ML for Smart Monkeys

Amit Sethi

Faculty member, IIT Bombay

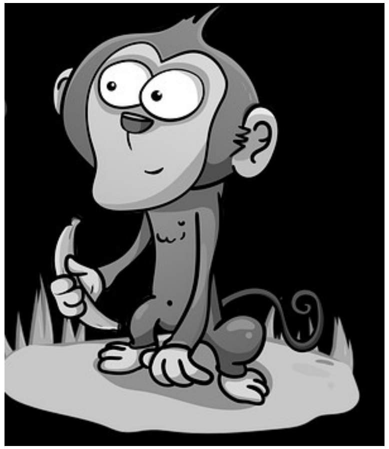


Image source: Pixabay.com

ML is…

- The practice of automating the use of related data to estimate models that make useful predictions about new data, where the model is too complex for standard statistical analysis, e.g.
  - Improve accuracy of classification of images using labeled images
  - Improve win percentage on alpha-go using several simulated game move sequences and their results
  - Improve the Turing test confusion between human and machine for NLP Q&A using a large sample of text including Q&A

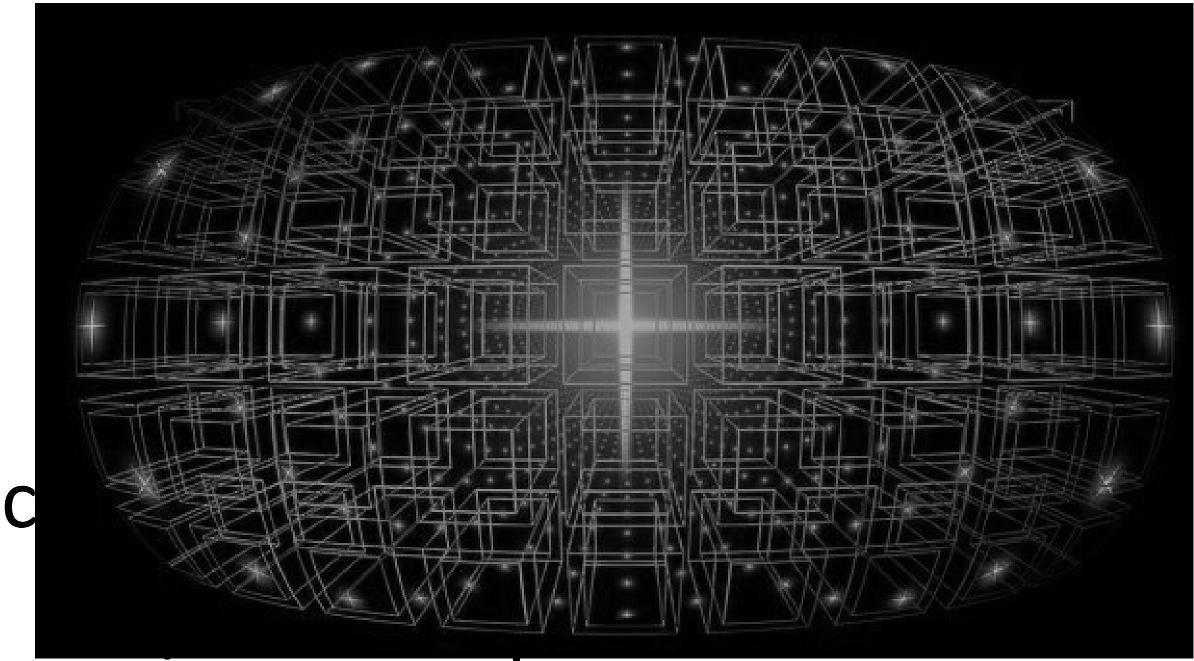
When not to use ML

- Possible inputs are countable and few
  - Use look up tables
- Algorithm is well-known and efficient
  - E.g. sorting, Dijkstra's shortest path
- Model is well-known and tractable
  - Use statistical estimation
- There is no notion of contiguity
  - Use dicrete variable methods or give up
- Lack of data
  - Use transfer learning or few-shot learning, or give up

When to use ML

- Possible inputs are many or continuous
- No well-known or efficient algorithm
- Model is not well-known or tractable
- Strong notion of contiguity
- Good amount of data
- Desired output known
- Well-defined inputs

Sweet spot for ML

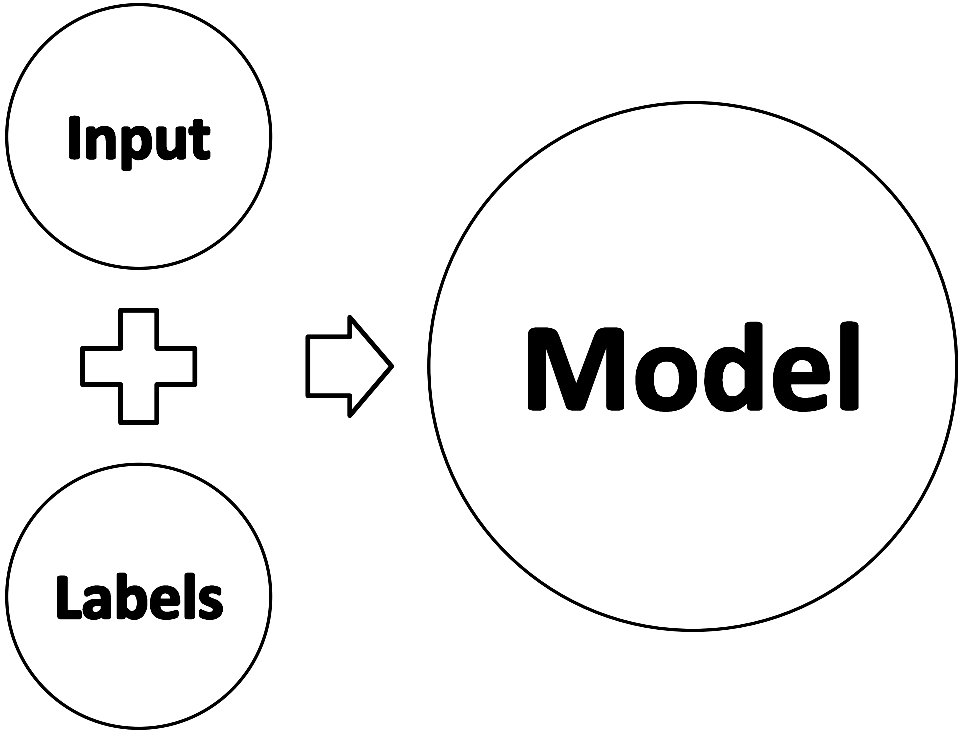


- Lots of structured data
- Explainability is not critic
- Prediction accuracy is the primary goal
- Underlying model is complex but stationary

Image courtesy: Pixabay.com

ML model training and deployment

Training on past data

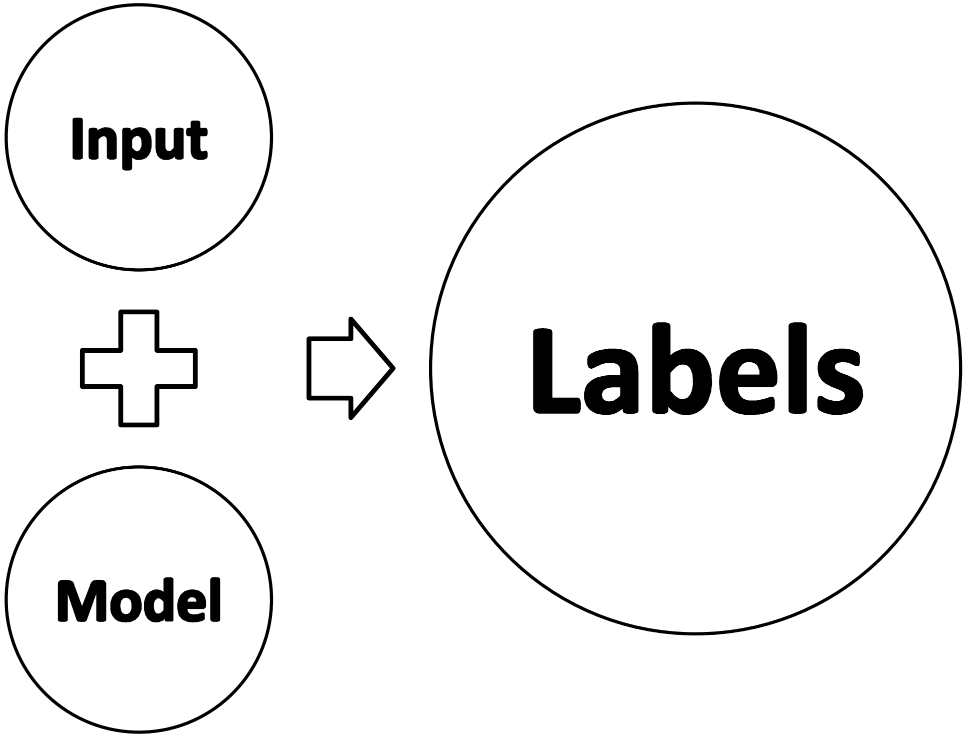


Input

Labels

Model

Prediction on future data



Input

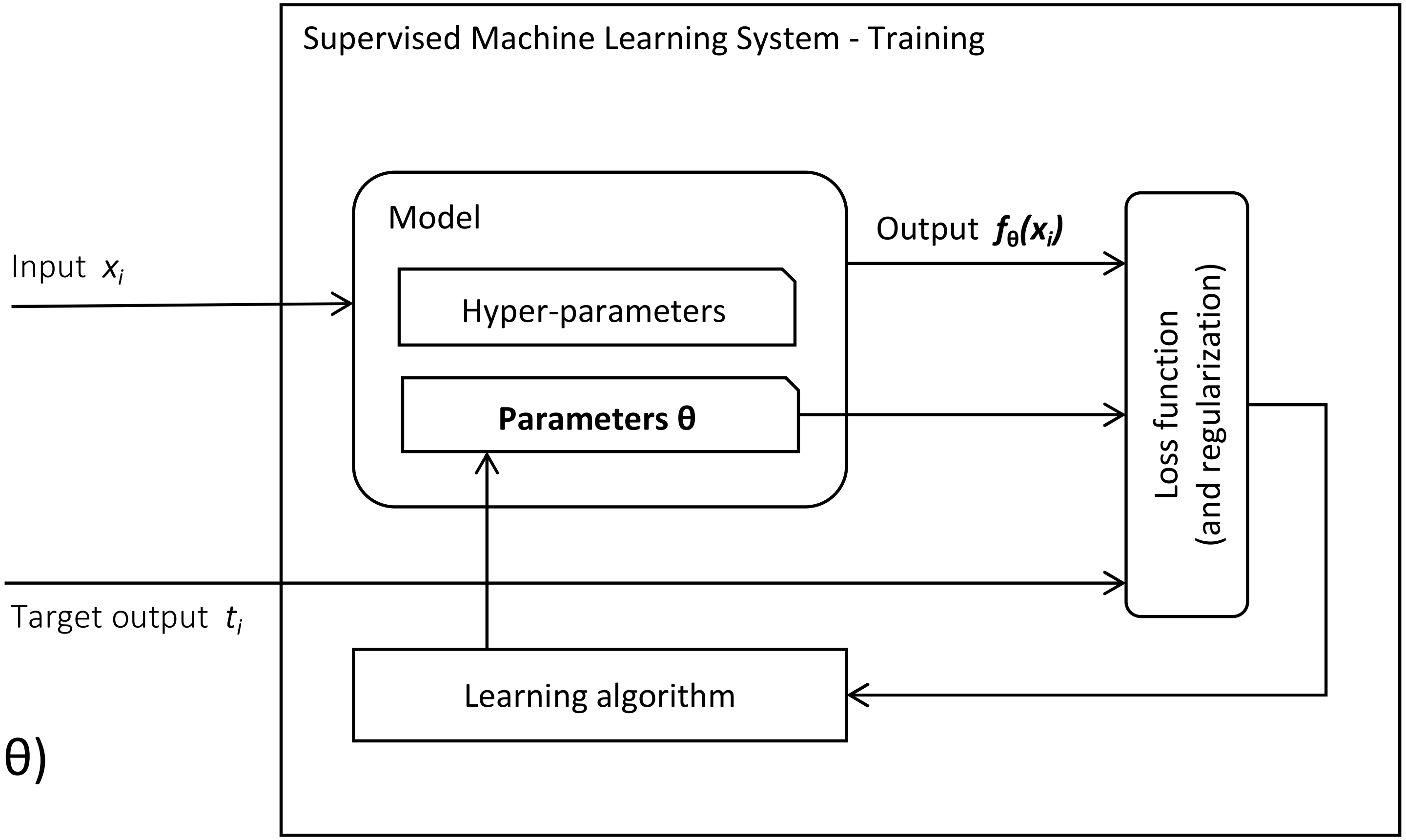
Model

Labels

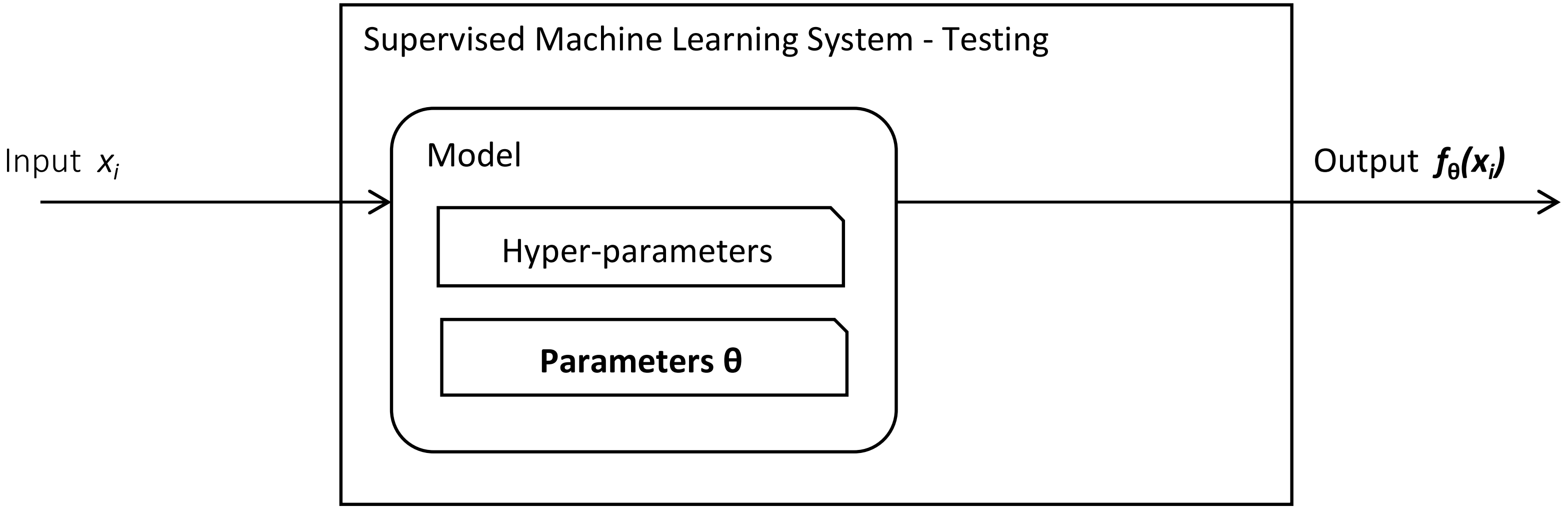
ML gives a model

- Elements of a model:
  - Input x_i
  - Function $f_\theta(x_\mathrm{i})$

- Utility of the model:
  - Target output t_i
  - Bring $f_\theta(x_\mathrm{i})$ close to t_i
  - Minimize loss $L(t_i, f_\theta(x_\mathrm{i}), \theta)$



Components of a Trained ML System


Supervised Machine Learning System - Testing
Input x_i
Model
Hyper-parameters
Parameters θ
Output $\boldsymbol{f_{\theta}(x_i)}$

Mathematically speaking…

- Determine f such that $t_i = f(x_i)$ and $g(T, X)$ is minimized for unseen set T and X pairs, where T is the ground truth that cannot be used
- Form of f is fixed, but some parameters can be tuned:
  - So, $y=f_\theta(x)$, where, x is observed, and y needs to be inferred
  - e.g. $y = 1$, if $mx > c$, $y = 0$ otherwise, so $\theta = (m,c)$
- Machine Learning is concerned with designing algorithms that learn "better" values of θ given "more" x (and t) for a given problem

Key Concept

Parameters and Hyperparameters

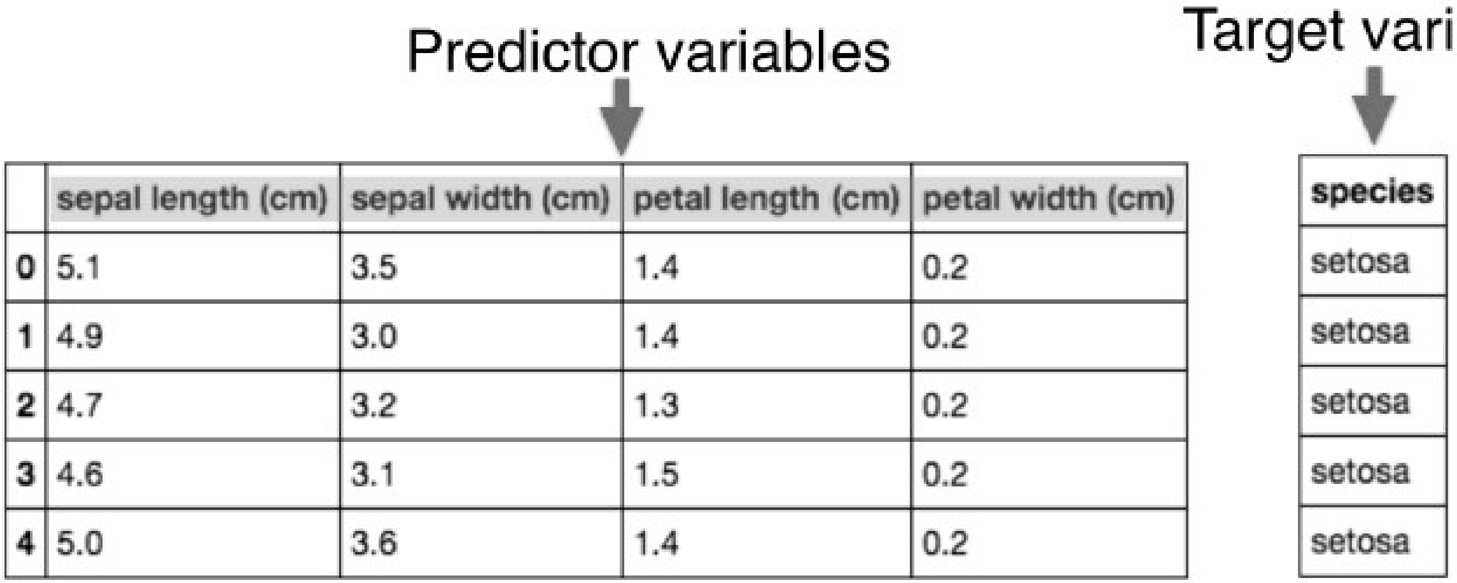
- Parameters: These are the variable whose values are updated during the training process of model.
  - Feature coefficient in regression model
  - Weights of a neural network
- Hyperparameters: These are the variables/ parameter whose values are fixed by model developer before the beginning of learning process.
  - Number of variables in a tree node
  - Height of a tree
  - Number of layers of a neural network

Type of ML problems

- Supervised learning: uses labeled data
  - Classification: Labels are discrete
  - Regression: Labels are continuous
  - Ranking: Labels are ordinal
- Unsupervised learning: uses unlabeled data
  - Clustering: Divide data into discrete groups
  - Dimension reduction: Represent data with fewer numbers
- Somewhere in between: fewer labels than one per example
  - Semi-supervised learning: some examples are labeled
  - Weakly supervised learning: groups of examples are labeled
  - Reinforcement learning: Label (reward) is available after a sequence of steps

Supervised Learning

- Predictor variables/features and a target variable (label)
- Aim: Predict the target variable (label), given the predictor variables
  - ○ **Classification**: Target variable (y) consists of categories
  - ○ **Regression**: Target variable is continuous



Predictor variables

Target variable (Label)

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)
0	5.1	3.5	1.4	0.2
1	4.9	3.0	1.4	0.2
2	4.7	3.2	1.3	0.2
3	4.6	3.1	1.5	0.2
4	5.0	3.6	1.4	0.2

species
setosa
setosa
setosa
setosa
setosa

Broad types of ML problems

Output →	**Categorical**	**Ordinal**	**Continuous**
Supervised	Classification	Ranking	Regression
(Examples)	{Cats, dogs}	{Low, Med, High}	[-20,+10)
Unsupervised	Clustering		Dimension reduction

Some popular ML frameworks

	Classification	Regression	Clustering	Dimension reduction
Vector	Logistic regression	Linear regression	K-means, Fuzzy C-means, DB-SCAN	PCA, k-PCA, LLE, ISOMAP
	SVM, RF, NN			
Series, text	RNN, LSTM, Transformer, 1-D CNN, HMM			
Images	2-D CNN, MRF			
Video, MRI	3-D CNN, CNN+LSTM, MRF			

Recipe for ML training

- Decide on the type of the ML problem
- Prepare data
- Shortlist ML frameworks
- Prepare training, validation, and test sets
- Train, validate, repeat
- Use test data only once

Preparing data

- Remove useless data
    - No variance
    - Falsely assumed to be available
- Reduce redundancy
    - Correlated
        - Pearson and Spearman

- Handle missing data
    - Impute, if sporadic
    - Drop, if too frequent
- Transform variables
    - Convert discrete to one-hot-bit
    - Normalize continuous variables

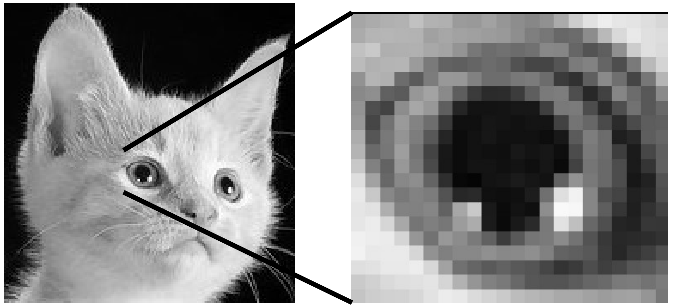
Examples of structure in the data

- Records

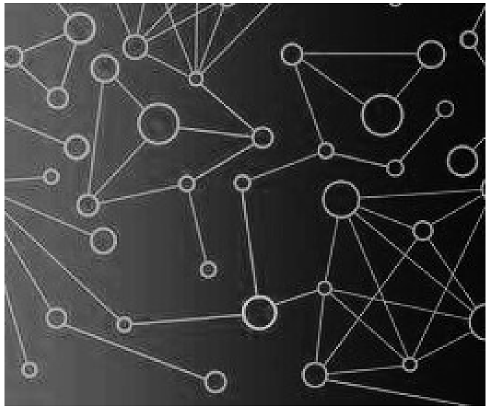
Product SKU	Price	Margin	Volume
A123ajkhdf	$ 120	30%	1,000,000
B456ddsjh	$200	10%	2,000,000

- Temporal order

- Spatial order

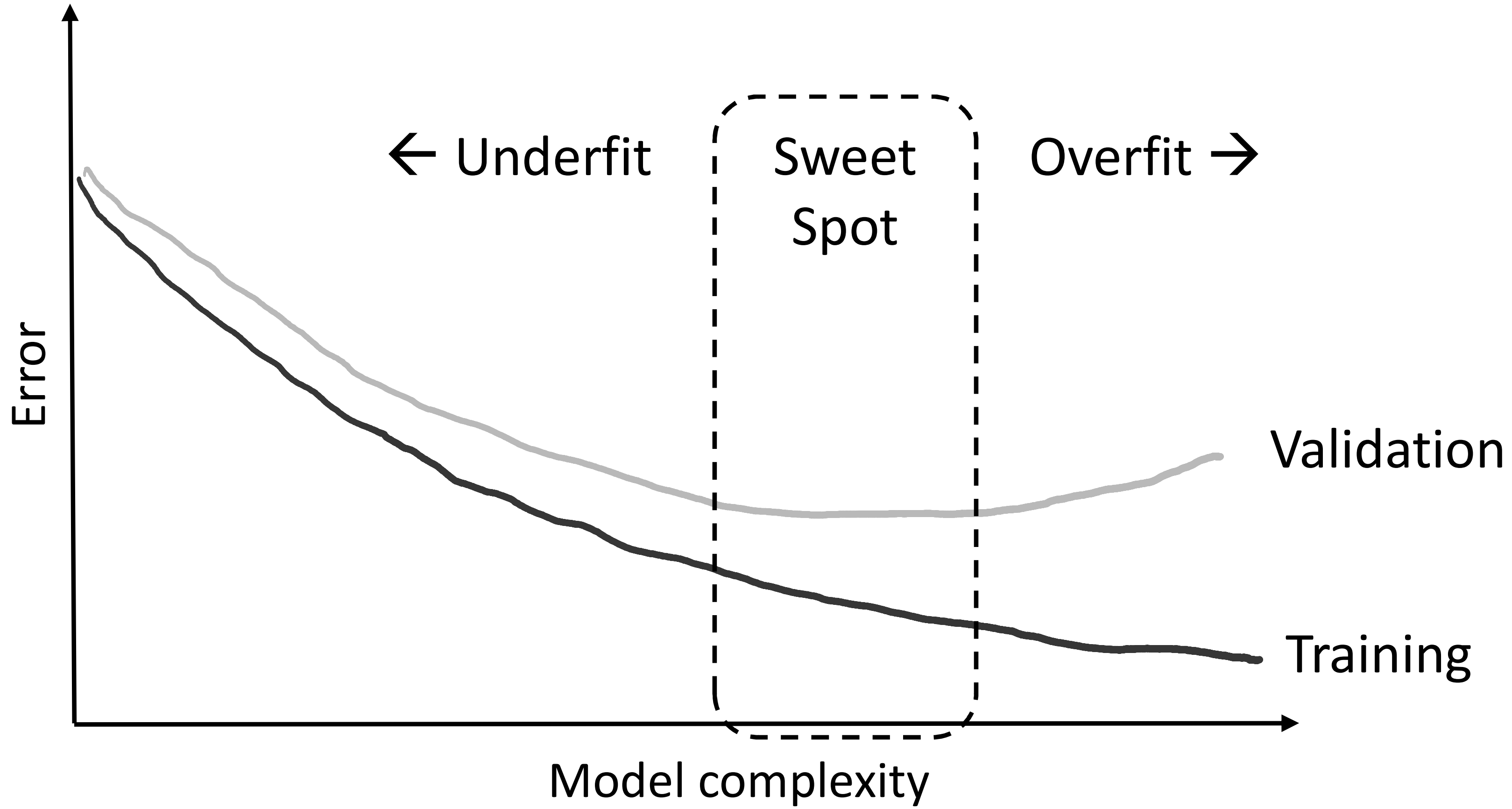


- Web of relationships



Images courtesy: Pixabay.com

Model choice and rigorous validation are very important



Bias-variance trade-off

Generalization of model is bounded by the two undesirable outcomes high bias and high variance.

- Underfitting: High bias, Low variance
- Overfitting: Low bias, High variance

Bias occurs when an algorithm has limited flexibility to learn the true signal from the dataset. High bias can cause an algorithm to miss the relevant relations between features and target outputs (underfitting).

Variance is an error from sensitivity to small fluctuations in the training set. High variance can cause an algorithm to model the random noise in the training data, rather than the intended outputs (overfitting).

Regularization is a key concept in ML

- Regularization means constraining the model
- More constraints may reduce model fit on training data
- However, it may improve fit on validation and test data
- Training performance of more constrained models are more likely to reflect test performance

Loss versus performance metric

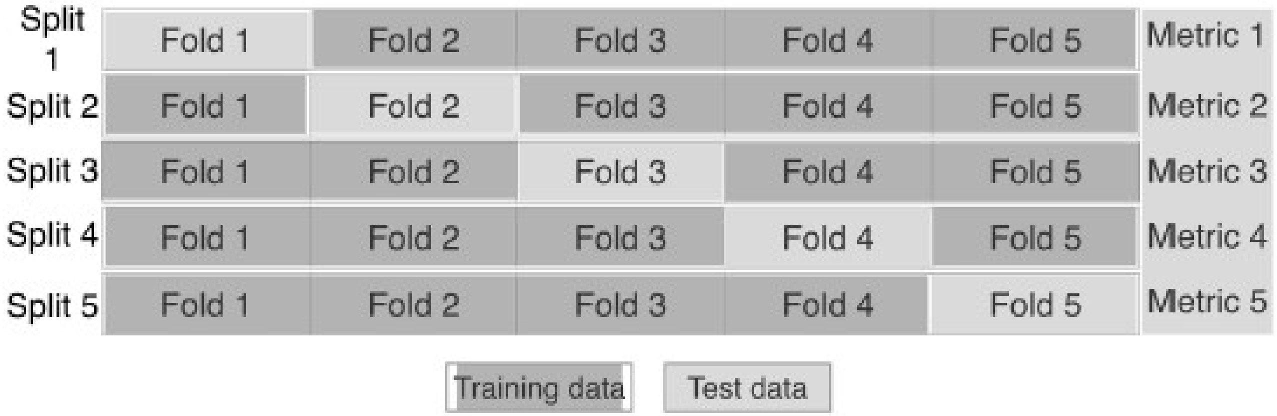
- Loss is a convenient expression used for guiding the learning (optimization)
- Loss is related to performance metric, **but** it is not the same
- Loss also includes regularization
- Performance metric is what is used to judge the model
- Performance metric on only the held-out (validation or test) data makes sense

Preparing data for training and validation

- Data splits:
  - Training → Used to optimize the parameters (e.g. random 70%)
  - Validation → Used to compare models (e.g. random 15%)
  - Testing → One final check after multiple rounds of validation (e.g. random 15%)
- Cross-validation:
  - K-folds: One fold for validation, K-1 folds for training
  - Rotate folds K times
  - Select framework (hyperparameters) best average performance
  - Re-train best framework on entire data
  - Test one final time on held-out data that was not a part of any fold

Cross-validation

- Model performance measurement is dependent on way the data is split
- Not representative of the model's ability to generalize
- Solution: Cross-validation, especially when data is less
- Con: more computations



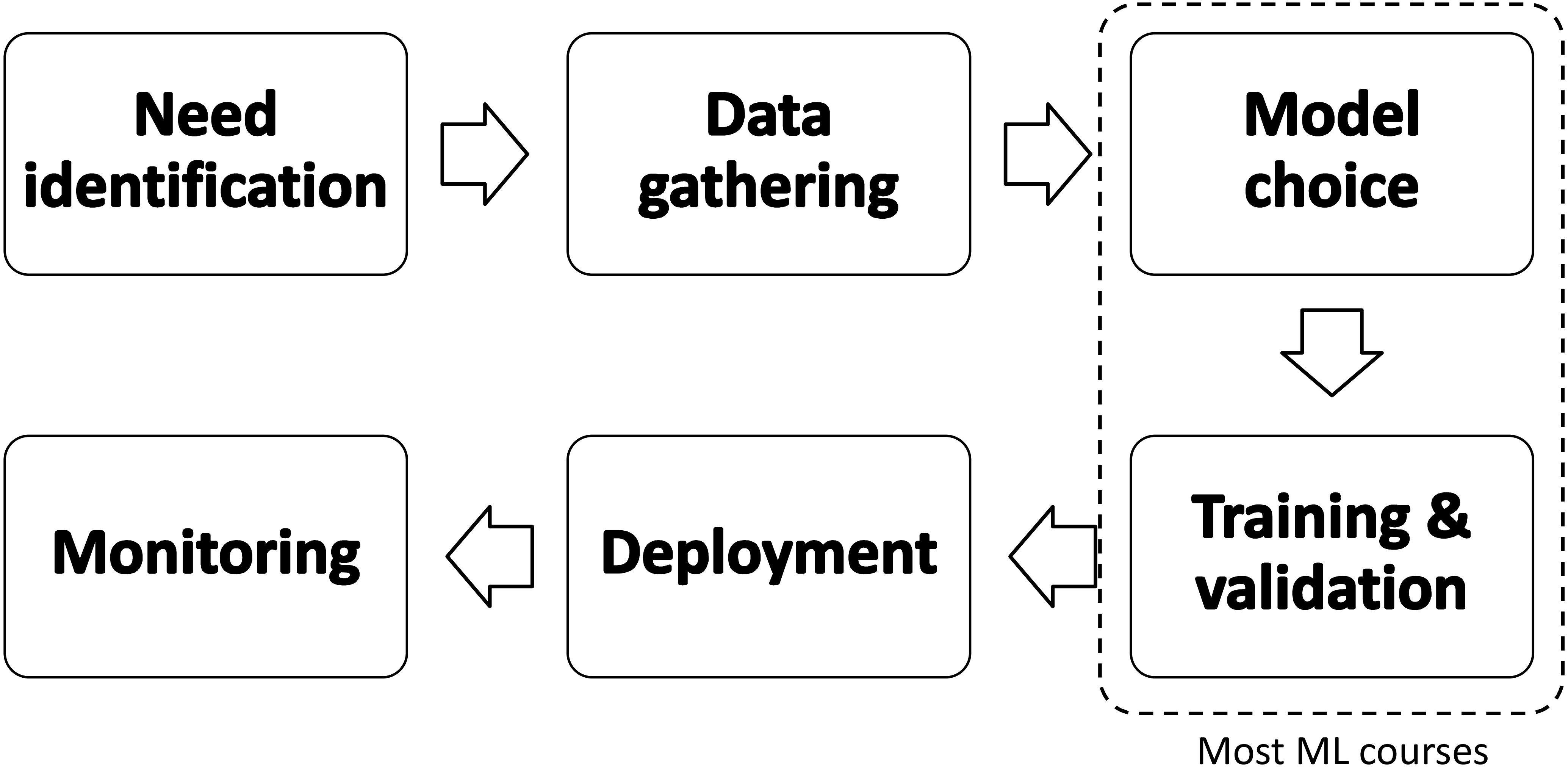
Split 1	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Metric 1
Split 2	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Metric 2
Split 3	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Metric 3
Split 4	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Metric 4
Split 5	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Metric 5

Training data Test data

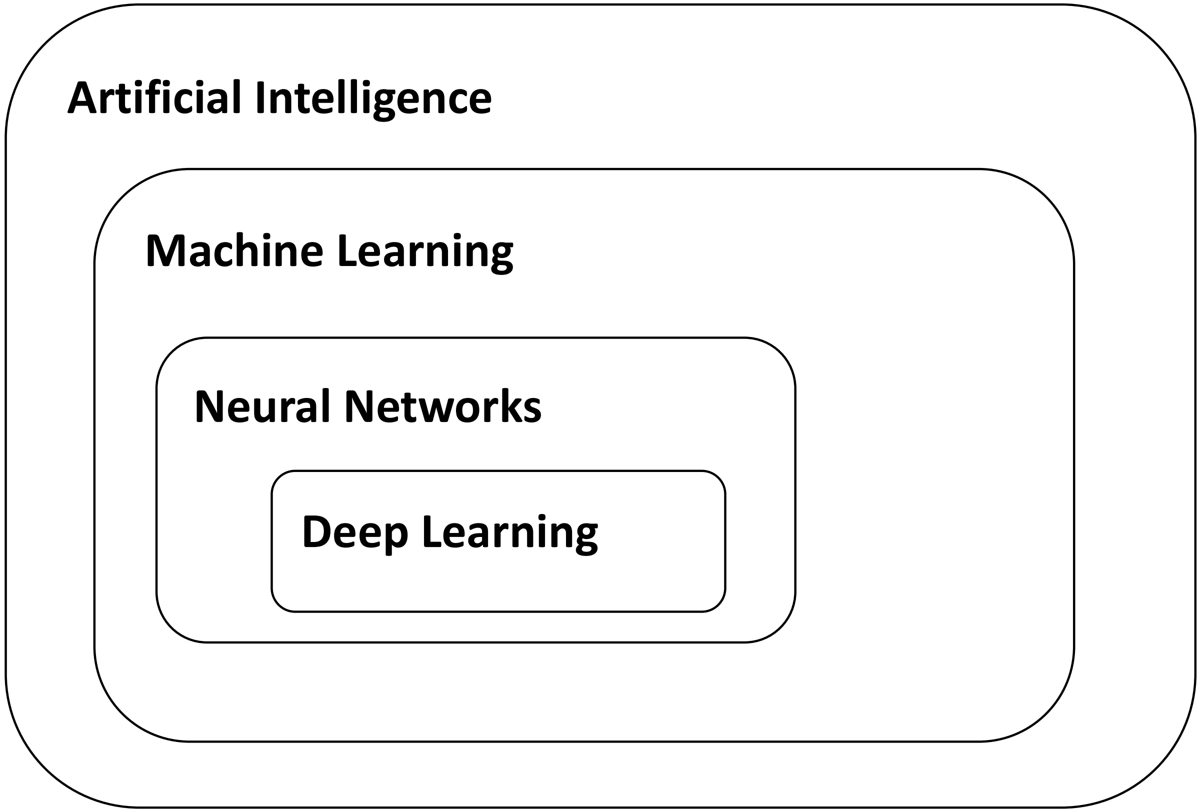
ML can fail to perform in deployment

- Lack of training diversity: data had limited confounders
  - Single speaker, author, camera, background, accent, ethnicity, etc.
  - Data imbalance between high-value rare and more common examples
- Proxy label leak during training:
  - E.g. Only speakers A and B provide emotion “anger,” so ML confused their voice characteristics with “anger”
- Too much manual cleansing of training data
- Too little training data, and very complex models
- Concept drift: The assumptions behind training are no longer valid

ML life stages



Relation of ML to other fields



Artificial Intelligence

Machine Learning

Neural Networks

Deep Learning

Relation of ML to other fields

Machine Learning

Probability and Statistics

Optimization

Linear Algebra

Programming Data Science