Instituto Superior Técnico
Technical University of Lisbon

Big Data

Bruno Lopes
Catarina Moreira
João Pinho

Motivation

2 220 PetaBytes

Of data that people create **every day**!

The Chartered Institute for IT
Enabling the information society

Motivation

90 % of Data

UNSTRUCTURED

Which is difficult to **manage** and **interpret!**

Motivation

2 hours per day

Spent searching for the right **information**

The Chartered Institute for IT
Enabling the information society

Motivation

$5 million per year

Money lost due to **data-related** problems

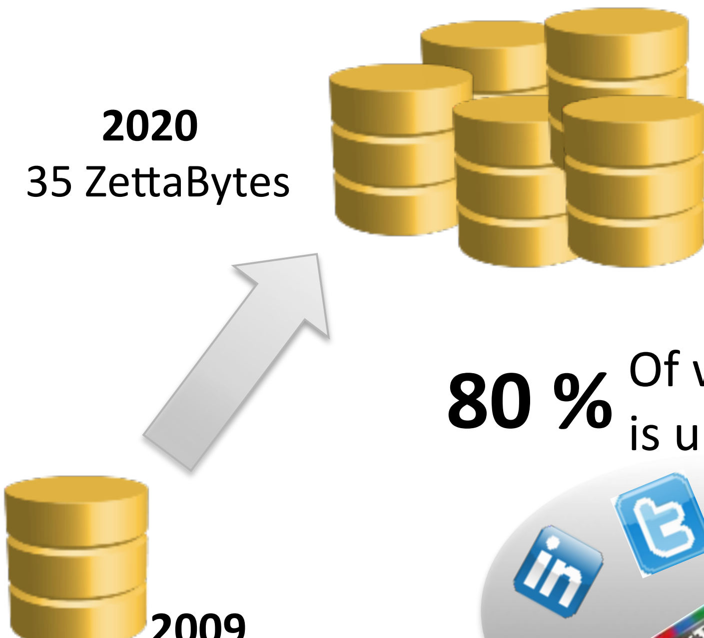
The Chartered Institute for IT
Enabling the information society

Motivation

Information is at the Center of a New Wave of Opportunity

44 % as much Data and Content Over the Coming Decade

2020
35 ZettaBytes



2009
800 000 PetaBytes

80 % Of world's data is unstructured

Organizations Need Deeper Insights …

1 in 3 Business leaders make **decisions** based on **information** they **don't trust** or **don't have**

1 in 2 Business leaders say they don't have **access** to the **information** they need to do their jobs

60% CEO's need to do a better job **capturing** and **understanding** information rapidly in order to make swift **business decisions**

Source: *Big Data by Ami Redwan Haq, Founder at Sentinel Solutions Ltd*

What is Big Data?

Big Data (video)

What is Big Data?

“ Data that **exceeds** the processing **capacity** of a conventional database. The data is too **big**, moves too **fast** or doesn’t fit the structures of your **database architectures**. ”

Source: *J. Hurwitz, A. Nugent, F. Halper & M. Kaufman, Big Data for Dummies, Willey & Sons, 2013*

Why Big Data?

" Big data enables organizations to **store**, **manage** and **manipulate** vast amounts of **data** at the right **speed** and at the right **time** to gain the right **insights**. "

Source: *J. Hurwitz, A. Nugent, F. Halper & M. Kaufman, Big Data for Dummies, Willey & Sons, 2013*

Why Big Data?

- The value of Big Data falls into 2 categories:

1. **Analytical Use**.
   - Reveal new **insights** that were **hidden** in too costly to process massive amounts of data;
   - Example: Peer influence among customers;

2. **Enabling New Products.**
   - Combining large number of signals from user's actions and friends
   - Example: Facebook

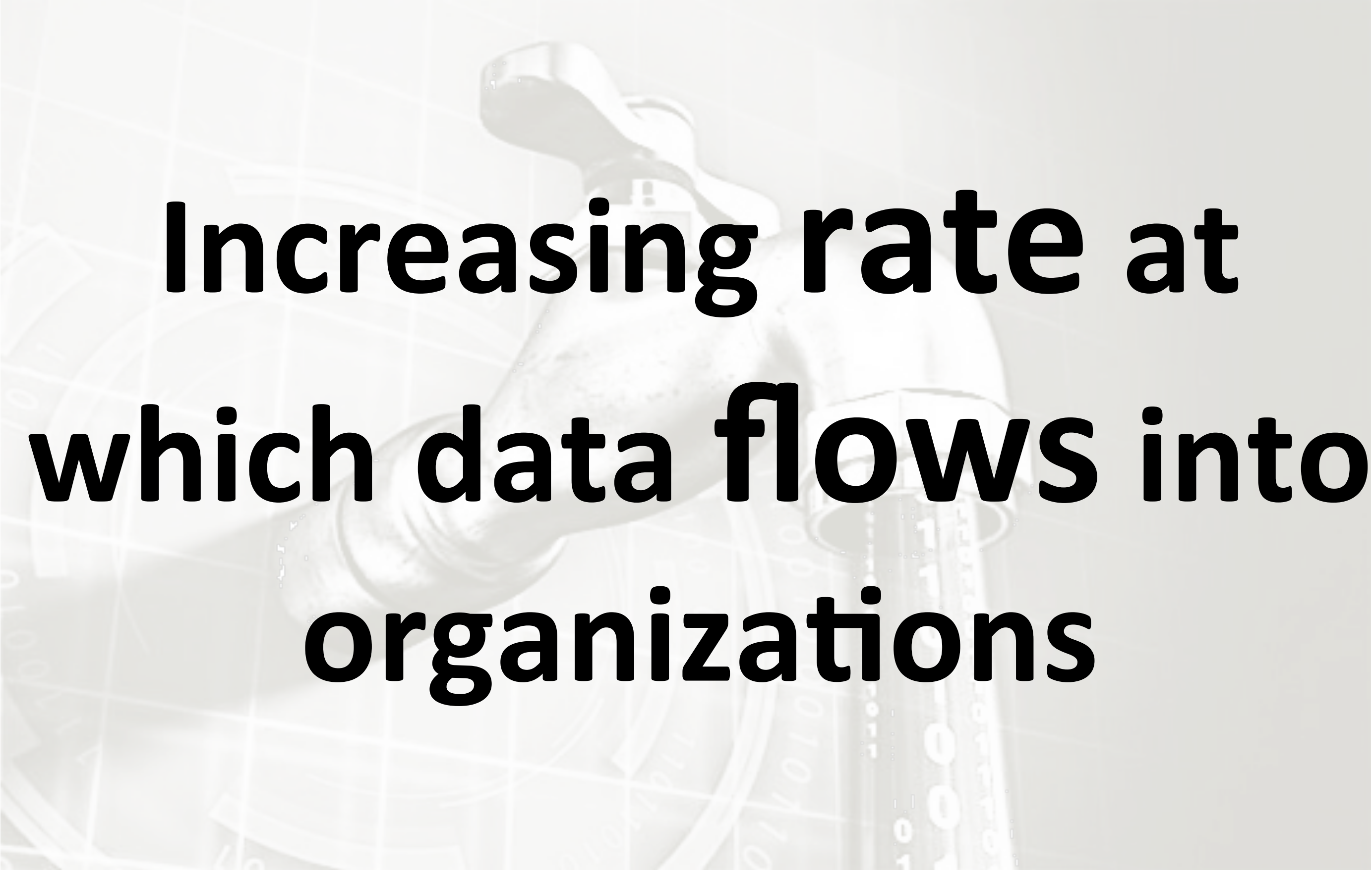
The V's

Big Data is defined as any kind of data source that has at least three shared characteristics:

- Velocity
- Variety
- Volume

- Viability
- Value

Velocity



Increasing rate at which data flows into organizations

Velocity

The Large Hadron Collider at CERN generates

25 PetaBytes per Year!

Scientists are forced to

Discard Massive Data

In order to keep **storage** requirements practical

Velocity

The way we deliver and consume products and services is generating a **massive data flow** to the provider

- It's possible to stream fast-moving data **into bulk storage** for later batch processing.

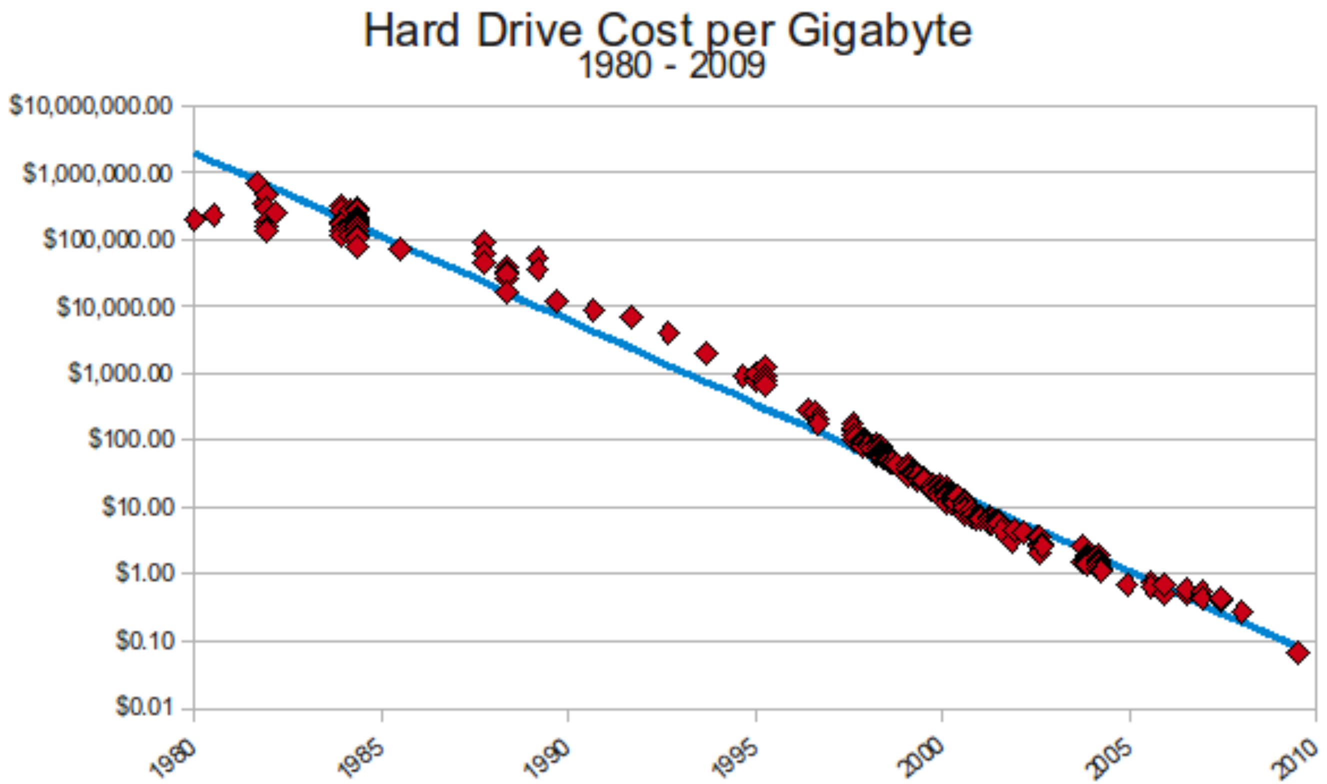
- The importance lies in taking data **from input through decision**.

Volume

Ability to **store massive** amounts of (un)structured **data**

Volume

No longer a **problem!**

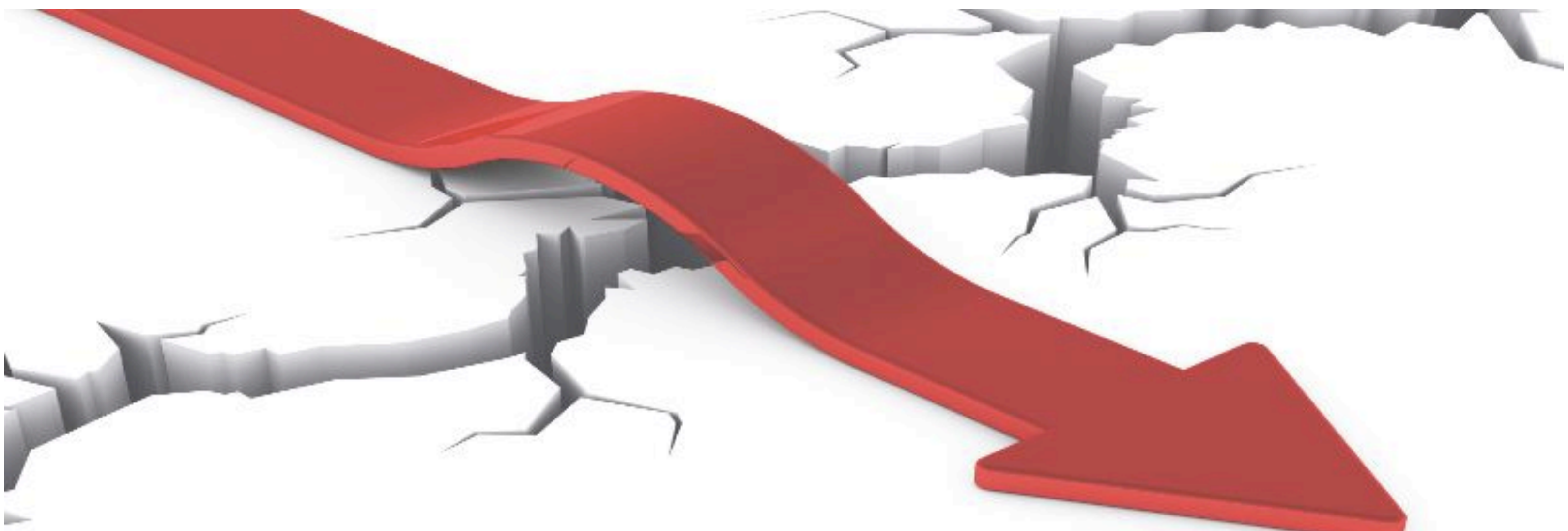


Volume

Don't forget the **cloud!**

Volume - Challenges

- How to determine **relevance** within large data volumes
- How to use analytics to create **value** from relevant data.



Variety

Managing, **merging** and **governing** different varieties of data.

Variety

- Source data is diverse and doesn't fall into neat relational structures
- Data comes in the form of emails, photos, videos, monitoring devices, PDFs, audio, etc.
- This variety of (un)structured data creates problems for **storage**, **mining** and **analyzing** data

Viability

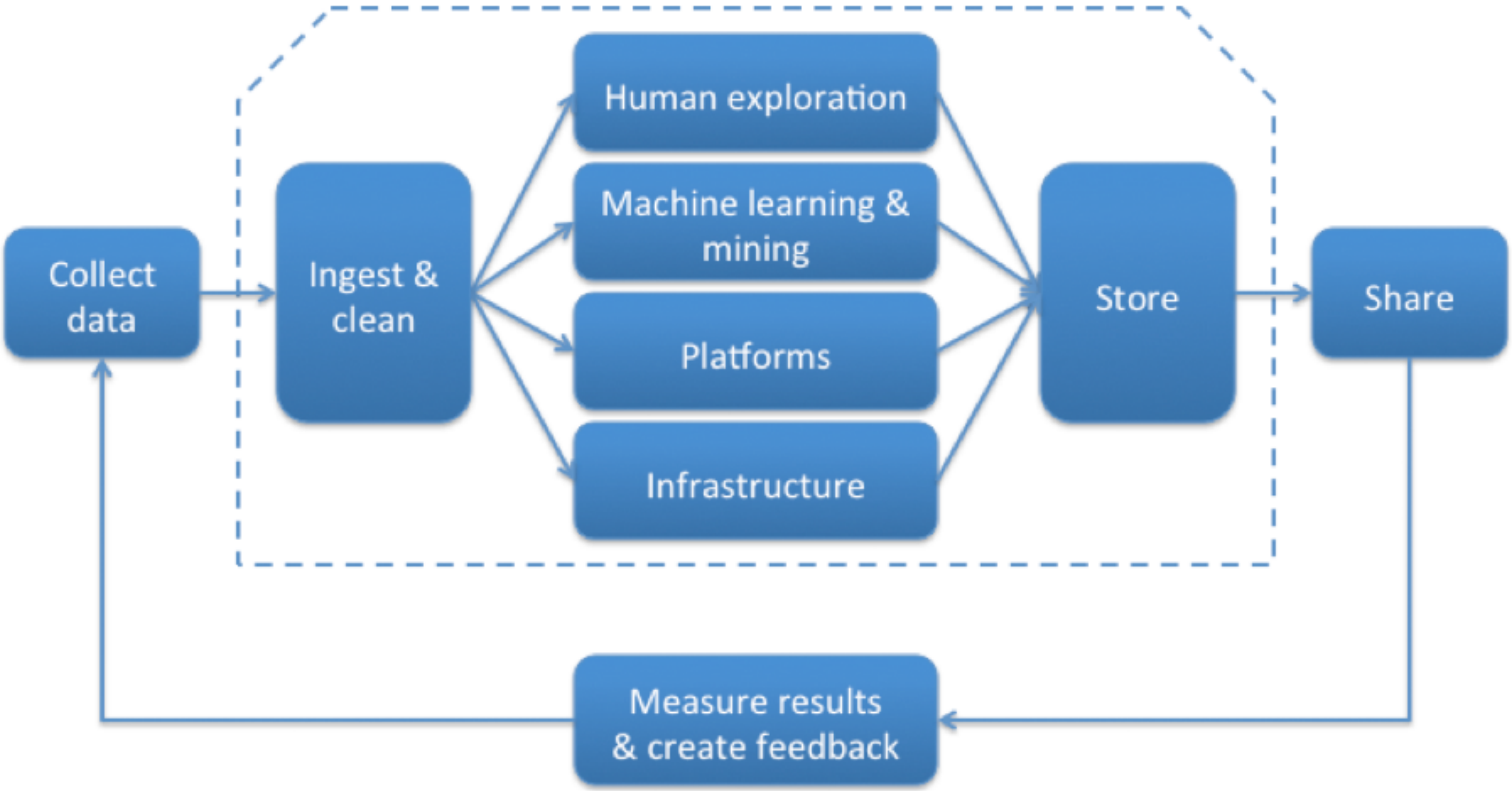
Quickly and **cost-effectively** test and confirm a particular variable's relevance before investing in the creation of a fully featured model.

Value

Gain **new insights** about data by analyzing variables that were **hidden**

So, **how** does it **work?**

The Big Data Supply Chain



Source: *Edd Dumbill, Planning for Big Data, O'Reilly, 2012*

Summary

- Motivation of why companies need to learn how to deal with Big Data!
- Presented a definition of Big Data and analyzed the V's.
- Presented an analysis of the Big Data Supply chain for enterprises.

Technologies for Big Data

The Reach of Big Data

- So far, the analytical use of Big Data has been in reach only to leading cooperations:

Walmart Google

But at a fantastic **cost!**

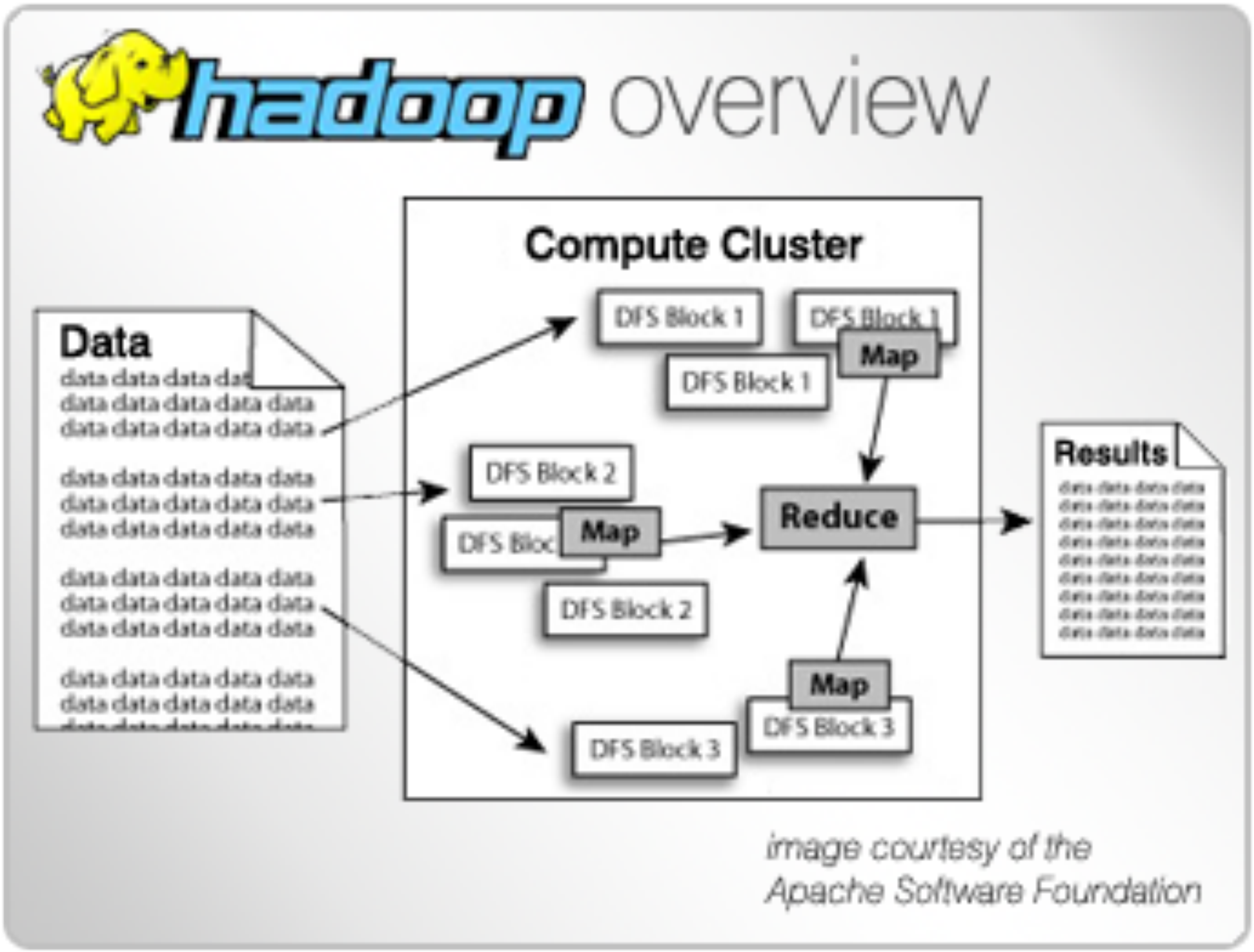
The Reach of Big Data

But now, there are several **open source** applications than can tackle Big Data related challenges!

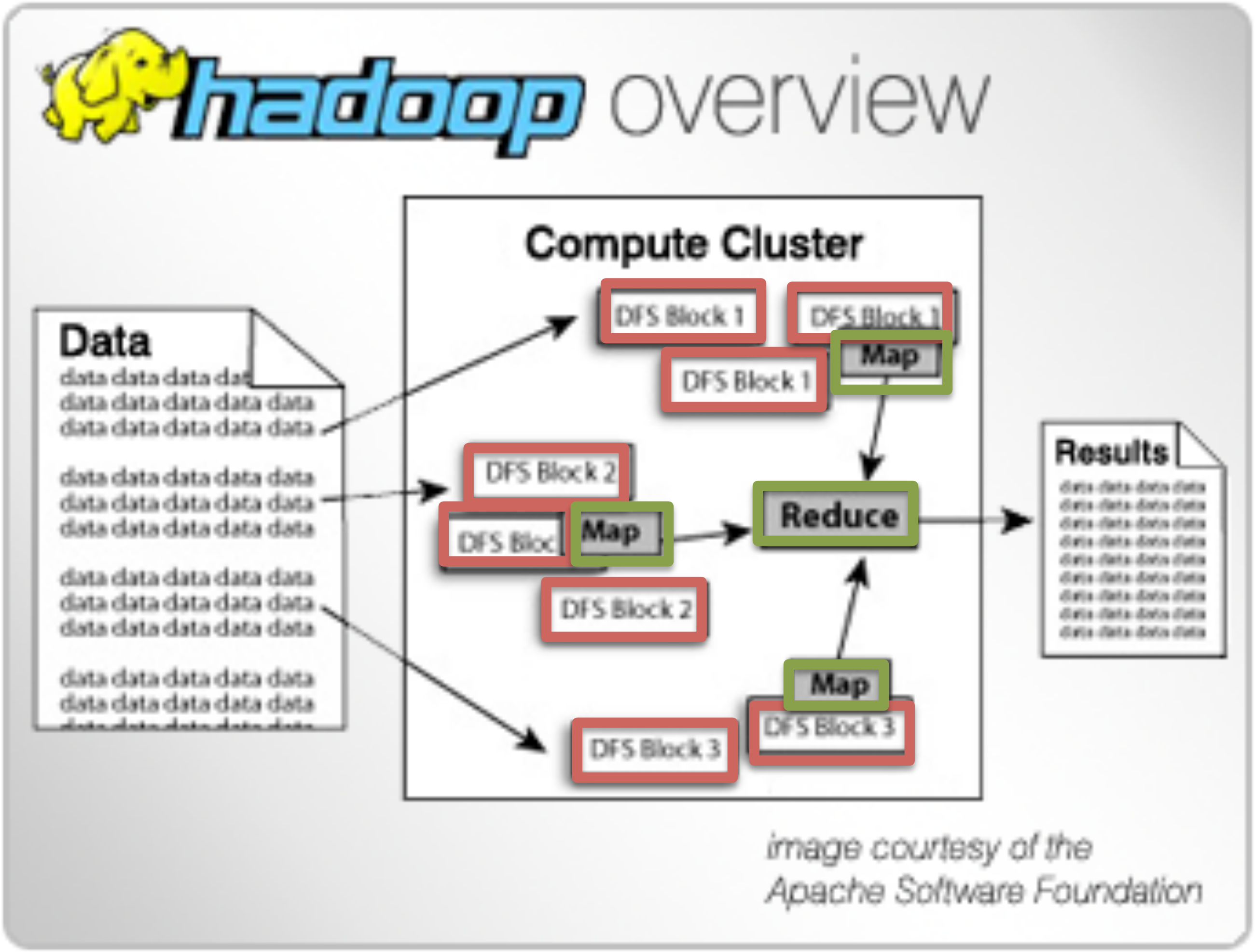
Apache Hadoop

- **Developed** and **released** by Yahoo!
- Open source framework for **storage** and **large scale parallel processing** of datasets on clusters.
- Ability to **cheaply** process large amounts of data, regardless of its **structure**.

Apache Hadoop



Apache Hadoop



(Video - MapReduce)

- IBM – MapReduce
- Health Care - Real Time Alerts

HDFS – Hadoop Distributed File System

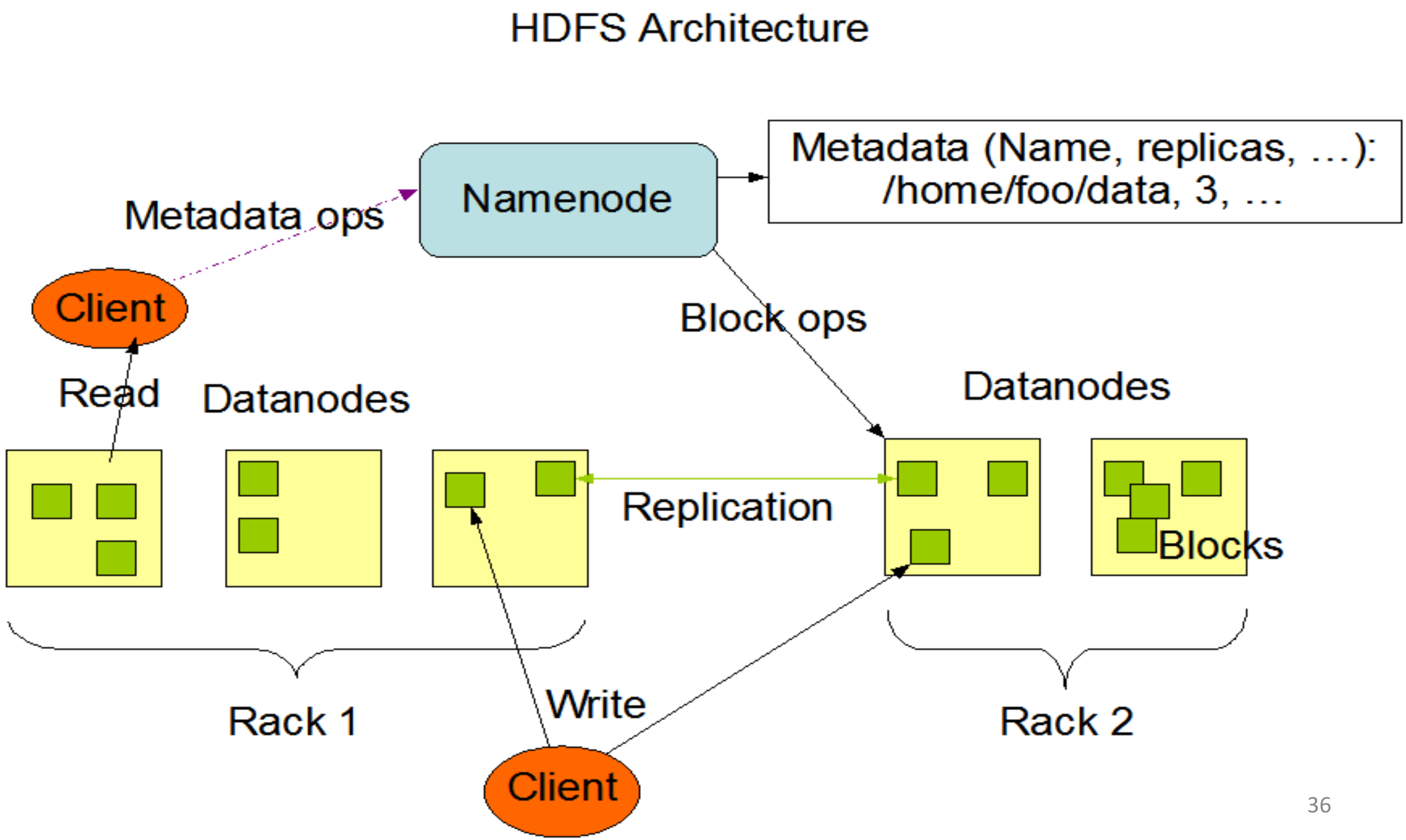
HDFS – Hadoop Distributed File System

- Detection of faults and quick, automatic recovery
- Emphasis on high throughput of data access
- Tuned to support large datasets
- Write-once-read-many access model: once a file is created, written and closed, needs not to be changed

HDFS – Hadoop Distributed File System

- Moving computation is cheaper than moving data
  - minimizes network congestion
  - Increases the overall throughput of the system
- Portability Across Heterogeneous Hardware and Software Platforms

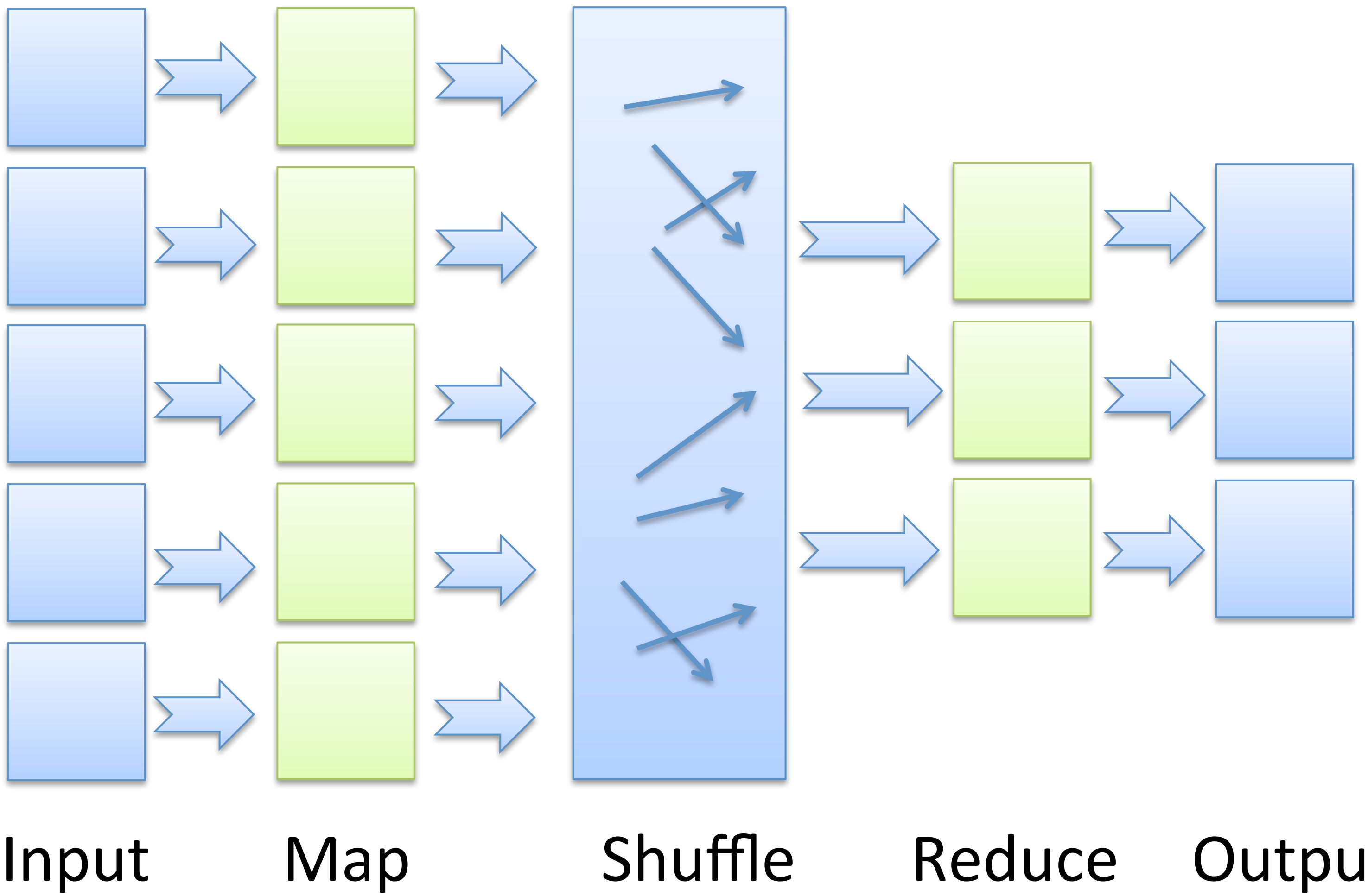
HDFS – Hadoop Distributed File System



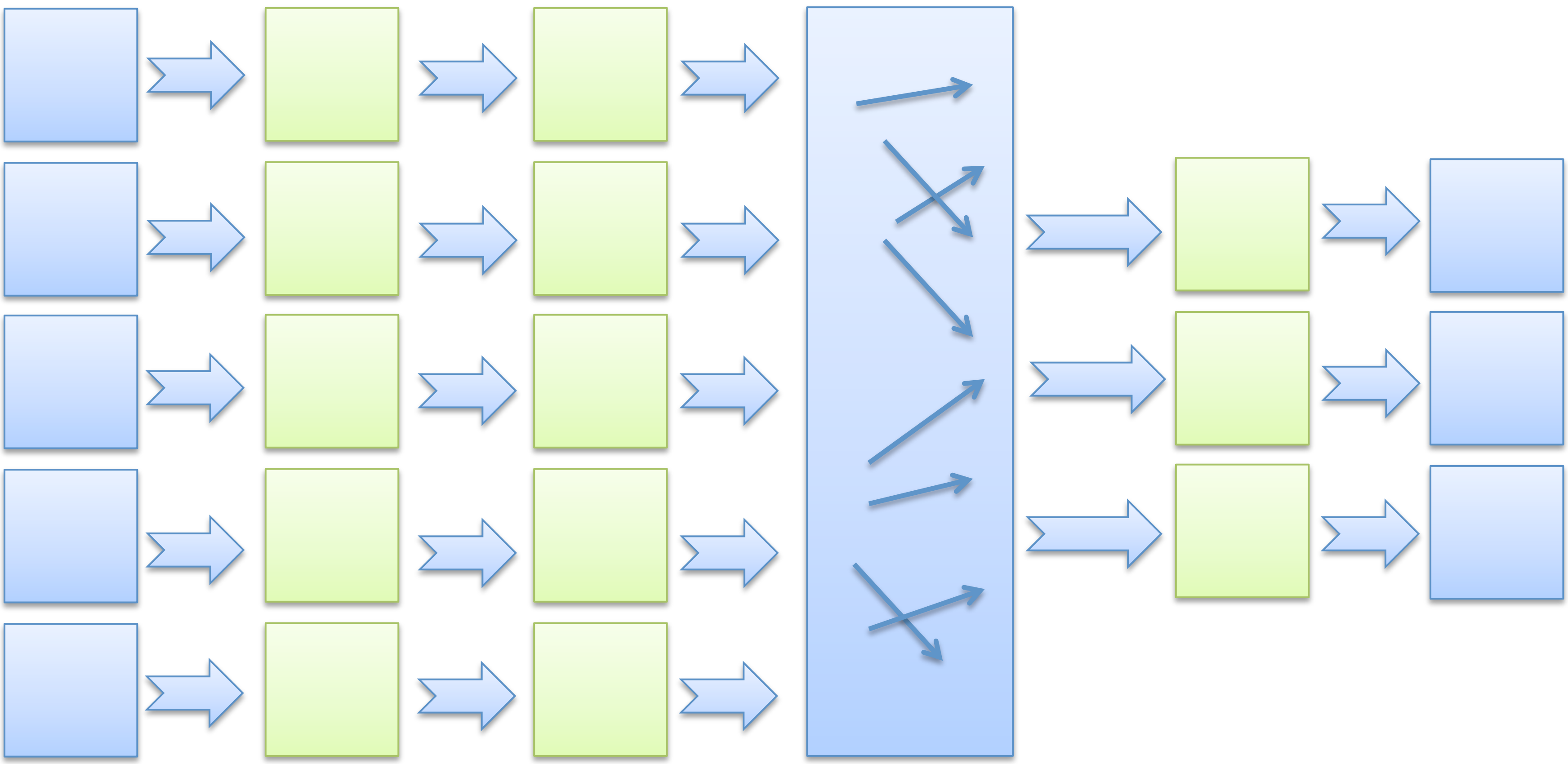
The Core of Hadoop - MapReduce

- Created by Google for web search indexes
- Ability to take a query over a dataset, **divide** it and run it in **parallel** over multiple nodes
- Consists in three stages:
  - Map
  - Shuffle
  - Reduce

The Core of Hadoop - MapReduce



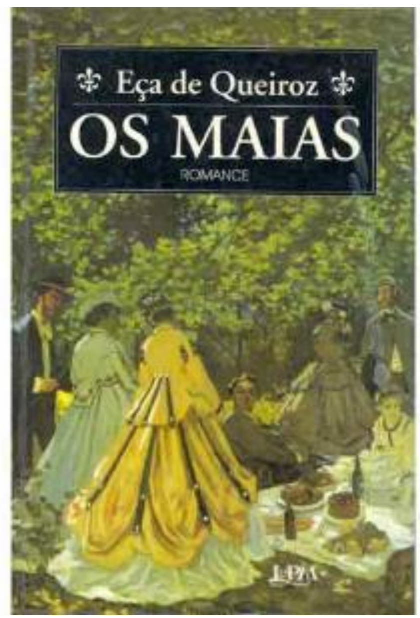
The Core of Hadoop - MapReduce

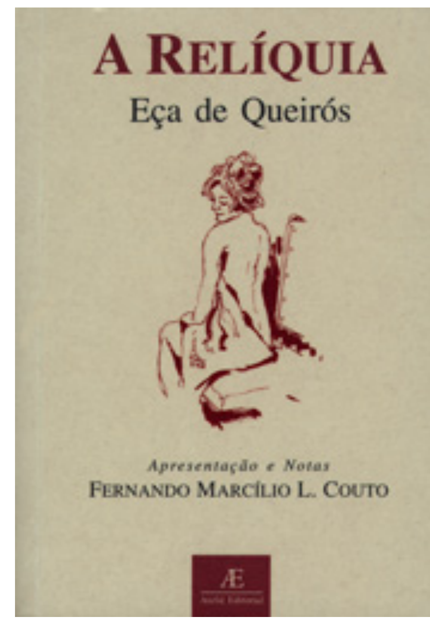


Input Map Reduce Shuffle Reduce Output

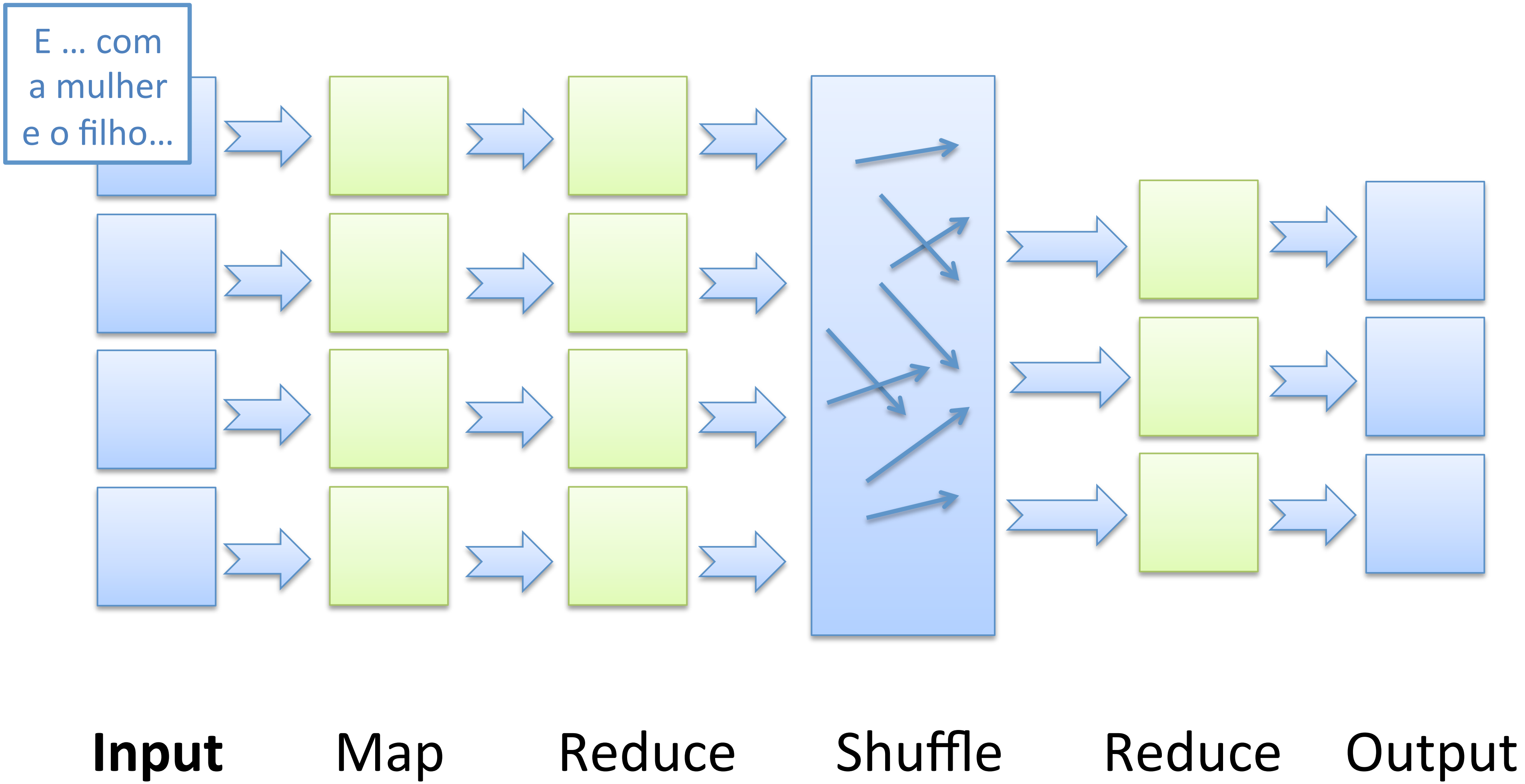
The Core of Hadoop - MapReduce

- Imagine that you want to apply MapReduce to word counting in books

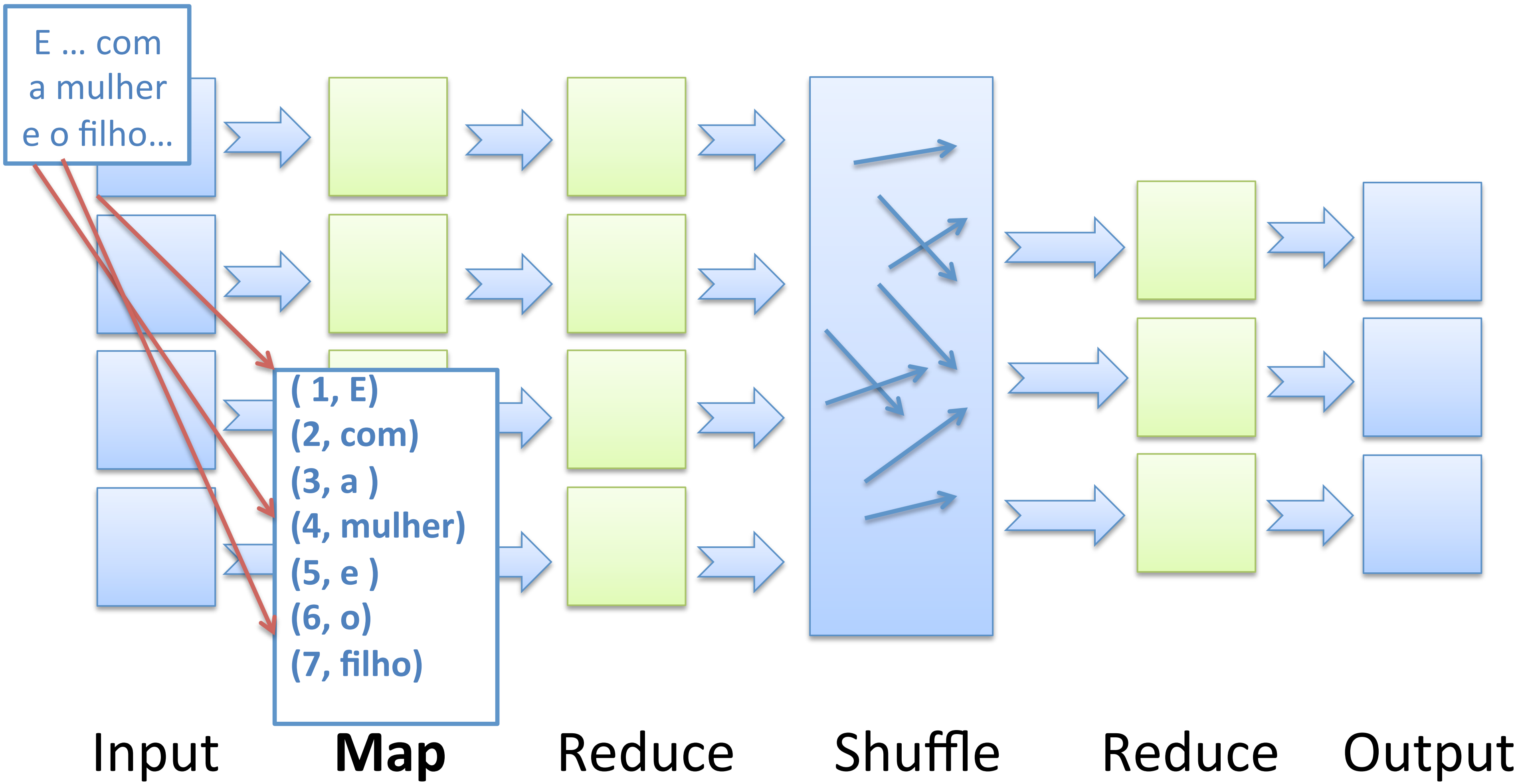




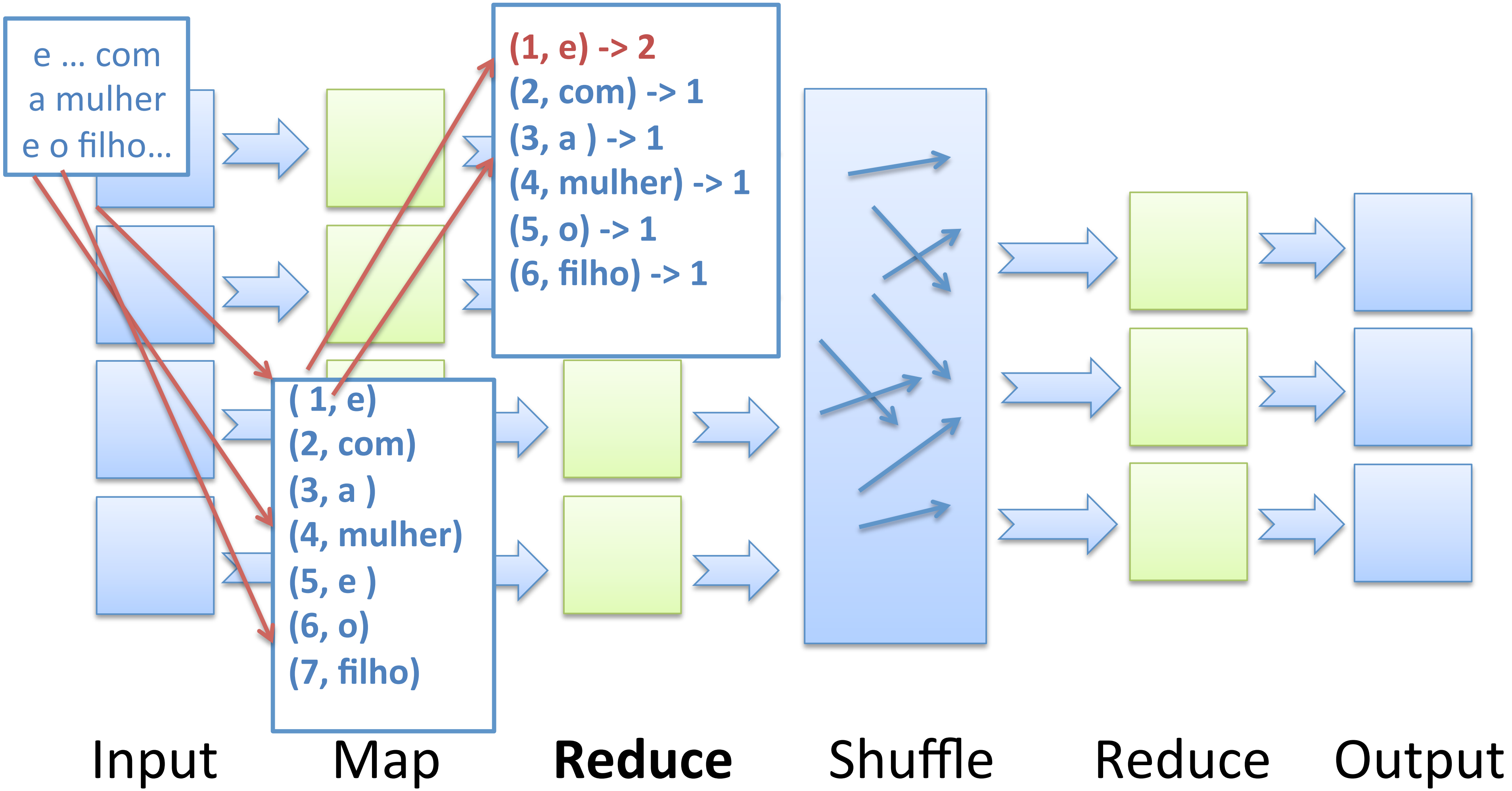
The Core of Hadoop - MapReduce



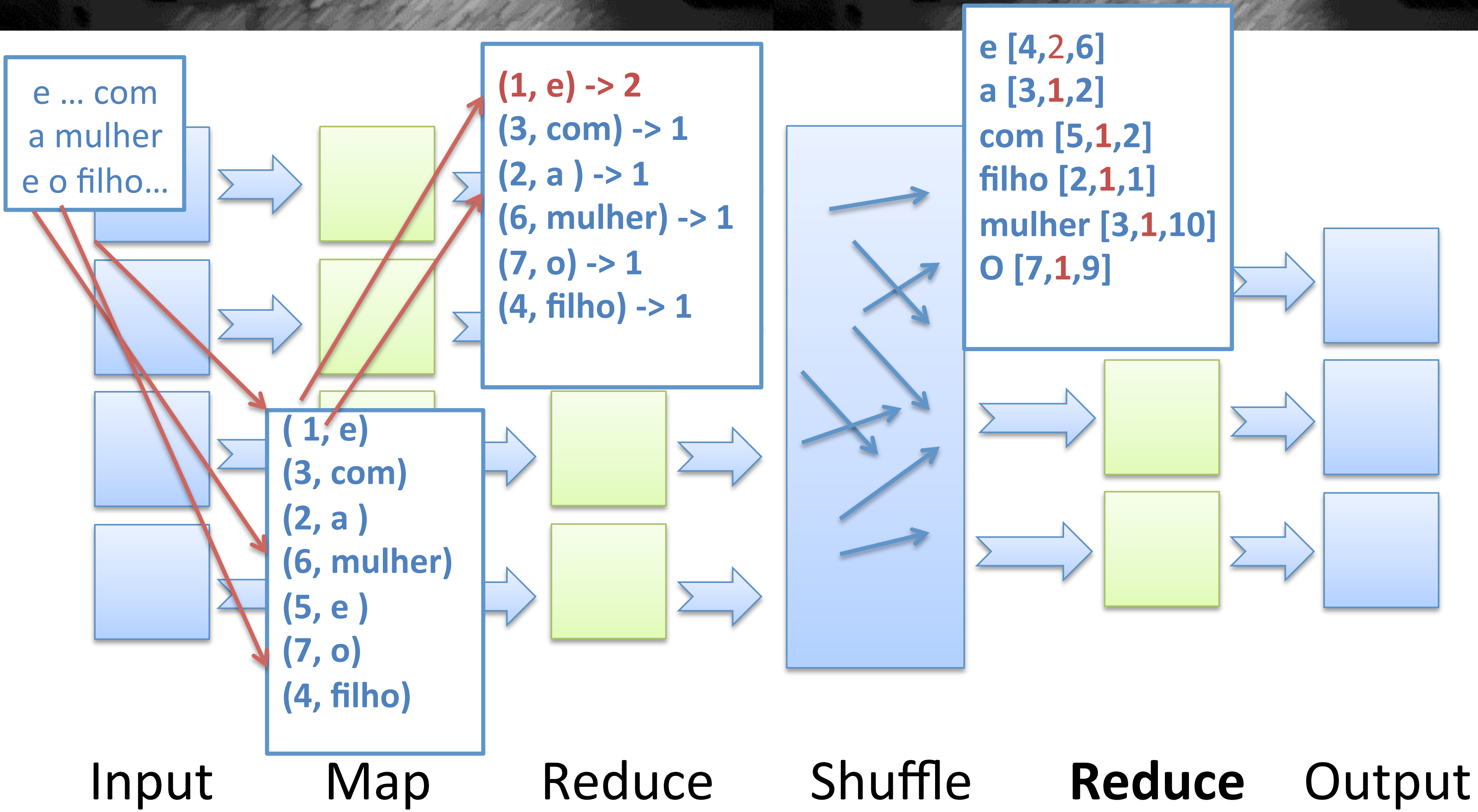
The Core of Hadoop - MapReduce



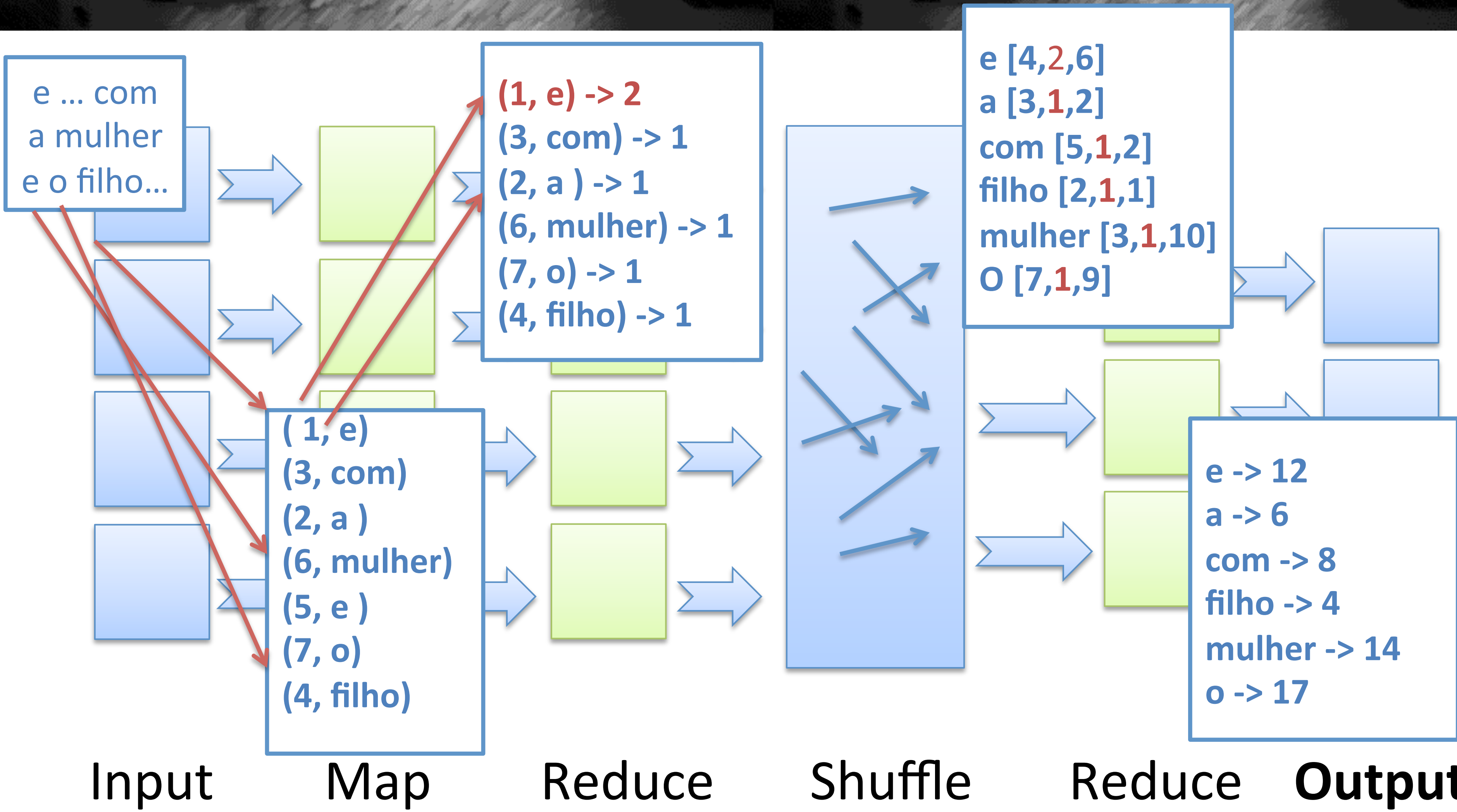
The Core of Hadoop - MapReduce



The Core of Hadoop - MapReduce



The Core of Hadoop - MapReduce



Who Uses Hadoop?

Walmart (Real Time)

The **Social Genome** product:

- Reach customers, or friends of customers, who have mentioned product and include a discount
- Combines public data from the web, social data and proprietary data

Walmart (Real Time)

The **Social Genome** product:

- Helps Walmart to understand the context of what their customers are saying online

- When a person tweets *I love Salt*, Walmart can understand that she is talking about the movie *Salt* and not the condiment.

Apache Hive

- Data warehouse software that facilitates querying and managing large datasets residing in a distributed storage
- Uses HiveSQL to analyze data
- Allows programmers to use their own map reduce functions

Apache Cassandra

- Database with high scalability and high availability without compromising performance.

- Offers
  - Denormalization
  - Materialized views
  - Powerful built in cache

Google Big Query

Built on GoogleFS

Table Info

Table ID	382129041633:sample.july2nd
Table Size	12.6 GB
Number of Rows	76,704,834
Creation Time	6:42pm, 2 Jul 2012
Last Modified	6:42pm, 2 Jul 2012

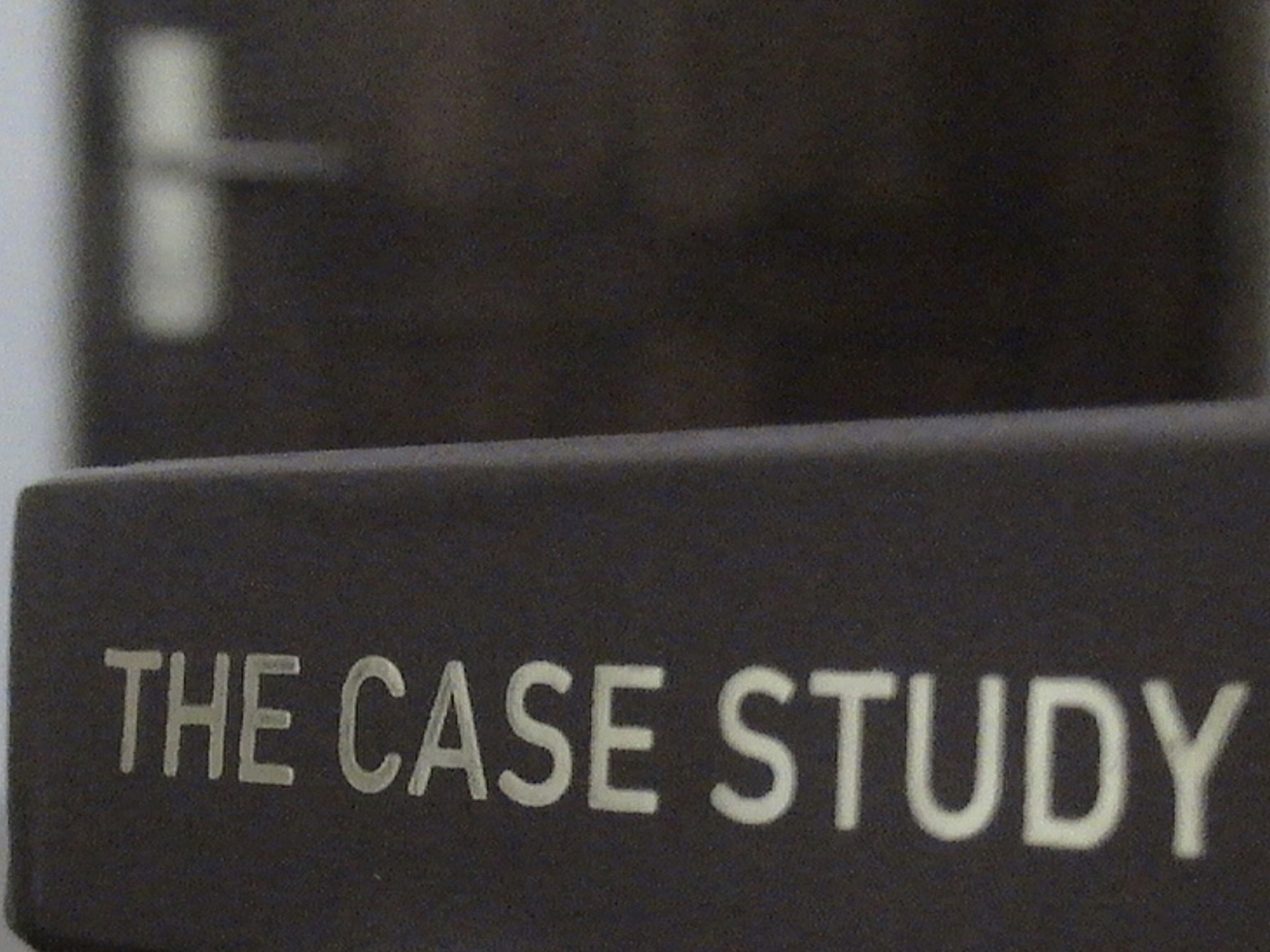
Google bigquery

Preview

Row	ROWTIME	RE	reID	rePosition	reLatitude
1	2012-07-01 03:02:03	1425297920\0765672417C	11265065	36468.89234957474	10.12794998163305
2	2012-07-01 03:02:03	1425297920\0765672417C	11265064	36478.90602057377	10.12786064593382

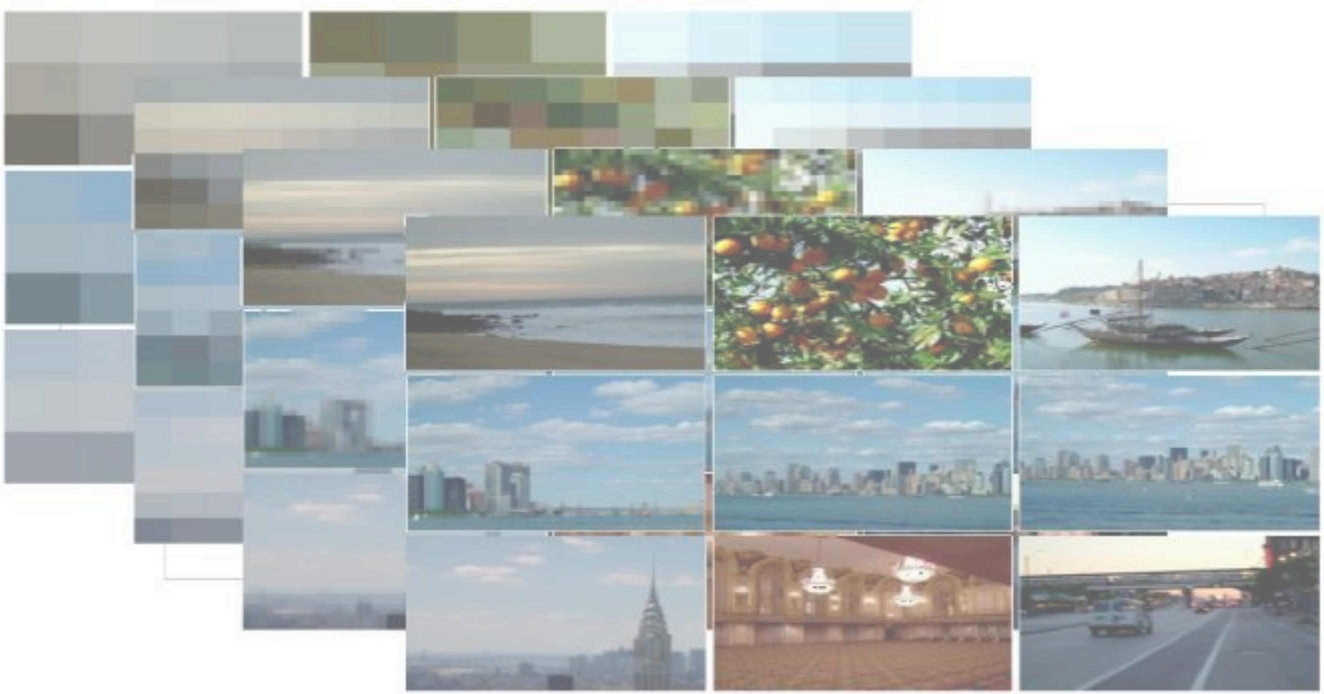
Summary

- Apache Hadoop techhnology
  - HDFS file system
  - MapReduce paradigm
- Apache Hive
- Apache Cassandra
- Google BigQuery


THE CASE STUDY

Project HEIDI

How to index **high dimensional** data?



Project HEIDI

- Multimedia datasets are growing!
  - Ex: Audio, video, medical images, photos, etc.

- We need techniques to efficiently **manage** and **access** information in such large datasets!

- Multimedia datasets **cannot** be searched like traditional databases (e.g. Text search). They are searched in **metric spaces**.

Project HEIDI

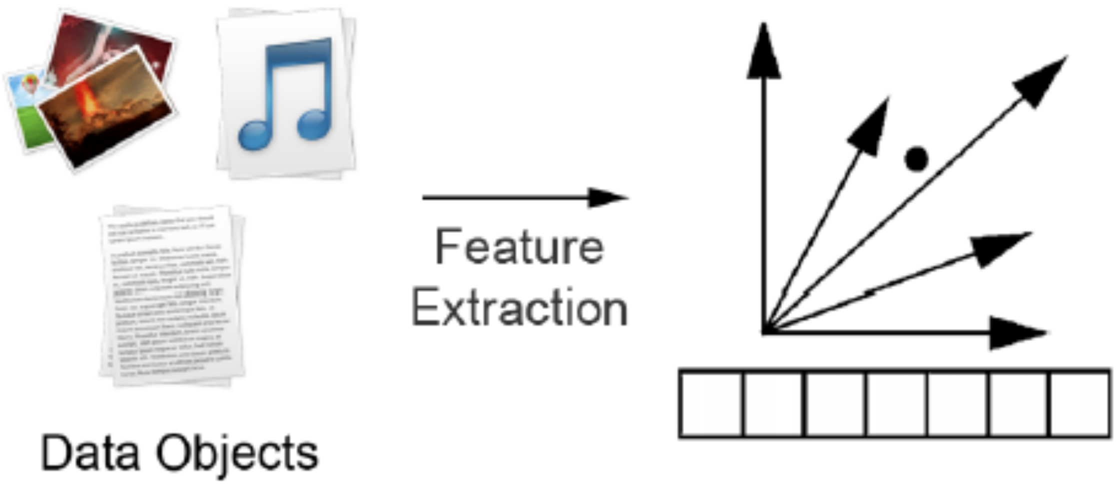
Given a **multimedia object** as input, return the most **similar** objects from a multimedia database

Project HEIDI

- In content-based image retrieval, images are described by their properties:
    - Colors
    - Texture
    - Shape
    - Shadows
    - etc

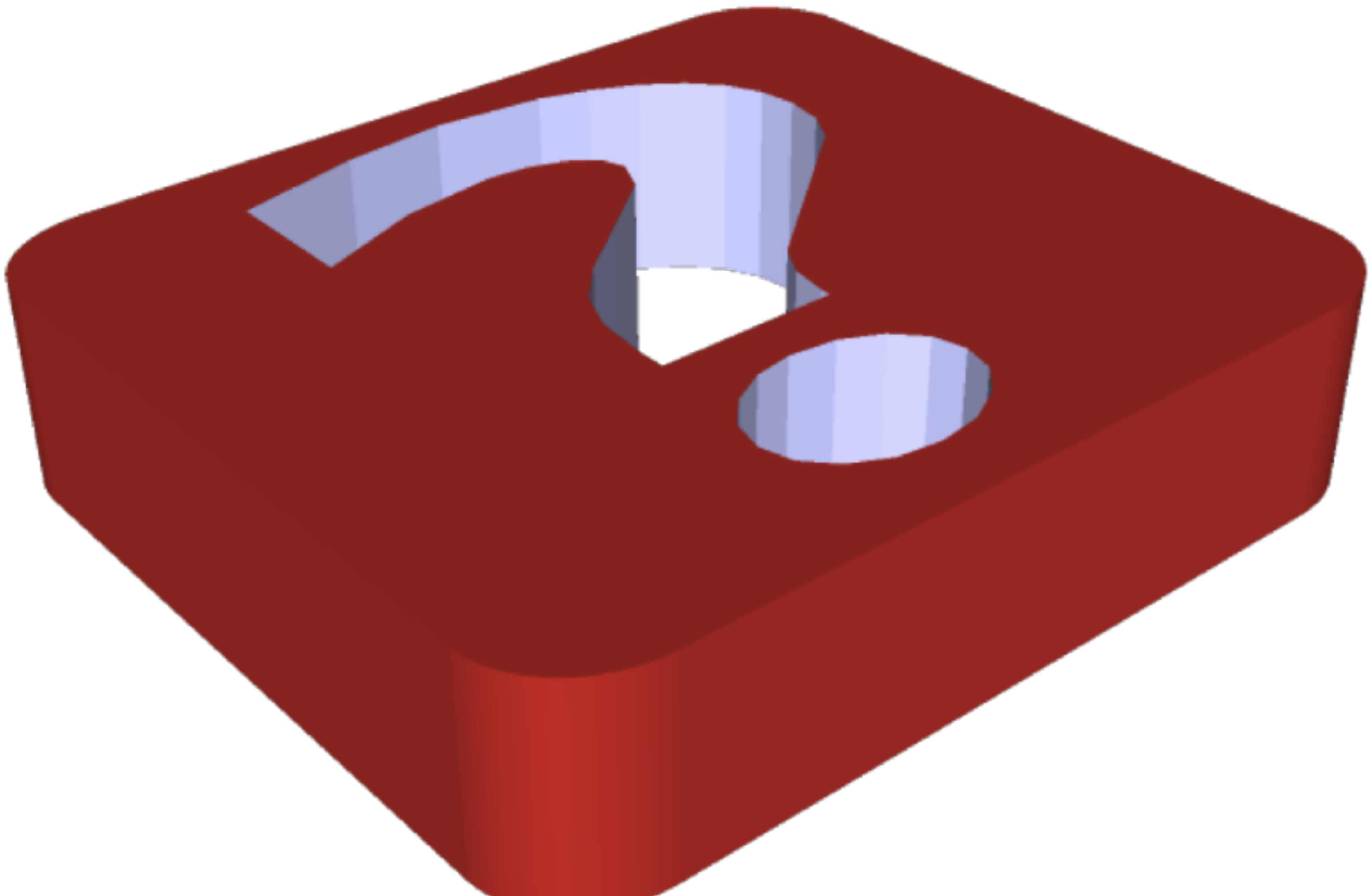
Project HEIDI

- The process of converting a multimedia object into a vector with its contents/features is called ***feature extraction***.



What's your

PROBLEM



Project HEIDI

The problem in high dimensional indexing is the

CURSE

Project HEIDI

The problem in high dimensional indexing is the

CURSE

OF DIMENSIONALITY!!!

The Curse of Dimensionality

- A phenomena that arises when analyzing and organizing data in high dimensional spaces.

- It does not occur in lower dimensions!

The Curse of Dimensionality

- When dimensionality increases, the **volume** of the space **increases** so fast that the data become **sparse**.

- Traditional tree indexing structures do not provide any advantages in the index process (outperformed by linear scan)…

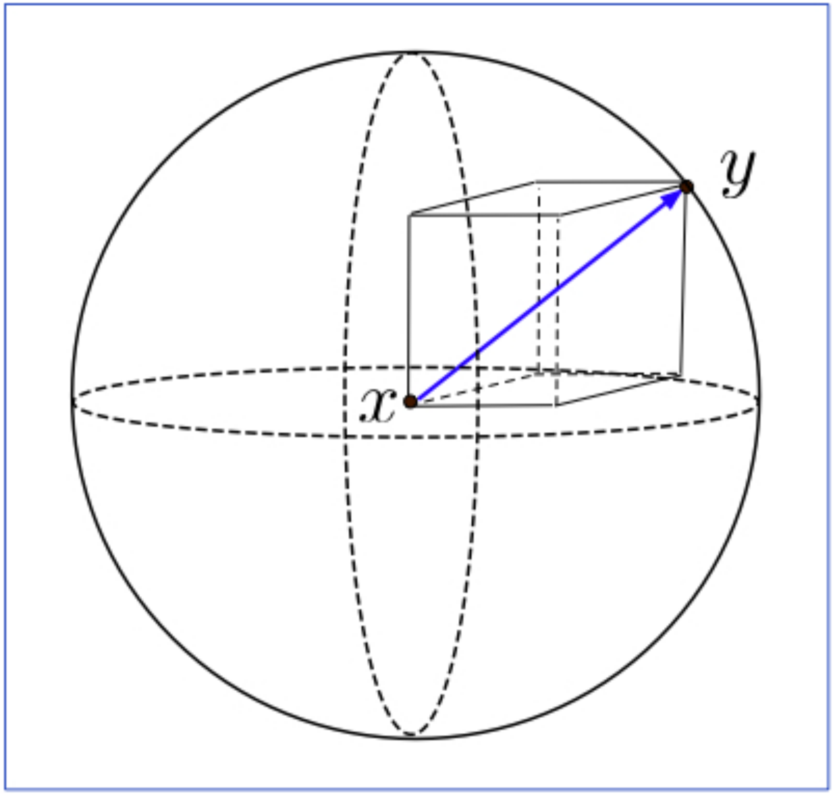
The Curse of Dimensionality

- Consequences:
  - (hyper) cubes
  - (hyper) spheres

- **Volume** increases **exponentially** with **growing dimension** (while the edge length remains constant).

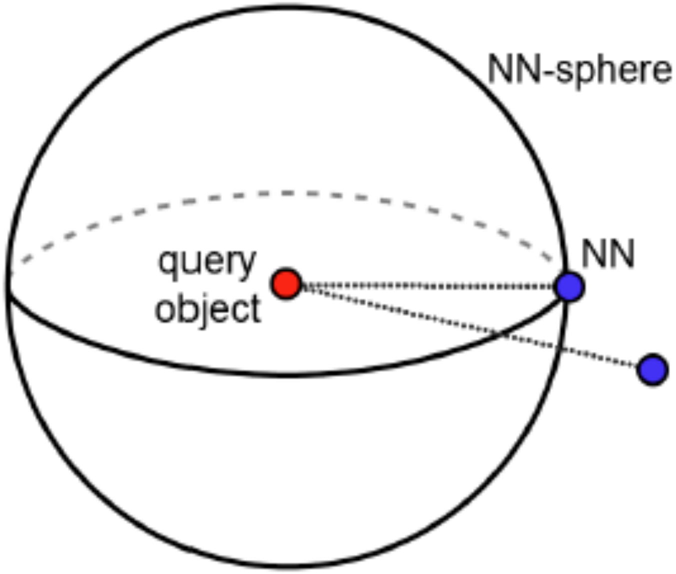
The Curse of Dimensionality

How can this curse affect datasets in metric spaces?



The Curse of Dimensionality

- In metric spaces, there is the Nearest Neighbor paradigm.



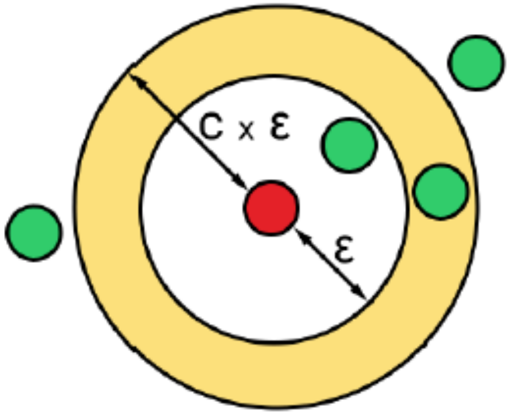
- The NN-sphere contains one point.

- For higher dimensions, the radius of the NN-sphere will be larger than the size of the database.

The Curse of Dimensionality

Literature addresses this issue in two ways:

1. Approximate Nearest Search



2. Reduce the dimensionality of the multimedia objects

Project HEIDI

The algorithm that is proposed in this project does not suffer from the **curse of dimensionality!**

Project HEIDI

The Linear SubSpace Indexing Algorithm

- Based on the idea of subspaces!
  - Feature extraction function that maps the **high dimensiona**l objects into **low dimensional** equivalent objects.

Project HEIDI

The Quick and Dirty Paradigm

- Discard data objects in lower dimensional spaces
- Massive comparison operations are less expensive in lower dimensions

Project HEIDI

However, this approach has

CONSEQUENCES

Project HEIDI

The Quick and Dirty Paradigm

- Can produce lots of **false hits!**

- The mapping to lower dimensions can lead to closer data points

Project HEIDI

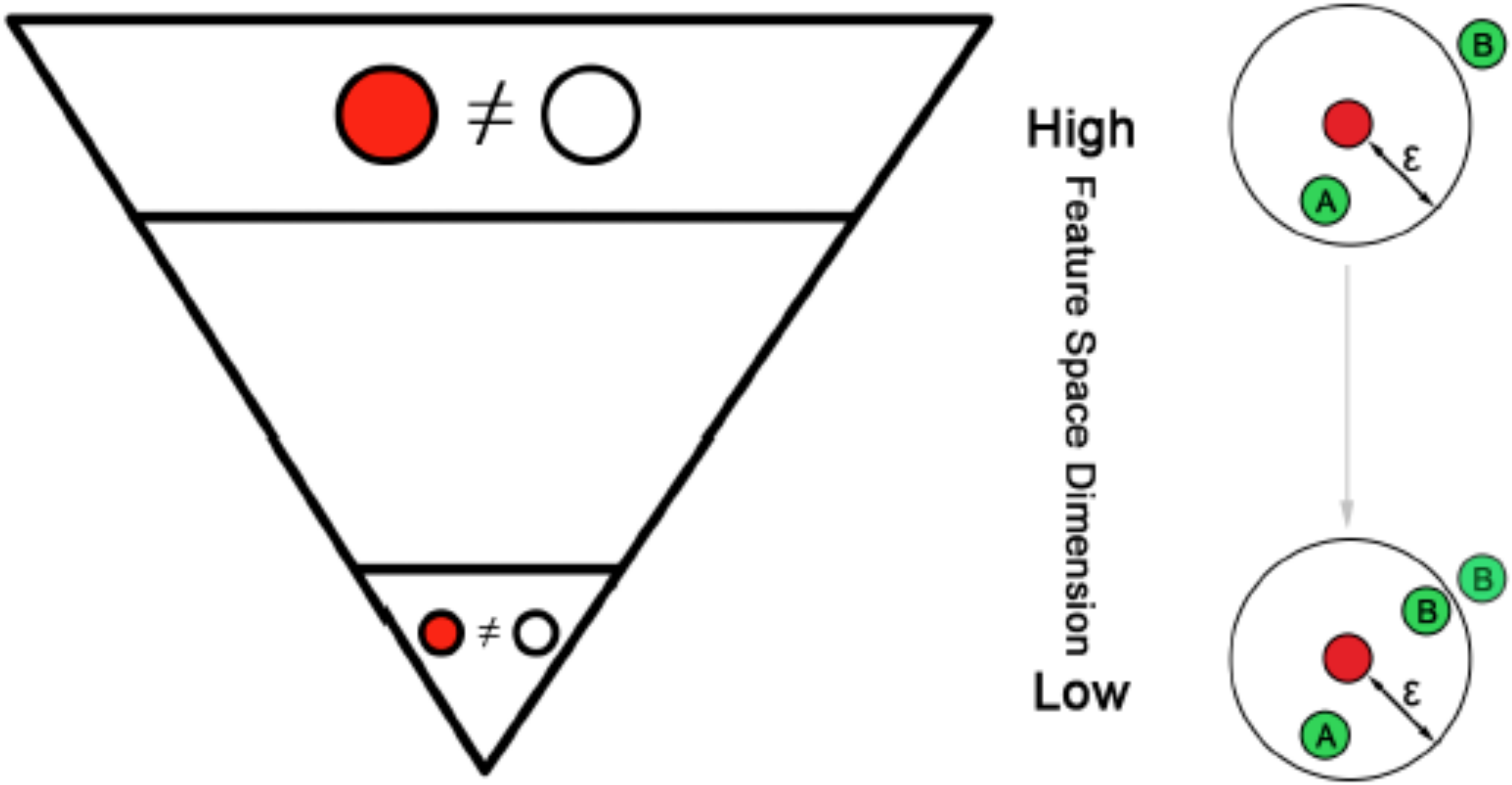
Are false hits really a

Project HEIDI

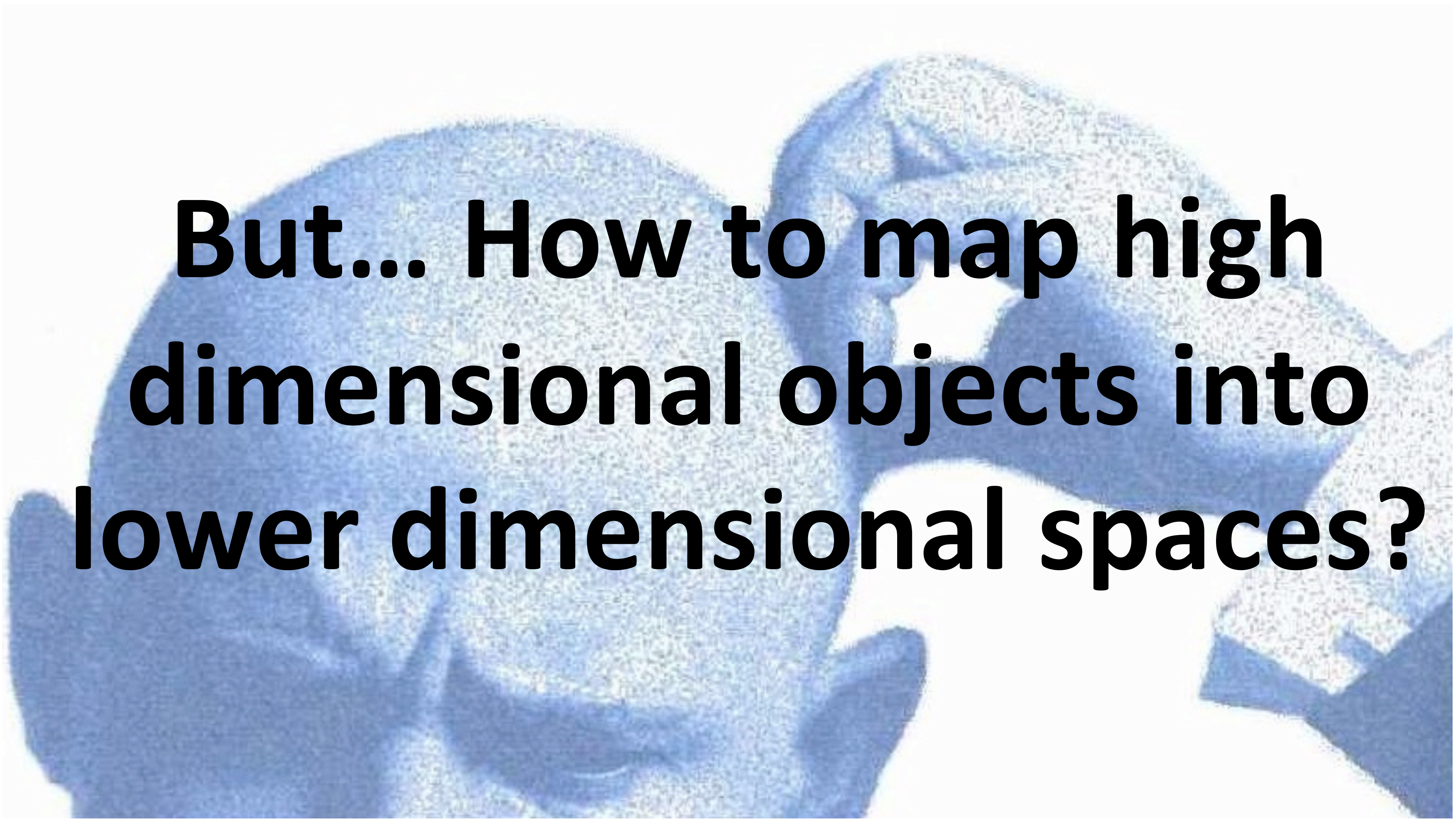
Project HEIDI

- Just perform a final comparison in the **original space**!
- All false hits will be **discarded**!
- The computation is **quick**, because the massive amount of objects were **discarded** in the **lower dimensions.**

Project HEIDI



Project HEIDI



But… How to map high dimensional objects into lower dimensional spaces?

Project HEIDI

The Quick and Dirty Paradigm

- **Lower bonding lemma**:
  - Distances between objects in the feature space are always smaller or equal that in the original space

$$d_{feature}(F(A), F(B)) \leq d(A, B)$$

Project HEIDI

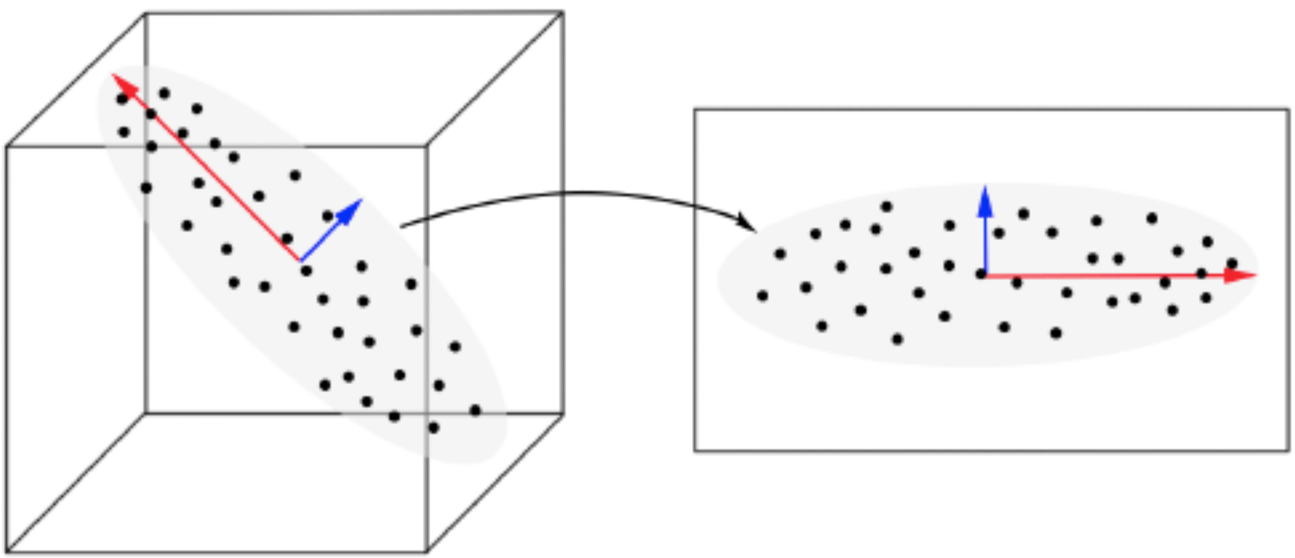
The Quick and Dirty Paradigm

- The Principal Component Analysis algorithm maps objects into lower dimensions

- Satisfies the **lower bounding lemma**

Project HEIDI

The Principal Component Analysis



Project HEIDI

The Principal Component Analysis

1. Compute Co-Variance Matrix

$$C_{ij} = \frac{\sum_{k=1}^{n} (x_i^k - m_i)(x_j^k - m_j)}{n-1}$$

Project HEIDI

The Principal Component Analysis

2. Compute eigen vectors and eigen values from co-variance matrix

3. Get the most significant eigen vectors and create a projection matrix

4. Project data

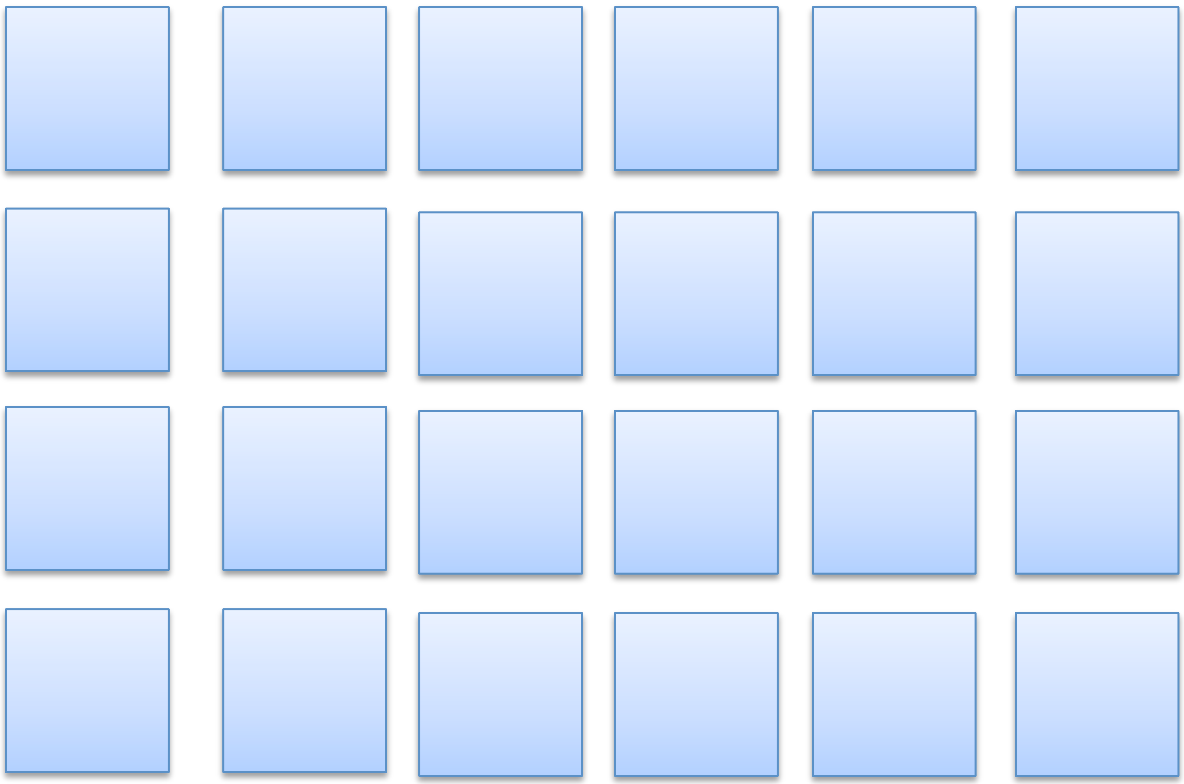
Project HEIDI

The Hierarchical Linear SubSpace Indexing Method

1. Index Phase (off-line)
   - Partition large dataset into chunks of data
   - Iteratively apply the Principal Component Analysis to project the dataset into lower dimensions
   - This action can be parallelized in many servers

Project HEIDI

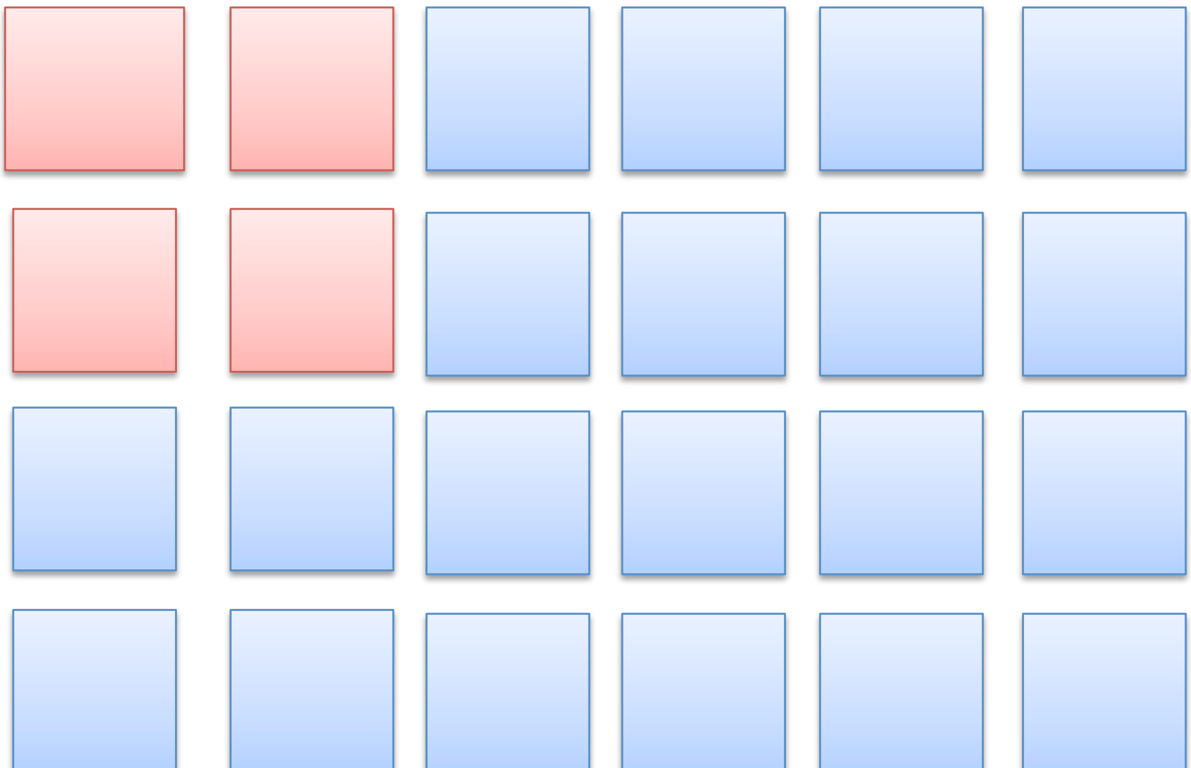
The Hierarchical Linear SubSpace Indexing Method



Project HEIDI

The Hierarchical Linear SubSpace Indexing Method

Original Space **Lower Space**

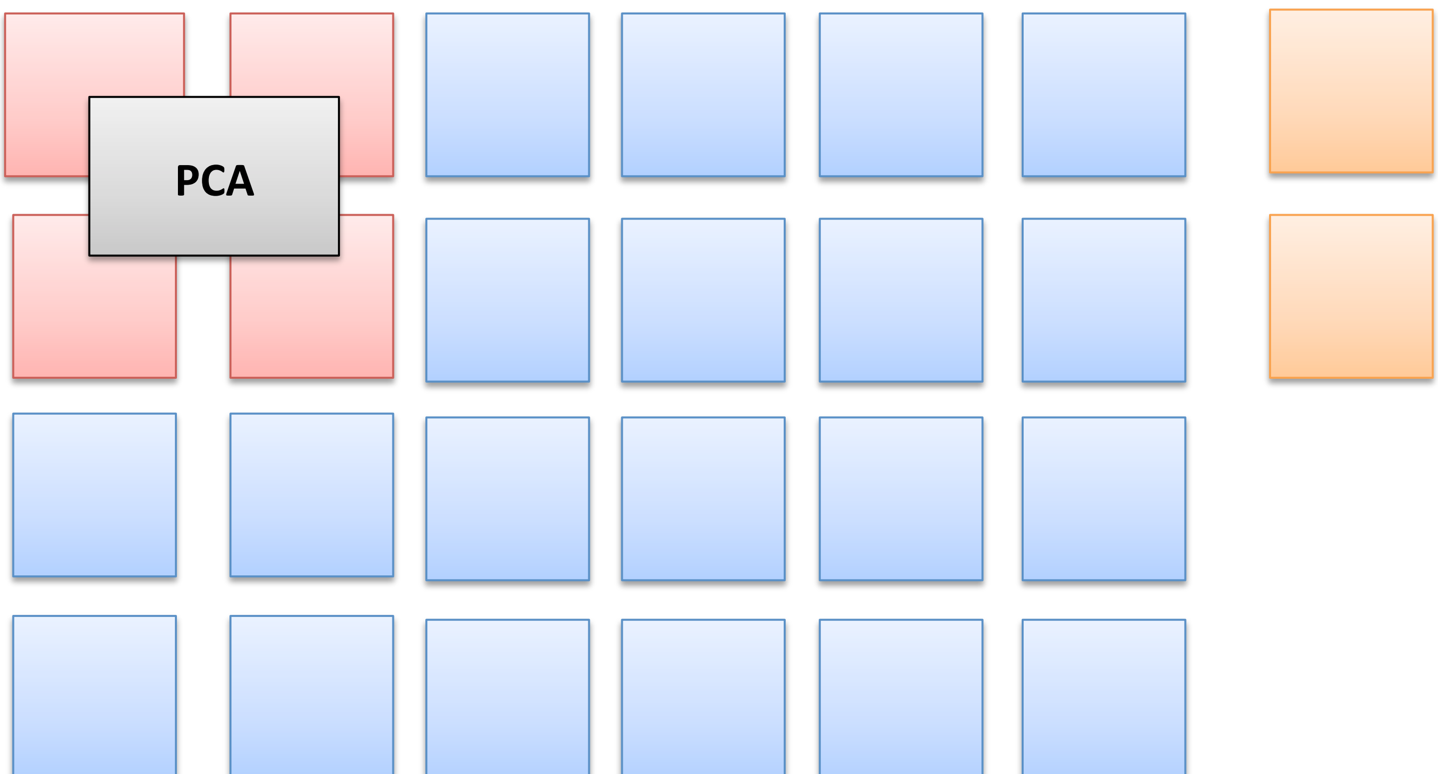


Project HEIDI

The Hierarchical Linear SubSpace Indexing Method

Original Space

Lower Space

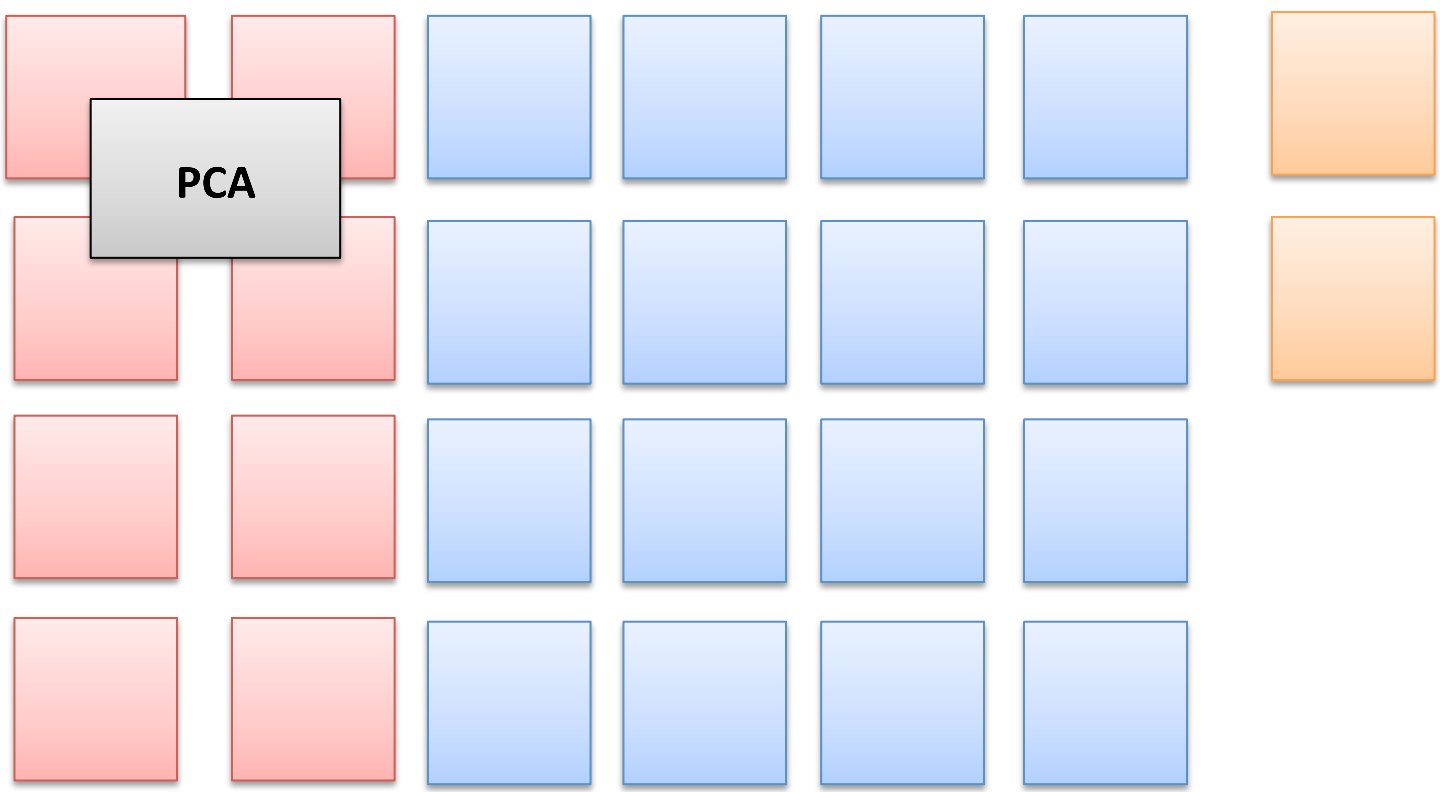


Project HEIDI

The Hierarchical Linear SubSpace Indexing Method

Original Space

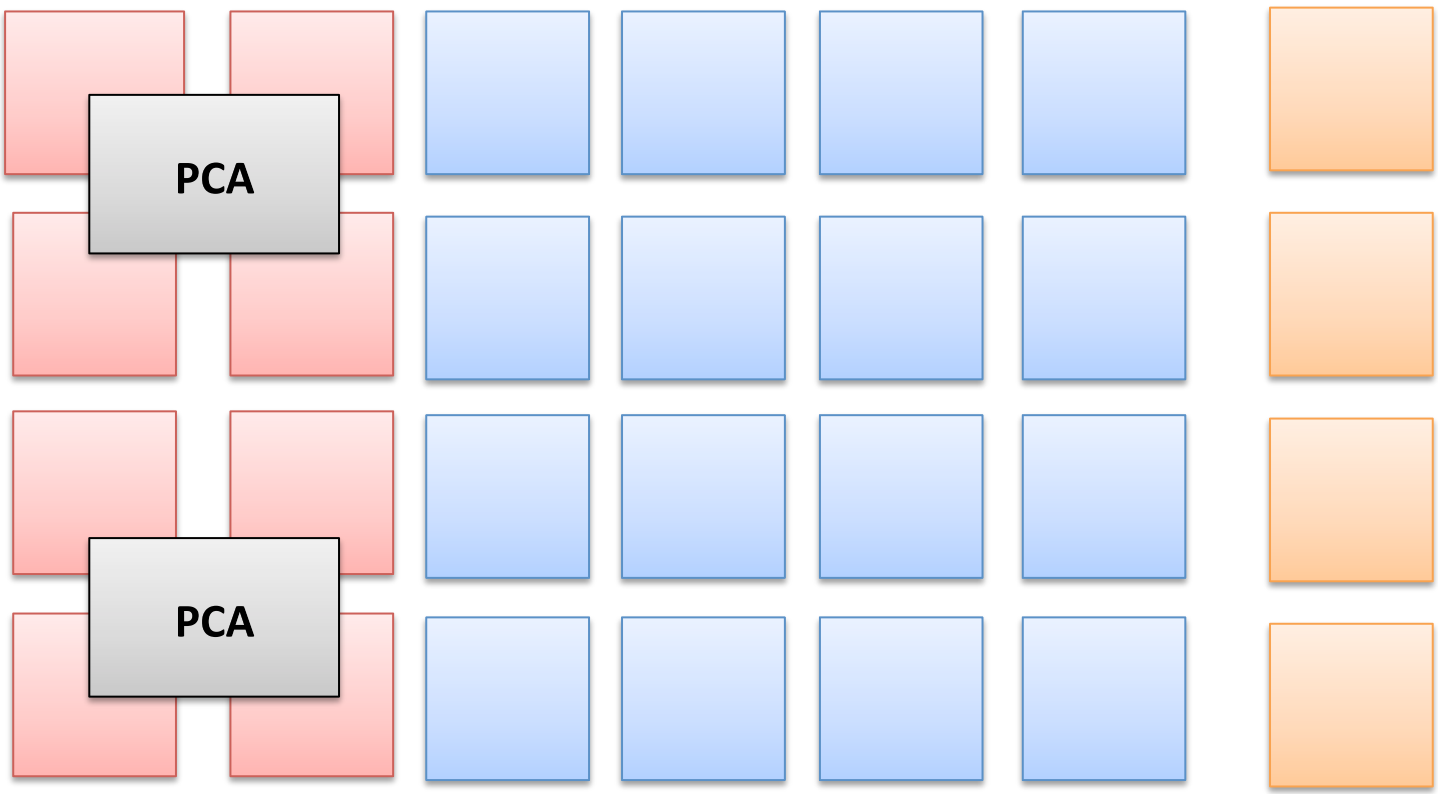
Lower Space



Project HEIDI

The Hierarchical Linear SubSpace Indexing Method

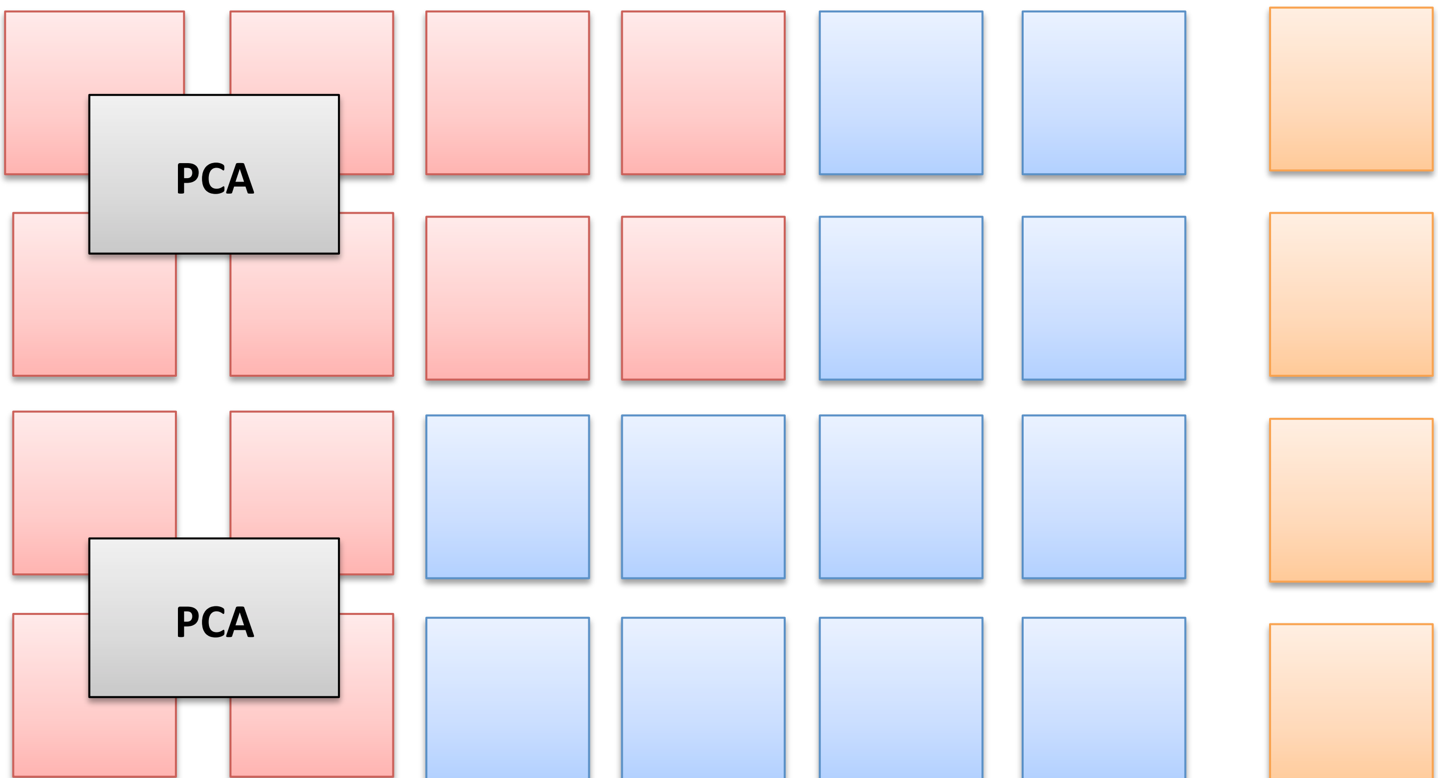
Original Space **Lower Space**



Project HEIDI

The Hierarchical Linear SubSpace Indexing Method

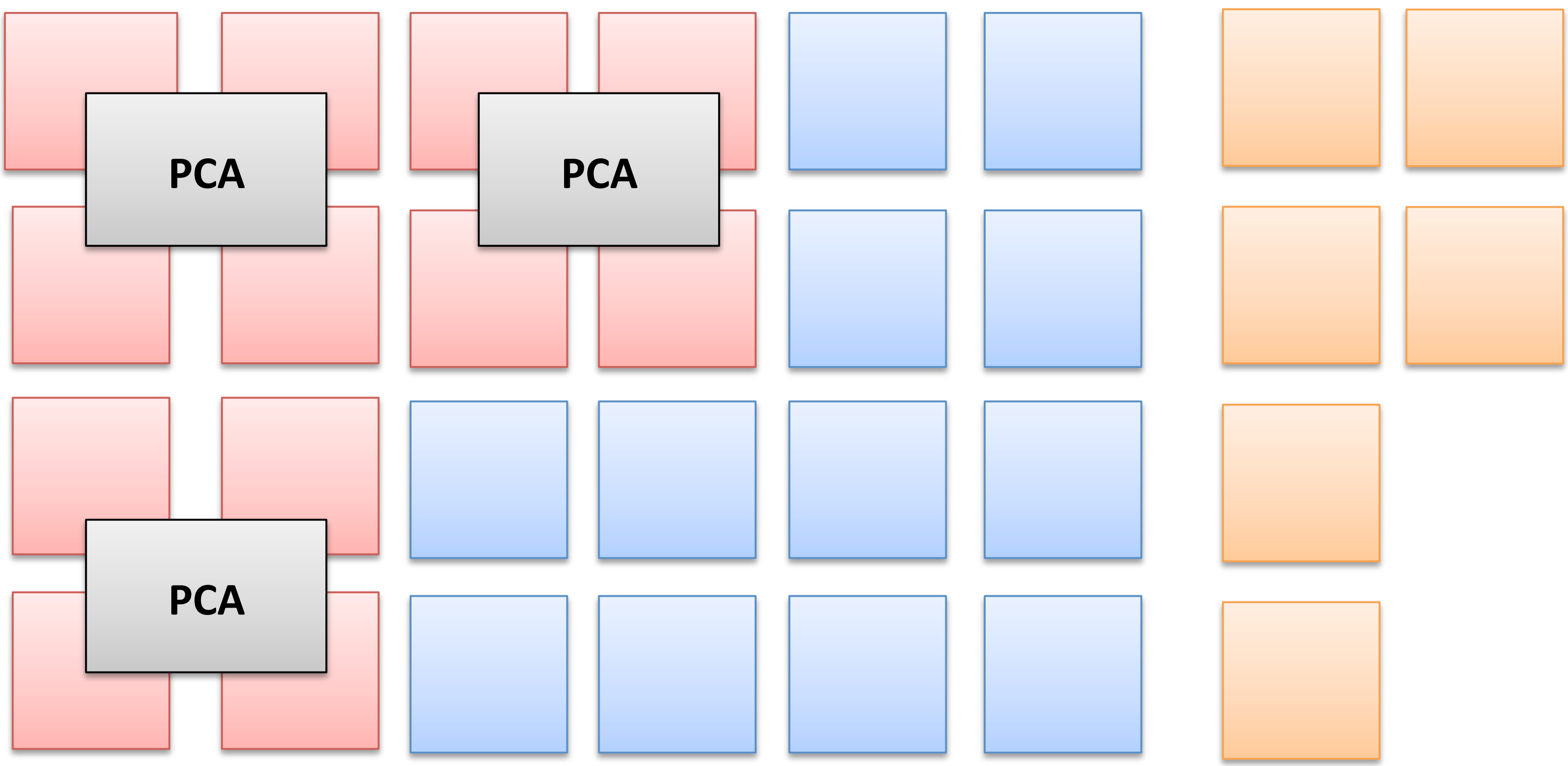
Original Space Lower Space



Project HEIDI

The Hierarchical Linear SubSpace Indexing Method

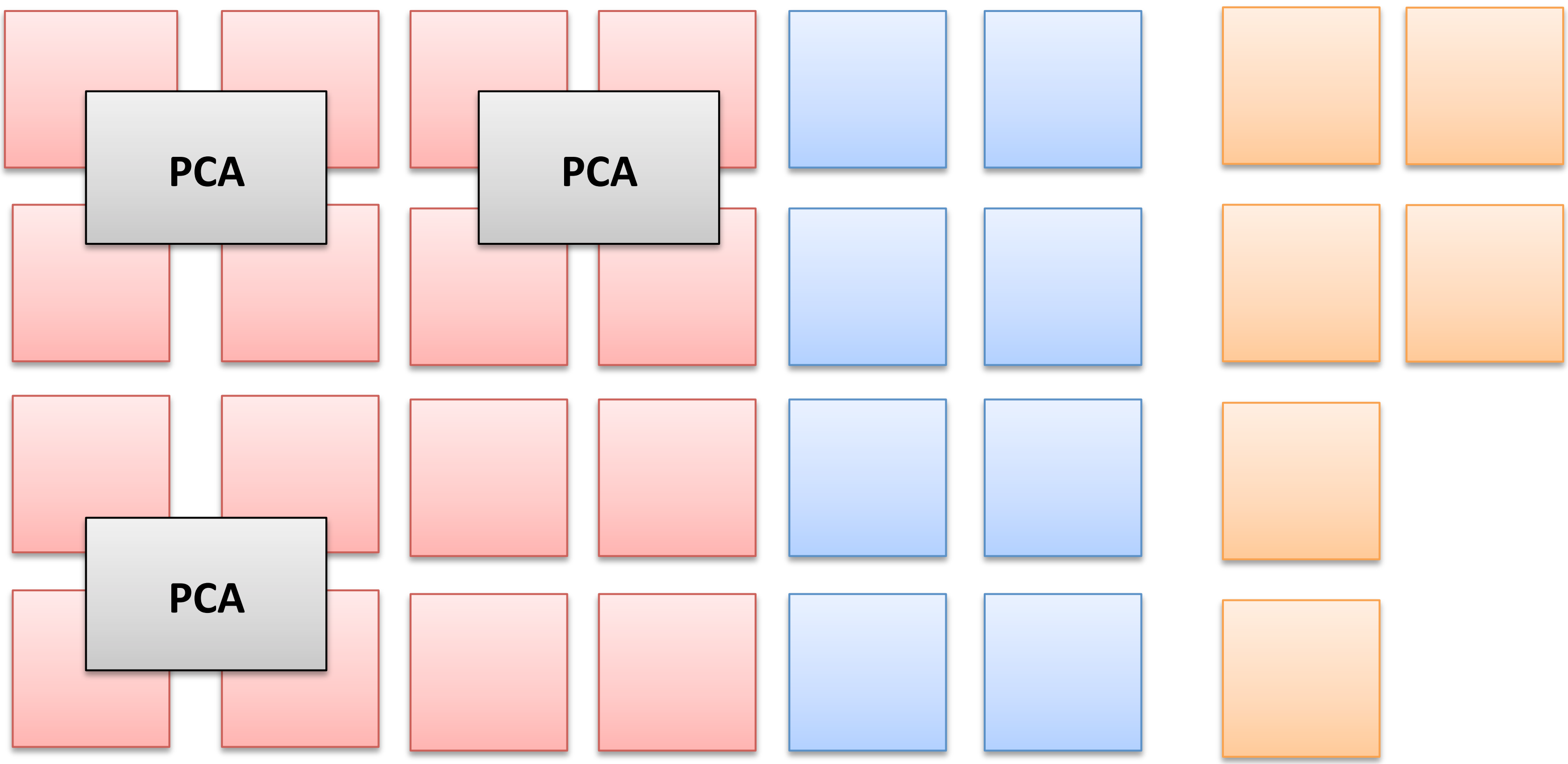
Original Space **Lower Space**



Project HEIDI

The Hierarchical Linear SubSpace Indexing Method

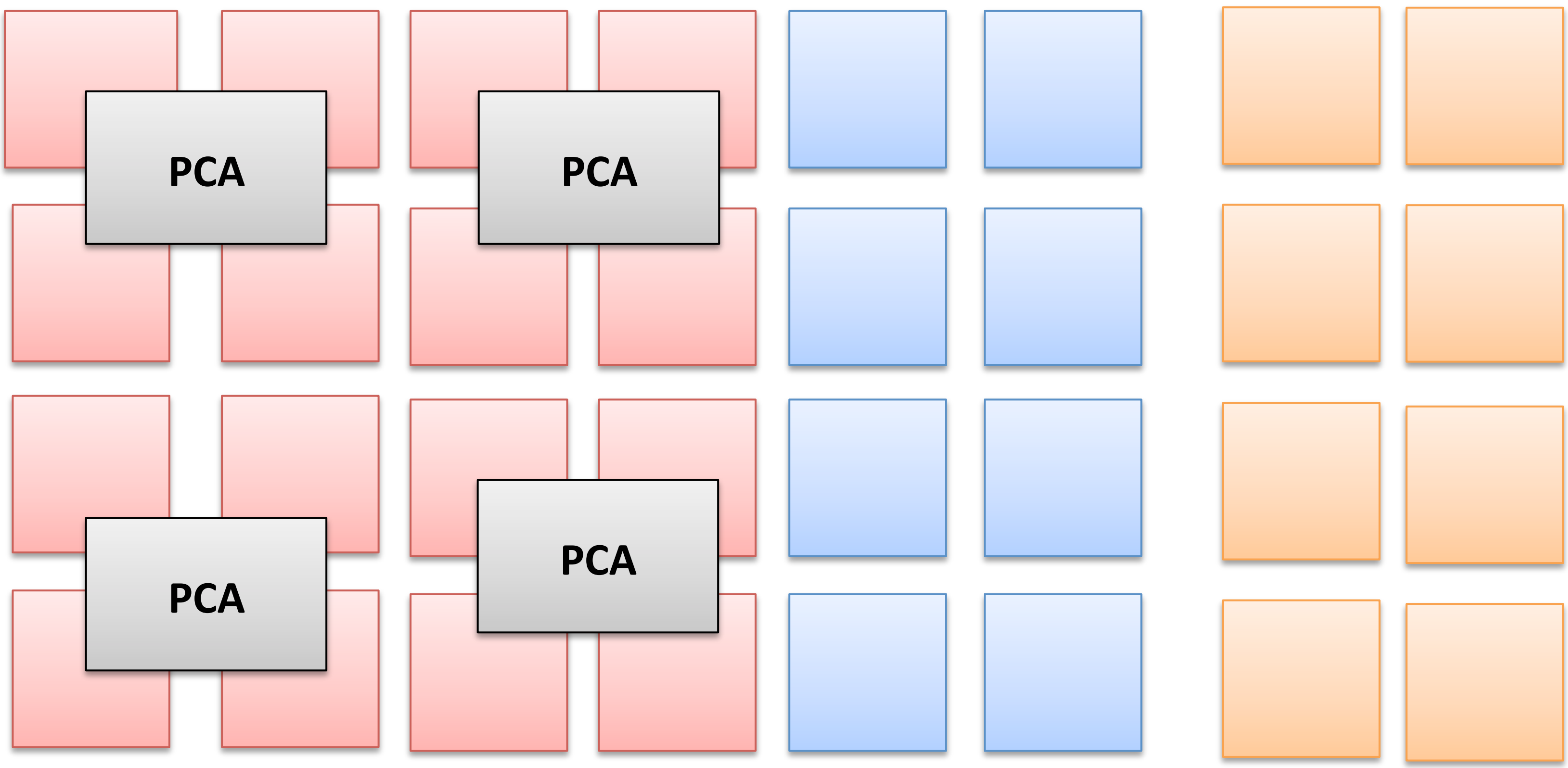
Original Space Lower Space



Project HEIDI

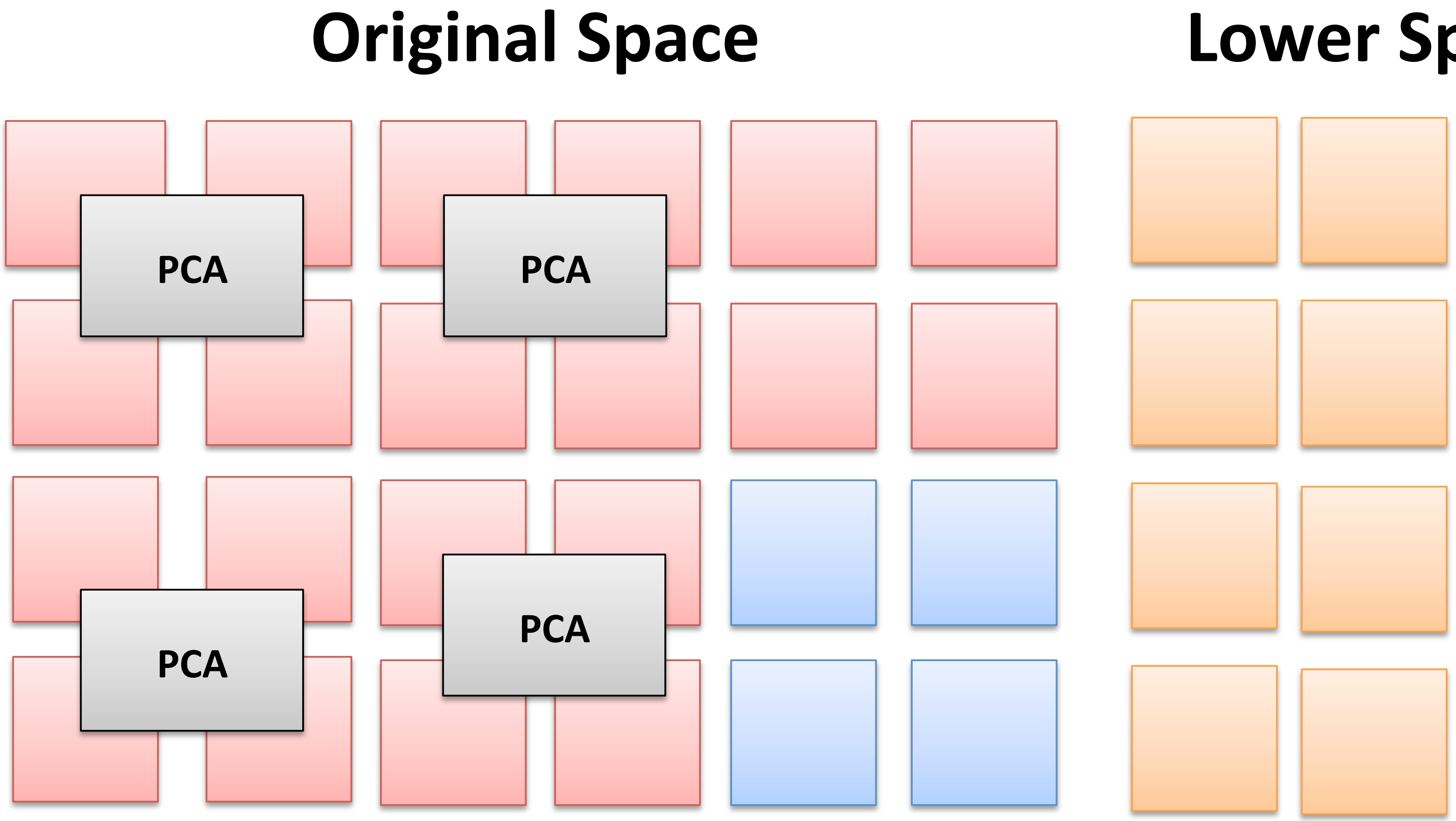
The Hierarchical Linear SubSpace Indexing Method

Original Space — Lower Space



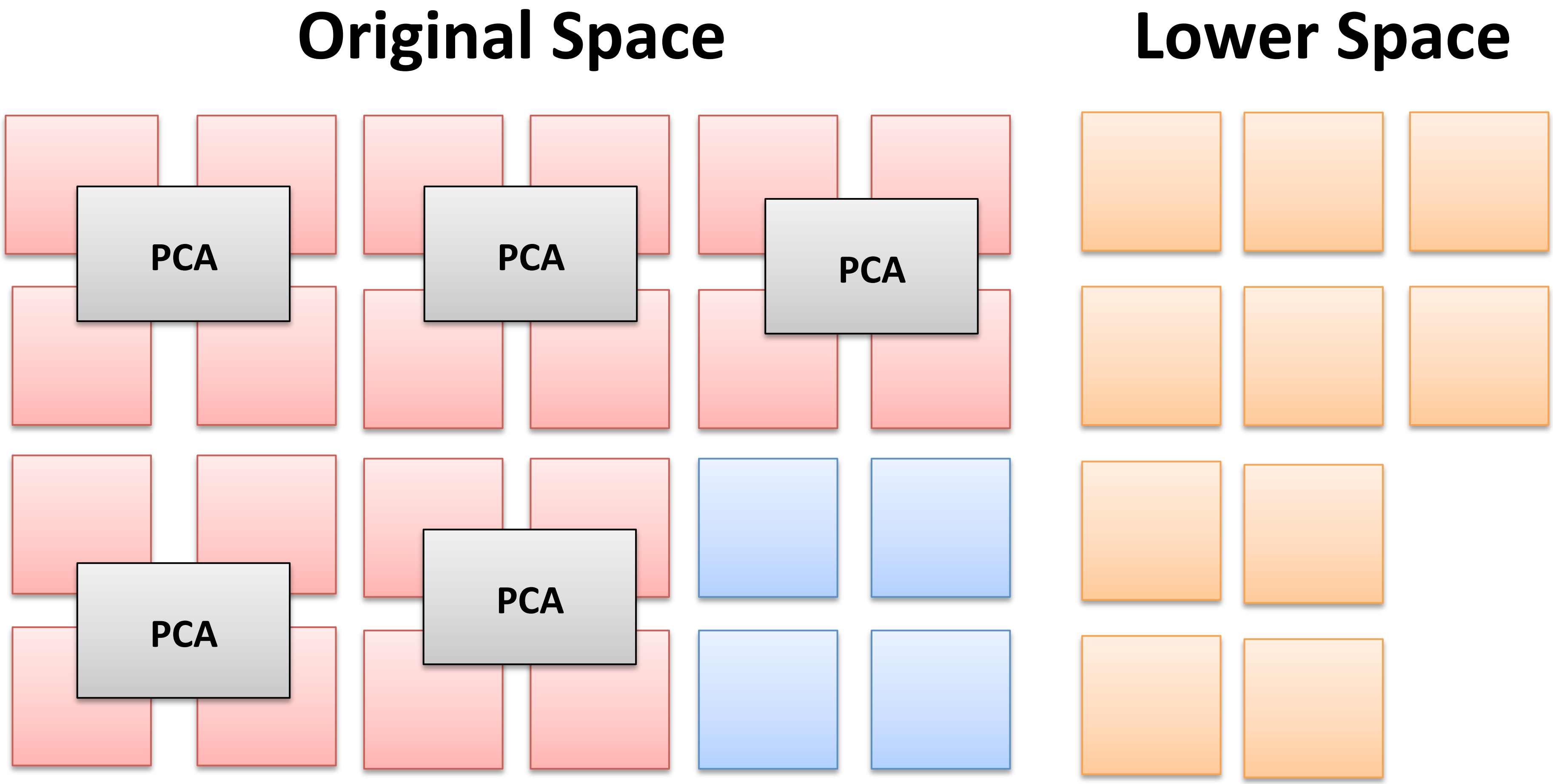
Project HEIDI

The Hierarchical Linear SubSpace Indexing Method



Project HEIDI

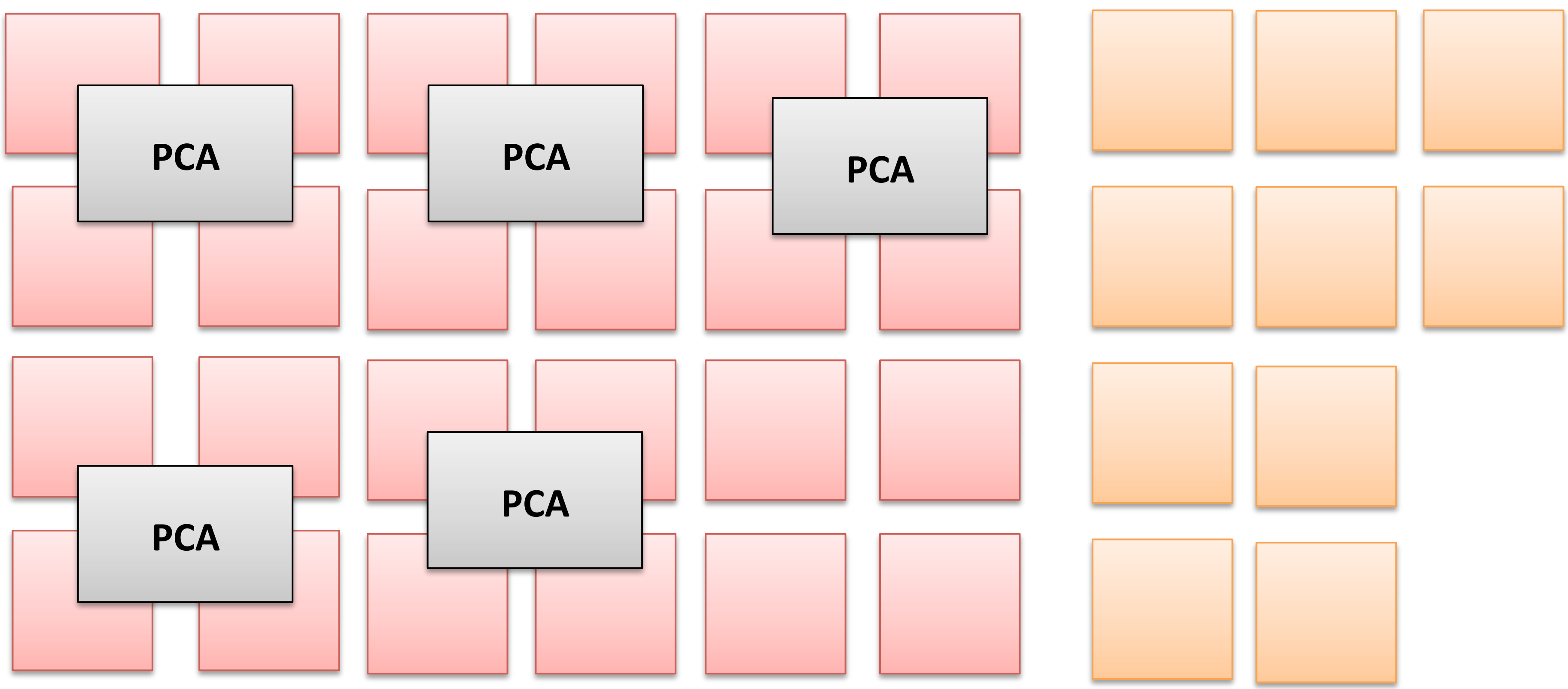
The Hierarchical Linear SubSpace Indexing Method



Project HEIDI

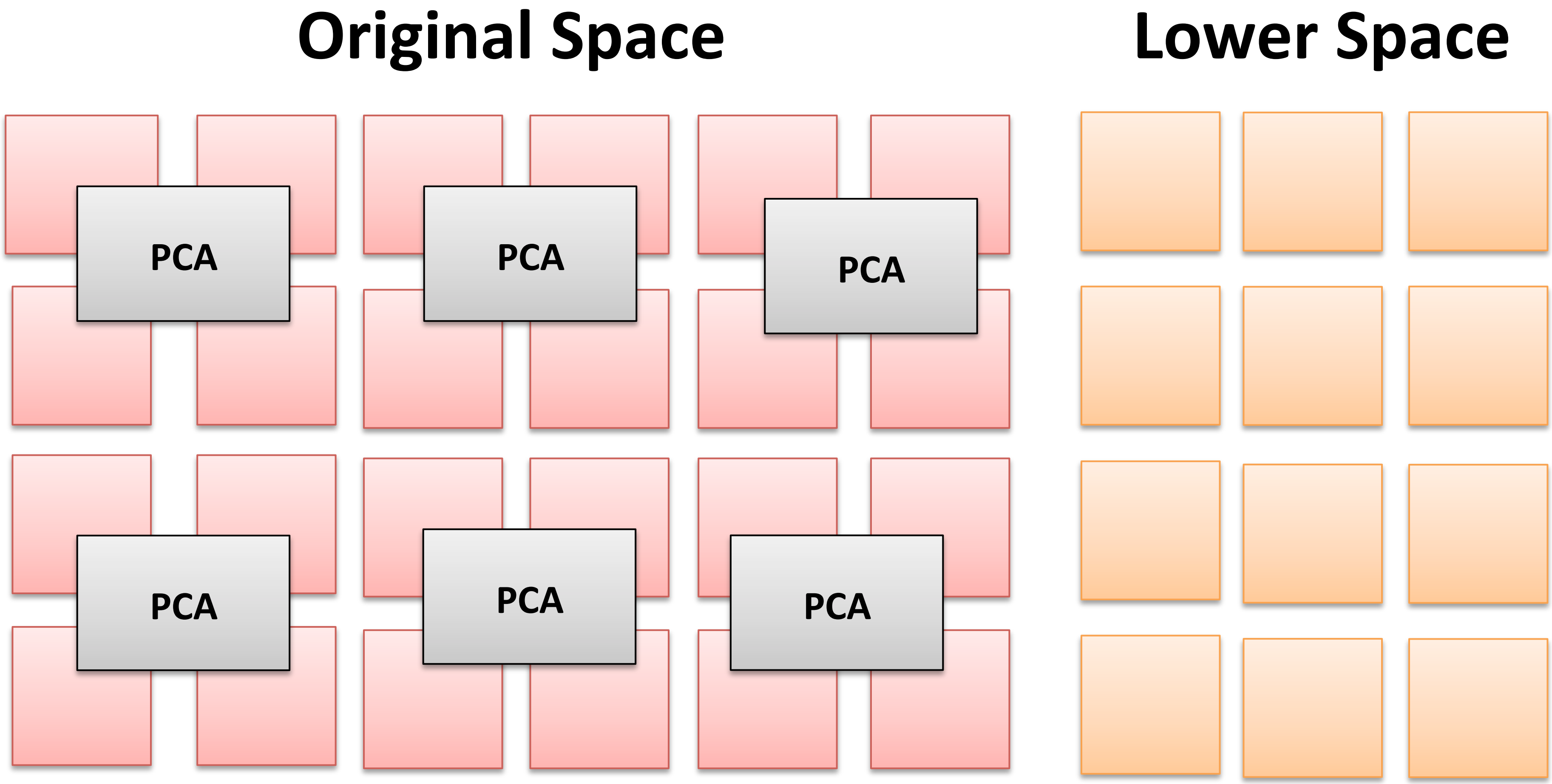
The Hierarchical Linear SubSpace Indexing Method

Original Space **Lower Space**



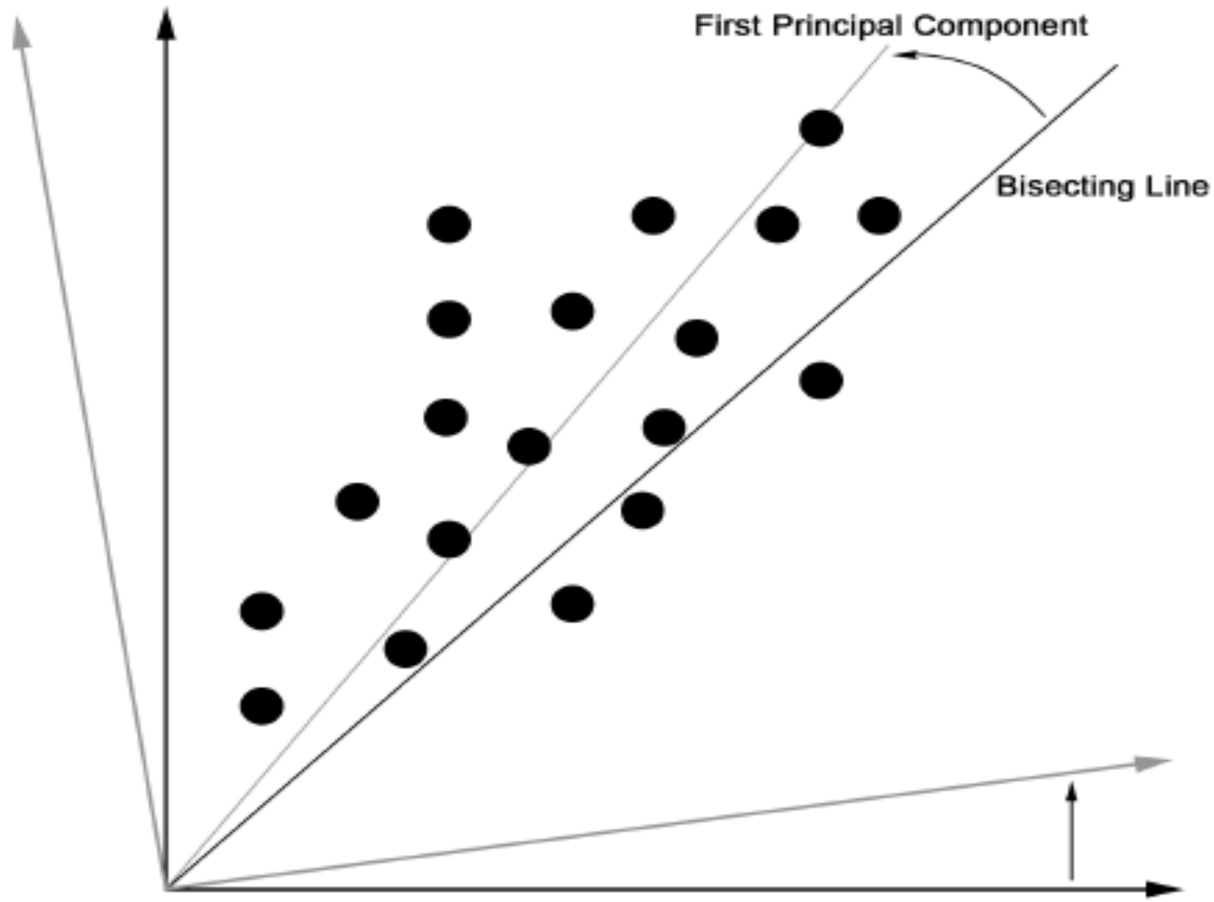
Project HEIDI

The Hierarchical Linear SubSpace Indexing Method



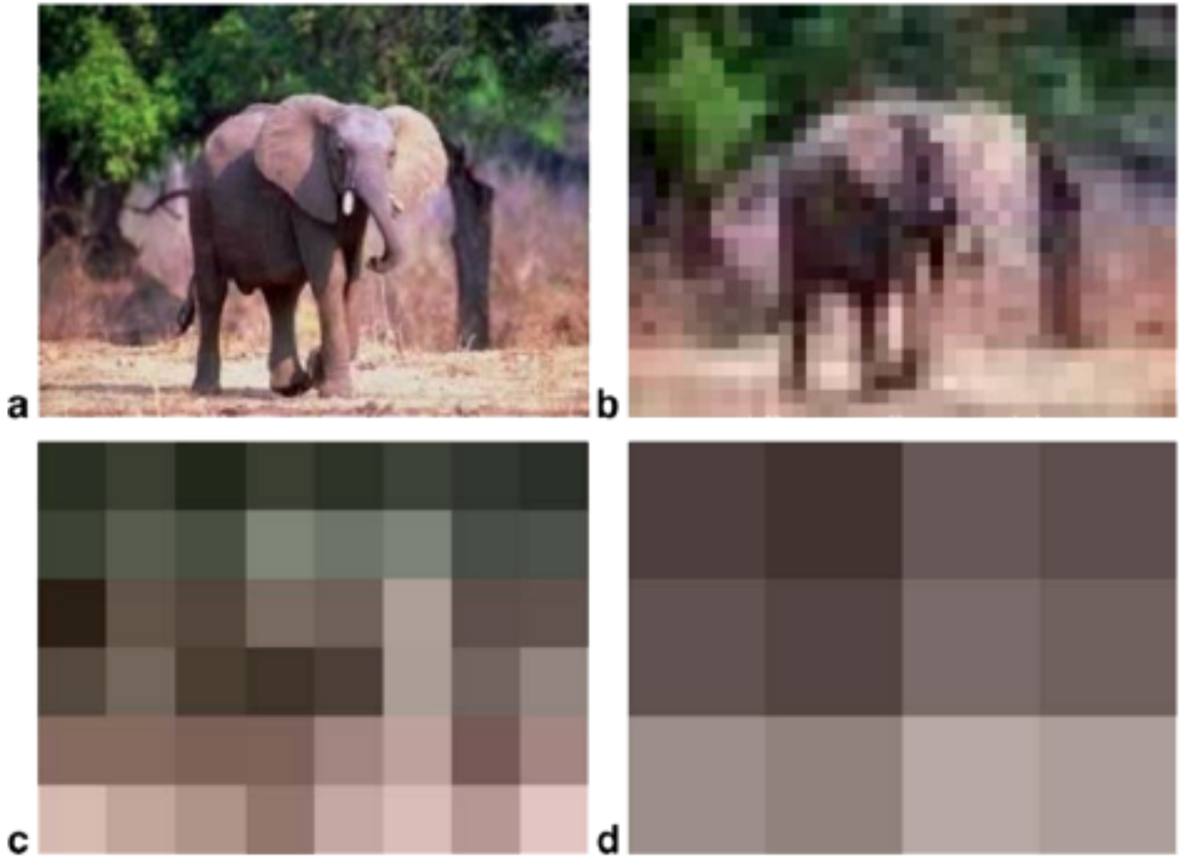
Project HEIDI

The Hierarchical Linear SubSpace Indexing Method



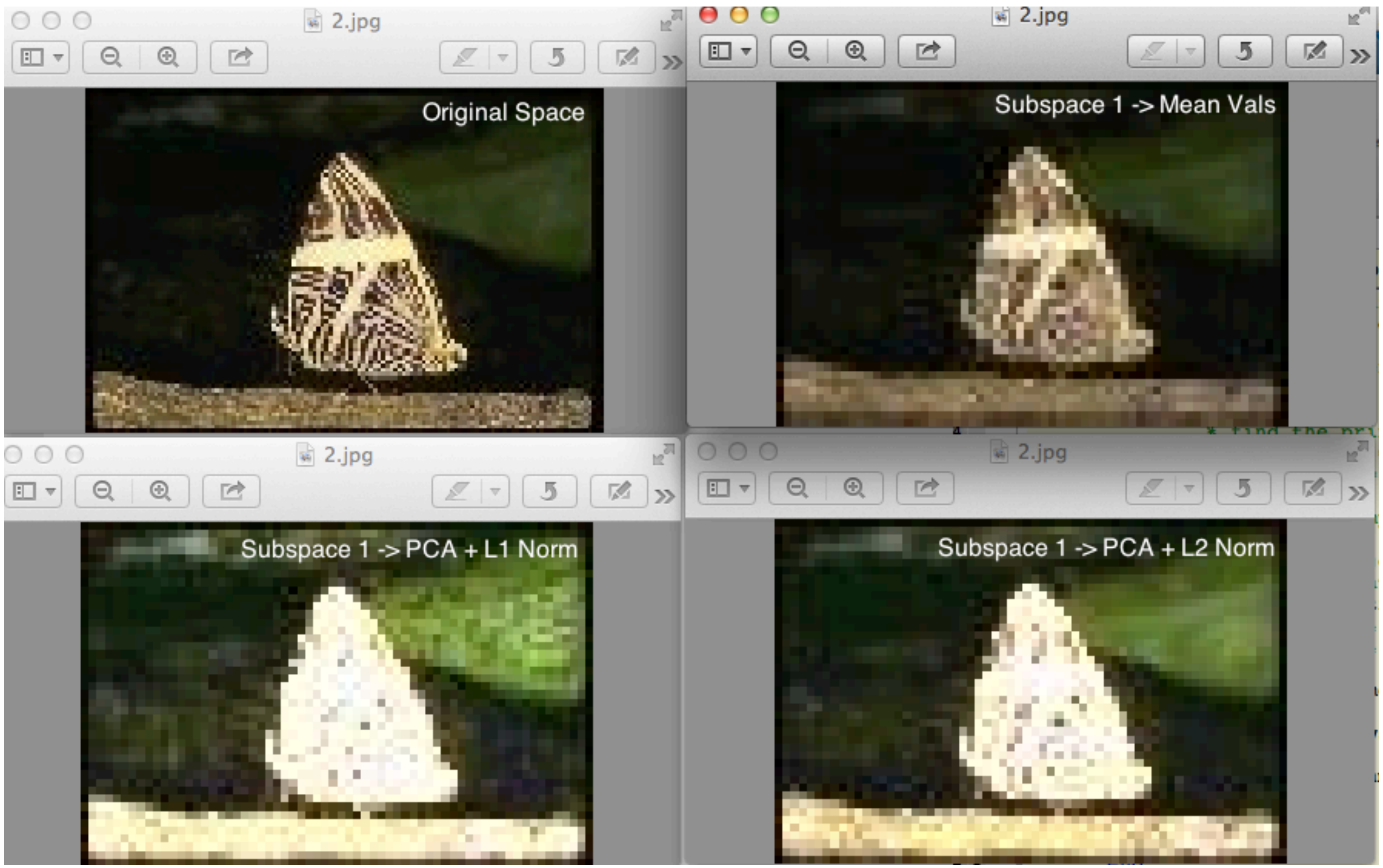
Project HEIDI

The Hierarchical Linear SubSpace Indexing Method



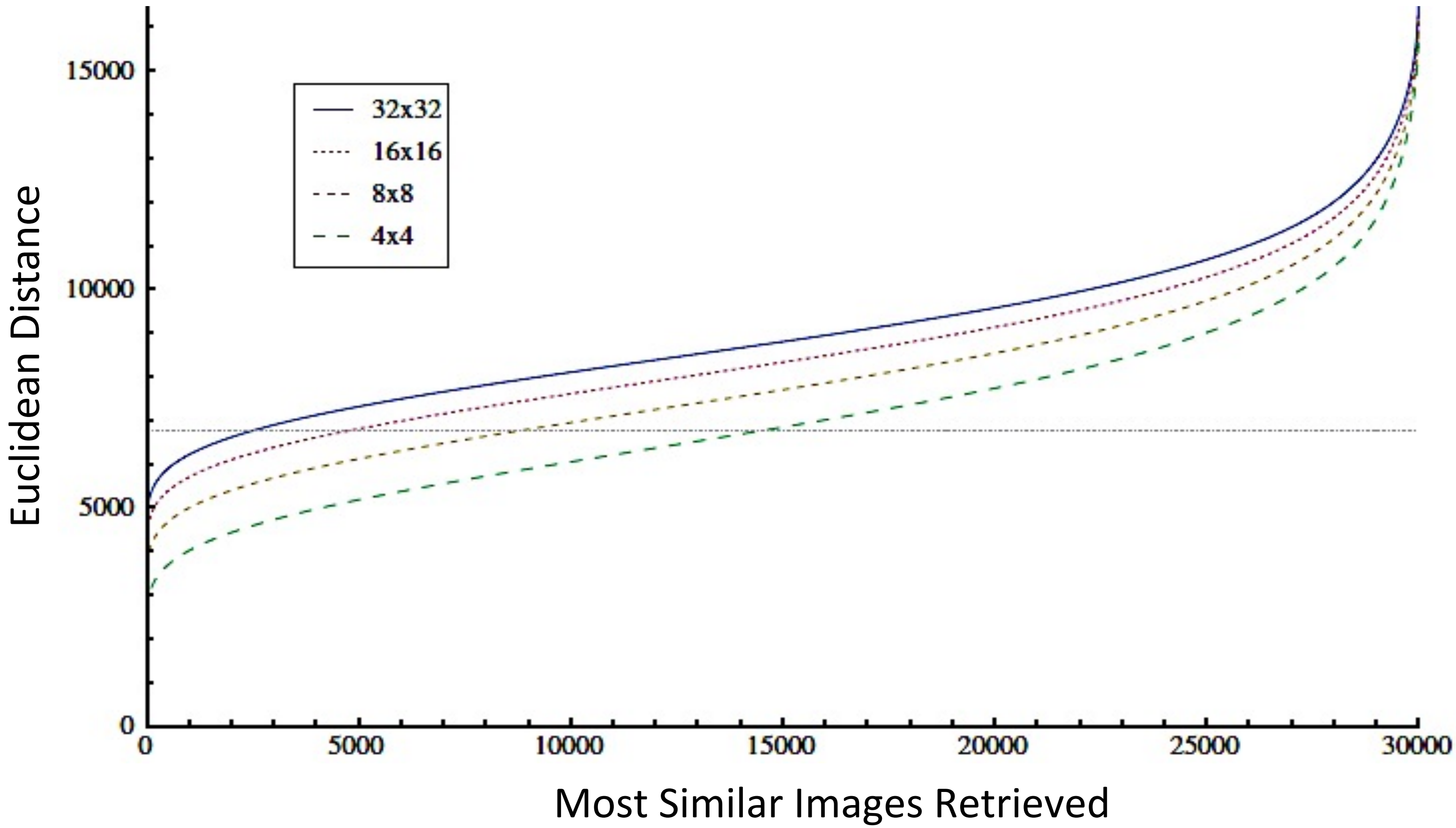
Project HEIDI

The Hierarchical Linear SubSpace Indexing Method



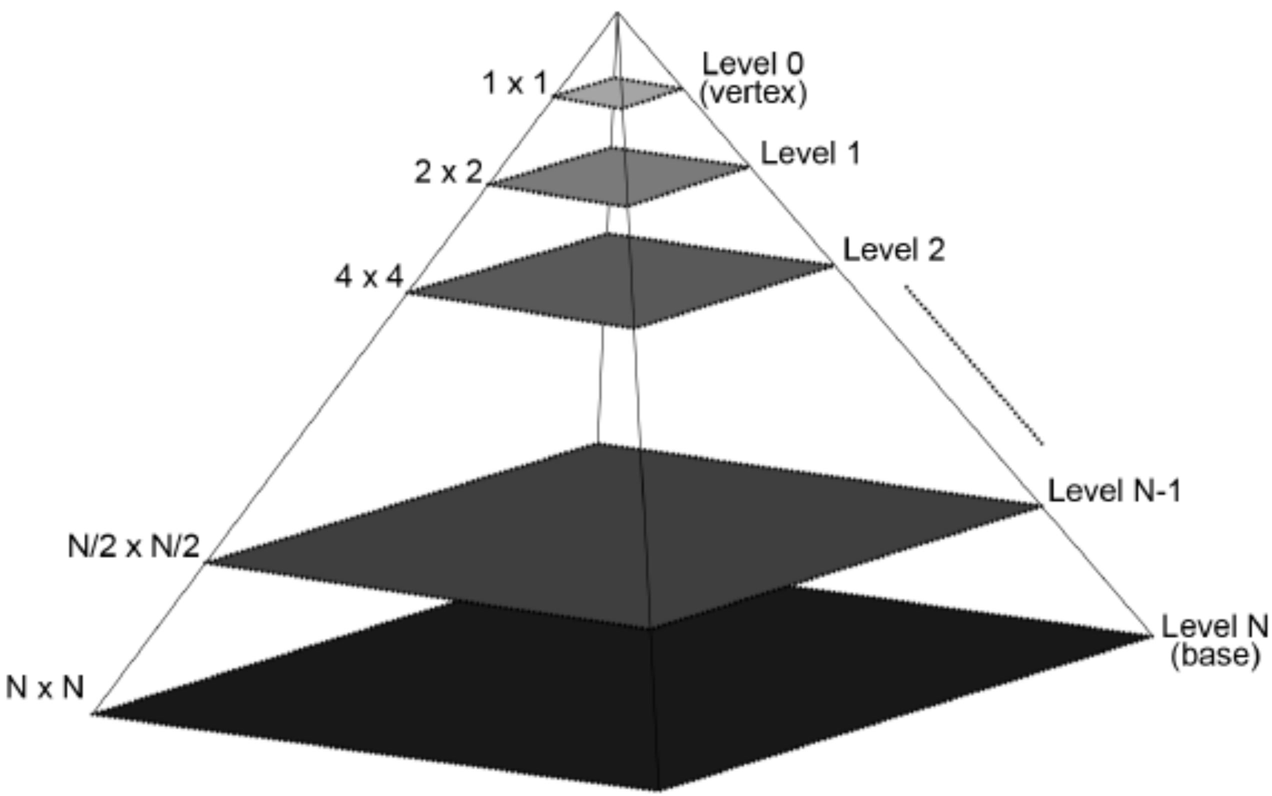
Project HEIDI

The Lower Bounding Lemma is satisfied



Project HEIDI

The Hierarchical Linear SubSpace Indexing Method



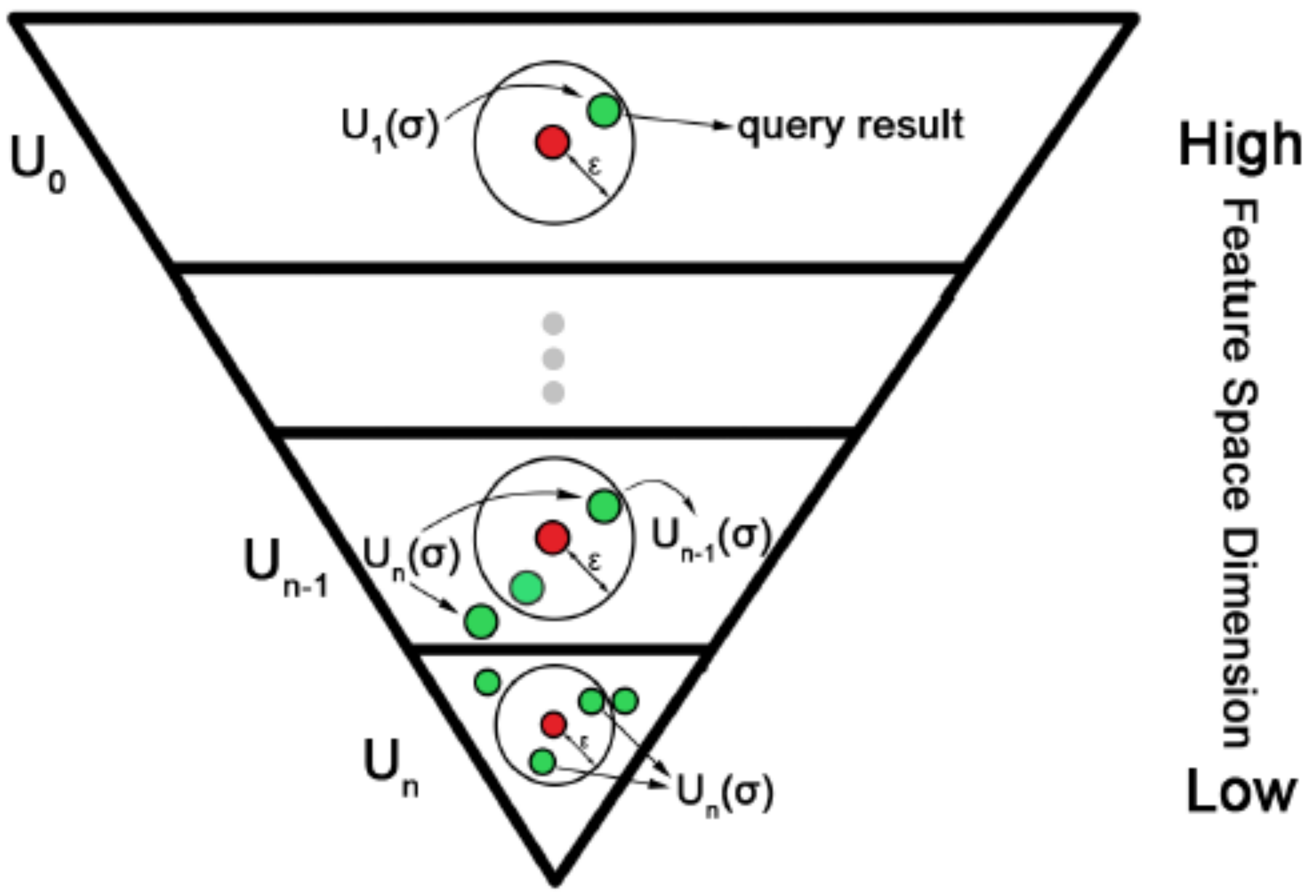
Project HEIDI

The Hierarchical Linear SubSpace Indexing Method

2. Query Phase (on-line)
   – Project queries using projection matrices from step 1
   – Starting in the lowest dimensional space, iteratively discard all objects that are different from the query
   – Keep doing this until the original space is reached and the false hits are discarded.

Project HEIDI

The Hierarchical Linear SubSpace Indexing Method



Summary

- Important concepts in Big Data for High Dimensional Datasets

- The curse of dimensionality
- Projection techniques
- The Lower Bounding Lemma
- The Quick and Dirty paradigm (some relation to the MapReduce paradigm)

Summary

Mathematics is more important than **engineering**!

Now…
Time to See This Working for REAL!!!