Desmitificando la Ciencia de Datos

¿Es la Ciencia de Datos para Todos?

José Charango Munizaga-Rosas
13 Noviembre 2020

Contenidos

- Introducción al prontuario de Charango
- Algunas definiciones básicas
- El proceso de la ciencia de datos
- Algunas consideraciones prácticas
- Botando algunos mitos comunes
- ¿Ciencia de datos para todos?
  - Propuesta de modelo de entrenamiento y desarrollo de competencias
- Conclusiones

El Presentador

¿Quién es José Charango Munizaga-Rosas?

- Ingeniero Matemático y Magíster en Gestión de Operaciones (Universidad de Chile)
- Ph.D. in Natural Resources Engineering (Laurentian university, Sudbury, ON, Canada)
  - Ex profesor de Planificación Minera, Estimación de Recursos e Investigación de Operaciones Mineras en WASM, Kalgoorlie.
  - Ex Senior Lecturer en DMEE, Curtin University, Perth, Australia.
- Actualmente Principal Consultant and Chief Data Scientist en Coalesce Group y Chief Data Scientist en BlueSky Labs, Perth, Western Australia
- Fundador y CEO de Minformatics SpA, Santiago, Chile
- Profesor Adjunto en el Departmento de Mineral and Energy Economics, Curtin University, Western Australia, Australia
- Profesor Adjunto, Departamento de Ingeniería de Minas, Director Académico Diplomado de Economía de Minerales, Universidad de Chile, Santiago, Chile
- Profesor Extraordinario, Facultad de Ingeniería de Minas, Universidad Nacional del Altiplano, Puno, Perú

¿Qué he Hecho en el Pasado? (Prontuario Criminal)

- Modelos de Automatas Celulares para Flujo Gravitacional
  - Simulación de método Block caving, simulación de subsidencia para operaciones de carbón long-wall
- Investigación de Operaciones Aplicado a Minería
  - Algoritmos genéticos, programación entera mixta, planificación minera
- Geoestadística y Análisis de Riesgo en Mineria
  - Uso de simulaciones condicionales para incorporar incertidumbre de leyes a la planificación minera, localización óptima de drillholes
- Ciencia de Datos en Distintos Dominios:
  - Salud
  - Seguridad Social
  - Logística
  - Finanzas
  - Distribución Eléctrica,
  - Planeamiento y Control de Riego (Agroindustria)

Algunas Definiciones Básicas

Acerca de los Datos (circa 2016, USA)

- Menos del 0.5% de todos los datos que creamos son alguna vez analizados y usados
- Hacia fines de 2016, un 73% de las organizaciones estaban ya invirtiendo o habían invertido en **big data**
- Google utiliza alrededor de 1,000 computadores para responder una única consulta
- Se proyectaba que para 2020, habrían mas de 50.000 millones de dispositivos inteligentes conectados en el mundo recolectando, analizando y compartiendo datos (probablemente el número quedó chico...)
- En 2015, un estimado de 1 trillón de fotos fueron tomadas y miles de millones de ellas fueron compartidas online
- En 2016, 40,000 consultas de búsqueda fueron hechas cada segundo solamente en Google alone, lo que dá del orden de 1.2 trillones de búsqeudas por año
- Se estimaba que para el 2020, aproximadamente 1.7 Mb de nueva información sería creada cada segundo por cada humano en el planeta
- Los datos deficientes le cuestan solamente a los negocios de USA del orden de US$600.000 millones anualmente
- 70% de los datos son creados por individuos, sin embargo, las empresas son responsables de almacenar y gestionar 80% de dicho volumen
- Hay aproximadamente tantas piezas de información digital como estrellas hay en el universo

Los datos **están en todos lados** y se **recolectan indiscriminadamente**

Conceptos Generales (1)

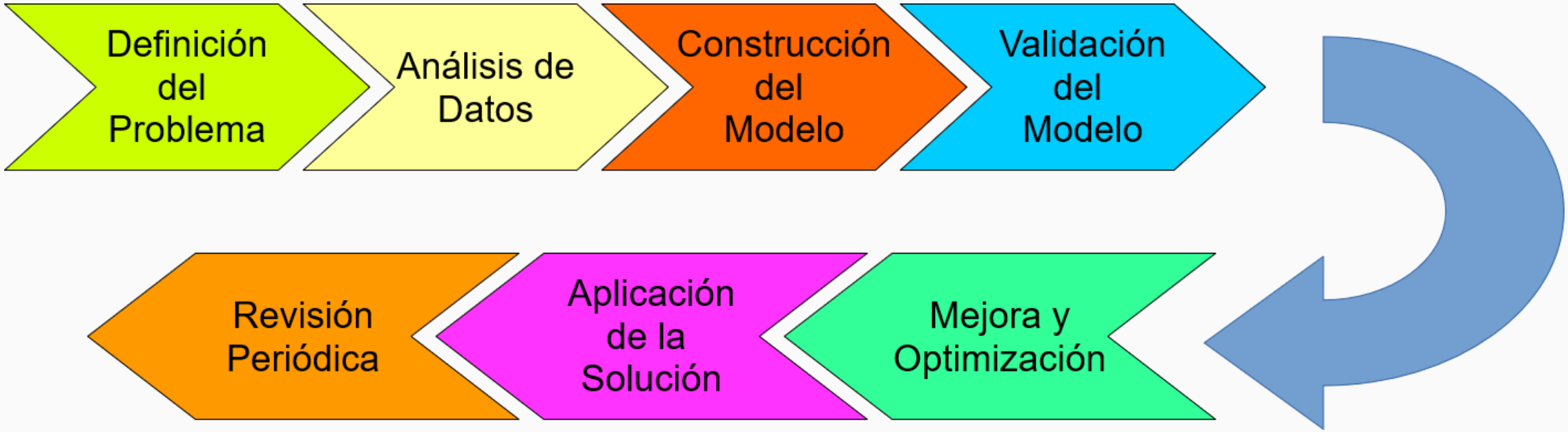
- Todas las compañías/organizaciones “racionales” quisieran:
  - Ahorrar dinero
  - Aumentar la eficiencia
  - Reducir los costos
  - Etc.
- Para satisfacer cualquiera de esos deseos, una *estrategia* es necesaria y debe ser aplicada:
  - Esta estrategia puede ir desde la aplicación de reglas simples a modelos super complejos
- ¿Qué se necesita?
  - Primero, una **definición apropiada del problema**:
    - El hecho de que una compañía quiera hacer más dinero no garantiza que ese dinero sea obtenible
  - Tambien es **esencial definir un mecanismo que permita discriminar buenas soluciones** (bajo el supuesto de que el problema a resolver vale la pena, i.e., hay un problema)

“Si tienes un problema que puede ser resuelto, no tiene sentido preocuparse. Si tienes un problema que no puede ser resuelto, no tiene sentido preocuparse” – Proverbio budista

Conceptos Generales (2)

- Todo esto suena muy bonito, sin embargo:
  - Las compañías/organizaciones no tienen información completa, perfecta y transparente
  - El futuro es díficil (de hecho **imposible**) de predecir
    - Las compañías/organizaciones usualmente ni siquiera conocen los datos que poseen
    - Menos aún conocen los problemas que tienen
  - Hay **problemas metodológicos** relacionados a la recolección de datos y el como se usan
    - El proceso de curatoría de datos debe garantizar que la recolección y almacenamiento no produce sesgos
    - Cualquier error en este aspecto inhabilita el uso de los datos (**Rubbish In = Rubbish Out**)
- Los datos por sí mismos son de poco valor:
  - Para entregar valor, los datos tienen que ser usados de manera inteligente
  - Usualmente los datos deben ser transformados para ser utilizados por los modelos matemáticos
    - Y esta es la parte *aburrida pero fundamental* del trabajo donde se gasta del orden de un **80%** del tiempo y que aparentemente no produce resultados vistosos

Workflow para Resolver un Problema



El Proceso de la Ciencia de Datos

¿Qué es el Data Science?

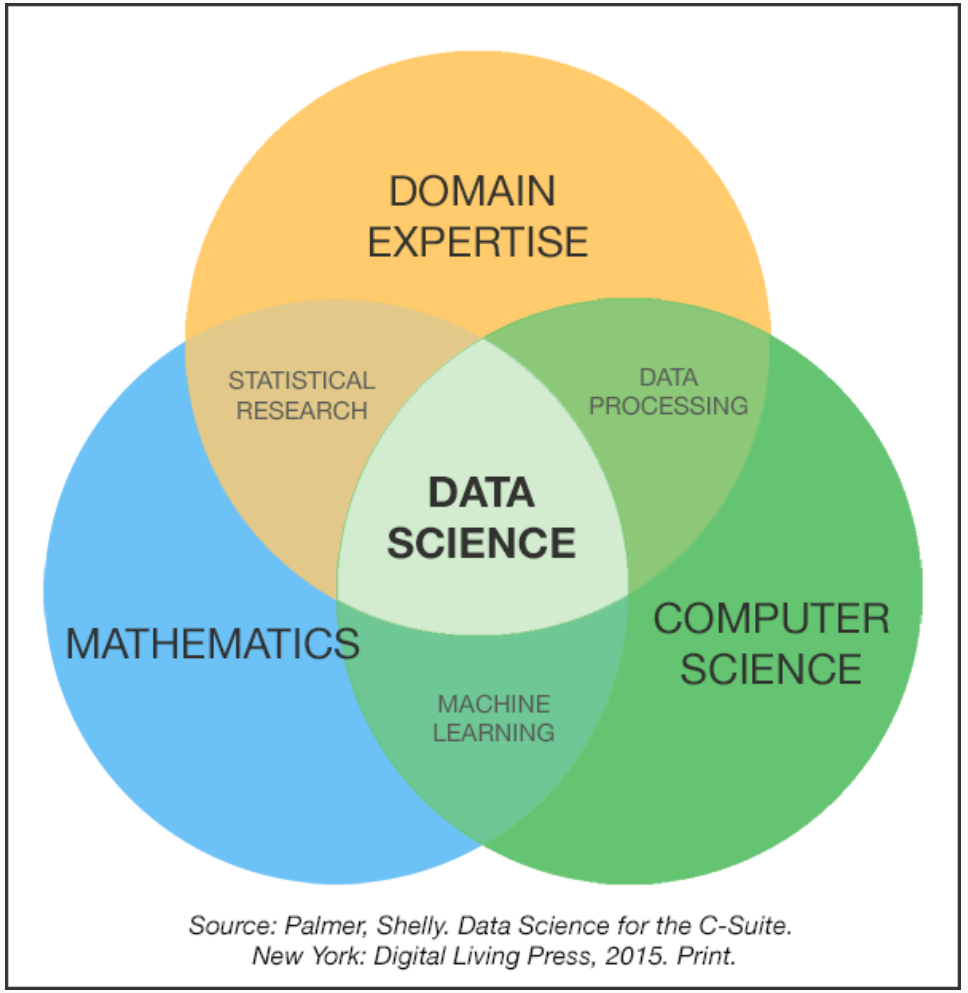
Definición adaptada de Wikipedia:

"El Data Science, también conocido como la ciencia conducida por datos, es un campo interdisciplinario que usa el método científico, procesos y sistemas parta extraer conocimiento o visión desde los datos en varias formas, ya sean estructurados o no estructurados, similar al descubrimiento de conocimiento en bases de datos (KDD sus siglas en Inglés)"

- La ciencia de datos está en la intersección de tres dominios distintos:
  - *Matemáticas, Ciencias de la Computación y Conocimiento del Negocio*
- Matemáticos y estadísticos hay disponibles en general, si bien el conocimiento es especializado
  - Muchos de los algoritmos están encapsulados en librerías y se pueden utilizar con algunas limitaciones dadas por los supuestos de los modelos que implementan
- Gente que entienda los modelos computacionales y que sea capaz de implementarlos de manera eficiente se facilita con las librerías existentes para ingestión y transformación de datos así como de algoritmos de aprendizaje
- **La comprensión de los aspectos del negocio, en particular los relacionados a la forma en que este opera y todos los detalles asociados a él, es la parte díficil**

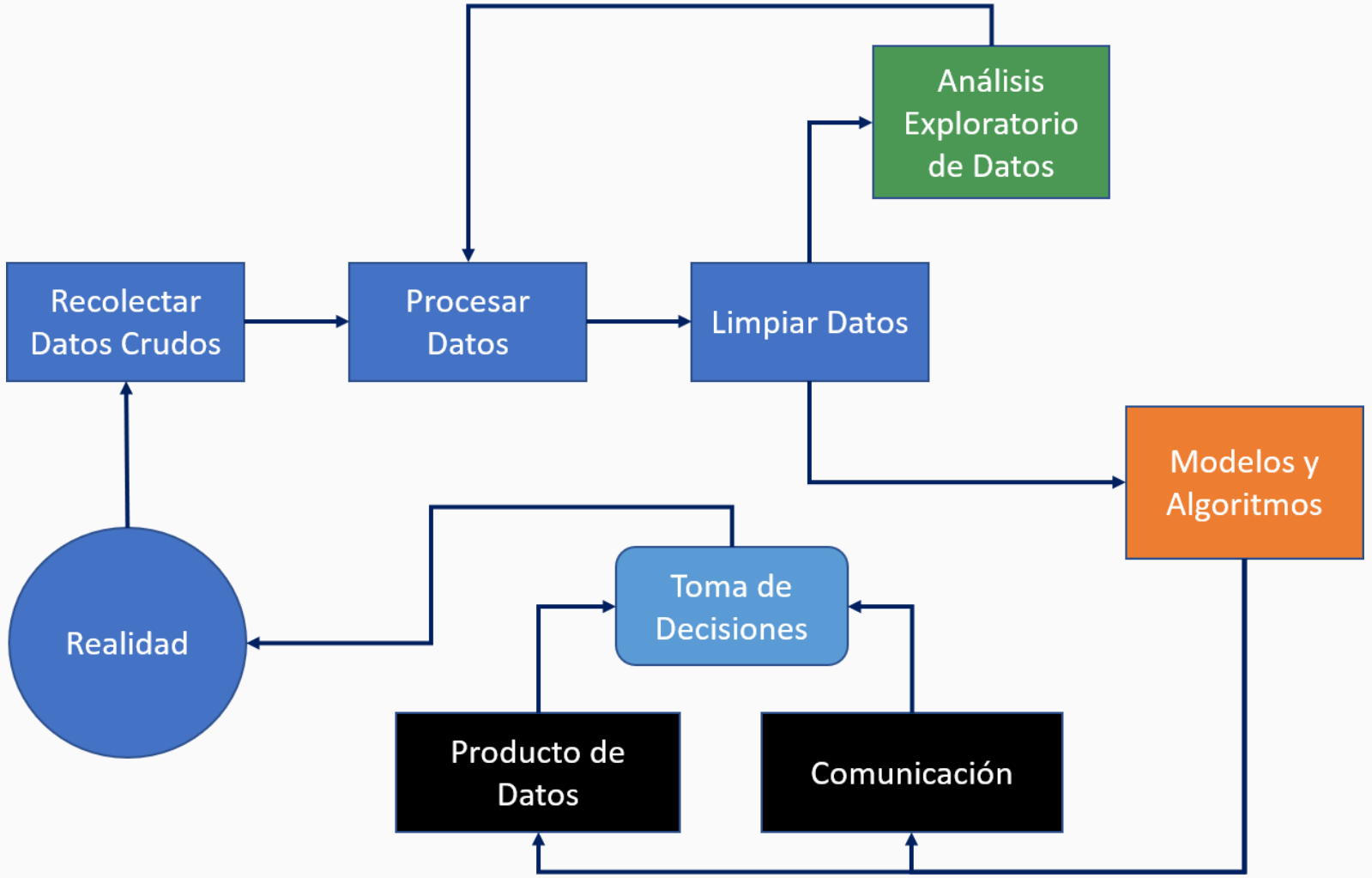
El Rol del Científico de Datos

- Dar con una persona que reúna las tres habilidades simultáneamente es díficil
  - En mi opinión no existe nadie que sea 100% bueno en cada una de las áreas:
    - Siendo el dominio de conocimiento del negocio es el que más complica ¿Porqué no mejor **entrenar** a quienes conocen el negocio para así poder identificar los problemas que hay que resolver? Sobre esto un poco más adelante...
  - Siempre es posible complementar debilidades a través de apoyo de un equipo
- La ciencia de datos **no** es **Analítica de Datos**, tampoco **Machine Learning**, menos **data mining**



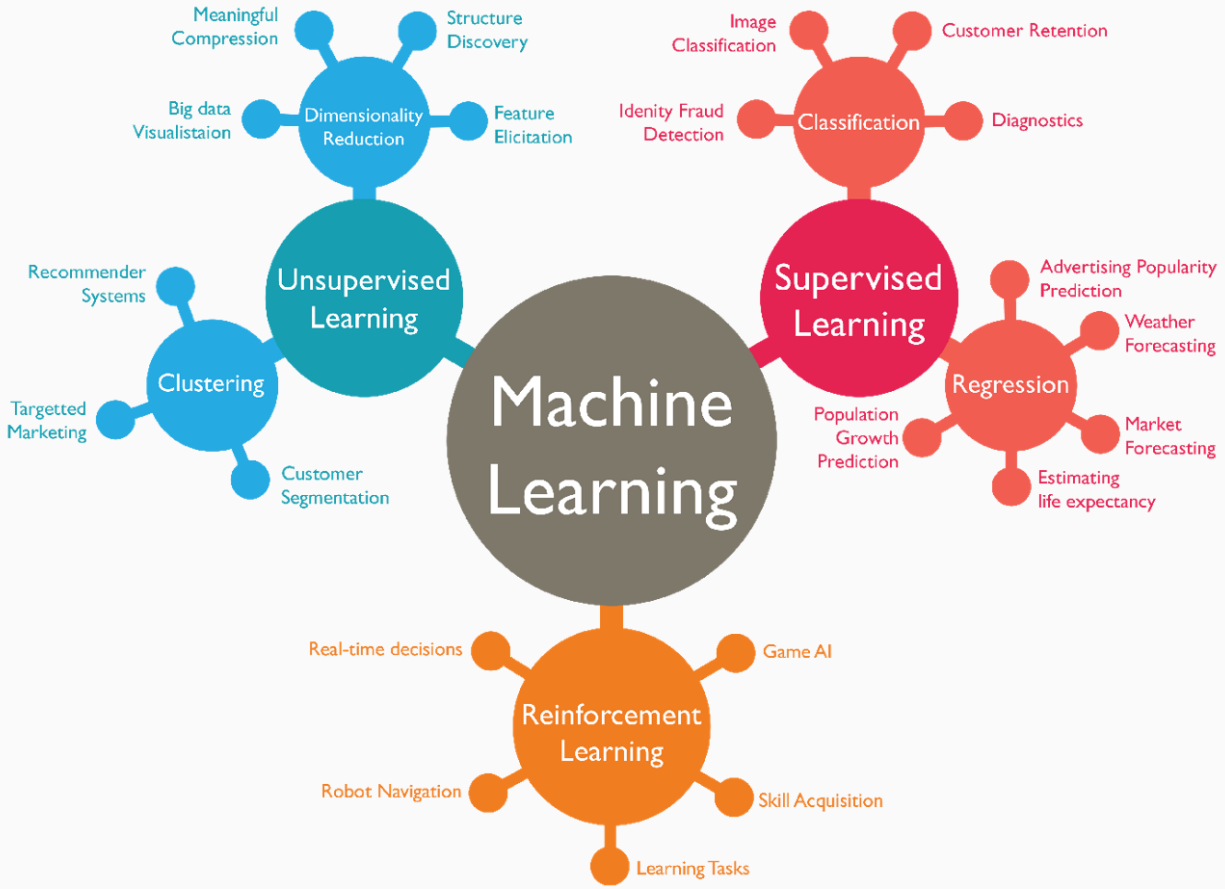
Source: Palmer, Shelly. Data Science for the C-Suite. New York: Digital Living Press, 2015. Print.

El Proceso de Ciencia de Datos



El Aprendizaje Automático

La parte de ajuste de modelos en Ciencia de Datos se lleva a cabo de manera natural utilizando algoritmos automatizados la mayoría de las veces, la taxonomía típica es (*Fuente: http://www.isaziconsulting.co.za/machinelearning.html*):



Algunas Consideraciones Prácticas

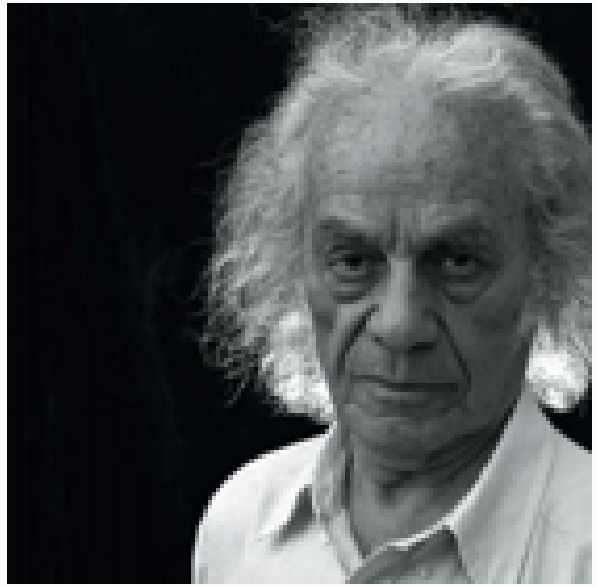
Los Promedios son el Enemigo

"El ser humano promedio tiene un seno y un testículo"

(**Des McHale**, año desconocido)

“Hay dos panes. Usted se come dos. Yo ninguno. Consumo promedio: un pan por persona”

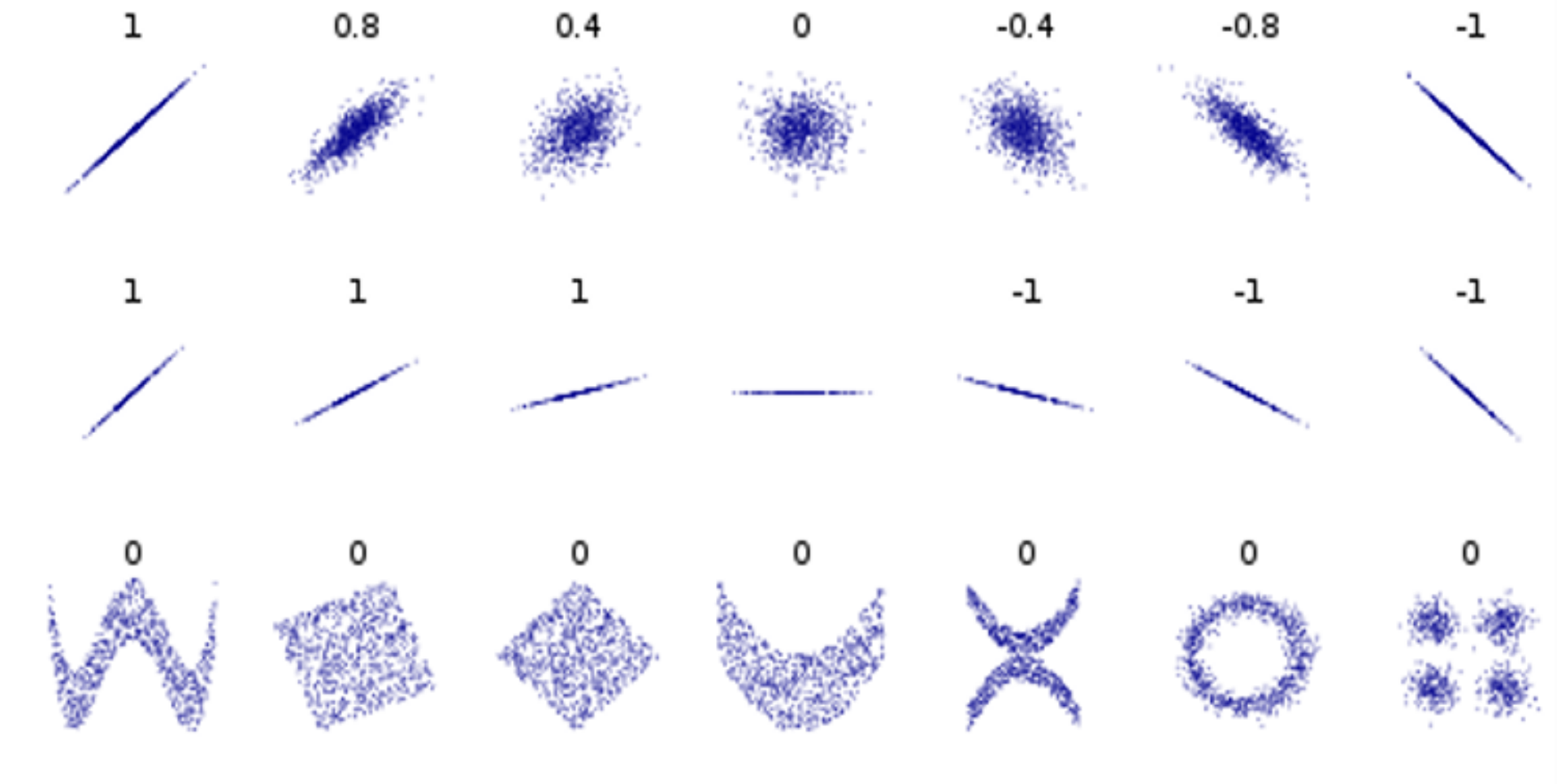
(**Nicanor Parra**, 2013)



Correlación no es Causalidad (1)

- Usualmente queremos obtener una expresión de la forma $Q = f(x, y, z)$ donde Q es la variable **dependiente** y x, y, z son las variables **dependientes**
- Un resultado estadísticamente significativo **no necesariamente implica causalidad**, también necesitamos:
  - Soporte teórico para la relación
  - Sentido común (Este no se puede enseñar, hay gente que nace con él y otros que no)
- La causalidad se define como la relación entre causa y efecto
  - Una relación de causalidad entre los eventos A y B implica que o bien A causa B o viceversa
- En la mente de los profesionales e ingenieros, la causalidad se mide usualmente a través de correlaciones
  - Sin embargo hay que notar que esta correlación es sólo **lineal**

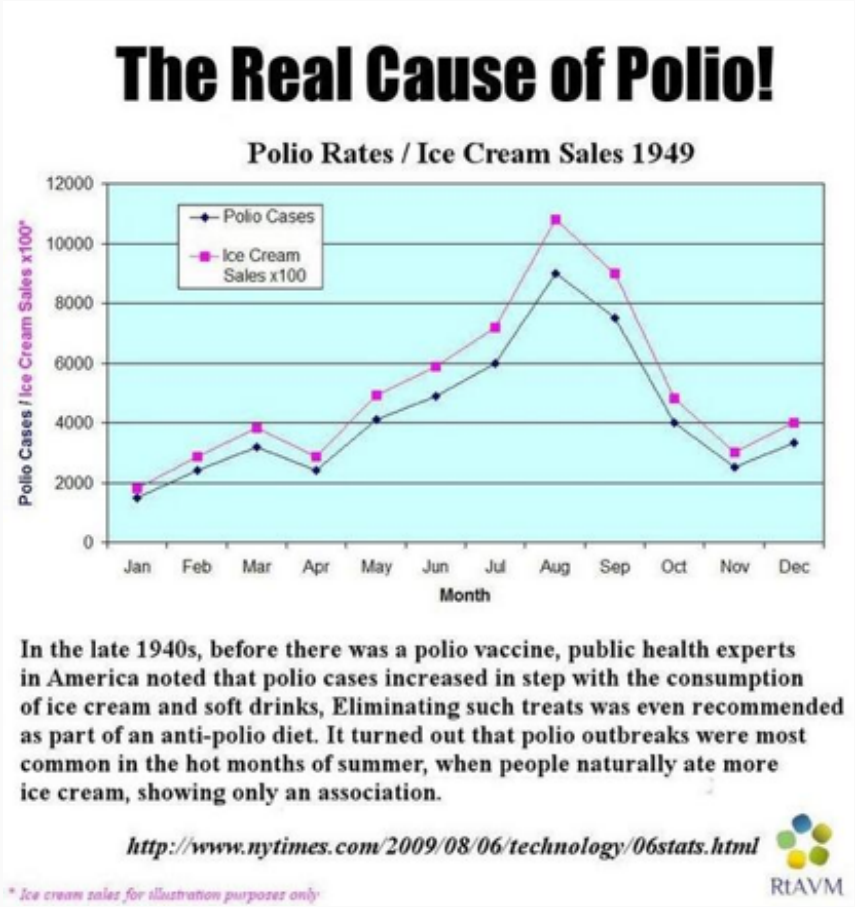
Correlación no es Causalidad (2)



Fuente: Wikipedia

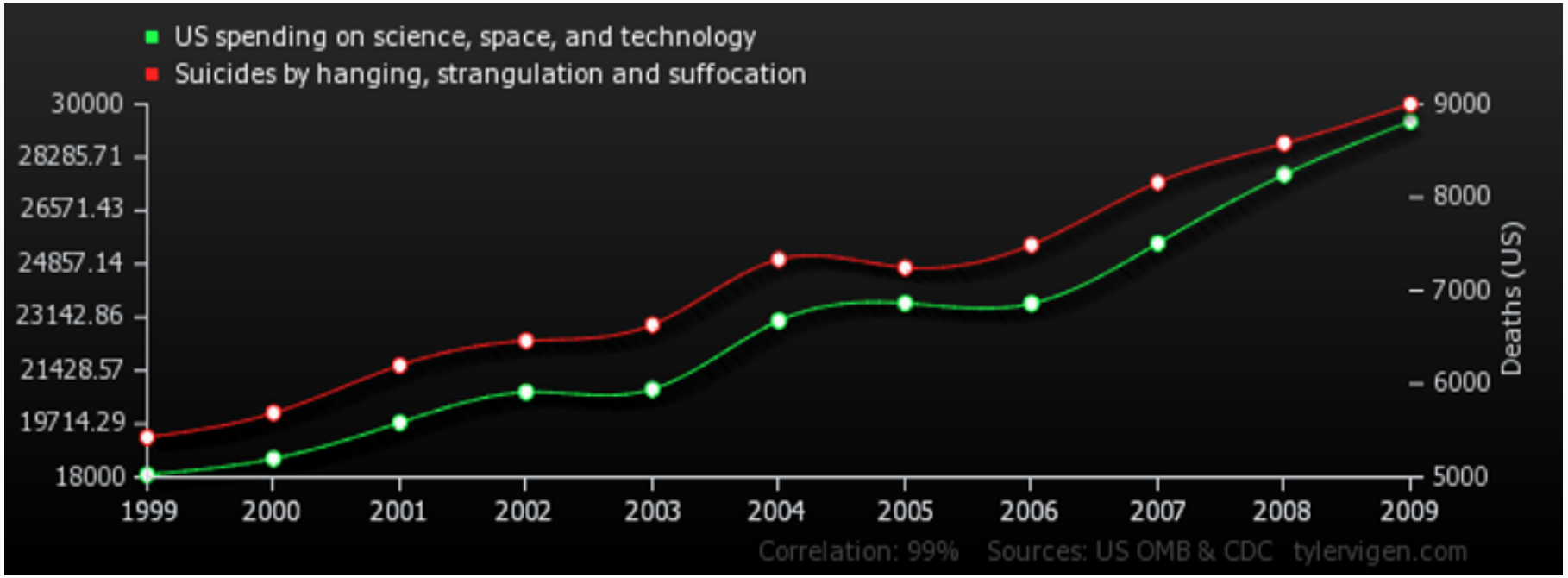
Correlación no es Causalidad (3)

Ejemplo clásico de mala relación de causalidad en presencia de correlación



La correlación (casi perfecta en este caso) no implica causalidad. es decir, la poliomelitis no se origina en el consumo de helado!!!

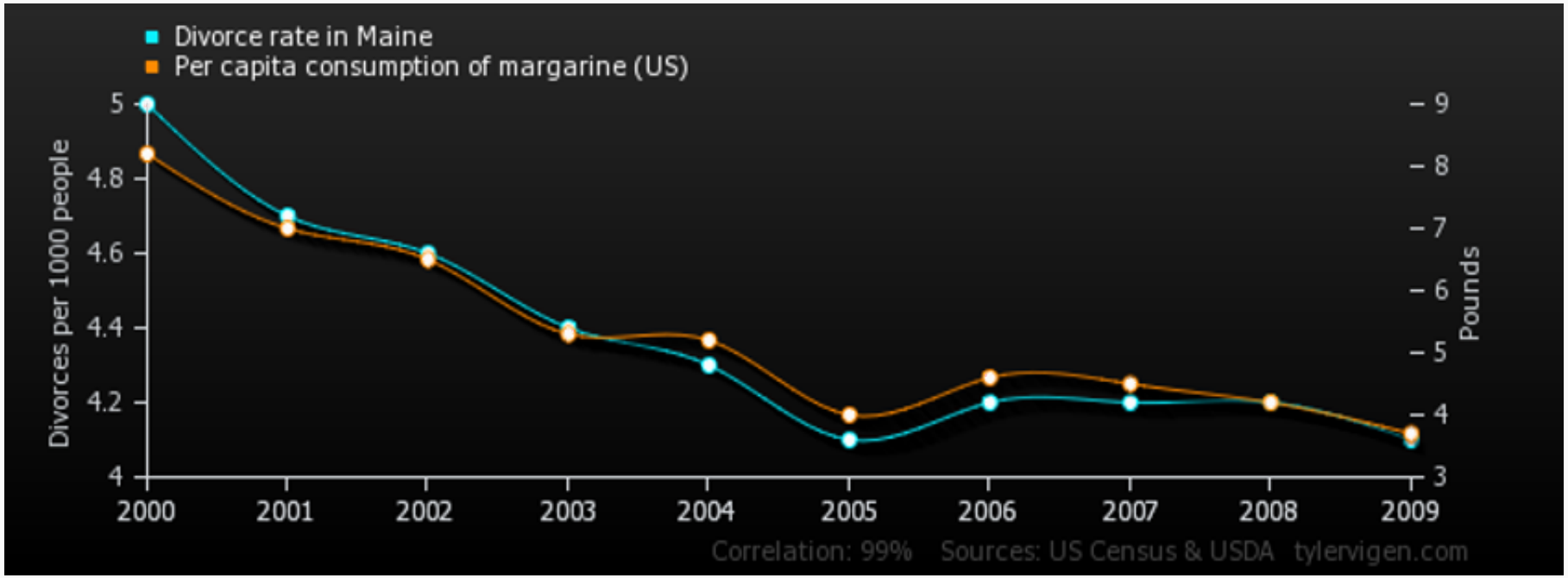
Correlación no es Causalidad (4)



Correlación: 0.992082

Fuente: http://www.tylervigen.com/

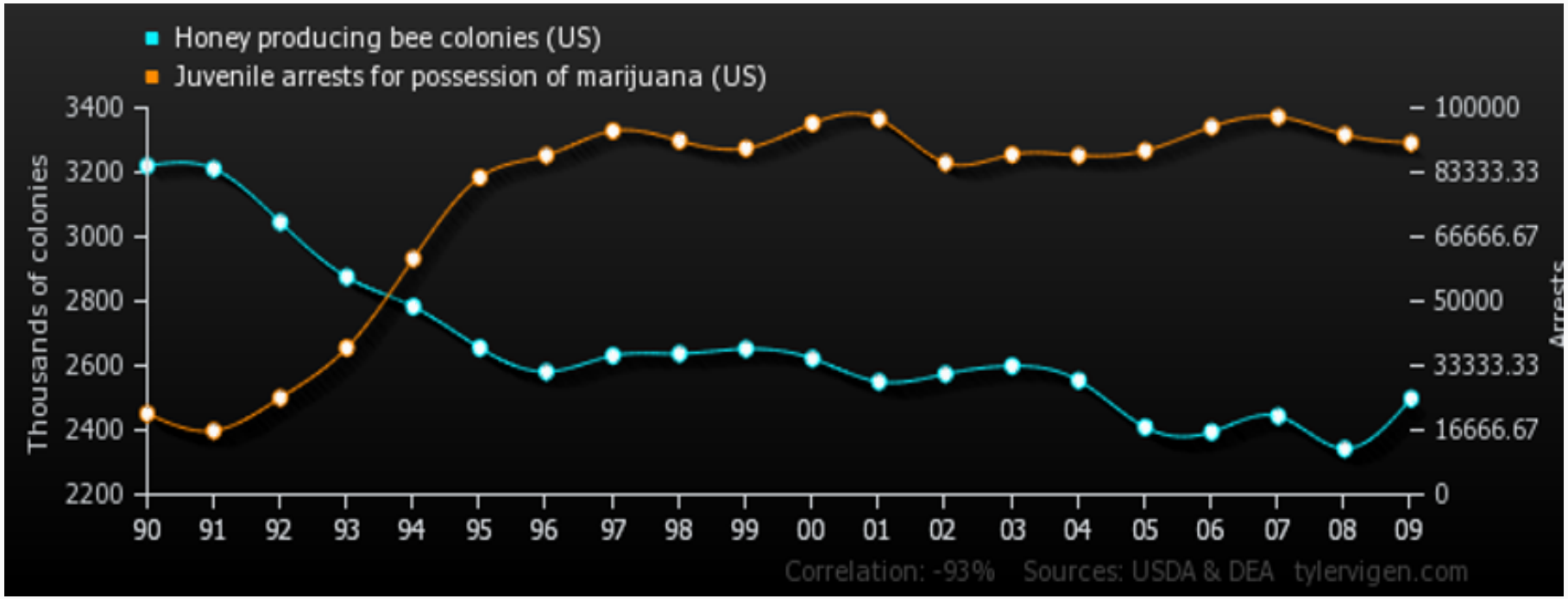
Correlación no es Causalidad (5)



Correlación: 0.992558

Fuente: http://www.tylervigen.com/

Correlación no es Causalidad (6)



Correlación: -0.933389

Fuente: http://www.tylervigen.com/

Existencia de Posibles Sesgos

El escoger una técnica de recolección de datos condiciona las conclusions posibles de obtener desde esos datos

Pregunta:

- Si un investigador utiliza una malla de 10 por 10 cm. para estudiar el tamaño de los peces en un lago... ¿Cúal es la conclusión obvia?

Respuesta a la pregunta:

- No hay peces menores en tamaño que un cuadrado de 10 por 10 cm.

Ejemplo Adicional:

- Si realizamos una encuesta en un supermercado a las 11.00 AM de un día Lunes... ¿A quién representan estas respuestas a la encuesta?
  - Probablemente a dueñas de casa, estudiantes haciendo la cimarra y jubilados

Idealmente se debe evitar el sesgo de selección de información...

Medir las Características de un Sistema Dado lo Afecta

Pregunta:

- Si quisieramos estimar el ingreso de una familia y le decimos a esa familia que el ingreso será estimado a partir de el consumo del grupo familiar... ¿Qué creen que pasará?

Respuesta a la pregunta:

- La gente tiende naturalmente a mentir acerca de su ingreso, o bien lo aumentan o lo disminuyen
  - En estos casos, organizaciones como el FMI que tienen que caracterizar ingresos lo aproximan a través del consumo, sin embargo, si es que se le dice al grupo familiar que esto es lo que se hará, el consumo del grupo familiar cambiará de su estado natural (puesto que existe la tendencia natural a aumentar o disminuir los ingresos)

Ejemplo Adicional:

- A nivel atómico, el uso de microscopios electrónicos afectan la trayectoria del objeto a observar puesto que los electrones “golpean” al observable cambiando así su trayectoria y velocidad

El Problema con Los Problemas (1)

“Si me dieran *una* hora para salvar el planta, gastaría *cincuenta y nueve* minutos definiendo el problema y *un* minuto resolviéndolo” Albert Einstein? (Fecha desconocida)

“Dame *seis* horas para cortar un árbol y gastaré las primeras *cuatro* afilando el hacha.” Abraham Lincoln? (Fecha desconocida)

La definición de problemas es clave en todo el proceso de la Ciencia de Datos, la siguiente lista muestra los aspectos mínimos que se cree deben ser cubiertos al identificar problemas:

- Hay que establecer la necesidad de solución
- Justificar la necesidad
- Hay que contextualizar el problema
- Escribir la declaración del problema

Algunos Mitos Comunes

Mitos Comunes (1)

- **La Ciencia de Datos es Para Genios Matemáticos o Ph.D.**
  - El gran requerimiento es en realidad saber **algo de estadística** o razonar en términos estadísticos
- **La Ciencia de Datos es Solo Aprender Herramientas**
  - El científico de datos debe **pensar fuera de esquemas tradicionales** para obtener soluciones y además conocer el negocio
- **El Científico de Datos Será Reemplazado muy Pronto por la Inteligencia Artificial**
  - A pesar de que las máquinas lleven el peso del trabajo, no son **capaces de interpretar** con criterio o sentido común los resultados
- **Los Cienítifcos de Datos Trabajan en Herramientas Sofisticadas Todo el Tiempo**
  - La herramienta no hace al científico de datos, más bien **es el proceso** que se sigue para obtener revelaciones de los datos

Mitos Comunes (2)

- **Los Científicos de Datos Sólo Trabajan en Volúmenes Grandes de Información**
  - Si bien es cierto que **volúmenes** grandes de información tienen potencial de producir mejores modelos, la estrategia para enfrentar un problema no está sólo circunscrita al uso de grandes volúmenes de información, también tenemos **velocidad**, **variedad** y **veracidad**
  - Con cualquiera de estas **V** presentes se puede utilizar ciencia de datos
- **Si Sabes Programar Entonces Puedes ser Científico de Datos**
  - Hay que recordar que el científico de datos está en la intersección de tres disciplinas. Si bien programar es una herramienta útil, aún se requieren **conocimientos matemáticos** y de **negocio** para ser científico de datos
- **Ciencia de Datos e Inteligencia de Negocios son lo Mismo**
  - Este es uno de los mitos más extendidos. A pesar de ciertas similitudes, la inteligencia de neogocios se concentra más en los aspectos operacionales y contextuales de una organización (clientes y audiencia por ejemplo). Por otro lado, la ciencia de datos se concentra mucho más en la **analítica predictiva** y la **identificación de patrones y revelaciones**.

Mitos Comunes (3)

- **Una Mayor Cantidad de Datos se Traduce en Mayor Precisión**
  - Más datos con el **análisis erróneo no implican mejor precisión**. Menos datos pero con mejor curatoría son a veces preferibles a bases de datos masivas de baja calidad
- **La Ciencia de Datos no Entrega Beneficios Económicos**
  - El proceso en sí permite no sólo caracterizar fenómenos en la organización sino que además permite una mejor **toma de decisiones** que se vé reflejada en mejores **eficiencias** y mayores **utilidades**
- **Los Científicos de Datos no son Científicos en Ningún Sentido Razonable**
  - Los científicos de datos utlizan el **método científico** en el desarrollo del trabajo investigativo, sin embargo, las herramientas que se utilizan requieren un cambio de pensamiento en la forma en que los modelos se crean y los descubrimientos se obtienen

¿Ciencia de Datos para Todos?

Discusión Acerca del Desarrollo de Habilidades

- Con todo lo que hemos discutido, podemos darnos cuenta de los siguientes hechos fundamentales:
  - La ciencia de datos ha llegado para quedarse
  - El proceso no puede ser automatizado a un 100%, aún posee un elemento de buen juicio que no puede ser reemplazado por máquinas o algoritmos
  - Se requiere un gran conocimiento del negocio donde se aplicarán los modelos, conocimiento muchas veces basado en la experiencia y díficil de replicar por herramientas automatizadas
  - Hay herramientas de fácil acceso que pueden ser utilizadas o en su defecto capacidad de desarrollo disponible
  - Muchos modelos que agregan valor no son demasiado complicados
    - De hecho hay toda una tendencia que apunta al uso de modelos **explicables**, pero eso es tema de otra presentación...
  - Lo más importante al aplicar los modelos es una buena compresión conceptual de los mismos, más que las matemáticas incolucradas puesto que las librerías usualmente encapsulan esos detalles y los ocultan del científico de datos

Aprender Haciendo

- Basado en lo anteriormente discutido, se cree que aprender haciendo es la mejor forma de desarrollar competencias dentro de las organizaciones
- Un esquema de desarrollo de competencias de este tipo comprende los siguientes elementos:
  - Sesiones de entrenamiento
  - En paralelo y extendiéndose más allá del término de las clases, se crean grupos, de preferencia heterogéneos y/o transversales a la organización
    - Cada grupo definirá un problema, recolectará datos, ajustará modelos y finalmente presentará los resultados de su investigación en un foro público
    - Cada grupo es asesorado por el consultor mediante lo que se denomina "Acompañamiento Experto" para reuniones semanales de avance y apoyo en la programación de soluciones
  - La actividad de cierre cuenta con un jurado conformado por el equipo de liderazgo de la organización quien determina la mejor solución de entre todas las presentadas
- Duración total del proceso: **6 meses en promedio**

Beneficios de Aprender Haciendo

La metodogía de desarrollo de competencias se ha desplegado en varias oportunidades tanto en Chile como en Perú y Australia, invariablemente se han observado los siguientes benficios comunes:

- La conformación de grupos heterogéneos permite que áreas en general disímiles puedan interactuar y conocerse
  - Esto entrega de manera instántanea un reconocimiento mutuo de habilidades dentro de las organizaciones y conocimiento de lo que el "otro" hace
- En la práctica, el modelo de aprendizaje haciendo se constituye en un proceso de **consultoría efectiva** en que varios problemas se abordan de manera simultánea
  - Los problemas al ser definidos desde las bases, son problemas que tiene sentido resolver porque son dolores existentes en la organización, lo que remueve muchas barreras de adopción
- No mucho tiempo después de terminada la actividad comienzan a aparecer los "campeones" que empiezan a buscar otros problemas y a implementar soluciones para ellos, usualmente se mantiene la relación con el consultor
- Algunos se motivan lo suficiente para aprender nuevas herramientas e incluso completar estudios de diplomado o magister
  - Esto es una increíble oportunidad de desarrollo de capital humano dentro de las organizaciones

Potenciales Obstáculos de Implementación

- El primer problema y tal vez el más importante que hay que sortear se relaciona a la organización semana a semana de los espacios de "acompañamiento experto"
  - En ciertas ocasiones aparecen reuniones de última hora que echan los planes por la borda con el consiguiente reagendamiento, esta parte del proceso es un poco desgastante
- El equipo de liderazgo de la organización debe entender esta actividad como una que forma parte de la jornada de trabajo y no como algo adicional a lo que ya se hace
  - En caso contrario, es muy probable que la gente no avance
- El compromiso en el tiempo hace que hayan posibilidades reales de que algunos estudiantes no terminen la actividad:
  - Por ejemplo en ciertos casos algunos enferman
  - Otros asumen responsabilidades laborales nuevas
- Un último problema no menor tiene que ver con las personalidades de los integrantes de los grupos, en ocasiones simplemente no es posible hacerlos trabajar
  - En esos casos, lo mejor es separar los grupos en subgrupos
  - Siempre hay llaneros solitarios que no quieren trabajar con nadie más, extremadamente brillantes y eficientes pero aislados

Conclusiones

Conclusiones

- Hay datos en todas partes
  - Esto no significa que podamos hacer frente a este océano de datos, todo depende del **problema** que queremos resolver
- Apretar botones en los paquetes de software no es díficil
  - Sin embargo, entender lo que se está haciendo y las implicaciones de los resultados saliendo del paquete de software es usualmente no trivial
- Los datos por sí solos no tienen valor, lo que hacemos con ellos es lo que agrega valor
- Hay una gran oportunidad, sin embargo, requiere *información transparente* y *modelos/herramientas* especializadas
- El elemento más importante que no debe olvidarse es la **gente**:
  - Ellos son los que últimamente ayudan a implementar las mejoras de los procesos
  - Ellos entienden el corazón del negocio a profundidad
  - Ellos son los que saben donde está la piedra en el zapato de la organización