Curación de datos

Por qué (y cómo) mantener los datos

Máster en Data Science

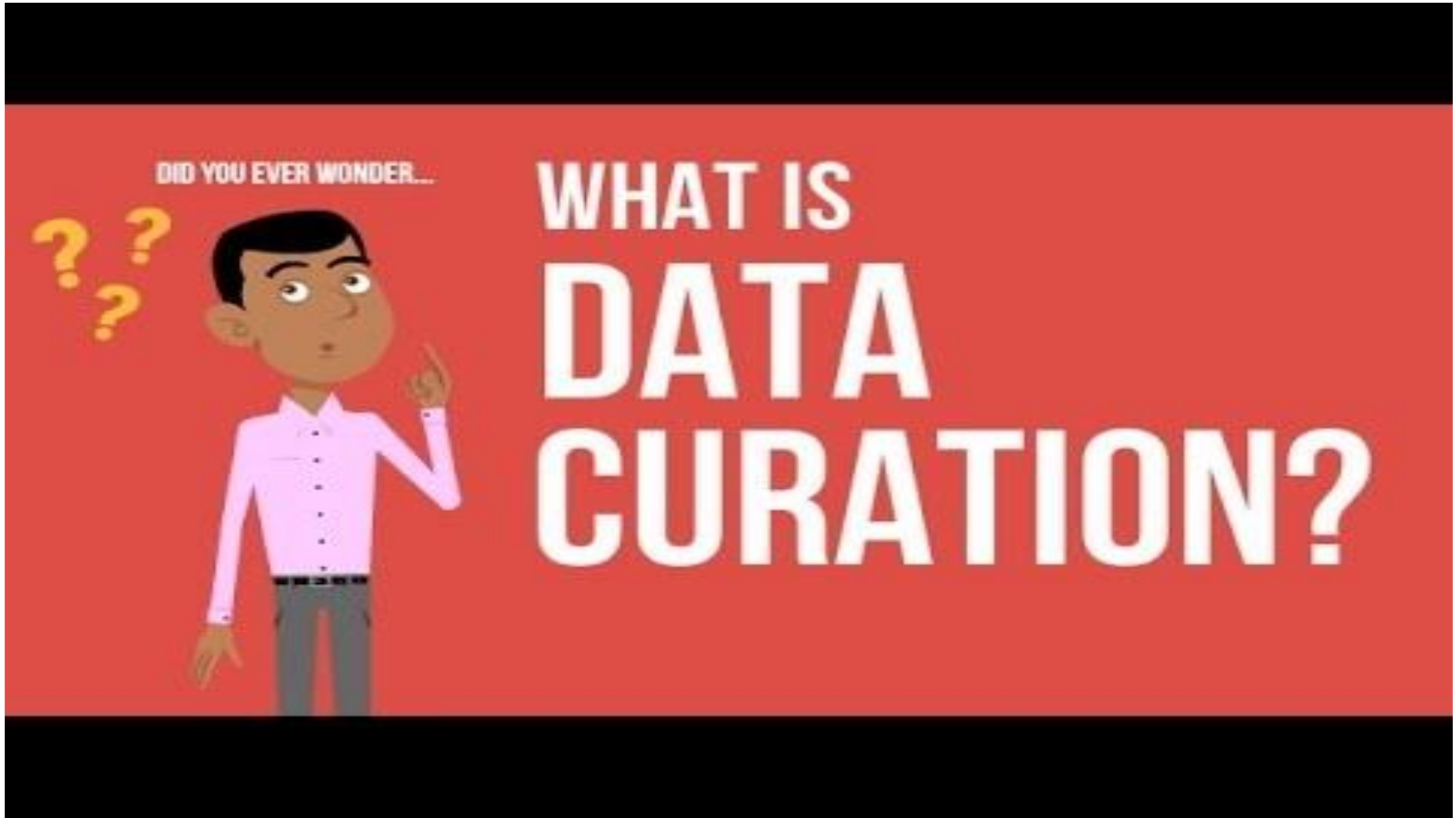
M1968 - El Ciclo de Vida de los Datos: de la Adquisición a la Presentación (2018-2019)

Diego García Saiz
Grupo de Ingeniería de Software y Tiempo Real
Departamento de Ingeniería Informática y Electrónica

¿Qué es la curación de datos?

- Si buscamos en internet, podemos encontrar diferentes definiciones:
  - "Data Curation is the active and ongoing management of data through its life cycle of interest and usefulness." (The University of Illinois' Graduate School of Library and Information Science**)***
  - "The processes of collecting data from diverse sources and integrating it into repositories that are many more times more valuable than the independent parts." (techrepublic)*
  - "Digital curation involves maintaining, preserving and adding value to digital research data throughout its lifecycle." (Digital Curation Centre)*
  - "The process of "caring" for Data, including to organizing, describing, cleaning, enhancing and preserving data for public use. Through curation the ICPSR (the International Leader in Data Stewardship) provides meaningful and enduring access to data." (ICPSR)*
  - "A means of managing data that makes it more useful for users engaging in data discovery and analysis." (Alation)*

**Definiciones extraídas de: http://www.dataversity.net/what-is-data-curation/*

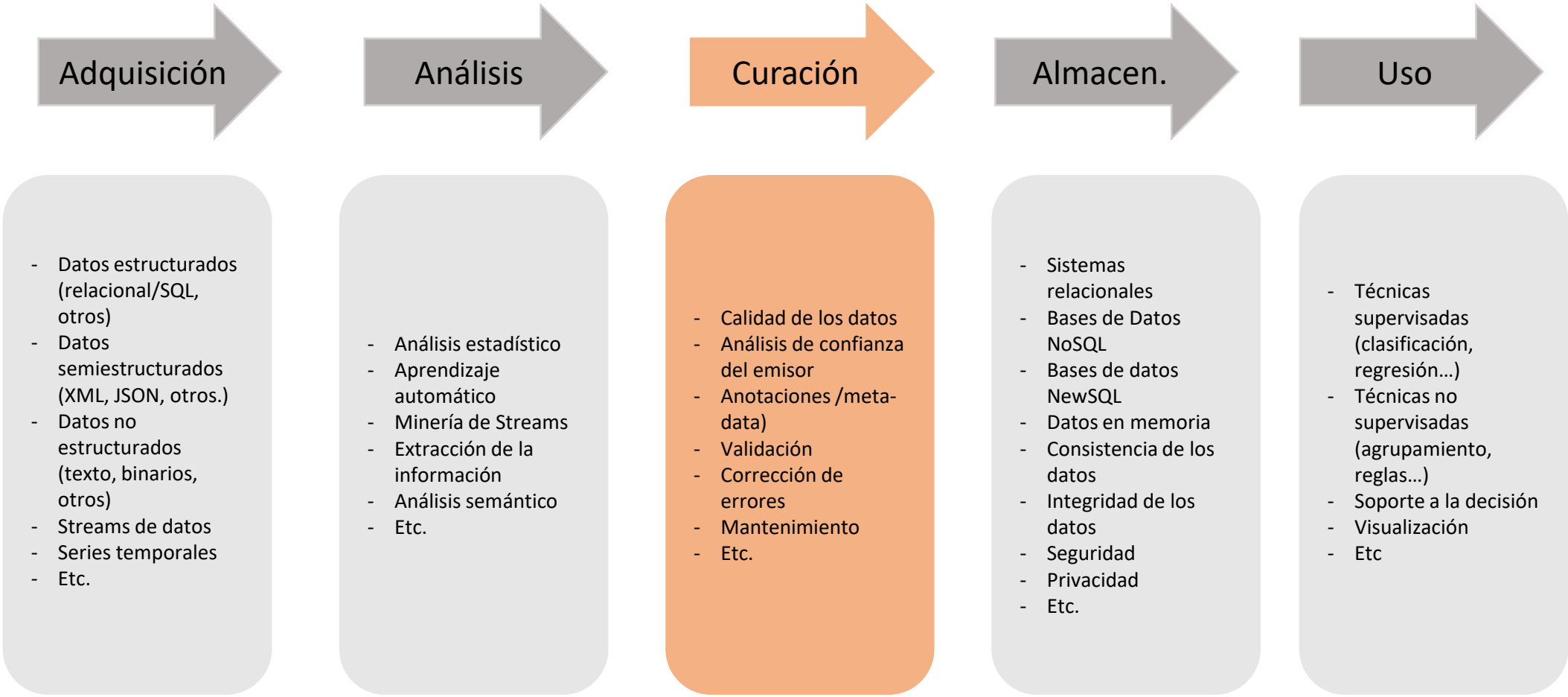

DID YOU EVER WONDER...
WHAT IS
DATA
CURATION?

¿Qué es la curación de datos?

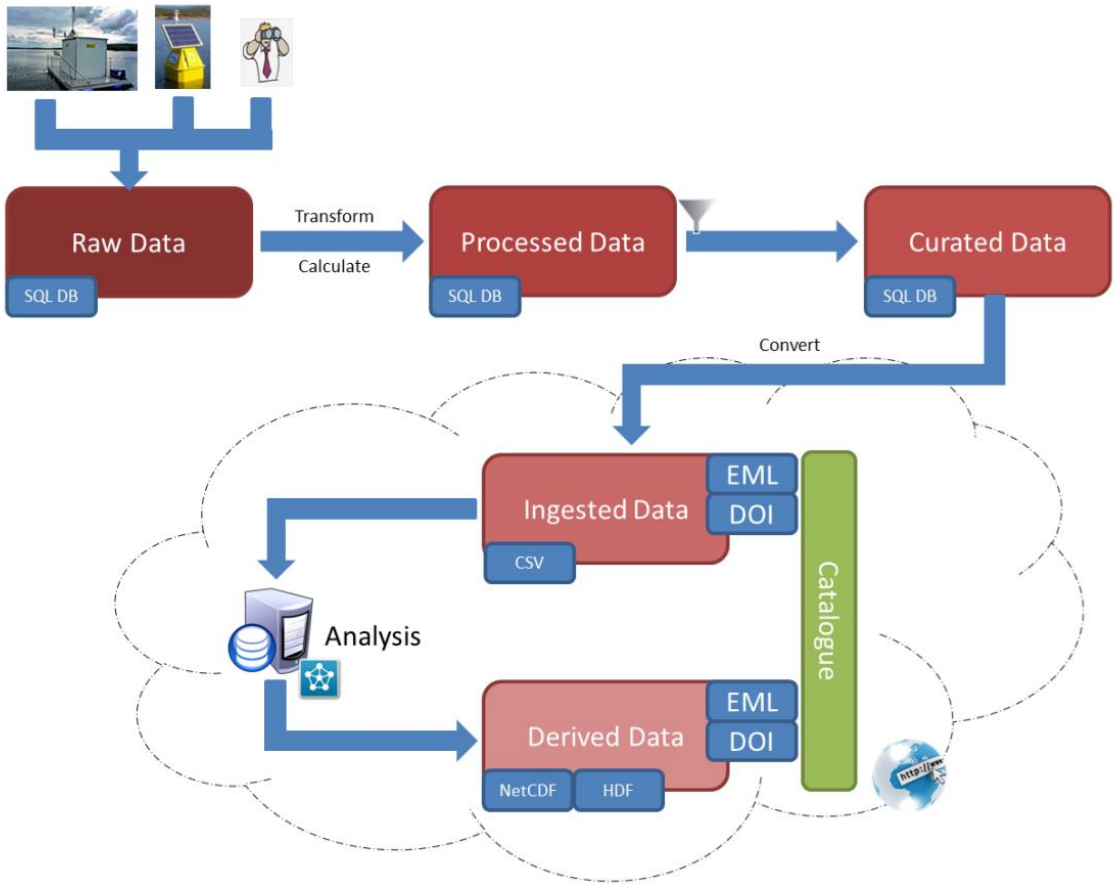
- Podemos definir la curación de datos como un proceso, dentro del ciclo de vida de los datos, enfocado al mantenimiento de los datos. Este proceso se compone a su vez de varios “sub-procesos” para garantizar que los datos están listos para ser utilizados en tareas de soporte a la decisión, minería… es decir, que los datos son correctos, no tienen errores, están completos y actualizados.

- En definitiva, es un proceso dentro del ciclo de vida de los datos que está a su vez presente en todas las fases del mismo, desde que se recolectan los datos iniciales hasta que se almacenan para su futuro uso.

¿Qué es la curación de datos?



¿Qué es la curación de datos?

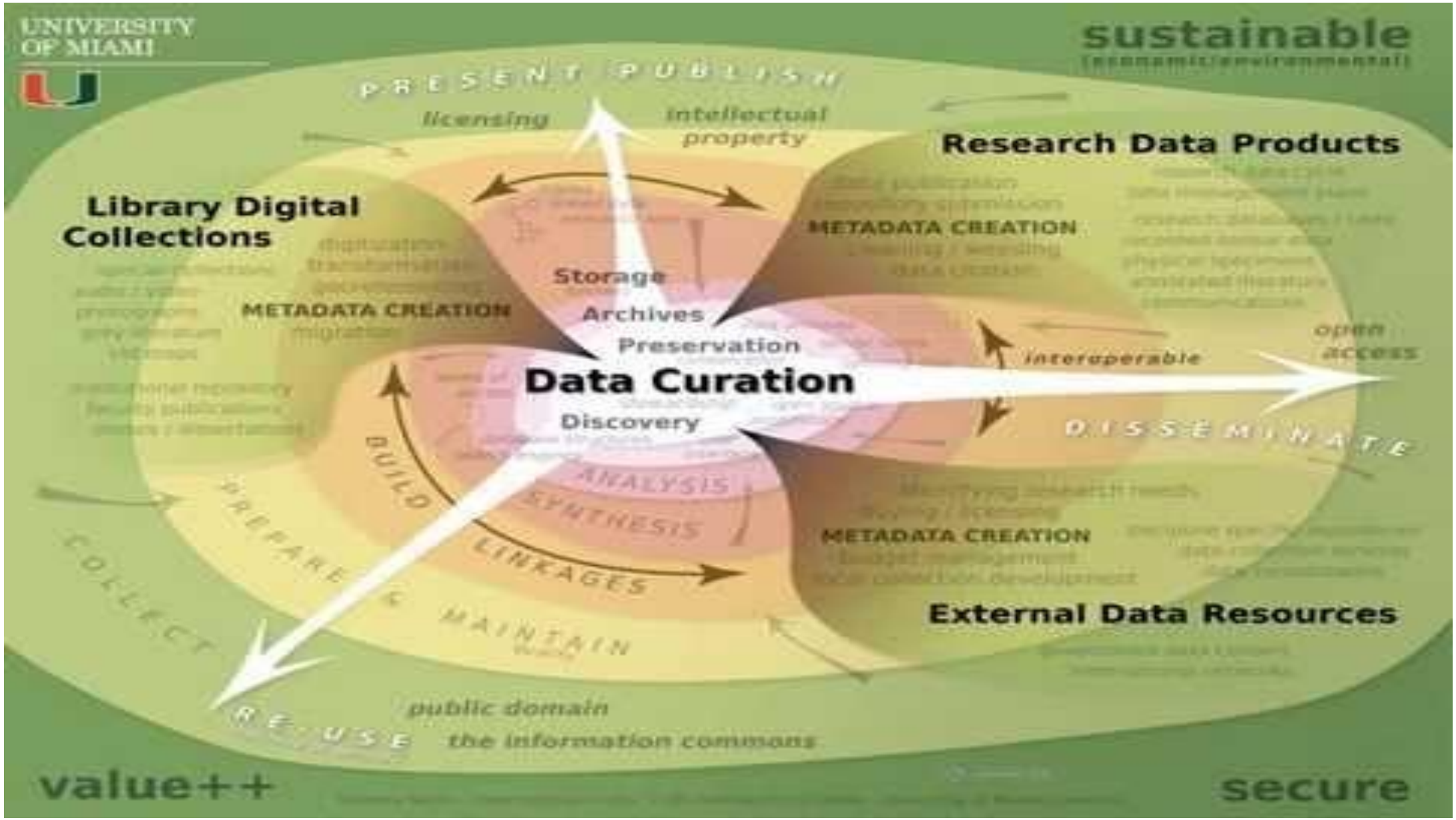


Curación digital (Digital Curation - DCC)

- Termino relacionado con el Data Curation o Curación de los Datos.

- La Curación Digital es un término muy ligado al mundo científicoy bibliográfico (http://www.dcc.ac.uk/digital-curation/what-digital-curation)

- Se refiere a los procesos relativos al mantenimiento y adición de valor sobre contenidos digitales confiables para su uso en el presente y en el futuro. Esto es, la Curación Digital consiste en el mantenimiento activo de los contenidos (texto académicos, artículos, otros) a lo largo del ciclo de vida de los datos.


UNIVERSITY OF MIAMI
sustainable
(economic/environmental)
PRESENT PUBLISH
licensing
intellectual property
Research Data Products
Library Digital Collections
METADATA CREATION
Storage
Archives
Preservation
METADATA CREATION
Data Curation
Discovery
interoperable
open access
DISSEMINATE
BUILD
ANALYSIS
SYNTHESIS
LINKAGES
PREPARE & MAINTAIN
COLLECT
METADATA CREATION
External Data Resources
RE-USE
public domain
the information commons
value++
secure

“Key-values” en la Curación de Datos

- Claves principales de la Curación de Datos:
  - La curación es un proceso que permanece activo en el tiempo (mantenimiento activo de los datos a lo largo del tiempo)
  - Además, la curación también consiste en añadir valor a los datos (por ejemplo, con meta-datos descriptivos)
  - Dentro de los procesos de curación, existen gran cantidad de sub-procesos para garantizar su calidad y confiabilidad (preprocesado y limpieza)

¿Por qué curar los datos?

- La curación de datos es, al fin y al cabo, parte normal de los procesos del ciclo de vida de los datos en área como la ciencia y la investigación, entre otras:
  - Se evita llegar a conclusiones erróneas debido a errores en los datos.
  - Otros pueden (y seguramente quieran) validar, en incluso replicar la investigación y sus resultados y las conclusiones en base a los mismos (análisis de minería de datos, otros).
    - En algunas disciplinas, incluso, los datos son accesibles para los revisores de los artículos e incluso se ofrecen al público general interesado.
  - Los principios de compartición y publicación de datos en abierto están cada vez más extendidos, tanto en las disciplinas científicas como en otros ámbitos (gobierno, IoT, otros).
  - Propicia el cumplimiento de los principios FAIR (FAIR data principles), asegurando además la calidad de los datos.
    - FAIR principles: http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

¿Por qué curar los datos?

- La curación de datos proporciona valor tanto extrínseco como intrínseco:
  - La ciencia no es barata, hay disciplinas con grandes inversiones.
    - El CIS, por ejemplo, tiene una inversión de aproximadamente 8 millones de euros, y se ha visto incrementado en los últimos meses.
  - Los procesos tanto de captura como el análisis de datos pueden ser muy costosos.
  - Hay datos que, si se pierden, son imposibles de recuperar como ,por ejemplo, datos temporales (time-series).
  - Los datos obtenidos por observación (Observational Data) son, por definición, imposibles de replicar.
  - Actualmente, es usual que almacenemos muchos más datos de los que podemos analizar en el momento, lo que hace necesario un buen mantenimiento de los mismos.

¿Por qué curar los datos?

- La curación aumenta el potencial de los datos para crear nuevo conocimiento del ya existente:
  - Los datos son reusables para aplicar diferentes técnicas de análisis, minería, etc.
    - E incluso para objetivos diferentes de los iniciales.
  - Permite combinar conjuntos de datos de forma “innovadora”.
  - Fomenta el concepto de “Science 2.0”, consistente en fomentar la colaboración mediante la compartición de información sobre los experimentos los datos…

¿Por qué curar los datos?

- Actualmente el interés en la curación de datos por parte de numerosos inversores y organismos está creciendo:
  - Algunos de estos organismos tienen políticas de retención de datos que necesitan de la curación de los mismos.
  - Hay una creciente necesidad de dar acceso abierto a los datos financiados por fondos públicos.
  - Existen unos principios y guías que deben cumplirse para dar acceso público a los datos de investigación (https://data.oecd.org/)
  - Cumplimiento de los principios FAIR

¿Por qué curar los datos?

- La curación de datos es indispensable para la correcta gestión de los activos institucionales:
  - Las universidades y otras organizaciones dedicadas a la investigación invierten grandes sumas de dinero en actividades investigadoras.
  - La investigación es su actividad principal.
  - Por ello, la propia curación de datos es un activo importante.
  - Ayuda además a visibilizar la actividad de estas instituciones, y a aumentar su impacto.

Curación de datos y problema de escalabilidad

- En la actualmente llamada “era de los datos”, existe un problema de escalibilidad:
  - e-Science
  - Nuevas generaciones de instrumentos y herramientas
  - Simulaciones por computador complejas
  - Gran cantidad de datos (terabytes, petabytes…)
  - Diferentes disciplinas generando, almacenando y analizando sus propios datos.
- EL problema de escalabilidad es más destacado en las disciplinas relacionadas con el Big Data:
  - Física de Partículas (por ejemplo, el colisionador de Hadrones)
  - Astronomía
  - Biomedicina y genética
  - Análisis en redes sociales.
  - Y muchas otras
- Con tal cantidad de datos, es difícil asegurar la calidad de los mismos. La curación, se hace indispensable.

Curación de datos y problema de complejidad

- Y no solamente existe un problema de escalabilidad, sino también de complejidad.

- Los datos para investigación y análisis son muy diversos:
  - Datos tabulares, temporales, geográficos…
  - Datos “en bruto”, datos derivados…
  - Diferentes formatos, como el relacional, los semiestructurados…
  - Diferentes estándares.

- Estos datos, por tanto, no son homogéneos.

Curación de datos y diversidad de contextos

- Diferentes culturas y costumbres en la investigación en Data Science
  - Las prácticas varían mucho, incluso en las mismas disciplinas. Por ejemplo:
    - Los datos de secuenciación genómica son usualmente abiertos al público
    - Sin embargo, en la proteómica esta práctica no está tan extendida, debido en parte a falta de estándares y a los derechos de explotación.
- Los intereses comerciales también tienen un peso específico en el mantenimiento de los datos.
- Otro factor importante es la interdisciplinariedad de algunas áreas, que combina diferentes prácticas y fuentes de datos.