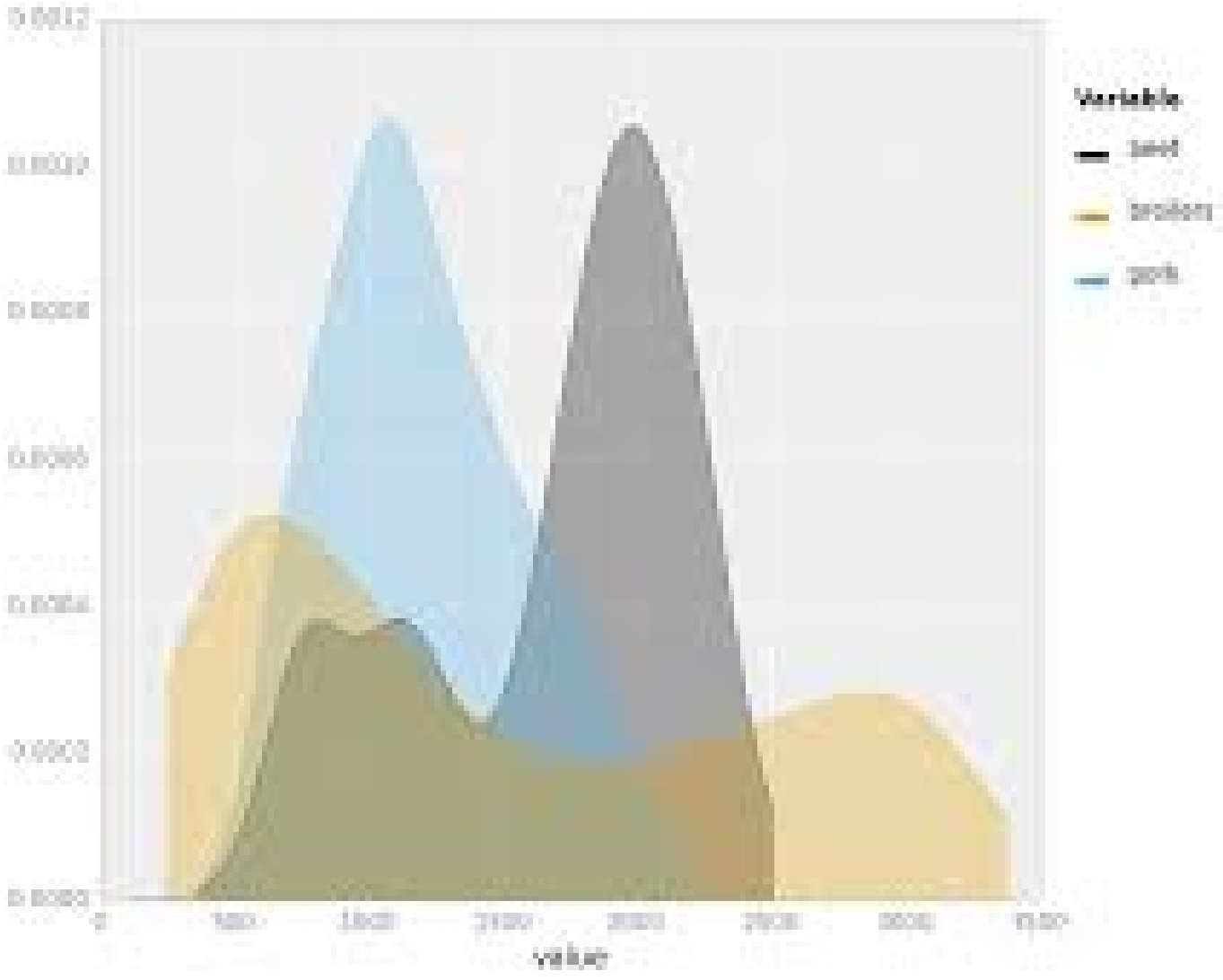
# Other Interesting Python Packages

# statsmodels

- `statsmodels` provides a large range of cross-sectional models aswell assometime-series models.
- statsmodels uses a model descriptive language (provided via the Python package patsy) to formulate the model when working with pandas `DataFrames`.
- Models supported include linear regression, generalized linear models, limited dependent variable models, ARMA and VAR models.

# Bokeh Data Visualization

# Bokeh Data Visualization



# Bokeh Data Visualization

**Bokeh Data Visualization**

- interactive graphics for the web
- designed for large data sets
- Designed for streaming data
- Native interface in pythin
- Fast javascript components
- DARPA funded
- v.01 relase imminent


scikit
learn

# scikit.learn


- scikit-learn is an open source machine learning library for the Python programming language.
- scikit-learn features various classification, regression and clustering algorithms including support vector machines, logistic regression, naive Bayes, random forests, gradient boosting, k-means and DBSCAN.
- scikit-learn is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.

# Sci-Kit Learn Site info

## Classification

Identifying to which category an object belongs to.

**Applications**: Spam detection, Image recognition.
**Algorithms**: *SVM*, *nearest neighbors*, *random forest*, ... — *Examples*

## Regression

Predicting a continuous-valued attribute associated with an object.

**Applications**: Drug response, Stock prices.
**Algorithms**: *SVR*, *ridge regression*, *Lasso*, ... — *Examples*

## Clustering

Automatic grouping of similar objects into sets.

**Applications**: Customer segmentation, Grouping experiment outcomes
**Algorithms**: *k-Means*, *spectral clustering*, *mean-shift*, ... — *Examples*

## Dimensionality reduction

Reducing the number of random variables to consider.

**Applications**: Visualization, Increased efficiency
**Algorithms**: *PCA*, *feature selection*, *non-negative matrix factorization*. — *Examples*

## Model selection

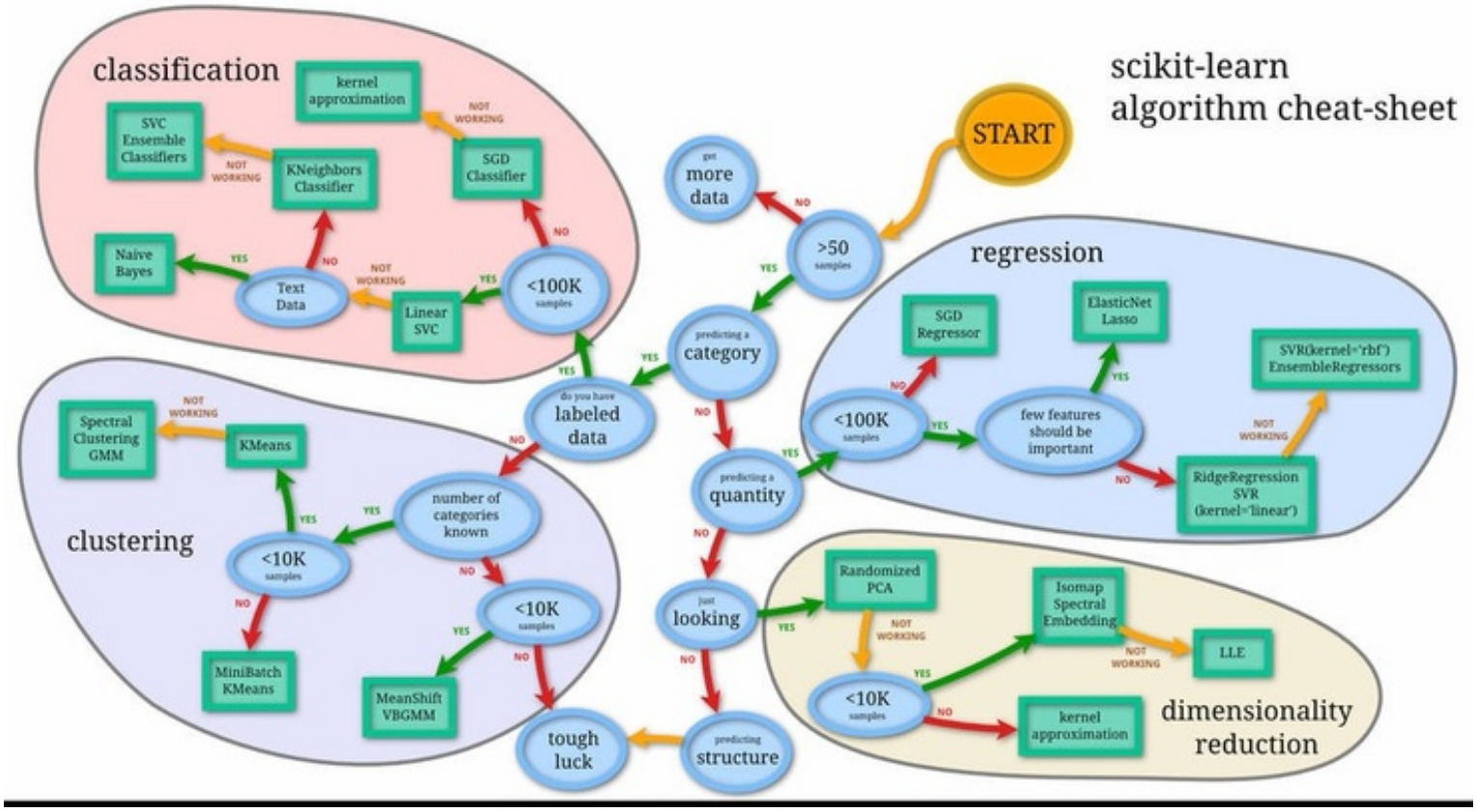
Comparing, validating and choosing parameters and models.

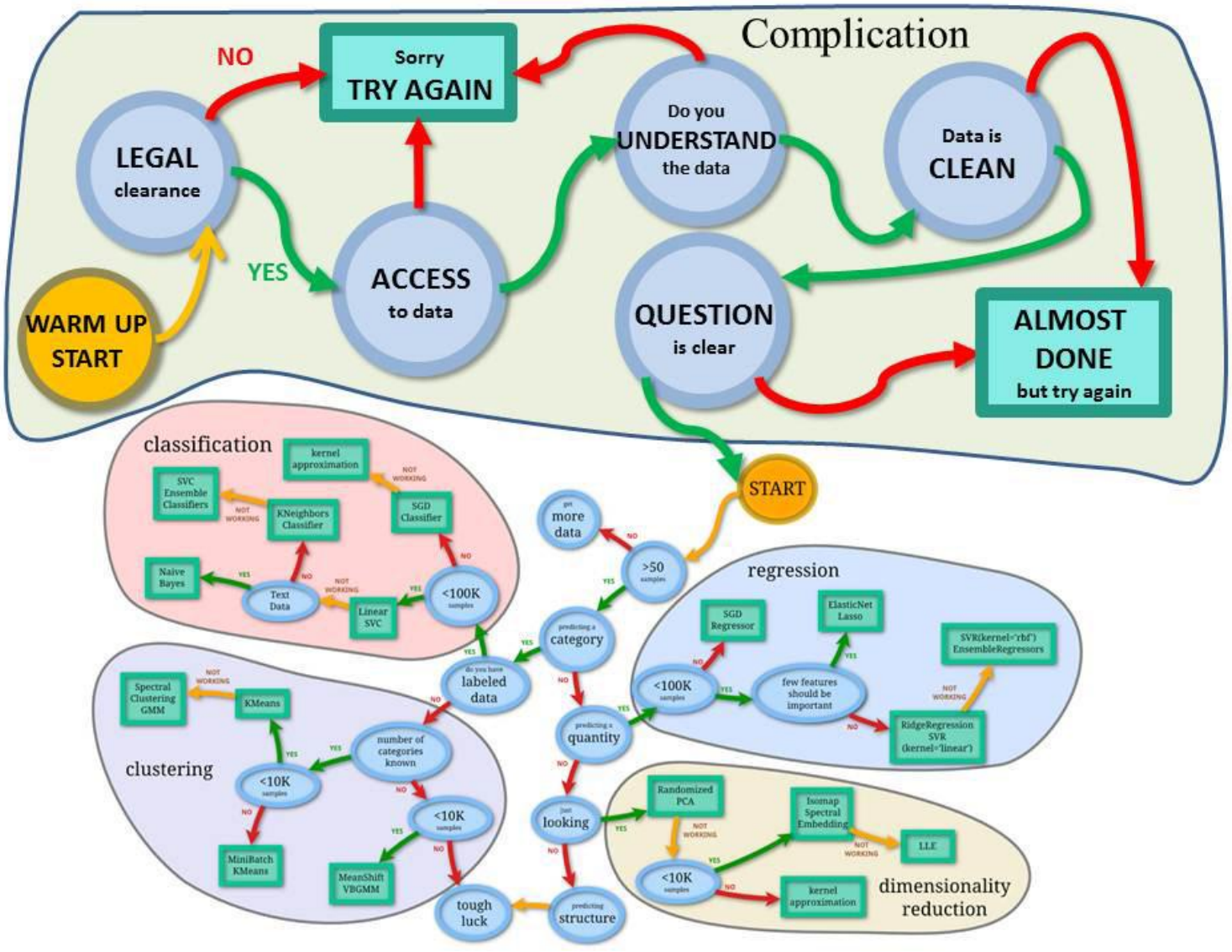
**Goal**: Improved accuracy via parameter tuning
**Modules**: *grid search*, *cross validation*, *metrics*. — *Examples*

## Preprocessing

Feature extraction and normalization.

**Application**: Transforming input data such as text for use with machine learning algorithms.
**Modules**: *preprocessing*, *feature extraction*. — *Examples*


scikit-learn
algorithm cheat-sheet
START
classification
kernel approximation
SVC Ensemble Classifiers
KNeighbors Classifier
SGD Classifier
Naive Bayes
Text Data
Linear SVC
<100K samples
NOT WORKING
YES
NO
get more data
>50 samples
predicting a category
do you have labeled data
regression
SGD Regressor
ElasticNet Lasso
SVR(kernel='rbf') EnsembleRegressors
<100K samples
few features should be important
RidgeRegression SVR (kernel='linear')
clustering
Spectral Clustering GMM
KMeans
number of categories known
<10K samples
MiniBatch KMeans
MeanShift VBGMM
predicting a quantity
just looking
predicting structure
tough luck
dimensionality reduction
Randomized PCA
Isomap Spectral Embedding
LLE
kernel approximation


Complication
NO
Sorry
TRY AGAIN
LEGAL
clearance
YES
ACCESS
to data
Do you
UNDERSTAND
the data
Data is
CLEAN
WARM UP
START
QUESTION
is clear
ALMOST
DONE
but try again
START
classification
kernel approximation
SVC Ensemble Classifiers
KNeighbors Classifier
SGD Classifier
Naive Bayes
Text Data
Linear SVC
<100K
more data
>50
predicting a category
labeled data
regression
SGD Regressor
ElasticNet Lasso
SVR(kernel='rbf') EnsembleRegressors
<100K
few features should be important
RidgeRegression SVR (kernel='linear')
predicting a quantity
clustering
Spectral Clustering GMM
KMeans
number of categories known
<10K
MiniBatch KMeans
<10K
MeanShift VBGMM
just looking
Randomized PCA
Isomap Spectral Embedding
LLE
<10K
kernel approximation
dimensionality reduction
tough luck
predicting structure

Simon Blomberg:

> From R's fortunes package: To paraphrase provocatively, 'machine learning is statistics minus any checking of models and assumptions'. -- Brian D. Ripley (about the difference between machine learning and statistics) useR! 2004, Vienna (May 2004) :-) Season's Greetings!

Andrew Gelman:

> In that case, maybe we should get rid of checking of models and assumptions more often. Then maybe we'd be able to solve some of the problems that the machine learning people can solve but we can't!

Machine Learning is statistics minus any checking of models or assumptions

# The Data Science Profession

Data Science Retreat (Berlin)

> *MOOC have not decreased the barrier of entry to machine-learning.*
> *Nowadays, you cannot be 'the guy who knows how to run (insert off-the-shelf-algo-here)'.*
> *In dataland, that's the equivalent to being a code monkey. MOOCs and superb libraries (scikit-learn, R's ecosystem) made sure there is plenty of people who can throw say a random forest to a problem. In the modern world, this is not adding that much value.*

# Other Packages

**pytz and babel**
ptyz and babel provide extended support for time zones and formatting information.

**rpy2**
rpy2 provides an interface for calling R 3.0.x in Python, as well as facilities for easily moving data between the two platforms.

**PyTables and h5py**
PyTables and h5py both provide access to HDF5 files, a flexible data storage format optimized for numeric data.