

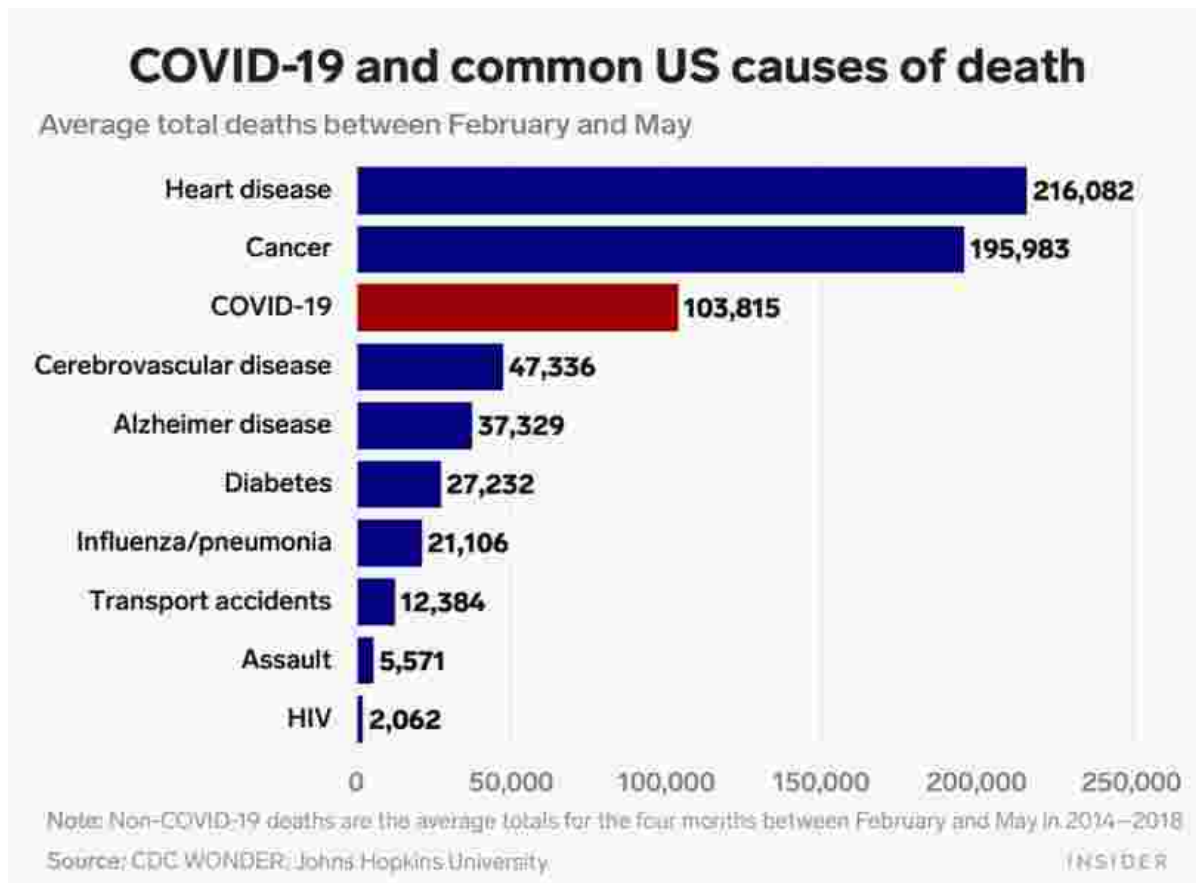
DATA COLLECTION AND VISUALIZATION

Pierre-Alexandre Balland

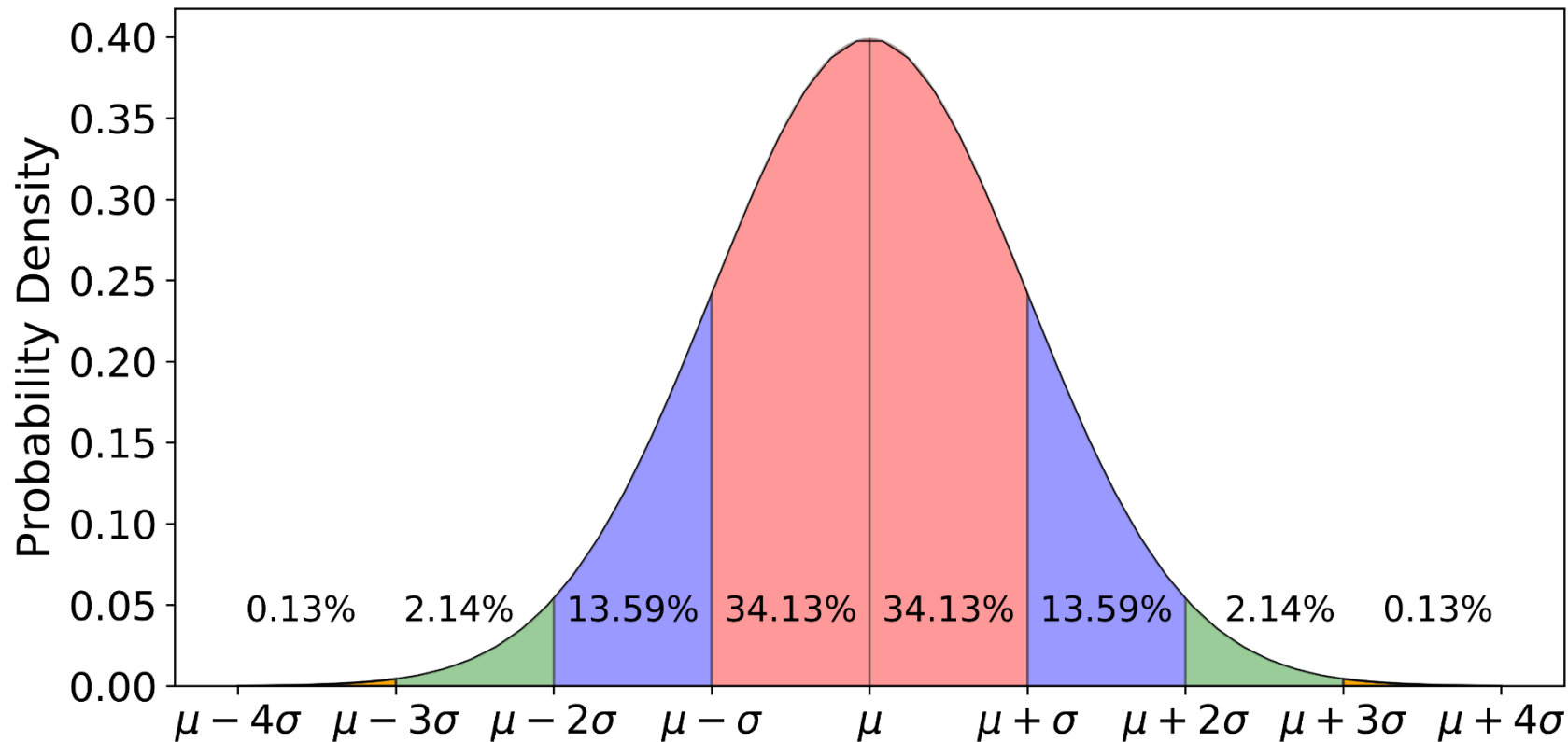
What is new in the DS revolution

```
35 self.logger = logging.getLogger(__name__)
36 if path:
37     self.file = open(os.path.join(path, 'requests.log'), 'a')
38     self.file.seek(0)
39     self.fingerprints.update(request)
40
41
42 @classmethod
43 def from_settings(cls, settings):
44     debug = settings.getbool('SUPERFILTER_DEBUG')
45     return cls(job_dir(settings), debug)
46
47 def request_seen(self, request):
48     fp = self.request_fingerprint(request)
49     if fp in self.fingerprints:
50         return True
51     self.fingerprints.add(fp)
52     if self.file:
53         self.file.write(fp + os.linesep)
54
55 def request_fingerprint(self, request):
56     return request_fingerprint(request)
```

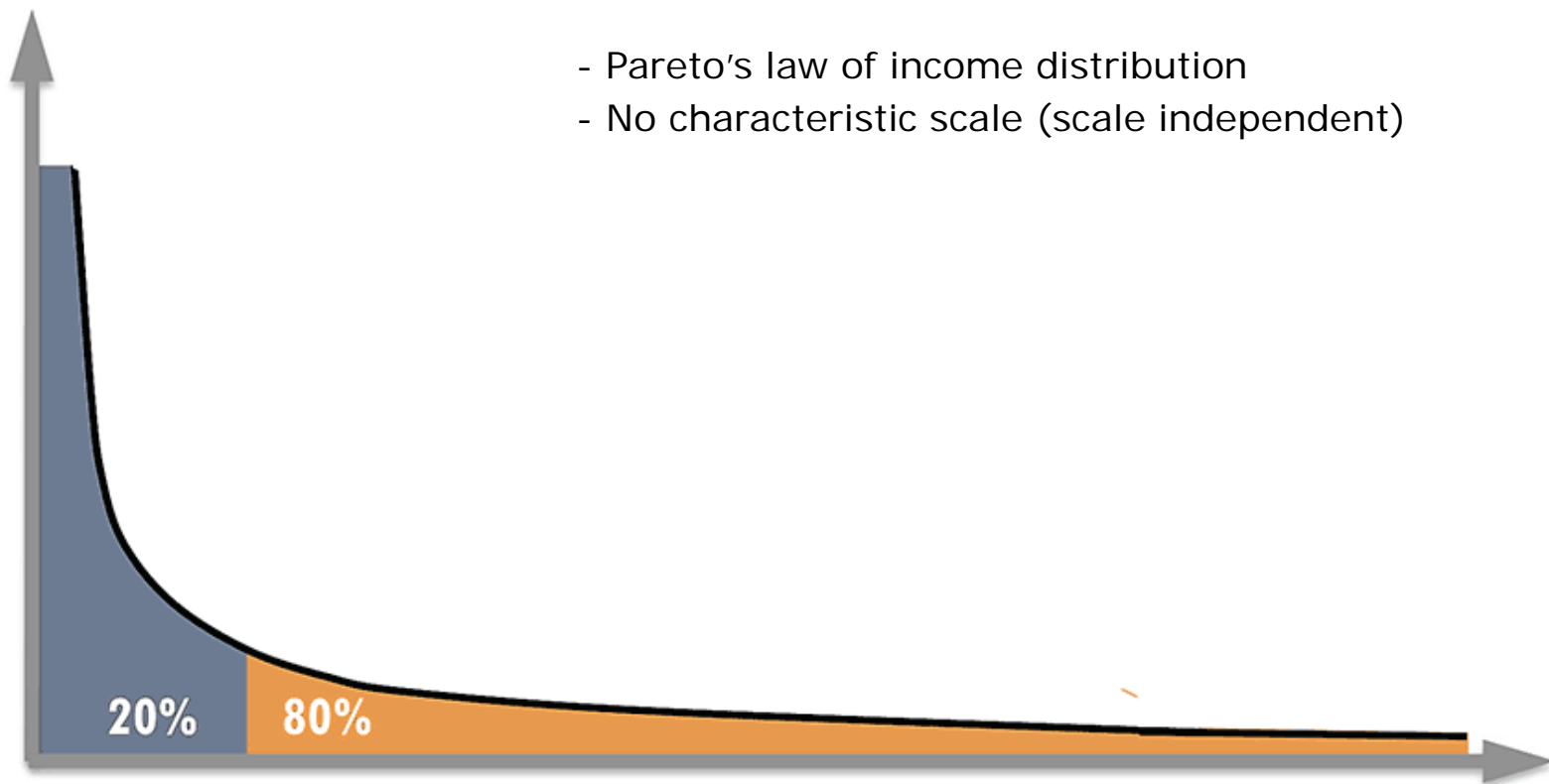
What is the take-away?



Normal Distribution



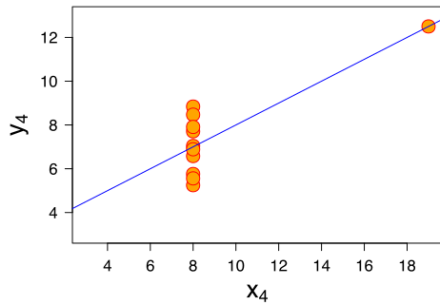
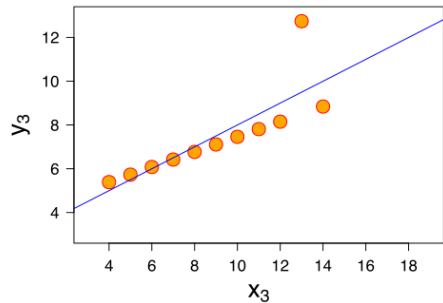
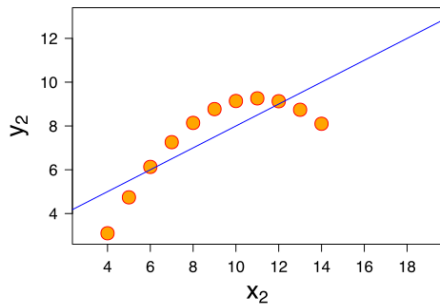
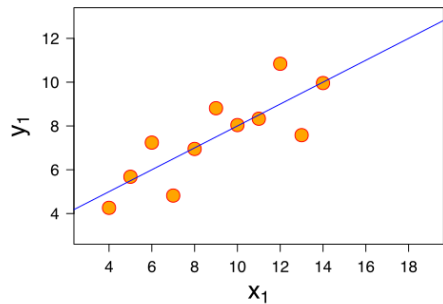
Fat tails (power laws for instance)



Where is the data?

- Government sources: www.cbs.nl
- Intergovernmental organisations: www.oecd.org
- University datasets: www.dataverse.harvard.edu
- Data products: www.crunchbase.com
- Privately owned data
- Collect your own data: surveys
- Collect your own data: create a data harvesting product

Anscombe's quartet



Mean of $x = 9$

Mean of $y = 7.5$

Correlation between x and $y = 0.816$

Linear regression line: $y = 3.00 + 0.5x$

$R^2 = 0.67$