



Stefan Kojouharov

[Follow](#)

Founder of Chatbots Life. I help Companies Create Great Chatbots & AI Systems and share my Insights along the way.

Jul 9, 2017 · 7 min read

## Cheat Sheets for AI, Neural Networks, Machine Learning, Deep Learning & Big Data

The Most Complete List of Best AI Cheat Sheets

Over the past few months, I have been collecting AI cheat sheets. From time to time I share them with friends and colleagues and recently I have been getting asked a lot, so I decided to organize and share the entire collection. To make things more interesting and give context, I added descriptions and/or excerpts for each major topic.

This is the most complete list and the Big-O is at the very end, enjoy...

If you like this list, you can let me know [here](#).

### Neural Networks

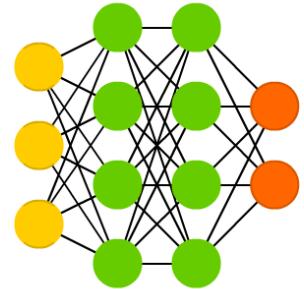
-  Backfed Input Cell
-  Input Cell
-  Noisy Input Cell
-  Hidden Cell
-  Probabilistic Hidden Cell
-  Spiking Hidden Cell
-  Output Cell
-  Match Input Output Cell
-  Recurrent Cell
-  Memory Cell
-  Different Memory Cell
-  Kernel
-  Convolution or Pool

*A mostly complete chart of*

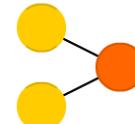
# Neural Networks

©2016 Fjodor van Veen - [asimovinstitute.org](http://asimovinstitute.org)

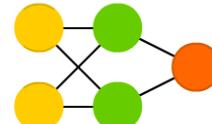
Deep Feed Forward (DFF)



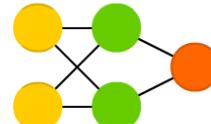
Perceptron (P)



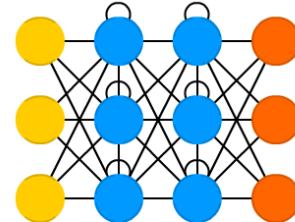
Feed Forward (FF)



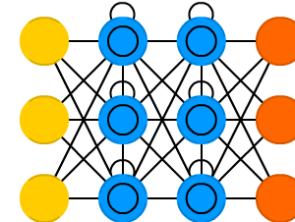
Radial Basis Network (RBF)



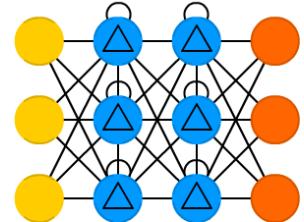
Recurrent Neural Network (RNN)



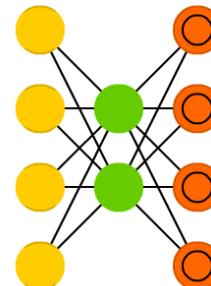
Long / Short Term Memory (LSTM)



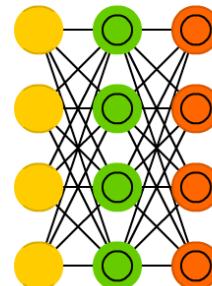
Gated Recurrent Unit (GRU)



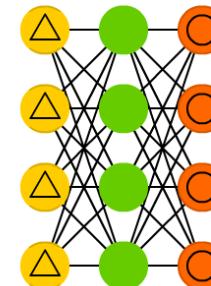
Auto Encoder (AE)



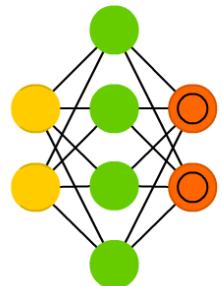
Variational AE (VAE)



Denoising AE (DAE)



Sparse AE (SAE)



Markov Chain (MC)



Hopfield Network (HN)



Boltzmann Machine (BM)

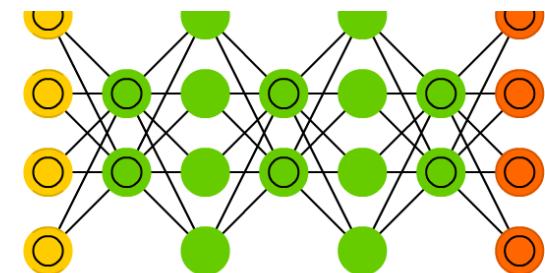
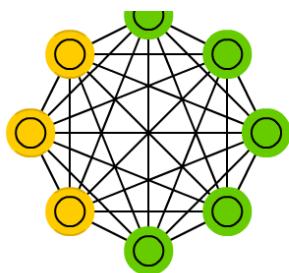
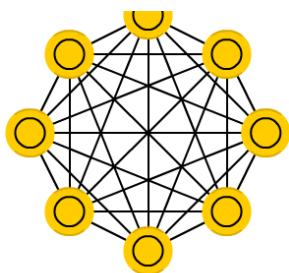
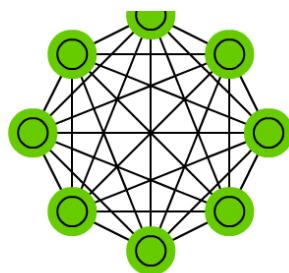


Restricted BM (RBM)

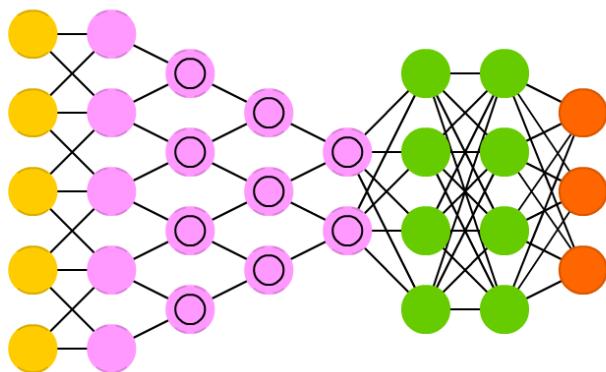


Deep Belief Network (DBN)

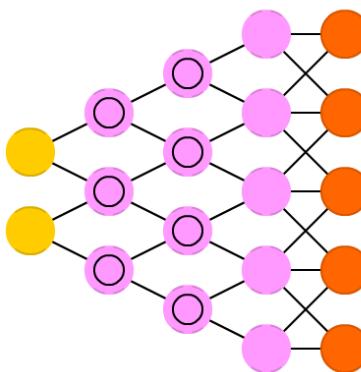




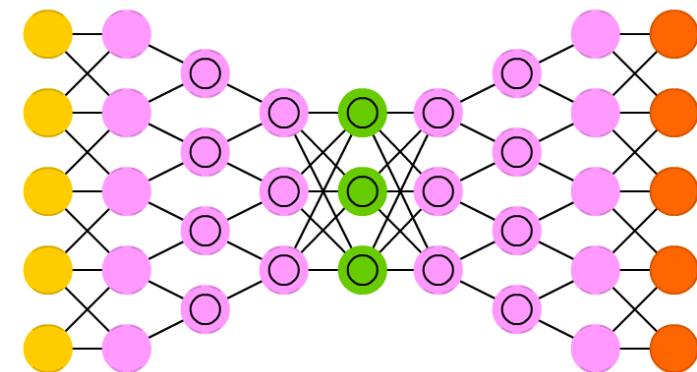
Deep Convolutional Network (DCN)



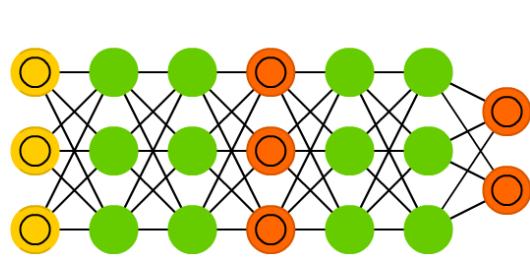
Deconvolutional Network (DN)



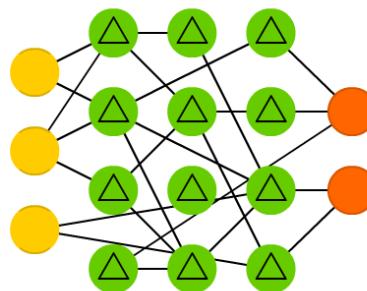
Deep Convolutional Inverse Graphics Network (DCIGN)



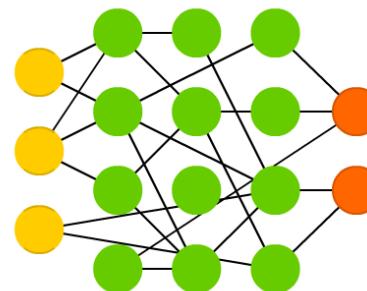
Generative Adversarial Network (GAN)



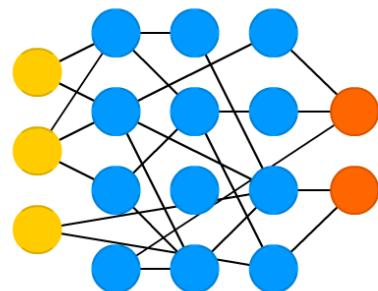
Liquid State Machine (LSM)



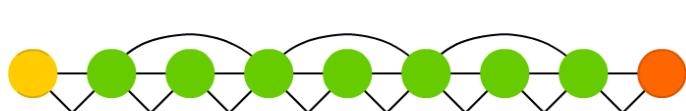
Extreme Learning Machine (ELM)



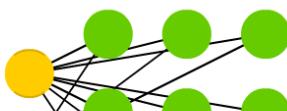
Echo State Network (ESN)



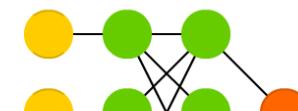
Deep Residual Network (DRN)



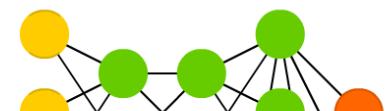
Kohonen Network (KN)

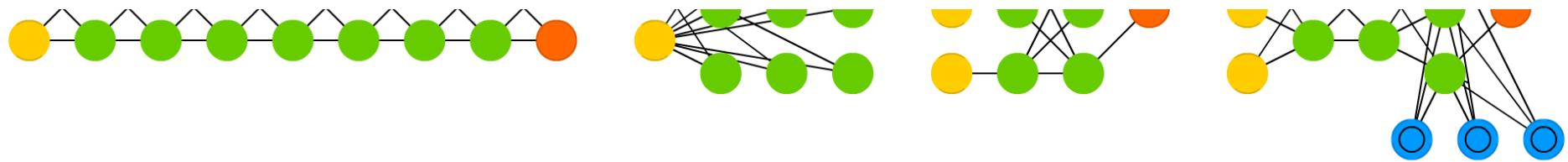


Support Vector Machine (SVM)



Neural Turing Machine (NTM)





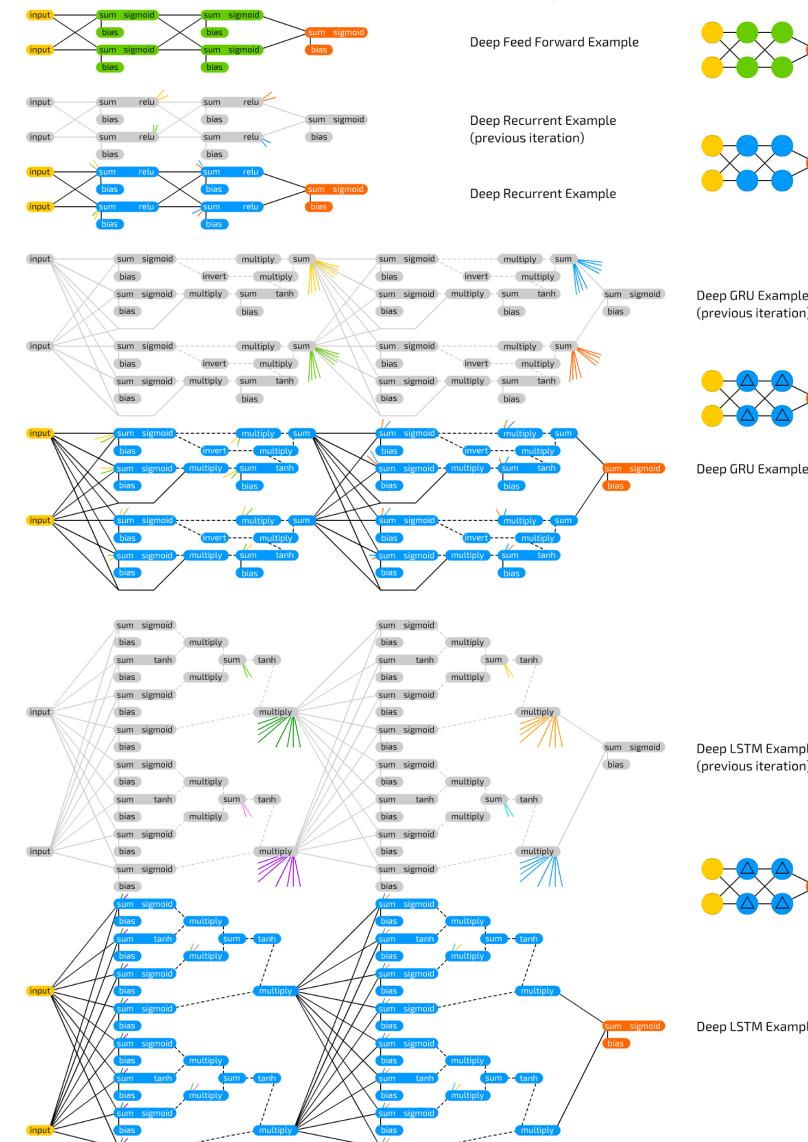
Neural Networks Cheat Sheet

## Neural Networks Graphs

An informative chart to build

# Neural Network Graphs

©2016 Fjodor van Veen - asimovinstitute.org



## Neural Networks Graphs Cheat Sheet

<b>Linear Vector Spaces:</b> Definition: A linear vector space, $X$ , is a set of elements (vectors) defined over a scalar field, $F$ , that satisfies the following conditions: 1) if $x \in X$ and $y \in X$ then $x+y \in X$ . 2) $\lambda \cdot x = y \in X$ . 3) $(x+y)+z = x+(y+z)$ 4) There is a unique vector $0 \in X$ , such that $x+0=x$ for all $x \in X$ . 5) For each vector $x \in X$ there is a unique vector in $X$ , to be called $(-x)$ , such that $x+(-x)=0$ . 6) multiplication, for all scalars $a \in F$ , and all vectors $x \in X$ , 7) For any $x \in X$ , $1x=x$ (for scalar 1). 8) For any two scalars $a, b \in F$ and $x \in X$ , $a(bx)=(ab)x$ . 9) $(a+b)x=a(x)+b(x)$ . 10) $a(x+y)=a x+a y$ .
<b>Linear Independence:</b> Consider $n$ vectors $\{x_1, x_2, \dots, x_n\}$ . If there exists $n$ scalars $a_1, a_2, \dots, a_n$ , at least one of which is nonzero, such that $a_1x_1+a_2x_2+\dots+a_nx_n=0$ , then the $\{x_i\}$ are linearly dependent.
<b>Spanning a Space:</b> Let $X$ be a linear vector space and let $\{u_1, u_2, \dots, u_m\}$ be a subset of vectors in $X$ . This subset spans $X$ if and only if for every vector $x \in X$ there exist scalars $x_1, x_2, \dots, x_m$ such that $x = x_1u_1+x_2u_2+\dots+x_mu_m$ .
<b>Inner Product:</b> Consider a scalar function of $x$ and $y$ . 1. $\langle x, y \rangle = \langle y, x \rangle$ 2. $\langle x, ay + bz \rangle = a\langle x, y \rangle + b\langle x, z \rangle$ 3. $\langle x, x \rangle \geq 0$ , where equality holds iff $x$ is the zero vector. <b>Norm:</b> A scalar function $\ x\ $ is called a norm if it satisfies: 1. $\ x\  \geq 0$ 2. $\ x\  = 0$ if and only if $x = 0$ . 3. $\ ax\  =  a \ x\ $ 4. $\ x+y\  \leq \ x\  + \ y\ $
<b>Angle:</b> The angle $\theta$ bet. 2 vectors $x$ and $y$ is defined by $\cos \theta = \frac{\langle x, y \rangle}{\ x\  \ y\ }$
<b>Orthogonality:</b> 2 vectors $x, y \in X$ are said to be orthogonal if $\langle x, y \rangle = 0$ .
<b>Gram Schmidt Orthogonalization:</b> Assume that we have $n$ independent vectors $y_1, y_2, \dots, y_n$ . From these vectors we will obtain $n$ orthogonal vectors $v_1, v_2, \dots, v_n$ . $v_1 = y_1, \quad v_k = y_k - \sum_{i=1}^{k-1} \frac{\langle v_i, y_k \rangle}{\langle v_i, v_i \rangle} v_i,$ where $\frac{\langle v_i, y_k \rangle}{\langle v_i, v_i \rangle} v_i$ is the projection of $y_k$ on $v_i$
<b>Vector Expansions:</b> $x = \sum_{i=1}^n x_i v_i = x_1 v_1 + x_2 v_2 + \dots + x_n v_n,$ for orthogonal vectors, $x_j = \frac{\langle v_j, x \rangle}{\langle v_j, v_j \rangle}$
<b>Reciprocal Basis Vectors:</b> $(v_i, v_j) = \begin{cases} 0 & i \neq j \\ 1 & i = j \end{cases}, \quad x_j = (v_j, x)$ To compute the reciprocal basis vectors: set $B = [v_1 \ v_2 \ \dots \ v_n]$ , $R = [r_1 \ r_2 \ \dots \ r_n]$ , $R^T = B^{-1}$ In matrix form: $x^p = B^{-1} x^s$
<b>Transformations:</b> A transformation consists of three parts: domain: $X = \{x_i\}$ , range: $Y = \{y_i\}$ , and a rule relating each $x_i \in X$ to an element $y_i \in Y$ . <b>Linear Transformations:</b> transformation $A$ is linear if: 1. for all $x_1, x_2 \in X$ , $A(x_1+x_2) = A(x_1) + A(x_2)$ 2. for all $x \in X$ , $a \in R$ , $A(ax) = aA(x)$
<b>Matrix Representations:</b> Let $\{v_1, v_2, \dots, v_n\}$ be a basis for vector space $X$ , and let $\{u_1, u_2, \dots, u_n\}$ be a basis for vector space $Y$ . Let $A$ be a linear transformation with domain $X$ and range $Y$ : $A(x) = y$ The coefficients of the matrix representation are obtained from $A(v_i) = \sum_{j=1}^m a_{ij} u_j$
<b>Change of Basis:</b> $B_t = [t_1 \ t_2 \ \dots \ t_n]$ , $B_w = [w_1 \ w_2 \ \dots \ w_n]$ $A' = [B_t^{-1} A B_t]$
<b>Eigenvalues &amp; Eigenvectors:</b> $Az = \lambda z$ , $[(A - \lambda I)] = 0$
<b>Diagonalization:</b> $B = [z_1 \ z_2 \ \dots \ z_n]$ , where $\{z_1, z_2, \dots, z_n\}$ are the eigenvectors of a square matrix $A$ , $[B^{-1}AB] = \text{diag}(\lambda_1 \ \lambda_2 \ \dots \ \lambda_n)$

<b>Perceptron Architecture:</b> $\mathbf{a} = \text{hardlim}(\mathbf{W}\mathbf{p} + \mathbf{b}), \quad \mathbf{W} = [\mathbf{l}_1 \mathbf{w}^T \ \mathbf{l}_2 \mathbf{w}^T \ \dots \ \mathbf{l}_s \mathbf{w}^T]^T,$ $a_t = \text{hardlim}(n_t) = \text{hardlim}(\mathbf{l}_t \mathbf{w}^T \mathbf{p} + b_t)$ <b>Decision Boundary:</b> $\mathbf{l}^T \mathbf{w} + b = 0$ The decision boundary is always orthogonal to the weight vector. Single-layer perceptrons can only classify linearly separable vectors.
<b>Perceptron Learning Rule:</b> $\mathbf{W}^{\text{new}} = \mathbf{W}^{\text{old}} + \epsilon \mathbf{p}^T, \quad \mathbf{b}^{\text{new}} = \mathbf{b}^{\text{old}} + \epsilon,$ where $\mathbf{e} = \mathbf{t} - \mathbf{a}$ ,
<b>Hebb's Postulate:</b> "When an axon of cell A is near enough to excite a cell B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A's efficiency, as one of the cells firing B, is increased."
<b>Linear Associator:</b> $\mathbf{a} = \text{purelin}(\mathbf{W}\mathbf{p})$
<b>The Hebb Rule:</b> Supervised Form: $w_{ij}^{\text{new}} = w_{ij}^{\text{old}} + t_{qi}P_{qi}$ $\mathbf{W} = t_1 \mathbf{P}^T + t_2 \mathbf{P}_2^T + \dots + t_q \mathbf{P}_q^T$ $\mathbf{W} = [\mathbf{t}_1 \ \mathbf{t}_2 \ \dots \ \mathbf{t}_Q] \begin{bmatrix} \mathbf{P}_1^T \\ \mathbf{P}_2^T \\ \vdots \\ \mathbf{P}_Q^T \end{bmatrix} = \mathbf{T} \mathbf{P}^T$
<b>Pseudoinverse Rule:</b> $\mathbf{W} = \mathbf{T} \mathbf{P}^+$ When the number, $R$ , of rows of $\mathbf{P}$ is greater than the number of columns, $Q$ , of $\mathbf{P}$ and the columns of $\mathbf{P}$ are independent, then the pseudoinverse can be computed by $\mathbf{P}^+ = (\mathbf{P}^T \mathbf{P})^{-1} \mathbf{P}^T$
<b>Variations of Hebbian Learning:</b>
<b>Filtered Learning</b> (Ch.16): $\mathbf{W}^{\text{new}} = (1-\gamma)\mathbf{W}^{\text{old}} + \alpha t \mathbf{p}^T$
<b>Delta Rule</b> (Ch.10): $\mathbf{W}^{\text{new}} = \mathbf{W}^{\text{old}} + \alpha(t_q - a_q) \mathbf{p}_q^T$
<b>Unsupervised Hebb</b> (Ch.13): $\mathbf{W}^{\text{new}} = \mathbf{W}^{\text{old}} + \alpha a_q \mathbf{p}_q^T$
<b>Taylor:</b> $F(\mathbf{x}) = F(\mathbf{x}^*) + \nabla F(\mathbf{x})^T _{\mathbf{x}=\mathbf{x}^*}(\mathbf{x} - \mathbf{x}^*) + \frac{1}{2}(\mathbf{x} - \mathbf{x}^*)^T \nabla^2 F(\mathbf{x})^T _{\mathbf{x}=\mathbf{x}^*}(\mathbf{x} - \mathbf{x}^*) + \dots$
<b>Grad</b> $\nabla F(\mathbf{x}) = \left[ \frac{\partial}{\partial x_1} F(\mathbf{x}) \ \frac{\partial}{\partial x_2} F(\mathbf{x}) \ \dots \ \frac{\partial}{\partial x_n} F(\mathbf{x}) \right]^T$
<b>Hessian:</b> $\nabla^2 F(\mathbf{x}) = \begin{bmatrix} \frac{\partial^2}{\partial x_1^2} F(\mathbf{x}) & \frac{\partial^2}{\partial x_1 \partial x_2} F(\mathbf{x}) & \dots & \frac{\partial^2}{\partial x_1 \partial x_n} F(\mathbf{x}) \\ \frac{\partial^2}{\partial x_2 \partial x_1} F(\mathbf{x}) & \frac{\partial^2}{\partial x_2^2} F(\mathbf{x}) & \dots & \frac{\partial^2}{\partial x_2 \partial x_n} F(\mathbf{x}) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial x_n \partial x_1} F(\mathbf{x}) & \frac{\partial^2}{\partial x_n \partial x_2} F(\mathbf{x}) & \dots & \frac{\partial^2}{\partial x_n^2} F(\mathbf{x}) \end{bmatrix}$
<b>Directional Derivatives:</b>
<b>1<sup>st</sup> Dir.Der.:</b> $\frac{\mathbf{p}^T \nabla F(\mathbf{x})}{\ \mathbf{p}\ }$ , <b>2<sup>nd</sup> Dir.Der.:</b> $\frac{\mathbf{p}^T \nabla^2 F(\mathbf{x}) \mathbf{p}}{\ \mathbf{p}\ ^2}$
<b>Minima:</b> Strong Minimum: if a scalar $\delta > 0$ exists, such that $F(x) < F(x+\Delta x)$ for all $\Delta x$ such that $\delta > \ \Delta x\  > 0$ . Global Minimum: if $F(x) < F(x+\Delta x)$ for all $\Delta x \neq 0$ Weak Minimum: if it is not a strong minimum, and a scalar $\delta > 0$ exists, such that $F(x) \leq F(x+\Delta x)$ for all $\Delta x$ such that $\delta > \ \Delta x\  > 0$ .
<b>Necessary Conditions for Optimality:</b>
<b>1<sup>st</sup>-Order Condition:</b> $\nabla F(\mathbf{x}) _{\mathbf{x}=\mathbf{x}^*} = 0$ (Stationary Points) <b>2<sup>nd</sup>-Order Condition:</b> $\nabla^2 F(\mathbf{x}) _{\mathbf{x}=\mathbf{x}^*} \geq 0$ (Positive Semi-definite Hessian Matrix).
<b>Quadratic fn.:</b> $F(\mathbf{x}) = \frac{1}{2} \mathbf{x}^T \mathbf{A} \mathbf{x} + \mathbf{b}^T \mathbf{x} + c$ $\nabla F(\mathbf{x}) = \mathbf{A} \mathbf{x} + \mathbf{d}, \quad \nabla^2 F(\mathbf{x}) = \mathbf{A}, \quad \lambda_{\min} \leq \frac{\mathbf{p}^T \mathbf{A} \mathbf{p}}{\ \mathbf{p}\ ^2} \leq \lambda_{\max}$

<b>General Minimization Algorithm:</b> $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{p}_k \quad \text{or} \quad \Delta \mathbf{x}_k = (\mathbf{x}_{k+1} - \mathbf{x}_k) = \alpha_k \mathbf{p}_k$ <b>Steepest Descent Algorithm:</b> $\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \mathbf{g}_k \quad \text{where, } \mathbf{g}_k = \nabla F(\mathbf{x}) _{\mathbf{x}=\mathbf{x}_k}$ <b>Stable Learning Rate:</b> ( $\alpha_k = \alpha$ , constant) $\alpha < \lambda_{\max}^{-2}$ $\{\lambda_1, \lambda_2, \dots, \lambda_n\}$ Eigenvalues of Hessian matrix $\mathbf{A}$ <b>Learning Rate to Minimize Along the Line:</b> $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{p}_k \Rightarrow \alpha_k = -\frac{\mathbf{g}_k^T \mathbf{p}_k}{\mathbf{p}_k^T \mathbf{A} \mathbf{p}_k}$ (For quadratic fn.)
<b>After Minimization Along the Line:</b> $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{p}_k \Rightarrow \mathbf{g}_{k+1}^T \mathbf{p}_k = 0$ <b>ADALINE:</b> $\mathbf{a} = \text{purelin}(\mathbf{W}\mathbf{p} + \mathbf{b})$ <b>Mean Square Error:</b> (for ADALINE it is a quadratic fn.) $F(\mathbf{x}) = E[e^2] = E[(t - a)^2] = E[(t - \mathbf{x}^T \mathbf{z})^2]$ $F(\mathbf{x}) = c - 2\mathbf{x}^T \mathbf{h} + \mathbf{x}^T \mathbf{R} \mathbf{x},$ $c = E[t^2], \quad \mathbf{h} = E[tz]^T \quad \text{and} \quad \mathbf{R} = E[z z^T] \Rightarrow \mathbf{A} = 2\mathbf{R}, \quad \mathbf{d} = -2\mathbf{h}$ Unique minimum, if it exists, is $\mathbf{x}^* = \mathbf{R}^{-1} \mathbf{h}$ , where $\mathbf{x} = \begin{bmatrix} 1 & \mathbf{w} \\ b & 1 \end{bmatrix}$ and $\mathbf{z} = \begin{bmatrix} \mathbf{p} \\ 1 \end{bmatrix}$
<b>LMS Algorithm:</b> $\mathbf{W}(k+1) = \mathbf{W}(k) + 2\alpha e(k) \mathbf{p}^T(k)$ $b(k+1) = b(k) + \alpha e(k)$ <b>Convergence Point:</b> $\mathbf{x}^* = \mathbf{R}^{-1} \mathbf{h}$ <b>Stable Learning Rate:</b> $0 < \alpha < 1/\lambda_{\max}$ where $\lambda_{\max}$ is the maximum eigenvalue of $\mathbf{R}$
<b>Adaptive Filter ADALINE:</b> $a(k) = \text{purelin}(\mathbf{W}\mathbf{p}(k) + b) = \sum_{i=1}^R \mathbf{w}_{1,i} y_i(k-i+1) + b$
<b>Backpropagation Algorithm:</b> <b>Performance Index:</b> Mean Square error: $F(\mathbf{x}) = E[\mathbf{e}^T \mathbf{e}] = E[(t - a)^2]$ Approximate Performance Index: (single sample) $\hat{F}(\mathbf{x}) = \mathbf{e}^T \mathbf{e}(k) \text{ or } (\mathbf{t}(k) - \mathbf{a}(k))^T (\mathbf{t}(k) - \mathbf{a}(k))$ Sensitivity: $\mathbf{s}^m = \frac{\partial \hat{F}}{\partial \mathbf{n}^m} = \left[ \frac{\partial \hat{F}}{\partial n_1^m} \ \frac{\partial \hat{F}}{\partial n_2^m} \ \dots \ \frac{\partial \hat{F}}{\partial n_s^m} \right]^T$ Forward Propagation: $\mathbf{a}^0 = \mathbf{p}$ , $\mathbf{a}^{m+1} = \mathbf{f}^{m+1}(\mathbf{W}^{m+1} \mathbf{a}^m + \mathbf{b}^{m+1}) \text{ for } m = 0, 1, \dots, M-1$ $\mathbf{a} = \mathbf{a}^M$ Backward Propagation: $\mathbf{s}^M = -2\mathbf{F}^M(\mathbf{n}^M)(\mathbf{t} - \mathbf{a})$ , $\mathbf{s}^m = \hat{F}^m(\mathbf{n}^m)(\mathbf{W}^{m+1})^T \mathbf{s}^{m+1} \text{ for } m = M-1, \dots, 2, 1, \text{ where}$ $\hat{F}^m(\mathbf{n}^m) = \text{diag}([\hat{f}^m(n_1^m) \ \hat{f}^m(n_2^m) \ \dots \ \hat{f}^m(n_s^m)])$ $\hat{f}^m(n_j^m) = \frac{\partial \hat{f}^m(n_j^m)}{\partial n_j^m}$ <b>Weight Update (Approximate Steepest Descent):</b> $\mathbf{W}^{m+1} = \mathbf{W}^m - \alpha \mathbf{s}^m (\mathbf{a}^{m+1})^T$ $\mathbf{b}^{m+1} = \mathbf{b}^m - \alpha \mathbf{s}^m$
<b>hardlim:</b> $a = \begin{cases} 1 & n < 0 \\ 0 & n \geq 0 \end{cases}$ , <b>hardlims:</b> $a = \begin{cases} -1 & n < 0 \\ +1 & n \geq 0 \end{cases}$ , <b>purelin:</b> $a = n$ , <b>Logsig:</b> $a = \frac{1}{1+e^{-n}}$ , <b>tansig:</b> $a = \frac{e^{n-e^{-n}}}{e^{n+e^{-n}}}$ , <b>postlin:</b> $a = \begin{cases} 0 & n < 0 \\ n & n \geq 0 \end{cases}$
<b>compt:</b> $a = \begin{cases} 1 & \text{neuron with max } n \\ 0 & \text{all other neurons} \end{cases}$ , <b>satlin:</b> $a = \begin{cases} 0 & n < 0 \\ 1 & -1 \leq n \leq 1, \\ 1 & n > 1 \end{cases}$ , <b>satlns:</b> $a = \begin{cases} -1 & n < 0 \\ 1 & -1 \leq n \leq 1, \\ 1 & n > 1 \end{cases}$
<b>Delay:</b> $a(t) = u(t-1)$ , <b>Integrator:</b> $a(t) = \int_0^t u(\tau) d\tau + a(0)$
<b>**HINT:</b> $\text{diag}([1 \ 2 \ 3]) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}$

## Neural Network Cheat Sheet

Ultimate Guide to Leveraging NLP & Machine Learning for your Chatbot

Code Snippets and Github Included

[chatbotslife.com](http://chatbotslife.com)



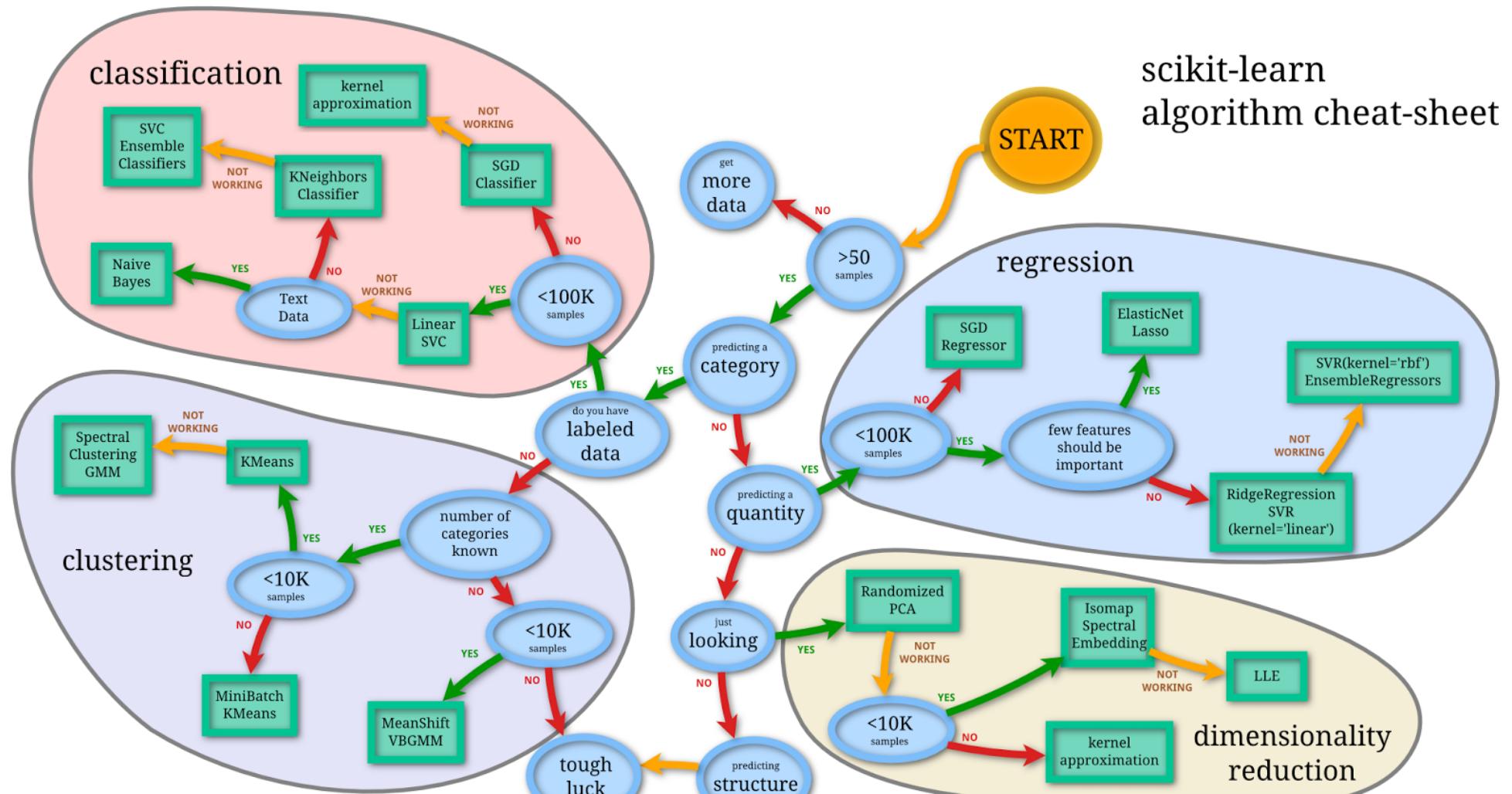
## Machine Learning Overview



Machine Learning Cheat Sheet

## Machine Learning: Scikit-learn algorithm

This machine learning cheat sheet will help you find the right estimator for the job which is the most difficult part. The flowchart will help you check the documentation and rough guide of each estimator that will help you to know more about the problems and how to solve it.



Machine Learning Cheat Sheet

## Scikit-Learn

**Scikit-learn** (formerly `scikits.learn`) is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.

# Python For Data Science Cheat Sheet

## Scikit-Learn

Learn Python for data science interactively at [www.DataCamp.com](http://www.DataCamp.com)



### Scikit-learn

Scikit-learn is an open source Python library that implements a range of machine learning, preprocessing, cross-validation and visualization algorithms using a unified interface.



#### A Basic Example

```
>>> from sklearn import neighbors, datasets, preprocessing
>>> from sklearn.cross_validation import train_test_split
>>> from sklearn.metrics import accuracy_score
>>> iris = datasets.load_iris()
>>> X, y = iris.data[:, :2], iris.target
>>> X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=33)
>>> scaler = preprocessing.StandardScaler().fit(X_train)
>>> X_train = scaler.transform(X_train)
>>> X_test = scaler.transform(X_test)
>>> knn = neighbors.KNeighborsClassifier(n_neighbors=5)
>>> knn.fit(X_train, y_train)
>>> y_pred = knn.predict(X_test)
>>> accuracy_score(y_test, y_pred)
```

#### Loading The Data

Also see NumPy & Pandas

Your data needs to be numeric and stored as NumPy arrays or SciPy sparse matrices. Other types that are convertible to numeric arrays, such as Pandas DataFrame, are also acceptable.

```
>>> import numpy as np
>>> X = np.random.random((10,5))
>>> y = np.array(['M','M','F','F','M','F','M','M','F','F'])
>>> X[X < 0.7] = 0
```

#### Training And Test Data

```
>>> from sklearn.cross_validation import train_test_split
>>> X_train, X_test, y_train, y_test = train_test_split(X,
y,
random_state=0)
```

#### Preprocessing The Data

##### Standardization

```
>>> from sklearn.preprocessing import StandardScaler
>>> scaler = StandardScaler().fit(X_train)
>>> standardized_X = scaler.transform(X_train)
>>> standardized_X_test = scaler.transform(X_test)
```

##### Normalization

```
>>> from sklearn.preprocessing import Normalizer
>>> scaler = Normalizer().fit(X_train)
>>> normalized_X = scaler.transform(X_train)
>>> normalized_X_test = scaler.transform(X_test)
```

##### Binarization

```
>>> from sklearn.preprocessing import Binarizer
```

## Create Your Model

### Supervised Learning Estimators

#### Linear Regression

```
>>> from sklearn.linear_model import LinearRegression
>>> lr = LinearRegression(normalize=True)
```

#### Support Vector Machines (SVM)

```
>>> from sklearn.svm import SVC
>>> svc = SVC(kernel='linear')
```

#### Naive Bayes

```
>>> from sklearn.naive_bayes import GaussianNB
>>> gnb = GaussianNB()
```

#### KNN

```
>>> from sklearn import neighbors
>>> knn = neighbors.KNeighborsClassifier(n_neighbors=5)
```

### Unsupervised Learning Estimators

#### Principal Component Analysis (PCA)

```
>>> from sklearn.decomposition import PCA
>>> pca = PCA(n_components=0.95)
```

#### K Means

```
>>> from sklearn.cluster import KMeans
>>> k_means = KMeans(n_clusters=3, random_state=0)
```

## Model Fitting

### Supervised learning

```
>>> lr.fit(X, y)
>>> knn.fit(X_train, y_train)
>>> svc.fit(X_train, y_train)
```

### Unsupervised Learning

```
>>> k_means.fit(X_train)
>>> pca_model = pca.fit_transform(X_train)
```

Fit the model to the data

Fit the model to the data

Fit to data, then transform it

## Prediction

### Supervised Estimators

```
>>> y_pred = svc.predict(np.random.random((2,5)))
>>> y_pred = lr.predict(X_test)
>>> y_pred = knn.predict_proba(X_test)
```

### Unsupervised Estimators

```
>>> y_pred = k_means.predict(X_test)
```

Predict labels

Predict labels

Estimate probability of a label

Predict labels in clustering algos

### Encoding Categorical Features

```
>>> from sklearn.preprocessing import LabelEncoder
>>> enc = LabelEncoder()
>>> y = enc.fit_transform(y)
```

### Imputing Missing Values

```
>>> from sklearn.preprocessing import Imputer
>>> imp = Imputer(missing_values=0, strategy='mean', axis=0)
>>> imp.fit_transform(X_train)
```

### Generating Polynomial Features

```
>>> from sklearn.preprocessing import PolynomialFeatures
```

## Evaluate Your Model's Performance

### Classification Metrics

#### Accuracy Score

```
>>> knn.score(X_test, y_test)
>>> from sklearn.metrics import accuracy_score
>>> accuracy_score(y_test, y_pred)
```

Estimator score method  
Metric scoring functions

#### Classification Report

```
>>> from sklearn.metrics import classification_report
>>> print(classification_report(y_test, y_pred))
```

Precision, recall, f1-score and support

#### Confusion Matrix

```
>>> from sklearn.metrics import confusion_matrix
>>> print(confusion_matrix(y_test, y_pred))
```

### Regression Metrics

#### Mean Absolute Error

```
>>> from sklearn.metrics import mean_absolute_error
>>> y_true = [3, -0.5, 2]
>>> mean_absolute_error(y_true, y_pred)
```

#### Mean Squared Error

```
>>> from sklearn.metrics import mean_squared_error
>>> mean_squared_error(y_test, y_pred)
```

#### R<sup>2</sup> Score

```
>>> from sklearn.metrics import r2_score
>>> r2_score(y_true, y_pred)
```

### Clustering Metrics

#### Adjusted Rand Index

```
>>> from sklearn.metrics import adjusted_rand_score
>>> adjusted_rand_score(y_true, y_pred)
```

#### Homogeneity

```
>>> from sklearn.metrics import homogeneity_score
>>> homogeneity_score(y_true, y_pred)
```

#### V-measure

```
>>> from sklearn.metrics import v_measure_score
>>> metrics.v_measure_score(y_true, y_pred)
```

### Cross-Validation

```
>>> from sklearn.cross_validation import cross_val_score
>>> print(cross_val_score(knn, X_train, y_train, cv=4))
>>> print(cross_val_score(lr, X, y, cv=2))
```

## Tune Your Model

### Grid Search

```
>>> from sklearn.grid_search import GridSearchCV
>>> params = {"n_neighbors": np.arange(1,3),
"metric": ["euclidean", "cityblock"]}
>>> grid = GridSearchCV(estimator=knn,
param_grid=params)
>>> grid.fit(X_train, y_train)
>>> print(grid.best_score_)
>>> print(grid.best_estimator_.n_neighbors)
```

### Randomized Parameter Optimization

```
>>> from sklearn.grid_search import RandomizedSearchCV
>>> params = {"n_neighbors": range(1,5),
"weights": ["uniform", "distance"]}
>>> rsearch = RandomizedSearchCV(estimator=knn,
param_distributions=params,
cv=4,
n_iter=8,
random_state=5)
>>> rsearch.fit(X_train, y_train)
>>> print(rsearch.best_score_)
```

```
>>> binarizer = Binarizer(threshold=0.0).fit(X)  
>>> binary_X = binarizer.transform(X)
```

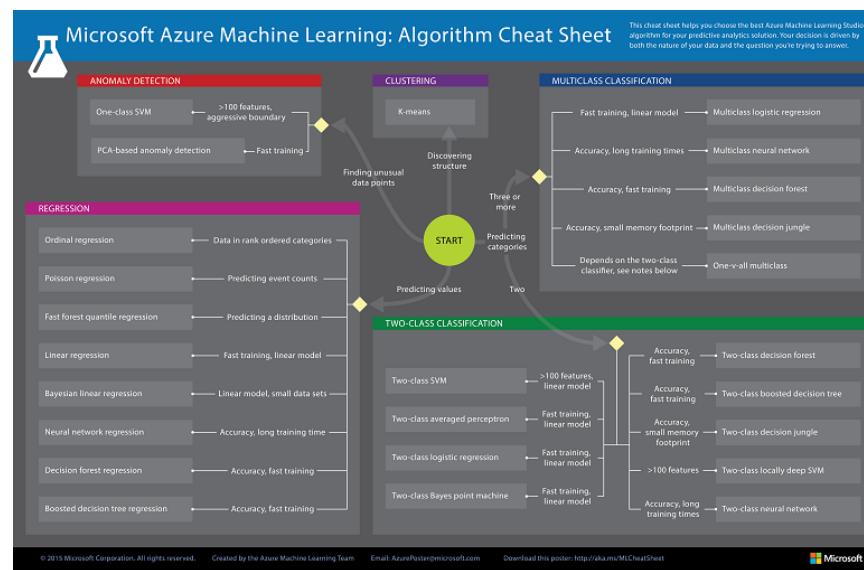
```
>>> poly = PolynomialFeatures(5)  
>>> poly.fit_transform(X)
```



## Scikit-Learn Cheat Sheet

# MACHINE LEARNING : ALGORITHM CHEAT SHEET

*This machine learning cheat sheet from Microsoft Azure will help you choose the appropriate machine learning algorithms for your predictive analytics solution. First, the cheat sheet will ask you about the data nature and then suggests the best algorithm for the job.*



MACHINE LEARNING ALGORITHM CHEAT SHEET

>>> If you like this list, you can let me know [here](#). <<

Ultimate Guide to Leveraging NLP & Machine Learning for your Chatbot

Code Snippets and Github Included

[chatbotslife.com](http://chatbotslife.com)



## Python for Data Science

**Python For Data Science Cheat Sheet**

Python Basics  
Learn More Python for Data Science interactively at [www.datacamp.com](http://www.datacamp.com)

**Variables and Data Types**

Variable Assignment
>>> a = 'is' >>> b = 'nice' >>> my_list = ['my', 'list', a, b] >>> my_list2 = [[4,5,6,7], [3,4,5,6]]

**Calculations With Variables**

Operations	Description
>>> x+2	Sum of two variables
>>> x-2	Subtraction of two variables
>>> x*2	Multiplication of two variables
>>> x**2	Exponentiation of a variable
>>> x%2	Remainder of a variable
>>> x/float(2)	Division of a variable

**Types and Type Conversion**

Type	Value	Description
str()	'5', '3.45', 'True'	Variables to strings
int()	5, 3, 1	Variables to integers
float()	5.0, 1.0	Variables to floats
bool()	True, True, True	Variables to booleans

**Asking For Help**

```
>>> help(str)
```

**Strings**

```
>>> my_string = 'thisStringIsAwesome'  
>>> my_string  
'thisStringIsAwesome'
```

**String Operations**

```
>>> my_string * 2  
'thisStringIsAwesomethisStringIsAwesome'  
>>> my_string + "Init"  
'thisStringIsAwesomeInit'  
>>> "m" in my_string  
True
```

**Lists**

Selecting List Elements Index starts at 0

Operation	Description
>>> my_list[1]	Select item at index 1
>>> my_list[-3]	Select 3rd last item
>>> my_list[1:3]	Select items at index 1 and 2
>>> my_list[:1]	Select items after index 0
>>> my_list[1:-1]	Select items before index 3
>>> my_list[[1,2]]	Copy my_list
>>> my_list2[1][0]	my_list[[list][itemOfList]]
>>> my_list2[1][1:2]	

**List Operations**

```
>>> my_list + my_list  
'[my', 'list', 'is', 'nice', 'my', 'list', 'is', 'nice']'  
>>> my_list * 2  
'[my', 'list', 'is', 'nice', 'my', 'list', 'is', 'nice']'  
>>> my_list2 > 4
```

**List Methods**

Method	Description
>>> my_list.index('a')	Get the index of an item
>>> my_list.count('a')	Count the items
>>> my_list.append('c')	Append an item at a time
>>> my_list.remove('b')	Remove an item
>>> del(my_list[0:1])	Remove an item
>>> my_list.reverse()	Reverse the list
>>> my_list.extend('d')	Append an item
>>> my_list.insert(0, 'e')	Remove an item
>>> my_list.sort()	Sort the list

**Numpy Arrays**

Selecting Numpy Array Elements Index starts at 0

Operation	Description
>>> my_array[1]	Select item at index 1
>>> my_array[0:2]	Select items at index 0 and 1
>>> my_2darray[1,0]	Select items in array
>>> my_2darray[1,0:2]	my_2darray[rows, columns]

**Numpy Array Operations**

```
>>> my_array > 3  
array([False, False, False, True], dtype=bool)  
>>> my_array * 2  
array([ 0,  4,  6, 12])  
>>> my_array + np.array([5, 6, 7, 8])  
array([ 5,  6,  7, 12])
```

**Numpy Array Functions**

Function	Description
>>> my_array.shape	Get the dimensions of the array
>>> np.append(my_array, other_array)	Append items to an array
>>> np.insert(my_array, 1, 5)	Insert items in an array
>>> np.delete(my_array, [1])	Delete items in an array
>>> np.mean(my_array)	Mean of the array
>>> np.median(my_array)	Median of the array
>>> my_array.corrcoef()	Correlation coefficient
>>> np.std(my_array)	Standard deviation

## Python Data Science Cheat Sheet

# Python For Data Science Cheat Sheet

## Bokeh

Learn Bokeh [Interactively](#) at [www.DataCamp.com](http://www.DataCamp.com), taught by Bryan Van de Ven, core contributor

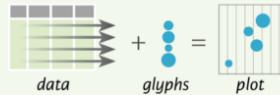


### Plotting With Bokeh

The Python interactive visualization library **Bokeh** enables high-performance visual presentation of large datasets in modern web browsers.



Bokeh's mid-level general purpose `bokeh.plotting` interface is centered around two main components: data and glyphs.



The basic steps to creating plots with the `bokeh.plotting` interface are:

#### 1. Prepare some data:

Python lists, NumPy arrays, Pandas DataFrames and other sequences of values

#### 2. Create a new plot

#### 3. Add renderers for your data, with visual customizations

#### 4. Specify where to generate the output

#### 5. Show or save the results

```
>>> from bokeh.plotting import figure
>>> from bokeh.io import output_file, show
>>> x = [1, 2, 3, 4, 5]           Step 1
>>> y = [6, 7, 2, 4, 5]
>>> p = figure(title="simple line example",      Step 2
              x_axis_label='x',
              y_axis_label='y')
>>> p.line(x, y, legend="Temp.", line_width=2)    Step 3
>>> output_file("lines.html")                  Step 4
>>> show(p)                                    Step 5
```

## 1 Data

### Also see Lists, NumPy & Pandas

Under the hood, your data is converted to Column Data Sources. You can also do this manually:

```
>>> import numpy as np
>>> import pandas as pd
>>> df = pd.DataFrame(np.array([[33.9, 4, 65, 'US'],
>                            [32.4, 4, 66, 'Asia'],
>                            [21.4, 4, 109, 'Europe']]),
>                            columns=['mpg', 'cyl', 'hp', 'origin']),
>                            index=['Toyota', 'Fiat', 'Volvo'])
```

## 2 Plotting

```
>>> from bokeh.plotting import figure
>>> p1 = figure(plot_width=300, tools='pan,box_zoom')
>>> p2 = figure(plot_width=300, plot_height=300)
```

## 3 Renderers & Visual Customizations

### Glyphs

#### Scatter Markers

```
>>> p1.circle(np.array([1,2,3]), np.array([3,2,1]),
             fill_color='white')
>>> p2.square(np.array([1.5,3.5,5.5]), [1,4,3],
             color='blue', size=1)
```

#### Line Glyphs

```
>>> p1.line([1,2,3,4], [3,4,5,6], line_width=2)
>>> p2.multi_line(pd.DataFrame([[1,2,3],[5,6,7]]),
                  pd.DataFrame([[3,4,5],[3,2,1]]),
                  color="blue")
```

### Rows & Columns Layout

#### Rows

```
>>> from bokeh.layouts import row
```

```
>>> layout = row(p1,p2, p3)
```

#### Columns

```
>>> from bokeh.layouts import columns
```

```
>>> layout = column(p1,p2,p3)
```

### Grid Layout

```
>>> from bokeh.layouts import gridplot
>>> row1 = [p1,p2]
>>> row2 = [p3]
>>> layout = gridplot([[p1,p2],[p3]])
```

### Tabbed Layout

```
>>> from bokeh.models.widgets import Panel, Tabs
>>> tab1 = Panel(child=p1, title="tab1")
>>> tab2 = Panel(child=p2, title="tab2")
>>> layout = Tabs(tabs=[tab1, tab2])
```

### Legends

#### Legend Location

##### Inside Plot Area

```
>>> p.legend.location = 'bottom_left'
```

##### Outside Plot Area

```
>>> r1 = p2.asterisk(np.array([1,2,3]), np.array([3,2,1]))
>>> r2 = p2.line([1,2,3,4], [3,4,5,6])
>>> legend = Legend(items=[("One", [p1, r1]), ("Two", [r2])], location=(0, -30))
>>> p.add_layout(legend, 'right')
```

## 4 Output

### Output to HTML File

```
>>> from bokeh.io import output_file, show
>>> output_file('my_bar_chart.html', mode='cdn')
```

### Notebook Output

```
>>> from bokeh.io import output_notebook, show
>>> output_notebook()
```

### Embedding

#### Standalone HTML

```
>>> from bokeh.embed import file_html
```

```
>>> html = file_html(p, CDN, "my_plot")
```

#### Components

```
>>> from bokeh.embed import components
>>> script, div = components(p)
```

## 5 Show or Save Your Plots

### Customized Glyphs

#### Also see Data

#### Selection and Non-Selection Glyphs

```
>>> p.circle('mpg', 'cyl', source=cds_df,
             selection_color='red',
             nonselection_alpha=0.1)
```

#### Hover Glyphs

```
>>> hover = HoverTool(tooltips=None, mode='vline')
>>> p.add_tools(hover)
```

#### Colormapping

```
>>> color_mapper = CategoricalColorMapper(
            factors=['Europe', 'Asia', 'US'],
            palette=['red', 'green', 'blue'])
>>> p.circle('mpg', 'cyl', source=cds_df,
             color=dict(field='origin',
                        transform=color_mapper),
             legend='Origin'))
```

#### Also see Data

### Linked Plots

#### Linked Axes

```
>>> p2.x_range = p1.x_range
>>> p2.y_range = p1.y_range
```

#### Linked Brushing

```
>>> p4 = figure(plot_width = 100, tools='box_select,lasso_select')
>>> p4.circle('mpg', 'cyl', source=cds_df)
>>> p5 = figure(plot_width = 200, tools='box_select,lasso_select')
>>> p5.circle('mpg', 'hp', source=cds_df)
>>> layout = row(p4,p5)
```

### Legend Orientation

```
>>> p.legend.orientation = "horizontal"
>>> p.legend.orientation = "vertical"
```

#### Legend Background & Border

```
>>> p.legend.border_line_color = "navy"
>>> p.legend.background_fill_color = "white"
```

## Statistical Charts With Bokeh

#### Also see Data

Bokeh's high-level `bokeh.charts` interface is ideal for quickly creating statistical charts

### Bar Chart

```
>>> from bokeh.charts import Bar
>>> p = Bar(df, stacked=True, palette=['red','blue'])
```

### Box Plot

```
>>> from bokeh.charts import BoxPlot
>>> p = BoxPlot(df, values='vals', label='cyl',
                legend='bottom_right')
```

### Histogram

```
>>> from bokeh.charts import Histogram
>>> p = Histogram(df, title='Histogram')
```

### Scatter Plot

```
>>> from bokeh.charts import Scatter
>>> p = Scatter(df, x='mpg', y='hp', marker='square',
                xlabel='Miles Per Gallon',
                ylabel='Horsepower')
```

```
>>> px = plt.figure(figsize=(8, 8), facecolor='white', edgecolor='black',  
x_range=(0, 8), y_range=(0, 8))  
>>> p3 = figure()
```

```
>>> show(px)  
>>> show(layout)
```

```
>>> save(px)  
>>> save(layout)
```



## Big Data Cheat Sheet

## TensorFlow

In May 2017 Google announced the second-generation of the TPU, as well as the availability of the TPUs in [Google Compute Engine](#).[12] The second-generation TPUs deliver up to 180 teraflops of performance, and when organized into clusters of 64 TPUs provide up to 11.5 petaflops.



## About

### TensorFlow

TensorFlow™ is an open source software library for numerical computation using data flow graphs. TensorFlow was originally developed for the purposes of conducting machine learning and deep neural networks research, but the system is general enough to be applicable in a wide variety of other domains as well.

### Skflow

Scikit Flow provides a set of high level model classes that you can use to easily integrate with your existing Scikit-learn pipeline code. Scikit Flow is a simplified interface for TensorFlow, to get people started on predictive analytics and data mining. Scikit Flow has been merged into TensorFlow since version 0.8 and now called TensorFlow Learn.

### Keras

Keras is a minimalist, highly modular neural networks library, written in Python and capable of running on top of either TensorFlow or Theano

## Installation

### How to install new package in Python:

`pip install <package-name>`

Example: `pip install requests`

### How to install tensorflow?

### `id(object)`

Return the identity of an object. This is guaranteed to be unique among simultaneously existing objects.

### `import __builtin__ dir(__builtin__)`

Other built-in functions

## TensorFlow

### Main classes

`tf.Graph()`  
`tf.Operation()`  
`tf.Tensor()`  
`tf.Session()`

### Some useful functions

`tf.get_default_session()`  
`tf.get_default_graph()`  
`tf.reset_default_graph()`  
`ops.reset_default_graph()`  
`tf.device("/cpu:0")`  
`tf.name_scope(value)`  
`tf.convert_to_tensor(value)`

### TensorFlow Optimizers

`GradientDescentOptimizer`  
`AdadeltaOptimizer`  
`AdagradOptimizer`  
`MomentumOptimizer`  
`AdamOptimizer`  
`FtrlOptimizer`  
`RMSPropOptimizer`

### Reduction

### TensorFlowEstimator

### Each classifier and regressor have following fields

`n_classes=0 (Regressor), n_classes are expected to be input (Classifiers)`

`batch_size=32,`  
`steps=200, // except`  
`TensorFlowRNNClassifier - there is 50`  
`optimizer='Adagrad',`  
`learning_rate=0.1,`

```
device = cpu/gpu
python_version = cp27/cp34
sudo pip install
https://storage.googleapis.com/
tensorflow/linux/$device/tensorflow-
0.8.0-$python_version-none-linux_x86
_64.whl
```

### How to install Skflow

`pip install sklearn`

### How to install Keras

`pip install keras`

update `~/.keras/keras.json` - replace  
“theano” by “tensorflow”

## Helpers

### Python helper

#### Important functions

**`type(object)`**

Get object type

**`help(object)`**

Get help for object (list of available  
methods, attributes, signatures and so on)

**`dir(object)`**

Get list of object attributes  
(fields, functions)

**`str(object)`**

Transform an object to string

**`object?`**

Shows documentations about the object

**`globals()`**

Return the dictionary containing the  
current scope's global variables.

**`locals()`**

Update and return a dictionary containing  
the current scope's local variables.

`reduce_sum`  
`reduce_prod`  
`reduce_min`  
`reduce_max`  
`reduce_mean`  
`reduce_all`  
`reduce_any`  
`accumulate_n`

### Activation functions

`tf.nn?`

`relu`  
`relu6`  
`elu`  
`softplus`  
`softsign`  
`dropout`  
`bias_add`  
`sigmoid`  
`tanh`  
`sigmoid_cross_entropy_with_logits`  
`softmax`  
`log_softmax`  
`softmax_cross_entropy_with_logits`  
`sparse_softmax_cross_entropy_with_logits`  
`weighted_cross_entropy_with_logits`  
etc.

## Skflow

### Main classes

`TensorFlowClassifier`  
`TensorFlowRegressor`  
`TensorFlowDNNClassifier`  
`TensorFlowDNNRegressor`  
`TensorFlowLinearClassifier`  
`TensorFlowLinearRegressor`  
`TensorFlowRNNClassifier`  
`TensorFlowRNNRegressor`



### TesorFlow Cheat Sheet

## Keras

In 2017, Google's TensorFlow team decided to support Keras in TensorFlow's core library. Chollet explained that Keras was conceived to be an interface rather than an end-to-end machine-learning framework. It presents a higher-level, more intuitive set of abstractions that make it easy to configure neural networks regardless of the backend scientific computing library.

# Python For Data Science Cheat Sheet

## Keras

Learn Python for data science interactively at [www.DataCamp.com](http://www.DataCamp.com)



### Keras

Keras is a powerful and easy-to-use deep learning library for Theano and TensorFlow that provides a high-level neural networks API to develop and evaluate deep learning models.

#### A Basic Example

```
>>> import numpy as np
>>> from keras.models import Sequential
>>> from keras.layers import Dense
>>> data = np.random.random((1000,100))
>>> labels = np.random.randint(2,size=(1000,1))
>>> model = Sequential()
>>> model.add(Dense(32,
    activation='relu',
    input_dim=100))
>>> model.add(Dense(1, activation='sigmoid'))
>>> model.compile(optimizer='rmsprop',
    loss='binary_crossentropy',
    metrics=['accuracy'])
>>> model.fit(data,labels,epochs=10,batch_size=32)
>>> predictions = model.predict(data)
```

### Data

[Also see NumPy, Pandas & Scikit-Learn](#)

Your data needs to be stored as NumPy arrays or as a list of NumPy arrays. Ideally, you split the data in training and test sets, for which you can also resort to the `train_test_split` module of `sklearn.cross_validation`.

### Keras Data Sets

```
>>> from keras.datasets import boston_housing,
    mnist,
    cifar10,
    imdb
>>> (x_train,y_train), (x_test,y_test) = mnist.load_data()
>>> (x_train2,y_train2), (x_test2,y_test2) = boston_housing.load_data()
>>> (x_train3,y_train3), (x_test3,y_test3) = cifar10.load_data()
>>> (x_train4,y_train4), (x_test4,y_test4) = imdb.load_data(num_words=20000)
>>> num_classes = 10
```

### Other

```
>>> from urllib.request import urlopen
>>> data = np.loadtxt(urlopen("http://archive.ics.uci.edu/ml/machine-learning-databases/pima-indians-diabetes/pima-indians-diabetes.data"), delimiter=",")
>>> X = data[:,0:8]
>>> y = data[:,8]
```

### Preprocessing

#### Sequence Padding

```
>>> from keras.preprocessing import sequence
>>> x_train4 = sequence.pad_sequences(x_train4,maxlen=80)
>>> x_test4 = sequence.pad_sequences(x_test4,maxlen=80)
```

#### One-Hot Encoding

```
>>> from keras.utils import to_categorical
>>> Y_train = to_categorical(y_train, num_classes)
>>> Y_test = to_categorical(y_test, num_classes)
```

## Model Architecture

### Sequential Model

```
>>> from keras.models import Sequential
>>> model = Sequential()
>>> model2 = Sequential()
>>> model3 = Sequential()
```

### Multilayer Perceptron (MLP)

#### Binary Classification

```
>>> from keras.layers import Dense
>>> model.add(Dense(12,
    input_dim=8,
    kernel_initializer='uniform',
    activation='relu'))
>>> model.add(Dense(8,kernel_initializer='uniform',activation='relu'))
>>> model.add(Dense(1,kernel_initializer='uniform',activation='sigmoid'))
```

#### Multi-Class Classification

```
>>> from keras.layers import Dropout
>>> model.add(Dense(512,activation='relu',input_shape=(784,)))
>>> model.add(Dropout(0.2))
>>> model.add(Dense(512,activation='relu'))
>>> model.add(Dropout(0.2))
>>> model.add(Dense(10,activation='softmax'))
```

#### Regression

```
>>> model.add(Dense(64,activation='relu',input_dim=train_data.shape[1]))
>>> model.add(Dense(1))
```

### Convolutional Neural Network (CNN)

```
>>> from keras.layers import Activation,Conv2D,MaxPooling2D,Flatten
>>> model2.add(Conv2D(32,(3,3),padding='same',input_shape=x_train.shape[1:]))
>>> model2.add(Activation('relu'))
>>> model2.add(Conv2D(32,(3,3)))
>>> model2.add(Activation('relu'))
>>> model2.add(MaxPooling2D(pool_size=(2,2)))
>>> model2.add(Dropout(0.25))
>>> model2.add(Conv2D(64,(3,3), padding='same'))
>>> model2.add(Activation('relu'))
>>> model2.add(Conv2D(64,(3, 3)))
>>> model2.add(Activation('relu'))
>>> model2.add(MaxPooling2D(pool_size=(2,2)))
>>> model2.add(Dropout(0.25))
>>> model2.add(Flatten())
>>> model2.add(Dense(512))
>>> model2.add(Activation('relu'))
>>> model2.add(Dropout(0.5))
>>> model2.add(Dense(num_classes))
>>> model2.add(Activation('softmax'))
```

### Recurrent Neural Network (RNN)

```
>>> from keras.layers import Embedding,LSTM
>>> model3.add(Embedding(20000,128))
>>> model3.add(LSTM(128,dropout=0.2,recurrent_dropout=0.2))
>>> model3.add(Dense(1,activation='sigmoid'))
```

[Also see NumPy & Scikit-Learn](#)

### Train and Test Sets

```
>>> from sklearn.model_selection import train_test_split
>>> X_train5,X_test5,y_train5,y_test5 = train_test_split(X,
    y,
    test_size=0.33,
    random_state=42)
```

### Standardization/Normalization

```
>>> from sklearn.preprocessing import StandardScaler
>>> scaler = StandardScaler().fit(x_train5)
```

## Inspect Model

```
>>> model.output_shape
>>> model.summary()
>>> model.get_config()
>>> model.get_weights()
```

Model output shape	Model output shape
Model summary	Model summary representation
Model configuration	List all weight tensors in the model

## Compile Model

### MLP: Binary Classification

```
>>> model.compile(optimizer='adam',
    loss='binary_crossentropy',
    metrics=['accuracy'])
```

### MLP: Multi-Class Classification

```
>>> model.compile(optimizer='rmsprop',
    loss='categorical_crossentropy',
    metrics=['accuracy'])
```

### MLP: Regression

```
>>> model.compile(optimizer='rmsprop',
    loss='mse',
    metrics=['mae'])
```

### Recurrent Neural Network

```
>>> model3.compile(loss='binary_crossentropy',
    optimizer='adam',
    metrics=['accuracy'])
```

## Model Training

```
>>> model3.fit(x_train4,
    y_train4,
    batch_size=32,
    epochs=15,
    verbose=1,
    validation_data=(x_test4,y_test4))
```

## Evaluate Your Model's Performance

```
>>> score = model3.evaluate(x_test,
    y_test,
    batch_size=32)
```

## Prediction

```
>>> model3.predict(x_test4, batch_size=32)
>>> model3.predict_classes(x_test4, batch_size=32)
```

## Save/ Reload Models

```
>>> from keras.models import load_model
>>> model3.save('model_file.h5')
>>> my_model = load_model('my_model.h5')
```

## Model Fine-tuning

### Optimization Parameters

```
>>> from keras.optimizers import RMSprop
>>> opt = RMSprop(lr=0.0001, decay=1e-6)
>>> model2.compile(loss='categorical_crossentropy',
    optimizer=opt,
    metrics=['accuracy'])
```

## Early Stopping

```
>>> from keras.callbacks import EarlyStopping
>>> early_stopping_monitor = EarlyStopping(patience=2)
>>> model3.fit(x_train4,
    y_train4,
    batch_size=32,
    epochs=15,
    validation_data=(x_test4,y_test4),
    callbacks=[early_stopping_monitor])
```

```
>>> Y_train3 = to_categorical(y_train3, num_classes)
>>> Y_test3 = to_categorical(y_test3, num_classes)
```

```
>>> scaler = StandardScaler()
>>> standardized_X = scaler.transform(x_train2)
>>> standardized_X_test = scaler.transform(x_test2)
```



## Keras Cheat Sheet

## Numpy

NumPy targets the [CPython](#) reference [implementation](#) of Python, which is a non-optimizing [bytecode](#) interpreter. Mathematical algorithms written for this version of Python often run much slower than [compiled](#) equivalents. NumPy address the slowness problem partly by providing multidimensional arrays and functions and operators that operate efficiently on arrays, requiring rewriting some code, mostly inner loops using NumPy.

# Python For Data Science Cheat Sheet

## NumPy Basics

Learn Python for Data Science [Interactively](#) at [www.DataCamp.com](http://www.DataCamp.com)



### NumPy

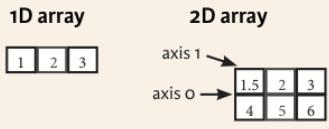
The NumPy library is the core library for scientific computing in Python. It provides a high-performance multidimensional array object, and tools for working with these arrays.

Use the following import convention:

```
>>> import numpy as np
```



### NumPy Arrays



### Creating Arrays

```
>>> a = np.array([1,2,3])
>>> b = np.array([(1.5,2,3), (4,5,6)], dtype = float)
>>> c = np.array([(1.5,2,3), (4,5,6)], [(3,2,1), (4,5,6)]),
      dtype = float)
```

### Initial Placeholders

```
>>> np.zeros((3,4))
>>> np.ones((2,3,4),dtype=np.int16)
>>> d = np.arange(10,25,5)

>>> np.linspace(0,2,9)

>>> e = np.full((2,2),7)
>>> f = np.eye(2)
>>> np.random.random((2,2))
>>> np.empty((3,2))
```

Create an array of zeros  
Create an array of ones  
Create an array of evenly spaced values (step value)  
Create an array of evenly spaced values (number of samples)  
Create a constant array  
Create a 2X2 identity matrix  
Create an array with random values  
Create an empty array

### I/O

#### Saving & Loading On Disk

```
>>> np.save('my_array', a)
>>> np.savetxt('array.npz', a, b)
>>> np.load('my_array.npy')
```

#### Saving & Loading Text Files

```
>>> np.loadtxt("myfile.txt")
>>> np.genfromtxt("my_file.csv", delimiter=',')
>>> np.savetxt("myarray.txt", a, delimiter=" ")
```

### Data Types

```
>>> np.int64
>>> np.float32
>>> np.complex
>>> np.bool_
>>> np.object
```

Signed 64-bit integer types  
Standard double-precision floating point  
Complex numbers represented by 128 floats  
Boolean type storing TRUE and FALSE values  
Python object type

### Inspecting Your Array

```
>>> a.shape
>>> len(a)
>>> b.ndim
>>> e.size
>>> b.dtype
>>> b.dtype.name
>>> b.astype(int)
```

Array dimensions  
Length of array  
Number of array dimensions  
Number of array elements  
Data type of array elements  
Name of data type  
Convert an array to a different type

### Asking For Help

```
>>> np.info(np.ndarray.dtype)
```

### Array Mathematics

#### Arithmetic Operations

```
>>> g = a - b
array([[-0.5,  0.,  0.],
       [-3., -3., -3.]])
>>> np.subtract(a,b)
>>> b + a
array([[ 2.5,  4.,  6.],
       [ 5.,  7.,  9.]])
>>> np.add(b,a)
>>> a / b
array([[ 0.66666667,  1.,
        [ 0.25,  0.4,  1. ],
        [ 0.5,  0.6,  0.5 ]])
>>> np.divide(a,b)
>>> a * b
array([[ 1.5,  4.,  9.],
       [ 4., 10., 18.]])
>>> np.multiply(a,b)
>>> np.exp(b)
>>> np.sqrt(b)
>>> np.sin(a)
>>> np.cos(b)
>>> np.log(a)
>>> e.dot(f)
array([[ 7.,  7.],
       [ 7.,  7.]])
```

#### Subtraction

Subtraction  
Addition

Addition  
Division

Division  
Multiplication

Multiplication  
Exponentiation  
Square root  
Print sines of an array  
Element-wise cosine  
Element-wise natural logarithm  
Dot product

### Comparison

```
>>> a == b
array([[False,  True,  True],
       [False, False, False]], dtype=bool)
>>> a < 2
array([True, False, False], dtype=bool)
>>> np.array_equal(a, b)
```

Element-wise comparison  
Element-wise comparison  
Array-wise comparison

### Aggregate Functions

```
>>> a.sum()
>>> a.min()
>>> b.max(axis=0)
>>> b.cumsum(axis=1)
>>> a.mean()
>>> b.median()
>>> a.corrcoef()
>>> np.std(b)
```

Array-wise sum  
Array-wise minimum value  
Maximum value of an array row  
Cumulative sum of the elements  
Mean  
Median  
Correlation coefficient  
Standard deviation

### Copying Arrays

```
>>> h = a.view()
>>> np.copy(a)
>>> h = a.copy()
```

Create a view of the array with the same data  
Create a copy of the array  
Create a deep copy of the array

### Sorting Arrays

### Subsetting, Slicing, Indexing

#### Subsetting

```
>>> a[2]
3
>>> b[1,2]
6.0
```

1	2	3
1.5	2	3
4	5	6

#### Slicing

```
>>> a[0:2]
array([1, 2])
>>> b[0:2,1]
array([ 2.,  5.])
```

1	2	3
1.5	2	3
4	5	6

```
>>> b[[1]]
array([1.5, 2., 3.])
>>> c[1,...]
array([[[ 3.,  2.,  1.],
        [ 4.,  5.,  6.]]])
```

```
>>> a[:, :-1]
array([3, 2, 1])
```

#### Boolean Indexing

```
>>> a[a<2]
array([1])
```

#### Fancy Indexing

```
>>> b[[1, 0, 1, 0], [0, 1, 2, 0]]
array([ 4.,  2.,  6., 1.5])
```

```
>>> b[[1, 0, 1, 0], [0, 1, 2, 0]]
array([[[ 4.,  5.,  6.,  4.],
        [ 1.5,  2.,  3.,  1.5],
        [ 4.,  5.,  6.,  4.],
        [ 1.5,  2.,  3.,  1.5]]])
```

### Also see Lists

Select the element at the 2nd index

Select the element at row 0 column 2 (equivalent to `b[1][2]`)

Select items at index 0 and 1

Select items at rows 0 and 1 in column 1

Select all items at row 0 (equivalent to `b[0:1, :]`)  
Same as `[1, :, :]`

Reversed array `a`

Select elements from `a` less than 2

Select elements `(1,0), (0,1), (1,2)` and `(0,0)`

Select a subset of the matrix's rows and columns

### Array Manipulation

#### Transposing Array

```
>>> i = np.transpose(b)
>>> i.T
```

Permute array dimensions  
Permute array dimensions

#### Changing Array Shape

```
>>> b.ravel()
>>> g.reshape(3,-2)
```

Flatten the array  
Reshape, but don't change data

#### Adding/Removing Elements

```
>>> h.resize((2,6))
>>> np.append(h,g)
>>> np.insert(a, 1, 5)
>>> np.delete(a, [1])
```

Return a new array with shape `(2,6)`  
Append items to an array  
Insert items in an array  
Delete items from an array

#### Combining Arrays

```
>>> np.concatenate((a,d),axis=0)
array([ 1,  2,  3, 10, 15, 20])
>>> np.vstack((a,b))
array([[[ 1.,  2.,  3.],
        [ 1.5,  2.,  3.],
        [ 4.,  5.,  6.]]])
>>> np.r_[e,f]
>>> np.hstack((e,f))
array([[ 7.,  7.,  1.,  0.],
       [ 7.,  7.,  0.,  1.]])
>>> np.column_stack((a,d))
array([[ 1,  10],
       [ 2,  15],
       [ 3,  20]])
>>> np.c_[a,d]
```

Concatenate arrays

Stack arrays vertically (row-wise)

Stack arrays vertically (row-wise)  
Stack arrays horizontally (column-wise)

Create stacked column-wise arrays

Create stacked column-wise arrays

#### Splitting Arrays

```
>>> np.hsplit(a,3)
[array([1],array([2]),array([3]))]
>>> np.vsplit(c,2)
[array([[ 1.5,  2.,  3.],
        [ 4.,  5.,  6.]]),
 array([[ 3.,  2.,  1.],
        [ 4.,  5.,  6.]])]
```

Split the array horizontally at the 3rd index  
Split the array vertically at the 2nd index

```
>>> np.string_
>>> np.unicode_
```

Fixed-length string type  
Fixed-length unicode type

```
>>> a.sort()
>>> c.sort(axis=0)
```

Sort an array  
Sort the elements of an array's axis



## Numpy Cheat Sheet

# Pandas

The name ‘Pandas’ is derived from the term “panel data”, an econometrics term for multidimensional structured data sets.

# Python For Data Science Cheat Sheet

## Pandas Basics

Learn Python for Data Science [Interactively](#) at [www.DataCamp.com](http://www.DataCamp.com)



### Pandas

The Pandas library is built on NumPy and provides easy-to-use data structures and data analysis tools for the Python programming language.



Use the following import convention:

```
>>> import pandas as pd
```

### Pandas Data Structures

#### Series

A one-dimensional labeled array capable of holding any data type

	A	3
Index	B	-5
	C	7
	D	4

```
>>> s = pd.Series([3, -5, 7, 4], index=['a', 'b', 'c', 'd'])
```

#### DataFrame

##### Columns

	Country	Capital	Population
1	Belgium	Brussels	11190846
2	India	New Delhi	1303171035
3	Brazil	Brasilia	207847528

```
>>> data = {'Country': ['Belgium', 'India', 'Brazil'],
   'Capital': ['Brussels', 'New Delhi', 'Brasilia'],
   'Population': [11190846, 1303171035, 207847528]}

>>> df = pd.DataFrame(data,
   columns=['Country', 'Capital', 'Population'])
```

### I/O

#### Read and Write to CSV

```
>>> pd.read_csv('file.csv', header=None, nrows=5)
>>> pd.to_csv('myDataFrame.csv')
```

#### Read and Write to Excel

```
>>> pd.read_excel('file.xlsx')
>>> pd.to_excel('dir/myDataFrame.xlsx', sheet_name='Sheet1')
  Read multiple sheets from the same file
>>> xlsx = pd.ExcelFile('file.xls')
```

### Asking For Help

```
>>> help(pd.Series.loc)
```

### Selection

Also see NumPy Arrays

#### Getting

```
>>> s['b']
-5
>>> df[1:]
   Country    Capital  Population
1  India     New Delhi      1303171035
2  Brazil    Brasilia      207847528
```

Get one element

Get subset of a DataFrame

### Selecting, Boolean Indexing & Setting

#### By Position

```
>>> df.iloc([0], [0])
'Belgium'
>>> df.iat([0], [0])
'Belgium'
```

Select single value by row & column

#### By Label

```
>>> df.loc([0], ['Country'])
'Belgium'
>>> df.at([0], ['Country'])
'Belgium'
```

Select single value by row & column labels

#### By Label/Position

```
>>> df.ix[2]
   Country    Brazil
   Capital  Brasilia
   Population  207847528
>>> df.ix[:, 'Capital']
0    Brussels
1   New Delhi
2   Brasilia
```

Select single row of subset of rows

```
>>> df.ix[1, 'Capital']
'New Delhi'
```

Select a single column of subset of columns

```
>>> df.ix[1, 'Capital']
'New Delhi'
```

Select rows and columns

**Boolean Indexing**

```
>>> s[~(s > 1)]
>>> s[(s < -1) | (s > 2)]
>>> df[df['Population']>1200000000]
```

Series s where value is not > 1  
s where value is <-1 or > 2  
Use filter to adjust DataFrame

#### Setting

```
>>> s['a'] = 6
```

Set index a of Series s to 6

### Dropping

```
>>> s.drop(['a', 'c'])
>>> df.drop('Country', axis=1)
```

Drop values from rows (axis=0)  
Drop values from columns (axis=1)

### Sort & Rank

```
>>> df.sort_index(by='Country')
>>> s.order()
>>> df.rank()
```

Sort by row or column index  
Sort a series by its values  
Assign ranks to entries

### Retrieving Series/DataFrame Information

#### Basic Information

```
>>> df.shape
>>> df.index
>>> df.columns
>>> df.info()
>>> df.count()
```

(rows,columns)  
Describe index  
Describe DataFrame columns  
Info on DataFrame  
Number of non-NA values

#### Summary

```
>>> df.sum()
>>> df.cumsum()
>>> df.min() / df.max()
>>> df.idmin() / df.idmax()
>>> df.describe()
>>> df.mean()
>>> df.median()
```

Sum of values  
Cumulative sum of values  
Minimum/maximum values  
Minimum/Maximum index value  
Summary statistics  
Mean of values  
Median of values

### Applying Functions

```
>>> f = lambda x: x*2
>>> df.apply(f)
>>> df.applymap(f)
```

Apply function  
Apply function element-wise

### Data Alignment

#### Internal Data Alignment

NA values are introduced in the indices that don't overlap:

```
>>> s3 = pd.Series([7, -2, 3], index=['a', 'c', 'd'])
>>> s + s3
a    10.0
b    NaN
c     5.0
d     7.0
```

### Arithmetic Operations with Fill Methods

You can also do the internal data alignment yourself with the help of the fill methods:

```
>>> s.add(s3, fill_value=0)
a    10.0
b    -5.0
c     5.0
d     7.0
>>> s.sub(s3, fill_value=2)
>>> s.div(s3, fill_value=4)
>>> s.mul(s3, fill_value=3)
```

```
>>> df = pd.read_excel(xlsx, 'Sheet1')
```

```
>>> pd.to_sql('myDf', engine)
```



Pandas Cheat Sheet

## Data Wrangling

The term “data wrangler” is starting to infiltrate pop culture. In the 2017 movie Kong: Skull Island, one of the characters, played by actor Marc Evan Jackson is introduced as “Steve Woodward, our data wrangler”.

# Data Wrangling with pandas Cheat Sheet

<http://pandas.pydata.org>

## Syntax – Creating DataFrames

	a	b	c
1	4	7	10
2	5	8	11
3	6	9	12

```
df = pd.DataFrame(
    {"a": [4, 5, 6],
     "b": [7, 8, 9],
     "c": [10, 11, 12]},
    index = [1, 2, 3])
Specify values for each column.
```

```
df = pd.DataFrame(
    [[4, 7, 10],
     [5, 8, 11],
     [6, 9, 12]],
    index=[1, 2, 3],
    columns=['a', 'b', 'c'])
Specify values for each row.
```

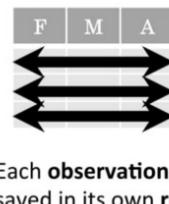
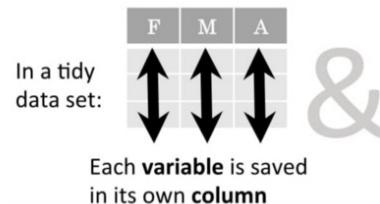
	a	b	c
n	v		
d	1	4	7
e	2	5	11

```
df = pd.DataFrame(
    {"a": [4, 5, 6],
     "b": [7, 8, 9],
     "c": [10, 11, 12]},
    index = pd.MultiIndex.from_tuples(
        [('d',1),('d',2),('e',2)],
        names=['n', 'v']))
Create DataFrame with a MultiIndex
```

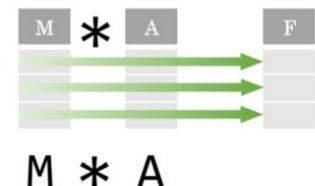
## Method Chaining

Most pandas methods return a DataFrame so that another pandas method can be applied to the

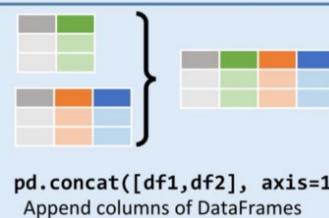
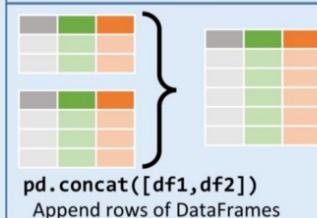
## Tidy Data – A foundation for wrangling in pandas



Tidy data complements pandas's **vectorized operations**. pandas will automatically preserve observations as you manipulate variables. No other format works as intuitively with pandas.



## Reshaping Data – Change the layout of a data set



```
df.sort_values('mpg')
Order rows by values of a column (low to high).
```

```
df.sort_values('mpg', ascending=False)
Order rows by values of a column (high to low).
```

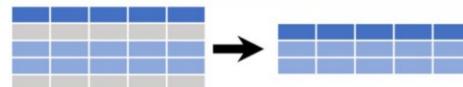
```
df.rename(columns = {'y': 'year'})
Rename the columns of a DataFrame
```

```
df.sort_index()
Sort the index of a DataFrame
```

```
df.reset_index()
Reset index of DataFrame to row numbers, moving index to columns.
```

```
df.drop(['Length', 'Height'], axis=1)
Drop columns from DataFrame
```

## Subset Observations (Rows)



```
df[df.Length > 7]
Extract rows that meet logical criteria.
df.drop_duplicates()
Remove duplicate rows (only considers columns).
df.head(n)
Select first n rows.
df.tail(n)
Select last n rows.
```

```
df.sample(frac=0.5)
Randomly select fraction of rows.
df.sample(n=10)
Randomly select n rows.
df.iloc[10:20]
Select rows by position.
df.nlargest(n, 'value')
Select and order top n entries.
df.nsmallest(n, 'value')
Select and order bottom n entries.
```

## Subset Variables (Columns)



```
df[['width', 'length', 'species']]
Select multiple columns with specific names.
df['width'] or df.width
Select single column with specific name.
df.filter(regex='regex')
Select columns whose name matches regular expression regex.
```

### regex (Regular Expressions) Examples

'.'	Matches strings containing a period '.'
'Length\$'	Matches strings ending with word 'Length'
'^Sepal'	Matches strings beginning with the word 'Sepal'
'^x[1-5]\$'	Matches strings beginning with 'x' and ending with 1,2,3,4,5
'^(?!Species\$).*'	Matches strings except the string 'Species'

## Logic in Python (and pandas)

result. This improves readability of code.

```
df = (pd.melt(df)
      .rename(columns={
          'variable' : 'var',
          'value' : 'val'})
      .query('val >= 200')
)
```

<	Less than	<code>!=</code>	Not equal to
>	Greater than	<code>df.column.isin(values)</code>	Group membership
<code>==</code>	Equals	<code>pd.isnull(obj)</code>	Is NaN
<code>&lt;=</code>	Less than or equals	<code>pd.notnull(obj)</code>	Is not NaN
<code>&gt;=</code>	Greater than or equals	<code>&amp;,  , ~, ^, df.any(), df.all()</code>	Logical and, or, not, xor, any, all

<http://pandas.pydata.org/> This cheat sheet inspired by Rstudio Data Wrangling Cheatsheet (<https://www.rstudio.com/wp-content/uploads/2015/02/data-wrangling-cheatsheet.pdf>) Written by Irv Lustig, Princeton Consultants

`df.loc[:, 'x2':'x4']`

Select all columns between x2 and x4 (inclusive).

`df.iloc[:, [1,2,5]]`

Select columns in positions 1, 2 and 5 (first column is 0).

`df.loc[df['a'] > 10, ['a', 'c']]`

Select rows meeting logical condition, and only the specific columns .

### Data Wrangling Cheat Sheet

## Summarize Data

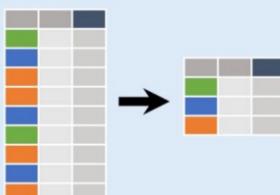
```
df['w'].value_counts()
Count number of rows with each unique value of variable
len(df)
# of rows in DataFrame.
df['w'].nunique()
# of distinct values in a column.
df.describe()
Basic descriptive statistics for each column (or GroupBy)
```



pandas provides a large set of **summary functions** that operate on different kinds of pandas objects (DataFrame columns, Series, GroupBy, Expanding and Rolling (see below)) and produce single values for each of the groups. When applied to a DataFrame, the result is returned as a pandas Series for each column. Examples:

<b>sum()</b>	<b>min()</b>
Sum values of each object.	Minimum value in each object.
<b>count()</b>	<b>max()</b>
Count non-NA/null values of each object.	Maximum value in each object.
<b>median()</b>	<b>mean()</b>
Median value of each object.	Mean value of each object.
<b>quantile([0.25, 0.75])</b>	<b>var()</b>
Quantiles of each object.	Variance of each object.
<b>apply(function)</b>	<b>std()</b>
Apply function to each object.	Standard deviation of each object.

## Group Data



```
df.groupby(by="col")
Return a GroupBy object, grouped by values in column named "col".
df.groupby(level="ind")
Return a GroupBy object, grouped by values in index level named "ind".
```

All of the summary functions listed above can be applied to a group. Additional GroupBy functions:

<b>size()</b>	<b>agg(function)</b>
Size of each group.	Aggregate group using function.

## Windows

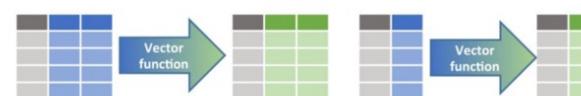
## Handling Missing Data

```
df.dropna()
Drop rows with any column having NA/null data.
df.fillna(value)
Replace all NA/null data with value.
```

## Make New Columns



```
df.assign(Area=lambda df: df.Length*df.Height)
Compute and append one or more new columns.
df['Volume'] = df.Length*df.Height*df.Depth
Add single column.
pd.qcut(df.col, n, labels=False)
Bin column into n buckets.
```



pandas provides a large set of **vector functions** that operate on all columns of a DataFrame or a single selected column (a pandas Series). These functions produce vectors of values for each of the columns, or a single Series for the individual Series. Examples:

<b>max(axis=1)</b>	<b>min(axis=1)</b>
Element-wise max.	Element-wise min.
<b>clip(lower=-10,upper=10)</b>	<b>abs()</b>
Trim values at input thresholds	Absolute value.

The examples below can also be applied to groups. In this case, the function is applied on a per-group basis, and the returned vectors are of the length of the original DataFrame.

<b>shift(1)</b>	<b>shift(-1)</b>
Copy with values shifted by 1.	Copy with values lagged by 1.
<b>rank(method='dense')</b>	<b>cumsum()</b>
Ranks with no gaps.	Cumulative sum.
<b>rank(method='min')</b>	<b>cummax()</b>
Ranks. Ties get min rank.	Cumulative max.
<b>rank(pct=True)</b>	<b>cummin()</b>
Ranks rescaled to interval [0, 1].	Cumulative min.
<b>rank(method='first')</b>	<b>cumprod()</b>
Ranks. Ties go to first value.	Cumulative product.

## Plotting

## Combine Data Sets

adf	bdf
x1   x2	x1   x3
A 1	A T
B 2	B F
C 3	D T



### Standard Joins

x1	x2	x3
A 1	T	
B 2	F	
C 3	NaN	

```
pd.merge(adf, bdf,
        how='left', on='x1')
Join matching rows from bdf to adf.
```

x1	x2	x3
A 1.0	T	
B 2.0	F	
D NaN	T	

```
pd.merge(adf, bdf,
        how='right', on='x1')
Join matching rows from adf to bdf.
```

x1	x2	x3
A 1	T	
B 2	F	

```
pd.merge(adf, bdf,
        how='inner', on='x1')
Join data. Retain only rows in both sets.
```

x1	x2	x3
A 1	T	
B 2	F	
C 3	NaN	
D NaN	T	

```
pd.merge(adf, bdf,
        how='outer', on='x1')
Join data. Retain all values, all rows.
```

### Filtering Joins

x1	x2
A 1	
B 2	

```
adf[adf.x1.isin(bdf.x1)]
All rows in adf that have a match in bdf.
```

x1	x2
C 3	

```
adf[~adf.x1.isin(bdf.x1)]
All rows in adf that do not have a match in bdf.
```

ydf	zdf
x1   x2	x1   x2
A 1	B 2
B 2	C 3
C 3	D 4



### Set-like Operations

x1	x2
B 2	
C 3	

```
pd.merge(ydf, zdf)
Rows that appear in both ydf and zdf (Intersection).
```

x1	x2
A 1	
B 2	

```
pd.merge(ydf, zdf, how='outer')
Rows that appear in either or both ydf and zdf (Union).
```

**df.expanding()**

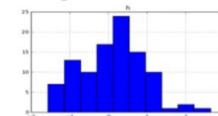
Return an Expanding object allowing summary functions to be applied cumulatively.

**df.rolling(n)**

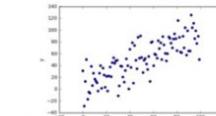
Return a Rolling object allowing summary functions to be applied to windows of length n.

**df.plot.hist()**

Histogram for each column

**df.plot.scatter(x='w',y='h')**

Scatter chart using pairs of points



C	3
D	4
x1	x2
A	1

```
pd.merge(ydf, zdf, how='outer',
         indicator=True)
.query('_merge == "left_only"')
.drop(['_merge'],axis=1)
Rows that appear in ydf but not zdf (Setdiff).
```

<http://pandas.pydata.org/> This cheat sheet inspired by Rstudio Data Wrangling Cheatsheet (<https://www.rstudio.com/wp-content/uploads/2015/02/data-wrangling-cheatsheet.pdf>) Written by Irv Lustig, Princeton Consultants

Pandas Data Wrangling Cheat Sheet

Ultimate Guide to Leveraging NLP & Machine Learning for your Chatbot

Code Snippets and Github Included

chatbotslife.com



## Data Wrangling with dplyr and tidyr

# Data Wrangling with dplyr and tidyverse

Cheat Sheet



## Syntax - Helpful conventions for wrangling

### dplyr::tbl\_df(iris)

Converts data to tbl class. tbl's are easier to examine than data frames. R displays only the data that fits onscreen:

```
Source: local data frame [150 x 5]
      Sepal.Length Sepal.Width Petal.Length
1           5.1       3.5          1.4
2           4.9       3.0          1.4
3           4.7       3.2          1.3
4           4.6       3.1          1.5
5           5.0       3.6          1.4
...
Variables not shown: Petal.Width (dbl), Species (fctr)
```

### dplyr::glimpse(iris)

Information dense summary of tbl data.

### utils::View(iris)

View data set in spreadsheet-like display (note capital V).

	iris				
	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa

### dplyr::%>%

Passes object on left hand side as first argument (or . argument) of function on righthand side.

`x %>% f(y)` is the same as `f(x, y)`

`y %>% f(x, ., z)` is the same as `f(x, y, z)`

"Piping" with %>% makes code more readable, e.g.

## Tidy Data - A foundation for wrangling in R

In a tidy data set:



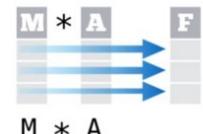
Each **variable** is saved in its own **column**

&

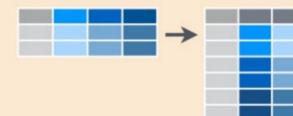


Each **observation** is saved in its own **row**

Tidy data complements R's **vectorized operations**. R will automatically preserve observations as you manipulate variables. No other format works as intuitively with R.



## Reshaping Data - Change the layout of a data set



### tidy::gather(cases, "year", "n", 2:4)

Gather columns into rows.

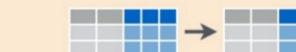


### tidy::spread(pollution, size, amount)

Spread rows into columns.

### tidy::separate(storms, date, c("y", "m", "d"))

Separate one column into several.



### tidy::unite(data, col, ..., sep)

Unite several columns into one.

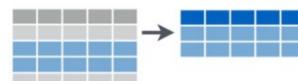
**dplyr::data\_frame(a = 1:3, b = 4:6)**  
Combine vectors into data frame (optimized).

**dplyr::arrange(mtcars, mpg)**  
Order rows by values of a column (low to high).

**dplyr::arrange(mtcars, desc(mpg))**  
Order rows by values of a column (high to low).

**dplyr::rename(tb, y = year)**  
Rename the columns of a data frame.

## Subset Observations (Rows)



### dplyr::filter(iris, Sepal.Length > 7)

Extract rows that meet logical criteria.

### dplyr::distinct(iris)

Remove duplicate rows.

### dplyr::sample\_frac(iris, 0.5, replace = TRUE)

Randomly select fraction of rows.

### dplyr::sample\_n(iris, 10, replace = TRUE)

Randomly select n rows.

### dplyr::slice(iris, 10:15)

Select rows by position.

### dplyr::top\_n(storms, 2, date)

Select and order top n entries (by group if grouped data).

## Subset Variables (Columns)



### dplyr::select(iris, Sepal.Width, Petal.Length, Species)

Select columns by name or helper function.

### Helper functions for select - ?select

#### select(iris, contains("."))

Select columns whose name contains a character string.

#### select(iris, ends\_with("Length"))

Select columns whose name ends with a character string.

#### select(iris, everything())

Select every column.

#### select(iris, matches(".t.))

Select columns whose name matches a regular expression.

#### select(iris, num\_range("x", 1:5))

Select columns named x1, x2, x3, x4, x5.

#### select(iris, one\_of(c("Species", "Genus")))

Select columns whose names are in a group of names.

#### select(iris, starts\_with("Sepal"))

Select columns whose name starts with "Sepal".

**TIPING** Using %>% makes code more readable, e.g.

```
iris %>%
  group_by(Species) %>%
  summarise(avg = mean(Sepal.Width)) %>%
  arrange(avg)
```

RStudio® is a trademark of RStudio, Inc. • [CC BY](#) RStudio • [info@rstudio.com](mailto:info@rstudio.com) • 844-448-1212 • [rstudio.com](http://rstudio.com)

	Less than	%in%	Not equal to
>	Greater than	is.na	Group membership
==	Equal to	is.NA	Is NA
<=	Less than or equal to	!is.na	Is not NA
>=	Greater than or equal to	&,  , !, xor, any, all	Boolean operators

[devtools::install\\_github\("rstudio/EDAWR"\)](#) for data sets

Learn more with [browseVignettes\(package = c\("dplyr", "tidyR"\)\)](#) • dplyr 0.4.0 • tidyR 0.2.0 • Updated: 1/15

Select columns whose name starts with a character string.

**select(iris, Sepal.Length:Petal.Width)**

Select all columns between Sepal.Length and Petal.Width (inclusive).

**select(iris, -Species)**

Select all columns except Species.

## Data Wrangling with dplyr and tidyR Cheat Sheet

## Summarise Data



`dplyr::summarise(iris, avg = mean(Sepal.Length))`

Summarise data into single row of values.

`dplyr::summarise_each(iris, funs(mean))`

Apply summary function to each column.

`dplyr::count(iris, Species, wt = Sepal.Length)`

Count number of rows with each unique value of variable (with or without weights).



Summarise uses **summary functions**, functions that take a vector of values and return a single value, such as:

`dplyr::first`

First value of a vector.

`dplyr::last`

Last value of a vector.

`dplyr::nth`

Nth value of a vector.

`dplyr::n`

# of values in a vector.

`dplyr::n_distinct`

# of distinct values in a vector.

`IQR`

IQR of a vector.

`min`

Minimum value in a vector.

`max`

Maximum value in a vector.

`mean`

Mean value of a vector.

`median`

Median value of a vector.

`var`

Variance of a vector.

`sd`

Standard deviation of a vector.

## Group Data

`dplyr::group_by(iris, Species)`

Group data into rows with the same value of Species.

`dplyr::ungroup(iris)`

Remove grouping information from data frame.

`iris %>% group_by(Species) %>% summarise(...)`

Compute separate summary row for each group.

## Make New Variables



`dplyr::mutate(iris, sepal = Sepal.Length + Sepal.Width)`

Compute and append one or more new columns.

`dplyr::mutate_each(iris, funs(min_rank))`

Apply window function to each column.

`dplyr::transmute(iris, sepal = Sepal.Length + Sepal.Width)`

Compute one or more new columns. Drop original columns.



Mutate uses **window functions**, functions that take a vector of values and return another vector of values, such as:

`dplyr::lead`

Copy with values shifted by 1.

`dplyr::lag`

Copy with values lagged by 1.

`dplyr::dense_rank`

Ranks with no gaps.

`dplyr::min_rank`

Ranks. Ties get min rank.

`dplyr::percent_rank`

Ranks rescaled to [0, 1].

`dplyr::row_number`

Ranks. Ties got to first value.

`dplyr::ntile`

Bin vector into n buckets.

`dplyr::between`

Are values between a and b?

`dplyr::cume_dist`

Cumulative distribution.

`dplyr::cumall`

Cumulative **all**

`dplyr::cumany`

Cumulative **any**

`dplyr::cummean`

Cumulative **mean**

`cumsum`

Cumulative **sum**

`cummax`

Cumulative **max**

`cummin`

Cumulative **min**

`cumprod`

Cumulative **prod**

`pmax`

Element-wise **max**

`pmin`

Element-wise **min**

`iris %>% group_by(Species) %>% mutate(...)`

Compute new variables by group.

## Combine Data Sets



### Mutating Joins

x1	x2	x3
A	1	T
B	2	F
C	3	NA

x1	x3	x2
A	T	1
B	F	2
D	T	NA

`dplyr::left_join(a, b, by = "x1")`

Join matching rows from b to a.

`dplyr::right_join(a, b, by = "x1")`

Join matching rows from a to b.

`dplyr::inner_join(a, b, by = "x1")`

Join data. Retain only rows in both sets.

`dplyr::full_join(a, b, by = "x1")`

Join data. Retain all values, all rows.

### Filtering Joins

x1	x2
A	1
B	2

x1	x2
C	3
D	T

`dplyr::semi_join(a, b, by = "x1")`

All rows in a that have a match in b.

`dplyr::anti_join(a, b, by = "x1")`

All rows in a that do not have a match in b.



### Set Operations

x1	x2
B	2
C	3

x1	x2
A	1
B	2

`dplyr::intersect(y, z)`

Rows that appear in both y and z.

`dplyr::union(y, z)`

Rows that appear in either or both y and z.

`dplyr::setdiff(y, z)`

Rows that appear in y but not z.

### Binding

x1	x2
A	1
B	2

x1	x2
C	3
B	2

`dplyr::bind_rows(y, z)`

Append z to y as new rows.



This image is a screenshot of a cheat sheet for data wrangling with dplyr and tidyr. It features a flowchart showing the transformation of data frames through various operations. Below the flowchart, there is copyright information for RStudio, a link to install the EDAWR package, and a section for the `dplyr::bind_cols(y, z)` function.

RStudio® is a trademark of RStudio, Inc. • [CC BY](#) RStudio • [info@rstudio.com](mailto:info@rstudio.com) • 844-448-1212 • [rstudio.com](http://rstudio.com)

`devtools::install_github("rstudio/EDAWR")` for data sets

Learn more with `browseVignettes(package = c("dplyr", "tidyverse"))` • dplyr 0.4.0 • tidyverse 0.2.0 • Updated: 1/15

Data Wrangling with dplyr and tidyr Cheat Sheet

**dplyr::bind\_cols(y, z)**  
Append z to y as new columns.  
Caution: matches rows by position.

## Scipy

SciPy builds on the [NumPy](#) array object and is part of the NumPy stack which includes tools like [Matplotlib](#), [pandas](#) and [SymPy](#), and an expanding set of scientific computing libraries. This NumPy stack has similar users to other applications such as [MATLAB](#), [GNU Octave](#), and [Scilab](#). The NumPy stack is also sometimes referred to as the SciPy stack.[\[3\]](#)

# Python For Data Science Cheat Sheet

## SciPy - Linear Algebra

Learn More Python for Data Science [Interactively](#) at [www.datacamp.com](http://www.datacamp.com)



### SciPy

The SciPy library is one of the core packages for scientific computing that provides mathematical algorithms and convenience functions built on the NumPy extension of Python.



### Interacting With NumPy

[Also see NumPy](#)

```
>>> import numpy as np
>>> a = np.array([1,2,3])
>>> b = np.array([(1+5j,2j,3j), (4j,5j,6j)])
>>> c = np.array([[1.5,2,3], (4,5,6)], [(3,2,1), (4,5,6)])
```

### Index Tricks

>>> np.mgrid[0:5,0:5]	Create a dense meshgrid
>>> np.ogrid[0:2,0:2]	Create an open meshgrid
>>> r_[:3,[0]*5,-1:1:10j]	Stack arrays vertically (row-wise)
>>> np.c_[b,c]	Create stacked column-wise arrays

### Shape Manipulation

>>> np.transpose(b)	Permute array dimensions
>>> b.flatten()	Flatten the array
>>> np.hstack((b,c))	Stack arrays horizontally (column-wise)
>>> np.vstack((a,b))	Stack arrays vertically (row-wise)
>>> np.hsplit(c,2)	Split the array horizontally at the 2nd index
>>> np.vsplit(d,2)	Split the array vertically at the 2nd index

### Polynomials

```
>>> from numpy import poly1d
>>> p = poly1d([3,4,5])
```

Create a polynomial object

### Vectorizing Functions

```
>>> def myfunc(a):
    if a < 0:
        return a**2
    else:
        return a/2
>>> np.vectorize(myfunc)
```

Vectorize functions

### Type Handling

```
>>> np.real(b)
>>> np.imag(b)
>>> np.real_if_close(c,tol=1000)
>>> np.cast['f'](np.pi)
```

Return the real part of the array elements  
Return the imaginary part of the array elements  
Return a real array if complex parts close to 0  
Cast object to a data type

### Other Useful Functions

```
>>> np.angle(b,deg=True)
>>> g = np.linspace(0,np.pi,num=5)
>>> g[3:] += np.pi
>>> np.unwrap(g)
>>> np.logspace(0,10,3)
>>> np.select([c<4],[c*2])
>>> misc.factorial(a)
>>> misc.comb(10,3,exact=True)
```

Return the angle of the complex argument  
Create an array of evenly spaced values (number of samples)  
Unwrap  
Create an array of evenly spaced values (log scale)  
Return values from a list of arrays depending on conditions  
Factorial  
Combine N things taken at k time

## Linear Algebra

You'll use the linalg and sparse modules. Note that `scipy.linalg` contains and expands on `numpy.linalg`.

[Also see NumPy](#)

### Creating Matrices

```
>>> A = np.matrix(np.random((2,2)))
>>> B = np.asmatrix(b)
>>> C = np.mat(np.random((10,5)))
>>> D = np.mat([[3,4], [5,6]])
```

### Basic Matrix Routines

#### Inverse

```
>>> A.I
>>> linalg.inv(A)
```

#### Transposition

```
>>> A.T
```

```
>>> A.H
```

#### Trace

```
>>> np.trace(A)
```

#### Norm

```
>>> linalg.norm(A)
>>> linalg.norm(A,1)
>>> linalg.norm(A,np.inf)
```

#### Rank

```
>>> np.linalg.matrix_rank(C)
```

#### Determinant

```
>>> linalg.det(A)
```

#### Solving linear problems

```
>>> linalg.solve(A,b)
>>> E = np.mat(a).T
>>> linalg.lstsq(F,E)
```

#### Generalized inverse

```
>>> linalg.pinv(C)
```

```
>>> linalg.pinv2(C)
```

### Creating Sparse Matrices

```
>>> F = np.eye(3, k=1)
>>> G = np.mat(np.identity(2))
>>> C[C > 0.5] = 0
>>> H = sparse.csr_matrix(C)
>>> I = sparse.csc_matrix(D)
>>> J = sparse.dok_matrix(A)
>>> E.todense()
>>> sparse.isspmatrix_csc(A)
```

Create a 2x2 identity matrix  
Create a 2x2 identity matrix

Compressed Sparse Row matrix  
Compressed Sparse Column matrix  
Dictionary Of Keys matrix  
Sparse matrix to full matrix  
Identify sparse matrix

### Sparse Matrix Routines

#### Inverse

```
>>> sparse.linalg.inv(I)
```

#### Norm

```
>>> sparse.linalg.norm(I)
```

#### Solving linear problems

```
>>> sparse.linalg.spsolve(H,I)
```

### Sparse Matrix Functions

```
>>> sparse.linalg.expm(I)
```

#### Inverse

#### Norm

Solver for sparse matrices

Sparse matrix exponential

### Asking For Help

## Matrix Functions

### Addition

```
>>> np.add(A,D)
```

### Subtraction

```
>>> np.subtract(A,D)
```

### Division

```
>>> np.divide(A,D)
```

### Multiplication

```
>>> A @ D
```

```
>>> np.multiply(D,A)
```

```
>>> np.dot(A,D)
```

```
>>> np.vdot(A,D)
```

```
>>> np.inner(A,D)
```

```
>>> np.outer(A,D)
```

```
>>> np.tensordot(A,D)
```

```
>>> np.kron(A,D)
```

### Exponential Functions

```
>>> linalg.expm(A)
```

```
>>> linalg.expm2(A)
```

```
>>> linalg.expm3(D)
```

### Logarithm Function

```
>>> linalg.logm(A)
```

### Trigonometric Functions

```
>>> linalg.sinm(D)
```

```
>>> linalg.cosm(D)
```

```
>>> linalg.tanm(A)
```

### Hyperbolic Trigonometric Functions

```
>>> linalg.sinhm(D)
```

```
>>> linalg.coshm(D)
```

```
>>> linalg.tanhm(A)
```

### Matrix Sign Function

```
>>> np.signm(A)
```

### Matrix Square Root

```
>>> linalg.sqrtm(A)
```

### Arbitrary Functions

```
>>> linalg.funm(A, lambda x: x*x)
```

### Decompositions

#### Eigenvalues and Eigenvectors

```
>>> la, v = linalg.eig(A)
```

```
>>> l1, l2 = la
```

```
>>> v[:,0]
```

```
>>> v[:,1]
```

```
>>> linalg.eigvals(A)
```

#### Singular Value Decomposition

```
>>> U,s,Vh = linalg.svd(B)
```

```
>>> M,N = B.shape
```

```
>>> Sig = linalg.diagsvd(s,M,N)
```

#### LU Decomposition

```
>>> P,L,U = linalg.lu(C)
```

### Sparse Matrix Decompositions

```
>>> la, v = sparse.linalg.eigs(F,1)
```

```
>>> sparse.linalg.svds(H, 2)
```

Solve ordinary or generalized eigenvalue problem for square matrix  
Unpack eigenvalues

First eigenvector

Second eigenvector

Unpack eigenvalues

Singular Value Decomposition (SVD)

Construct sigma matrix in SVD

LU Decomposition

Eigenvalues and eigenvectors SVD

```
>>> misc.central_diff_weights(3)  weights for n-point central derivative  
>>> misc.derivative(myfunc,1.0)  Find the n-th derivative of a function at a point
```

```
>>> help(scipy.linalg.diagsvd)  
>>> np.info(np.matrix)
```

DataCamp  
Learn Python for Data Science Interactively



## Scipy Cheat Sheet

# Matplotlib

**matplotlib** is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK+. There is also a procedural “pylab” interface based on a state machine (like OpenGL), designed to closely resemble that of MATLAB, though its use is discouraged.<sup>[2]</sup> SciPy makes use of matplotlib.

pyplot is a matplotlib module which provides a MATLAB-like interface. <sup>[6]</sup> matplotlib is designed to be as usable as MATLAB, with the ability to use Python, with the advantage that it is free.

# Python For Data Science Cheat Sheet

## Matplotlib

Learn Python [Interactively](#) at [www.DataCamp.com](#)



### Matplotlib

Matplotlib is a Python 2D plotting library which produces publication-quality figures in a variety of hardcopy formats and interactive environments across platforms.



### 1 Prepare The Data

[Also see Lists & NumPy](#)

#### 1D Data

```
>>> import numpy as np
>>> x = np.linspace(0, 10, 100)
>>> y = np.cos(x)
>>> z = np.sin(x)
```

#### 2D Data or Images

```
>>> data = 2 * np.random.random((10, 10))
>>> data2 = 3 * np.random.random((10, 10))
>>> Y, X = np.mgrid[-3:3:100j, -3:3:100j]
>>> U = -1 - X**2 + Y
>>> V = 1 + X - Y**2
>>> from matplotlib.cbook import get_sample_data
>>> img = np.load(get_sample_data('axes_grid/bivariate_normal.npy'))
```

### 2 Create Plot

```
>>> import matplotlib.pyplot as plt
```

#### Figure

```
>>> fig = plt.figure()
>>> fig2 = plt.figure(figsize=plt.figaspect(2.0))
```

#### Axes

All plotting is done with respect to an Axes. In most cases, a subplot will fit your needs. A subplot is an axes on a grid system.

```
>>> fig.add_axes()
>>> ax1 = fig.add_subplot(221) # row-col-num
>>> ax3 = fig.add_subplot(212)
>>> fig3, axes = plt.subplots(nrows=2, ncols=2)
>>> fig4, axes2 = plt.subplots(ncols=3)
```

### 3 Plotting Routines

#### 1D Data

```
>>> fig, ax = plt.subplots()
>>> lines = ax.plot(x,y)
>>> ax.scatter(x,y)
>>> axes[0,0].bar([1,2,3],[3,4,5])
>>> axes[1,0].barh([0.5,1.2,2.5],[0,1,2])
>>> axes[1,1].axhline(0.45)
>>> axes[0,1].axvline(0.65)
>>> ax.fill(x,y,color='blue')
>>> ax.fill_between(x,y,color='yellow')
```

#### 2D Data or Images

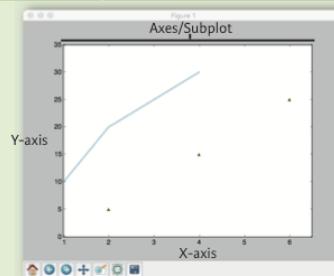
```
>>> fig, ax = plt.subplots()
>>> im = ax.imshow(img,
                  cmap='gist_earth',
                  interpolation='nearest')
```

Colormapped or RGB arrays

Draw points with lines or markers connecting them  
Draw unconnected points, scaled or colored  
Plot vertical rectangles (constant width)  
Plot horizontal rectangles (constant height)  
Draw a horizontal line across axes  
Draw a vertical line across axes  
Draw filled polygons  
Fill between y-values and 0

## Plot Anatomy & Workflow

### Plot Anatomy



Figure

### Workflow

The basic steps to creating plots with matplotlib are:

- 1 Prepare data
- 2 Create plot
- 3 Plot
- 4 Customize plot
- 5 Save plot
- 6 Show plot

```
>>> import matplotlib.pyplot as plt
>>> x = [1,2,3,4]                                Step 1
>>> y = [10,20,25,30]
>>> fig = plt.figure()                            Step 2
>>> ax = fig.add_subplot(111)                      Step 3
>>> ax.plot(x, y, color='lightblue', linewidth=3) Step 3, 4
>>> ax.scatter([2,4,6],
              [5,15,25],
              color='darkgreen',
              marker='^')                               Step 4
>>> ax.set_xlim(1, 6.5)
>>> plt.savefig('foo.png')
>>> plt.show()                                    Step 6
```

### 4 Customize Plot

#### Colors, Color Bars & Color Maps

```
>>> plt.plot(x, x, x**2, x, x**3)
>>> ax.plot(x, y, alpha = 0.4)
>>> ax.plot(x, y, c='k')
>>> fig.colorbar(im, orientation='horizontal')
>>> im = ax.imshow(img,
                  cmap='seismic')
```

#### Markers

```
>>> fig, ax = plt.subplots()
>>> ax.scatter(x,y,marker=".")                   Step 1
>>> ax.plot(x,y,marker="o")                      Step 2
```

#### LineStyles

```
>>> plt.plot(x,y,linewidth=4.0)
>>> plt.plot(x,y,ls='solid')
>>> plt.plot(x,y,ls='--')
>>> plt.plot(x,y,'--',x**2,y**2,'-.')
>>> plt.setp(lines,color='r',linewidth=4.0)
```

#### Text & Annotations

```
>>> ax.text(1, -2.1, 'Example Graph', style='italic')
>>> ax.annotate("Sine", xy=(8, 0), xycoords='data',
               xytext=(10.5, 0), textcoords='data',
               arrowprops=dict(arrowstyle="->", connectionstyle="arc3"),)
```

#### Vector Fields

```
>>> axes[0,1].arrow(0,0,0.5,0.5)
>>> axes[1,1].quiver(y,z)
>>> axes[0,1].streamplot(X,Y,U,V)
```

#### Mathtext

```
>>> plt.title(r'$\sigma_i=15$', fontsize=20)
```

#### Limits, Legends & Layouts

```
>>> ax.margins(x=0,y=0.1)
>>> ax.axis('equal')
>>> ax.set(xlim=[0,10.5], ylim=[-1.5,1.5])
>>> ax.set_xlims(0,10.5)
```

#### Legends

```
>>> ax.set(title='An Example Axes',
           ylabel='Y-Axis',
           xlabel='X-Axis')
>>> ax.legend(loc='best')
```

#### Ticks

```
>>> ax.xaxis.set(ticks=range(1,5),
                 ticklabels=[3,100,-12,"foo"])
>>> ax.tick_params(axis='y',
                  direction='inout',
                  length=10)
```

#### Subplot Spacing

```
>>> fig3.subplots_adjust(wspace=0.5,
                        hspace=0.3,
                        left=0.125,
                        right=0.9,
                        top=0.9,
                        bottom=0.1)
```

#### Axis Spines

```
>>> ax1.spines['top'].set_visible(False)
>>> ax1.spines['bottom'].set_position(('outward',10))
```

Add padding to a plot  
Set the aspect ratio of the plot to 1  
Set limits for x-and y-axis  
Set limits for x-axis

Set a title and x-and y-axis labels

No overlapping plot elements

Manually set x-ticks  
Make y-ticks longer and go in and out

Adjust the spacing between subplots

Fit subplot(s) in to the figure area

Make the top axis line for a plot invisible  
Move the bottom axis line outward

### 5 Save Plot

#### Save figures

```
>>> plt.savefig('foo.png')
```

#### Save transparent figures

```
>>> plt.savefig('foo.png', transparent=True)
```

### 6 Show Plot

```
>>> plt.show()
```

### Close & Clear

```
>>> plt.clf()
```

```
>>> plt.cla()
```

```
>>> plt.close()
```

Clear an axis

Clear the entire figure

Close a window

```
interpolation='nearest',  
vmin=-2,  
vmax=2)
```

```
>>> axes2[2].contourf(data1)  
>>> axes2[2]= ax.clabel(CS)
```

Plot filled contours  
Label a contour plot

DataCamp  
Learn Python for Data Science interactively



## Matplotlib Cheat Sheet

>>> *If you like this list, you can let me know [here](#).* <<<

# Data Visualization

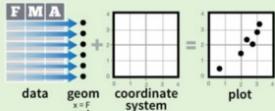
# Data Visualization with ggplot2

Cheat Sheet

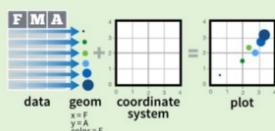


## Basics

**ggplot2** is based on the **grammar of graphics**, the idea that you can build every graph from the same few components: a **data set**, a set of **geoms**—visual marks that represent data points, and a **coordinate system**.



To display data values, map variables in the data set to aesthetic properties of the geom like **size**, **color**, and **x** and **y** locations.



Build a graph with **qplot()** or **ggplot()**

**aesthetic mappings**    **data**    **geom**  
`qplot(x = cty, y = hwy, color = cyl, data = mpg, geom = "point")`  
Creates a complete plot with given data, geom, and mappings. Supplies many useful defaults.

**ggplot(data = mpg, aes(x = cty, y = hwy))**

Begins a plot that you finish by adding layers to. No defaults, but provides more control than qplot().

**data**  
`ggplot(mpg, aes(hwy, cty)) +  
geom_point(aes(color = cyl)) +  
geom_smooth(method = "lm") +  
coord_cartesian() +  
scale_color_gradient() +  
theme_bw()`  
**add layers, elements with +**  
`layer = geom +  
default stat +  
layer specific  
mappings`  
**additional elements**  
`last_plot()`

Add a new layer to a plot with a **geom\_\***() or **stat\_\***() function. Each provides a geom, a set of aesthetic mappings, and a default stat and position adjustment.

**Geoms** - Use a geom to represent data points, use the geom's aesthetic properties to represent variables. Each function returns a layer.

### One Variable

#### Continuous

`a <- ggplot(mpg, aes(hwy))`



**a + geom\_area(stat = "bin")**

`x, y, alpha, color, fill, linetype, size`

**a + geom\_density(kernel = "gaussian")**

`x, y, alpha, color, fill, linetype, size, weight`

**a + geom\_dotplot()**

`x, y, alpha, color, fill`



**a + geom\_freqpoly()**

`x, y, alpha, color, linetype, size`

**b + geom\_freqpoly(aes(y = ..density..))**



**a + geom\_histogram(binwidth = 5)**

`x, y, alpha, color, fill, linetype, size, weight`

**b + geom\_histogram(aes(y = ..density..))**

#### Discrete

`b <- ggplot(mpg, aes(fl))`



**b + geom\_bar()**

`x, alpha, color, fill, linetype, size, weight`

### Graphical Primitives

`c <- ggplot(map, aes(long, lat))`



**c + geom\_polygon(aes(group = group))**

`x, y, alpha, color, fill, linetype, size`



`d <- ggplot(economics, aes(date, unemploy))`

**d + geom\_path(lineend = "butt",**

`linejoin = "round", linemitre = 1)`

`x, y, alpha, color, linetype, size`



**d + geom\_ribbon(aes(ymin = unemploy - 900,**

`ymax = unemploy + 900))`

`x, ymax, ymin, alpha, color, fill, linetype, size`



`e <- ggplot(seals, aes(x = long, y = lat))`

**e + geom\_segment(aes(**

`xend = long + delta_long,`

`yend = lat + delta_lat))`

`x, xend, y, yend, alpha, color, linetype, size`



**e + geom\_rect(aes(xmin = long, ymin = lat**

`))`

`x, xmin, ymin, alpha, color, fill, linetype, size`

### Two Variables

#### Continuous X, Continuous Y

`f <- ggplot(mpg, aes(cty, hwy))`



**f + geom\_blank()**

`x, y, alpha, color, fill, shape, size`



**f + geom\_jitter()**

`x, y, alpha, color, fill, shape, size`



**f + geom\_point()**

`x, y, alpha, color, fill, shape, size`



**f + geom\_quantile()**

`x, y, alpha, color, linetype, size, weight`



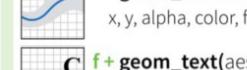
**f + geom\_rug(sides = "bl")**

`alpha, color, linetype, size`



**f + geom\_smooth(model = lm)**

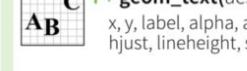
`x, y, alpha, color, fill, linetype, size, weight`



**f + geom\_text(aes(label = cty))**

`x, y, label, alpha, angle, color, family, fontface,`

`hjust, lineheight, size, vjust`



**f + geom\_step(direction = "hv")**

`x, y, alpha, color, linetype, size`

#### Discrete X, Continuous Y

`g <- ggplot(mpg, aes(class, hwy))`



**g + geom\_bar(stat = "identity")**

`x, y, alpha, color, fill, linetype, size`



**g + geom\_boxplot()**

`lower, middle, upper, x, ymax, ymin, alpha,`

`color, fill, linetype, shape, size, weight`



**g + geom\_dotplot(binaxis = "y",**

`stackdir = "center")`

`x, y, alpha, color, fill`



**g + geom\_violin(scale = "area")**

`x, y, alpha, color, fill, linetype, size`

#### Discrete X, Discrete Y

`h <- ggplot(diamonds, aes(cut, color))`



**h + geom\_jitter()**

`x, y, alpha, color, fill, shape, size`

### Three Variables



`seals$z <- with(seals, sqrt(delta_long^2 + delta_lat^2))`

`m + geom_raster(aes(fill = z), hjust = 0.5,`

#### Continuous Bivariate Distribution

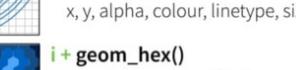
`i <- ggplot(movies, aes(year, rating))`



**i + geom\_hex(binwidth = c(5, 0.5))**

`xmax, xmin, ymax, ymin, alpha, color, fill,`

`linetype, size, weight`

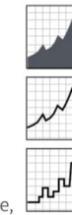


**i + geom\_hex()**

`x, y, alpha, colour, fill size`

#### Continuous Function

`j <- ggplot(economics, aes(date, unemploy))`



**j + geom\_area()**

`x, y, alpha, color, fill, linetype, size`



**j + geom\_line()**

`x, y, alpha, color, linetype, size`



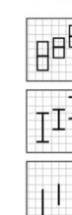
**j + geom\_step(direction = "hv")**

`x, y, alpha, color, linetype, size`

#### Visualizing error

`df <- data.frame(grp = c("A", "B"), fit = 4:5, se = 1:2)`

`k <- ggplot(df, aes(grp, fit, ymin = fit-se, ymax = fit+se))`



**k + geom\_crossbar(fatten = 2)**

`x, y, ymax, ymin, alpha, color, fill, linetype,`

`size`



**k + geom\_errorbar()**

`x, ymax, ymin, alpha, color, linetype, size,`

`width (also geom_errorbarh())`



**k + geom\_linerange()**

`x, ymin, ymax, alpha, color, linetype, size`



**k + geom\_pointrange()**

`x, y, ymin, ymax, alpha, color, fill, linetype,`

`shape, size`

#### Maps

`data <- data.frame(murder = USArests$Murder,`

`state = tolower(rownames(USArests)))`

`map <- map_data("state")`

`l <- ggplot(data, aes(fill = murder))`



**l + geom\_map(aes(map\_id = state), map = map) +**

`expand_limits(x = map$long, y = map$lat)`

`map_id, alpha, color, fill, linetype, size`

Returns the last plot

```
ggsave("plot.png", width = 5, height = 5)
```

Saves last plot as 5' x 5' file named "plot.png" in working directory. Matches file type to file extension.



```
ggsave("plot.png",  
       width = 5, height = 5,  
       xmax = long + delta_long,  
       ymax = lat + delta_lat)  
xmax, xmin, ymax, ymin, alpha, color, fill,  
linetype, size
```

```
m <- ggplot(seals, aes(long, lat))
```

```
m + geom_contour(aes(z = z))
```

x, y, z, alpha, colour, linetype, size, weight



```
vjust=0.5, interpolate=FALSE)  
x, y, alpha, fill
```

```
m + geom_tile(aes(fill = z))
```

x, y, alpha, color, fill, linetype, size

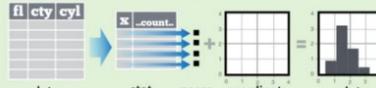
RStudio® is a trademark of RStudio, Inc. • [CC BY](#) RStudio • [info@rstudio.com](mailto:info@rstudio.com) • 844-448-1212 • [rstudio.com](http://rstudio.com)

Learn more at [docs.ggplot2.org](http://docs.ggplot2.org) • ggplot2 0.9.3.1 • Updated: 3/15

## Data Visualization Cheat Sheet

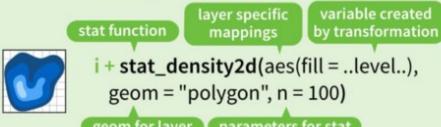
## Stats - An alternative way to build a layer

Some plots visualize a **transformation** of the original data set. Use a **stat** to choose a common transformation to visualize, e.g. `a + geom_bar(stat = "bin")`



Each stat creates additional variables to map aesthetics to. These variables use a common `..name..` syntax.

stat functions and geom functions both combine a stat with a geom to make a layer, i.e. `stat_bin(geom="bar")` does the same as `geom_bar(stat="bin")`



```
a + stat_bin(binwidth = 1, origin = 10)          1D distributions
x, y | ..count.,..ncount.,..density.,..ndensity..
a + stat_bindot(binwidth = 1, binaxis = "x")
x, y, | ..count.,..ncount..
a + stat_density(adjust = 1, kernel = "gaussian")
x, y, | ..count.,..density.,..scaled..
```

```
f + stat_bin2d(bins = 30, drop = TRUE)           2D distributions
x, y | ..count.,..density..
f + stat_bhex(bins = 30)
x, y, fill | ..count.,..density..
f + stat_density2d(contour = TRUE, n = 100)
x, y, color, size | ..level..
```

```
m + stat_contour(aes(z = z))                  3 Variables
x, y, z, order | ..level..
m + stat_spoke(aes(radius = z, angle = z))
angle, radius, x, xend, y, yend | ..x.,..xend.,..y.,..yend..
m + stat_summary_hex(aes(z = z), bins = 30, fun = mean)
x, y, z, fill | ..value..
m + stat_summary2d(aes(z = z), bins = 30, fun = mean)
x, y, z, fill | ..value..
```

```
g + stat_boxplot(coef = 1.5)                   Comparisons
x, y | ..lower.,..middle.,..upper.,..outliers..
g + stat_ydensity(adjust = 1, kernel = "gaussian", scale = "area")
x, y | ..density.,..scaled.,..count.,..n.,..violinwidth.,..width..
```

```
f + stat_ecdf(n = 40)                          Functions
x, y | ..x.,..y..
f + stat_quantile(quartiles = c(0.25, 0.5, 0.75), formula = y ~ log(x),
method = "rq")
x, y | ..quantile.,..x.,..y..
f + stat_smooth(method = "auto", formula = y ~ x, se = TRUE, n = 80,
fullrange = FALSE, level = 0.95)
x, y | ..se.,..x.,..y.,..ymin.,..ymax..
```

```
ggplot() + stat_function(es(x = -3),
fun = dnorm, n = 101, args = list(sd = 0.5))
x | ..y..
f + stat_identity()
ggplot() + stat_qq(aes(sample = 1:100), distribution = qt,
```

## Scales

Scales control how a plot maps data values to the visual values of an aesthetic. To change the mapping, add a custom scale.



### General Purpose scales

Use with any aesthetic: alpha, color, fill, linetype, shape, size

`scale_*_continuous()` - map cont' values to visual values  
`scale_*_discrete()` - map discrete values to visual values  
`scale_*_identity()` - use data values as visual values  
`scale_*_manual(values = c())` - map discrete values to manually chosen visual values

### X and Y location scales

Use with x or y aesthetics (x shown here)

`scale_x_date(labels = date_format("%m/%d"),
breaks = date_breaks("2 weeks"))` - treat x values as dates. See ?strptime for label formats.  
`scale_x_datetime()` - treat x values as date times. Use same arguments as `scale_x_date()`.  
`scale_x_log10()` - Plot x on log10 scale  
`scale_x_reverse()` - Reverse direction of x axis  
`scale_x_sqrt()` - Plot x on square root scale

### Color and fill scales

**Discrete**

```
n <- b + geom_bar(aes(fill = fl))
```

**Continuous**

```
o <- a + geom_dotplot(aes(fill = ..x..))
o + scale_fill_gradient(low = "red", high = "yellow")
o + scale_fill_gradient2(low = "red", high = "blue",
mid = "white", midpoint = 25)
o + scale_fill_gradientn(colours = terrain.colors(6))
Also: rainbow(), heat.colors(), topo.colors(), cm.colors(),
RColorBrewer::brewer.pal
```

### Shape scales

**Manual shape values**

0	6	12	18	24
1	7	13	19	25
2	8	14	20	*
3	9	15	21	
4	10	16	22	0

**Shape values shown in**

```
p <- f + geom_point(aes(shape = fl))
p + scale_shape(solid = FALSE)
p + scale_shape_manual(values = c(3:7))
```

## Coordinate Systems

`r + b + geom_bar()`

`r + coord_cartesian(xlim = c(0, 5))`  
`xlim, ylim`  
The default cartesian coordinate system

`r + coord_fixed(ratio = 1/2)`  
ratio, xlim, ylim  
Cartesian coordinates with fixed aspect ratio between x and y units

`r + coord_flip()`

xlim, ylim  
Flipped Cartesian coordinates

`r + coord_polar(theta = "x", direction = 1)`  
theta, start, direction  
Polar coordinates

`r + coord_trans(xtrans = "sqrt")`  
xtrans, ytrans, limx, limy  
Transformed cartesian coordinates. Set extras and strains to the name of a window function.

`z + coord_map(projection = "ortho",
orientation = c(41, -74, 0))`  
projection, orientation, xlim, ylim  
Map projections from the mapproj package (mercator (default), aequalarea, lagrange, etc.)

## Position Adjustments

Position adjustments determine how to arrange geoms that would otherwise occupy the same space.

`s <- ggplot(mpg, aes(fl, fill = drv))`

`s + geom_bar(position = "dodge")`  
Arrange elements side by side

`s + geom_bar(position = "fill")`  
Stack elements on top of one another, normalize height

`s + geom_bar(position = "stack")`  
Stack elements on top of one another

`f + geom_point(position = "jitter")`  
Add random noise to X and Y position of each element to avoid overplotting

Each position adjustment can be recast as a function with manual `width` and `height` arguments

`s + geom_bar(position = position_dodge(width = 1))`

## Themes

`r + theme_bw()`  
White background with grid lines

`r + theme_classic()`  
White background no gridlines

## Faceting

Facets divide a plot into subplots based on the values of one or more discrete variables.

`t <- ggplot(mpg, aes(cty, hwy)) + geom_point()`



Set `scales` to let axis limits vary across facets

`t + facet_grid(~ x, scales = "free")`  
x and y axis limits adjust to individual facets  
• `"free_x"` - x axis limits adjust  
• `"free_y"` - y axis limits adjust

Set `labeler` to adjust facet labels

`t + facet_grid(. ~ fl, labeler = label_both)`  
fl: c fl: d fl: e fl: p fl: r  
`t + facet_grid(. ~ fl, labeler = label_bquote(alpha ^ .(x)))`  
 $\alpha^c \alpha^d \alpha^e \alpha^p \alpha^r$   
`t + facet_grid(. ~ fl, labeler = label_parsed)`  
c d e p r

## Labels

`t + ggtitle("New Plot Title")`  
Add a main title above the plot  
`t + xlab("New X label")`  
Change the label on the X axis  
`t + ylab("New Y label")`  
Change the label on the Y axis  
`t + labs(title = " New title", x = "New x", y = "New y")`  
All of the above

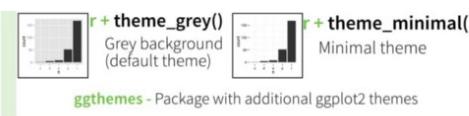
## Legends

`t + theme(legend.position = "bottom")`  
Place legend at "bottom", "top", "left", or "right"  
`t + guides(color = "none")`  
Set legend type for each aesthetic: colorbar, legend, or none (no legend)  
`t + scale_fill_discrete(name = "Title",
labels = c("A", "B", "C"))`  
Set legend title and labels with a scale function.

## Zooming

Without clipping (preferred)  
`t + coord_cartesian(xlim = c(0, 100), ylim = c(10, 20))`

```
dparams = list(df=5)
sample, x, y | ..x., ..y..
f + stat_sum()
x, y, size | ..size...
f + stat_summary(fun.data = "mean_cl_boot")
f + stat_unique()
```



**With clipping** (removes unseen data points)

`t + xlim(0, 100) + ylim(10, 20)`

`t + scale_x_continuous(limits = c(0, 100)) + scale_y_continuous(limits = c(0, 100))`

RStudio® is a trademark of RStudio, Inc. • [CC BY](#) RStudio • [info@rstudio.com](mailto:info@rstudio.com) • 844-448-1212 • [rstudio.com](http://rstudio.com)

Learn more at [docs.ggplot2.org](http://docs.ggplot2.org) • ggplot2 0.9.3.1 • Updated: 3/15

ggplot cheat sheet

# PySpark

# Python For Data Science Cheat Sheet

## PySpark Basics

Learn Python for data science interactively at [www.DataCamp.com](http://www.DataCamp.com)



### Spark

PySpark is the Spark Python API that exposes the Spark programming model to Python



### Initializing Spark

#### SparkContext

```
>>> from pyspark import SparkContext
>>> sc = SparkContext(master = 'local[2]')
```

#### Inspect SparkContext

>>> sc.version	Retrieve SparkContext version
>>> sc.pythonVer	Retrieve Python version
>>> sc.master	Master URL to connect to
>>> str(sc.sparkHome)	Path where Spark is installed on worker nodes
>>> str(sc.sparkUser())	Retrieve name of the Spark User running SparkContext
>>> sc.appName	Return application name
>>> sc.applicationId	Retrieve application ID
>>> sc.defaultParallelism	Return default level of parallelism
>>> sc.defaultMinPartitions	Default minimum number of partitions for RDDs

#### Configuration

```
>>> from pyspark import SparkConf, SparkContext
>>> conf = (SparkConf()
...     .setMaster("local")
...     .setAppName("My app")
...     .set("spark.executor.memory", "1g"))
>>> sc = SparkContext(conf = conf)
```

### Using The Shell

In the PySpark shell, a special interpreter-aware SparkContext is already created in the variable called `sc`.

```
$ ./bin/spark-shell --master local[2]
$ ./bin/pyspark --master local[4] --py-files code.py
```

Set which master the context connects to with the `--master` argument, and add Python .zip, .egg or .py files to the runtime path by passing a comma-separated list to `--py-files`.

### Loading Data

#### Parallelized Collections

```
>>> rdd = sc.parallelize([('a', 7), ('a', 2), ('b', 2)])
>>> rdd2 = sc.parallelize([('a', 2), ('d', 1), ('b', 1)])
>>> rdd3 = sc.parallelize(range(100))
>>> rdd4 = sc.parallelize([('a', ['x', 'y', 'z']), ('b', ['p', 'r'])])
```

#### External Data

Read either one text file from HDFS, a local file system or any Hadoop-supported file system URI with `textFile()`, or read in a directory of text files with `wholeTextFiles()`.

## Retrieving RDD Information

### Basic Information

```
>>> rdd.getNumPartitions()
>>> rdd.count()
3
>>> rdd.countByKey()
defaultdict(<type 'int'>, {'a':2,'b':1})
>>> rdd.countByValue()
defaultdict(<type 'int'>, {'b':2,1,('a',2):1,('a',7):1})
>>> rdd.collectAsMap()
{'a': 2, 'b': 2}
>>> rdd3.sum()
4950
>>> sc.parallelize([]).isEmpty()
True
```

List the number of partitions  
Count RDD instances  
Count RDD instances by key  
Count RDD instances by value  
Return (key,value) pairs as a dictionary  
Sum of RDD elements  
Check whether RDD is empty

### Summary

```
>>> rdd3.max()
99
>>> rdd3.min()
0
>>> rdd3.mean()
49.5
>>> rdd3.stdev()
28.8607004772218
>>> rdd3.variance()
833.25
>>> rdd3.histogram(3)
([0,33,66,99],[33,33,34])
>>> rdd3.stats()
```

Maximum value of RDD elements  
Minimum value of RDD elements  
Mean value of RDD elements  
Standard deviation of RDD elements  
Compute variance of RDD elements  
Compute histogram by bins  
Summary statistics (count, mean, stdev, max & min)

## Applying Functions

```
>>> rdd.map(lambda x: x+[x[1],x[0]])
...     .collect()
[('a',7,7,'a'),('a',2,2,'a'),('b',2,2,'b')]
>>> rdd5 = rdd.flatMap(lambda x: x+[x[1],x[0]])
>>> rdd5.collect()
['a',7,7,'a','a',2,2,'a','b',2,2,'b']
>>> rdd4.flatMapValues(lambda x: x)
...     .collect()
[('a','x'),('a','y'),('a','z'),('b','p'),('b','r')]
```

Apply a function to each RDD element  
Apply a function to each RDD element and flatten the result  
Apply a flatMap function to each (key,value) pair of rdd4 without changing the keys

## Selecting Data

### Getting

```
>>> rdd.collect()
[('a', 7), ('a', 2), ('b', 2)]
>>> rdd.take(2)
[('a', 7), ('a', 2)]
>>> rdd.first()
('a', 7)
>>> rdd.top(2)
[('b', 2), ('a', 7)]
```

### Sampling

```
>>> rdd3.sample(False, 0.15, 81).collect()
[3,4,27,31,40,41,42,43,60,76,79,80,86,97]
```

### Filtering

```
>>> rdd.filter(lambda x: "a" in x)
...     .collect()
[('a',7),('a',2)]
>>> rdd5.distinct().collect()
['a',2,'b',7]
>>> rdd.keys().collect()
['a', 'a', 'b']
```

## Iterating

```
>>> def g(x): print(x)
>>> rdd.foreach(g)
('a', 7)
```

Return a list with all RDD elements

Take first 2 RDD elements

Take first RDD element

Take top 2 RDD elements

Return sampled subset of rdd3

Filter the RDD

Return distinct RDD values

Return (key,value) RDD's keys

## Reshaping Data

### Reducing

```
>>> rdd.reduceByKey(lambda x,y : x+y)
...     .collect()
[('a',9),('b',2)]
>>> rdd.reduceByKey(lambda a, b: a + b)
('a',7,'a',2,'b',2)
```

Merge the rdd values for each key

Merge the rdd values

### Grouping by

```
>>> rdd3.groupByKey()
...     .mapValues(list)
...     .collect()
[('a',[(2,2)])
...  ,('b',[2,2])]
```

Return RDD of grouped values

Group rdd by key

### Aggregating

```
>>> seqOp = (lambda x,y: (x[0]+y,x[1]+1))
>>> combOp = (lambda x,y:(x[0]+y[0],x[1]+y[1]))
>>> rdd3.aggregate((0,0),seqOp,combOp)
(4950,100)
>>> rdd3.aggregateByKey((0,0),seqOp,combOp)
...     .collect()
[('a',(9,2)),('b',(2,1))]
>>> rdd3.fold(0,add)
4950
>>> rdd.foldByKey(0, add)
...     .collect()
[('a',9),('b',2)]
>>> rdd3.keyBy(lambda x: xx)
...     .collect()
```

Aggregate RDD elements of each partition and then the results

Aggregate values of each RDD key

Aggregate the elements of each partition, and then the results

Merge the values for each key

Create tuples of RDD elements by applying a function

## Mathematical Operations

```
>>> rdd.subtract(rdd2)
...     .collect()
[('b',2),('a',7)]
>>> rdd2.subtractByKey(rdd)
...     .collect()
[('d', 1)]
```

Return each rdd value not contained in rdd2

Return each (key,value) pair of rdd2 with no matching key in rdd

Return the Cartesian product of rdd and rdd2

## Sort

```
>>> rdd2.sortBy(lambda x: x[1])
...     .collect()
[('d',1),('b',1),('a',2)]
>>> rdd2.sortByKey()
...     .collect()
[('a',2),('b',1),('d',1)]
```

Sort RDD by given function

Sort (key,value) RDD by key

## Repartitioning

```
>>> rdd.repartition(4)
...     .coalesce(1)
```

New RDD with 4 partitions

Decrease the number of partitions in the RDD to 1

## Saving

```
>>> rdd.saveAsTextFile("rdd.txt")
>>> rdd.saveAsHadoopFile("hdfs://namenodehost/parent/child",
...                         'org.apache.hadoop.mapred.TextOutputFormat')
```

## Stopping SparkContext

```
>>> sc.stop()
```

## Execution

```
$ ./bin/spark-submit examples/src/main/python/pi.py
```

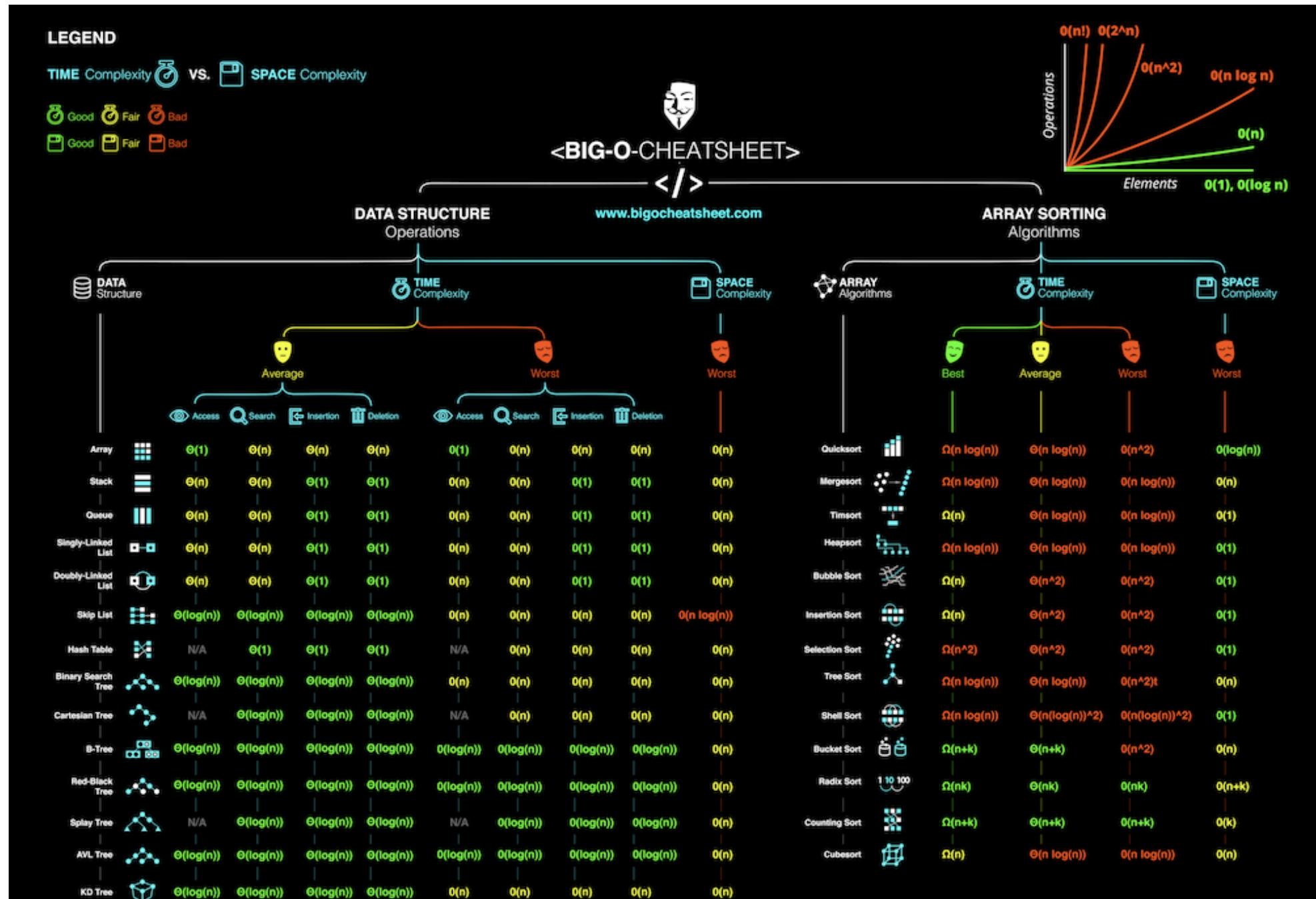
```
>>> textFile = sc.textFile("/my/directory/*.txt")
>>> textFile2 = sc.wholeTextFiles("/my/directory/")
```

```
('b', 2)
('a', 2)
```



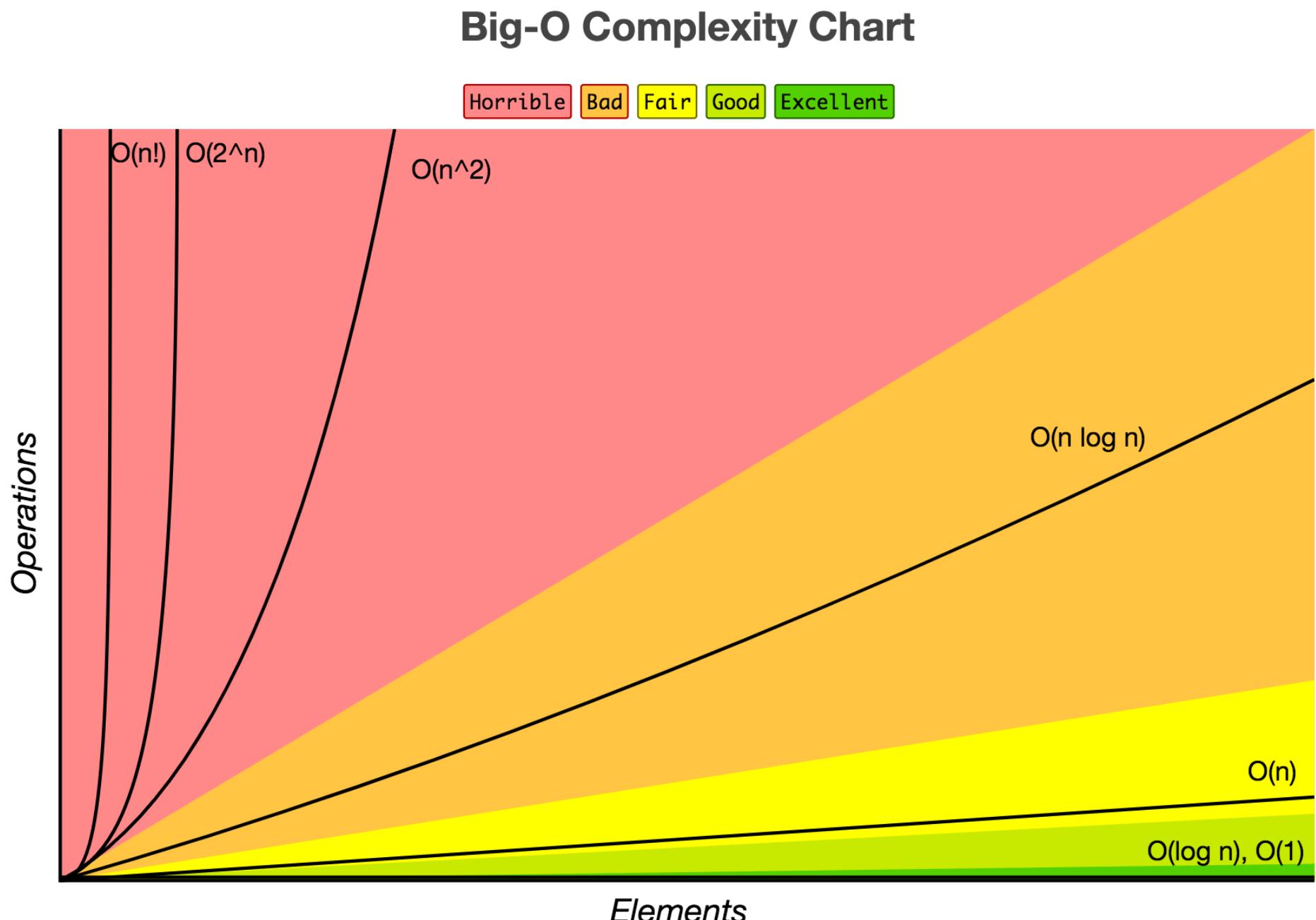
## Pyspark Cheat Sheet

# Big-O





### Big-O Algorithm Cheat Sheet



Big-O Algorithm Complexity Chart

## Common Data Structure Operations

Data Structure	Time Complexity								Space Complexity	
	Average				Worst					
	Access	Search	Insertion	Deletion	Access	Search	Insertion	Deletion		
<u>Array</u>	$\Theta(1)$	$\Theta(n)$	$\Theta(n)$	$\Theta(n)$	$O(1)$	$O(n)$	$O(n)$	$O(n)$	$O(n)$	
<u>Stack</u>	$\Theta(n)$	$\Theta(n)$	$\Theta(1)$	$\Theta(1)$	$O(n)$	$O(n)$	$O(1)$	$O(1)$	$O(n)$	
<u>Queue</u>	$\Theta(n)$	$\Theta(n)$	$\Theta(1)$	$\Theta(1)$	$O(n)$	$O(n)$	$O(1)$	$O(1)$	$O(n)$	
<u>Singly-Linked List</u>	$\Theta(n)$	$\Theta(n)$	$\Theta(1)$	$\Theta(1)$	$O(n)$	$O(n)$	$O(1)$	$O(1)$	$O(n)$	
<u>Doubly-Linked List</u>	$\Theta(n)$	$\Theta(n)$	$\Theta(1)$	$\Theta(1)$	$O(n)$	$O(n)$	$O(1)$	$O(1)$	$O(n)$	
<u>Skip List</u>	$\Theta(\log(n))$	$\Theta(\log(n))$	$\Theta(\log(n))$	$\Theta(\log(n))$	$O(n)$	$O(n)$	$O(n)$	$O(n)$	$O(n \log(n))$	
<u>Hash Table</u>	N/A	$\Theta(1)$	$\Theta(1)$	$\Theta(1)$	N/A	$O(n)$	$O(n)$	$O(n)$	$O(n)$	
<u>Binary Search Tree</u>	$\Theta(\log(n))$	$\Theta(\log(n))$	$\Theta(\log(n))$	$\Theta(\log(n))$	$O(n)$	$O(n)$	$O(n)$	$O(n)$	$O(n)$	
<u>Cartesian Tree</u>	N/A	$\Theta(\log(n))$	$\Theta(\log(n))$	$\Theta(\log(n))$	N/A	$O(n)$	$O(n)$	$O(n)$	$O(n)$	
<u>B-Tree</u>	$\Theta(\log(n))$	$\Theta(\log(n))$	$\Theta(\log(n))$	$\Theta(\log(n))$	$O(\log(n))$	$O(\log(n))$	$O(\log(n))$	$O(\log(n))$	$O(n)$	
<u>Red-Black Tree</u>	$\Theta(\log(n))$	$\Theta(\log(n))$	$\Theta(\log(n))$	$\Theta(\log(n))$	$O(\log(n))$	$O(\log(n))$	$O(\log(n))$	$O(\log(n))$	$O(n)$	
<u>Splay Tree</u>	N/A	$\Theta(\log(n))$	$\Theta(\log(n))$	$\Theta(\log(n))$	N/A	$O(\log(n))$	$O(\log(n))$	$O(\log(n))$	$O(n)$	
<u>AVL Tree</u>	$\Theta(\log(n))$	$\Theta(\log(n))$	$\Theta(\log(n))$	$\Theta(\log(n))$	$O(\log(n))$	$O(\log(n))$	$O(\log(n))$	$O(\log(n))$	$O(n)$	
<u>KD Tree</u>	$\Theta(\log(n))$	$\Theta(\log(n))$	$\Theta(\log(n))$	$\Theta(\log(n))$	$O(n)$	$O(n)$	$O(n)$	$O(n)$	$O(n)$	

BIG-O Algorithm Data Structure Operations

# Array Sorting Algorithms

Algorithm	Time Complexity			Space Complexity
	Best	Average	Worst	
<u>Quicksort</u>	$\Omega(n \log(n))$	$\Theta(n \log(n))$	$O(n^2)$	$O(\log(n))$
<u>Mergesort</u>	$\Omega(n \log(n))$	$\Theta(n \log(n))$	$O(n \log(n))$	$O(n)$
<u>Timsort</u>	$\Omega(n)$	$\Theta(n \log(n))$	$O(n \log(n))$	$O(n)$
<u>Heapsort</u>	$\Omega(n \log(n))$	$\Theta(n \log(n))$	$O(n \log(n))$	$O(1)$
<u>Bubble Sort</u>	$\Omega(n)$	$\Theta(n^2)$	$O(n^2)$	$O(1)$
<u>Insertion Sort</u>	$\Omega(n)$	$\Theta(n^2)$	$O(n^2)$	$O(1)$
<u>Selection Sort</u>	$\Omega(n^2)$	$\Theta(n^2)$	$O(n^2)$	$O(1)$
<u>Tree Sort</u>	$\Omega(n \log(n))$	$\Theta(n \log(n))$	$O(n^2)$	$O(n)$
<u>Shell Sort</u>	$\Omega(n \log(n))$	$\Theta(n(\log(n))^2)$	$O(n(\log(n))^2)$	$O(1)$
<u>Bucket Sort</u>	$\Omega(n+k)$	$\Theta(n+k)$	$O(n^2)$	$O(n)$
<u>Radix Sort</u>	$\Omega(nk)$	$\Theta(nk)$	$O(nk)$	$O(n+k)$
...	...	...	...	...

Counting Sort $\Omega(n+k)$  $\Theta(n+k)$  $O(n+k)$  $O(k)$ Cubesort $\Omega(n)$  $\Theta(n \log(n))$  $O(n \log(n))$  $O(n)$ 

Big-O Array Sorting Algorithms

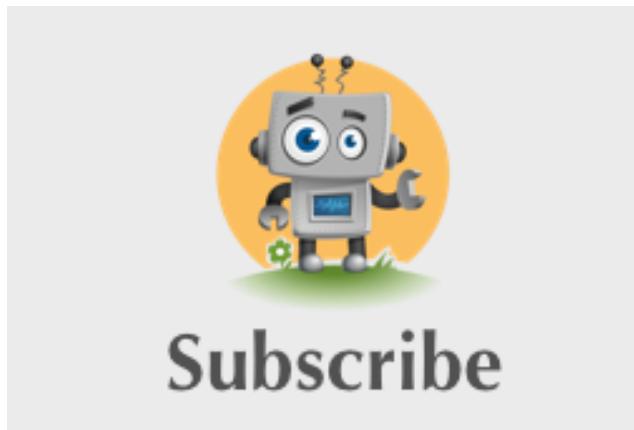
## About Stefan

Stefan is the founder of [Chatbot's Life](#), a Chatbot media and consulting firm. Chatbot's Life has grown to over 150k views per month and has become the premium place to learn about Bots & AI online. Chatbot's Life has also consulted many of the top Bot companies like Swelly, Instavest, OutBrain, NearGroup and a number of Enterprises.

176 Q 29



...



## Resources

Big-O Algorithm Cheat Sheet: <http://bigocheatsheet.com/>

Bokeh Cheat Sheet:

[https://s3.amazonaws.com/assets.datacamp.com/blog\\_assets/Python\\_Bokeh\\_Cheat\\_Sheet.pdf](https://s3.amazonaws.com/assets.datacamp.com/blog_assets/Python_Bokeh_Cheat_Sheet.pdf)

Data Science Cheat Sheet:

<https://www.datacamp.com/community/tutorials/python-data-science-cheat-sheet-basics>

Data Wrangling Cheat Sheet: <https://www.rstudio.com/wp-content/uploads/2015/02/data-wrangling-cheatsheet.pdf>

Data Wrangling: [https://en.wikipedia.org/wiki/Data\\_wrangling](https://en.wikipedia.org/wiki/Data_wrangling)

Ggplot Cheat Sheet: <https://www.rstudio.com/wp-content/uploads/2015/03/ggplot2-cheatsheet.pdf>

Keras Cheat Sheet:

<https://www.datacamp.com/community/blog/keras-cheat-sheet#gs.DRKeNMs>

Keras: <https://en.wikipedia.org/wiki/Keras>

Machine Learning Cheat Sheet: <https://ai.icymi.email/new-machinelearning-cheat-sheet-by-emily-barry-abdsc/>

Machine Learning Cheat Sheet: <https://docs.microsoft.com/en-in/azure/machine-learning/machine-learning-algorithm-cheat-sheet>

ML Cheat Sheet:: <http://peekaboo-vision.blogspot.com/2013/01/machine-learning-cheat-sheet-for-scikit.html>

Matplotlib Cheat Sheet:

<https://www.datacamp.com/community/blog/python-matplotlib-cheat-sheet#gs.uEKySpY>

Matplotlib: <https://en.wikipedia.org/wiki/Matplotlib>

Neural Networks Cheat Sheet: <http://www.asimovinstitute.org/neural-network-zoo/>

Neural Networks Graph Cheat Sheet:

<http://www.asimovinstitute.org/blog/>

Neural Networks: <https://www.quora.com/Where-can-find-a-cheat-sheet-for-neural-network>

Numpy Cheat Sheet:

<https://www.datacamp.com/community/blog/python-numpy-cheat-sheet#gs.AK5ZBgE>

NumPy: <https://en.wikipedia.org/wiki/NumPy>

Pandas Cheat Sheet:

<https://www.datacamp.com/community/blog/python-pandas-cheat-sheet#gs.oundfxM>

Pandas: [https://en.wikipedia.org/wiki/Pandas\\_\(software\)](https://en.wikipedia.org/wiki/Pandas_(software))

Pandas Cheat Sheet:

<https://www.datacamp.com/community/blog/pandas-cheat-sheet-python#gs.HPFoRIC>

Pyspark Cheat Sheet:

<https://www.datacamp.com/community/blog/pyspark-cheat-sheet-python#gs.L=J1zxQ>

Scikit Cheat Sheet:

<https://www.datacamp.com/community/blog/scikit-learn-cheat-sheet>

Scikit-learn: <https://en.wikipedia.org/wiki/Scikit-learn>

Scikit-learn Cheat Sheet: <http://peekaboo-vision.blogspot.com/2013/01/machine-learning-cheat-sheet-for-scikit.html>

Scipy Cheat Sheet:  
<https://www.datacamp.com/community/blog/python-scipy-cheat-sheet#gs.JDSg3OI>

SciPy: <https://en.wikipedia.org/wiki/SciPy>

TesorFlow Cheat Sheet: <https://www.altoros.com/tensorflow-cheat-sheet.html>

Tensor Flow: <https://en.wikipedia.org/wiki/TensorFlow>



