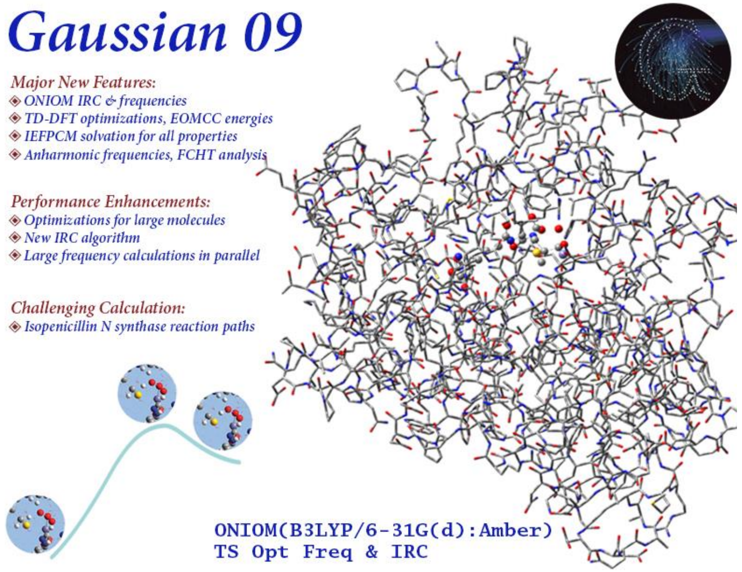
# Statistical learning framework

Seongok Ryu
KAIST Chemistry

# Contents

- Motivation
- Statistical modeling and inference
- Likelihood function
- Maximum likelihood estimation
- Bayes' theorem and Bayesian inference
- Generative methods vs Discriminvative methods
- Beyond likelihoodism

# Motivation




VASP
b-initio
ackage
imulation
ienna

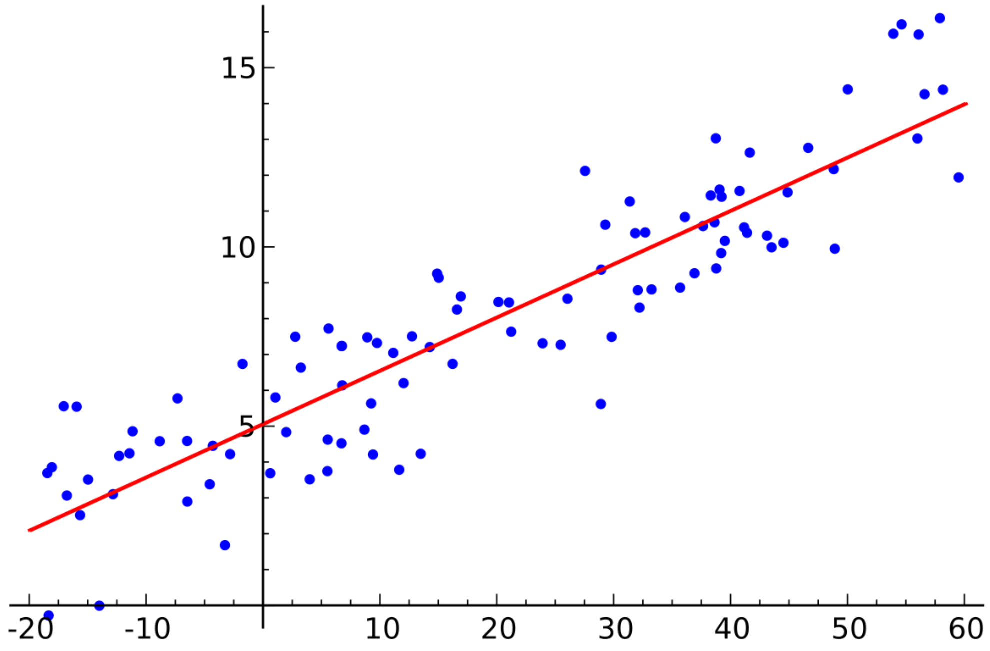
ACE-Molecule

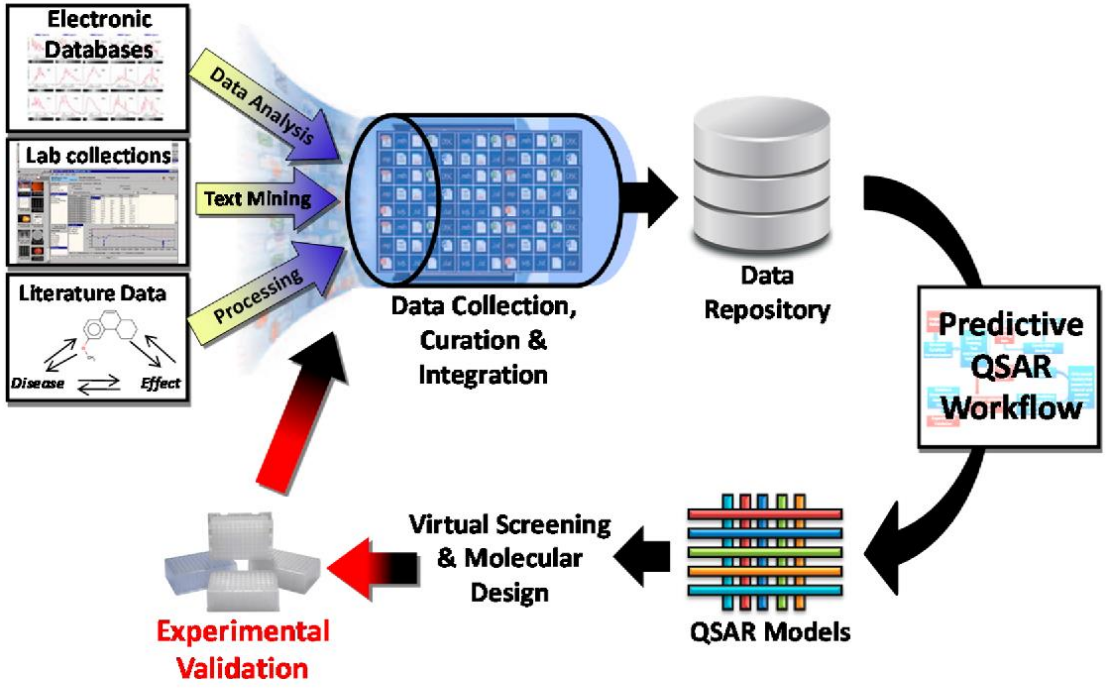
Our group has devoted to develop the computational tools based on quantum chemistry theory, *ACE-Molecule*. This package aims to calculate interesting physical properties by solving physics equations, such as Schrodinger equation.

Such physics principle-based computations have been regarded as a standard tool in physics/chemistry research communities.

# Motivation

In recent two years, our group also has studied statistical modeling frameworks, which find a hypothesis that best explains given phenomena.



# Motivation

| Physics theory based simulation | Statistical modeling |
|---|---|
| Constructing models with basic physics equation<br>• Schrodinger equation<br>• Newton's equation | Developing hypotheses from observations (data) |
| Pros)<br>• Very accurate for some systems<br>• Theoretically guaranteed<br><br>Cons)<br>• Difficult to calculate large systems, e.g. proteins<br>• Not easily applicable to very complex systems, e.g. strongly correlated materials | Pros)<br>• High scalability<br>• Able to learn patterns inherent in a complex system<br><br>Cons)<br>• Necessitate sufficient amount of data<br>• Not easy to generalize obtained hypothesis |

# Viewpoints on machine learning

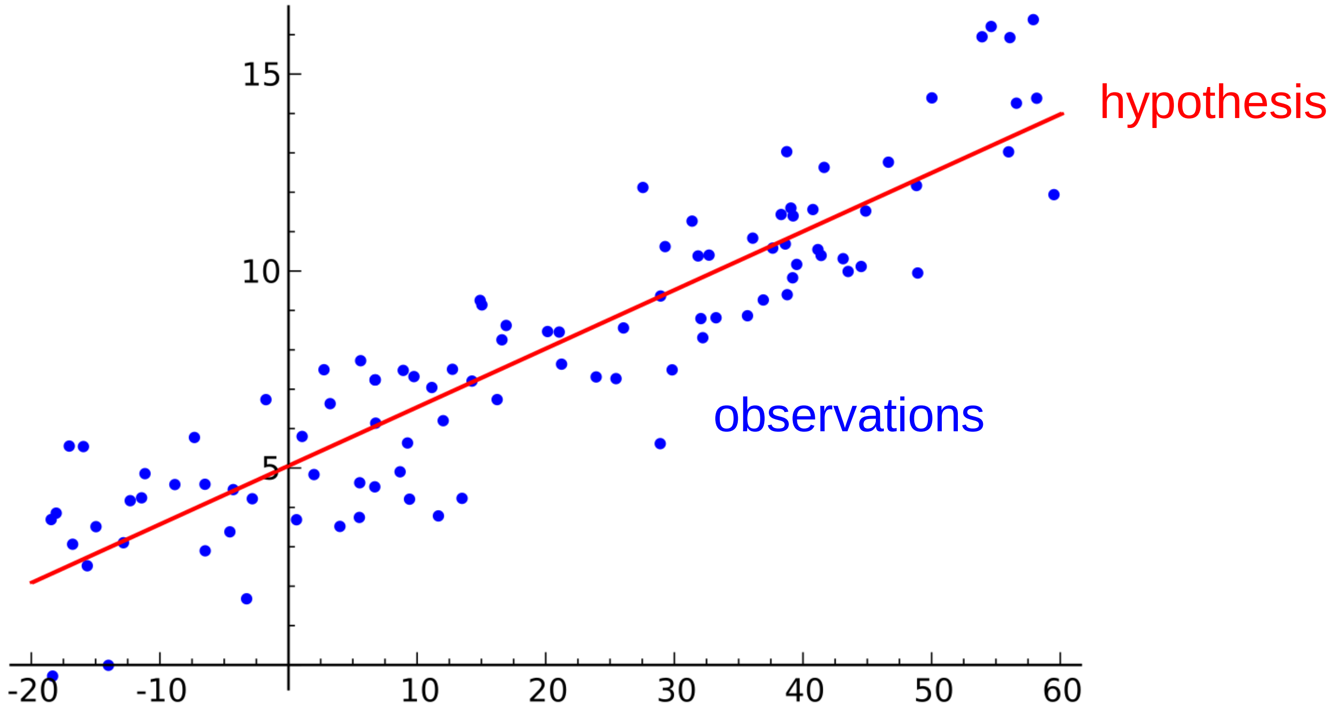
1. Statistical learning theory
    - What hypothesis do we want to obtain?
    - What can we learn and not?
    - Learning algorithms including prediction models, generative models, data-efficient learning approaches

2. Model architecture
    - Efficient parameterization of hypothesis
    - So-called *inductive biases* (we will investigate very detail later)

3. Generalization
    - To avoid over-fitting problems
    - Regularization: to reduce number of parameters consisting a hypothesis

# Statistical modeling and inference

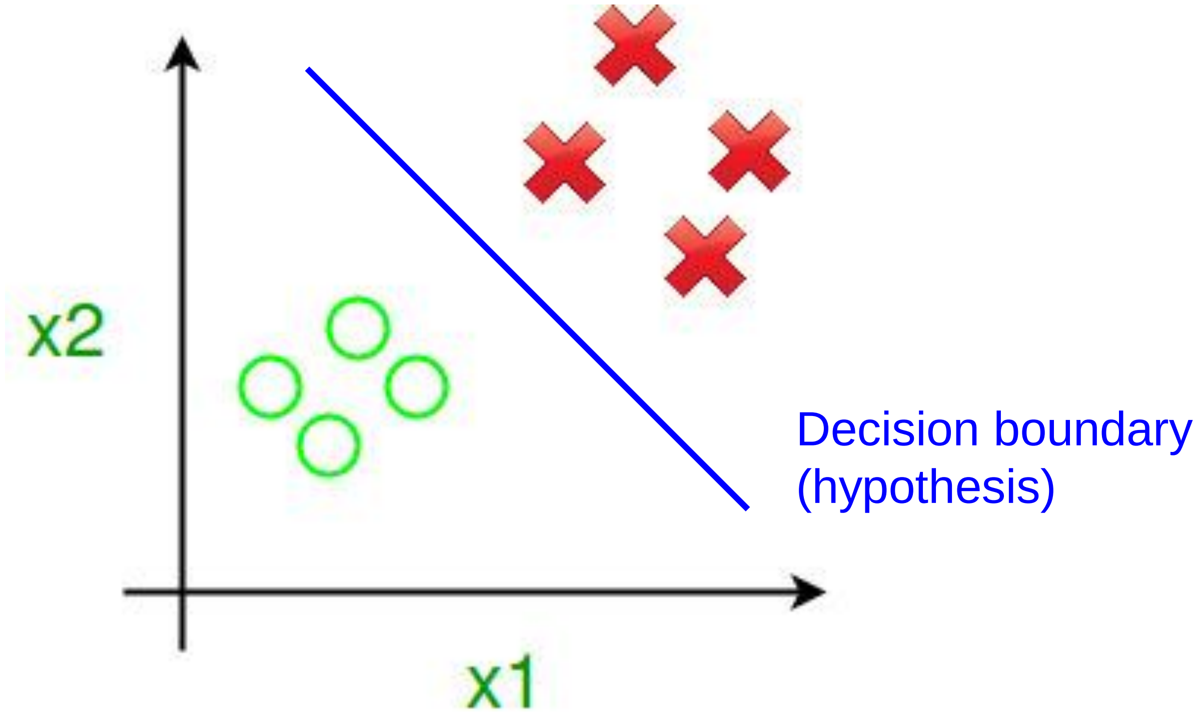
*In this presentation, we concentrate on supervised learning algorithms which are usually used for prediction models.*




- *Statistical modeling* aims to obtain a *hypothesis* with training samples (observations) $\{\mathbf{X}, \mathbf{Y}\}$, where we consider that samples are *independent and identically distributed (i.i.d.) random variables* sampled from the joint distribution $p(\mathbf{X}, \mathbf{Y})$.
- *Statistical inference* is a procedure of inferring a new output $\boldsymbol{y}^*$ of a new input $\boldsymbol{x}^*$ by using the obtained hypothesis.

# Statistical modeling and inference

*In this presentation, we concentrate on supervised learning algorithms which are usually used for prediction models.*




- *Statistical modeling* aims to obtain a *hypothesis* with training samples (observations) $\{\mathbf{X}, \mathbf{Y}\}$, where we consider that samples are *independent and identically distributed (i.i.d.) random variables* sampled from the joint distribution $p(\mathbf{X}, \mathbf{Y})$.
- *Statistical inference* is a procedure of inferring infer the new output $\boldsymbol{y}^*$ of the new input $\boldsymbol{x}^*$ by using the obtained hypothesis.

# Likelihood function

"In statistics, a *likelihood* function $\mathcal{L}(\theta, D)$ is a function of parameters $\theta$ within the parameter space $\Theta$ that describes the probability of obtaining the data $D$." – Wikipedia, Likelihood function.

- For the random variables sampled from a ***discrete probability distribution***, the function

$$\mathcal{L}(\theta|x) = p_\theta(x) = P_\theta(X = x)$$

  is the likelihood of $\theta$, given the outcome of the $x$ of the random variable $X$.

- For the random variables following an ***absolutely continuous probability distribution*** with ***density function*** $f$ depending on parameter $\theta$,

$$\mathcal{L}(\theta|x) = f_\theta(x)$$

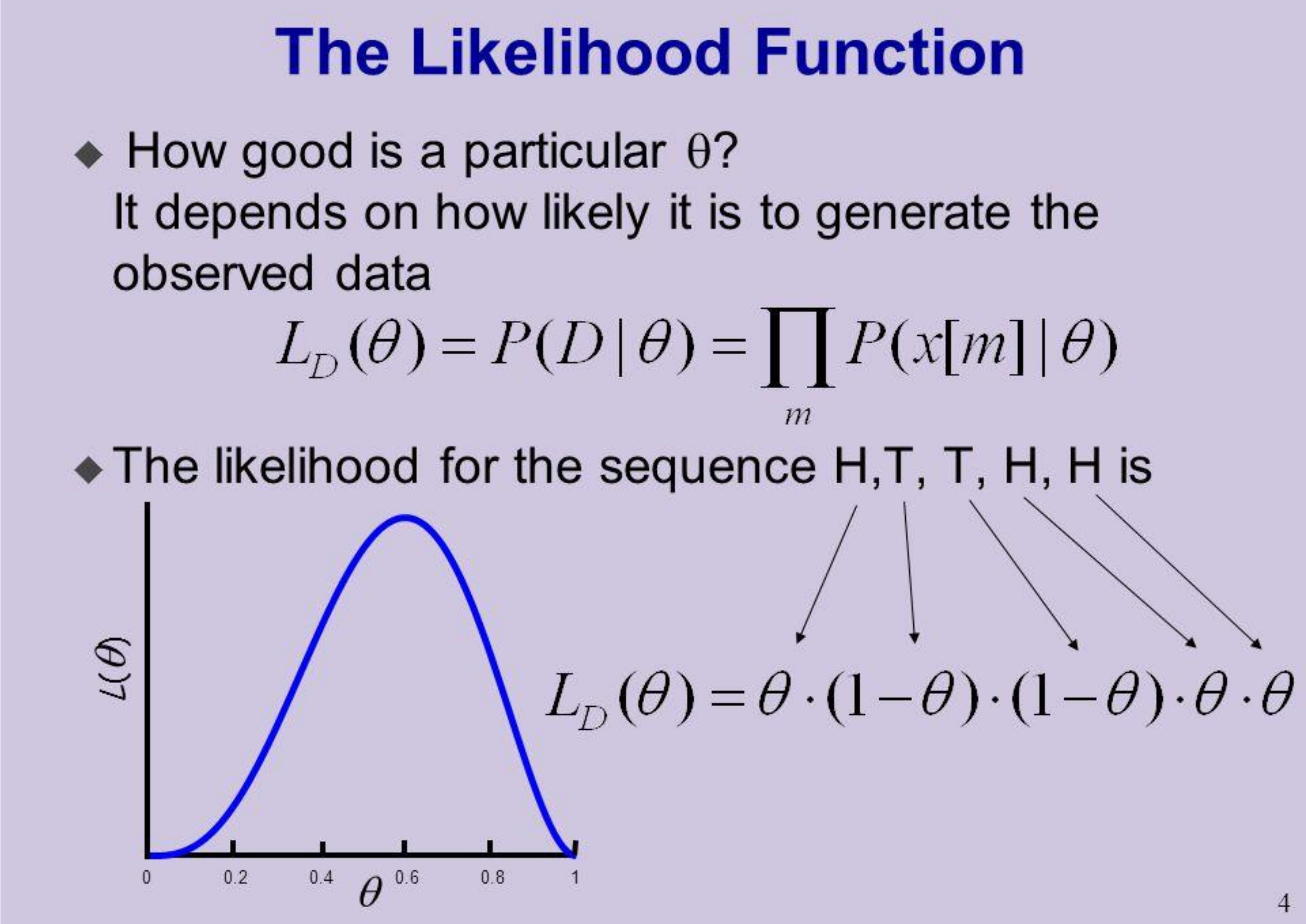
  is the likelihood function of $\theta$, given the outcome of $x$ of the random variable $X$.

  Note that a likelihood function is defined under assuming the existence of a density function, which is not ensured for the high-dimensional probability distribution function.

# Likelihood function

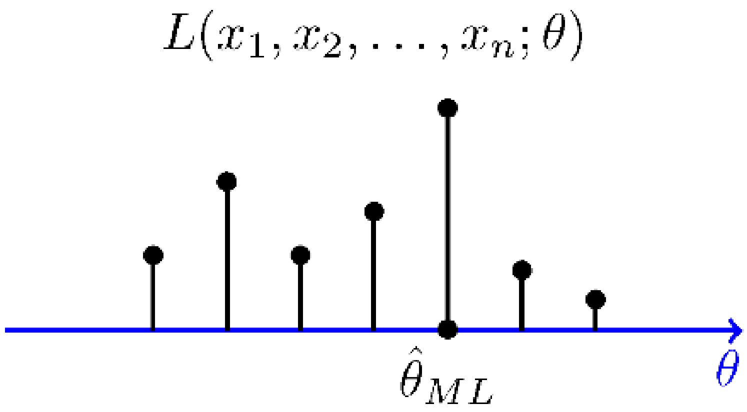
"In statistics, a *likelihood* function $\mathcal{L}(\theta, D)$ is a function of parameters $\theta$ within the parameter space $\Theta$ that describes the probability of obtaining the data $D$." – Wikipedia, Likelihood function.

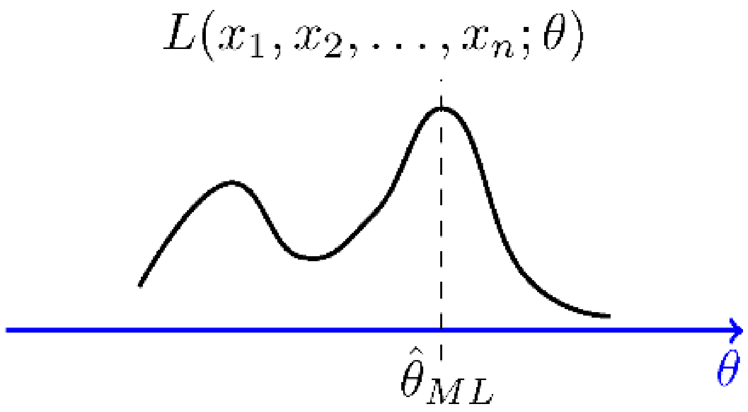
https://slideplayer.com/slide/4663314/



# Maximum likelihood estimation

Maximum Likelihood Estimation (MLE or ML-estimation)
: Choose $\hat{\theta}_{ML} = \text{argmax}_{\theta \in \Theta} \mathcal{L}(\theta, D)$ which maximizes a likelihood function $\mathcal{L}(\theta, D)$.

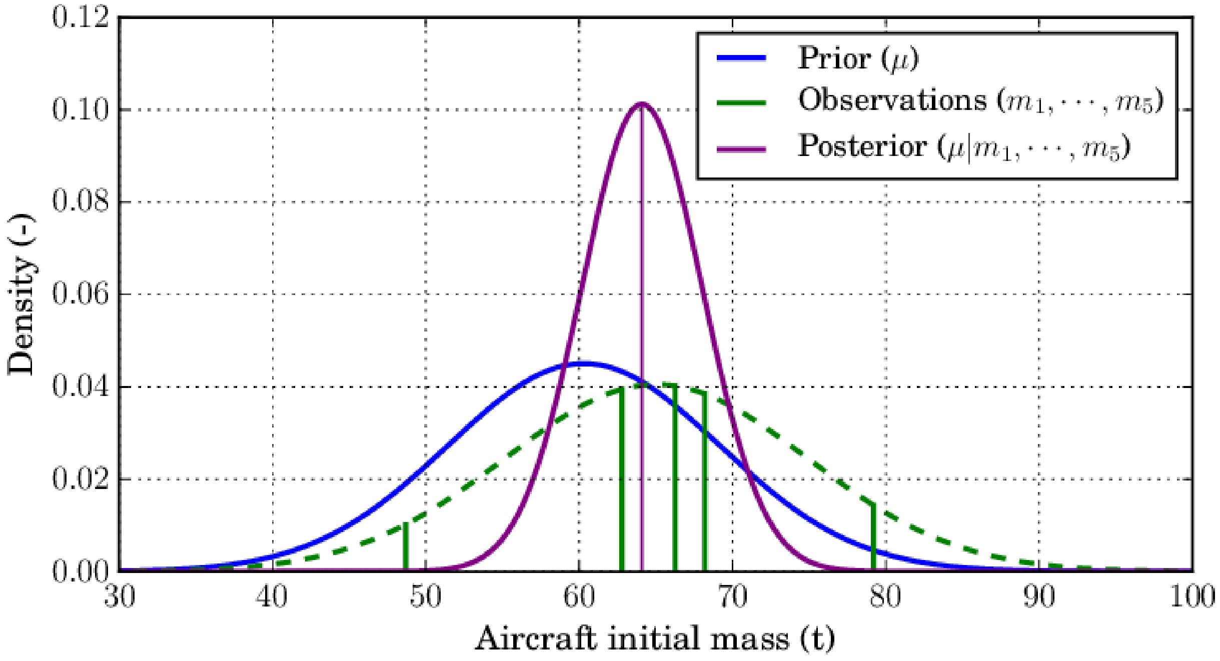



In practice, a *log-likelihood* function $l(\theta, D) = \log \mathcal{L}(\theta, D)$ is more conveniently used, since logarithmic function is strictly increasing function.

# Bayes' theorem and Bayesian inference

"*Bayesian inference* is a method of statistical inference in which Bayes' theorem is used to update the probability for a hypothesis as more evidence or information becomes available."

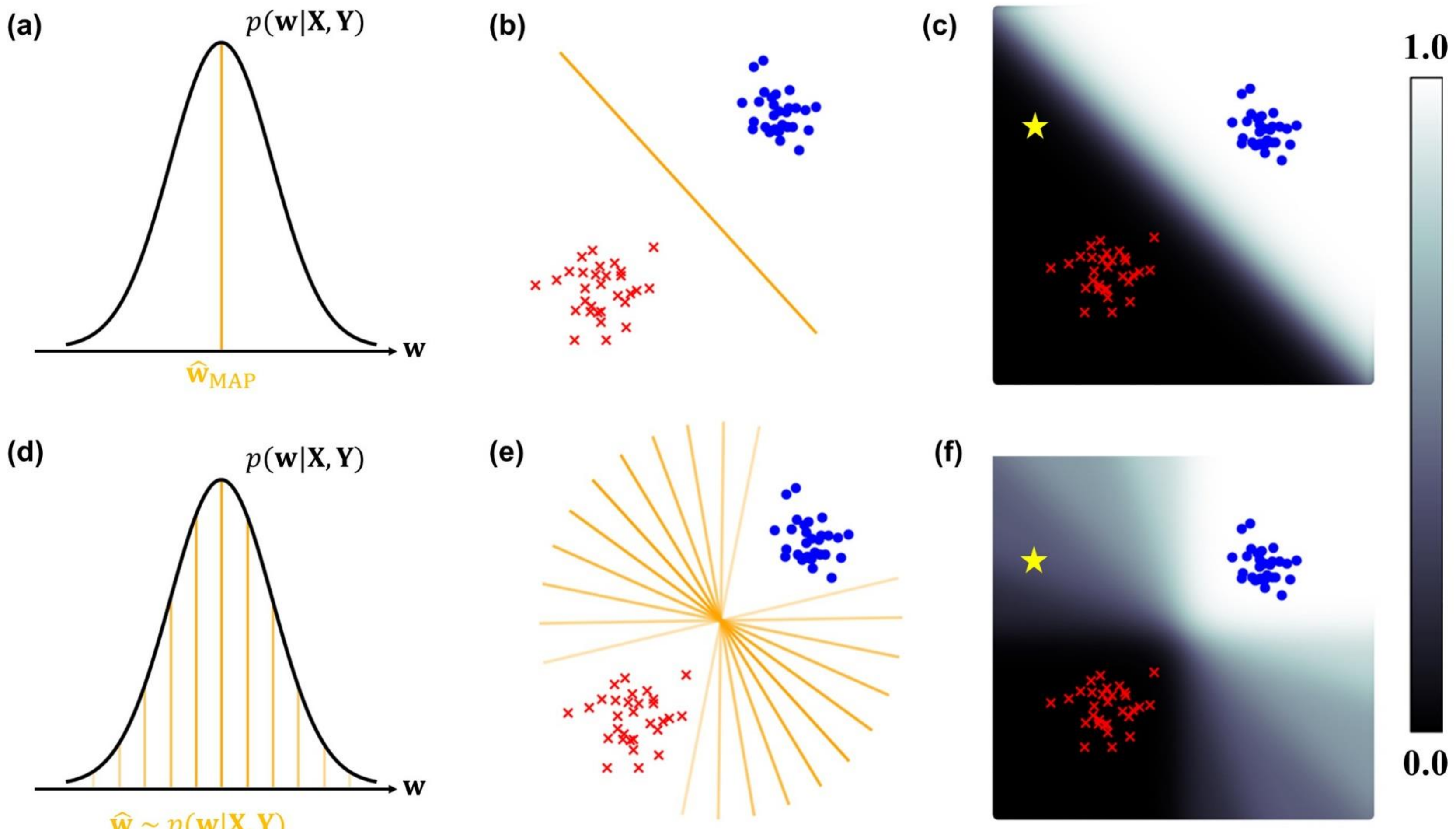



Likelihood    Prior

*Bayes' theorem*: $$p(\theta|D) = \frac{p(D|\theta) \cdot p(\theta)}{p(D)}$$

Posterior    Evidence

# Bayes' theorem and Bayesian inference




Maximum-a-posteriori estimation
: $\hat{\mathrm{w}} = \mathrm{argmax}_{\mathrm{w}}\, p(x|\mathrm{X},\mathrm{Y})$
→ point-estimation

Bayesian inference of model parameter
→ aiming to compute a posterior distribution
→ however, computation is usually intractable

# Beyond Likelihoodism

Recall the definition of likelihood

- For the random variables following an ***absolutely continuous probability distribution*** with ***density function*** $f$ depending on parameter $\theta$,

$$\mathcal{L}(\theta|x) = f_\theta(x)$$

  is the likelihood function of $\theta$, given the outcome of $x$ of the random variable $X$.

  Note that a likelihood function is defined under assuming the existence of a density function, which is not ensured for the high-dimensional probability distribution function.

For a high-dimensional distribution space, the existence of a likelihood function is not ensured.
This leads to ambiguity in defining the measure between the two distribution, such as the divergence between a data distribution and a predictive distribution.

# Beyond Likelihoodism

Therefore, another framework is needed to interpret learning procedures.

- *Empirical risk minimization (ERM)* principle defines a family of learning algorithms and gives theoretical bounds on model performance.
- We assume that there is a joint probability distribution $P(x, y)$ over $X$ and $Y$, and that the training dataset consists of $n$ instances $(x_1, y_1), \ldots, (x_n, y_n)$ drawn i.i.d. from $P(x, y)$. We also assume that we are given a non-negative loss function $L(\hat{y}, y)$ which measures difference between the prediction $\hat{y}$ of the hypothesis and true outcome $y$.
- The risk associated with hypothesis $h(x)$ is then defined as the expectation of the loss function:

$$R(h) = \mathbb{E}[L(h(x), y)] = \int L(h(x), y)\,dP(x, y)$$

- The ultimate goal of a learning framework is to find the optimal hypothesis $h^*$ among a fixed class of functions $\mathcal{H}$ for which the risk $R(h)$ is minimal:

$$h^* = \operatorname*{argmin}_{h \in \mathcal{H}} R(h)$$

# Beyond Likelihoodism

Therefore, another framework is needed to interpret learning procedures.

- In general, the risk cannot be computed because the (true) distribution $P(x, y)$ is unknown to the learning algorithm. However, we can compute an approximation, so-called the empirical risk, by averaging the loss function on the training set:

$$R_{emp}(h) = \frac{1}{n}\sum_{i=1}^{n} L(h(x_i), y_i)$$

- The ERM principle states that the learning algorithm should choose a hypothesis $\hat{h}$ which minimizes the empirical risk:

$$\hat{h} = \underset{h \in \mathcal{H}}{\operatorname{argmin}}\, R_{emp}(h)$$