


Repositorios, retos y preparación de datos

Oliver Díaz Montesdeoca
Universidad de Barcelona

INTELIGENCIA ARTIFICIAL

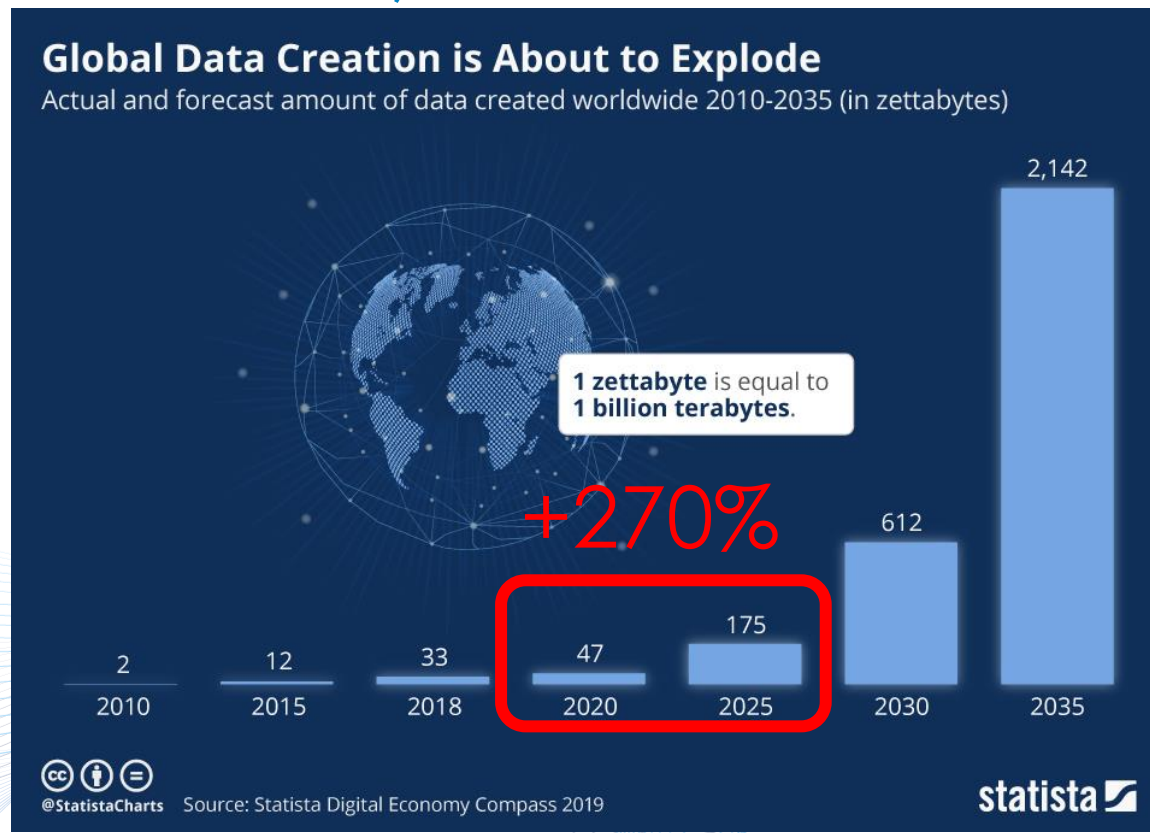
ÍNDICE

- La importancia de los datos hoy en día
 - Datos médicos: Datos de imagen y no de imagen
 - Escala de preparación de los datos de imagen médica
 - Preparación y procesamiento de datos
 - De-identificación de datos
 - Curación de datos
 - Almacenamiento de datos
 - Anotación de datos
 - Generación de datos sintéticos
- 

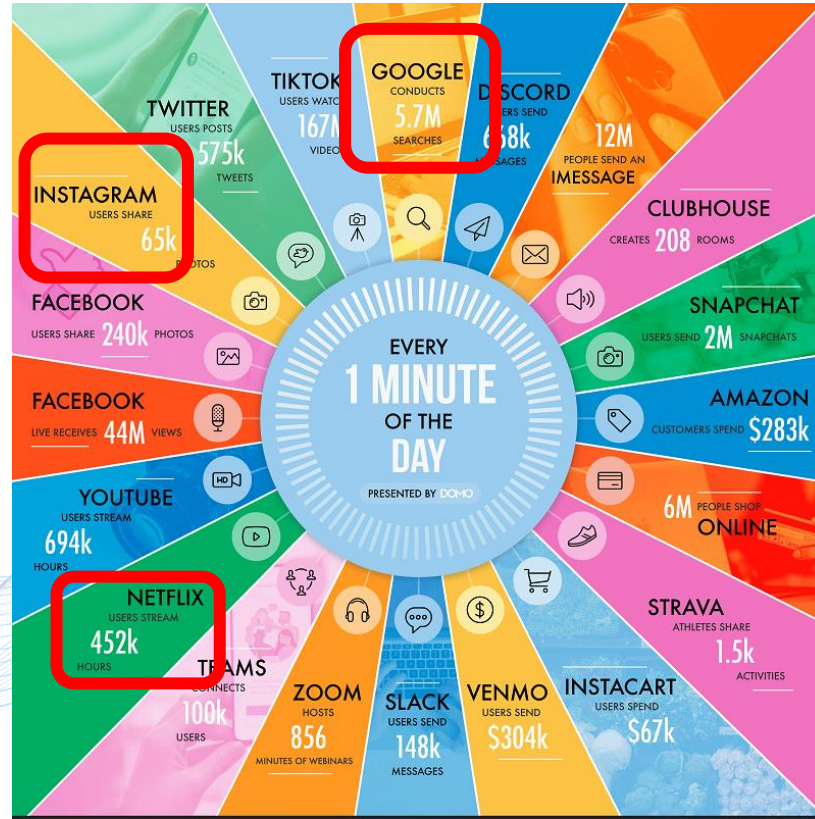
REVOLUCIÓN DIGITAL



DATOS, DATOS Y MÁS DATOS



LOS DATOS NUNCA MUEREN



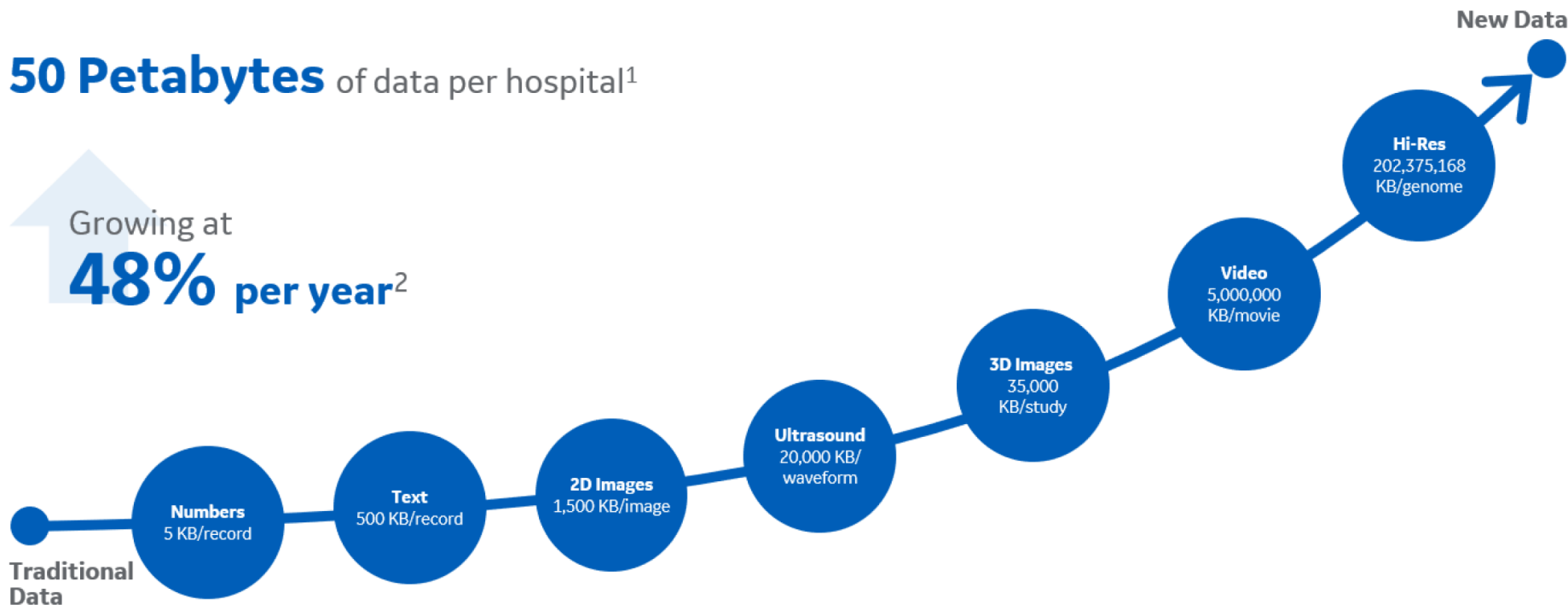
<https://www.domo.com/learn/infographic/data-never-sleeps-9>

DATOS MASIVOS TAMBIÉN EN SALUD

50 Petabytes of data per hospital¹

Growing at

48% per year²



^{1,2} Source: IDC & EMC Study – <https://www.cycloneinteractive.com/cyclone/assets/File/digital-universe-healthcare-vertical-report-ar.pdf>

¿QUÉ PODEMOS HACER CON ESTOS DATOS?

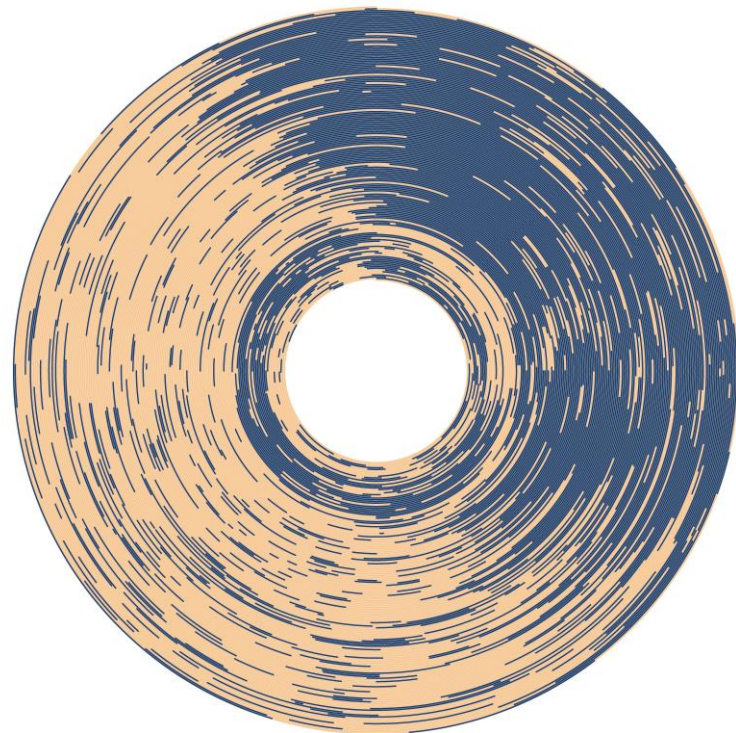
La digitalización de datos nos ha permitido la aplicación/uso/análisis de diferentes tecnologías en los últimos años para mejorar (de alguna manera) nuestras vidas:

- Almacenamiento/computación en la nube
- Big Data/Ciencia de datos
- Tecnologías de la información (TIC)
- Plataformas móviles / Apps
- **Inteligencia Artificial**
- Blockchain
- Realidad Aumentada/Virtual
- Internet de las cosas médicas (IoMT)
- ...

EJEMPLO DEL USO DE DATOS EN SALUD

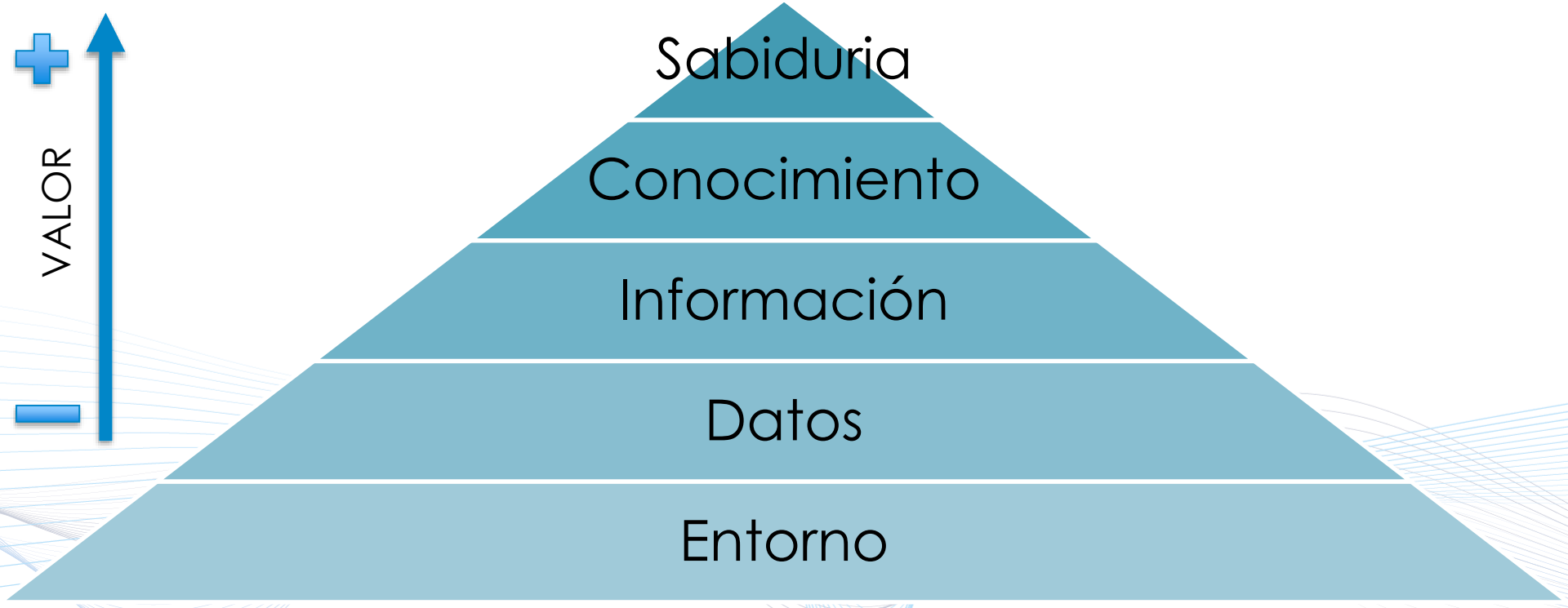
Patrón de sueño de un bebé durante los primeros 4 meses.

Espiral comenzando desde el interior. Cada revolución representa 24 horas. Medianoche en la parte superior.



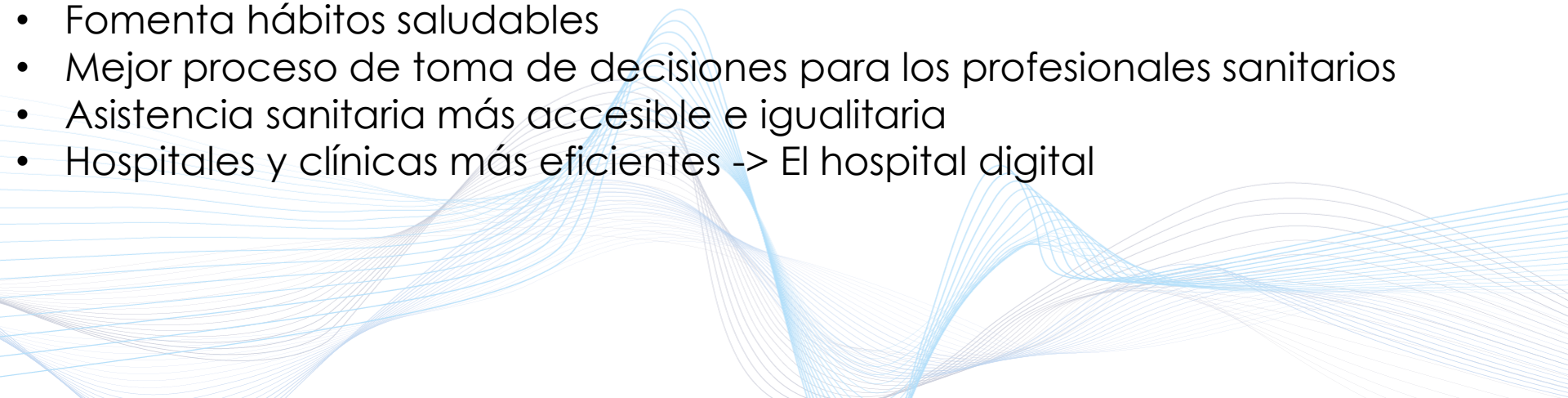
Fuente:
[reddit](#)

PIRÁMIDE DEL CONOCIMIENTO

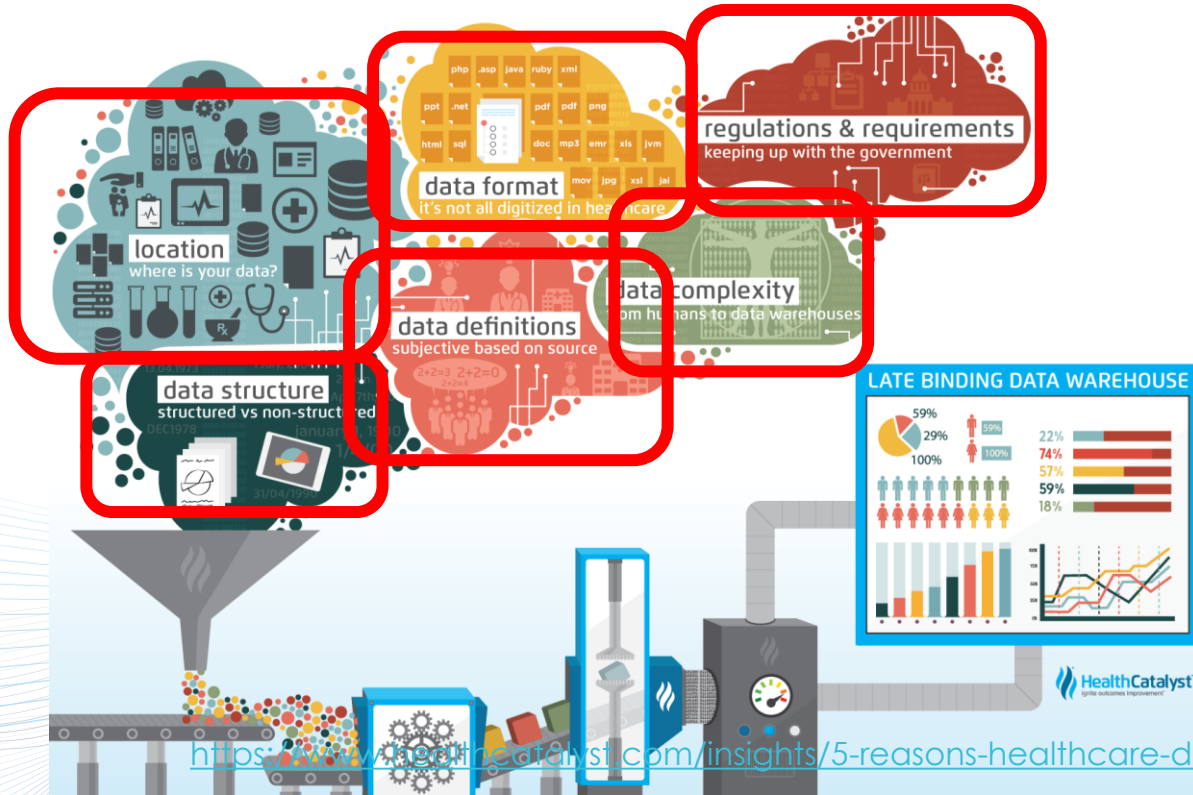


¿QUÉ NOS APORTAN LOS DATOS?

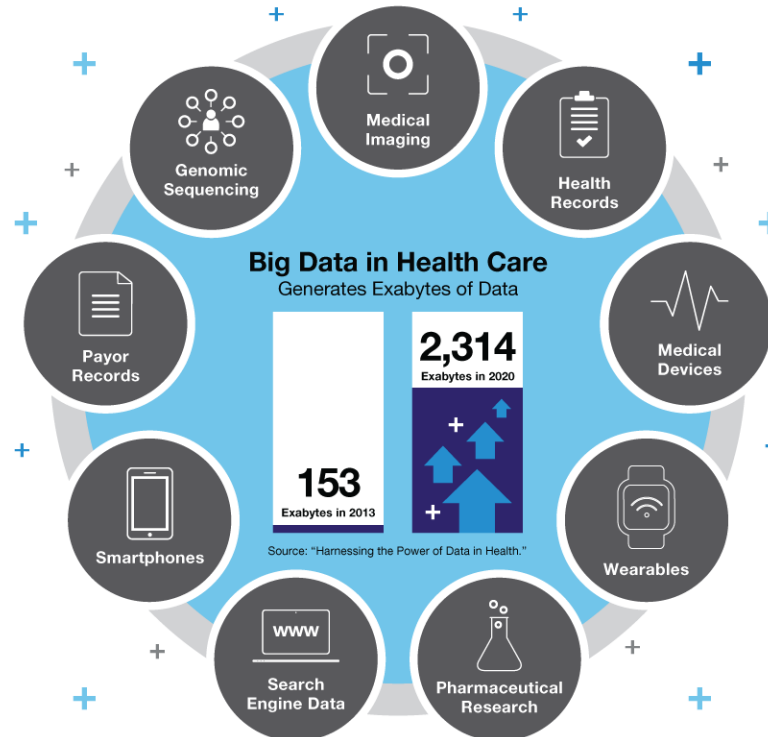
"El **conocimiento** (datos) es una gran fuente de información que puede salvar vidas"

- Mejora en la monitorización de pacientes
 - Pacientes mejor informados (empoderamiento)
 - Fomenta hábitos saludables
 - Mejor proceso de toma de decisiones para los profesionales sanitarios
 - Asistencia sanitaria más accesible e igualitaria
 - Hospitales y clínicas más eficientes -> El hospital digital
- 

RETOS (PENDIENTE) DE LOS DATOS EN SALUD



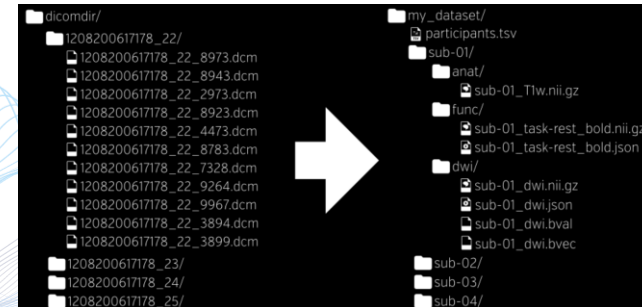
¿DE DÓNDE VIENEN LOS DATOS?



¿CÓMO SON LOS DATOS MÉDICOS?

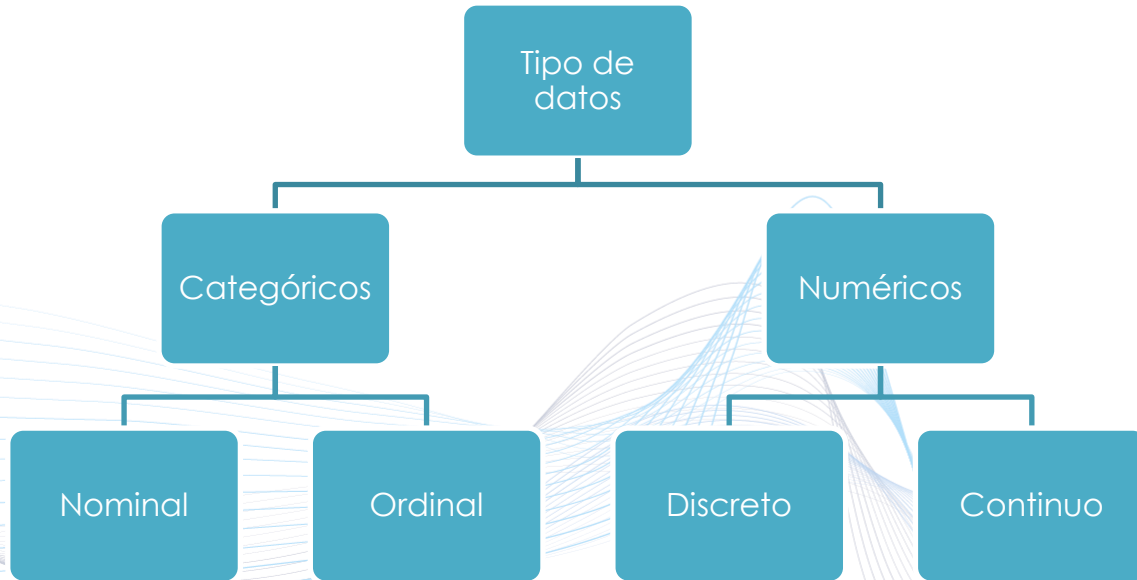
Imágenes

- DICOM
- NIFTI
- MINC
- ANALYZE
- BIDS

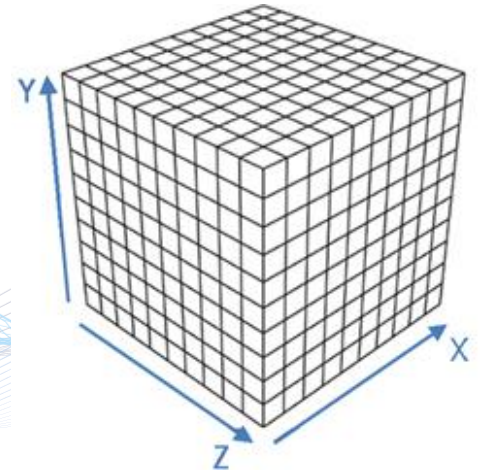


¿CÓMO SON LOS DATOS MÉDICOS?

No Imagen



- CSV
- JSON
- XML
- ...



¿CÓMO SON LOS DATOS MÉDICOS?

No Imagen

Dx

Diagnosis

Demographic: Age diagnosis, menopausal status, country, First or previous history of cancer,

Imaging: **Mammography/MRI pre-treatment** (T2 contrast enhanced): (multiple vendors)
(Date, site (Right, Left, Bilateral), BIRADS, Location)

Pathology (Date, ID, molecular subtypes* (see biological markers) size, nodal involvement, cancer staging, metastatic status, info previous aesthetic breast surgery (augmentation or reduction).

Staging (Stage, TNM, Sentinel LN)

Biological markers (ER, PgR, HER2, Ki67, TN subtype*) (BRCA1/2, related genes)

NAC

Treatment

NAC treatment (Scheme and duration: Date initiation/ending

Type Surgery:
Tumorectomy/Mastectomy \pm
Linfadenectomy/Sentinel Node)

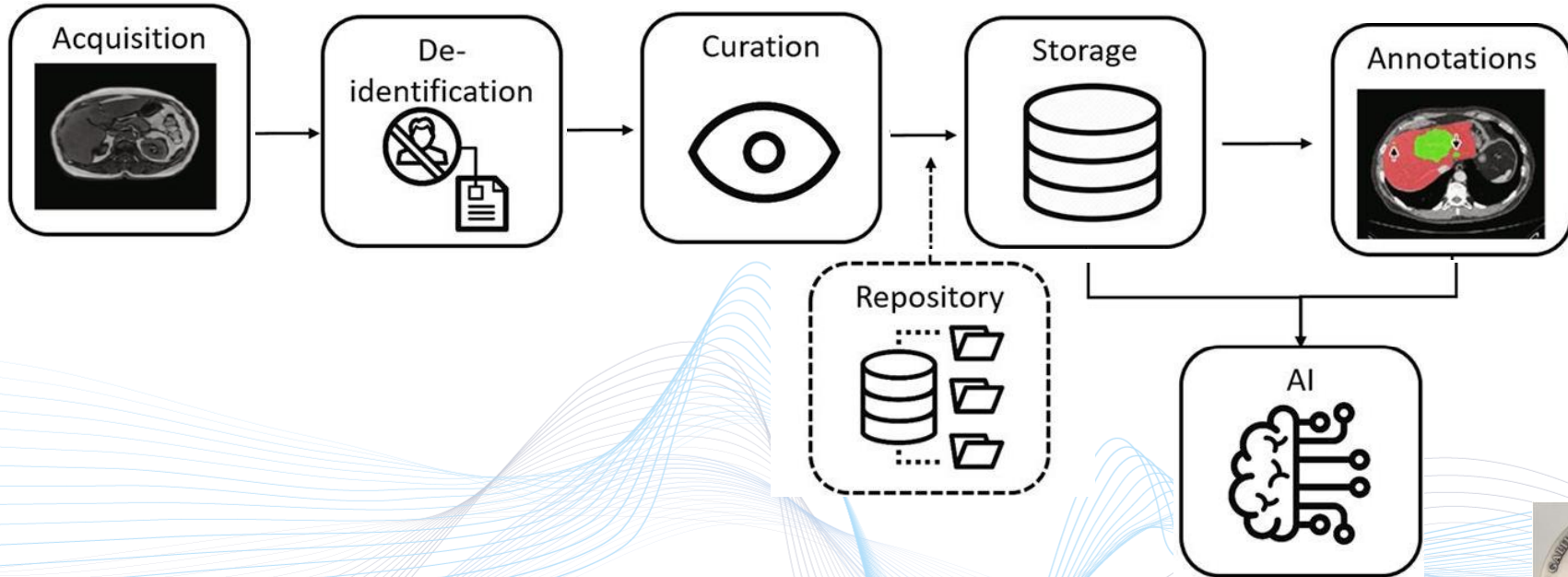
FUP

Follow-up

Pathology Report post-treatment
(response to NAC (pathological Complete Response pCR vs no response).
yPTNM after treatment, size lesion

Mammography FUP: date, result

CICLO DE VIDA HERRAMIENTAS IA (antes del desarrollo)

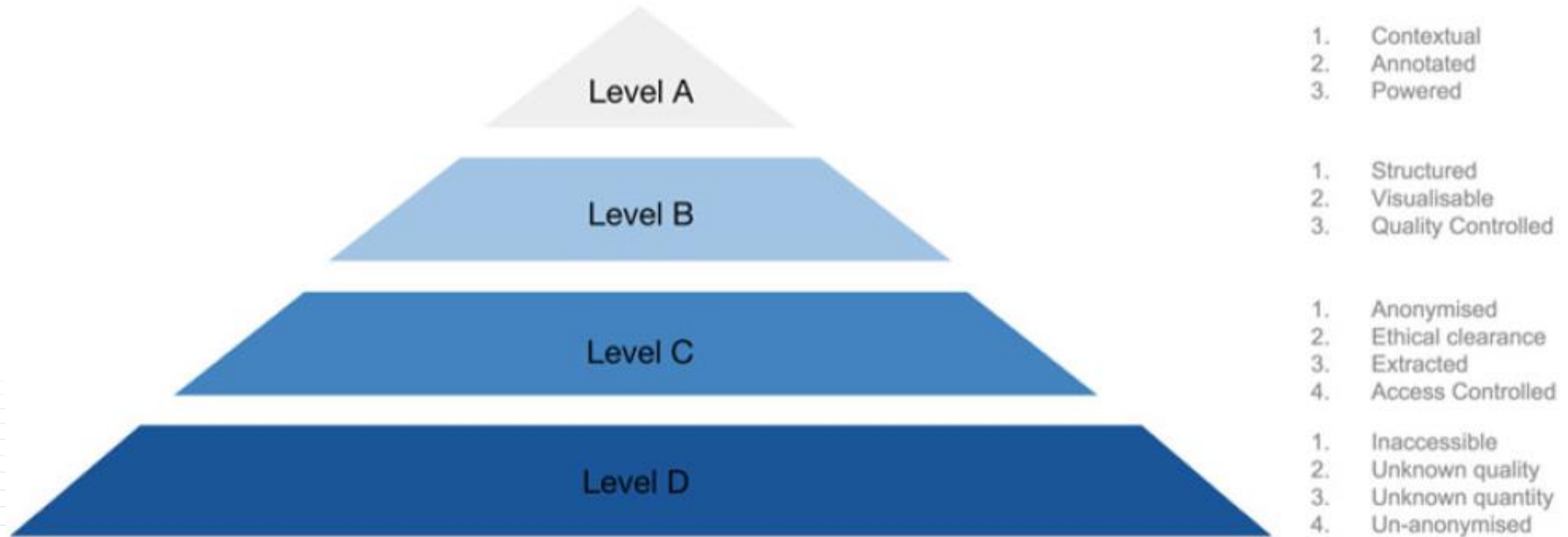


Díaz, O et al (2021). Data preparation for artificial intelligence in medical imaging: a comprehensive guide to open-access platforms and tools. *Physica Medica*, 83, 25-37

<https://www.sciencedirect.com/science/article/pii/S1120179721000958>



Medical Imaging data readiness (MIDaR) scale



Harvey, H., & Glocker, B. (2019). A standardised approach for preparing imaging data for machine learning tasks in radiology. In Artificial intelligence in medical imaging (pp. 61-72). Springer, Cham.

DE-IDENTIFICACIÓN



- Preservación de la privacidad individual
- Información sensible que puede identificar al paciente.
 - Nombre, dirección, identificación del paciente,...
- ISO 25237



El Reglamento General de Protección de Datos (GDPR) (Reglamento 2016/679) es una normativa con la que el Parlamento Europeo, el Consejo de la Unión Europea y la Comisión Europea pretenden reforzar y unificar la **protección de datos** de todas las personas dentro de la Unión Europea (UE). También se ocupa de la **exportación de datos personales fuera de la UE**.

DE-IDENTIFICACIÓN

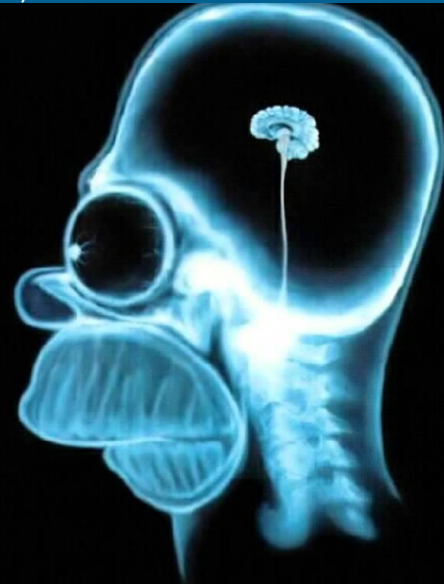
De-
identification



Herramientas:

- DicomClearnear
- Posda tools
- DicomAnonymazer
- PrivacyGuard
- ...

```
(0010,0010): Anonymised00034  
(0010,0020): XXX_72947  
(0010,0030): 19000101  
(0010,0040): M  
(0010,1010): 039Y  
(0018,0015): HEAD
```



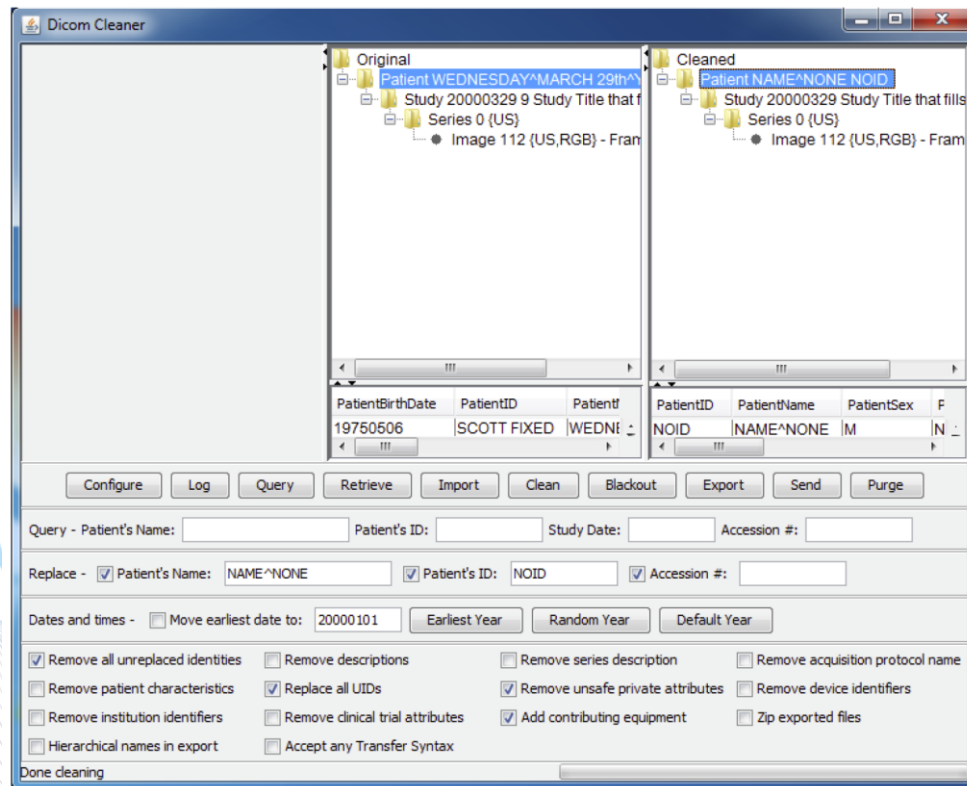
DE-IDENTIFICACIÓN

De-
identification



Herramientas:

- **DicomCleaner**
- Posda tools
- DicomAnonymizer
- PrivacyGuard
- ...



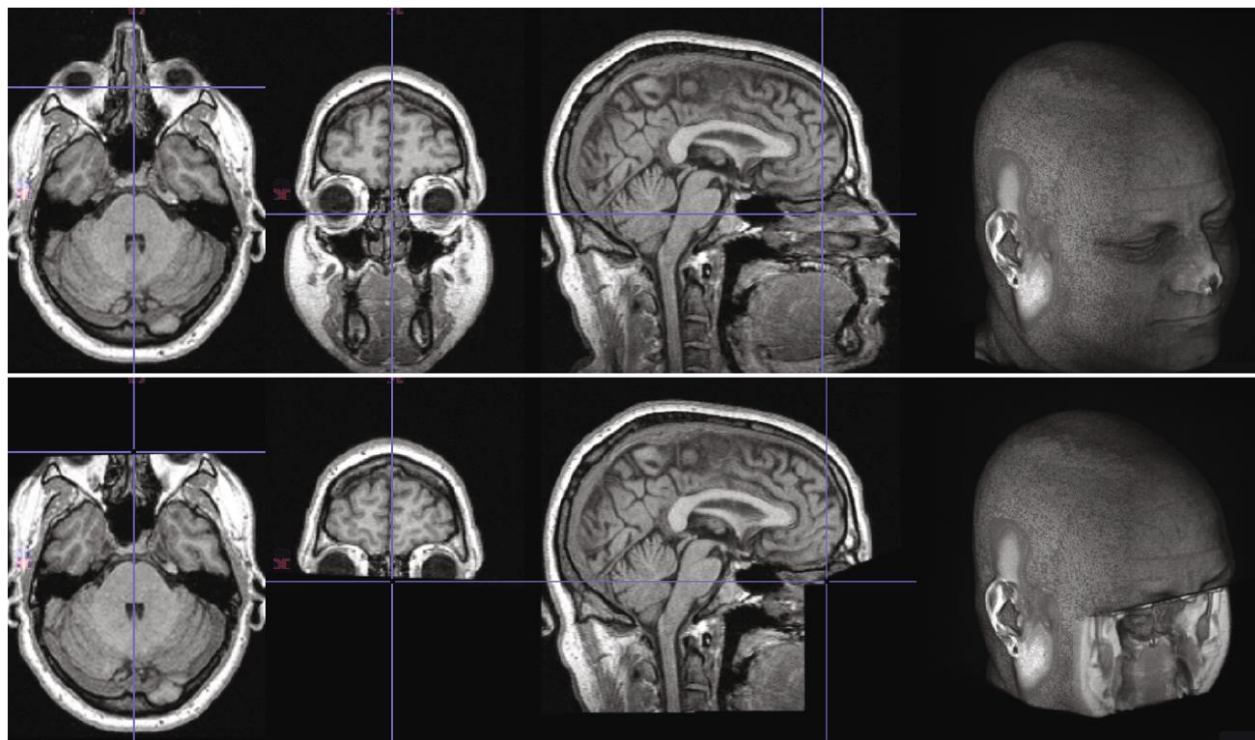
DE-IDENTIFICACIÓN

De-
identification

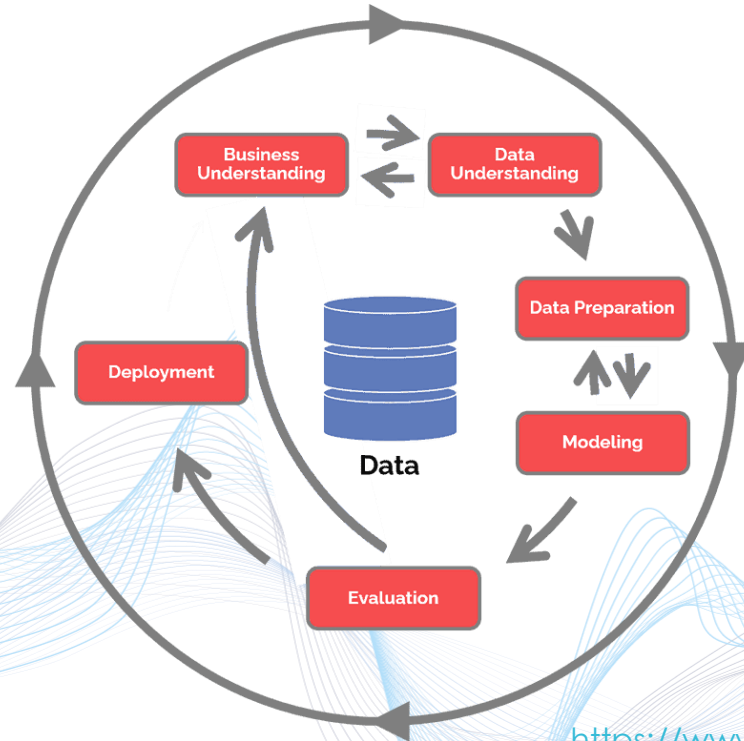


Deface:

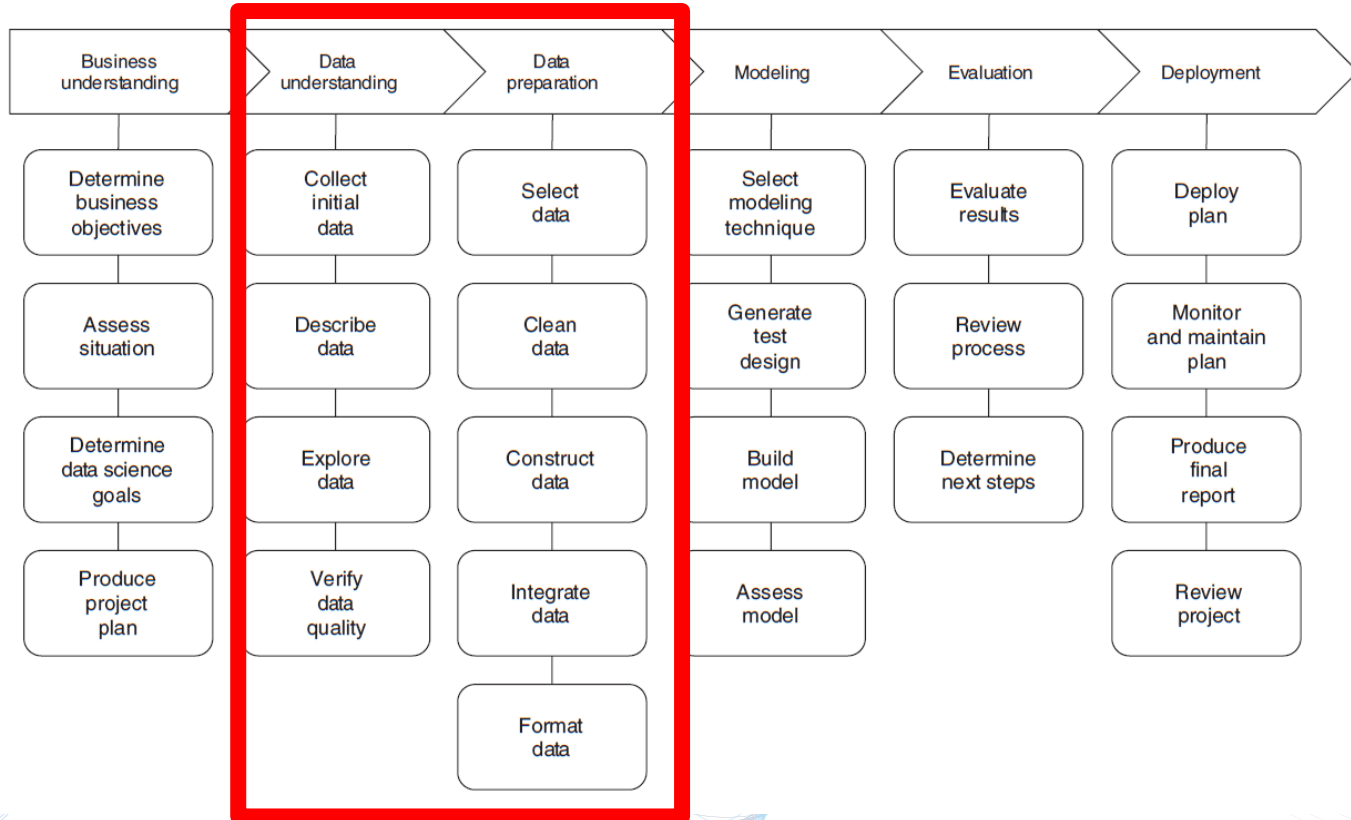
- pydeface
- mrideface
- ...



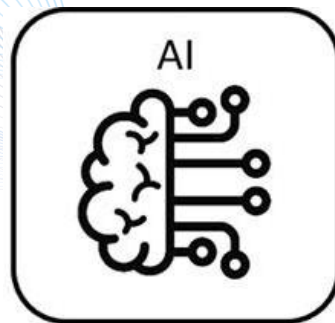
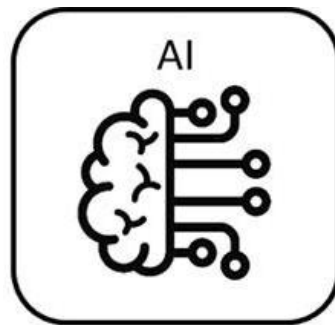
PREPARACIÓN DE LOS DATOS



PREPARACIÓN DE LOS DATOS



PREPARACIÓN DE LOS DATOS



CURACIÓN DE LOS DATOS

Curation



Definición: Conjunto de procedimientos y acciones que se refieren a la gestión, creación, modificación y verificación de la calidad, integridad, validación, trazabilidad y reproducibilidad de los datos.

Problemas:

- Duplicación de datos
- Inconsistencia en las cabeceras DICOM en pacientes, estudios, series...
- Normalización de los datos
- ...

Las herramientas de curación de datos permiten investigar, detectar, prevenir y resolver problemas en los repositorios de datos (Diaz et al Physica Medica, 83, 25-37, 2021)

CURACIÓN DE LOS DATOS

Curation



Herramientas:

- Postda Tools
- Dicom3tools
- DCMTK
- **Dcm4ache**
- ...

dcm4ache Folder Trash AE Management Offline Storage Worklist Console MPPS Console User Admin Audit Repository Logout

Select dominant Patient :

Patient ID	Issuer	Patient Name	Patient Sex	Birth Date
210406		Mustermann^Max	M	1983/07/19
200406		Test^Moritz	M	1973/07/05

Merge Cancel

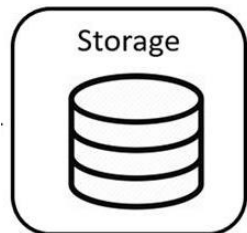
dcm4ache Folder Trash AE Management Offline Storage Worklist Console MPPS Console User Admin Audit Repository Logout

Displaying studies 1 to 20 of 20 matching studies.

Patient Name:	Patient ID:	Study ID:	Study Date:	Accession No.:	Modality:
Mustermann^Max	210406	1983/07/19	M		
Test^Moritz	200406	1973/07/05	M		

Study ID (@Media):	Series Description / Body Part / IUID :	Acc.No.:	Ref. Physician:	Status:	NoS:	NoI:
2002/11/17 06:11:59	2K-RESOL	OT	Resolution 2K	20022002	AAPM	6 14
2002/11/17 12:13:35	1K-RESOL	OT	Resolution 1K	20022002	AAPM	6 14
2002/11/17 12:34:55	1K-NOISE	OT	Noise 1K	20022002	AAPM	2 4
2002/11/17 16:23:33	1K-LUMIN	OT	Luminance 1K	20022002	AAPM	6 25
2002/11/17 17:12:05	1K-GLARE	OT	Glare 1K	20022002	AAPM	3 13
2002/11/17 18:55:52	2K-MULTI	OT	Multi Purpose 2K	20022002	AAPM	3 3
2002/11/17 19:06:19	1K-MULTI	OT	Multi Purpose 1K	20022002	AAPM	3 3
2002/11/17 19:26:13	2K-ANATM	OT	Anatomical 2K	20022002	AAPM	3 4
Buc^Jérôme	SCSFREN	OT				1 1
GSD^Test^Pattern	GSD^PATTERN	OT	GSD^Test Pattern		Riesmeier^Joerg	1 17
2000/06/09 12	GSD^PATTERN	OT				
Mustermann^Max	210406	1983/07/19	M			
2000/02/17 11:28:16	1	CT	Abdomen AbdRoutine			1 228
2004/05/05 12:07:55	2500000000455053	CT	Head^@EGN Ptg.	2500000000455053	EGN	1 172
2004/05/05 15:51:55	419171576	MR	MR des Gehirns	2500000000456090	BAL	2 185
Test^Moritz	200406	1973/07/05	M			
2001/04/08 11:27:32	18871	CT	CTBK Becken	1884504	UN6N	2 403

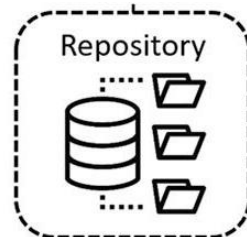
ALMACENAMIENTO DE DATOS



PACS, RIS, HIS,...

Almacenamiento:

- XNAT
- Kheops
- Orthanc-Server
- ...



Repositories:

<https://www.cancerimagingarchive.net/>
<https://grand-challenge.org/>
<https://www.kaggle.com/>

...



THE  **CANCER**
IMAGING ARCHIVE

ANOTACIÓN DE LOS DATOS

Annotations



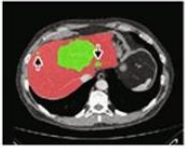
Herramientas:

- ITK-SNAP
- 3D slicer
- ImageJ
- ...

Plataformas
colaborativas

ANOTACIÓN DE LOS DATOS

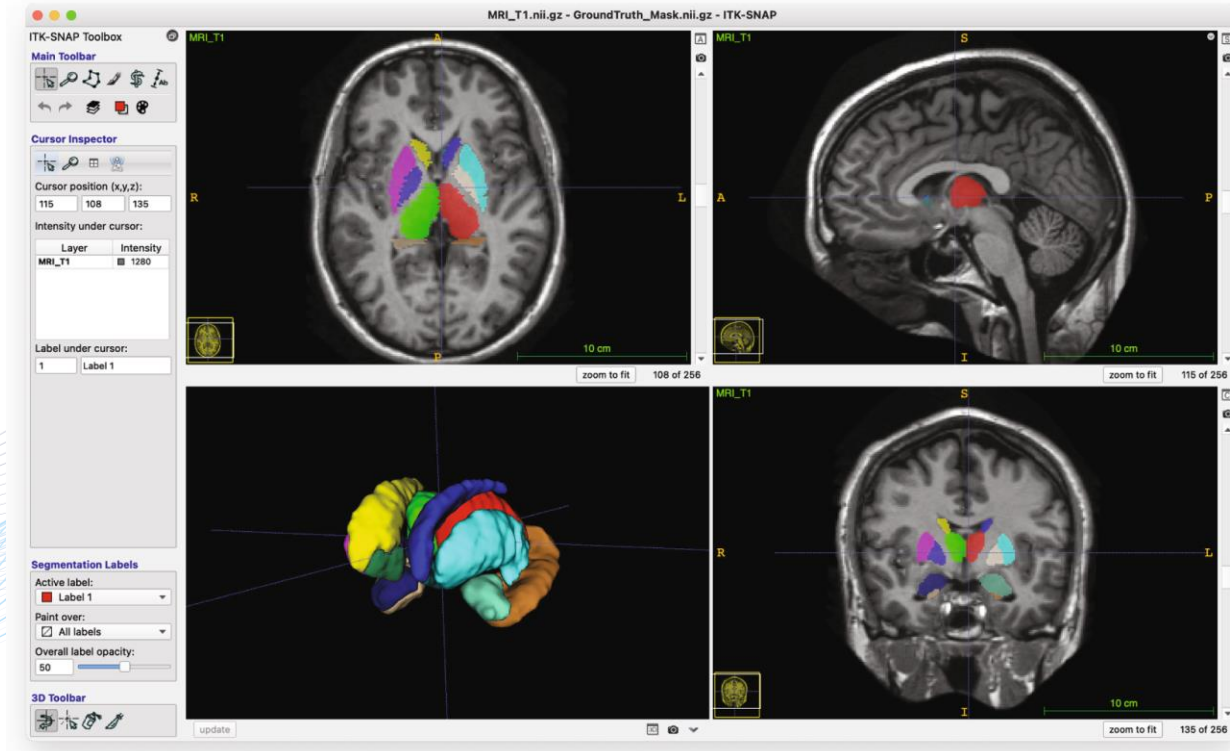
Annotations



Herramientas:

- **ITK-SNAP**
- 3D slicer
- ImageJ
- ...

Plataformas
colaborativas



ANOTACIÓN DE LOS DATOS

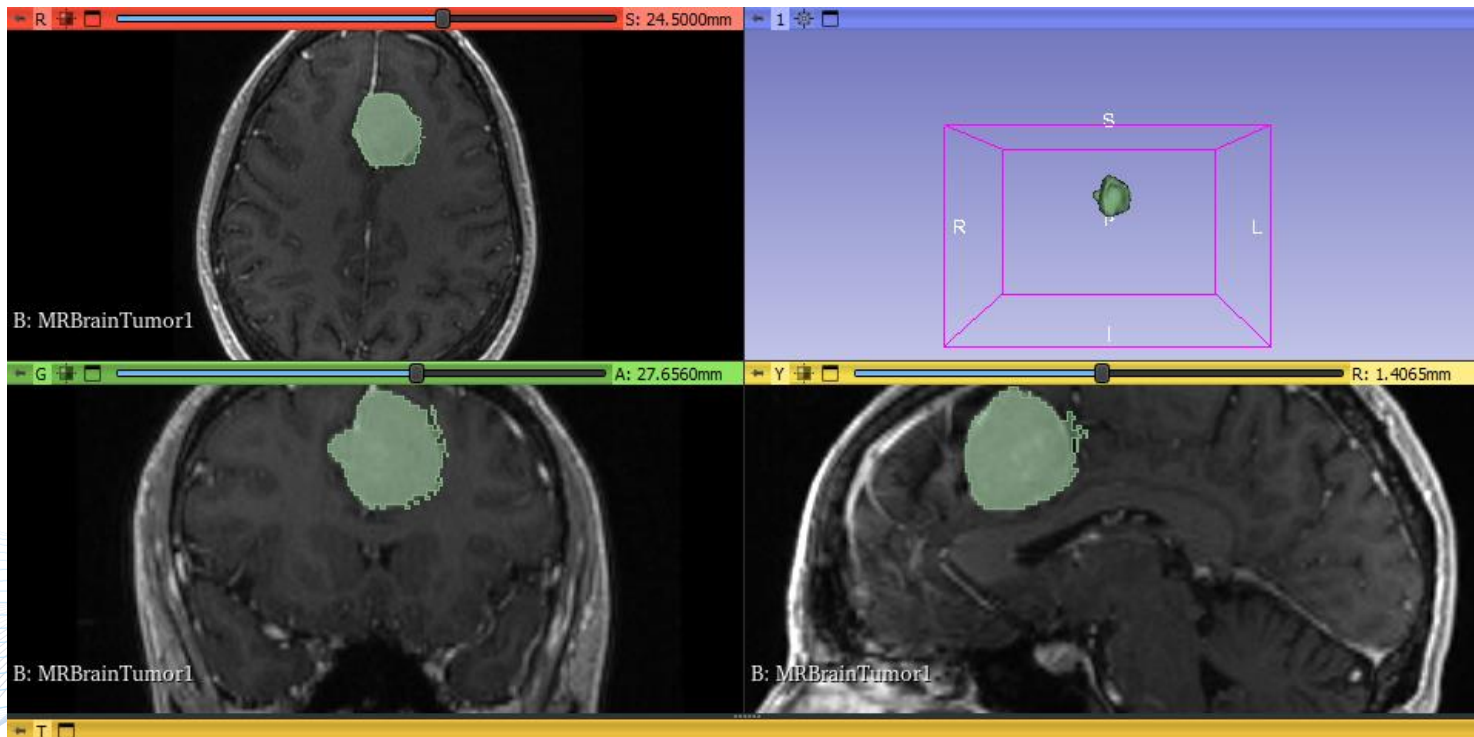
Annotations



Herramientas:

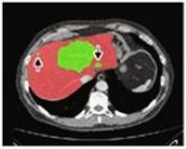
- ITK-SNAP
- **3D slicer**
- ImageJ
- ...

Plataformas
colaborativas



ANOTACIÓN DE LOS DATOS

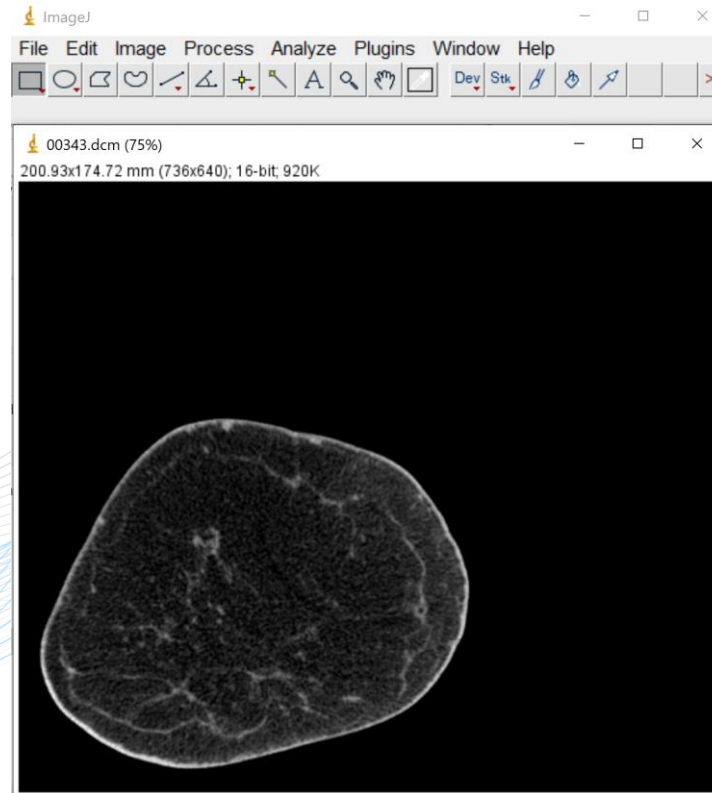
Annotations



Herramientas:

- ITK-SNAP
- 3D slicer
- **ImageJ**
- ...

Plataformas
colaborativas



ANOTACIÓN DE LOS DATOS

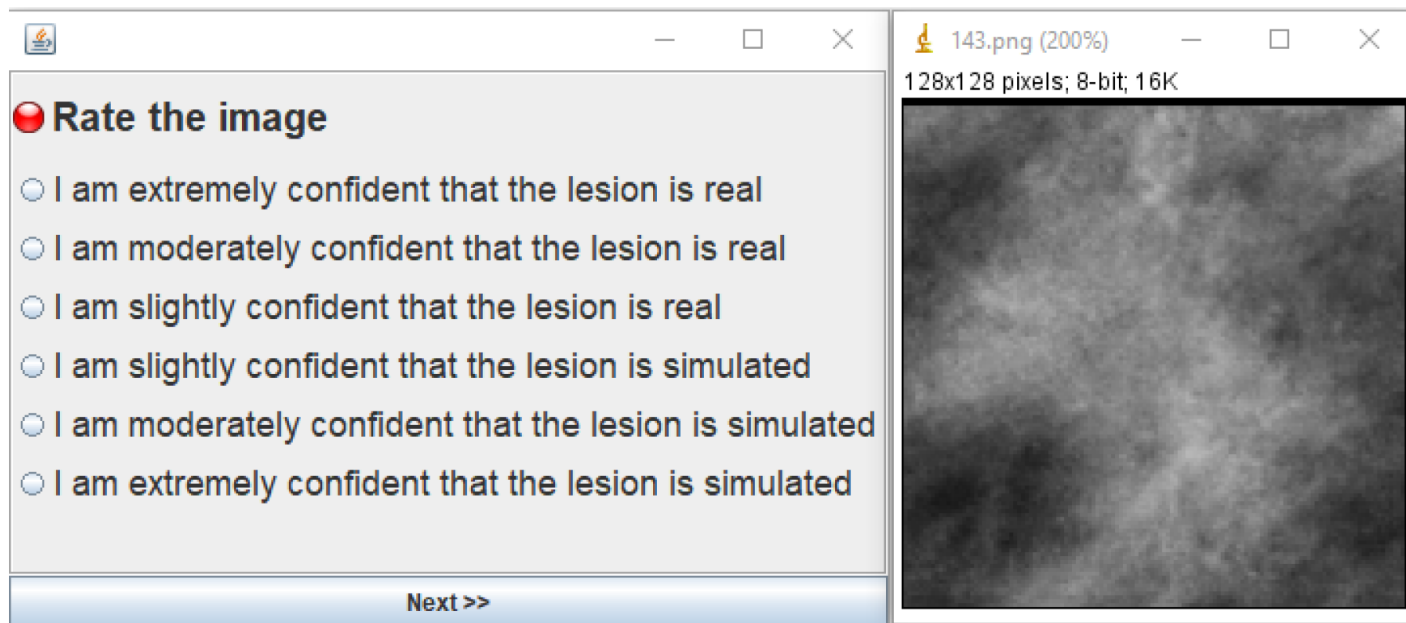
Annotations



Herramientas:

- ITK-SNAP
- 3D slicer
- **ImageJ**
- ...

Plataformas
colaborativas



Rate the image

- ☐ I am extremely confident that the lesion is real
- ☐ I am moderately confident that the lesion is real
- ☐ I am slightly confident that the lesion is real
- ☐ I am slightly confident that the lesion is simulated
- ☐ I am moderately confident that the lesion is simulated
- ☐ I am extremely confident that the lesion is simulated

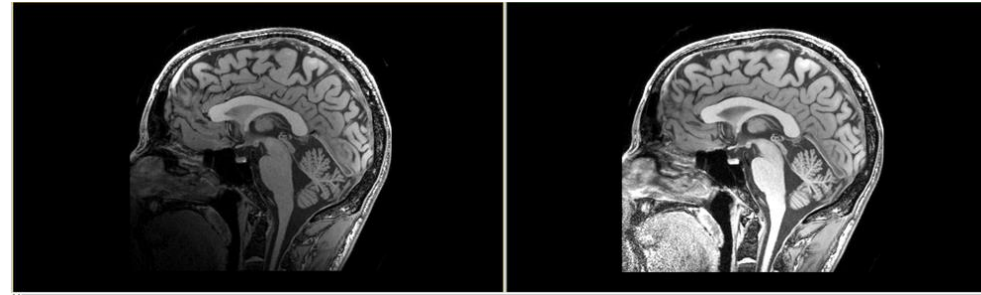
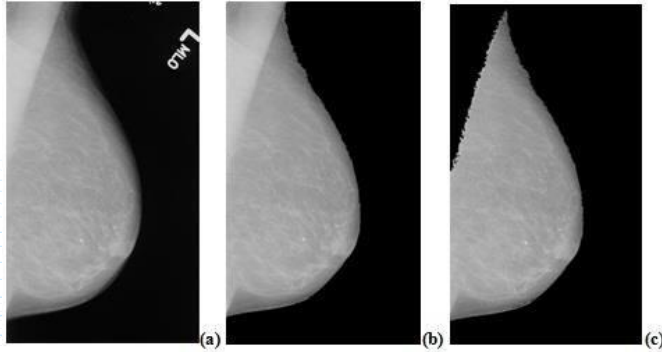
Next >>

143.png (200%)
128x128 pixels; 8-bit; 16K

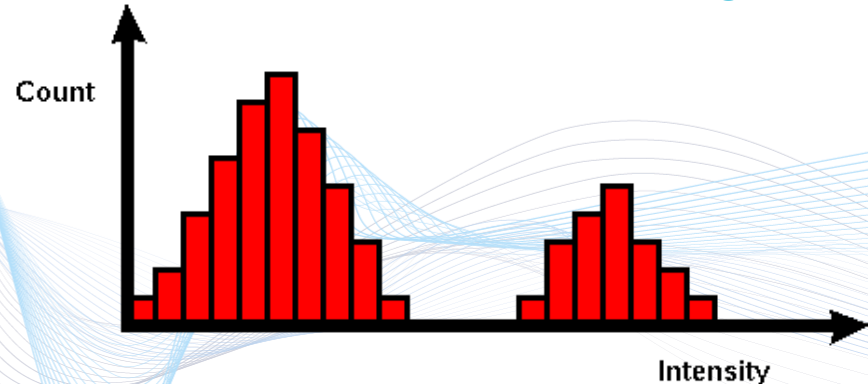
Alyafi, B., Diaz, O., et al. (2020). Quality analysis of DCGAN-generated mammography lesions. In *15th International workshop on breast imaging (IWBI2020)* (Vol. 11513, pp. 80-85). SPIE.

PREPROCESADO DE IMAGEN

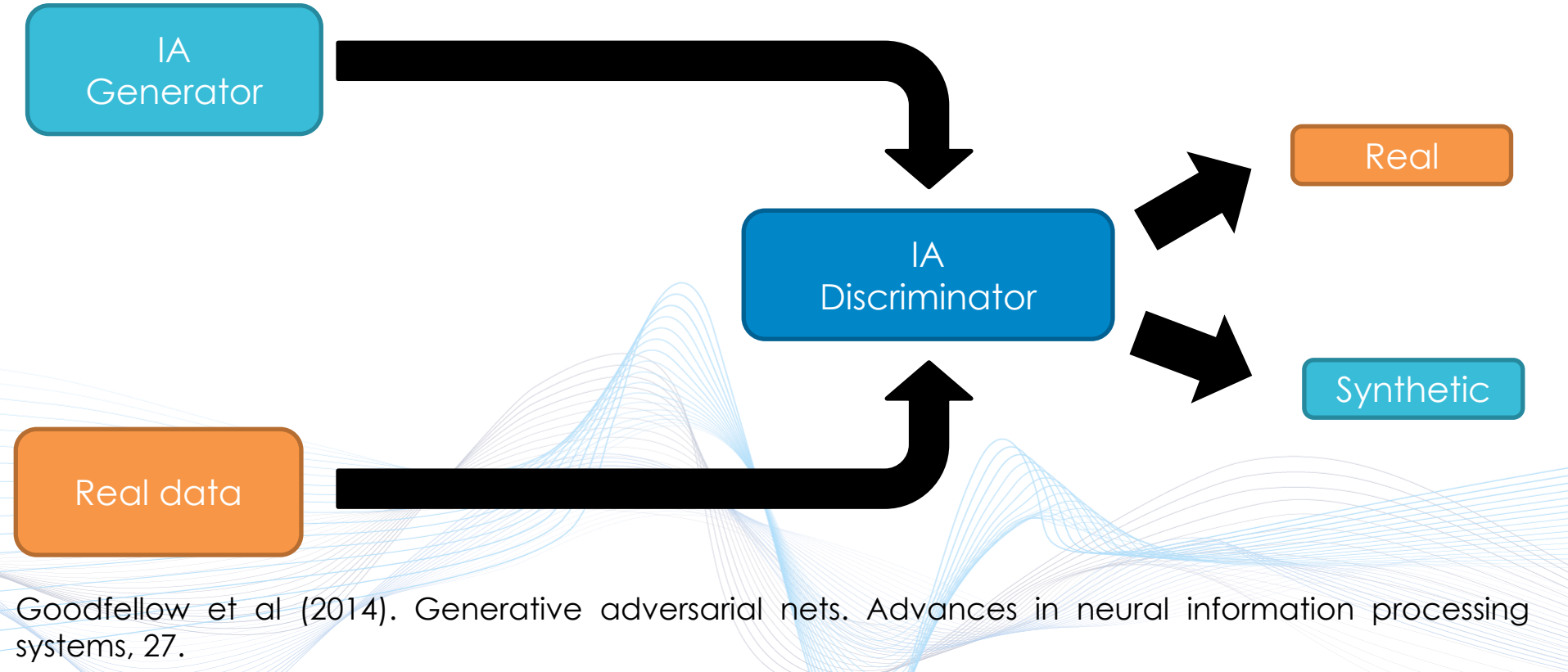
- Corrección de inhomogeneidades de intensidad (bias field correction)
- Normalización de histogramas
- Reducción de ruido
- Reducción del tamaño
- Recortar regiones
- ...



<https://www.slicer.org/>

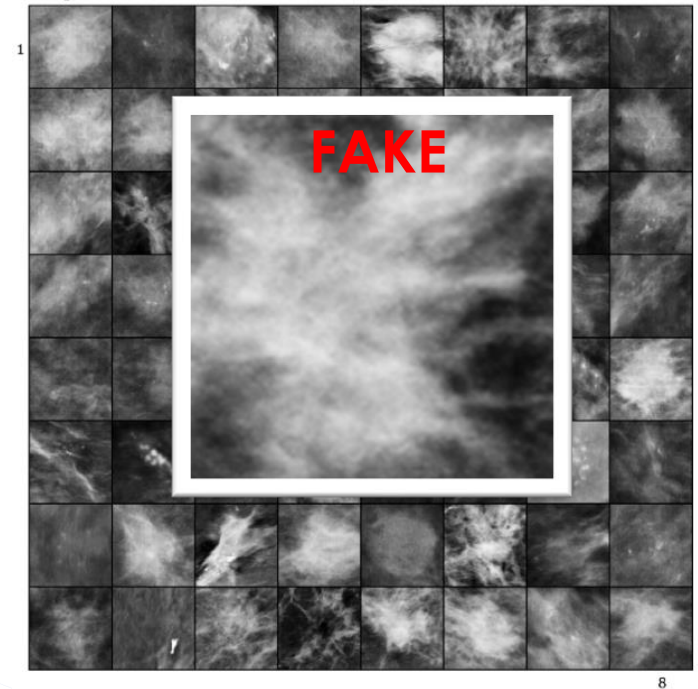
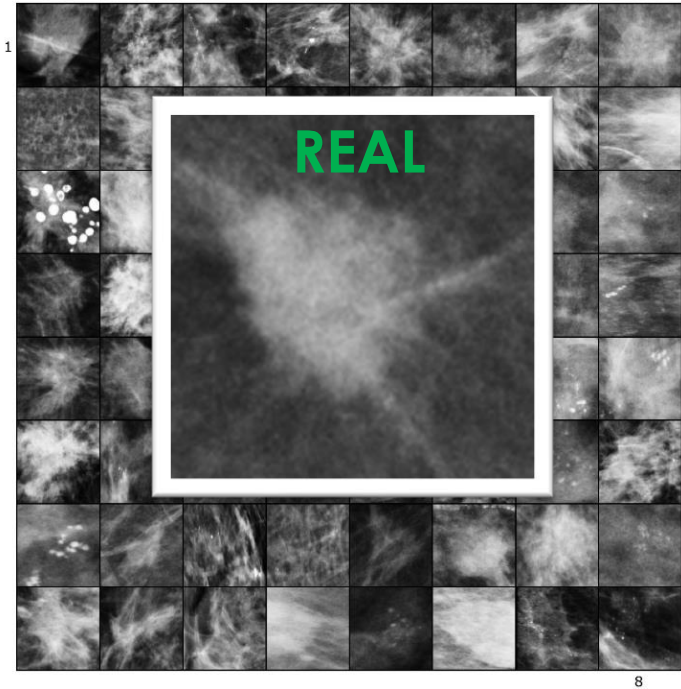


GENERACIÓN DE DATOS SINTÉTICOS



Goodfellow et al (2014). Generative adversarial nets. Advances in neural information processing systems, 27.

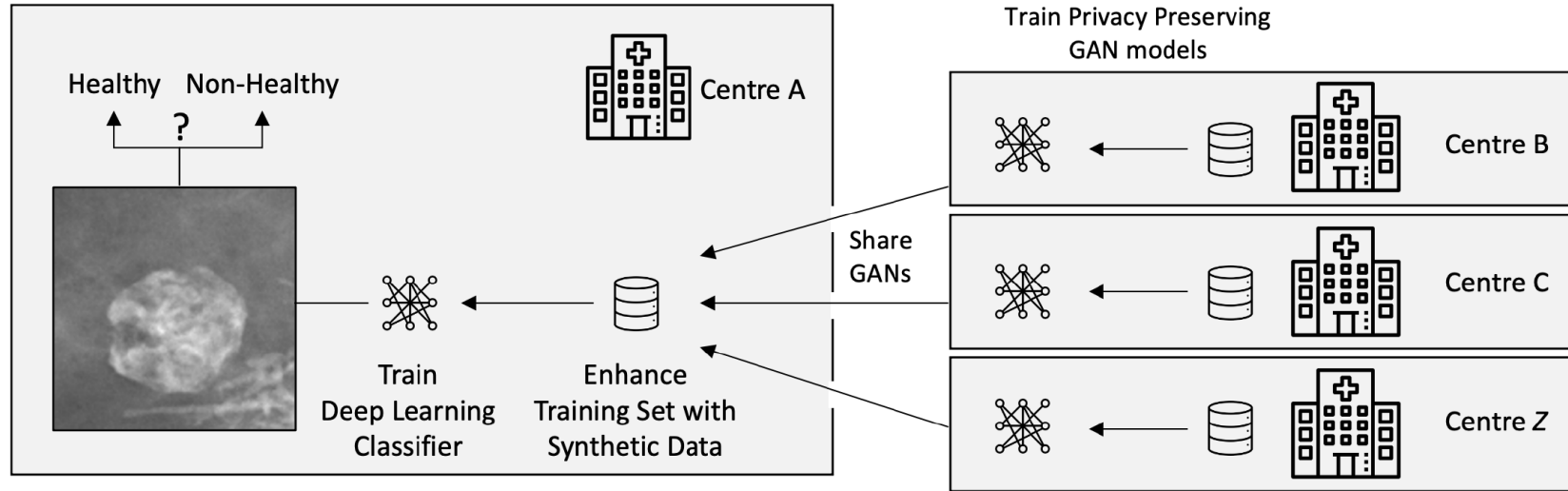
GENERACIÓN DE DATOS SINTÉTICOS



Alyafi et al (2020). Quality analysis of DCGAN-generated mammography lesions. *IWBI2020*, 11513, 115130B

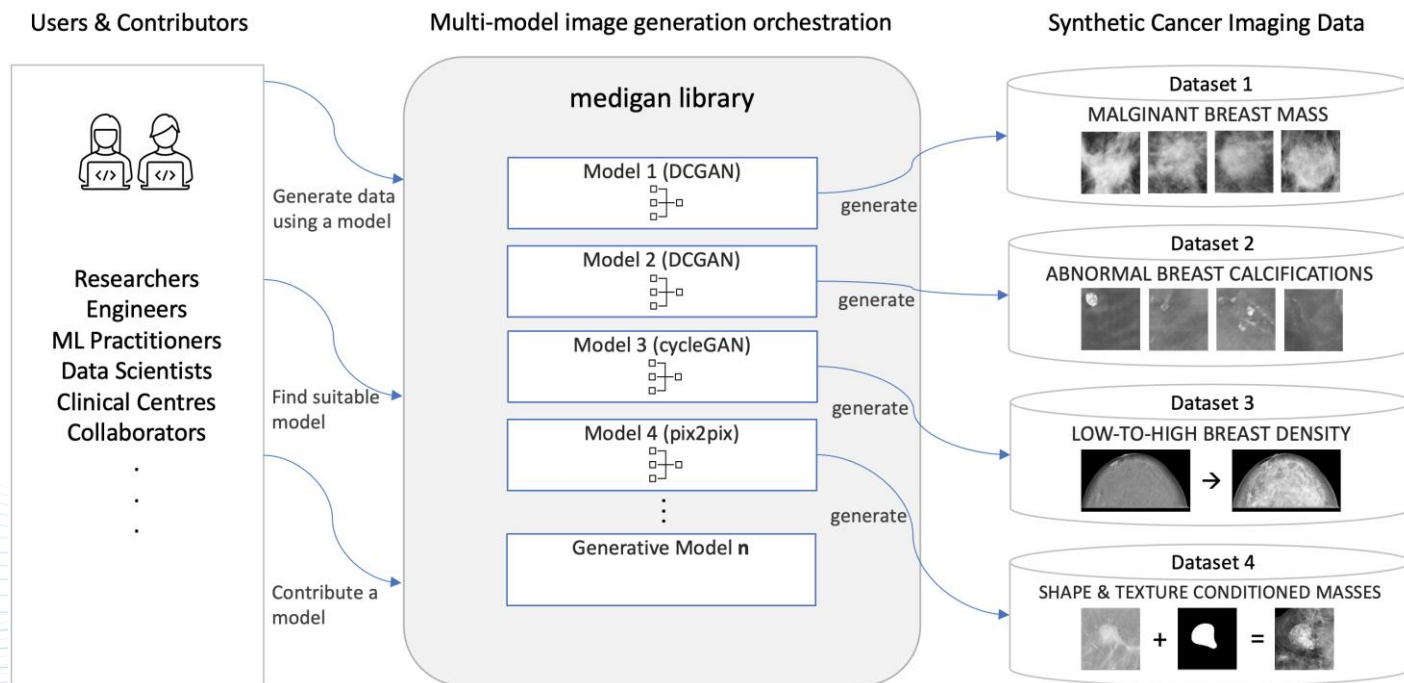
Alyafi et al (2020). DCGANs for realistic breast mass augmentation in x-ray mammography. In *Medical Imaging 2020: CAD*, 11314, 1131420

GENERACIÓN DE DATOS SINTÉTICOS



Szafranowska, Z., Osuala, R., Breier, B., Kushibar, K., Lekadir, K., & Diaz, O. (2022). Sharing Generative Models Instead of Private Data: A Simulation Study on Mammography Patch Classification. arXiv preprint arXiv:2203.04961.

GENERACIÓN DE DATOS SINTÉTICOS

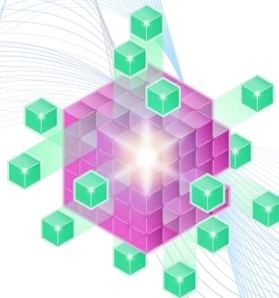


RETOS ACTUALES (Y POSIBLES SOLUCIONES)

- **Falta de datos** (datos sintéticos)
- **Privacidad de los datos** (anonimización, aprendizaje federado, datos sintéticos)
- **Análisis de datos multicéntricos** (homogeneización de datos)
- **Análisis de datos de múltiples fuentes** (!científico de datos!)
 - Multiomics o panomics
- BONUS: **Analfabetismo digital** (formación en competencias digitales)

CONCLUSIONES

- Estamos en la era de la **Transformación Digital**
- Los **datos** necesitan ser **procesados** para mejorar su **rendimiento**
- La **Ciencia de Datos** permitirá un mejor seguimiento, predicción de enfermedades y tratamiento
- La **preparación/procesado de datos** es esencial para el desarrollo de algoritmos de IA
- **Herramientas de acceso abierto** (open source) representan un buen punto de partida
- Enriquece tus datos



Repositorios, retos y preparación de datos

Oliver Díaz Montesdeoca
Universidad de Barcelona

INTELIGENCIA ARTIFICIAL