Syllabus

Time Series Econometrics (Vasja)

1-3: Univariate Time Series

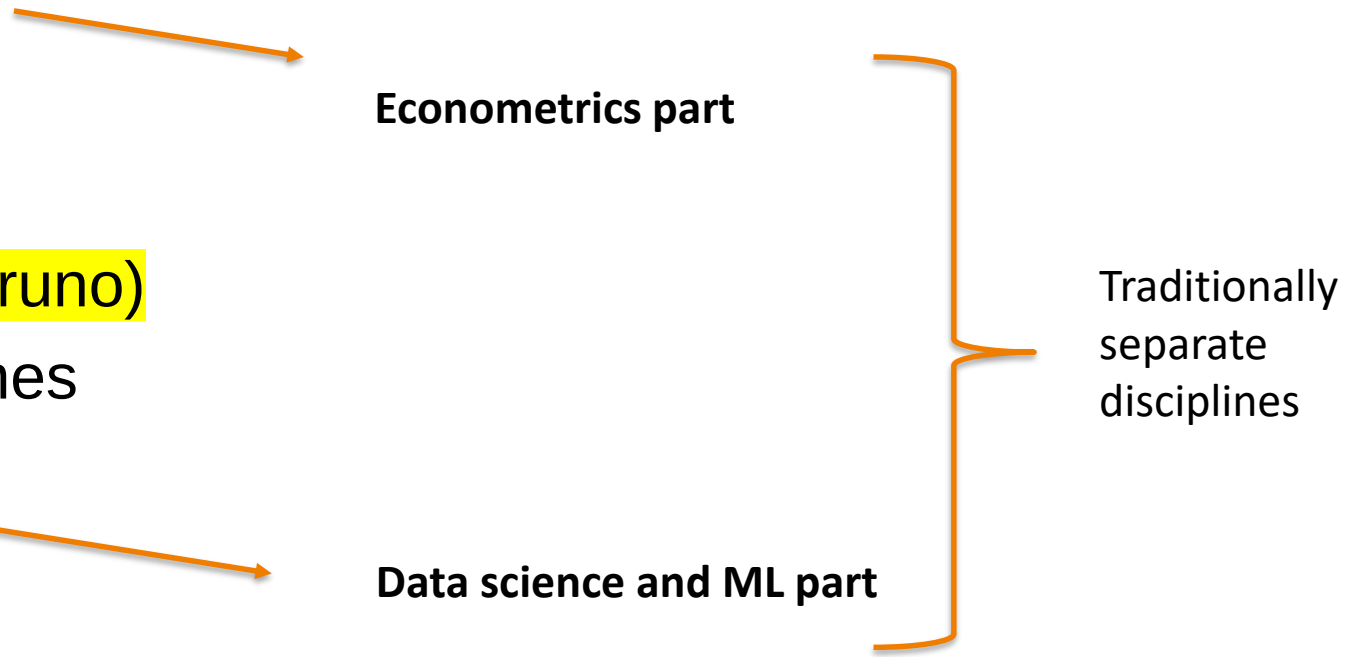
4 : Multivariate Time Series

Reproducible Analytical Pipelines (Bruno)

5-12: Reproducible Analytical Pipelines

Neural Networks (Vasja)

13-15: Basics of neural networks

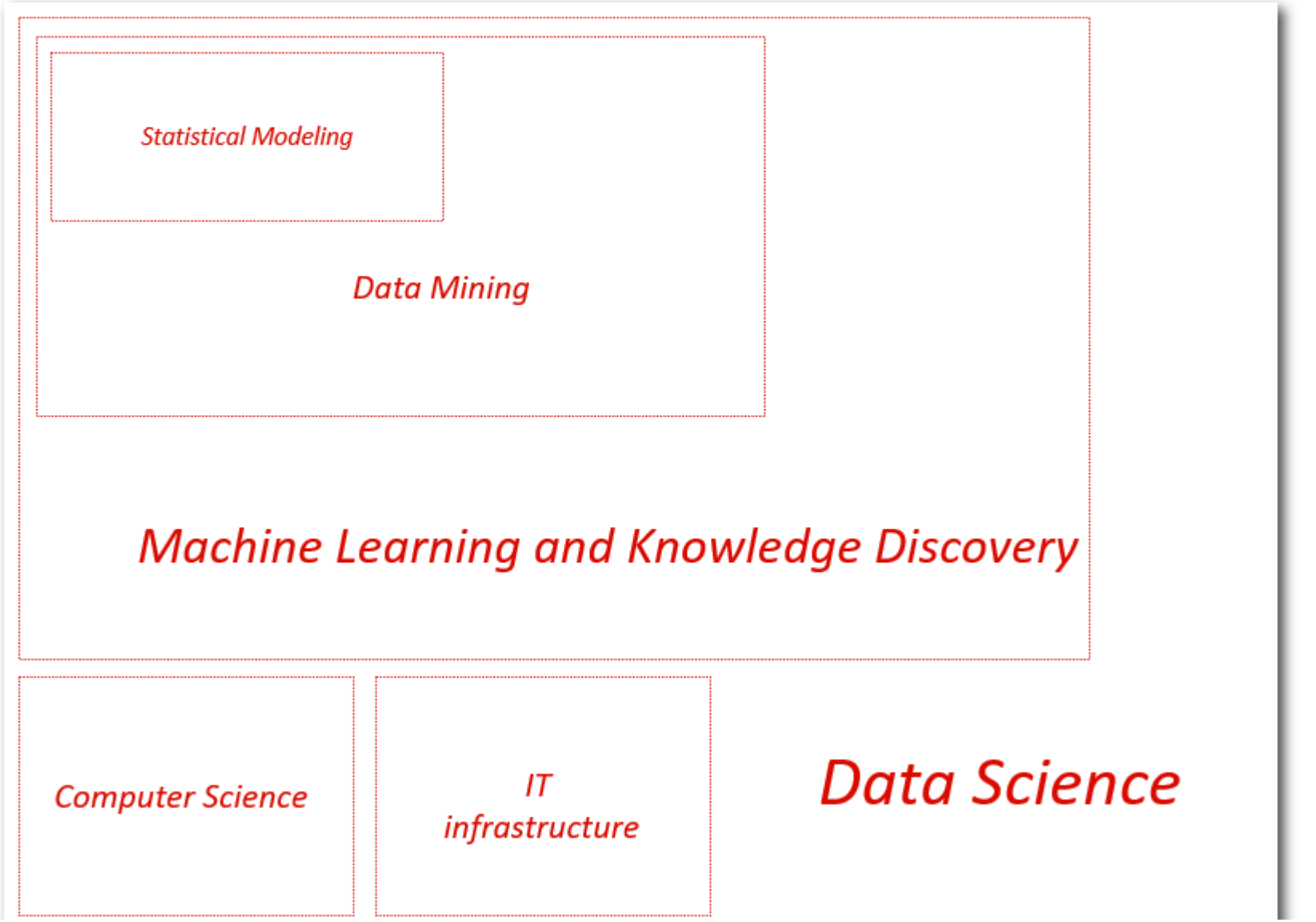


Data Science

- Data science is a broad and loosely defined term (an example from Wiki):

"Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from noisy, structured and unstructured data, and apply knowledge and actionable insights from data across a broad range of application domains."

- Covers topics like: data retrieval (web scraping), databases & data architectures, machine learning, data mining, statistics.
- Includes Big Data and Machine Learning.
- Used in industry (Amazon, Google, hedge funds,…) and public sector lately (central banks, statistical offices, European Commission, IMF,…)



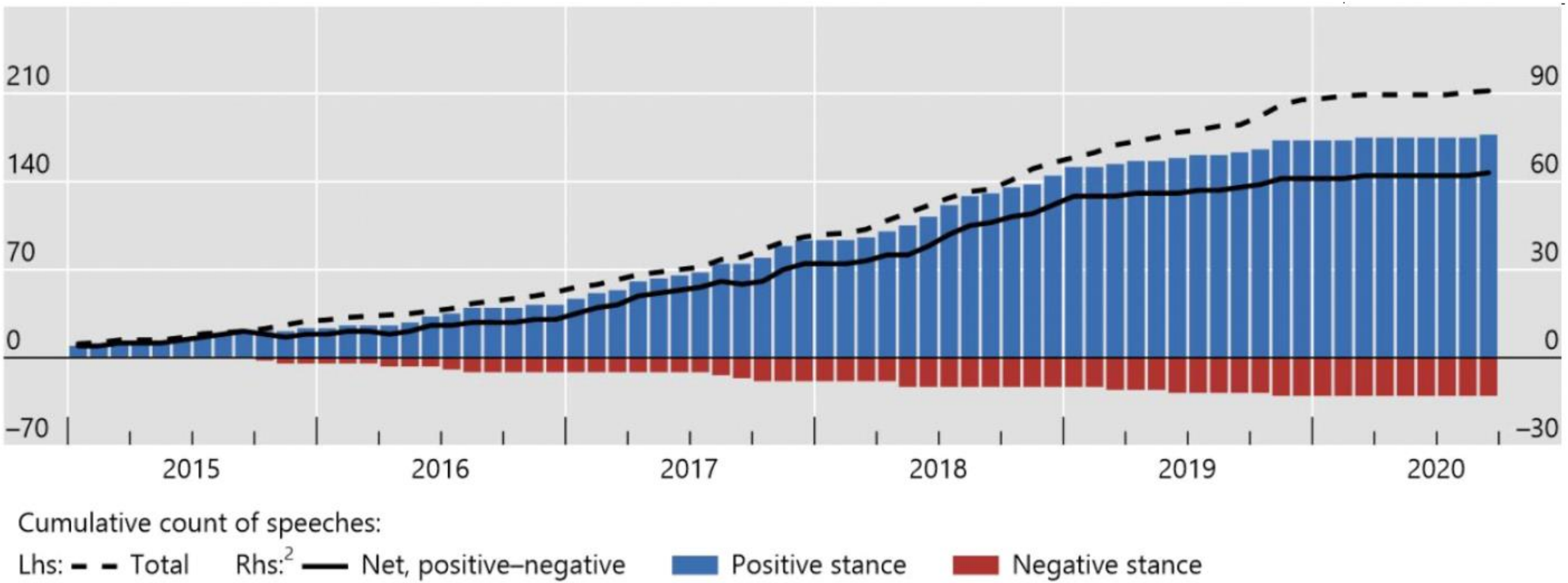
SOURCE: Data Science in Aviation presentation (Scope, opportunities and challenges in Data Science – Innaxis. David Pérez, Director)

Doerr, S., Gambacorta, L. and Garralda, J.M.S., 2021. Big data and machine learning in central banking. *BIS Working Papers*, (930).

Central banks' interest in big data is mounting[1]

Number of speeches

Graph 1



Big Data

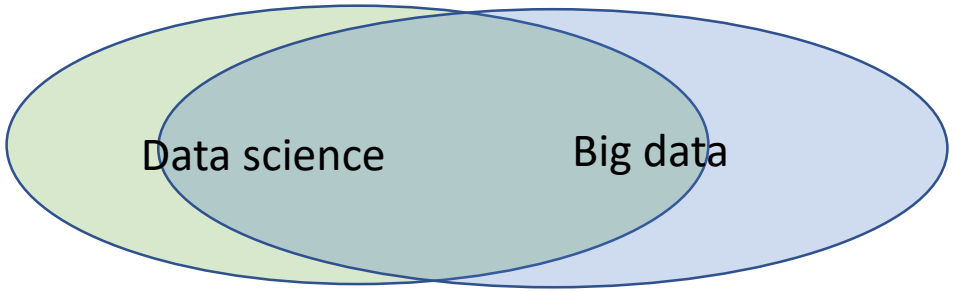
- Big Data is also a loosely defined term.
- Commonly defined with 3Vs:

Volume – larger than traditional structured data sets (e.g. set of all ticker transaction)

Velocity – data generated quickly (in minutes or even seconds)

Variety – data of different shapes (ordinal and numerical, in seconds or quarterly,…)

- Sometimes 4th V is added (veracity – often collected from open sources, unstructured and not “analysis ready”)
- Data science vs. Big data: Data science can be done on small data. Not everything in big data is data science. Substantial overlap and the division is not clear.

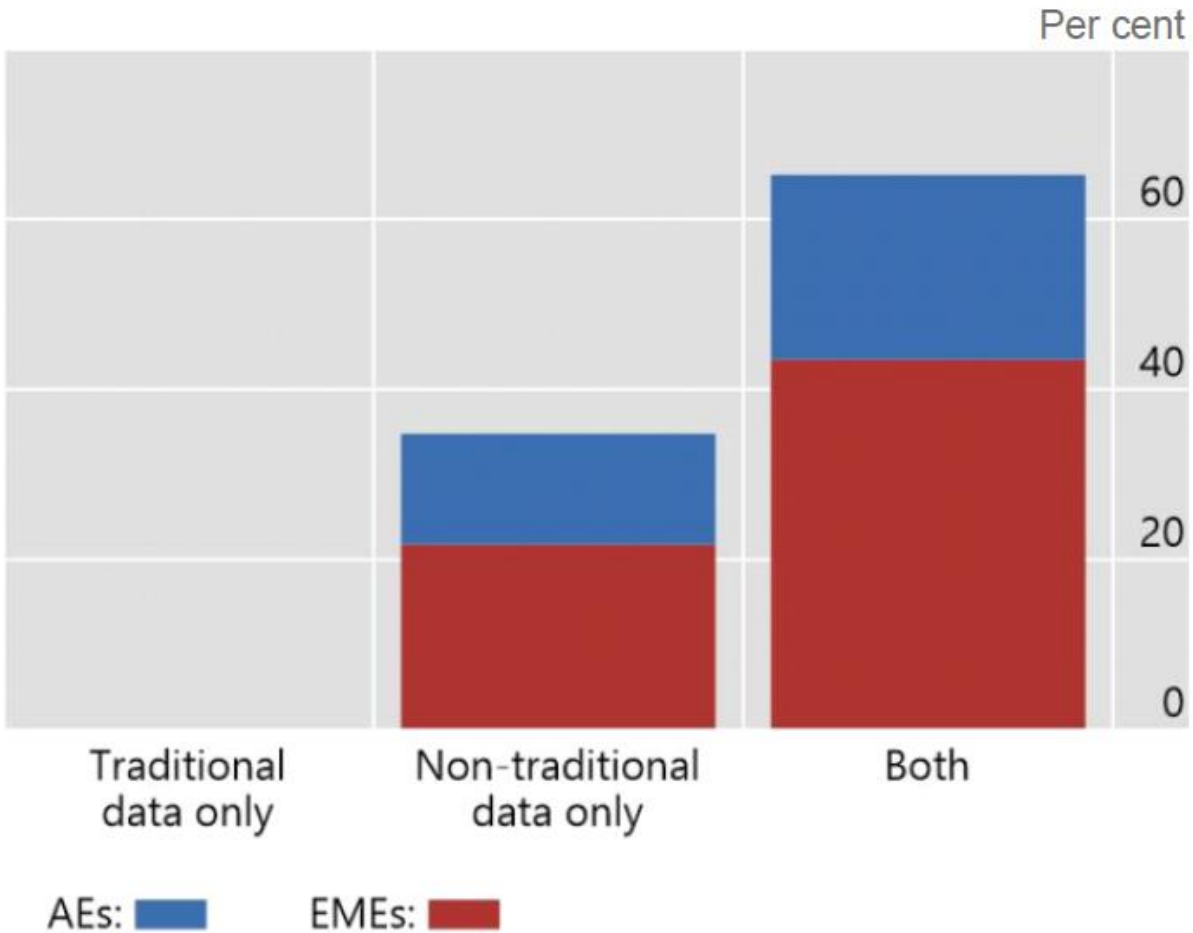


Doerr, S., Gambacorta, L. and Garralda, J.M.S., 2021. Big data and machine learning in central banking. *BIS Working Papers*, (930).

Central bank definitions of big data and main sources

Graph 2

How does your institution define big data?[1]



Word count on sources[2]

Machine Learning

- Part of AI field that focuses on learning and extracting knowledge from data:
- AI (azure.Microsoft): “Artificial intelligence is the capability of a computer system to mimic human cognitive functions such as learning and problem-solving.”
- ML (azure.Microsoft): “…the process of using mathematical models of data to help a computer learn without direct instruction. ”

Machine Learning cont.

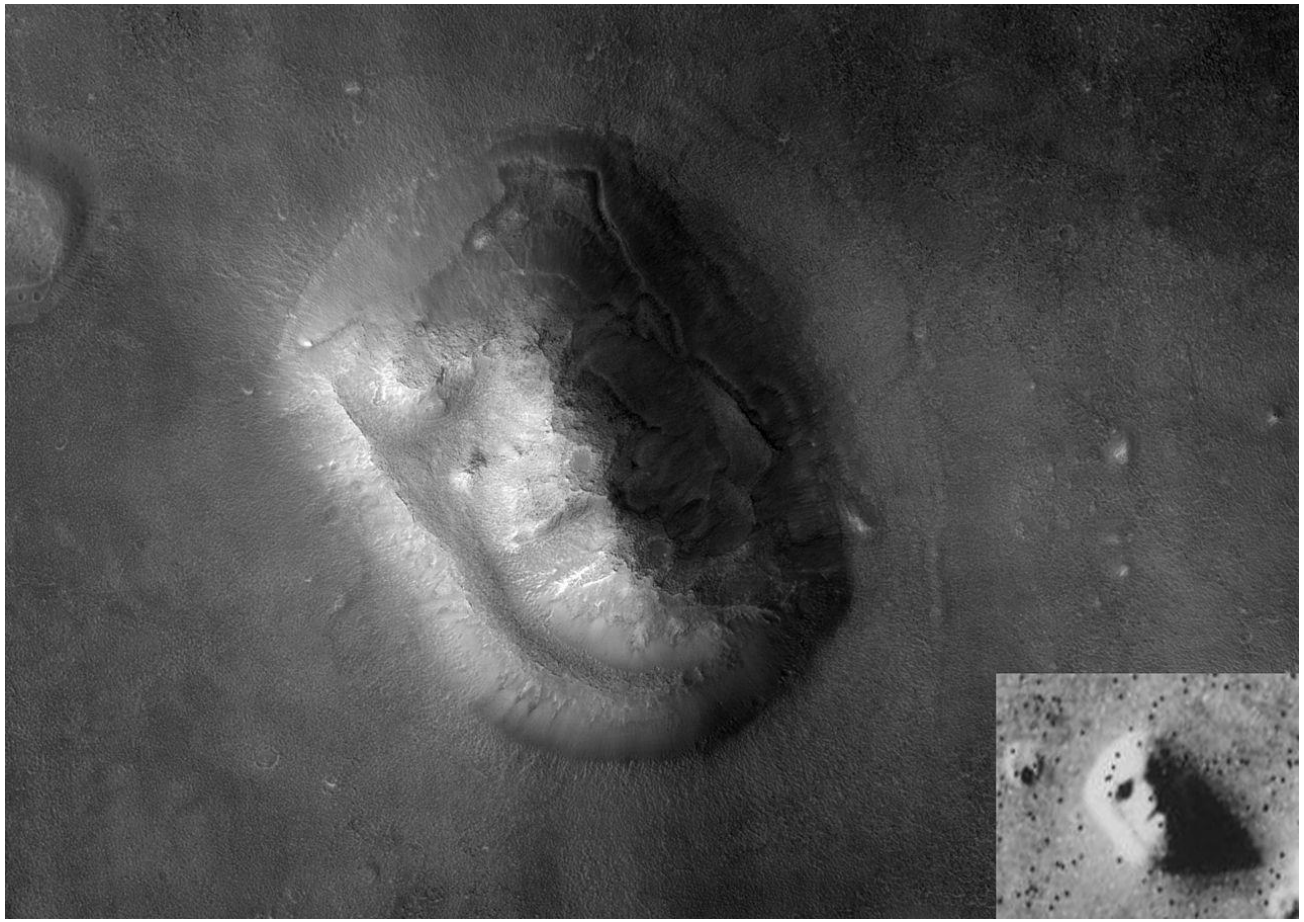
- **Unsupervised**: exploratory data analysis to find interesting patterns, problem is unknown, relies on data without a labelled target variable
- **Supervised** learning: when problem is known, take a labeled input to predict its most likely outcome
- **Classifier** (outcome is countable) or **regression** (otherwise)

- Despite the name “learning” there is no learning in the human sense, these are function minimization/maximization programs

Word of caution

- Machine Learning model does not imply causality merely correlation
- Lately have the two fields (econometrics&ML) started to overlap
- Why?
- ML was traditionally about prediction and econometrics both (prediction or causality).
- A model that predicts well can do poorly for causality and vice versa (variance&bias trade off)
- 2 examples

https://en.wikipedia.org/wiki/Cydonia_(Mars)



ML algorithm could recognize a face here. ML was about common patterns or co-movements. There is no face on Mars.

Why a model that predicts well can be poor for causality

- Sometimes it’s good to drop a noisy variable to improve prediction (smaller beta variance).
- However, that variable might be important for causal analysis (omitted variable bias).

ML gone wrong – human input is still very much needed

- Google Flu trends (https://en.wikipedia.org/wiki/Google_Flu_Trends)
- Helathcare prediction algorithm (used in hospitals and by insurance companies) used to single out patients in need w
  black patients
  (https://science.sciencemag.org/content/366/64
- Microsoft bot spew out racial insults
  (https://en.wikipedia.org/wiki/Tay_(bot)#Suspen
- NN trained to classify wolves vs. dogs was in fact
  classifier (https://arxiv.org/abs/1602.04938)

(a) Husky classified as wolf (b) Explanation

Figure 11: Raw data and explanation of a bad model's prediction in the "Husky vs Wolf" task.

	Before	After
Trusted the bad model	10 out of 27	3 out of 27
Snow as a potential feature	12 out of 27	25 out of 27

Table 2: "Husky vs Wolf" experiment results.

ML vs. Econo (Bio-Psiho-etc..) -metrics

- Traditionally the data sets were small.
- Computing capacity surged
- Too many different data -> impossible to estimate a model (non-invertibility)
- Solution to deal with big data is Machine Learning (includes regularization and other tricks to make models ”estimateable”)
- Before ML there was econometrics
- Strong overlap between the two for supervised learning (econometric models are often a building block of ML algorithms)
- E-B-P-metrics more focused on causality than forecasting. Lately causality analysis is being introduced in ML and ML methods are being introduced in econometrics.
- Machine Learning fine tunes and “tests” parameters with validation and test set. Econometrics often uses full data to estimate parameters and uses other techniques to asses the model (statistical tests)