A taste of Big Data

Raúl Ramos Pollán

Universidad Industrial de Santander

Cómputo escalable

Algoritmos/Aplicaciones

```
Initialize Q(s, a) arbitrarily
Repeat (for each episode):
   Initialize s
   Choose a from s using policy derived from Q
        (e.g., ε-greedy)
   Repeat (for each step of episode):
      Take action a, observe r, s'
      Choose a' from s' using policy derived from Q
        (e.g., ε-greedy)
      Q(s, a) <-- Q(s, a) + α[r + γQ(s', a')- Q(s, a)]
      s <-- s'; a <-- a';
   until s is terminal
```

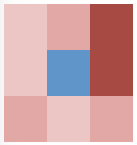
Datos

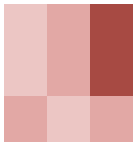
Infraestructura de cómputo

Recurso humano/ Comunidades científicas

Paralelización

Procesamiento de imágenes → operaciones en regiones (i.e. contraste)



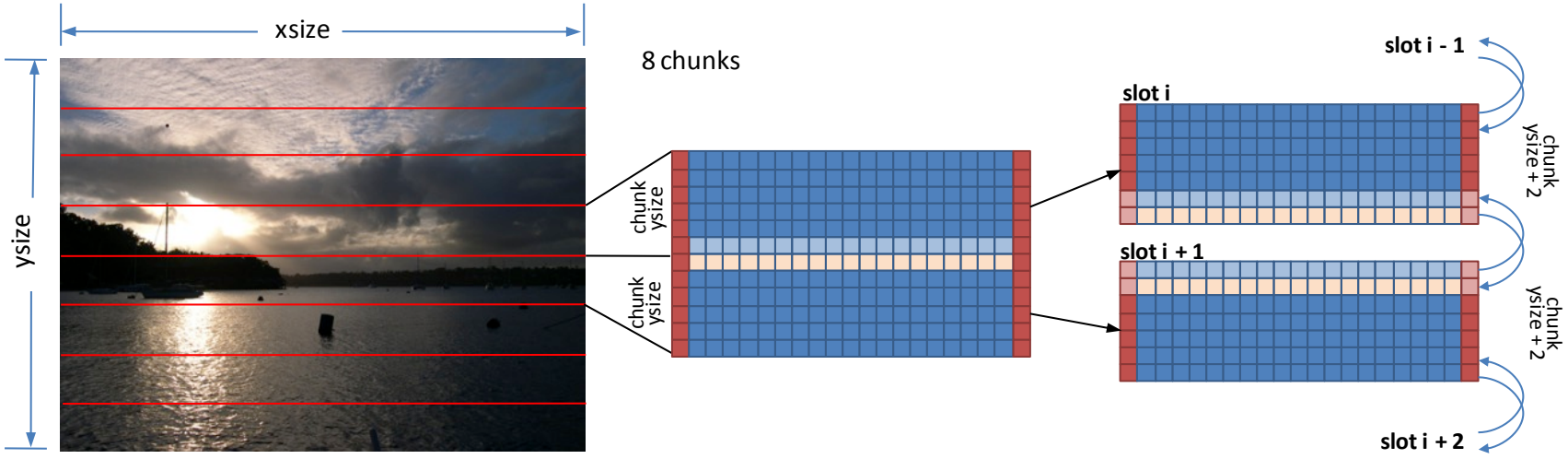


and perform n iterations

Ejemplos de algoritmos

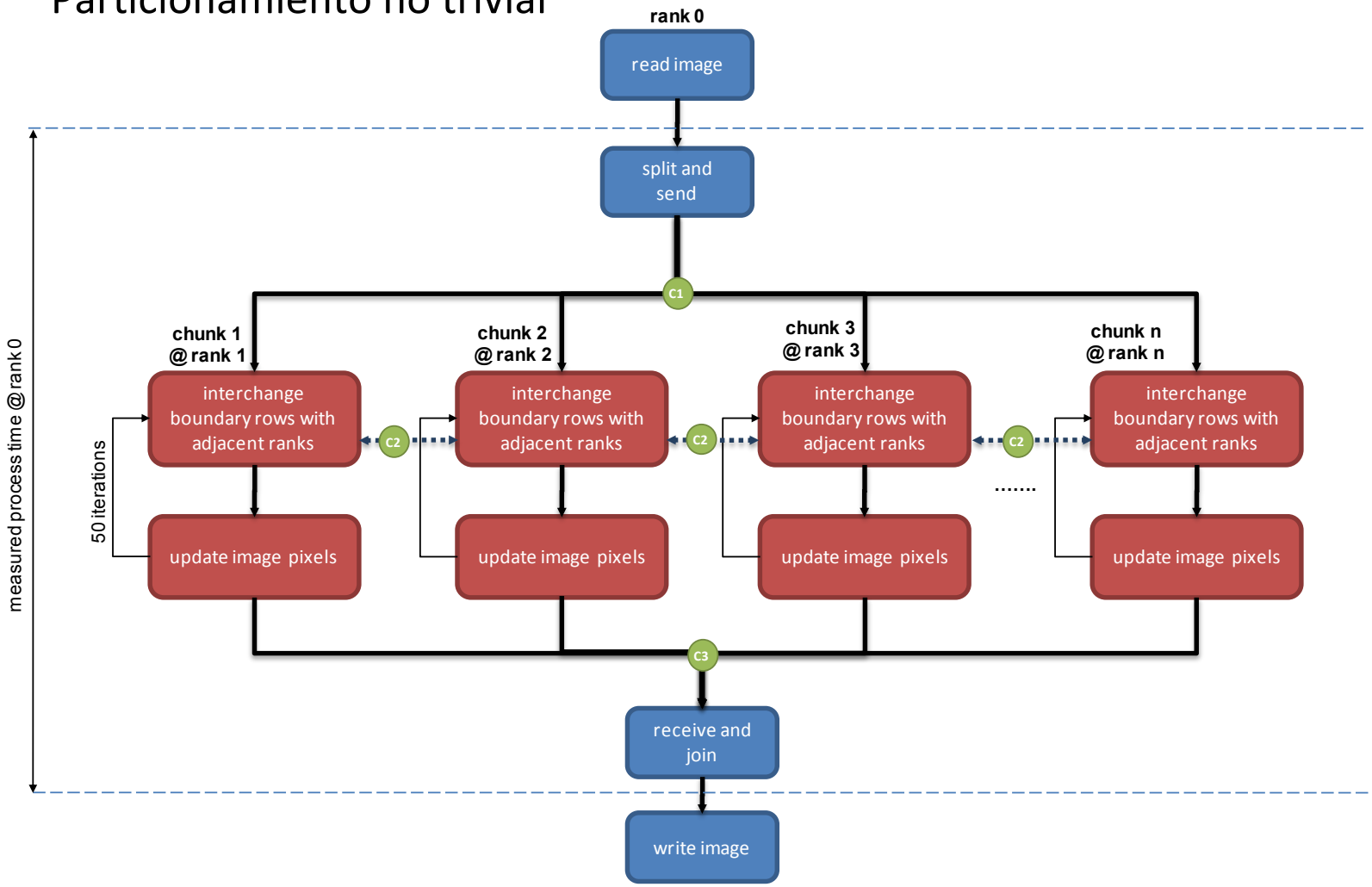
Particionamiento no trivial → por algoritmo

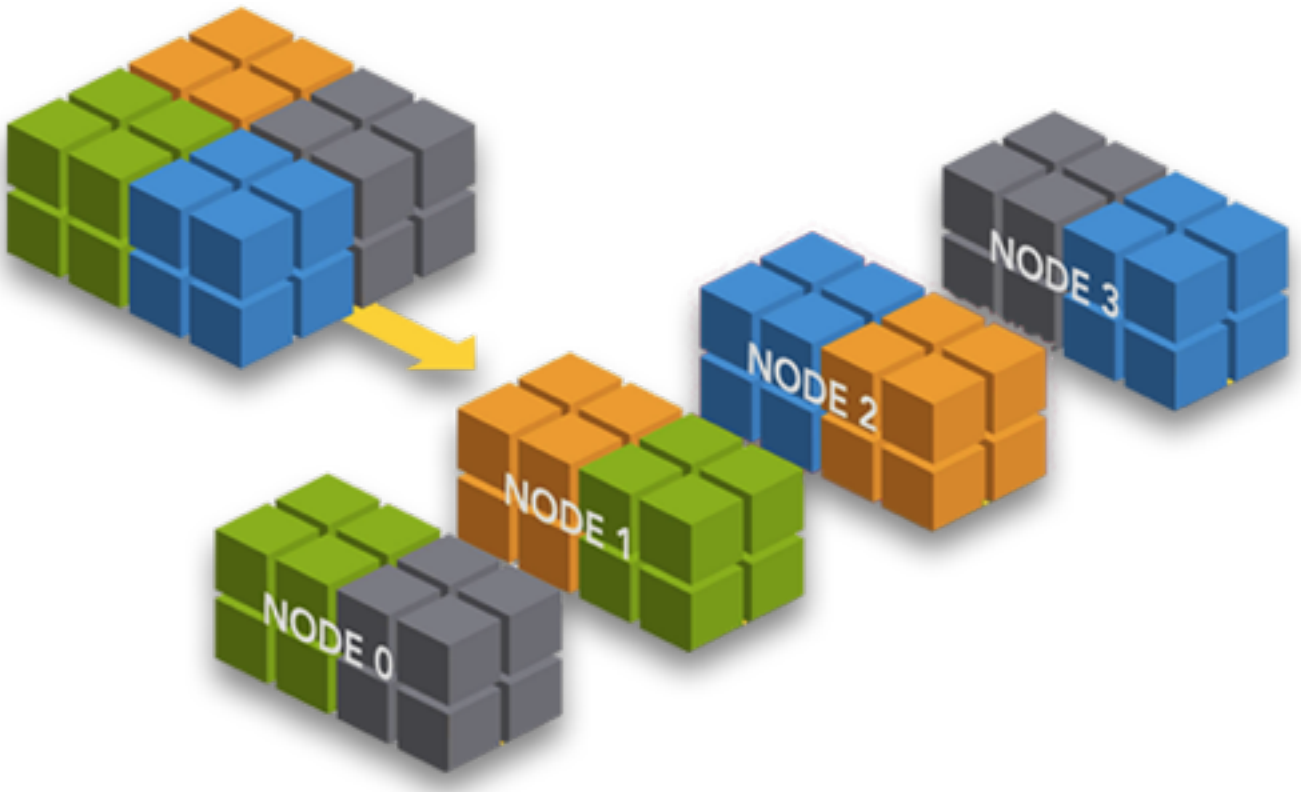
(otra estrategia → por datos para muchas imágenes, pero cada una tiene que caber en memoria)

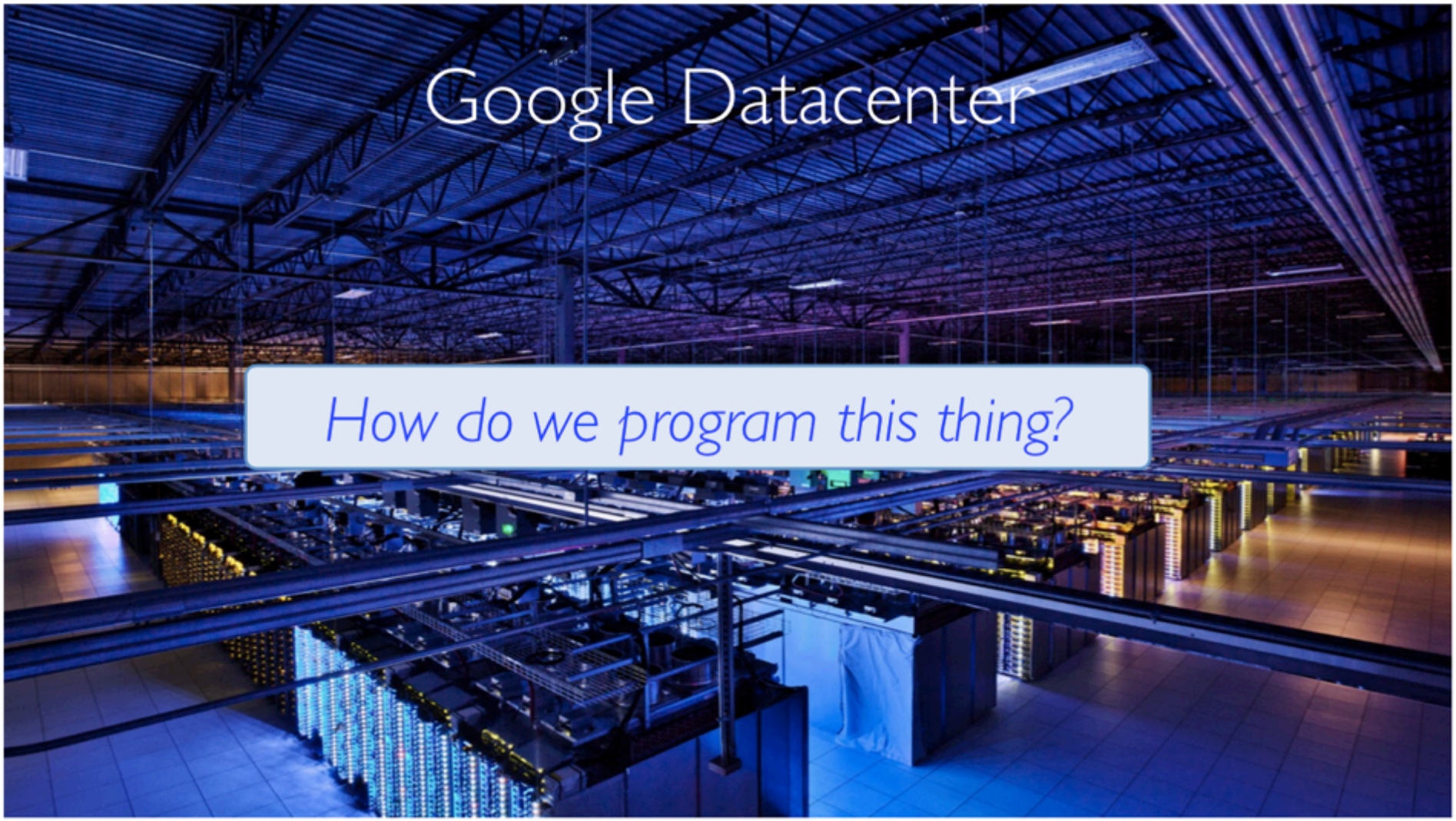


Ejemplos de algoritmos

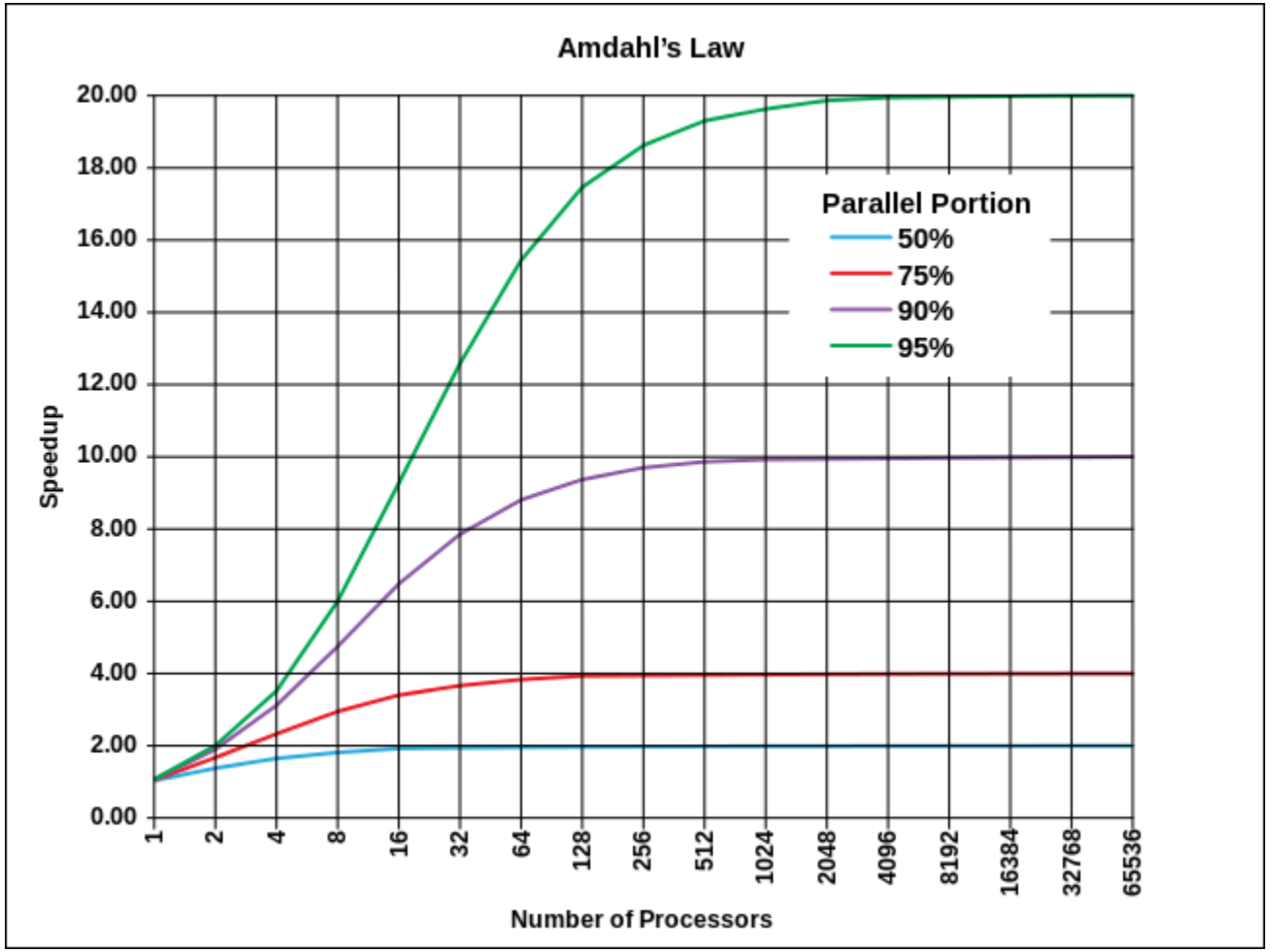
Particionamiento no trivial




NODE 0
NODE 1
NODE 2
NODE 3


Google Datacenter
How do we program this thing?

Amdahl's law


Amdahl's Law
Speedup
20.00
18.00
16.00
14.00
12.00
10.00
8.00
6.00
4.00
2.00
0.00
Parallel Portion
50%
75%
90%
95%
1
2
4
8
16
32
64
128
256
512
1024
2048
4096
8192
16384
32768
65536
Number of Processors

Parallelize computation of the mean

$$X = \{x_i\} \quad i \in [1, n]$$

$$\bar{x} = mean(X) = \frac{1}{m} \sum_{i=1}^{n} x_i$$

Parallelize computation of the mean

$$X = \{x_i\} \quad i \in [1, n]$$

$$\bar{x} = mean(X) = \frac{1}{m} \sum_{i=1}^{n} x_i$$

$$\bar{x} = \frac{1}{m} \left[\sum_{i=1}^{n_1} x_i + \sum_{i=n_1+1}^{n_2} x_i \; + \ldots + \sum_{i=n_{t-1}+1}^{n} x_i \right]$$

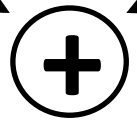
Parallelize computation of the mean

$$X = \{x_i\} \quad i \in [1, n]$$

$$\bar{x} = mean(X) = \frac{1}{m} \sum_{i=1}^{n} x_i$$

$$\bar{x} = \frac{1}{m} \left[\sum_{i=1}^{n_1} x_i + \sum_{i=n_1+1}^{n_2} x_i \; + \ldots + \sum_{i=n_{t-1}+1}^{n} x_i \right]$$

$$\sum_{i=1}^{n_1} x_i \qquad \sum_{i=n_1+1}^{n_2} x_i \qquad \ldots \qquad \sum_{i=n_{t-1}+1}^{n} x_i$$



How to scale?

$$X = \{x_i\} \quad i \in [1, n]$$

$$\bar{x} = mean(X) = \frac{1}{m}\sum_{i=1}^{n} x_i \qquad var(X) = \frac{1}{m}\sum_{i=1}^{n} (\bar{x} - x_i)^2$$

How to scale?

$$X = \{x_i\} \quad i \in [1, n]$$

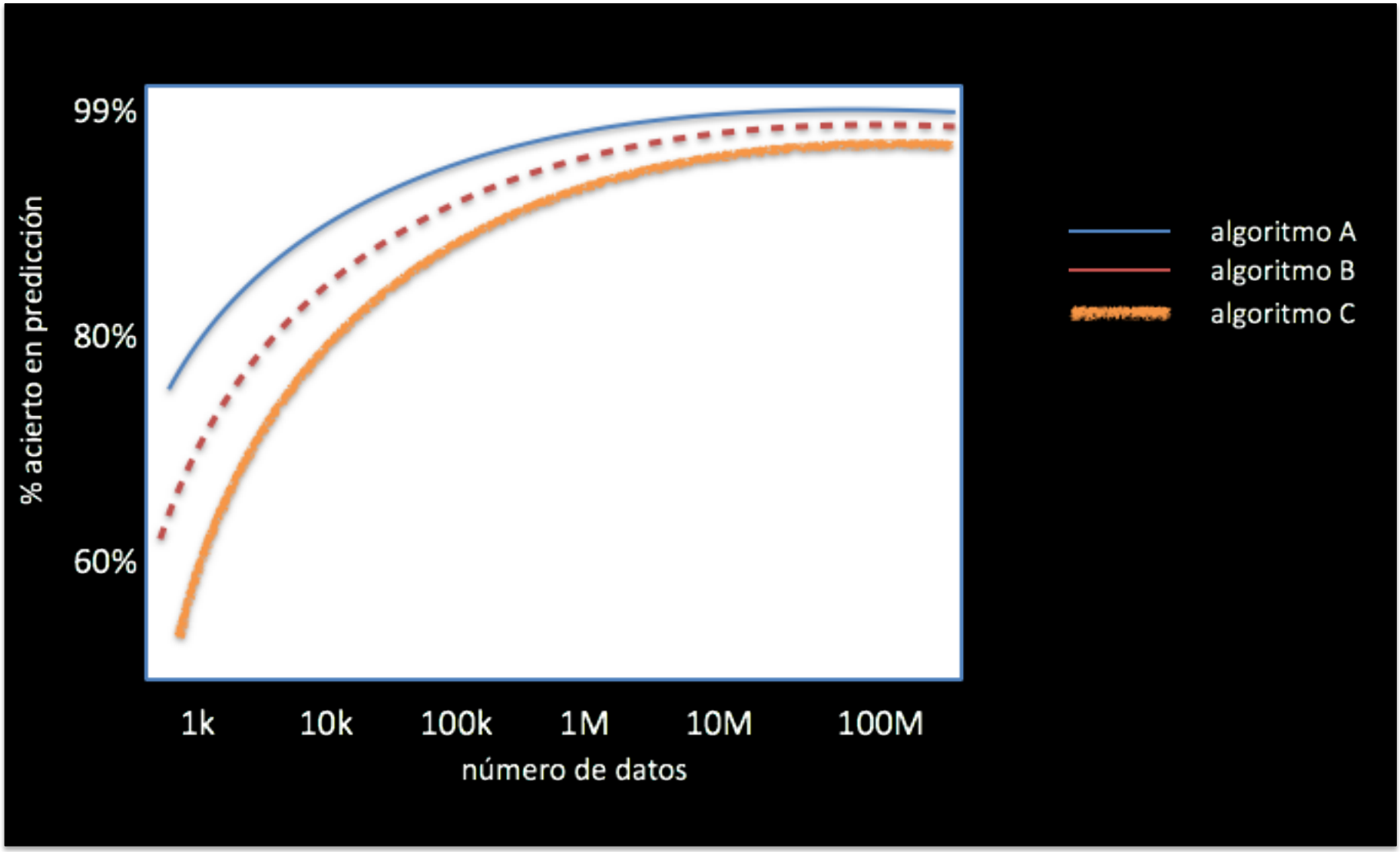
$$\bar{x} = mean(X) = \frac{1}{m}\sum_{i=1}^{n} x_i \qquad var(X) = \frac{1}{m}\sum_{i=1}^{n}(\bar{x} - x_i)^2$$

$$var(X) = \frac{1}{m}\left[\sum \bar{x}^2 - 2\bar{x}\sum x_i + \sum x_i^2\right]$$

$$var(X) = \frac{1}{m}\left[\frac{1}{m}\left(\sum x_i\right)^2 - \frac{1}{m}2\left(\sum x_i\right)^2 + \sum x_i^2\right]$$

$$var(X) = \frac{1}{m}\sum x_i^2 - \frac{1}{m^2}\left(\sum x_i\right)^2$$

Predictive Analytics in Big Data

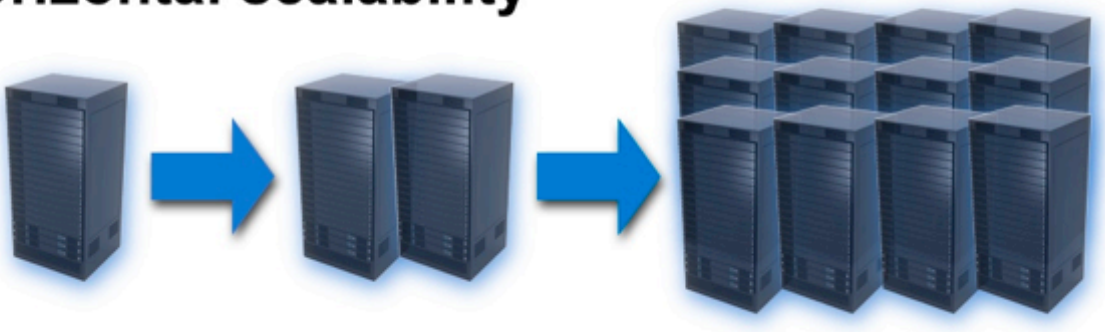


Scalability in Big Data

Vertical "scalability"

scale up
traditional DBs

Horizontal scalability



scale out
noSQL

seek "triviality" for appropriate sw+hw architectures

recent technologies (virtualization, etc.) tend to favor the cost of scaling out

How to scale?

Parallelize the code yourself!!

Lots of flexibility …. but HARD!!!

map reduce

```
2012-01-01 09:08 BOG Libros 88.56 Discover
2012-01-01 09:09 BGA Libros 337.71 Efectivo
2012-01-01 09:52 BGA Libros 62.41 Discover
2012-01-01 10:08 MED Musica 93.37 Visa
2012-01-01 10:22 BGA Musica 369.94 MasterCard
2012-01-01 10:58 MED Musica 119.12 Efectivo
2012-01-01 11:36 BOG Musica 296.76 Discover
```

map reduce

```
BOG 88.56
BGA 337.71
BGA 62.41
MED 93.37
BGA 369.94
MED 119.12
BOG 296.76
```

```
BOG 88.56 296.76

BGA 337.71 62.41 369.94

MED 93.37 119.12
```

```
BOG 385.32

BGA 770.06

MED 212.49
```

map reduce

```
map (input record):
    yield key=record[2], value=record[4]
```

shuffle → gathers all values for each key

```
reduce (key, [values]):
    yield key, sum(values)
```

map reduce

```
map (input record):
    yield key1, value1

shuffle


reduce (key1, [list of values]):
   yield key2, value2
```

How to crunch 1PB?

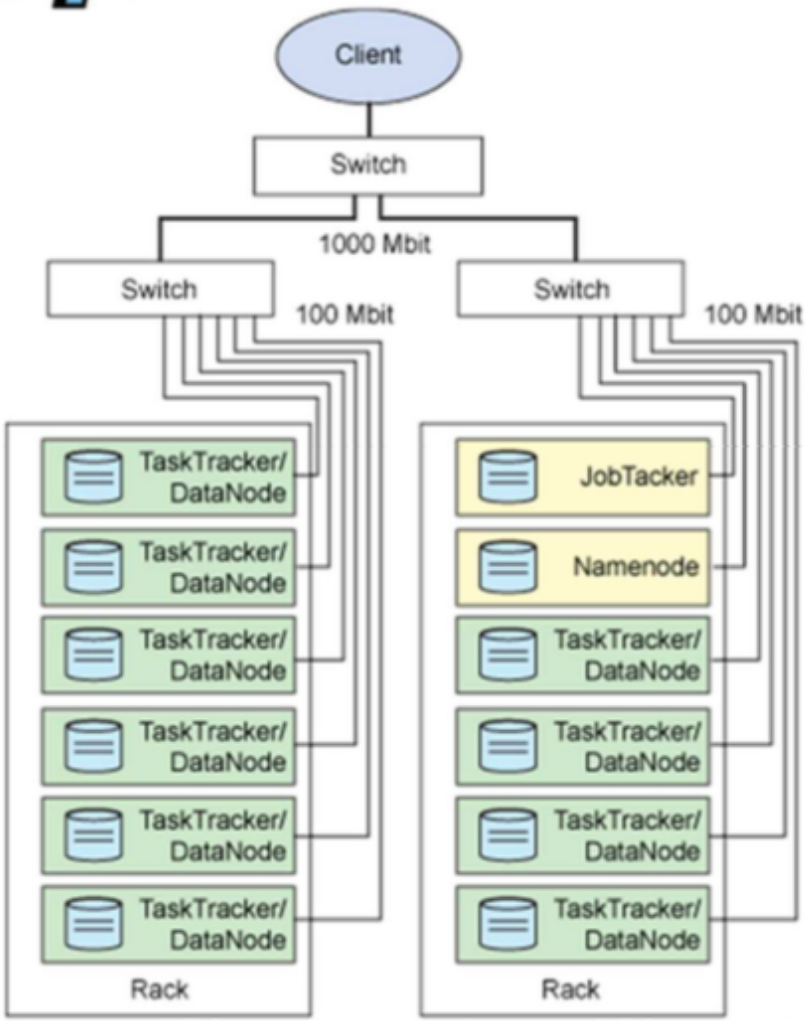
Lots of disks spinning all the time
Redundancy, since disks die
Lots of CPU cores, working all the time
Retry, since network errors happen

Design Qualities

Scalable – many servers with cores and disks
Reliable – redundant storage
Fault-tolerant – auto retry, self healing

Computation to Data

Very simple computing model Map-Reduce
Each computing node is also a storage node
HDFS on top of ext3, fixed 64MB file blocks
write once, read many



Computing goes to data

Data goes to computing

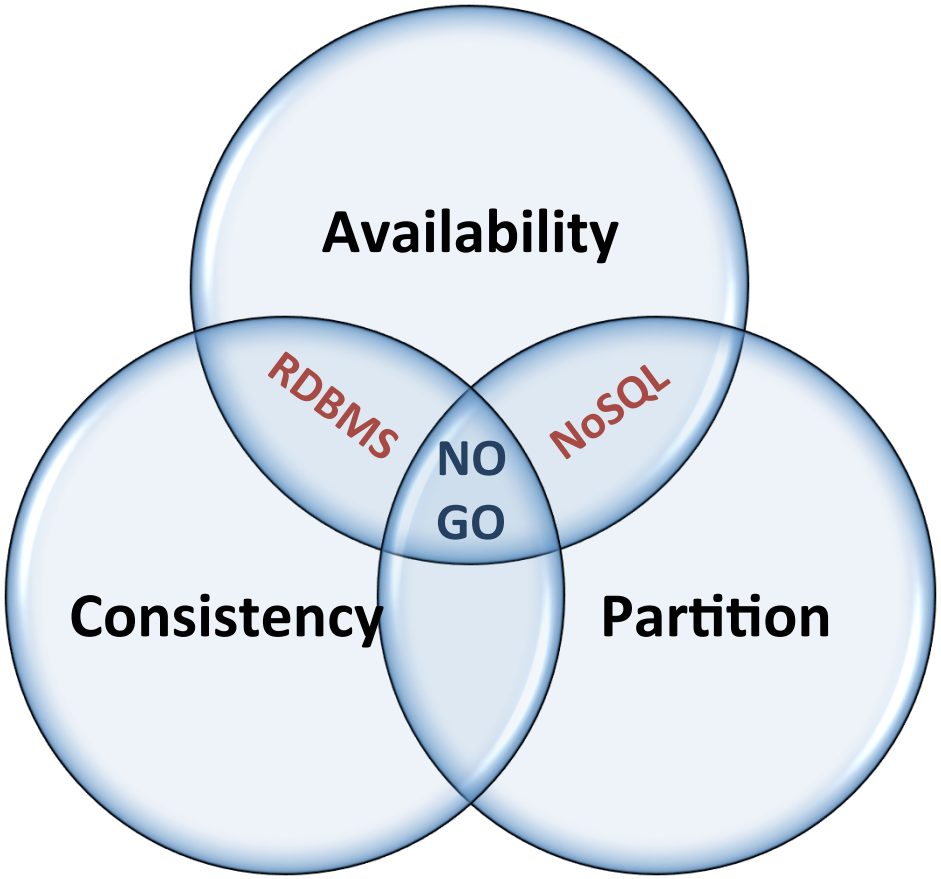
NoSQL

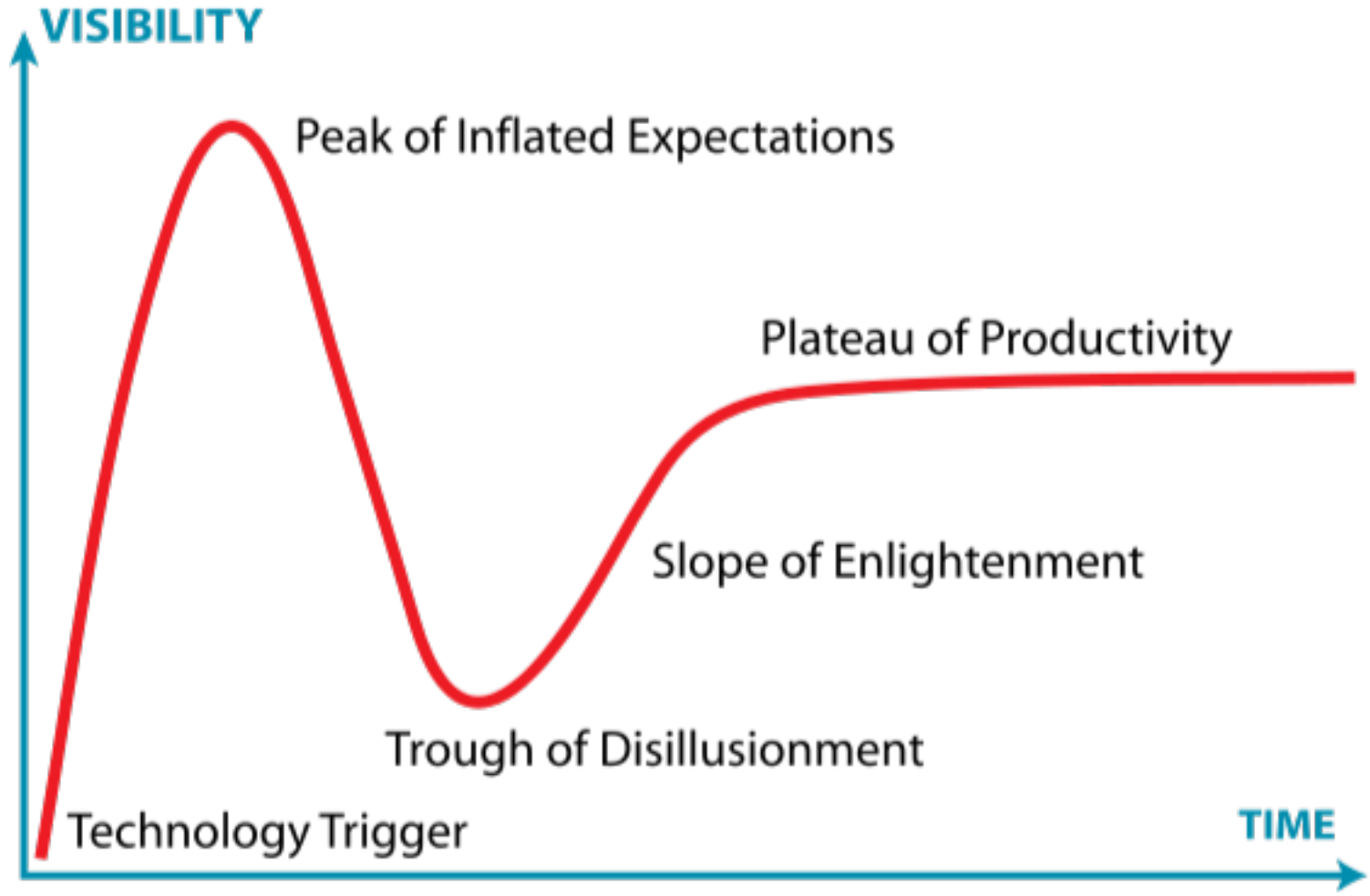
- Expressivity SQL vs. Scalability

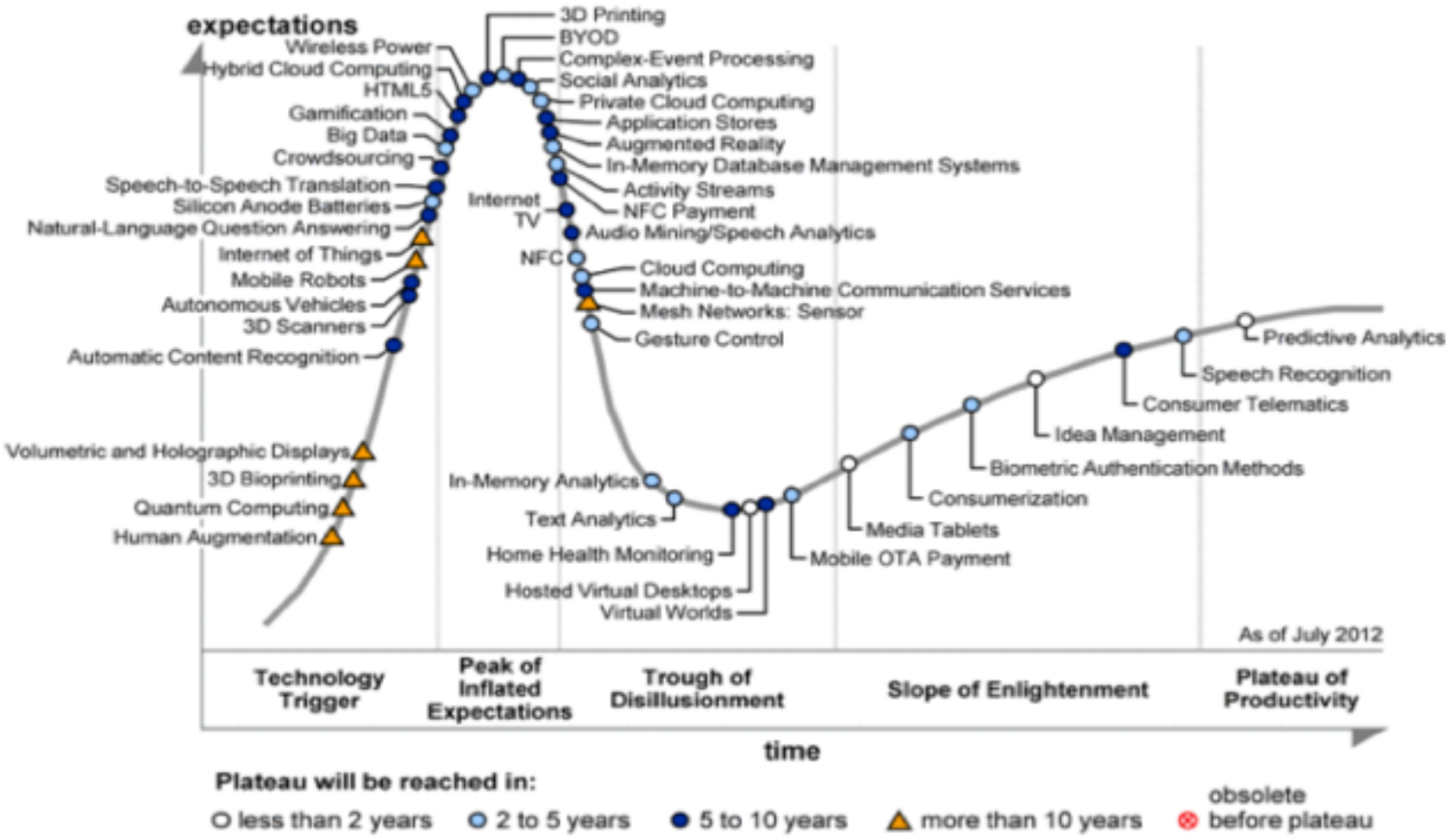
- Simpler data model (key, values)
- Simpler operations
  - Scan/access per key, basic transactions (check&put)
  - No joins, no SQL language
- Simple failover and scale up
- Big table, Hbase, DynamoDB, Azure, Cassandra, etc.

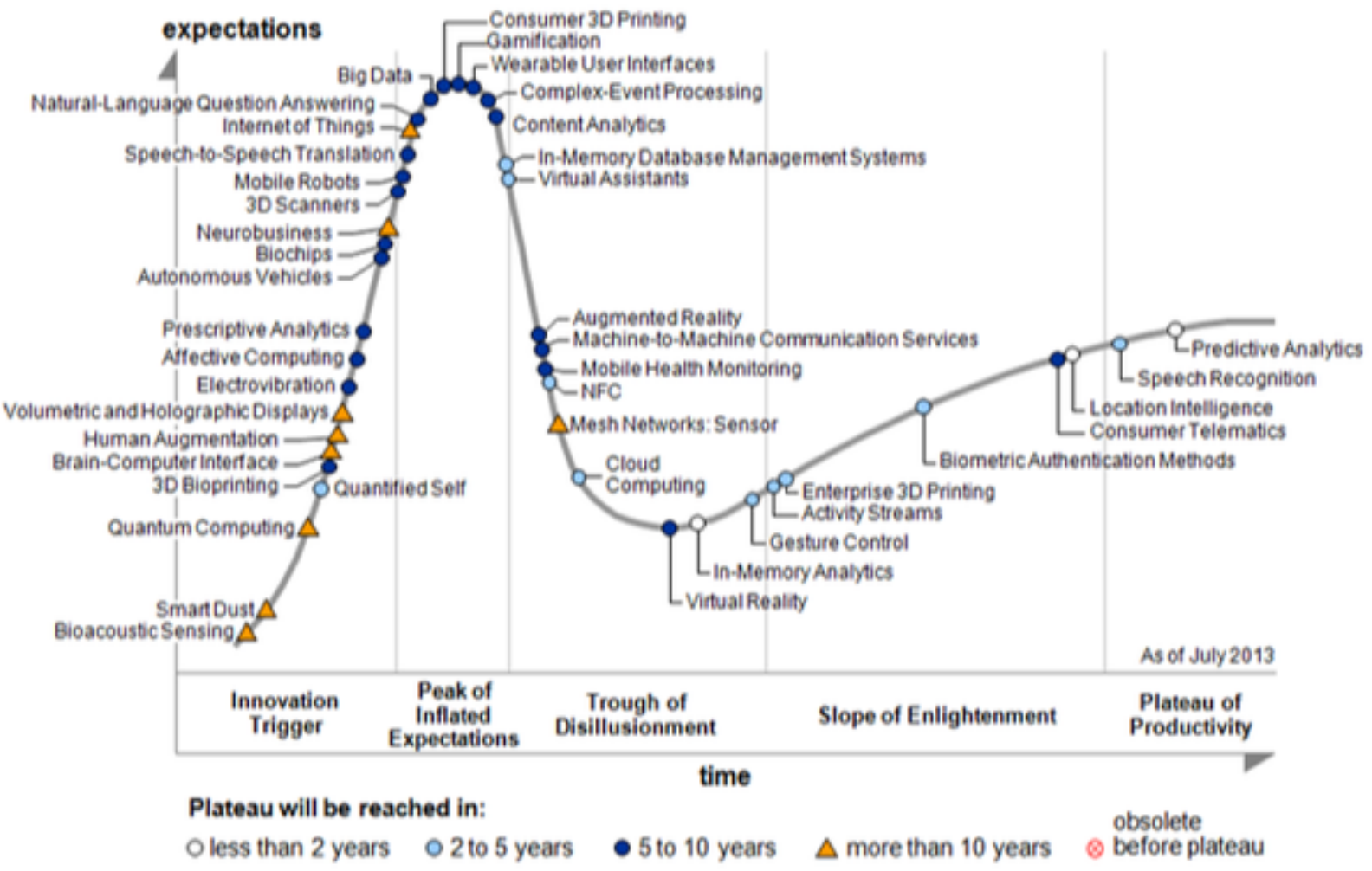
→ more work for programmers!!!

NoSQL Eventual Consistency




VISIBILITY
Peak of Inflated Expectations
Plateau of Productivity
Slope of Enlightenment
Trough of Disillusionment
Technology Trigger
TIME


expectations
3D Printing
BYOD
Complex-Event Processing
Social Analytics
Private Cloud Computing
Application Stores
Augmented Reality
In-Memory Database Management Systems
Activity Streams
NFC Payment
Audio Mining/Speech Analytics
Cloud Computing
Machine-to-Machine Communication Services
Mesh Networks: Sensor
Gesture Control
Internet TV
NFC
Wireless Power
Hybrid Cloud Computing
HTML5
Gamification
Big Data
Crowdsourcing
Speech-to-Speech Translation
Silicon Anode Batteries
Natural-Language Question Answering
Internet of Things
Mobile Robots
Autonomous Vehicles
3D Scanners
Automatic Content Recognition
Volumetric and Holographic Displays
3D Bioprinting
Quantum Computing
Human Augmentation
In-Memory Analytics
Text Analytics
Home Health Monitoring
Hosted Virtual Desktops
Virtual Worlds
Media Tablets
Mobile OTA Payment
Consumerization
Biometric Authentication Methods
Idea Management
Consumer Telematics
Speech Recognition
Predictive Analytics
As of July 2012
Technology Trigger
Peak of Inflated Expectations
Trough of Disillusionment
Slope of Enlightenment
Plateau of Productivity
time
Plateau will be reached in:
less than 2 years
2 to 5 years
5 to 10 years
more than 10 years
obsolete before plateau


expectations
Consumer 3D Printing
Gamification
Wearable User Interfaces
Big Data
Complex-Event Processing
Natural-Language Question Answering
Content Analytics
Internet of Things
In-Memory Database Management Systems
Speech-to-Speech Translation
Virtual Assistants
Mobile Robots
3D Scanners
Neurobusiness
Biochips
Autonomous Vehicles
Augmented Reality
Prescriptive Analytics
Machine-to-Machine Communication Services
Affective Computing
Mobile Health Monitoring
Electrovibration
NFC
Volumetric and Holographic Displays
Mesh Networks: Sensor
Human Augmentation
Brain-Computer Interface
Cloud Computing
3D Bioprinting
Quantified Self
Enterprise 3D Printing
Activity Streams
Quantum Computing
Gesture Control
In-Memory Analytics
Virtual Reality
Smart Dust
Bioacoustic Sensing
Predictive Analytics
Speech Recognition
Location Intelligence
Consumer Telematics
Biometric Authentication Methods
As of July 2013
Innovation Trigger
Peak of Inflated Expectations
Trough of Disillusionment
Slope of Enlightenment
Plateau of Productivity
time
Plateau will be reached in:
less than 2 years
2 to 5 years
5 to 10 years
more than 10 years
obsolete before plateau

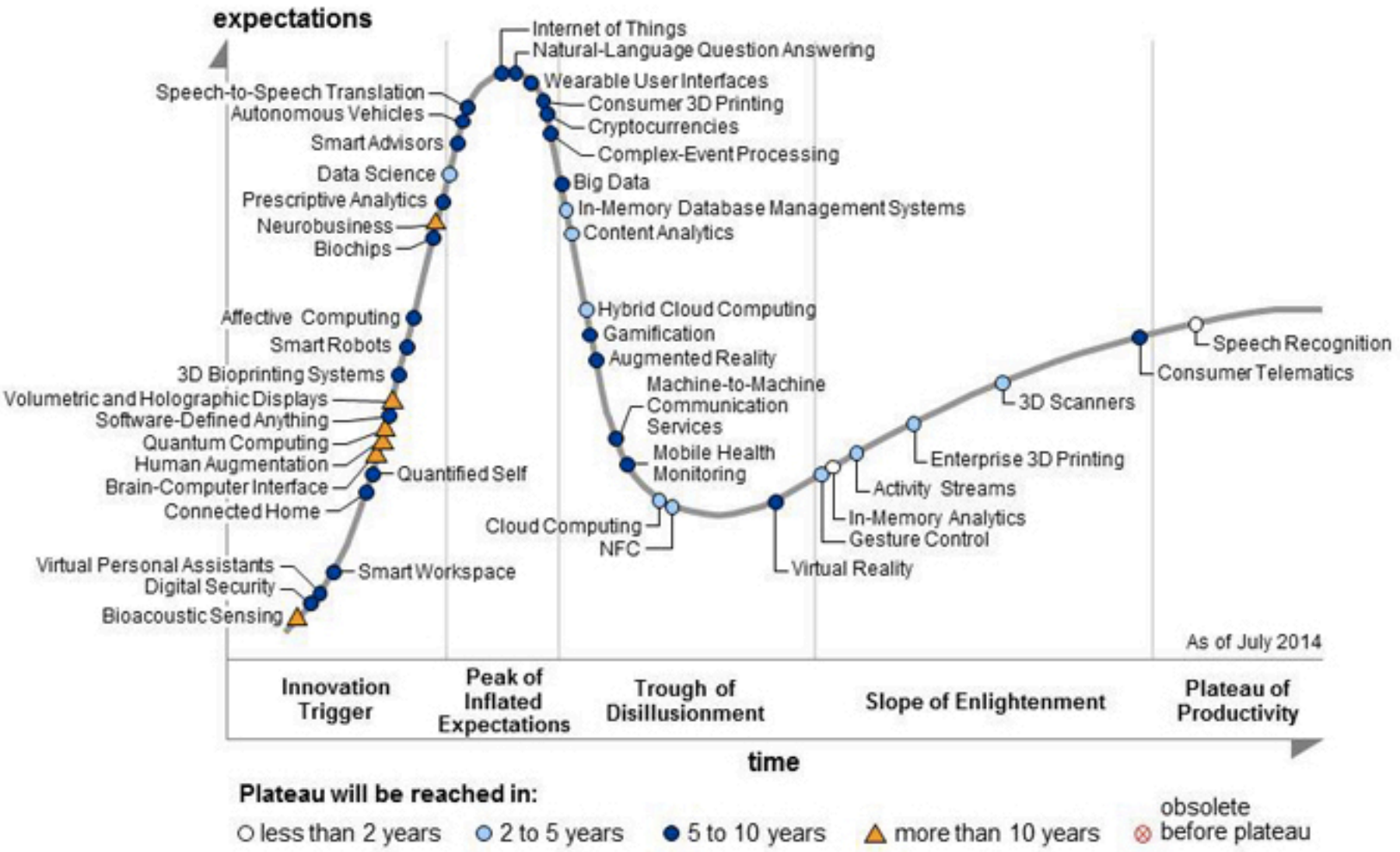
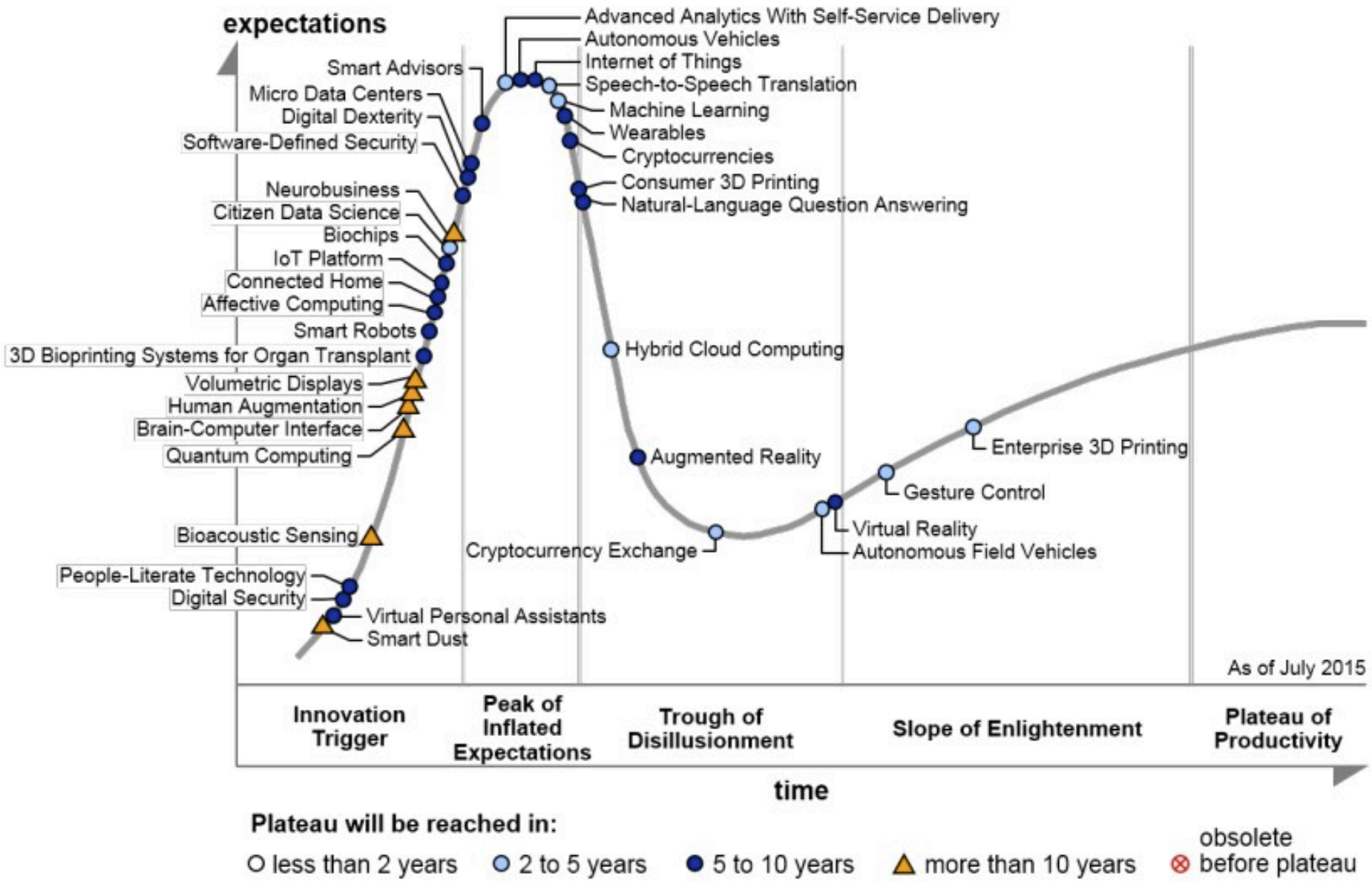

expectations
Internet of Things
Natural-Language Question Answering
Wearable User Interfaces
Consumer 3D Printing
Cryptocurrencies
Complex-Event Processing
Big Data
In-Memory Database Management Systems
Content Analytics
Hybrid Cloud Computing
Gamification
Augmented Reality
Machine-to-Machine Communication Services
Mobile Health Monitoring
Cloud Computing
NFC
Virtual Reality
In-Memory Analytics
Gesture Control
Activity Streams
Enterprise 3D Printing
3D Scanners
Consumer Telematics
Speech Recognition
Speech-to-Speech Translation
Autonomous Vehicles
Smart Advisors
Data Science
Prescriptive Analytics
Neurobusiness
Biochips
Affective Computing
Smart Robots
3D Bioprinting Systems
Volumetric and Holographic Displays
Software-Defined Anything
Quantum Computing
Human Augmentation
Brain-Computer Interface
Connected Home
Quantified Self
Virtual Personal Assistants
Digital Security
Bioacoustic Sensing
Smart Workspace
As of July 2014
Innovation Trigger
Peak of Inflated Expectations
Trough of Disillusionment
Slope of Enlightenment
Plateau of Productivity
time
Plateau will be reached in:
less than 2 years
2 to 5 years
5 to 10 years
more than 10 years
obsolete before plateau

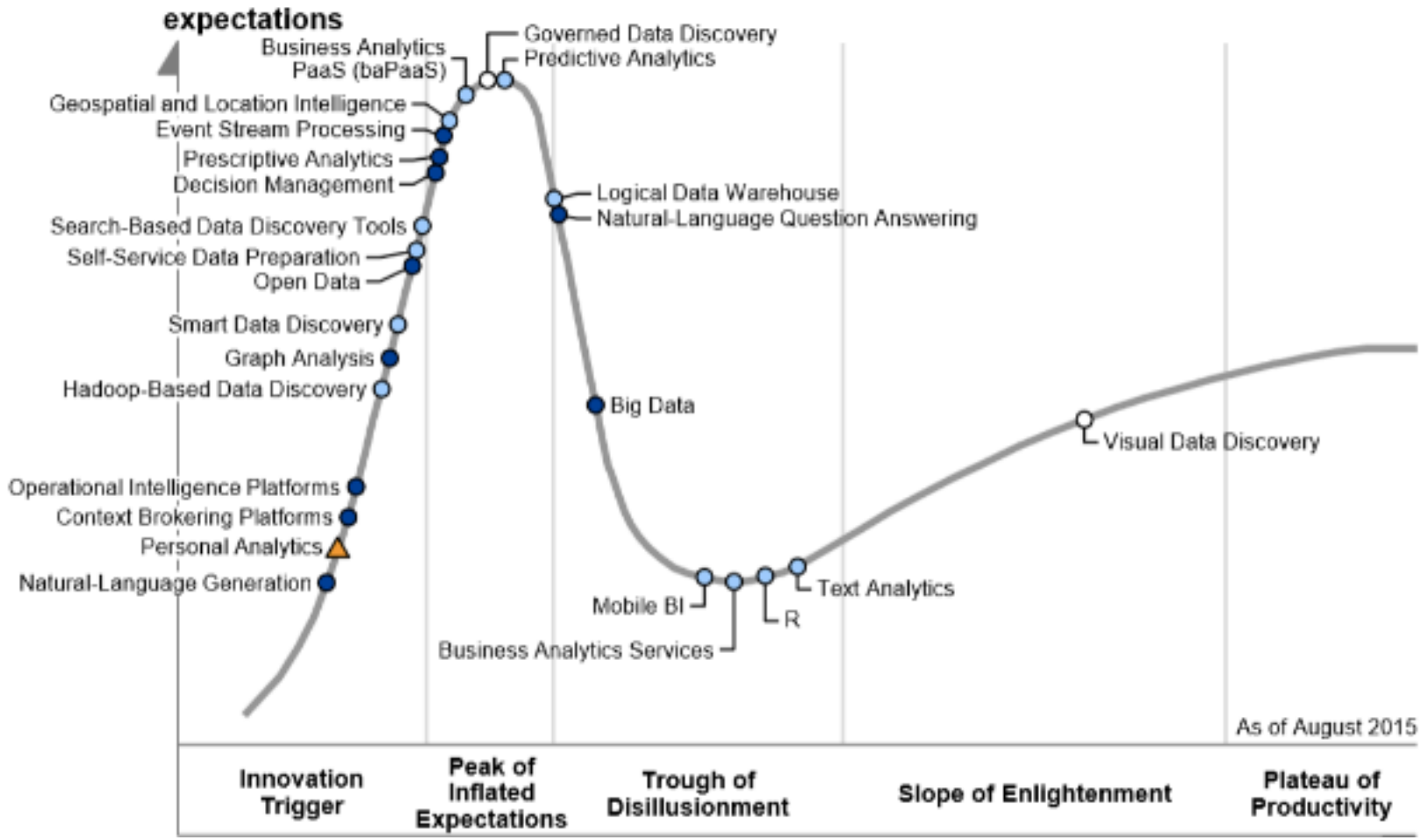
Figure 2. Hype Cycle for Emerging Technologies, 2015



Source: Gartner (July 2015)

BIG DATA OUT OF EMERGING TECHNOLOGIES

Figure 1. Hype Cycle for Business Intelligence and Analytics, 2015



BIG DATA IN ANALYTICS