La Revolucion de los Datos: Retos del Siglo XXI para la Ciencia, la Sociedad, la UNAM y el C3

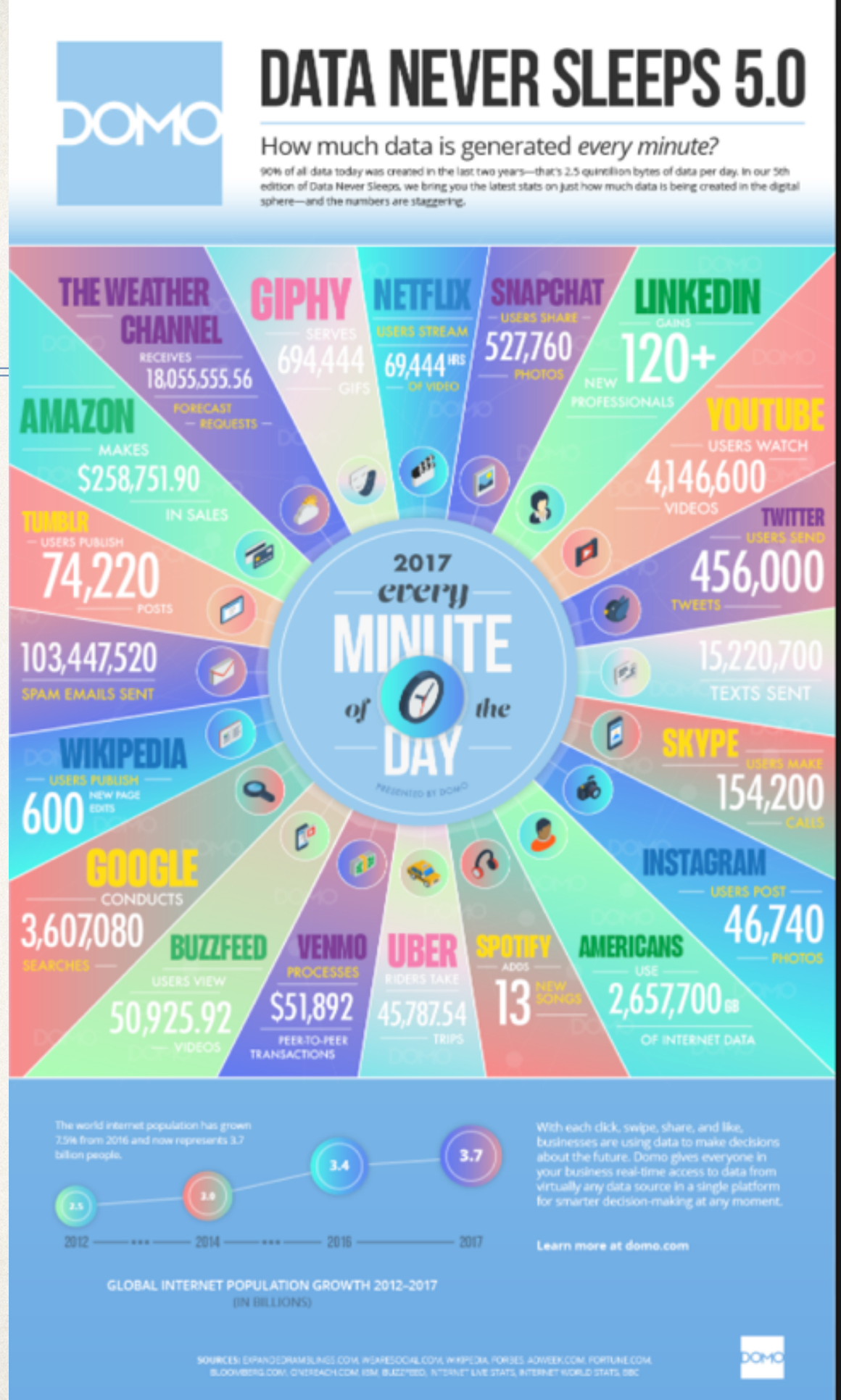
Chris Stephens
C3-Centro de Ciencias de la Complejidad y Instituto de Ciencias Nucleares, UNAM

CViCom 2018
UNAM 8-10/08/2018

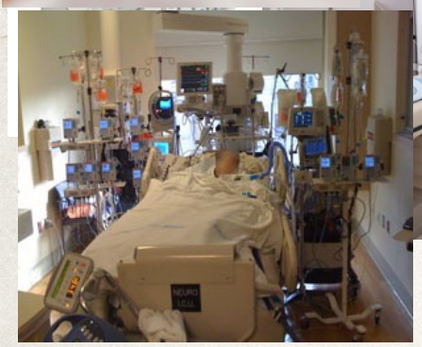
Ha habido una "Revolución de los Datos"

Afecta profundamente a nuestras vidas como científicos y ciudadanos

¿Pero, que exactamente es revolucionario?



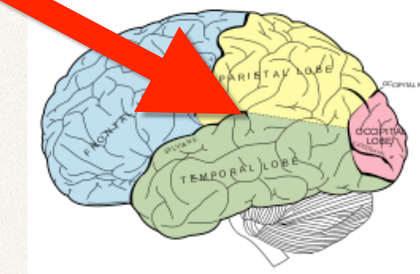
Datos profundos: La Revolución de los Datos y la Toma de Decisión



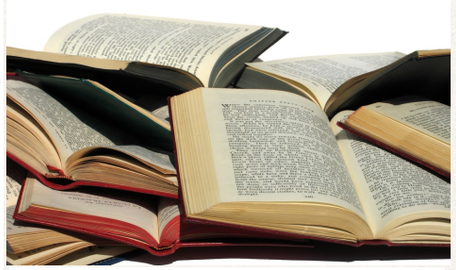
Una revolución en la generación de datos

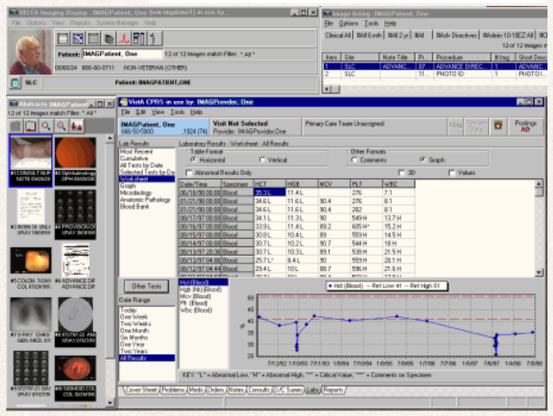
1 genoma humana
= 1GB (200)
Imagen de CT
= 10MB
Imagen de MRI
= 40MB

Cerebro humano
10-100 Terrabytes



Todos los libros en el mundo 30-50 Terrabytes



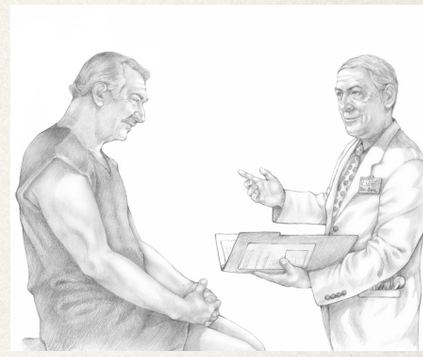


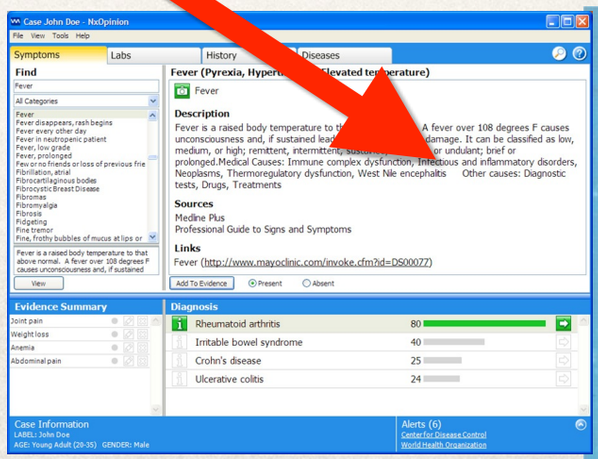
En forma electrónica
1 zettabyte

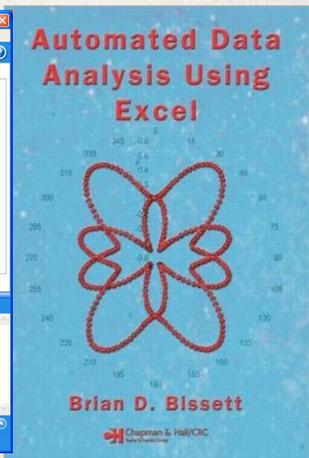
Una revolución en el almacenamiento de datos

No ha habido una revolución en:
Tipo de datos
Velocidad
Propósito

Una revolución en el análisis de datos





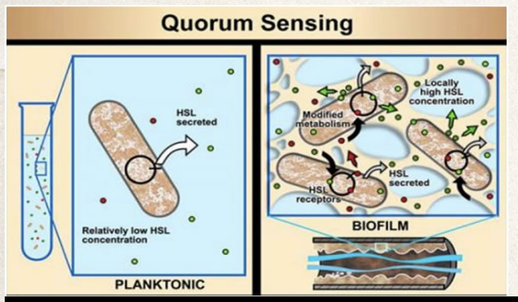


¿Porqué la Revolución de los Datos es y será tan importante?

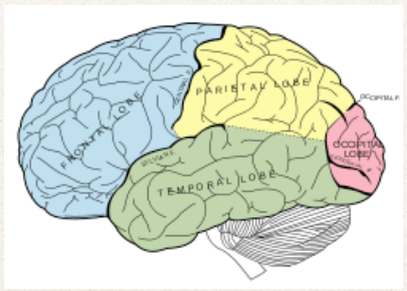
El propósito principal de los seres vivos (y un propósito principal de la ciencia) es… **Predecir** para la

Toma de Decisiones

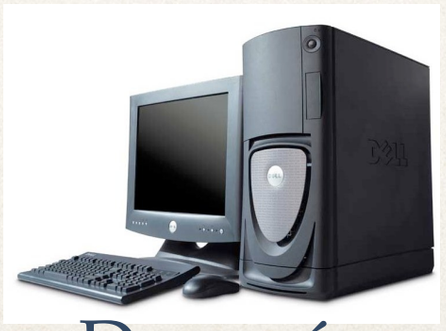
Todos los seres vivos son “mineros de datos” y minamos BIG data



¿Dónde están estos datos?



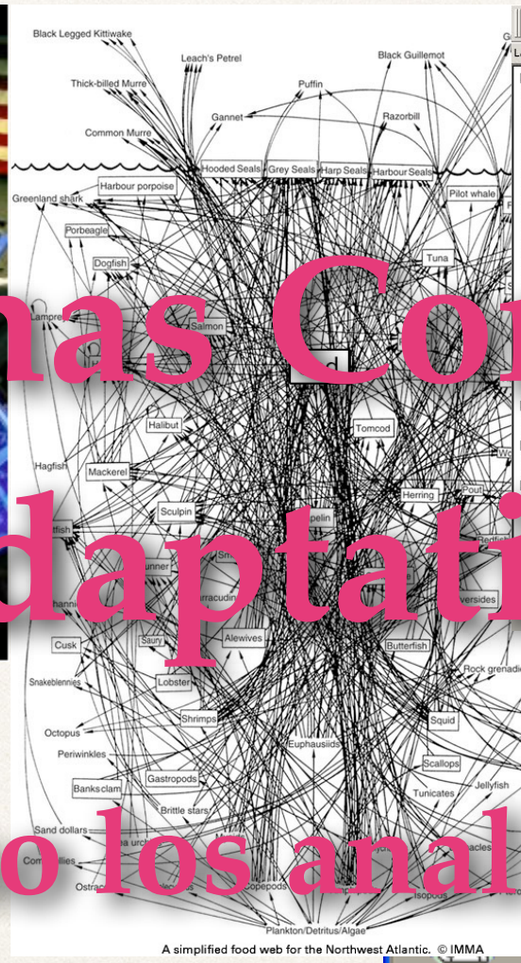
Antes



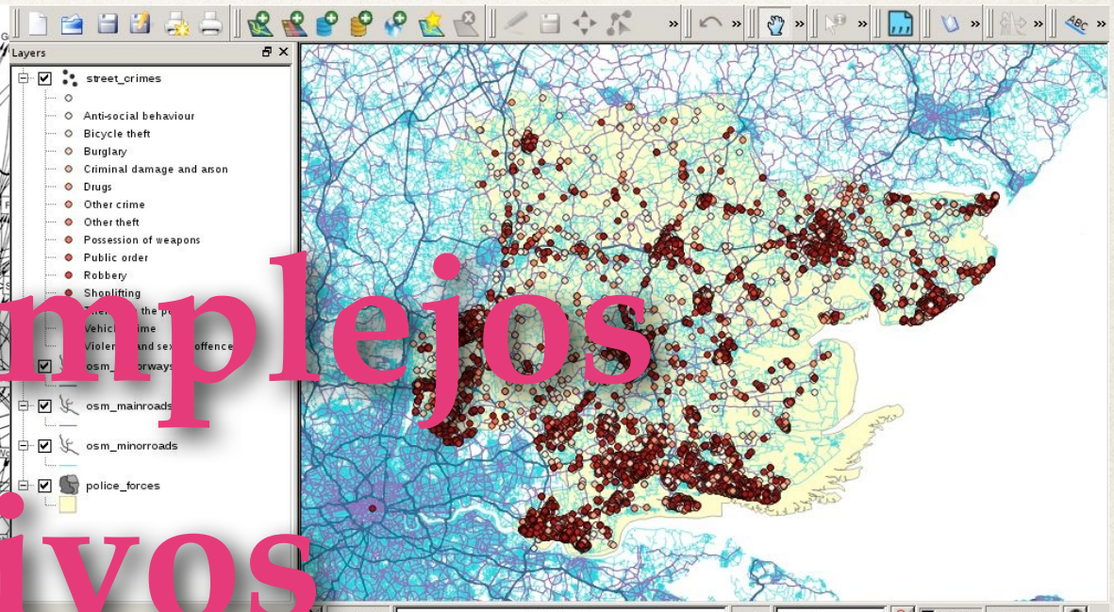
Después

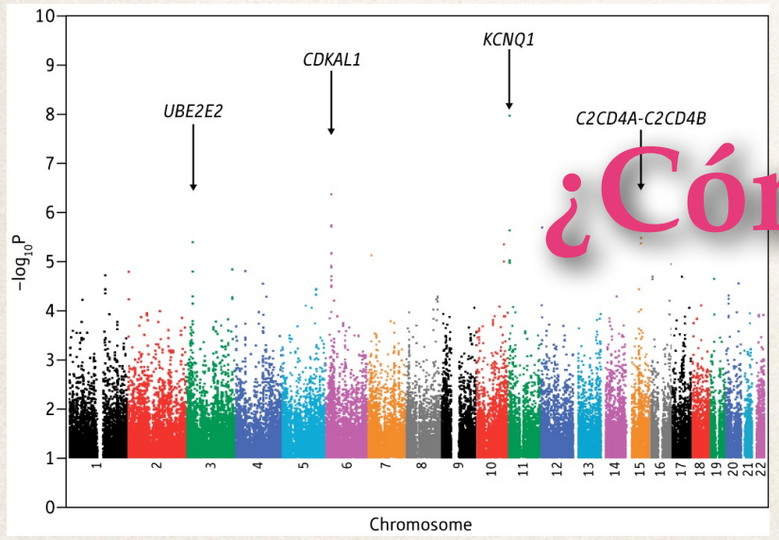
de la Revolución de los Datos

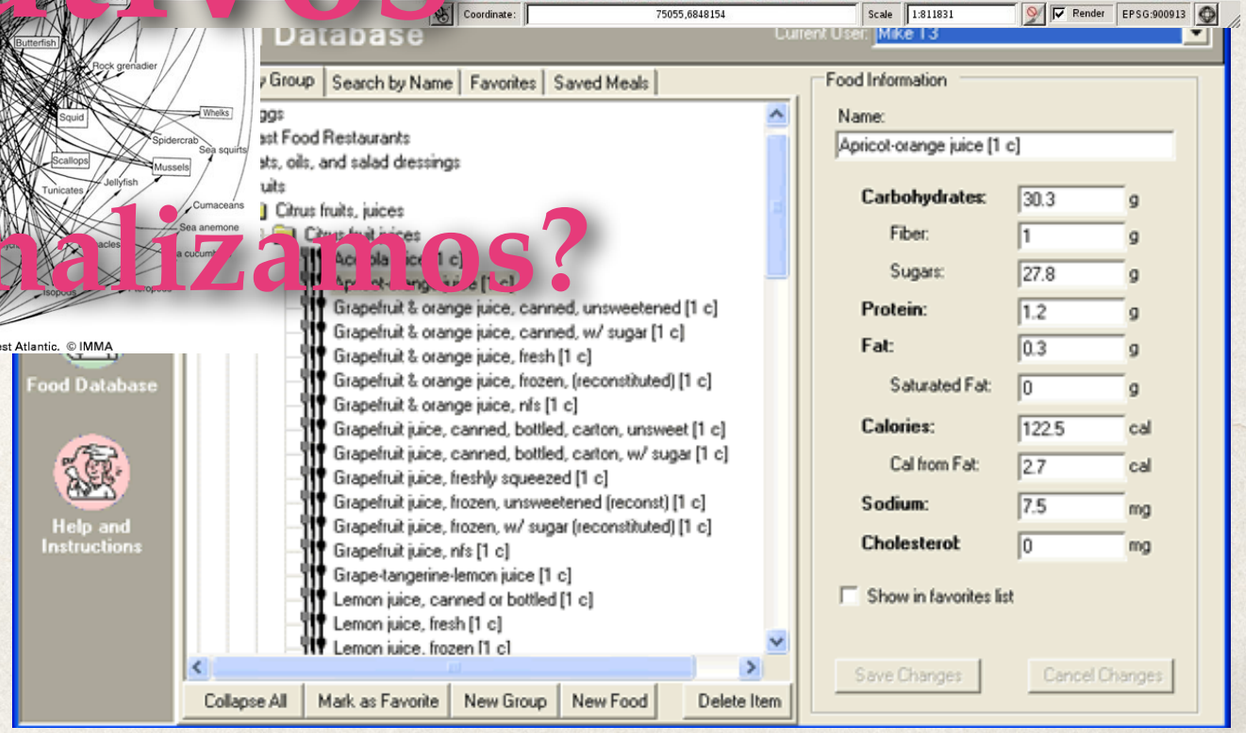
¿Qué representan los datos de la Revolución?



A simplified food web for the Northwest Atlantic. © IMMA



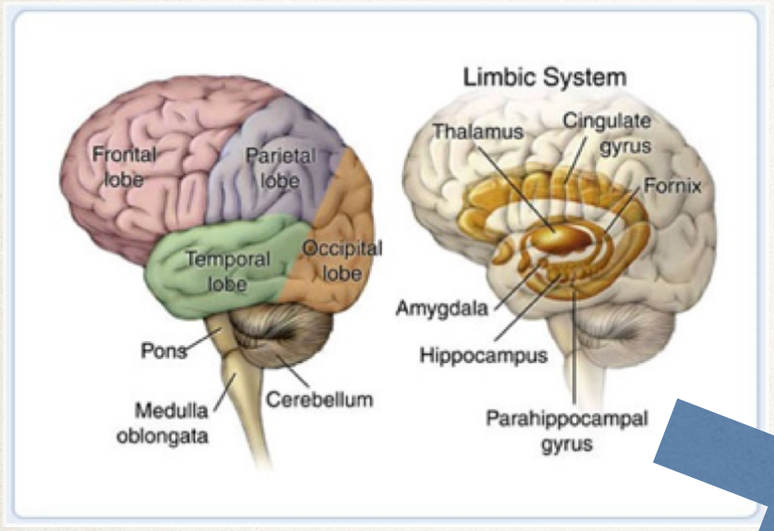


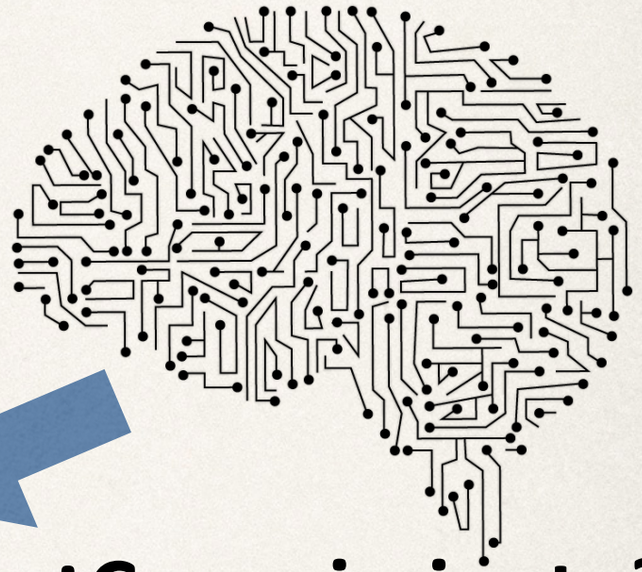


Sistemas Complejos Adaptativos

¿Cómo los analizamos?

Representación Algorítmica de una Predicción/Decisión





Conocimiento

$$P(C|X(t))$$

¿Conocimiento?

Heurística

Datos + Información +

Conocimiento

¿Qué formato?

Lenguaje natural, numérico, categórico, espacio-temporal

Los P(C|X) son “maquinas” - como maquinas físicas hacen una cosa bien

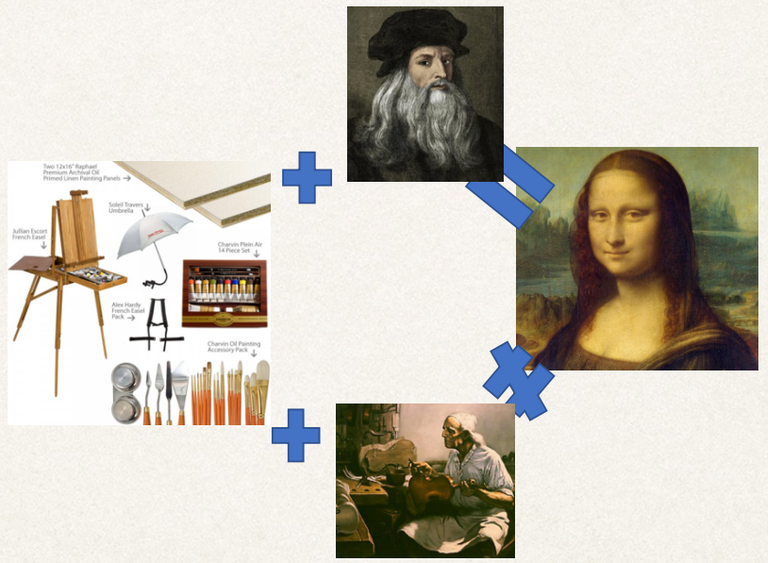
Cada decisión/acción que tomamos require y usa una maquina

DIFFERENTE, independientemente de si o no usamos IH or IA

La Importancia de Conocimiento de Dominio

Hay muchas cajas de herramientas

Seleccionando un P(C|X) es como seleccionar una herramienta. Pero…



Para problemas complejos - multi-factorial y multi-escala - se requiere múltiples disciplinas y trabajo en equipo

Misión del C3 -Centro de Ciencias de la Complejidad

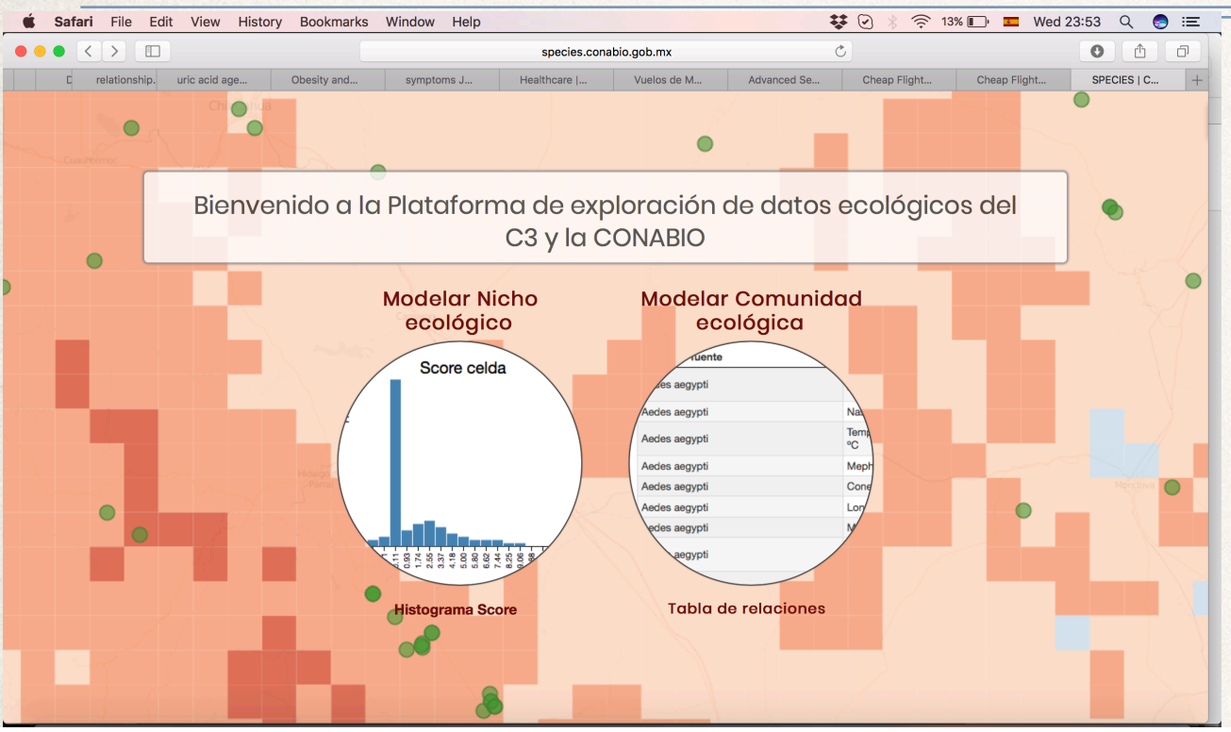
… realizar investigación científica transdisciplinaria de frontera en las ciencias de la complejidad, creando un espacio en donde expertos de muy diversas áreas puedan interactuar y contribuir a la solución de problemas trascendentes y de importancia nacional

Es también misión del Centro formar científicos entrenados en el trabajo transdisciplinario en equipo y en el fortalecimiento de los métodos modernos asociados a la ciencia computacional.

Importancia de buscar sinergias y colaboración con el CViCom

Algunos Proyectos del C3.

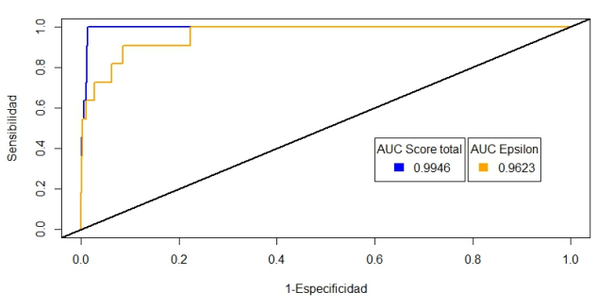
La computación juega un papel importante en todos



Detección de interacciones biológicas

- Depredación
- *Lynx rufus*

	Presa	Score total
Sylvilagus audubonii	1	32.9163696
Sylvilagus bachmani	0	32.9163696
Nasua nasua	0	30.5265046
Reithrodontomys microdon	1	29.2840247
Dipodomys ordii	0	29.1183394
Peromyscus eremicus	0	29.0314133
Peromyscus levipes	1	28.1092097
Chaetodipus penicillatus	1	27.9424153
Neotoma floridana	0	26.9401445
Sigmodon fulviventer	0	26.8817102
Sylvilagus graysoni	0	26.4597068
Reithrodontomys montanus	0	26.3436345
Callipepla gambelii	0	26.3158638
Picoides stricklandi	0	25.8401668
Antrostomus arizonae	0	25.5477411
Columbina inca	0	25.5359324
Spilogale angustifrons	0	25.0583737
Picoides villosus	0	24.9179631
Xerospermophilus spilosoma	1	24.6575532
Sorex emarginatus	0	23.9509356



Enfermedades Crónicas

"Salud y enfermedad: un enfoque desde las Ciencias de la Complejidad en la búsqueda de alarmas tempranas". PAPIIT

"Enfermedad y Salud: un enfoque desde las ciencias de la complejidad". Fronteras de la Ciencia, CONACyT

"Salud y enfermedad y las Alertas Tempranas". Fronteras de la Ciencia, CONACyT

"Obesidad y Diabetes y Médico en tu Casa - Colaboración C3, Lab CDMX, SEDESA

Enfermedades Emergentes

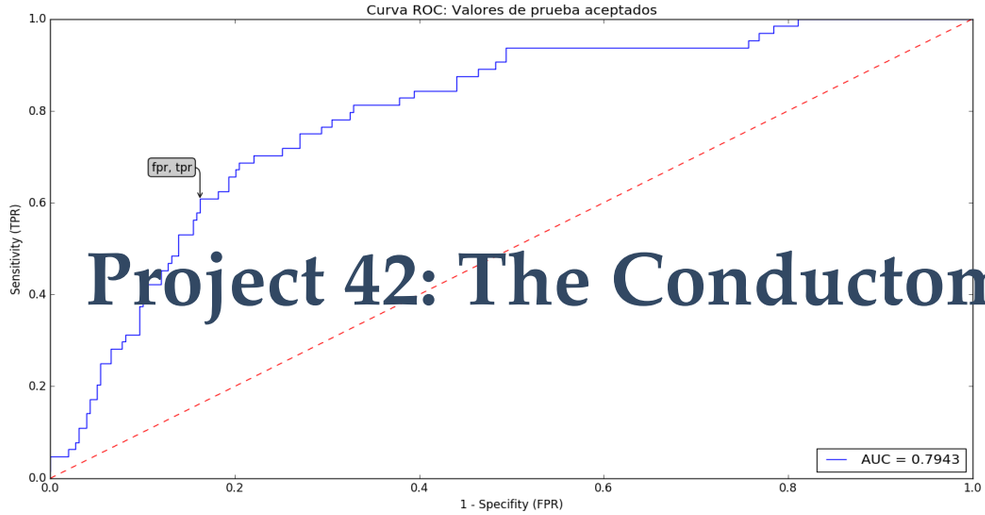
"Complejidad ecológica de las enfermedades emergentes". PAPIIT

"SPECIES", CONABIO

"Vacuna para la Enfermedad de Chagas", Fundación Carlos Slim

Cancer (Biología de Sistemas)

"Cancer como una enfermedad compleja: leucemia y cáncer epitelial, de lo micro a lo macro"



Project 42: The Conductome

Resultados de modelos predictivos basados en datos de un estudio de 1,800 no-académicos, académicos y estudiantes de la UNAM: 2,524 variables - Genético, epidemiológico, fisiológico,…

Conclusiones

- La Revolución de los Datos esta generando ~ 1 Zettabyte de datos cada año. Estamos ahogando en datos.

- No tenemos la más mínima idea de cuanto “conocimiento” hay en estos datos. No hablamos “base de datos”.

- En esos datos hay posibilidad de salvar/quitar vidas, evitar/fomentar conflicto, reducir/aumentar la pobreza,…, representar/distorsionar realidad

- El uso (y abuso) de estos datos requiere la Ciencia de la Computación, pero también requiere muchas más disciplinas (semántica de los datos).

- Para la solución de los grandes problemas complejos que enfrentamos, la sociedad necesita que la UNAM (y otros) produce nuevas maneras de hacer ciencia donde la computación juega un papel metodológico fundamental en proyectos transdisciplinarios de gran importancia.

- Hay grandes oportunidades para colaboraciones fructíferas entre el C3 y el CViCom en estos problemas.