**Dataset and data analysis**

The original dataset is available as SQLite database and csv file. There are 2001 measurements (rows) for 20 organisms, of 33 variables (columns), out of which 19 are continuous, 7 - categorical, 2 represent count data, 2 – time data, and 4 can be considered experimental metadata. Data set is not orthogonal; some data entries are missing (~8%, Figure S5) and not all categories (e.g. organisms) are equally represented.

*The following variables were dropped due to significant* ***collinearity****:*

Good et al. demonstrated that changes in cell diameter lead to changes in spindle size and that this effect is linear within certain cell size range (our analysis suggests this range to be below 140 µm). Piecewise regression suggests

**piecewise regression**

was applied [fig 1.]

**Tree models**

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 our data set. Regression trees use forward selection

(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. It is worth noting that the data set for meiotic spindle is smaller and sparser, thus estimates inferred from it are likely far less accurate.

Having used the tree models

Tree models tend to overfit the data. Therefore, we used them mostly to guide us in choosing the variables for multiple regression models. We used backward selection approach to eliminate non-significant terms from the model.

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

The maximal model can be described with a function: y=β0+ β1x1+…+ βixi, where y is meiotic/mitotic metaphase spindle length (pole to pole). We chose variables according to the rank provided by the tree models and evaluated potential curvature in the behavior of variables through generalized additive model (**Figure S6? – use re**levant gam\_ files from R\_and\_plots folder).

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

**Unsupervised learning**

For unsupervised learning data was scaled and centered and various techniques were tested for missing values imputation using (kNN, bagged trees, median or random sampling of data grouped by organism and spindle type, lower rank matrix approximation). The choice of technique was made on the basis of KS test (before and after) for each of the variables. Moreover, given the

Such preprocessed data was used for clustering and meiotic/mitotic data partitions evaluated with Chi square test.

For clustering across species, the optimal number of clusters was evaluated using and elbow method on a plot of cluster numbers

**Supervised learning**

We verified, whether our data set contains sufficient information to predict class assignment through supervised 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 fill in missing values.

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/>).