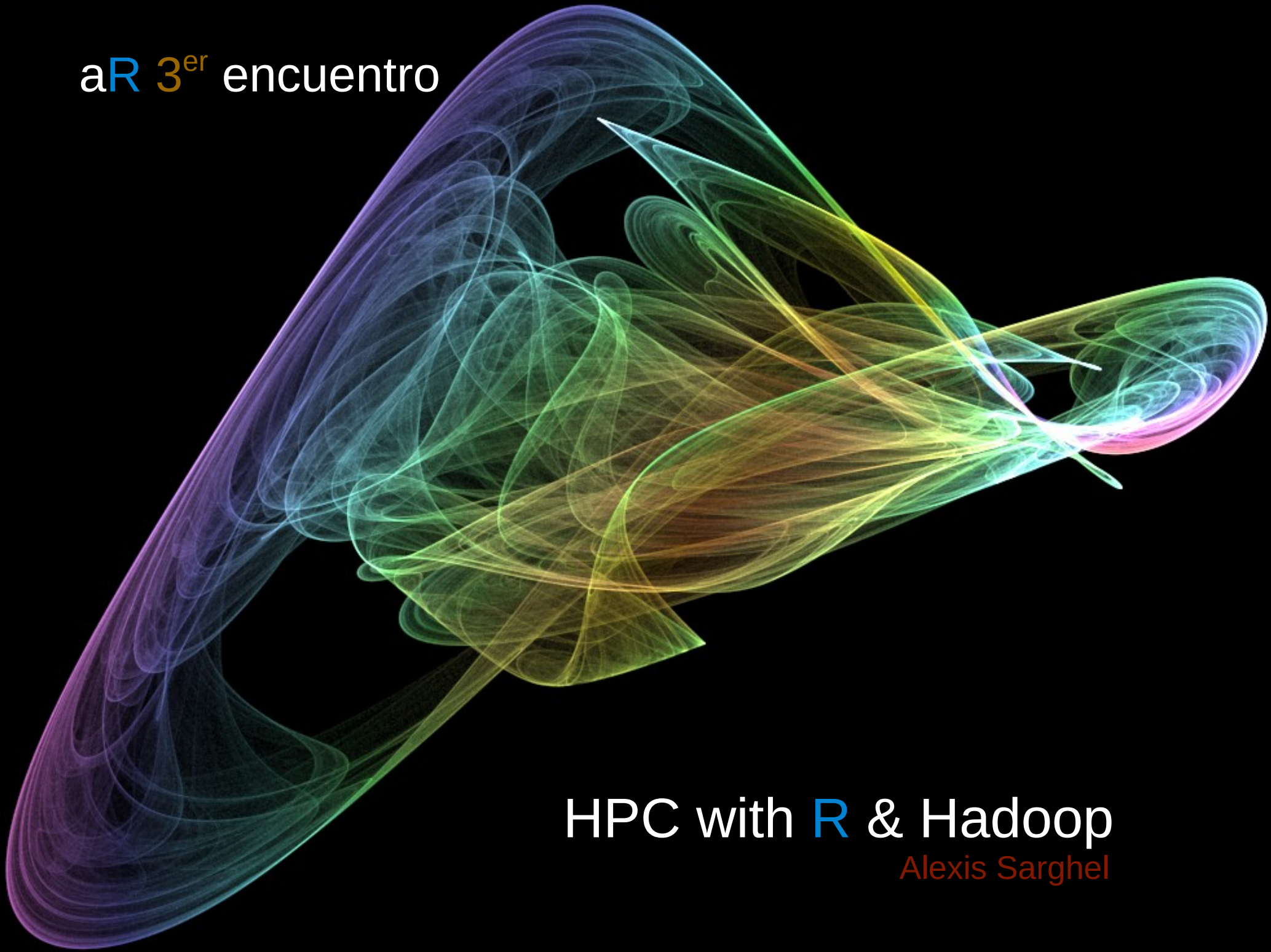
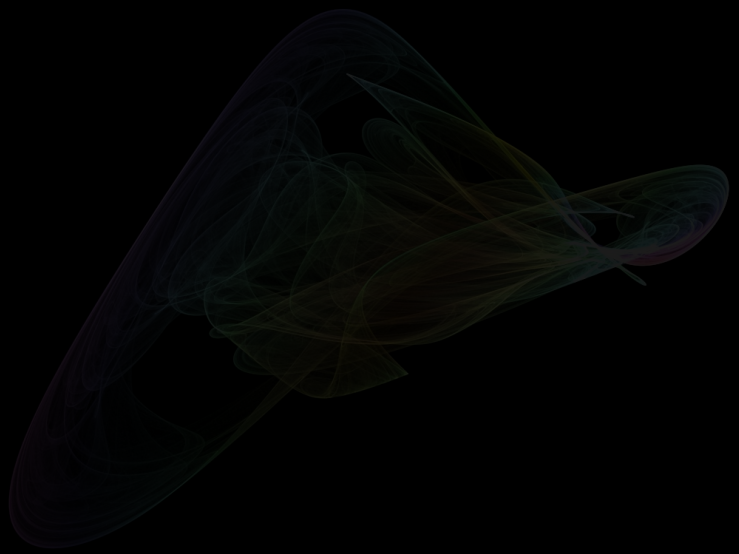
aR 3er encuentro

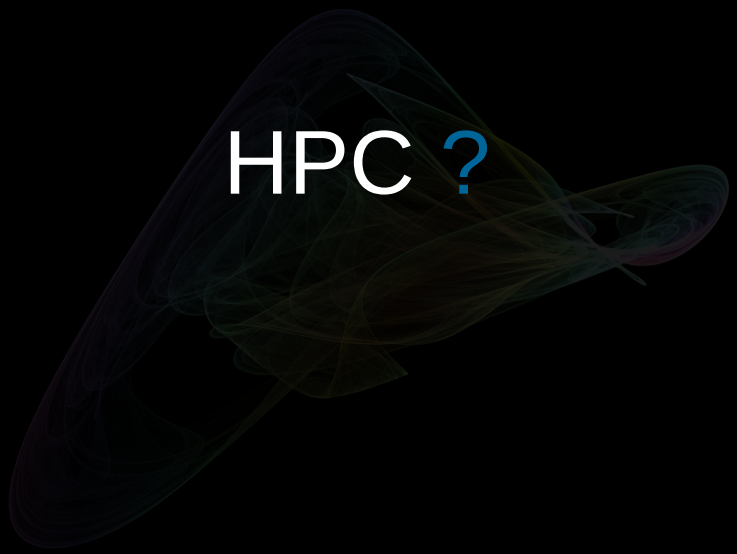


HPC with R & Hadoop

Alexis Sarghel



HPC ?

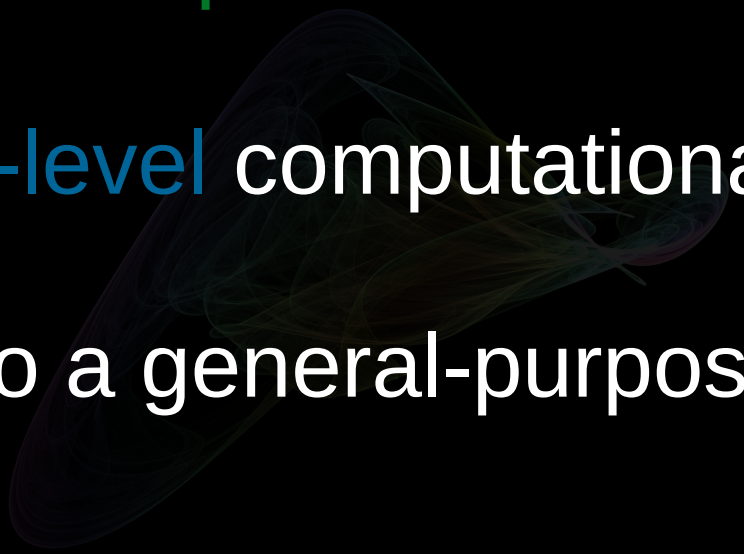


Hight-performance computing

A supercomputer is a computer

with a high-level computational capacity

compared to a general-purpose computer



Performance of a supercomputer is

measured in floating point operations per second (FLOPS)

instead of million instructions per second (MIPS)

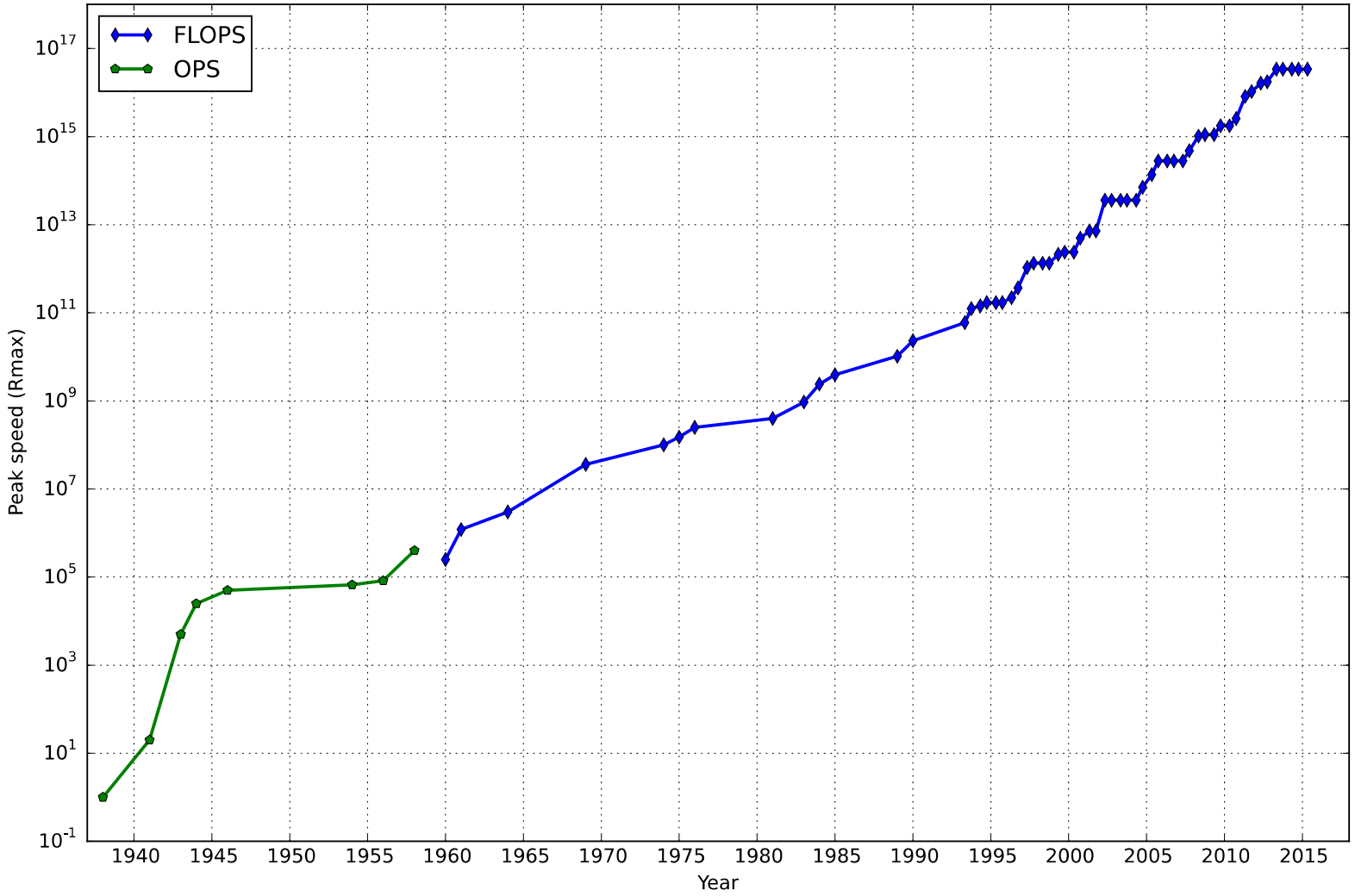
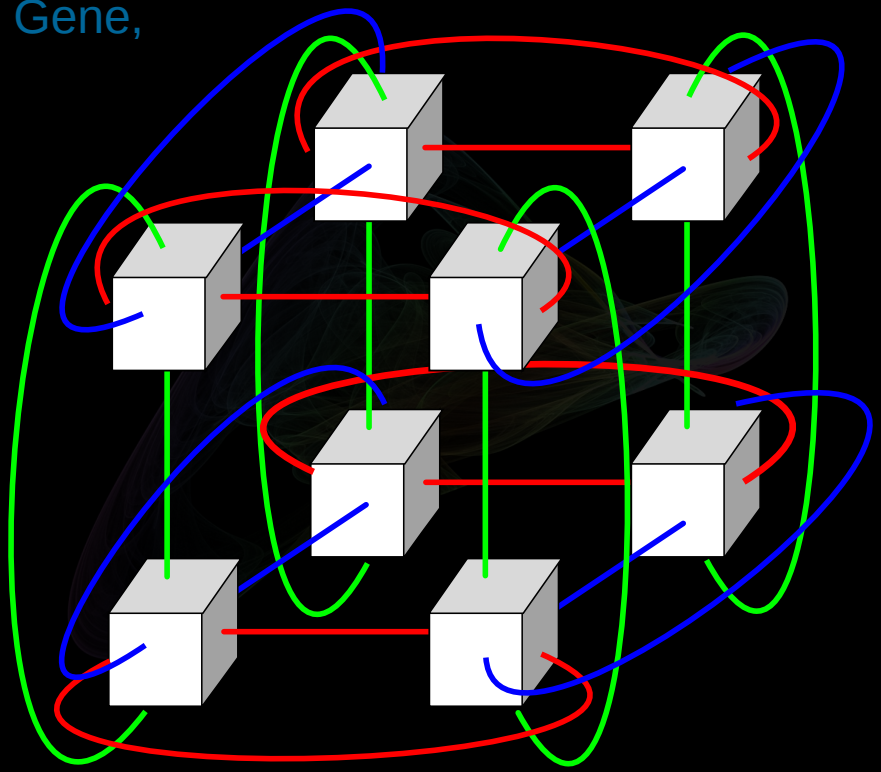

FLOPS
OPS
Peak speed (Rmax)
Year
10^{17}
10^{15}
10^{13}
10^{11}
10^{9}
10^{7}
10^{5}
10^{3}
10^{1}
10^{-1}
1940
1945
1950
1955
1960
1965
1970
1975
1980
1985
1990
1995
2000
2005
2010
2015

Diagram of a 3-dimensional torus interconnect used by systems such as Blue Gene, Cray XT3, etc.



Hight-performance computing

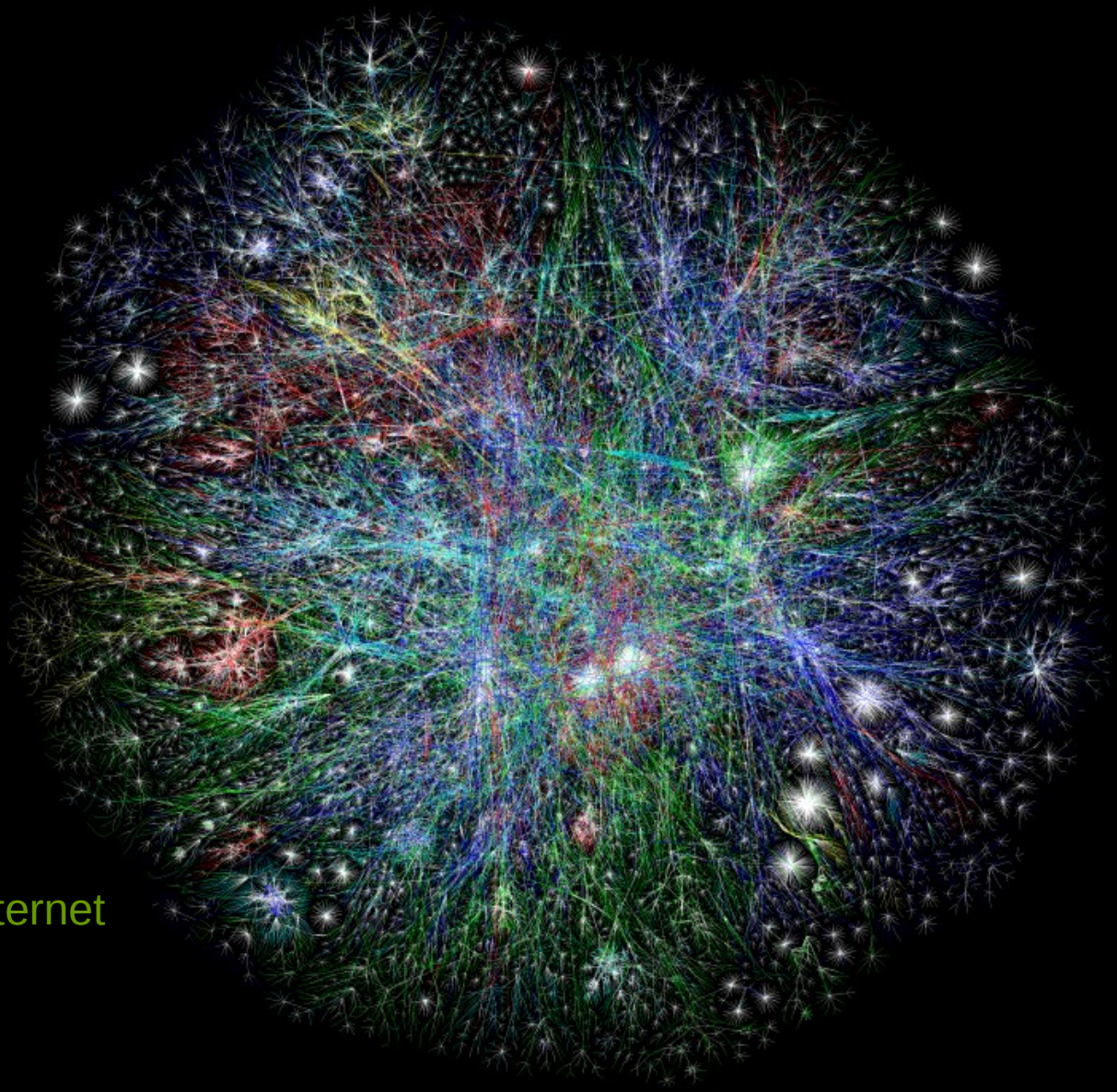
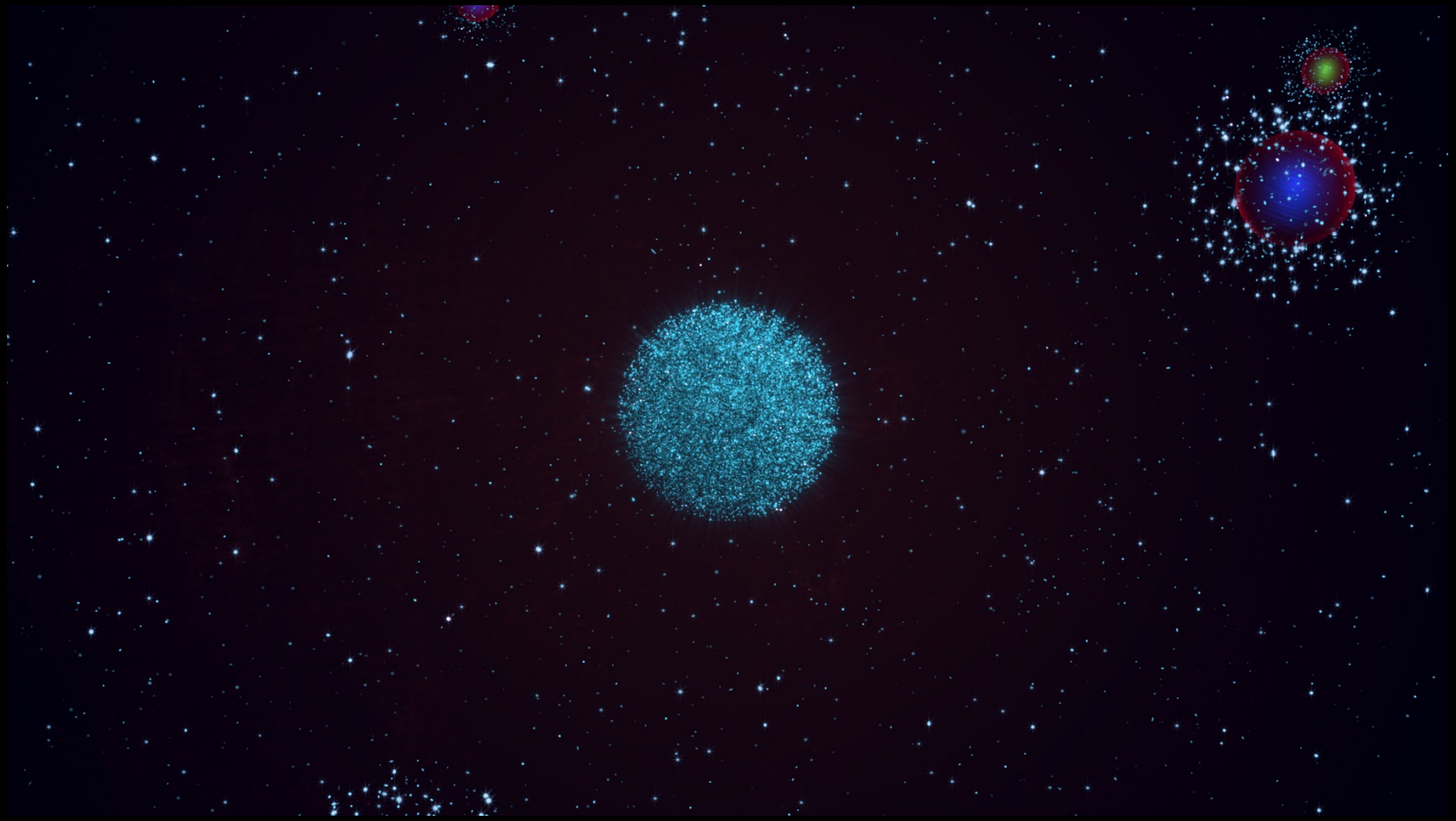


Diagram of internet

2015



Boson de Higgs

Big Data ?

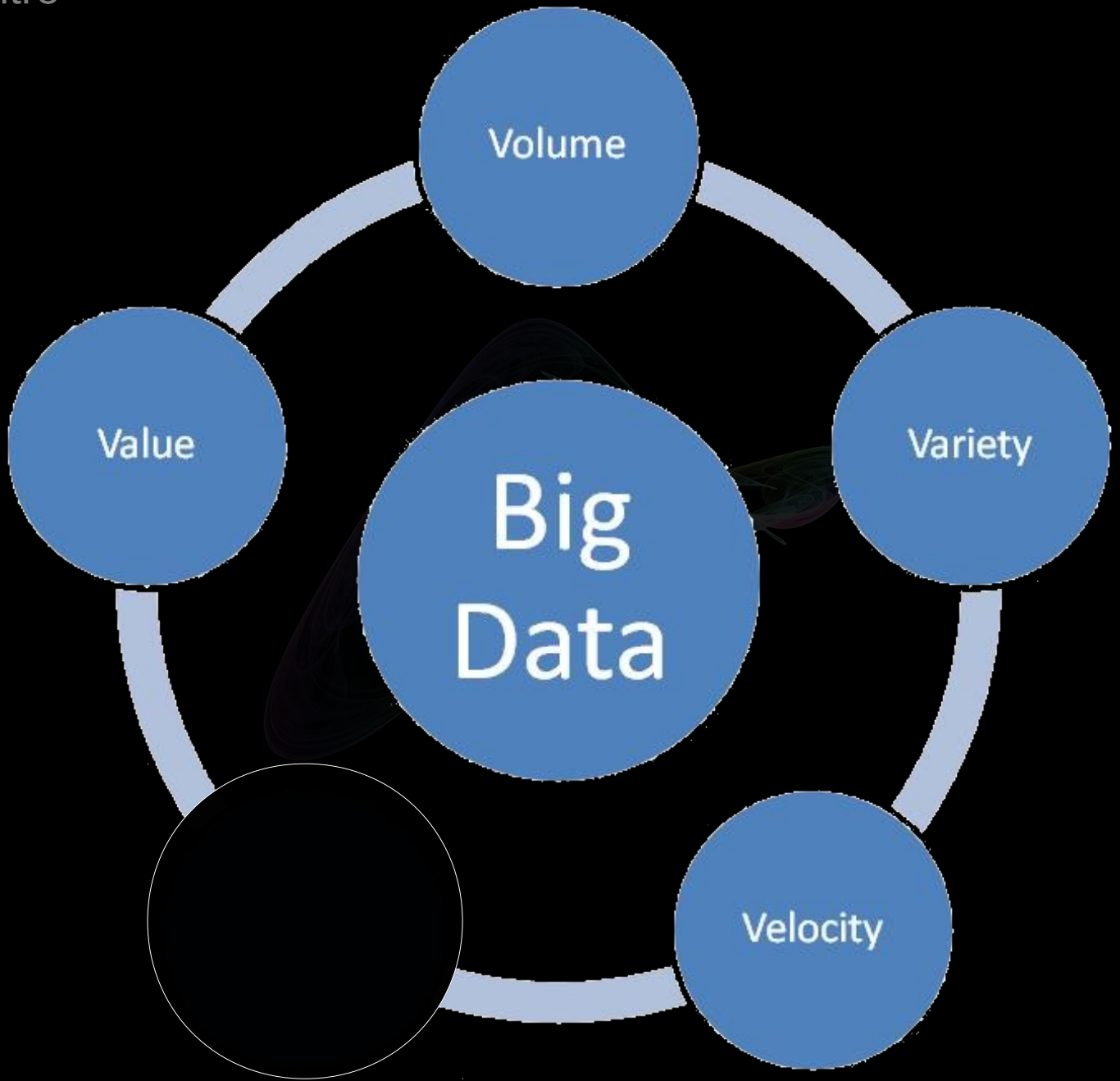
Big data is a broad term

for data sets so large or complex

that traditional data processing

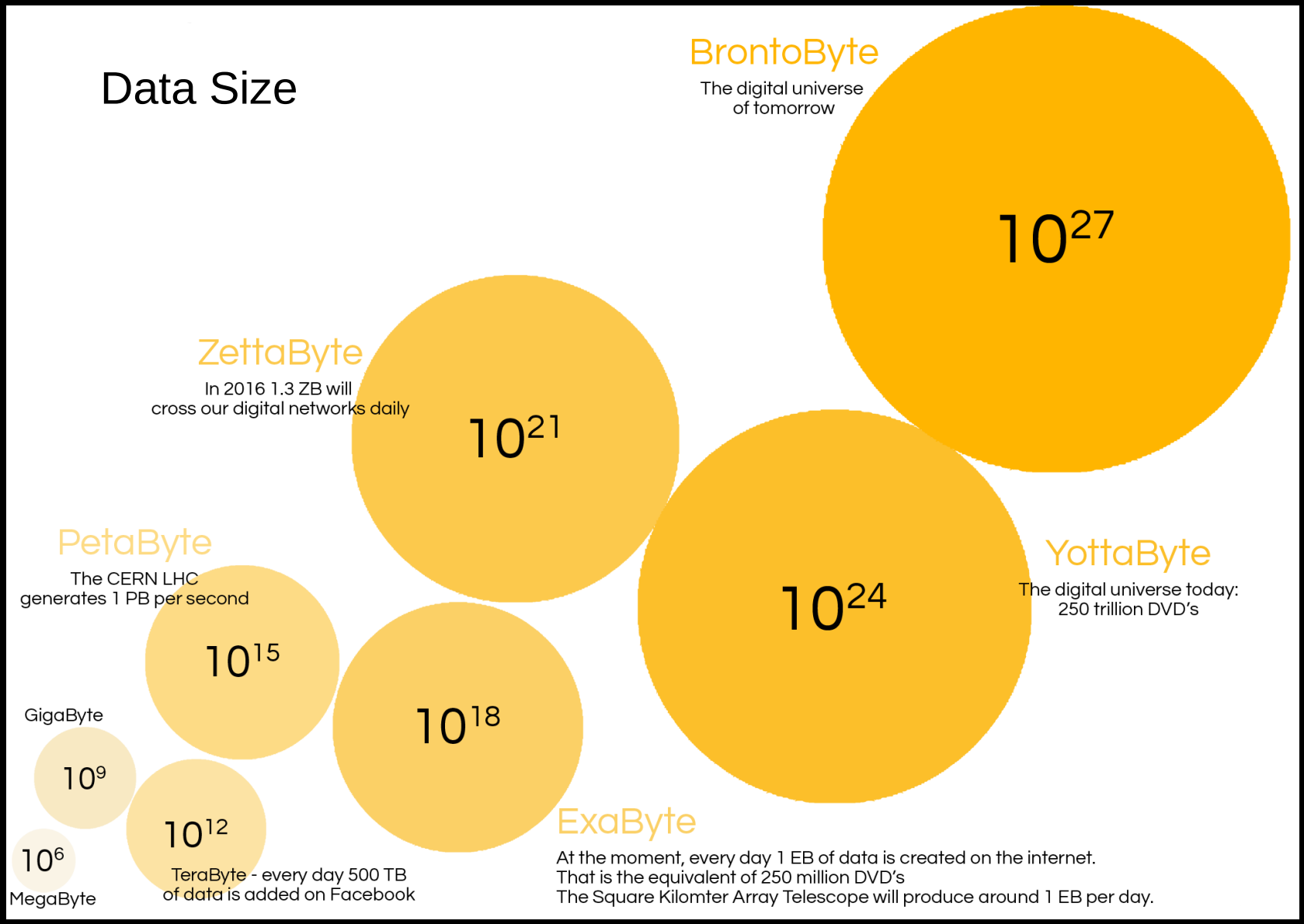
applications are inadequate

3 V's



4 V's

5 V's


Data Size
BrontoByte
The digital universe of tomorrow
10^{27}
ZettaByte
In 2016 1.3 ZB will cross our digital networks daily
10^{21}
YottaByte
The digital universe today: 250 trillion DVD's
10^{24}
PetaByte
The CERN LHC generates 1 PB per second
10^{15}
10^{18}
GigaByte
10^{9}
10^{12}
TeraByte - every day 500 TB of data is added on Facebook
10^{6}
MegaByte
ExaByte
At the moment, every day 1 EB of data is created on the internet.
That is the equivalent of 250 million DVD's
The Square Kilomter Array Telescope will produce around 1 EB per day.

Hadoop ?

Apache Hadoop is an open-source software framework

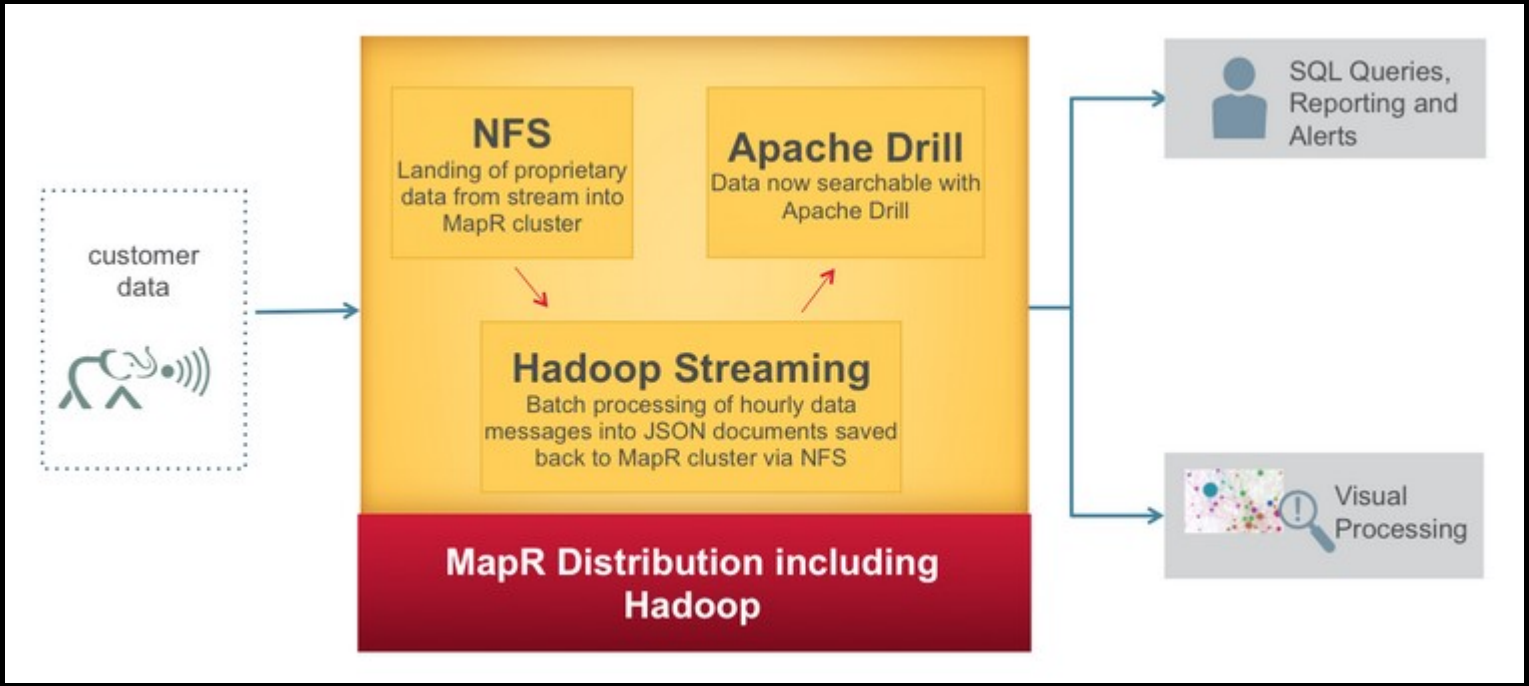
written in Java for distributed storage

and distributed processing

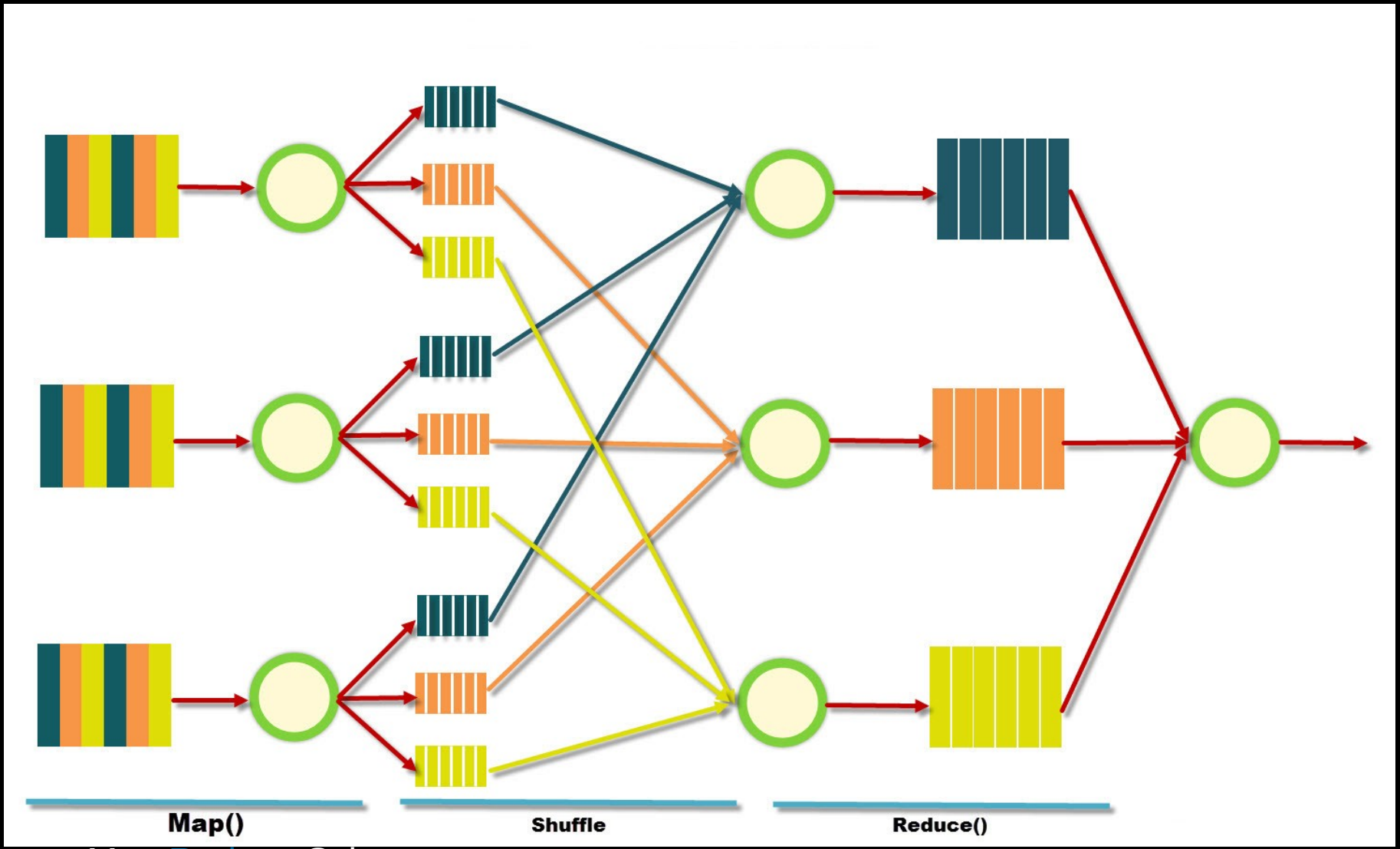
of very large data sets

on computer clusters built from commodity hardware

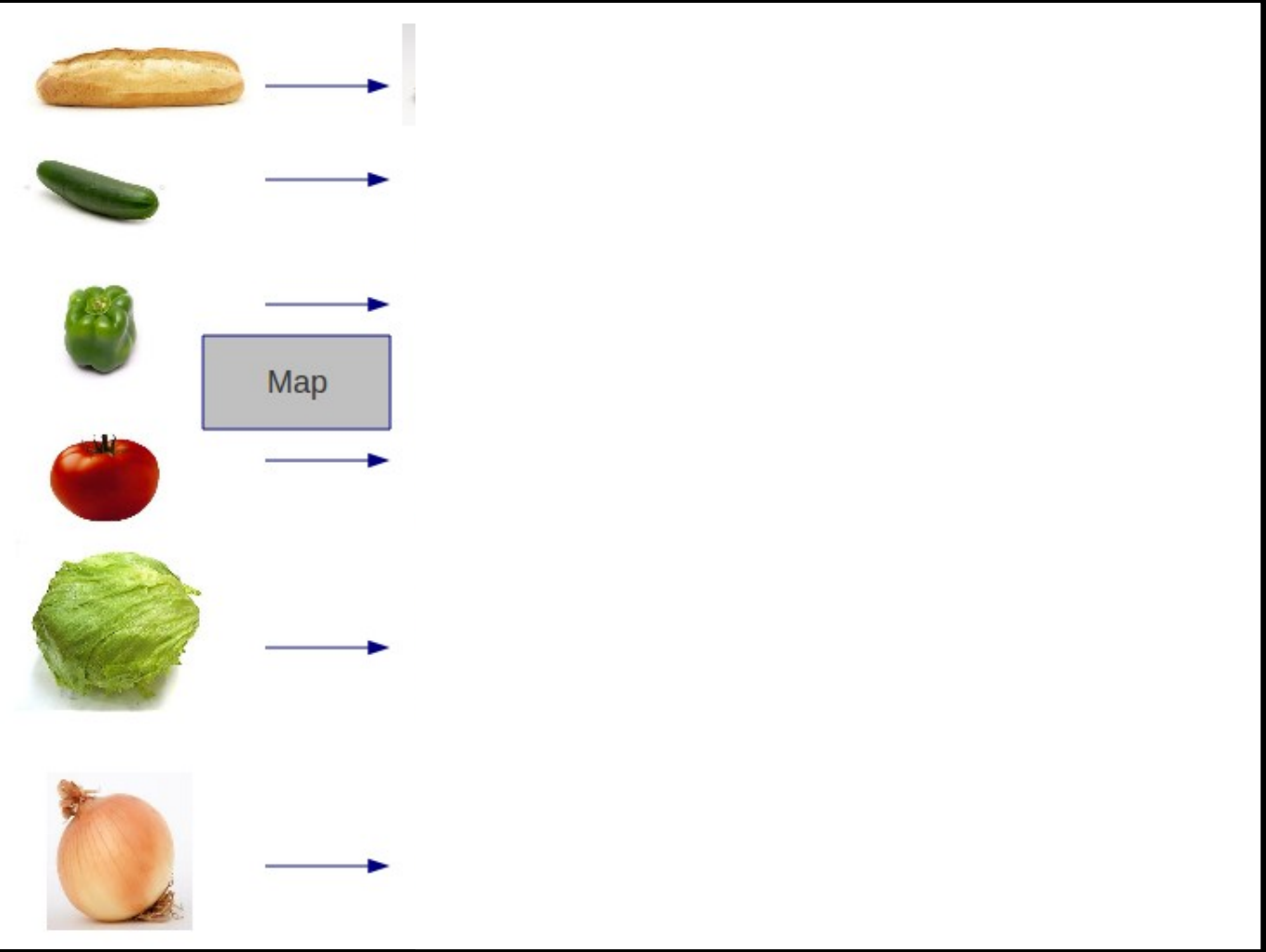
Map Reduce ?



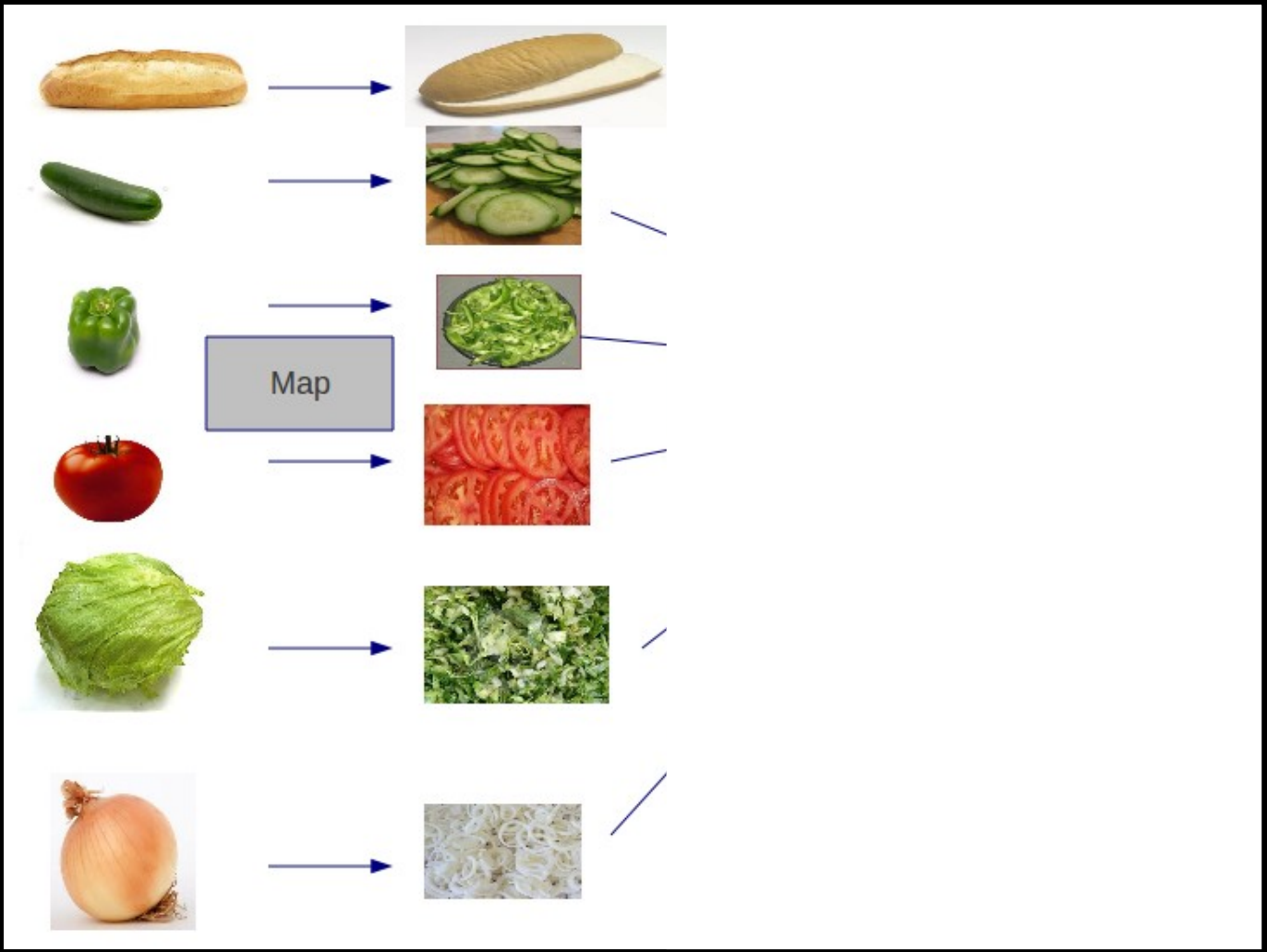
Map Reduce Schema



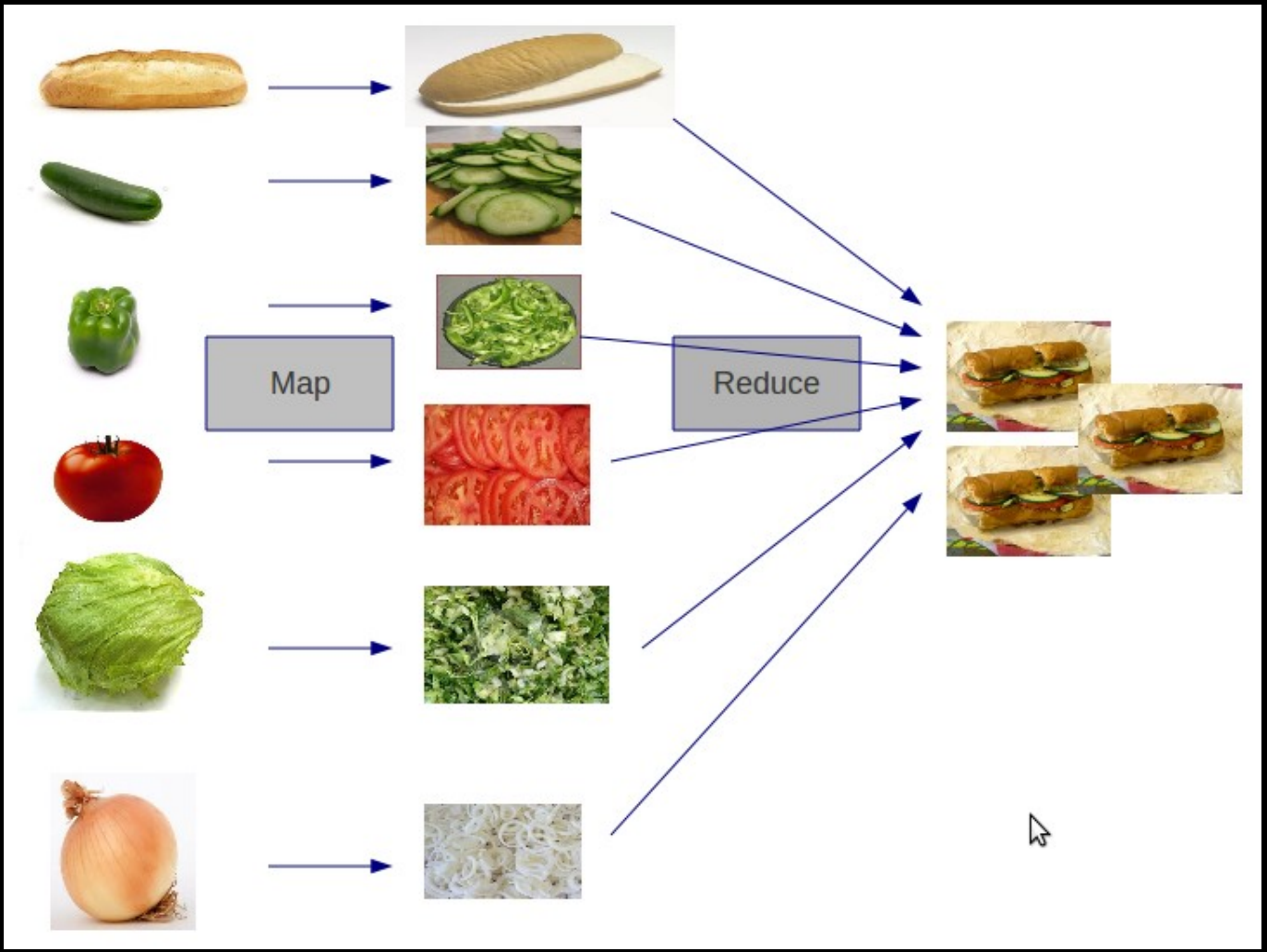
Map Reduce Schema



Map Reduce Schema



Map Reduce Schema



Map Reduce Schema

R Studio ?


R Studio®

library(rhdfs) → Hadoop

library (plyr) → distributed processing

library(rmr2) → Map Reduce

Happy Hacking