Presentation

Lecture 1: Introduction

Dr. Irene Vrbik
University of British Columbia Okanagan

Welcome!

Welcome to DATA 311: Machine Learning TEST!

> **DATA 311 (3) Machine Learning**
>
> Regression, classification, resampling, model selection and validation, fundamental properties of matrices, dimension reduction, tree-based methods, unsupervised learning. [3-2-0]
>
> Prerequisite: Either (a) one of STAT 205, STAT 230 or (b) a score more than 75% in one of APSC 254, BIOL 202, PSYO 373; and one of COSC 111, APSC 177.

A little about me

- I am currently a Tenure-track Assistant Professor of Teaching
- I have taught a variety of courses (from introductory data science and to graduate courses in statistics) at several institutions (Guelph, McGill, MDS Program)
- I am currently the Data Science Program advisor, Articulation, and curriculum representative

Where can you find me?

Office: SCI 104 **email:** irene.vrbik@ubc.ca

Websites: irene.quarto.pub, irene.vrbik.ok.ubc.ca

Educational Background

1. **McMaster University**, BSc (Mathematics & Statistics)
2. **University of Guelph**, MSc (Applied Statistics)

   Thesis: Using Individual-level Models to Model Spatio-temporal Combustion Dynamics. This involved modelling the spatio-temporal combustion dynamics of fire in a Bayesian framework. *Supervisors:* Rob Deardon and Zeng Feng.
3. **University of Guelph**, PhD (Applied Statistics)

   Thesis: Non-Elliptical and Fractionally-Supervised Classification. This involved model-based classification with a particular emphasis on non-elliptical distributions.
   Supervisor: Paul D. McNicholas.

Experience

Postdoctoral Fellow at McGill University Under the supervision of Dr. David Stephens, this work focused on the statistical and computational challenges associated with analyzing genetic data. It involved clustering and modeling HIV DNA sequences.

Postdoctoral Fellow at UBCO Awarded by NSERC (Natural Sciences and Engineering Research Council of Canada), this research involved collaborations with faculty from several disciplines (eg. Medical Physics, Biology, and Chemistry) and was supervised by Dr. Jason Loeppky.

Instructor at UBCO a three-year contract position in the Department of Computer Science, Mathematics, Physics, and Statistics.

Research Interests

Statistics and Machine Leaning in Curriculum Design

- e.g. topics modeling in Data Science course calendars

Curricular Analytics the systematic analysis and evaluation of educational curricula to gain insights into various aspects of curriculum design, delivery, and assessment.

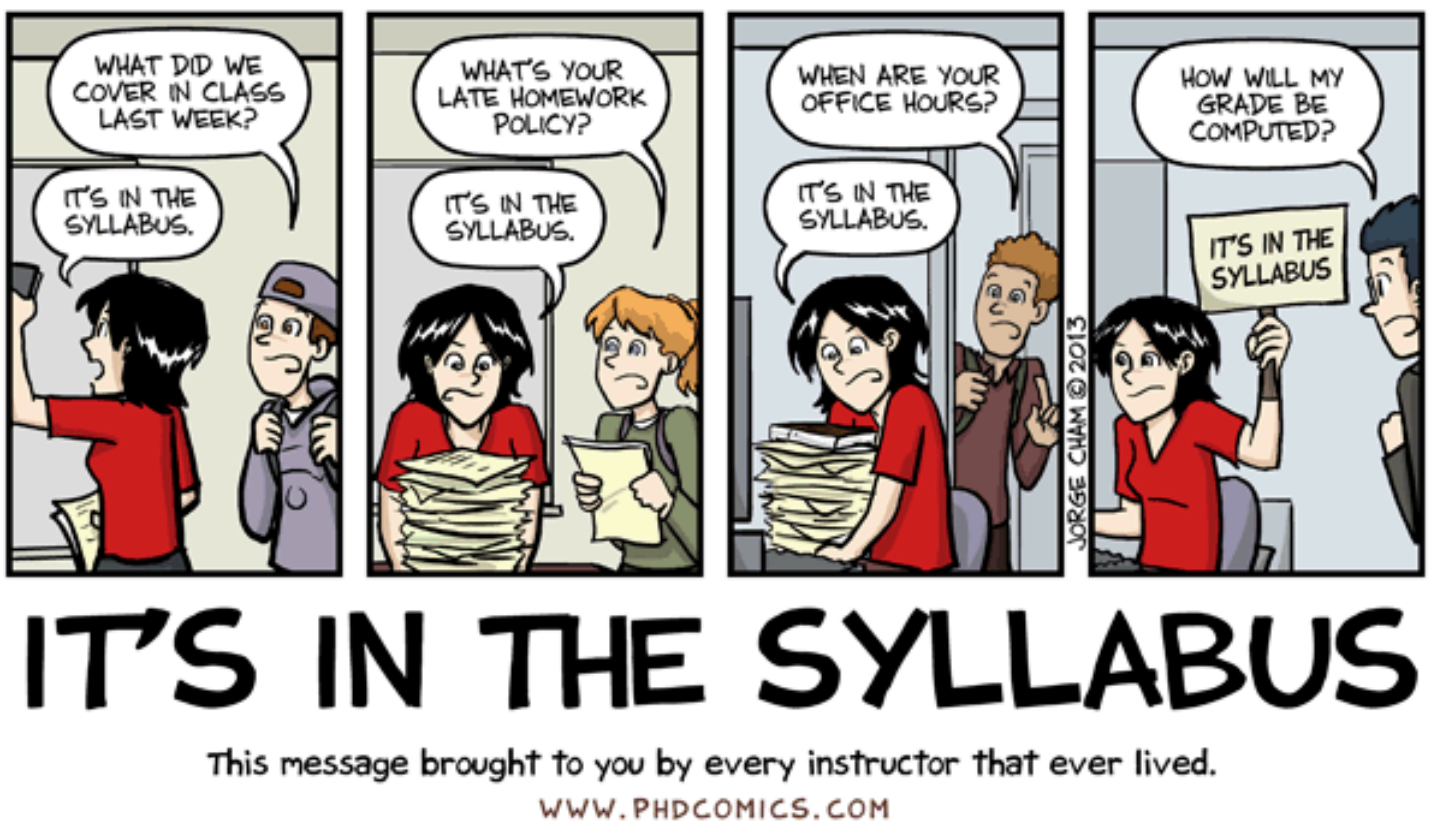
- e.g. metric calculation for various pathways, curriculum visualization, course recommendation systems

Tools for teaching, learning, and technology

- e.g. Prairie Learn: online problem-driven learning system for creating homework and tests

Course Syllabus

The course syllabus is a *dynamic* document which has been posted to Canvas and course website. Many administrative questions can be answered there.



Course Tools

Canvas will be use for most course related material:

- Grades
- Assignments (downloading/submitting)
- Course announcements/discussions
- Supplementary files (eg. data sets, code, etc…)

Lectures

- Lectures will be posted on our course webpage.
- Take time to learn how to:
  - navigate through the slides
  - export to PDF (good for tablet annotation)
  - use the clipboard (example Clipboard code HTML only)

Programming Language

- Any necessary coding will be done in R:
- Relevant code will be posted to *Supplmentary Files* Canvas when necessary. Most relevant pieces will be embedded in the slides and/or included in Labs
- It is also recommended that you complete assignments using Rmarkdown in RStudio.

Clipboard code

How to use the clipboard

Hover over the code block below and you will see a copy icon in the top-right corner:

```
1 head(iris)
```

```
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
```

Warning

This interactivity will only work on the HTML version of the slides, not PDF.

Why R?

Pros

- exposure to R in Statistics prerequisite course
- Rich Ecosystem
- Reproducibility
- Textbook

Cons

- Steep learning curve
- Performance
- Package Quality
- Limited Industry Adoption

Lab Delivery

- Labs will be held in person; students **must** be enrolled in a lab (which cannot conflict with other courses)
- TAs provide guidance on carrying out analyses in R for the techniques discussed in lecture.
- Knowledge of commands and programming techniques will be evaluated throughout the course.
- Follow the instructions carefully and practice skills by completing (and redoing!) labs and assignments

Textbook

The main textbook reference for this course is:

> **ISLR**: An Introduction to Statistical Learning with Applications *in R (Second Edition)*. By: Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani.

This book is available for free at statlearning.com; see resources here.

A secondary (less referenced) textbook is:

> **ESL**: The Elements of Statistical Learning: data mining, inference, and prediction, 2nd edition. By: Hastie, Tibshirani, Friedman.

Can be downloaded for free at: hastie.su.domains/ElemStatLearn.

Lecture format

- Slides will occasionally be supplemented with handwritten material.
- Aside for doodling, substantial written material will be done digitally (on my iPad) and uploaded to Canvas.
- Lectures may also include discussions which you will only gain access to by attending class.
- You will not get the whole story by reading the slides!

Class Etiquette

1. Please be respectful, especially to other students
2. Please be present. Attendance will not be taken, but you are encouraged to come and learn together.
3. Please restrict the use of electronic devices to course related material; other content could be distracting.
4. Please be forgiving; instructors are people too, we *will* make mistakes.

Course Questions

In class

- If you are stuck on a concept during lecture, please feel free to raise your hand and ask for clarification.
- If you are needing help understanding something, chances are, other students are too!
- I will do my best to answer questions on the fly or organize a more thoughtful answer to be presented first thing next class or posted to Canvas.

Course Questions

Outside of class

Outside of class, the general order in which I would suggest you asking course-related questions is:

1. Consult the course syllabus
2. Post your question on the public forum on Canvas*
3. Come see me during student hours or visit your TA during lab (whichever comes first)
4. e-mail (weekdays are best)

Machine Learning

Machine learning (ML) is a subfield of artificial intelligence (AI) that uses algorithms and statistical models to learn from data to perform complex tasks.

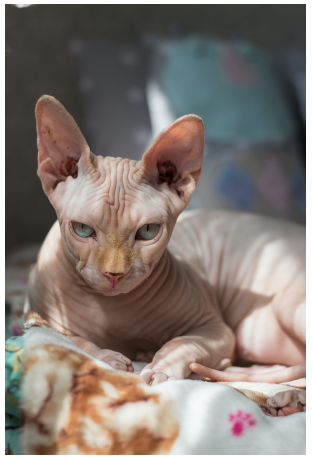
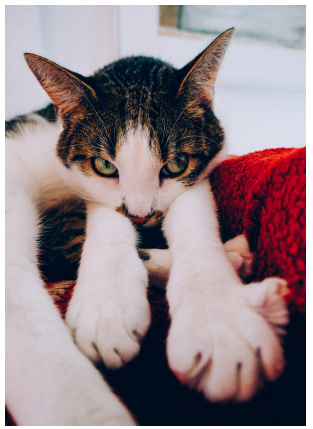
- e.g. recommend TV shows you might like, determine if an e-mail is spam or not, predict the selling price of a home.

How does Machine Learning work

Machine Learning has been described[1] as:

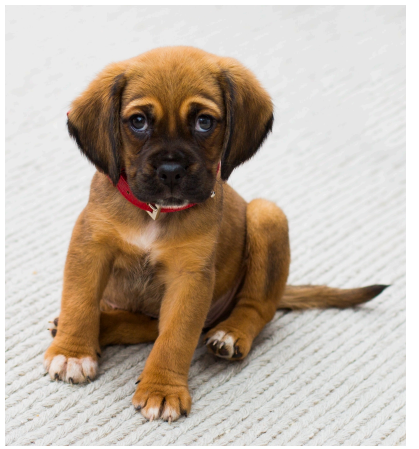
> “The field of study that makes computers capable of learning without being explicitly programmed.”

Cats

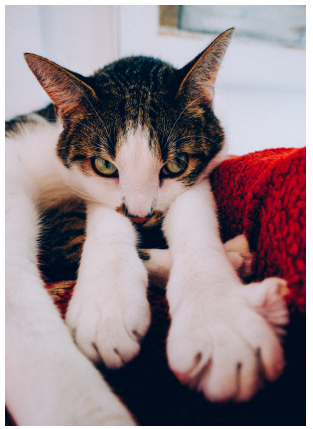




Dogs



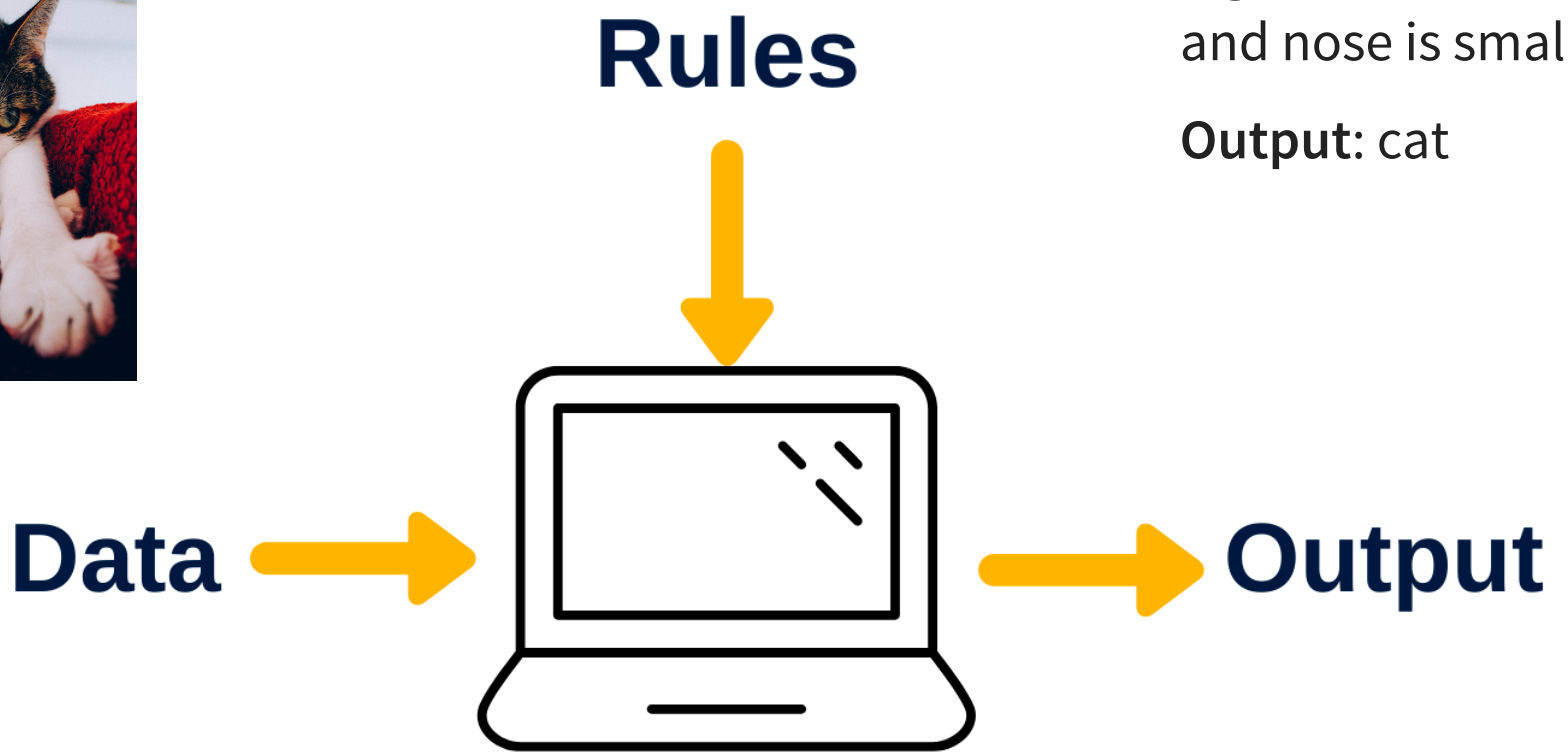



Traditional Programing

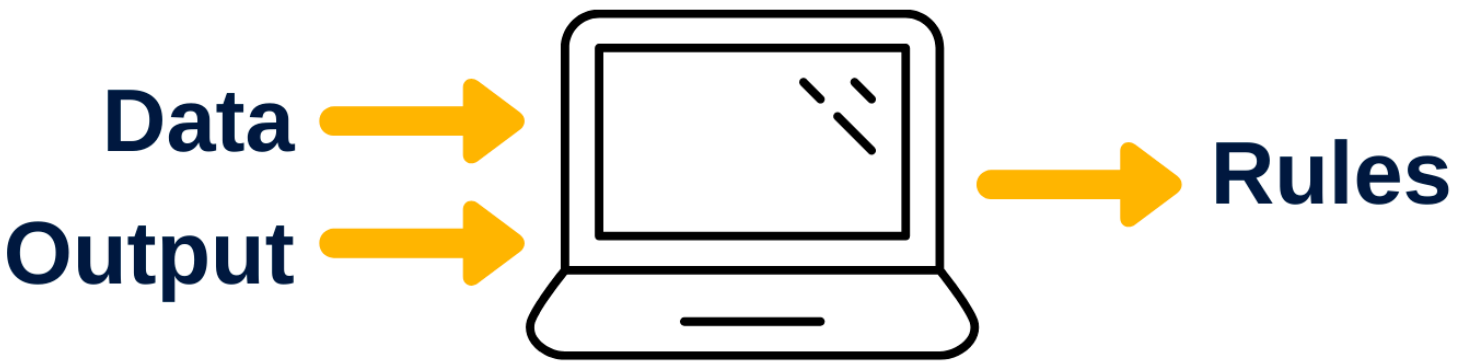
Input: 

e.g. If claws are sharp and nose is small …

Output: cat



Machine Learning



In supervised machine learning the computer must learn these distinguishing patterns for itself in order to determine a set of rules by which future data will be sorted.

General concepts

- To continue with the cats and dogs, the more examples a human is given, the better they would become at distinguishing between the two species.

- The more variety in the samples, the easier it may become in detecting patterns and ultimately predicting the result.

- This process is often iterative.

Why is ML important?

- Many of the statistical techniques you've (probably) learned thus far are either completely inapplicable to much of this data, or only applicable on a subset.

- With growing access to data and computing power, models can be built faster than ever and used in countless fields to gain useful insights.

- This course will guide you through a few of the classical approaches in Machine Learning (ML)

Supervised ML

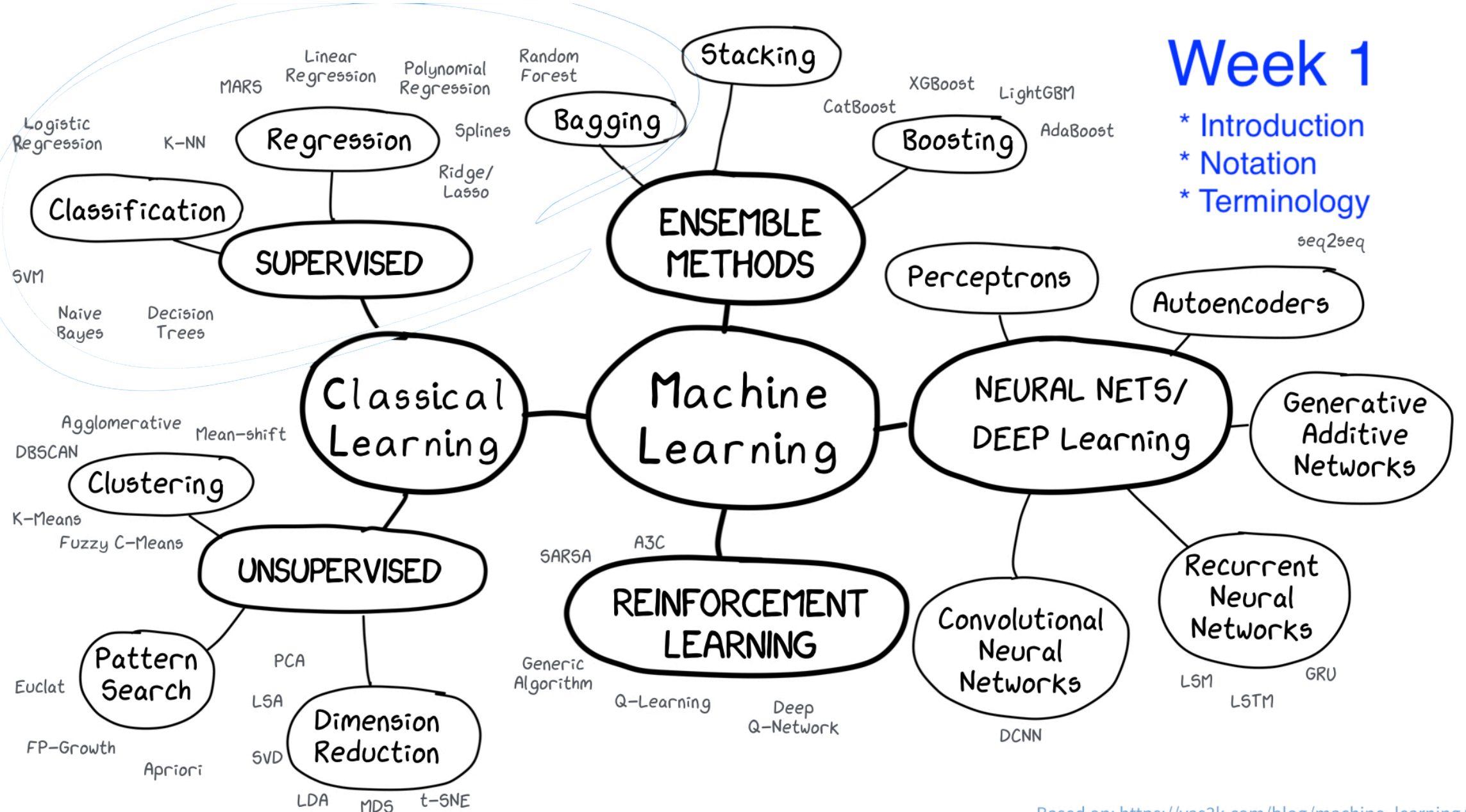
Involves training a model on a labeled examples. There are two main goals:

- **Classification**: Assigning unseen examples into to predefined categories (e.g., spam vs. not spam).
- **Regression**: Predicting continuous values (e.g., predicting house prices based on features).

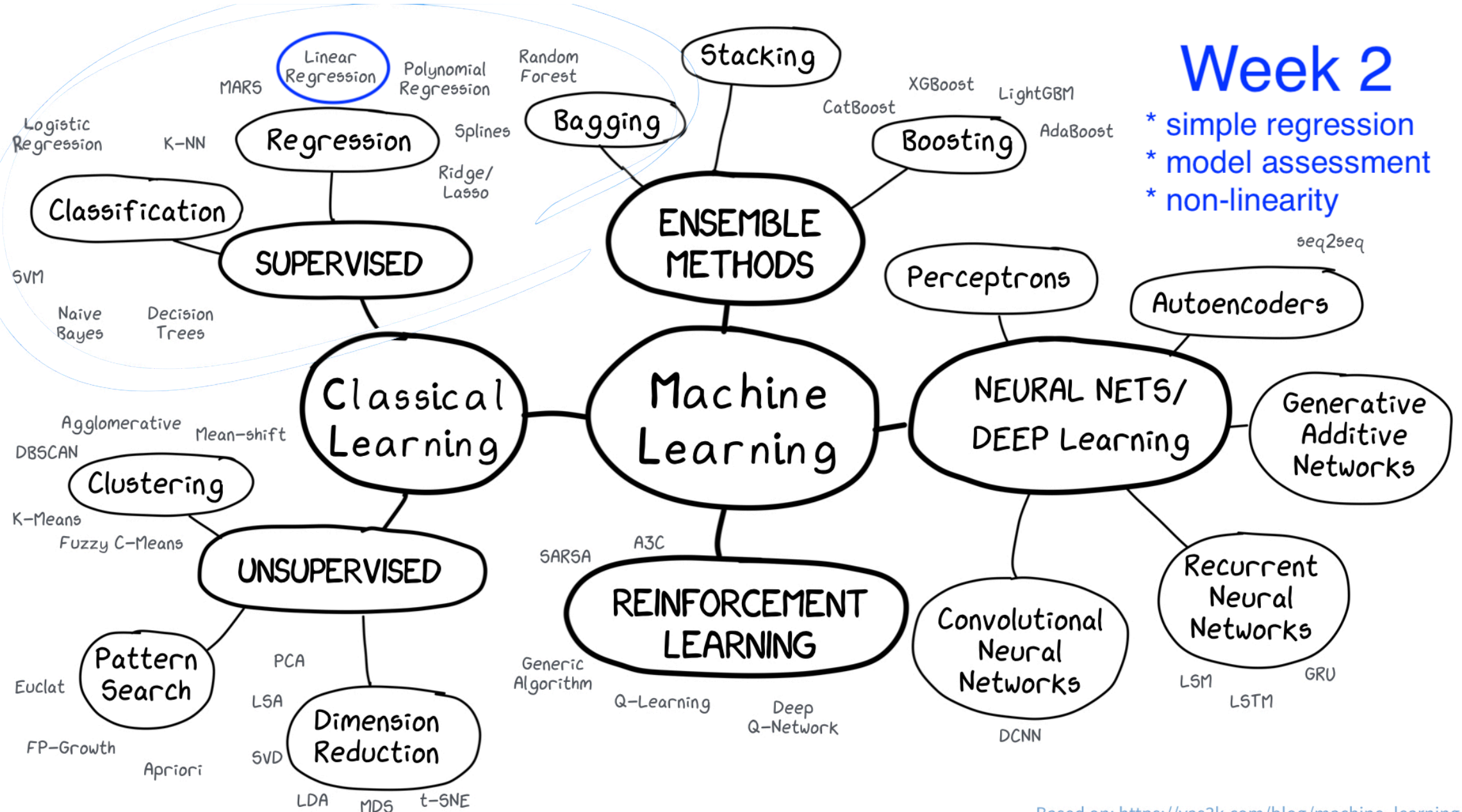
Unsupervised ML

Involves training a model on unlabeled examples.

- **Clustering:** Grouping similar data points into clusters (e.g., customer segmentation).
- **Dimensionality Reduction**: Reducing the number of features while retaining the most important information (e.g., reduce image size).



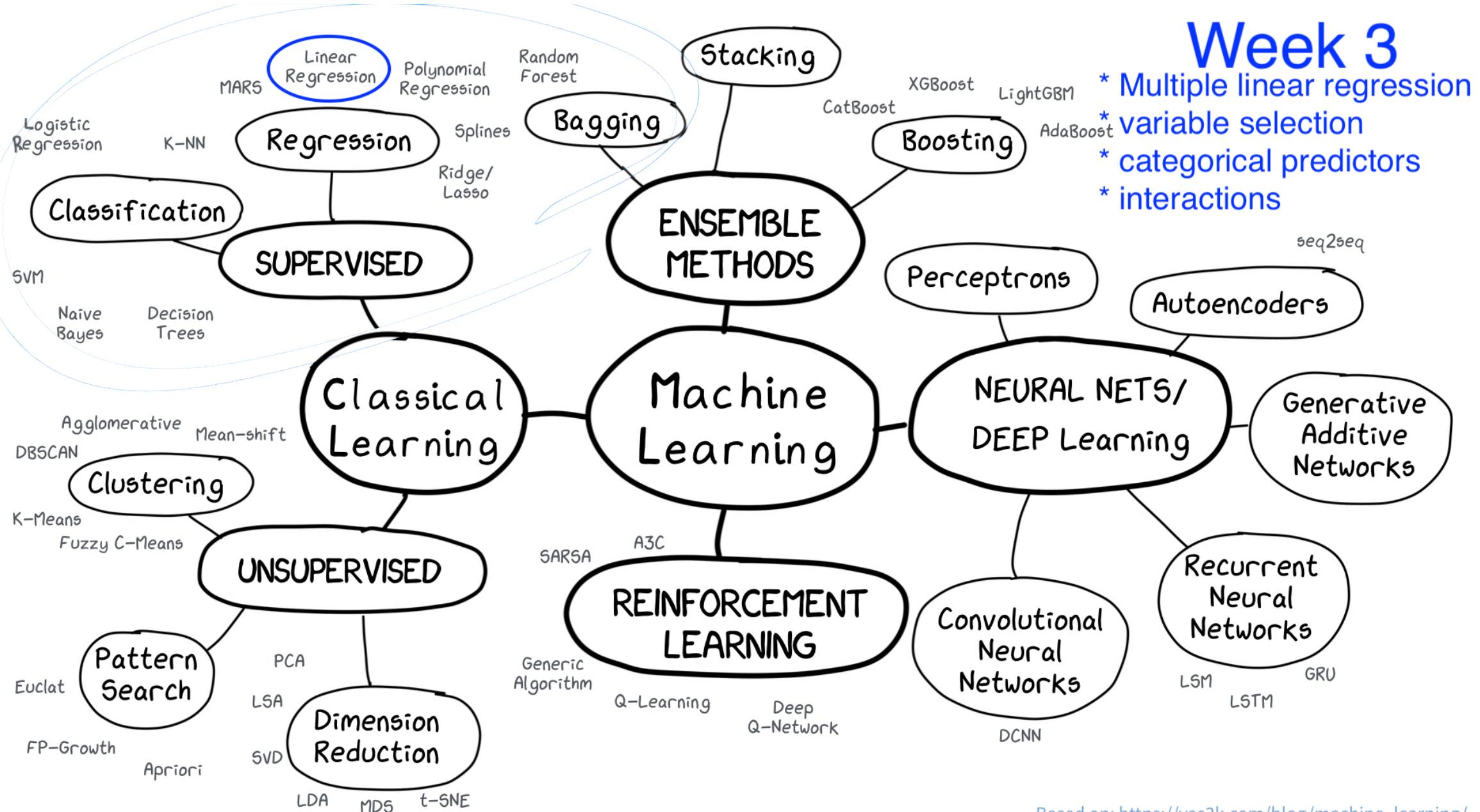
Based on: https://vas3k.com/blog/machine_learning/

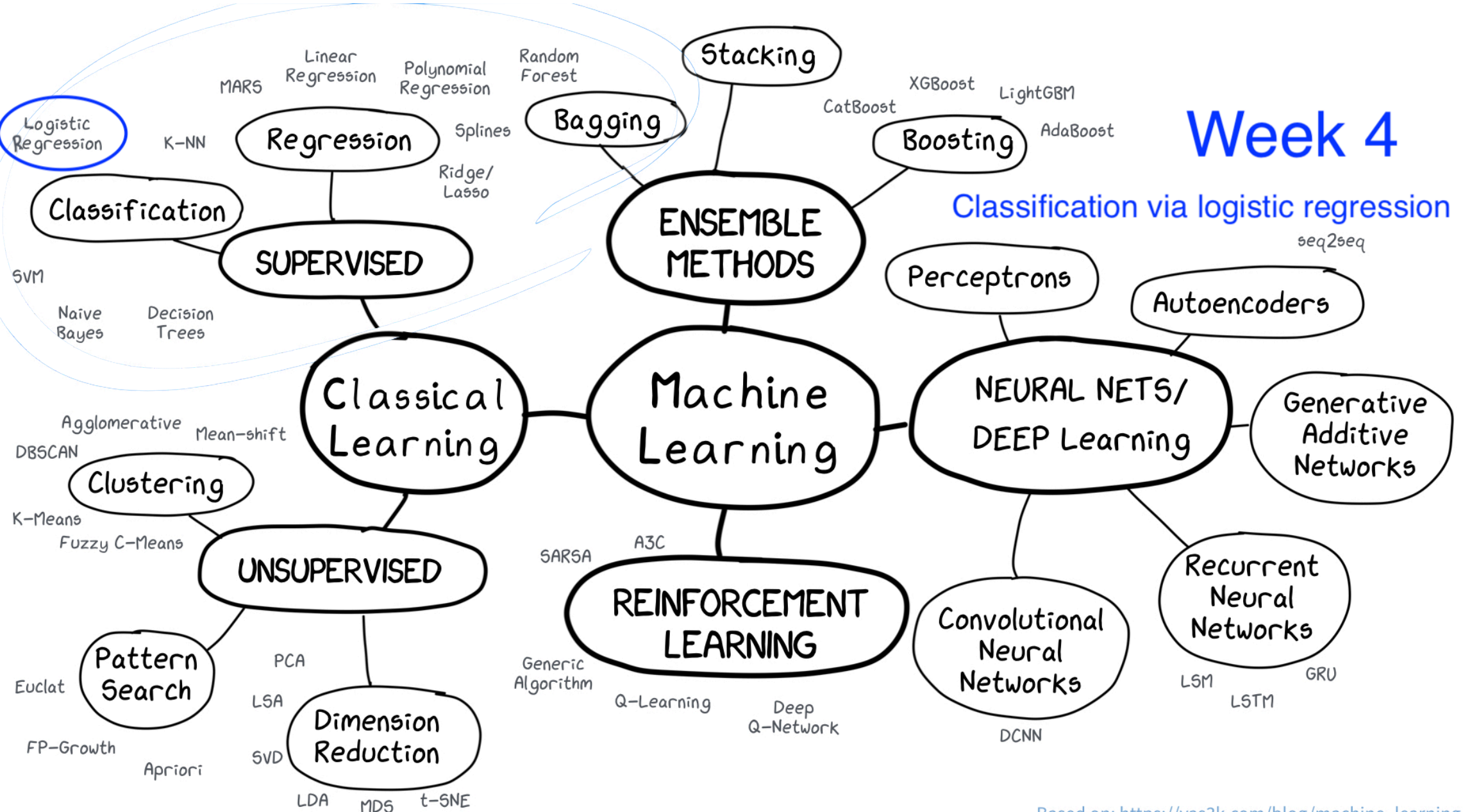


Week 2

* simple regression
* model assessment
* non-linearity

Based on: https://vas3k.com/blog/machine_learning/

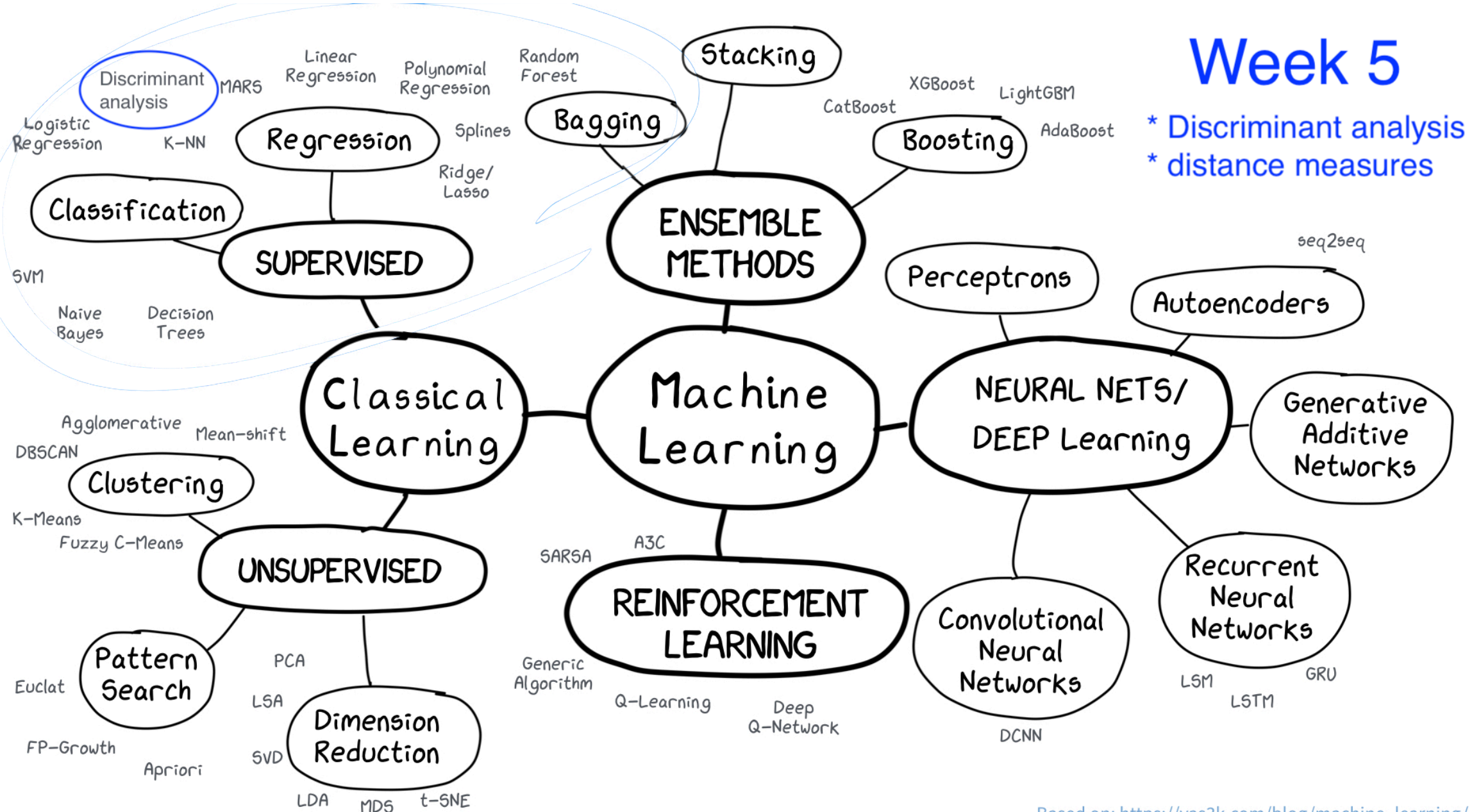

Week 3
* Multiple linear regression
* variable selection
* categorical predictors
* interactions
Machine Learning
Classical Learning
Regression
Linear Regression
Polynomial Regression
MARS
Splines
Ridge/ Lasso
K-NN
Logistic Regression
Classification
SUPERVISED
SVM
Naive Bayes
Decision Trees
Clustering
Agglomerative
Mean-shift
DBSCAN
K-Means
Fuzzy C-Means
UNSUPERVISED
Pattern Search
Euclat
FP-Growth
Apriori
Dimension Reduction
PCA
LSA
SVD
LDA
MDS
t-SNE
ENSEMBLE METHODS
Stacking
Bagging
Random Forest
Boosting
XGBoost
LightGBM
CatBoost
AdaBoost
REINFORCEMENT LEARNING
SARSA
A3C
Generic Algorithm
Q-Learning
Deep Q-Network
NEURAL NETS/ DEEP Learning
Perceptrons
Autoencoders
seq2seq
Generative Additive Networks
Convolutional Neural Networks
DCNN
Recurrent Neural Networks
LSM
LSTM
GRU
Based on: https://vas3k.com/blog/machine_learning/

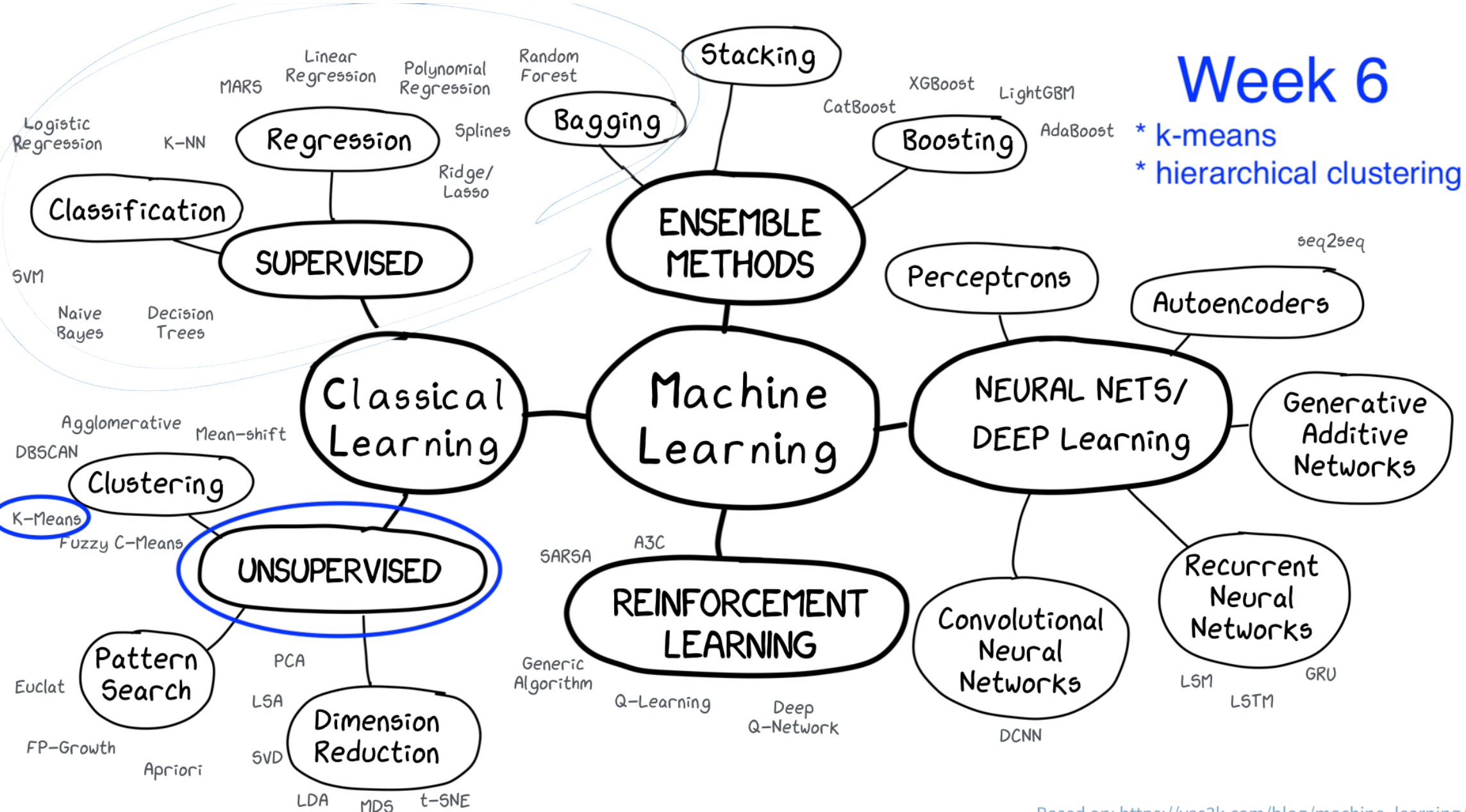


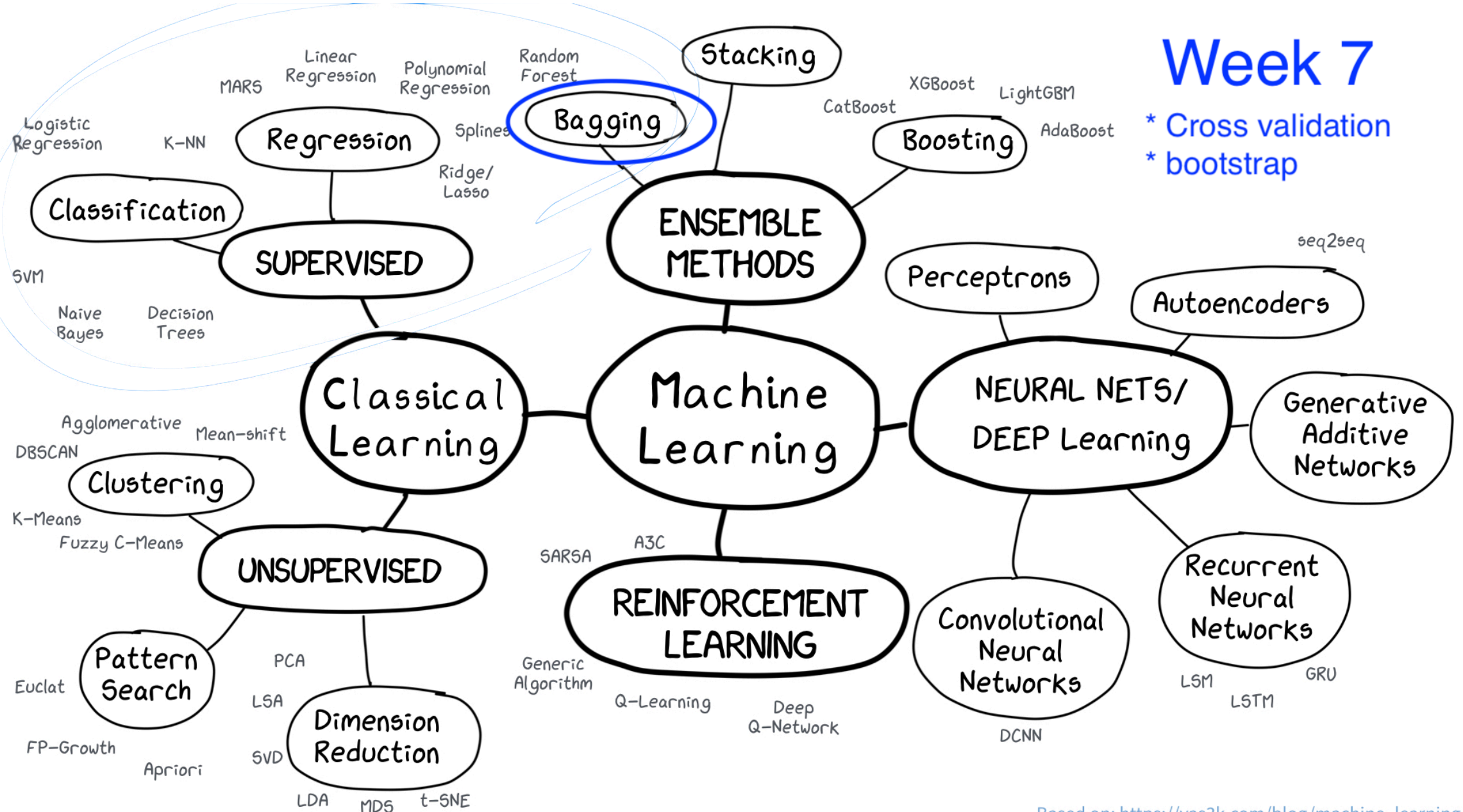
Week 4

Classification via logistic regression

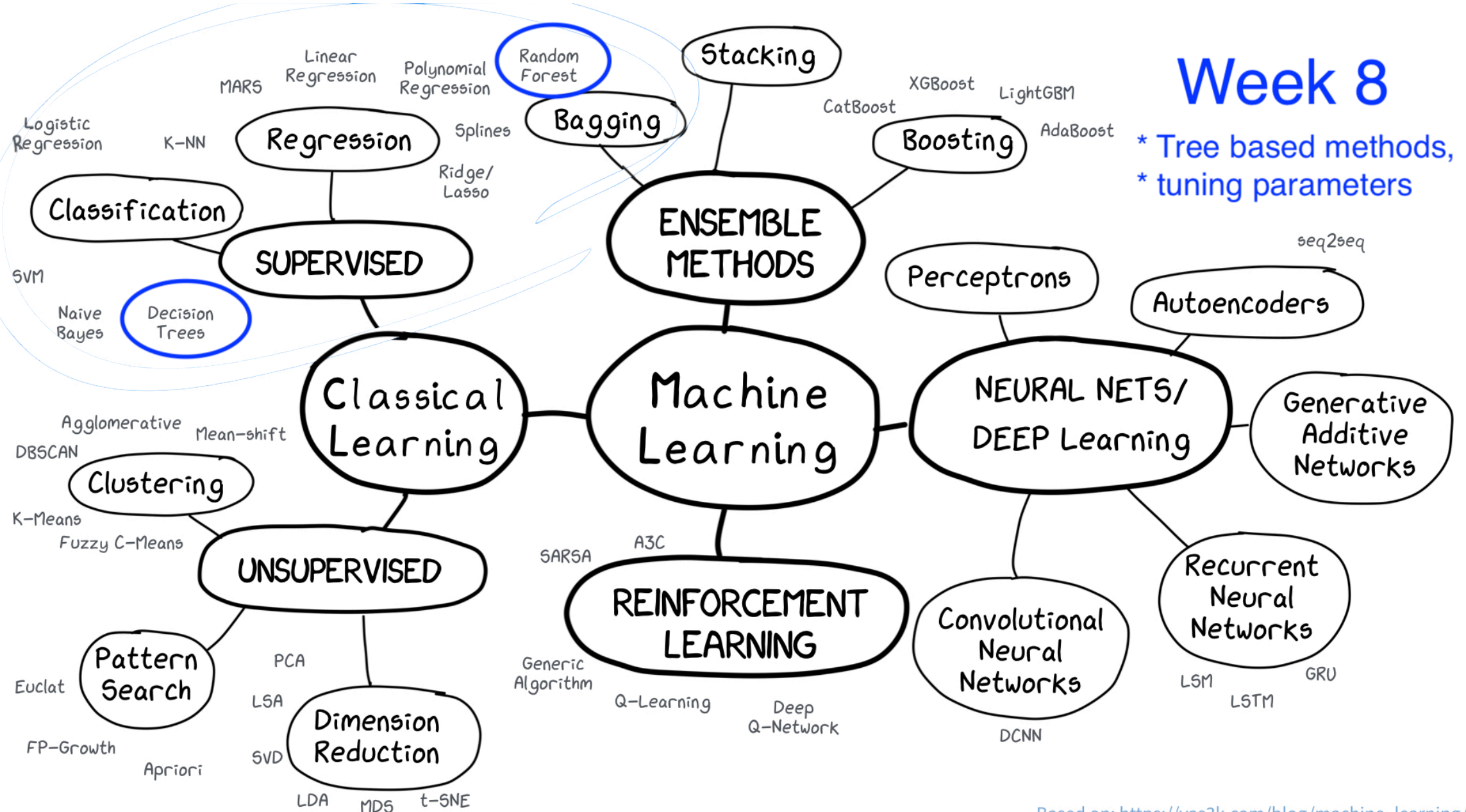
Based on: https://vas3k.com/blog/machine_learning/

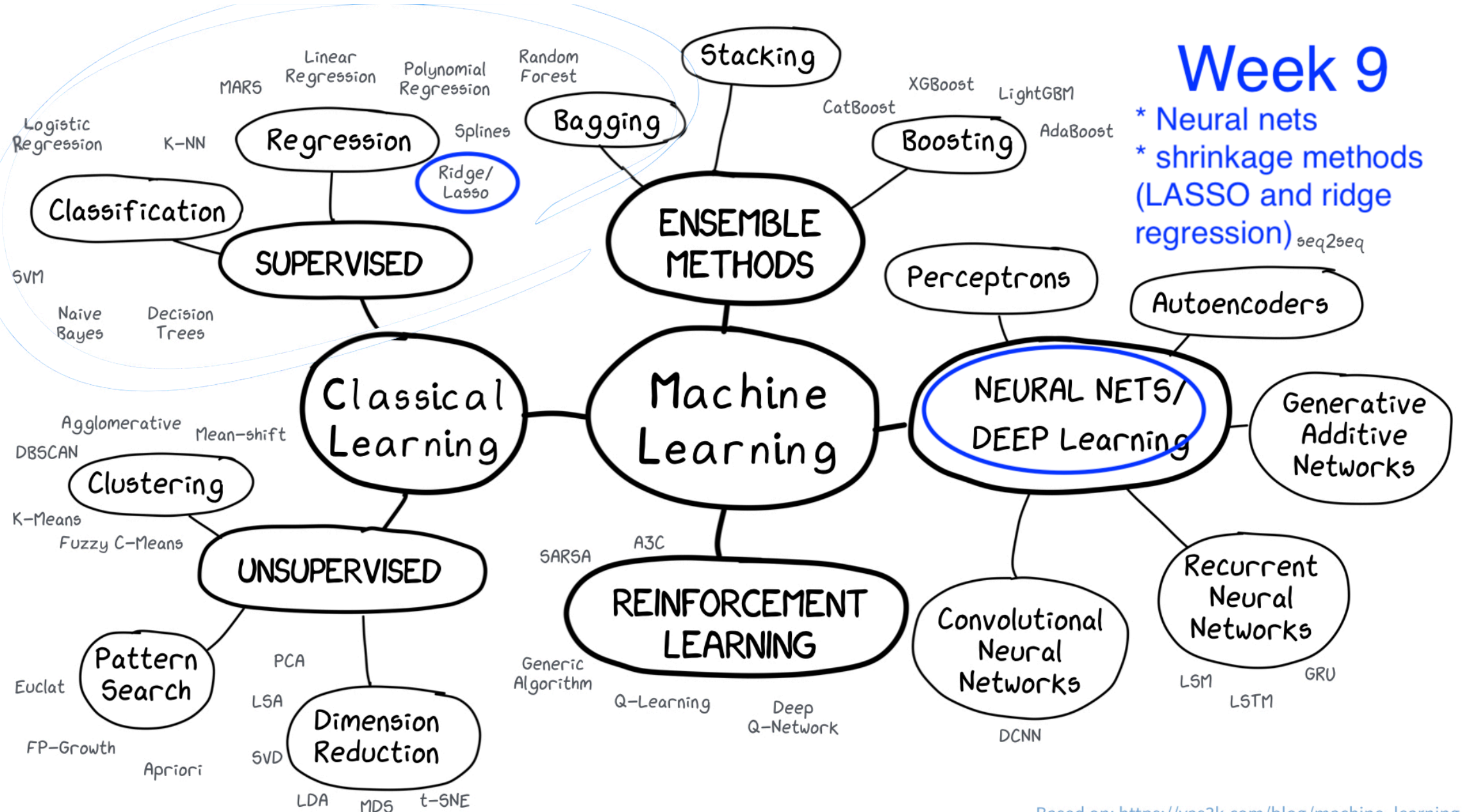

Week 5
* Discriminant analysis
* distance measures
Machine Learning
Classical Learning
Regression
Linear Regression
Polynomial Regression
Splines
Ridge/ Lasso
Logistic Regression
K-NN
Discriminant analysis
MARS
Classification
SVM
Naive Bayes
Decision Trees
SUPERVISED
Clustering
Agglomerative
Mean-shift
DBSCAN
K-Means
Fuzzy C-Means
UNSUPERVISED
Pattern Search
Euclat
FP-Growth
Apriori
Dimension Reduction
PCA
LSA
SVD
LDA
MDS
t-SNE
ENSEMBLE METHODS
Stacking
Bagging
Random Forest
Boosting
XGBoost
LightGBM
CatBoost
AdaBoost
REINFORCEMENT LEARNING
SARSA
A3C
Generic Algorithm
Q-Learning
Deep Q-Network
NEURAL NETS/ DEEP Learning
Perceptrons
Autoencoders
seq2seq
Generative Additive Networks
Convolutional Neural Networks
DCNN
Recurrent Neural Networks
LSM
LSTM
GRU
Based on: https://vas3k.com/blog/machine_learning/


Week 6
* k-means
* hierarchical clustering
Machine Learning
Classical Learning
Regression
Linear Regression
Polynomial Regression
MARS
Logistic Regression
K-NN
Splines
Ridge/ Lasso
Classification
SUPERVISED
SVM
Naive Bayes
Decision Trees
Clustering
Agglomerative
Mean-shift
DBSCAN
K-Means
Fuzzy C-Means
UNSUPERVISED
Pattern Search
Euclat
FP-Growth
Apriori
Dimension Reduction
PCA
LSA
SVD
LDA
MDS
t-SNE
ENSEMBLE METHODS
Stacking
Bagging
Random Forest
Boosting
XGBoost
LightGBM
CatBoost
AdaBoost
REINFORCEMENT LEARNING
SARSA
A3C
Generic Algorithm
Q-Learning
Deep Q-Network
NEURAL NETS/ DEEP Learning
Perceptrons
Autoencoders
seq2seq
Generative Additive Networks
Convolutional Neural Networks
DCNN
Recurrent Neural Networks
LSM
LSTM
GRU
Based on: https://vas3k.com/blog/machine_learning/

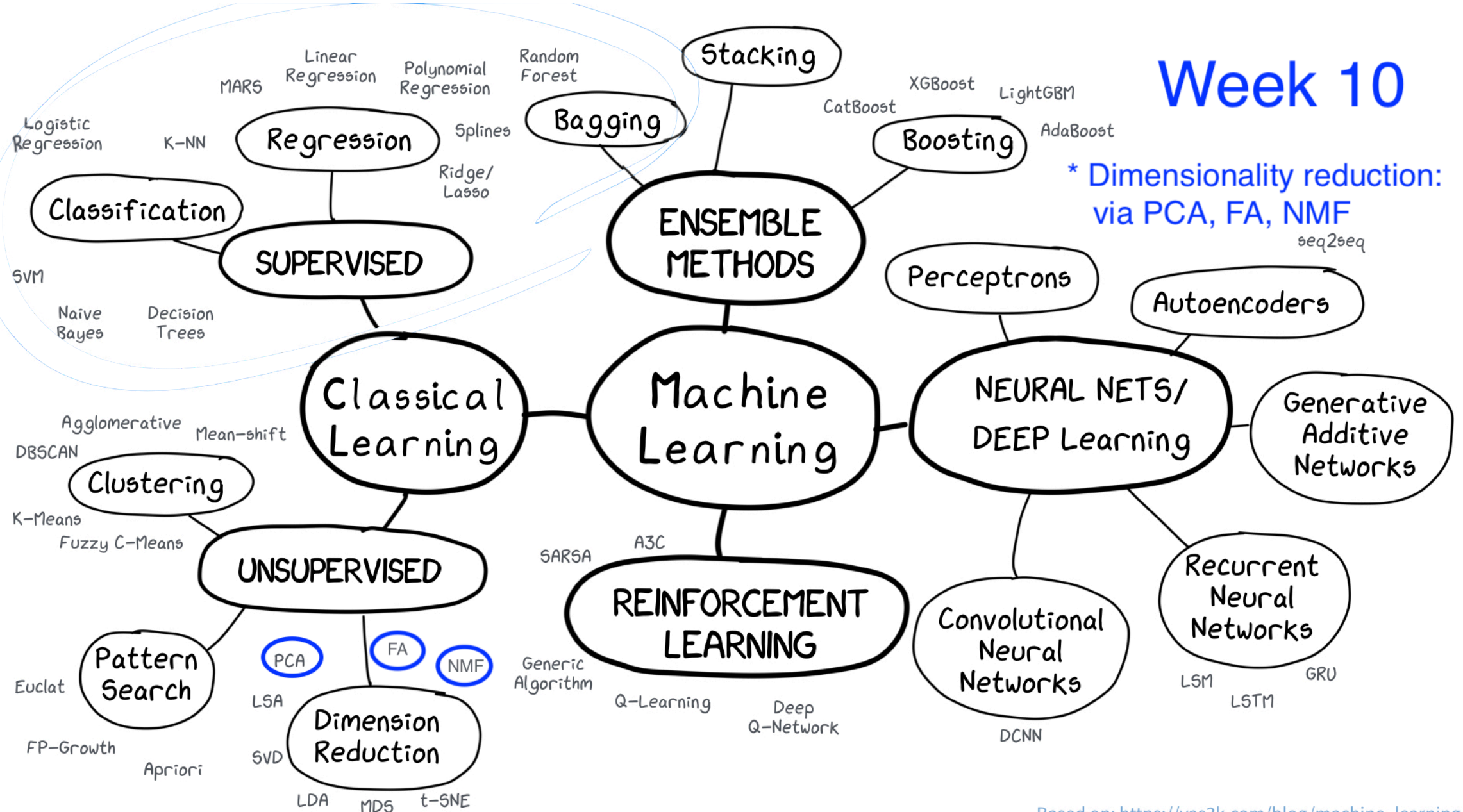


Based on: https://vas3k.com/blog/machine_learning/


Week 8
* Tree based methods,
* tuning parameters
Machine Learning
Classical Learning
Regression
Linear Regression
Polynomial Regression
MARS
Logistic Regression
K-NN
Splines
Ridge/ Lasso
Classification
SVM
Naive Bayes
Decision Trees
SUPERVISED
Clustering
Agglomerative
Mean-shift
DBSCAN
K-Means
Fuzzy C-Means
UNSUPERVISED
Pattern Search
Euclat
FP-Growth
Apriori
Dimension Reduction
PCA
LSA
SVD
LDA
MDS
t-SNE
ENSEMBLE METHODS
Random Forest
Bagging
Stacking
Boosting
XGBoost
LightGBM
CatBoost
AdaBoost
REINFORCEMENT LEARNING
SARSA
A3C
Generic Algorithm
Q-Learning
Deep Q-Network
NEURAL NETS/ DEEP Learning
Perceptrons
Autoencoders
seq2seq
Generative Additive Networks
Convolutional Neural Networks
DCNN
Recurrent Neural Networks
LSM
LSTM
GRU
Based on: https://vas3k.com/blog/machine_learning/



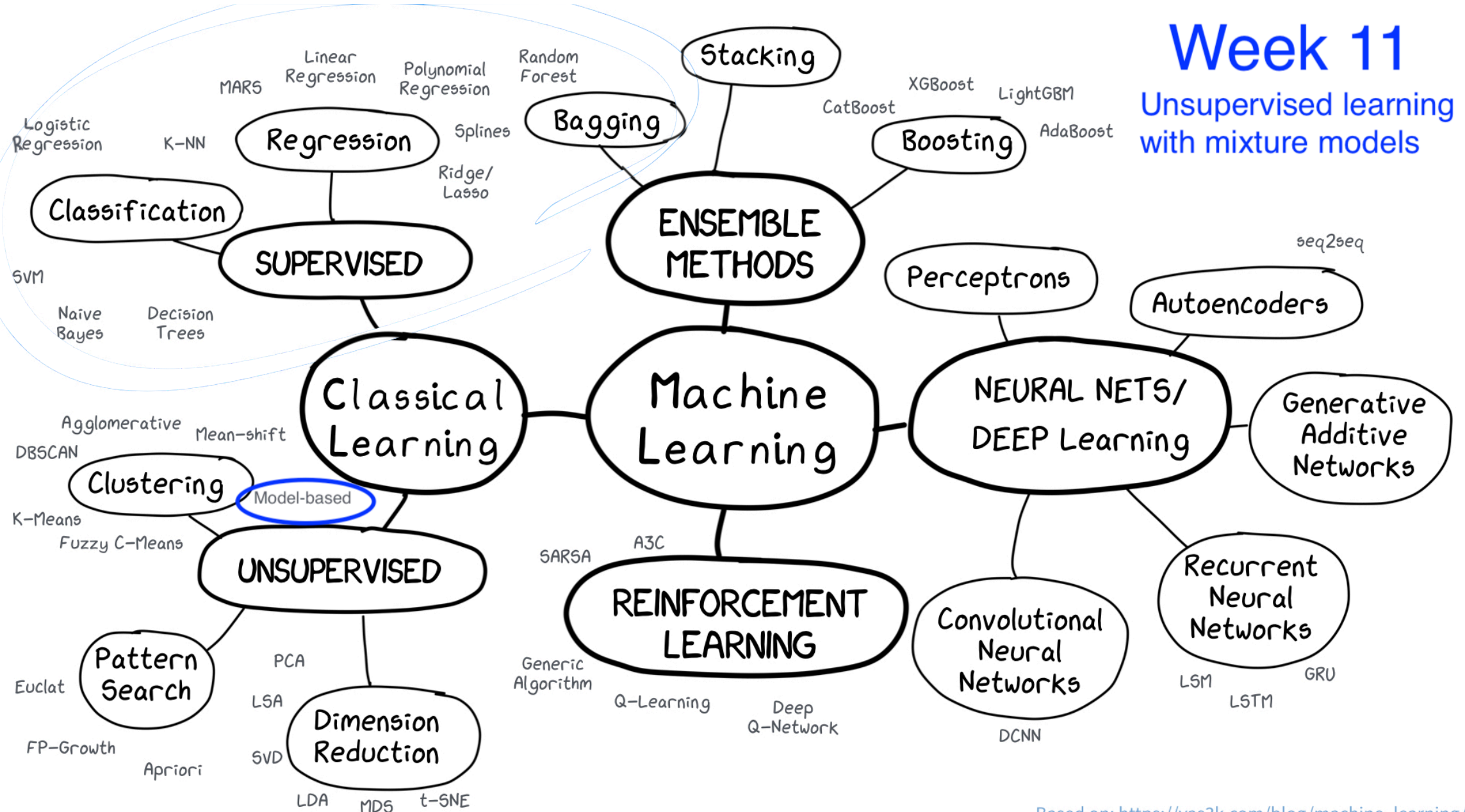
Based on: https://vas3k.com/blog/machine_learning/



Based on: https://vas3k.com/blog/machine_learning/

Week 11

Unsupervised learning with mixture models



Based on: https://vas3k.com/blog/machine_learning/

iClicker

Students may choose to participate in iClicker questions by enrolling at: https://join.iclicker.com/BODT