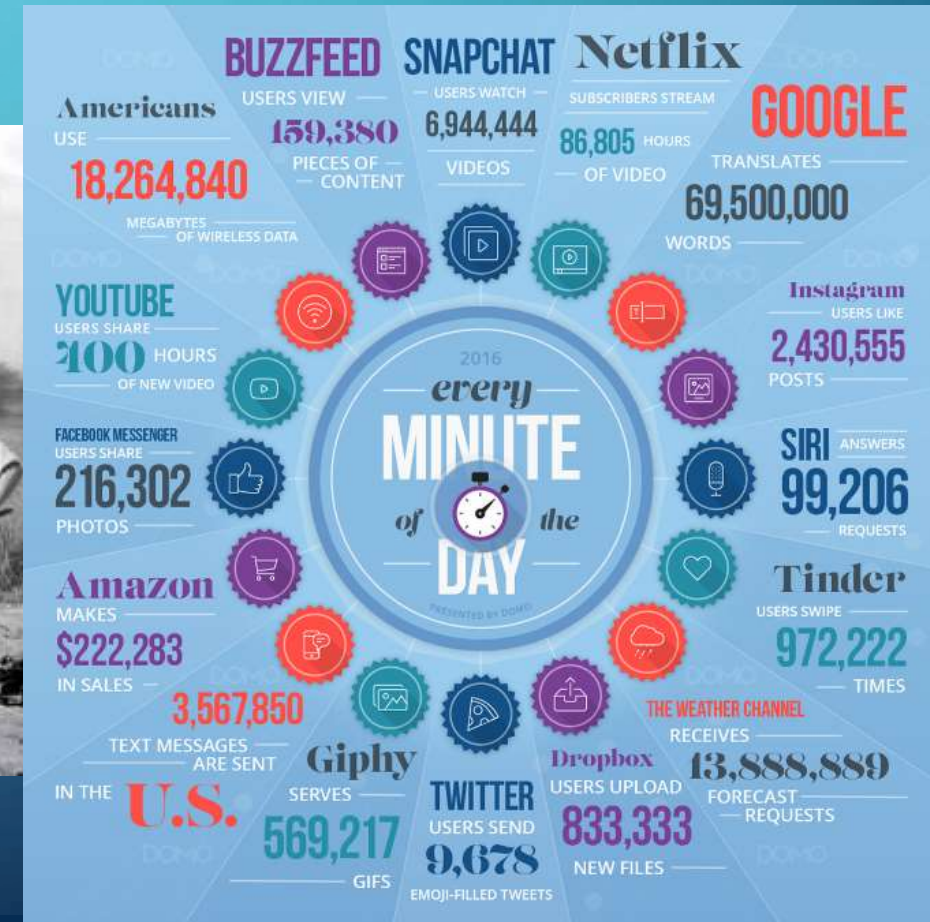
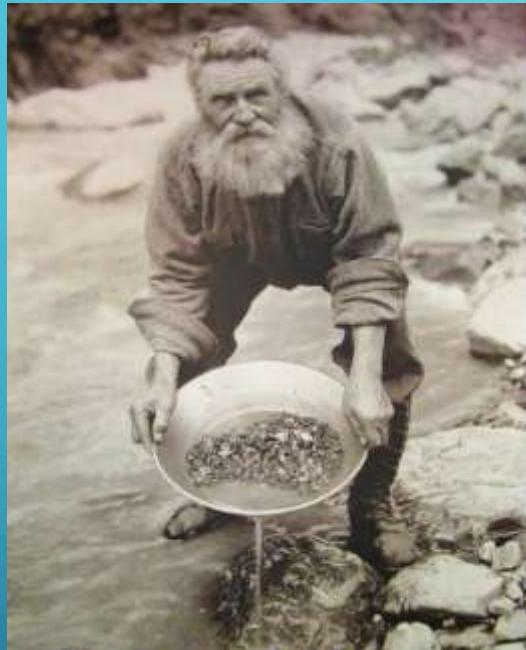




AVANCÉES EN ANALYSES D'IMAGES: RADIOMICS, BIG DATA ET MACHINE LEARNING

L FOURNIER, A BOUCHOUICHA, D BALVAY, A BURGUN, B RANCE
SERVICE DE RADIOLOGIE, SERVICE DE BIOINFORMATIQUE, BIOSTATISTIQUES
HÔPITAL EUROPÉEN GEORGES POMPIDOU
TEAM 2 “IMAGING OF ANGIOGENESIS”
INSERM UMR_S 1138 EQ22 “INFORMATION SCIENCES TO SUPPORT
PERSONALIZED MEDICINE”

BIG DATA : LA NOUVELLE RUEE?



BIG DATA: LES 4V

VOLUME

MAIS AUSSI

VARIETY

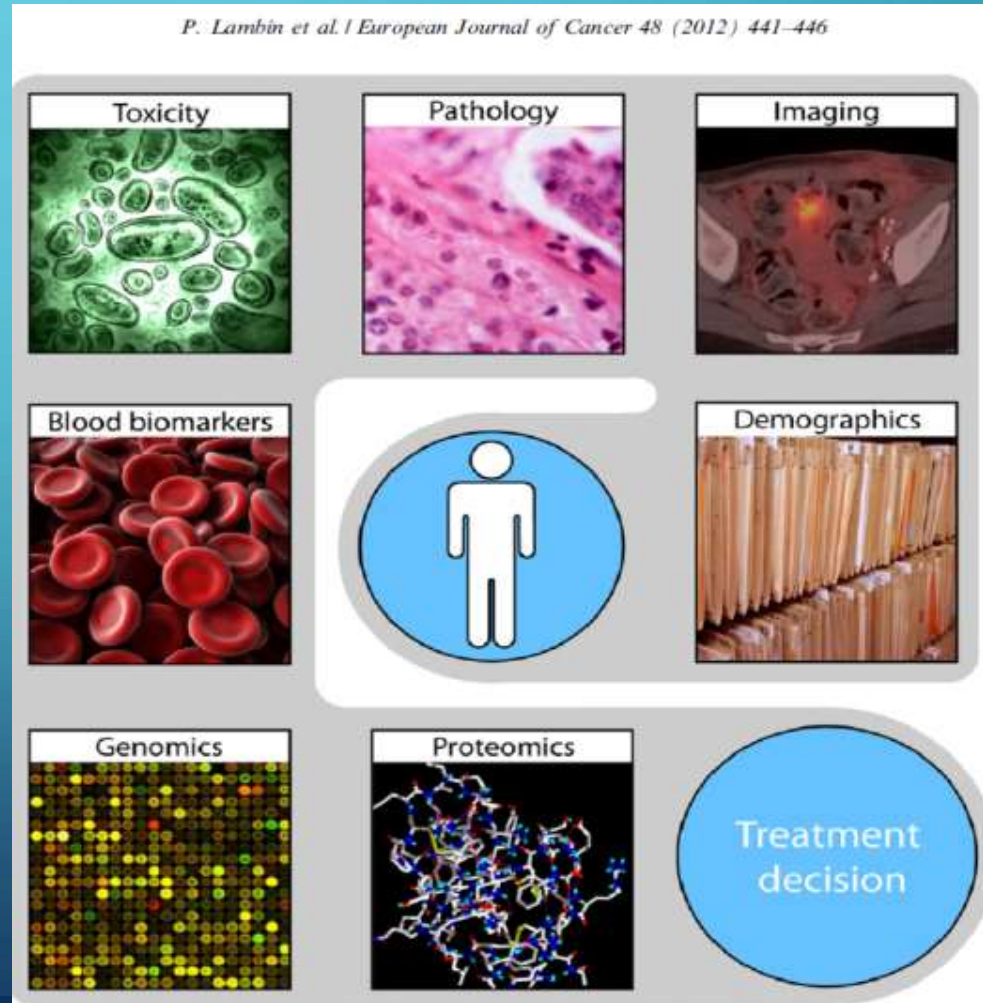
VELOCITY

VERACITY

IMAGERIE & BIG DATA

- Imagerie = élément du big data

→ Intégration de l'imagerie dans les bases de données



VOLUME

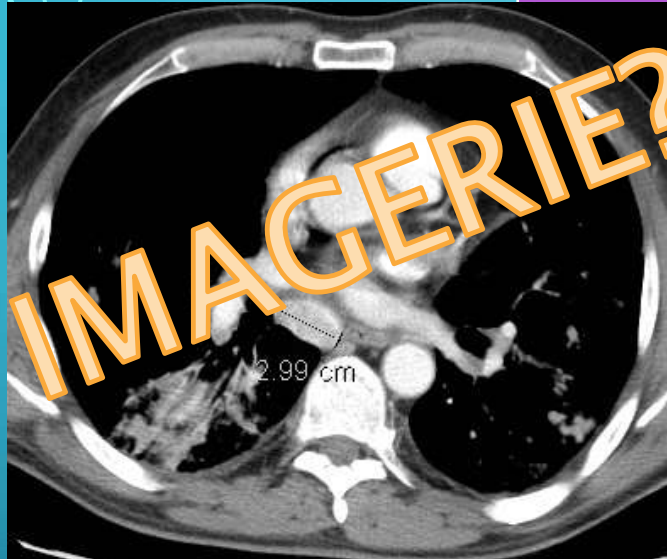
Entrepôt de données HEGP



Type	# instances	# patients
Demographics (age, sex, Hospital vital status)		742 487
Vital signs (temperature, blood pressure, weight, ...)	14 million	141 725
Diagnoses (ICD 10 codes)	6 million	338 871
Medical Procedures (French CCAM codes)	4 million	285 097
Clinical data (EHR DxCare forms)	70 million	468 057
Free Text reports (Hospitalization, Surgery, Imaging, Pathology ...)	3.5 million	359 547
Lab test results	103 million	379 483
Drug prescriptions	3.5 million	121 441
Omics data	80 million	173

VARIETY

Entrepôt de données HEGP

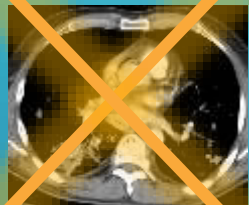


Type	# instances	# patients
Demographics (age, sex, Hospital vital signs)		742 487
Vital signs (temperature, blood pressure, heart rate, ...)	14 million	141 725
Diseases (ICD 10 codes)	6 million	338 871
Surgical Procedures (French CCAM codes)	4 million	285 097
Clinical data (EHR DxCare forms)	70 million	468 057
Free Text reports (Hospitalization, Surgery, Imaging, Pathology ...)	3.5 million	359 547
Lab test results	103 million	379 483
Drug prescriptions	3.5 million	121 441
Omics data	80 million	173

Entrepôt de données HEGP



IMAGERIE?



Information
dérivée
de l'image

Type	# instances	# patients
Demographics (age, sex, Hospital vital status)		742 487
temperature, blood pressure,	14 million	141 725
ICD 10 codes)	6 million	338 871
Procedures (French CCAM codes)	4 million	285 097
Clinical data (EHR DxCare forms)	70 million	468 057
Free Text reports (Hospitalization, Surgery, Imaging, Pathology ...)	3.5 million	359 547
Lab test results	103 million	379 483
Drug prescriptions	3.5 million	121 441
Omics data	80 million	173

COMMENT INTÉGRER L'IMAGERIE DANS LES ENTREPOTS/BASES DE DONNÉES?

- Informations dérivées de l'imagerie → pas les données brutes
- Ne représentent que l'état des connaissances à un moment donné
- Solution = banques d'images
- PROBLEMATIQUES A RESOUDRE
 - Ethique? → *Session JFR Teleradiologie et bases de données, vendredi 8h30*
 - Espace de stockage?
 - Interfaçage image / données? → B Rance

COHORTES DE PATIENTS – BANQUES D'IMAGES

1. Espace de stockage!

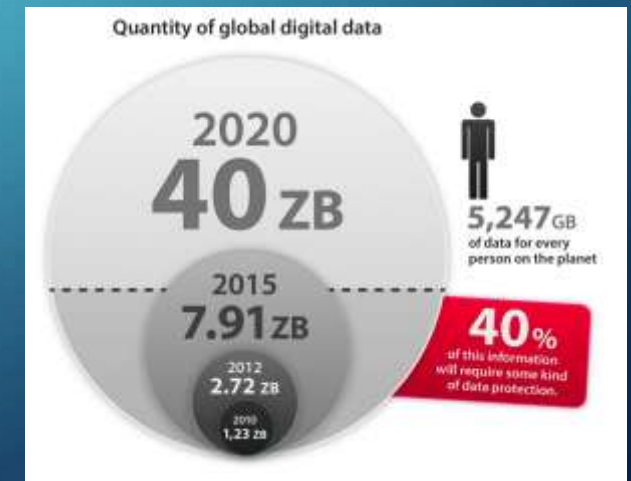
QUESTION

- Quel est le volume de données produit par personne par jour dans le monde?
 - a) 1 megabyte (10^6)
 - b) 1 gigabyte (10^9)
 - c) 1 terabyte (10^{12})

QUESTION

- Quel est le volume de données produit par personne par jour dans le monde?
 - a) 1 megabyte (10^6)
 - b) 1 gigabyte (10^9)
 - c) 1 terabyte (10^{12})

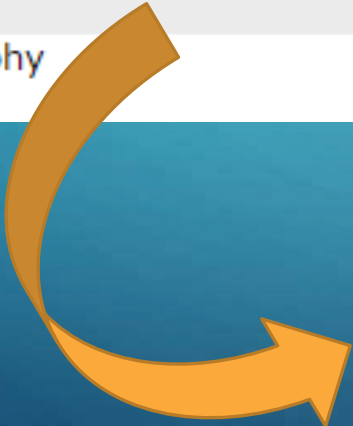
Soit 2,5 exabytes (10^{18}) au total par jour!
(3 milliards d'individus connectés dans le



QUESTION

The most common features of study

Value	Description	Matric(px)	Bit per px	Size	Study count
CD	Color flow Doppler	768 x 576	8	0,442 MB	
CR	Computed radiography	3520 x 4280	12	30 MB	2
CT	Computed tomography	512 x 512	16	0,524 MB	40-3000
DSA	Digital Subtraction Angiography	512 x 512	8		15-40
DX	Digital Radiography	2048 x 2048	12		2
MG	Mammography	4608 x 5200	14	45,7 MB	1
MR	Magnetic Resonance	256 x 256	16	0,131 MB	60-3000
NM	Nuclear Medicine	256 x 256		0,128 MB	
PET	Positron Emission Tomography - PET	128 x 128		32 MB	
US	Ultrasound	512 x 512	8	0,262 MB	20-240
XA	X-Ray Angiography	512 x 512	16		



Tomosynthèse

TDM cardiaque, de perfusion...

IRM neurologique fonctionnelle...

→ Plusieurs GB/examen

COHORTES DE PATIENTS – BANQUES D'IMAGES

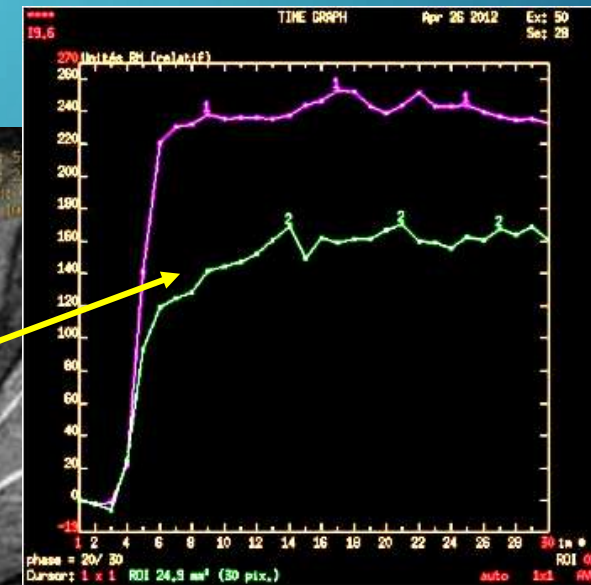
1. Espace de stockage!
2. Annotation des données!

VERACITY

VELOCITY

COHORTES DE PATIENTS – BANQUES D'IMAGES

1. Espace de stockage!
2. Annotation des données!



Le radiologue annote les images dans le PACS → information perdue!

COHORTES DE PATIENTS – BANQUES D'IMAGES

1. Espace de stockage!
2. Annotation des données!

RESULTATS :

Il existe effectivement un volumineux myome en position corporeale postérieure, sous séreuse de type 6 de la classification de FIGO. Celui-ci mesure 90 x 90 x 80mm.

Il présente un signal T2 hétérogène, isosignal T1, hypersignal diffusion avec un ADC élevé aux alentours de 1,2 à 1,5 évoquant plutôt un myome cellulaire qu'une dégénérescence sarcomateuse.

Après injection dynamique de produit de contraste le myome se rehausse de manière intense et similaire à celle du myomètre en périphérie avec une zone centrale hypovascularisée.

Par ailleurs, pas d'anomalie annexielle.
Pas d'épanchement.
Pas d'adénomégalie ilio-inguinale.

Taille

Localisation

Signal T2

Signal T1

Signal DWI

ADC

Diagnostic

Le radiologue fait des comptes rendus → information perdue!

COHORTES DE PATIENTS – BANQUES D'IMAGES

1. Espace de stockage!

2. Annotation des données!

- DONNEES NON STRUCTUREES, NON EXHAUSTIVES
- Outils de sémantique standardisés

2 projets

→ DICOM Structured Reporting (DICOM-SR)

→ Annotation and Image Markup (AIM) project (Stanford, NCI)

COHORTES DE PATIENTS – BANQUES D'IMAGES

- Exemple: Cancer Imaging Archive : portail d'imagerie correspondant à des patients du TCGA (the Cancer Genome Atlas) colligé par plusieurs institutions américaines: Henry Ford Hospital, University of California San Francisco, MD Anderson Cancer Center, Emory University, and Thomas Jefferson University Hospital
- Images anonymisées liées à certaines données cliniques

COHORTES DE PATIENTS – BANQUES D'IMAGES

- Exemple: Cancer Atlas correspondant à (The Cancer Atlas) colligé par le Memorial Sloan-Kettering Cancer Center, EMMES Corporation, University Hospital
- Images anonymisées

research manuscript. You can subscribe to our [Email List](#) or social media feeds to be notified of new collections and changes to existing collections.

Show entries Filter table:

Cancer Type	Collection	Location	Subjects	Modalities
Ovarian Serous Cystadenocarcinoma	TCGA-OV	Ovary	111	CT, MR
Lung Squamous Cell Carcinoma	TCGA-LUSC	Lung	31	CT, NM, PT
Colon Adenocarcinoma	TCGA-COAD	Colon	25	CT
Kidney Chromophobe	TCGA-KICH	Kidney	15	CT, MR
Rectum Adenocarcinoma	TCGA-READ	Rectum	3	CT, MR
Head and Neck Squamous Cell Carcinoma	TCGA-HNSC	Head-Neck	148	CT, MR, PT
Thyroid Cancer	TCGA-THCA	Thyroid	4	CT, PT
Glioblastoma Multiforme	TCGA-GBM	Brain	261	MR, CT, DX
Low Grade Glioma	TCGA-LGG	Brain	199	MR, CT
Liver Hepatocellular Carcinoma	TCGA-LIHC	Liver	65	MR, CT, PT
Lung Adenocarcinoma	TCGA-LUAD	Chest	64	CT, PT, NM
Kidney Renal Clear Cell Carcinoma	TCGA-KIRC	Renal	267	CT, MR, CR
Bladder Endothelial Carcinoma	TCGA-BLCA	Bladder	6	CT, CR
Breast Cancer	TCGA-BRCA	Breast	122	MR, MG
Kidney Renal Papillary Cell Carcinoma	TCGA-KIRP	Renal	31	CT, MR, PT
Prostate Cancer	Prostate-3T	Prostate	64	MR
Head and Neck Carcinomas	QIN-HeadNeck	Head-Neck	156	PT, CT
Prostate Cancer	NaF Prostate (Limited Access)	Prostate	9	PT, CT
Varies	QIN Collections (Limited Access)	Varies	-	Varies

e
Ford
on

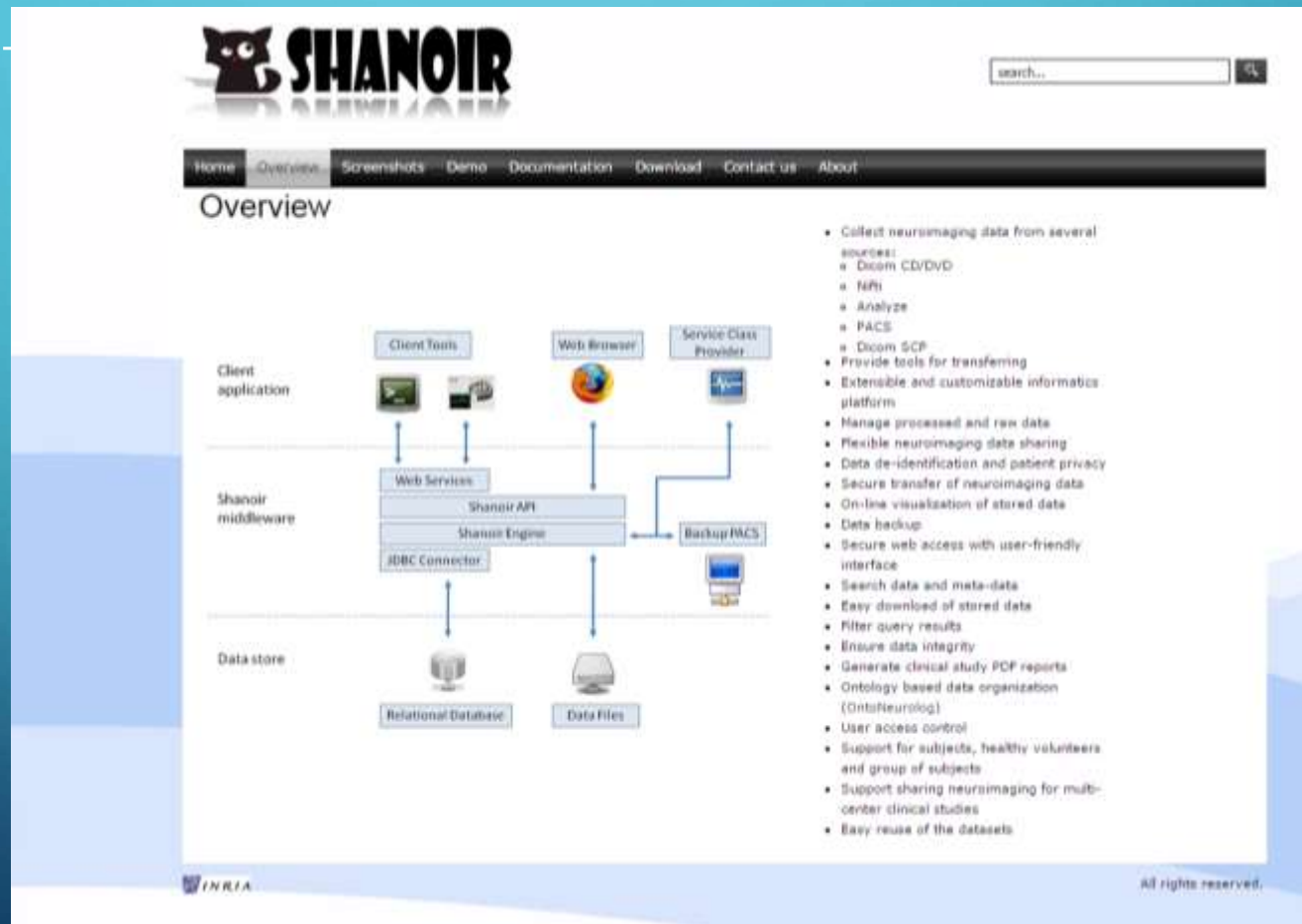
CANCER IMAGING ARCHIVE–CROWDS CURE CANCER



Annotation
d'images de la
banque par des
radiologues
pendant le RSNA

IMAGERIE ET BIG DATA – BANQUES D'IMAGES

- Peu de grandes banques d'images annotées avec métadonnées
- Neurologie ++





IMAGERIE = ELEMENT DE BIG DATA ?

OU

IMAGERIE = SOURCE DE BIG DATA ?



DÉMARCHES DE RECHERCHE

Hypothesis Driven Research



Formuler une hypothèse



Collecter les données



Analyser & Interpréter



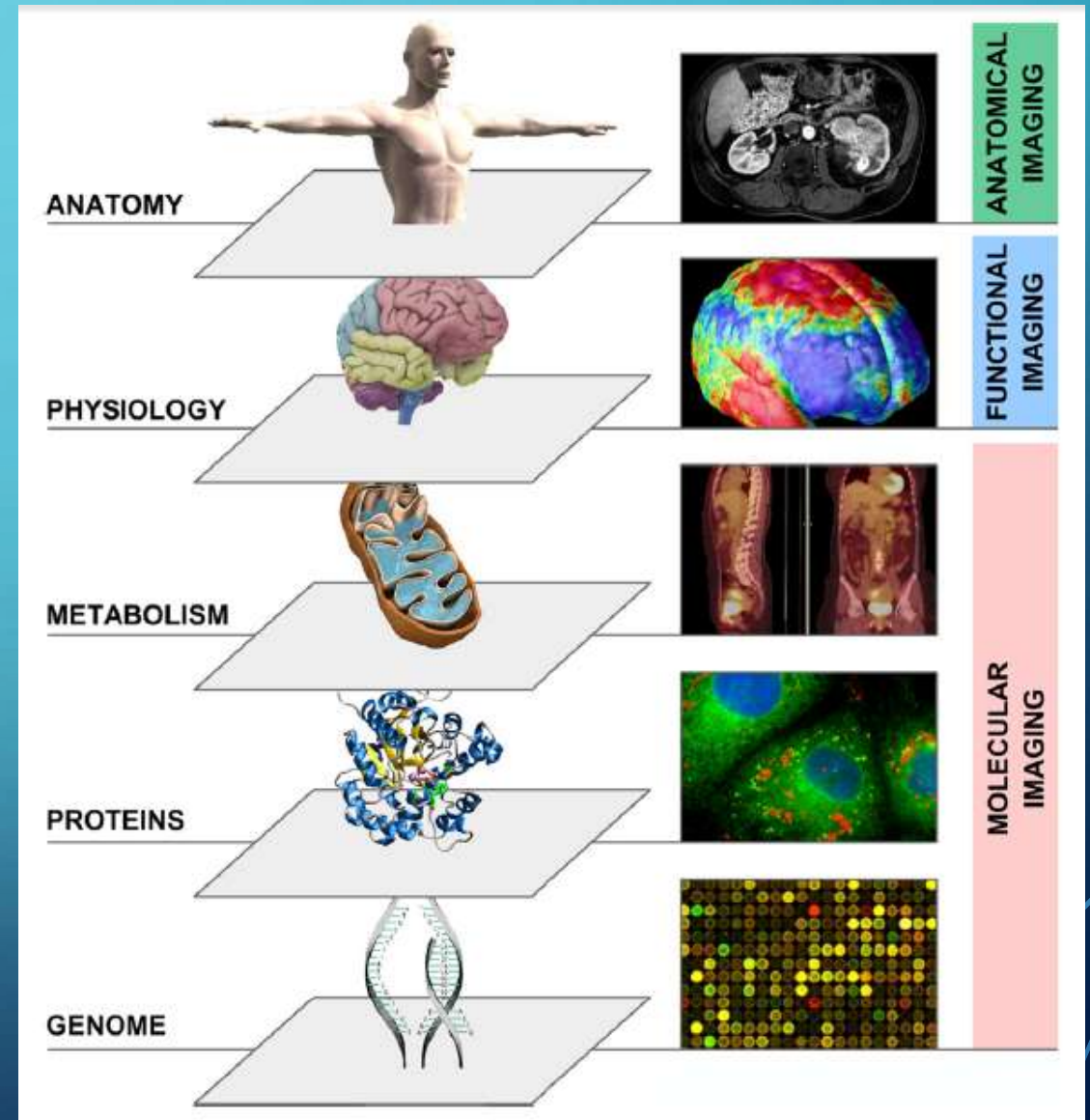
Nouveaux biomarqueurs

LIMITES DE LA DÉMARCHE

- Démarche actuelle: Tester chaque biomarqueur = coûteux en argent et en temps

ET

- L'image n'est qu'une expression phénotypique de la biologie sous-jacente
- Probablement plus d'information dans l'image que ce qu'on extrait
- On n'a pas besoin d'en comprendre le sens pour qu'un paramètre soit utile/utilisable



CHANGEMENT DE PARADIGME

Hypothesis Driven Research



Formuler une hypothèse



Collecter les données



Analyser & Interpréter



Nouveaux biomarqueurs

Data Driven Research



Collecter un grand nombre de données



Analyser & Interpréter



Générer une hypothèse



ALLER PLUS LOIN → RADIOMICS

- Radiomique

- Extraction (+/- automatisée) d'un grand nombre de paramètres quantitatifs en imagerie sans hypothèse a priori



Corrélation à un
« outcome »

Corrélation aux profils
d'expression des gènes
« Radiogenomique »

ETAPES

1. Collecter les données

→ logiciel permettant d'extraire un très grand nombre de paramètres **A PARTIR D'IMAGES OBTENUES EN ROUTINE CLINIQUE**

2. Nettoyer, uniformiser, enrichir les données

→ “data curation” **= SOUVENT ETAPE LIMITANTE**

3. Traiter les données

→ méthodes bio-informatiques et bio-statistiques

1. REDUCTION DU NOMBRE DE DESCRIEURS SELECTIONNES POUR LEUR CARACTERE ROBUSTE ET INFORMATIF
2. TESTES POUR LEUR PERTINENCE CLINIQUE

TYPES DE DESCRIPTEURS

- Géométrie
 - Forme, contours...
- Niveaux de gris
 - moyenne, histogramme
- Organisation spatiale



Paramètres quantitatifs



TYPES DE DESCRIPTEURS

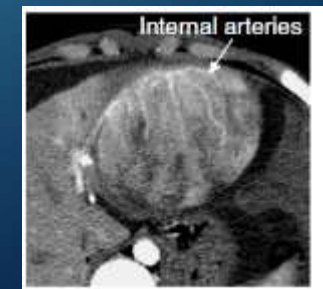
Daniel Balvay

- Géométrie
 - Forme, contours...
- Niveaux de gris
 - moyenne, histogramme
- Organisation spatiale
- Annotations sémantiques



Paramètres quantitatifs

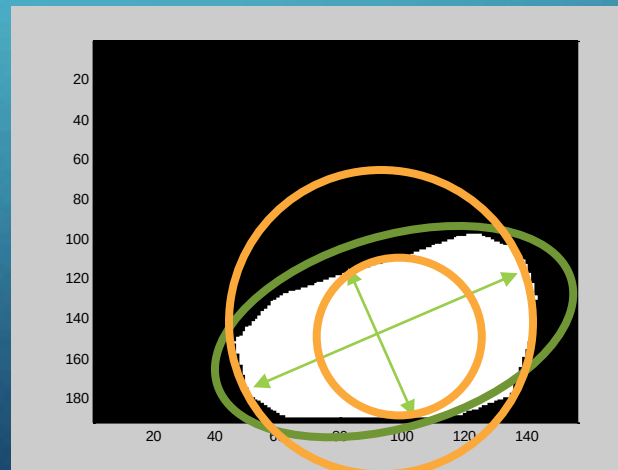
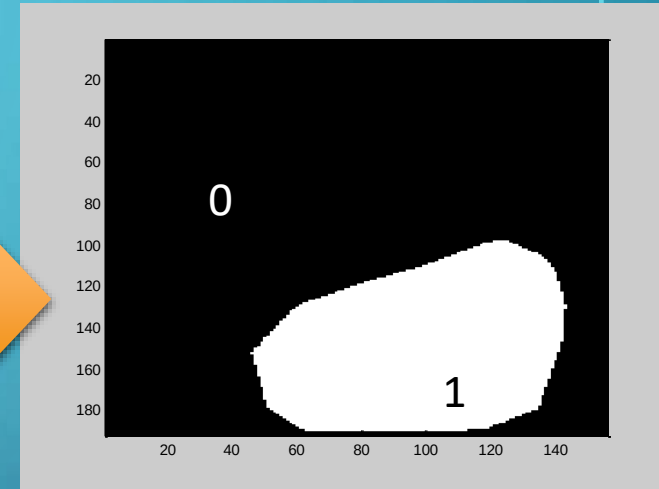
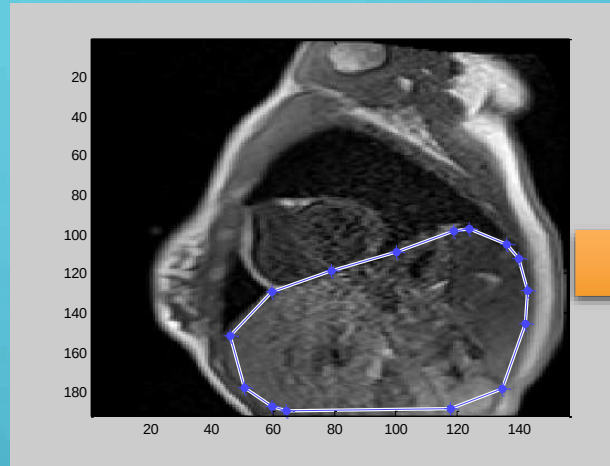
TEXTURE



TYPES DE DESCRIPTEURS: GEOMETRIE

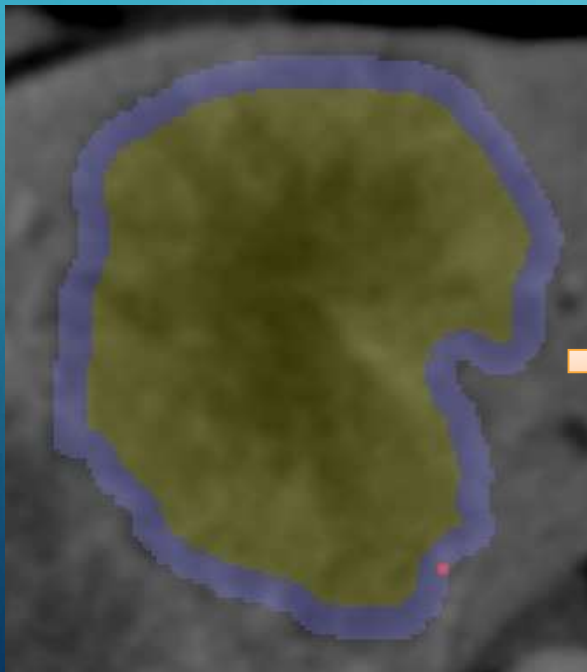
Daniel Balvay

- Analyse de contours
- Indice de compacité
- Surface, volume, direction dans l'espace...



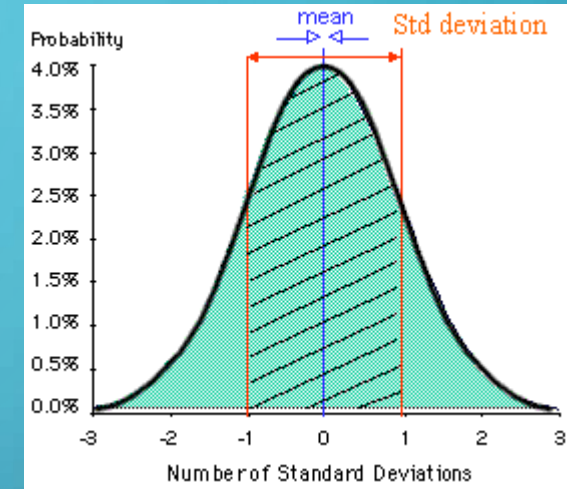
TYPES DE DESCRIPTEURS: NIVEAUX DE GRIS

- Analyse de niveaux de gris
 - Extraction de paramètres quantitatifs décrivant la « composition » des pixels
 - Non ou peu visible à l'œil

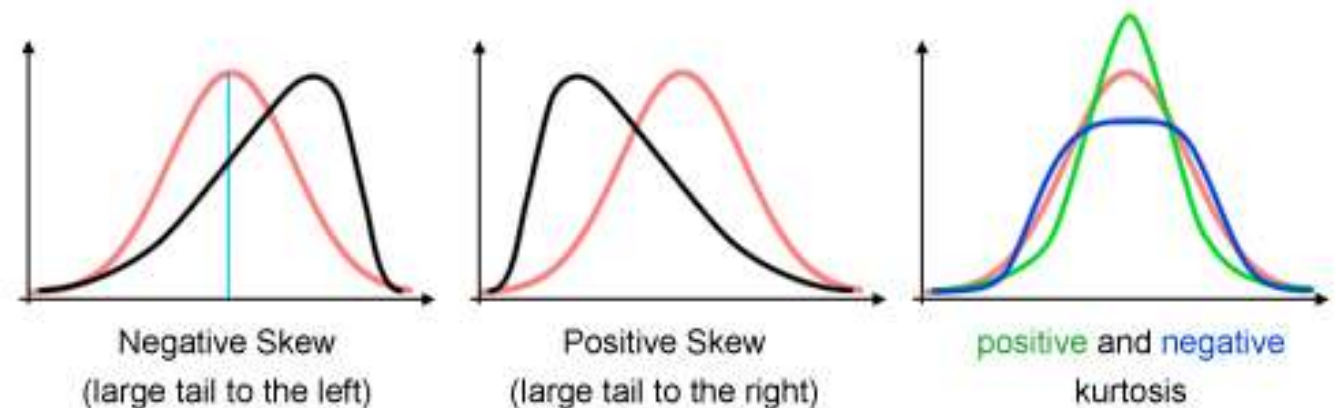


TYPES DE DESCRIPTEURS: HISTOGRAMME

- **Moyenne** → centre de la distribution
- **Déviati**on standard ou **Variance** → échelle de la distribution
- **Skewness** → symétrie de la distribution
- **Kurtosis** → « pointu » de la distribution

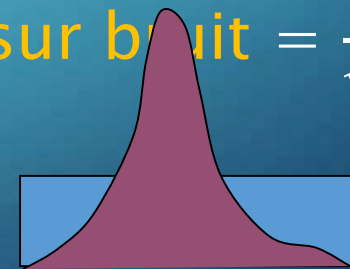


© www.scratchapixel.com



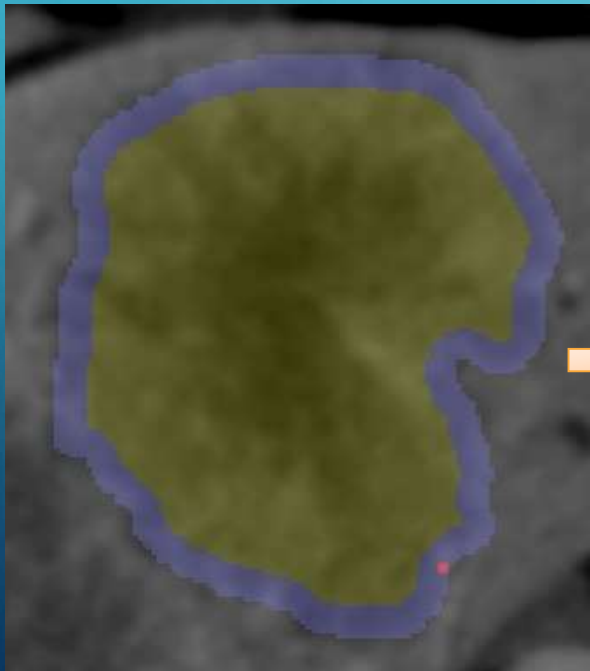
TYPES DE DESCRIPTEURS: HISTOGRAMME

- **Moyenne** → centre de la distribution
- **Déviati**on standard ou **Variance** → échelle de la distribution
- **Skewness** → symétrie de la distribution
- **Kurtosis** → « pointu » de la distribution
- **Entropie** = $-\sum \log(p(i))p(i)$ → mesure l'irrégularité ou complexité, indiquée par une valeur élevée
- **Energie** (uniformité) = $\sum p(i)^2$
- **Signal sur bruit** = $\frac{MOY}{\sqrt{Var}}$



LIMITES DE L'HISTOGRAMME

- Analyse de niveaux de gris
 - Description de la « **composition** » des pixels mais pas de la « **disposition** »



TYPES DE DESCRIPTEURS: DISTRIBUTION SPATIALE

- **Gray level co-occurrence matrix** : changement en niveaux de gris / mouvement de translation
- **Gray level Run-Length matrix** : nombre de lignes de même niveau de gris de différentes longueurs / orientations
- **Gray level Size-Zone matrix** : nombre d'objets de même niveaux de gris de différentes tailles
- **Neighbourhood gray-tone difference matrix** : contrastes locaux / mouvement de translation

TYPES DE DESCRIPTEURS: DISTRIBUTION SPATIALE

- Matrice de co-occurrence

Pour une translation donnée: exemple, 1 pixel

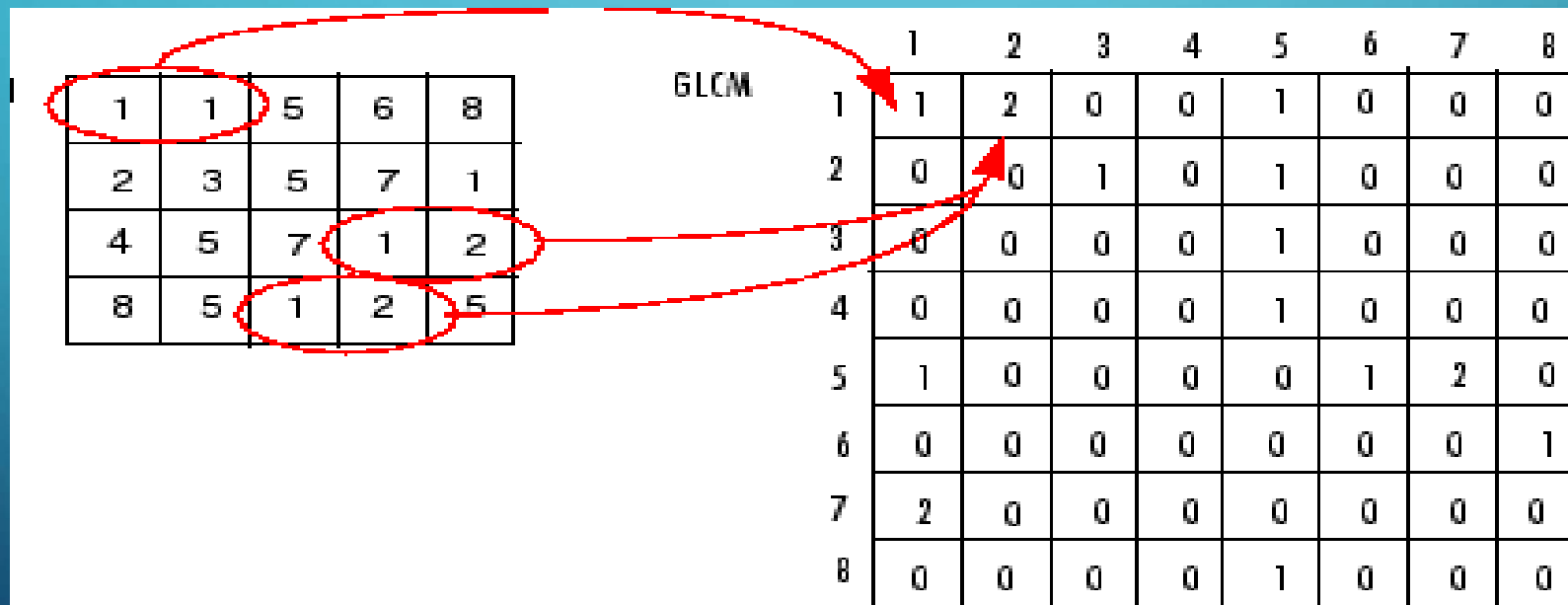


Diagram illustrating the calculation of a GLCM (Gray-Level Co-occurrence Matrix) for a 5x5 input grid. The input grid is shown on the left, and the resulting GLCM is shown on the right. Red circles highlight specific pixel pairs in the input grid, and red arrows point to their corresponding entries in the GLCM matrix.

Input Grid (5x5):

1	1	5	6	8
2	3	5	7	1
4	5	7	1	2
8	5	1	2	5

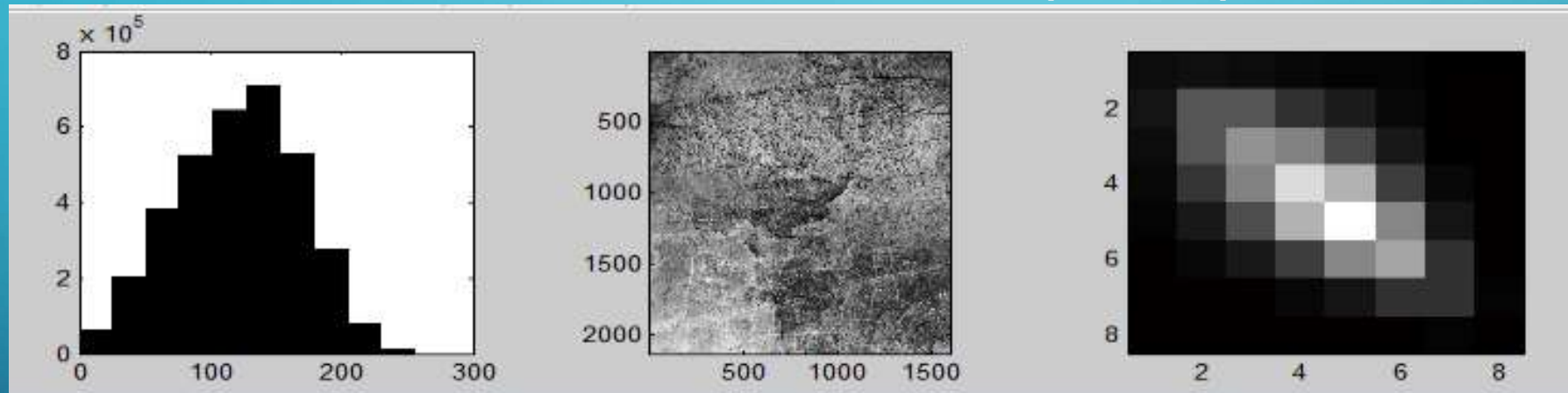
GLCM (8x8):

	1	2	3	4	5	6	7	8
1	1	2	0	0	1	0	0	0
2	0	0	1	0	1	0	0	0
3	0	0	0	0	1	0	0	0
4	0	0	0	0	1	0	0	0
5	1	0	0	0	0	1	2	0
6	0	0	0	0	0	0	0	1
7	2	0	0	0	0	0	0	0
8	0	0	0	0	1	0	0	0

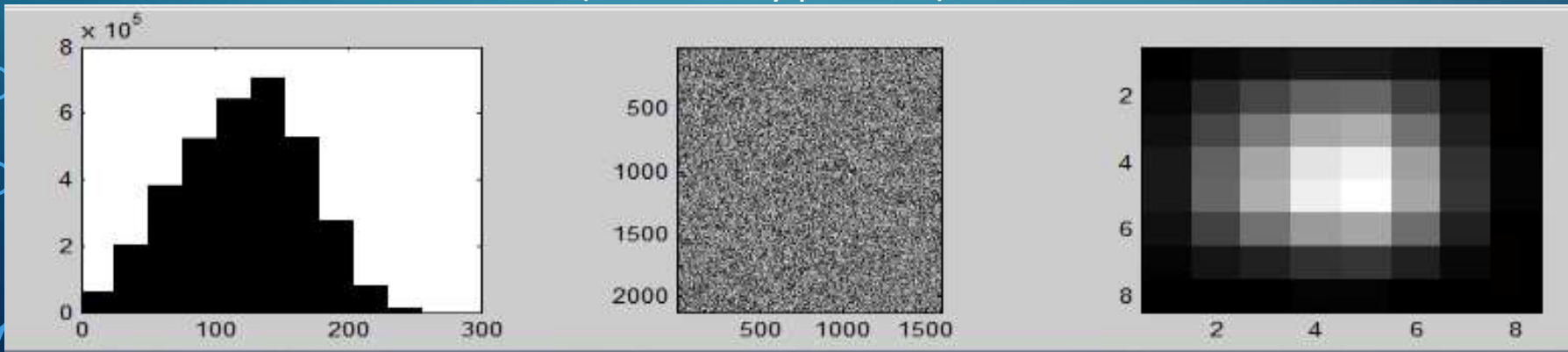
TYPES DE DESCRIPTEURS: DISTRIBUTION SPATIALE

- Matrice de co-occurrence

Pour une translation donnée: exemple, 1 pixel



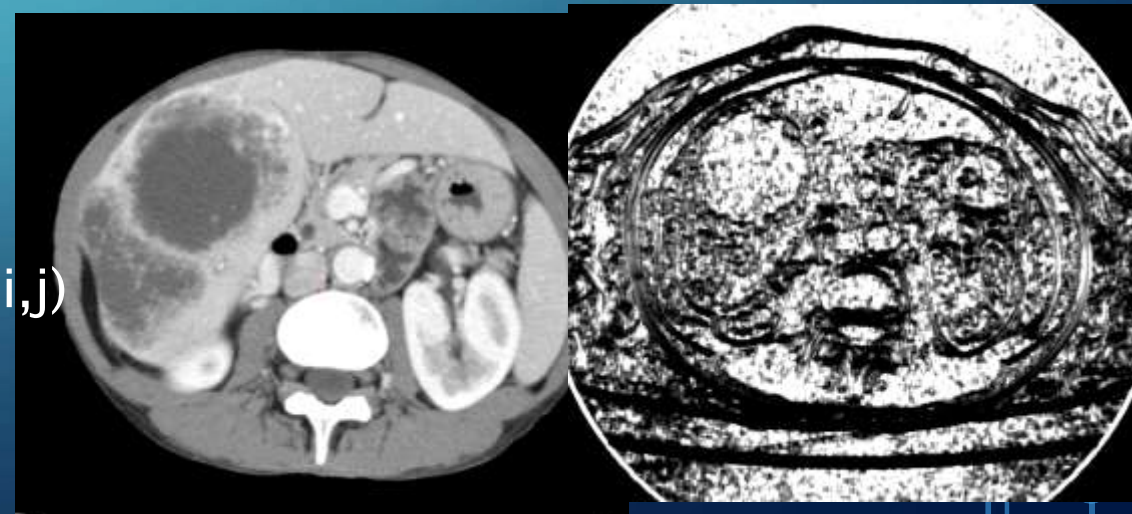
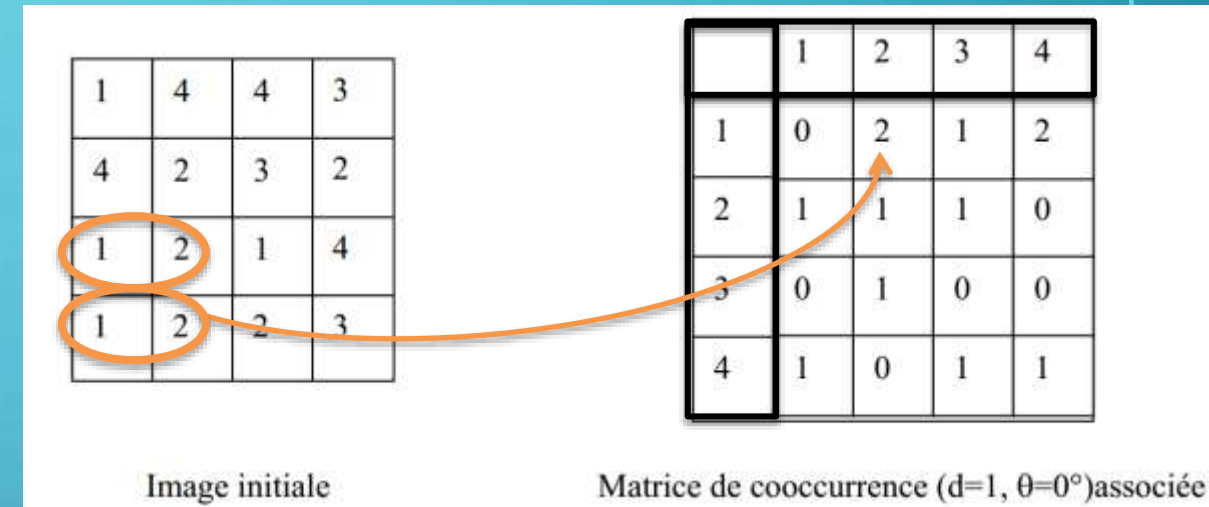
LOSS OF SPATIAL INFORMATION (random x/y position)



TYPES DE DESCRIPTEURS: DISTRIBUTION SPATIALE

- **Contrast** = $\sum \sum (i-j)^2 p(i,j)$
- **Sum Average** = $\sum \sum \frac{i+j}{2} p(i,j)$
- **Inverse different moment** = $\sum \sum \frac{p(i,j)}{1+(i-j)^2}$
- **Homogénéité** = $\sum \sum \frac{p(i,j)}{1+|i-j|}$
- **Cluster Tendency** = $\sum \sum (i+j-mx-my)^2 p(i,j)$
- **Corrélation** = $\sum \sum \frac{(i-mx)(j-my)}{\sigma_x \sigma_y} p(i,j)$
- **Variance d'Haralick** = $\sum \sum [(i-mx)^2 + (j-my)^2] p(i,j)$
- **Probabilité maximale** = $\max p(i,j)$

La matrice de co-occurrence



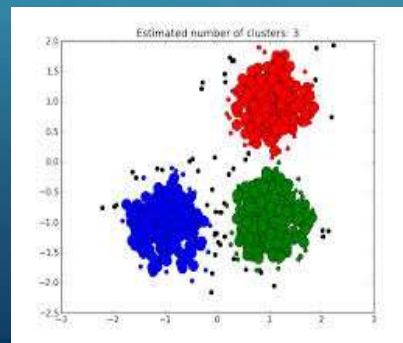
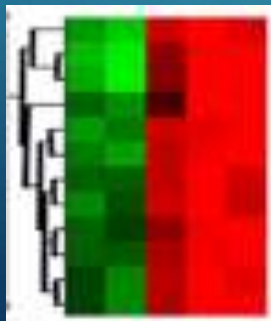
TRAITEMENT DES DONNEES

Afef Bouchouicha

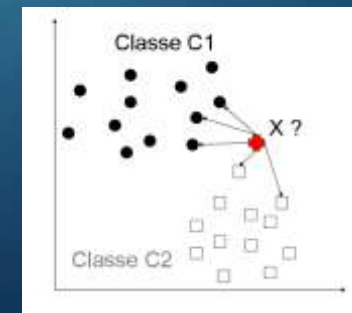
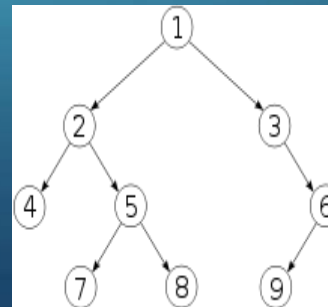


Clustering
hiérarchique
K-means

Approche non supervisée
(décision a posteriori)



Approche supervisée
(décision a priori)



Association rules
C4.5
Random Forest
Naive Bayes
Neural network (aka Deep Learning)

RADIOGENOMICS

**nature
biotechnology**

Decoding global gene expression programs in liver cancer by noninvasive imaging

Eran Segal¹, Claude B Sirlin², Clara Ooi⁴, Adam S Adler⁵, Jeremy Gollub⁶, Xin Chen⁸, Bryan K Chan², George R Matcuk⁷, Christopher T Barry³, Howard Y Chang⁵ & Michael D Kuo²

- Population d'apprentissage: 38 patients avec CHC
 - 138 caractéristiques en imagerie sur des scanners
 - Analyse microarray de 6,732 genes
- Population test: 32 patients avec CHC

RADIOGENOMICS: POPULATION D'APPRENTISSAGE

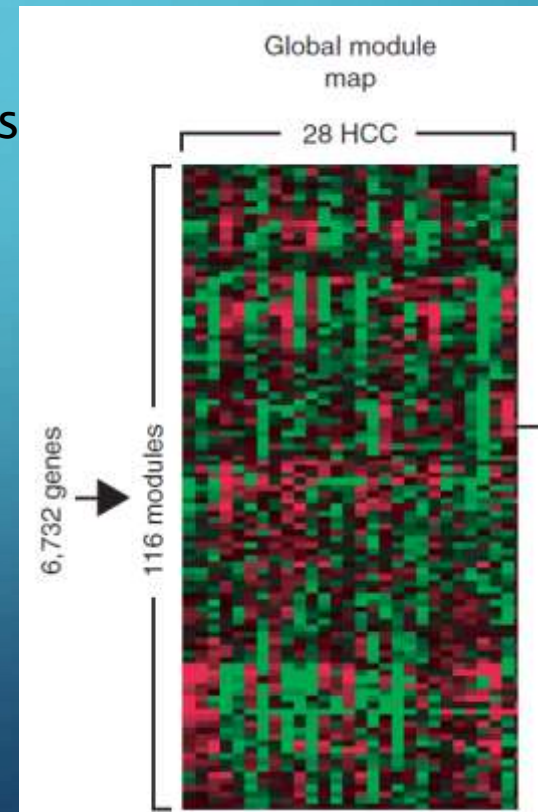
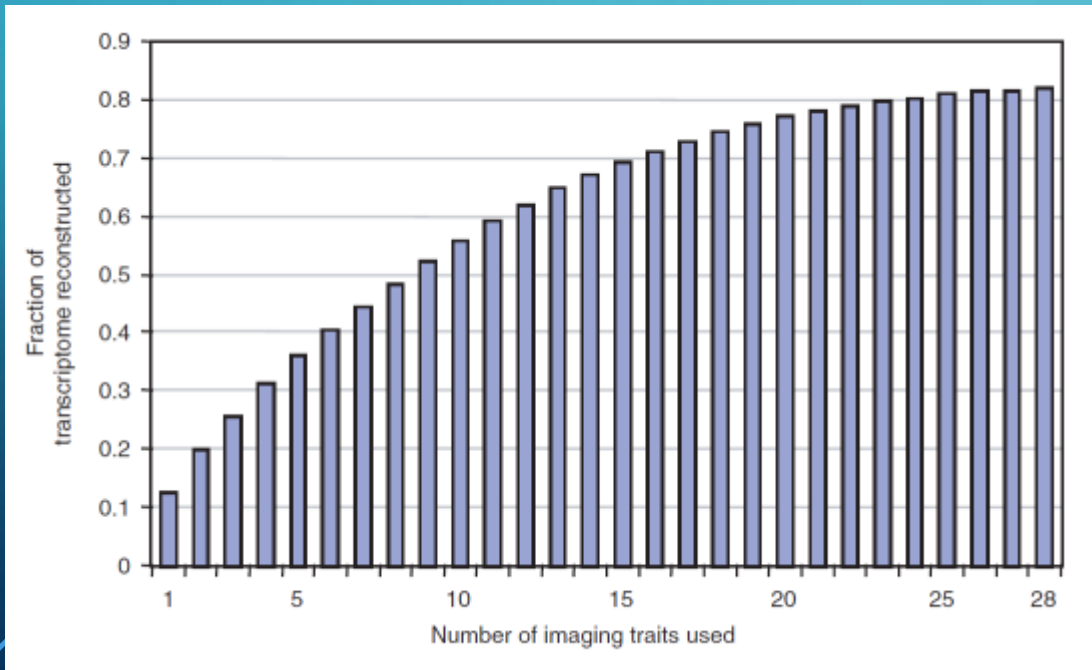
138 imaging traits

6,732 gene expression levels
(microarray)

fréquence and proéminence,
interobserver agreement,
indépendance des autres
paramètres

32 imaging traits

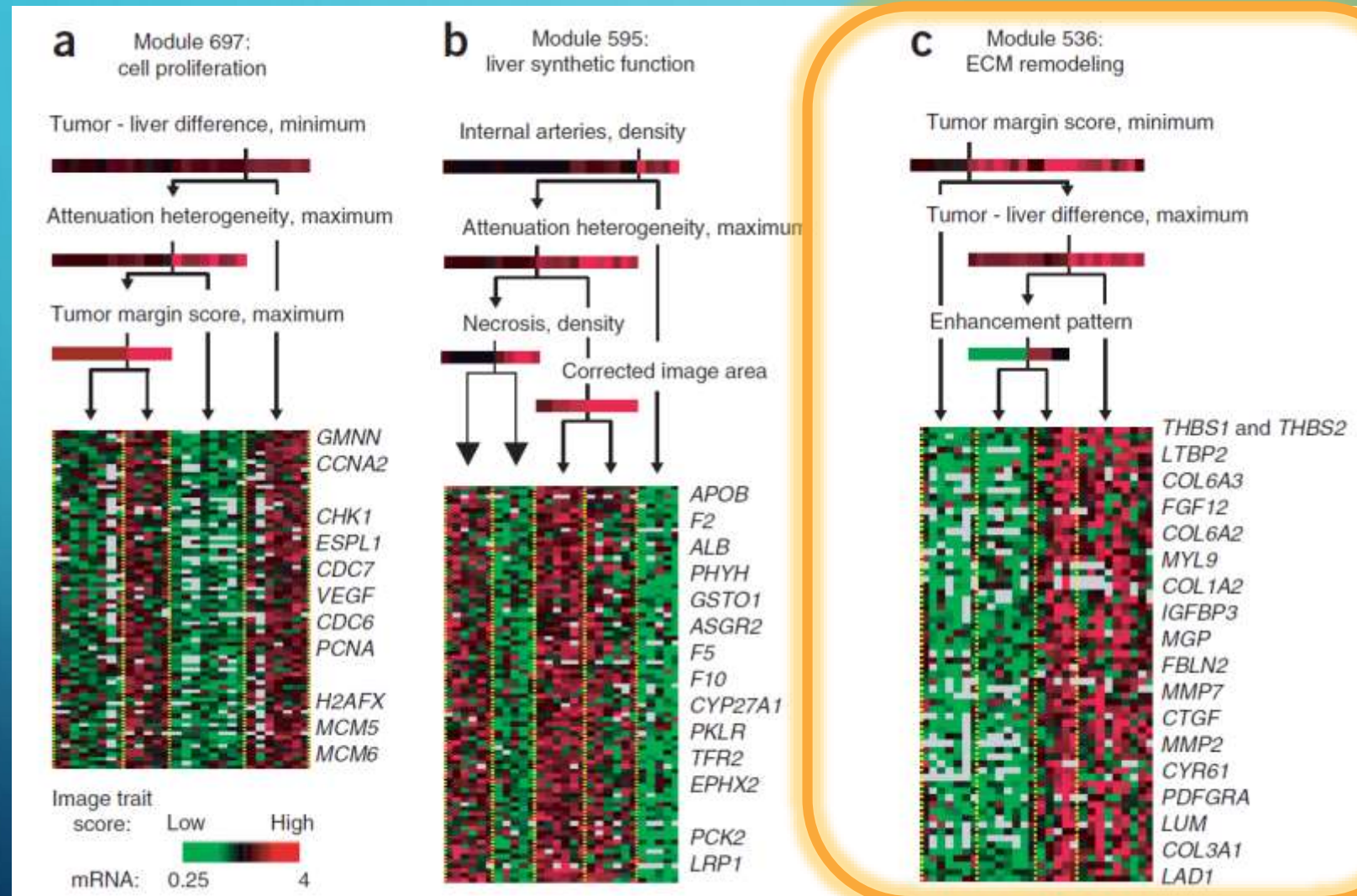
116 gene modules



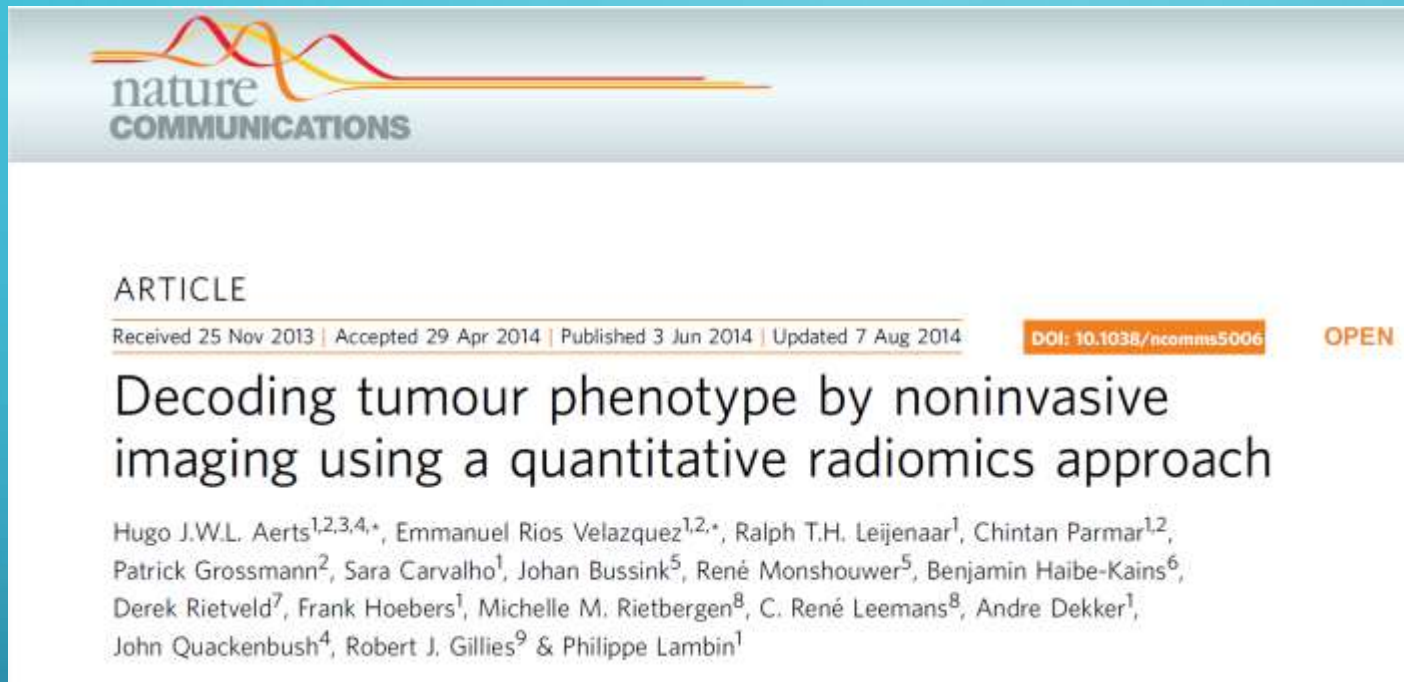
Necrosis, percent
Internal septa
Texture heterogeneity, portal-venous phase
Tumor margin score, minimum
Enhancement pattern
Internal arteries, rank
Hypodense halo
Wash-out, maximum
Internal arteries, density
Tumor - liver difference, maximum
Corrected imaging area
Necrosis, density
Tumor margin score, maximum
Texture heterogeneity, arterial phase
Internal arteries, necrosis edge
Capsule
Wash-in, maximum
Infiltration, percent
Tumor - liver difference, minimum
Attenuation - heterogeneity score, maximum

Segal, Nature Biotechnol 2007

RADIOGENOMICS

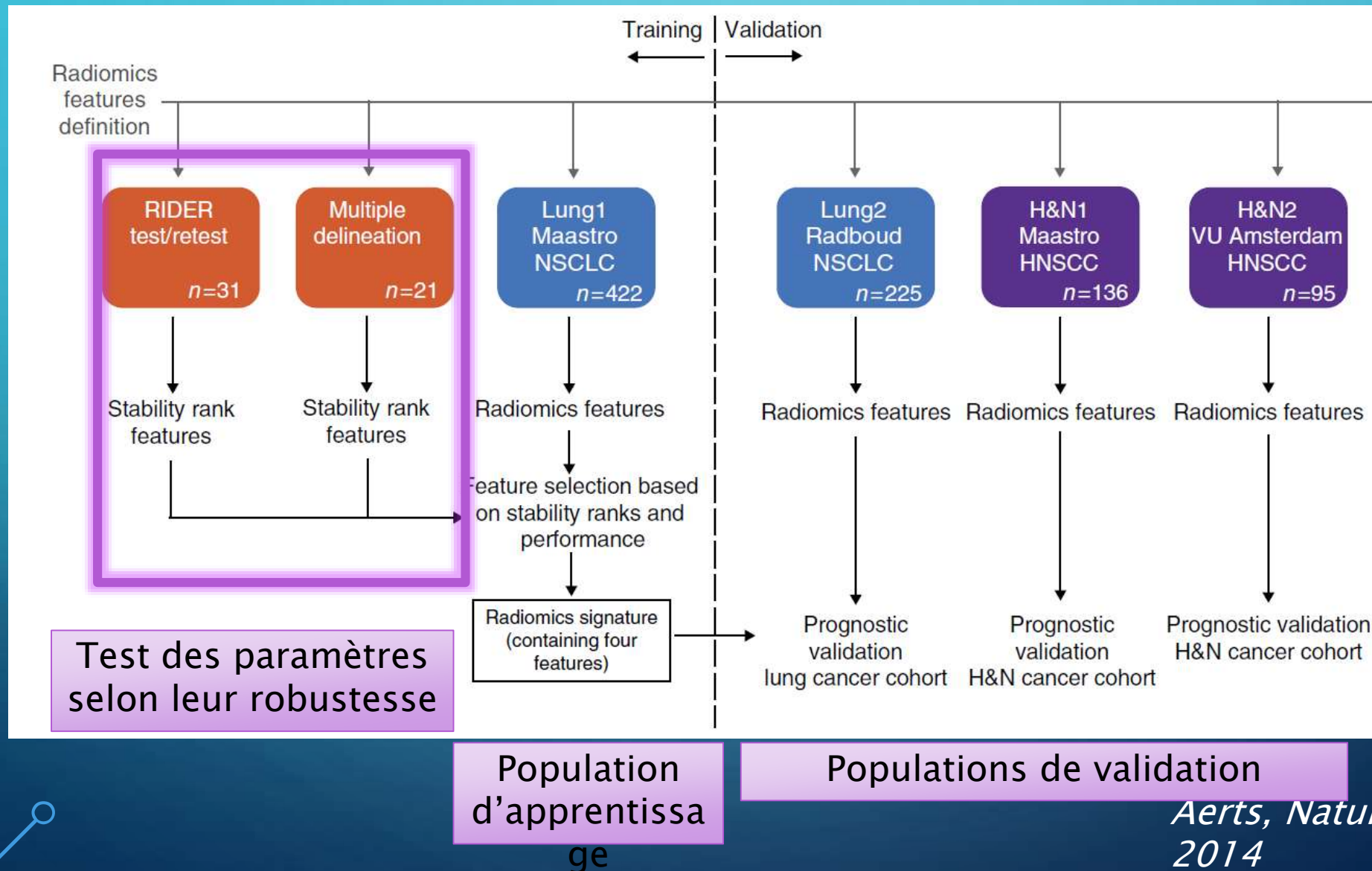


RADIOMICS



- 7 sets de data : cancers poumon et ORL = 1,019 patients
- 440 caractéristiques de radiomique
- Corrélation à la survie et aux profils d'expression génique

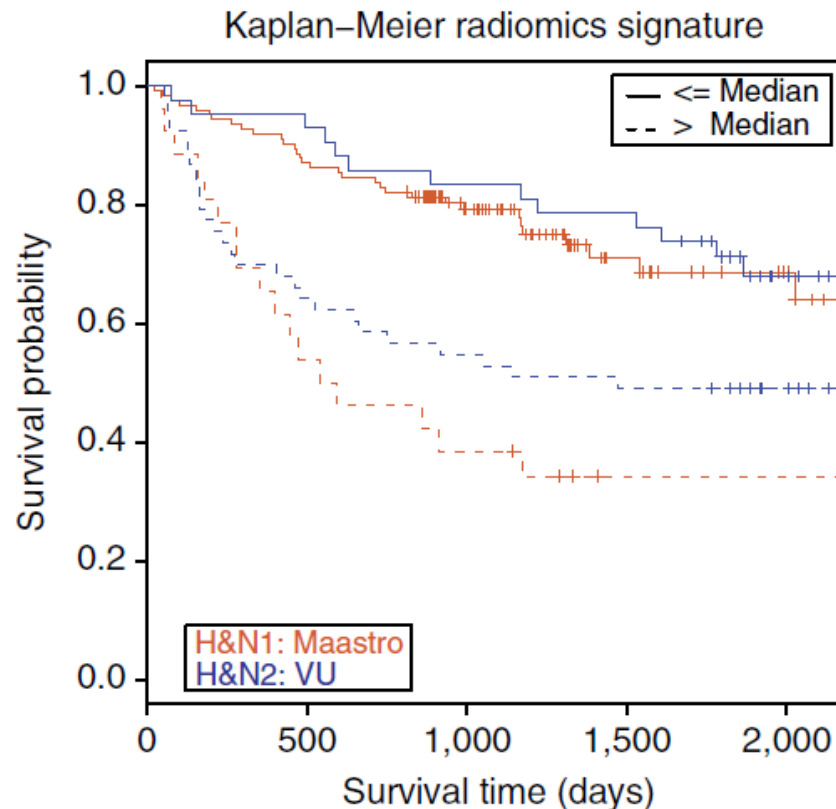
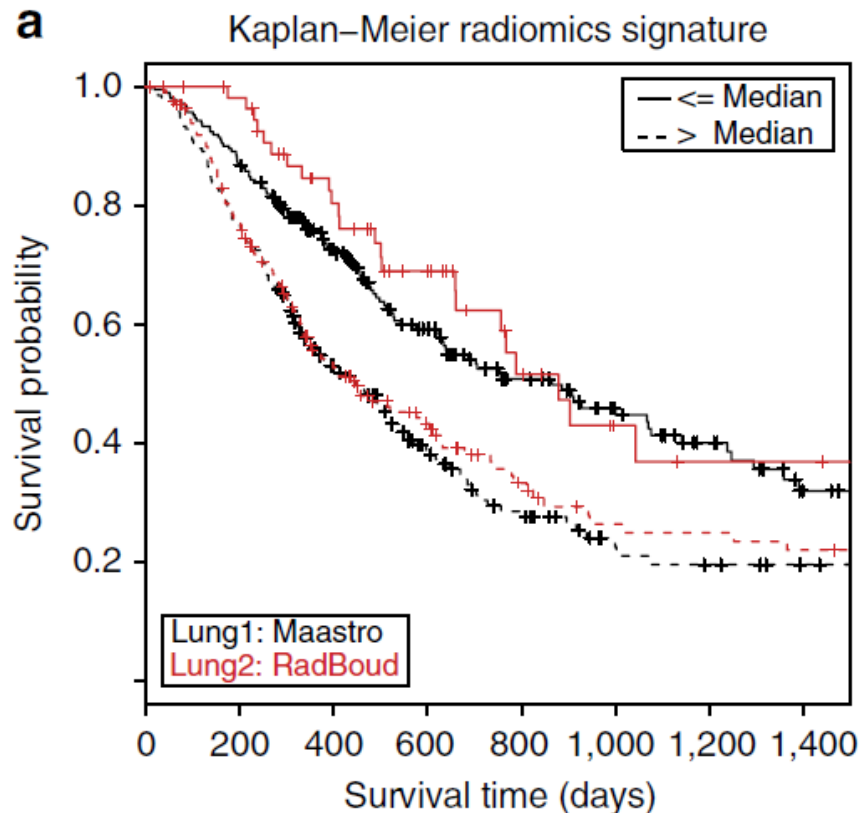
RADIOMICS: ANALYSIS WORKFLOW



RADIOMICS: ANALYSIS WORKFLOW

Features:

- (I) 'Statistics Energy' = overall density of the tumour volume,
- (II) 'Shape Compactness' = how compact the tumour shape is,
- (III) 'Grey Level Nonuniformity' = intratumour heterogeneity
- (IV) wavelet 'Grey Level Nonuniformity HLH' = intratumour heterogeneity



- Les mêmes quatre paramètres prédisent la survie:

- Localisations différentes
- Traitement différents (chimo vs radiothérapie)

- Indépendants de TNM, volume tumoral

- Associés à un pattern d'expression génique
- ature Commun*

RADIOMIQUE: POURQUOI FAIRE?

Pour tester simultanément un grand nombre de nouveaux paramètres extraits des images

→ biomarqueurs potentiels qui auront besoin ensuite d'être compris et testés

Pour générer de nouvelles hypothèses

→ Différents phénotypes en imagerie pourront révéler des nouveaux mécanismes moléculaires qu'il faudra comprendre



Machine learning "gives
computers the ability to learn
without being explicitly
programmed"



Arthur Samuel, 1959

QUID DE L'INTELLIGENCE ARTIFICIELLE?



Geoffrey Hinton

"I think that if you work as a radiologist, you are like Wile E. Coyote in the cartoon. You're already over the edge of the cliff, but you haven't yet looked down. There's no ground underneath. People should stop training radiologists now. It's just completely obvious that in five years deep learning is going to do better than radiologists."

Nov 24, 2016

QUID DE L'INTELLIGENCE ARTIFICIELLE?



Geoffrey



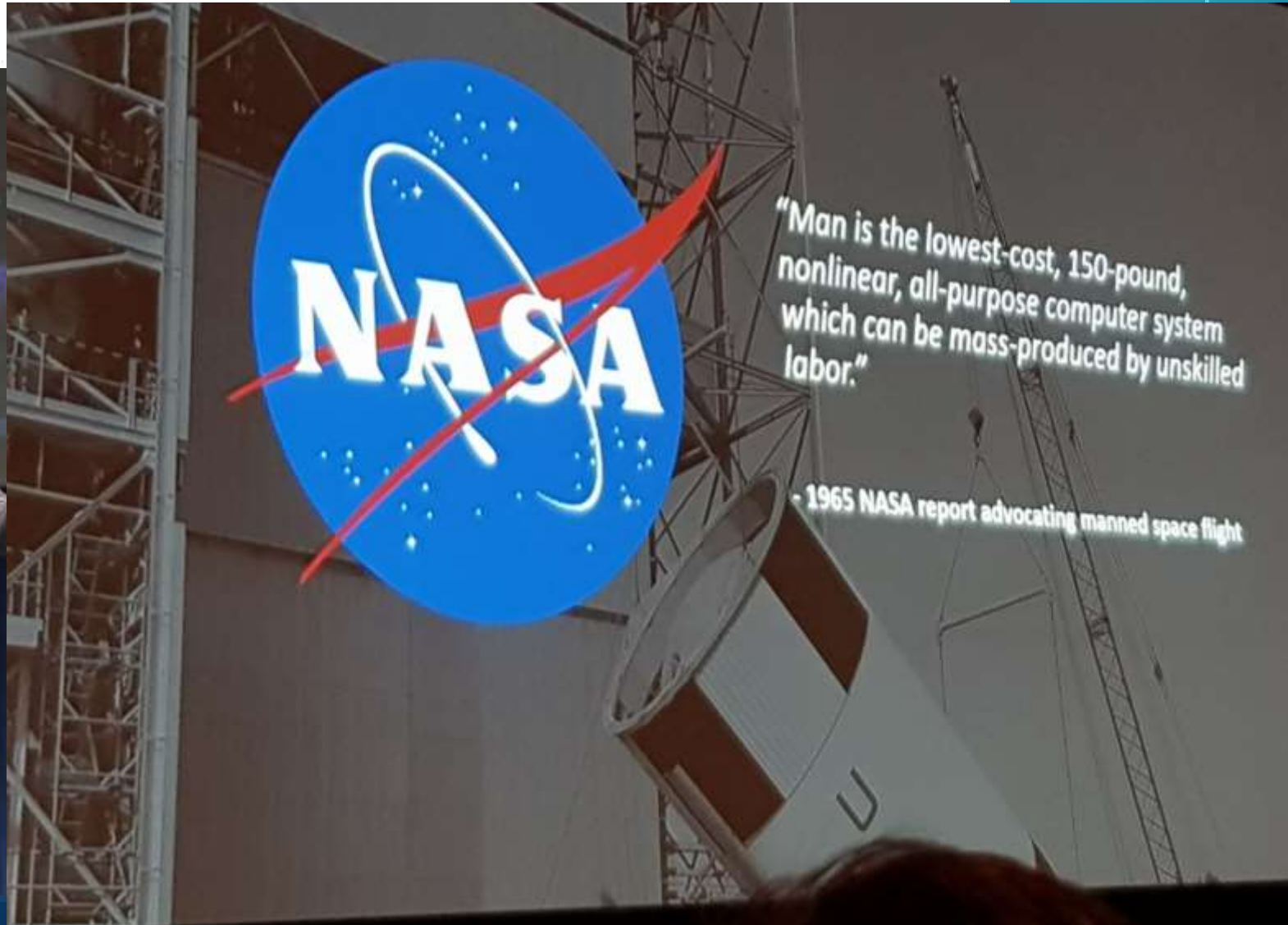
« Radiologists who use AI will replace those who

Bertalan Mesko *The Medical Futurist*

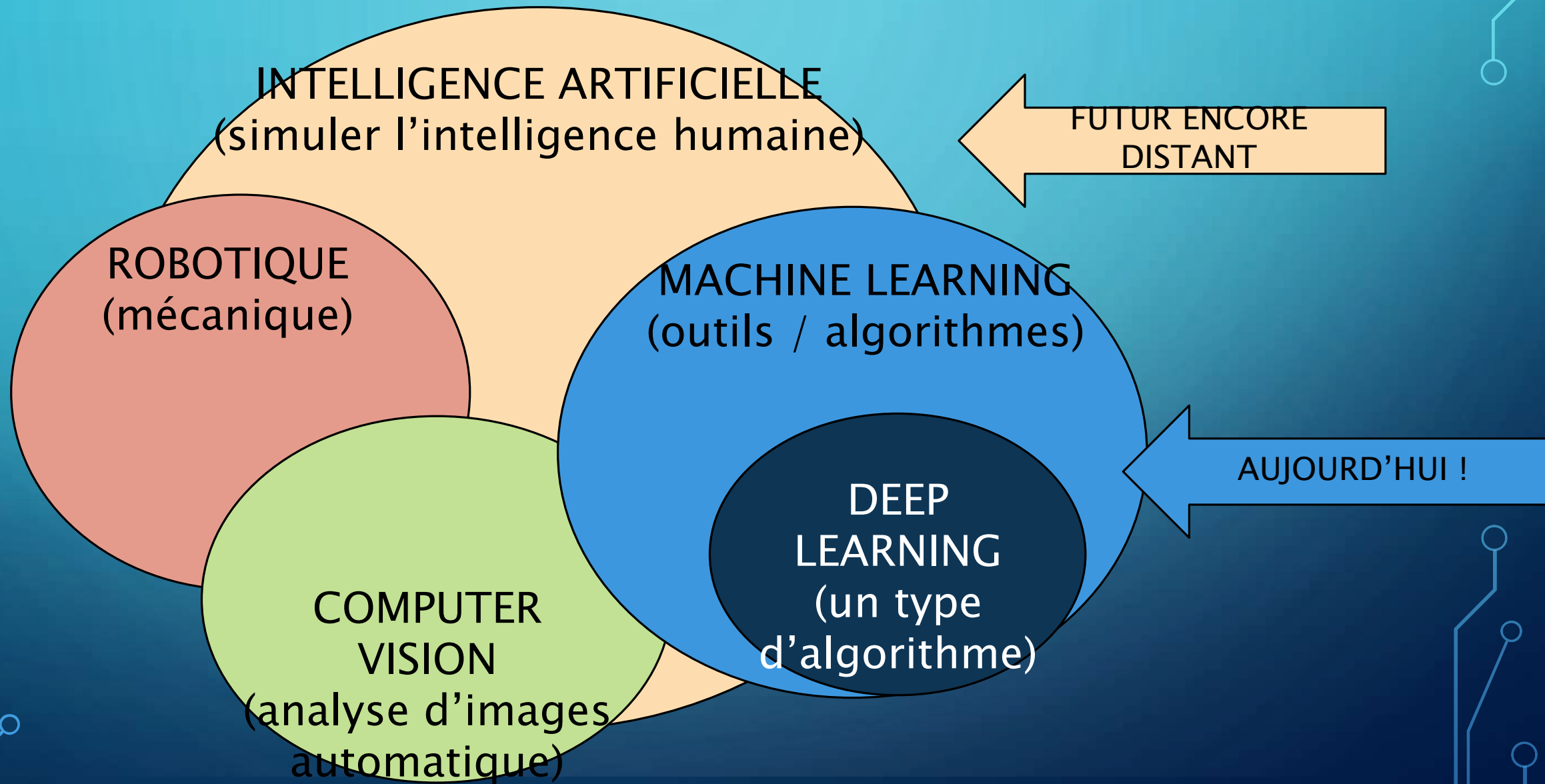
QUID DE L'INTELLIGENCE ARTIFICIELLE?

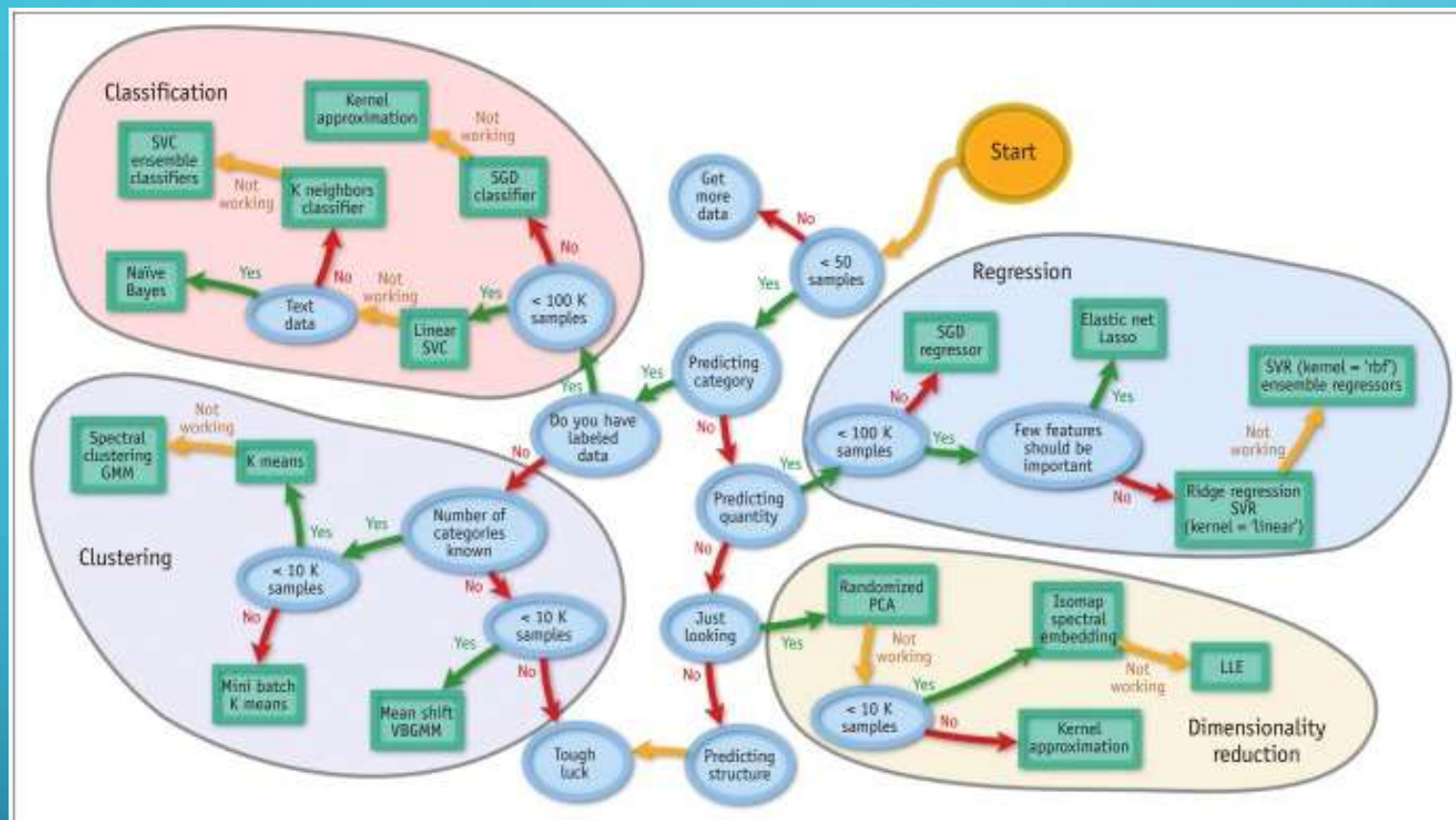


Geoffrey



INTELLIGENCE ARTIFICIELLE: CONCEPTS





Deep Learning in Medical Imaging: General Overview

June-Goo Lee, PhD,¹ Sanghoon Jun, PhD,^{2,3} Young-Won Cho, MS,^{2,3} Hyunna Lee, PhD,^{2,3} Guk Bae Kim, PhD,^{2,3} Joon Beom Seo, MD, PhD,^{2,*} and Namgug Kim, PhD^{2,3,*}

¹Biomedical Engineering Research Center, University of Ulsan College of Medicine, Asan Medical Center, Seoul 05505, Korea.

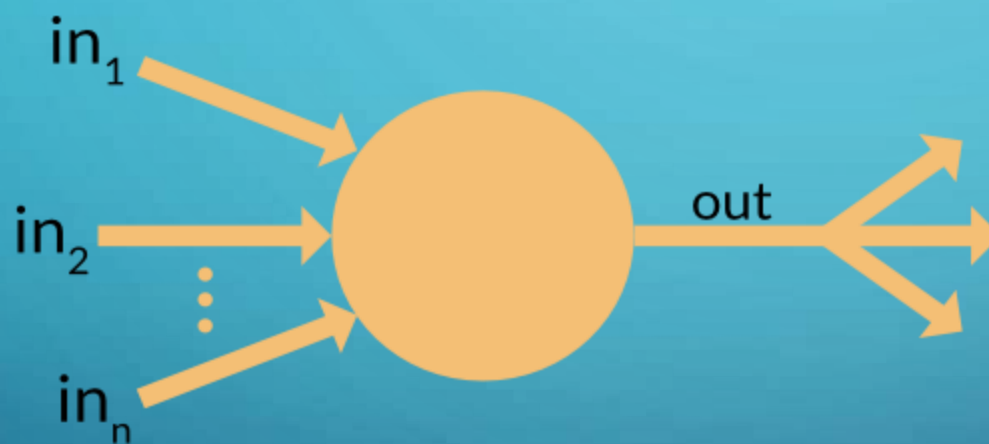
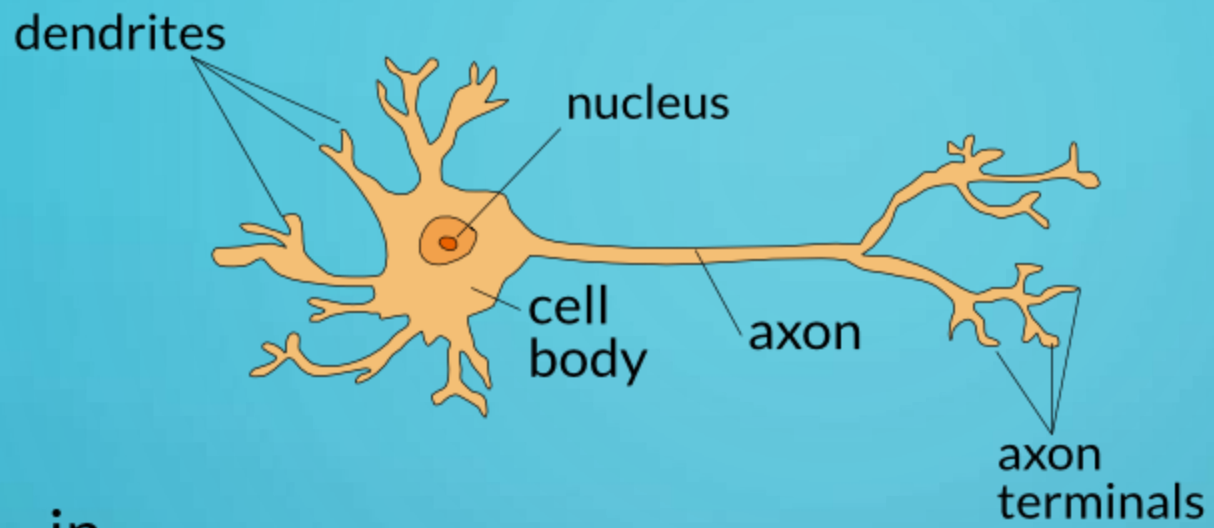
(TRÈS) COURTE HISTOIRE DU DEEP-LEARNING

1958 – F. Rosenblatt (1958). « The Perceptron : a Probabilistic Model for Information Storage and Organisation in the brain », *Psychological Review*, vol. 65 (1958).

1969 – M. Minsky, S. Papert, « Perceptrons », *MIT press* (1969).

1982 – J. J. Hopfield, « Neural networks and physical systems with emergent collective computational abilities », *Proc. Natl. Acad Sci.*, vol. 79, p. 2554 (1982).

2015 – Y. Le Cun, Y. Bengio, G. Hinton « Deep Learning », *Nature*, vol. 521, p. 436 (2015).



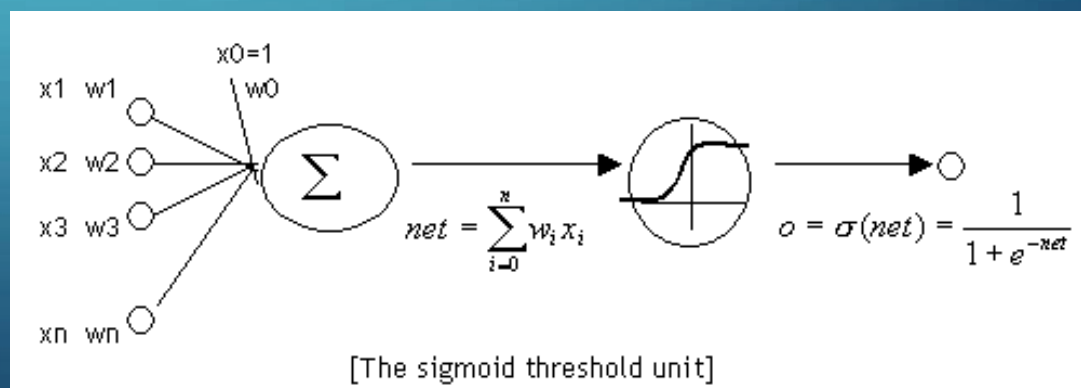
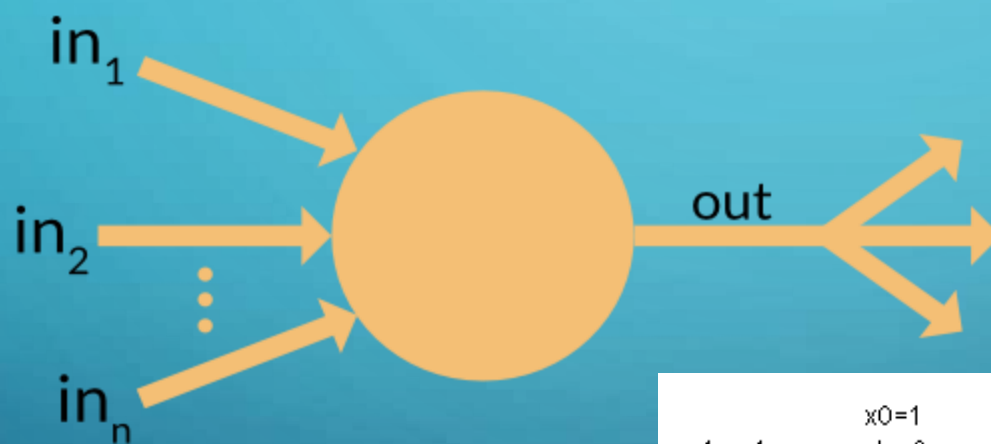
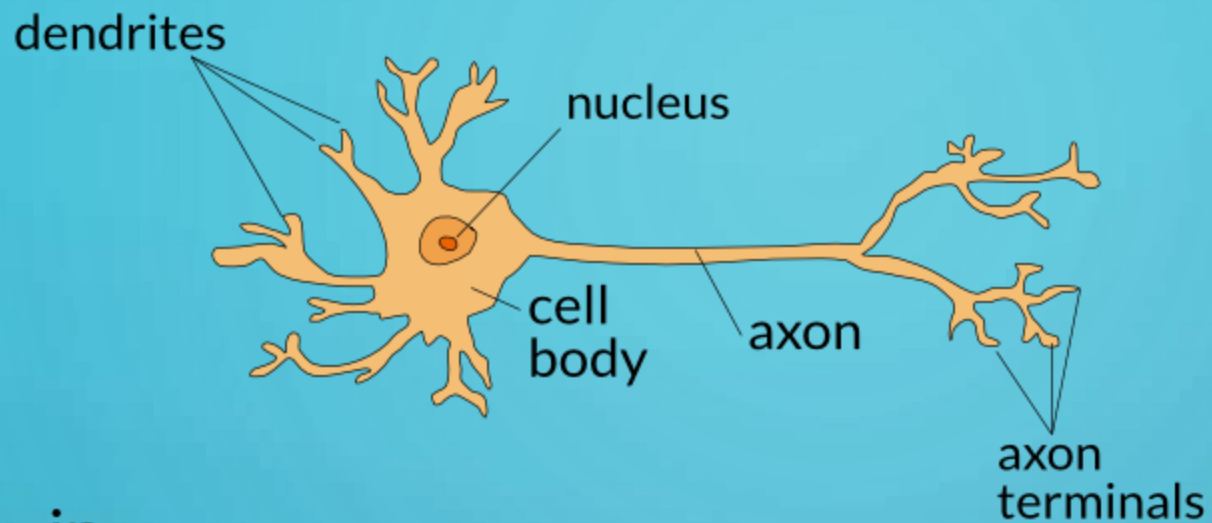
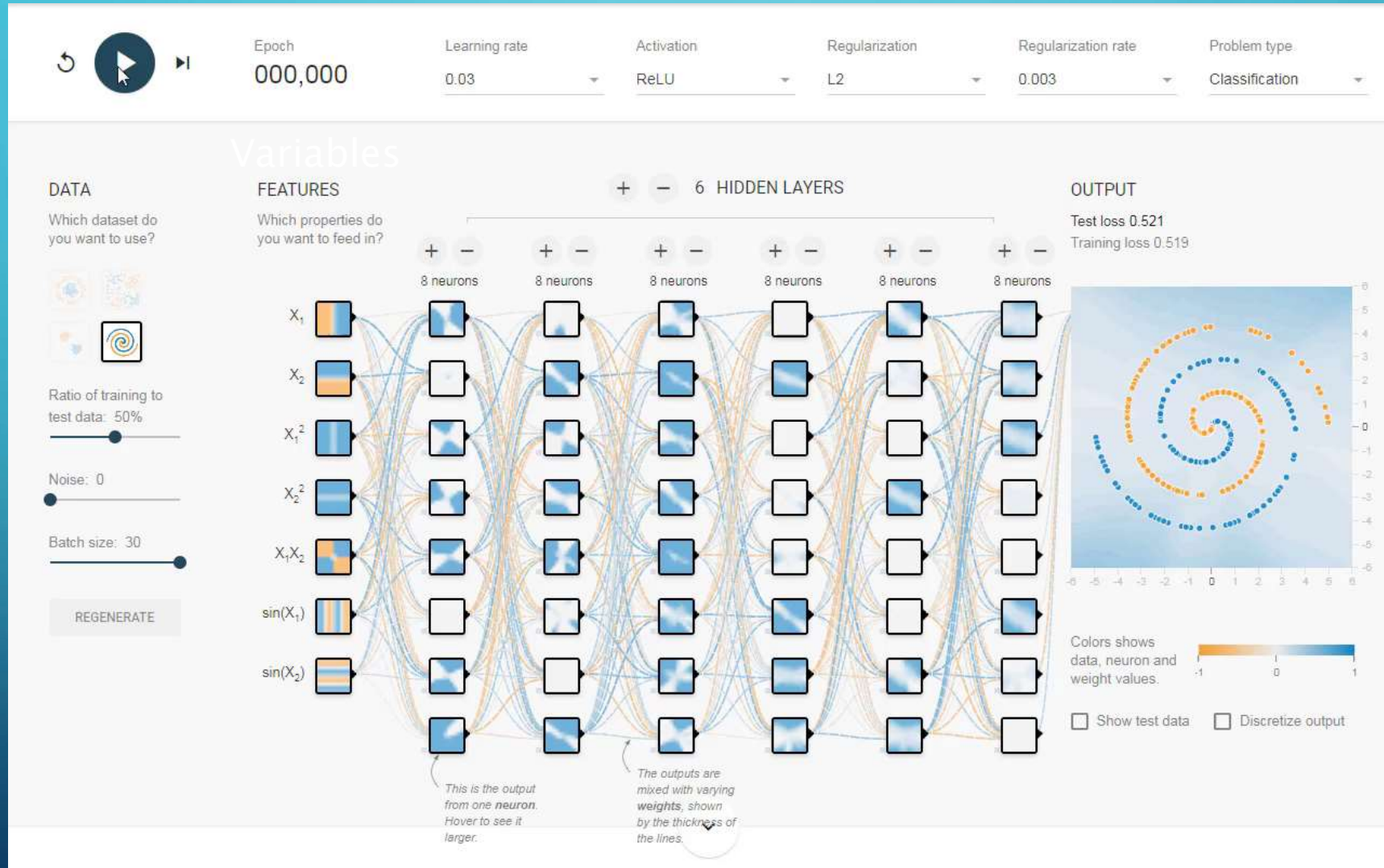


Image: <https://appliedgo.net/perceptron/>

DEEP-LEARNING – DÉMONSTRATION

<http://playground.tensorflow.org/>



INTELLIGENCE ARTIFICIELLE: AMBITIONS



*Keith J Dreyer, PhD Harvard
RSNA 2017*

INTELLIGENCE ARTIFICIELLE EN RADIOLOGIE

Trier les indications

Améliorer les techniques d'imagerie (reconstruction, échantillonnage de l'espace K...)
↑ rapidité d'acquisition IRM, ↓ irradiation scanner...

Optimiser la planification

Présentation « intelligente » d'images (pertinence, contenu...)

INDICATION

EXAMEN
(technique)

AFFICHAGE
(visualisation)

DETECTION

CLASSIFI-
CATION

COMPTE-
RENDU

DECISION

« CAD »

Algorithme de décision

Algorithme d'interprétation

Comptes-rendus standardisés
extraits des annotations
d'images



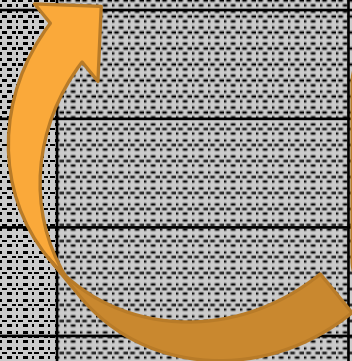
INTELLIGENCE ARTIFICIELLE: DIAGNOSTIC

SYSTEMS	FINDINGS	CONDITIONS
RESPIRATORY	260	2 968
HEAD AND NECK	258	2 628
NEUROSCIENCE	361	2 774
CARDIAC	93	1 179
PEDIATRICS	189	2 456
GASTROINTESTINAL	352	2 866
GENITOURINARY	234	1 905
REPRODUCTIVE	74	714
VASCULAR	159	1 800
MUSCULOSKELETAL	635	4 083
TOTAL	2 613	23 373
ARTIFICIAL INTELLIGENCE	DETECTION	INTERPRETATION

GAMUTS

INTELLIGENCE ARTIFICIELLE: DIAGNOSTIC

	COMPUTED TOMOGRAPH Y	MAGNETIC RESONANCE	POSITRON EMISSION	RADIOGRAPH Y	ANGIOGRAPH Y	ULTRASOUND	FLUOROSCOPI Y
ABDOMINAL IMAGING							
BREAST IMAGING							
CARDIAC IMAGING							
EMERGENCY IMAGING							
MUSCULOSKELETAL							
NEURORADIOLOGY							
NUCLEAR MEDICINE							
PEDIATRIC IMAGING							
THORACIC IMAGING							
INTERVENTIONAL							



Chaque point
correspond à un
diagnostic et donc à un
algorithme à développer
pour l'atteindre.

CONCLUSION

Imagerie et big data

- Intégration des données brutes dans les bases de données
 - Financement? Propriété?
- Qualité des données
 - Volume important généré par la clinique disponible mais peu exploitable
 - Outils d'annotation nécessaires, intégrés dans le workflow clinique
 - Structuration des comptes rendus
- Radiomique
 - Outil prometteur – Validation? Standardisation?
- Intelligence artificielle
 - Un outil à apprivoiser – Validation? Réglementation?