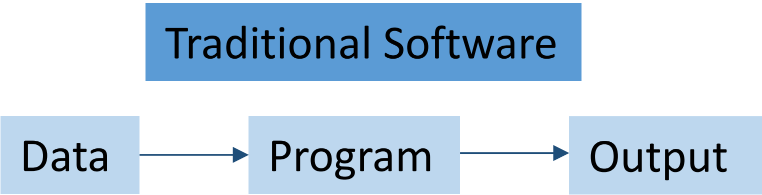
Supervised learning methods for gut microbiota signature identification

Sam Zhu
Third year Ph.D. student
Supervisor: Professor Margaret IP
Joint Graduate Seminar
Department of Microbiology
7th December 2017

“What is machine learning?”

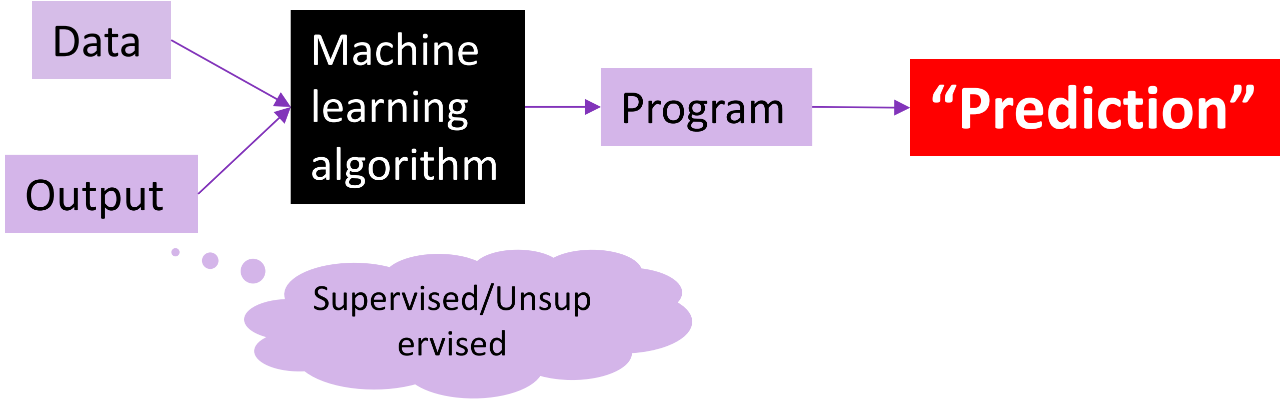


Traditional Software

Data

Program

Output



Data

Output

Machine learning algorithm

Program

“Prediction”

Supervised/Unsupervised

Characters of microbiome data:

- High dimension (OTUs, MLGs)
- Labeled (Clinical features)
- Predictions

Common analysis methods:

- α and β diversity
- Classical statistics testing
- Non-supervised Learning: PCoA, clustering…

Why supervised machine learning?

Applications:

- Text, image classification
- Microarray
- Biological image
- Cancer prediction (susceptibility, recurrence, survival)

Supervised Learning methods:

- Decision trees
- Ensembles (**Random forest**…)
- Native Bayes
- **Linear model**
- Support vector machine
- Neural networks

…….

Basic concepts of modelling

- Training and Testing

- Underfit and Overfit

- Evaluation parameters:
  - **AUC**: area under ROC (receiver operating characteristic)
  - Expect Prediction Error
  - Matthews correlation coefficient (MCC)

Random forest (RF)

- Decision tree
    - Simple, fast, interpretable
    - Overfitting
    - Non-robust

Random forest (RF)

- RF → Bootstrap aggregating (bagging) → Ensemble learning
- Bootstrap sampling
- Random picking features
- Voting for the best
- Application: Microarray

Cell Metabolism

Gut Microbiome-Based Metagenomic Signature for Non-invasive Detection of Advanced Fibrosis in Human Nonalcoholic Fatty Liver Disease

Data:

- Metadata
- Gut microbiome (Metagenome)

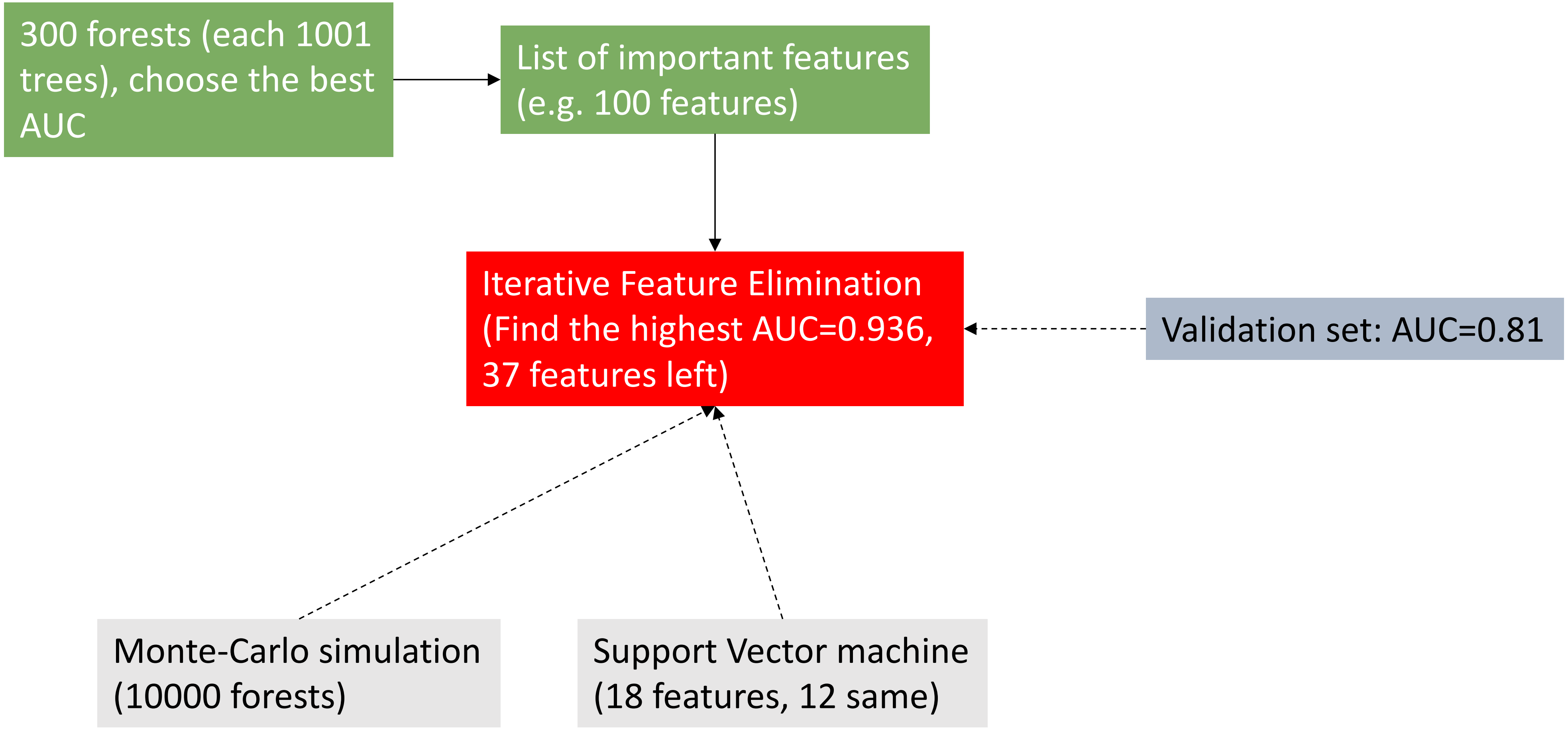
Result:

- Biopsy

Target:

- Predicting Advanced Fibrosis in non-alcoholic fatty liver disease (NAFLD)

Methods and results



300 forests (each 1001 trees), choose the best AUC

List of important features (e.g. 100 features)

Iterative Feature Elimination (Find the highest AUC=0.936, 37 features left)

Validation set: AUC=0.81

Monte-Carlo simulation (10000 forests)

Support Vector machine (18 features, 12 same)

Rohit Loomba, Victor Seguritan, 2017

Colorectal cancer (CRC) and gut microbiome

Year	Author	Data type	Method	AUC (Validation set)	Features selected
2014	Zackular	16S	Bayesian	0.798	6 OTU
2014	Zellar	16S	LASSO	0.84 (0.85)	Ranking features
2015	Qiang feng	metagenome	**Random Forest**	**0.98 (0.96)**	15
2017	Ai Luoyan	16S	**Random Forest**	**0.94 (0.86)**	--

- Four studies since 2014
- Ai has evaluated some supervised classifiers, with random forest and Bayes Net the best
- RF and LASSO found lack of ***Streptococcus salivarius*** (Bacteriocin-like Inhibitory Substances)

Limitations of Random Forest:

Not explicitly do feature selection

“The purpose of models is not to fit the data but to sharpen the question.”

- S Karlin, 11 th R A Fisher Memorial Lecture, 1983

LASSO
(Least absolute shrinkage and selection operator)

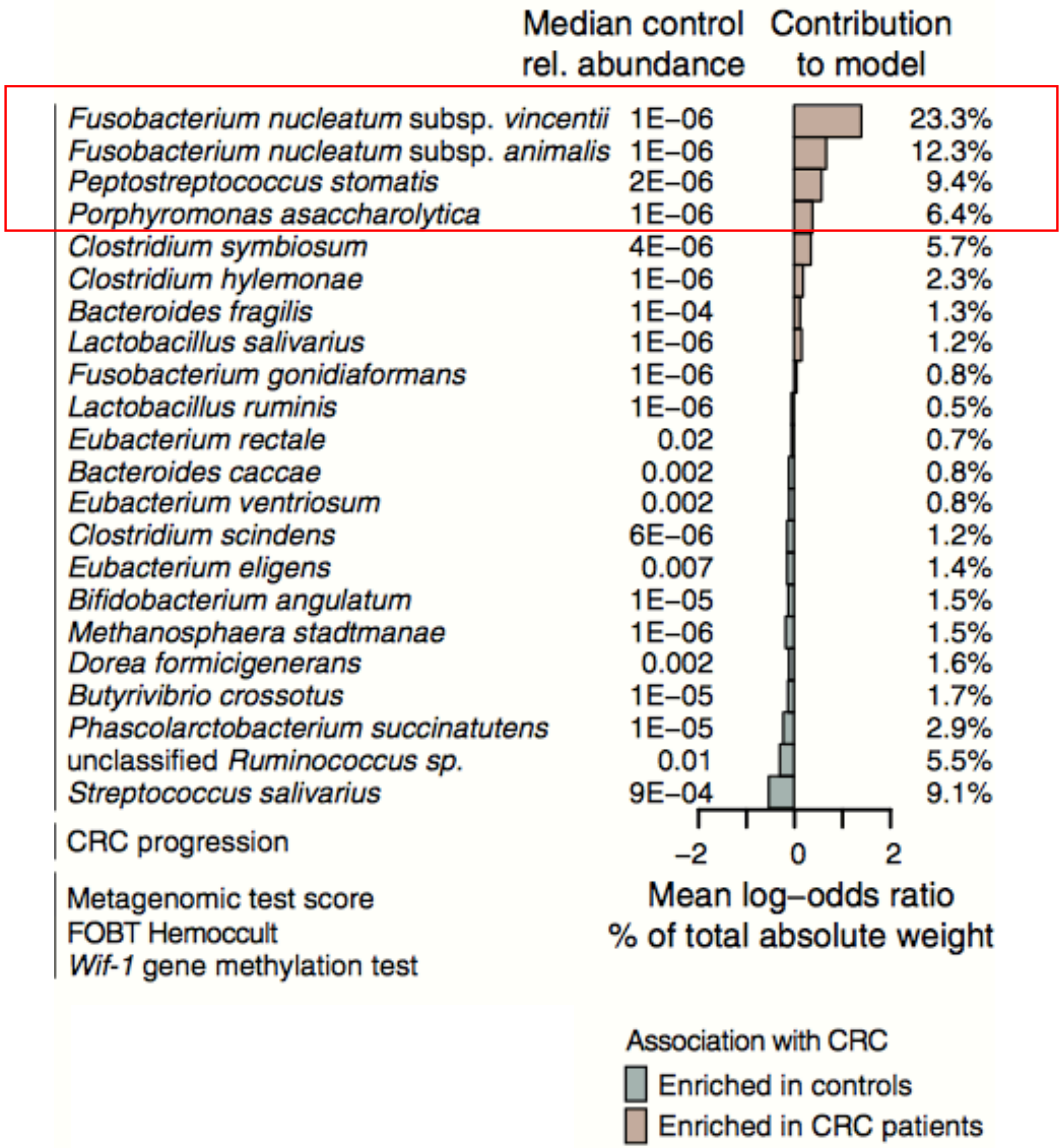
Elastic Net Regularization

- Coefficient problem in Linear or Logistic regression
- ‘Ridge’ & ‘LASSO’ **Penalty**
- ‘LASSO’ Penalty: Set many coefficients to zero, catch ‘Big Fish’

Potential of fecal microbiota for early-stage detection of colorectal cancer

- 10 times resampling
  10-fold cross-validation
- 100 LASSO models, existed in >50%

51%



	Median control rel. abundance	Contribution to model
Fusobacterium nucleatum subsp. *vincentii*	1E−06	23.3%
Fusobacterium nucleatum subsp. *animalis*	1E−06	12.3%
Peptostreptococcus stomatis	2E−06	9.4%
Porphyromonas asaccharolytica	1E−06	6.4%
Clostridium symbiosum	4E−06	5.7%
Clostridium hylemonae	1E−06	2.3%
Bacteroides fragilis	1E−04	1.3%
Lactobacillus salivarius	1E−06	1.2%
Fusobacterium gonidiaformans	1E−06	0.8%
Lactobacillus ruminis	1E−06	0.5%
Eubacterium rectale	0.02	0.7%
Bacteroides caccae	0.002	0.8%
Eubacterium ventriosum	0.002	0.8%
Clostridium scindens	6E−06	1.2%
Eubacterium eligens	0.007	1.4%
Bifidobacterium angulatum	1E−05	1.5%
Methanosphaera stadtmanae	1E−06	1.5%
Dorea formicigenerans	0.002	1.6%
Butyrivibrio crossotus	1E−05	1.7%
Phascolarctobacterium succinatutens	1E−05	2.9%
unclassified *Ruminococcus sp.*	0.01	5.5%
Streptococcus salivarius	9E−04	9.1%

Georg Zeller, Julien Tap, 2014

Clostridium difficile

1. Pre-FMT(Fecal Microbiota Transplant) microbiota & clinical response to an FMT
2. Post-FMT microbiome & additional FMTs

Pre-FMT AUC=0.865

OTU	Family	Genus	β
7	*Lactobacillaceae*	*Lactobacillus*	349
3	*Streptococcaceae*	*Streptococcus*	-304

Post-FMT AUC=0.961

OTU	Family	Genus	β(SE)
5	*Enterococcaceae*	*Enterococcus*	98.5
17	*Bacteroidaceae*	*Bacteroides*	109

Stephen T. A. Rush, PhD Thesis, 2017

Performance of different classifiers

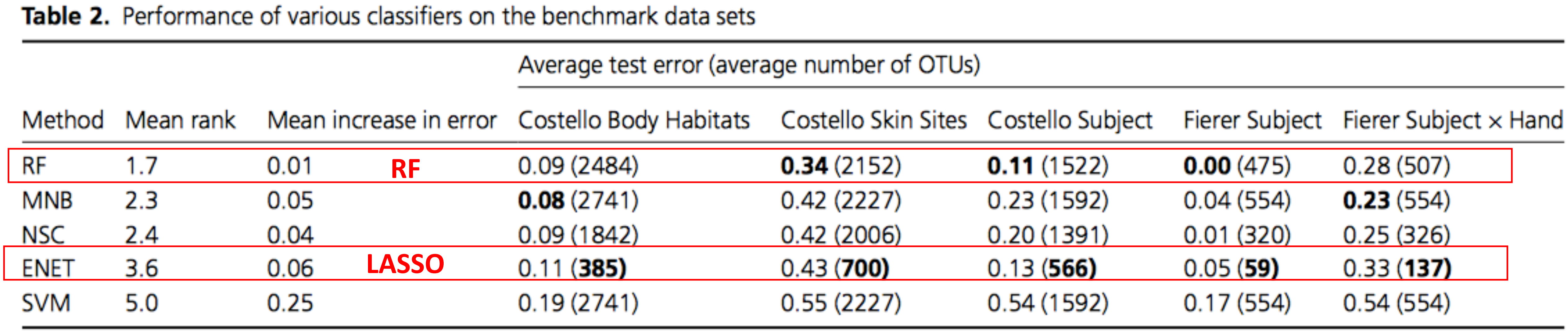


Table 2. Performance of various classifiers on the benchmark data sets

			Average test error (average number of OTUs)				
Method	Mean rank	Mean increase in error	Costello Body Habitats	Costello Skin Sites	Costello Subject	Fierer Subject	Fierer Subject × Hand
RF	1.7	0.01	0.09 (2484)	**0.34** (2152)	**0.11** (1522)	**0.00** (475)	0.28 (507)
MNB	2.3	0.05	**0.08** (2741)	0.42 (2227)	0.23 (1592)	0.04 (554)	**0.23** (554)
NSC	2.4	0.04	0.09 (1842)	0.42 (2006)	0.20 (1391)	0.01 (320)	0.25 (326)
ENET	3.6	0.06	0.11 (**385**)	0.43 (**700**)	0.13 (**566**)	0.05 (**59**)	0.33 (**137**)
SVM	5.0	0.25	0.19 (2741)	0.55 (2227)	0.54 (1592)	0.17 (554)	0.54 (554)

RF

LASSO

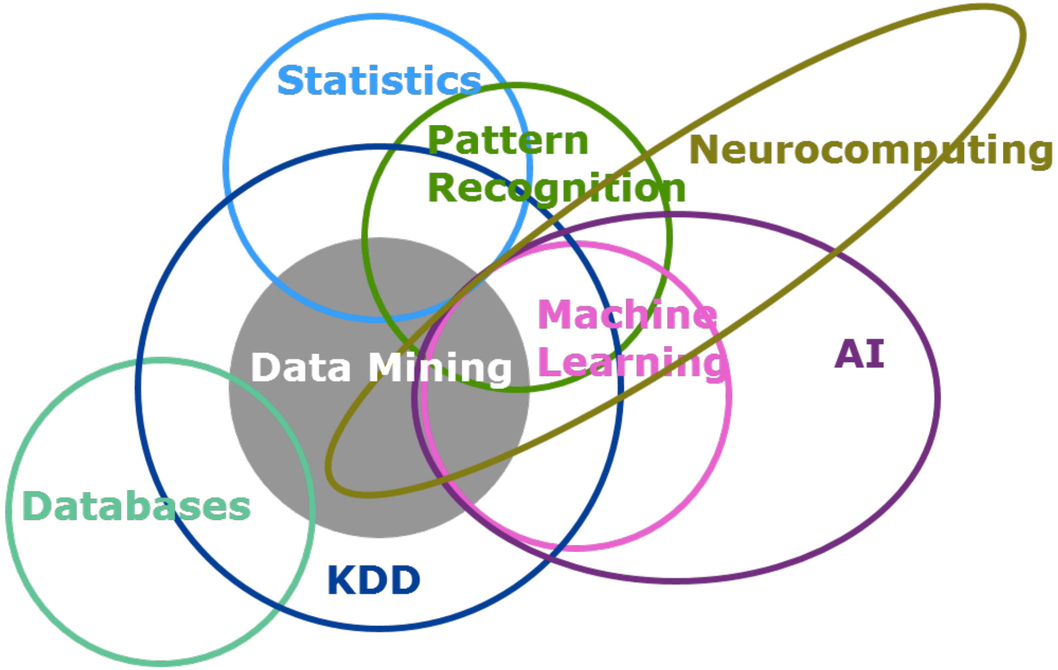
Random forest is a good classifier while LASSO is the first choice of feature selection

Dan Knights, Elizabeth K. Costello, 2011

Careful! Not statistics! 200+ years vs 30 years

- Statistician: Significant! I have p value! I can show you inference!

- Machine learning expert: I don't know what happened, I repeated 1000000 times and that's it.



Machine learning : Based on data (More!), no requirement of data type (tricked by data). Learn fast, handle complex data, accurate.

SAS institute

Thank you