Los datos en una investigación

Claudio Gutiérrez
DCC, Universidad de Chile

¿Qué es un dato? Un dataset?

Definición **extensional**:

- Un archivo de datos
- Un conjunto (red) de archivos de datos

Definición **intensional**:

- Una URI (o una dirección o una API)
- La salida (el output) de un sensor
- Un concepto: “los jugadores de la *Premier League*”

Ejemplo: El ppto. de Chile del año 2019 es extensional.
“El valor de cambio de pesos por dólares” es intensional.

Dimensiones básicas de datasets

1. **Volumen**: escala
2. **Formatos**: valores, texto, imagen, grafos
3. **Tipo**: estático/aislado dinámico/continuo
4. **Contenido**: naturaleza / humano / organizacional
5. **Temas legales**: tipo propiedad y usos
6. **Temas éticos**: sesgos, privacidad, etc.

1. Volumen

Escala humana: KB, MB, GB

- texto, imagen, video
- procesable (aún) con los sentidos humanos

Escala más allá de lo humano: TB, PB, EX, ZB

- Necesidad de automatización
- Consideraciones de

Escalas

Nombre	***Standard***	***Binary use***
Kilobyte	10 e 3	2 e 10
Megabyte	10 e 6	2 e 20
Gigabyte	10 e 9	2 e 30
Terabyte	10 e 12	2 e 40
Petabyte	10 e 15	2 e 50
Exabyte	10 e 18	2 e 60
Zettabyte	10 e 21	2 e 70

2.a. Formatos clásicos

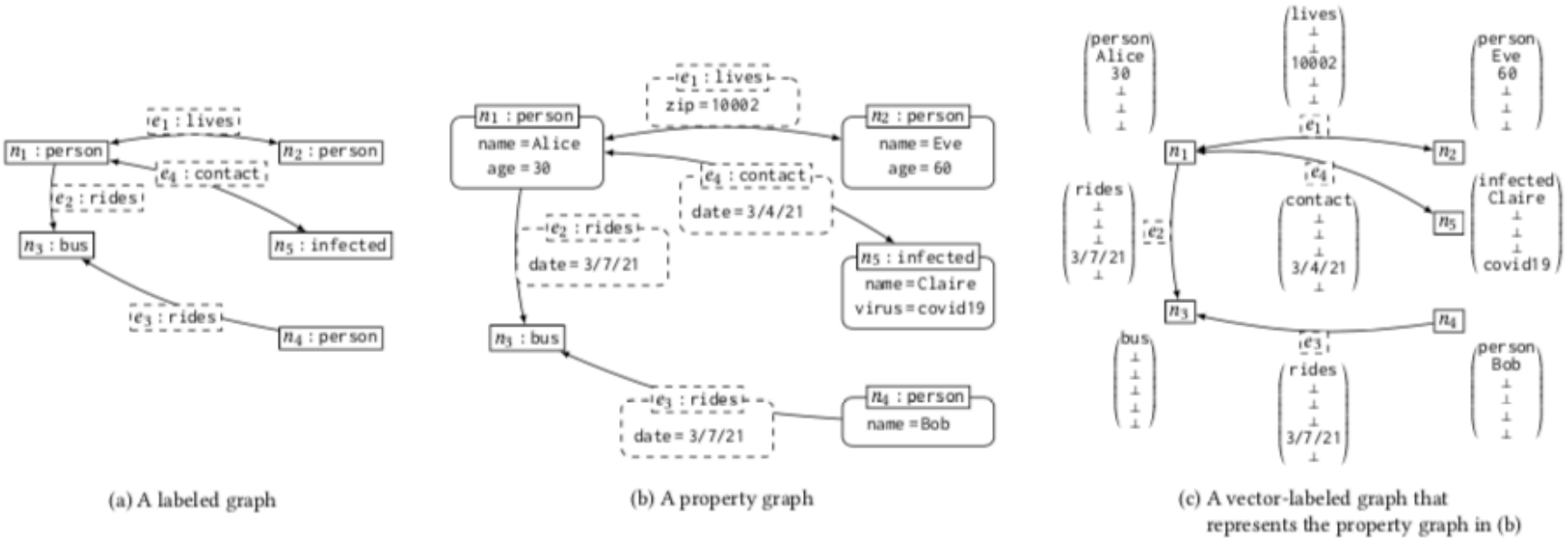
Binario (eficiente, ilegible)
Texto (natural para humanos, ineficiente, poca o nada estructura)

Tablas (buena organización, históricamente popular, eficiente, excelente soporte: RDBMS; versión "reguleque": CSV)

Documentos (natural para humanos, semi – estructurado, formalizado como XML, JSON, et al.)

Grafos (excelente expresividad, difícil –aun- de procesar, poco soporte. Es lo que viene)

Datos en formato grafos



(a) A labeled graph

(b) A property graph

(c) A vector-labeled graph that represents the property graph in (b)

Figure 2: Three graph data models.

el formato vectorial

Representación de objetos y fenómenos como arreglos de números

Espacios vectoriales (reales): Espacios de arreglos de números de diferentes dimensiones

- vectores
- matrices

2.b. Formatos media

Formatos no solo determinan la calidad y el tamaño del archivo, sino también con qué dispositivos de reproducción será compatible.

Audio/sonido: bit depth (profundidad de bits), sample rate

Imagen: con/sin pérdida, vectoriales, mapas de bits

Video: contenedor (medatados, audio) y códecs (codificación, compresión)

- Antiguos, pero a gran escala muy reciente. Parte sustancial de nuevas arquitecturas de conocimiento (Kowledge Graphs).
- Necesidad de dominar herramientas para su procesamiento

3. Tipos

Estático / discreto.

- para informes, etc.

Dinámico / Continuo: Flujos.

- sensores, cámaras vigilancia, tipo de cambio, temperatura,

4. Tres clases fundamentales de datos

Datos Organizacionales/Empresariales: generados por procesos pre-diseñados para ello.

Bases de datos, tablas, archivos, informes, etc.

Datos naturales: generados por la naturaleza

metereológicos, astronómicos, sísmicos, geológicos, de materiales, experimentos físicos, químicos, cámaras de monitereo de la naturaleza, etc.

Datos humanos: generados por la presencia humana

médicos, relojes “inteligentes”, celulares, georeferencia, fotos, videos, cámaras de vigilancia, escritos, música, redes sociales digitales, applicaciones web, etc. etc.

Datos “naturales”

Por largo tiempo los datos más usados.

Ciencia y métodos clásicos se basan en este modelo (ojo!)

Datos astronómicos, satelitales, geológicos, metereológicos, geográficos, flora y fauna, etc.

Legalmente el tema es la apropiación y su uso. Lógica del derecho tradicional funciona aquí con modificaciones menores

Datos humanos y sociales

El gran impulsor del grueso de las aplicaciones más “novedosas” hoy día.

Datos personales, de comportamiento, biomédicos, relaciones sociales, etc.

Estos son el problema fundamental de la ética y del derecho de datos hoy. Los grandes temas aquí son cómo accesarlos, cuánto almacenarlos, cómo procesarlos, qué uso darles. La propiedad es sólo una faceta y no la más relevante.

Datos humanos y sociales(cont.)

Datos explícitos:

Contactos, amig@s, a quien sigues, seguidores, datos de reuniones, agenda, compras, créditos, pagos, datos identificación: email, teléfono, dirección, etc.

Datos implícitos:

Ubicación geográfica, permanencia y movimientos, sitios y páginas visitados, lecturas, acciones realizadas, clicks, intereses, gustos, deseos, etc.

Datos humanos y sociales(cont.)

Datos explícitos:

– tradicionales, larga historia de regulación
– con conciencia de la persona
– con consentimiento (informado) de la persona

Datos implícitos:

– esto es esencialmente lo nuevo
– sin conciencia de la persona
– evita problemas de autorización
– problemas con noción tradicional de consentimiento (informado)
– ventajas: no está “perturbado” por el sujeto; permite recolección automatizada masiva

ETICA

Ética clásica

El grado en que una entidad posea agencia moral determina la responsabilidad de la entidad

Agencia moral:

- *Causalidad*: agente tiene responsabilidad por las consecuencias de sus acciones
- *Conocimiento*: agente tiene responsabilidad si conocía las consecuencias de sus acciones
- *Elección*: agente tiene responsabilidad si tenía la libertad de elegir alternativa que causara menos daño

Datos y tecnología

Hans Jonas: el radio de alcance de nuestras acciones hoy sobrepasa lejos el ámbito de acción de un individuo

- Tiene consecuencias en el futuro
- Tiene consecuencias en otras geografías
- Amplifica las consecuencias tradicionales

Muchas de estas consecuencias son desconocidas o no calculadas al hacer algo

Algunos temas

- Privacidad:
- Privacidad de Grupo:
- Propensidad: predicciones sobre lo que la gente podría hacer
- Investigación

Cuestionario

Principles and Guidelines for Companies, Authorities & Organisations
DataEthics.eu 1. Edition 2018

Sigue requerimientos de GDPR (General Data Protection Regulation)

El humano al centro

1. ¿Dependen sus datos de que los toma/captura de usuarios no consultados?
2. ¿Se privilegia a los usuarios por sobre intereses comerciales o institucionales?
3. ¿Se asegura que los usuarios más que la institución son los beneficiarios de los datos?
4. ¿Usa principios de privacidad-por-diseño y puede describirlos clara y transparentemente?

Control sobre datos individuales

1. Procesamiento en el lugar:
   - ¿Se asegura que, en la medida en que es posible, los datos son procesados en los propios aparatos de los usuarios?
   - ¿En caso contrario, se asegura que los datos recolectados no identificarán a una persona?
2. Perfilamiento:
   - ¿Usa perfilamiento? ¿Permite que el usuario incida en los valores, reglas, y la entrada que subyace al perfilamiento?
3. Predicción:
   - ¿Usa datos para predecir comportamiento al nivel individual?

Transparencia

1. Almacenamiento
   - En qué país se almacenan sus datos?
   - Dónde se ubica su proveedor de almacenamiento?
   - La transmisión de datos va por fuera de la UE?
2. Inteligencia Artificial
   - Usa AI? En caso positivo, puede explicar los criterios algorítmicos y parámetros usados?
3. Diseño de comportamiento
   - Usa datos personales para influenciar comportamiento?
   - Es transparente cuándo se usa de esa manera?
   - Se asegura que el diseño no crea adicciones y así influencia la autodeterminación de las personas?
4. Fuente abierta
   - Usa software abierto, de tal forma que otros pueden desarrollarlo más?

Accountability

1. Anonimidad
   - ¿Cuándo anonimiza los datos?
   - ¿Usa encriptación de punto a punto?
   - ¿Minimiza el uso de metadatos y lo explica?
2. Cero-conocimiento
   - ¿Usa cero-conocimiento como principio de diseño?
3. Venta de datos
   - ¿Vende datos a terceros?
   - ¿Vende datos como datos personales identificables?
   - ¿Vende datos como patrones sobre datos agregados?
   - Si vende datos, ¿se asegura que es información completamente anonimizada y que sólo describe patrones, no individuos?

Accountability (cont.)

4. Compartir datos
   - ¿Usa cookies de terceras partes?
   - ¿Ello incluye Social Media cookies y logins?
   - ¿Usa Google Analytics y trackers similares?
   - Si usa cookies de terceras partes, ¿está seguro que los usuarios saben que su uso de cookies lleva a compartir datos de sus usuarios con terceros?
5. Enriquecimiento de datos
   - Enriquece sus datos con datos externos (SM, scraping, comprados)
   - ¿Este enriquecimiento ocurre en respuesta o cooperación con sus usuarios?
6. Organización
   - ¿Tiene una unidad o depto. responsable del manejo ético de datos?
   - ¿Cómo se maneja la ética de datos en la organización?
   - ¿Cómo se asegura que sus lineamientos de ética de datos sean respetados?
7. Control externo
   - ¿Puede su manejo de datos ser auditado independiente y externamente?
   - ¿Exige y controla la ética de datos de sus subcontratos y partners?

Igualdad

1. Plataformas públicas
   - ¿Tiene diálogos con sus usuarios en plataformas públicas?
   - ¿Tiene lineamientos para el uso de esas plataformas?
   - ¿Modera la plataforma para remover datos personales sensibles?
   - Si sus servicios son ofrecidos a niños, ¿se asegura el consentimiento de los padres?
2. Reuso de datos
   - ¿Se usan los datos para desarrollar o entrenar un algoritmo?
   - ¿Se asegura que el uso de los datos no lleva a discriminación?
   - ¿Se asegura que el uso de datos no expone vulnerabilidades de individuos?
3. Inteligencia Artificial
   - ¿Se asegura que el uso de AI/ML es para el beneficio de los individuos y no causa daño físico, psicológico, social o financiero al individuo?

Anexos

Caso de estudio 1

Detección de transacciones fraudulentas en un banco

- Se tiene datos de tx sospechosas y fraudulentas investigadas en el pasado
- Se tiene sistema basado en reglas que emite sospechas y son estudiadas manualmente
- Usando los resultados de esas investigaciones se quiere implementar un clasificador de riesgos (predecir tx. sospechosas)

Caso de estudio 2

Análisis de las mermas de supermercados

- Se usan datos de los registros de mermas de los supermercados.
- Se enfoca en productos perecibles (e.g. lácteos) para identificar causas
- Objetivo: hacer un predictor de mermas

Caso de estudio 3

Análisis de sensores de una turbina generadora de electricidad para anticipar posibles fallos

- Se usan datos de lecturas de aprox. 30 sensores de temperatura, vibración y revoluciones de la turbina
- Se tiene etiquetado períodos donde la turbina funcionaba bien y mal
- Objetivo: predecir nuevos fallos de la turbina

Caso de estudio 4

Análisis del mercado de jugadores de fútbol

- Se tienen datos de las ligas de fútbol europea,
- Datos de jugadores, sus valoraciones y formas de juego
- Interesa analizar el comportamiento del precio de jugadores respecto de sus atributos técnicos
- Objetivo: predecir los jugadores jóvenes que aumentarán su valor

Caso de estudio 5

Análisis de la deserción de estudiantes de primer año de una Universidad

- Se tienen datos de estudiantes de primer año de los últimos 3 años
- Se quiere analizar causas de la deserción en primer año
- Objetivo: hacer un predictor de estudiantes que desertan

Caso de estudio 6

Análisis del crecimiento de árboles cítricos

- Se tiene los datos de una plantación de naranjos de la cuarta región durante 2 años
- Se registran diariamente datos atmosféricos y de suelo
- Se tiene registros mensuales de crecimiento y estado de las plantas
- Objetivo: analizar formas de riego y necesidades de agua para el crecimiento de las plantas