Hands-On Data Science with R

Desde el Procesamiento a la Minería de Datos

Camilo A Herrera R

18 de Junio de 2016

Extracción y Generación de Conocimiento a través de los Datos

Estadístico - Universidad del Valle

Sp. Data Science & Sp. Executive Data Science - Johns Hopkins Bloomberg

Msc candidate - Biometria - Universidad de Buenos Aires

twitter: @hr_mr_zork - web: http://camiloherrera.co/

Email: ch@camiloherrera.co

Cronograma

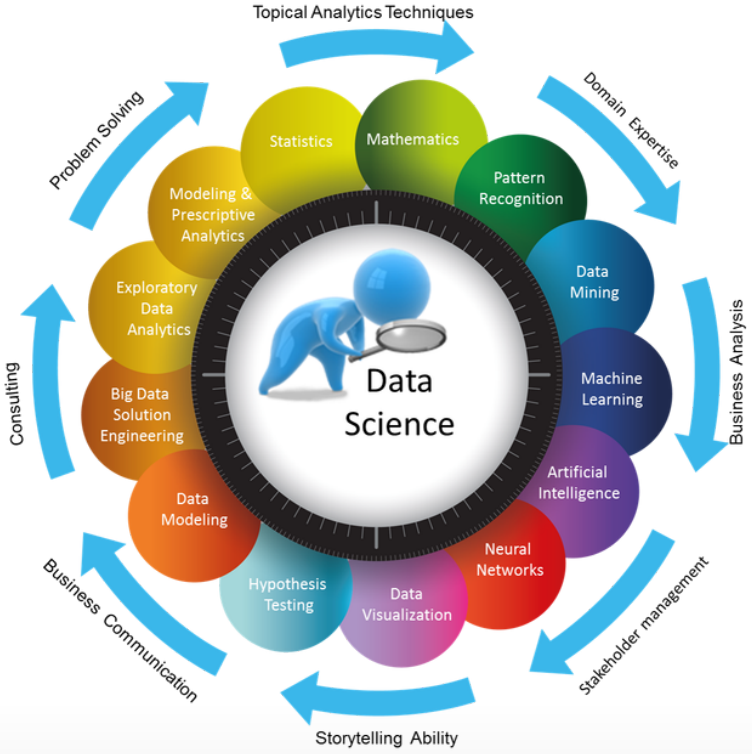
- Apertura del Evento
- Socialización R-Users MeetUp
- Sesión Introductoria Data Science con R
- Conferencias MeetUp:
    - Cleaning Data and Merging Data Sources with R -> Daniel Valencia
    - Visualizando datos con ggplot -> Maria Isabel Arce
- Taller en R (Hands-on)
- Cierre

Usa el Hashtag en twitter #RUsersCali

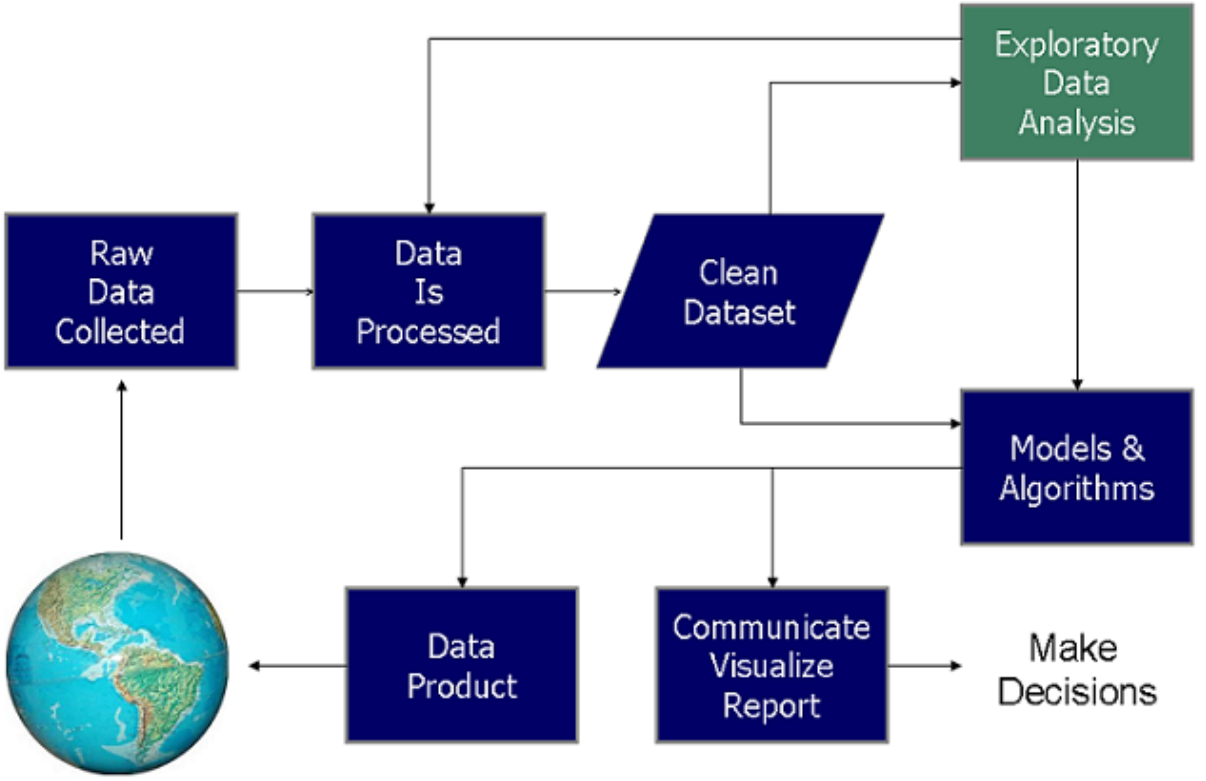
Ciencia de Datos

Que es Ciencia de Datos

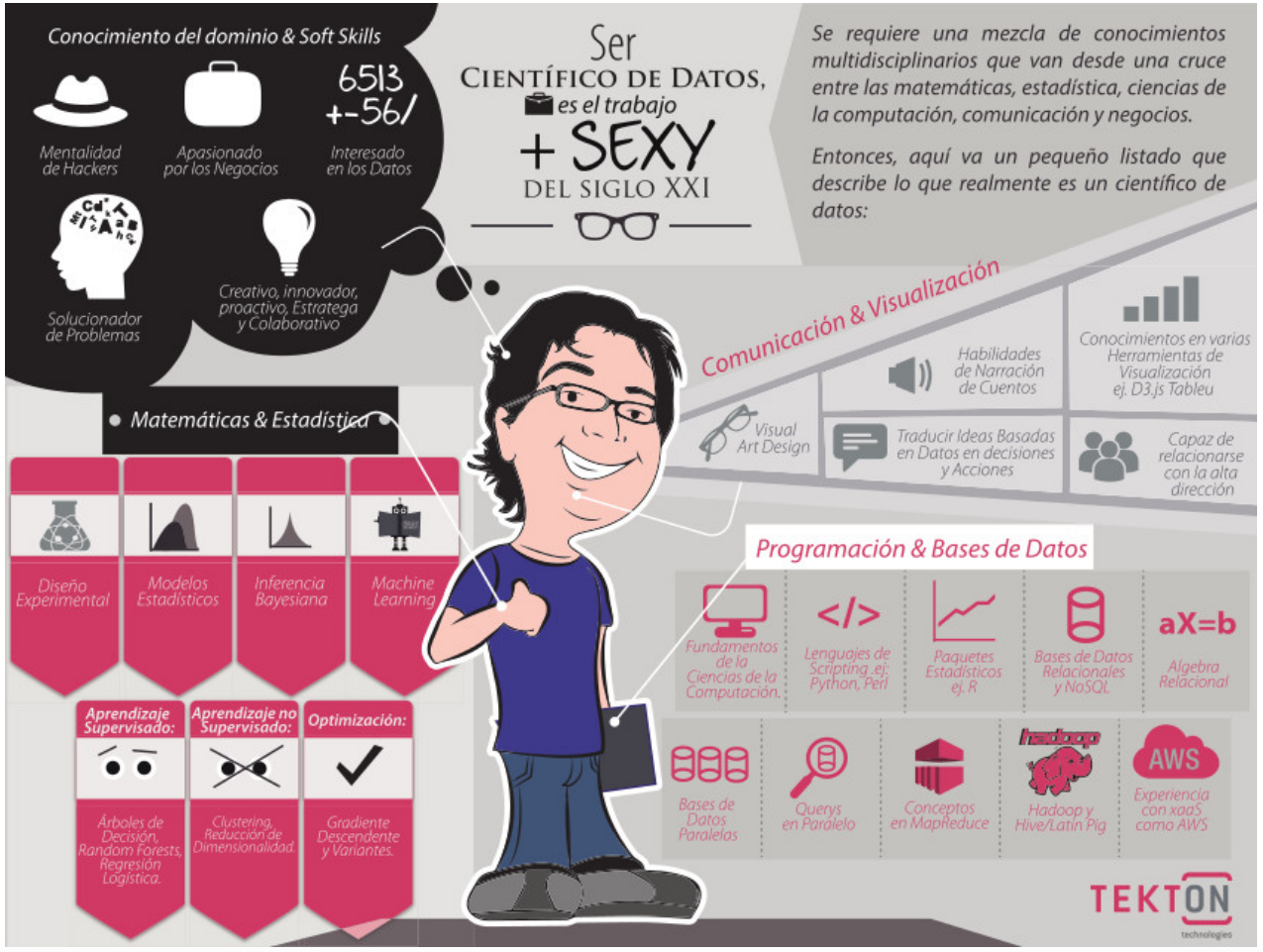
La Ciencia de datos es un campo interdisciplinario que involucra los procesos y sistemas para extraer conocimiento o un mejor entendimiento de grandes volúmenes de datos.



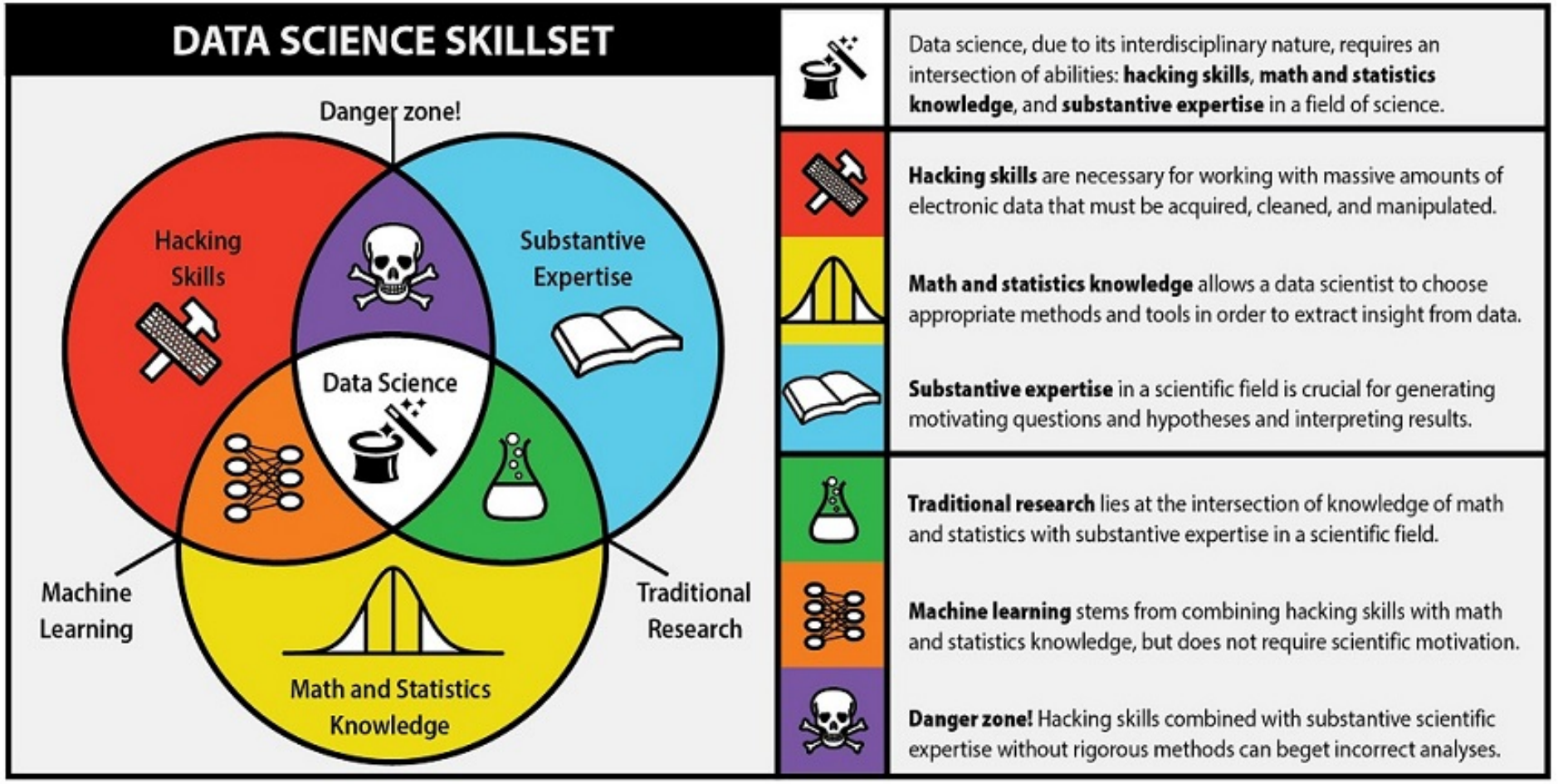
Data Science Process



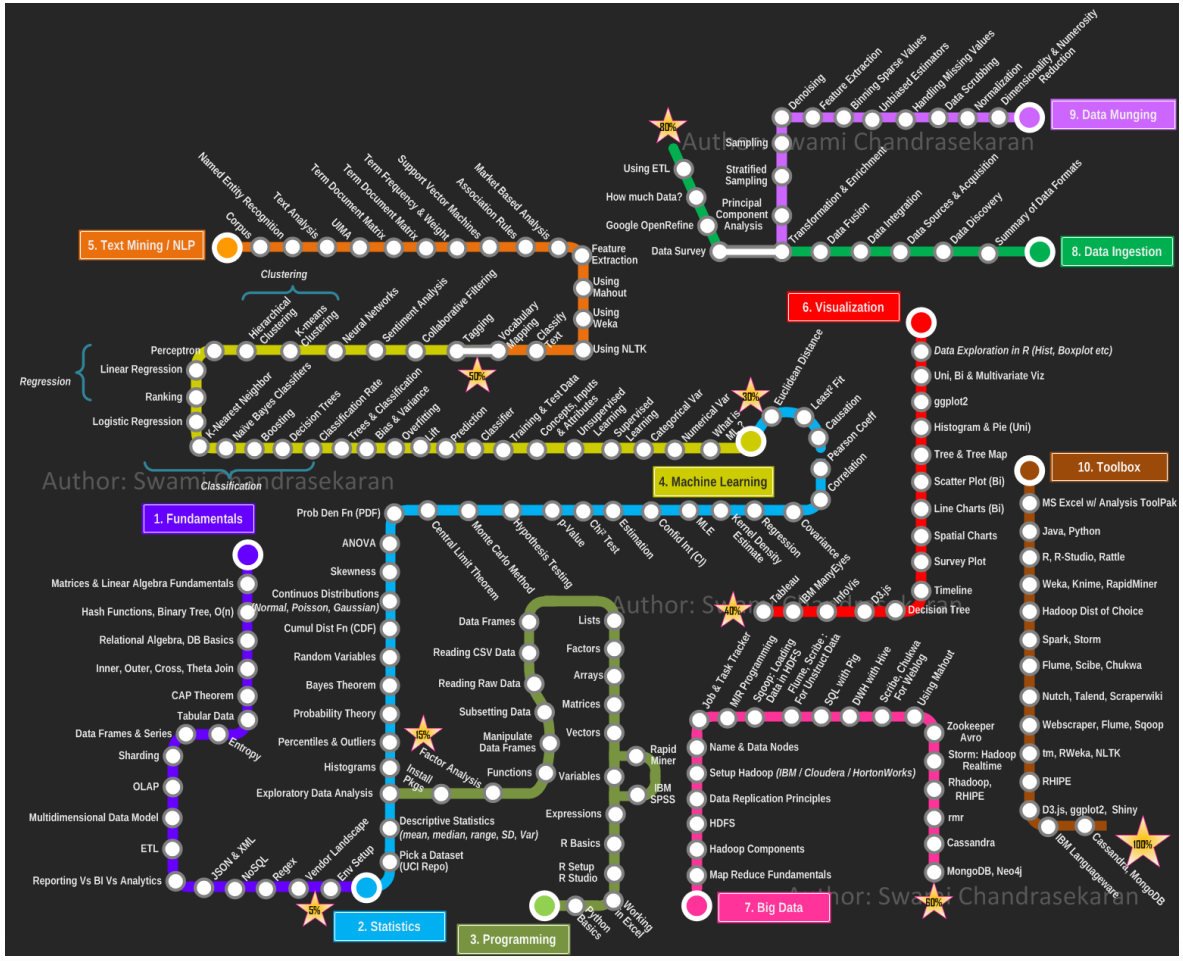
El Científico de Datos



Habilidades de un Científico de Datos



Una Ruta para ser Científico de Datos



Equipos de Data Science

La ciencia de datos es un proceso que requiere un esfuerzo importante, por lo tanto se necesita de un grupo con un comportamiento equivalente a un equipo deportivo:

Equipos de Data Science

Un equipo de ciencia de datos está compuesto por:

1. Ingenieros de Datos
   - Arquitectura de Datos
   - Infraestructura de Datos
2. Data science
   - Limpieza de Datos
   - Análisis y Comunicación
3. Líder del Equipo de Datos

R
Lenguaje de Programación

¿Por qué R?

- R Es gratis
- Cuenta con un amplio conjunto de paquetes
- Acceso a los datos
- Limpieza de datos
- Análisis
- Generacion de Reportes
- Tiene uno de los mejores entornos de desarrollo - Rstudio http://www.rstudio.com/
- Tiene un increíble ecosistema de desarrolladores
- Los paquetes son fáciles de instalar y "juegan muy bien juntos"

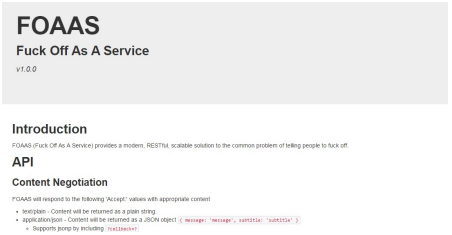
¿Por qué R?

R es considerado la "lingua franca" de la ciencia de datos, por esta razón las empresas se están adaptando rápidamente a R para desarrollar sus programas y productos de "Data Science".

Paquetes en R

Paquetes R en CRAN

Existen paquetes para todo!!

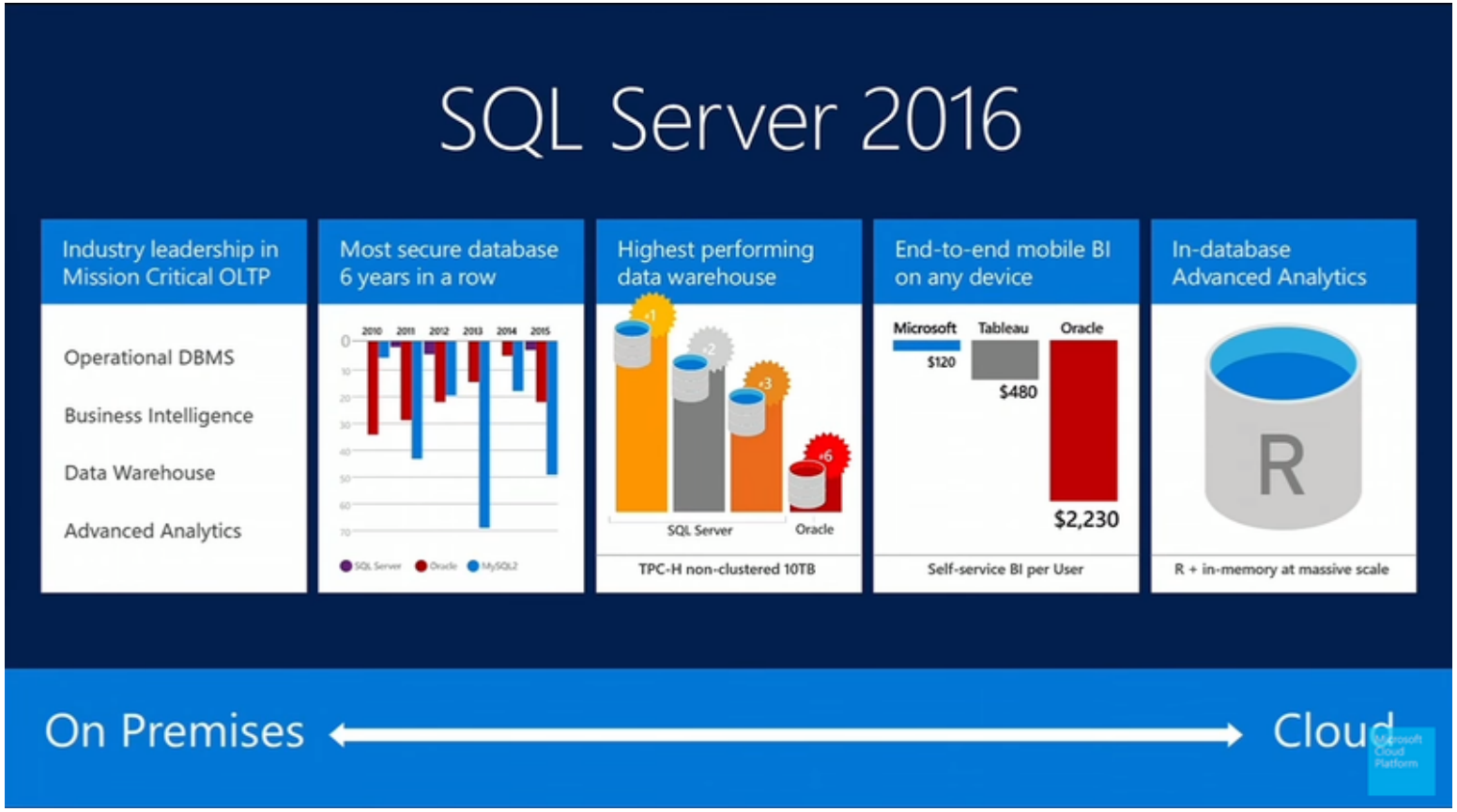



```
R> library(rfoaas)
R> greed("R Programming", "Random R Hacker")
The point is, ladies and gentleman, that r programming -- for lack of a
better word -- is good. R Programming is right. R Programming works. R
Programming clarifies, cuts through, and captures the essence of the
evolutionary spirit. R Programming, in all of its forms -- R
Programming for life, for money, for love, knowledge -- has marked the
upward surge of mankind. - Random R Hacker
R>
```

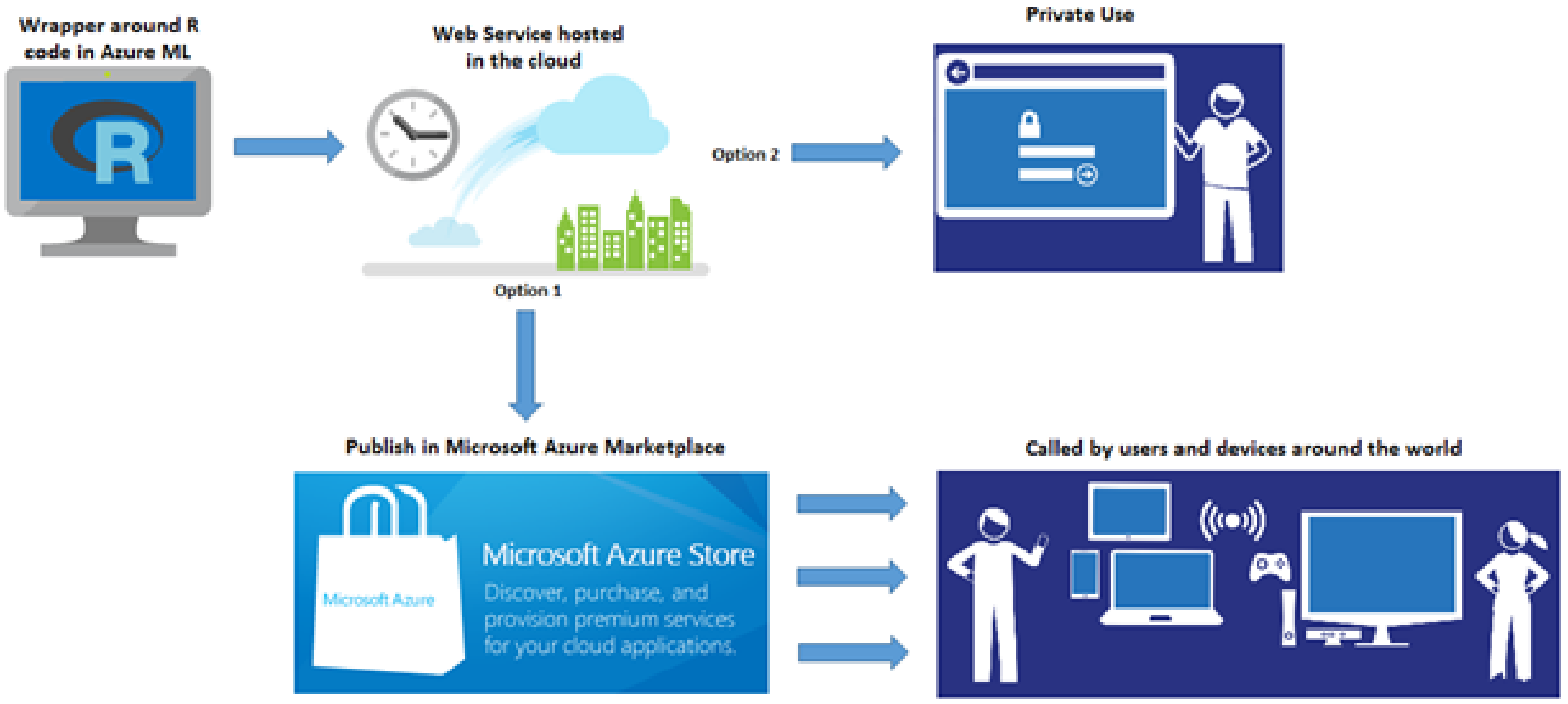
Visualizador

Plataformas de Analítica y Bigdata que apuestan por R

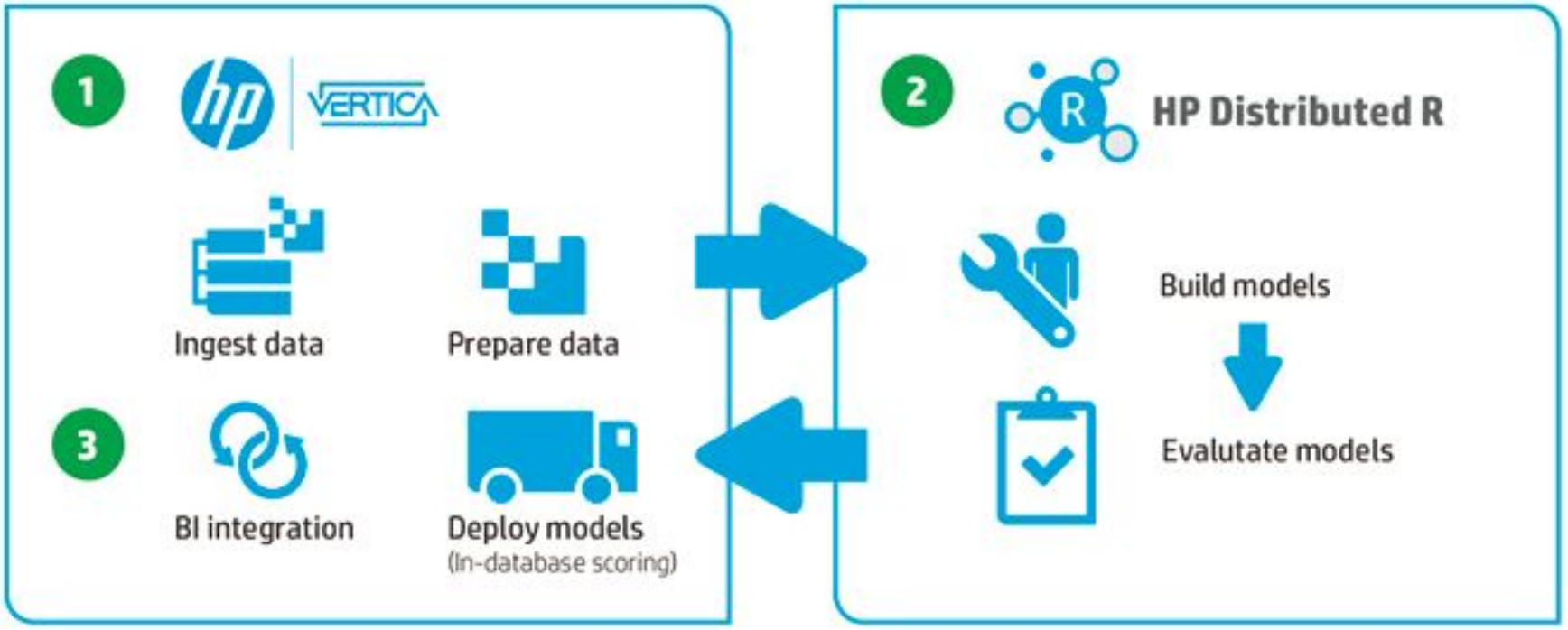
SQL


SQL Server 2016
Industry leadership in Mission Critical OLTP
Operational DBMS
Business Intelligence
Data Warehouse
Advanced Analytics
Most secure database 6 years in a row
2010 2011 2012 2013 2014 2015
Highest performing data warehouse
SQL Server
Oracle
TPC-H non-clustered 10TB
End-to-end mobile BI on any device
Microsoft
$120
Tableau
$480
Oracle
$2,230
Self-service BI per User
In-database Advanced Analytics
R
R + in-memory at massive scale
On Premises
Cloud

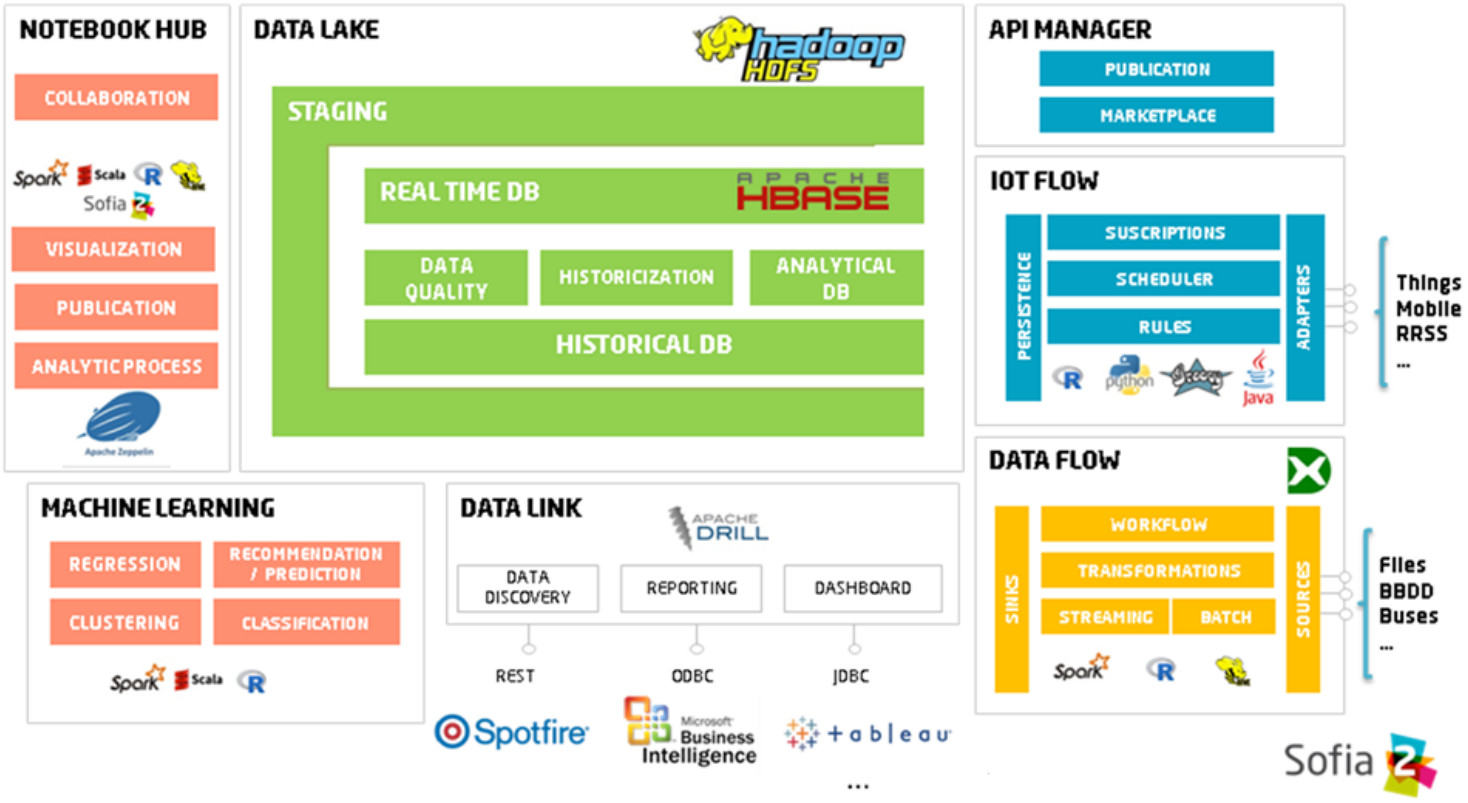
Azure


Wrapper around R code in Azure ML
Web Service hosted in the cloud
Private Use
Option 2
Option 1
Publish in Microsoft Azure Marketplace
Called by users and devices around the world
Microsoft Azure Store
Discover, purchase, and provision premium services for your cloud applications.
Microsoft Azure

HP

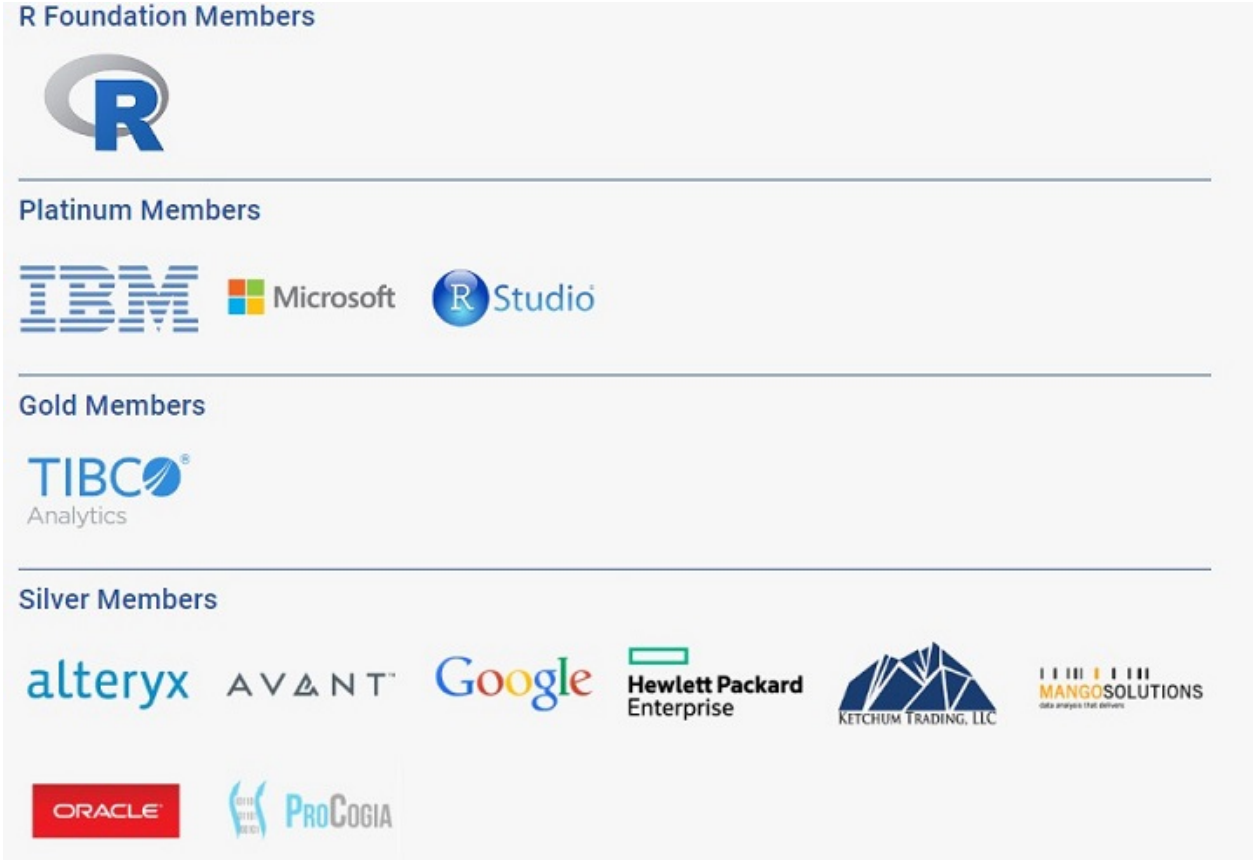

1
VERTICA
Ingest data
Prepare data
3
BI integration
Deploy models
(In-database scoring)
2
HP Distributed R
Build models
Evalutate models

Sofia2 - Indra



La Apuesta de las Industrias por R

R Consortium



La misión explícita del R Consortium es "avanzar en la promoción mundial y el apoyo para el lenguaje de código abierto R"

Big Data

Guía del Viajero Intergaláctico
The Hitchhiker's Guide to the Galaxy

Una de las Historias cuenta, que una raza de **seres hiperinteligentes pandimensionales** construyeron una computadora llamada Pensamiento Profundo («Deep Thought») fabricada con el único objetivo de descifrar la respuesta definitiva.

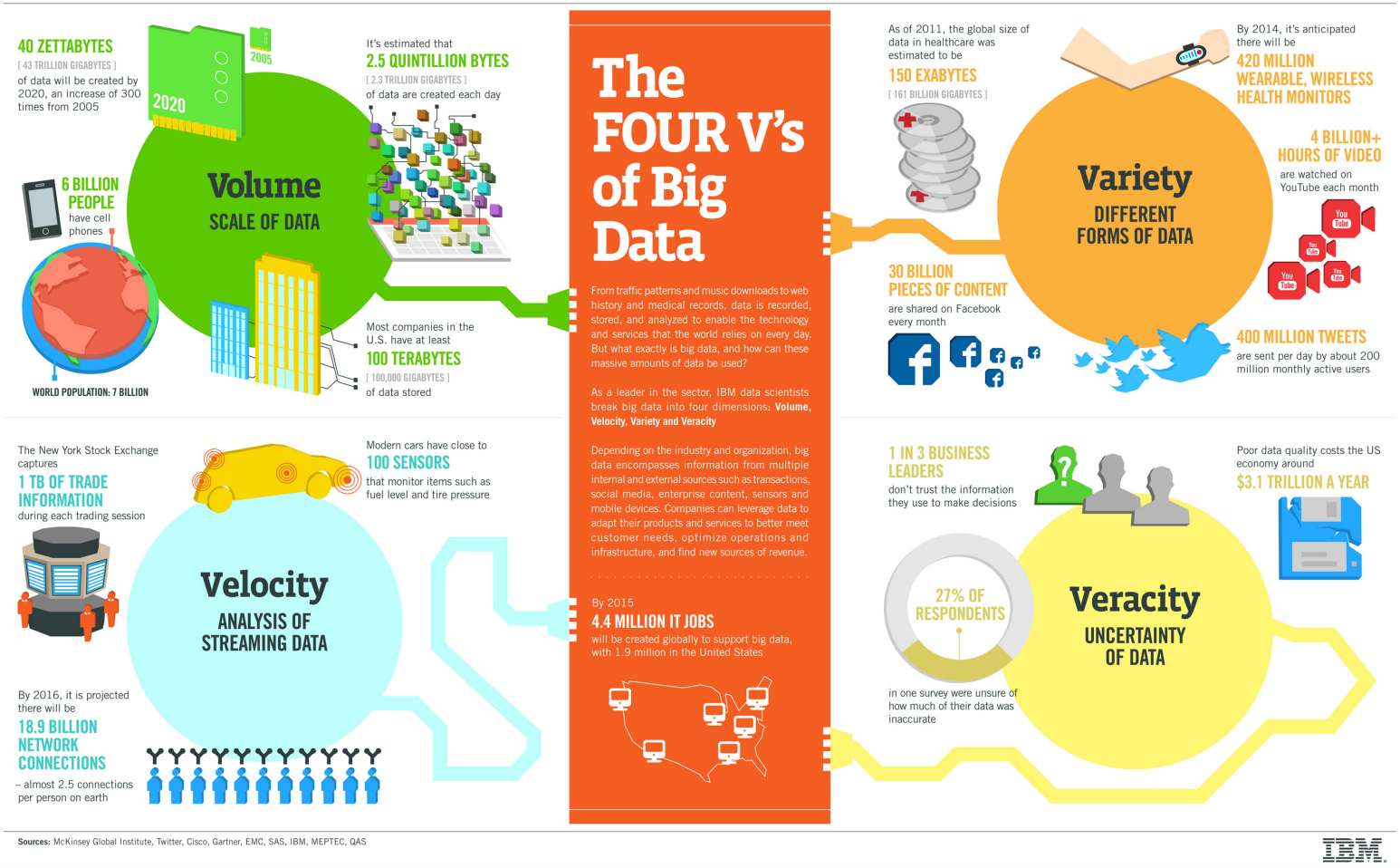
«el sentido de la vida, el universo y todo lo demás»

Guía del Viajero Intergaláctico
The Hitchhiker's Guide to the Galaxy

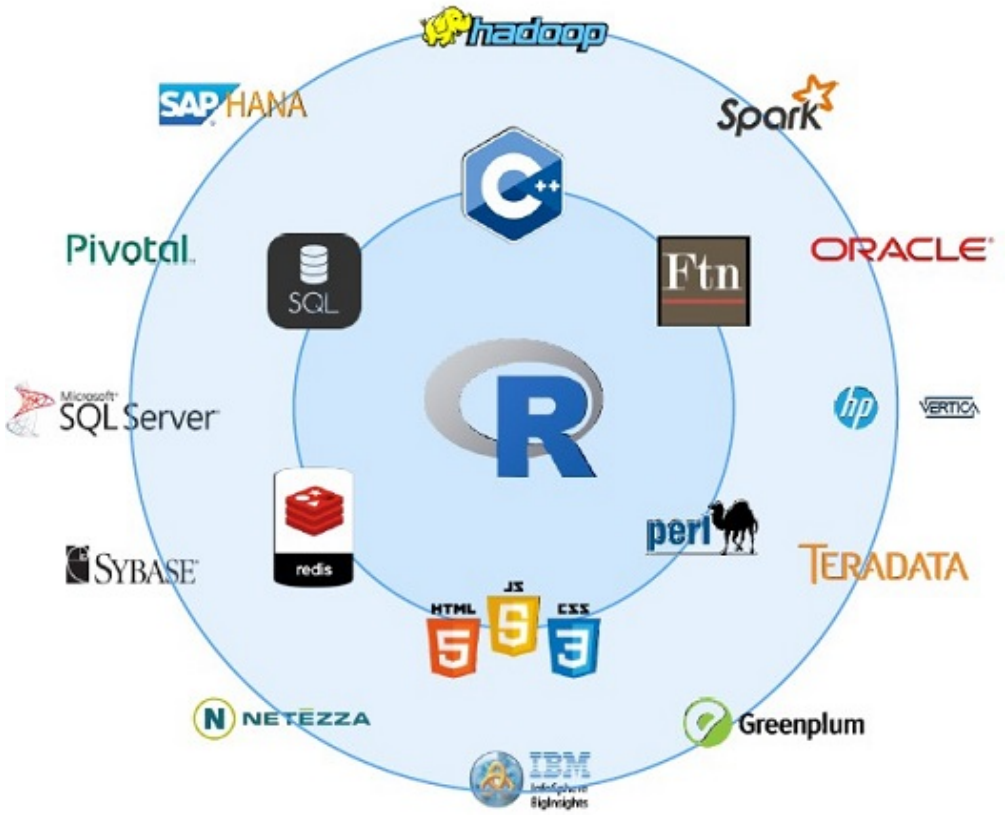
Pensamiento Profundo se toma **siete millones y medio de años** para dar esa respuesta, la cual, para pesar de muchos, resulta ser **42** sin lugar a dudas.

Cuando la respuesta se revela como 42, **se ven forzados a construir una computadora aún más poderosa para calcular la «Pregunta máxima»**, pero sus planes nunca culminan…

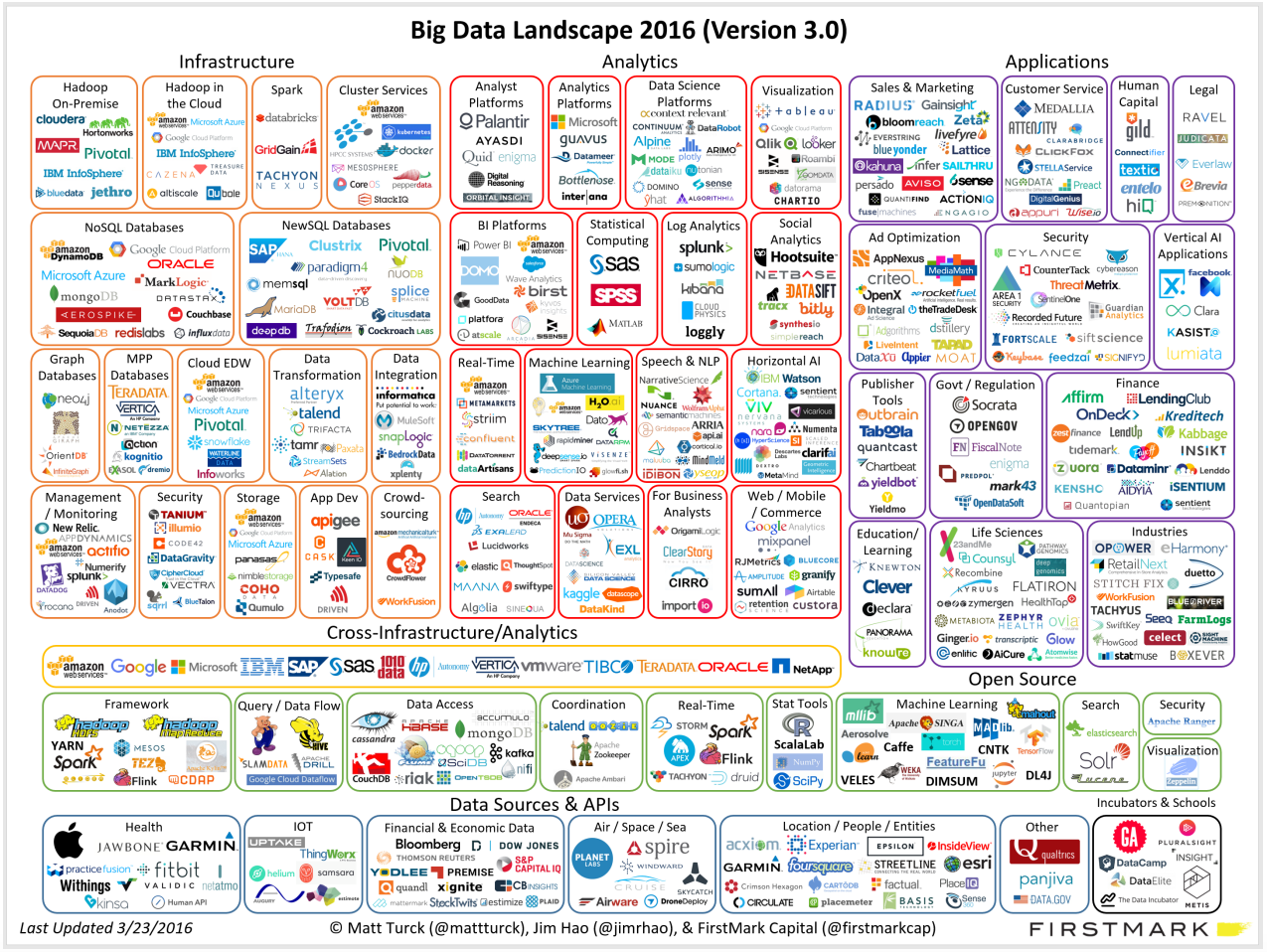
Big Data


40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005
2005
2020
6 BILLION PEOPLE
have cell phones
WORLD POPULATION: 7 BILLION
Volume
SCALE OF DATA
It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day
Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored
The Four V's of Big Data
From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?
As a leader in the sector, IBM data scientists break big data into four dimensions: Volume, Velocity, Variety and Veracity
Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.
By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States
As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]
30 BILLION PIECES OF CONTENT
are shared on Facebook every month
Variety
DIFFERENT FORMS OF DATA
By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS
4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month
400 MILLION TWEETS
are sent per day by about 200 million monthly active users
The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session
Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure
Velocity
ANALYSIS OF STREAMING DATA
By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections per person on earth
1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions
27% OF RESPONDENTS
in one survey were unsure of how much of their data was inaccurate
Veracity
UNCERTAINTY OF DATA
Poor data quality costs the US economy around
$3.1 TRILLION A YEAR
Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS
IBM

R y Big Data

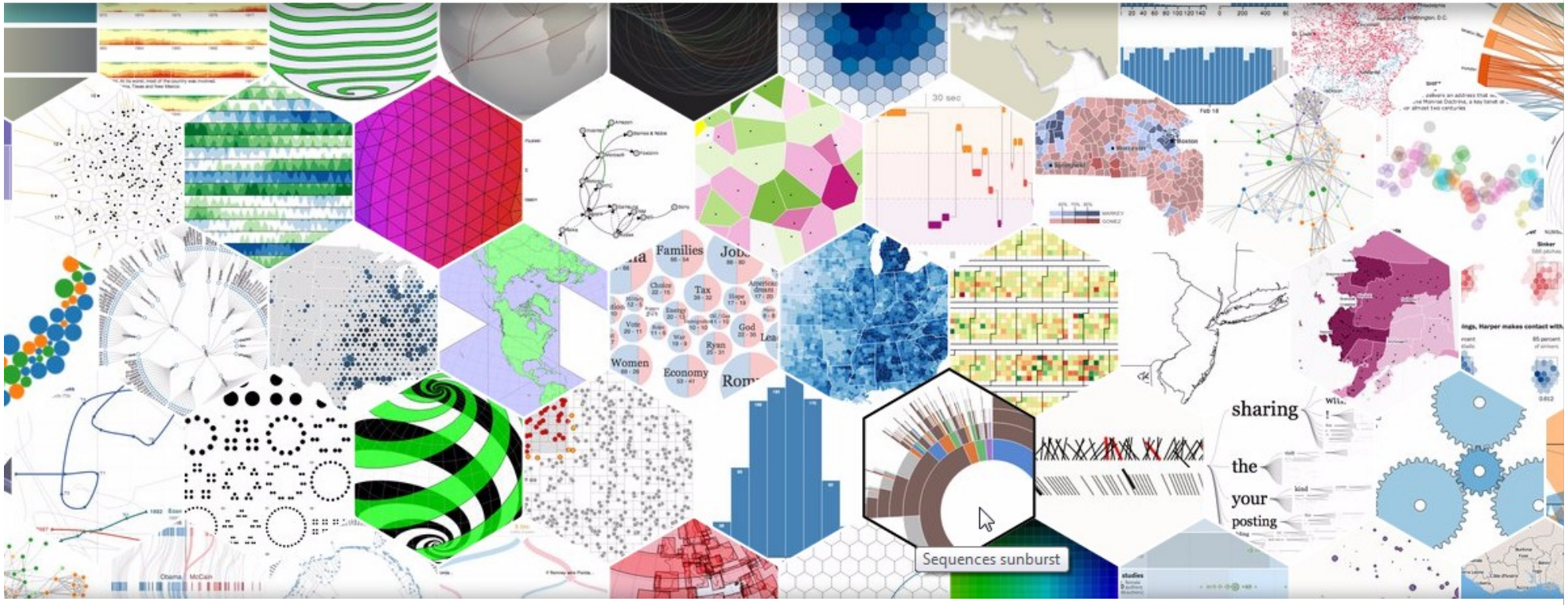

hadoop
SAP HANA
Spark
C++
Pivotal
SQL
Ftn
ORACLE
Microsoft SQL Server
R
hp
VERTICA
redis
perl
SYBASE
TERADATA
HTML
JS
CSS
NETEZZA
Greenplum
IBM

Panorama del Big Data 2016



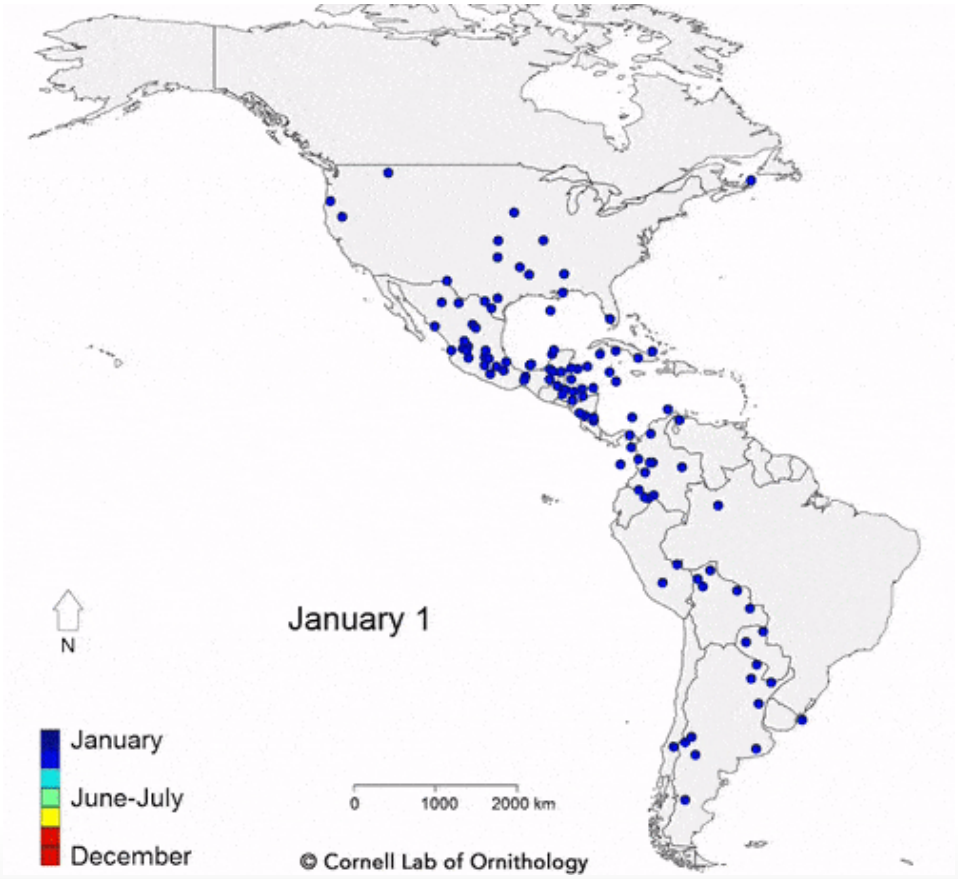
Visualización y Productos de Datos

D3.Js


Families
Tax
Women
Economy
sharing
the
your
posting
Sequences sunburst

Visualización y Productos de Datos

Generar nuestras propias visualizaciones.



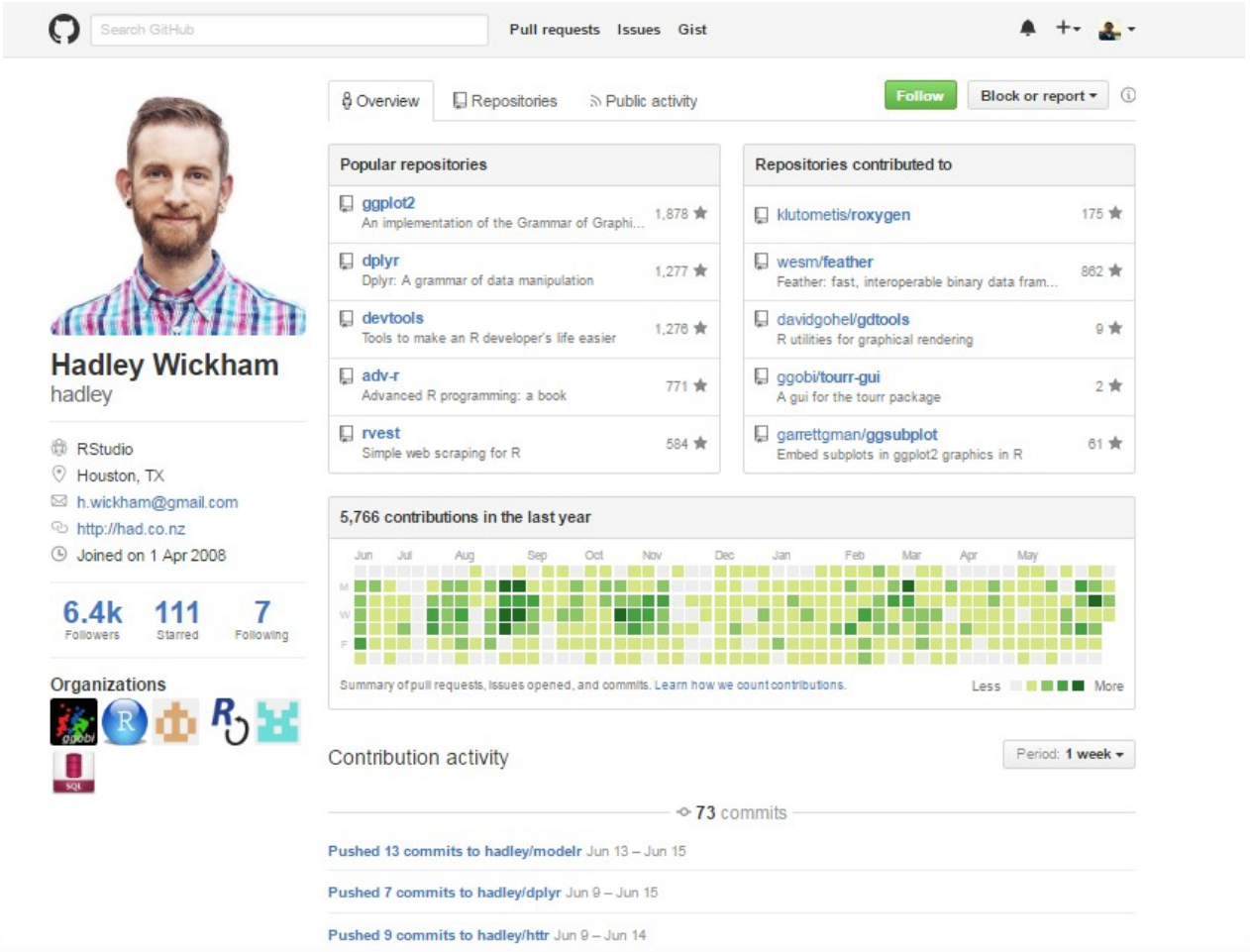
Visualización y Productos de Datos

Generar nuestras propias visualizaciones.

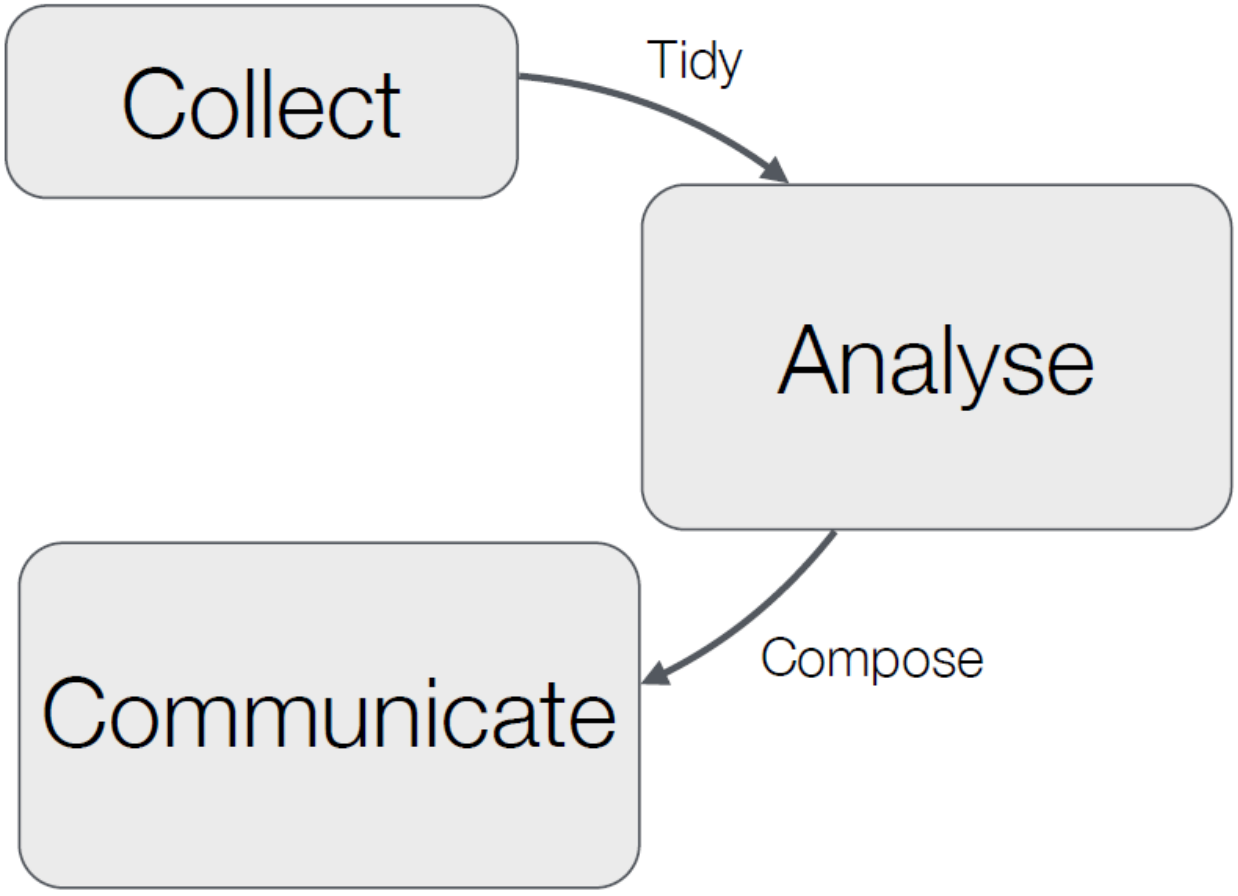
- Gráficos varios dentro de Dashboards
- Gráficos de Redes (nodos)
- Get data y plot (Ejemplo MIO)
- **Publica un Tweet con el hashtag: #RUsersCali**
- Gráficos y Social tracking (Partido)
- Ejemplo Twitter
- Ejemplo linet (Descargas eléctricas)
- Data Mining - "What if"

Rstudio y el Hadleyverse

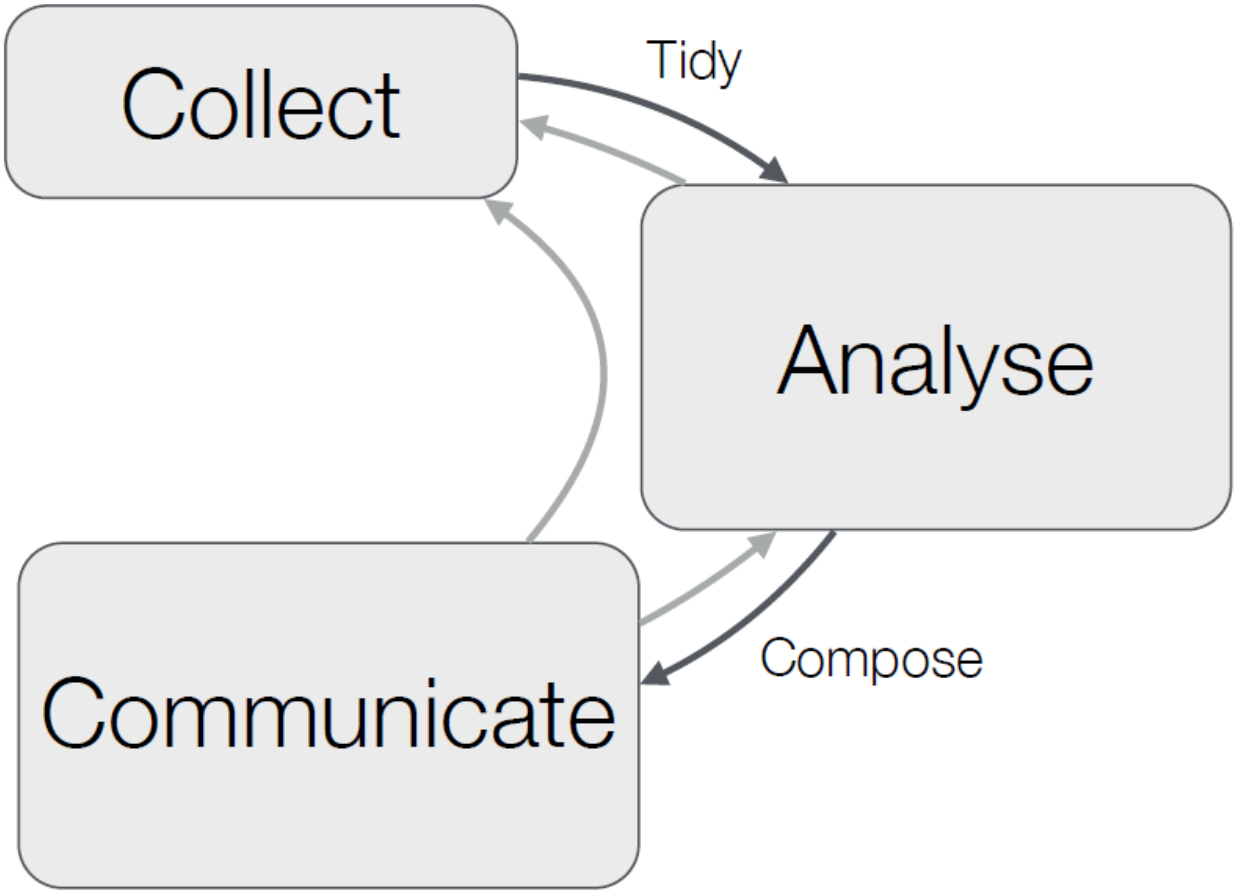
Rstudio y el Hadleyverse



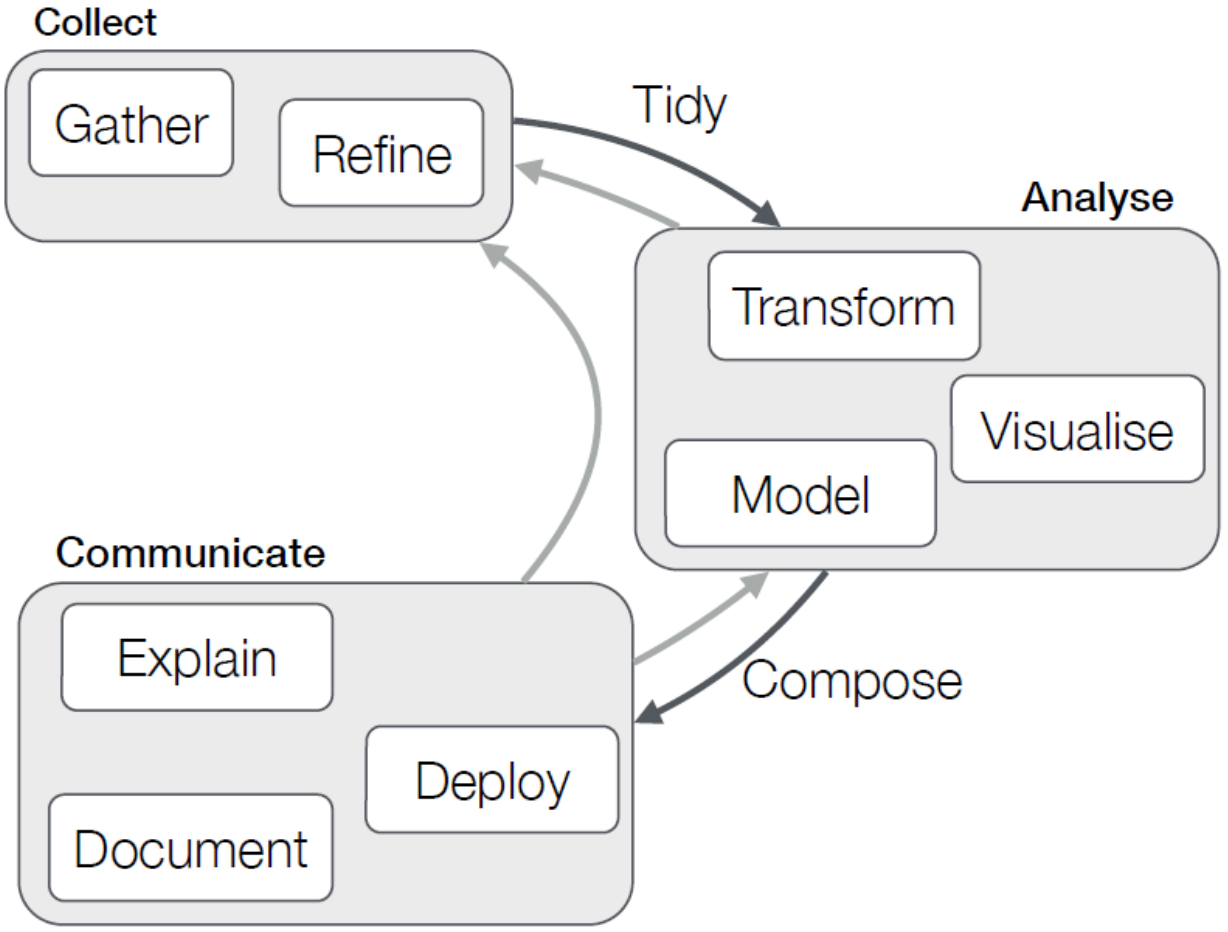
Procesos con Datos



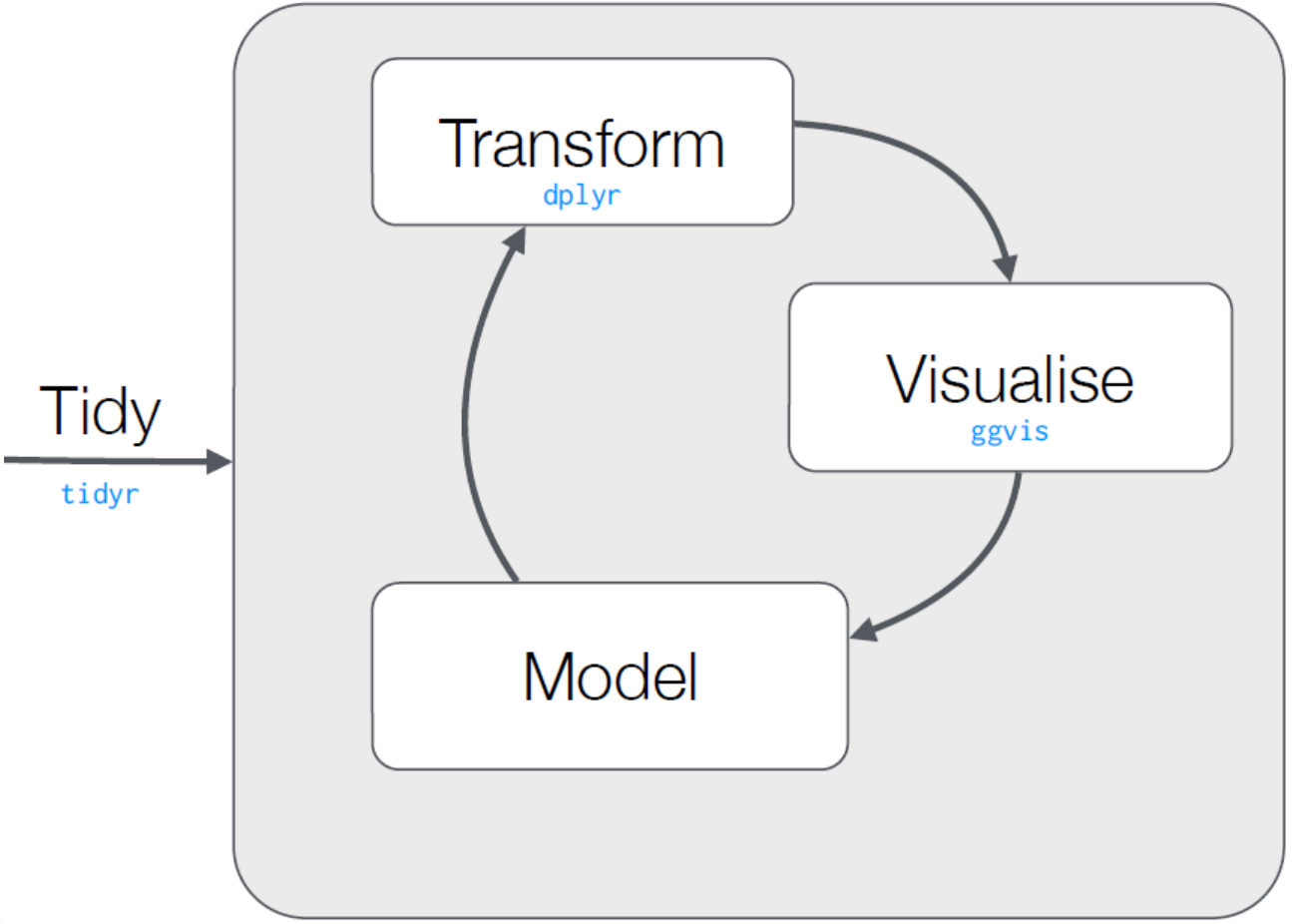
Procesos con Datos



Procesos con Datos



Procesos con Datos

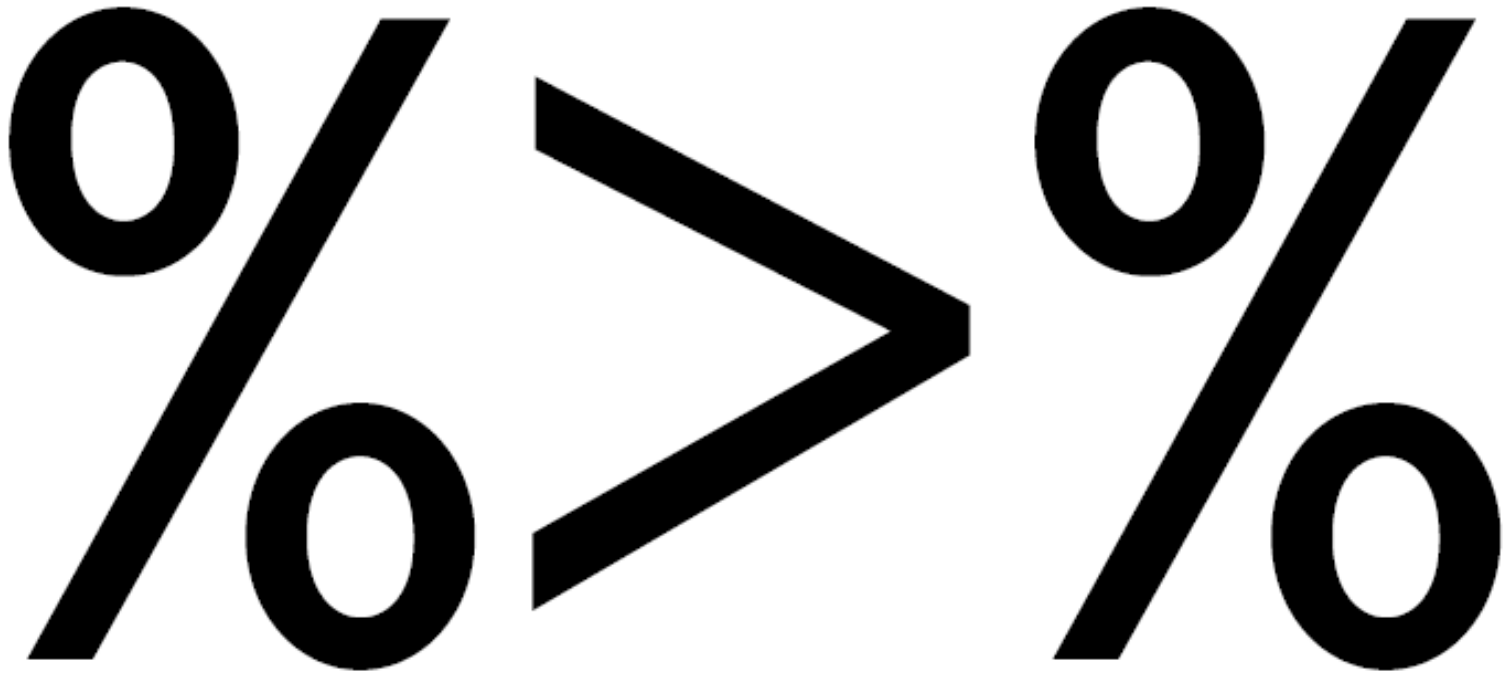


dplyr

+ group by

- **filter**: keep rows matching criteria
- **select**: pick columns by name
- **arrange**: reorder rows
- **mutate**: add new variables
- **summarise**: reduce variables to values

Pipe

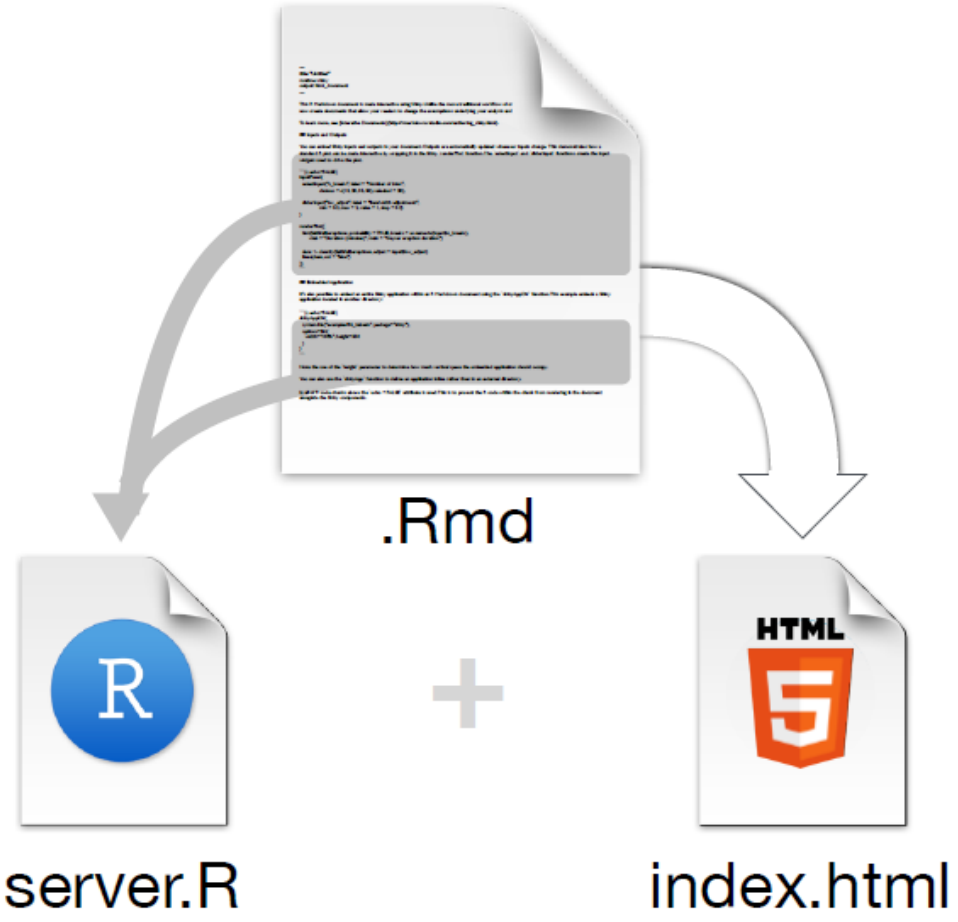


`tidyr, dplyr, ggvis, ...`

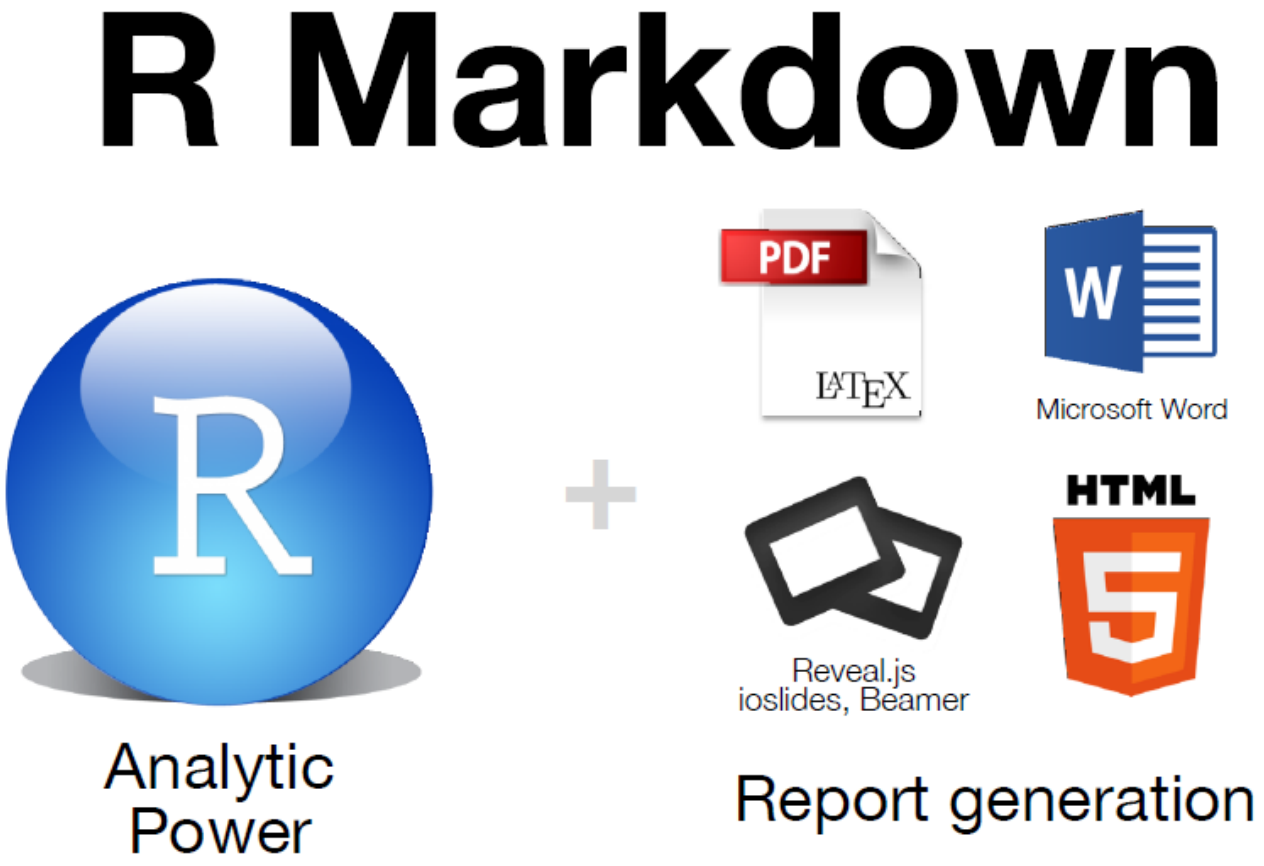
Apps Shiny

Typical Shiny App

Apps Shiny



Investigación Reproducible



Fin Sesión Introductoria