**Week 1**

* Storing data has become incredibly cheap over recent years
* Shift in Data & Analytics: People used to look back into the past with their data. But nowadays data is used in a predictive forward-facing function – Data is being stored in the cloud and the end users can directly access and analyse data
* The four V’s of Big Data:
  + Volume:
  + Velocity
  + Variety
  + Veracity

**Parallelism & Scalability**

**Parallelism**

* Execute many independent tasks at once – for example grid search
  + **Data parallelism**: Execute the same task in parallel on different slices of the data
  + Example: Query processing in modern cloud databases which store partitions of the data on different machines
  + **Pipeline parallelism**: Break task into a sequence of processing stages – each stage takes results from previous stage as input, with results being passed downstream immediately

**Scalability**

Ability of a system to handle a growing amount of wark by adding resources to the system

* Scale-up: Replace machine with better machine (easier – especially in the cloud)
* Scale-out: Add more machine of the same type

Desired goal in practice:

* **Linear scalability with number of machines / cores** in scale out settings
* Elastic scaling in cloud environments

Scalability != Performance

* A common misperception is that scalable systems are automatically performant
* Scalability often comes with increased overheads