Data Analytics and Mathematical Modeling for Psychiatric Diagnosis in a Big Data Processing Environment

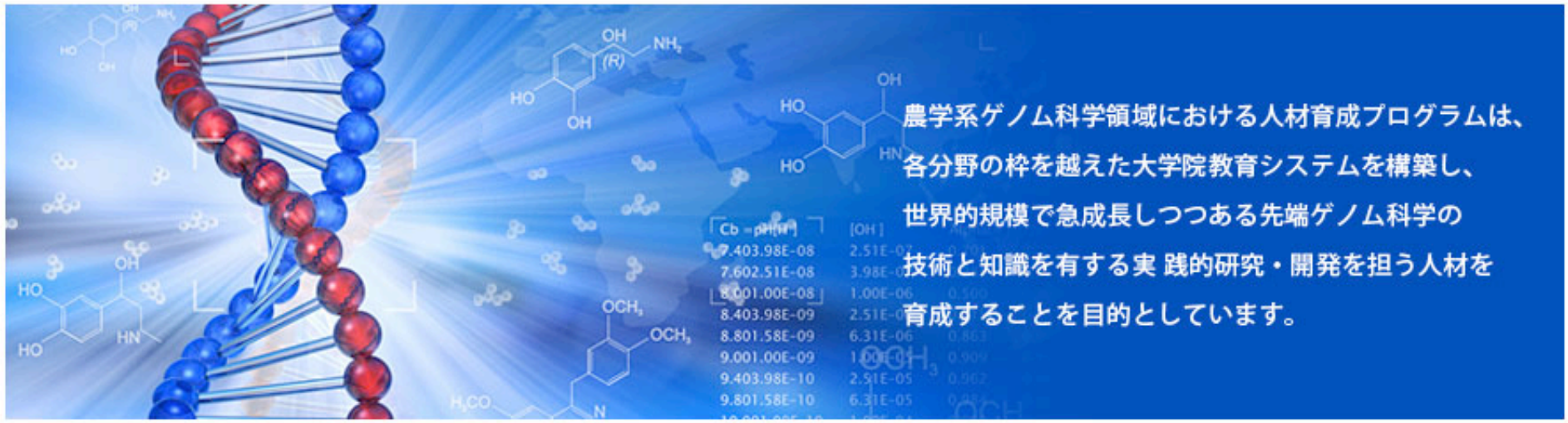
Kazuo Ishii, PhD, Professor of Genomic Sciences

Kazuo Ishii[1*], Shusuke Numata[2], Makoto Kinoshita[2]and Tetsuro Ohmori[2]

1 Tokyo University of Agriculture and Technology, Tokyo, Japan

2 University of Tokushima School of Medicine, Tokushima, Japan

*E-mail: kazuoishii@cc.tuat.ac.jp



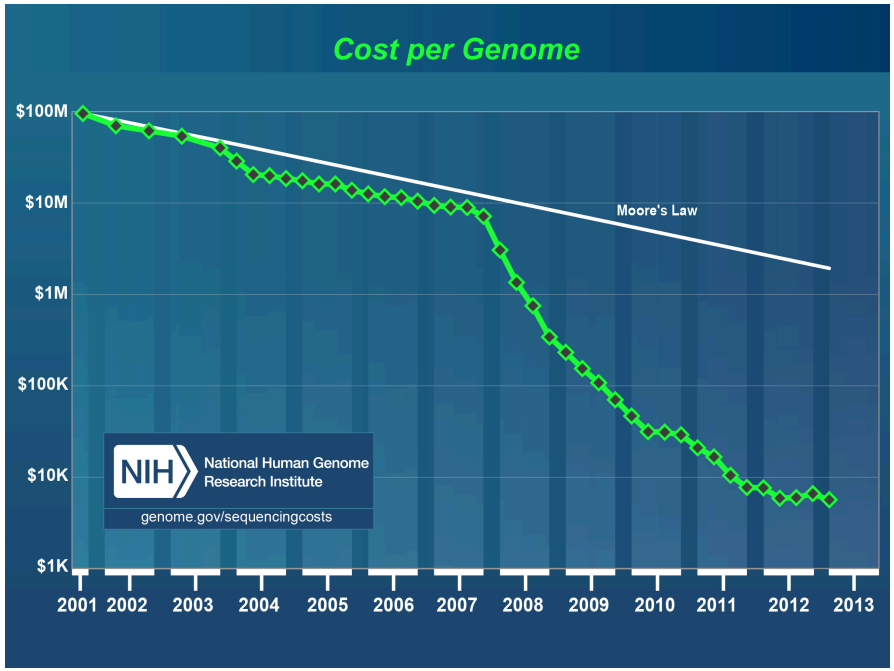
Agenda

- Back ground
- Research Aim and Target
- Research Scheme
- Practical Case Study
- Summary

Back ground

Era of Genomic Big Data

- **Genomic Big Data production by Next Generation Sequencing Technologies is increasing year after year.**







Next Generation Sequencers

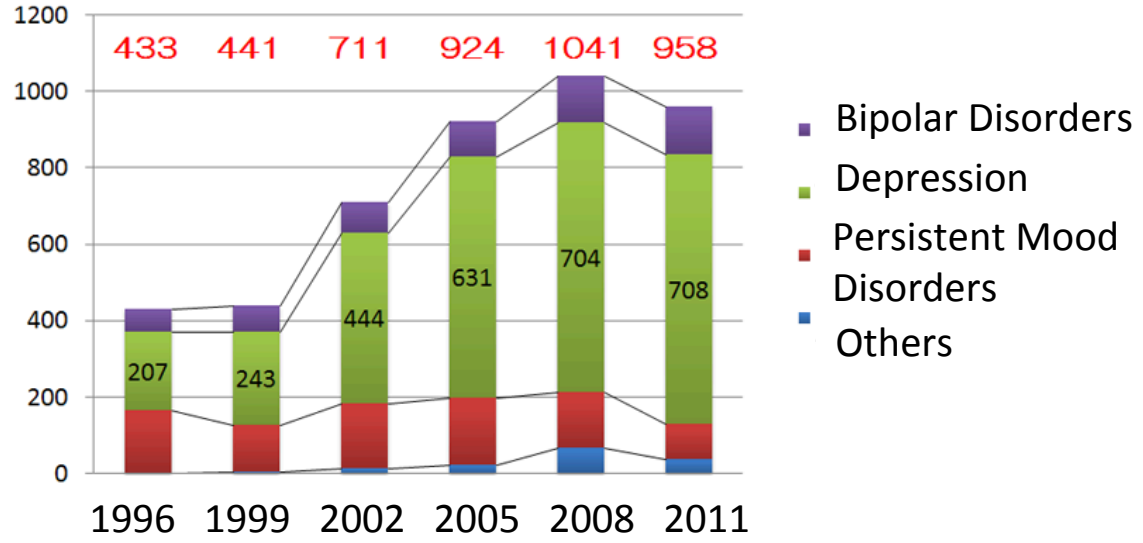
Back ground

Mental Health

- Neuropsychiatric Disorders, such as depression, bipolar disorders are increasing year after year.
- But, no effective evidence based-diagnosis.
- Big Data-based new diagnosis system is expected to provide revolutionary innovation in mental health.

Increasing Number of Mental Illness

(x 1000 persons)



From Japanese Government Documents (2012)

Research Aim and Target

- **Aim:**

Development of Big Data Mining Method

Development of optimized algorithm and mathematical modeling methods for genomic big data; from

500,000 - 10,000,000 explanatory variables

(biological markers)

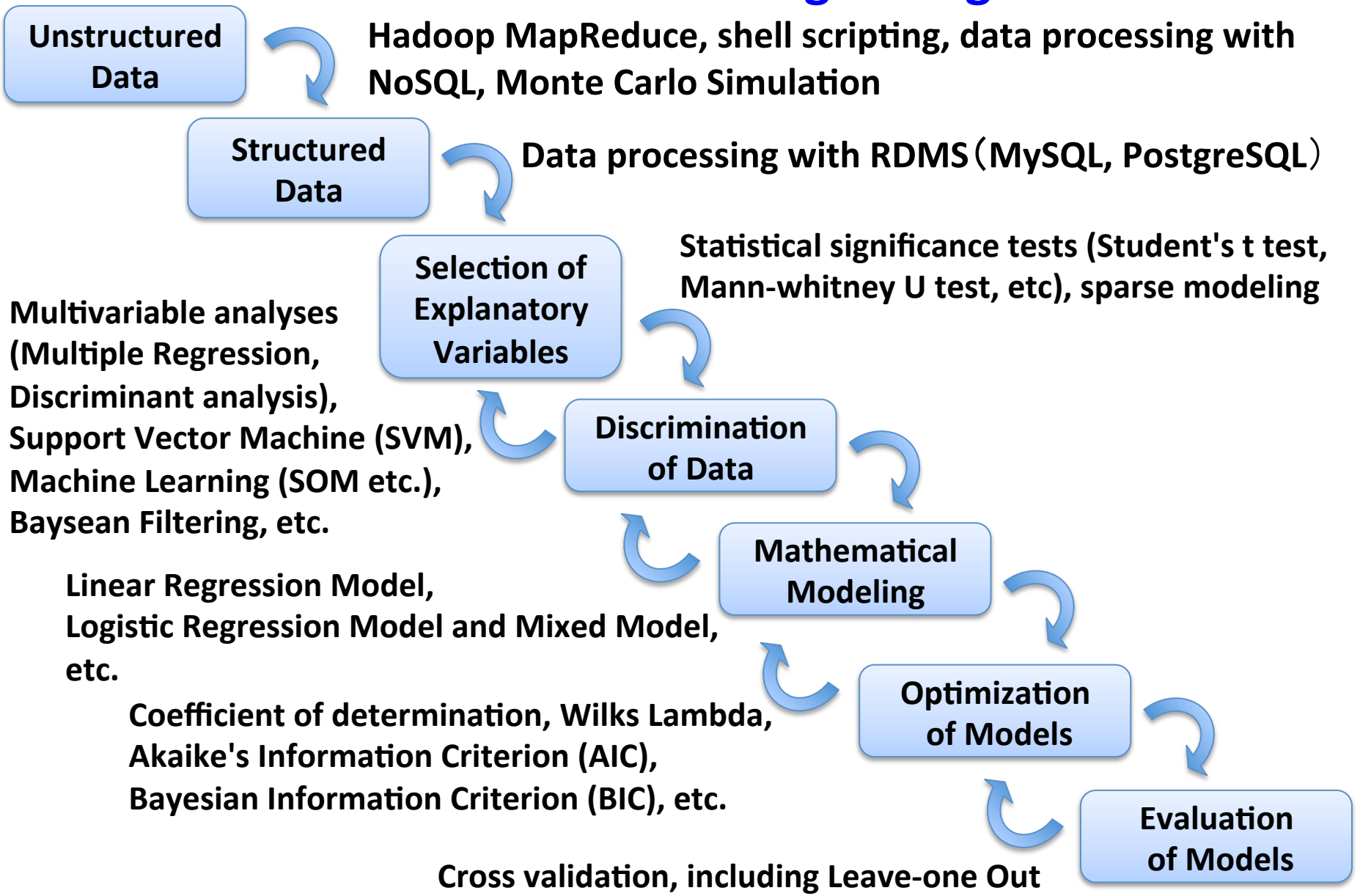
- **Target (Data is provided by Tokushima Univ.)**

Diagnosis system for three major mental disorders;
depression, etc

Research Scheme

Overview of Research Process

Mathematical Modeling for Big Data



Research Scheme

HPC and Cloud (Amazon)

Powerful and High Performance

- HPC

Very Large Memory and Many Core CPUs

4TB Memory, 80 core CPU

- Cloud (Amazon)

Many Core CPUs

but memory is not so large

244 GB Memory, 32 core CPU x n

More core CPUs available by using many instances.

Platform should be selected based on its purpose

Example of Methylation Calling Software

- **Bismark** – Mapping with bowtie
- **PASH** – small memory and fast
- **BSMAP** – Mapping with SOAP
- **Methylcoder**
- **BS-Seq** – for plants
- **Kismeth** – for plants, web-based

HPC: High Performance Computing



HP ProLiant DL900 Servers

HP ProLiant DL980 G7 Server

Starting from **$ 33,585**
Lease for as low as $ 877
Financial Services ›

Select a model

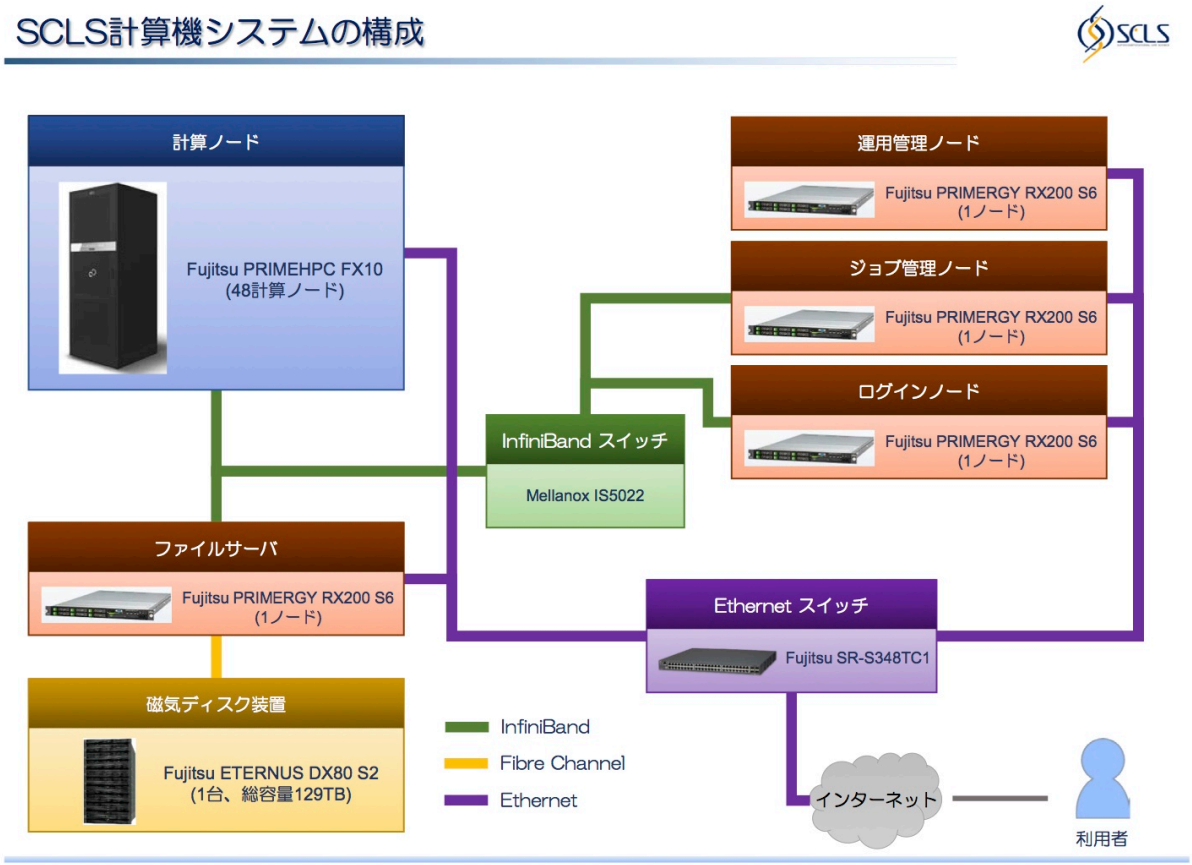
Customize and buy

Like 34

Learn more ›

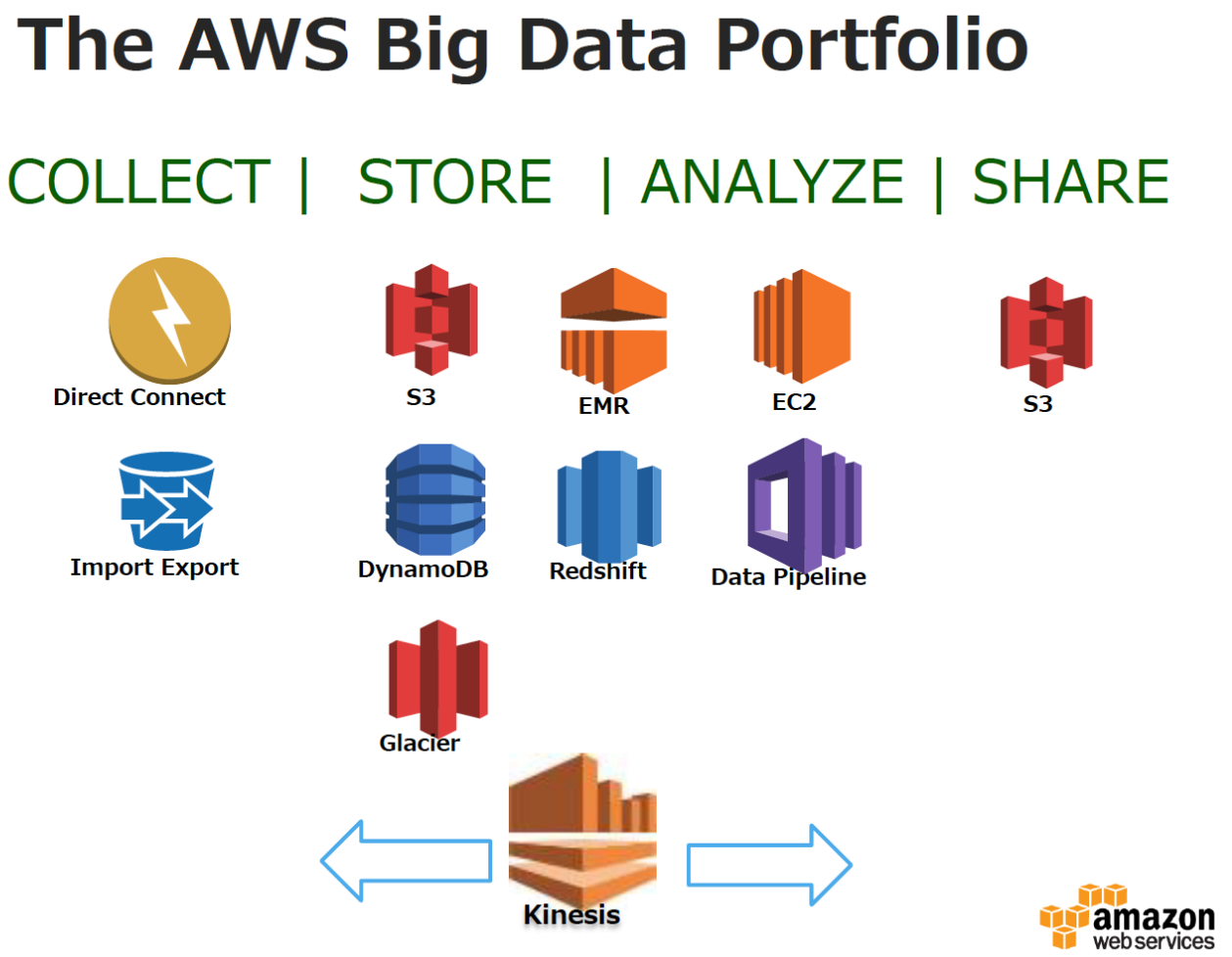
4TB Memory, 80 Core CPU

HPC: RIKEN “K Computer” Compatible Computer: SCLS supercomputer system



1.5TB Memory, 48 Nodes

Utilization of Amazon Web Services (AWS)



Provided from Mr.Y. Yoshiara, AWS JAPAN

Available Open Source Tools on Amazon Web Services

- NCBI BLAST
- Crossbow
- CloudBurst
- Myrna
- Clovr
- BioPerl Max
- VIPDAC
- Superfamily
- Cloud-Coffee
- BioNimbus
- GMOD
- CloudAligner
- CRdata
- SeqWare
- Blend
- StormSeq
- BioConductor

MIT StarCluster	Sun Grid Engine	Condor
Torque	Slurm	Rocks
Chef	Puppet	

Get links to AMIs at:
https://github.com/mndoci/mndoci.github.com/wiki/Life-Science-Apps-on-AWS

Provided from Mr.Y. Yoshiara, AWS JAPAN

Research Scheme

Platform should be selected based on its purpose

- Data Analysis of Methyl-Seq requires extremely large memory
- ex.

BisMark (Methylation site calling soft)

-> 870 GB in one process

R -> 900 GB in one process

requires about 1TB memory

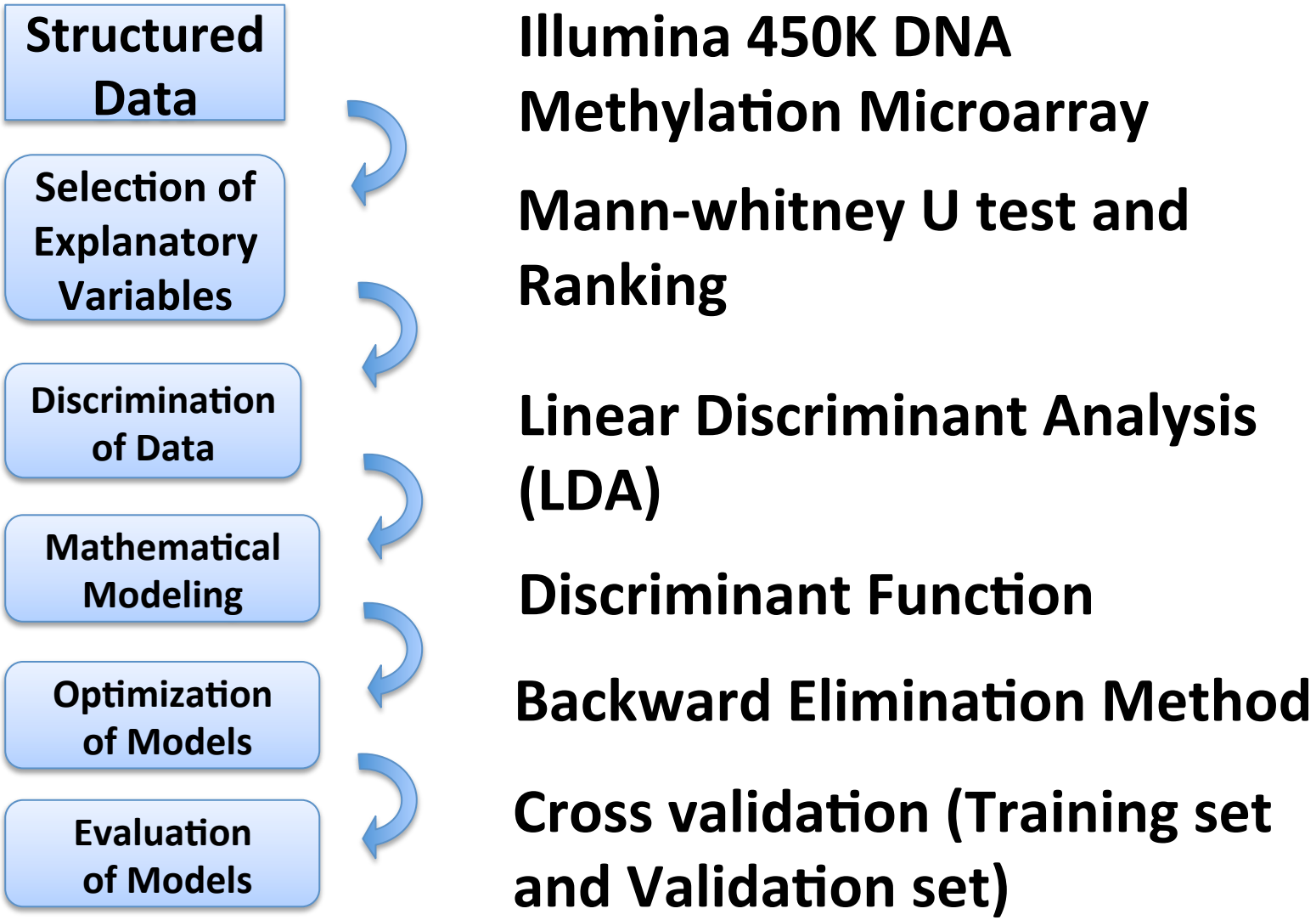
Amazon – Cloud could not analyze methylation calling with BisMark

Practical Case Study

Here, we only show the case of 450K MicroArray in this presentation.
Results of NGS will be shown elsewhere.

Research Process in This Method

Mathematical Modeling for Big Data



DNA Methylation rate does not show a normal distribution

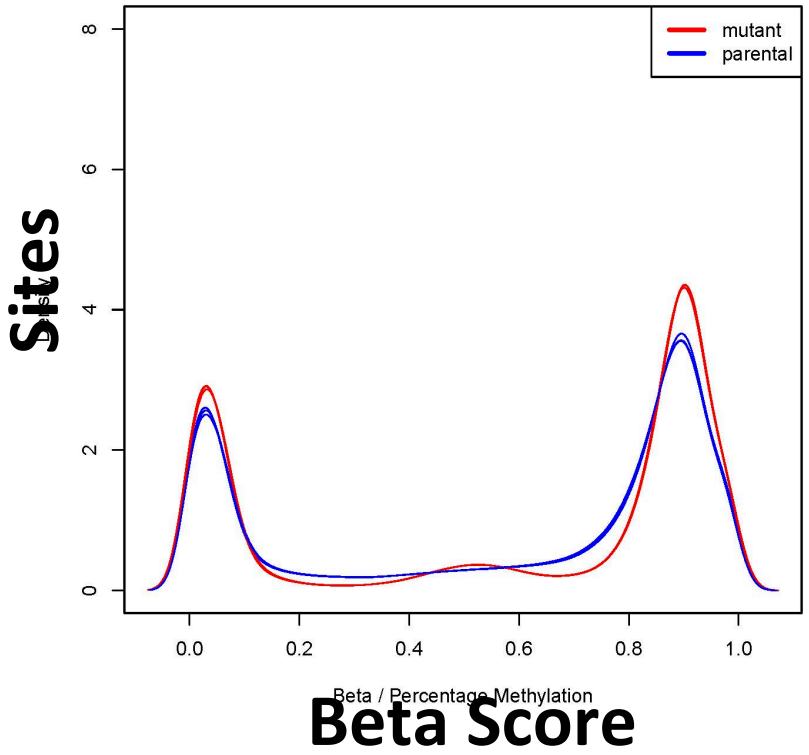
Beta-value for an i^{th} interrogated CpG site is defined as:

$$Beta_i = \frac{\max(y_{i,methy},0)}{\max(y_{i,unmethy},0)+\max(y_{i,methy},0)+\alpha}$$

where $y_{i,menty}$ and $y_{i,unmenty}$ are the intensities measured by the i^{th} methylated and unmethylated probes, respectively

Both Next Generation Sequencing Data and Methylation MicroArray Data

DNA Methylation rate does not show a normal distribution



No equal variances

Range:
0 <= Beta <= 1

Both Next Generation Sequencing Data and Methylation MicroArray Data

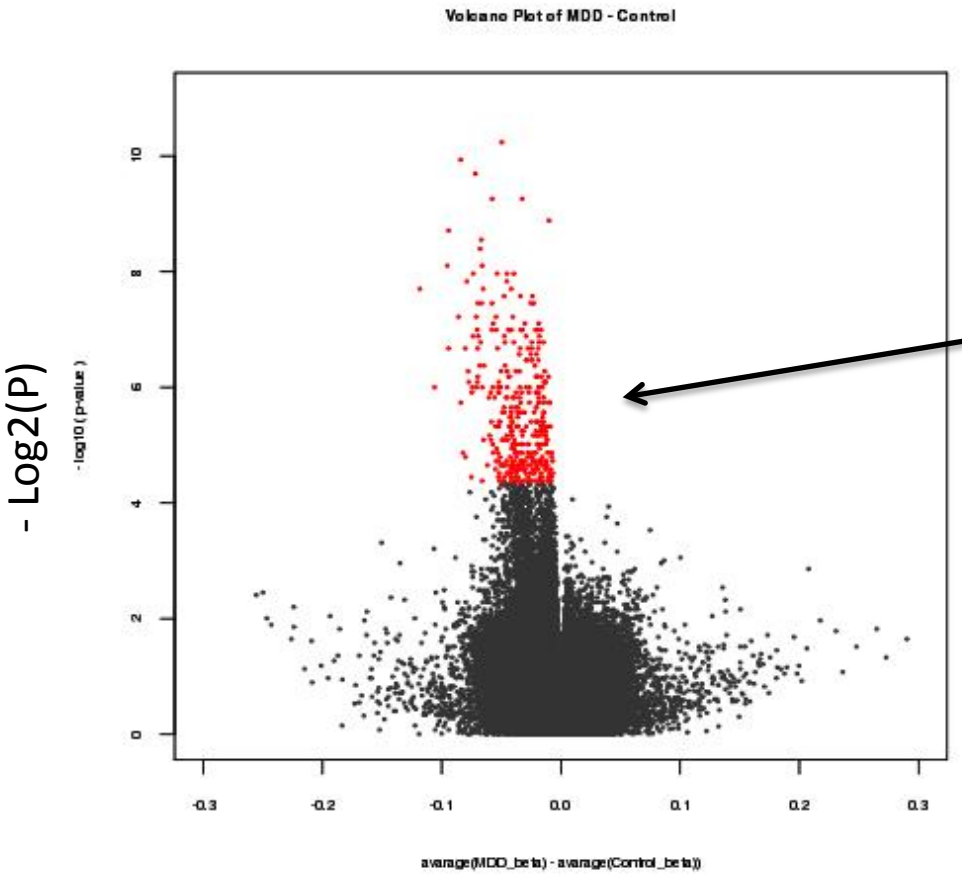
Protocol Exchange (2014) doi:10.1038/protex.2014.002

Ratio Data

- The distribution of errors does not show normal distribution *(binominal distribution)*
- Non equal variances $(\sigma^2 = npq)$
- Dependent variables have upper and lower limits (0 <= Beta <= 1)
- Sometimes shows overdispersion

So the significance should test by Non Parametric testing

Mon Parametric Test is Required
Mann–Whitney U test

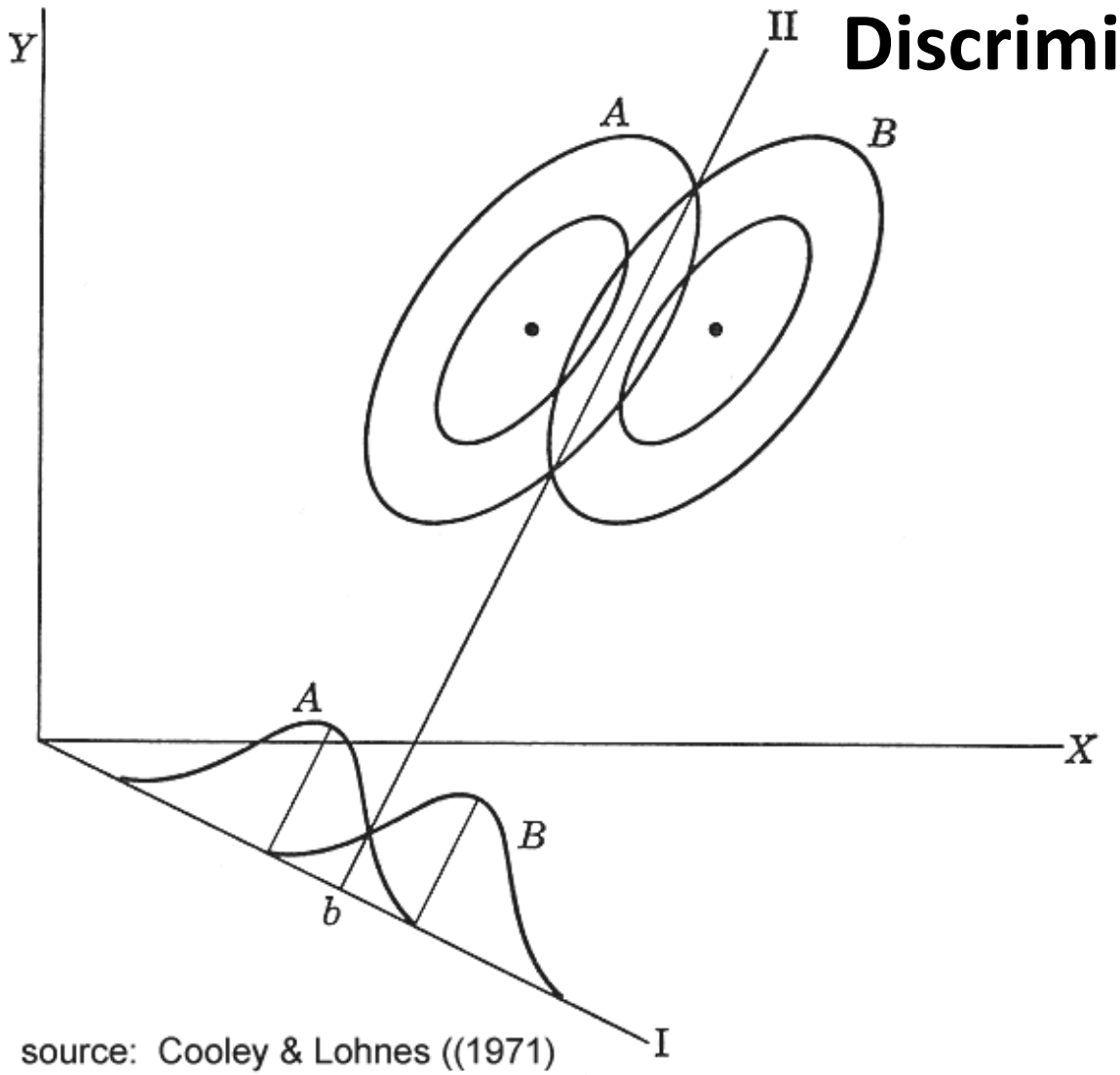


Selected Sites

20 patients and
19 healthy volunteers

This is the example of one neuropsychiatric diseases.
20 patients and 19 healthy volunteers were tested with 500, 000 explanatory variables.

Linear Discriminant Analysis



Discriminant Function

source: Cooley & Lohnes ((1971)

Figure 9.1.

Discriminant Function

Discriminant Score

$$f_{km} = u_0 + u_1 X_{1km} + u_2 X_{2km} + \ldots + u_p X_{pkm}$$

where

f_{km} = the value (score) on the canonical discriminant function for case m in the group *k*.

X_{ikm} = the value on discriminant variable *Xi* for case m in group *k*; and

u_i = coefficients which produce the desired characteristics in the function.

Evaluation of the Discrimination Sensitivity and Specificity

Sensitivity *= true positives / (true positive + false negative)*

= Diagnosed as patients / Patients

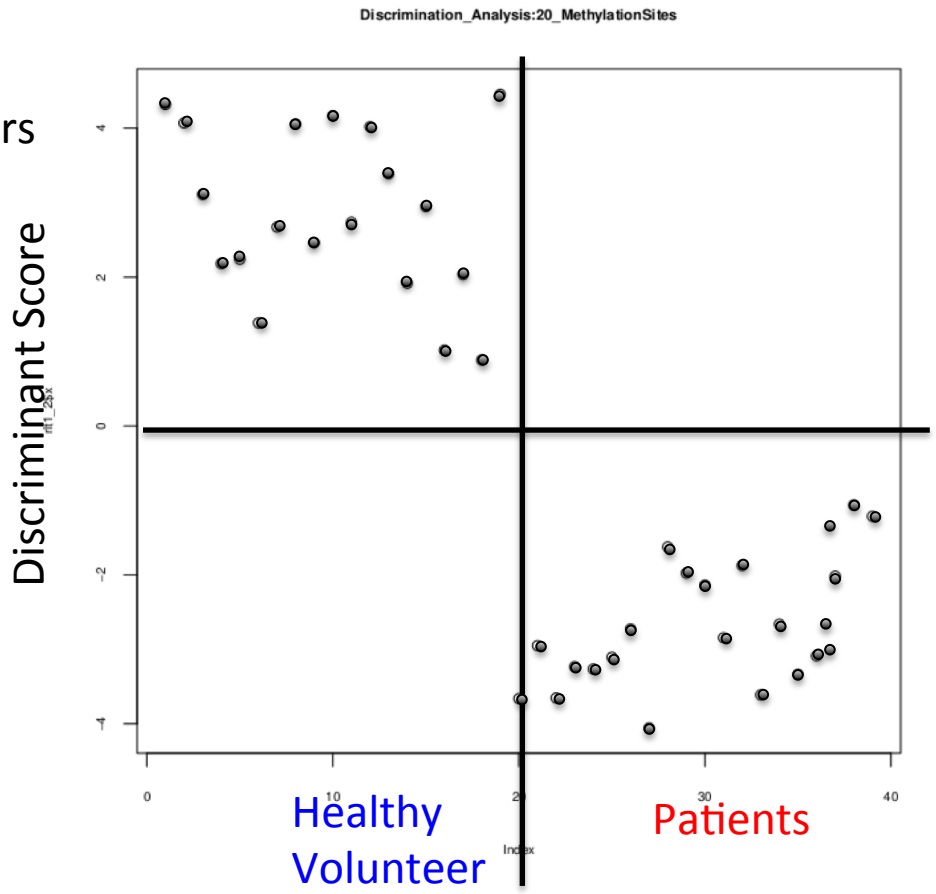
Specificity *= true negatives / (true negative + false positives)*

= Diagnosed as non patients / Healthy Volunteers

Discriminant Analysis of a Psychiatric Disorder with DNA Methylation Markers in a **Training group**

Discriminant analysis with 20 patients and 19 healthy volunteers (Training group)
With methylation rate of DNA Markers top20 ranked by Mann-whitny U test

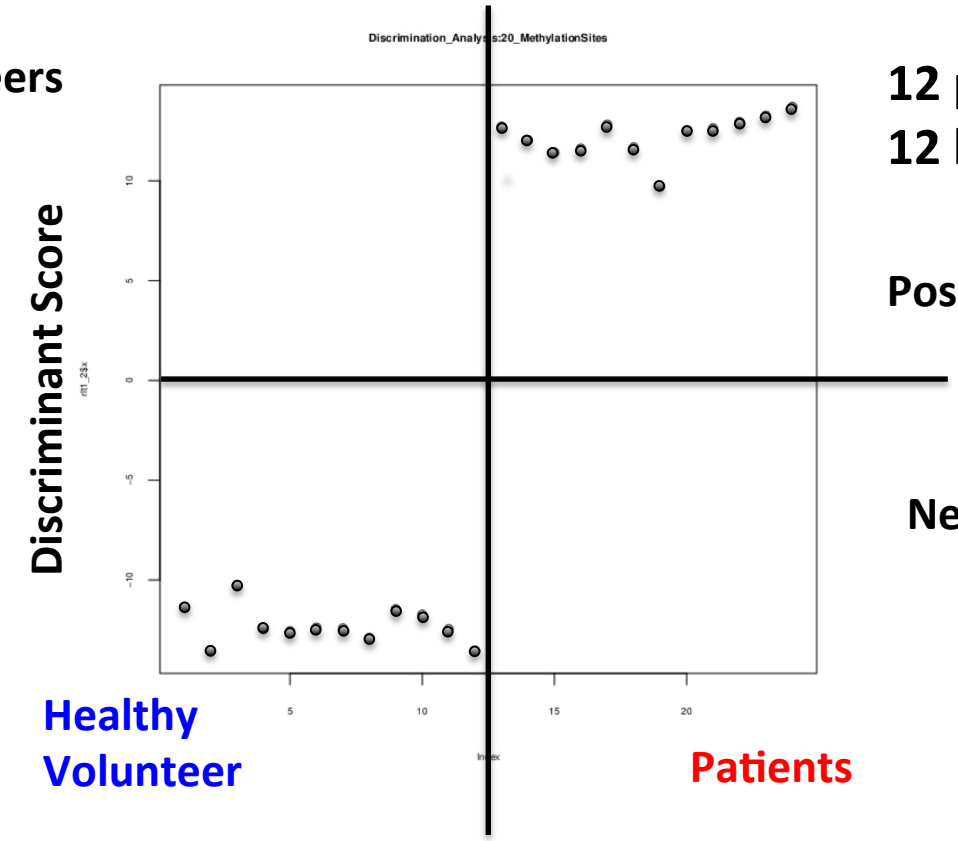
20 patients and
19 healthy volunteers



Discriminant Analysis of a Psychiatric Disorder with DNA Methylation Markers in a Validation group

Discriminant Analysis with 12 patients and 12 healthy volunteers (Validation group)
With Methylation rate of DNA Markers top20 ranked by Mann-whitny U test

12 patients and
12 healthy volunteers



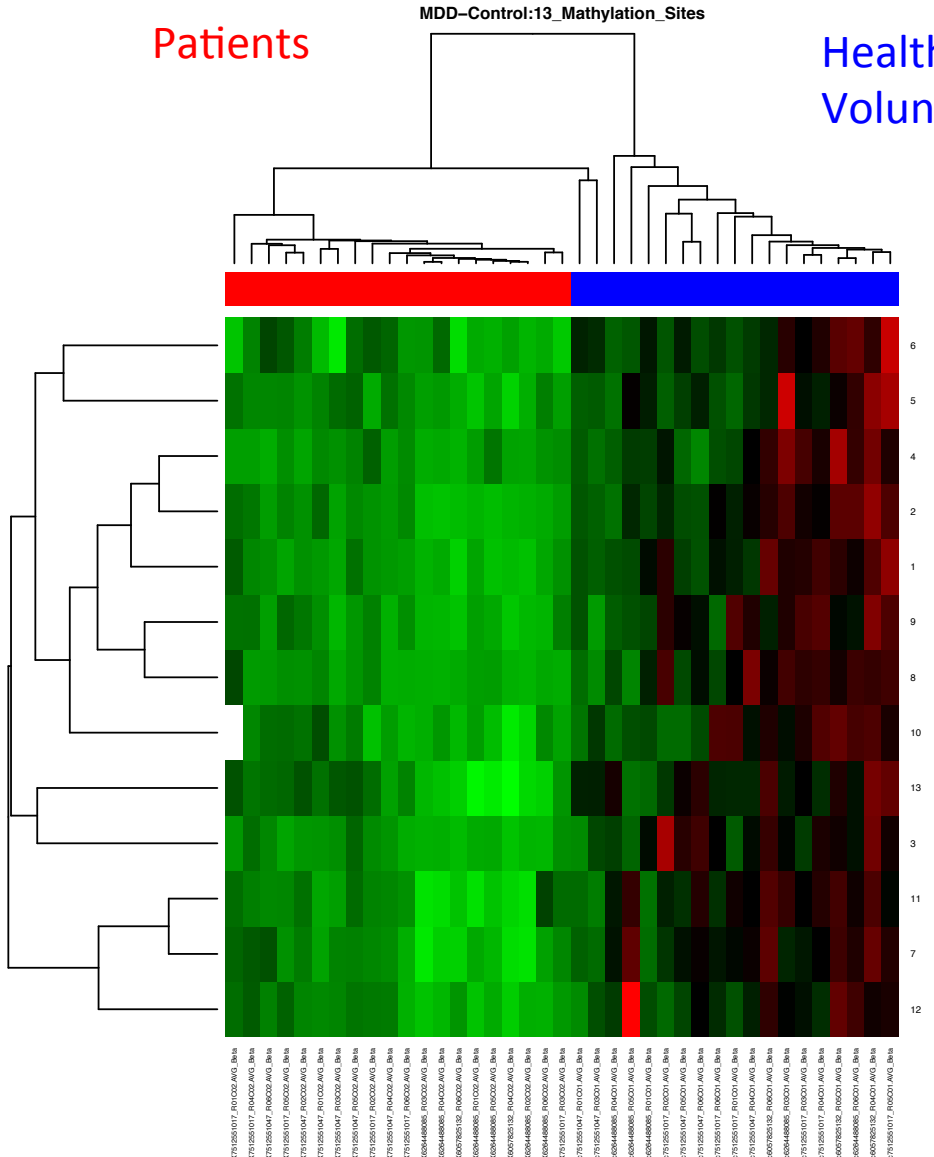
12 patients and
12 healthy volunteers

Positive

Negative

The discriminant function was reconstructed for evaluation of variables.

Cluster Analysis of a Psychiatric Disorder with DNA Methylation Markers in a Training group



20 patients and
19 healthy volunteers

Summary

- **Big Data processing environment should be selected based on its performance and purpose of data analysis**
- **Multivariable diagnosis methods** using DNA methylation ratio works well for **Diagnosis of Psychiatric Diseases**
- **Selection with a non parametric test** and **multivariable analysis** is extremely effective