Essence of Machine Learning (and Deep Learning)

Hoa M. Le

Data Science Lab, HUST

hoamle.github.io

Examples

- https://www.youtube.com/watch?v=BmkA1ZsG2P4
- http://www.r2d3.us/visual-intro-to-machine-learning-part-1/

Machine Learning is about …

… a computer program (machine) *<u>learns</u>* to do a task (problem) from experience (data)

- *learning* ≜ improved *performance* with more experience

- Tom Mitchell

$\Uparrow$

predictive modelling with **sample data**

$\Uparrow$

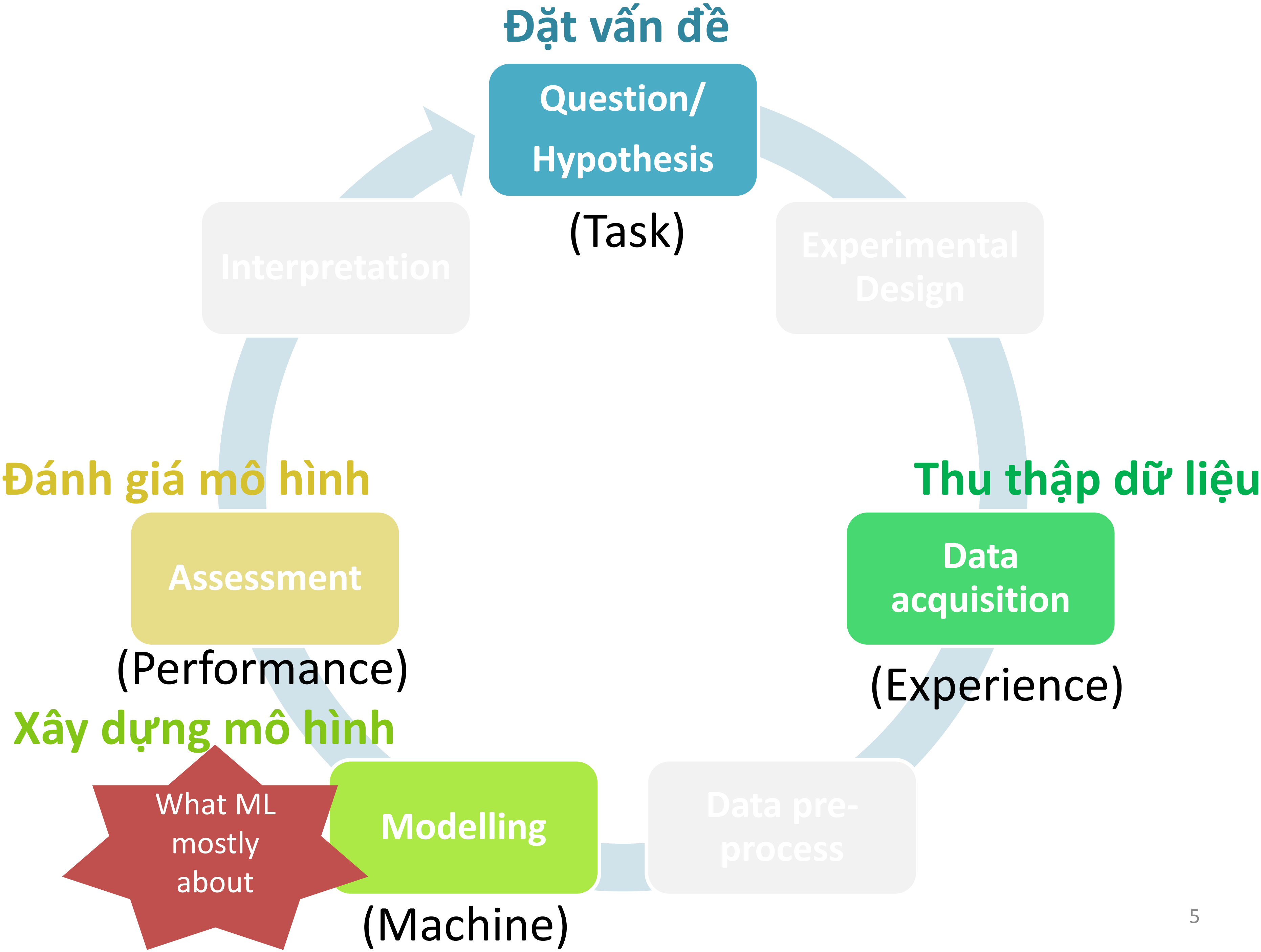
$\boxed{\text{"heurestics" \& statistical modelling}}$

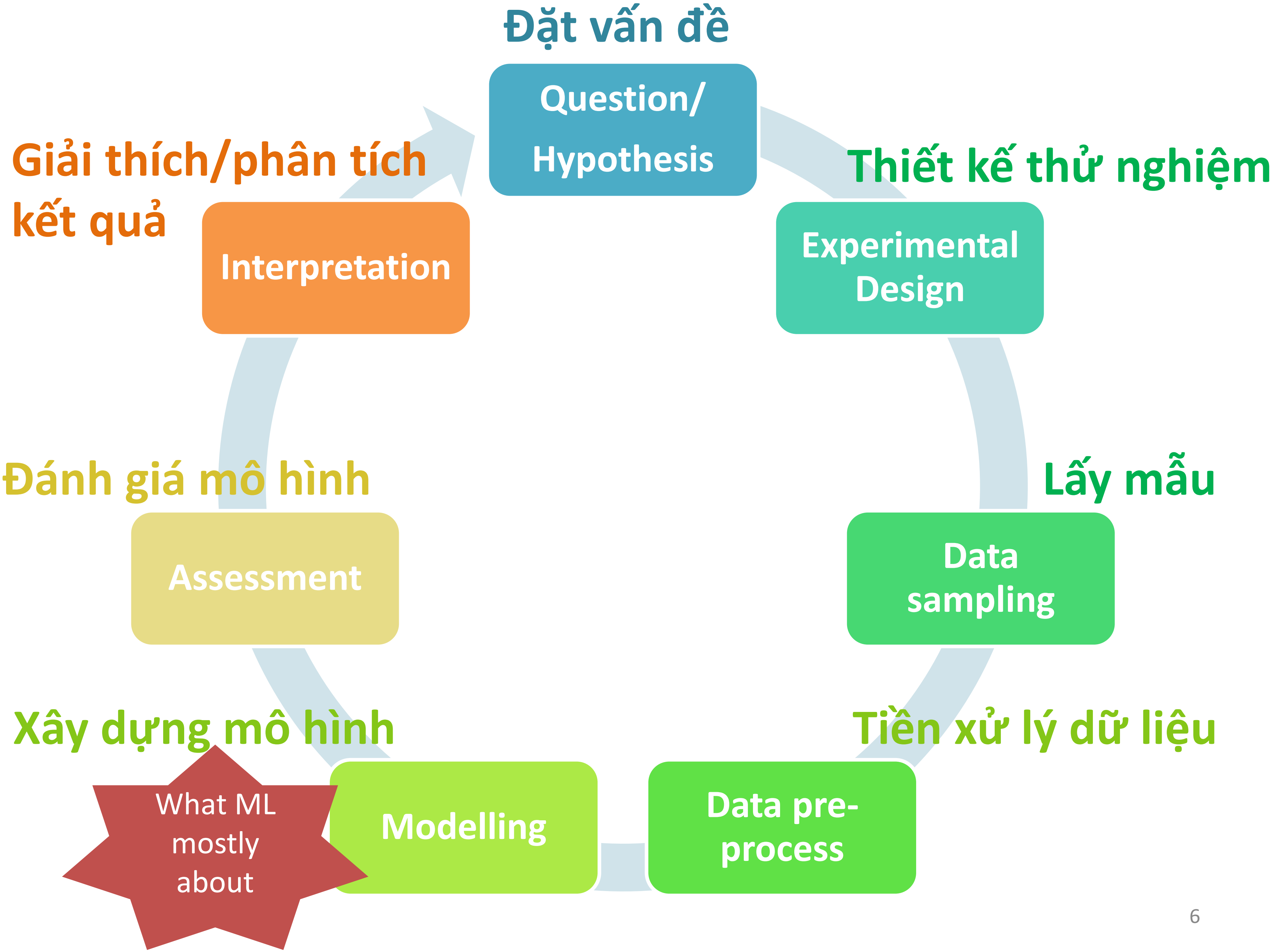
note 1: “heurestic” as in “intuitive, but not (yet!) rigorously proven by mathematical tools at some extend”

note 2: predictive modelling can also be in the form of rule-based systems, models in physics, etc

BUILD
A MACHINE LEARNING SOLUTION

the Pipeline


Đặt vấn đề
Question/
Hypothesis
(Task)
Interpretation
Experimental
Design
Đánh giá mô hình
Thu thập dữ liệu
Assessment
Data
acquisition
(Performance)
(Experience)
Xây dựng mô hình
What ML
mostly
about
Modelling
Data pre-
process
(Machine)


Đặt vấn đề
Question/
Hypothesis
Thiết kế thử nghiệm
Experimental
Design
Lấy mẫu
Data
sampling
Tiền xử lý dữ liệu
Data pre-
process
Xây dựng mô hình
Modelling
What ML
mostly
about
Đánh giá mô hình
Assessment
Giải thích/phân tích
kết quả
Interpretation

Đặt vấn đề

Question/
Hypothesis

Q.a. What are there in an **abitrary photo?**
Q.b. What is there in an **abitrary photo?**
Q.c. Is there any puppy an **abitrary photo?**

cat
flower
dog
jet
ground
grass
…

Other questions:
- Where are the puppies in a photo?
- How confident can I assure that there is a cat a photo?
- For what reasons can I know that there is a cat in a photo?

Thiết kế thử nghiệm

Experimental Design

(i.e. planning)

Machine Learning

i.e.

Automatic data-driven predictive models

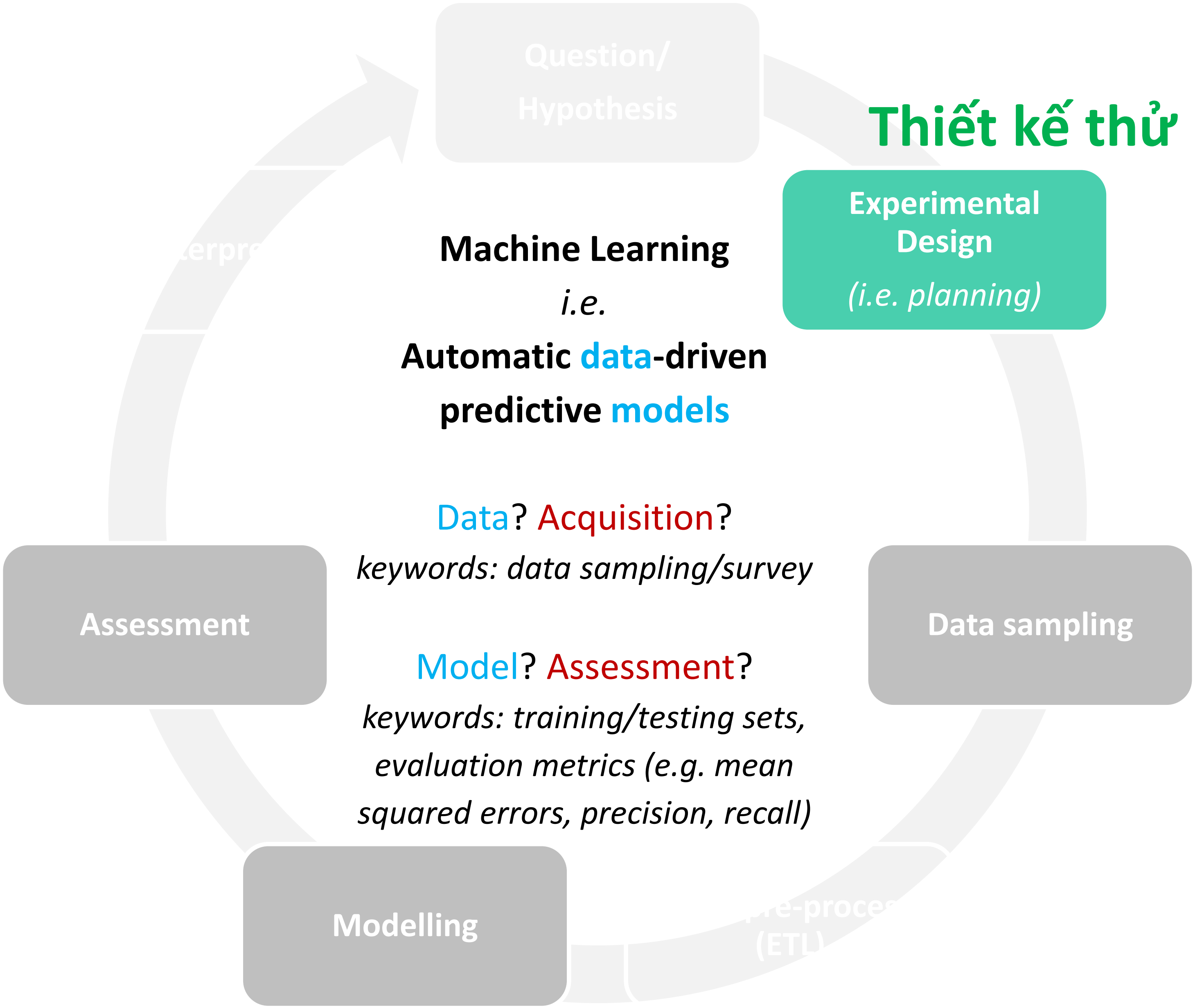
Data? Acquisition?

keywords: data sampling/survey

Model? Assessment?

keywords: training/testing sets, mean squared errors, precision, recall, …

Thiết kế thử nghiệm



Machine Learning

i.e.

Automatic data-driven predictive models

Data? Acquisition?

keywords: data sampling/survey

Model? Assessment?

keywords: training/testing sets, evaluation metrics (e.g. mean squared errors, precision, recall)

Avoid as many sampling biases as possible
http://norvig.com/experiment-design.html

Question/ Hypothesis

Experimental Design

Data Sampling

Representative sample

- *How many photos, categories, photos in each category, …?*
- *(If time-series data: eg videos)*

Sample at which time points?

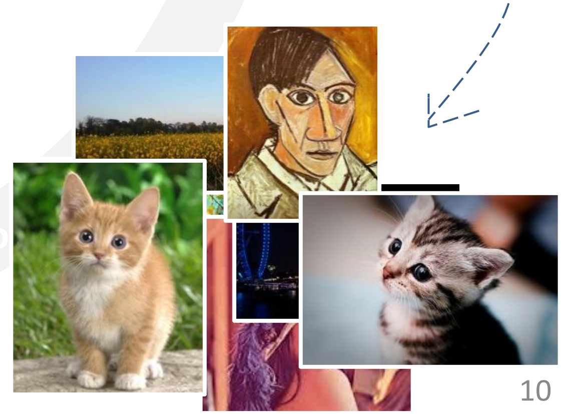
- *Imbalance class?*
- *Selection bias?*

Lấy mẫu

Assessment

Data sampling

Modelling



Which metrics to use depend on which problem

http://scikit-learn.org/stable/modules/model_evaluation.html

Question/ Hypothesis

Experimental Design

Model Assessment

Evaluation metrics

- *Accuracy*
- *Precision, Recall*
- *Area Under Curve (AUC)*
- *Mean squared errors (MSE)*
- *…*

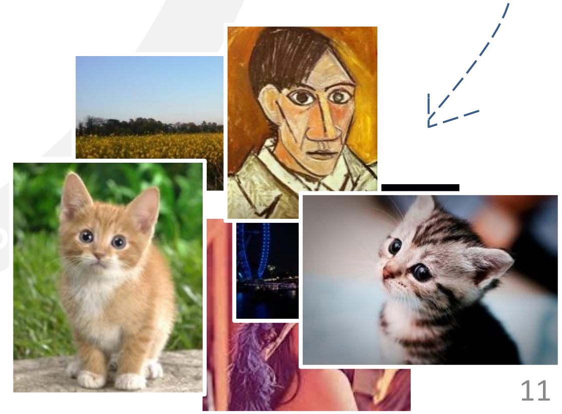
(If hypothesis testing problem)

- *t-statistic, z-statistic, χ^2-statistic, …*

Đánh giá mô hình

Assessment

Data sampling



Modelling

cat
flower
dog
jet
ground
grass

If training/testing set split is well designed with sufficient examples, we might not need to repeat many experiments.

Question/ Hypothesis

Experimental Design

Model Assessment

Evaluation setup

Evaluation (i.e.report results) on *unseen* data

- Training/testing set split: follows data sampling principles
- Repeat experiment: gives measurable confidence to the reported results

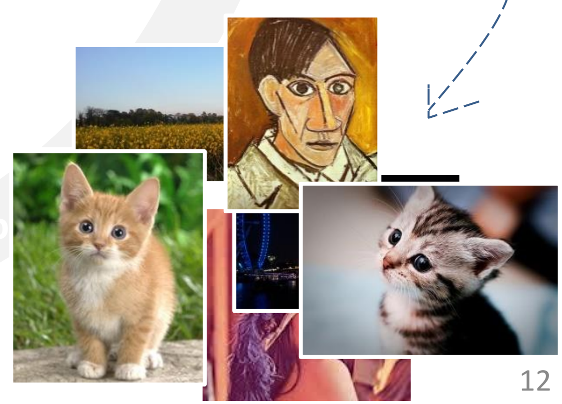
Đánh giá mô hình

Assessment

Data sampling

Modelling

```
cat
flower
dog
jet
ground
grass
```



"All models are wrong, but some are useful."
- Box and Drape, 1987

Model Building

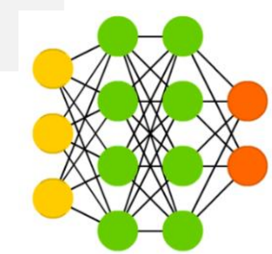
Model = a simplification of **reality**

(e.g. map of Hanoi)

Keywords: Linear models, Graphical models, Neural networks, SVM, Gaussian Process, Random forest …

Modelling tip: building model goes from the ***most simplified*** forms to the ***more complex*** to describe reality more precisely

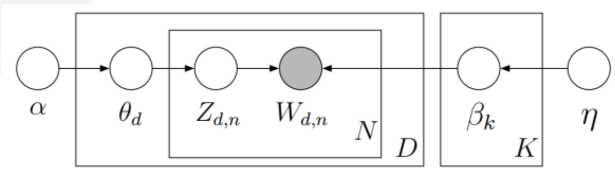
(e.g. building from Linear models to Latent variable models / Deep neural networks)



Xây dựng mô hình

What ML mostly about

Modelling



Raw data → Post-processed data

- *Data ETL: extract, transform, load*
- *Data standardisation / normalisation*
- *Data imputation (if missing values)*

-0.34	-0.46	-0.87
1.47	-0.24	2.21
-1.05	0.02	-1.74
0.09	-0.58	1.02
1.63	-0.53	0.06
1.11	-0.63	-0.93
-0.34	-0.46	-0.87
1.47	-0.24	2.21
-1.05	0.02	-1.74
0.09	-0.58	1.02
1.63	-0.53	0.06
1.11	-0.63	-0.93
0.09	-0.58	1.02
1.63	-0.53	0.06
1.11	-0.63	-0.93
....	....	....

Assessment

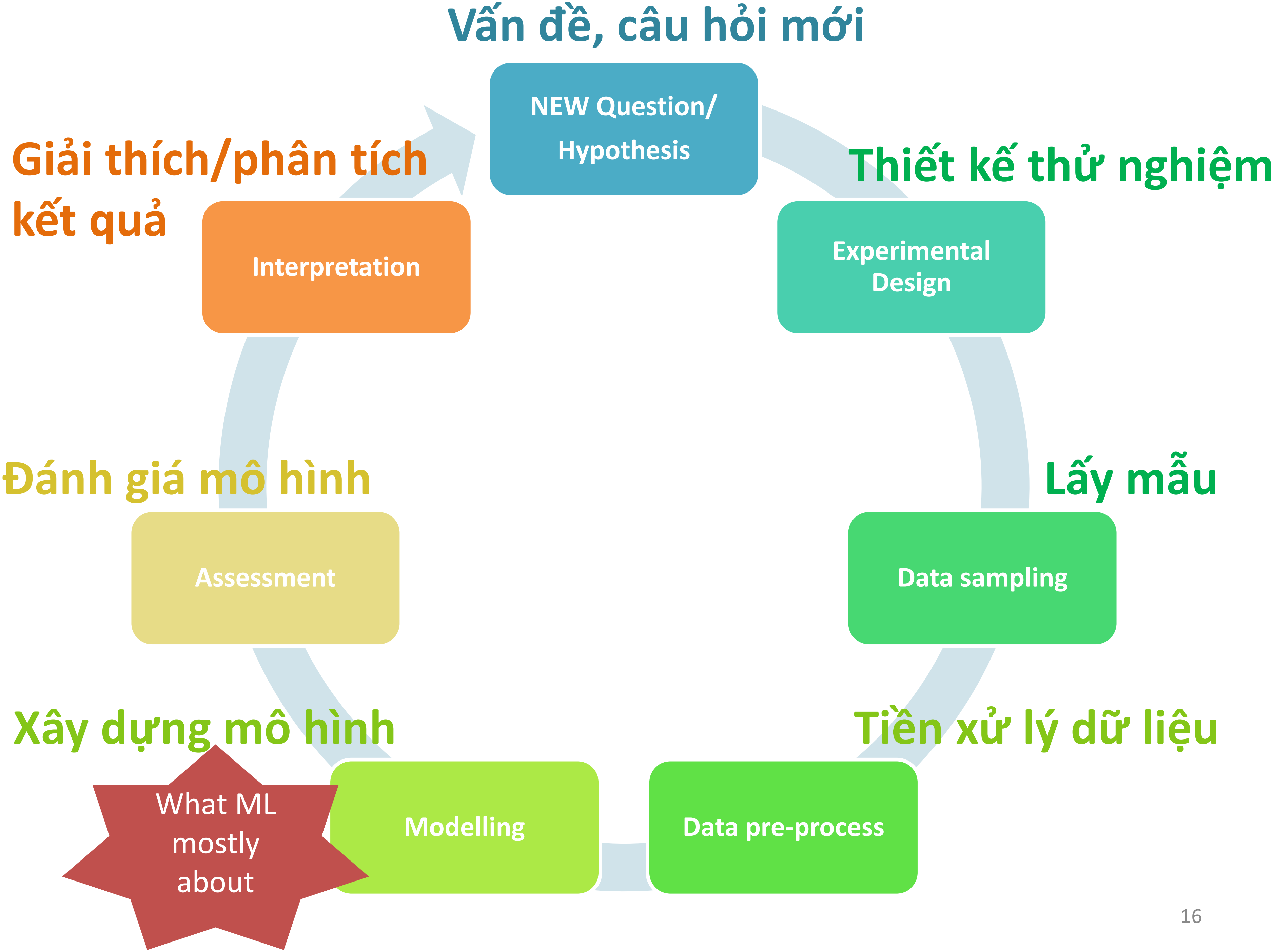
Feature extraction

Tiền xử lý dữ liệu

Modelling

Data pre-process


Đặt vấn đề
Question/
Hypothesis
Thiết kế thử nghiệm
Experimental
Design
Lấy mẫu
Data sampling
Tiền xử lý dữ liệu
Data pre-process
Xây dựng mô hình
Modelling
What ML
mostly
about
Đánh giá mô hình
Assessment
Interpretation

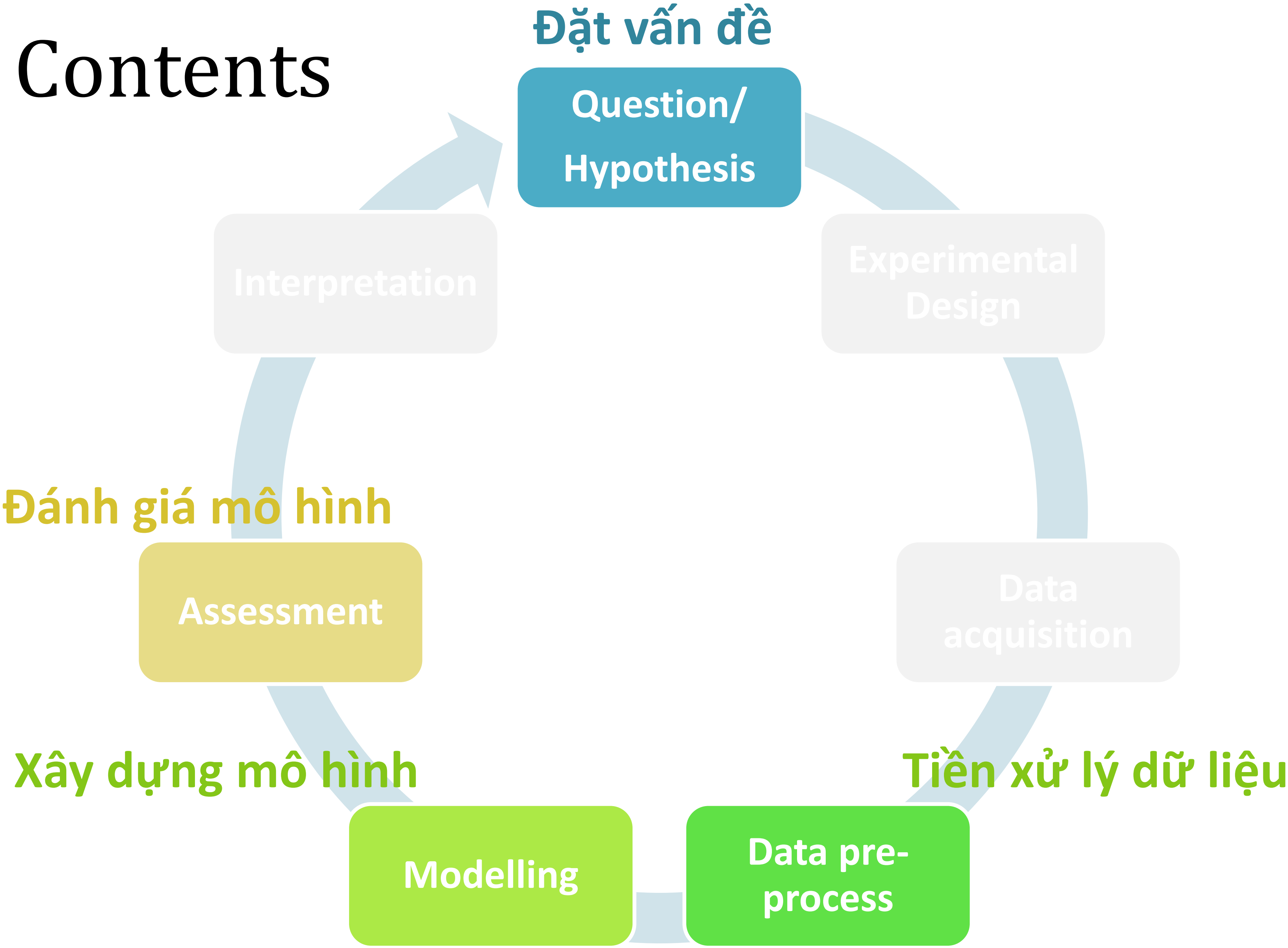

Vấn đề, câu hỏi mới
NEW Question/
Hypothesis
Thiết kế thử nghiệm
Experimental
Design
Lấy mẫu
Data sampling
Tiền xử lý dữ liệu
Data pre-process
Xây dựng mô hình
Modelling
What ML
mostly
about
Đánh giá mô hình
Assessment
Giải thích/phân tích
kết quả
Interpretation

PRINCIPLES OF MODELLING

Statistical reasoning (*)

*(*) A machine learning algorithm does not necessarily have a probabilistic interpretation, or developed from a statistical framework. Nevertheless, statistical reasoning provides a rigorous mathematical tool for estimation and inference to make optimal decision (e.g. prediction, action) under **uncertainty,** which is one of the ultimate objectives in ML**.***

Contents


Đặt vấn đề
Question/
Hypothesis
Experimental
Design
Data
acquisition
Tiền xử lý dữ liệu
Data pre-
process
Xây dựng mô hình
Modelling
Đánh giá mô hình
Assessment
Interpretation

ML problem: Classification

Question

Is there any cat in an abitrary photo?

Experience: dataset of {image, label} pairs $\mathcal{D} = \{x_n, y_n\}_{n=1}^{N}$

Modelling

predict $\hat{y}_n$ – *cat existence* – *given arbitrary* x_n

Cat?
Not cat?

supervised learning

Image

x_n

$\mathbb{N}^{400\times600\times3}$

Prediction

$\hat{y}_n$

$\{\text{True}, \text{False}\}$

(single-class)

binary classification problem

Assessment

Accuracy = $\frac{1}{N}\sum_n \mathbb{I}(\hat{y}_n = y_n)$

Precision, Recall, F1-score

Area Under Curve (AUC)

...

Example models:
Logistic regression (linear model)
Neural Net with sigmoid output (nonlinear model)

ML problem: Classification

Question

What is there in an abitrary photo?

Experience: dataset of {image, label} pairs $\mathcal{D} = \{x_n, y_n\}_{n=1}^{N}$

Modelling

predict $\hat{y}_n$ – *object identity* – *given arbitrary* x_n

Image

x_n

$\mathbb{N}^{400\times600\times3}$

cat
flower
dog
jet
ground
grass

Prediction

$\hat{y}_n$

$\{1,2,3,4,5,6\}$

supervised learning

(multi-class)

categorical classification problem

Assessment

Accuracy = $\frac{1}{N}\sum_n \mathbb{I}(\hat{y}_n = y_n)$

Precision, Recall, F1-score

Area Under Curve (AUC)

...

Example models:
Softmax classification (linear model)
Neural Net with softmax output (nonlinear model)

ML problem: Regression

Question **How much is the price of a house given …**

Experience: dataset of {(area, location, #rooms), price} pairs $\mathcal{D} = \{x_n, y_n\}_{n=1}^{N}$

Modelling predict $\hat{y}_n$ – *house price* – *given arbitrary* x_n

Area	100m^2
Location	24.7^{0}N 183.0^{0}E
#Rooms	3

→ \$150,000

supervised learning

Features/Predictors

x_n

$\mathbb{R} \times \mathbb{R}^2 \times \mathbb{N}$

Prediction

$\hat{y}_n$

$\mathbb{R}$

regression problem

Assessment squared_errors = $\frac{1}{N}\sum_n(\hat{y}_n - y_n)^2$

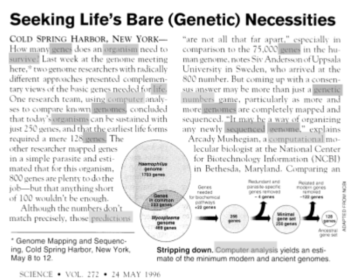
Example models/algorithms:
Linear regression (linear model)
Neural Net with linear output (nonlinear model)
Curve fitting algorithm

ML problem: Clustering

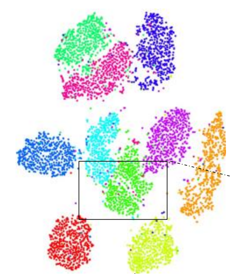
Question **What is the "topic" that a news article is talking about?**
Experience: dataset of article content *only* $\mathcal{D} = \{x_n\}_{n=1}^{N}$

Modelling predict $\boxed{z_n}$ – *"topic" (cluster) identity – given arbitrary* x_n

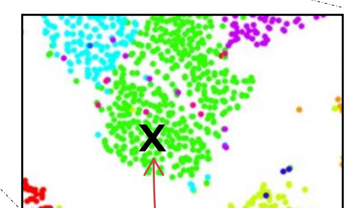
unsupervised learning



Article (text)
x_n
$\mathbb{N}^{1500}$



Prediction
z_n
$\{1,2,\ldots,10\}$



x_n
z_n = **green**

Assessment mean_distance_to_clusters $= \frac{1}{N}\sum_n \left(x_n - \mu_{z_n}\right)^2$

Example models/algorithms:
k-means algorithm
Generative models: Mixture models, Topic models

Note: "topic" = group/cluster in this context, and is ***not*** *pre-defined*
We will meet the term "topic" again when visiting Topic models

A ML problem can also be:

- both supervised and unsupervised *(semi-supervised)*
- combination of regression and classification sub-problems *e.g. image localisation*

Classification: C classes
Input: Image
Output: Class label
Evaluation metric: Accuracy

→ CAT

Localization:
Input: Image
Output: Box in the image (x, y, w, h)
Evaluation metric: Intersection over Union

→ (x, y, w, h)

Classification + Localization

CAT

PRINCIPLES OF MODELLING

1. Model structure - constructs relationships (*stochastic and/or deterministic*) between model elements: data, parameters, and hyper-parameters.

   Keywords: graphical model

2. Learning principle - defines a framework to estimate unknown parameters (and unobserved i.e. hidden/latent variables)

   Keywords: Maximum Likelihood criterion, Bayesian inference, ++ others

3. Regularisation

   Keywords: over-fitting, Bayesian inference, ++ others

   Relevant keywords: L2-regularisation (Ridge), L1-regularisation (LASSO)

⇒ ALGORITHM - implements 1 + 2 + 3 to train the model

Keywords: (stochastic) gradient descent, Expectation-Maximisation (EM), Variational Inference (VI), sampling-based inference methods

4. Model selection

   Keywords: cross-validation

Before we get going…

"Mathematics is the art of giving the same name to different things ."

-Henri Poincaré.

"The purpose of computation is insight, not numbers."

-Richard Hamming

$$p(\mathbf{w} \mid \alpha, \beta) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \int \left(\prod_{i=1}^{k} \theta_i^{\alpha_i - 1} \right) \left(\prod_{n=1}^{N} \sum_{i=1}^{k} \prod_{j=1}^{V} (\theta_i \beta_{ij})^{w_n^j} \right) d\theta,$$

$$p(D \mid \alpha, \beta) = \prod_{d=1}^{M} \int p(\theta_d \mid \alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn} \mid \theta_d) p(w_{dn} \mid z_{dn}, \beta) \right) d\theta_d.$$