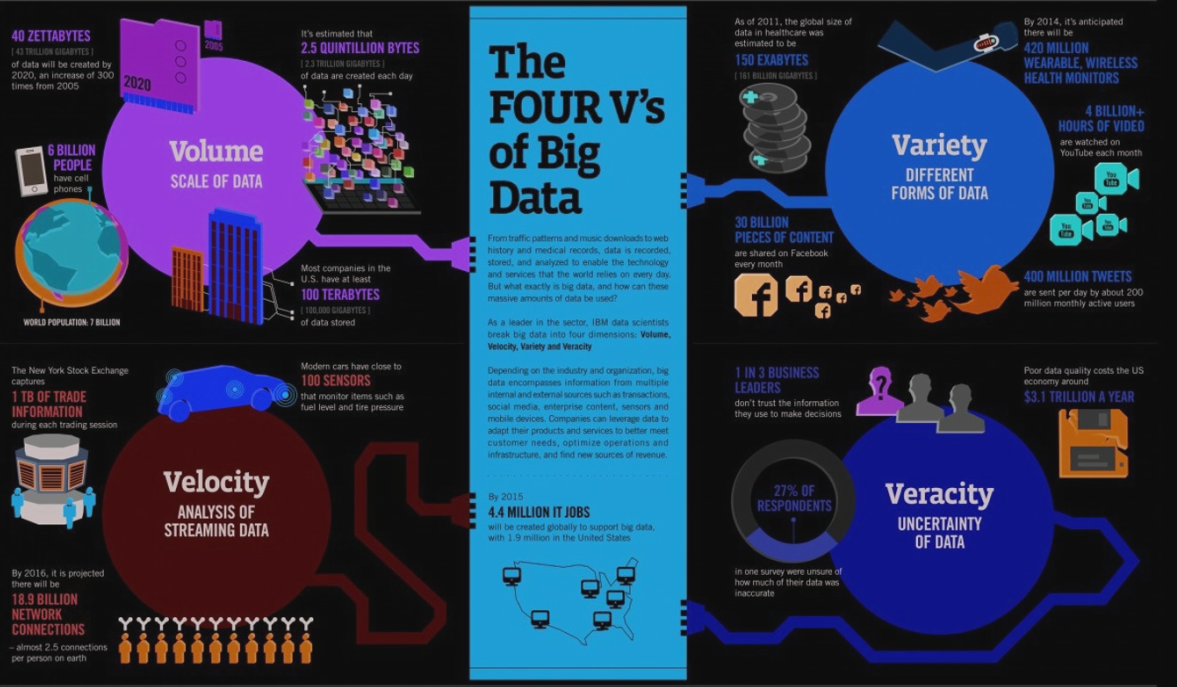
# Characterising Data

## The V’s

The first charactisations of big data were by someone with a penchant for alliteration ... others followed



Big Data- includes data sets with sizes beyond the ability of commonly used software tools

BIG DATA is ANY attribute that challenges CONSTRAINTS of a system CAPABILITY or BUSINESS NEED

* These characterise bigness, adequately
* 3V: Volume, Velocity and Variety
* Other V’s characterise problems with analysis and understanding
  + Veracity: correctness, truth
  + Variability: change in meaning over time, e.g., natural language
* Other V’s characterise aspirations
  + Visualisation: one method for analysis
  + Value: what we want to get out of the data

● Volume is size of data.

● Velocity is the frequency/Pace of incoming data that needs to be processed.

● Variety refers to different types of data.

● Veracity refers to the fact that how accurate or truthful a data set may be. More specifically, how accurate and reliable the data is?

## Metadata

Is:

* Data about data is critical to understanding
* Structured so that a computer can process

Can:

* Descriptive- Describes content for identification e.g. title, author of a book
* Structural- Documents relationships and links e.g. chapters in a book
* Administrative- Helps to manage information e.g. version number

==: Structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use or manage an information resource

### Why Use Metadata?

* Facilitate data discovery (find more discovery)
* Help users determine the applicability of the data
* Enable interpretation and reuse
* Clarify ownership and restrictions on reuse

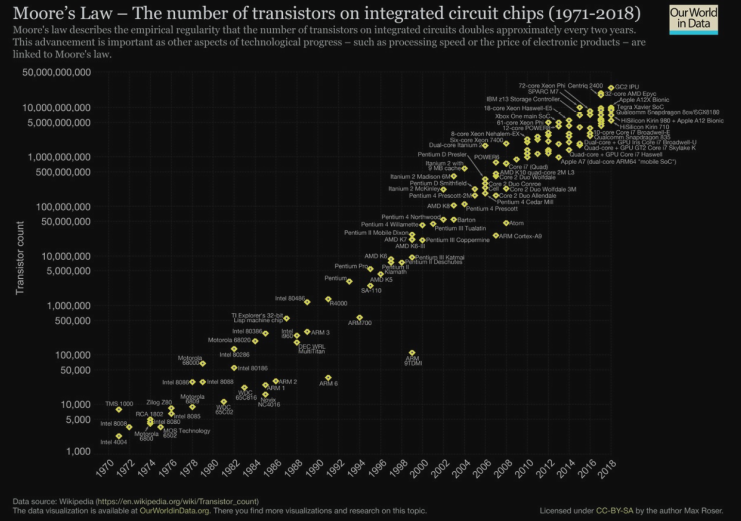
## Dimensions of data

Infographics on data dimensions

## Growth laws

Understanding the exponential growth（指数性）

### Moore’s Law



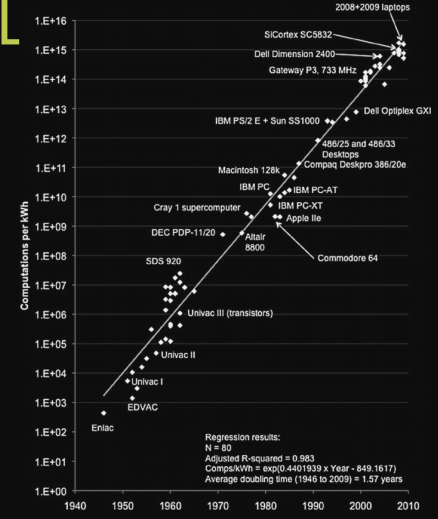
Number of transistors per chip doubles every 2 years

o More memory

o Bigger CPUs

o Faster memory, CPUs

### Koomey’s



Corollary of Moores Law

Amount of battery needed will fall by a factor of 100 every decade

Leads to ubiquitous computing

### Bell’s Law

Roughly every decade a new, lower priced computer class forms based on a new programming platform, network, and interface resulting in new usage and the establishment of a new industry

* Yes: PCs, mobile computing, cloud, internet-of things
* No: Java, big data, Hadoop, flash memory

### Zimmerman’s Law

* Surveillance is constantly increasing
* Privacy constantly decreasing

# Unix Shell

Command line interface to a Unix computer

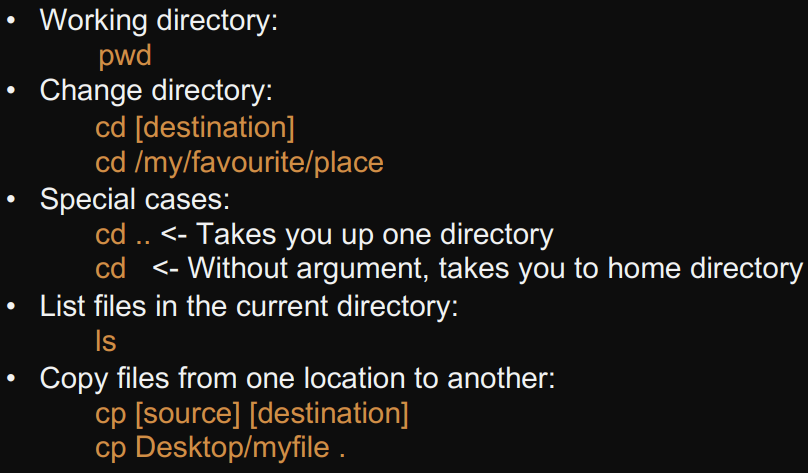
Why are shells interesting for Data Scientists?

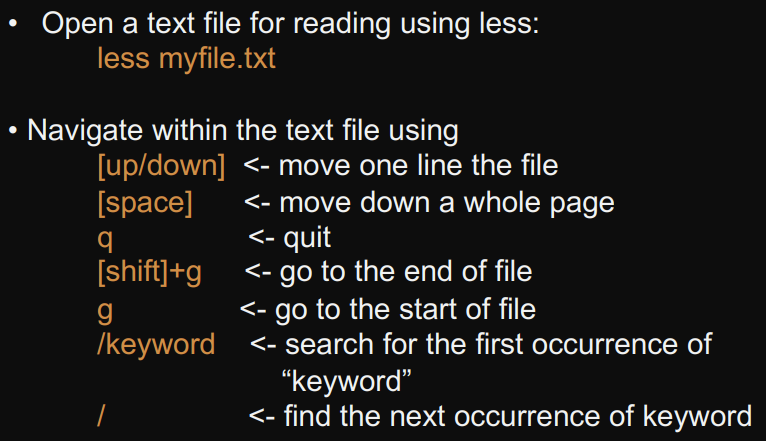
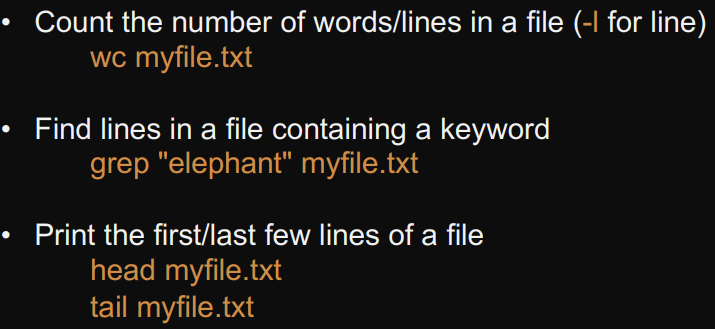
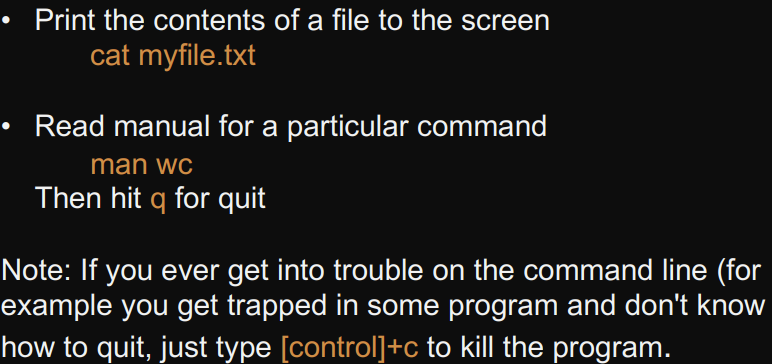
* Provide powerful & easy way to manipulate large data files
* And move data around a network

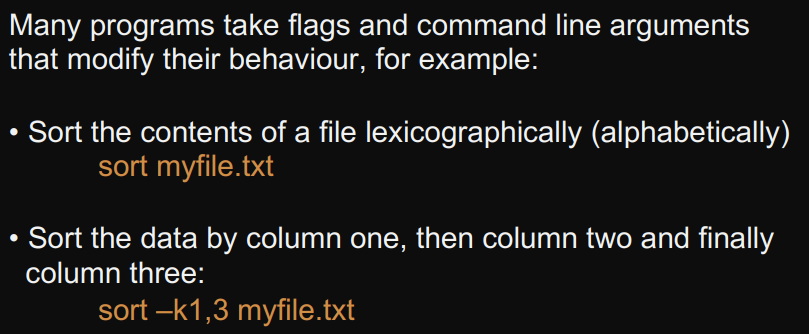
Super-computers are typically UNIX based

* Easier to manipulate and wrangle Big Data
  + Simple and easy to learn.
  + Ideal for textual data, e.g. unstructured data for social networks, life sciences, system logs, etc.
  + Quick to sort, search, match, replace and clean your data.
* Explore data before you use it in Python or R

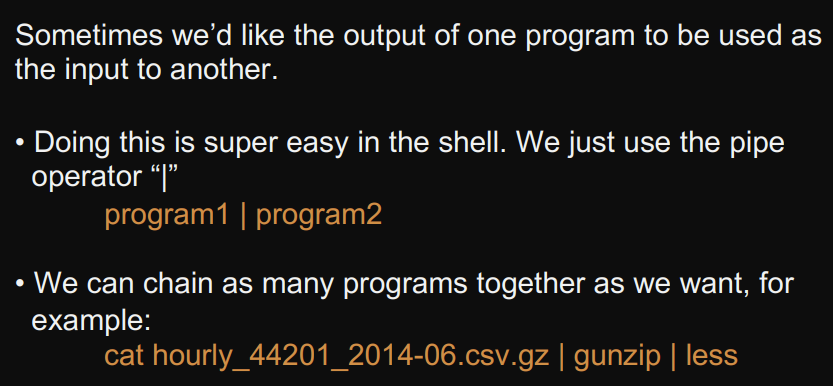
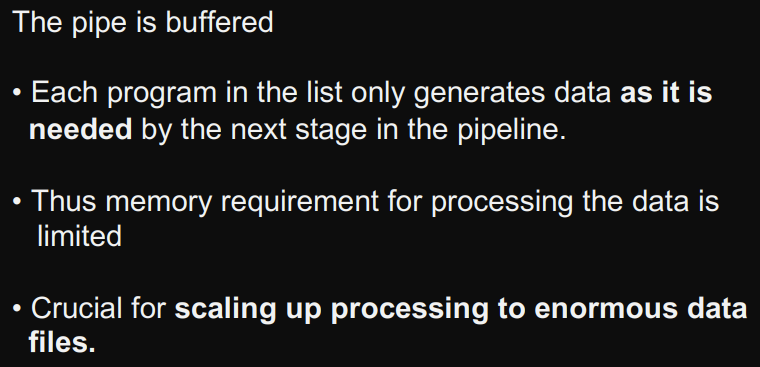
Navigating the Filesystem



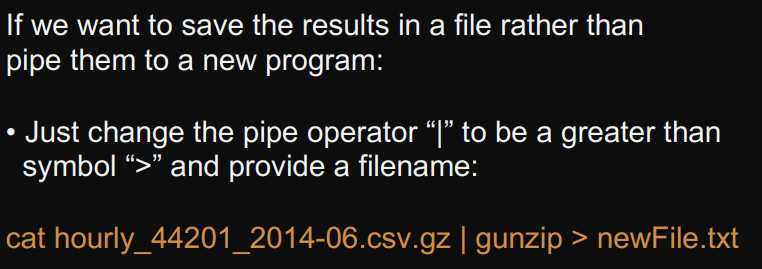
Reading a Text File  

Flags and Arguments

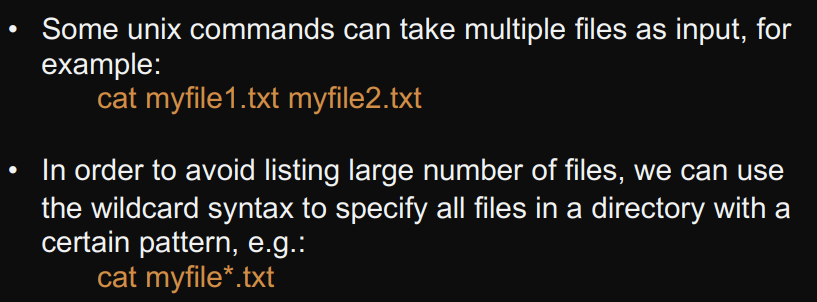
Pipes

Redirects



Wildcards



awk

