Taller de Proyecto

Diplomado de Ciencia e Ingeniería de Datos
Departamento de Ciencias de la Computación
Universidad de Chile

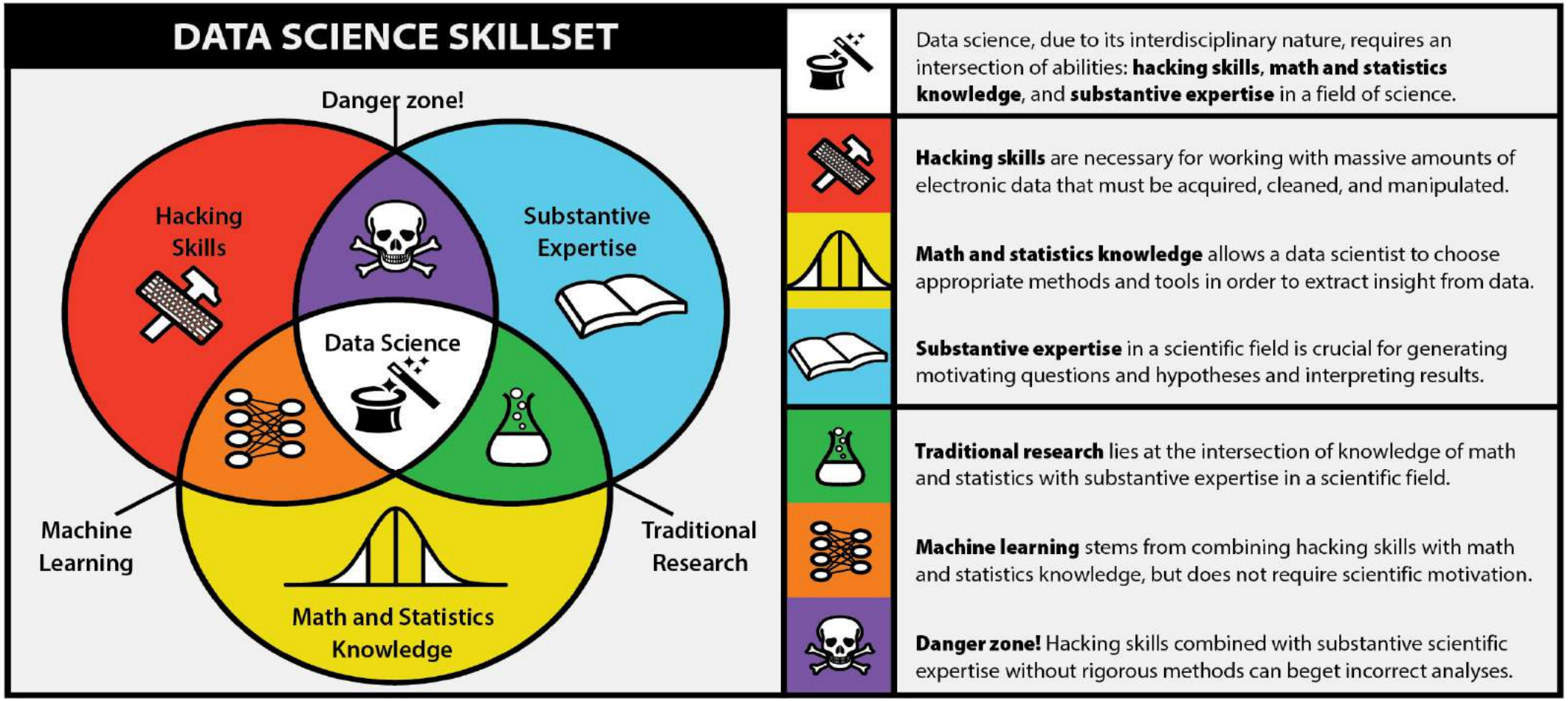
Juan Manuel Barrios
juan.barrios@impresee.com

Camila Álvarez
camila.alvarez@impresee.com

30 de agosto de 2018

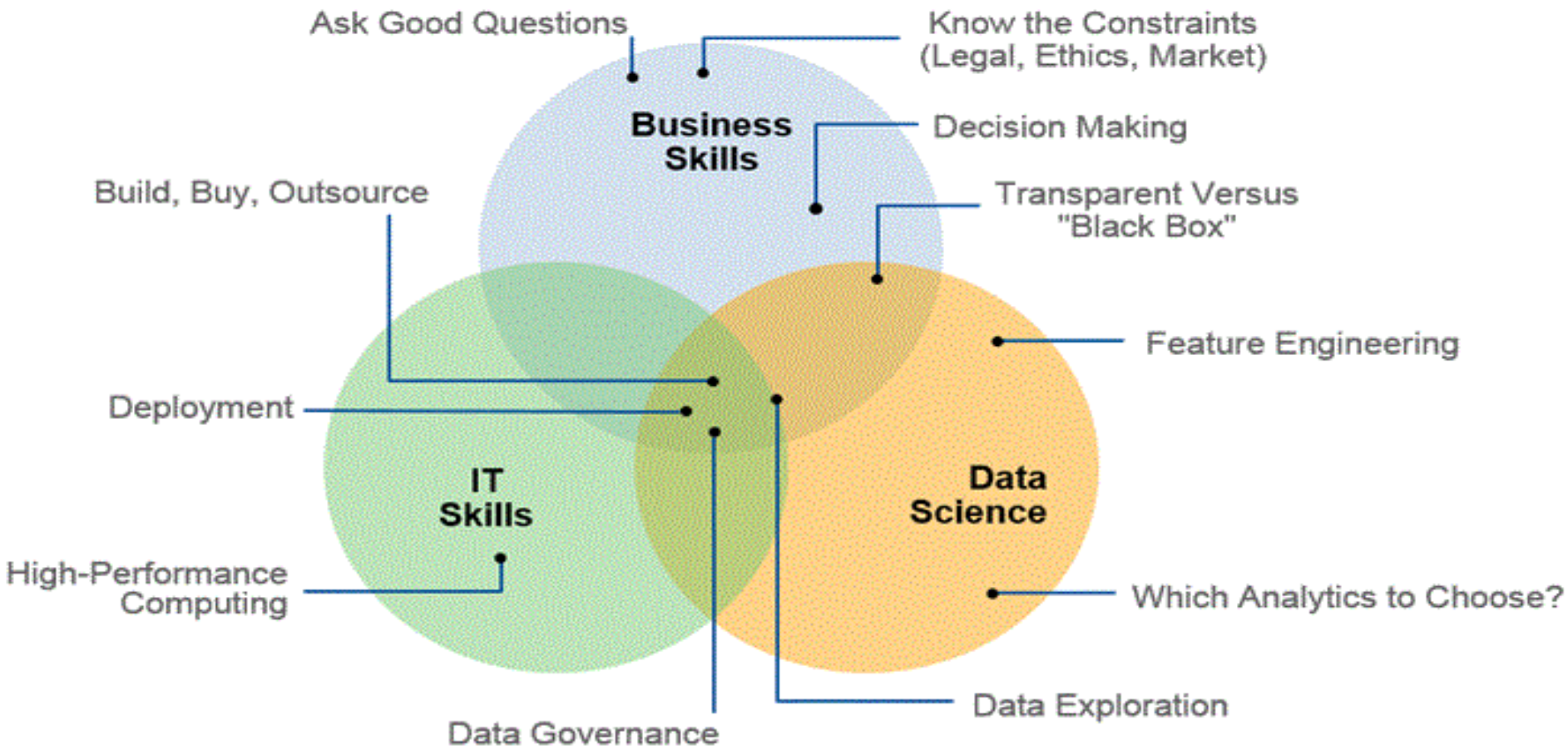
Introducción

Data Science



Fuente: berkeleysciencereview.com

Data Science en la Industria



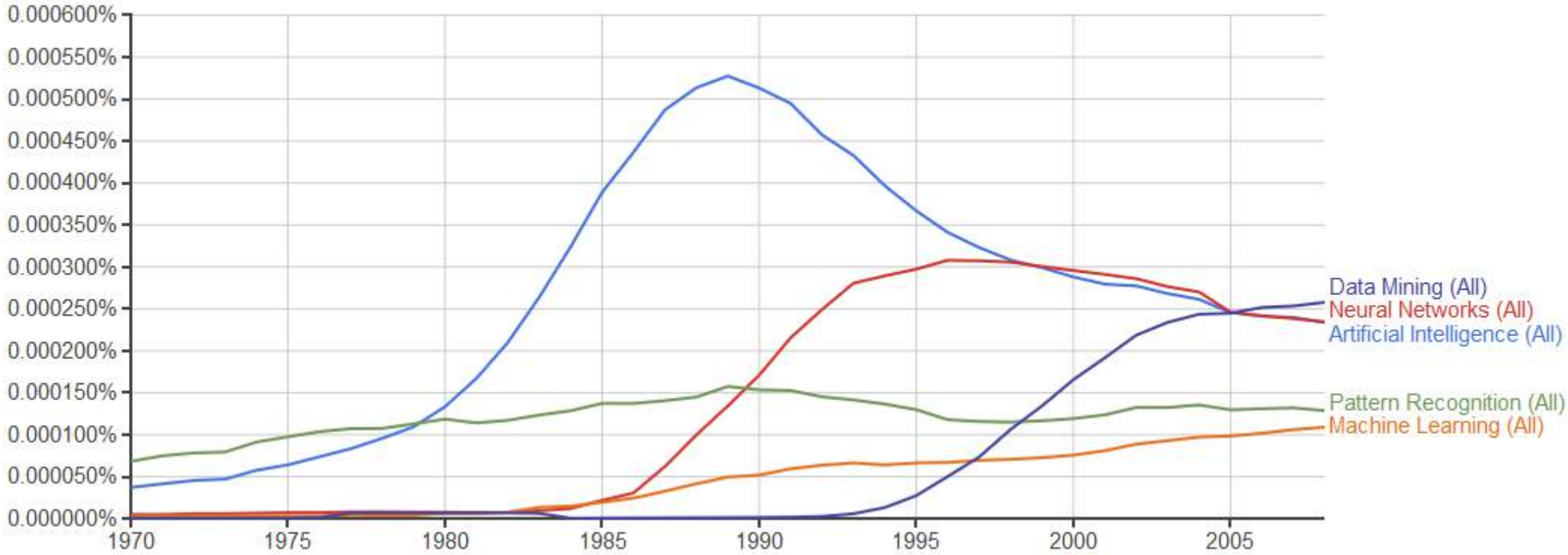
Source: Gartner (June 2015)

Muchas Áreas Relacionadas

- Artificial Intelligence
- Pattern Recognition
- Machine Learning
- Deep Learning
- Data Mining
- Data Science
- Data Visualization
- Information Retrieval
- Big Data

Uso de Términos (1970-2008)

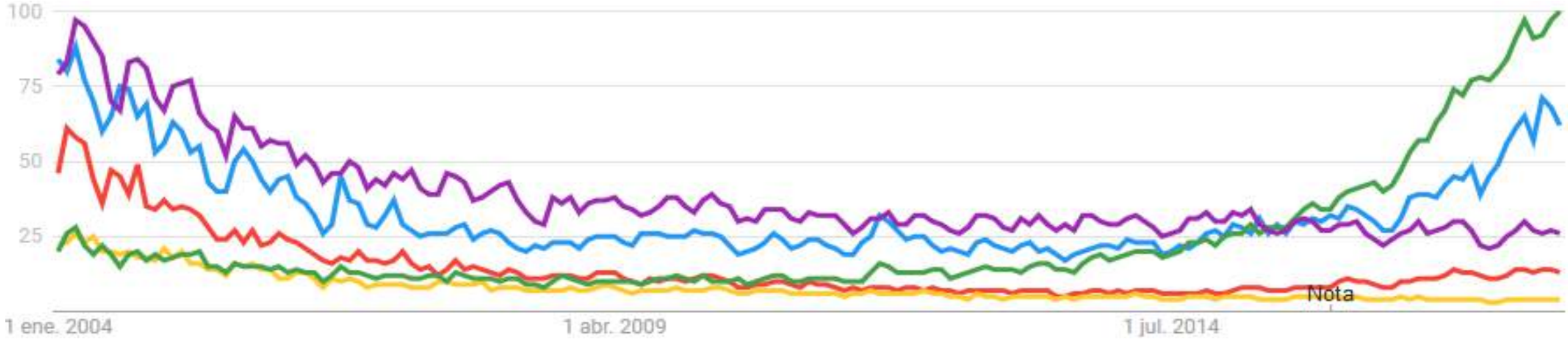
- Google Books (apariciones en libros)



Fuente: https://books.google.com/ngrams

Uso de Términos (desde 2004)

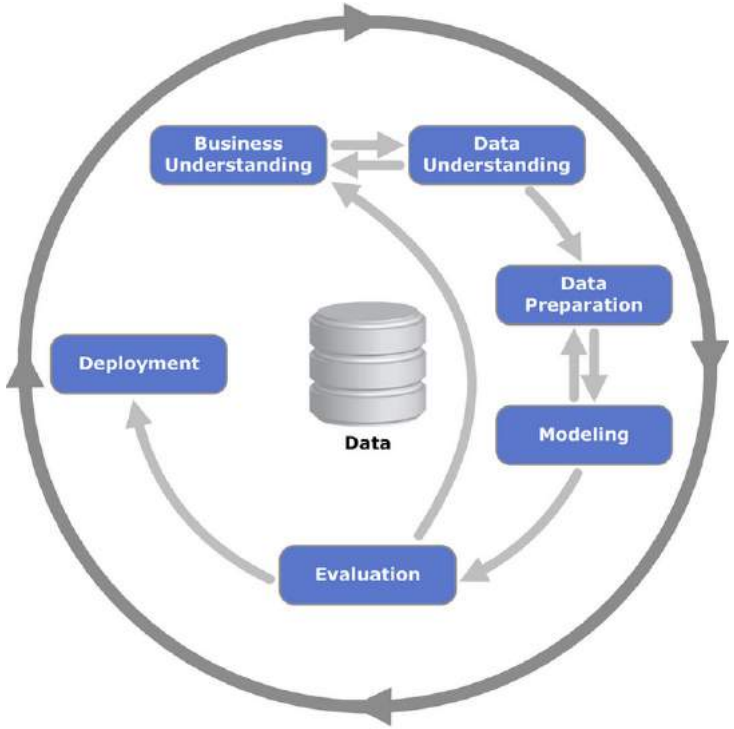
- Google Trends (términos en el buscador)



Fuente: https://trends.google.com/trends/

Proyectos de Análisis de Datos

- **Cross-Industry Standard Process for Data Mining (CRISP-DM)**
  - Define un ciclo de vida típico para proyectos de análisis de datos
    - Entendimiento del Negocio (Business Understanding)
    - Entendimiento de los Datos (Data Understanding)
    - Preparación de Datos (Data Preparation)
    - Modelamiento (Modeling)
    - Evaluación (Evaluation)
    - Instalación (Deployment)



Preparación de Datos

- Se dice que hasta un 80% del esfuerzo de un proyecto de análisis datos corresponde a preparar y limpiar datos
  - Algoritmos no funcionan muy bien con datos crudos (sin limpiar)
  - La limpieza de datos no puede ser totalmente automatizada
  - Algoritmos no funcionan muy bien con demasiados atributos ("Curse of Dimensionality")
  - La integración y enriquecimiento de datos se debe realizar con sabiduría
- Data Management: Todas las disciplinas relacionadas a la administración y gestión de los datos: mantención de versiones, permisos (governance), seguridad, etc.

Corrección de Datos

- **Objetivo**: Modificar los datos originales (raw data) para poder ser cargados y procesados
  - Cambiar formato de archivos (csv, xls, xml, json)
  - Corregir codificación de texto (ASCII, ISO-8859-1, UTF-8)
  - Definir tipos de dato (número, string, fechas, valores nominales, binarios, texto libre)
  - Formato de fechas (dd-mm-yyyy, yyyymmdd, mm/dd/yy)
  - Muestreo aleatorio del dataset (si es muy grande)
- Se debe tener precaución de no eliminar información valiosa ni introducir sesgos de selección
- **Resultado**: Datos técnicamente correctos que pueden ser leídos y cargados con algún software

Limpieza de Datos (1)

- **Objetivo**: Convertir datos técnicamente correctos en datos consistentes
  - Corregir valores especiales (debido a restricciones not-null o datos de pruebas).
    - Fecha: 1-1-1900, 9-9-1999, 1-1-1970
    - RUT: 1-9, 2-7, 11111111-1
  - Corregir valores inválidos semánticamente. Ej. montos negativo, edad mayor a 100 años
  - Convertir unidades (metros, grados, dólares/pesos)
  - Corregir series numéricas (muestrear, suavizar, etc.) Ej. suavizar la temperatura del día o valor del dólar
  - Decidir el completado automático de valores faltantes (imputación)
  - …

Limpieza de Datos (2)

  - Conciliar valores/nombres/códigos inconsistentes, corregir typos
    - "Aeropuerto Comodoro Arturo Merino Benítez" vs "Aeropuerto de Santiago" vs "SCL"
    - "Avenida Libertador General Bernardo O'Higgins" vs "Alameda" vs "Av. Ohigins" Normalización de valores. Ej: convertir en porcentajes o escalar valores a [0,1]
  - Buscar valores outliers y decidir si se deben eliminar (caso irrelevante o erróneo) o se deben mantener (caso relevante al negocio)
  - Buscar registros incompletos y decidir si se deben eliminar o mantener
- Se debe tener precaución de no eliminar información valiosa ni introducir sesgos de selección
- **Resultado**: Datos limpios que pueden ser usados para análisis. Notar que durante la exploración se pueden descubrir más valores a limpiar y/o corregir.

Integración de Datos

- **Objetivo**: Enriquecer los datos con fuentes externas
  - Conciliar referencias a objetos idénticos (Uso de URIs)
  - Conciliar inconsistencias de unidades. Ej.: dólar vs pesos
  - Conciliar inconsistencias de resolución. Ej.: Datos por trimestre vs datos anuales
  - Conciliar inconsistencias de datos. Ej.: Datos contradictorios de pobreza o cesantía
- Mantener la provenance de cada dataset
  - Provenance: Historia de dueños (ownership) y el tipo de licenciamiento
  - Manejar la actualización de datos
- **Resultado**: Datos que consideran información exógena para lograr un mejor análisis

Fuentes de Datos

- Notar las variedad de posibles fuentes de datos y sus características:
  - Datos obtenidos de sistemas de información (bd relacionales)
  - Datos generados por aparatos físicos, e.g. IoT (series)
  - Datos generados por personas manualmente (textos libres)
  - Datos generados por logs de uso de software
  - Datos obtenidos desde streams online, e.g. twitter
  - Datos obtenidos por descargas de archivos desde sitios públicos como data.gob.cl
- Data Curation: Se refiere a las actividades realizadas para la integración, selección y publicación de datos
- Data Lineage: Metadatos del dataset, como origen, movimientos, transformaciones que permiten su trazabilidad

Taller de Proyecto 2018

Taller de Proyecto

- **Objetivo del Taller**: Desarrollar un proyecto práctico de análisis de datos reales, aplicando las técnicas y herramientas aprendidas en los cursos del diplomado
- Grupos de máximo **cuatro** personas
- ¿Qué se espera del proyecto?
  - **Analizar** un conjunto de datos de tamaño razonable
  - **Aplicar** conceptos vistos en el diplomado
  - **Extraer información** interesante de los datos

Datos a usar en el Proyecto

- Los datos pueden ser **públicos** o **privados**
  - En caso de datos privados al menos se debe poder mostrar estadísticas, análisis y conclusiones
- Los datos se deben **enriquecer** con datos externos públicos
  - Ejemplo: información de comunas, ingresos promedios, noticias, variables macroeconómicas del país, etc.
- Notar que el proyecto se enfoca en **buscar** y **descubrir** patrones en vez de comprobar hipótesis preconcebidas
- ¿Qué herramientas usar para el proyecto?
  - Es posible usar **cualquier herramienta** a elección del grupo

Tareas del Proyecto

- El Taller se compone de cuatro tareas:
  - **Tarea 0**: Definición del proyecto
  - **Tarea 1**: Preparación y exploración de datos
  - **Tarea 2**: Métodos no supervisados y visualización
  - **Tarea 3**: Métodos supervisados

Tarea 0: Definición del Proyecto

- Definir los datos a usar y objetivo del proyecto
  - Definir el **conjunto de datos** sobre los que desea trabajar
  - Definir una **entidad** interesante de estudiar: cliente, paciente, factura, etc.
  - Definir un **atributo**, **evento** o **comportamiento** de esa entidad que se desee entender mejor, e.g.: clientes que compran una sola vez, pacientes con cierta enfermedad, facturas que no son pagadas, etc.
- Resultados esperados:
  - Mostrar **características** del dataset, **volúmen** de datos, **atributos** existentes y posibles datos externos a utilizar
  - Definir **entidad de interés** a usar para clustering y para clasificación junto con **posibles atributos** para sus vectores característicos

Tarea 1: Exploración de Datos

- Definir la estructura de los datos, corregir y limpiar datos
- Conocer los datos disponibles: rangos, distribuciones, correlaciones
  - Usar estadísticas simples y/o exploración visual
- Resultados esperados:
  - Entender la información que proporciona el dataset
  - Identificar los valores que podrían contener información interesante
  - Identificar los rangos de valores posibles
  - Identificar potenciales correlaciones entre valores
  - **Crear gráficos simples** (histogramas, scatterplots, boxplots, tablas de frecuencias, etc.) mostrando información de atributos y valores outliers

Tarea 2: Métodos No Supervisados y Visualización

- Uso de algoritmos de clustering y evaluación visual de los resultados
- Crear agrupaciones por medio de algoritmos de clustering:
  - Organizar el dataset como matriz: cada fila representa una instancia de la entidad a estudiar y las columnas son sus atributos
  - Definir un criterio de comparación entre instancias
  - Ejecutar un algoritmo de clustering para agrupar instancias parecidas
  - Medir la calidad de la agrupación obtenida
  - Agregar un atributo **"M"** con el identificador de grupo al que fue asignado cada instancia
- Análisis visual de los grupos obtenidos:
  - Visualización de las instancias en el dataset usando varios atributos (incluyendo **"M"**)
  - Visualización estática y/o interactiva
  - Verificar si **"M"** tienen algún sentido para el negocio

Tarea 3: Métodos Supervisados

- Uso de métodos de clasificación y evaluación objetiva de resultados
- Se debe usar un clasificador para predecir un atributo nominal en función de los otros atributos:
  - Organizar el dataset como matriz: cada fila representa una instancia de la entidad a estudiar y las columnas son sus atributos
  - Elegir un atributo **"X"** como objetivo de la predicción
    - **"X"** debe contener datos nominales (Si/No) o categóricos (Nivel 1,2,3)
  - Dividir las instancias en dos subconjuntos: entrenamiento y test
  - Con datos de entrenamiento crear un clasificador **C** que utiliza **"X"** como valor de la clase
- Evaluación objetiva del clasificador:
  - Con datos de test, ignorar **"X"** y utilizar el clasificador **C** para calcular un atributo **"Xpredicted"** con la clase entregada por **C**
  - Comparar los valores **"X"** y **"Xpredicted"** y evaluar la calidad de la predicción de **C** usando matrices de confusión y otros indicadores

Evaluación

- Las Tareas 1, 2 y 3 se evaluará con dos notas:
  - Nota de trabajo en clases (incluye trabajo en sesiones y presentación oral)
  - Nota de informe escrito
- El informe escrito es un PDF que se entrega al final del taller
  - Corresponde a las mismas slides de la presentación oral incluyendo posibles correcciones y mejoras
- La nota final del Taller es el promedio de las seis notas
- **Requisito de Asistencia**: Para realizar cada presentación oral el grupo **debe** estar completo

Calendario y Presentaciones

1. **Jueves 30-Agosto** (después de programación estadística)
2. **Martes 25-Septiembre** (después de minería de datos)
   - Tarea 0: Definición del Proyecto
3. **Miércoles 17-Octubre** (después de aprendizaje de máquinas)
   - Tarea 1: Exploración de Datos
4. **Martes 6-Noviembre** (después de recuperación de información)
5. **Martes 28-Noviembre** (después de visualización de datos)
   - Tarea 2: Clustering y Visualización
6. **Lunes 17-Diciembre** (después de big data)
7. **Martes 18-Diciembre**
8. **Jueves 20-Diciembre**
   - Tarea 3: Clasificación
9. **(Máximo) Jueves 27-Diciembre**
   - Envío de Informe Escrito