

Analysing Phenotype MicroArray data with *opm*

Benjamin Hofner, FAU, Erlangen/Germany

Lea A.I. Vaas, Fraunhofer Institute, Hamburg/Germany

María del Carmen Montero Calasanz, Newcastle University, Newcastle/UK

Markus Göker, DSMZ, Braunschweig/Germany

Why did we implement *opm*?

Part 1:

- **Free**, easily extendable software
- Flexible production of **high-quality graphics** and **robust statistical analysis**
- **Interactive or fully automated** usage possible

Part 2:

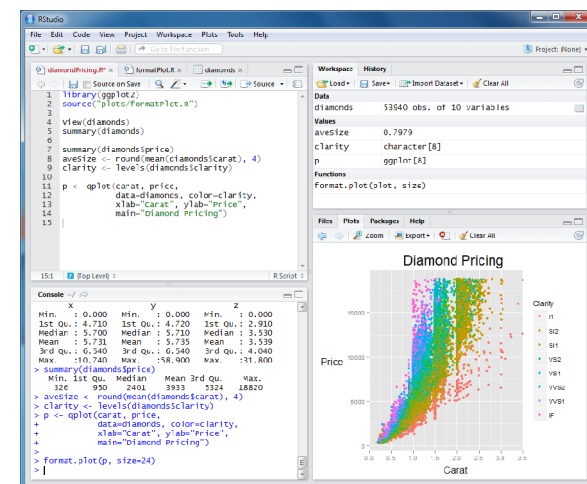
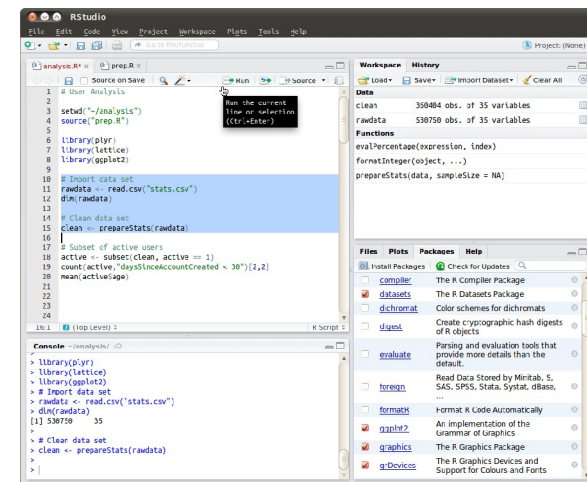
- Flexible **metadata** management
- **Reproducible** research
- Easy **interaction** with other software (and databases)



www.r-project.org

Why R?

- *de facto* standard for free-software statistical computing
- all operating systems
- flexible and clean coding
- non-interactive and interactive use
- powerful GUIs/IDEs (e.g. RStudio™)



www.rstudio.com

Parts of this workshop

1. **Reading PM data** into opm
2. Adding and manipulating **metadata**
3. **Visualizing** PM data in opm
4. Estimating and visualizing **curve parameters**
5. **Statistical comparisons** of experimental groups

Metadata

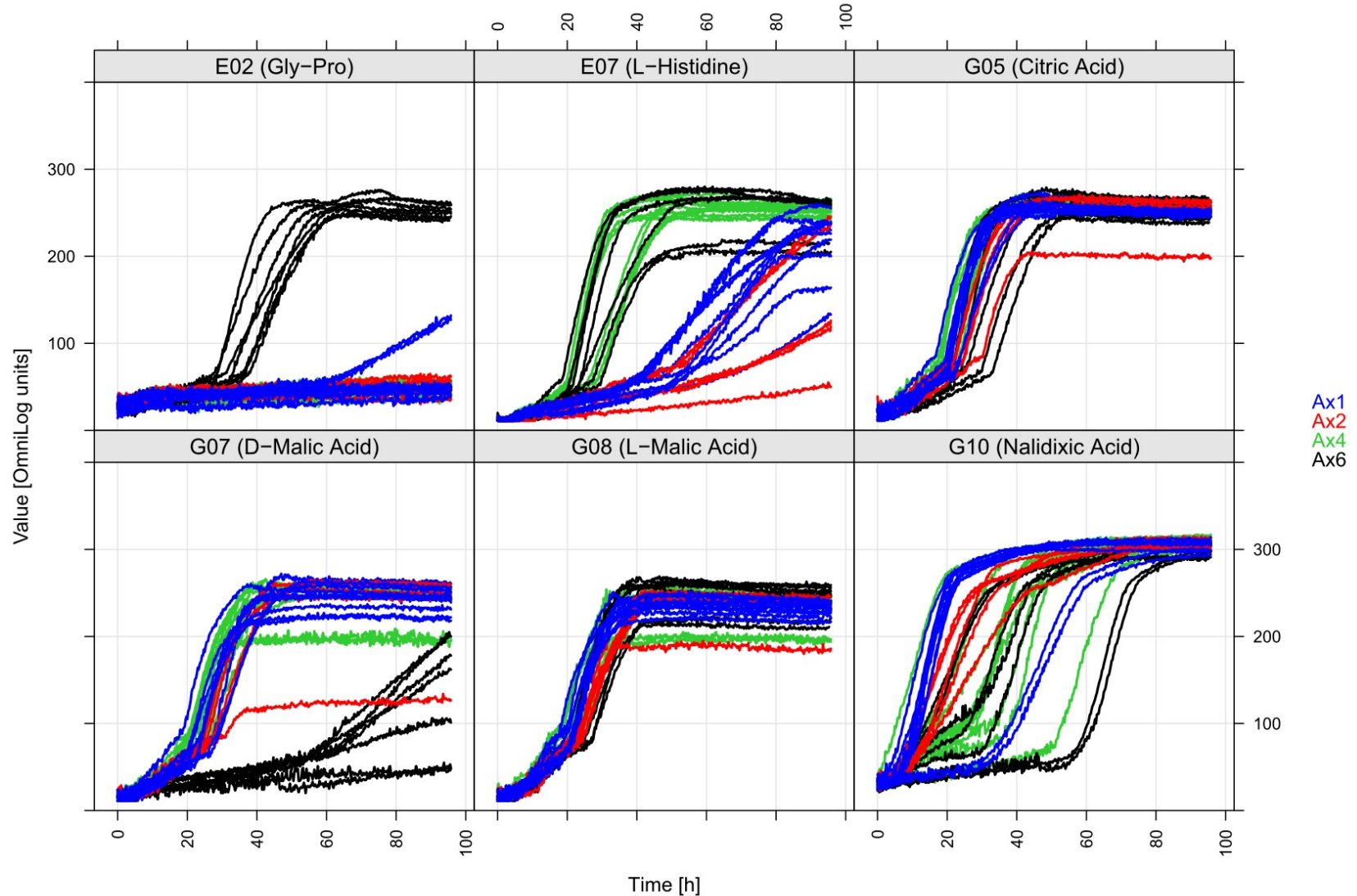
- “data about data”
- structural or (here) **descriptive**: organism measured, growth conditions etc.

Goal of *opm*:

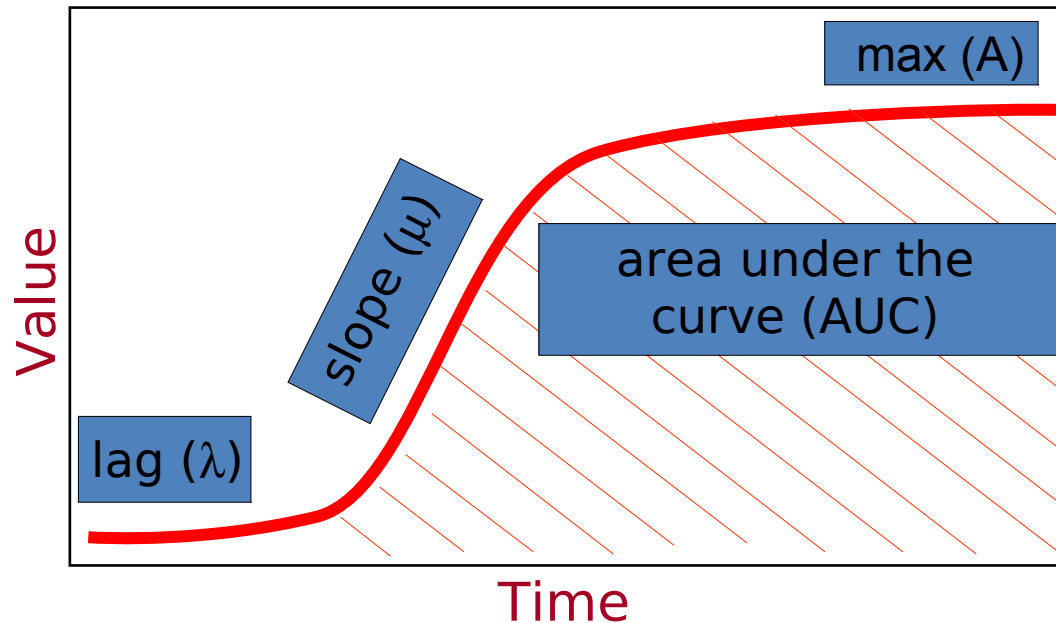
- **self-describing objects** that carry all user-relevant information (data and their metadata)
- **sufficient** for subsequent plots and statistical tests
- can be written to files and **distributed** between systems
- use a **light-weight**, flexible file format

The data: multi-dimensional, need aggregation

Six wells, 32 plates



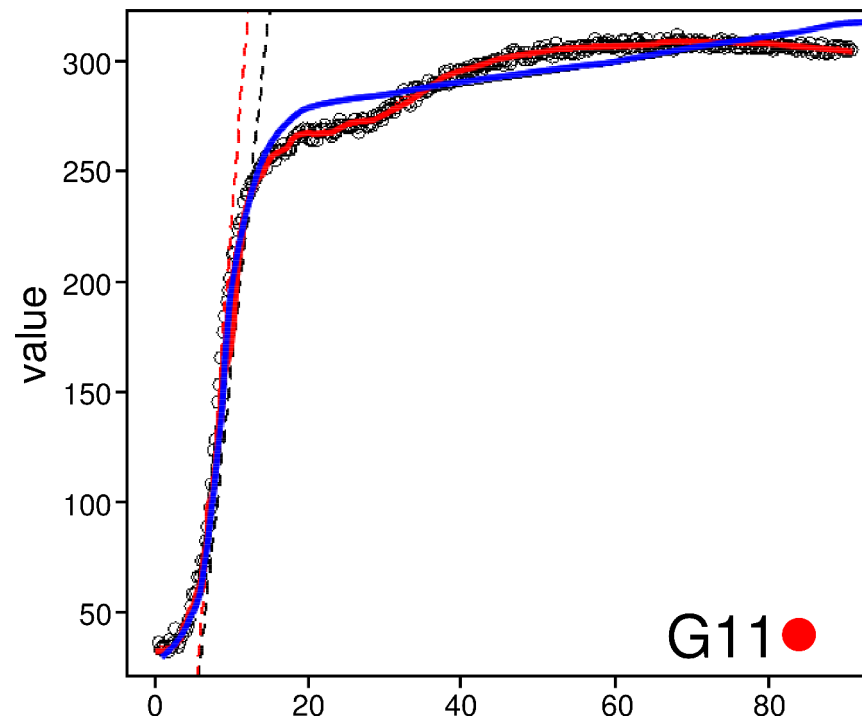
Aggregating: estimating curve parameters



Options

- **parametric models** (Gompertz, logistic etc.): explicitly estimate the four parameters or parameters from which the four can be derived
- **splines**: estimate smooth curves non-parametrically and analytically derive the four parameters from the smooth curves

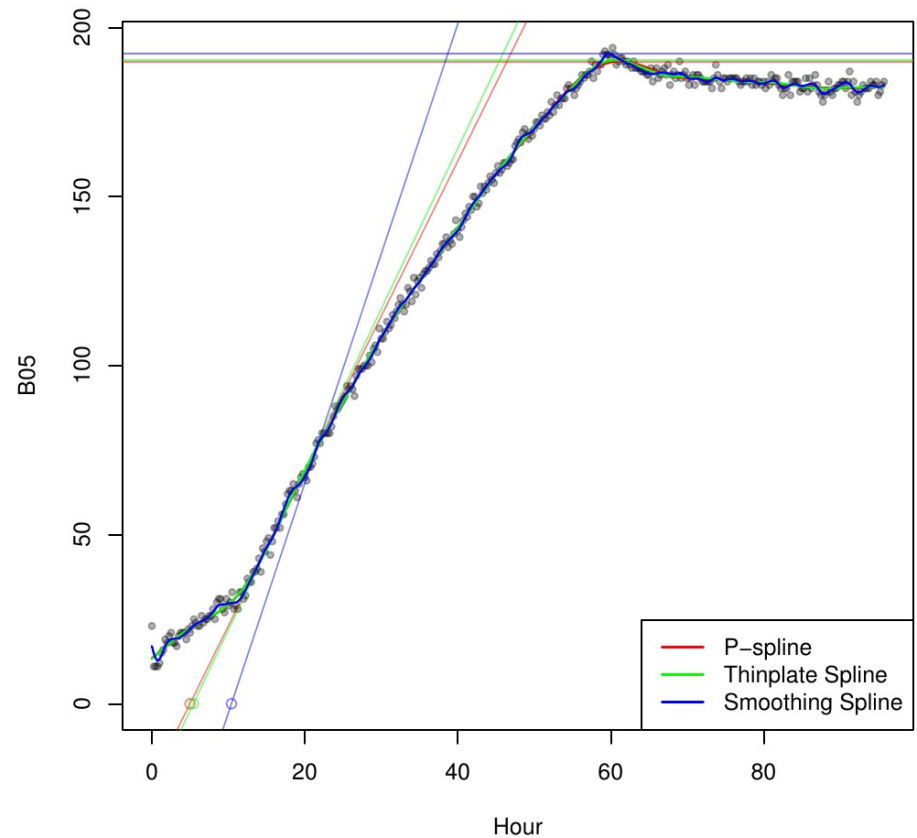
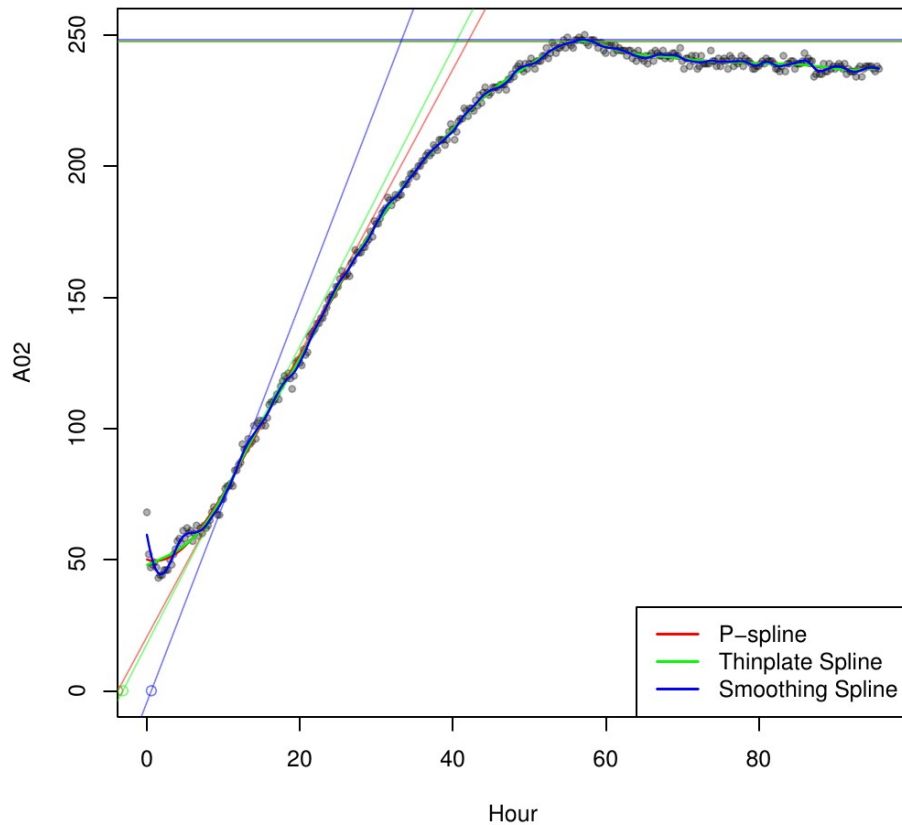
Aggregating: estimating curve parameters



- **parametric models** (Gompertz, logistic etc.) work only well for rather regular curve shapes
- **splines** usually work as well in these situations and outperform them in other cases

Aggregating: estimating curve parameters

Some spline methods perform better than others:



What is the correct statistical procedure to compare four groups pairwise to each other?

Data situation: Typical approach:
 six separate t-tests

7 plates:	Ax1	Ax2 - Ax1 , $p = \dots$
10 plates:	Ax2	Ax4 - Ax1 , $p = \dots$
10 plates:	Ax4	Ax6 - Ax1 , $p = \dots$
5 plates:	Ax6	Ax4 - Ax2 , $p = \dots$
		Ax6 - Ax2 , $p = \dots$
		Ax6 - Ax4 , $p = \dots$

What is the correct statistical procedure to compare four groups pairwise to each other?

Data situation:

7 plates: Ax1
10 plates: Ax2
10 plates: Ax4
5 plates: Ax6

Typical approach:
six separate t-tests

~~Ax2 - Ax1, $p = \dots$~~

~~Ax4 - Ax1, $p = \dots$~~

~~Ax6 - Ax1, $p = \dots$~~

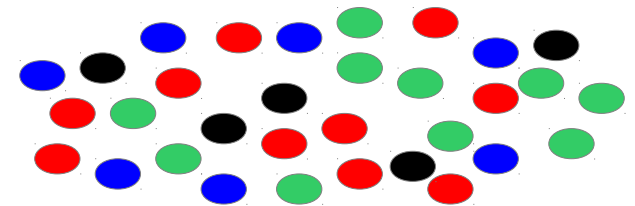
W R O N G

~~Ax4 - Ax2, $p = \dots$~~

~~Ax6 - Ax2, $p = \dots$~~

~~Ax6 - Ax4, $p = \dots$~~

Why the t-tests fail:



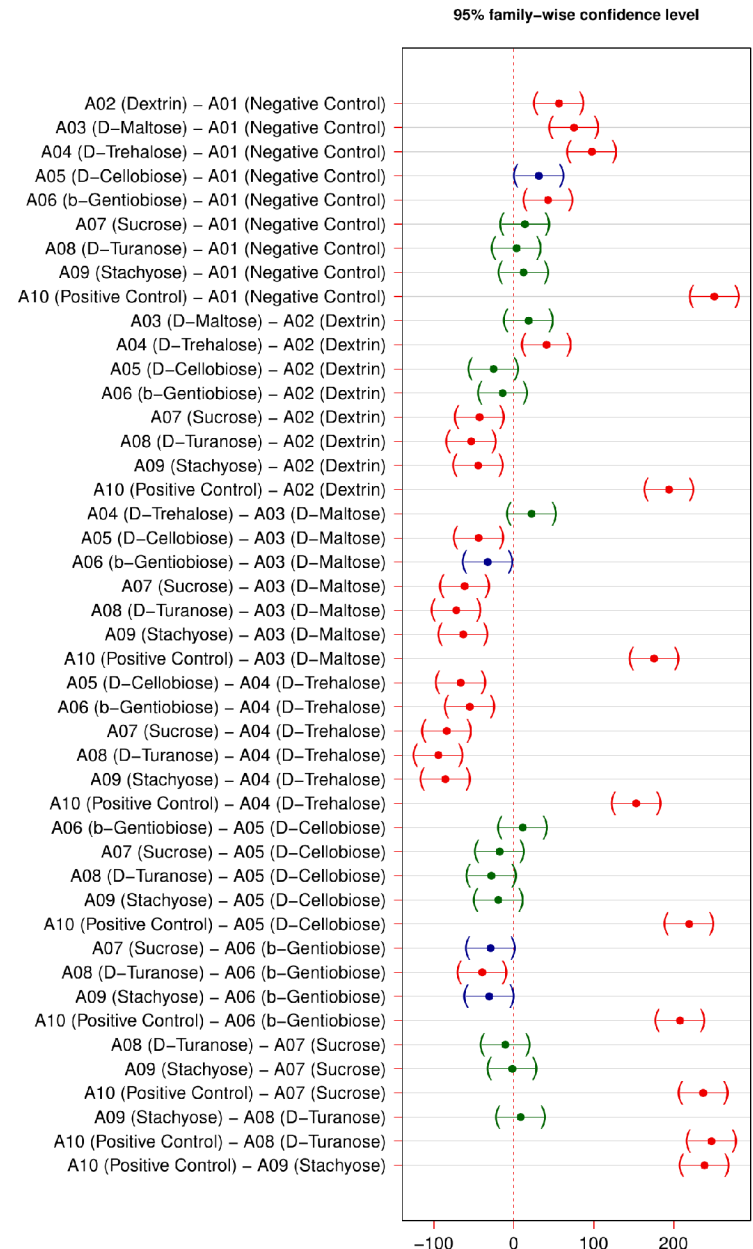
The more pairwise comparisons of multiple groups from the same sample, the larger the risk of identifying **false-positive** differences.

One must correct for the **family-wise error rate**

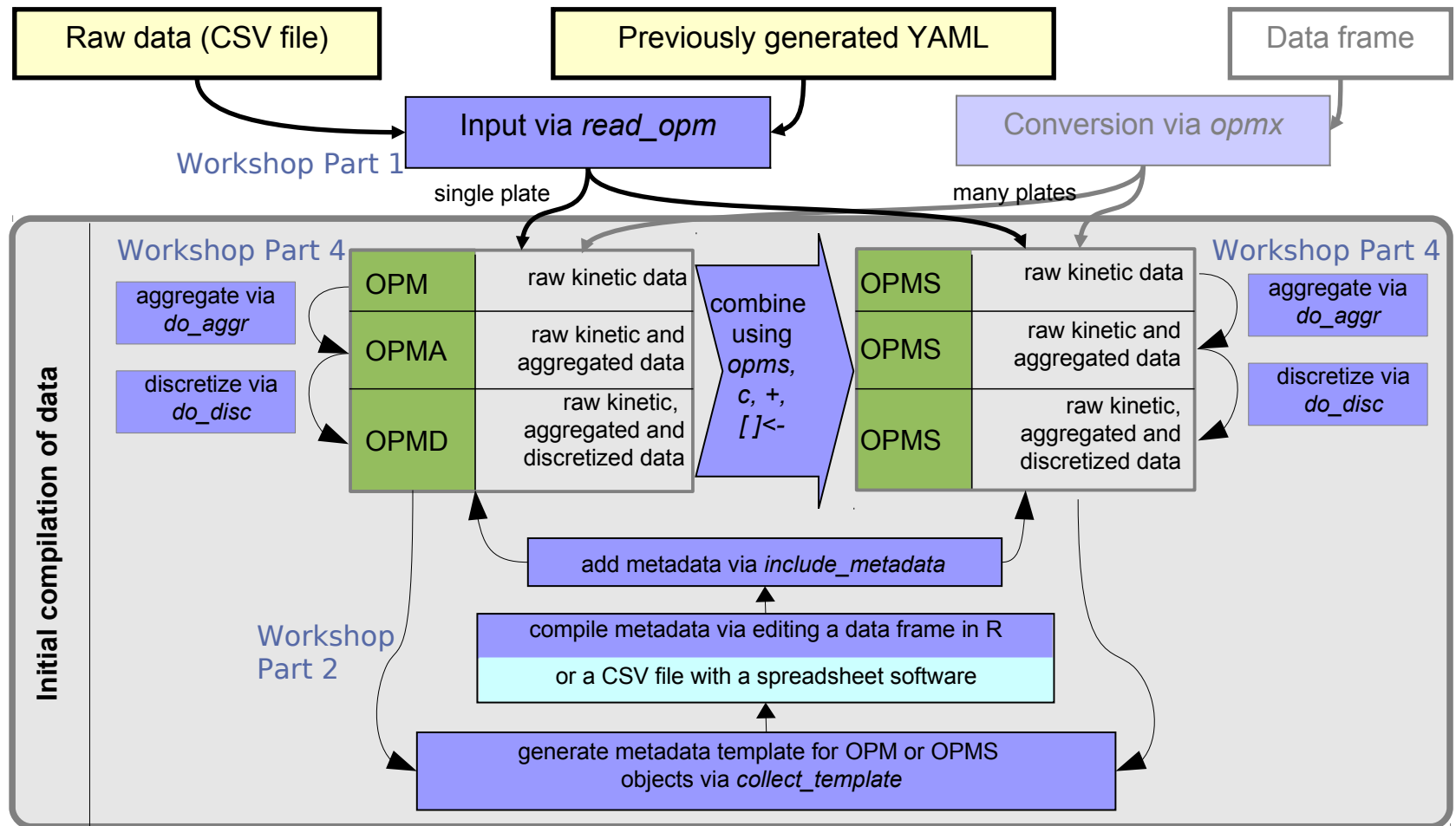
Multiple comparison of means

- user-defined comparisons
- inherent multiplicity adjustment
- significance of difference
- *and* effect size visible
- on original scale

Hothorn, T. et al. (2008) Simultaneous inference in general parametric models. *Biometr. J.* 50. 346-363.



Data manipulation in *opm*

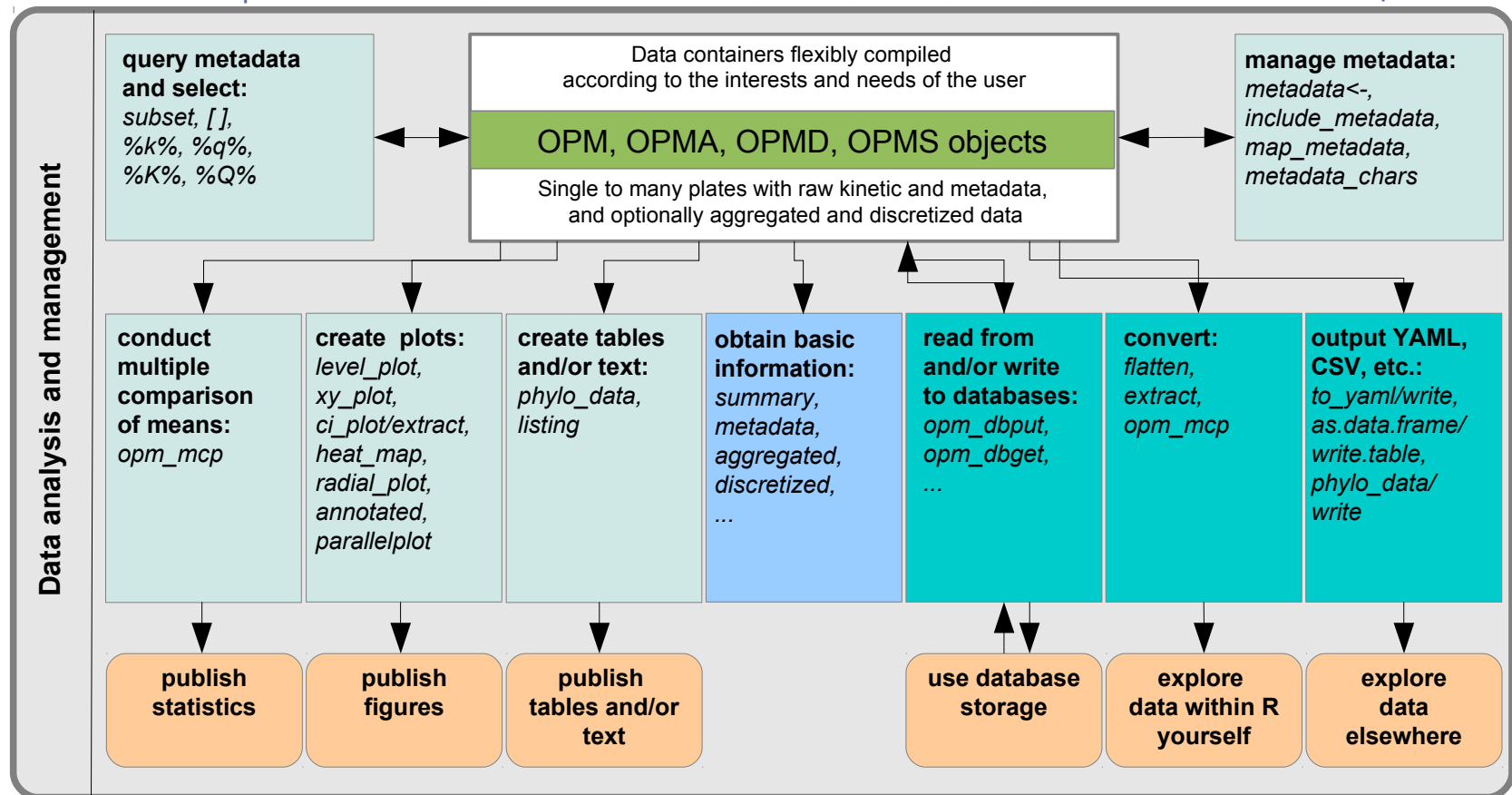


CSV := Character Separated Values; YAML, a data serialization format

Data analysis in *opm*

Workshop Part 2

Workshop Part 2



Workshop
Part 5

Workshop
Parts 3, 4

Further reading

- **Project homepage:**

<http://opm.dsmz.de>

- **Package and further resources:**

<http://www.goeker.org/opm>

- **Development version of the package:**

https://r-forge.r-project.org/R/?group_id=1573

- **Overview on opm:**

Lea A.I. Vaas, J. Sikorski, B. Hofner, N. Buddruhs, A. Fiebig, H.-P. Klenk and M. Göker. "opm: An R package for analysing OmniLog® Phenotype MicroArray Data". *Bioinformatics* 29 (14): 1823-1824, 2013.

- **Parameter comparison and visualization with opm:**

Lea A.I. Vaas, J. Sikorski, V. Michael, M. Göker and H.-P. Klenk. "Visualization and curve-parameter estimation strategies for efficient exploration of phenotype microarray kinetics". *PLoS ONE* 7 (4): e34846, 2012.

- **Advanced usage of opm to detect differential expressions:**

B. Hofner, L. Boccuto and M. Göker. "Controlling false discoveries in high-dimensional situations: Boosting with stability selection". *BMC Bioinformatics* 16 (6): 144, 2015.