# Week 3. Big data; machine learning

N.M. Milovantseva, Ph.D.
School of World Economy
NRU HSE

# The rise of big data

- Recent availability of “big data” has transformed business
- Businesses have been very successful in solving private-sector problems using technology and big data
- Big data can also be used in social science for research in economics, political science, sociology, etc.

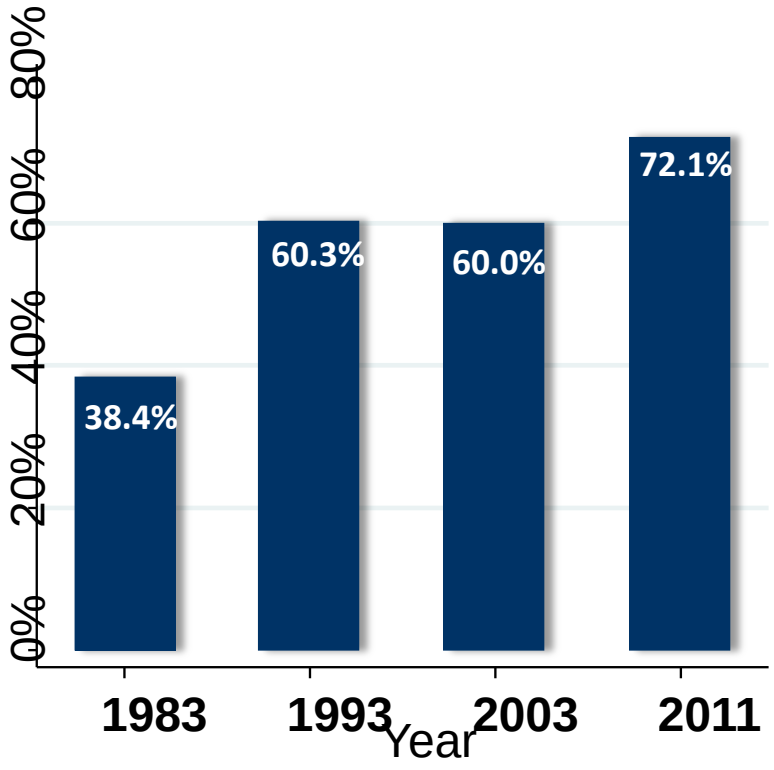
# Economic theories

- Economics has long been a theoretical field
  - Develop mathematical models
  - Use these theories to explain patterns and try to make policy recommendations for improvements
- Problem: untested theories
  - leads to a politicization of questions that in principle have scientific answers

# The rise of big data in research

- Large datasets are starting to transform social sciences
- Test and improve theories using real-world data
- Examples of big data sources
  - Government data: tax records, Medicare
  - Corporate data: Facebook, retailer data
  - Unstructured data: Twitter, newspapers

**Empirical articles in leading economics journals, 1983-2011**

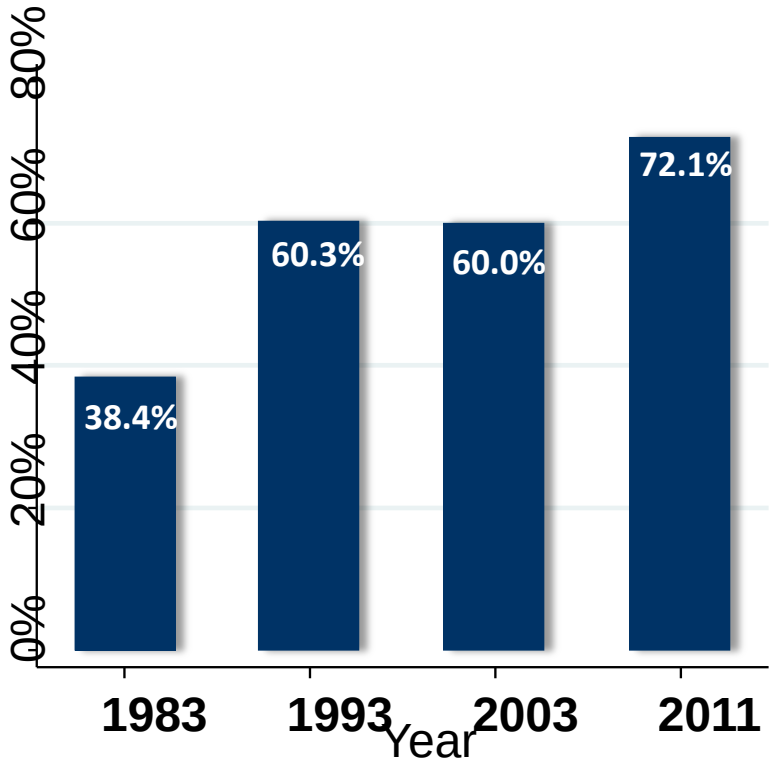



*Source: Hamermesh (JEL 2013)*

# The rise of big data in research - cont.

- Social science research
  - International relations: news from different sources, statements by officials, exchange rates and shares
  - Economy - transactions of individuals and legal entities, macroeconomic indicators, company reporting
- Business
  - Logistics - data on cargo transportation, warehouse utilization, retail sales
  - Marketing - the behavior of users on the site, online purchases and search requests, electronic correspondence, likes, comments, posts on social networks and blogs

**Empirical articles in leading economics journals, 1983-2011**




*Source: Hamermesh (JEL 2013)*

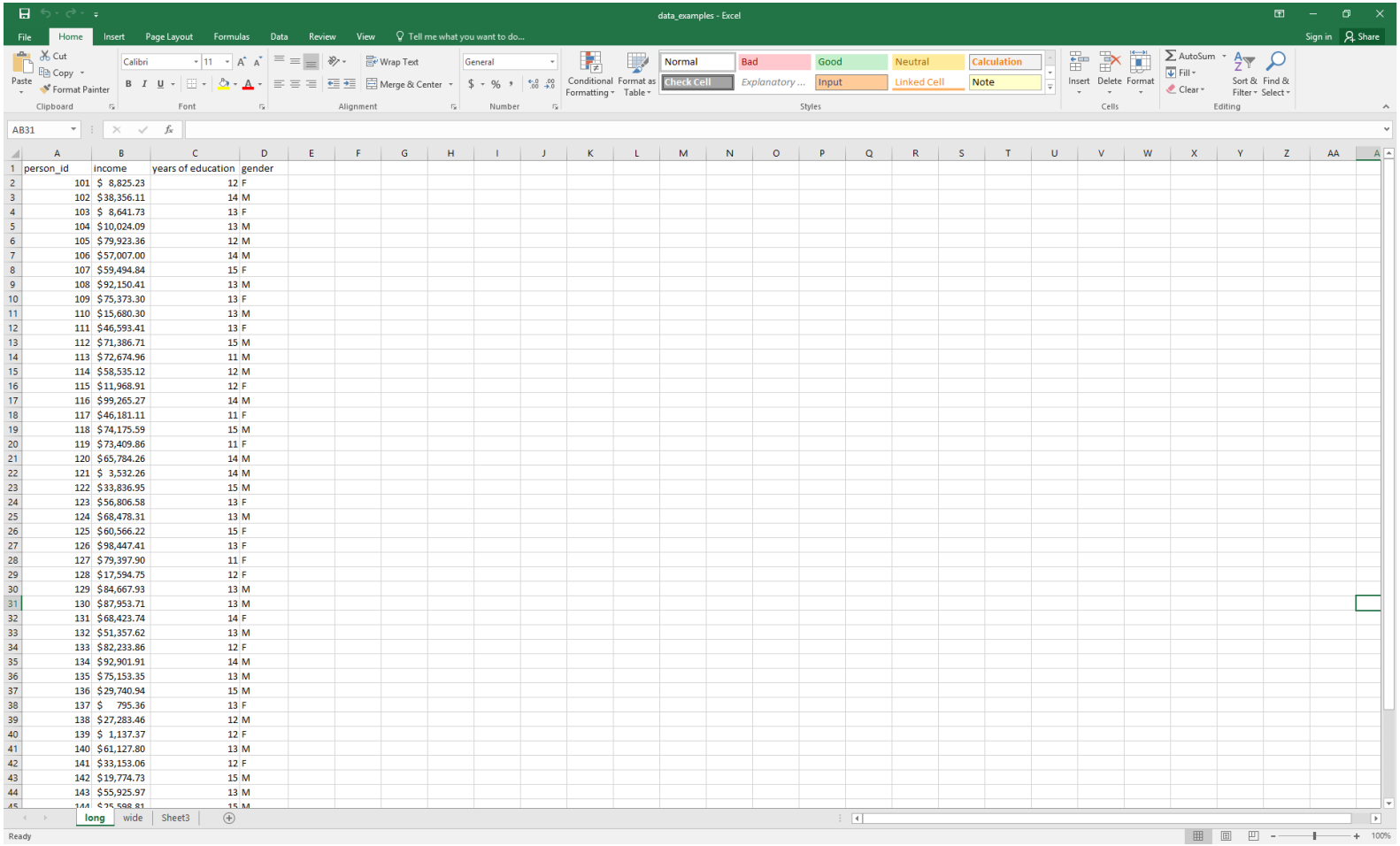
# Why is big data transformational?

- Highly reliable data on a large scale
- Ability to measure new variables
- Universal coverage
  - can “zoom in” to subgroups
- Large samples
  - can approximate scientific experiments

# Two types of big data

- “Long” data: many observations relative to variables (e.g., tax records)
- “Wide” data: few observations relative to variables (e.g. Amazon clicks, newspapers)

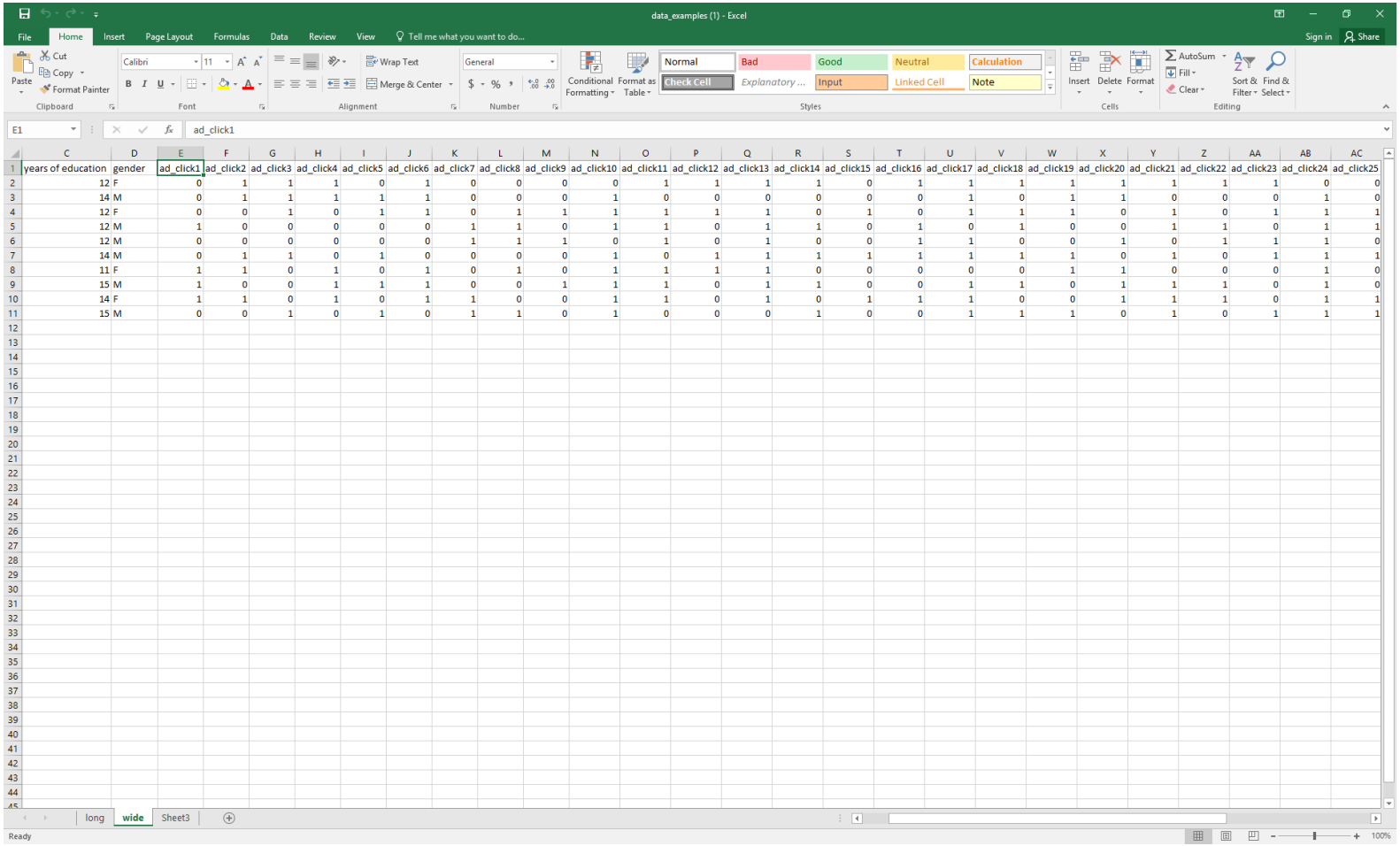
# “Long” data



data_examples - Excel

| person_id | income | years of education | gender |
|---|---|---|---|
| 101 | $ 8,825.23 | 12 | F |
| 102 | $38,356.11 | 14 | M |
| 103 | $ 8,641.73 | 13 | F |
| 104 | $10,024.09 | 13 | M |
| 105 | $79,923.36 | 12 | M |
| 106 | $57,007.00 | 14 | M |
| 107 | $59,494.84 | 15 | F |
| 108 | $92,150.41 | 13 | M |
| 109 | $75,373.30 | 13 | F |
| 110 | $15,680.30 | 13 | M |
| 111 | $46,593.41 | 13 | F |
| 112 | $71,386.71 | 15 | M |
| 113 | $72,674.96 | 11 | M |
| 114 | $58,535.12 | 12 | M |
| 115 | $11,968.91 | 12 | F |
| 116 | $99,265.27 | 14 | M |
| 117 | $46,181.11 | 11 | F |
| 118 | $74,175.59 | 15 | M |
| 119 | $73,409.86 | 11 | F |
| 120 | $65,784.26 | 14 | M |
| 121 | $ 3,532.26 | 14 | M |
| 122 | $33,836.95 | 15 | M |
| 123 | $56,806.58 | 13 | F |
| 124 | $68,478.31 | 13 | M |
| 125 | $60,566.22 | 15 | F |
| 126 | $98,447.41 | 13 | F |
| 127 | $79,397.90 | 11 | F |
| 128 | $17,594.75 | 12 | F |
| 129 | $84,667.93 | 13 | M |
| 130 | $87,953.71 | 13 | M |
| 131 | $68,423.74 | 14 | F |
| 132 | $51,357.62 | 13 | M |
| 133 | $82,233.86 | 12 | F |
| 134 | $92,901.91 | 14 | M |
| 135 | $75,153.35 | 13 | M |
| 136 | $29,740.94 | 15 | M |
| 137 | $ 795.36 | 13 | F |
| 138 | $27,283.46 | 12 | M |
| 139 | $ 1,137.37 | 12 | F |
| 140 | $61,127.80 | 13 | M |
| 141 | $33,153.06 | 12 | F |
| 142 | $19,774.73 | 15 | M |
| 143 | $55,925.97 | 13 | M |
| 144 | $25,598.81 | 15 | M |

# “Wide” data



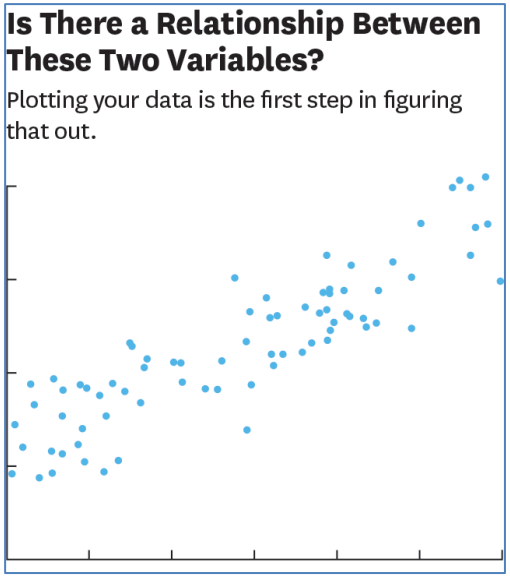
| years of education | gender | ad_click1 | ad_click2 | ad_click3 | ad_click4 | ad_click5 | ad_click6 | ad_click7 | ad_click8 | ad_click9 | ad_click10 | ad_click11 | ad_click12 | ad_click13 | ad_click14 | ad_click15 | ad_click16 | ad_click17 | ad_click18 | ad_click19 | ad_click20 | ad_click21 | ad_click22 | ad_click23 | ad_click24 | ad_click25 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 12 | F | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 |
| 14 | M | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 |
| 12 | F | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 |
| 12 | M | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 1 |
| 12 | M | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 0 |
| 14 | M | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 |
| 11 | F | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 |
| 15 | M | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 0 |
| 14 | F | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 1 |
| 15 | M | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 |

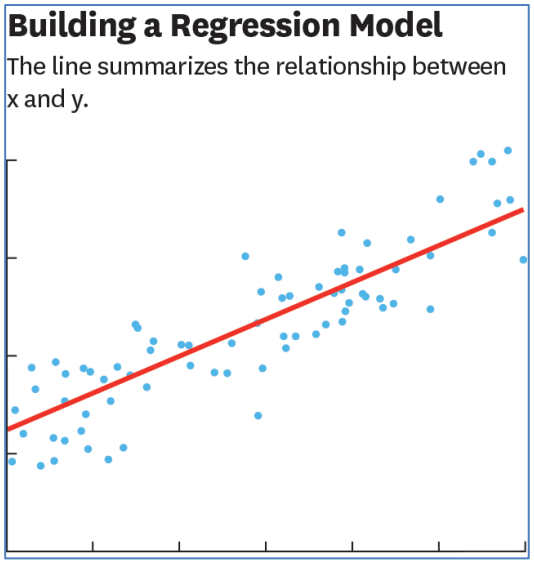
# Methods

- Social science has focused on “long” data
  - Application: identifying causal effects
  - Example: effects of improving schools on income
- Computer science has focused on “wide” data
  - Application: prediction
  - Example: predicting income to target ads

# Regression analysis

- Answers the questions:
  - Which factors matter most?
  - Which can be ignored?
  - How do those factors interact with each other?
  - How certain are we about all of these factors?



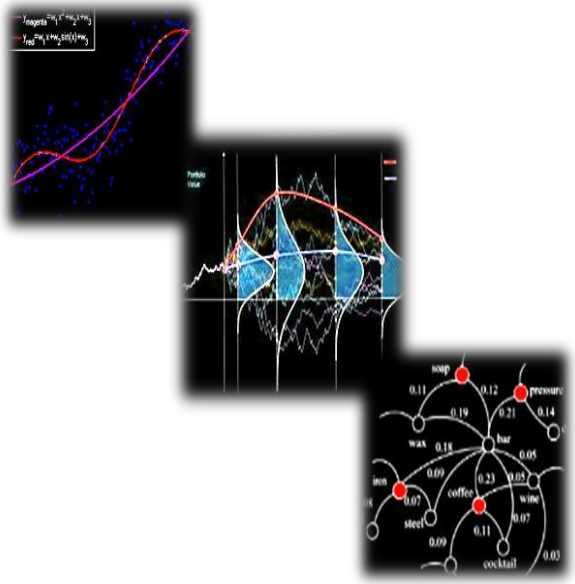

Red line - best explanation of relationship between independent and dependent variables

The method of **least squares**

$Y = 200 + 5X + \text{error term}$

# Machine learning

- Machine learning - a division of computer science engaged in developing methods for automatically searching hidden dependencies in data




Mathematical methods

+

Technology

# ML problem formulation

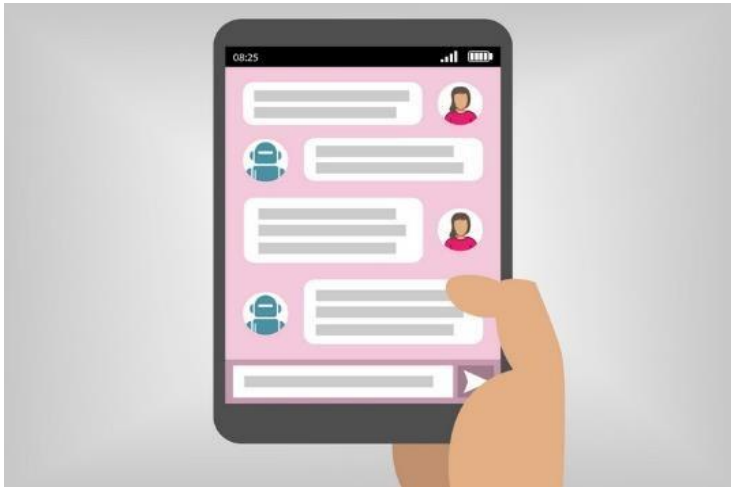
- **Data** - collection of objects
- **Objects** - are described by a set of observables and target variables
- **Observed variables** - can easily be measured for an arbitrary object
- **Target variables** - are known for a limited number of objects (the so-called training sample)
- **Task** - to predict the value of target object variables from the observed variables

# Simple example of ML problem

- Data - the aggregate of borrowers
- Objects –borrower, described by a set of documents (*observable variables*) and a binary variable, whether he repaid the loan or not (the *target variable)*
- Task – to understand from submitted set of documents, whether to issue a loan to an applicant

# More complex problem

- Data - a set of questions and answers
- Object is a pair of question (*observable variable*) and answer (*target variable*)
- Task - to build an algorithm for answering questions automatically



# What ML needs to solve a problem

- Big data
  - the more objects with known *target variable*, the more accurately prediction algorithm is constructed
  - the more data, the *more complex target variables space can be*

- Computing power
  - Great volumes of processed data need speedy hardware and new mathematical methods
  - Modern computers can easily process large data sets

# Artificial intelligence

- Replication of human intelligence in computers
- Developers introduce number of rules that the computer needs to follow
- The computer has a specific list of possible actions, and make decisions based on those rules

# ML vs AI

- Machine learning - the ability of a machine to learn using large data sets
  - ML allows computers to learn *by themselves* taking advantage of the processing power of modern computers
- Artificial Intelligence is the replication of human intelligence in computers
  - *Coded rules* are required

# Supervised vs unsupervised learning

**Supervised learning**

- Labelled data sets
- AI is given them as inputs and told of the expected outputs
- If AI gives wrong output, it readjusts its calculations iteratively until no mistakes are made
- Example: predicting weather. AI is trained on historical data. Inputs: pressure, humidity, wind speed. Outputs: temperature.

**Unsupervised learning**

- Data sets have specified structure
- AI is allowed to make logical classifications of data
- Example: predicting behavior for an e-commerce website. Instead of learning by using labelled data set of inputs and outputs, AI classifies the input data. Based on this classification, it will tell which kind of users are most likely to buy certain products.

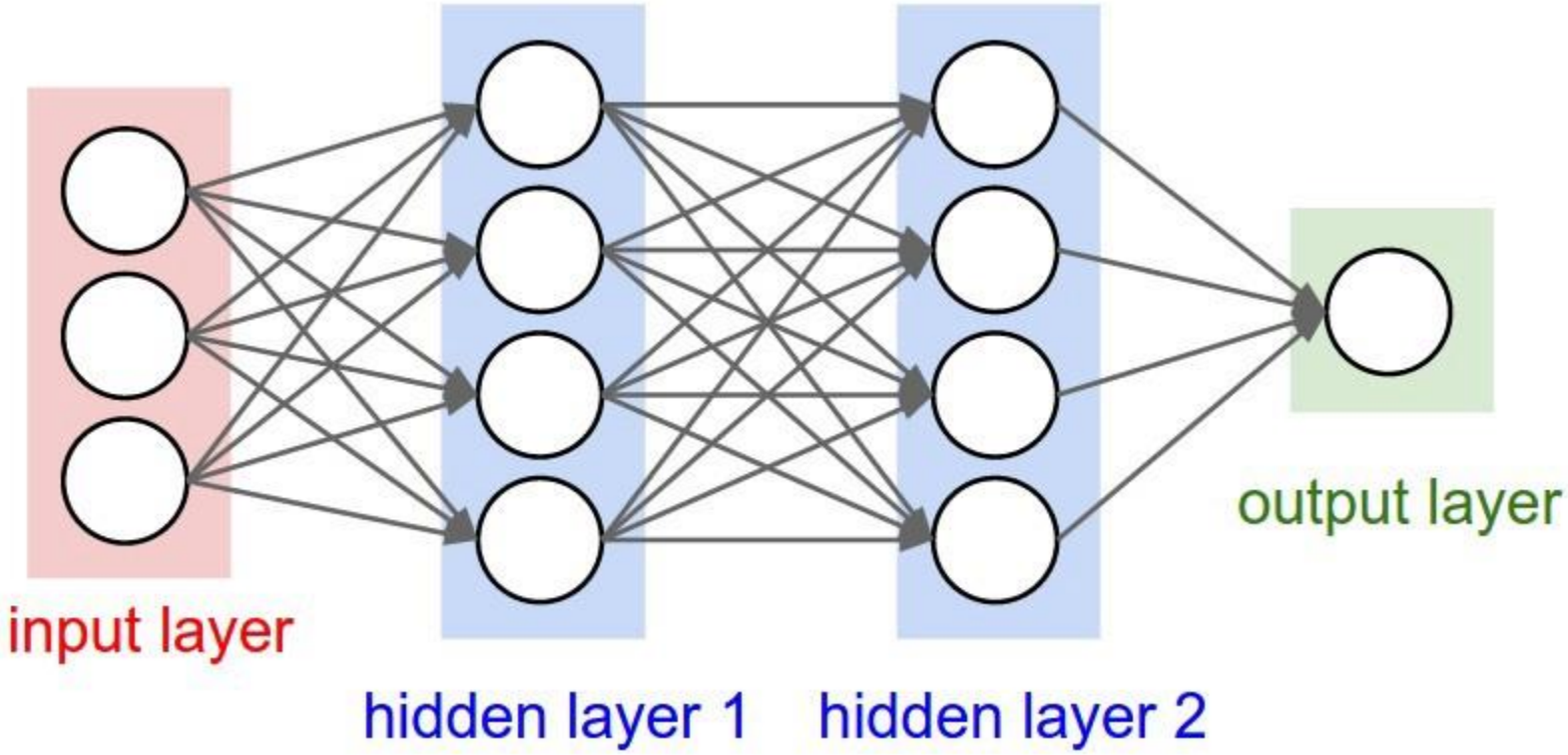
# Deep learning

- AI is trained to predict based on a given a set of inputs
- Supervised and unsupervised learning can be used
- Deep learning uses a neural network structure to imitate animal intelligence

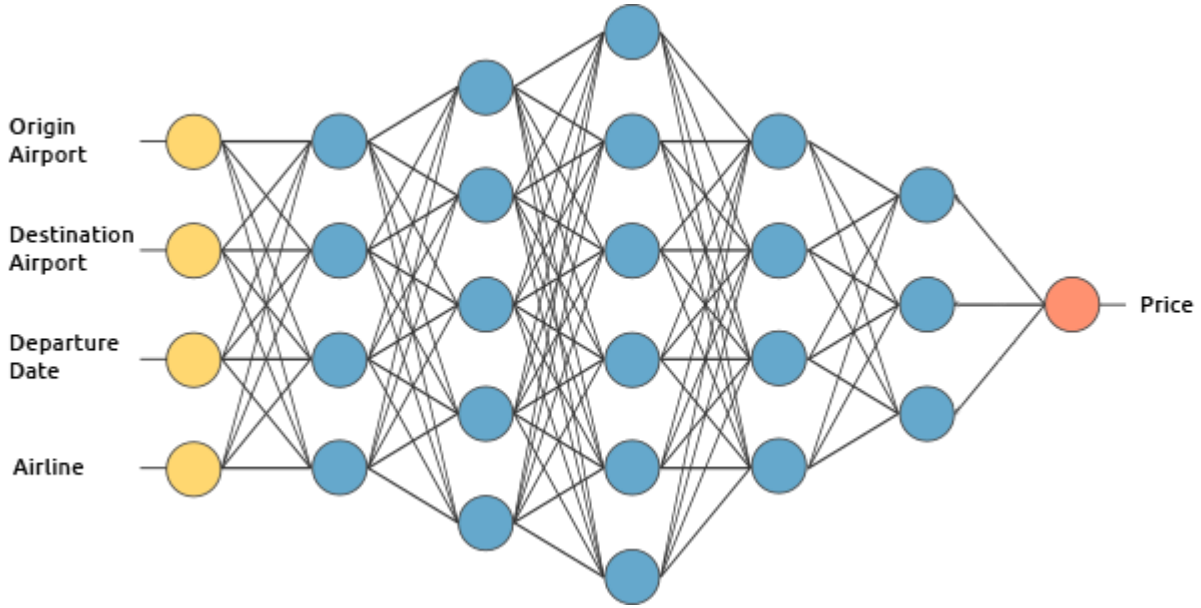
# Deep learning – sample task

- Task is to build a service that will estimate price of airline tickets
- Input: origin airport, destination airport, departure date, airline
- Output: predict the price of one-way ticket

# AI's neural networks



# AI computes price prediction by deep learning



# Training AI

- Our big data: historical data of ticket prices. Need a very large list of ticket prices because of the large amount of possible airports and departure date combinations
- Untrained AI goes through entire data set
- Compare its outputs with data set
- Create a cost function
- When cost function=0 AI is trained (AI's outputs=data set outputs)

# Reducing the cost function (CF)

- Change the weights between neurons
- Gradient descent to minimize cost function
  - computes derivative (gradient) of the CF cost at a certain set of weight
- Deep Learning updates the weights using gradient descent automatically
- When CF is minimized our AI is trained and **can work as the airplane ticket price estimation service**

# Neural networks used for:

- *Convolutional neural networks* - for computer vision
- *Recurrent neural networks* - for natural language processing

# Key points of AI

- Deep learning is a technique of ML
- Deep learning uses neural network (NN)
- Three types of layers of neurons in NN
- Neurons' importance is dictated by weights
- Iterating through data set reduces AI's error

# AI's future

- A "game-changing" technology
- Singularity
  - the point where artificial super-intelligence surpasses human intelligence
  - still "relatively" far away

# Challenges and opportunities created by new AI technology

- AI is more than
  - a new technology to manufacturing processes
  - another step in compliance.
  - a new predictive model
- AI can potentially change the world as we know it:
  - How we live
  - Work
  - Do business
  - Govern

# Back to “trusting technology”

- We have to ask questions about AI, understand how AI systems are trained, where the data is coming from, etc.
- In particular, we need to think about the values or “ethics” that structure how AI operates.
- For example
  - How do we want an autonomous car to react when confronted with an unavoidable accident?
  - Should it minimize the loss of life, even if that means sacrificing the occupants of the car or should it prioritize the lives of the occupants at any cost?
  - Alternatively, should the choice be a random one?

# What we learned

1) Advantages of big data
2) Sources, use and types of big data
3) Big data for causal inferences vs prediction
4) Regression
5) Machine learning
6) Artificial intelligence