Machine Learning in Security: Attack Spaces & Defense

Yamuna Prasad
IIT Jammu

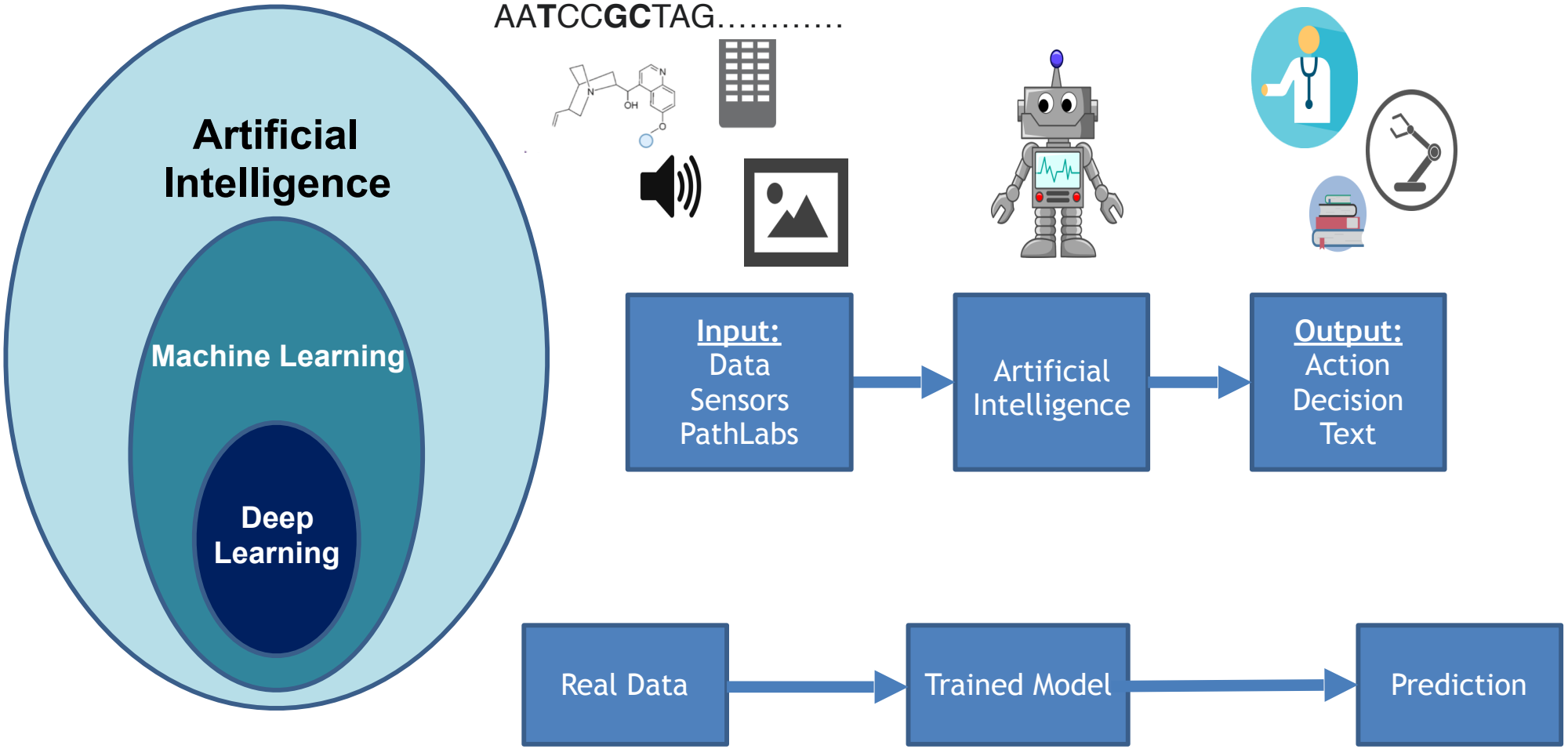
yamuna.prasad@iitjammu.ac.in

IIT JAMMU

Outline

- Introduction : Process Involved in ML (or AI Model)
- Sources of Data, Types and Challenges
- Problem Types
- Learning Process: Supervised & Unsupervised
- Approaches & Challenges in Learning Process
- ML in Security
- Attacks Spaces & Attack Types
- Attack Resistive Models
- Demo and Open Discussions

Introduction:Learning Process

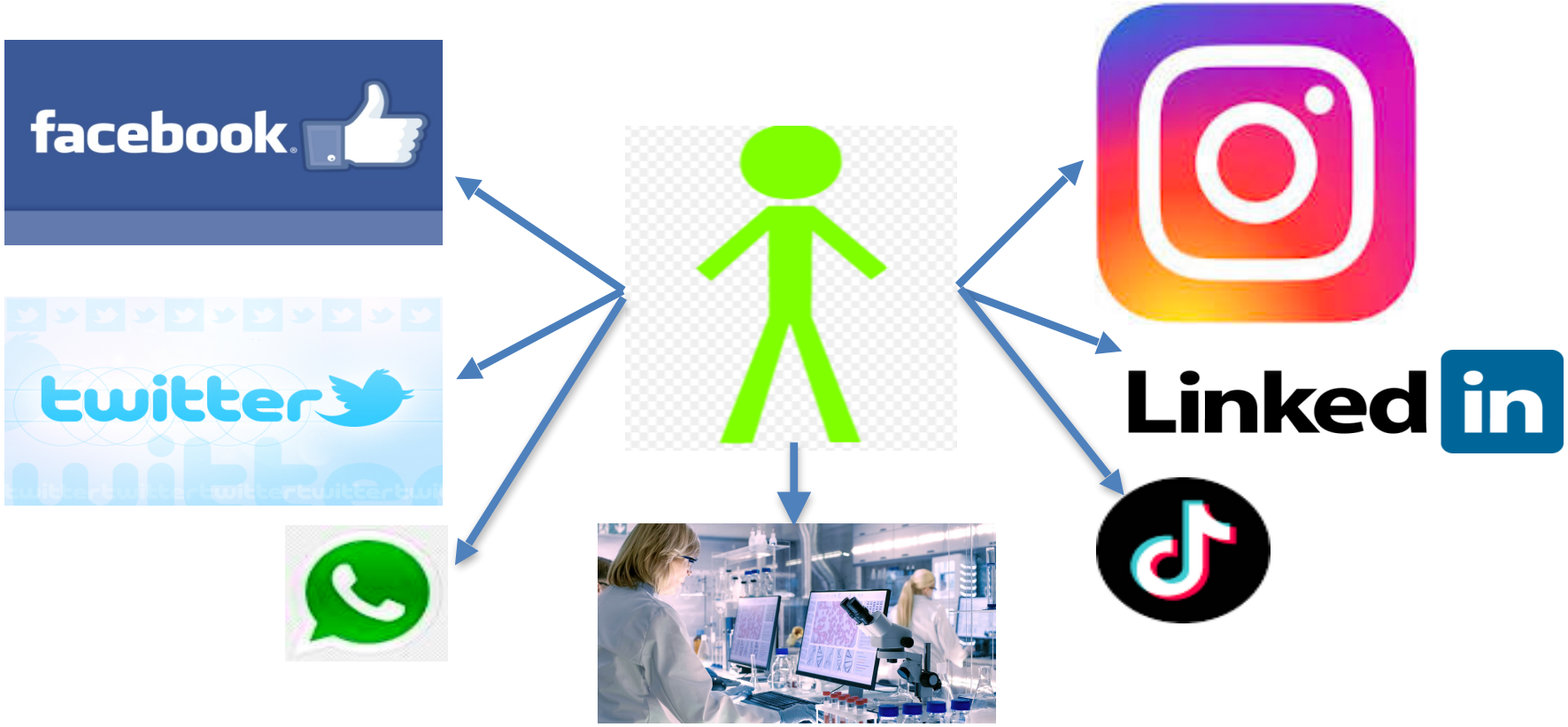


Img Ref: web

Introduction:Learning Process contd…

- Data Cleansing + Data Preparation + Data Analysis
- Data Cleansing: Noise Removal
- Data Preparation: *Missing Value Curation*, (Multi Dimensional) Vector Space Representation, Feature Identification, Feature Extraction, Feature Vector Normalisation, Data Augmentation without bias, Data Visualization
- Data Analysis: Statistical Inferencing, Machine Learning Models and Other Complex Predictive Models

Sources of Data

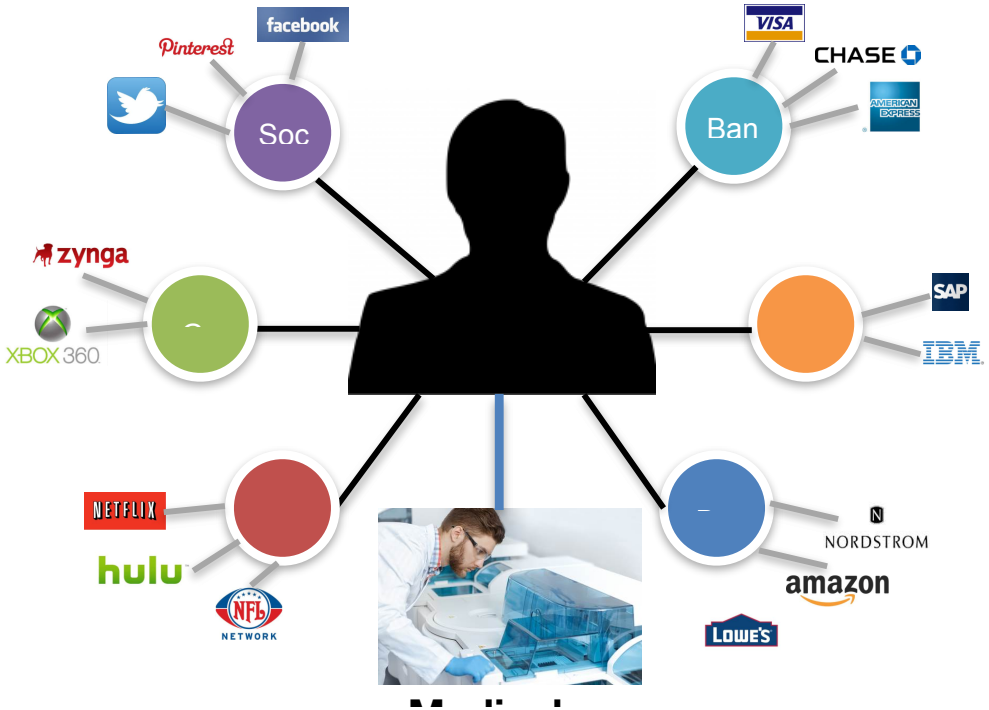


- User Location, Sentiment analysis, Personalisation, Product Recommendations, IoT and Security, Disease Diagnosis, Trend Prediction, Drug Discovery etc.
- Apps and Sensor

Img Ref: web sources & elsevier

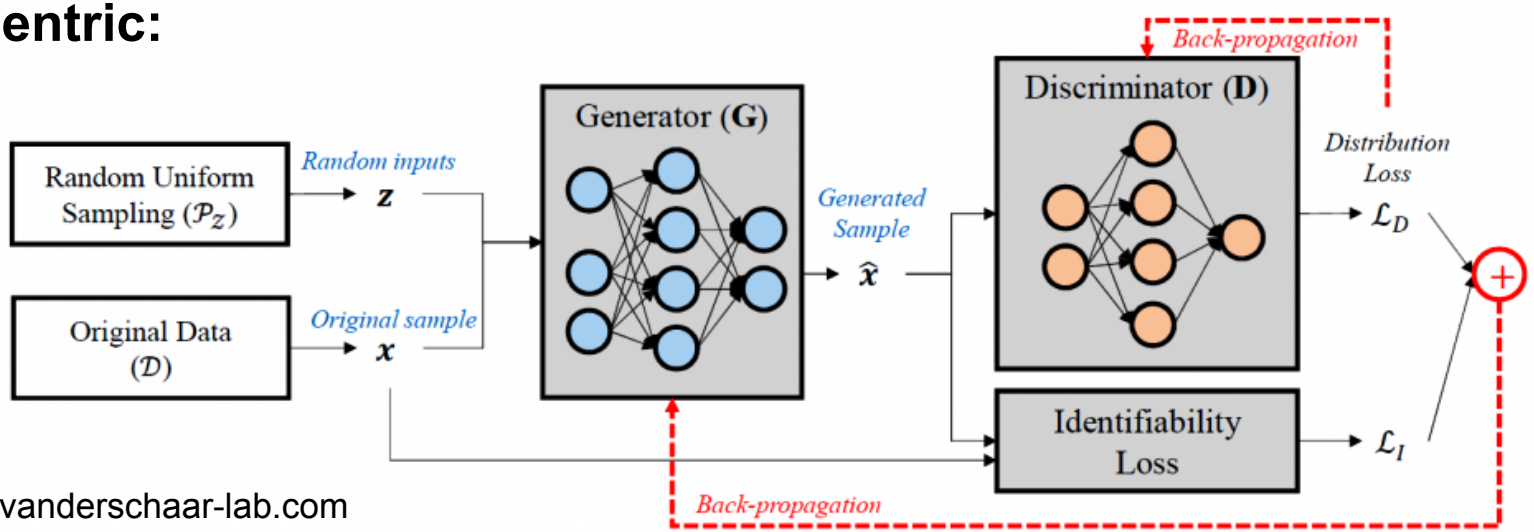
Data Generation: Detailed View

User Centric



Img Ref: web sources

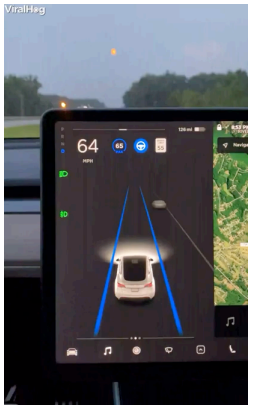
Model Centric:



Img Ref: www.vanderschaar-lab.com

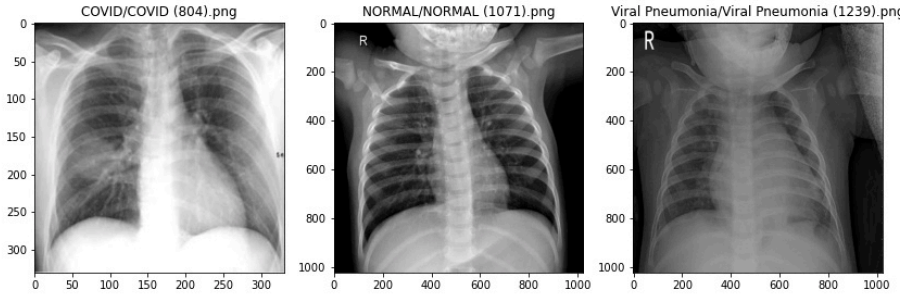
Challenges in Data

- Bioinformatics/Medical
  - Few samples, high dimensions/features etc.
  - Bias due to device noise, human error etc.
- Computer Vision:
  - Noises from sources, multimodality, human bias, IoTs
- Natural Language Processing
  - Noisy data, code-mixed language, discreteness, human bias

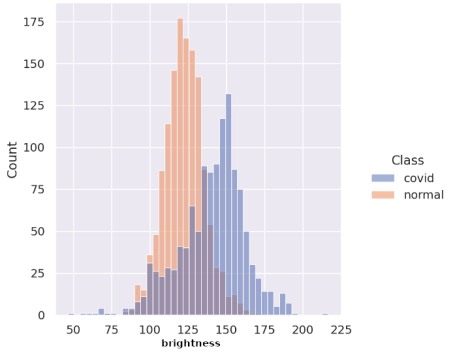


Challenges in Data: Bias Examples

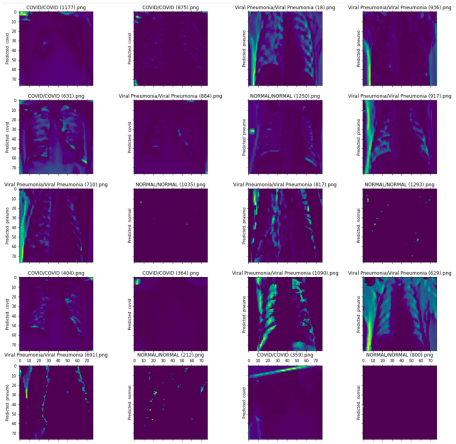
- COVID 19 Xray Images (COVID19 data at Kaggle)



Data



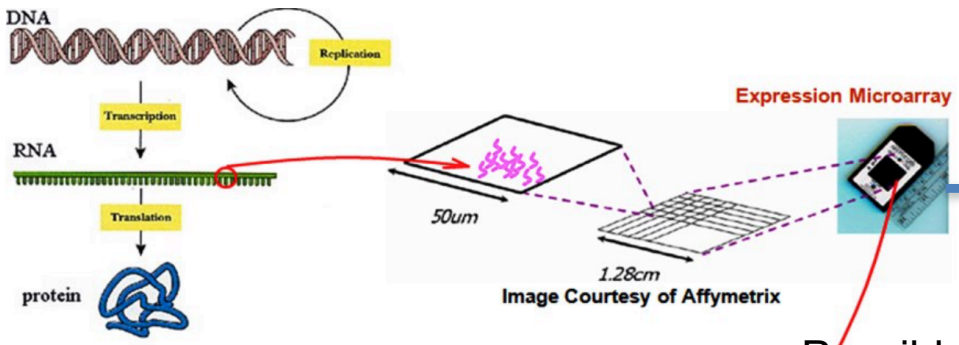
Histogram



Gradient Visualisation

IMg Ref: https://towardsdatascience.com

- Gene Expression Data



Possible Bias

Genes Exp values

Sample \ Gene	M23197_at	U66497_at	M92287_at	...	Class
Sample 1	261	88	4778	...	ALL
Sample 2	101	74	2700	...	ALL
Sample 3	1450	34	498	...	AML
.	.	.	.	...	.
.	.	.	.	...	.
.	.	.	.	...	.

Expression Microarray Data Set

Img Ref: web sources

Types of Data

- Structured Data: Databases (sample and feature (attribute) value pairs
  - Most of the learning Models needs Structured Data for Modelling
- Unstructured Data: Text, Audio, Image and Video
- Semi-Structured: Partially Structured data
- Unimodal/Multimodal: Behavioural

GOAL for Data Analysis: (Un/Semi)-Structured => Structured

IIT JAMMU

Types of Data: Examples

Unstructured data

The university has 5600 students.
John's ID is number 1, he is 18 years old and already holds a B.Sc. degree.
David's ID is number 2, he is 31 years old and holds a Ph.D. degree. Robert's ID is number 3, he is 51 years old and also holds the same degree as David, a Ph.D. degree.

AATCCGCTAG……

Semi-structured data

```
<University>
 <Student ID="1">
  <Name>John</Name>
  <Age>18</Age>
  <Degree>B.Sc.</Degree>
 </Student>
 <Student ID="2">
  <Name>David</Name>
  <Age>31</Age>
  <Degree>Ph.D. </Degree>
 </Student>
 ....
</University>
```

Structured data

ID	Name	Age	Degree
1	John	18	B.Sc.
2	David	31	Ph.D.
3	Robert	51	Ph.D.
4	Rick	26	M.Sc.
5	Michael	19	B.Sc.

Unstructured Data ——> Feature Extraction ——> Vector Space Representation

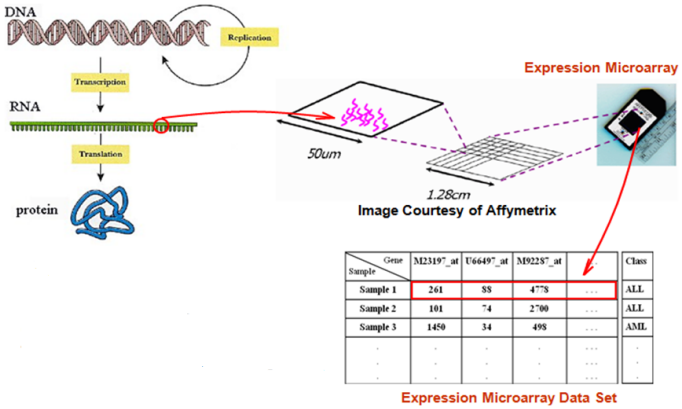
Eg.

Structure (DNA/RNA/Protein) ===> Nucleotide Sequences (enumerate character rep)
Text Data ===> Bag of Words Representation/ One-hot Vector
Image Data ===> Pixel based/feature extraction
Speech/Audio/Signal ===> Feature Extraction

IIT JAMMU

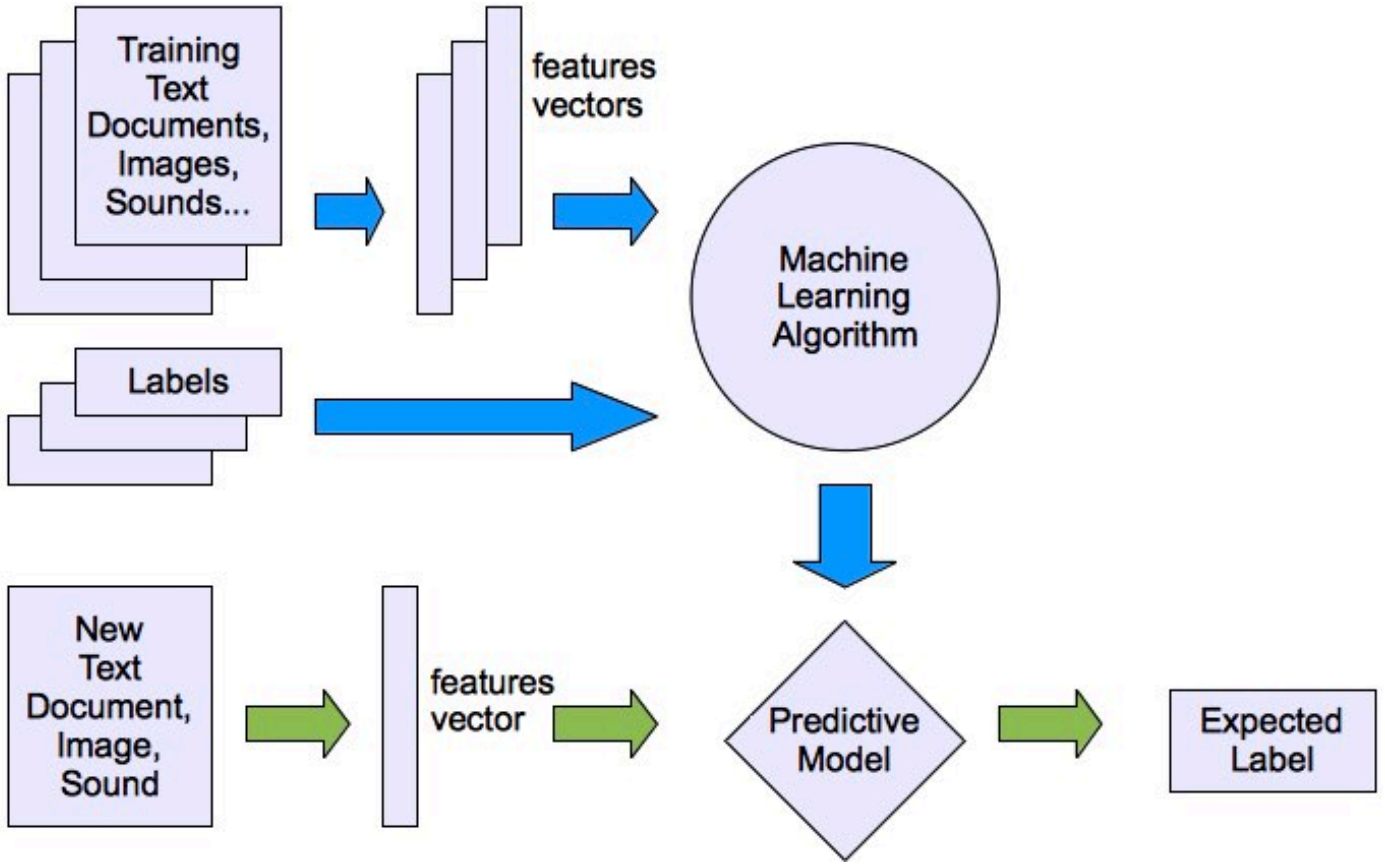
Problem Types

- Classification
  - Disease Diagnosis, Drug Target Identification, Drug Design and Discovery, Microbial Species Prediction,Environmental and Host Phenotypes Pred.,Interaction Pred
  - Intrusion Detection, Spam, Sentiment
  - Emotion, Product Reviews etc.
- Regression
  - Efficacy prediction
- Other problems in Various domains
  - Rankings
  - Translation
  - Summarisation etc.



Task: Classify novel samples into known disease type (disease diagnosis)
Challenge: Thousands of genes (columns), few samples
Solution: Dimensionality Reduction

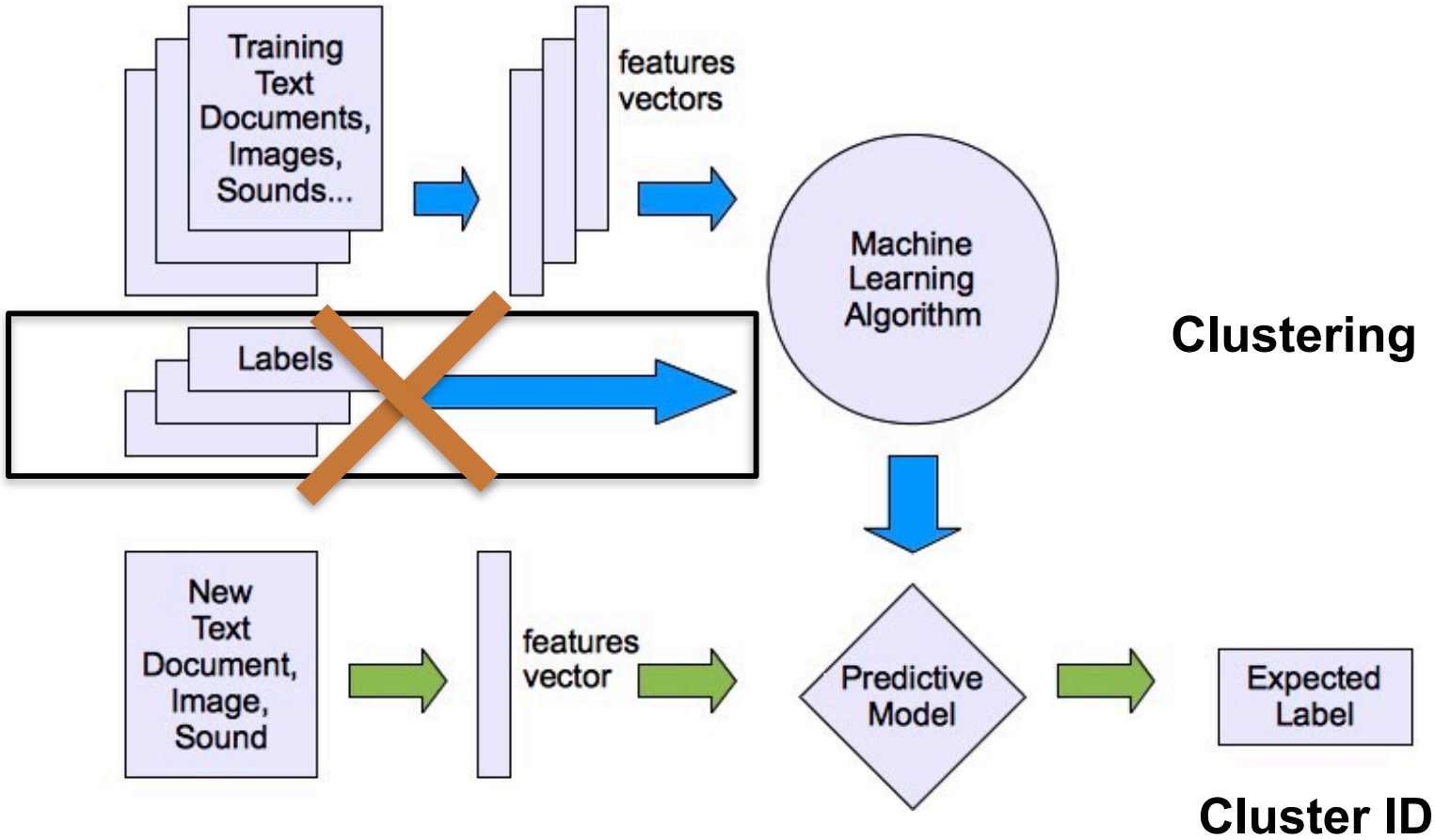
Supervised Learning



Img Ref: web

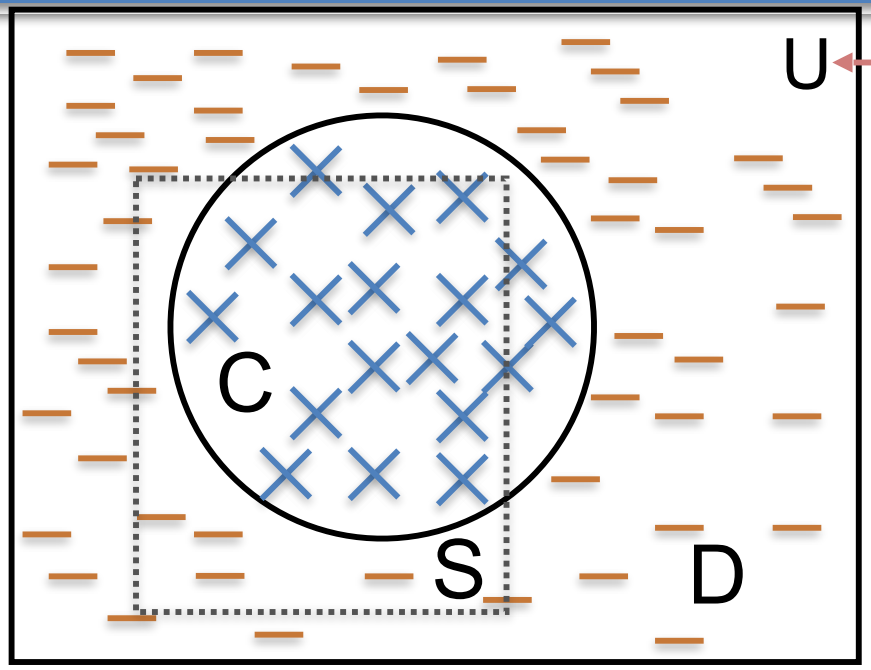
IIT JAMMU

Un-Supervised Learning



Img Ref: web

Learning Theory



Universe data distribution

Let C be the concept (true function) which is capable of modelling the unknown data distribution **D**
And, the training sample **S** (i.i.d from **D**) is available to model the concept **C.**
We look for a hypothesis $h \in H$ which can model C with low generalization error (at the most $0 \leq \epsilon \leq 1$).
[Acceptable Hypothesis!]
Confidence?

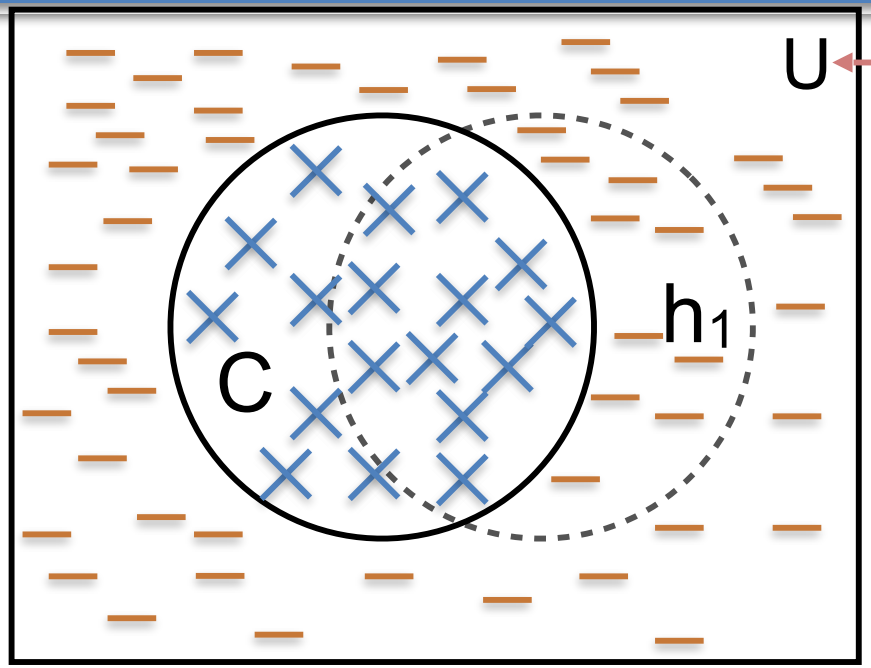
Approximately correct:

$$P(C \oplus h) <= \epsilon$$

Prob. distribution

Error

Learning Theory: Approximately correct hypothesis



Universe data distribution

$C \oplus h_1$ = Error region$_1$

$h_i \in H \quad \forall i$

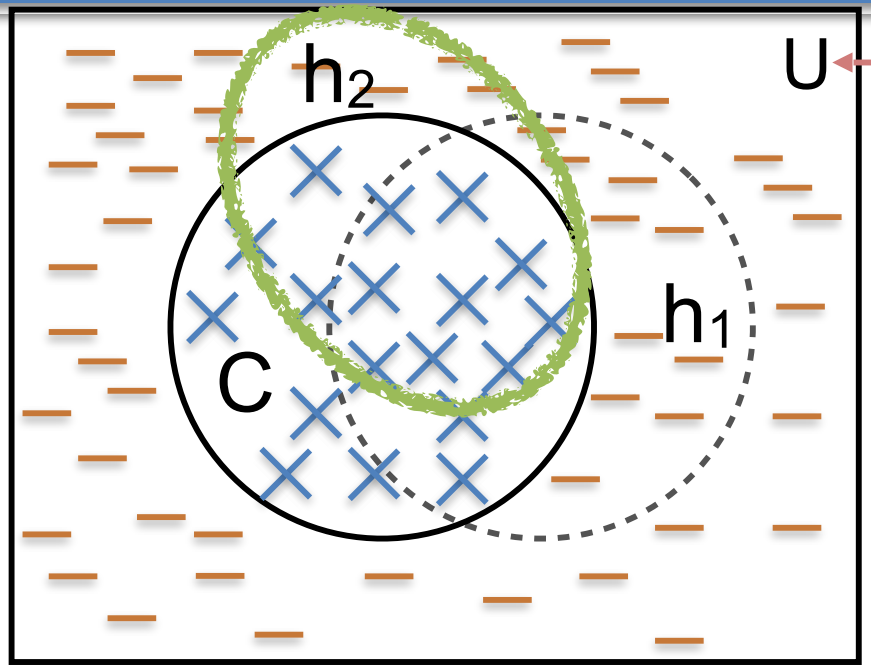
Approximately correct:

$$P(C \oplus h_1) <= \epsilon_1$$

Prob. distribution

Error

Learning Theory: PAC



U ← Universe data distribution

$C \oplus h_1$ = Error region$_1$

$C \oplus h_2$ = Error region$_2$

$$h_i \in H \quad \forall i$$

Approximately correct:

$$P(C \oplus h_i) <= \epsilon_i$$

Prob. distribution

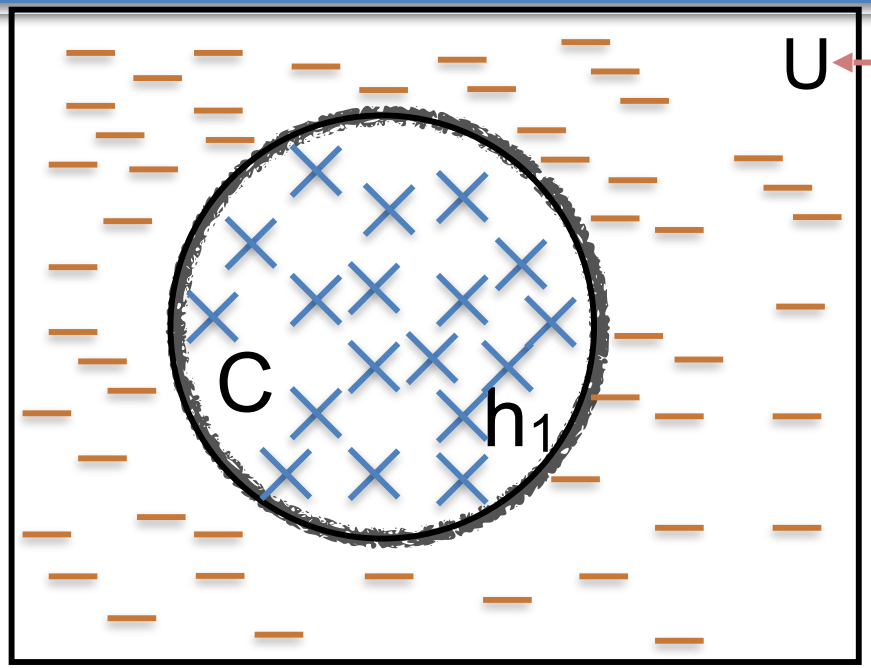
Error

Probably:

$$P(P(C \oplus h_1) > \epsilon_1)) < \delta$$

i.e. Probability that generalization error is less than ϵ_1 is at most $0 \leq \delta \leq 1$

Learning Theory: Consistent Hypothesis



Universe data distribution

$C \oplus h_1 = 0$

$h_i \in H \quad \forall i$

Approximately correct

$P(P(C \oplus h) <= \epsilon) >= (1 - \delta)$

Prob. distribution

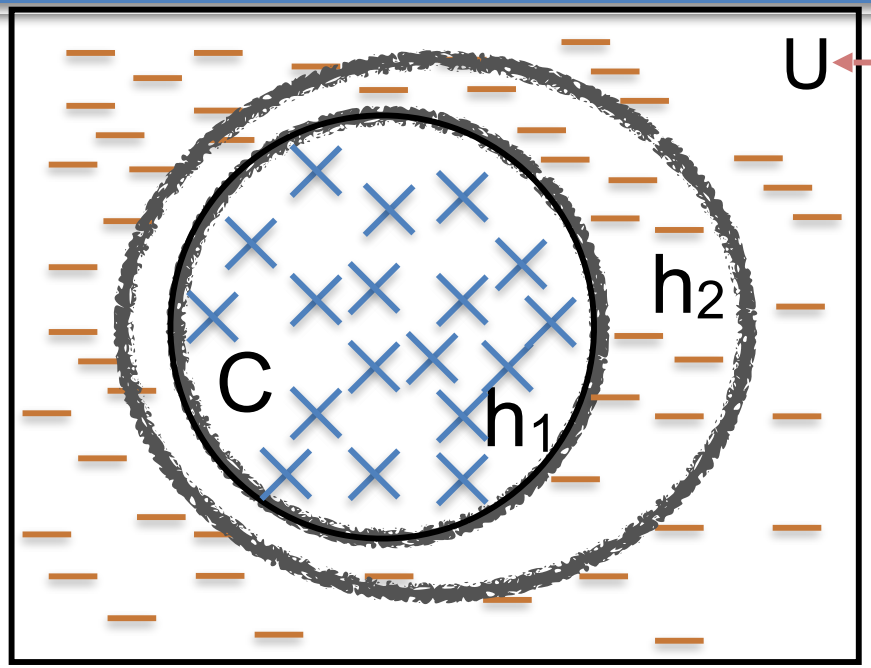
Confidence

OR

$P(P(C \oplus h) > \epsilon)) < \delta$

i.e. Probability that generalization error is less than ϵ is at most $0 \leq \delta \leq 1$

Learning Theory: Consistent Hypothesis



Universe data distribution

$C \oplus h_1 = 0$

$C \oplus h_2 = 0$

$h_i \in H \quad \forall i$

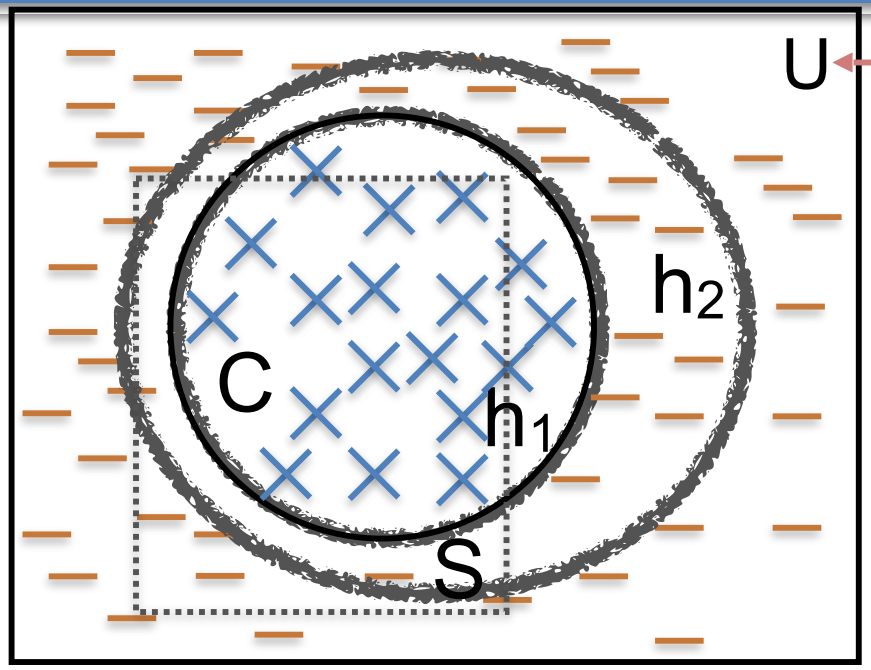
Approximately correct

$P(P(C \oplus h) <= \epsilon) >= (1 - \delta)$

Prob. distribution

Confidence

Learning Theory: Consistent Hypothesis



U ← Universe data distribution

Consistent Hypothesis (i.e. $error_S(h) = 0$)

The gap between training and true errors:

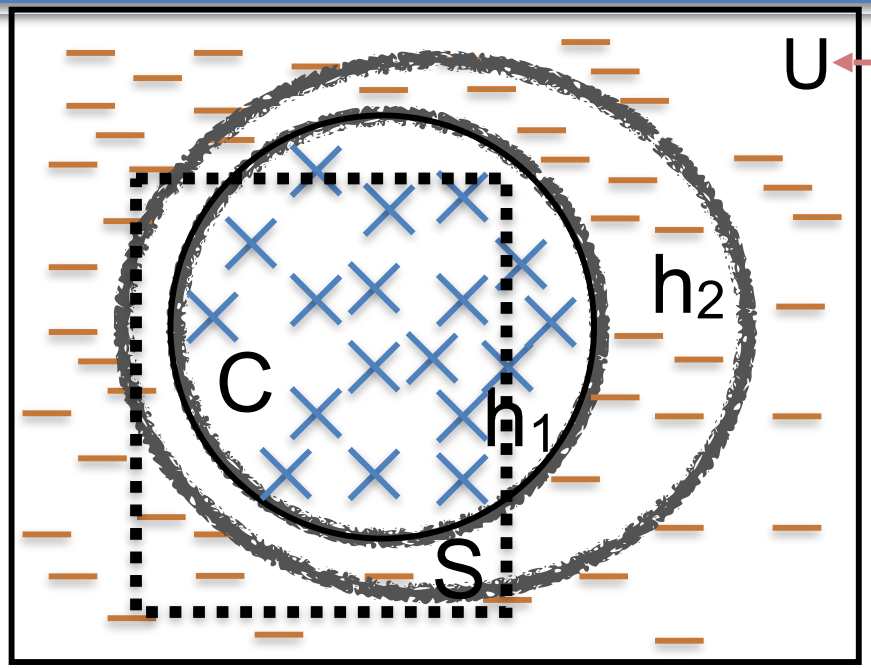
$$error_D(h) \leq error_S(h) + \epsilon$$

True error of a hypothesis h with respect to c ($error_D(h)$) is determined by how often h(x) and c(x) disagree (i.e. $h(x) \neq c(x)$) over future instances drawn at random (i.i.d) from D:

$$error_D(h) = P_{x \in D}[c(x) \neq h(x)]$$

$$error_D(h) = E_{x,c(x)}[L(c(x), h(x)] = \int_X \int_{c(X)} L(c(x), h(x)) P(x, c(x))\, dxdy$$

Learning Theory: Consistent Hypothesis



Universe data distribution

Consistent Hypothesis (i.e. $error_S(h) = 0$)

The gap between training and true errors:

$$error_D(h) \leq error_S(h) + \epsilon$$

Training error (empirical error) of a hypothesis h with respect to c ($error_S(h)$) is determined by how often h(x) and c(x) disagree (i.e. $h(x) \neq c(x)$) over training instances $x \in S$ ($\subseteq D$):

$$error_S(h) = P_{x \in S}[c(x) \neq h(x)] \equiv \frac{\sum_{x \in S} I(c(x) \neq h(x))}{|S|}$$

Learning Theory: Consistent Hypothesis (i.e. $error_S(h) = 0$)

The gap between training and true errors: $error_D(h) \leq error_S(h) + \epsilon$

$$P_r[(\exists h \in H)\ s.t.\ (error_S(h) = 0) \wedge (error_D(h) > \epsilon)] \leq |H| \exp^{-\epsilon m} \leq \delta$$

where, **m** is the number of samples in training data S (|S|)

Then:

$$m \geq \frac{1}{\epsilon}(\ln|H| + \ln(\frac{1}{\delta})$$

And, with probability at least (1- δ), the true error ($error_D(h)$) will be bounded as follows:

$$error_D(h) \leq \frac{1}{m}(\ln|H| + \ln(\frac{1}{\delta})$$

Learning Theory: Agnostic Learning (i.e. $error_S(h) \neq 0$)

The gap between training and true errors: $error_D(h) \leq error_S(h) + \epsilon$

$$P_r[(\exists h \in H)\ s.t.\ (error_D(h) > error_S(h) + \epsilon)] \leq |H| \exp^{-2\epsilon^2 m} \leq \delta$$

where, **m** is the number of samples in training data S (|S|)

Then:

$$m \geq \frac{1}{2\epsilon^2}(\ln|H| + \ln(\frac{1}{\delta})$$

And, with probability at least (1- δ), the true error ($error_D(h)$) will be bounded as follows:

$$error_D(h) \leq error_S(h) + \sqrt{\frac{1}{2m}(\ln|H| + \ln(\frac{1}{\delta}))}$$

Learning Theory: Infinite Hypothesis space (|H| = ∞)

Expressiveness of an infinite hypothesis space: ***Vapnik Chervonenkis DImension***

The gap between training and true errors: $error_D(h) \leq error_S(h) + \epsilon$

where, **m** is the number of samples in training data S (|S|)

Then:

$$m \geq \frac{1}{\epsilon}\left(8\ VC(H)\ \log_2\left(\frac{13}{\epsilon}\right) + 4\log_2\left(\frac{2}{\delta}\right)\right)$$

And, with probability at least (1- δ), the true error ($error_D(h)$) will be bounded as follows:

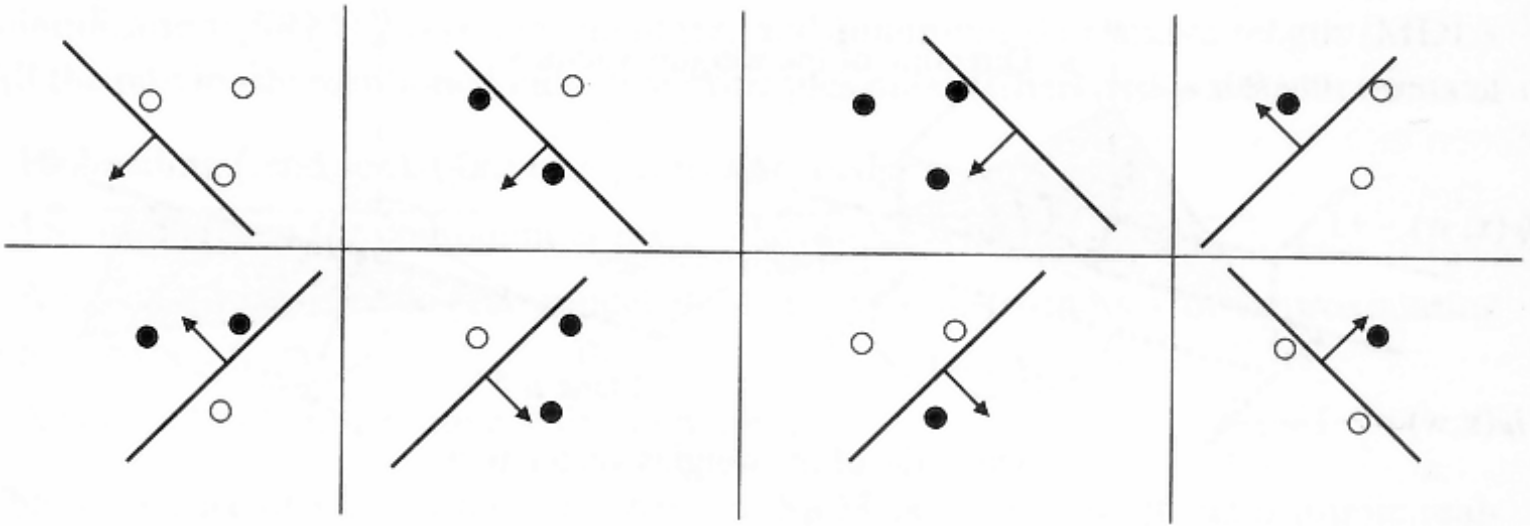
$$error_D(h) \leq error_S(h) + \sqrt{\frac{1}{m}\left(VC(H)\left(\ln\frac{2m}{VC(H)} + 1\right) + \ln(\frac{4}{\delta})\right)}$$

Learning Theory: Infinite Hypothesis space (|H| = ∞)

High VC dimension => better chance of approximating **h s.t.** **($error_S(h) = 0$)**

Low VC dimension => better chance of generalizing out of sample **($error_S(h) \approx error_D(h)$)**

The gap between training and true errors: $error_D(h) \leq error_S(h) + \Omega(VC(H))$



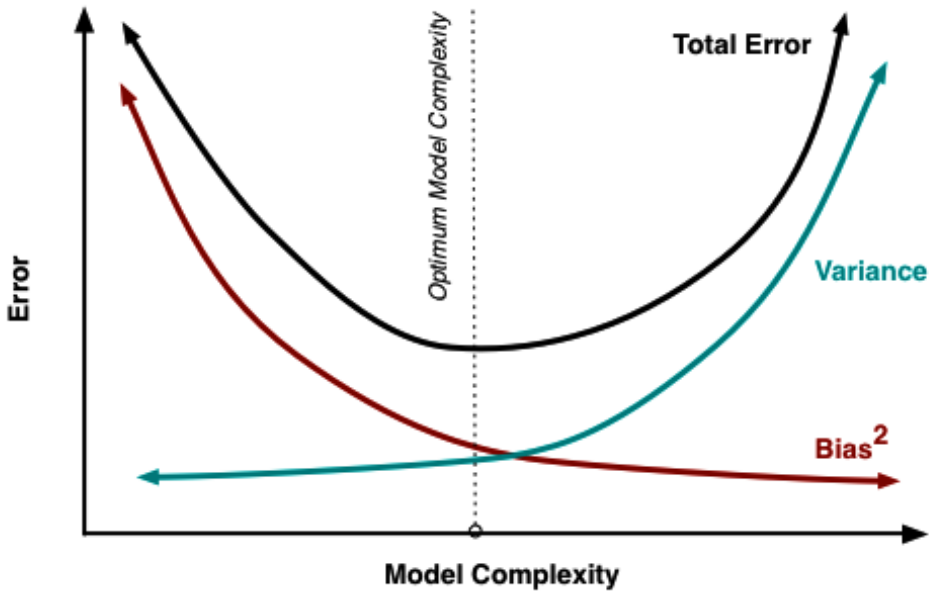
Learning Theory: Infinite Hypothesis space (|H| = ∞)

High VC dimension => better chance of approximating **h s.t.** **($error_S(h) = 0$)**

Low VC dimension => better chance of generalizing out of sample **($error_S(h) \approx error_D(h)$)**

The gap between training and true errors:

$$error_D(h) \leq error_S(h) + \Omega(VC(H))$$



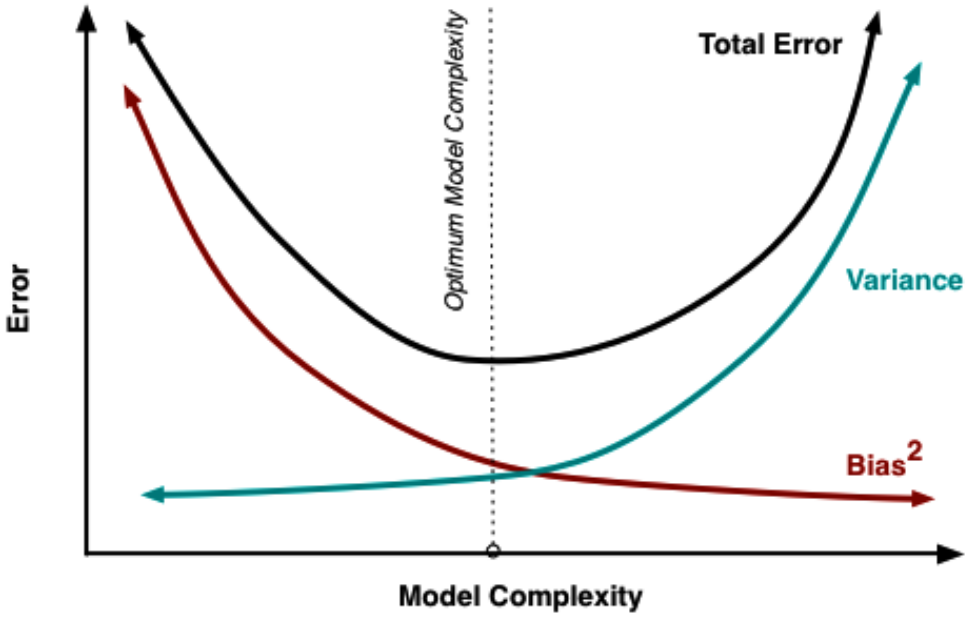
Img Ref: http://scott.fortmann-roe.com/docs/BiasVariance.html

Challenges in Learning

- Bias-Variance
  - Overfitting (Variance)
    - Regularization, Cross-validation, Data Augmentation
    - Reduce Model Complexity
  - Underfitting (Bias)
    - Increase the Model Complexity, Data curation, Domain Information

- Loss (Empirical and Structural)
- VC-dimensions

Explainable AI !



Img Ref: http://scott.fortmann-roe.com/docs/BiasVariance.html