INTRODUCCIÓN A LA CIENCIA DE DATOS

Juan Fernando Pérez

Departamento MACC

Universidad del Rosario

Universidad del Rosario

MACC

Matemáticas Aplicadas y Ciencias de la Computación

Agenda

- Bienvenida y Presentación
- El Equipo
- Introducción a la Ciencia de Datos
- El plan de trabajo

El Equipo

El Equipo

Andrés García Suaza

Profesor principal de la Facultad de Economía.

Doctor en Economía de la Universidad Carlos III de Madrid, España.

Experiencia en investigación y consultoría en temas de mercado laboral, estadística aplicada y econometría

El Equipo

Dora Suárez

Líder de Proyectos del Hub de Innovación y Transferencia – HINNT del Departamento de Matemáticas Aplicadas y Ciencias de la Computación (MACC) de la Universidad del Rosario.

Estadística de la Universidad Nacional, Colombia, y Maestría en Estadística de la Universidad Federal de Pernambuco, Brasil.

El Equipo

Juan Fernando Pérez

Profesor Principal del Departamento de Matemáticas Aplicadas y Ciencias de la Computación.

Doctor en Ciencias de la Computación de la University of Antwerp, Bélgica.

Experiencia como investigador en ciencias de la computación en Imperial College London, Reino Unido, y The University of Melbourne, Australia

Ciencia de Datos

Ciencia de Datos

¿Qué es la ciencia de datos?

Comprende la intersección de un número de disciplinas

- Estadística
- Minería de Datos
- Aprendizaje de Máquina (Machine Learning)
- Bases de Datos
- Big Data
- Analítica de Datos
- Inteligencia de Negocios
- **Matemáticas Aplicadas y Ciencias de las Computación**

Estadística

Recolección, interpretación, análisis, presentación y organización de datos

Población a estudiar

Toma de muestras para inferir rasgos de la población a partir de la muestra

Diseño de la muestra (representatividad, aleatoriedad)

Estadística (cont.)

Análisis e interpretación:

- Análisis descriptivo
- Estimación de ciertas características (e.g., proporciones)
- Pruebas hipótesis
- Construcción de modelos (e.g., regresión)

Estadística (cont.)

Modelos estadísticos para describir fenómenos

Capturar relaciones entre variables

Derivar Relaciones desde Principios Básicos

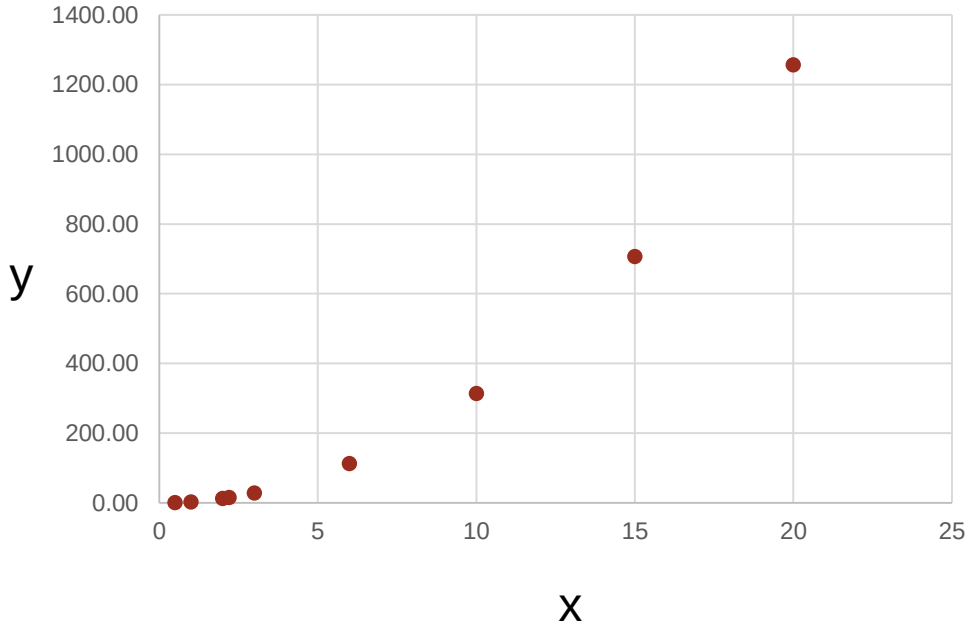
Derivar Relaciones a partir de Datos

Data-driven Science

Ejemplo de juguete

Datos:

x	y
1	3.14
0.5	0.79
3	28.27
6	113.10
2.2	15.21
2	12.57
10	314.16
15	706.86
20	1256.64



$$y = \pi x^2$$

Estadística y Modelos

Modelos estadísticos para describir fenómenos

Capturar relaciones entre variables, e.g.,

$$y = \alpha x + \beta x^2 + \gamma x^3 + \varepsilon$$

Modelos: abstracciones/representaciones/aproximaciones de/a la realidad

Suficientemente preciso pero relativamente sencillo

Intermezzo - Regresión

https://www.gsma.com/betterfuture/resources/telefonica-case-study-predicting-air-pollution-levels-24-to-48-hours-in-advance-in-sao-paulo-brazil

Intermezzo - Regresión

Ciudad con problemas de polución

Pocas estaciones de medición

Muchas estaciones de celular

¿Cómo estimar/predecir los niveles de polución en un día y zona específica?

Intermezzo - Regresión

Datos activos de celular (x_1)

+

Datos pasivos de celular (x_2)

+

Datos del tiempo (x_3)

=

Estimación de la polución (y)

Desarrollado por

Telefónica en Sao Paulo

$$y = \alpha\, x_1 + \beta\, x_2 + \gamma\, x_3 + \varepsilon$$

Minería de Datos

Procesar y Analizar datos para extraer nueva información útil

Descubrir patrones al explorar grandes cantidades de datos

Muchas veces incluye labores de extracción, pre-procesamiento, almacenamiento, análisis y visualización en grandes bases de datos

Intersección de estadística, aprendizaje de máquina y bases de datos

Minería de Datos (cont.)

Proceso de descubrimiento de conocimiento en bases de datos (Knowledge Discovery in Databases - KDD).

- Selección
- Pre-procesamiento
- Transformación
- Minería de datos
- Interpretación y evaluación

Minería de Datos (cont.)

Tareas de minería de datos:

- Detección de anomalías (e.g., tiempos excesivamente largos de proceso)
- Reglas de asociación: descubrir asociaciones entre variables (e.g., hábitos de compra asociados a ciertas características de los clientes)
- Clustering (agrupamiento): descubrir grupos de datos con características similares (e.g., muchas de las compras en una tienda son unitarias y menores a 20mil pesos)
- Clasificación: determinar si nuevos datos pertenecen a una de varias categorías (e.g., se tiene una enfermedad o no dados unos síntomas)

Aprendizaje de Máquina

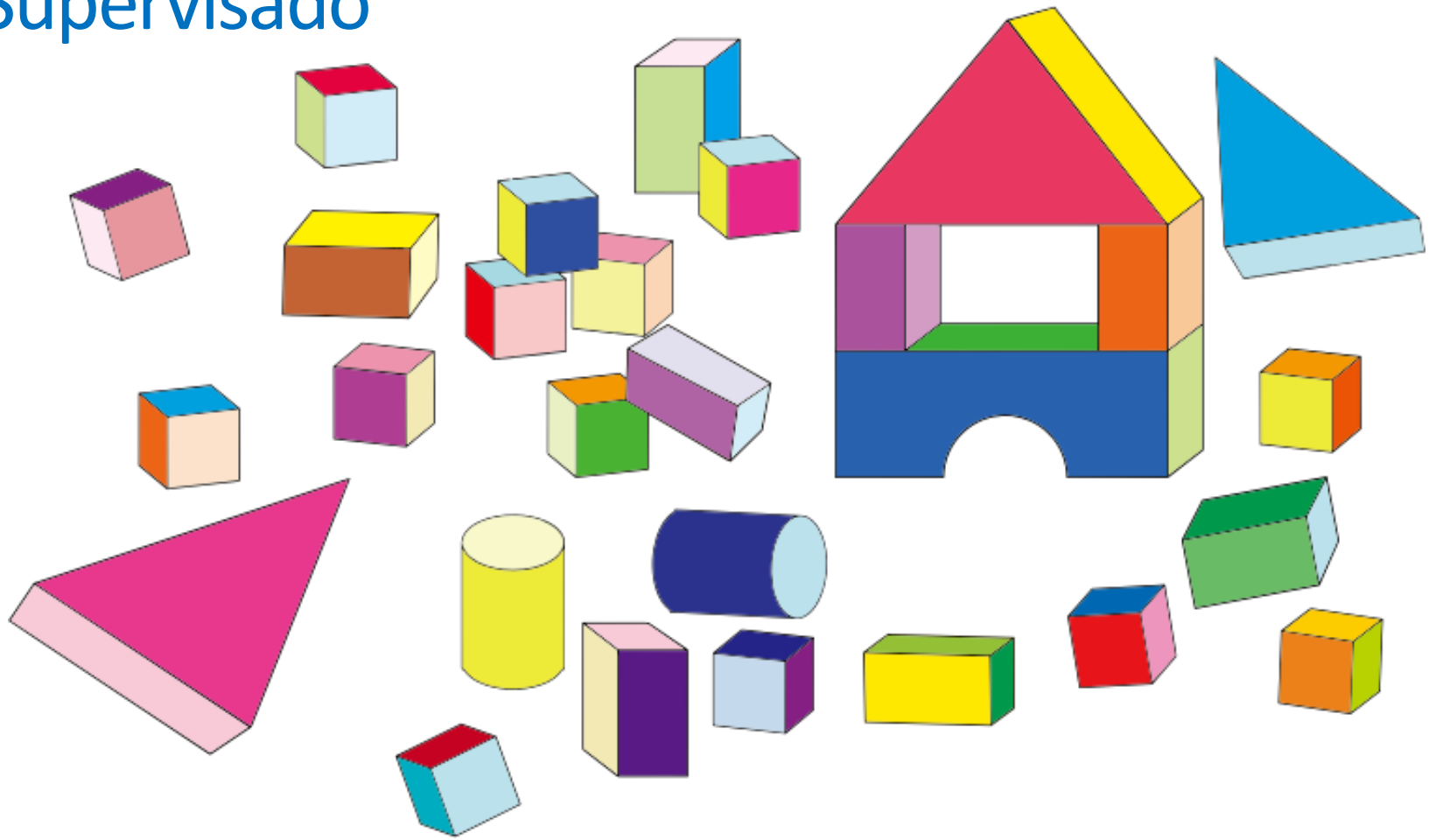
Diseñar e implementar métodos que le permitan a una máquina aprender a realizar tareas que no le fueron programadas explícitamente

Reconocimiento de patrones

Múltiples métodos, algunos relacionados con minería de datos

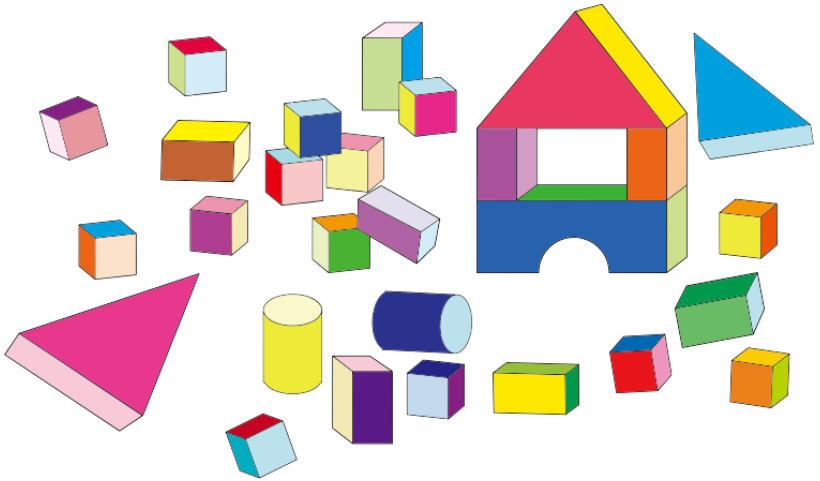
Clustering, clasificación, reglas de asociación, redes neuronales, support vector machines

Intermezzo – Aprendizaje Supervisado vs No Supervisado



https://openclipart.org/detail/254023/building-block-toys

Intermezzo – Aprendizaje Supervisado vs No Supervisado



Supervisado:

- “Esto es un cubo”
- “Esto es un cilindro”
- “¿Qué es esto?”

No Supervisado:

- Agrupamiento sin direccionamiento

Aprendizaje Supervisado

Aprender a partir de información previamente etiquetada

Ejemplo: datos (descriptores, imágenes) de personas con ciertas características y etiqueta del grupo al que pertenece (niño vs adulto)

Proceso de entrenamiento

Ante nueva información el computador es capaz de decidir qué etiqueta asignar a un nuevo dato

Aprendizaje Supervisado

Etiquetas:

- Valores continuos: problema de regresión
  - Ej. Predecir la polución (concentración de partículas PM2.5) a partir de información de tráfico, tiempo, etc.
- Valores categóricos: problema de clasificación
  - Ej. Predecir si un paciente tiene o no una enfermedad a partir de sus síntomas

Aprendizaje No Supervisado

Aprender a partir de información sin etiquetar

Ejemplo: datos (descriptores, imágenes) de personas con ciertas características. Agrupar similares.

Ante nueva información el computador es capaz de decidir qué etiqueta asignar a un nuevo dato

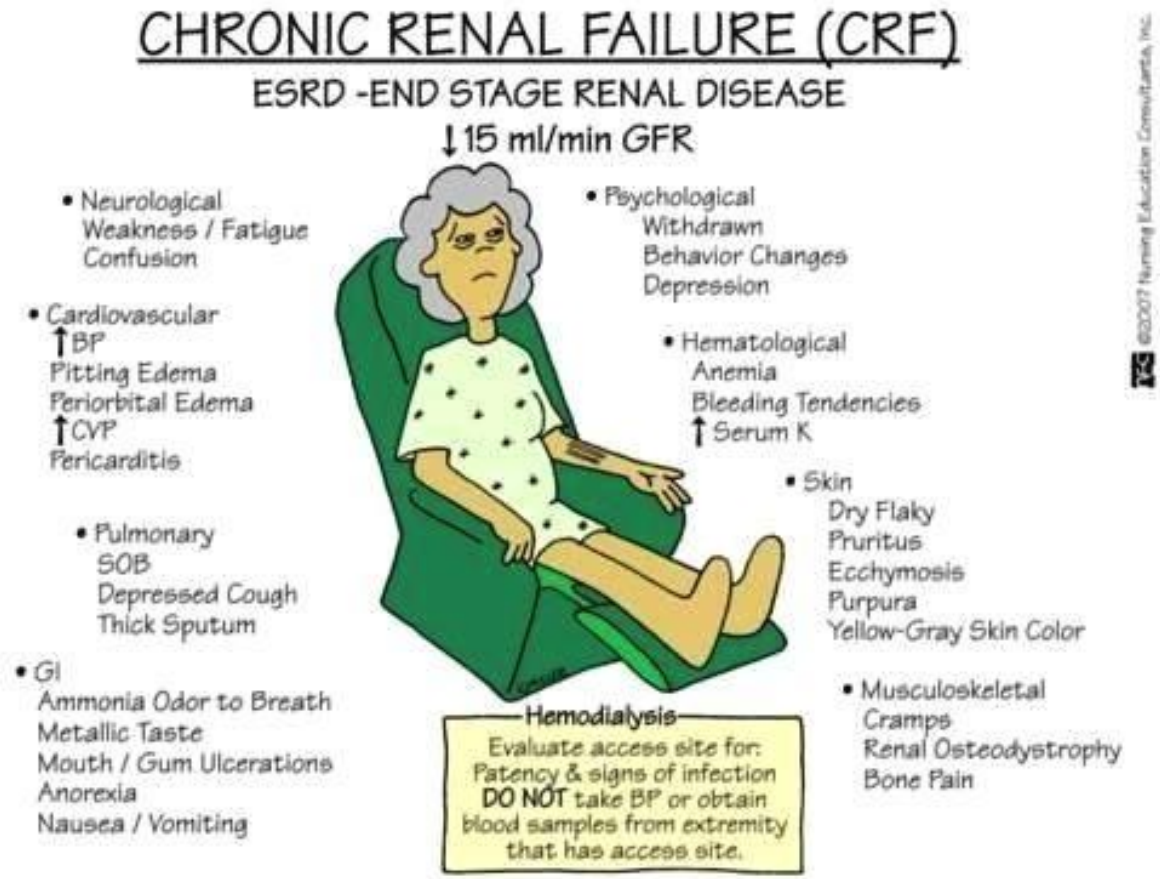
Aprendizaje No Supervisado

Clustering (agrupamiento):

- Identificar grupos de observaciones similares
- Ej. Identificar grupos similares de pacientes
- Ej. Identificar grupos de colegios con recursos y poblaciones similares

- Descubrir patrones no evidentes a priori

Intermezzo - Clustering



https://ar.pinterest.com/pin/864409722198329855

Intermezzo - Clustering

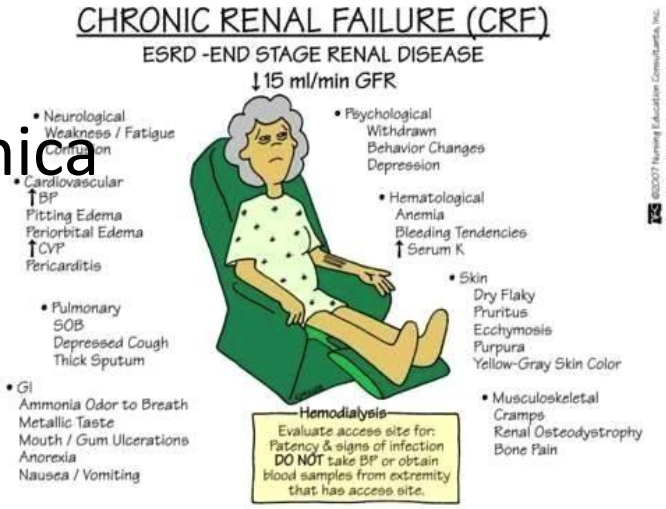
Pacientes con enfermedad renal crónica

Hemodiálisis (HD)

Identificar pacientes similares (clusters)

Comportamiento: antes de iniciar HD, después de iniciar HD

Otros factores: demográficos, comorbilidades



Intermezzo - Clustering

BMC Nephrol. 2016; 17: 25.
Published online 2016 Mar 2. doi: 10.1186/s12882-016-0238-2
PMCID: PMC4776444
PMID: 26936756

Cluster analysis and its application to healthcare claims data: a study of end-stage renal disease patients who initiated hemodialysis

Minlei Liao, Yunfeng Li,✉ Farid Kianifard, Engels Obi, and Stephen Arcona

Author information ▶ Article notes ▶ Copyright and License information ▶ Disclaimer

Identifican grupos de pacientes de altísimo costo con comorbilidades

Sugieren atención de comorbilidades en etapa temprana de HD

Estadística, Minería de Datos y Aprendizaje de Máquina

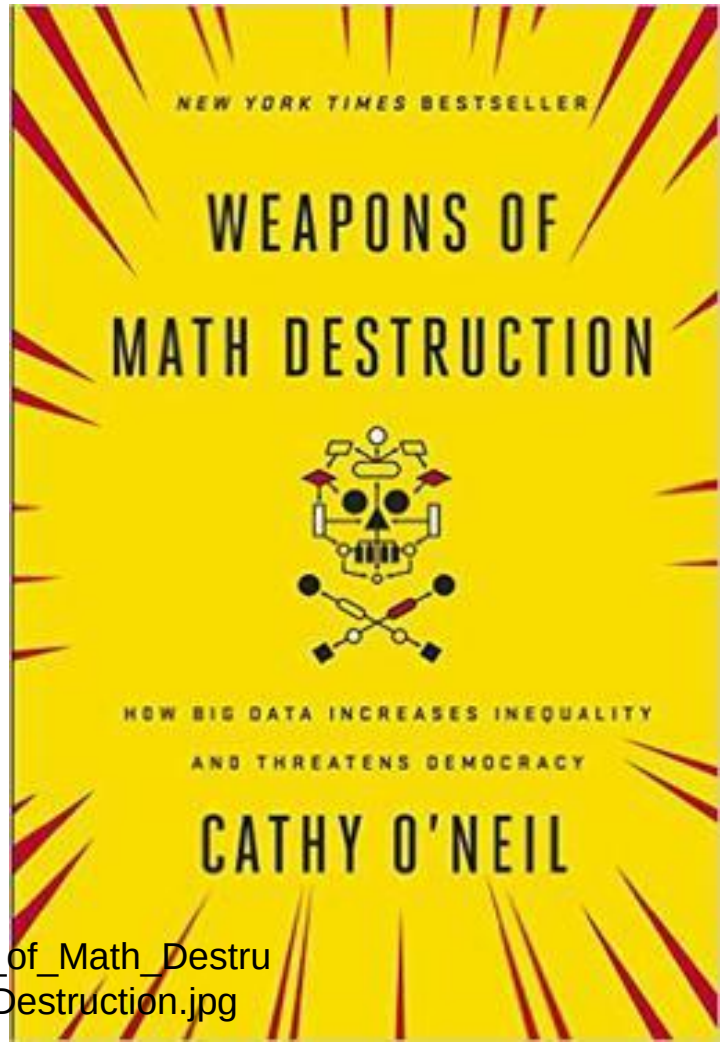
Disciplinas muy cercanas y con metodologías comunes

Extracción de información de grandes cantidades de dados

Modelos matemáticos y computacionales

Algoritmos

Intermezzo - Alerta



https://en.wikipedia.org/wiki/Weapons_of_Math_Destruction#/media/File:Weapons_of_Math_Destruction.jpg

Intermezzo - Alerta

- Todo modelo de ML es entrenado sobre datos existentes
- La selección de las variables a usar determina qué tiene el cuenta el modelo para generar la predicción/clasificación/agrupamiento
- Puede hacer que sesgos previos se mantengan: racismo, discriminación social, etc
- **Abrir las cajas negras** (algoritmos de ML)

Ciencia de Datos

¿Qué es la ciencia de datos?

Comprende la intersección de un número de disciplinas

- Estadística
- Minería de Datos
- Aprendizaje de Máquina (Machine Learning)
- Bases de Datos
- Big Data
- Analítica de Datos
- Inteligencia de Negocios
- **Matemáticas Aplicadas y Ciencias de las Computación**

Big Data

Grandes cantidades de datos disponibles

- Universo digital
  - 2008: ~1 zettabytes
  - 2013: 4.4 zettabytes
  - 2017: 15 zettabytes
  - 2020: 44 zettabytes

Big Data

¿Qué es un zettabyte?

- 1,000 Exabytes
- 1,000,000 Petabytes
- 1,000,000,000 Terabytes
- 1,000,000,000,000 Gigabytes

Volumen

Big Data

¿Cómo se genera toda esta información?

- Número de dispositivos conectados a internet:
  - 2013: 13 mil millones
  - 2020: 50 mil millon

Big Data

¿Qué hacemos con estos dispositivos?

- Cada minuto
  - Enviamos 204 millones de correos
  - Damos 1.8 millones de likes en Facebook
  - Subimos 200,000 fotos a Facebook
  - Subimos 400 horas de video a You Tube (65 años de video en un día)

Variedad

Velocidad

Big Data

¿Quién genera esta información?

- Todos nosotros
- Voluntaria e involuntar

Veracidad

Big Data

Valor

Big Data

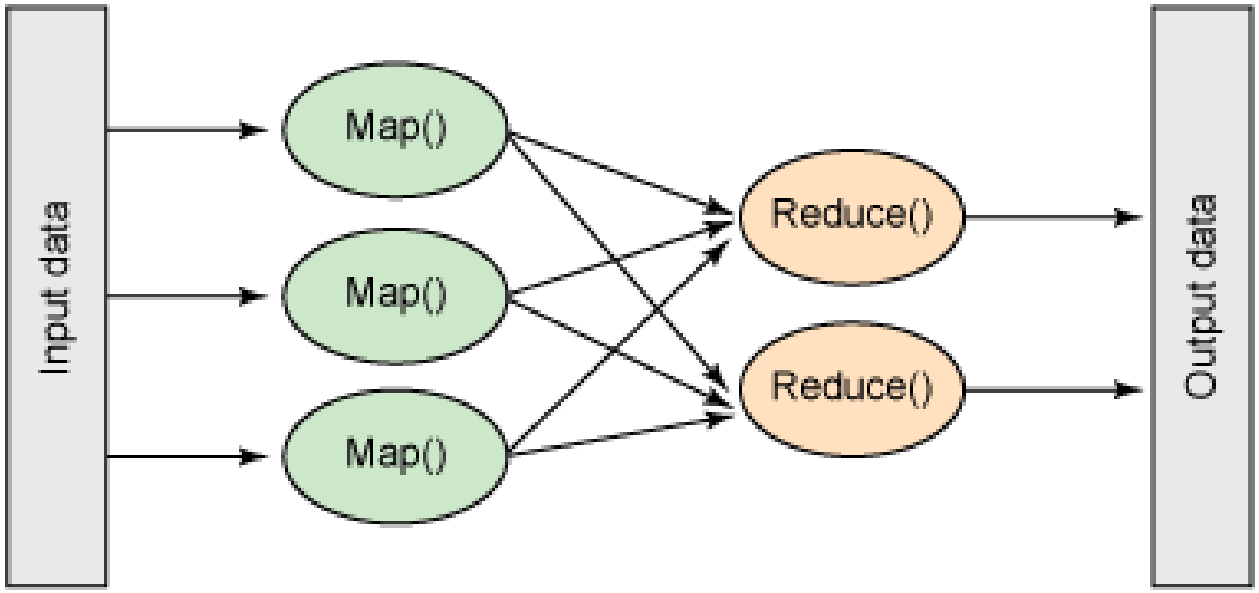
¿Por qué ahora?

- Gran disponilidad de información, bla, bla, bla,…
- Tecnología (hardware):
  - Capacidad de cómputo
  - Capacidad de almacenamiento
  - Capacidad de transmisión de información
  - Moore’s law
  - Redes y dispositivos inalámbricas
  - Transacciones
- Tecnología (Software y Algoritmos)

Big Data

Software y Algoritmos

- Map Reduce (2004, Google paper)
- Modelo de procesamiento de datos en paralelo



Big Data

¿Po

Núr

- 19
- 19
- 19
- 2005: 64,780,610 (YouTube)
- 2010: 206,956,763 (Pinterest)
- Hoy: > 1,259,155,000

Big Data

Map Reduce

- Ejemplo: contar palabras
- Contar cuántas veces aparece cada palabra en millones de documentos
- Cada nodo toma unos cuantos documentos y cuenta las palabras en ellos (Map)
- Cada nodo saca el total para un grupo de palabras (Reduce)

Big Data

De MapReduce a Hadoop y más allá

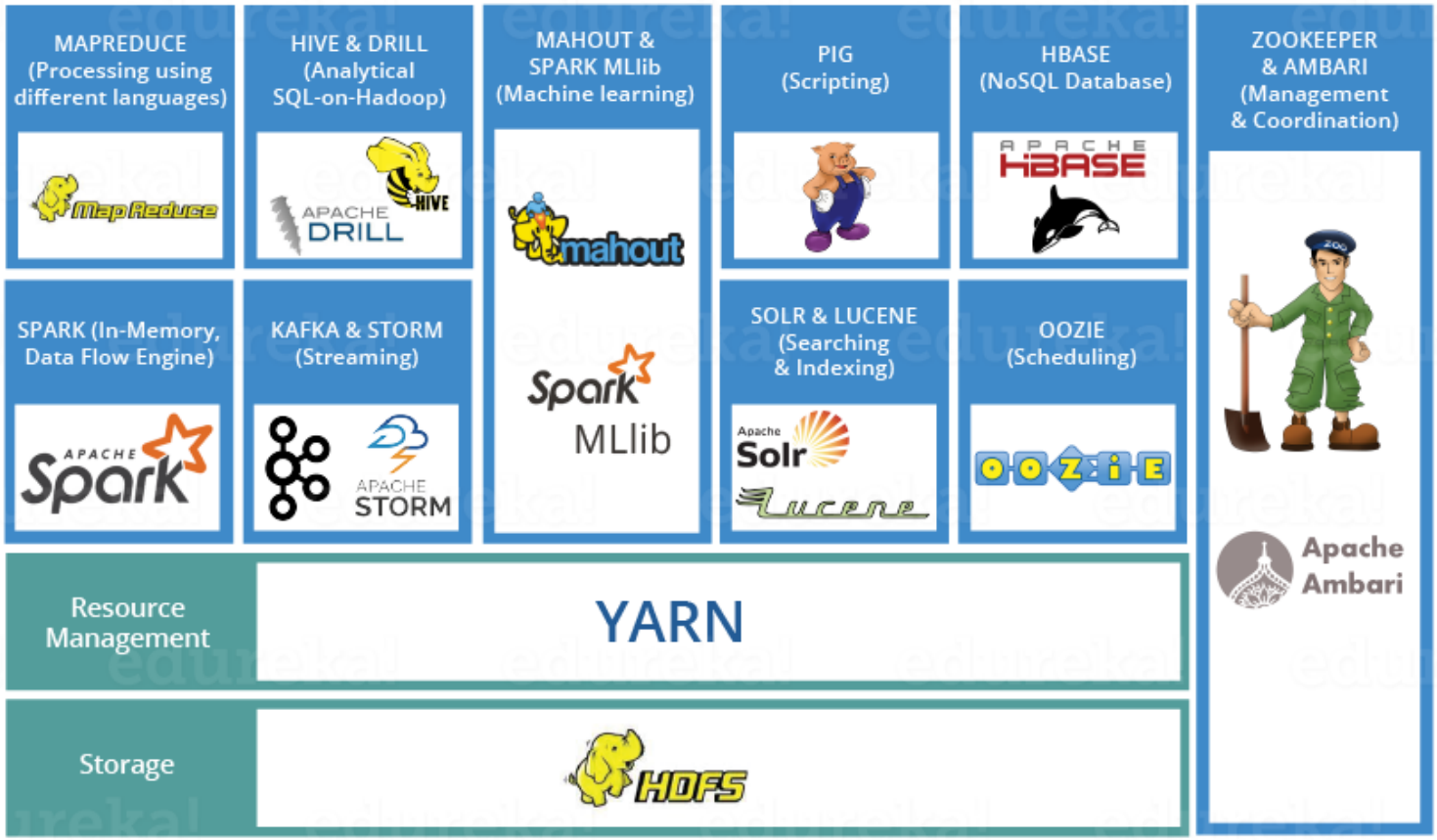
- Apache Hadoop 0.1.0 (Abril 2006)

- Yahoo corre cluster Hadoop con 1000 máquinas (Octubre 2006)

-

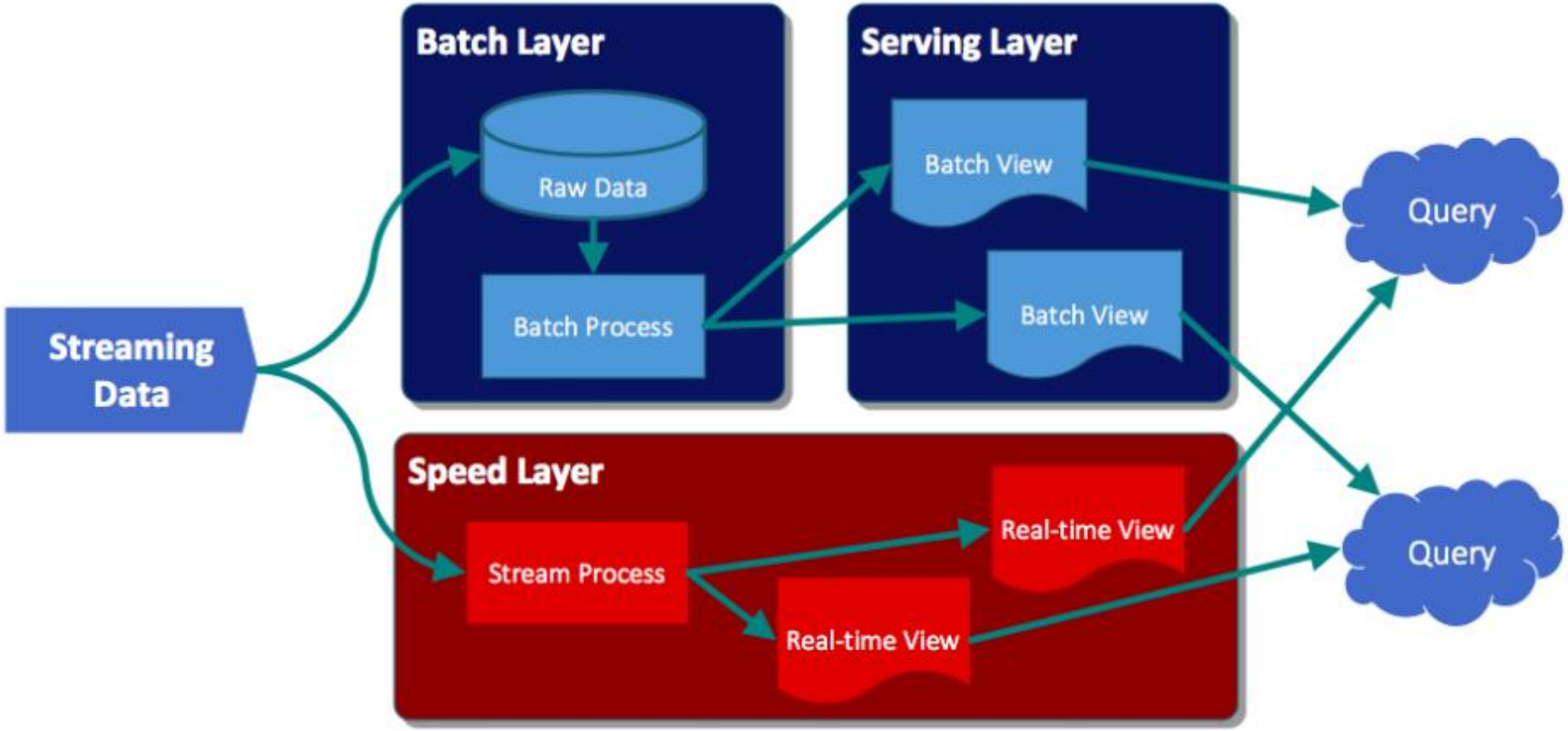
Big Data

El ecosistema Hadoop



Big Data

Arquitectura Lambda

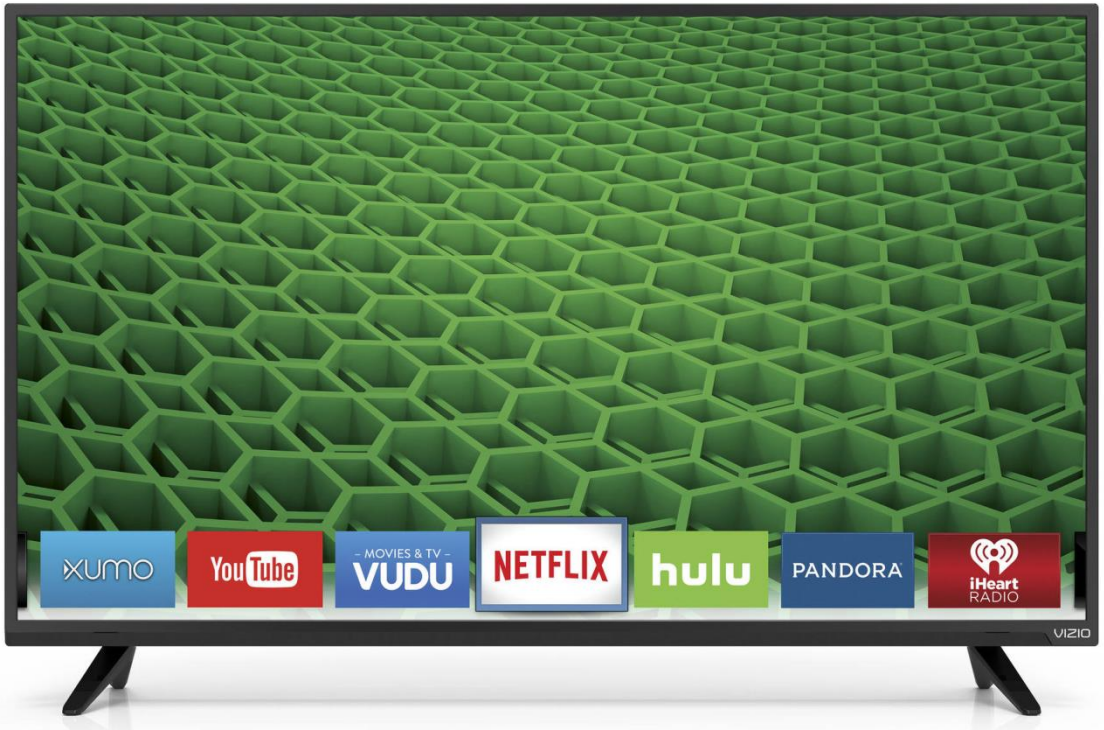


Internet de las Cosas

Teléfonos celulares

Internet de las Cosas

Televisores inteligentes



Internet de las Cosas

Asistentes virtuales para la casa

Internet de las Cosas

Parlantes

Internet de las Cosas

Monitores de actividad física

Internet de las Cosas

Internet de las Cosas

Dispositivos

- Conectados a internet
- Sensores
- Recolectan información
- Actuadores
- Inteligencia Artificial
- Procesamiento local y remoto

Internet de las Cosas

Dispositivos

- Capacidades limitadas de cómputo
- Consumo de energía limitado (baterías)
- Almacenamiento limitado

Procesamiento remoto

- Uso de capacidades ilimitadas en grandes centros de datos (nube)
- Emplear datos de muchas fuentes

Big Data

Ciencia de Datos

Otros nombres

- Ambientes empresariales (no necesariamente tecnológicos): Inteligencia de negocios
- Analítica de datos
- Big Data
- Muchos nombres que en muchos casos se refieren a la misma idea general con diferencias de campos de estudio y aplicación

Componente Tecnológico

No son solo métodos abstractos

Requerimos implementaciones

Efectivas y Eficientes

Herramientas computacionales

Herramientas Computacionales

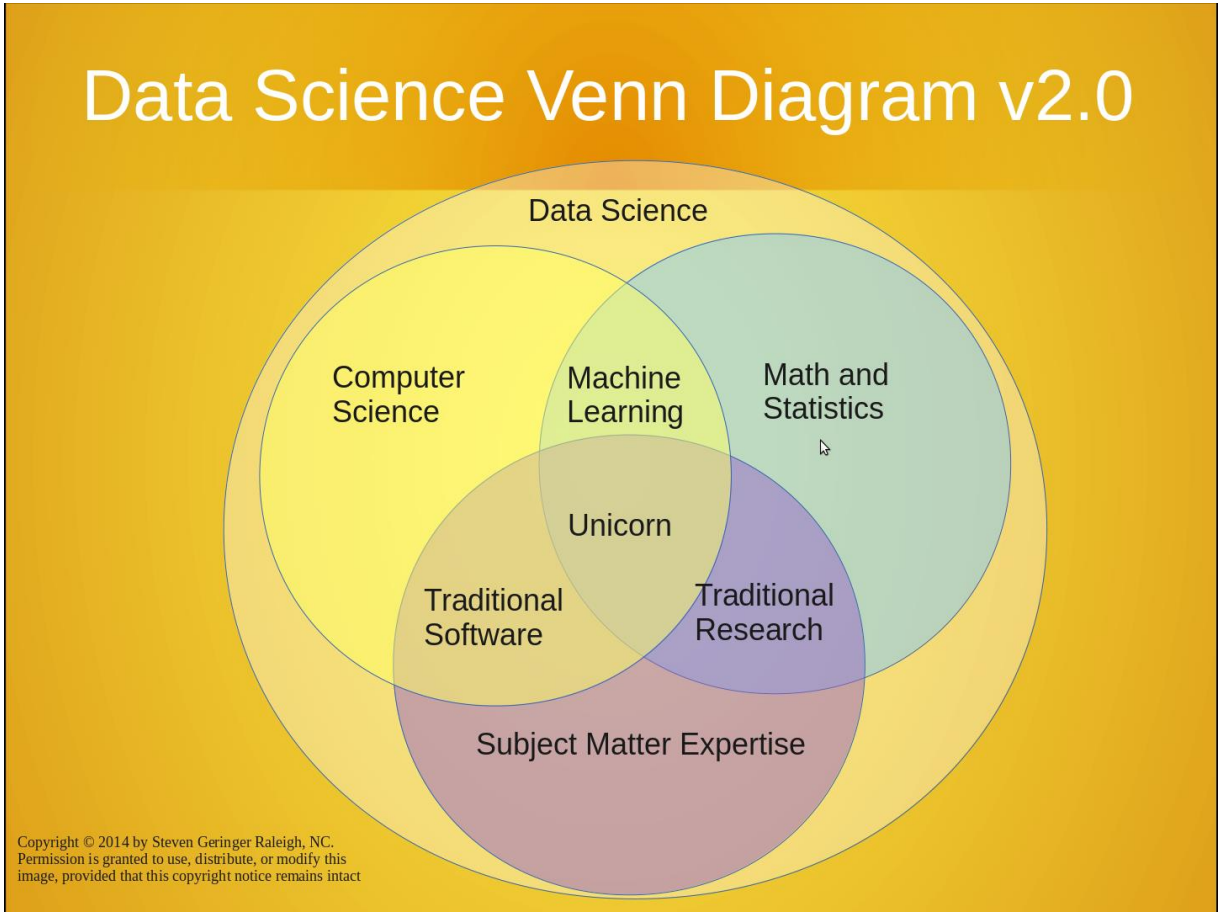
Herramientas de alto nivel

- No requieren programar
- Limitaciones sobre metodologías a usar y tratamiento de datos

Lenguajes de Programación:

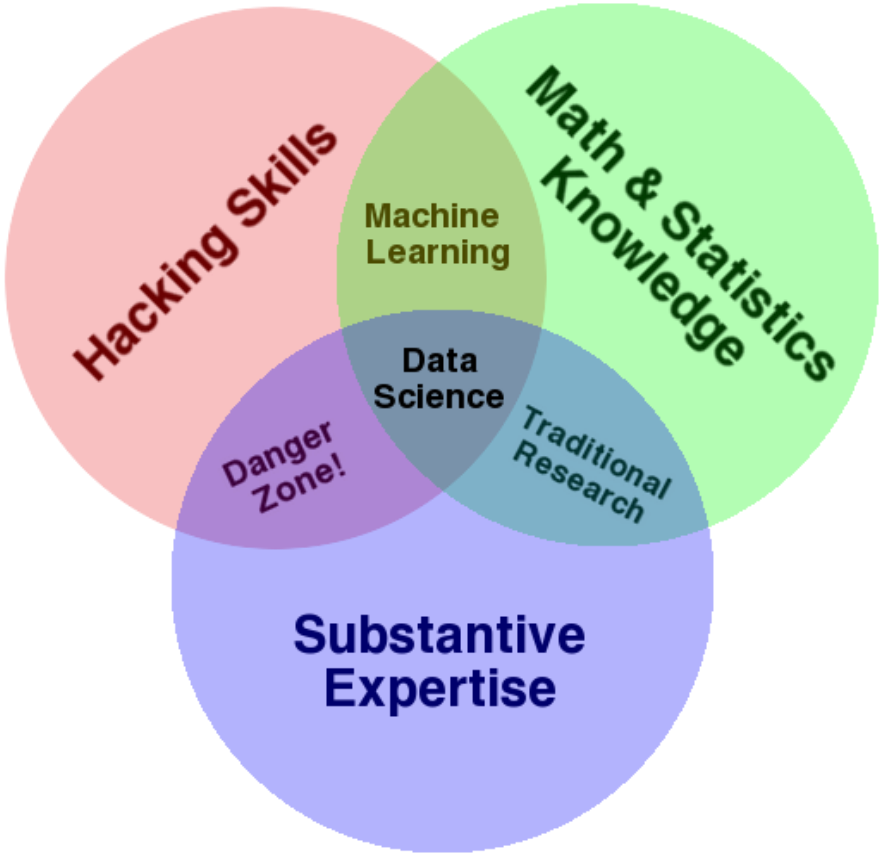
- R: estadística
- Python: aprendizaje de máquina

El/La Científico/a de Datos



Tomado de
https://www.kdnuggets.com/2016/10/battle-data-science-venn-diagrams.html

El/La Científico/a de Datos v2



Hoja de Ruta

Hoja de Ruta

M1: Introducción a la Ciencia de Datos

M2: Introducción al Análisis de Datos y R

M3: Python

M4: Map-Reduce y Hadoop

Hoja de Ruta (cont.)

M5: Aprendizaje supervisado

M6: Métodos de Regresión

M7: Aprendizaje no Supervisado

M8: Proyecto Ciencia de Datos

Hoja de Ruta

M1: Introducción a la Ciencia de Datos

M3: Python

M2: Introducción al Análisis de Datos y R

M4: Map-Reduce y Hadoop

M5: Aprendizaje supervisado

M6: Métodos de Regresión

M7: Aprendizaje no Supervisado

M8: Proyecto Ciencia de Datos

Gracias

@MACC_URosario

@macc_ur

www.urosario.edu.co/Departamento-MACC

@MACC.URosario