



Datos como materia prima esencial en el siglo XXI

**Claudio Gutiérrez • DCC, Universidad de Chile / Instituto
Milenio de Fundamentos de los Datos •
cgutierr@dcc.uchile.cl**





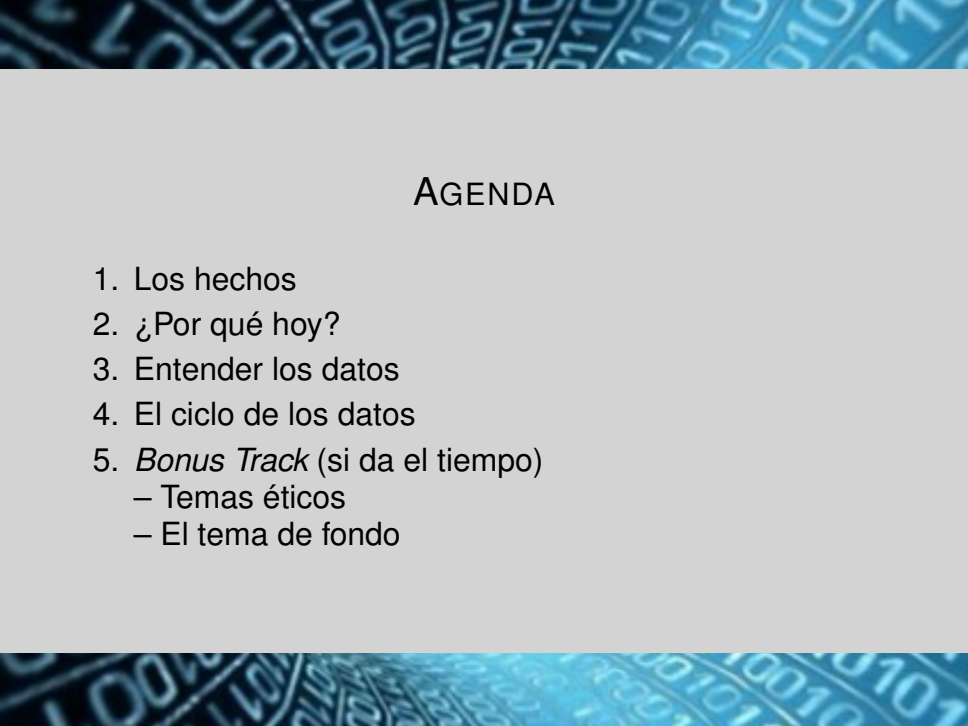
VERSIÓN 1 MINUTO

A decorative header featuring a blue background with white binary code (0s and 1s) arranged in a perspective that creates a sense of depth and movement.

La tecnología ha permitido crear un mundo virtual más allá de la escala humana gracias a inéditas técnicas de captura, almacenamiento y procesamiento de datos.

Los datos abren una etapa nueva en la vida de los humanos, esto es, no son solo un fenómeno tecnológico, social, productivo o empresarial. Entender esto es clave para saber cómo abordarlos y relacionarnos con ellos.

A decorative footer featuring a blue background with white binary code (0s and 1s) arranged in a perspective that creates a sense of depth and movement, mirroring the header.



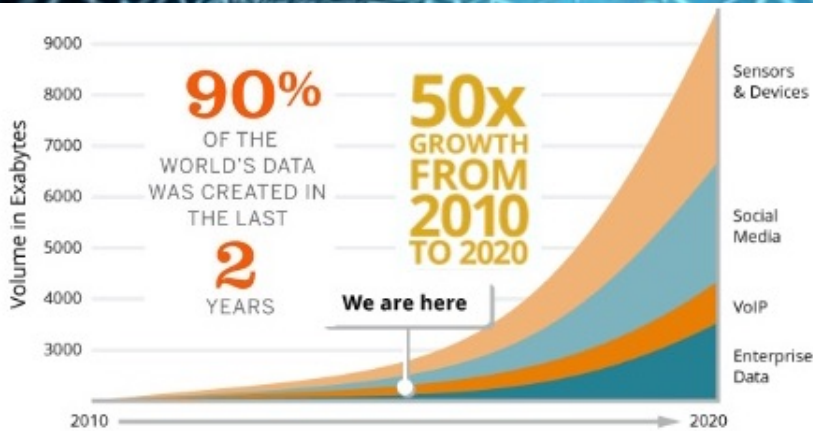
AGENDA

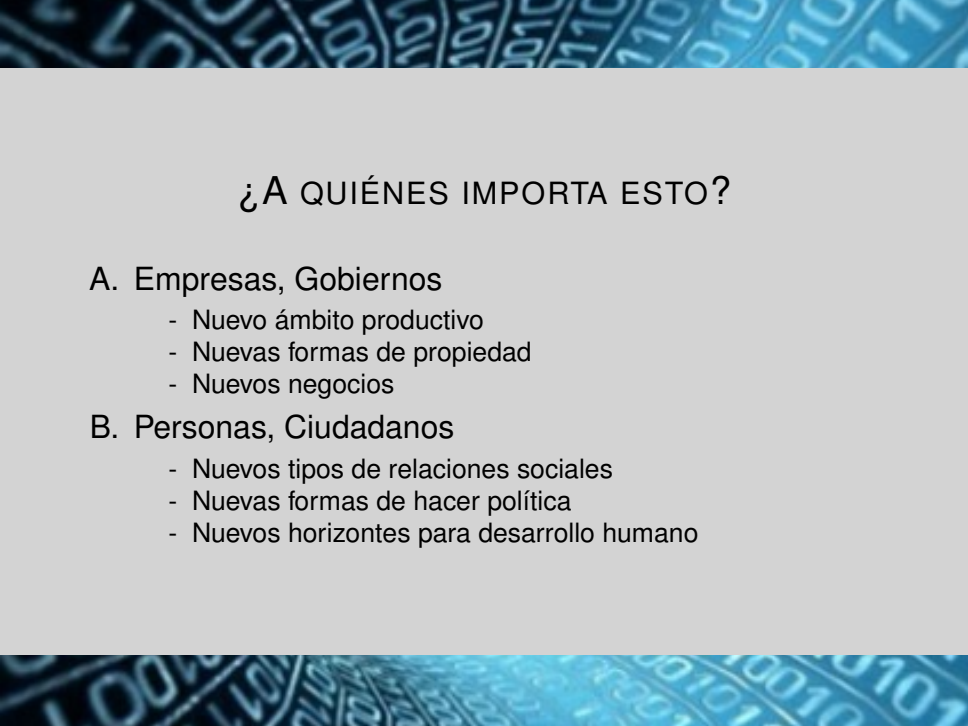
1. Los hechos
2. ¿Por qué hoy?
3. Entender los datos
4. El ciclo de los datos
5. *Bonus Track* (si da el tiempo)
 - Temas éticos
 - El tema de fondo



LOS HECHOS







¿A QUIÉNES IMPORTA ESTO?

A. Empresas, Gobiernos

- Nuevo ámbito productivo
- Nuevas formas de propiedad
- Nuevos negocios

B. Personas, Ciudadanos

- Nuevos tipos de relaciones sociales
- Nuevas formas de hacer política
- Nuevos horizontes para desarrollo humano



¿POR QUÉ HOY?



TECNOLOGÍAS NUEVAS

- Captura: sensores, cámaras, hardware, *pero sobre todo* redes sociales
 - Almacenamiento: grandes *data centers*
 - Análisis estadístico a gran escala (software)
 - Aprendizaje (redes neuronales) *pero sobre todo deep learning*
- 



El **fundamento** que posibilita lo que está ocurriendo, el fenómeno **realmente nuevo** que está cambiando nuestra forma de relacionarnos es la emergencia de un nuevo mundo:

→ LOS DATOS





ENTENDER LOS DATOS



DOS FENÓMENOS ACTUALES

- Datos más allá de la escala humana
- Un nuevo mundo está siendo generado

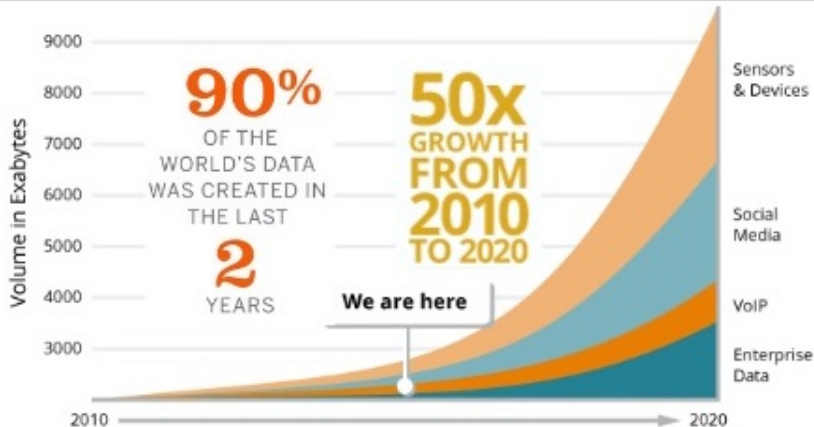




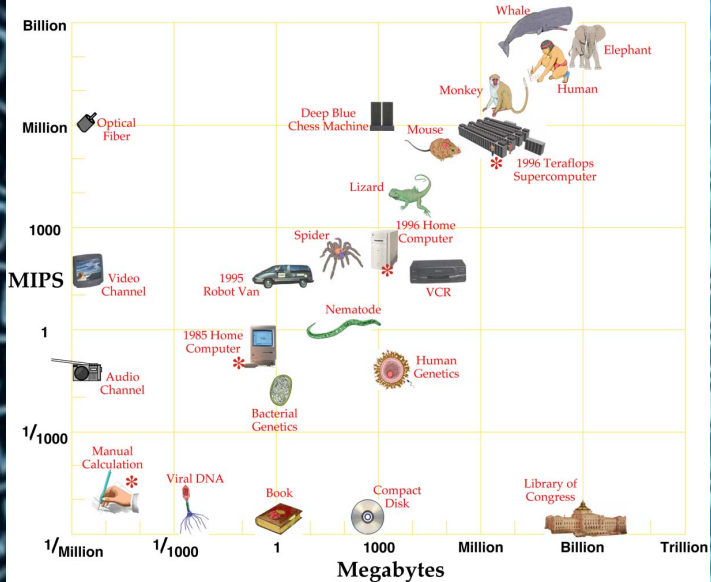
escala humana	{	Byte B	$\sim 10^0$	un carácter
		Kilo KB	$\sim 10^3$	texto escrito
		Mega MB	$\sim 10^6$	imagen, música
		Giga GB	$\sim 10^9$	videos

más allá	{	Tera TB	$\sim 10^{12}$	Biblioteca Congreso USA
		Peta PB	$\sim 10^{15}$	Data center grande
		Exa EB	$\sim 10^{18}$	Todas las palabras habladas
		Zetta ZB	$\sim 10^{21}$	Cantidad de datos globales





All Things, Great and Small



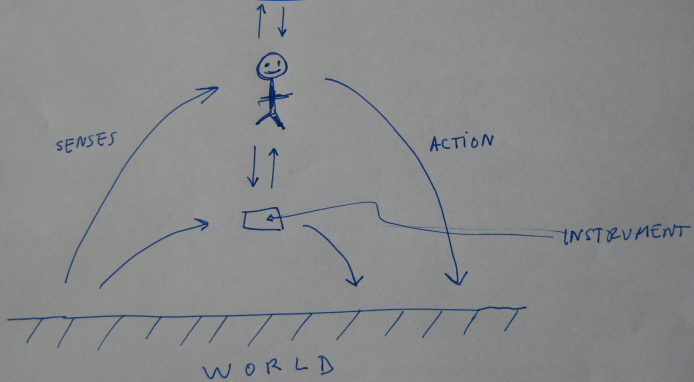


UNA BUENA METÁFORA

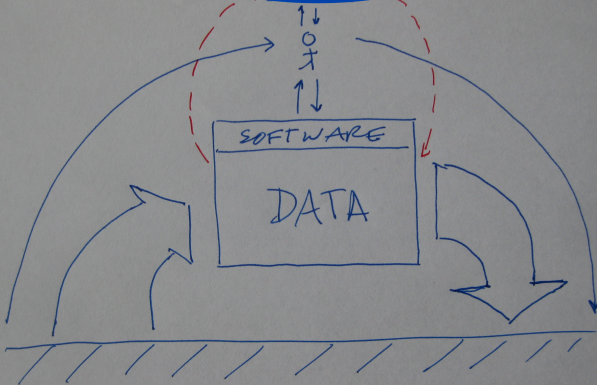
$$\frac{\text{Datos}}{\text{Mundo virtual}} = \frac{\text{Átomos}}{\text{Mundo material}}$$



ideas, conocimiento, etc.



ideas, conocimiento, etc.





EL CICLO DE LOS DATOS

EL CICLO DE LOS DATOS

1. Generación y captura
2. Almacenamiento y gestión
3. Integración y procesamiento
4. Disponibilización y usos

Cada una de estas etapas ha sido estudiada y conocida desde hace siglos. Lo nuevo hoy son:

- la **escala** (más allá de lo humano)
- las **técnicas de captura** (implícita)
- el **procesamiento** (aprendizaje de máquinas)

Data Lifecycle

Data is figuratively a living entity. Each juncture of the data lifecycle has different security needs, and carries a level of risk addressed by a mix of policy, people and technology.

10. DESTRUCTION

Risk level: ●

1. COLLECTION

Risk level: ●

2. RELEVANCE

Risk level: ●

3. CLASSIFICATION

Risk level: ●

4. HANDLING AND STORAGE

Risk level: ●

5. TRANSMISSION AND TRANSPORTATION

Risk level: ●

6. MANIPULATION, CONVERSION OR ALTERATION

Risk level: ●

7. RELEASE

Risk level: ●

8. BACKUP

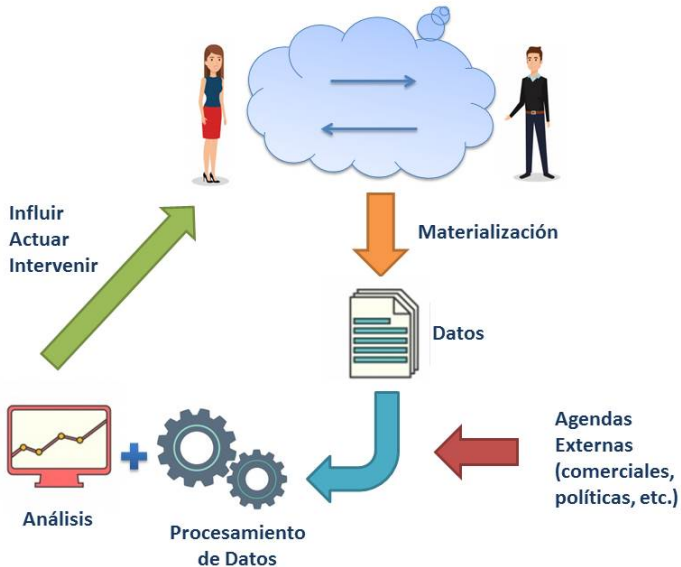
Risk level: ●

9. RETENTION

Risk level: ●

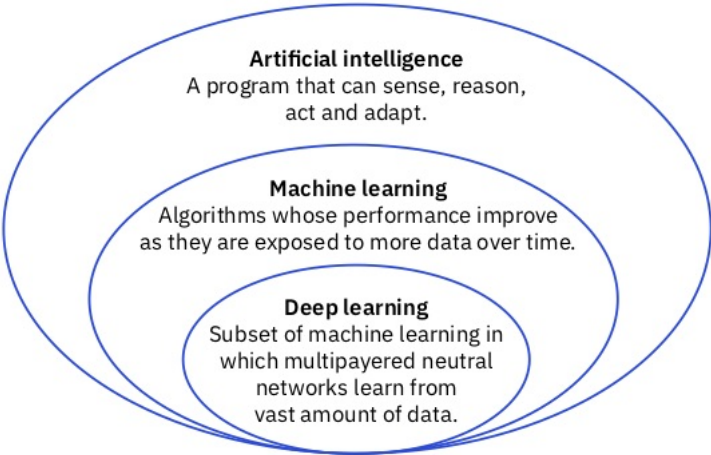
RISK LEVEL KEY

- High danger
- Moderate danger
- Low danger





UNA GRAN APLICACIÓN: IA



The diagram consists of three concentric ellipses. The outermost ellipse is labeled 'Artificial intelligence' and contains the text 'A program that can sense, reason, act and adapt.' Inside this ellipse is a smaller one labeled 'Machine learning' with the text 'Algorithms whose performance improve as they are exposed to more data over time.' Inside the 'Machine learning' ellipse is the smallest one, labeled 'Deep learning', with the text 'Subset of machine learning in which multipayered neutral networks learn from vast amount of data.'

Artificial intelligence

A program that can sense, reason,
act and adapt.

Machine learning

Algorithms whose performance improve
as they are exposed to more data over time.

Deep learning

Subset of machine learning in
which multipayered neutral
networks learn from
vast amount of data.



TRES CONCEPTOS CLAVES

1. Redes neuronales
2. Algoritmos de aprendizaje
3. Datos de entrenamiento



The image features a decorative header and footer with a blue and white binary code pattern (0s and 1s) arranged in a grid-like fashion. The central area of the slide is a solid light gray.

DATOS DE ENTRENAMIENTO

1. Datos son la base del aprendizaje (por eso las empresas andan detrás de los datos)
2. Muchos tipos: relacionales (tablas); texto; imágenes; video; grafos; etc.
3. Hoy son casi más importantes que los algoritmos (que se están estandarizando para aplicaciones comunes)

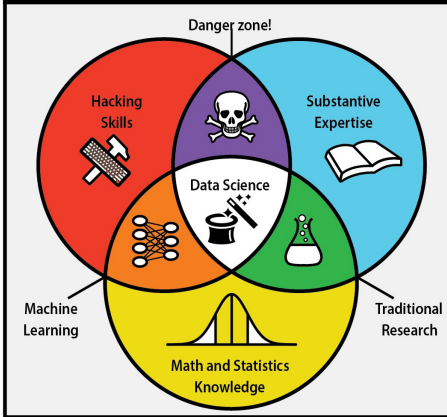
EJEMPLOS DE APLICACIONES EXITOSAS

1. Asuntos diarios: navegar un mapa del Metro, reconocer objetos en imágenes, traducir entre diferentes idiomas
2. Viajes: manejar, volar un dron, organizar agendas óptimas
3. Salud: detectar cáncer en tejidos, predecir hypoglicemia, detectar anomalías en imágenes
4. Ciencias: Predecir crecimiento cultivos, clasificar granos, detectar plagas, identificar clima amenazante
5. Seguridad: detectar malware, verificar identidad, anticipar pagos fraudulentos
6. Documentos: Identificar errores en documentos legales, identificar un autor, hacer un resumen
7. Juegos: Ganar en casi cualquier juego (poker, super mario, jeopardy, Go, etc.)



LXS DATA SCIENTISTS

DATA SCIENCE SKILLSET



Data science, due to its interdisciplinary nature, requires an intersection of abilities: **hacking skills**, **math and statistics knowledge**, and **substantive expertise** in a field of science.



Hacking skills are necessary for working with massive amounts of electronic data that must be acquired, cleaned, and manipulated.



Math and statistics knowledge allows a data scientist to choose appropriate methods and tools in order to extract insight from data.



Substantive expertise in a scientific field is crucial for generating motivating questions and hypotheses and interpreting results.



Traditional research lies at the intersection of knowledge of math and statistics with substantive expertise in a scientific field.



Machine learning stems from combining hacking skills with math and statistics knowledge, but does not require scientific motivation.



Danger zone! Hacking skills combined with substantive scientific expertise without rigorous methods can beget incorrect analyses.

MODERN DATA SCIENTIST

Data Scientist, the sexiest job of the 21st century, requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- ☆ Machine learning
- ☆ Statistical modeling
- ☆ Experiment design
- ☆ Bayesian inference
- ☆ Supervised learning: decision trees, random forests, logistic regression
- ☆ Unsupervised learning: clustering, dimensionality reduction
- ☆ Optimization: gradient descent and variants

DOMAIN KNOWLEDGE & SOFT SKILLS

- ☆ Passionate about the business
- ☆ Curious about data
- ☆ Influence without authority
- ☆ Hacker mindset
- ☆ Problem solver
- ☆ Strategic, proactive, creative, innovative and collaborative

PROGRAMMING & DATABASE

- ☆ Computer science fundamentals
- ☆ Scripting language e.g. Python
- ☆ Statistical computing packages, e.g., R
- ☆ Databases: SQL and NoSQL
- ☆ Relational algebra
- ☆ Parallel databases and parallel query processing
- ☆ MapReduce concepts
- ☆ Hadoop and Hive/Pig
- ☆ Custom reducers
- ☆ Experience with xaaS like AWS

COMMUNICATION & VISUALIZATION

- ☆ Able to engage with senior management
- ☆ Story telling skills
- ☆ Translate data-driven insights into decisions and actions
- ☆ Visual art design
- ☆ R packages like ggplot or lattice
- ☆ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau





TEMAS ÉTICOS

ALGUNOS USOS Y EJEMPLOS

- Juegos y adicción
- Burbujas sociales
- Publicidad personalizada (e.g. Google Ads, etc.)
- Creación de necesidades y *deseos*
- Conocimiento y testeo de políticas
- Manipulación de electores (UK, USA, Brasil, Chile?)
- Crédito social chino
- etc. etc.



TEMAS DE PROPIEDAD DATOS COMO BIEN PUBLICO

Un bien público tiene dos propiedades críticas:

1. NO-RIVALIDAD DE CONSUMO: el consumo de un individuo no desplaza del consumo a otro;
 2. NO-EXCLUSIÓN: es difícil si no imposible excluir a un individuo de gozar del bien [...]
- 



MUNDO DIGITAL COMO ECOSISTEMA

El principal desafío es cómo quisiéramos *gestionar* y *gobernar* estos bienes, incluyendo su ciclo completo, esto es, cómo se genera, accesa, almacena, procesa y se difunde.



A decorative header featuring a blue background with white binary code (0s and 1s) arranged in a perspective that creates a sense of depth and movement.

RECURSOS DIGITALES COMO COMUNES

Las cuestiones esenciales para cualquier análisis de comunes refieren inevitablemente a *equidad, eficiencia y sustentabilidad*.

Equidad refiere a asuntos de apropiación justa, igual contribución y mantenimiento de un recurso. *Eficiencia* se trata de la producción, gestión y uso óptimo de un recurso.

Sustentabilidad mira los resultados sobre el largo plazo.

Ch. Hess, E. Ostrom, 2006.

A decorative footer featuring a blue background with white binary code (0s and 1s) arranged in a perspective that creates a sense of depth and movement, mirroring the header.



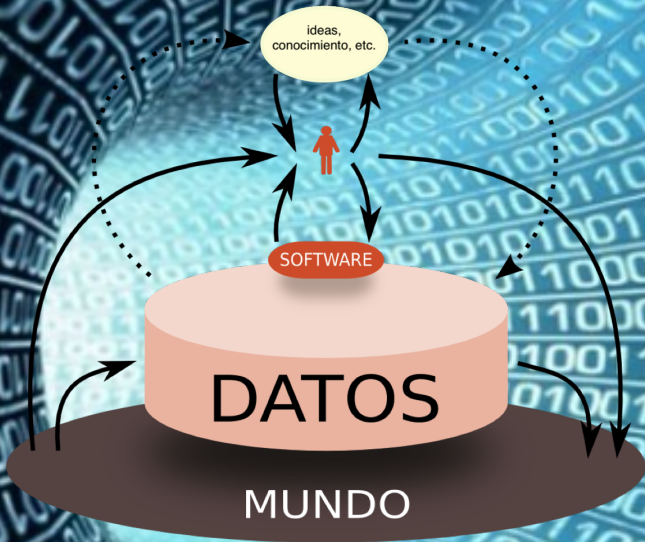
EL TEMA DE FONDO



LOS DATOS CREAN NUEVOS MUNDOS

La naturaleza de la acción humana cambió con los datos masivos porque ellos constituyen un **mundo nuevo** que modifica sustancialmente la relación del humano con su entorno





gracias!

cgutierr@dcc.uchile.cl



Créditos imágenes: Getty Images/iStockphoto, freepik.com, Ana Martínez,
infocux Technologies/Flickr, www.promptcloud.com