Lecture 01

What is Machine Learning? An Overview.

STAT 479: Machine Learning, Fall 2018

Sebastian Raschka

http://stat.wisc.edu/~sraschka/teaching/stat479-fs2018/

About this Course

When

- Tue 8:00-9:15 am
- Thu 8:00-9:15 am

Where

- SMI 331

Office Hours

- Sebastian Raschka:
  Tue 3:00-4:00, Room MSC 1171
- Shan Lu (TA):
  Wed 3:00-4:00 pm, Room MSC B248

For details -> http://stat.wisc.edu/~sraschka/teaching/stat479-fs2018/

What is Machine Learning?

"Machine learning is the hot new thing"

— John L. Hennessy, President of Stanford (2000–2016)

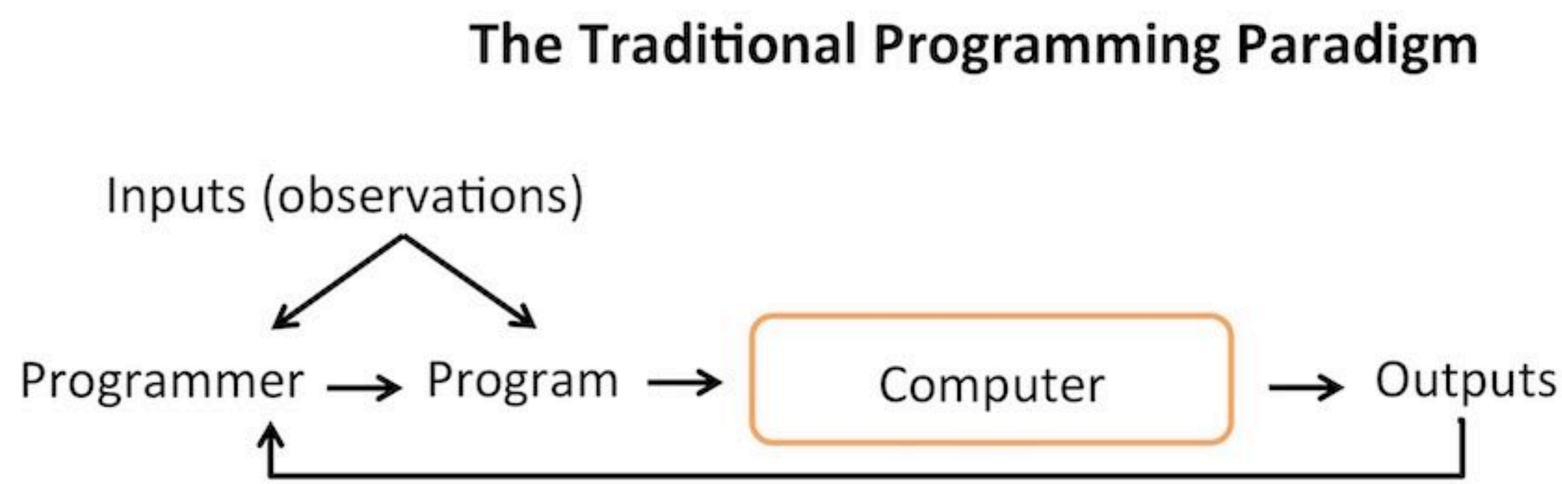
"A breakthrough in machine learning would be worth ten Microsofts"
— Bill Gates, Microsoft Co-Founder

"Machine learning is the field of study that gives computers the ability to learn without being explicitly programmed"

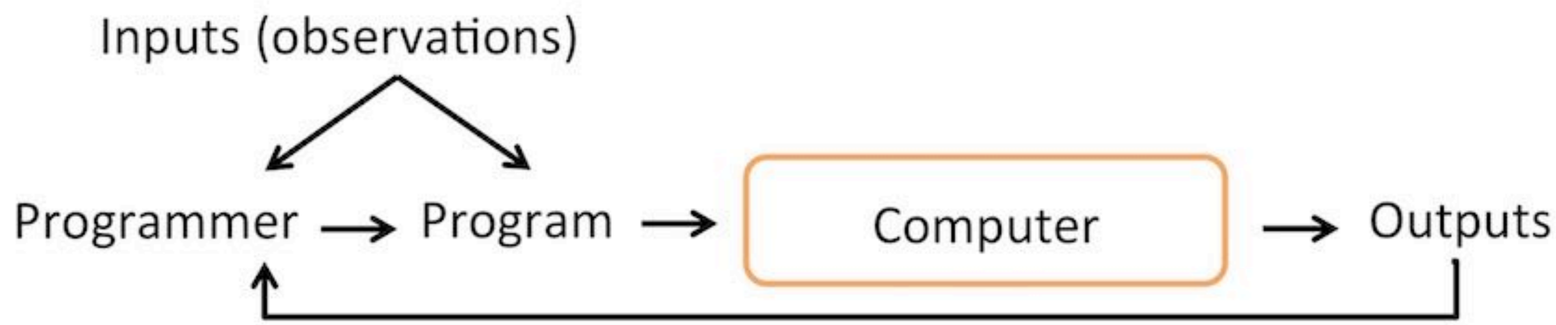
— Arthur L. Samuel, AI pioneer, 1959

(This is likely not an original quote but a paraphrased version of Samuel's sentence "Programming computers to learn from experience should eventually eliminate the need for much of this detailed programming effort.")

Arthur L Samuel. "Some studies in machine learning using the game of checkers". In: *IBM Journal of research and development* 3.3 (1959), pp. 210–229.

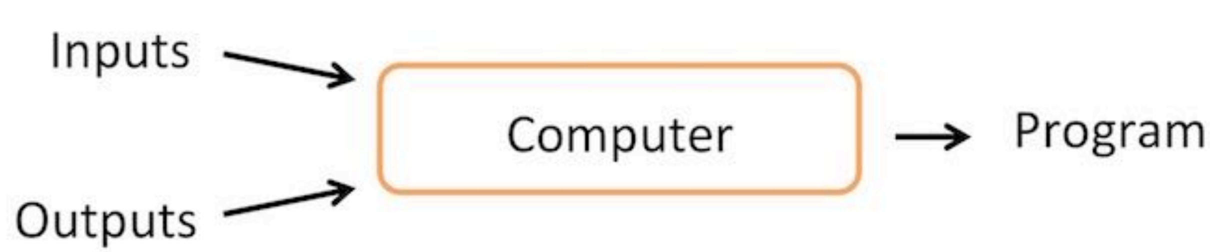

The Traditional Programming Paradigm
Inputs (observations)
Programmer
Program
Computer
Outputs

The Traditional Programming Paradigm



Machine Learning is the field of study that gives computers the ability to learn without being explicitly programmed
– Arthur Samuel (1959)

Machine Learning



Sebastian Raschka, 2016

"If software ate the world, models will run it"
— Steven A. Cohen and Matthew W. Granade, The Wallstreet Journal, 2018

"A computer program is said to **learn** from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."

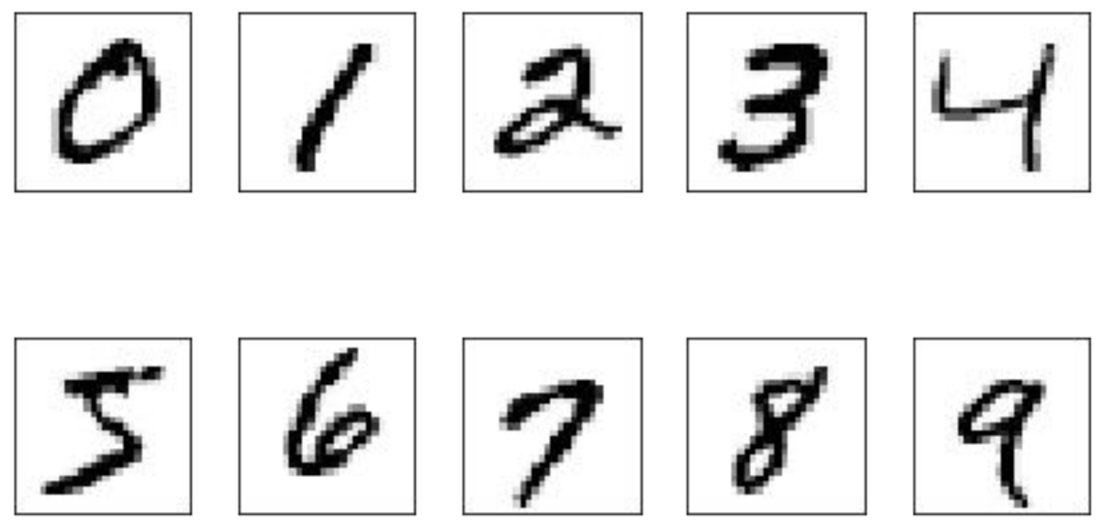
— Tom Mitchell, Professor at Carnegie Mellon University

Tom M Mitchell et al. "Machine learning. 1997". In: *Burr Ridge, IL: McGraw Hill* 45.37 (1997), pp. 870–877.

"A computer program is said to **learn** from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."

— Tom Mitchell, Professor at Carnegie Mellon University

Handwriting Recognition Example:



- Task T:
- Performance measure P :
- Training experience E:

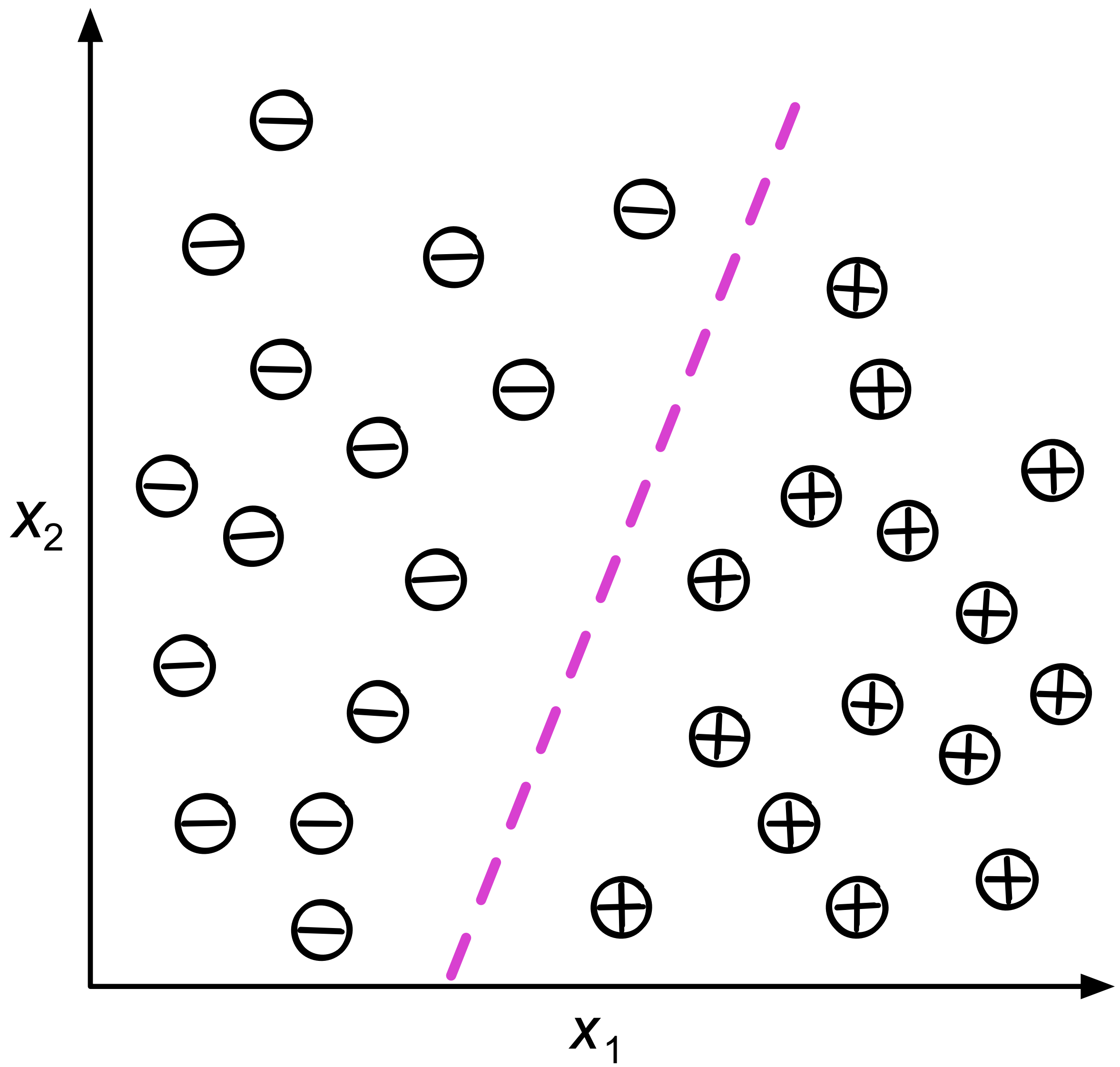
Some Applications of Machine Learning (1):

Some Applications of Machine Learning (2):

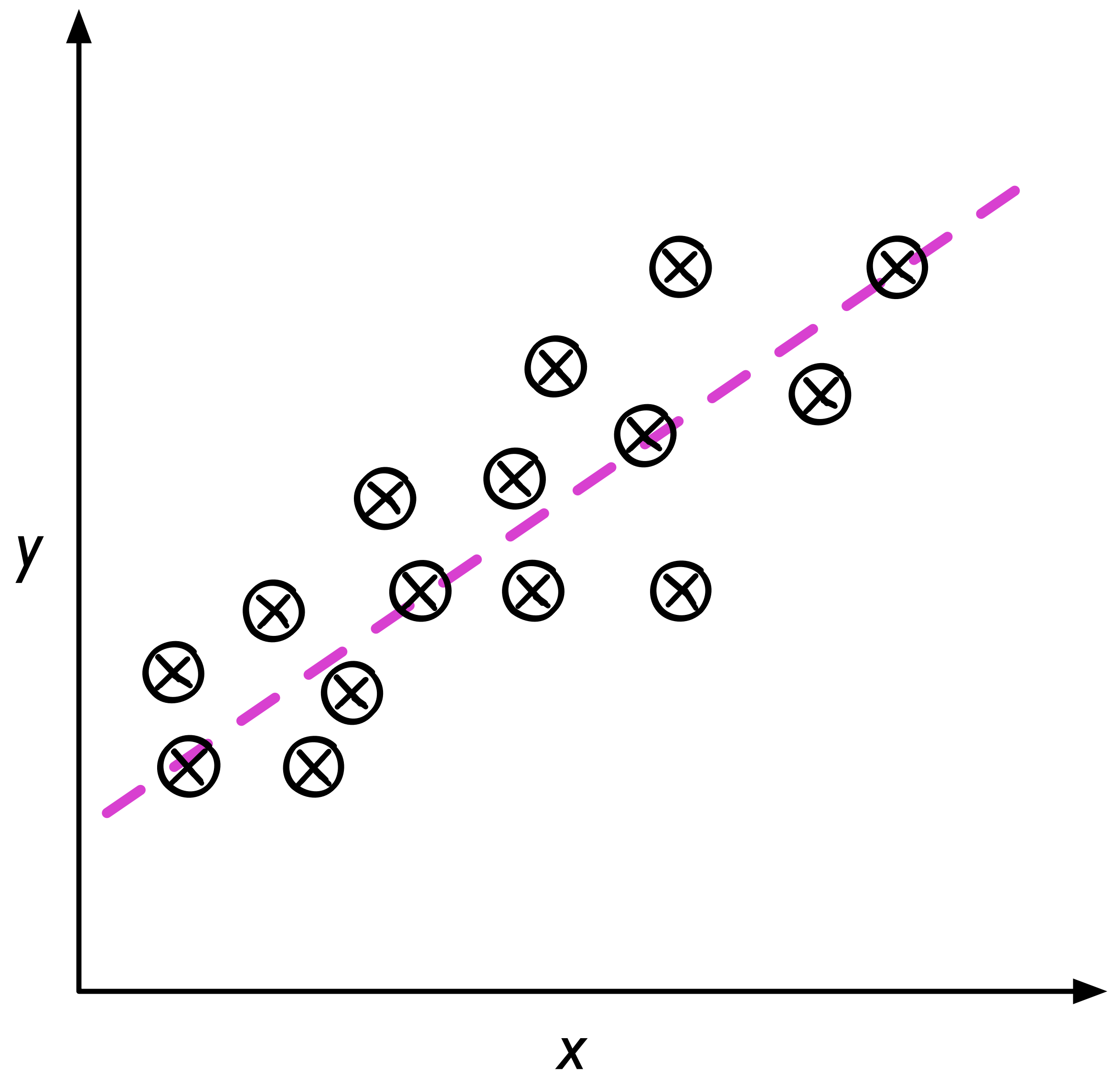
Categories of Machine Learning

Supervised Learning	- Labeled data - Direct feedback - Predict outcome/future

Supervised Learning: Classification



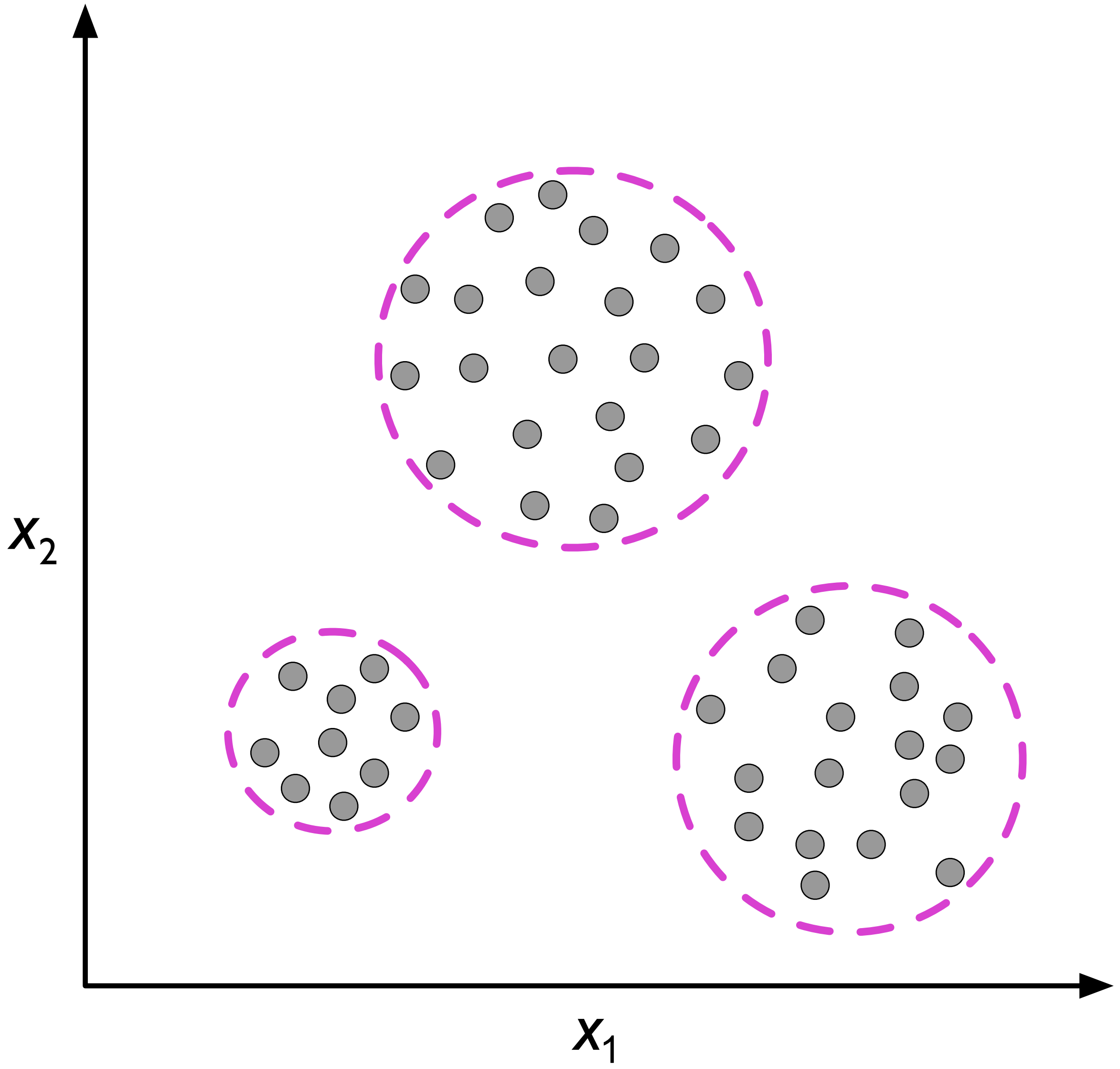
Supervised Learning: Regression



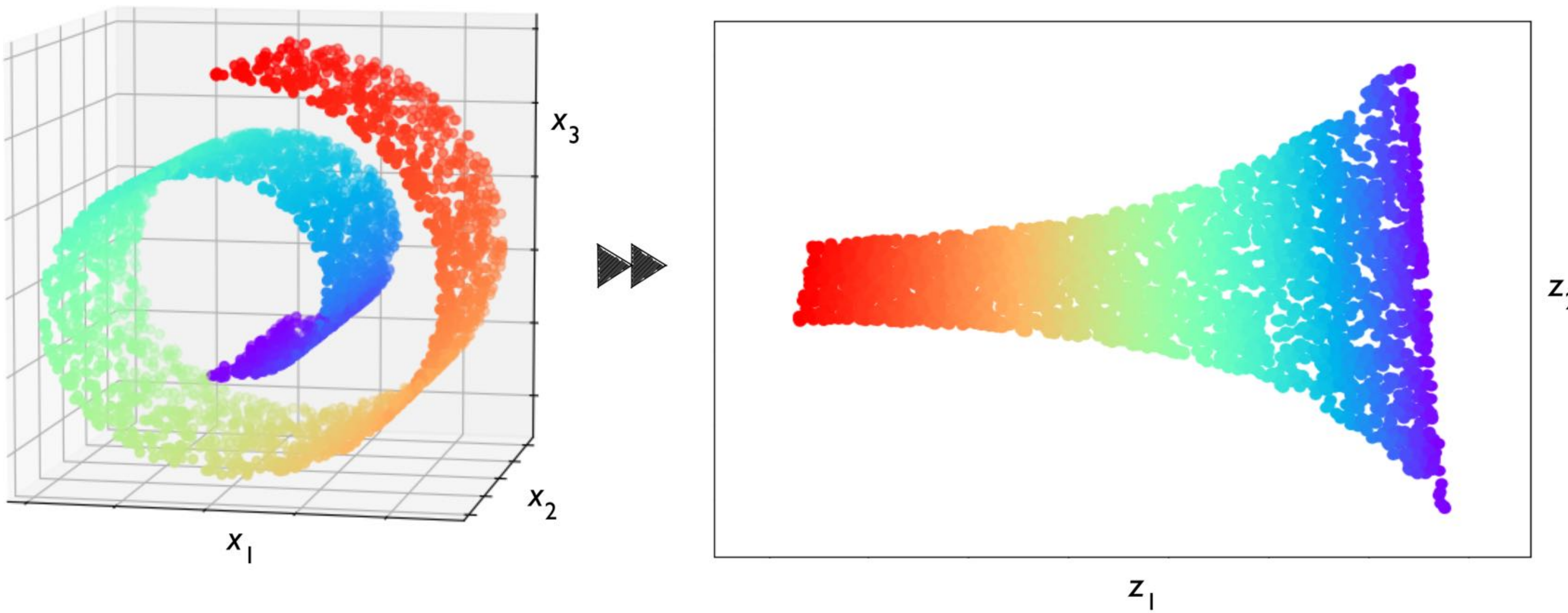
Categories of Machine Learning

Supervised Learning	Labeled data Direct feedback Predict outcome/future
Unsupervised Learning	No labels/targets No feedback Find hidden structure in data

Unsupervised Learning -- Clustering



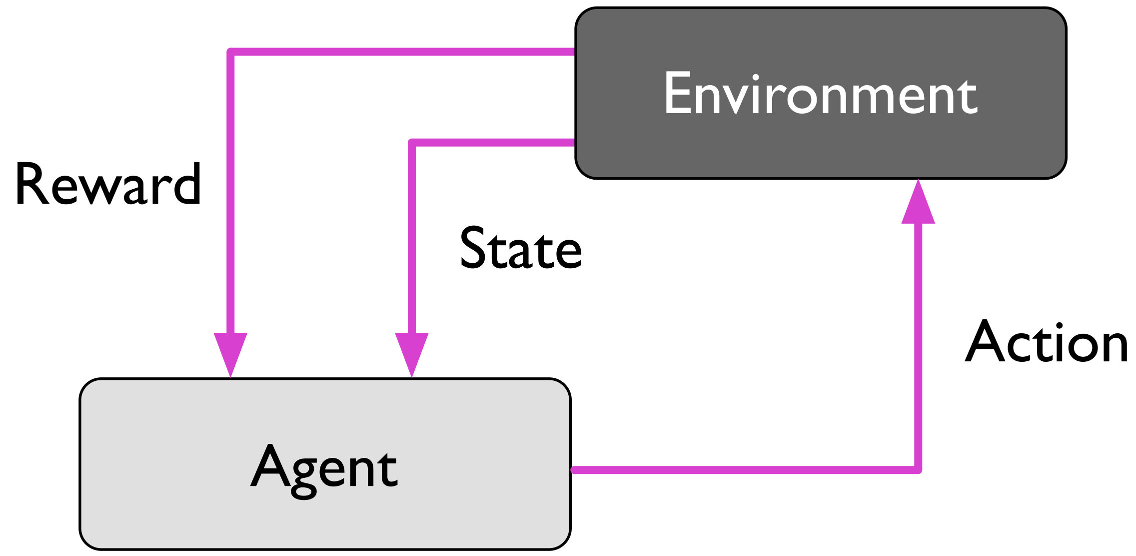
Unsupervised Learning
-- Dimensionality Reduction



Categories of Machine Learning

Supervised Learning	› Labeled data › Direct feedback › Predict outcome/future
Unsupervised Learning	› No labels/targets › No feedback › Find hidden structure in data
Reinforcement Learning	› Decision process › Reward system › Learn series of actions

Reinforcement Learning



Semi-Supervised Learning

Supervised Learning (Formal Notation)

Training set: $\mathscr{D} = \{ <\mathbf{x}^{[i]}, y^{[i]}>, i = 1, \dots, n\},$

Unknown function: $f(\mathbf{x}) = y$

Hypothesis: $h(\mathbf{x}) = \hat{y}$

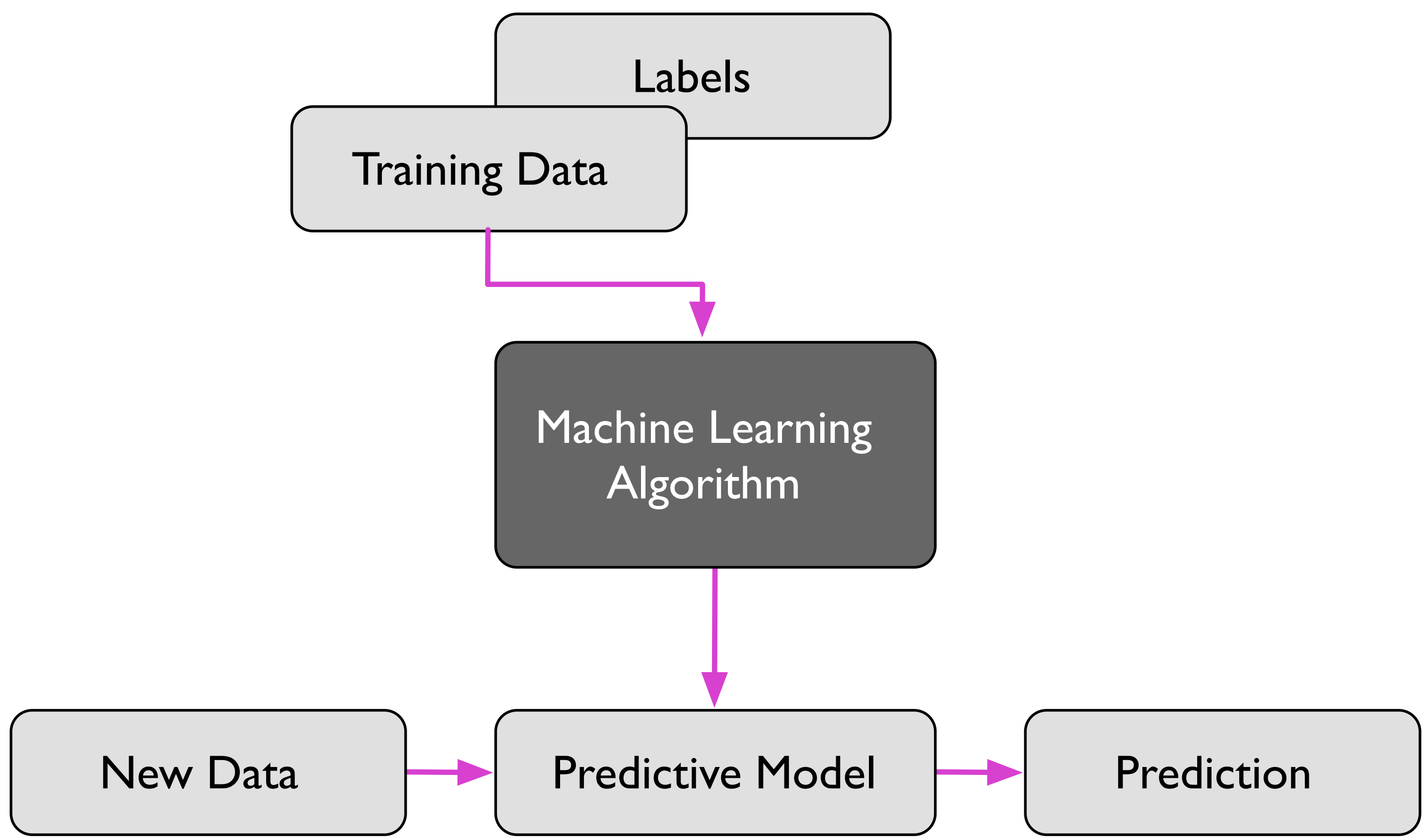
Classification

$h : \mathbb{R}^m \rightarrow \_\_$

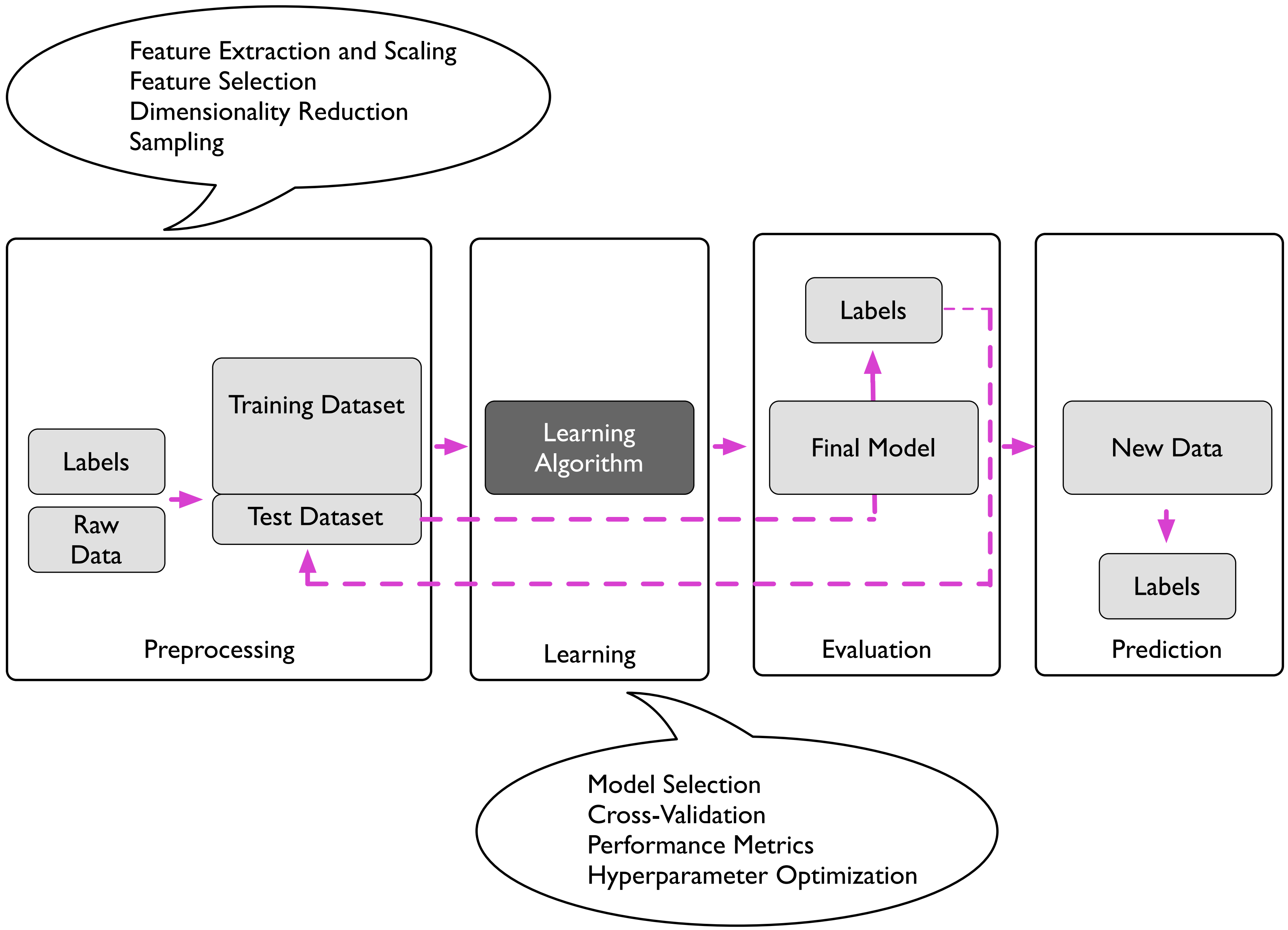
Regression

$h : \mathbb{R}^m \rightarrow \_\_$

Supervised Learning Workflow -- Overview



Supervised Learning Workflow -- More Detailed Overview



Data Representation

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix}$$

Feature vector

Data Representation

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix}$$

Feature vector

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix}$$

Data Representation

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} x_1^{[1]} & x_2^{[1]} & \cdots & x_m^{[1]} \\ x_1^{[2]} & x_2^{[2]} & \cdots & x_m^{[2]} \\ \vdots & \vdots & \ddots & \vdots \\ x_1^{[n]} & x_2^{[n]} & \cdots & x_m^{[n]} \end{bmatrix}$$

Feature vector

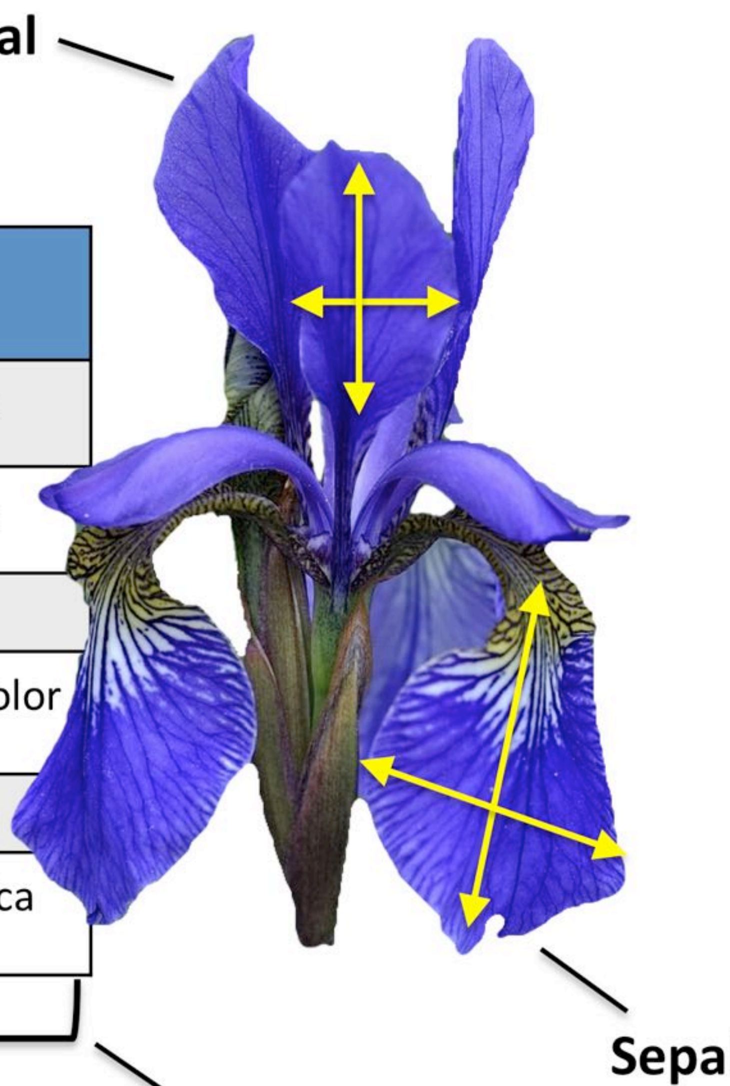
____________ ______________ ____________ ______________

Data Representation

$m =$ ____

$n =$ ____

	Sepal length	Sepal width	Petal length	Petal width	
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
			...		
50	6.4	3.5	4.5	1.2	Versicolor
			...		
150	5.9	3.0	5.0	1.8	Virginica

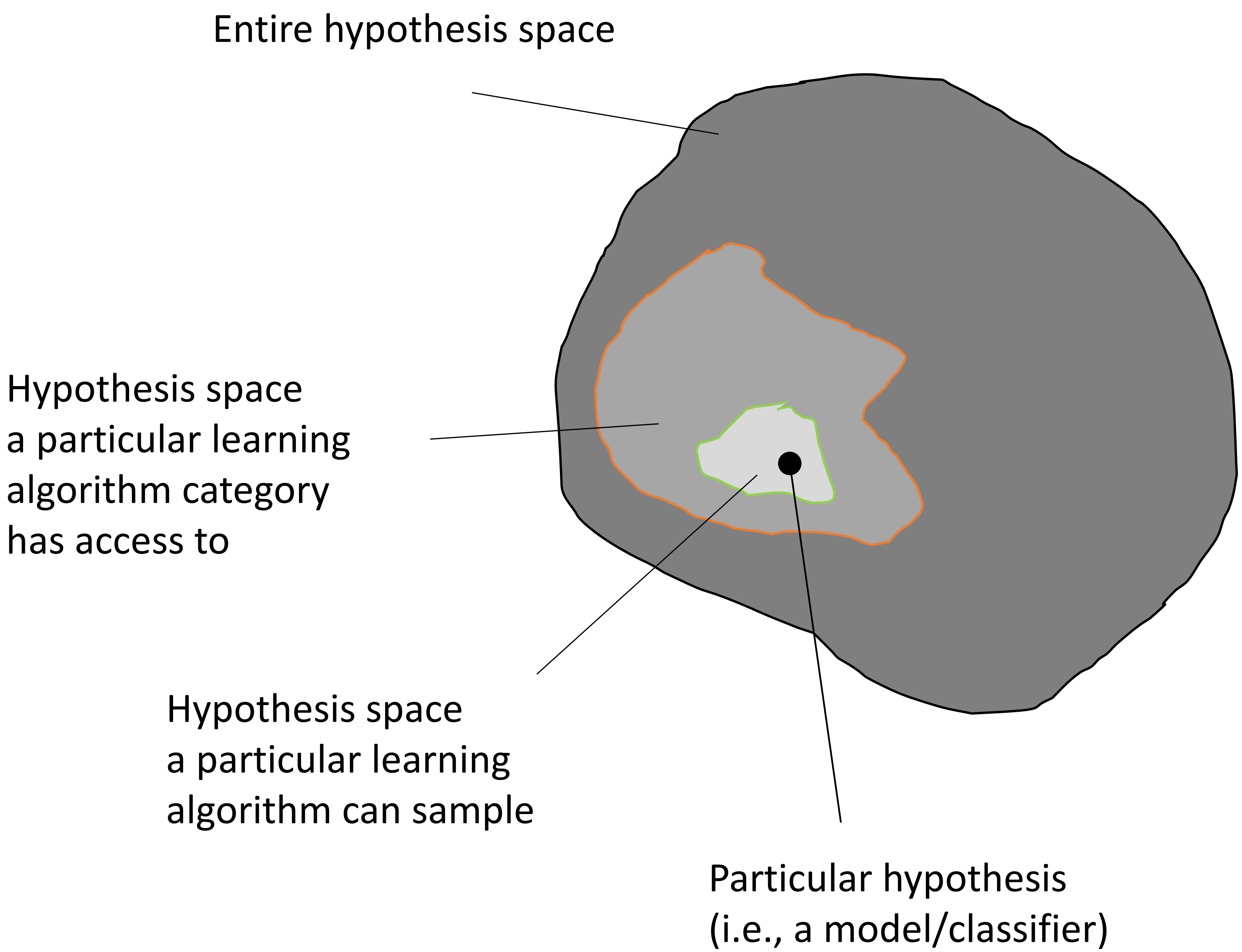


Data Representation

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} y^{[1]} \\ y^{[2]} \\ \vdots \\ y^{[n]} \end{bmatrix}$$

Input features ______________

Hypothesis Space



Hypothesis Space Size

sepal length < 5 cm	sepal width < 5 cm	petal length < 5 cm	petal width < 5 cm	Class Label
True	True	True	True	Setosa
True	True	True	False	Versicolor
True	True	False	True	Setosa
...	...	...	...	...

How many possible hypotheses?

4 binary features: ___________ different feature combinations

3 classes and (Setosa, Versicolor, Virginica) and _____________ rules,

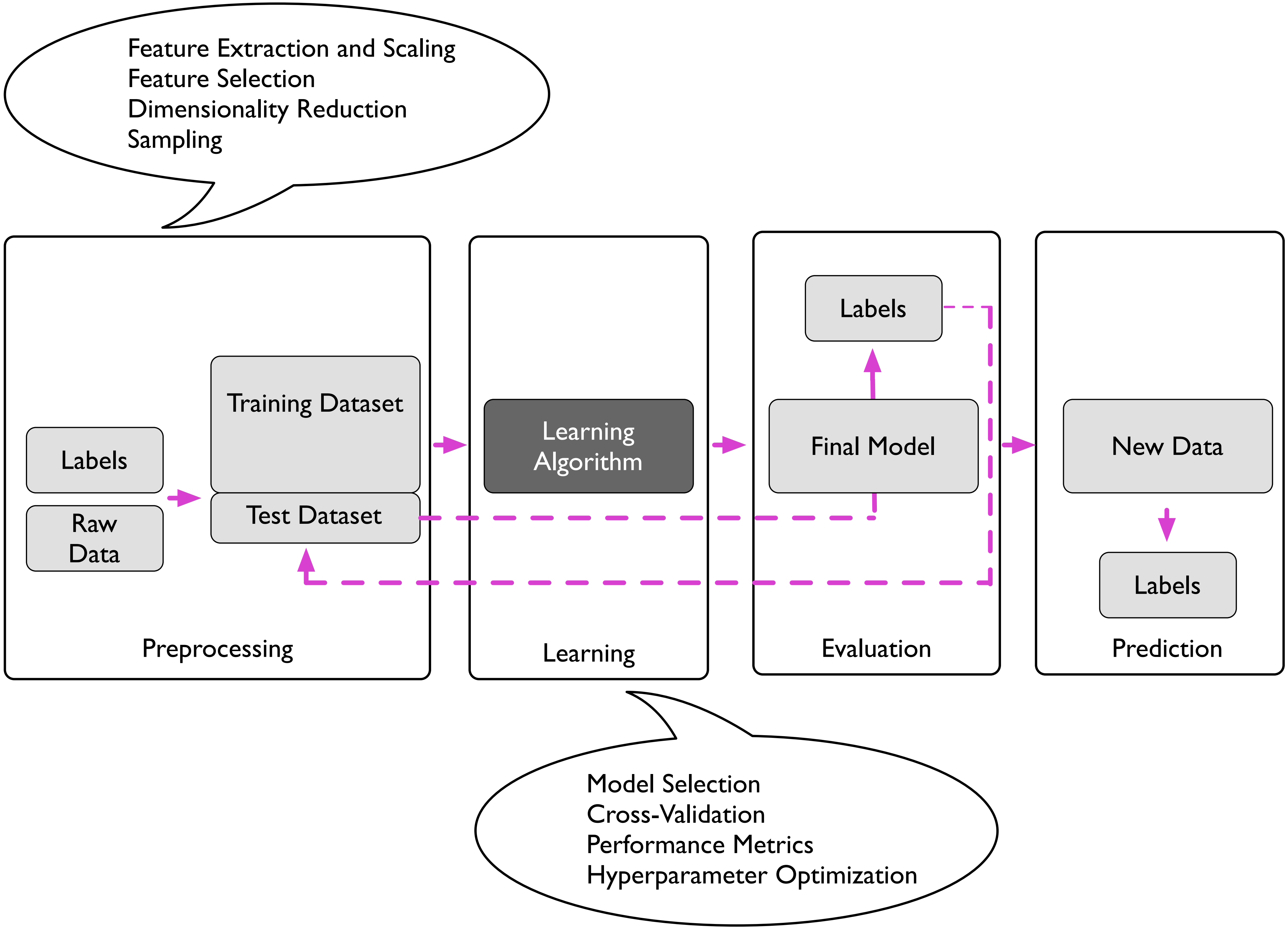
that is ________________ potential combinations

Classes of Machine Learning Algorithms

- Generalized linear models (e.g.,
- Support vector machines (e.g.,
- Artificial neural networks (e.g.,
- Tree- or rule-based models (e.g.,
- Graphical models (e.g.,
- Ensembles (e.g.,
- Instance-based learners (e.g.,

5 Steps for Approaching a Machine Learning Application

1. Define the problem to be solved.
2. Collect (labeled) data.
3. Choose an algorithm class.
4. Choose an optimization metric for learning the model.
5. Choose a metric for evaluating the model.


Feature Extraction and Scaling
Feature Selection
Dimensionality Reduction
Sampling
Labels
Raw Data
Training Dataset
Test Dataset
Preprocessing
Learning Algorithm
Learning
Labels
Final Model
Evaluation
New Data
Labels
Prediction
Model Selection
Cross-Validation
Performance Metrics
Hyperparameter Optimization

Objective Functions

- Maximize the posterior probabilities (e.g., naive Bayes)
- Maximize a fitness function (genetic programming)
- Maximize the total reward/value function (reinforcement learning)
- Maximize information gain/minimize child node impurities (CART decision tree classification)
- Minimize a mean squared error cost (or loss) function (CART, decision tree regression, linear regression, adaptive linear neurons, ...)
- Maximize log-likelihood or minimize cross-entropy loss (or cost) function
- Minimize hinge loss (support vector machine)

Optimization Methods

- Combinatorial search, greedy search (e.g.,
- Unconstrained convex optimization (e.g.,
- Constrained convex optimization (e.g.,
- Nonconvex optimization, here: using backpropagation, chain rule, reverse autodiff. (e.g.,
- Constrained nonconvex optimization (e.g.,

Evaluation -- Misclassification Error

$$L(\hat{y}, y) = \begin{cases} 0 & \textbf{if } \hat{y} = y \\ 1 & \textbf{if } \hat{y} \neq y \end{cases}$$

$$ERR_{\mathscr{D}_{\textbf{test}}} = \frac{1}{n} \sum_{i=1}^{n} L\left(\hat{y}^{[i]}, y^{[i]}\right)$$

Other Metrics in Future Lectures

- Accuracy (1-Error)
- ROC AUC
- Precision
- Recall
- (Cross) Entropy
- Likelihood
- Squared Error/MSE
- L-norms
- Utility
- Fitness
- ...

But more on other metrics in future lectures.

Categorizing Machine Learning Algorithms

- eager vs lazy;

Categorizing Machine Learning Algorithms

- eager vs lazy;
- batch vs online;

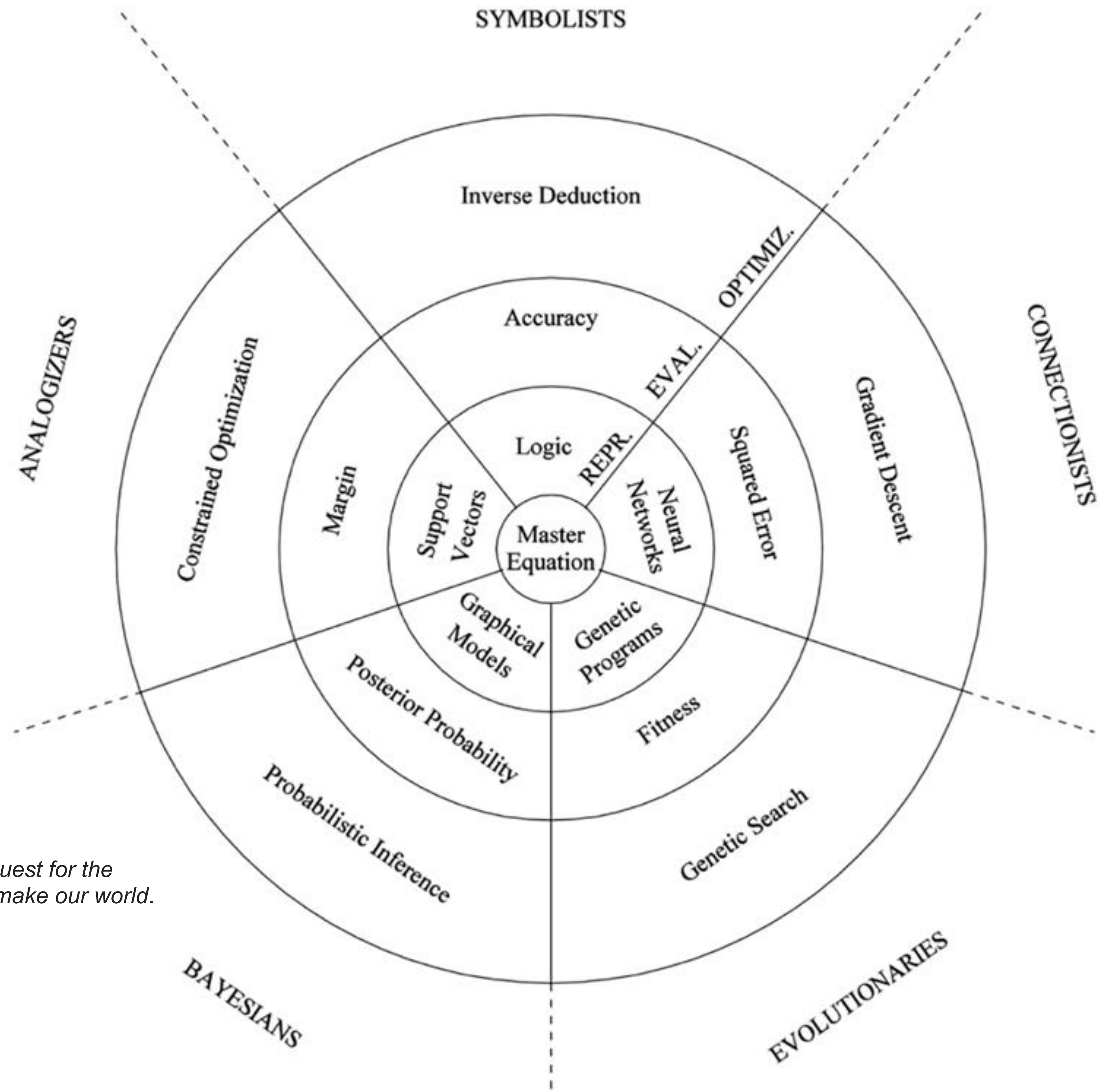
Categorizing Machine Learning Algorithms

- eager vs lazy;
- batch vs online;
- parametric vs nonparametric;

Categorizing Machine Learning Algorithms

- eager vs lazy;
- batch vs online;
- parametric vs nonparametric;
- discriminative vs generative.

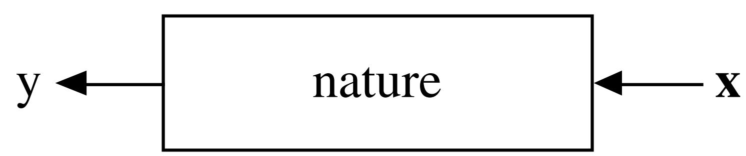
Pedro Domingo's 5 Tribes of Machine Learning



Source: Domingos, Pedro.
The master algorithm: How the quest for the ultimate learning machine will remake our world.
Basic Books, 2015.

Breiman, Leo. "Statistical modeling: The two cultures (with comments and a rejoinder by the author).
" *Statistical science* 16.3 (2001): 199-231.

A



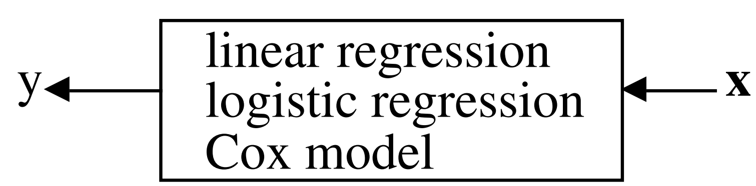
There are two goals in analyzing the data:

Prediction. To be able to predict what the responses are going to be to future input variables;
Information. To extract some information about how nature is associating the response variables to the input variables.

Breiman, Leo. "Statistical modeling: The two cultures (with comments and a rejoinder by the author). " *Statistical science* 16.3 (2001): 199-231.

B

The values of the parameters are estimated from the data and the model then used for information and/or prediction. Thus the black box is filled in like this:

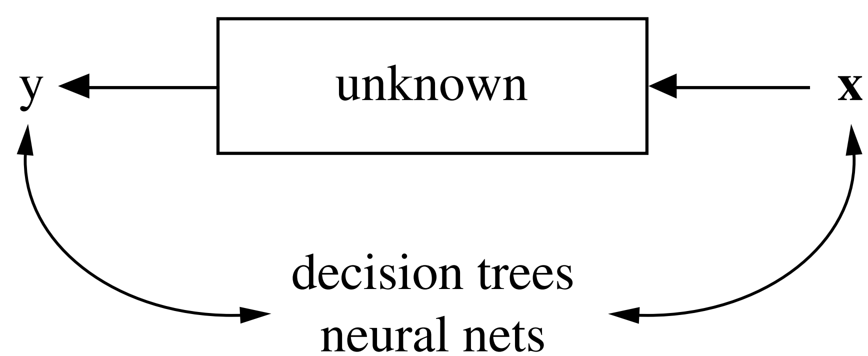


Model validation. Yes–no using goodness-of-fit tests and residual examination.

Breiman, Leo. "Statistical modeling: The two cultures (with comments and a rejoinder by the author). " *Statistical science* 16.3 (2001): 199-231.

C

The analysis in this culture considers the inside of the box complex and unknown. Their approach is to find a function $f(\mathbf{x})$—an algorithm that operates on $\mathbf{x}$ to predict the responses $\mathbf{y}$. Their black box looks like this:



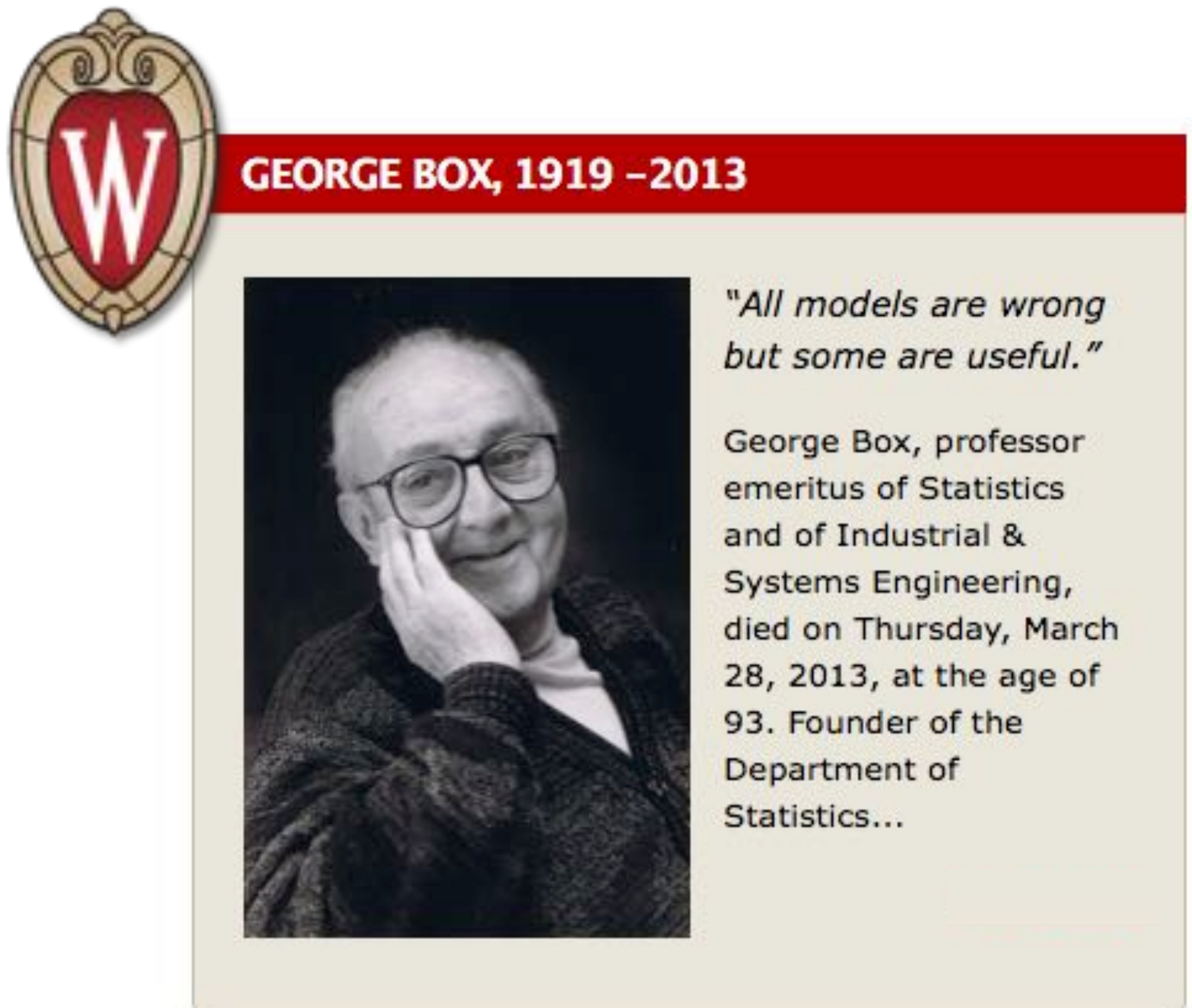
Model validation. Measured by predictive accuracy.


0
1
2 cm

Evolved antenna (Source: https://en.wikipedia.org/wiki/Evolved_antenna) via evolutionary algorithms; used on a 2006 NASA spacecraft.

Black Boxes vs Interpretability

Black Boxes vs Interpretability



Different Motivations for Studying Machine Learning

- Engineers:

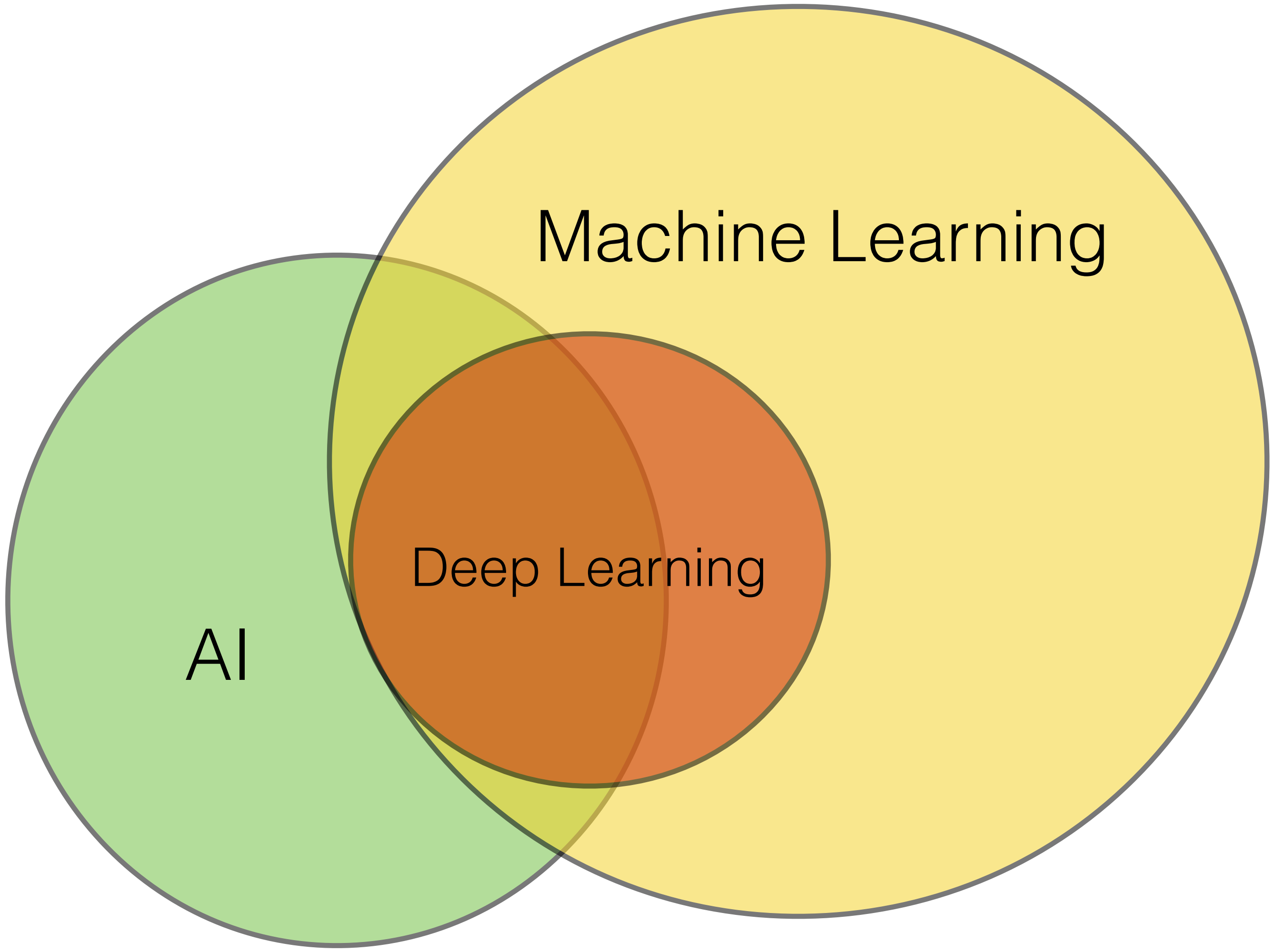
- Mathematicians, computer scientists, and statisticians:

- Neuroscientists:

The Relationship between Machine Learning and Other Fields

Machine Learning and Data Mining

Machine Learning, AI, and Deep Learning



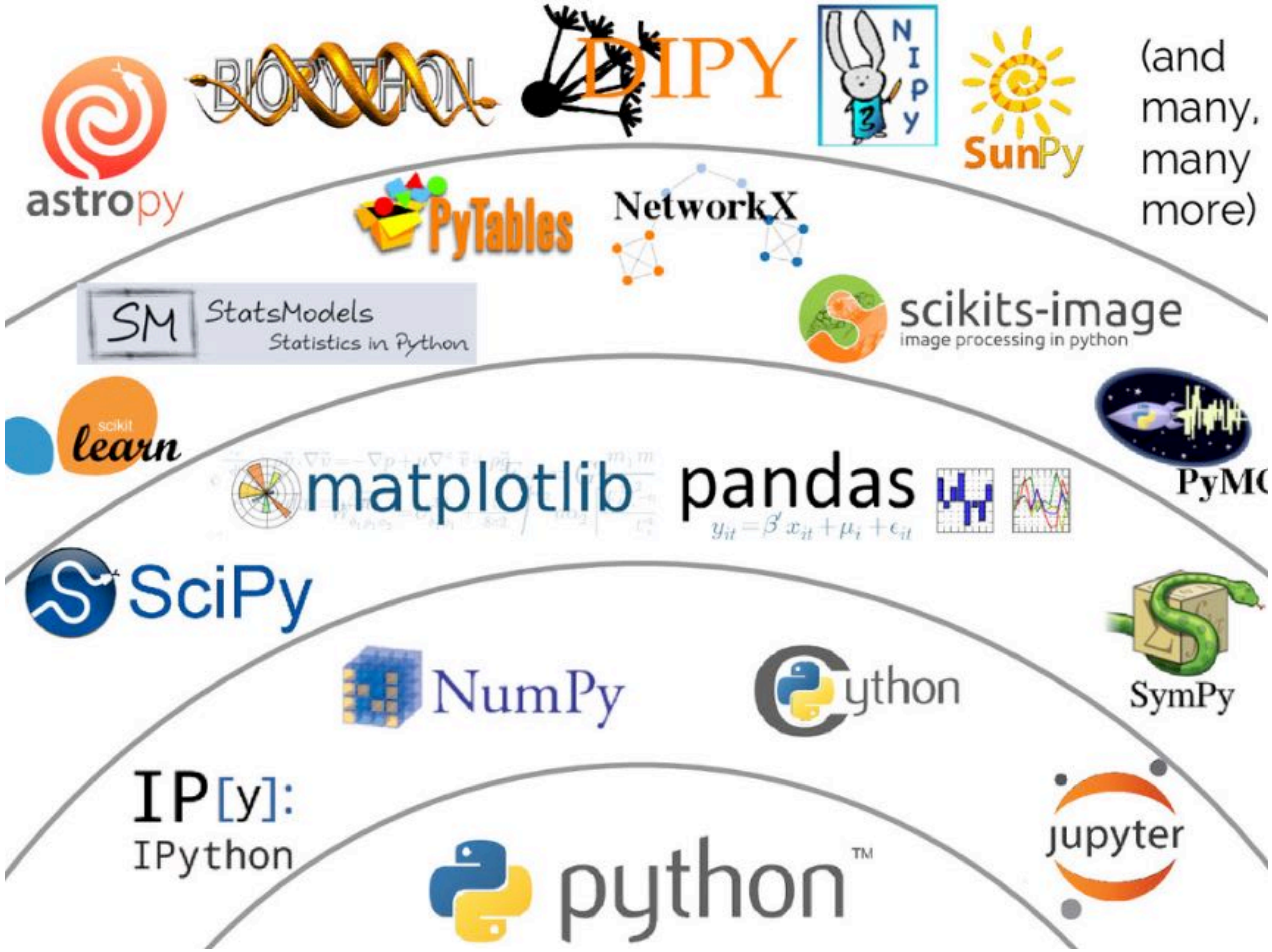


Image by Jake VanderPlas; Source:
https://speakerdeck.com/jakevdp/the-state-of-the-stack-scipy-2015-keynote?slide=8)

TIOBE Index for September 2018

Sep 2018	Sep 2017	Change	Programming Language	Ratings	Change
1	1		Java	17.436%	+4.75%
2	2		C	15.447%	+8.06%
3	5	↑	Python	7.653%	+4.67%
4	3	↓	C++	7.394%	+1.83%
5	8	↑	Visual Basic .NET	5.308%	+3.33%
6	4	↓	C#	3.295%	-1.48%
7	6	↓	PHP	2.775%	+0.57%
8	7	↓	JavaScript	2.131%	+0.11%
9	-	↑↑	SQL	2.062%	+2.06%
10	18	↑↑	Objective-C	1.509%	+0.00%
11	12	↑	Delphi/Object Pascal	1.292%	-0.49%
12	10	↓	Ruby	1.291%	-0.64%
13	16	↑	MATLAB	1.276%	-0.35%
14	15	↑	Assembly language	1.232%	-0.41%
15	13	↓	Swift	1.223%	-0.54%
16	17	↑	Go	1.081%	-0.49%
17	9	↓↓	Perl	1.073%	-0.88%
18	11	↓↓	R	1.016%	-0.80%
19	19		PL/SQL	0.850%	-0.63%
20	14	↓↓	Visual Basic	0.682%	-1.07%

Programming language "popularity"

https://www.tiobe.com/tiobe-index/

https://www.tiobe.com/tiobe-index/programming-languages-definition/

Roadmap for this Course

http://stat.wisc.edu/~sraschka/teaching/stat479-fs2018/#schedule

Reading Assignments

- Raschka and Mirjalili: Python Machine Learning, 2nd ed., Ch 1
- Elements of Statistical Learning, Ch 01 (https://web.stanford.edu/~hastie/ElemStatLearn/)