The data set is available as SQLite database and csv file. There are X measurements for X organisms of X variables. Data set is not orthogonal; some data entries are missing (X total, across the data set) and not all organisms are equally represented.

Check for collinearity

The following variables were dropped due to significant collinearity:

**Piecewise regression** was applied [fig 1.]

**Multiple regression** and **analysis of covariance** [fig 3 and S3].

The maximal model can be described with a function: y=β0+ β1x1+…+ β13x13

y – meiotic metaphase spindle length (pole to pole or aster to aster).

To determine, which variables should be included in a statistical model of meiotic/mitotic spindle length, we ran regression tree using all continuous variables in the data set.

(forward propagation)

Regression tree model accounts for approximately 90% of variation in mitotic spindle length and 79% - in meiotic spindle length (90% if spindle aspect ratio is included in the model).

Interestingly, in both cases, cell diameter alone accounts for the largest fraction of deviance explained - 55%, in the case of mitotic spindles and 40% in the case of meiotic spindles.

Tree models tend to overfit the data. Therefore we used it mostly to guide us in choosing the variables

Given that the model explains 85% of variance, we used the

Unsupervised learning and logistic regression [fig 4]

We applied logistic regression using meiotic/mitotic

Shapiro test for normality for continuous measurement

Low rank matrix approximation was used to

Data was scaled

Kmeans/hierarchical clustering and GMM

Chi-square

Histograms for the groups

Python and R code used to process, analyze and plot data is available on GitHub <https://github.com/emmaggie/SPINDLE_ZOO> (academic license to M.S.). The IPython notebooks can be viewed with Notebook Viewer (<http://nbviewer.ipython.org/>).