Experimentation Support Tools

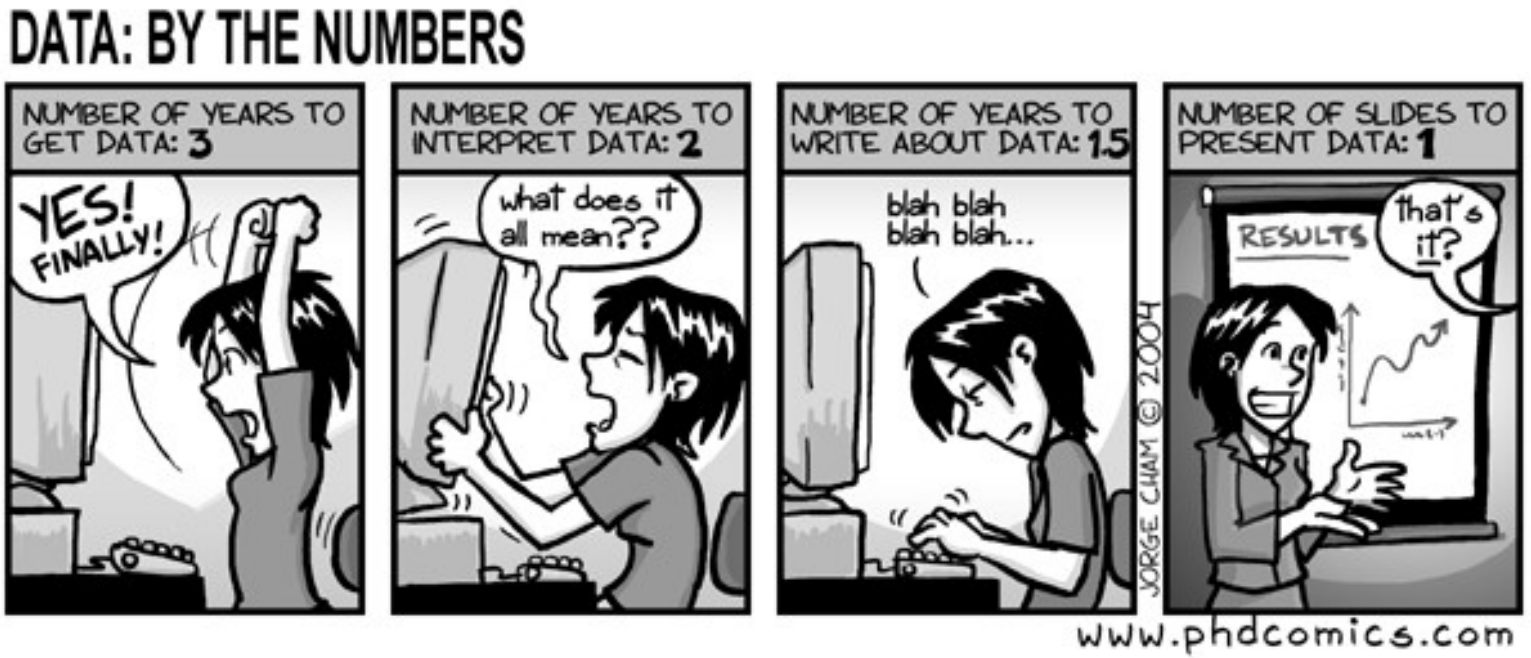
Víctor Jiménez

Barcelona Supercomputing Center
March 16th, 2012

Outline

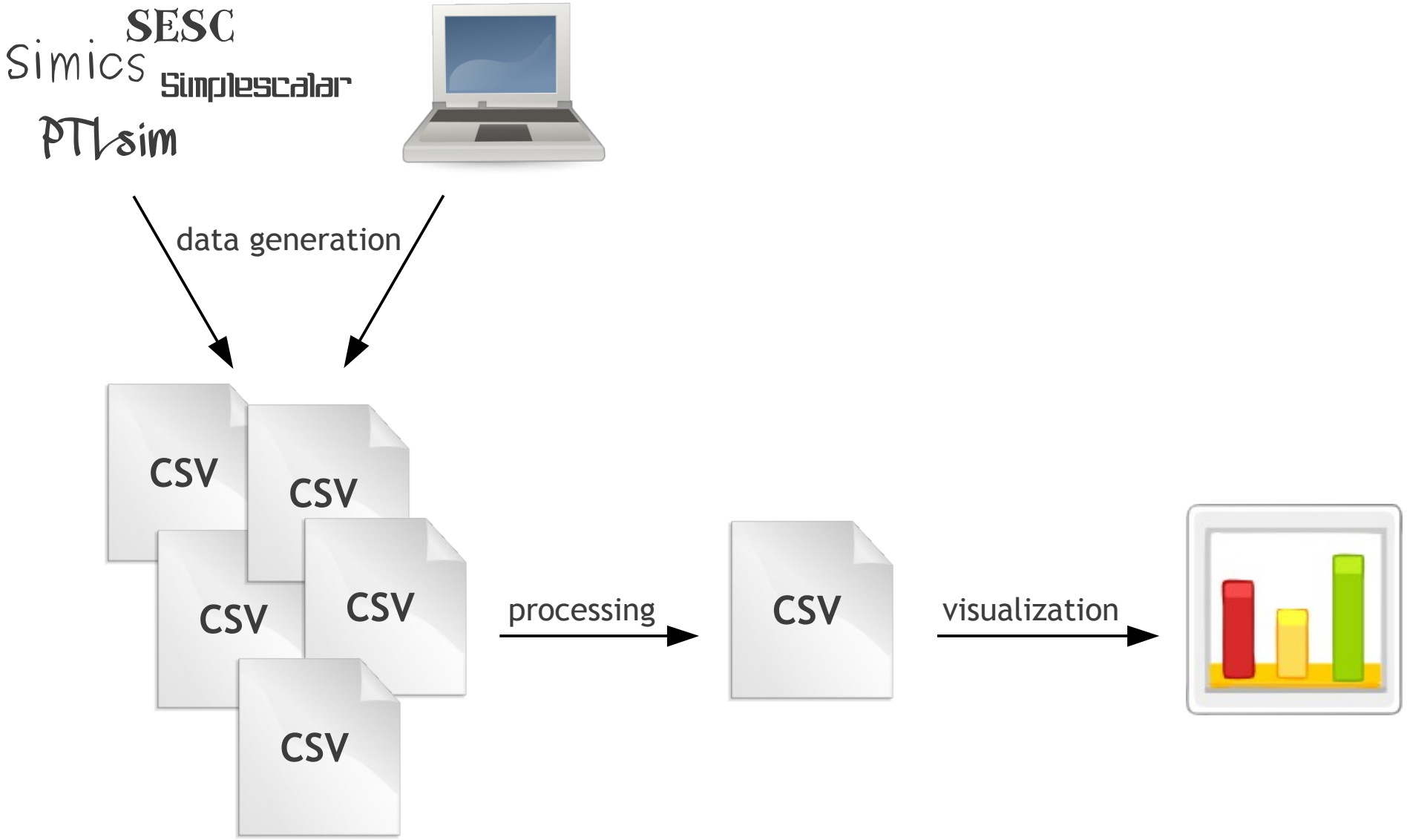
- Data processing
- Data visualization

Handling Data Is Difficult



It seems all PhD students have the same problems with data

Typical Data Flow



Tools for Data Processing

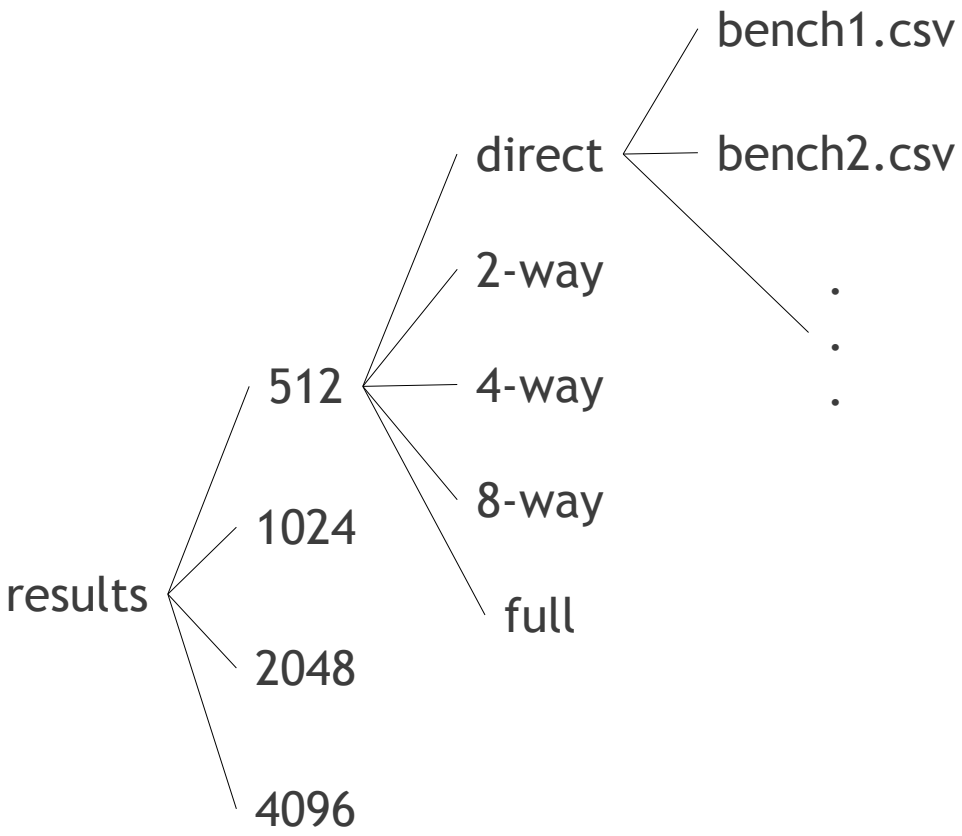
- Excel / LibreOffice
  - Do not scale (they are okay for small datasets)
  - Copy & paste is error prone
- Scripting (Python/R)
  - Great support for data analysis and visualization
  - User is responsible for data processing
    - Tedious script writing
- Others?
  - Which ones do you use?

PoTrA

- Data processing and visualization tool
  - Developed by Ramon Bertran (and Lluis Vilanova)
    - PhD students @BSC
- Collect multiple data sources (CSV files)
  - Generate an N-dimensional data object
- Different usages
  - Data transformation (processing)
  - Data visualization (plots)
- Behavior is defined through configuration files

Processing Examples

Study of cache size and associativity effect on performance



If #benchs=10 → 200 CSV files

Processing Examples

1. Using PoTrA to gather all CSV files into a single file

    Gathering regular expression:
    results/@SIZE@/@ASSOC@/@BENCH@.csv

    Reshape: SIZE, ASSOC, BENCH

SIZE	ASSOC	BENCH	IPC
512	direct	bench1	1.1
512	direct	bench2	0.6
…	…	…	…
512	2-way	bench1	1.2
…	…	…	…
1024	direct	bench1	1.2
…	…	…	…

2. Different dimension order when doing the reshape step

    Reshape: BENCH, SIZE, ASSOC

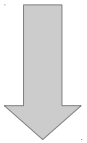
BENCH	SIZE	ASSOC	IPC
bench1	512	direct	1.1
bench1	512	2-way	1.2
…	…	…	…
bench1	1024	direct	1.2
…	…	…	…
bench2	512	direct	0.6
…	…	…	…

Processing Examples

3. Compute throughput (assume all $bench_i$ are executed in parallel)

   Reduce operation on variable BENCH

SIZE	ASSOC	BENCH	IPC
512	direct	bench1	1.1
512	direct	bench2	0.6
…	…	…	…
512	2-way	bench1	1.2
…	…	…	…
1024	direct	bench1	1.2
…	…	…	…



SIZE	ASSOC	THROUGHPUT
512	direct	5.4
512	2-way	5.8
…	…	…
1024	direct	6.1
…	…	…

Other Processing Features

- Many more operations are possible
  - Filtering values, creating new values based on expressions, normalizing values, etc.
- Support for big data sets
  - It can use HDF5 for better I/O performance
  - Final results can then be output to a CSV file

Tools for Data Visualization

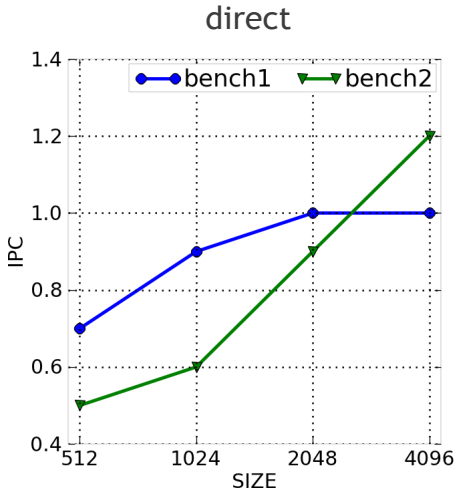
- Excel / LibreOffice
  - Difficult to automatize and obtain a consistent style
  - Ugly plots
- Gnuplot
- R
  - Full scientific environment similar to Matlab
  - Plenty of available libraries (analysis, plotting, etc.)
  - Good quality plots
- Matplotlib
  - Python library
  - Quality of plots is quite good
- Others?
  - Which ones do you use?

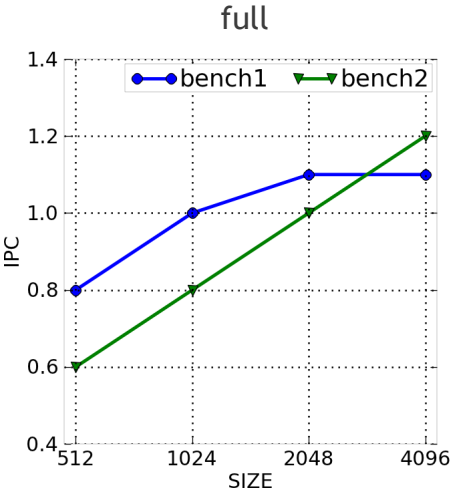
PoTrA

- Strengths
  - Plots are automatically generated from data
    - Very useful for design space exploration
  - No need to write code (e.g., scripts)
- Drawbacks
  - Not as flexible as using R or scripting
  - Quality of the plots is not best
    - Better than Excel, though

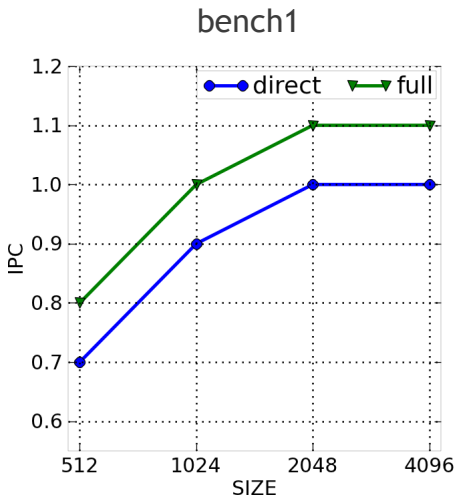
Visualization Examples

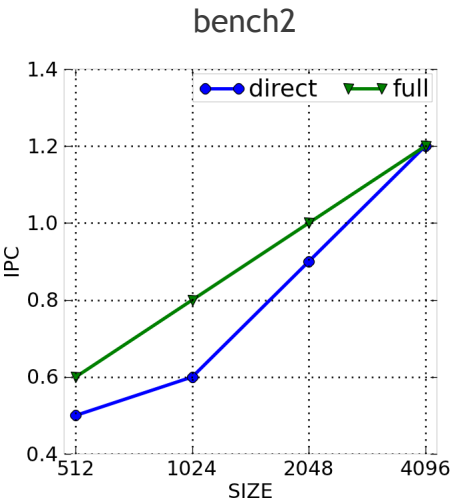
1. Plotting size effect per benchmark and associativity





2. Plotting size effect per associativity and per benchmark





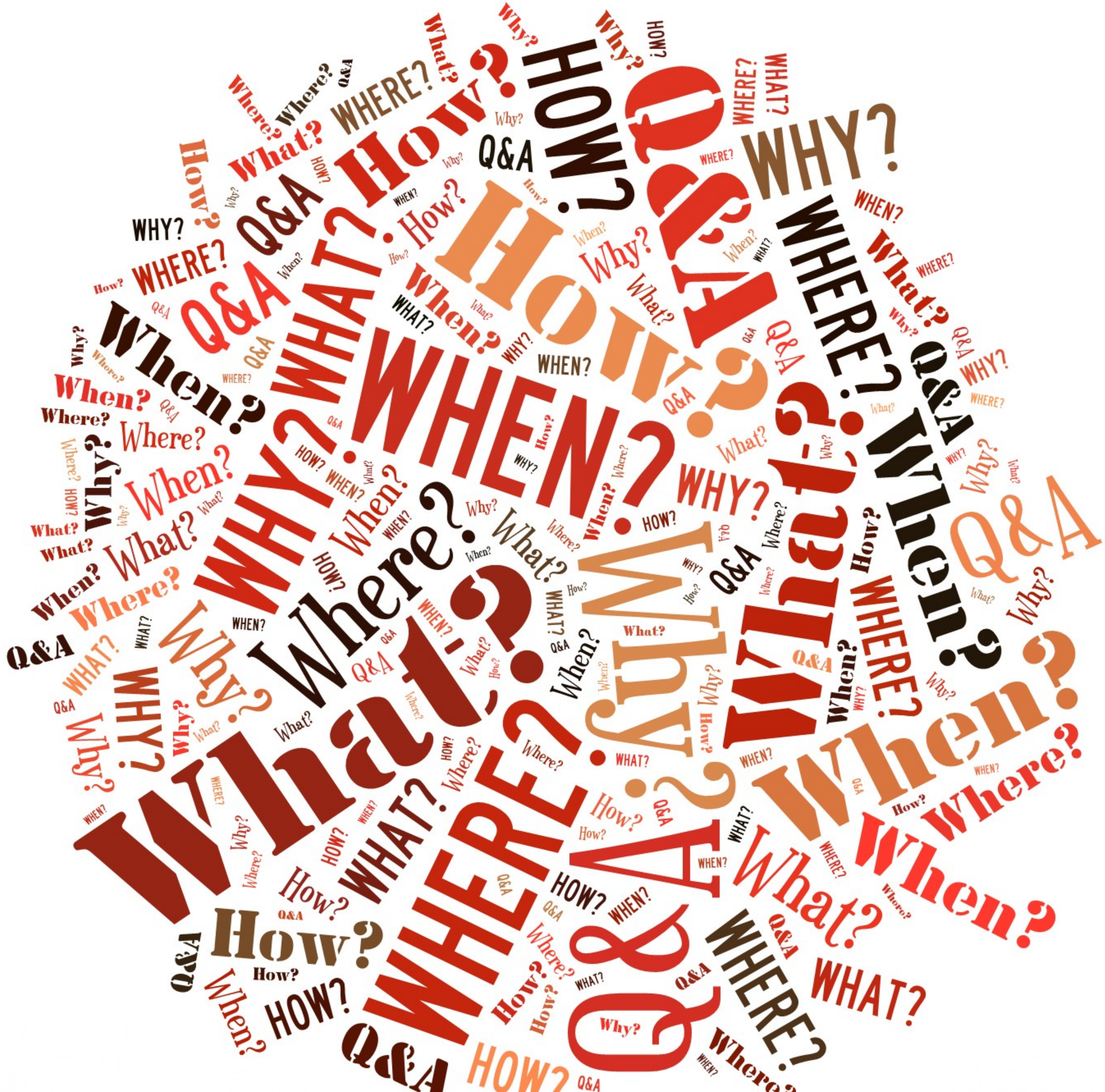
Other Visualization Features

- Support for many plot types
  - Line, bar, histogram, box
- Multiple plot values at the same time
  - Plotted on the same figure or in a different one
- Highly customizable
  - Font size, line size, margins, limits, etc.
  - Based on matplotlib
    - Many of the library's parameters are exposed

Configuration File Example

Full configuration file for the plots shown before

```
[Input Options]
input = .
inputplugin = ReaderCsv

[ReaderCsv]
csv-gather-regexp = ./results/@SIZE@/@ASSOC@/@BENCH@.csv

[Output Options]
output = ./results.csv
outputplugin = WriterCsv

[Processing Options]
plugin = Reshape
plugin = PlotDim

[Data Reshape]
reshape-dimension = ASSOC
reshape-dimension = BENCH
reshape-dimension = SIZE

[PlotDim]
pd-dimensionname = BENCH
pd-type = Line

[Common Plot Options]
plot-outdir = ./plots
plot-size = 4x4
plot-valuename = IPC
plot-autoxtics = True
plot-sort = True
plot-legendlocation = best
plot-legendtitle = True
plot-line-marker = True
plot-line-markersize = 6
plot-lw = 2
plot-glw = 1
plot-format = png
plot-dpi = 180
```


What?
WHERE?
Where?
WHEN?
When?
Why?
WHY?
How?
HOW?
WHAT?
Q&A