UNIVERSIDAD AUSTRAL | INGENIERÍA

Data ¿qué?

Mg. Lic. Gaston Pezzuchi, MSc.

2020

- Para debatir, tomaremos algunas ideas de:

Van der Aalst, Wil (2017). *Process Mining. Data Science in Action (2nd Edition).* Springer-Verlag

Data Science

- En los últimos años la ciencia de datos (*data science*) ha emergido como una nueva e importante disciplina:
  - Puede ser considerada como una amalgama de disciplinas clásicas:
    - Estadística
    - *Data mining*
    - DBMS
    - Sistemas Distribuidos
- Convertir datos en valor para individuos, organizaciones y la sociedad.
- Nuevos desafíos (“*Big Data*”) y nuevas preguntas a responder.

Internet of Events

- La Sociedad ha cambiado de ser predominantemente "analógica" a predominantemente "digital" (Hilbert y López (2011)) en muy pocos años.
- Actividades de Rutina
- Impacto en la forma de hacer negocios y de comunicar (Manyika, 2011)
- La Sociedad, las organizaciones y las personas están "*Always On*".
- Los datos ser reúnen *sobre cualquier cosa, a cualquier momento y en cualquier lugar.*
  - BIG DATA

M. Hilbert and P. Lopez. The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, **332**(6025):60–65, 2011.

J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. Byers. Big Data: The Next Frontier for Innovation, Competition, and Productivity. McKinsey Global Institute, 2011.

Internet of Events

- El estudio "*IDC Digital Universe Study*" (Abril 2014) confirma el crecimiento espectacular de los datos:
  - Ese estudio estima que la cantidad de datos digitales (en PC, cámaras digitales, servidores y sensores) almacenados en 2014 excedió 4 Zettabytes y predice que el "*universo digital*" va a crecer a 44 Zettabytes en el 2020.
- La tan anticipada *explosión de datos* ya ha ocurrido...

V. Turner, J.F. Gantz, D. Reinsel, and S. Minton. *The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things*. International Data Corporation, Framingham, MA, USA, 2014. http://www.emc.com/leadership/digital-universe/

De Bits a Zettabytes

- Un “bit” es la menor unidad de información posible. Tiene dos valores posibles 1 (on) y 0 (off).
- Un “byte” se compone de 8 bits y puede representar $2^8 = 256$ valores.
- Para las cantidades mas grandes de datos, se emplean múltiplos de 1000:
  - 1 Kilobyte (KB) = 1000 bytes,
  - 1 Megabyte (MB) = 1000 KB,
  - 1 Gigabyte (GB) = 1000 MB,
  - 1 Terabyte (TB) = 1000 GB,
  - 1 Petabyte (PB) = 1000 TB,
  - 1 Exabyte (EB) = 1000 PB, y
  - 1 Zettabyte (ZB) = 1000 EB.
- Por lo tanto, 1 Zettabyte = 10^{21} = 1,000,000,000,000,000,000,000 bytes.

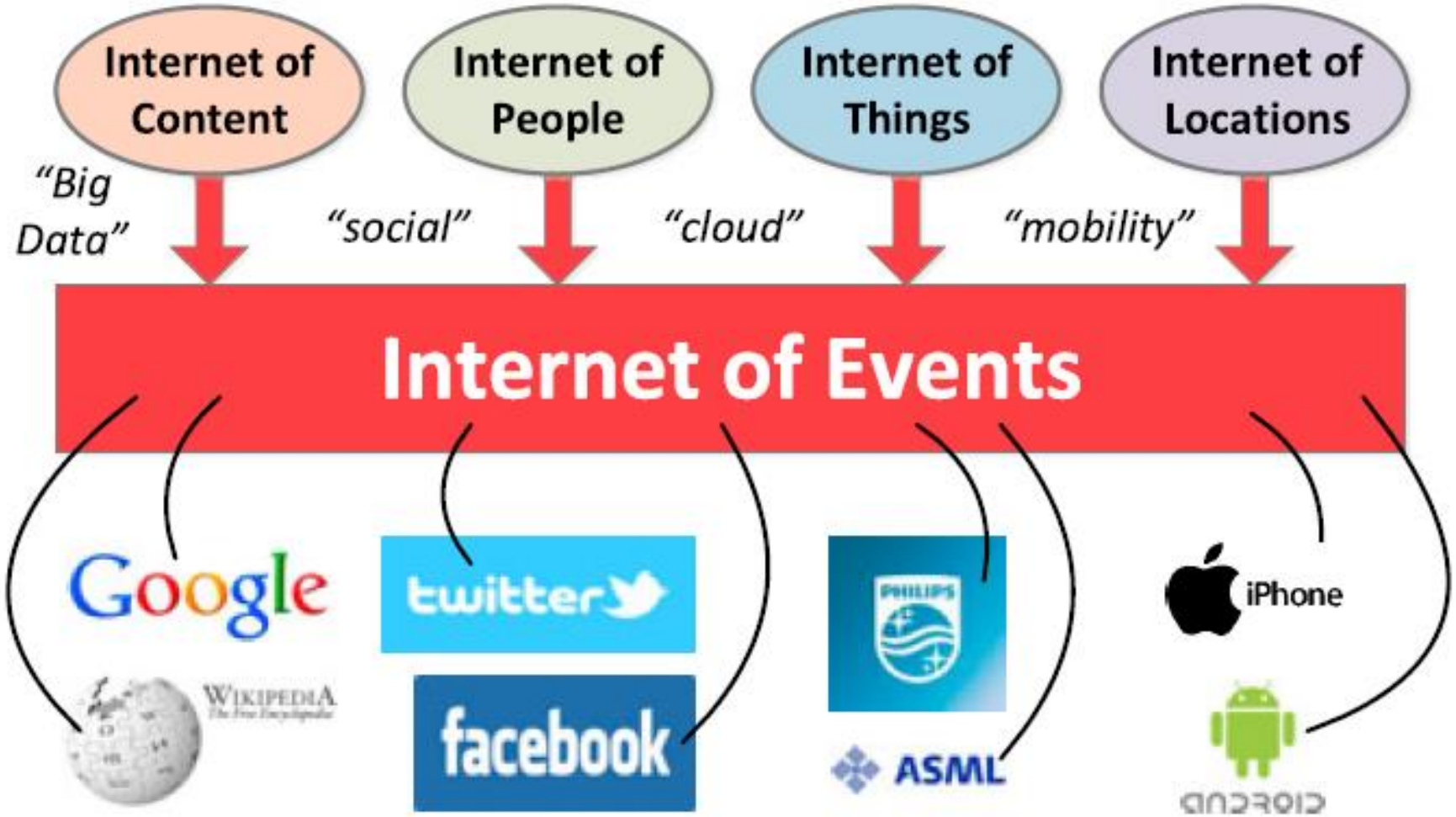
- En lo anterior usamos los prefijos del SI, no los prefijos binarios:
  - 1 Kilobyte = 2^{10} = 1024 bytes,
  - 1 Megabyte = 2^{20} = 1048576 bytes,
  - 1 Zettabyte = $2^{70} \approx 1.18 \times 1021$ bytes.

Datos no estructurados

- La gran mayoría de los datos almacenados en el *"universo digital"* es no-estructurada, y las organizaciones tienen problemas para tratar con estas grandes cantidades de datos.

- Uno de los principales desafíos actuales es *extraer información y valor a partir de los datos* almacenados en los sistemas de información.

Internet of Events



Internet de los Eventos (IoE)

- El termino *Internet of Events* (IoE), fue acuñado en 2014 y se refiere a todos los datos de eventos disponibles. En esencia la IoE esta compuesta por:

  - La ***Internet del Contenido*** (IoC), se refiere a toda la información creada por los seres humanos para incrementar el conocimiento sobre un tema particular. (incluye las paginas web tradicionales, los artículos de enciclopedia como Wkipedia, YouTube, e-books, newsfeeds, etc.)

  - La ***Internet de las Personas*** (IoP), se refiere a todos los datos relacionados con la interacción social (incluye los e-mails, Facebook, Twitter, foros, LinkedIn, etc.)

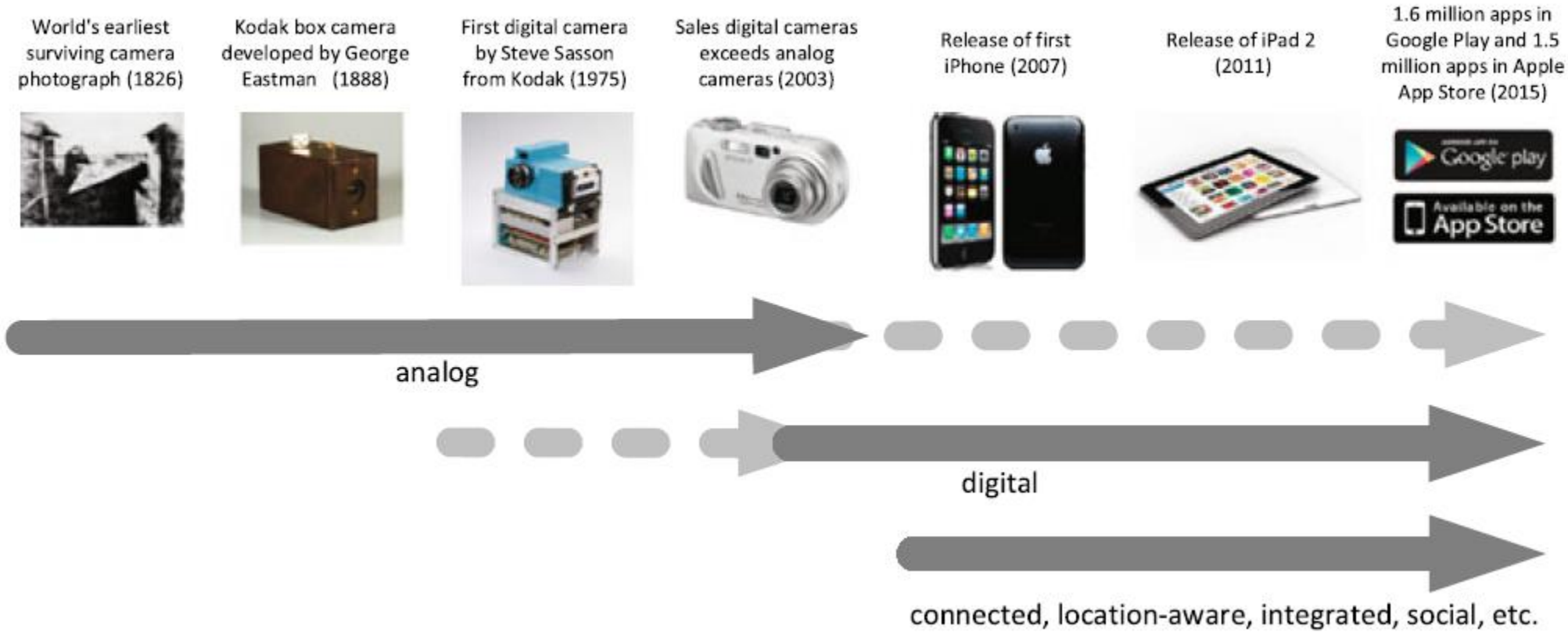
  - La ***Internet de las Cosas*** (IoT), se refiere a todos los objetos físicos conectados a la red. Incluye todas las cosas que tienen un único ID y una presencia en una estructura similar a la Internet.

  - La ***Internet de las Localizaciones*** (IoL) se refiere a todos los datos que tienen una dimensión geográfica o geoespacial. Algo que con la masividad de los dispositivos móviles (por ej. Smartphones) mas y mas eventos tienen atributos de localización o de movimiento.

Es importante notar que IoC, IoP, IoT e IoL tienen muchas superposiciones. Por ejemplo el nombre en una pagina web o la localización de la que un tweet se envía.

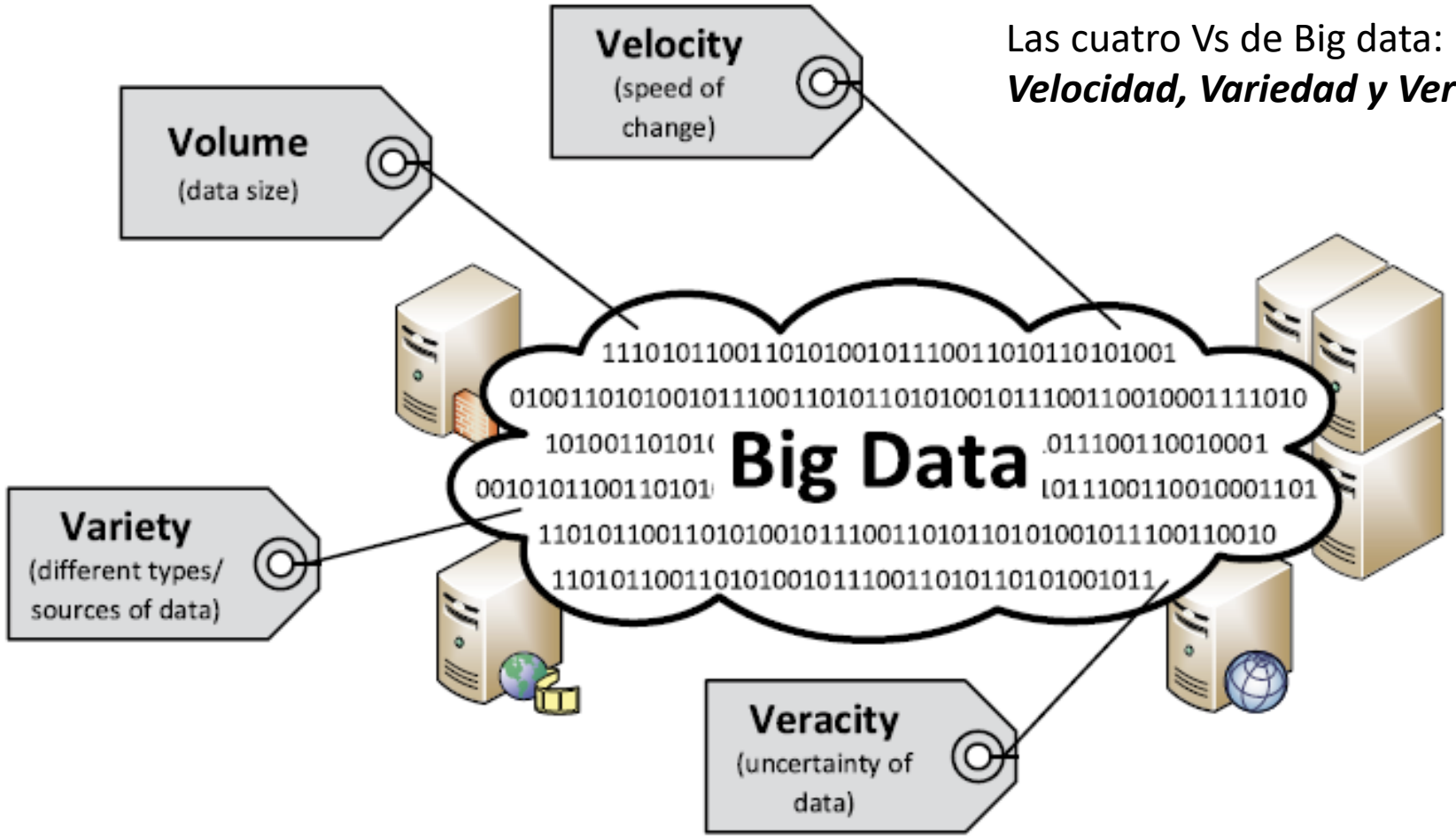
W.M.P. van der Aalst. Data Scientist: The Engineer of the Future. In K. Mertins, F. Benaben, R. Poler, and J. Bourrieres, editors, *Proceedings of the I-ESA Conference*, volume 7 of *Enterprise Interoperability*, pages 13–28. Springer, Berlin, 2014.

Transición de analógico a digital



Cambios dramáticos en la forma en que creamos y compartimos fotografías.

Big Data



Las cuatro Vs de Big data: ***Volumen, Velocidad, Variedad y Veracidad***

Las primeras tres V corresponden al trabajo de (Laney, 2001), luego se han propuesto varias otras Vs: *Variabilidad, Valor, Validez*, etc.

D. Laney. 3D Data Management: Controlling Data Volume, Velocity, and Variety (Research Note 949). Technical report, META Group, February 2001

Data Science

- *La Ciencia de Datos* ha emergido como una nueva disciplina en los últimos años. De hecho hay muchas definiciones que se han sugerido para ella. Ver por ejemplo (Donoho, 2015 y Forbes, 2013)

- ***Data science is an interdisciplinary field aiming to turn data into real value. Data may be structured or unstructured, big or small, static or streaming. Value may be provided in the form of predictions, automated decisions, models learned from data, or any type of data visualization delivering insights. Data science includes data extraction, data preparation, data exploration, data transformation, storage and retrieval, computing infrastructures, various types of mining and learning, presentation of explanations and predictions, and the exploitation of results taking into account ethical, social, legal, and business aspects.***

David Donoho. (2017) 50 Years of Data Science. *Journal of Computational and Graphical Statistics* 26:4, pages 745-766

G. Press. A Very Short History of Data Science. Forbes Technology, http://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/, 2013.

Data Science

- Esa definición implica que la *ciencia de datos* es en principio un concepto mas amplio que el de *estadística aplicada* y el de *minería de datos*.
- En líneas generales los "*cientistas*" de datos asisten a la organización para transformar datos en valor. Y se espera que respondan ciertas preguntas en base a los datos:

  - (Reporting) *What happened?*
  - (Diagnosis) *Why did it happen?*
  - (Prediction) *What will happen?*
  - (Recommendation) *What is the best that can happen?*

Data Science

- Ahora bien, la definición de ciencia de datos es bastante amplia, y algunos consideran a la ciencia de datos como una forma elaborada de decir *estadística.* Y es claro que hay una raíz fundante en la estadística.

- Ahora bien…

Data Analysis?

- John Tukey (1915–2000), conocido por la transformada rápida de Fourier y los "*box-plots*" entre otras cuestiones, pero en 1962 escribió:

  - "For a long time I have thought I was a statistician, interested in inferences from the particular to the general. But as I have watched mathematical statistics evolve, I have had cause to wonder and to doubt. . . . I have come to feel that my central interest is in data analysis, which I take to include, among other things: procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data."

J.W. Tukey. The Future of Data Analysis. *Annals of Mathematical Statistics*, **33**(1):1–67, 1962.

Data Analysis?

- Leo Breiman (1928–2005), otro eminente estadístico, escribió en 2001:
  - "This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics."
- David Donoho (2017) resume claramente esta discusión entre la escuela tradicional de estadística y los analistas de datos...

L. Breiman. Statistical Modeling: The Two Cultures. *Statistical Science*, **16**(3):199–231, 2001

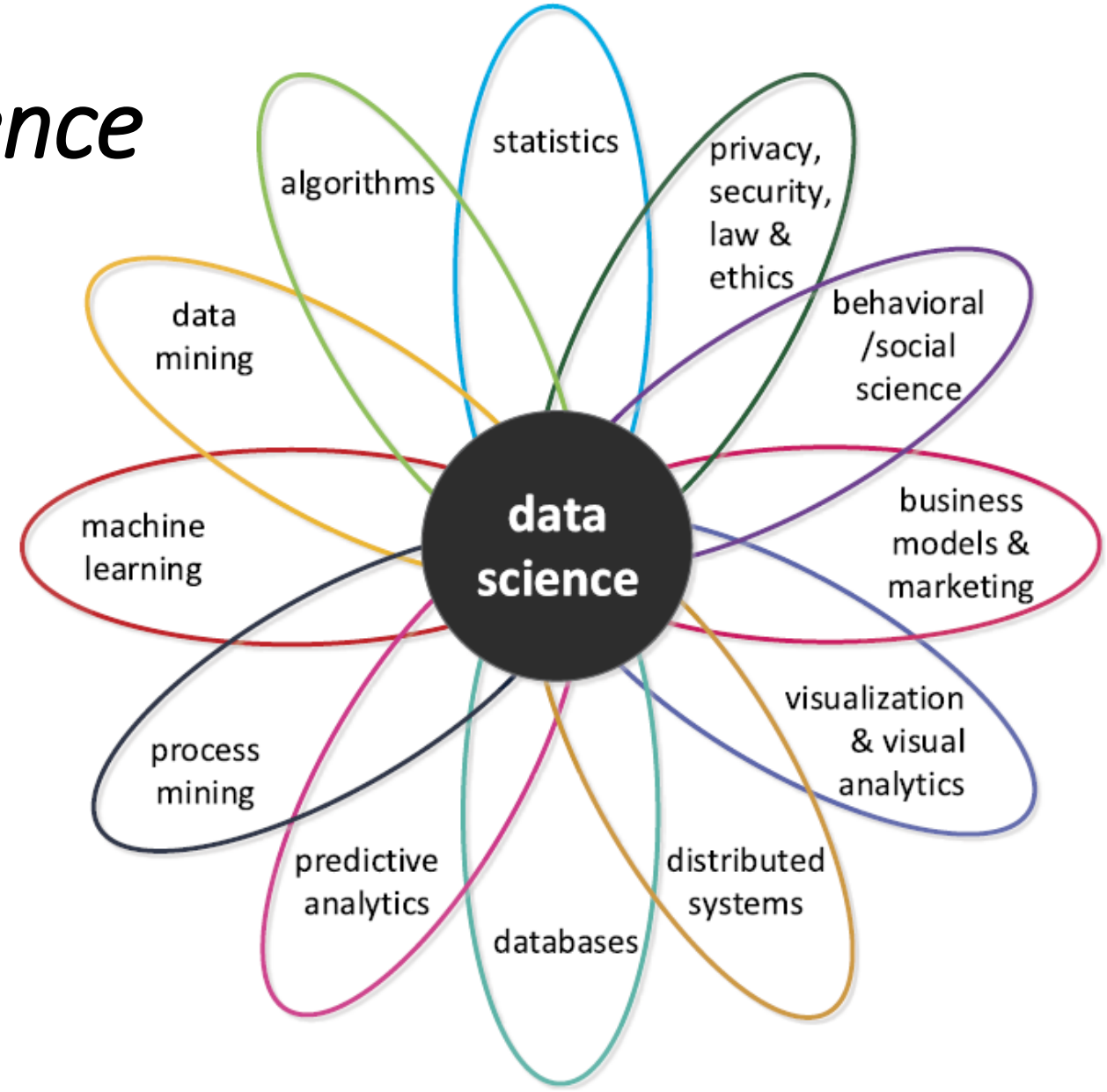
David Donoho. (2017) 50 Years of Data Science. *Journal of Computational and Graphical Statistics* 26:4, pages 745-766

Data Science

- La *ciencia de datos* está fuertemente relacionada con el procesamiento de datos. De hecho, Peter Naur (1928 – 2016), ganador del premio Turing empleó el concepto "*data science*" bastante antes que estuviera de moda. En 1974, Naur escribió:
  - "A basic principle of *data science*, perhaps the most fundamental that may be formulated, can now be stated: The data representation must be chosen with due regard to the transformation to be achieved and the data processing tools available"
- Incluso antes, Peter Naur también había definido *datalogy* como "science of the nature and use of data" y sugería emplear este termino en lugar de "*computer science*".
- Su libro de 1974, incluso tiene dos partes, considerando a "*large data*": "Part 5—Processes with Large Amounts of Data" and "Part 6—Large Data Systems". En ese libro, "*large amounts of data*" equivale a todos los datos que no pueden ser almacenados en la memoria de trabajo, y recordemos que la capacidad máxima de los discos magnéticos de almacenamiento considerados en la época variaba entre 1.25 y 250 Mb. Y no solo los discos son varios ordenes de magnitud mas pequeños que los discos actuales, sino que la noción de lo que es "*large/big*" ha cambiado dramáticamente desde principios de 1970s. Aun así, muchos de los principios centrales del procesamiento de datos han permanecido invariantes.

P. Naur. *Concise Survey of Computer Methods*. Studentlitteratur Lund, Akademisk Forlag, Kobenhaven, 1974.

Data Science



Ingredientes

Data Science

- *Data Science* es una amalgama de (sub)disciplinas diferentes y parcialmente superpuestas, cuyas fronteras no son estrictas y parecen cambiar con el tiempo.

- ¿Cual es la diferencia entre *data mining*, *machine learning* y *statistics*?
  - Sus raíces son muy diferentes:
    - *Data mining* emergió de la comunidad de base de datos, mientras que *machine learning* emergió de la comunidad de inteligencia artificial (AI), y ambos bastante desconectados de la comunidad de estadística
  - Pero las tres (sub)disciplinas definitivamente se superponen…

Data Science

- ***Estadística**:* puede ser considerada como el origen de la ciencia de datos. Tradicionalmente se la divide en estadística *descriptiva* (para sumarizar datos empleando nociones como la median, desviación standard y frecuencia) e *inferencial* (que emplea datos muestrales para estimar las características de la población o para testear una hipótesis)

- ***Algoritmos:*** son cruciales en cualquier aproximación al análisis de datos. Cuando el conjunto de datos se hace mas grande, la complejidad de los algoritmos es una cuestión importante. Por ejemplo el Algoritmo *Apriori* para encontrar conjuntos frecuentes, o la aproximación *MapReduce* para la paralelización algorítmica y el algoritmo *PageRank* de Google.

- ***Minería de Datos:*** de la que existen muchas definiciones, incluyendo: "*the analysis of (often large) data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner*" (Hand, Mannilla y Smyth, 2001). Usualmente los datos de entrada corresponden a una tabla y los resultados se expresan en forma de reglas, conglomerados, arboles, gráficos, ecuaciones, patrones, etc. Claramente la minería de datos se basa en estadística, bases de datos y algoritmos. Aunque comparada con la estadística el foco primario es en la escalabilidad y las aplicaciones practicas.

D. Hand, H. Mannila, and P. Smyth. *Principles of Data Mining*. MIT press, Cambridge, MA, 2001.

Data Science

- ***Machine learning**:* trata con la pregunta de como construir programas informáticos que automáticamente mejoren la experiencia. La diferencia entre *data mining y machine learning* es equivoca. El campo del aprendizaje automático emergió dentro de las técnicas de Inteligencia Artificial (AI), incluyendo por ejemplo, las redes neuronales. Es por eso que se emplea el termino *machine learning* para referirse a algoritmos que dan a las computadoras la capacidad de aprender SIN ser explícitamente programadas ("*learning from experience*"). Para aprender y adaptarse, un modelo se construye a partir de los datos de entrada (en lugar de emplear rutinas fijas). El modelo evoluciona para realizar predicciones o decisiones basadas en los datos.

- ***Process mining:*** agrega la perspectiva de procesos al aprendizaje automático y a la minería de datos. Busca la confrontación entre los datos de eventos (es decir, conducta observada) y los modelos de proceso (definidos de ante mano o descubiertos automáticamente). Los datos de eventos se relaciona con modelos explícitos de procesos, por ejemplo las redes de Petri o los modelos BPMN. De esta forma, los modelos de procesos se descubren a partir de los datos de eventos, o los datos de eventos se reproducen sobre los modelos para analizar "*compliance and performance*".

T.M. Mitchell. *Machine Learning*. McGraw-Hill, New York, 1997.

Data Science

- ***Predictive analytics**:* Es la practica de extraer información a partir de conjuntos de datos existentes para determinar patrones y predecir futuros resultados y tendencias. Para generar predicciones, las aproximaciones existentes de minería y aprendizaje se aplican en un contexto de negocios. Esta fuertemente relacionado con los conceptos de *Business Analytics* y *Business Intelligence.*

- ***Databases**:* Se emplean para almacenar datos. La disciplina de bases de datos forma uno de los pilares de la ciencia de datos. Los Sistemas de Gestión de Bases de Datos (DBMS) sirven a dos propósitos primarios: (i) estructurar los datos de manera que puedan ser administrados fácilmente, y (ii) proveer de escalabilidad y performance. Al utilizar tecnología de bases de datos, los programadores de aplicaciones no necesitan preocuparse por el almacenamiento de los datos. Hasta hace relativamente poco tiempo atrás, las bases de datos relacionales y el lenguaje SQL eran la norma. Ahora bien, dado el creciente volumen de datos, han emergido las bases de datos masivamente distribuidas y las denominadas NoSQL. Aun mas, el computo en memoria (SAP HANA, por ejemplo), puede ser empleado para responder preguntas en tiempo real. Un concepto relacionado es el de *OLAP* donde los datos se almacenan en cubos multidimensionales que facilitan el análisis desde diferentes puntos de vista.

Data Science

- ***Sistemas Distribuidos:*** Los sistemas distribuidos proveen la infraestructura sobre la que realizar los análisis. Un Sistema distribuido se compone de elementos que interactúan y que coordinan sus acciones para realizar un objetivo común *Cloud, grid y utility computing* dependen de sistemas distribuidos. Algunas tareas de análisis son demasiado grandes y demasiado complejas para ser realizadas en una única computadora. Esas tareas pueden ser divididas en tareas mas pequeñas que pueden ser realizadas concurrentemente sobre diferentes nodos de computo. La escalabilidad puede ser realizada al compartir y/o extender el conjunto de nodos de computo.

- ***Visualization & visual analytics****:* Se trata de elementos clave en la ciencia de datos. Finalmente las personas necesitan interpretar los resultados y guiar el análisis. El aprendizaje automático y las técnicas de minería de datos puede ser empleadas para conocimiento a partir de los datos. Ahora bien, hay muchos "*desconocidos desconocidos*" (cosas que no sabemos que no sabemos). EL análisis realmente depende muy fuertemente en el juicio humano y la interacción directa con los datos. Las capacidades de percepción de los sistemas cognitivos humanos pueden ser empleadas mediante visualizaciones adecuadas. *Visual analytics*, es un termino acuñado por Jim Thomas que combina las técnicas de análisis automatizado con visualizaciones interactivas para un mejor entendimiento, razonamiento y toma de decisiones sobre la base de datos muy grandes y complejos.

J. van Wijk. The Value of Visualization. In *Visualization 2005*, pages 79–86. IEEE CS Press, 2005.

D. Keim, J. Kohlhammer, G. Ellis, and F. Mansmann, editors. *Mastering the Information Age: Solving Problems with Visual Analytics*. VisMaster, http://www.vismaster.eu/book/, 79.2010.

- El 12 de Febrero de 2002, al hablar sobre las armas de destrucción masiva en Irak, el Secretario de Defensa de los EEUU empleo la siguiente clasificación:

- ***(i) "known knowns" (things we know we know),***
- ***(ii) "known unknowns" (things we know we don't know), and***
- ***(iii) "unknown unknowns" (things we don't know we don't know).***

Data Science

- ***Business models & marketing:*** La ciencia de datos trata en esencia de convertir los datos en valor, incluido el valor comercial o de negocios. Por ejemplo, la capitalización de Mercado de Facebook en Noviembre de 2015 era de aproximadamente US $300 billón y contaba con alrededor de 1500 usuarios activos mensuales. Por lo tanto, el valor promedio de un usuario de Facebook era de US $200. Al mismo tiempo, el valor promedio de un usuario de Twitter era de US $55 (capitalización de Mercado de aproximadamente US $17 billones con 307 millones de usuarios. Esto muestra el valor económico de los datos y el éxito de negocios de compañías jóvenes que se basan en un nuevo modelo de negocios. (*Airbnb, Uber, Alibaba*, etc.)

Data Science

- ***Behavioral/social science:*** Gran parte de los datos son (indirectamente) generados por las personas y los resultados de los análisis son muchas veces empleados para influir en las personas (guiando a un cliente a un producto, o alentando a un gerente a eliminar desperdicios). Las ciencias de la conducta implican un análisis sistemático e investigación de la conducta humana. Por otro lado, las ciencias sociales estudian los procesos de un Sistema social y las relaciones entre los individuos dentro de una Sociedad. Para interpretar los resultados de varios tipos de analíticas, es importante comprender la conducta humana y el contexto social en el que los seres humanos y las organizaciones operan. Aun mas, los análisis muchas veces plantean nuevas preguntas en relación con influir positivamente en las personas.

Data Science

- Ahora bien, la ***Privacidad, seguridad, la ética y la ley*** son ingredientes clave para proteger a los individuos y las organizaciones de “ malas” practicas de ciencia de datos.

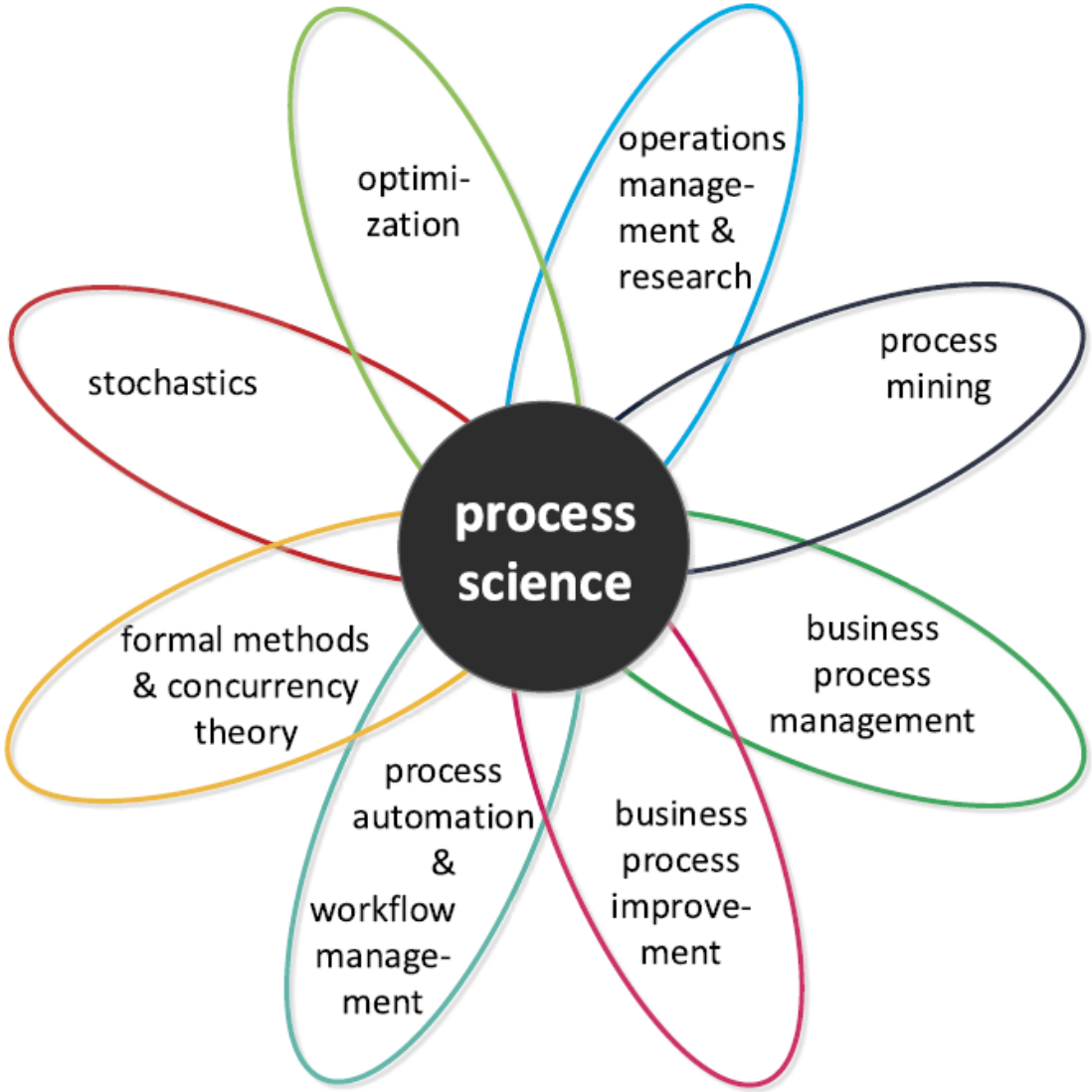
- *Privacidad* en relación a aislar información sensible, que depende muchas veces de mecanismos de seguridad que apuntan a asegurar la confidencialidad, integridad y disponibiliad de los datos.
  - Los datos debe ser precisos y almacenados con seguridad, evitando accesos no autorizados.

- La Privacidad y la seguridad deben ser cuidadosamente consideradas. Los individuos deben ser capaces de confiar en la forma en que sus datos son almacenados y transmitidos.

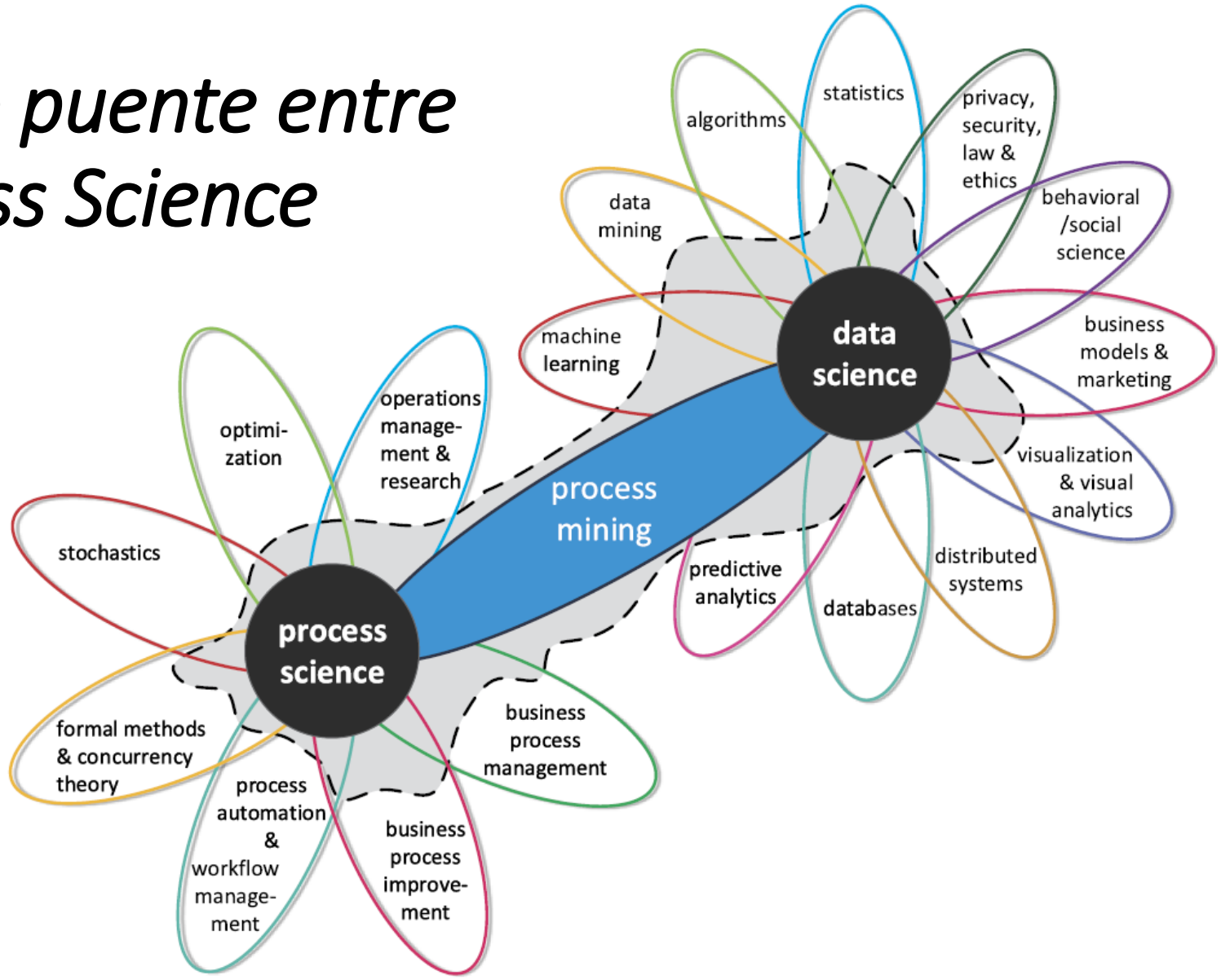
- La ética permite determinar cuales tipos de análisis son moralmente defendibles…

Process Science

Como concepto envolvente para la disciplina que combina el conocimiento de la tecnología de información y el conocimiento de las ciencias de la administración para mejorar y realizar procesos operacionales.



Process Mining como puente entre Data Science y Process Science



Process Science

- ***Process mining*** busca la confrontación entre los datos de eventos (conducta observada) y los modelos de procesos.

- La ciencia de datos tiende a ser agnóstica en términos de procesos. Tanto *Data Mining*, como la estadística o el *Machine Learning* no consideran procesos de extremo a extremo.

- *Process Science* muchas veces se enfoca en el modelado mas que en aprender de los datos de evento.