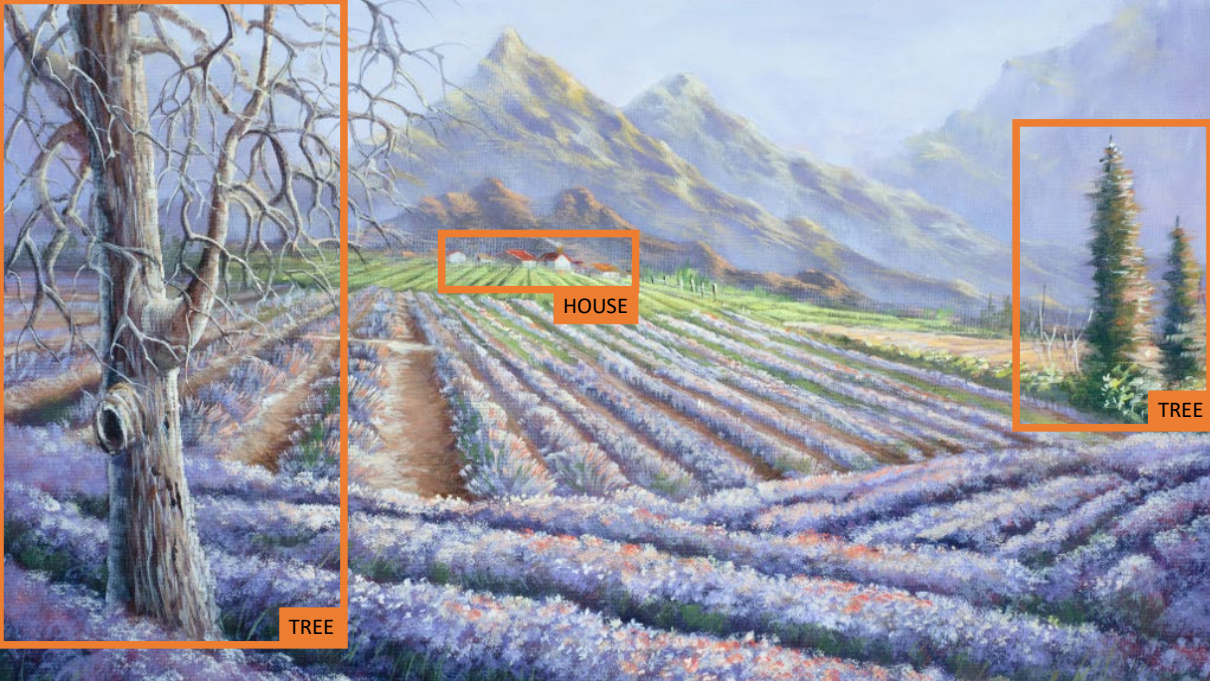
- Converting summer scenes to winter
- Landscape images to sea scapes
- Convert boats into ships

Missing Pixels

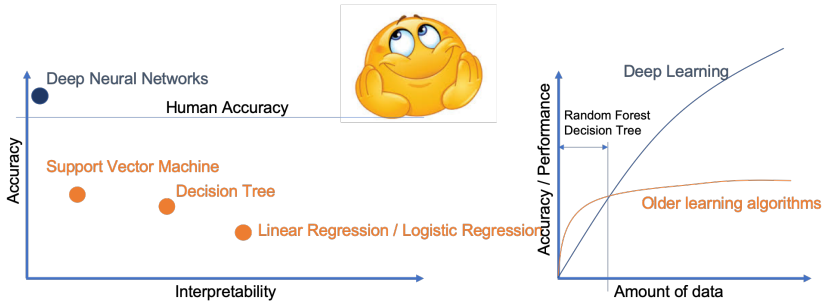
Computer Generated

Original


Outline
Computer generated
Original

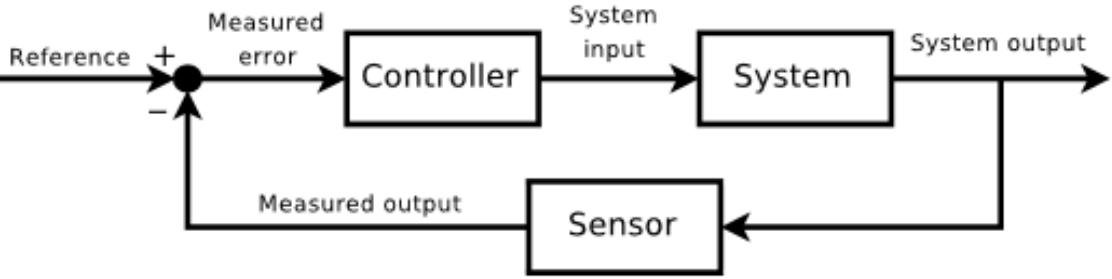

HOUSE
TREE
TREE

# Accuracy, interpretability and scalability

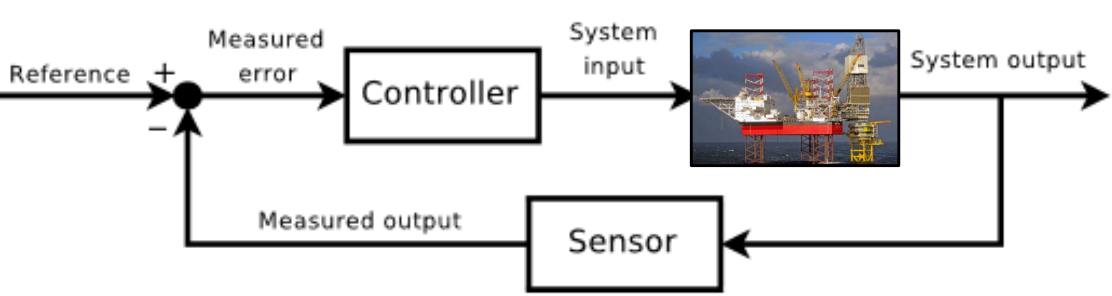


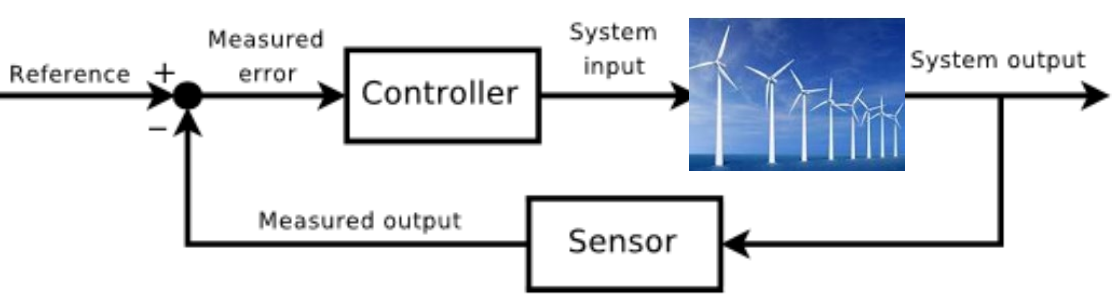
# Mechanistic vs data driven modeling

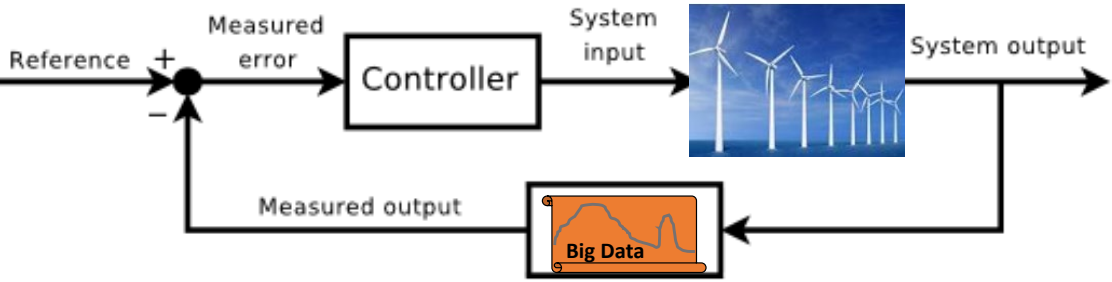
| Mechanistic modeling | Data-driven modeling |
|---|---|
| +Solid foundation based on physics | – Mostly black-boxes |
| – Difficult to assimilate very long term historical data into the computational models | + Takes into account long term historical data and experiences |
| – Computationally expensive, sensitive and susceptible to numerical instability | + Once the model is trained, it is very stable and efficient for making predictions / inferences |
| + Errors / uncertainties can be bounded and estimated | – Not possible to bound errors / uncertainties |
| + Less susceptible to bias | – Bias in data is reflected in the model prediction |
| + Generalizes well to new problems with similar physics | – Poor generalization on unseen problems |

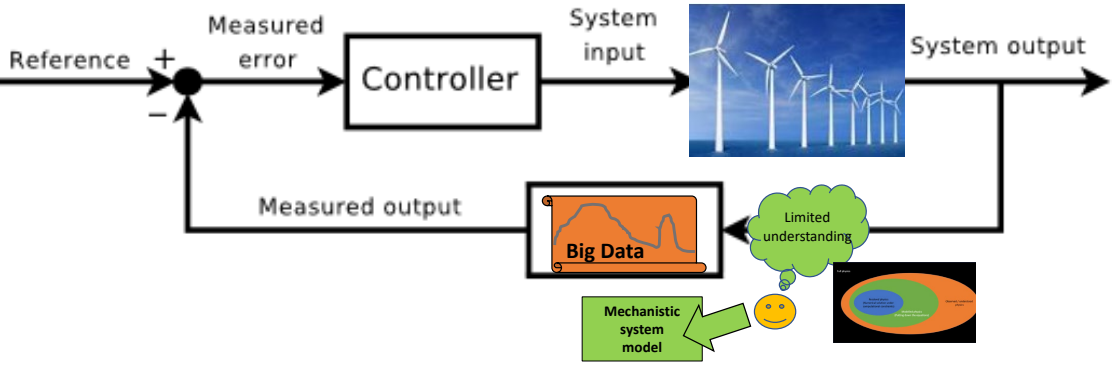


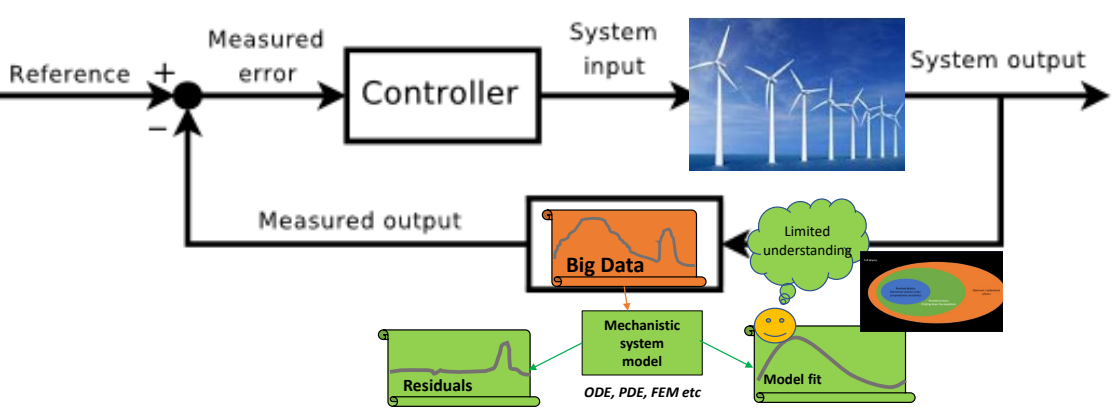

Modelling ***known*** **and** ***unknown*** **structures**
within the Framework of *Big Data Cybernetics*

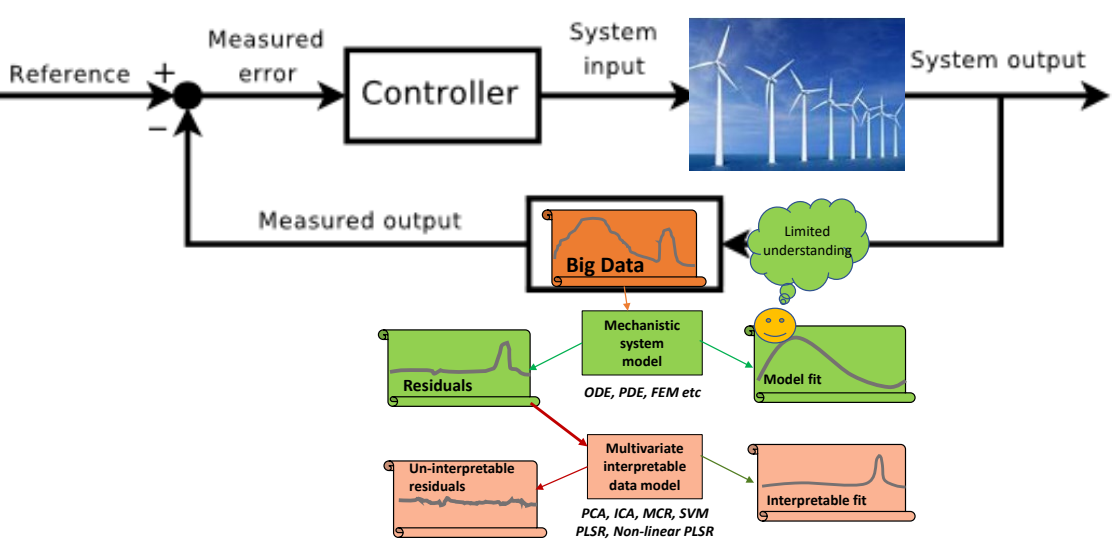

Reference
+
−
Measured
error
Controller
System
input
System output
Measured output
Sensor

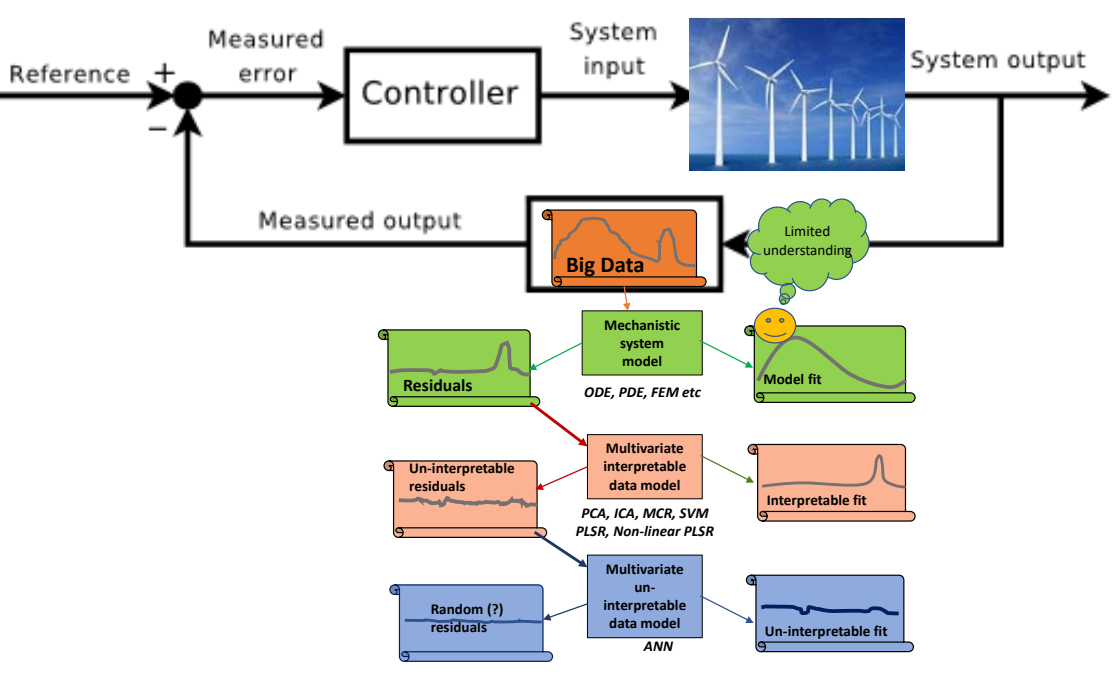

Reference
+
−
Measured
error
Controller
System
input
System output
Measured output
Sensor

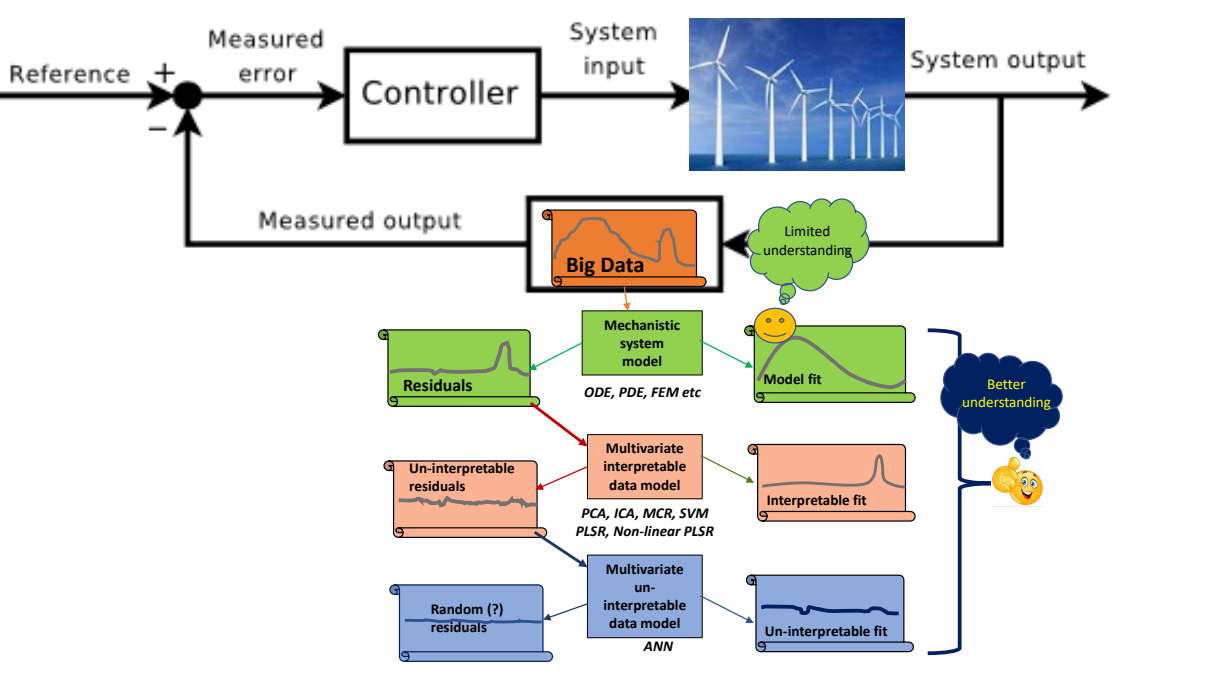

Reference
+
−
Measured error
Controller
System input
System output
Measured output
Big Data
$y_j$
time
Channels j=1,2,...,J
symbolized as
4Vs: Volume (amount), Velocity (rate), Veracity (quality), Variety

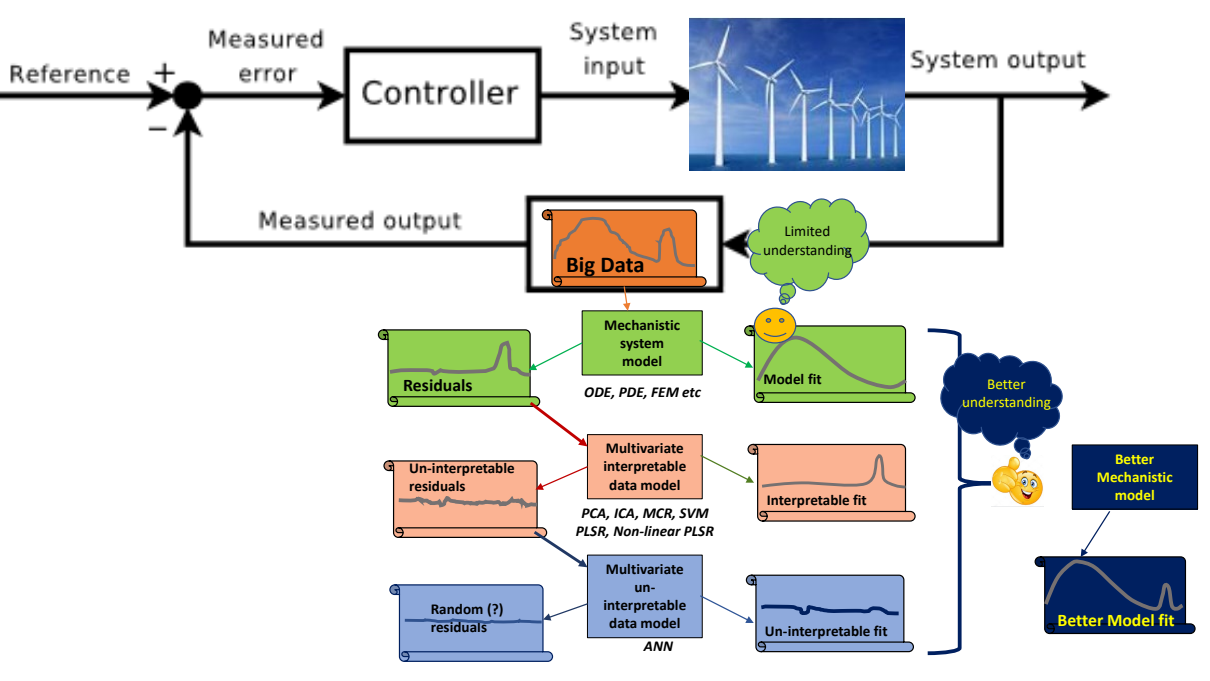

Reference
+
−
Measured error
Controller
System input
System output
Measured output
Big Data
Limited understanding
Mechanistic system model
ODE, PDE, FEM etc,

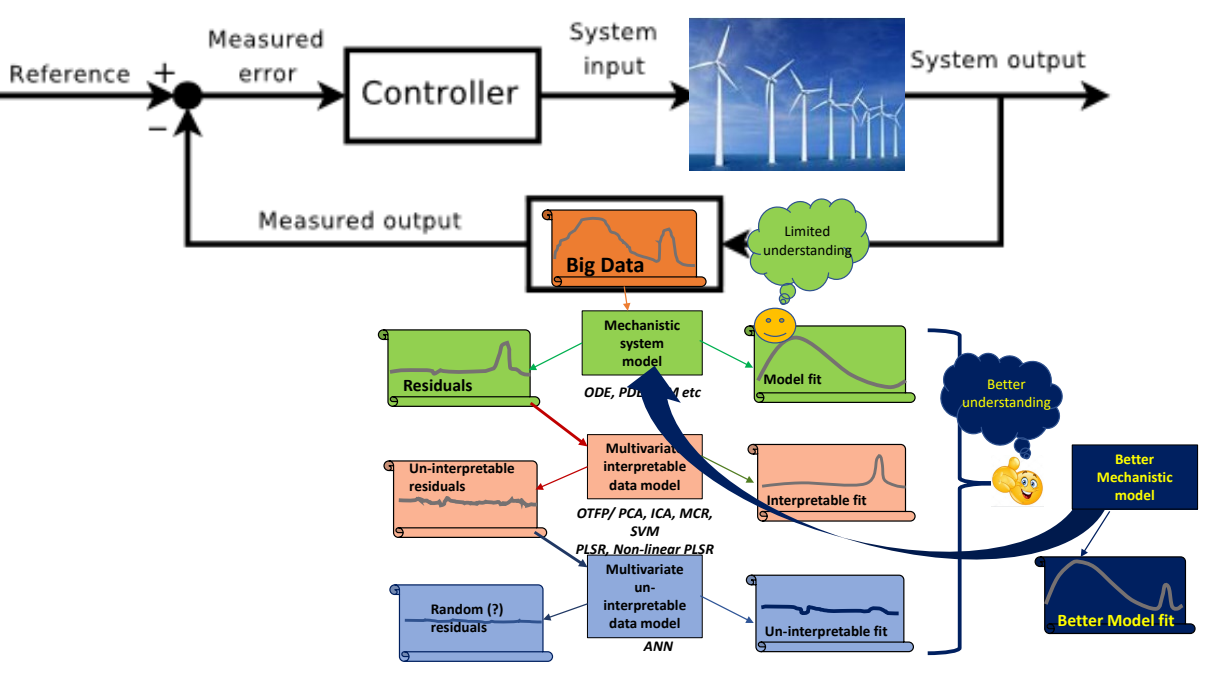

Reference
+
−
Measured
error
Controller
System
input
System output
Measured output
Big Data
Limited
understanding
Mechanistic
system
model
ODE, PDE, FEM etc
Residuals
Model fit


Reference
+
−
Measured error
Controller
System input
System output
Measured output
Big Data
Limited understanding
Mechanistic system model
ODE, PDE, FEM etc
Residuals
Model fit
Multivariate interpretable data model
PCA, ICA, MCR, SVM
PLSR, Non-linear PLSR
Un-interpretable residuals
Interpretable fit

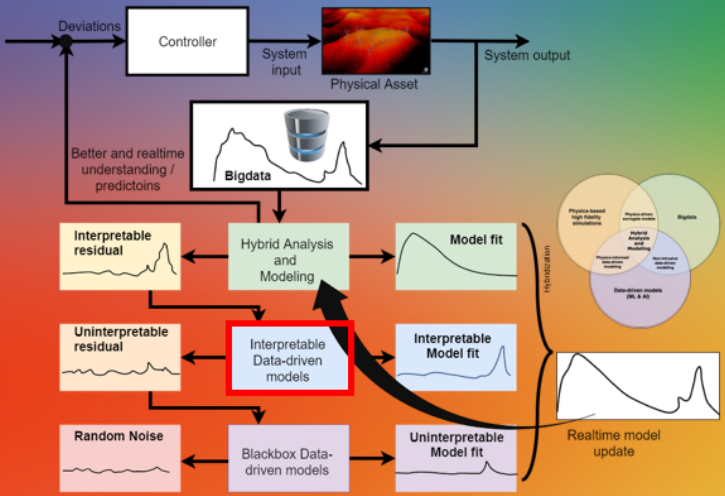

Reference
+
−
Measured error
Controller
System input
System output
Measured output
Big Data
Limited understanding
Mechanistic system model
ODE, PDE, FEM etc
Residuals
Model fit
Multivariate interpretable data model
PCA, ICA, MCR, SVM
PLSR, Non-linear PLSR
Un-interpretable residuals
Interpretable fit
Multivariate un-interpretable data model
ANN
Random (?) residuals
Un-interpretable fit


Reference
+
−
Measured error
Controller
System input
System output
Measured output
Big Data
Limited understanding
Mechanistic system model
ODE, PDE, FEM etc
Residuals
Model fit
Better understanding
Multivariate interpretable data model
PCA, ICA, MCR, SVM
PLSR, Non-linear PLSR
Un-interpretable residuals
Interpretable fit
Multivariate un-interpretable data model
ANN
Random (?) residuals
Un-interpretable fit


Reference
+
−
Measured error
Controller
System input
System output
Measured output
Big Data
Limited understanding
Mechanistic system model
ODE, PDE, FEM etc
Residuals
Model fit
Better understanding
Multivariate interpretable data model
PCA, ICA, MCR, SVM
PLSR, Non-linear PLSR
Un-interpretable residuals
Interpretable fit
Better Mechanistic model
Multivariate un-interpretable data model
ANN
Random (?) residuals
Un-interpretable fit
Better Model fit

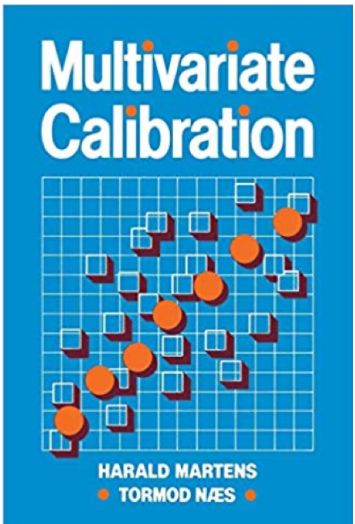

Reference
+
−
Measured error
Controller
System input
System output
Measured output
Big Data
Limited understanding
Mechanistic system model
ODE, PDE, ABM etc
Residuals
Model fit
Better understanding
Better Mechanistic model
Multivariate interpretable data model
OTFP/ PCA, ICA, MCR, SVM
PLSR, Non-linear PLSR
Un-interpretable residuals
Interpretable fit
Multivariate un-interpretable data model
ANN
Random (?) residuals
Un-interpretable fit
Better Model fit

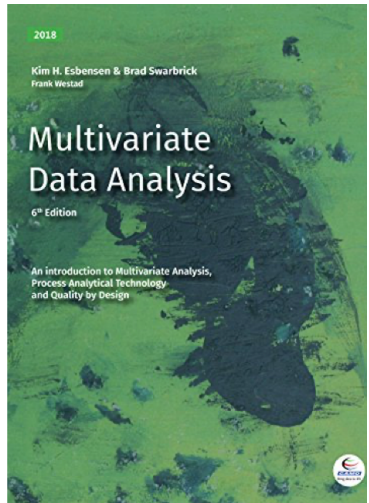
# Big Data Cybernetics

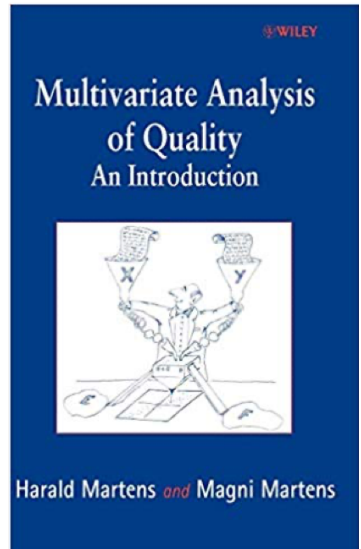



*A new paradigm in steering the world with big data characterized by high volume, high velocity, high vareity and high veracity*

# Additional Reading Materials




Additional articles, reports and papers will be shared

motivations and underlying points of view

overview of the data analysis methods types and traditions

# What is modelling?

Important topics to consider:

- The difference between significance and relevance
- Causality vs. indirect correlation
- "No prediction without interpretation"
- "No interpretation without a valid model"
- Is the model good enough to be applied on individual objects?
  - Classification: Accuracy in terms of sensitivity/selectivity
  - Quantification: Prediction error and uncertainty in prediction
- Estimation of model robustness by proper validation
- Do we find the true signal in our system?

# Modelling strategies

- First principle models: Physics, flow theory, reaction kinetics etc.
- Numerical simulations e.g. ‘Ab initio’, Molecular Dynamics, CFD, finite element
- Metamodelling
- Hybrid modelling
- Models based on empirical data (collected by means of proper designed experiments)

# Which variables are present in our process/system?

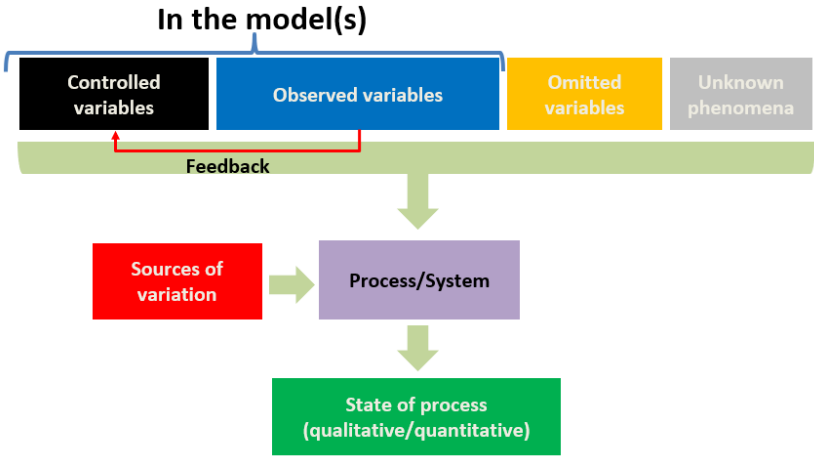



Figure: Variables in a system

# What's in our toolbox?

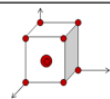
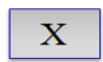
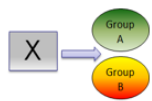
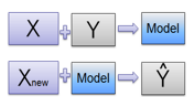
| Tool | Symbol | Purpose |
|---|---|---|
| Design of Experiments (DoE)<br><br>Full and Fractional Factorial Designs, Optimisation Designs |  | Gain maximum information from a minimum of experimental effort |
| Exploratory Data Analysis (EDA) |  | Find important sample groups or variable relationships in large data sets. Define classes. |
| Classification/Discrimination |  | Classify samples. Identify raw materials. Is the process under control (MSPC)? |
| Regression and Prediction |  | Predict quality for new samples and/or find out which variables that most influence a process |

Figure: Multivariate tools

# Considerations when choosing a specific method

Source, type and structure of data

- How to include theory and first principle models
- How to handle missing values
- How to deal with outliers
- Setting up validation schemes
- Interpretation using domain expertise and background information
- Dynamic model updates
- Applicability for prediction in real-time

# The dimensionality of a model

- Most systems are observed with more sensors/variables than the actual underlying dimensionality (rank)
- Methods based on latent variables will summarize the information in terms of "super variables" and reveal the individual variable's contribution to these
- What do we mean by *rank* ?
  - The numerical rank (this will for all real data be the smallest dimension of the data table)
  - The statistical rank (found by some statistical criterion, e.g. the chosen validation scheme or maximum likelihood)
  - The application specific rank based on experience and interpretation

# Redundancy and selectivity

- Most sensors are not measuring directly the inherent state or property of a system; 'das Ding an Sich'
- Although individual sensors are not selective, we can still use them for observing the system:
  - Qualitative: Is the system under control?
  - Quantitative: What is the water content in this rock on Mars?
- Sensor redundancy helps in making the models robust and to detect anomalies
- Using ten instead of two temperature sensors in e.g. a distillation process does not add eight new underlying dimensions to the system, but will describe the process better

Thanks for your attention and see you in the afternoon !!