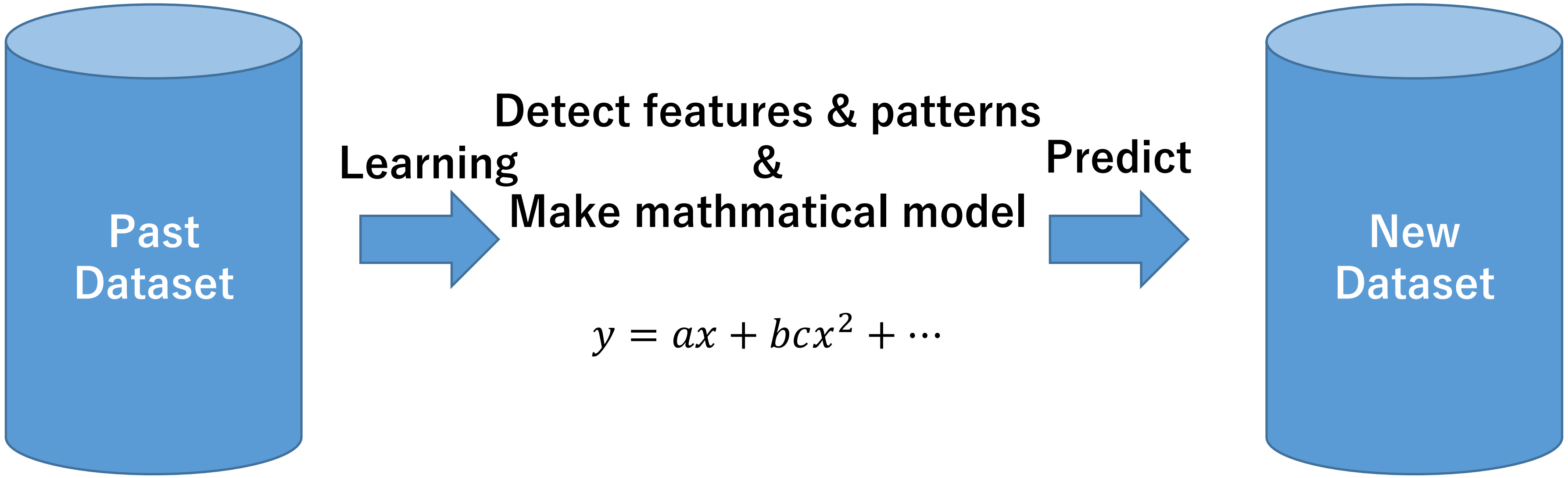
Difference between machine learning and statistics

&

Application to genome analysis

What is Machine Learning？

Machine learning (ML) is the study of algorithms and mathematical models
that computer systems use to progressively improve their performance on a specific task…(Wikipedia)



Past Dataset

Learning

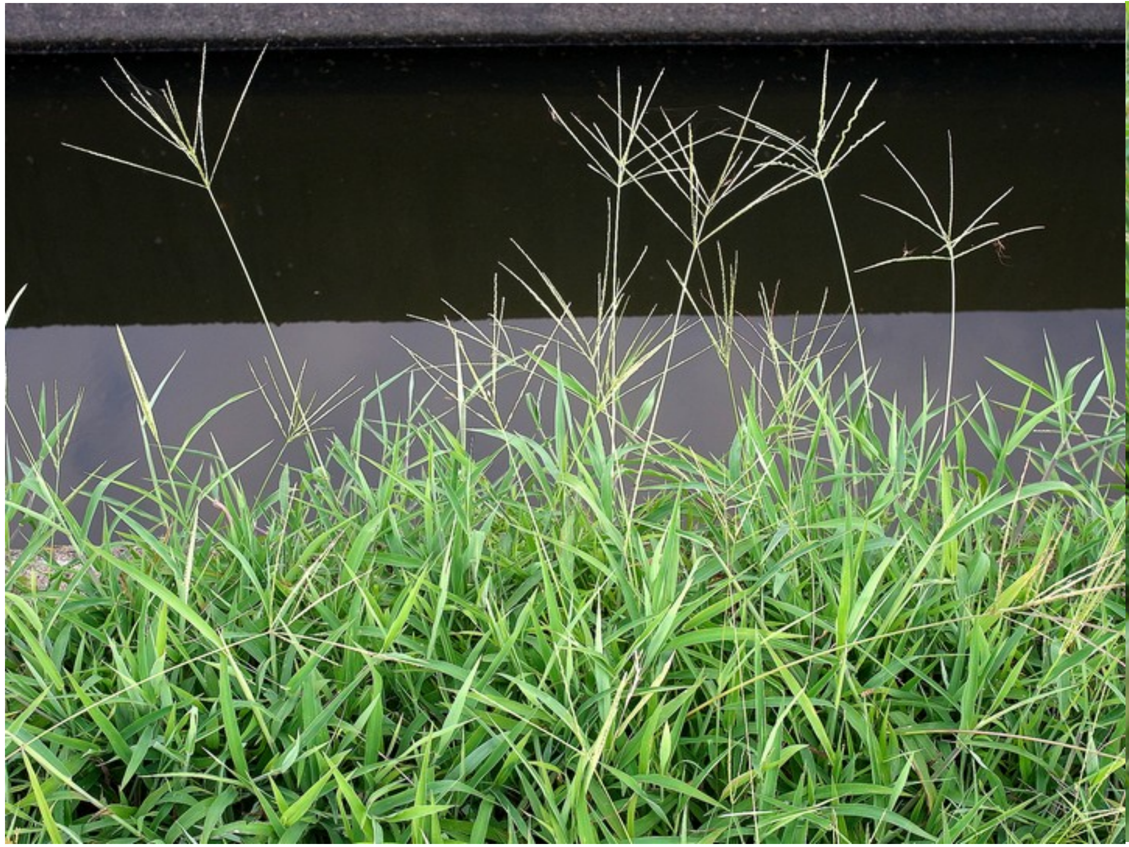
Detect features & patterns
&
Make mathmatical model

$$y = ax + bcx^2 + \cdots$$

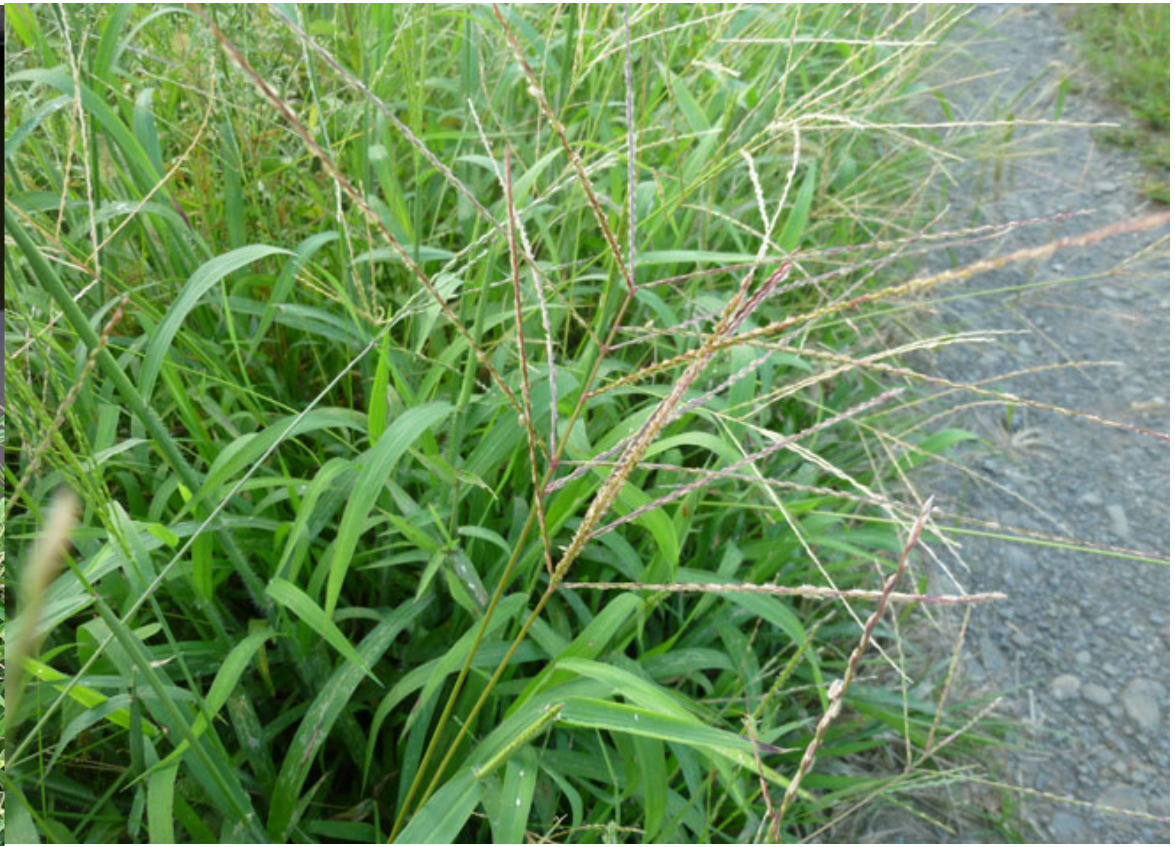
Predict

New Dataset

Example



Digitaria ciliaris



Digitaria violascens

Example

= *Digitaria ciliaris*

= *Digitaria ciliaris*

= *Digitaria violascens*

= *Digitaria violascens*

⋮ ⋮

Learning
From past data

Detect patterns

(ex. *Digitaria violascens*
= Hairless leaf sheath)

Prediction

New Data

= ？？？

97%:*Digitaria violascens*
3%: *Digitaria ciliaris*

Example

Past Data

Sequence	Crop yield
ATTGAC···	2,0kg
GAGGTA···	3,6kg
TGCCGC···	1,1kg
ATCGAA···	2.1kg
···	···

Learning

Detect features & patterns
&
Make mathmatical model

$$\boldsymbol{CropYield = ax + by + cz \ldots}$$

Prediction

New Dataset

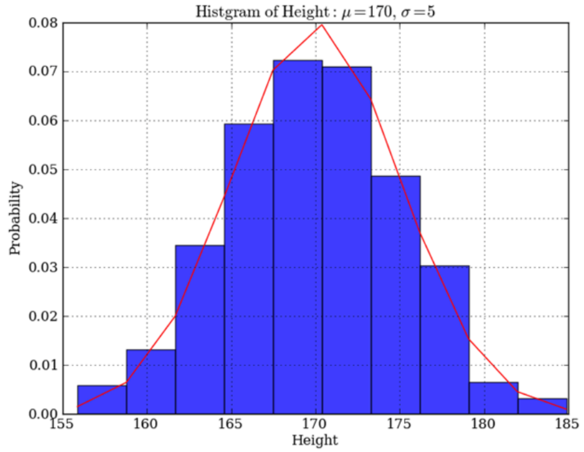
Sequence	Crop yield
GAAAAC···	???
TTAGGG···	???
···	···

Sequence	Predicted Crop yield
GAAAAC···	1.82kg
TTAGGG···	1.11kg
···	···

Difference between machine learning and statistics

…method & algorithm are common

- ## Statistics

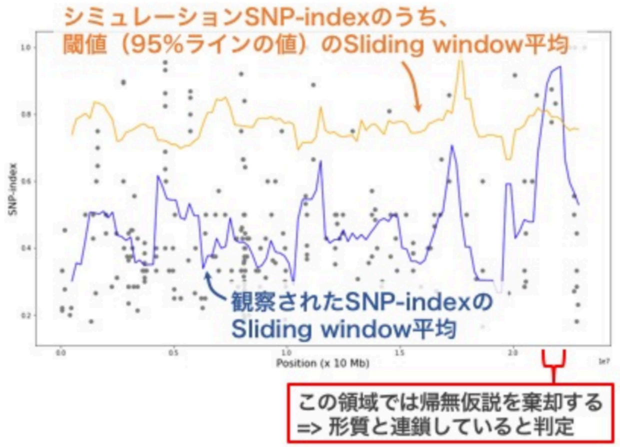


Description statistics … Make data easier for people to understand using statistics or visualize

ex) mean, variance

Estimated statistics …Consideration on data

ex) Estimate mean value of population
Estimate the position of related genes at SNP-index



- ## Machine learning

Make highly accurate prediction models from past data

Difference between machine learning and statistics

… Different purposes / strategies

- **Statistics** … Selection of models and methods and validity are important

ex) $\boldsymbol{CropYield = ax + by + cz \ldots}$

…**Good Accuracy**

➡

Which variable are important ?
Why this model get good accuracy ?

ex) $\boldsymbol{a}$ is big → $\boldsymbol{x}$ has large effect

- **Machine learning** … How to achieve high prediction accuracy is important

→ Focus on prediction accuracy even in various methods, black box model

ex) $\boldsymbol{CropYield = \sqrt[3]{a^8x^{27}z^{55}} + \lim_{n\to\infty}\left(\frac{\int b \sin \alpha n}{n}\right)^{n}} + \cdots$

…**Very high accuracy**

➡

Sometimes it's difficult to understand model

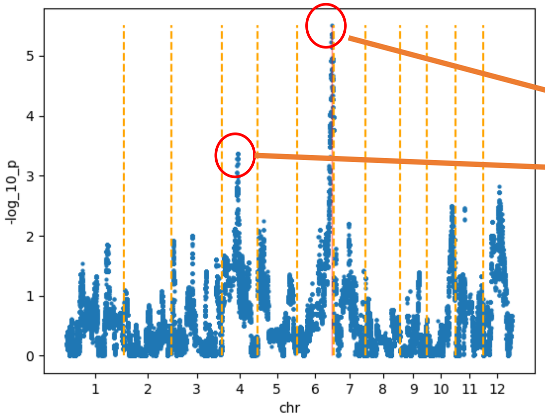
ex) $\boldsymbol{a}$ is big → ？ ？ ？

Application to genome analysis

ex1) To know gene function → It is important how easy it is to understand the model

$$\boldsymbol{CropYield = \beta_1 gene_1 + \beta_2 gene_2 + \beta_3 gene_3 \ldots}$$

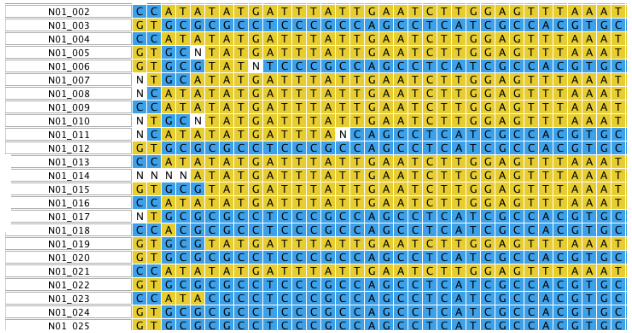
Gene Effects β



It shows that
Genes in this position affect Crop Yield

ex2) Apply Genomic Breeding → The prediction accuracy of new dataset is also important

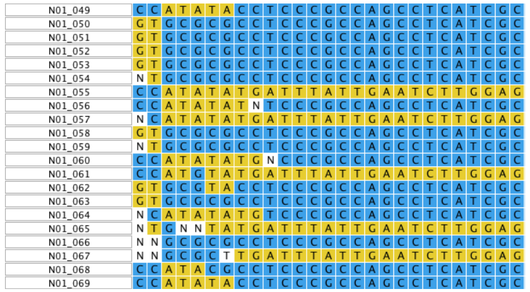
CropYield +



N01_002	CCATATATGATTTATTGAATCTTGGAGTTTAAAT
N01_003	GTGCGCGCCTCCCGCCAGCCTCATCGCCACGTGC
N01_004	CCATATATGATTTATTGAATCTTGGAGTTTAAAT
N01_005	GTGCNTATGATTTATTGAATCTTGGAGTTTAAAT
N01_006	GTGCGTATNTCCCGCCAGCCTCATCGCCACGTGC
N01_007	NTGCATATGATTTATTGAATCTTGGAGTTTAAAT
N01_008	NCATATATGATTTATTGAATCTTGGAGTTTAAAT
N01_009	CCATATATGATTTATTGAATCTTGGAGTTTAAAT
N01_010	NTGCNTATGATTTATTGAATCTTGGAGTTTAAAT
N01_011	NCATATATGATTTANCAGCCTCATCGCCACGTGC
N01_012	GTGCGCGCCTCCCGCCAGCCTCATCGCCACGTGC
N01_013	CCATATATGATTTATTGAATCTTGGAGTTTAAAT
N01_014	NNNNATATGATTTATTGAATCTTGGAGTTTAAAT
N01_015	GTGCGTATGATTTATTGAATCTTGGAGTTTAAAT
N01_016	CCATATATGATTTATTGAATCTTGGAGTTTAAAT
N01_017	NTGCGCGCCTCCCGCCAGCCTCATCGCCACGTGC
N01_018	CCACGCGCCTCCCGCCAGCCTCATCGCCACGTGC
N01_019	GTGCGTATGATTTATTGAATCTTGGAGTTTAAAT
N01_020	GTGCGCGCCTCCCGCCAGCCTCATCGCCACGTGC
N01_021	CCATATATGATTTATTGAATCTTGGAGTTTAAAT
N01_022	GTGCGCGCCTCCCGCCAGCCTCATCGCCACGTGC
N01_023	CCATACGCCTCCCGCCAGCCTCATCGCCACGTGC
N01_024	GTGCGCGCCTCCCGCCAGCCTCATCGCCACGTGC
N01_025	GTGCGCGCCTCCCGCCAGCCTCATCGCCACGTGC

Model

New Dataset



N01_049	CCATATACCTCCCGCCAGCCTCATCGC
N01_050	GTGCGCGCCTCCCGCCAGCCTCATCGC
N01_051	GTGCGCGCCTCCCGCCAGCCTCATCGC
N01_052	GTGCGCGCCTCCCGCCAGCCTCATCGC
N01_053	GTGCGCGCCTCCCGCCAGCCTCATCGC
N01_054	NTGCGCGCCTCCCGCCAGCCTCATCGC
N01_055	CCATATATGATTTATTGAATCTTGGAG
N01_056	CCATATATNTCCCGCCAGCCTCATCGC
N01_057	NCATATATGATTTATTGAATCTTGGAG
N01_058	GTGCGCGCCTCCCGCCAGCCTCATCGC
N01_059	NTGCGCGCCTCCCGCCAGCCTCATCGC
N01_060	CCATATATGNCCCGCCAGCCTCATCGC
N01_061	CCATGTATGATTTATTGAATCTTGGAG
N01_062	GTGCGTACCTCCCGCCAGCCTCATCGC
N01_063	GTGCGCGCCTCCCGCCAGCCTCATCGC
N01_064	NCATATATGTCCCGCCAGCCTCATCGC
N01_065	NTGNNTATGATTTATTGAATCTTGGAG
N01_066	NNGCGCGCCTCCCGCCAGCCTCATCGC
N01_067	NNGCGCTTGATTTATTGAATCTTGGAG
N01_068	CCATACGCCTCCCGCCAGCCTCATCGC
N01_069	CCATATACCTCCCGCCAGCCTCATCGC

Prediction Accuracy

Summary

- Statistics and machine learning are trying to solve problems by using data, but the goal and strategy is different.

- Especially genome analysis

Gene Data → To find gene function, evolutionary mechanism···etc

・What does data mean ?

・What can we detect from models ?