How to get started with Machine Learning

- Melanie Warrick

nyghtowl.io
@nyghtowl

Covering...

- Machine Learning Overview
- AI, Data Science, & Big Data Relationships
- Example Code - Linear Regression
- Algorithms & Tools
- Skills & Resources
- Questions

My Background

▶ **Data Science**

▶ **Software Engineering**

▶ **Business Domain Expertise**

Machine Learning

Computers...ability to learn without... explicit programing

-Arthur Samuel (1959)

- Build a model that finds patterns and/or predicts results

- Apply algorithm(s)
- Pick best result for pattern match or prediction

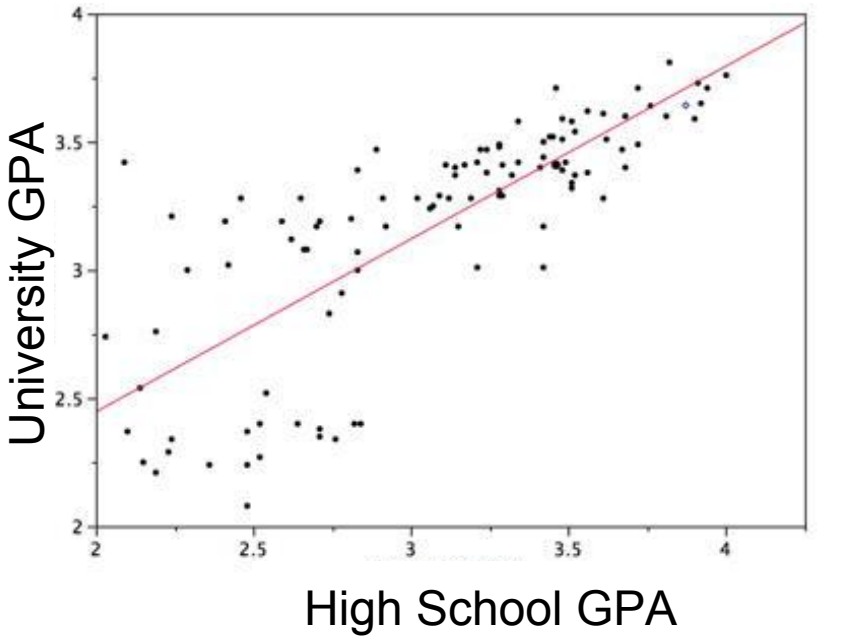
What is a Model?

y = mx + b

Find best fit m & b

algorithm to predict / pattern match

Linear Regression Model Example



@nyghtowl

Example Problems

- Handwritten address recognition
- Search engines - Google, Bing
- Twitter & Facebook Friend Recommender or Netflix
- Fraud detection
- Weather prediction
- Facial recognition

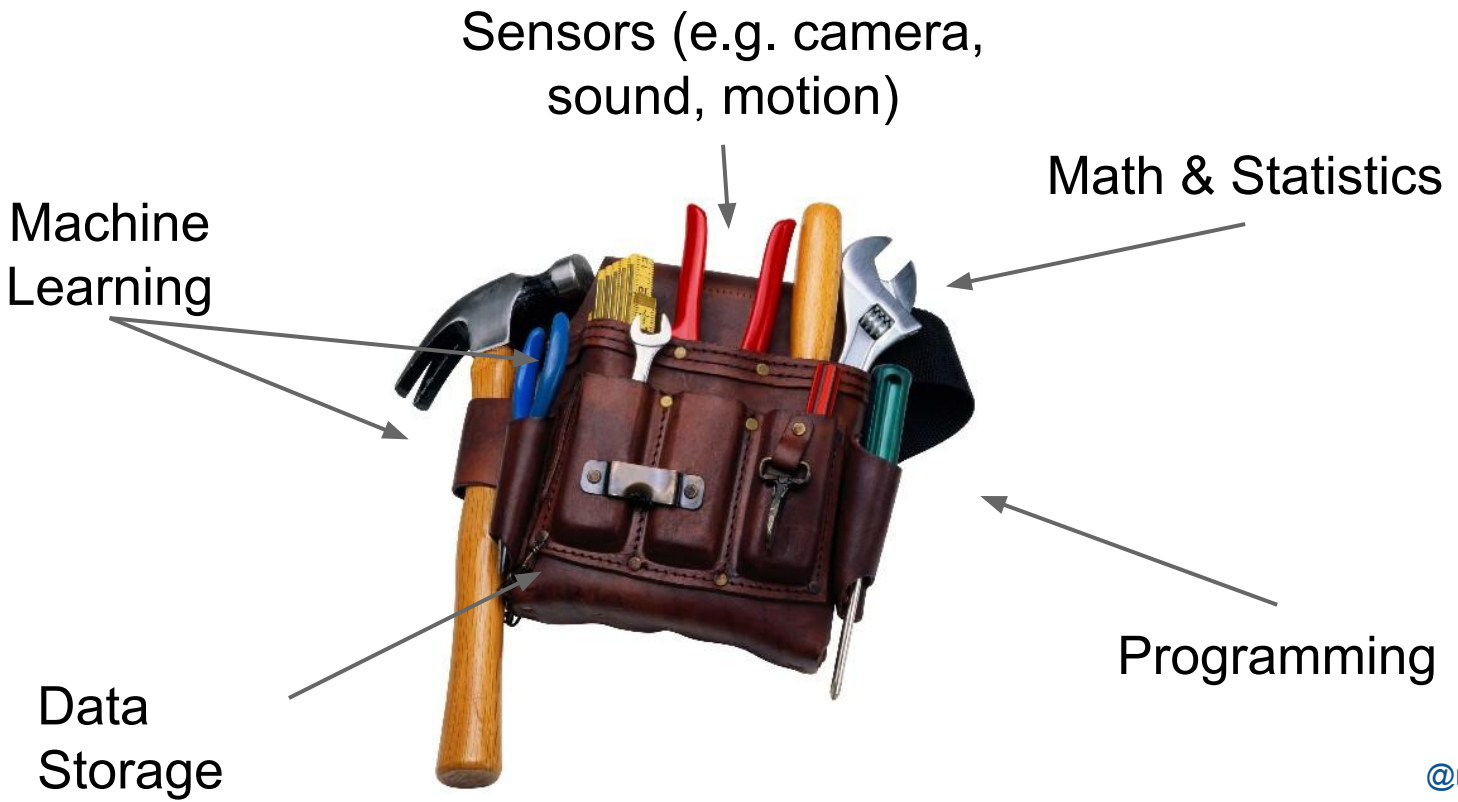
Trending Topics & Terminology

- **AI** = intelligence exhibited by machines or software

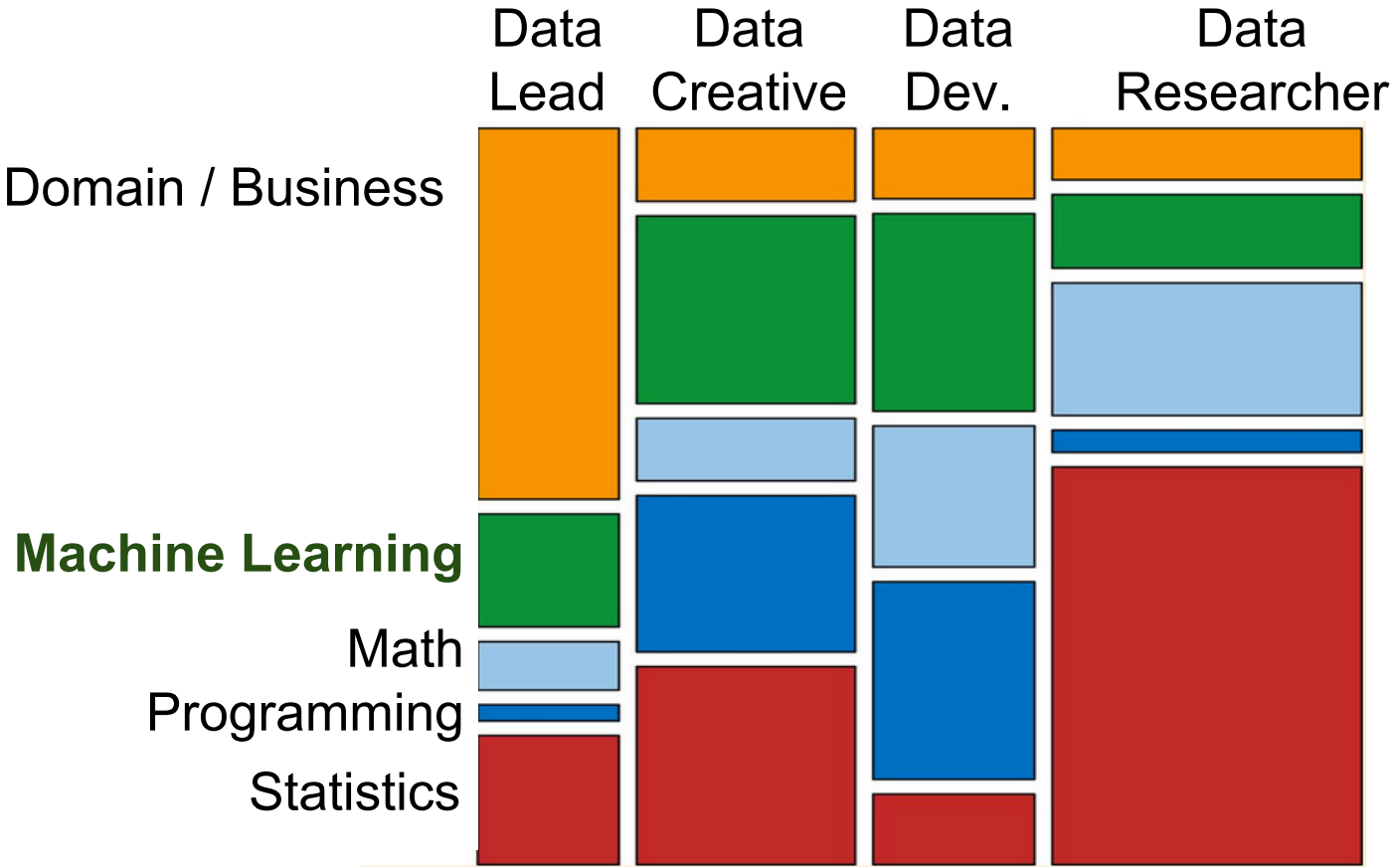
- **Data Science** = get knowledge from data and create products

- **Big Data** = beyond ability of common tech to capture and curate
  - *2 GB = 20 yrds of books | 50 PB = entire written works of humankind*

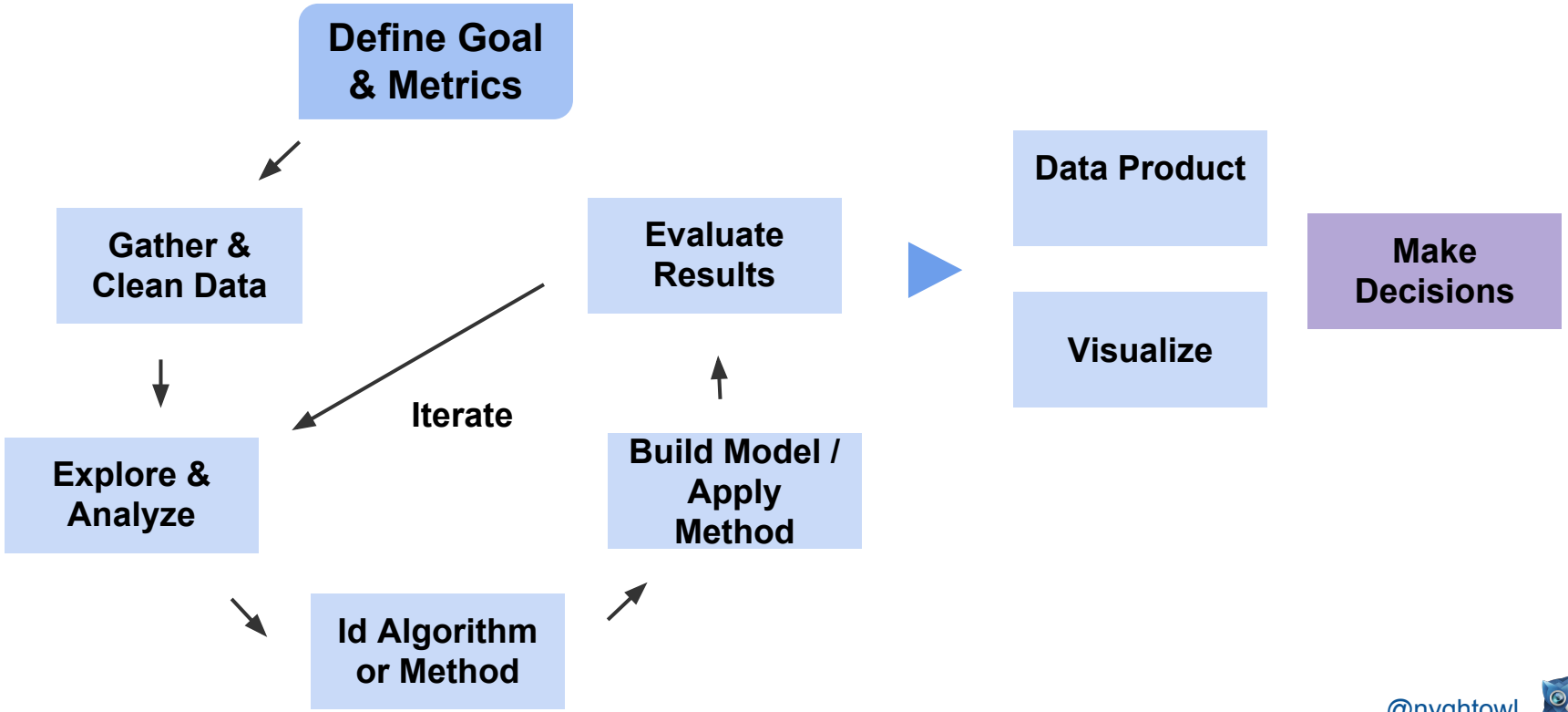
Tool in AI's Toolbelt



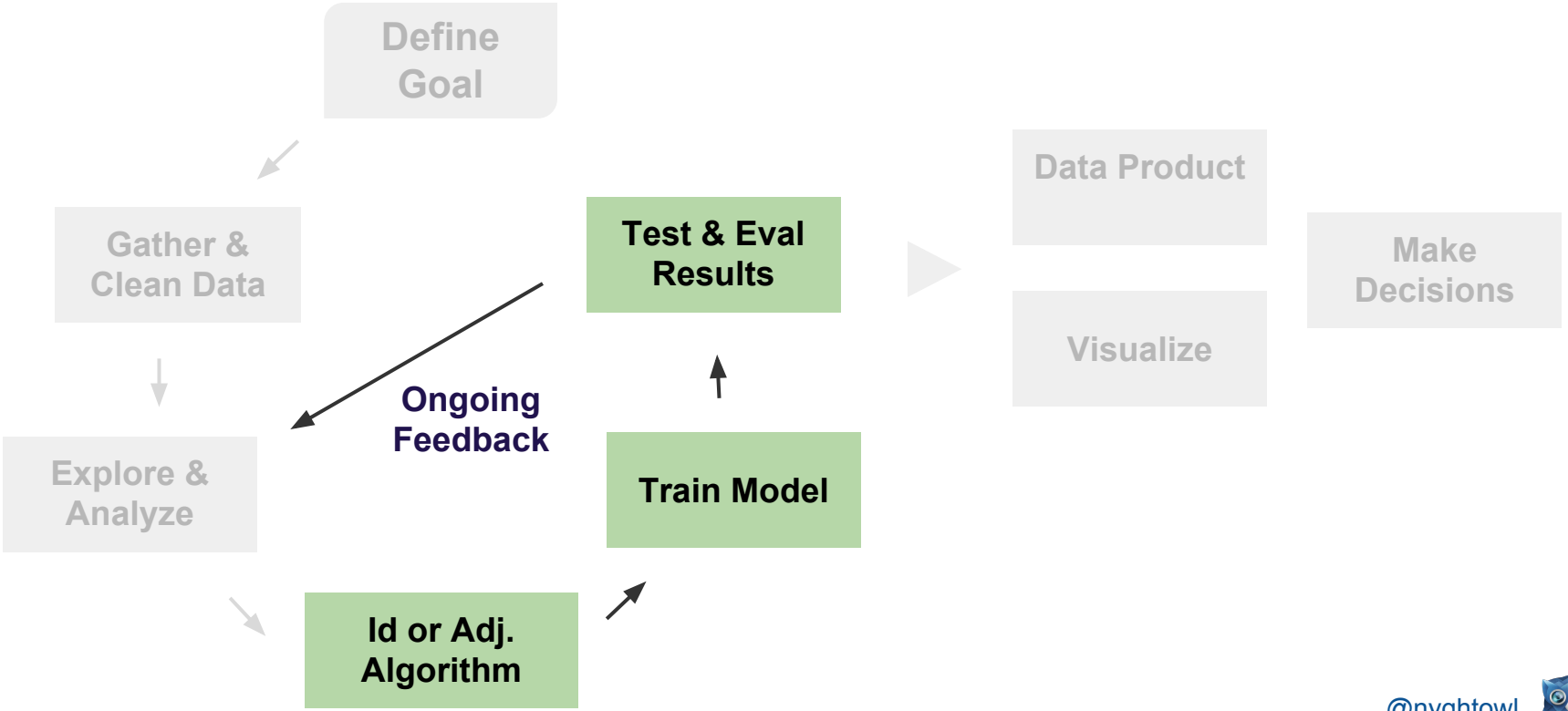
Data Science Roles & Skills



Data Science Project Flow



Machine Learning Flow



Machine Learning

Computers...ability to learn without... explicit programing

-Arthur Samuel (1959)

- Build a model that finds patterns and/or predicts results

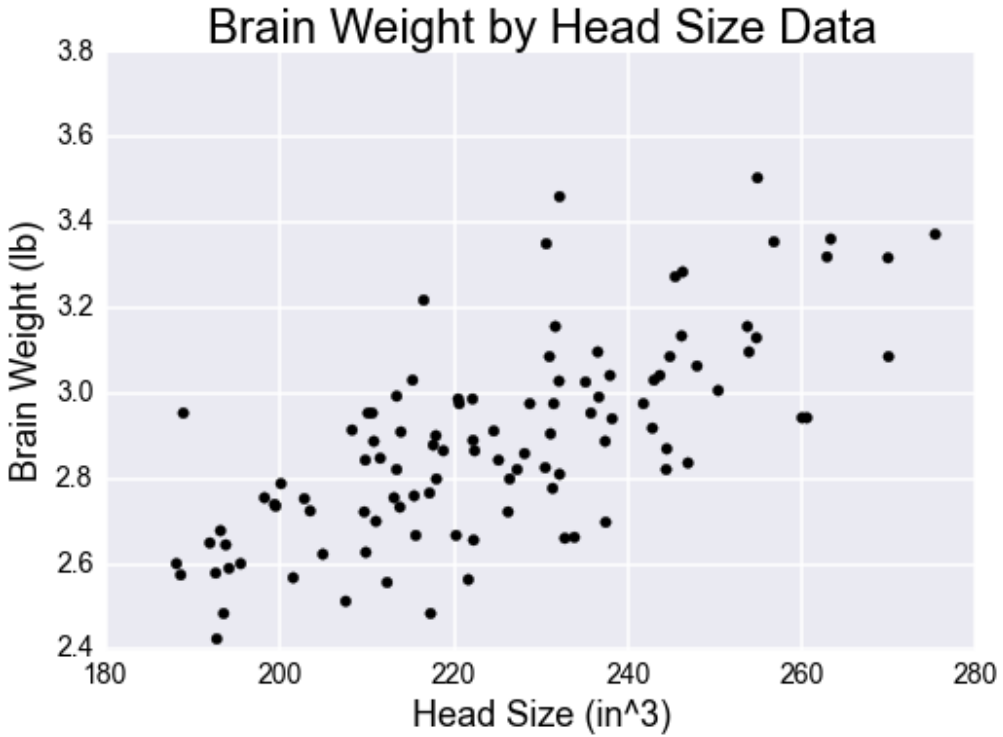
- Apply algorithm(s)
- Pick best result for prediction and pattern match

Ex: Linear Regression

Goal:

Predict Brain Weight

with Head Size



Ex: Get Data

```
import pandas as pd
from sklearn.cross_validation import
train_test_split

data = pd.read_csv(filename, sep="\t", header=0)
X, y = data['Head_Size'], data['Brain_Weight']

X_train, X_test, y_train, y_test =
    train_test_split(X, y, test_size=0.30)
```


Ex: Training & Test Data Split

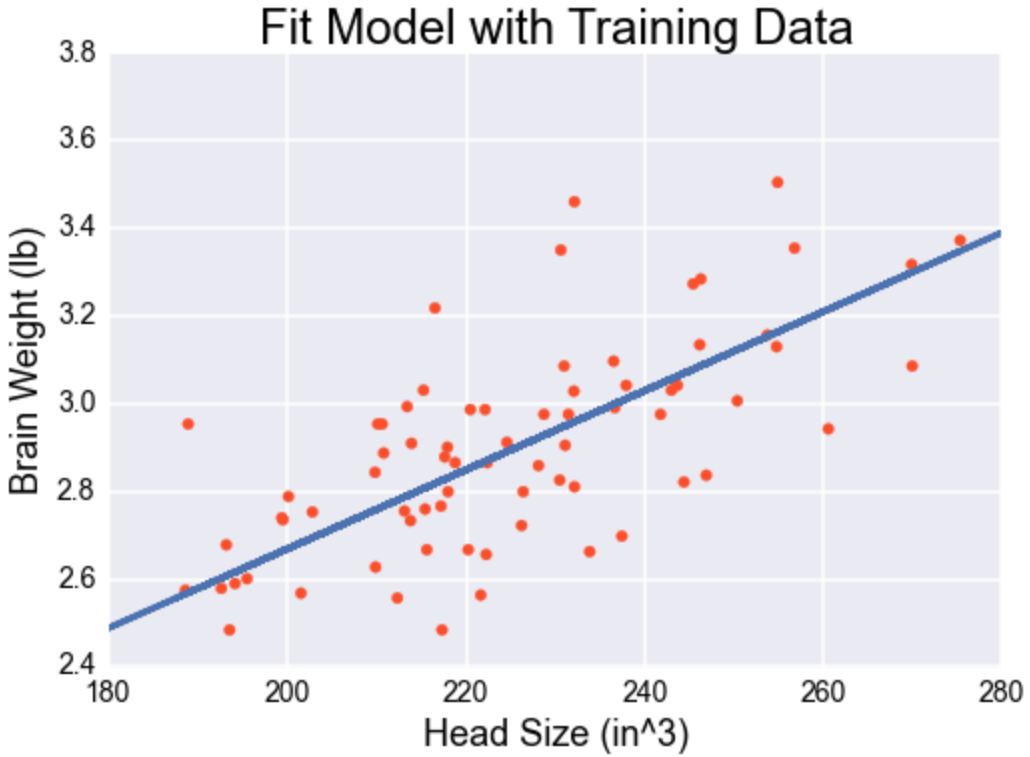
Ex: Create Model

```
from sklearn import datasets, linear_model


# Create y=mx+b template
model = linear_model.LinearRegression()


# Train the model - define m & b
model.fit(X_train, y_train)
```

m ~ 0.01

b ~ 0.89

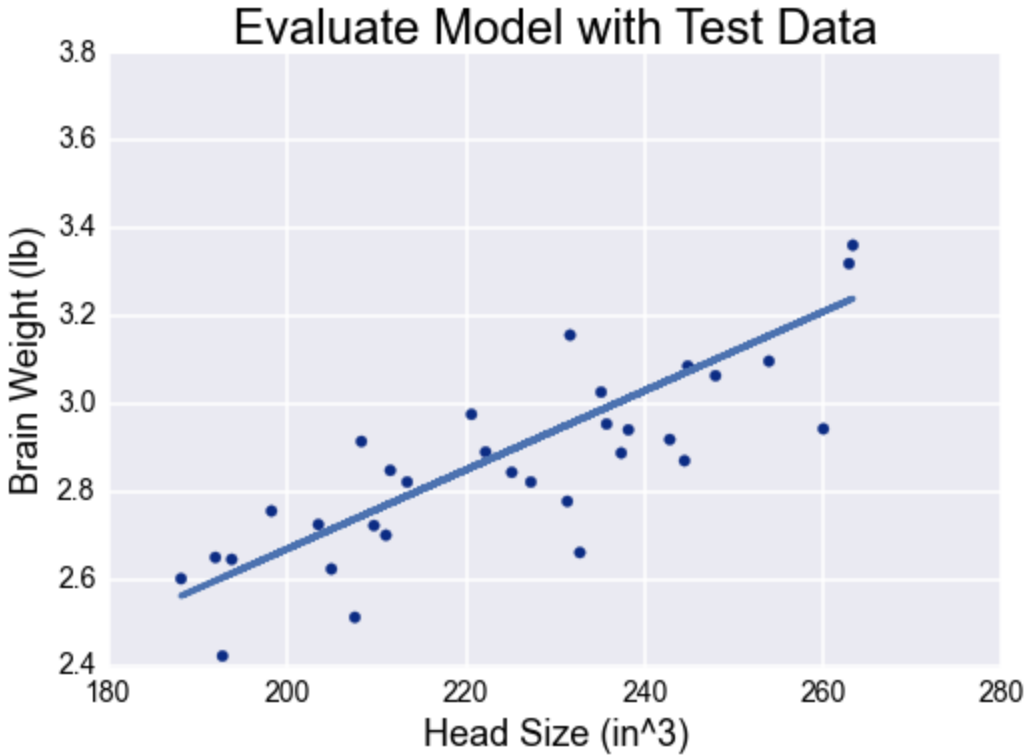
Ex: Fit Model to Training Data



y = 0.01x + .089

Ex: Evaluate Model with Test Data

```
y predictions =

    model.predict

    (X_test)
```



Ex: Sample Metric - Accuracy

```
# R squared: 1 is perfect prediction
```

```
print 'Accuracy', model.score(X_test, y_test)
```

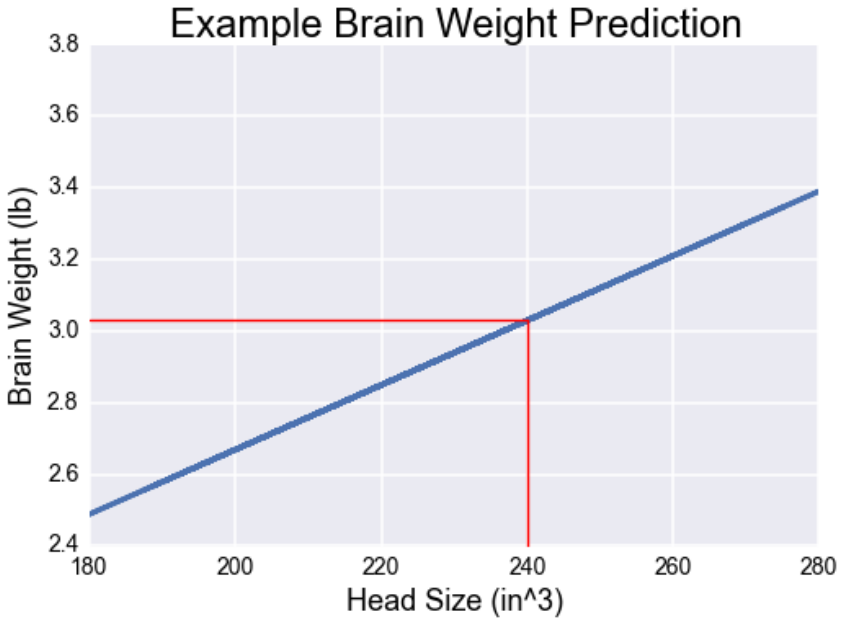
Accuracy: 0.63

Ex: Predict with New Data

```
model.predict(240) = 3.03
```

x = 240

y = 3.03



Ex: Visualize

```
import matplotlib.pyplot as plt
import seaborn
```

```
plt.scatter(X_test, y_test)
plt.plot(X_test,
    model.predict(X_test))
```


```
plt.title('Evaluate Model with...', fontsize=24)
plt.show()
```

Machine Learning Algorithms *(sample)*

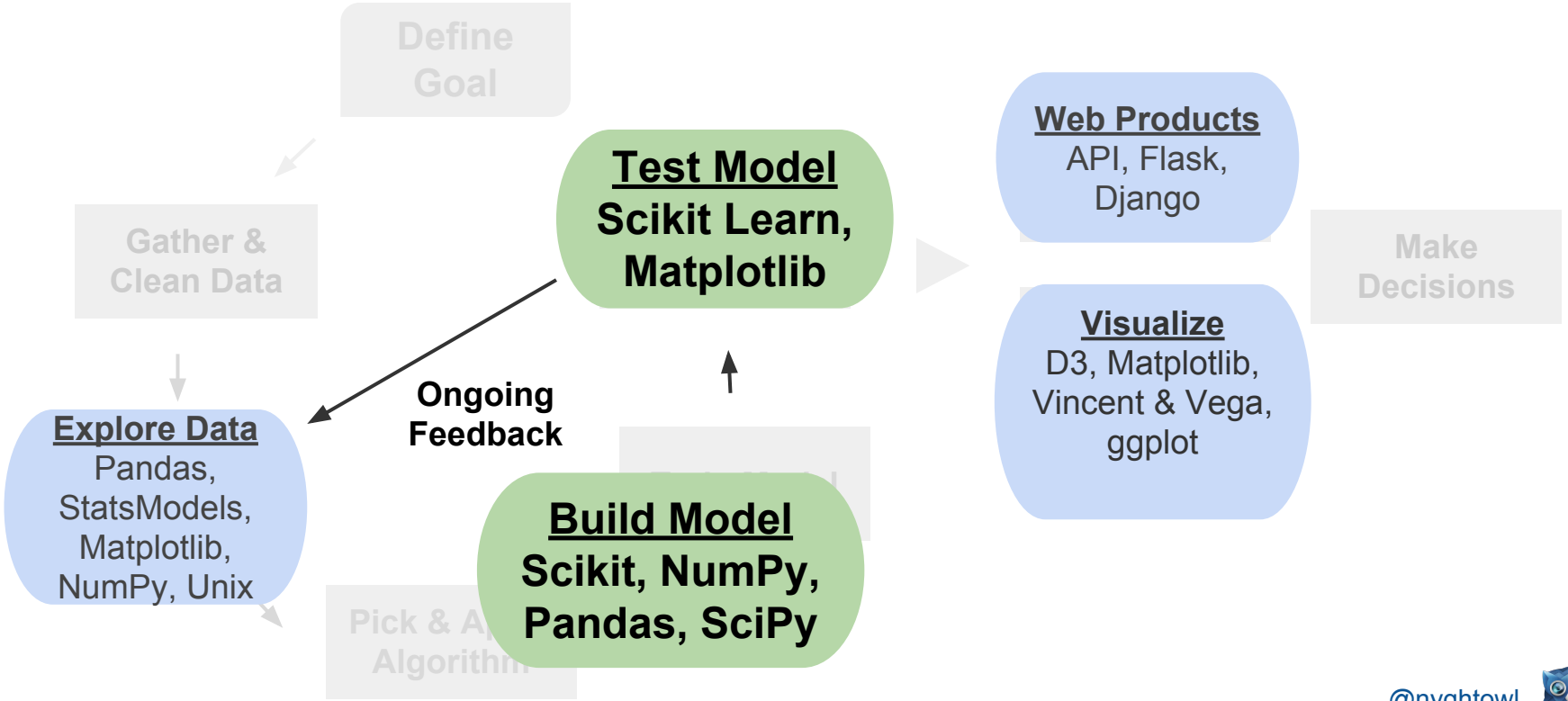
	Unsupervised	Supervised

Continuous	• Clustering & Dimensionality Reduction
	  ○ SVD
	  ○ PCA
	  ○ K-means
		• Regression
		  ○ Linear
		  ○ Polynomial
		• Decision Trees
		• Random Forests

Categorical	• Association Analysis
	  ○ Apriori
	  ○ FP-Growth
	• Hidden Markov Model
		• Classification
		  ○ KNN
		  ○ Trees
		  ○ Logistic Regression
		  ○ Naive-Bayes
		  ○ SVM

@nyghtowl

Machine Learning Key Tools



@nyghtowl

Machine Learning Skills to Build

- Algorithms
- Statistics (probability, inferential, descriptive)
- Linear Algebra (vectors & matrices)
- Data Analysis (intuition)
- SQL, Python, R, Java, Scala (programming)
- Databases & APIs (get data)

Machine Learning Resources

- Andrew Ng's Machine Learning on Coursera
- Khan Academy (linear algebra and stats)
- "Think Stats" - Allen Downey
- Zipfian's Practical Intro to Data Science
- Metacademy
- Open Source Data Science Masters
- StackOverflow, Data Tau, Kaggle
- Mentors

Last Thoughts

Help the machine learn without explicit programming

Tool used in AI, Data Science & big data

Key skills = algorithms, stats, programming and analytics

How to get started with Machine Learning

More info at:

nyghtowl.io

https://github.com/nyghtowl/PyCon_2014

Key References

- Zipfian
- Framed.io
- “Analyzing the Analyzers” - Harlan Harris, Sean Murphy, Marck Vaisman
- “Doing Data Science” - Rachel Schutt & Cathy O’Neil
- “Collective Intelligence” - Toby Segaran
- “Some Useful Machine Learning Libraries” (blog)
- University GPA Linear Regression Example
- Scikit-Learn (esp. linear regression)
- Mozy Blog
- StackOverflow
- Wiki