¿Qué es la ciencia de datos?

Julio Waissman Vilanova

Departamento de Matemáticas
Universidad de Sonora

14 de marzo de 2016

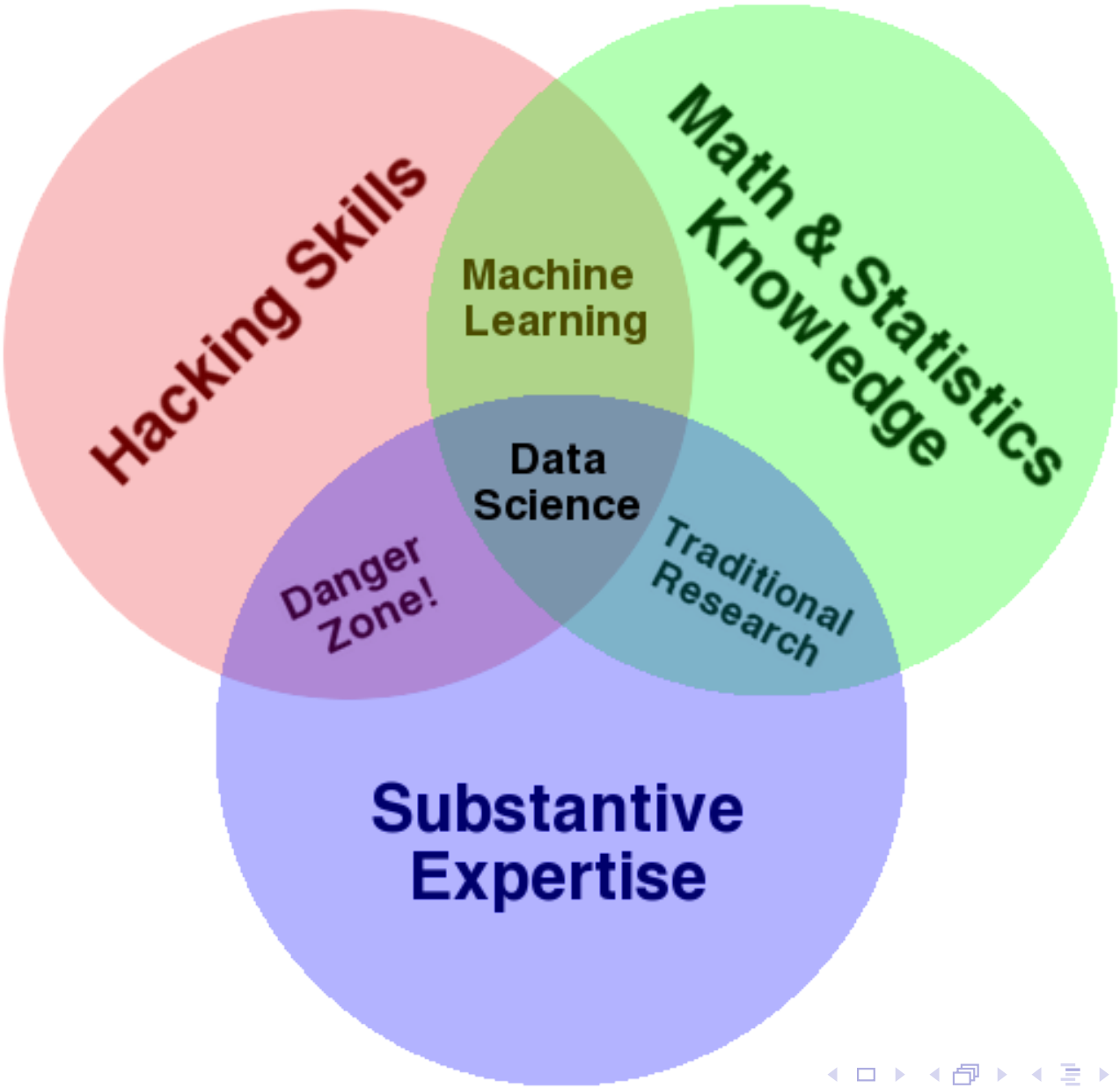
Plan de la presentación

1. Definición
2. ¿Que se hace en ciencia de datos?
3. Herramientas para Aprendizaje máquina y ciencia de datos

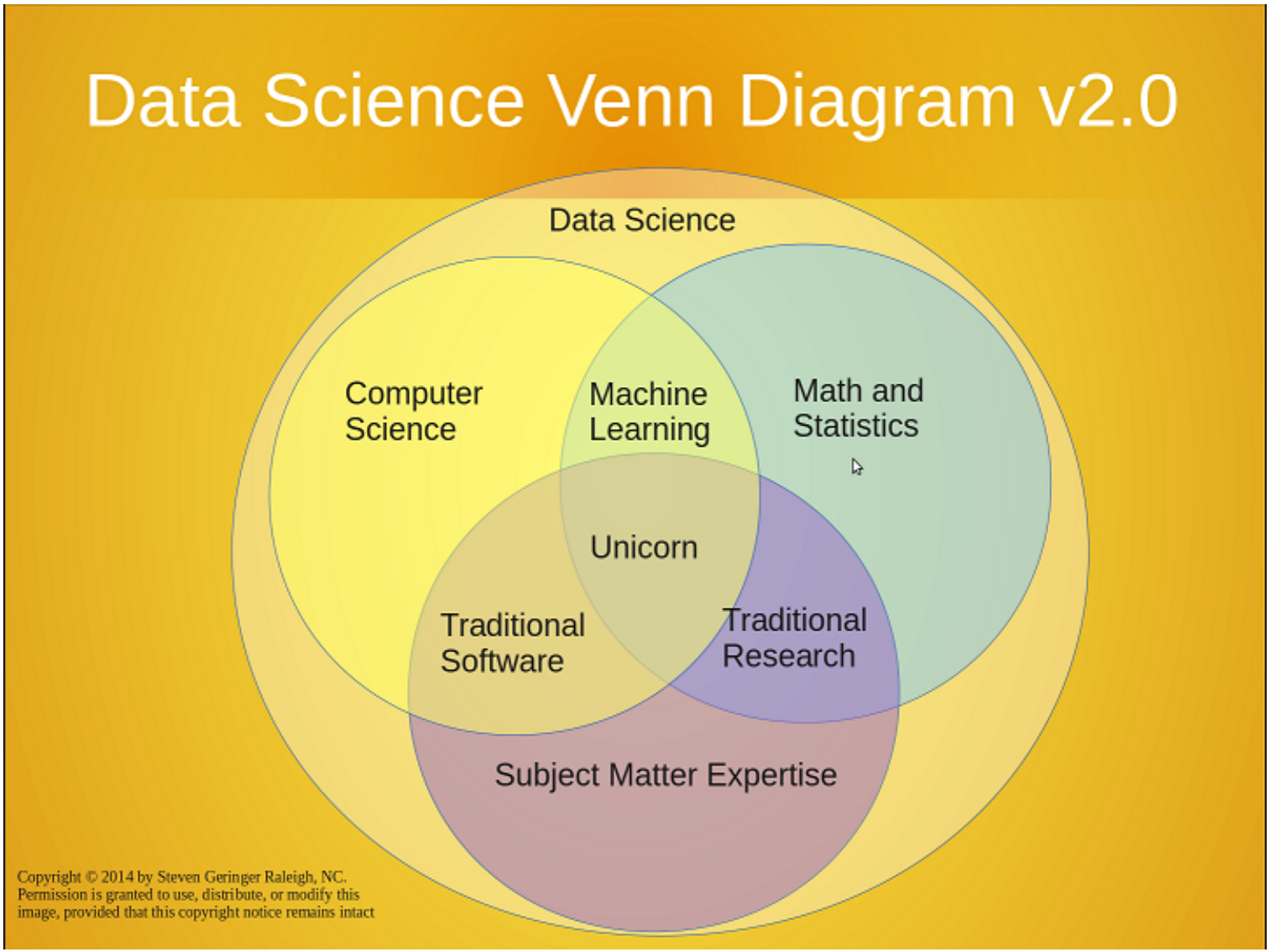
¿Que es la ciencia de datos?

- ¿Una palabra de moda?
- ¿Un ingeniero de software que sabe estadística o un estadístico que sabe ingeniería de software?
- ¿Usar reconocimiento de patrones?
- ¿Cualquier cosa que tenga que ver con “big data” y/o “data minning”?

Una definición esquemática



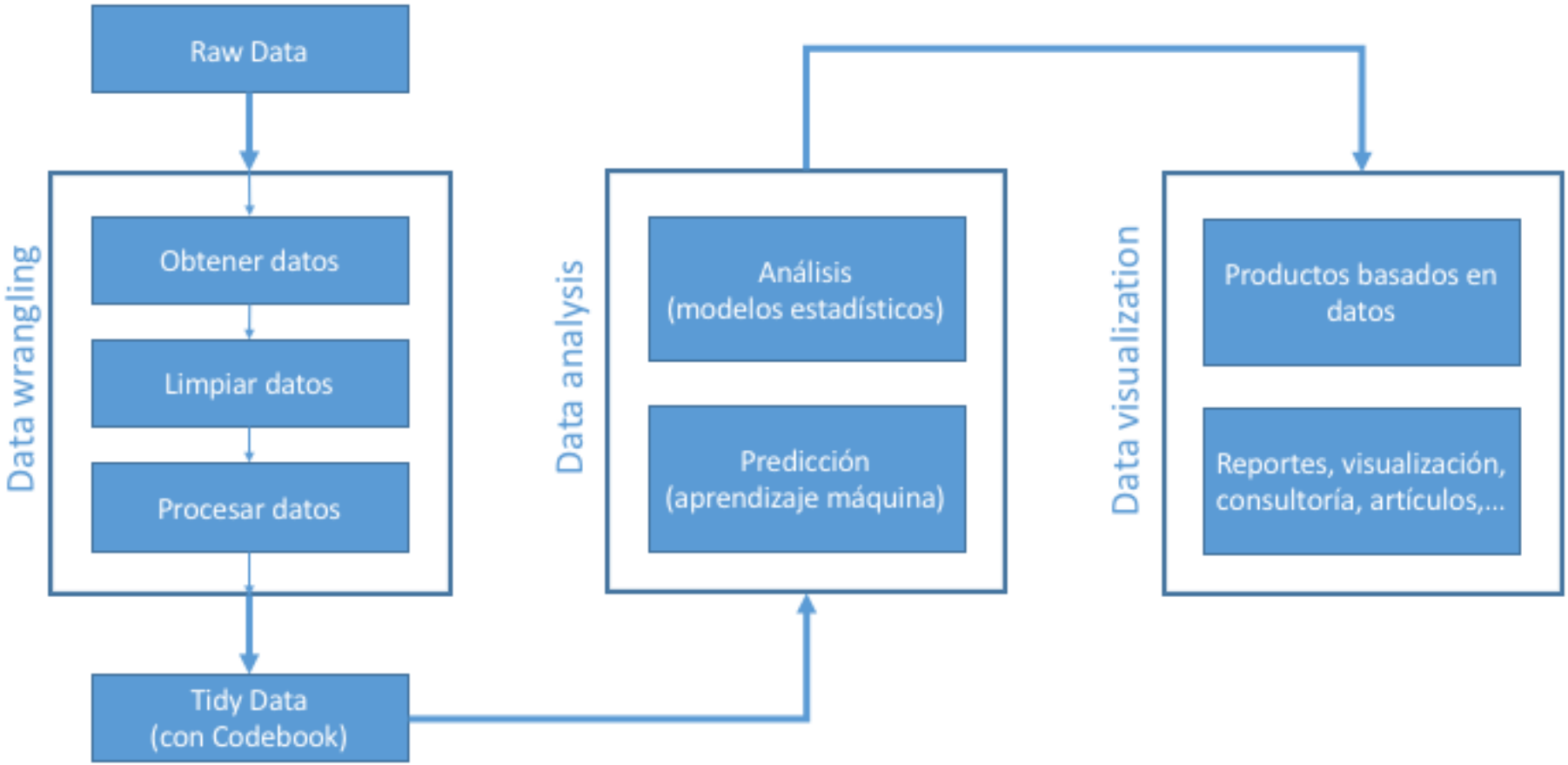
El diagrama revisado



¿Y que significa *substantive expertice*?

- Saber que preguntas realizar
- Interpretar correctamente los datos
- Comprender la estructura de los datos

¿Que se hace en ciencia de datos?



Data wrangling

- Obtener y normalizar datos de diversas fuentes
- La fase menos “glamorosa” de la ciencia de datos
- Ocupa posiblemente entre el 60 % y 70 % del tiempo de desarrollo
- La obtención y limpieza de datos debe ser muy bien documentada

Data analysis

- Mayor rigor matemático
- La fase más “glamorosa” de la ciencia de datos
- Cuidado con uso de algoritmos y métodos sin comprensión del proceso
- El análisis debe de ser reproducible

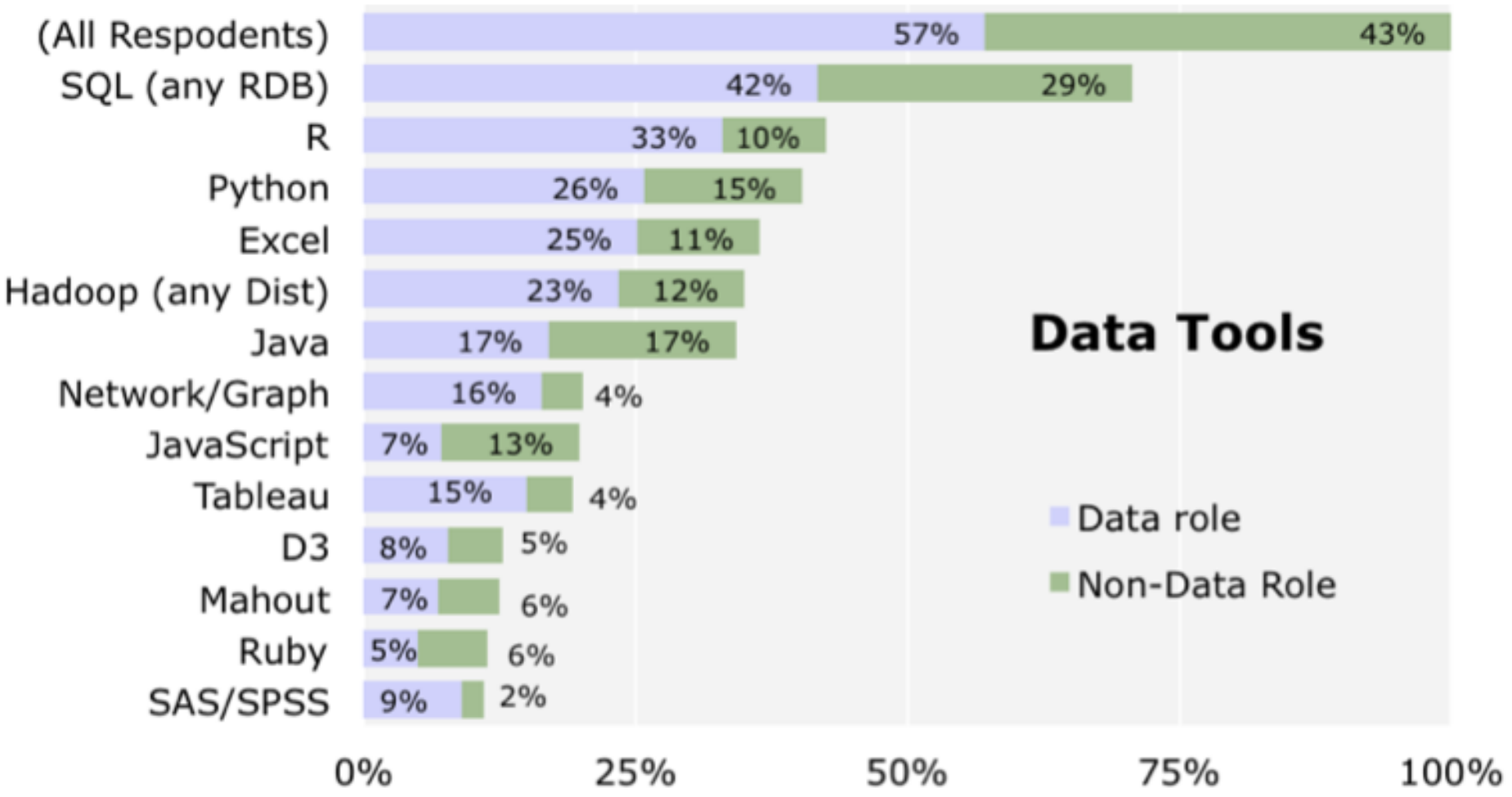
Data visualization

- Comunicación efectiva de ideas complejas con claridad precisión y eficiencia
- La fase menos formalizada (HMI, UX, ...)
- Típicamente no se le da la importancia que merece
- El éxito final de un producto basado en datos depende en gran medida de esta última fase

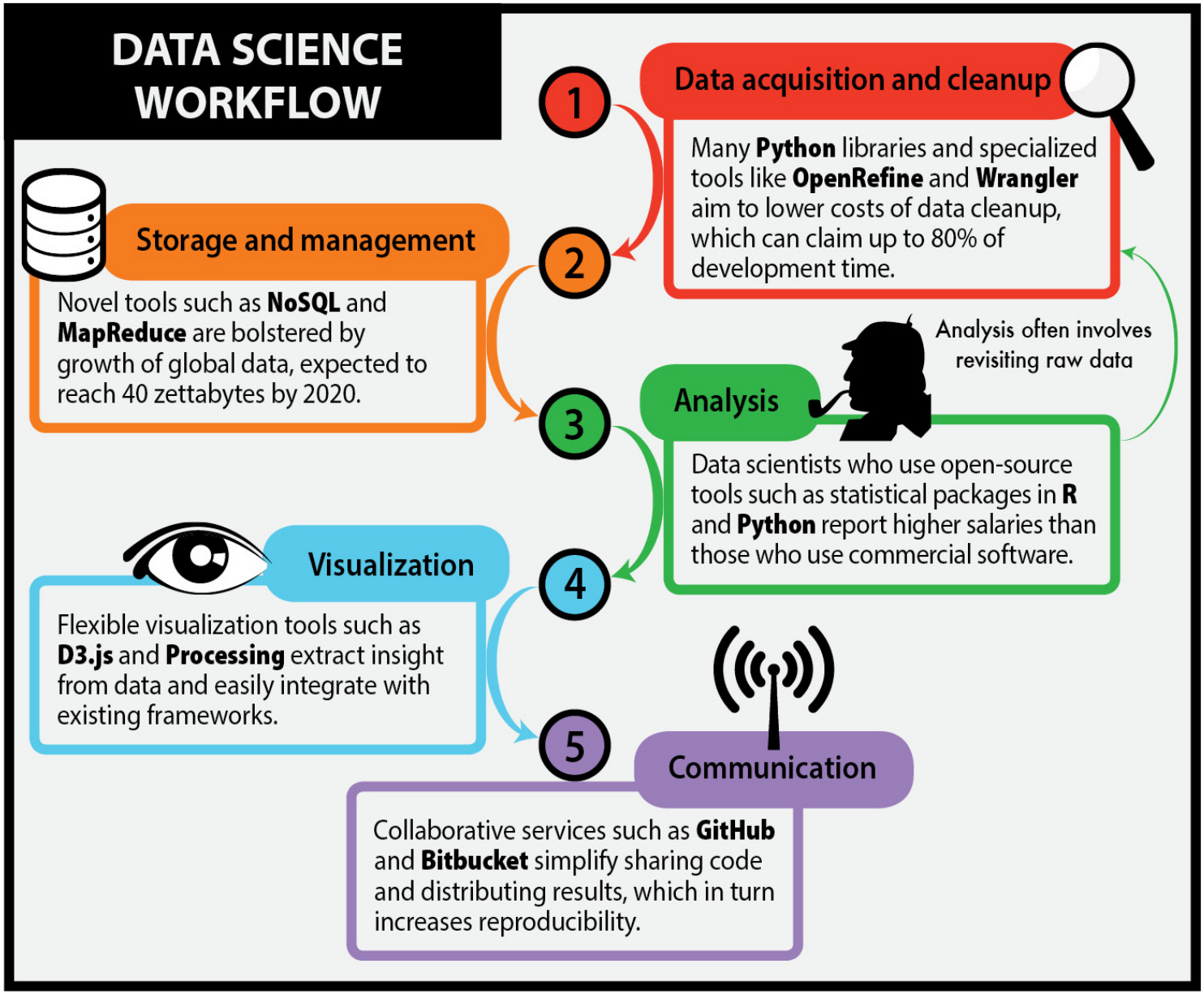
Grandes volumenes de datos

- El *Big Data* es uno de lo motivos por lo que la ciencia de datos se a vuelto popular, pero no el único.
- Es muy importante tener en cuenta la escalabilidad de un producto basado en datos en la fase de producción.
- Dos problemas básicos, el manejo de grande volumenes de datos, y el manejo de datos *semiestructurados*.
- Manejo de tecnologías emergentes (bases de datos noSQL y uso de métodos *Map Reduce*.

¿Que herramientas se usan?



¿Como se usan?


DATA SCIENCE WORKFLOW
1
Data acquisition and cleanup
Many Python libraries and specialized tools like OpenRefine and Wrangler aim to lower costs of data cleanup, which can claim up to 80% of development time.
2
Storage and management
Novel tools such as NoSQL and MapReduce are bolstered by growth of global data, expected to reach 40 zettabytes by 2020.
3
Analysis
Analysis often involves revisiting raw data
Data scientists who use open-source tools such as statistical packages in R and Python report higher salaries than those who use commercial software.
4
Visualization
Flexible visualization tools such as D3.js and Processing extract insight from data and easily integrate with existing frameworks.
5
Communication
Collaborative services such as GitHub and Bitbucket simplify sharing code and distributing results, which in turn increases reproducibility.

Python

- Lenguaje de uso general
- Libre, con gran cantidad de bibliotecas (modulos)
- Fácil de instalar y de mantener, sin embargo lento.
- Solución, uso de biblotecas con funciones precompiladas en lenguajes de bajo nivel (C y Fortran principalmente)
- Distribución Anaconda de *Continuum Analytics*

Empecemos con Python

Vamos a instalar y probar el lenguaje Python utilizando la distribución Anaconda y la utilidad `conda`

Las herramientas de base

- `numpy`
  - Arreglos multidimensionales (matrices y vectores)
  - Funciones matemáticas
- `pandas`
  - Manejo de datos para el análisis
  - Basado en la clase `ndarray` de `numpy`
- *Jupyter*
  - Documentos para análisis reproducible en ciencia de datos
  - Abierto para más de 40 lenguajes de programación