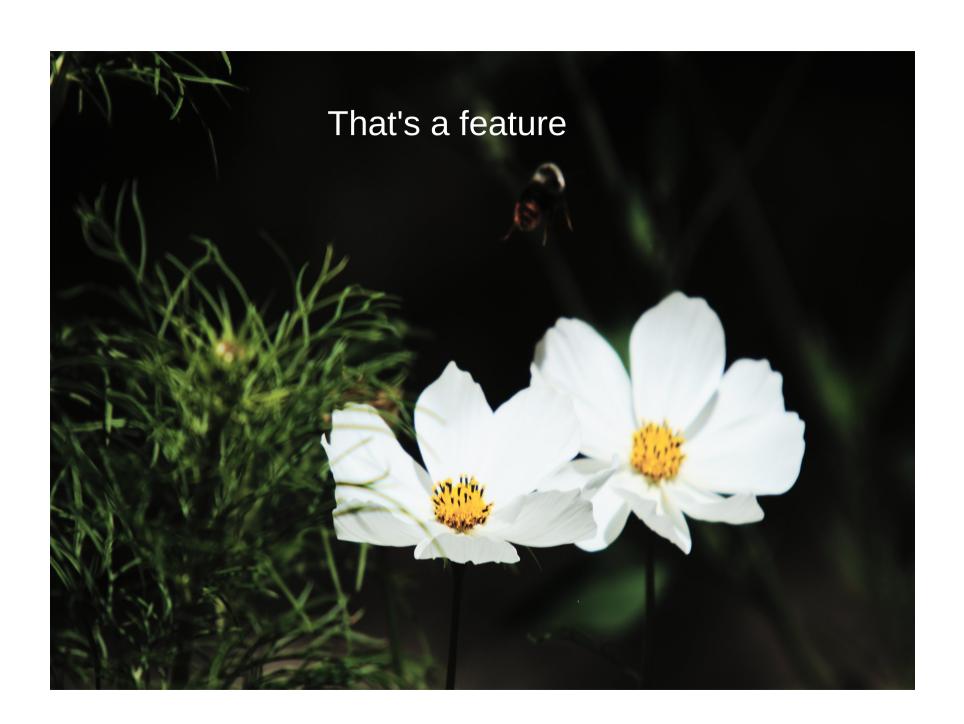
# Lies, Damn Lies, Statistics and Benchmarks



## The nature of the problem

Humans can't make sense of large numbers of numbers



#### We all know Statistics

- 87.5% of all statistics are made up on the spot
  - This one was.
- You can use statistics to prove anything
  - Including the truth
    - Sometimes
- Statistics are hard, let's go shopping
  - What numbers would you like to buy today?

# Understanding the data

#### Not all data is the same

- You have to ask yourself
  - What does this data represent?
  - What information can be extracted from this data?

## What can data represent?

- Baselines
- Progressions
- Comparative numbers
- Nothing at all
- Worse, incomplete data can be misleading.

#### Representing data

- The representation is as important as the data itself.
  - People evaluate the representation, not the data.
- Charts are good.
  - Choosing the right chart is important.
- Not all representations should be charts.
  - Some should be maps.
- Sometimes the right way of reporting is a simple table of data.

#### Benchmarks

- The sole purpose of a benchmark is to provide a marker for comparison
  - This is a single number
- Operations people don't need benchmarks
  - We need detailed statistics
  - Operations is often about tradeoffs
  - Not everything can be reduced to a single number.

## Running your own benchmarks

- Performance must be measured on the same hardware as production.
- Figure out the bottleneck of the system.
  - CPU
  - Disk I/O
  - Memory
  - Network bandwidth
  - Network latency
  - Data structures/algorithms.

# The result of any benchmark is typically a curve which closely follows the curve

$$1 - e^{-x}$$

## Benchmarking tools

In no particular order: sar, top, iostat, vmstat, dstat, tcpdump, vnstat, mrtg, cacti, htop, free, iptraf, ntop.

