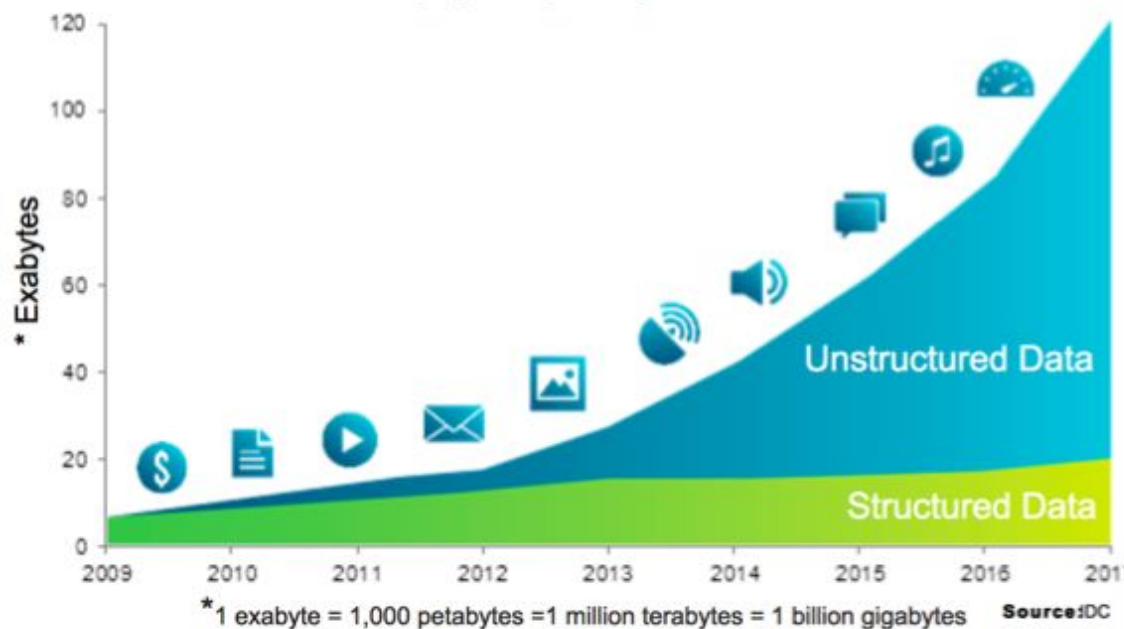




Introducción al PLN Clínico

Jocelyn Dunstan, PhD MSc
jdunstan@uc.cl

Problem - Traditional and Legacy Storage Designed for Transactional, Not Unstructured



La Corporación Internacional de Datos

proyecta que la cantidad de datos digitales generados anualmente en el mundo crecerá de 33 zettabytes en 2018 a 175 zettabytes el 2025, en donde un zettabyte es equivalente a 10^{21} bytes o un millón de millones de gigabytes.

<https://www.forbes.com/sites/tomcoughlin/2018/11/27/175-zettabytes-by-2025/#430a23055459>

Datos

No Estructurados



Texto



Audio



Imágenes



Video

Estructurados

Categoricos

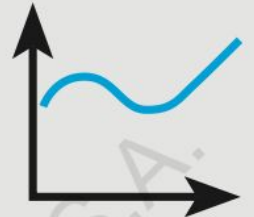
A

B

C

D

Numéricos

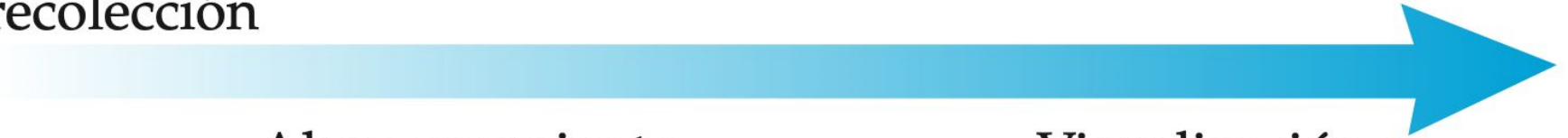


Generación y
recolección

Análisis

Almacenamiento
y gestión

Visualización e
interpretación



Inteligencia Artificial



Lenguaje

IA

Visión

Robótica

INTELIGENCIA ARTIFICIAL

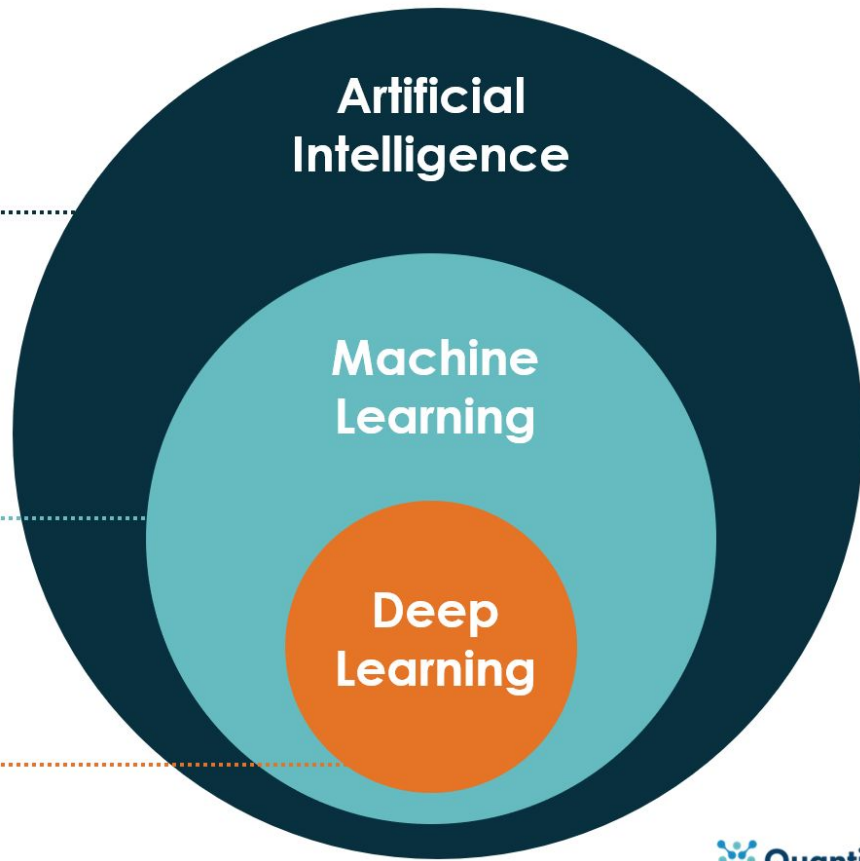
Programas computacionales que hacen tareas que usualmente requerirían inteligencia humana

APRENDIZAJE DE MÁQUINAS

Entrenar algoritmos para resolver tareas de reconocimiento de patrones en vez de programar reglas

APRENDIZAJE PROFUNDO

Entrenar algoritmos que usan redes neuronales profundas



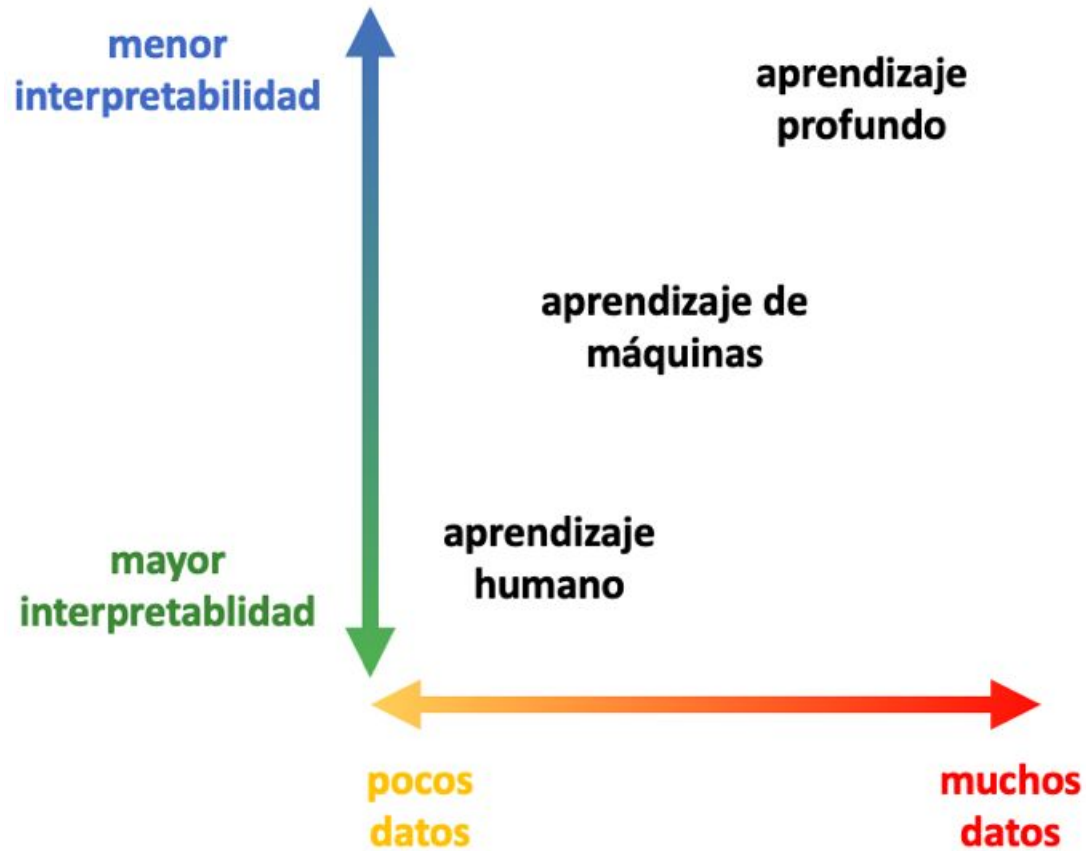
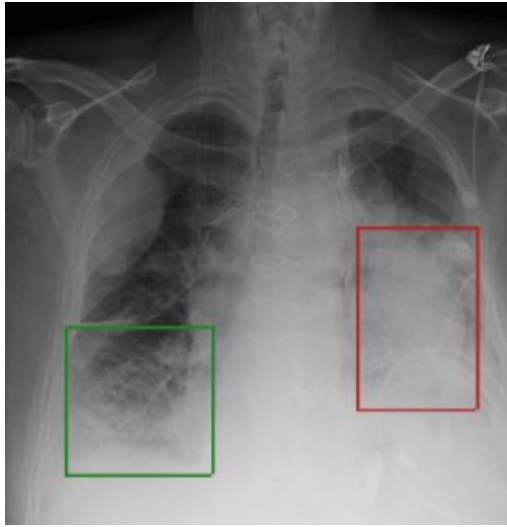


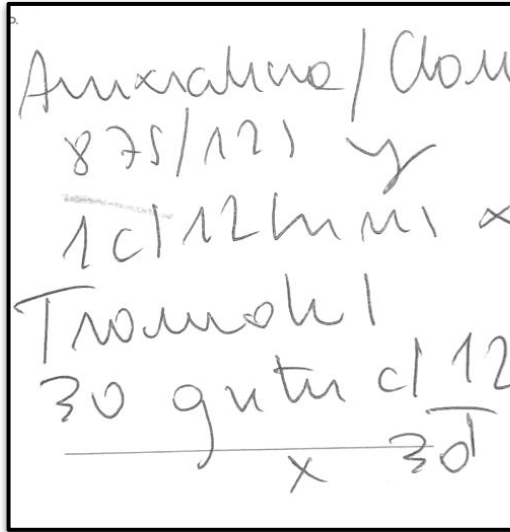
Figura adaptada de (Beam y Kohane, Big data and machine learning in healthcare. JAMA, 2018)

Procesamiento del Lenguaje Natural en Medicina



Visión Computacional

Medicina



Procesamiento del Lenguaje Natural



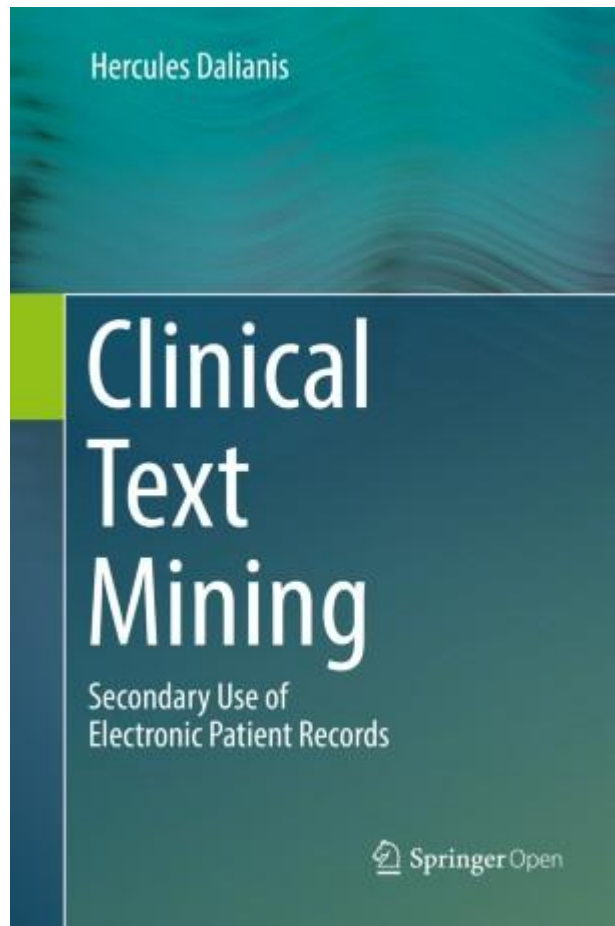
Robótica

Texto clínico

- El texto no estructurado representa una importante proporción de la narrativa clínica (e.g. resultados de exámenes, notas de pacientes hospitalizados, recetas)
- Mucha abreviación no estandarizada y errores de tipeo
- Es información que apoya toma de decisiones y uso secundario de datos.
- Existe disponibilidad restringida por razones de privacidad
- Falta de recursos lingüísticos para idiomas distintos del inglés

Algunos problemas clásicos de PLN en Medicina

- Detección de información clave (ej. enfermedades, medicamentos o dosis)
- Codificación automática (ej. GRD, CIE-10)
- Clasificación de textos (ej. Si una interconsulta pertenece al GES)
- Selección de cohortes de pacientes similares
- Anonimización de fichas clínicas



¡Hercules estuvo en Chile!

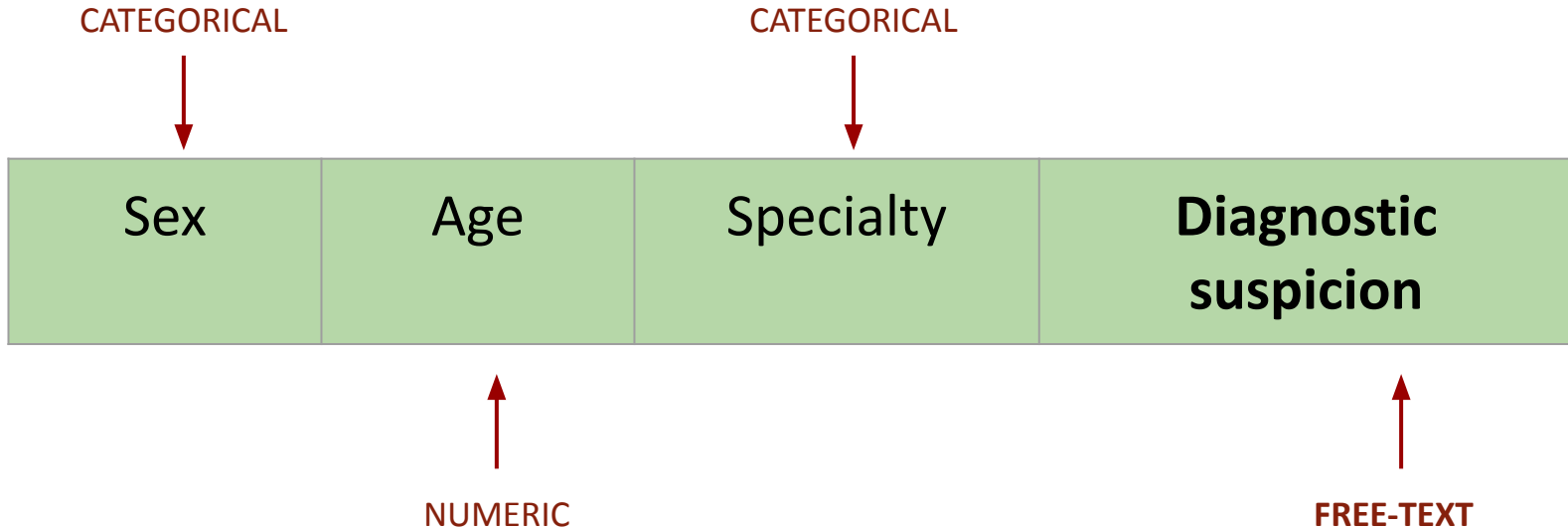
<https://link.springer.com/book/10.1007/978-3-319-78503-5>

Proyectos que hemos realizado

La lista de espera en hospitales públicos

Waiting List in Chilean Public Hospitals

- 75% of the Chilean population is in the public healthcare system
- To have a first consultation with a specialist you enter in a waiting list
- The median waiting time in the non-prioritized WL is 432 days.



Classification GES vs. non GES

Diagnostic suspicion	GES
RESTRICCION DEL CRECIMIENTO INTRAUTERINO	NO
IRC	NO
Ortesis	SI

Diagnostic
suspicion + age

GES label

Trabajo 46

Casos en conflicto pendientes de revisar

Acá se encuentran los casos que clasifiqué como GES, Procedimiento o Urgencia y creo que no deberían ser cargados a SIGTE. Por favor resuelva cada caso presionando el botón de la categoría correcta.

RUT	ID_LOCAL	EDAD	PRESTA_MIN	SOSPECHA_DIAG	G	P	U	Clasificación Definitiva
Ningún dato disponible en esta tabla								

Anterior Siguiente

Casos para eliminar de la planilla para subir a SIGTE

Acá se encuentran los casos que deben ser eliminados de la planilla a cargar a SIGTE. (También puede recorrerir cada caso.)

Descargar Planilla Corregida

RUT	ID_LOCAL	EDAD	PRESTA_MIN	SOSPECHA_DIAG	G	P	U	Clasificación Definitiva
	368831	47	18-02-021	GASTRECTOMIA TOTAL	No	No	SI	G P U S

Automatic **classifier to support decision making**. Used for 7 months at a public hospital

RESEARCH ARTICLE

Open Access

Supporting the classification of patients in public hospitals in Chile by designing, deploying and validating a system based on natural language processing

Fabián Villena^{1,2}, Jorge Pérez^{3,4}, René Lagos⁵ and Jocelyn Dunstan^{1,2*} 



Fabián Villena

The Chilean Waiting List Corpus

Since 2018 we have been collecting referrals written by primary care physicians. From the 11 million referrals collected, 10,000 were annotated.

Journals

[BMC Public Health](#) (2019)

[Revista Med. Chile](#) (2021)

[Revista Med. Clinica Las Condes](#) (2022)

[Clinical Dermatology](#) (2021)

[ACM Healthcare](#) (2022)

[BMC Med. Inf. Dec. Mak](#) (2021)

Conference Proceedings

[EMNLP Clinical Workshop](#) (2020)

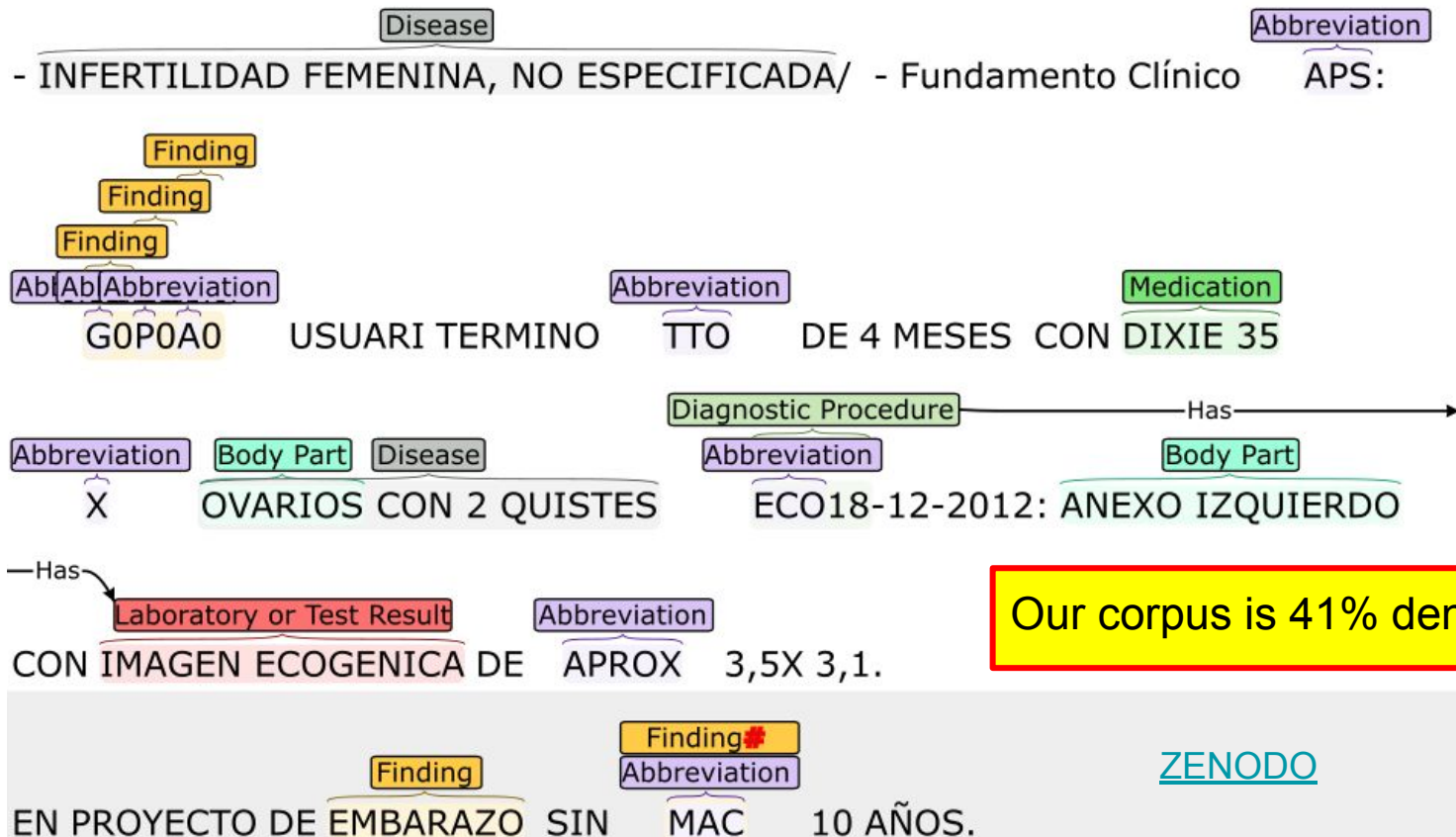
[ACL Clinical Workshop](#) (2022)

[Coling](#) (2022)

[EMNLP Clinical Workshop](#) (2022)

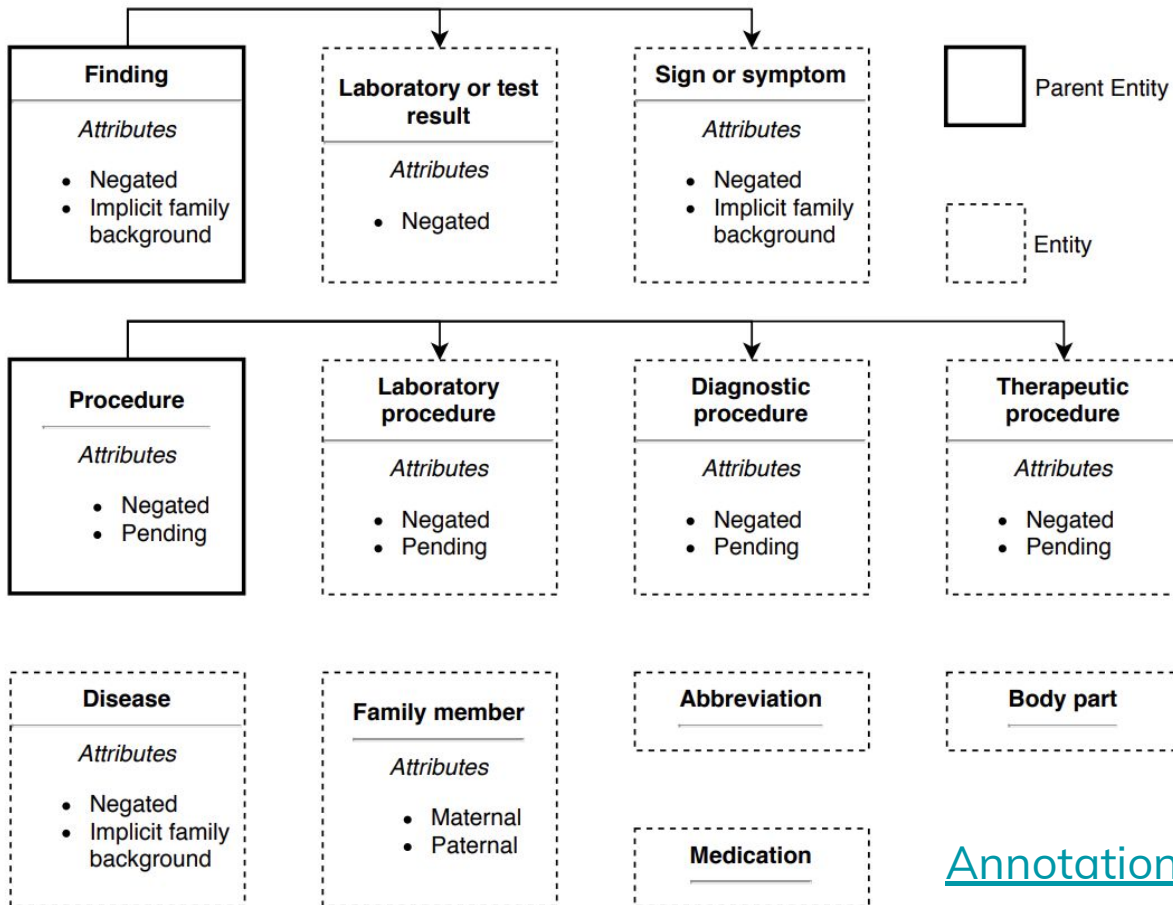


The Chilean Waiting List Corpus



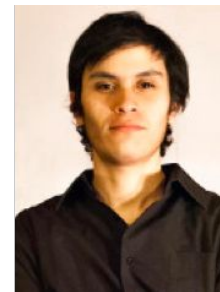
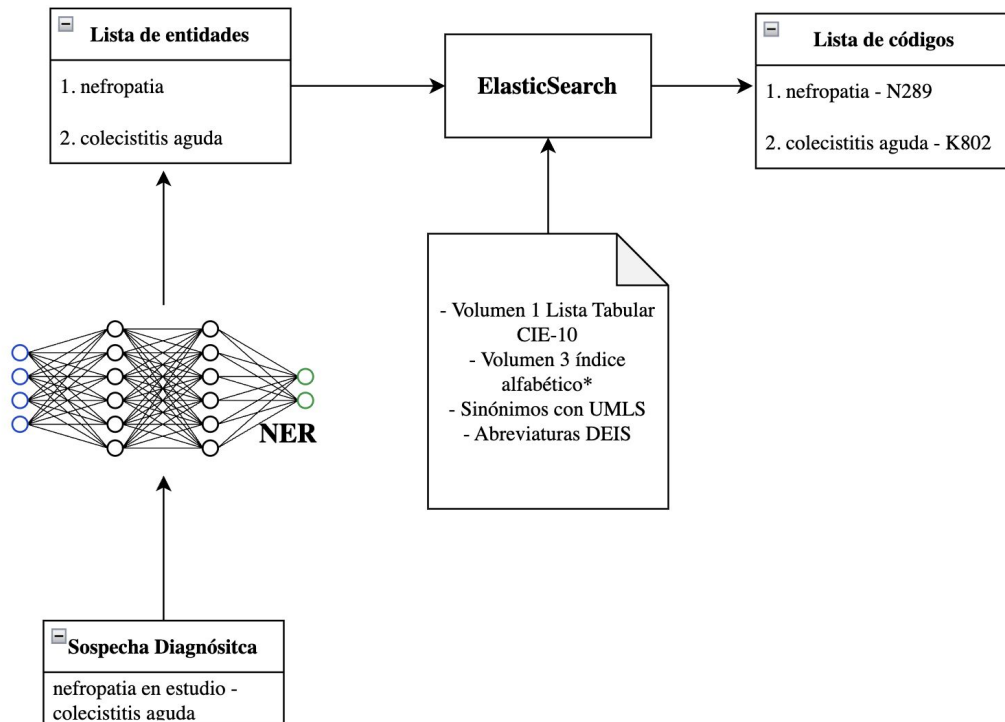
Our corpus is 41% dental

The Chilean Waiting List Corpus



[Annotation guidelines](#)

Automatic coding of the waiting list - nationwide



Fabián Villena

Transcription and information extraction



diente 1.1 con periodontitis apical asintomática Solicito
interconsulta endodoncia medicamento con hidróxido de calcio y
amoxicilina equivoco

PROCESAR BORRAR



- periodontitis apical asintomática

ENFERMEDADES

PARTES DEL CUERPO

- diente 1.1

MEDICAMENTOS

- hidróxido de calcio
- amoxicilina

COPIAR



Maicol Fernández



Fabián Villena

Fernandez et al, under review

Texto oncológico

Corpus of morphology & topography and ICD-O





We annotated morphology and topography mentions, adding CIE-O codes, in pathology reports from the largest oncological foundation in Chile. We enlarged the corpus using the BSC corpus Cantemist.



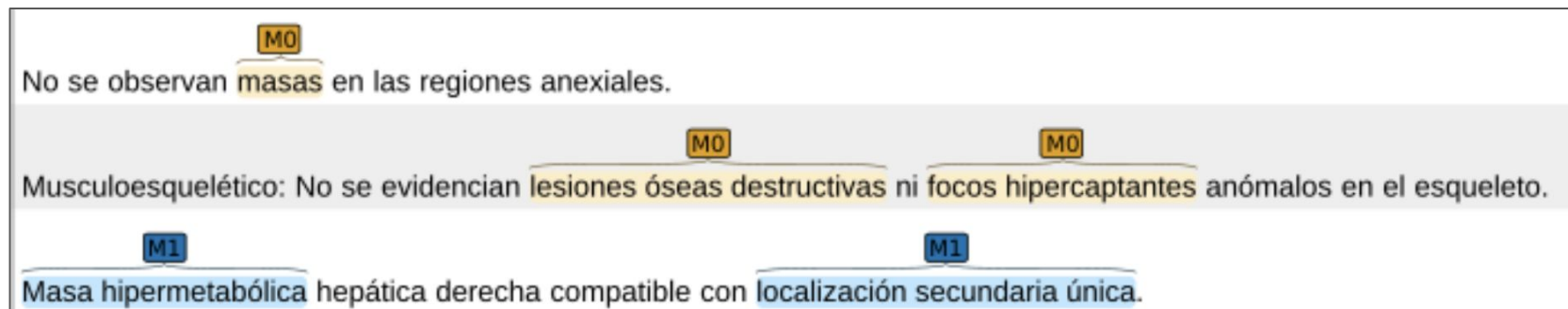
The image shows a login interface for two organizations: FALP (Juntos contra el cáncer) and CMM (Centro de Modelamiento Matemático). It includes a header with their logos, a title 'Iniciar sesión', an email input field with the placeholder 'jdunstan@uchile.cl', a password input field with a masked password '*****', and a button labeled 'INICIAR SESIÓN'. There is also a link for '¿OLVIDASTE TU CONTRASEÑA?'.

III.- Tiroidectomía total: **8260/3 | Morfologia** **Cáncer** del **C73.9 | Topografia** tiroides de 6 mm. de eje mayor con los caracteres de un **8260/3 | Morfologia** **8010/3 | Morfologia** carcinoma diferenciado papilar con esclerosis del estroma con infiltración de cápsula **C73.9 | Topografia** tiroidea .

Automatic Detection of Distant Metastasis Mentions in Radiology Reports in Spanish

Ricardo Ahumada, MSc¹ ; Jocelyn Dunstan, PhD² ; Matías Rojas, MSc³; Sergio Peñafiel, MSc⁴; Inti Paredes, MD, PhD⁴ ; and Pablo Báez, MD¹ 

[JCO Clinical Cancer Informatics](#)



FALP Radiology Reports: Annotated corpus for distant metastasis detection

[ZENODO](#)

Salud Ocupacional

Pre-trained language models in Spanish for health insurance coverage

Claudio Aracena^{1,2}, Nicolás Rodríguez³, Victor Rocco³, and Jocelyn Dunstan^{2,4,5,6}

Datasets	documents	tokens
Fine-tuning		
Admission	300 k	22.5 M
Medical	300 k	26.3 M
Admision+Medical	300 k	57.2 M
Continual Pre-training		
Admission	1.5 M	112.6 M
Medical	1.2 M	154.0 M
Admision+Medical	855 k	164.6 M
Pre-training		
Admision+Medical	7.1 M	1.03 B

The data was extracted from administrative and clinical records from an insurance and health provider that specialized in labor accidents. **Within this data, it is possible to find personal and sensitive information** such as personal and company names, addresses, health information, pre-existing conditions, and diagnoses, among others. **An anonymization process was not carried out since the model will be used for internal purposes and will not be released.** As a process of **memorization can occur** in the PLM, we believe it is best to keep the model private because privacy attacks can extract personal and sensitive information.

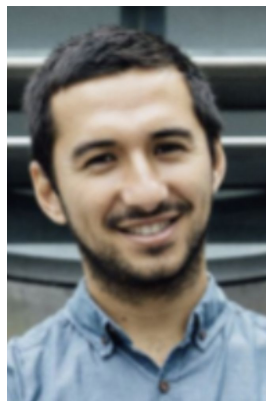
A Privacy-Preserving Corpus for Occupational Health in Spanish: Evaluation for NER and Classification Tasks

Claudio Aracena^{1,5}, Luis Miranda^{2,5}, Thomas Vakili³, Fabián Villena^{4,5},
Tamara Quiroga^{2,5}, Fredy Núñez-Torres⁶, Victor Rocco⁷, and Jocelyn Dunstan^{2,5}

Ingreso - 62 años **Age** , Am: asma, FA: 20/05/2032 **Full Date** , Alergias: No,
Ocupación: Director **Occupation** en Liceo del Sur **Institution** . PCTE refiere
que hoy miércoles 11/02 **Date Part** mientras trabajaba en sala de clases inicia
con ahogos, por lo que acude al hospital San Juan **Healthcare Unit** . Crisis
asmáticas a repetición el último tiempo (no usa inhalador).

ACHS-Privacy Corpus

[ZENODO](https://zenodo.org/record/10000000)



Claudio Aracena



Luis Miranda



Thomas Vakili

Clinical NLP Workshop this Friday!

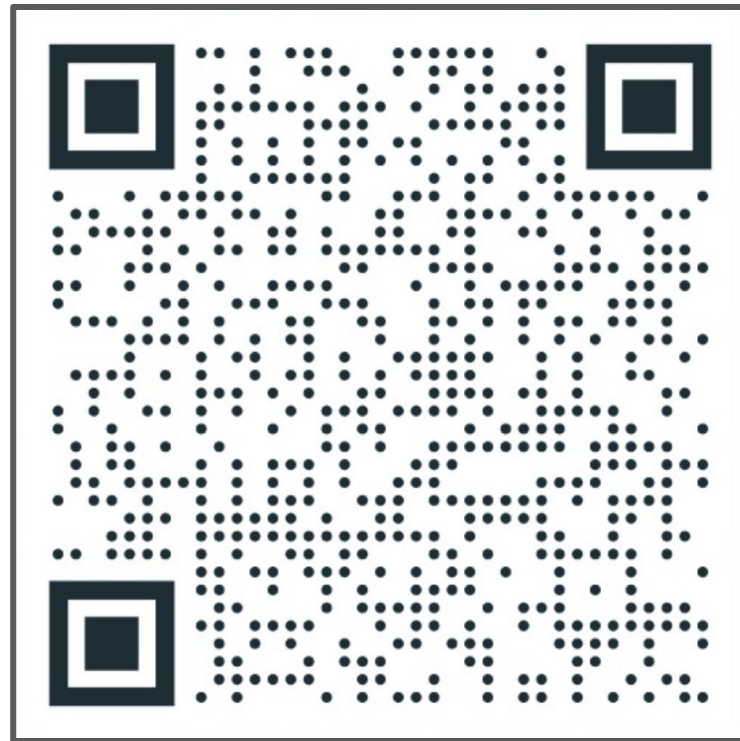
Conclusiones

- IA tiene variadas aplicaciones en medicina
- PLN apoya la extracción de información clave desde textos y dictado por voz
- PLN en medicina tiene características propias y es necesario crear recursos lingüísticos y computacionales para apoyar su uso en países que no hablan inglés
- El avance del área requiere el acceso a datos anonimizados y el apoyo a iniciativas interdisciplinarias

Ciencia de Datos

ON AIR

con Jocelyn
Dunstan Escudero



 jo_cientifica

Podcast Episode

¿Qué es ser radióloga?



Ciencia de Datos con Jocelyn Dunstan



Instituto Milenio
Fundamentos
de los datos



#PODCAST



Capítulo de hoy:
Electrónica y política

Invitado:
Dr. Héctor Ramírez
Investigador Titular y
Subdirector del AC3E

Capítulo también disponible en
el canal de YouTube del AC3E.

**Bonus: Productividad y mente académica en un
cuerpo casi sano**

IA hecha en Mexico



Ciencia de Datos con Jocelyn Dunstan



Ciencia de Datos

con Jocelyn
Dunstan Escudero



Especial
Memoria y datos:

Chile

Hugo Rojas
U. Alberto Hurtado
y VioDemos



o y por qué el modelo que
entrenaste en un hospital le va pésimo en otro

Bonus: Ciclo menstrual