BIG DATA AND MATHEMATICS FOR THE FUTURE

MATEITALY
VENEZIA, UNIVERSITÀ CA' FOSCARI, APRILE 2018

Emanuele Borgonovo
Bocconi University and AMASES
Director Bachelor in Economics, Management and Computer Science
Co-Editor in Chief, European Journal of Operational Research

Università Bocconi
MILANO

Digitalization

We are at the verge of the data driven economy

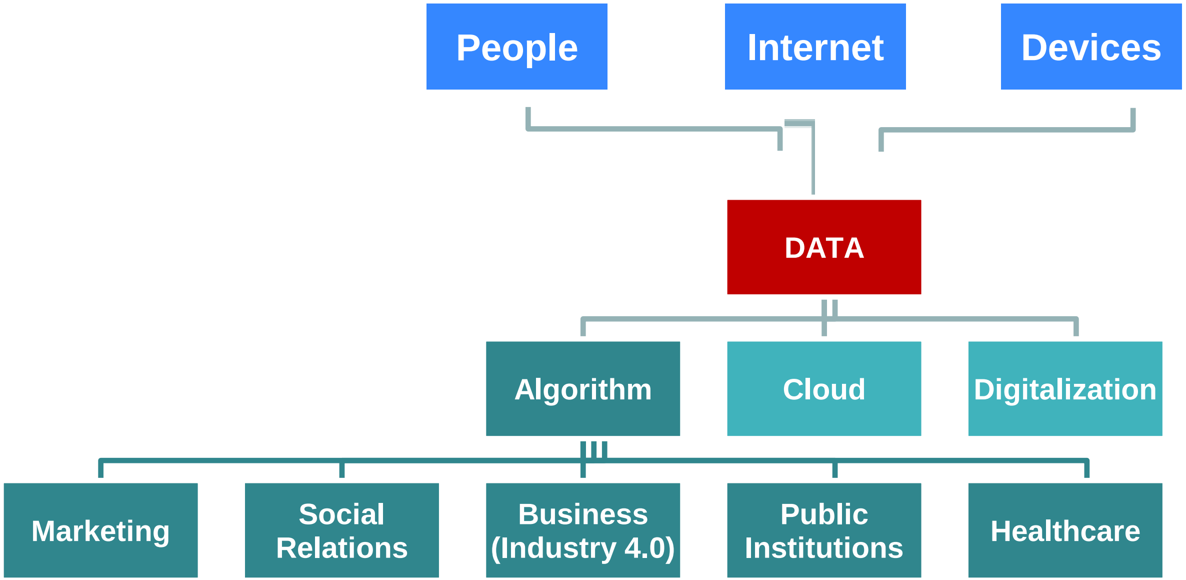
Globalization and increased connectivity make available to citizens, to corporations and scientists an amount of data unthinkable a few years ago

We register a large increase in the use of computers in everyday life, with new societal phenomena, entrepreneurial opportunities and new professions

http://www.webpagefx.com/internet-real-time/

Università Bocconi
MILANO

Prevasiveness



Communication EC-European Parliament

- July 2 2014
- Several players to create added value from the availability of big data
- EC document on cloud computing: big data and correlated services will reach 16,9 billion USD value in 2015, with an average growth rate of 40%, seven times higher the rate of growth of technology market

- In the UK number of big data specialists working in big firms will increase by 240% (Source: SAS report)

- EU Grand Colation for Digital Jobs

Università Bocconi
MILANO

EU Document

- *An adequate skills base:* The competence base addresses descriptive and predictive data analytics, data visualisation, artificial intelligence and decision-making software tools and algorithms.
- The EU document encourages *close cooperation between players* (i.e., industry and universities) to achieve the sharing of the desired competences
- *The training of* professionals who can perform in-depth thematic analysis, exploit machine findings, derive insight from data and use them for improved decision-making is considered crucial.

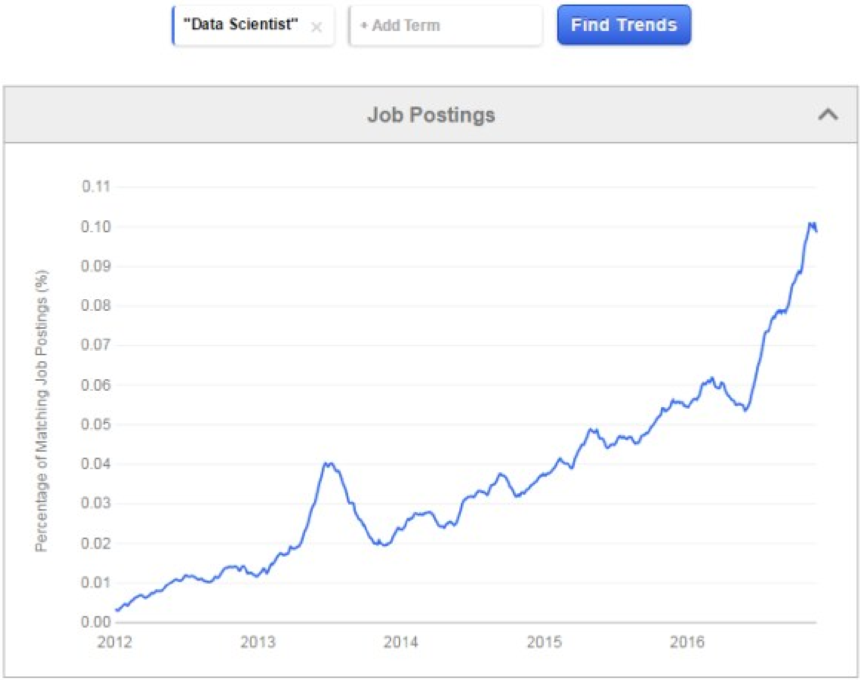
New Professions and New Skills

D.J. Patil nominated in February 2015

first White House chief data officer

But Mike Powell chief Data officer NYC

Italy: Diego Piacentini (Amazon) Digital Science group for the Italian Digital Agenda



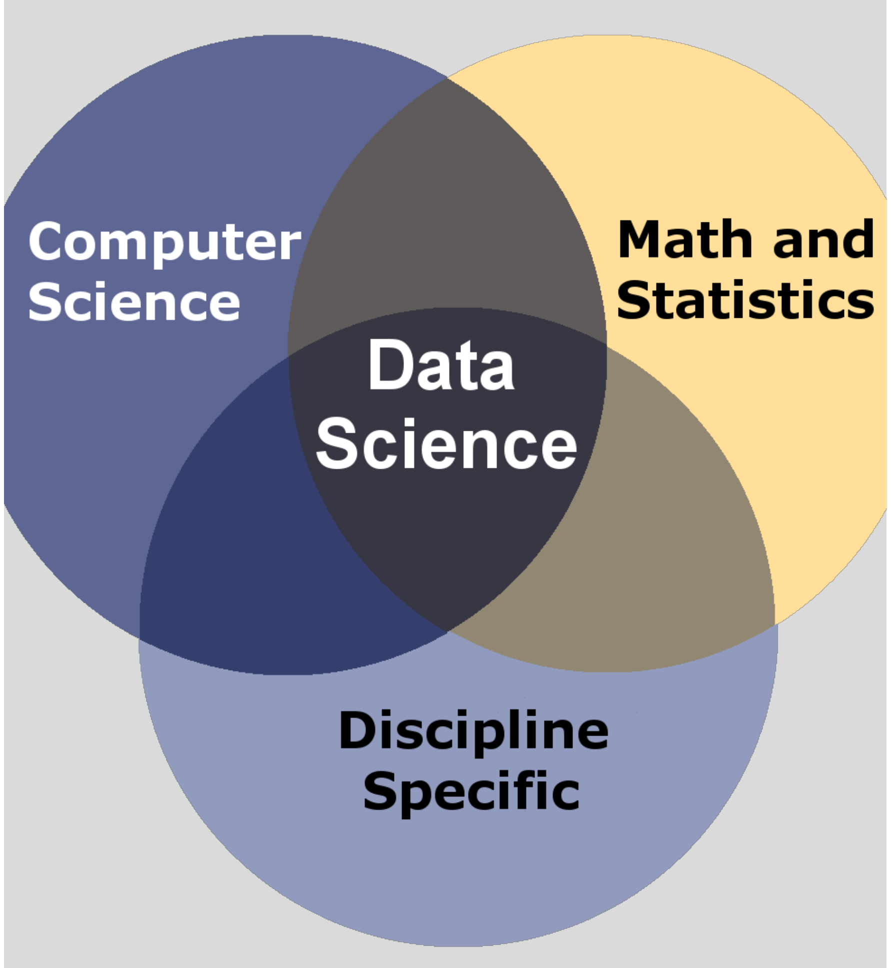
New Programs Nationally and Internationally

- Internationally: many (thousands) of MSc’s and BSc.
- MIT, The Infinite: “Two Sciences tie the knot” to announce a new major in Computer Science and Economics
- Italy (source: bachelorportal.eu)
  - Bachelor Level: a few programs (Bocconi, Milan; Bologna)
  - MSc: 11 programs (source: MastersPortal.eu): Bologna, Genova, Milan (Bocconi, Bicocca, Cattolica, Statale), Roma (Luiss, Tor Vergata, Sapienza), Padova, Torino

Università
Bocconi
MILANO

Pillars

Data science has three pillars: Mathematics, Statistics and Computer Science. It addresses challenging issues in a variety of applied fields by analyzing large-scale datasets and modeling the underlying complex phenomena.



Università Bocconi MILANO

What is Big Data?

Big Data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making (Source: Gartner)

- the technical skills to **gather and manage** data
- the ability to **interpret** data through the lens of economic models, thus enabling them to provide credible support to company decision making.

Computer Science

- Large Data Storage and Retreival
- Memory Requirements
- Parallel and Cloud Computing
- … an example…

Algorithms

— [Click Here]()

Machine Learning: Algorithms

- Supervised Learning
  - The dataset is divided into training and testing
- Unsupervised Learning
  - The data are not prepared or labelled, the algorithm learns by similarity
- Semi Supervised Learning

Types of Algorithms… (if a list is possible...)

- Regression and Regularization Algorithms
- Instance-based algorithms
- Decision Trees
- Bayesian Networks
- Neural Networks
- Deep Learning
- … more

Some most Recent Examples

—AI in optimization

- A. Lodi, Y. Banjo, E. Frejinger (Montreal)
- Using AI to help solve complex optimization problems
- Helps in Mixed Integer-Linear programs
- Helps in Integer programs improving branch and bound

— Making Simulations More Efficient

- Borgonovo, E., Lu, X. Rosasco A., Rudi A.
- Fast kriging

—Bayesian Classifier for Preference Learning

- Select your best camera
- Multicriteria problem
- Bayesian Approach: set a prior over a class of uncertain parameters of the algorithm
- Computer Implementation

Challenges

- Research for reducing memory requirements
- In statistics, for instance, large datasets: how can we estimate the same quantities in a reasonable time?
- Or how can we protect privacy in our “mathematical formulae”?
- Solving Large Scale Optimization Problems

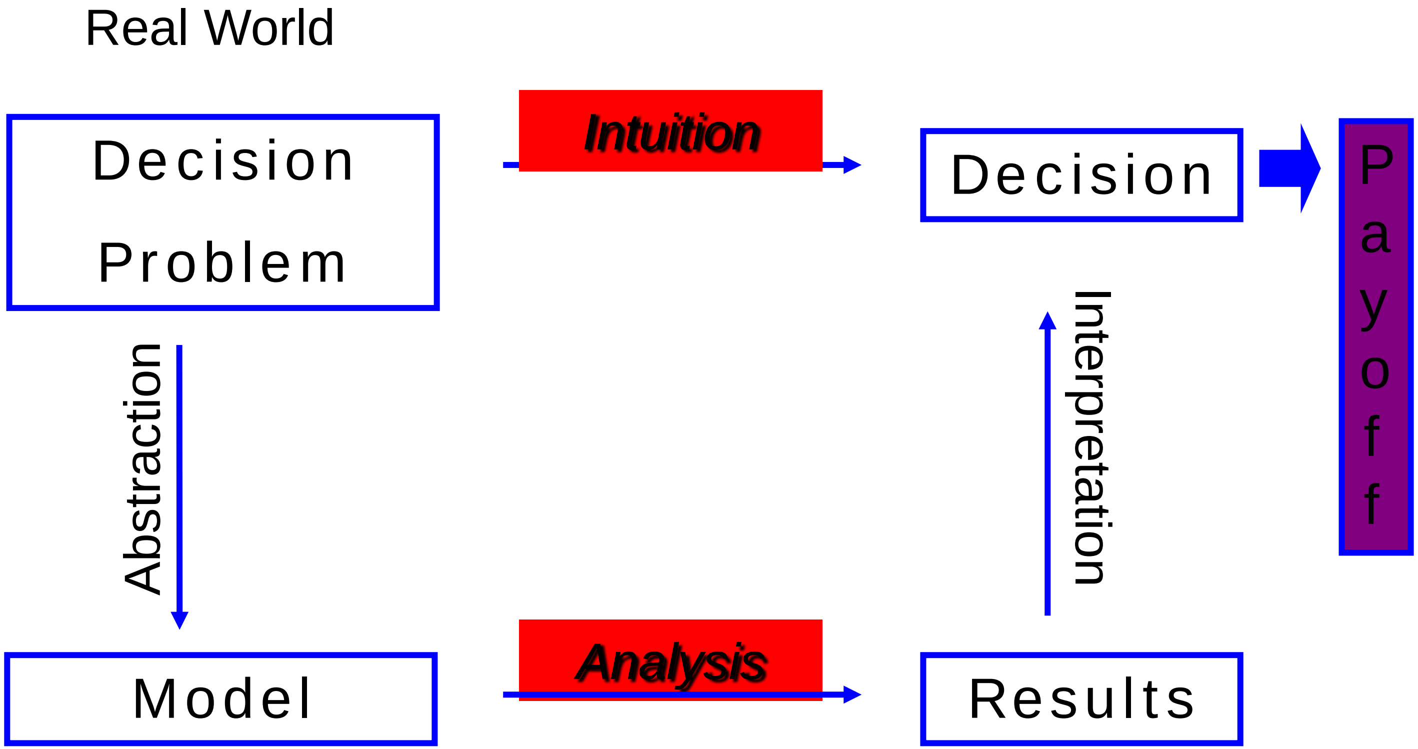
Large Donations

- MIT-IBM Watson lab 240MUSD (Sept. 2017)
- Accenture-MIT
- Several Private or Publicly Funded Chairs
- Facebook Research (Yann Le Cun)
- Microsoft Research (Ed Horvitz)
- Google Research (Peter Norvig)

Business Impact: Industry 4.0

- Increase automation
- Self Driving Cars
- Machines that communicate to machines

Modern Decision Making Process



Università Bocconi MILANO

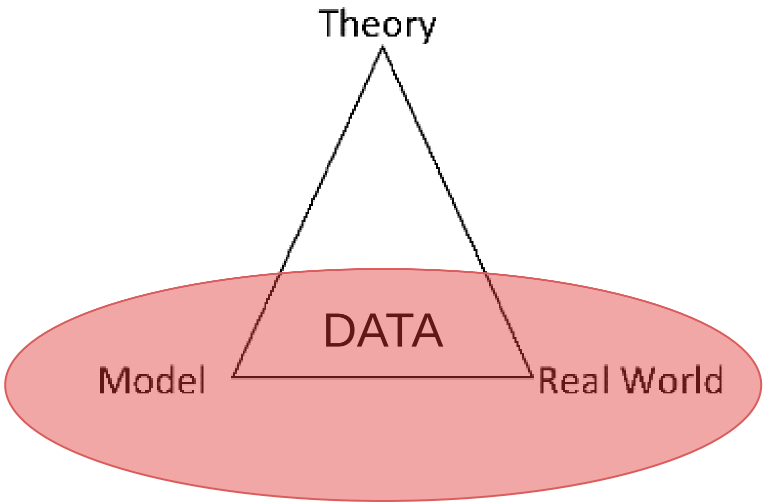
Social Impact

- New jobs for young generations?
- What about the intermediate generation?
- Need for additional education?
- How does society perceives AI?

Algorithms (models) can fail

- Google Flue successfully predicted flue diffusion in the US for 4 years
- At year 5 unsuccessful predictions due to overfitting

Epistemologically, what are we doing?



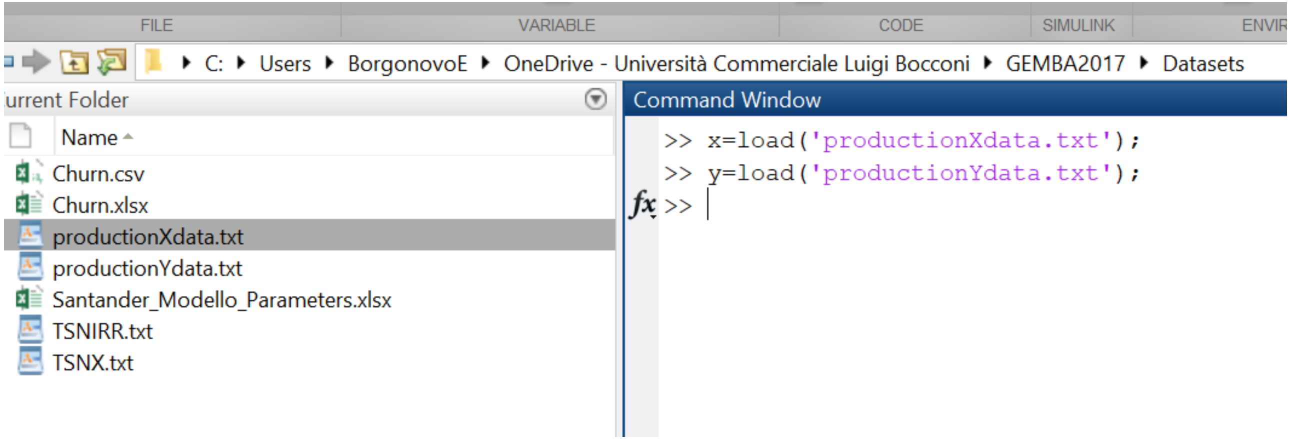
Scientific Challenges

- New Problems -> New Concepts -> New Algorithms
- Old Problems -> New Challenges -> New Algorithms
- Make Everything Fast and Large
- Investigate the profound meaning of what we are doing (let us not loose sight on the meaning)
- Causality, what is it?
- Interactions, what are they?
- Statistical dependence, what do we mean by it?

Training a Neural Network in Matlab

- We try here and train our own network for prediction.
- Suppose your variable of interest is the production of a given item. The production process is a function of three variables: the «geography» (the season, country, provider), quality of the supply, the reliability of the plant (number of planned and unplanned outages).
- You have collected data for these plants at a number of facilities. You have also rescaled the data on [0,1] (what you see is, in fact, an artificial dataset). Recall our goal is to train the network.

Load the Dataset



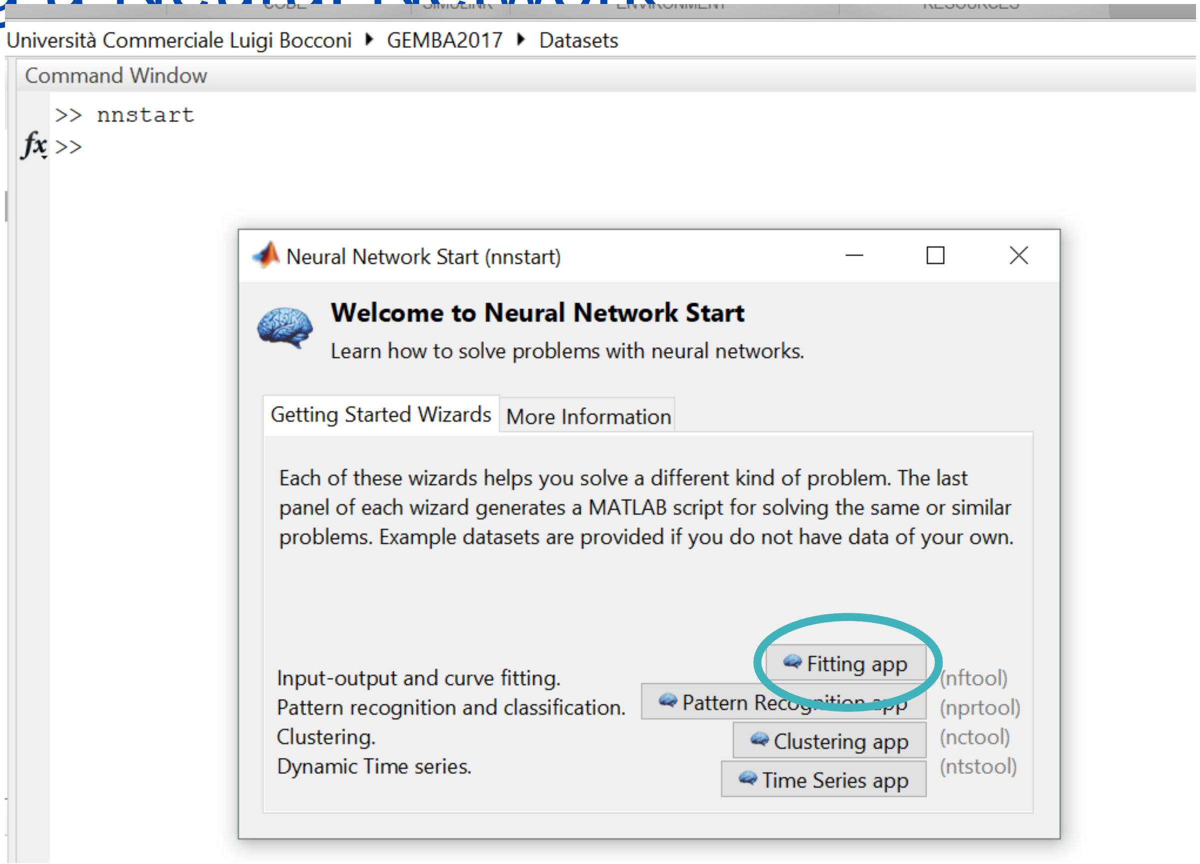
Fitting a Linear Regression

Jniversità Commerciale Luigi Bocconi ▸ GEMBA2017 ▸ Datasets

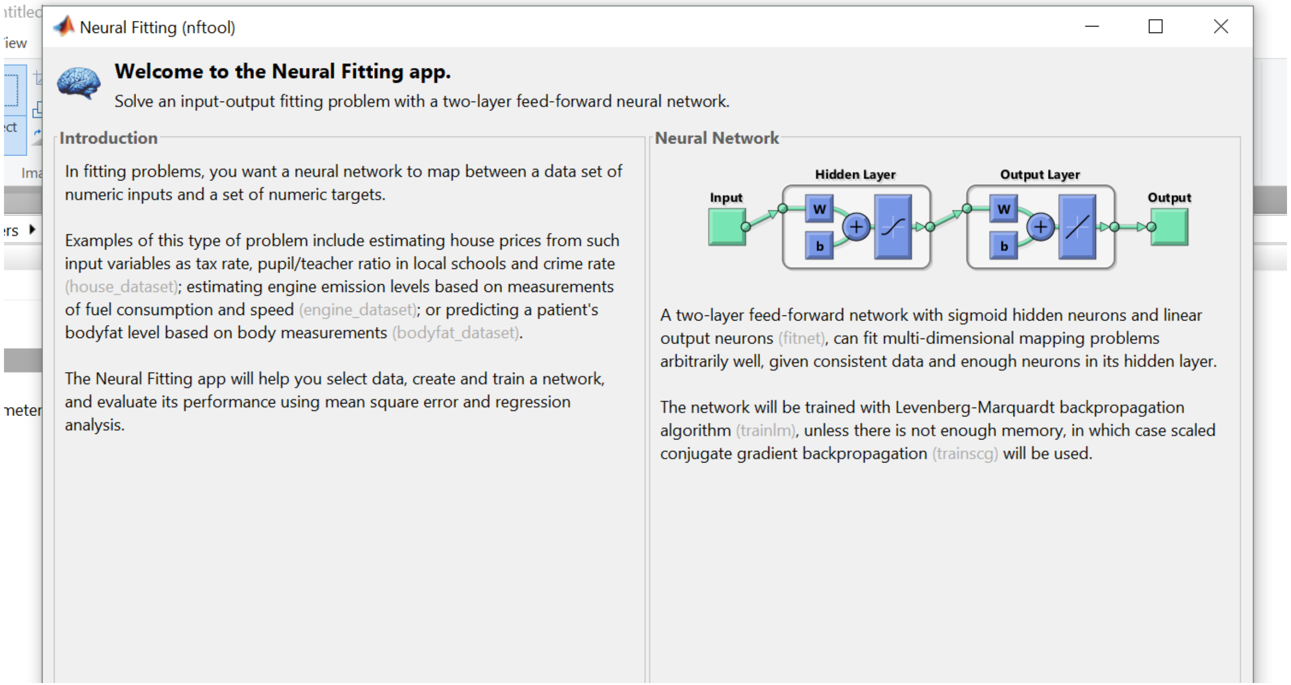
Command Window

```
>> fitlm(x,y)

ans =


Linear regression model:
    y ~ 1 + x1 + x2 + x3

Estimated Coefficients:
                   Estimate       SE         tStat       pValue
                   ________    _________    _______    __________

    (Intercept)    -0.24137    0.0069536    -34.712    1.1205e-173
    x1              0.24636     0.007855     31.364    9.5751e-151
    x2              0.23637     0.007463     31.672    7.4071e-153
    x3              0.24352    0.0074148     32.843     6.925e-161


Number of observations: 1000, Error degrees of freedom: 996
Root Mean Squared Error: 0.0684
R-squared: 0.752,  Adjusted R-Squared 0.751
F-statistic vs. constant model: 1.01e+03, p-value = 9.04e-301
>>
```

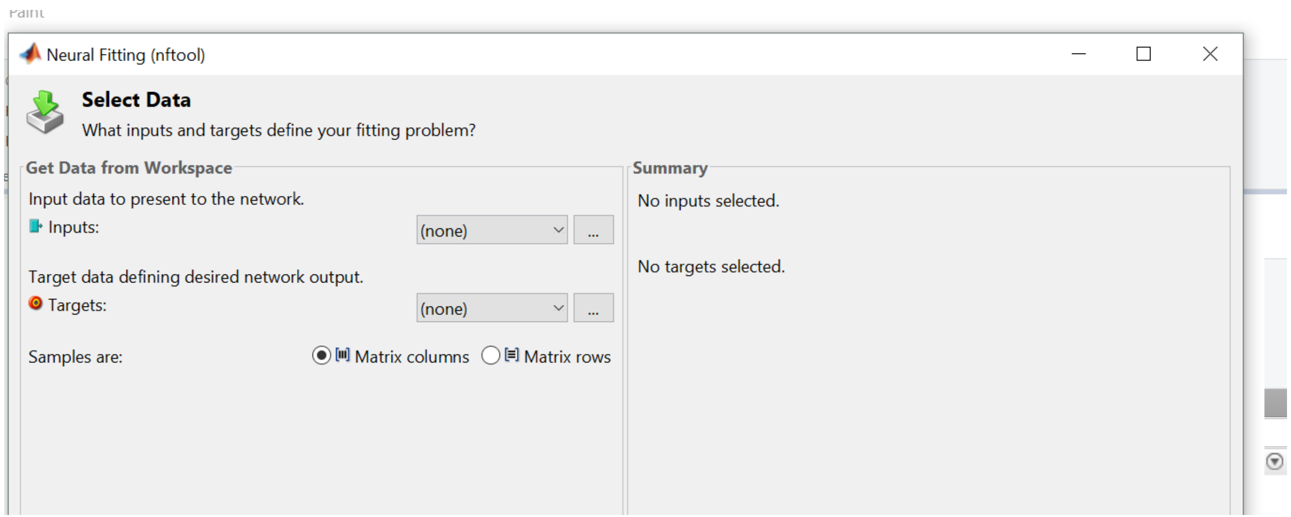
Fitting a Neural Network



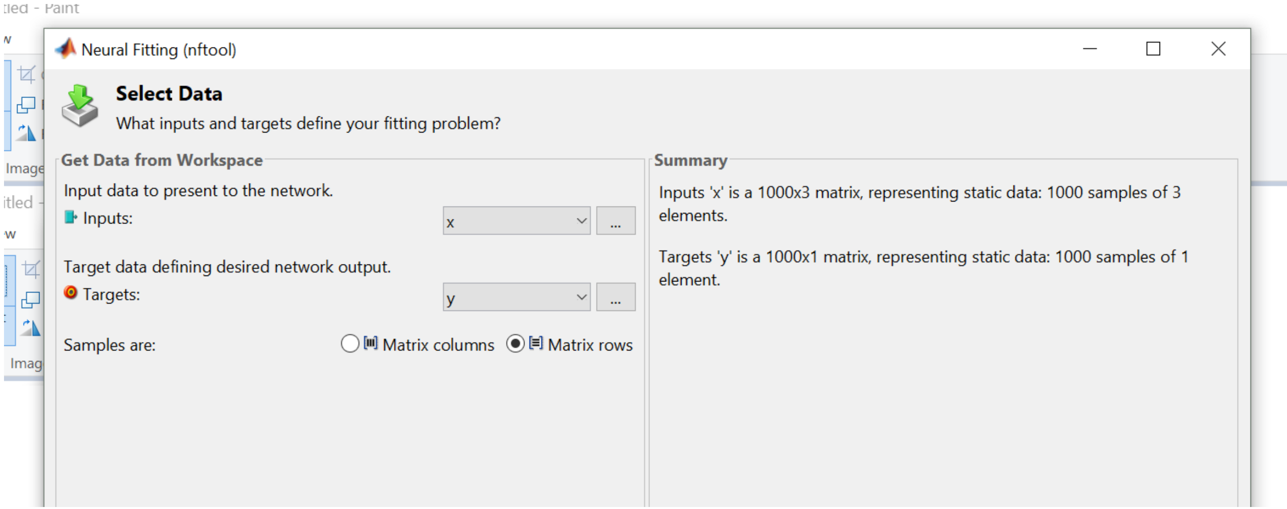
First screen



Import data menu

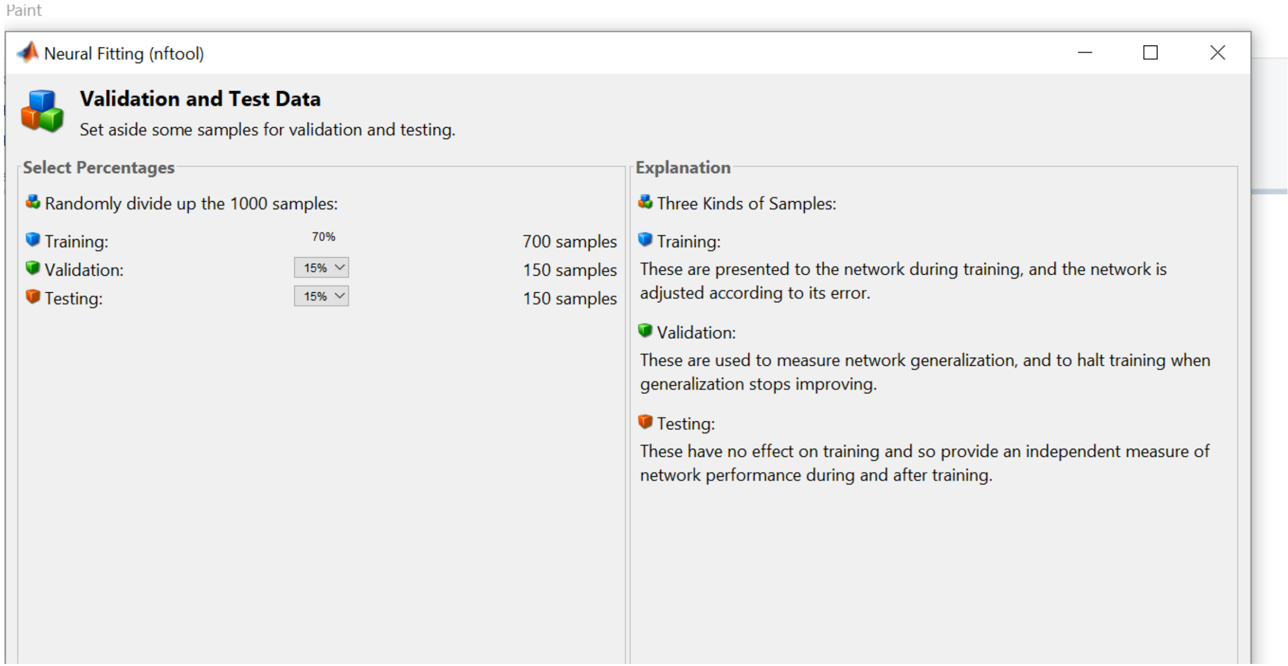


Import x and y

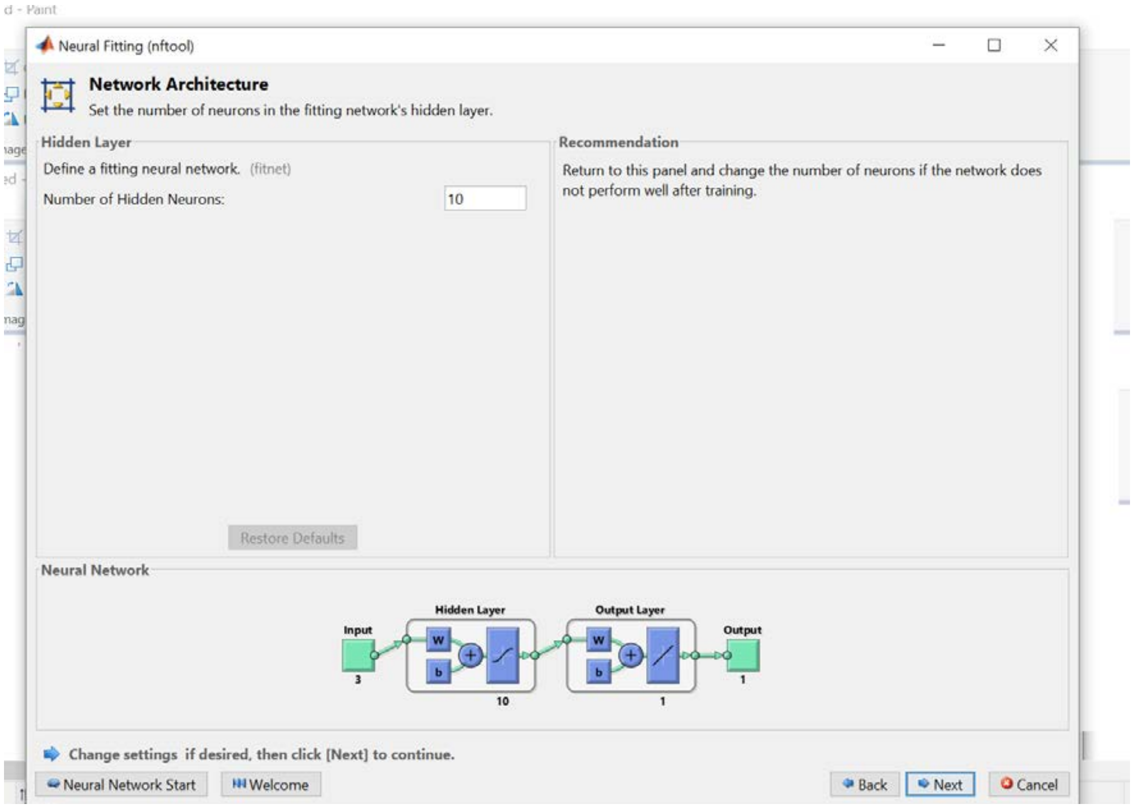


Click Next at the bottom left of the page

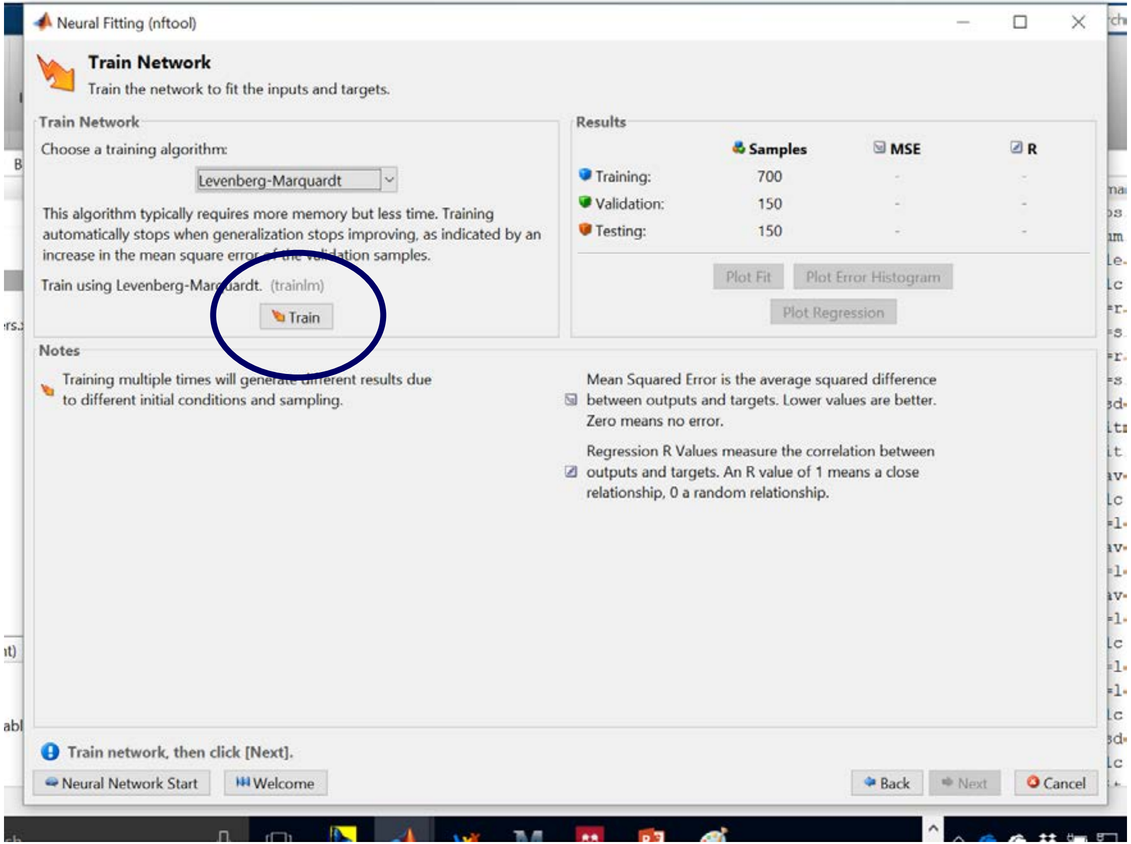
Train the Network



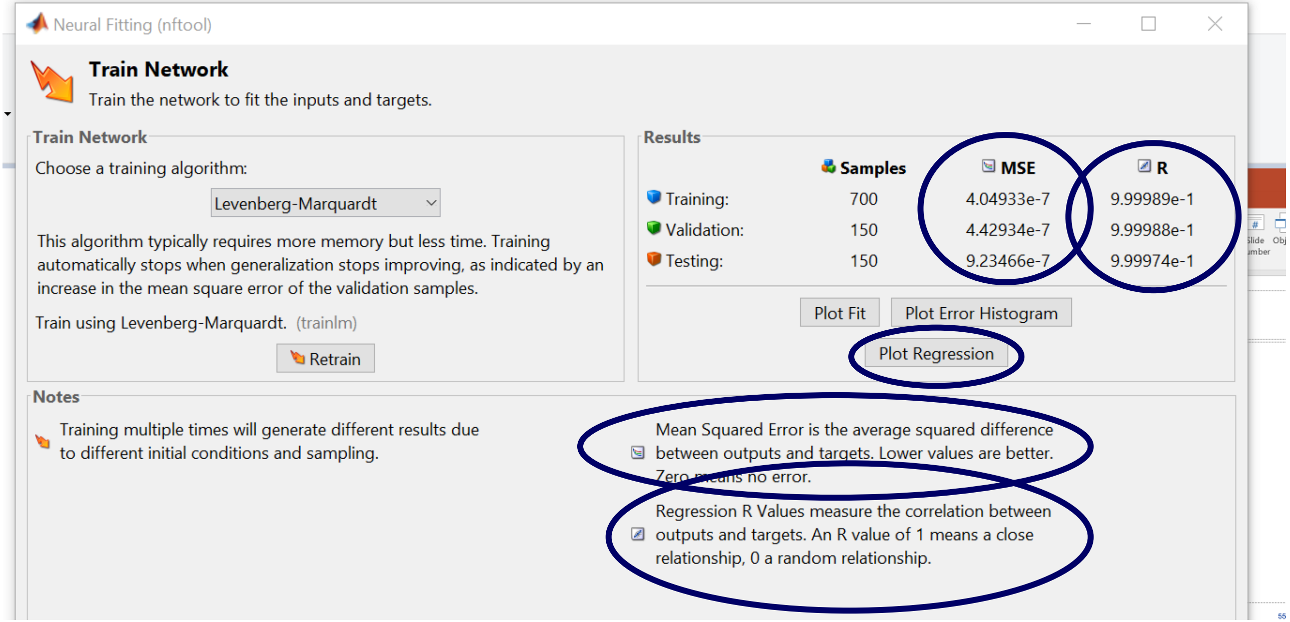
Network Structure

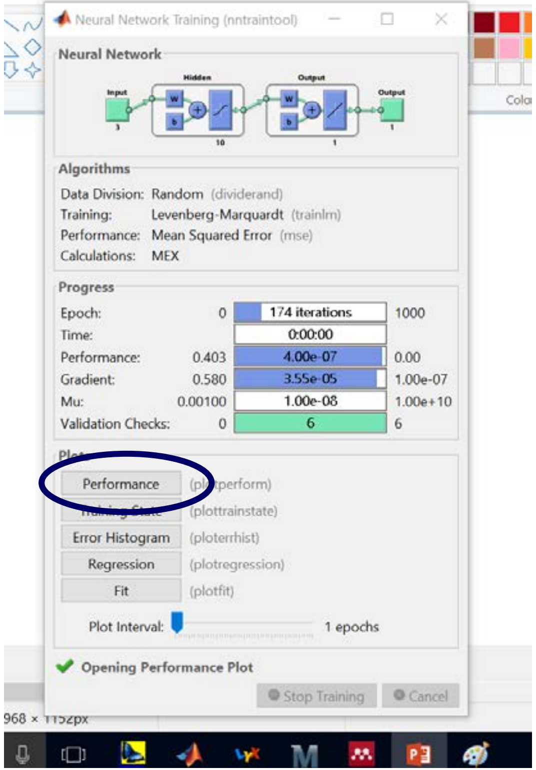


Training

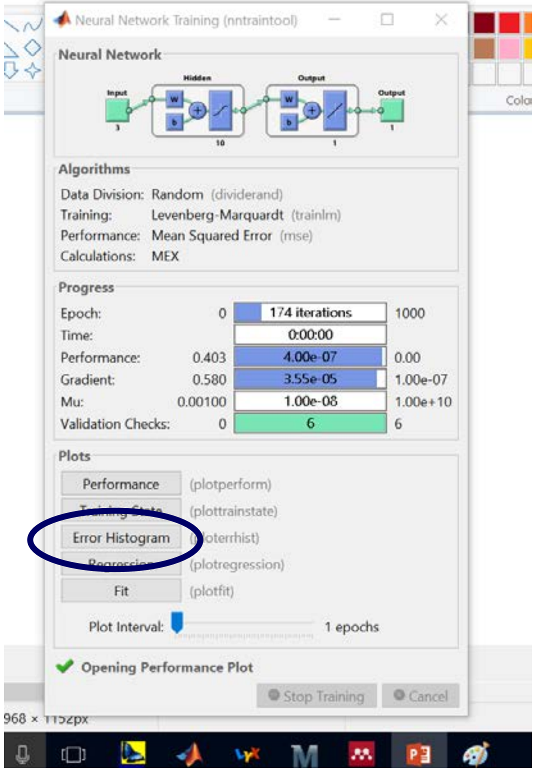

Neural Fitting (nftool)
Train Network
Train the network to fit the inputs and targets.
Train Network
Choose a training algorithm:
Levenberg-Marquardt
This algorithm typically requires more memory but less time. Training automatically stops when generalization stops improving, as indicated by an increase in the mean square error of the validation samples.
Train using Levenberg-Marquardt. (trainlm)
Train
Results
Samples
MSE
R
Training: 700
Validation: 150
Testing: 150
Plot Fit
Plot Error Histogram
Plot Regression
Notes
Training multiple times will generate different results due to different initial conditions and sampling.
Mean Squared Error is the average squared difference between outputs and targets. Lower values are better. Zero means no error.
Regression R Values measure the correlation between outputs and targets. An R value of 1 means a close relationship, 0 a random relationship.
Train network, then click [Next].
Neural Network Start
Welcome
Back
Next
Cancel

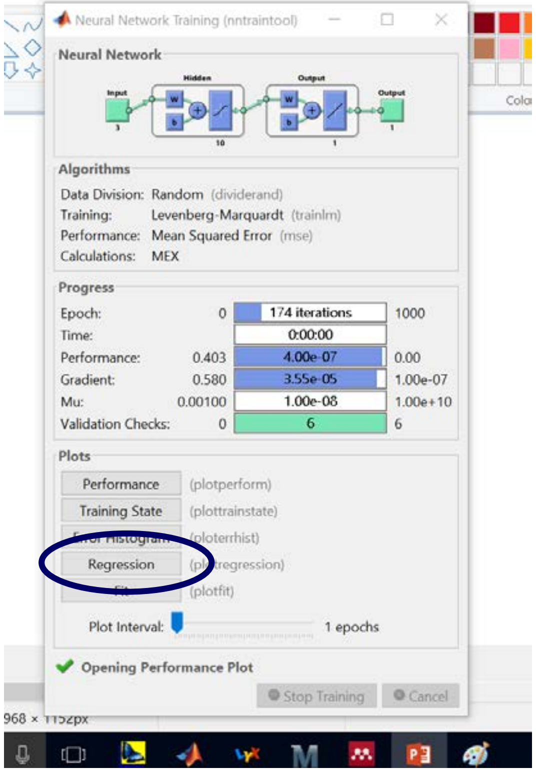
Reading Training Results





Neural Network Training (nntraintool)
Neural Network
Hidden
Output
Input
Output
Algorithms
Data Division: Random (dividerand)
Training: Levenberg-Marquardt (trainlm)
Performance: Mean Squared Error (mse)
Calculations: MEX
Progress
Epoch: 0 174 iterations 1000
Time: 0:00:00
Performance: 0.403 4.00e-07 0.00
Gradient: 0.580 3.55e-05 1.00e-07
Mu: 0.00100 1.00e-08 1.00e+10
Validation Checks: 0 6 6
Performance (plotperform)
(plottrainstate)
Error Histogram (ploterrhist)
Regression (plotregression)
Fit (plotfit)
Plot Interval: 1 epochs
Opening Performance Plot
Stop Training
Cancel


Neural Network Training (nntraintool)
Neural Network
Hidden
Output
Input
Output
Algorithms
Data Division: Random (dividerand)
Training: Levenberg-Marquardt (trainlm)
Performance: Mean Squared Error (mse)
Calculations: MEX
Progress
Epoch: 0 174 iterations 1000
Time: 0:00:00
Performance: 0.403 4.00e-07 0.00
Gradient: 0.580 3.55e-05 1.00e-07
Mu: 0.00100 1.00e-08 1.00e+10
Validation Checks: 0 6 6
Plots
Performance (plotperform)
(plottrainstate)
Error Histogram
(plotregression)
Fit (plotfit)
Plot Interval: 1 epochs
Opening Performance Plot
Stop Training
Cancel


Neural Network Training (nntraintool)
Neural Network
Hidden
Output
Input
Output
Algorithms
Data Division: Random (dividerand)
Training: Levenberg-Marquardt (trainlm)
Performance: Mean Squared Error (mse)
Calculations: MEX
Progress
Epoch: 0 174 iterations 1000
Time: 0:00:00
Performance: 0.403 4.00e-07 0.00
Gradient: 0.580 3.55e-05 1.00e-07
Mu: 0.00100 1.00e-08 1.00e+10
Validation Checks: 0 6 6
Plots
Performance (plotperform)
Training State (plottrainstate)
(ploterrhist)
Regression (plotregression)
(plotfit)
Plot Interval: 1 epochs
Opening Performance Plot
Stop Training
Cancel

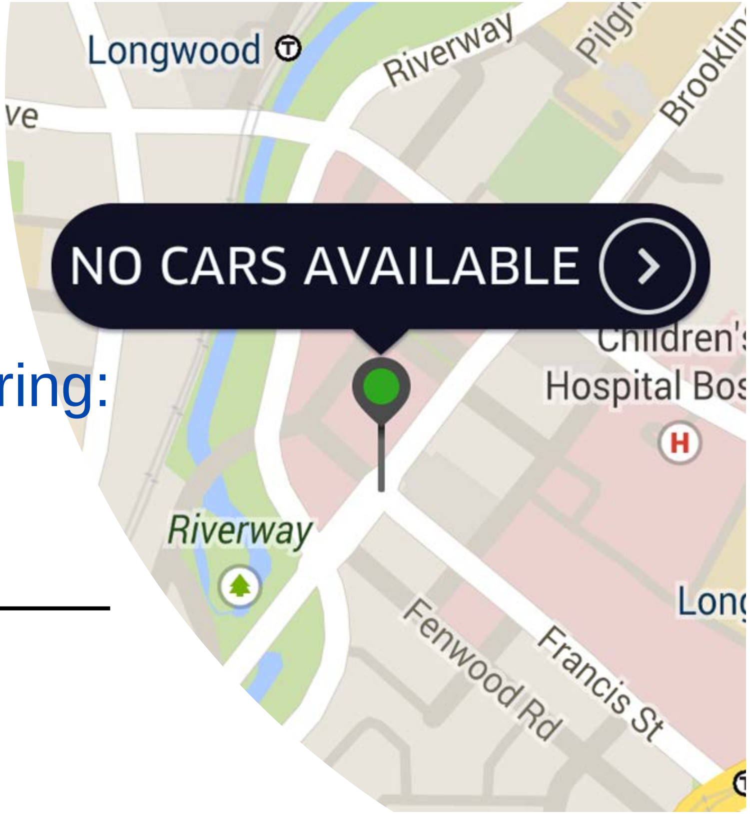
Discussion

- Advantages: does not require linear expression
- Very powerful
- Disadvantage: black box

Applications

- Forecasting demand for bike sharing
- Car Sharing Demand

Using Deep Learning to forecast the demand for car sharing: Milan case study



From Master of Science Thesis of Matteo Zunino, Bocconi University. Supervisors: E. Borgonovo and A. Rezzani

Car sharing: market overview*

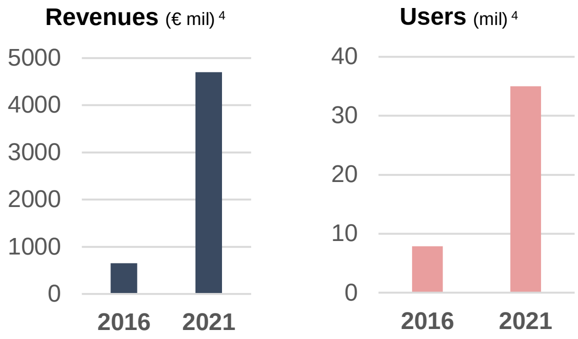
Urbanization is putting pressure on city governments to increase incentives for environmentally sustainable mobility services and **reduce traffic congestion in big cities**. The emergence of **new innovative mobility services** will ultimately lead to the **disappearance of private car ownership** in big cities in the future [1].

One of the most successful forms of shared-mobility services is **car sharing (CS)**. Car sharing is an innovative mobility service in which a car sharing operator (CSO) provides individuals and/or companies with the access to a fleet of vehicles on an as-needed basis through membership programs, without the cost and responsibilities associated with car ownership [2].

Car sharing services can be categorized in two main clusters: **1) Two-way** car sharing, in which users have to return the car in the place where it was originally accessed;
2) One-way car sharing, in which users are allowed to return cars in a different parking spot from the one they accessed it.

Car sharing services bring substantial benefits to citizens in terms of health, travel cost and time, increasing **access to flexible mobility without the downsides of private car ownership** [3]. Besides being an important **asset to deliver sustainable development** in urban centers, car sharing services represents also a **great market opportunity**., as the two graphs on the right show.

Figure 1: car sharing market revenues and users worldwide



2016: € 650 mil
2021 (projected): € 4,7 bill
CAGR: 48,6%

2016: 7,9 mil
2021 (projected): 35 mil
CAGR: 34,68%

Looking at **future developments**, CSOs already own fleets of connected vehicles that collect vast amounts of data, which is one of the most important assets for developing an **autonomous vehicle business**. Still, nowadays CSOs face **relevant challenges**, including low service penetration. As a consequence, they need to improve operations constantly to **increase service attractiveness and grow profitably** [5].

*From M. Zunino: Using Deep Learning to Forecast Car Sharing Demand, MSc thesis. Supervisors. E. Borgonovo and A. Rezzani

(1) Hampshire, R. C., Simek, C., Fabusuyi, T., Di, X. and Chen, X. (2017). "Measuring the Impact of an Unanticipated Suspension of Ride-Sourcing in Austin, Texas". University of Michigan at Dearborn.
(2) Millard-Ball, A., Murray, G., Ter Schure, J., Fox, C., Burkhardt, J. (2005). "Car-sharing: Where and How It Succeeds". Transportation Research Board of the National Academies, Washington.
(3) Lane, C., Zeng, H., Dhingra, C., Carrigan, A. (2015). "Carsharing: A Vehicle for Sustainable Mobility in Emerging Markets?". World Research Institute.
(4) Bert, J., Collie, B., Gerrits, M., and Xu, G. (2016). "What's Ahead for Car Sharing? The New Mobility and Its Impact on Vehicle Sales". BCG Perspectives.
(5) Ciari, F., Schuessler, N. and Axhausen, K. (2011). "Estimation of car-sharing demand using an activity-based microsimulation approach: Model discussion and preliminary results". Institute for Transport Planning and Systems (IVT), Zürich.

The Problem

Problem: vehicle imbalance in one-way systems

Demand for car sharing services shows temporal and spatial patterns. When users move to popular destinations, other areas remain without sufficient cars to serve demand. As a consequence, **areas with low demand might have a lot of idle cars, while areas with high demand have none**. This problem is known as **vehicle imbalance**.

If users experience continuous lack of available cars when trying to book the service, they switch to more reliable transportation services. **Vehicle imbalance can bring thousands of euros in daily operating losses**[1] to a car sharing business.

Solutions identified: vehicle relocation

With vehicle relocation, **cars are physically moved** from one location that has a vehicle surplus to another that experiences a vehicle shortage. Due to the evolving dynamics of service demand, **if CSOs don't update relocation strategies, costs increase without resolving vehicle imbalance** [2].

To solve vehicle imbalance and maximize profits in a car sharing system that employs relocation operations different **mathematical models known as optimization frameworks or decision support systems (DSS)** have been developed.

Research gap: demand forecast tool with high spatial and temporal resolution

The following quotes were taken from the European Journal of Operation Research and from the Transportation Research Journal:

"A **central issue** related to strategic decision-making for car-sharing systems is the **estimation of the spatial and temporal distribution of the demand**" [3]

"The **operational management** of carsharing systems is **complex** due to the stochastic and **dynamic nature of demand in time and space**" [4]

"Stronger focus on **reliable demand models which have not been covered comprehensively by literature in the past** is needed" [5]

(1) Jorge, D., Correia, G. H. A., and Barnhart, C. (2014). "Comparing optimal relocation operations with simulated relocation policies in one-way car sharing systems". IEEE Transactions on Intelligent Transportation Systems, 15(4): 1667-1775.
(2) Kek, A. G. H., Cheu, R. L., and Chor, M. L. (2006). "Relocation simulation model for multiple-station shared-use vehicle system". Transportation Research Record: Journal of the Transportation Research Board, 1986: 81-88.
(3) Boyaci, B., Zografos, K. G., Geroliminis N. (2014, pp. 735). "An optimization framework for the development of efficient one-way car-sharing systems". European Journal of Operational Research, 240: 718-733.
(4) Boyaci, B., Zografos, K. G. and Geroliminis N. (2017, pp. 215). "An integrated optimization-simulation framework for vehicle and personnel relocations of electric carsharing systems with reservations". Transportation Research Part B: Methodological, 95: 214-237.
(5) Weikl, S. (2016, pp. 30). "A Mesoscopic Relocation Model for Free-Floating Carsharing Systems". Working paper. Institute for Transport and Regional Planning Department of Traffic Engineering, Munich.

Objective: demand forecast in the city of Milan

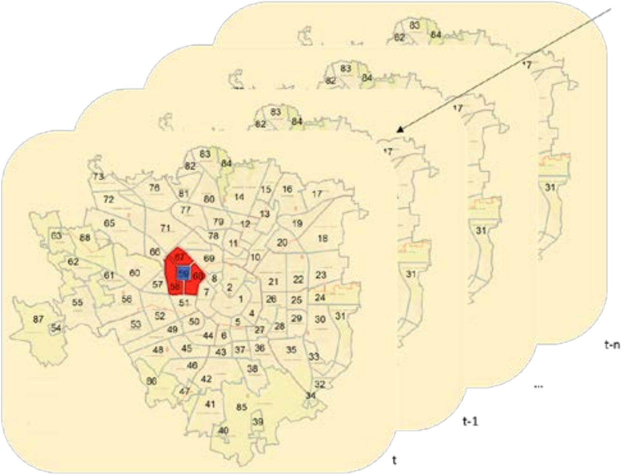
Objective: **forecast the demand** (i.e. the total number of bookings) **for a 1-hour period "t" in all the 86 areas of Milan** in which car sharing services are available, using historical booking data (from "t-1" to "t-n") of **each area** and its **adjacent areas**.

Dataset building. The original dataset contained a list of **810.000 rentals for all car sharing services in Milan**, corresponding to all trips in the month of May 2017. For each rental, the following data was available: 1) origin and destination area; 2) origin and destination location coordinates; 3) time of origin and arrival; 4) duration time.

To create the final dataset, for each of the 86 areas in which car sharing services are available the **80 variables below were calculated** for each one-hour period in the month of May 2017.

In particular, the following 40 variables were calculated both **for the area considered** and **for the areas adjacent to the one considered**: **A)** 24 variables representing the number of bookings in each of the previous 24 one-hour time periods; **B)** 7 variables representing the total number of bookings in the previous 24, 48, 72, 96, 120, 144 and 168 hours respectively; **C)** 7 variables representing the number of bookings in the same on-hour period in each of the previous 7 days; **D)** 2 variables representing the population of the area and the number of bookings in the considered one-hour period.

Figure 2: spatial and temporal division used for the analysis of car sharing demand

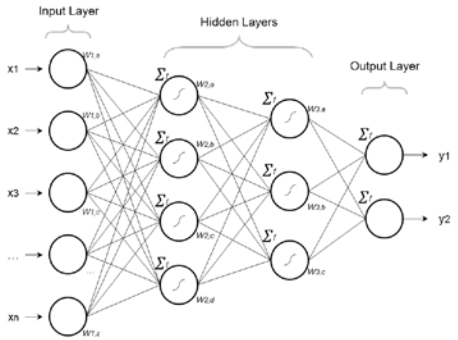


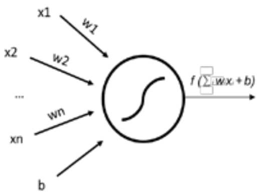
(1) The dataset didn't include any information about CS users or the CSOs that provided the service. However, because the CSO's financial performance and private information about users might be deducted from the data, the dataset is not sharable and the thesis cannot be published. Privacy matters were agreed by professor E. Borgonovo and Mr. E. Curto from Urbi for not sharing the dataset.

Method: Deep Neural Networks

Definition. **Machine learning** is an Artificial Intelligence (AI) discipline that aims to create **algorithms that are capable of learning from various type of examples presented to them**. One type of machine learning algorithms that had major success recently are **Neural Networks**[1] (NN), which consist in a series of interconnected layers formed by single units called neurons. Networks formed by hundreds of hidden layers are know as Deep Neural Networks, and the discipline that studies them is called **Deep Learning**.

Figure 1: simple form of neural network





Neurons. Each neuron is connected to the others in a series of synaptic connection that enable them to exchange information. From one layer to the other, neurons compute a weighted sum of their inputs and pass the result to the next layer. When the network tries to **find relationships between the input and output layer**, it modifies these weights to minimize the error, defined as the difference between the training data and the desired output [2].

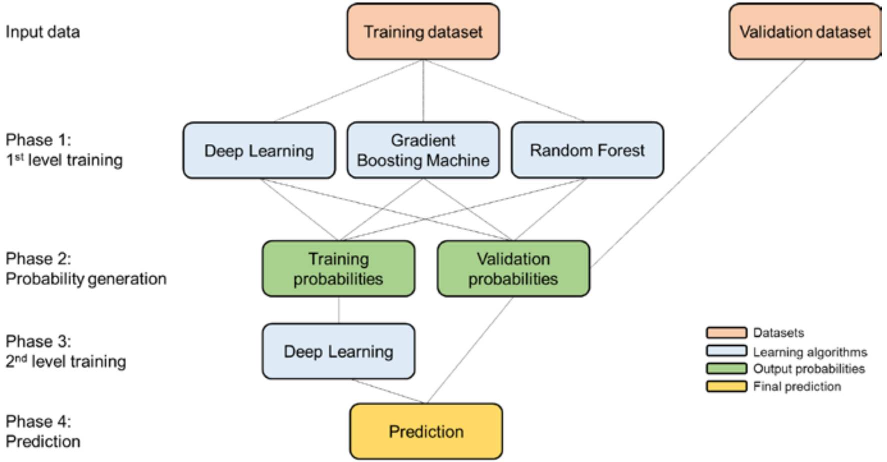
Backpropagation. During **training**, internal **parameters adjust by themselves to minimize the error** between the input variables and the output examples. This method, which enables feature extraction, is called backpropagation[3]. After extracting the features from the examples shown to it, the network is tested on a new set of data that was never shown to it in order to test his ability to predict (or categorize). This process is called **validation**.

Performance. If the error achieved during training and validation are sufficiently low, the model can be used in a real-case scenario. Deep Neural Networks (DNN) achieved incredible results in **finding both linear and non-linear relationships** in the vast sets of data in relatively little amounts of time **without the need of creating complex mathematical models from scratch**. They were able to resolve very complex problems with little need of human input [4].

(1) Bishop, C. M. (1995). "Neural networks for pattern recognition". Clarendon Press, Oxford.
(2) LeCun, Y., Bengio, Y. and Hinton, G. (2015). "Deep learning". Nature, 521: 436–444.
(3) Deng, L. and Yu, D. (2013). "Deep Learning Methods and Applications". Foundations and Trends in Signal Processing, 7(3–4): 197–387.
(4) Bengio, Y. (2016). "Machines who learn". Scientific American, 314: 46-51.

Deep Neural Network structure

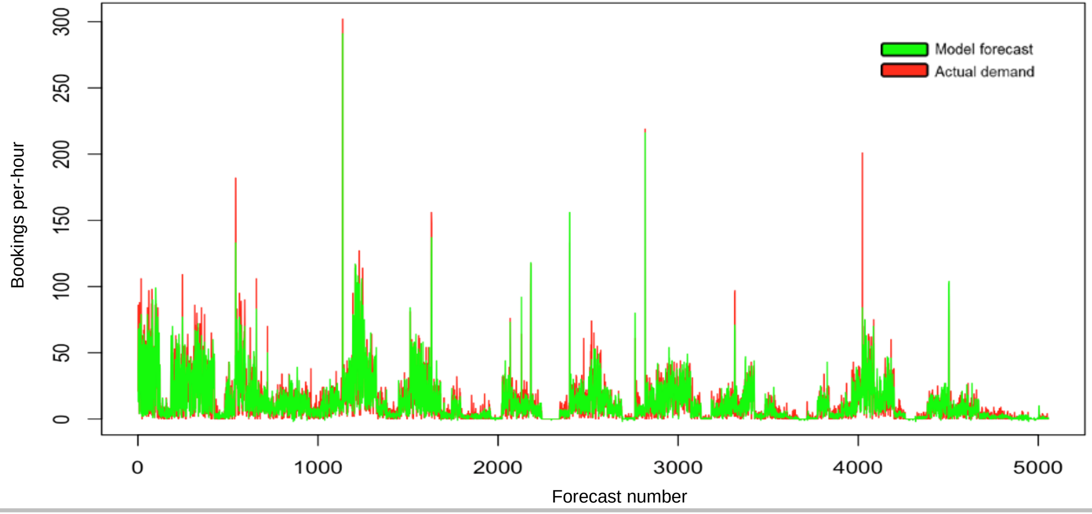
An open-source software called **H2O.ai** was used to create the forecasting tool. The first DNN structure created achieved a root mean square error (RMSE) of 4,69 (cars/hour). To lower the error, stacking was implemented. **Stacking** [1] allows to ensemble different machine learning algorithms. By implementing different learning techniques all at once, it is possible to achieve better performances compared to using one single learning method. **The final forecasting tool achieved an RMSE of 4,04** (cars/hour). A **visual representation** of this final model is presented below.



(1) Deng, L. and Yu, D. (2013). "Deep Learning Methods and Applications". Foundations and Trends in Signal Processing, 7(3–4): 197–387.

Deep Neural Network performance

The figure below compares the demand forecast of the tool with the actual demand for car sharing. It is evident that the **DNN created performs remarkably well in forecasting the demand for car sharing**. Still, the tool is not able to catch some **peaks in the demand**, which **compromises its accuracy** and therefore the possibility of using it on real car sharing operations. The reason behind these results is to be found in the **relatively small amount of data** available (1 month of car sharing trips in the city of Milan) not on the tool characteristics.



Comparison with previous NN forecasting tools

The table below compares the DNN created with previous attempts to forecast the demand for car sharing using neural networks to clarify which model is best suited to be used on a large car sharing system. The characteristics of each model are compared and rated using a three-color scale. In particular, the following **model characteristics** are highly rated: a) developed analyzing a large car sharing system with a considerable number of average daily bookings; b) trained using a large dataset coming from a system that provides great flexibility to its users; c) high spatial and temporal resolution; d) deep not traditional neural network. **Even if the RMSE of the tool created is higher compared to the ones of past models, its characteristics make it more suitable to be applied on a large-scale car sharing system**.

Model	CS System dimension	Average trips per day	Data type	Data amount	Spatial resolution	Time resolution	Deep neural network	RMSE
Cheu et al., 2006 [1]	12 stations, 50 vehicles	23	One-way station-based	n.a.	Station	3-hours	X	2,58 trips/3-hours
Xu and Lim, 2007 [2]	13 stations, 82 vehicles	n.a.	One-way station-based	n.a.	Station	1-hour	X	2,4 trips/hour
Zhu et al., 2015 [3]	n.a.	n.a.	Taxi trips	12000 trips	Area	24-hours	V	2,5 trips/day
Alfian et al., 2017 [4]	10 stations, 100 cars	100	Round-trip	X	Station	8-hours	X	3,52 trips/8-hours
DNN tool created	All car sharing systems in Milan	26132	All types of CS systems	810000 trips	Area	1-hour	V	4,04 vehicles/hour

Suited for any large system Needs refinement Not suited for any large system

(1) Cheu, R. L., Xu, J., Kek, A. G. H., Lim, W. P., and Chen, W. L. (2006). "Forecasting of shared-use vehicle trips using neural networks and support vector machines". Transportation Research Record: Journal of the Transportation Research Board, 1968(5): 40-46.
(2) Xu, J. X. and Lim, J. S. (2007). "A New Evolutionary Neural Network for Forecasting Net Flow of a Car Sharing System". Working paper. IEEE Congress on Evolutionary Computation (CEC).
(3) Zhu, X., Li, J., Liu, Z. and Yang, F. (2015). "Optimization approach to depot location in car sharing systems with big data". Working paper. IEEE Congress on Big Data, Beijing.
(4) Alfian, G., Rhee, J., Ijaz, M. F., Syafrudin, M. and Fitriyani, N. L. (2017). "Performance Analysis of a Forecasting Relocation Model for One-Way Carsharing". Applied Sciences, 7(6): 598.

GRAZIE PER LA VOSTRA ATTENZIONE!