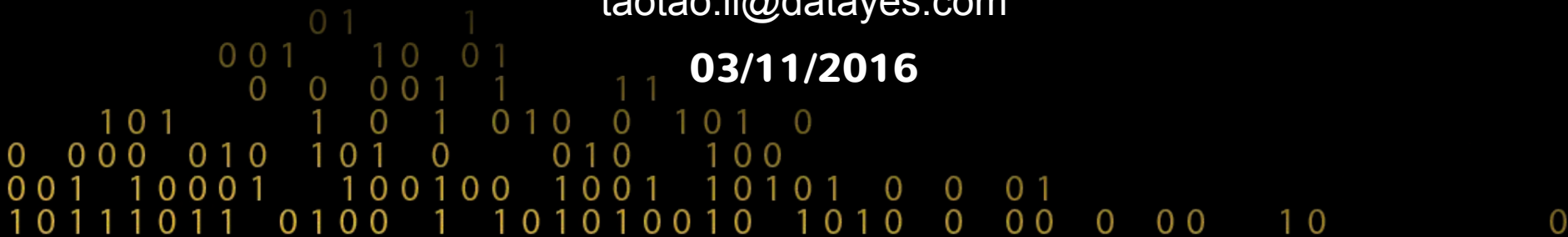


Introducing Spark

taotao.li@datayes.com

03/11/2016



Agenda

1

Spark ! When, What, Why ?

2

Basic Concepts in Spark

3

Programming Model in Spark

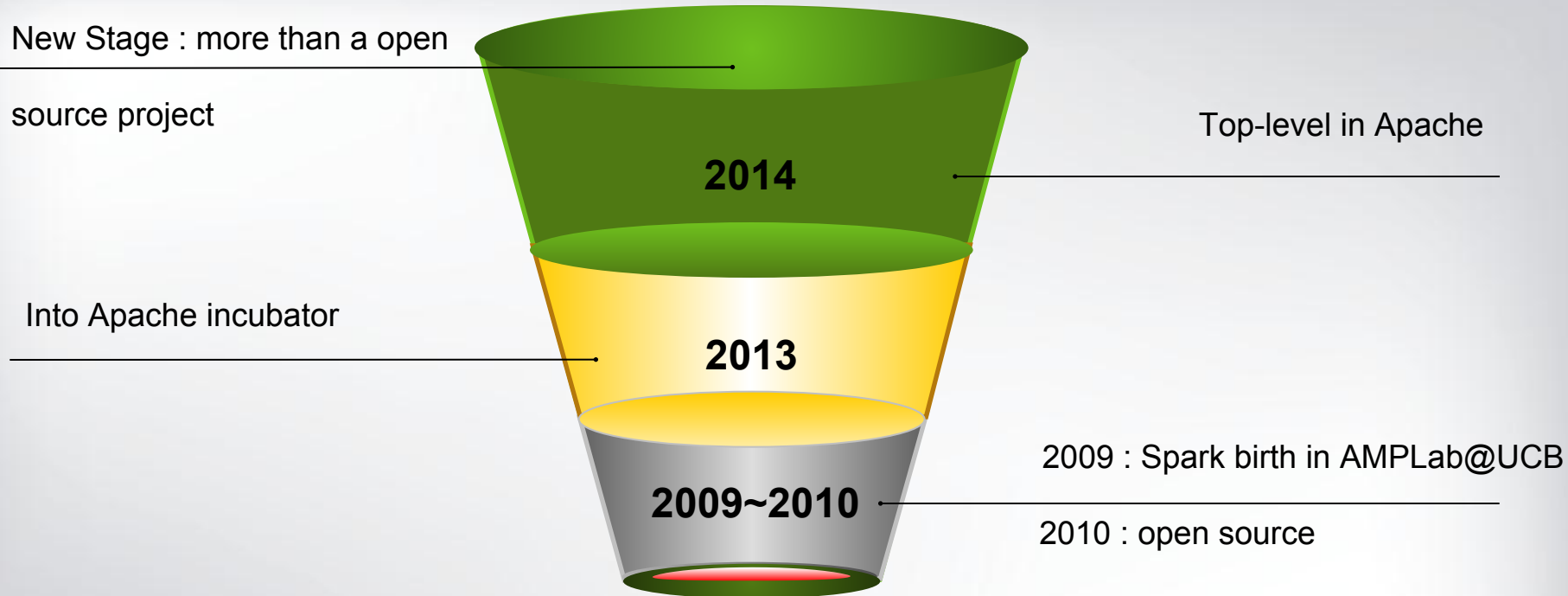
4

Demo, Demo, Demo

5

Q & A

Spark ! *When*, What, Why ?



Spark ! When, *What*, Why ?

From official: Apache Spark™ is a fast and general engine for large-scale data processing.

Key Points:

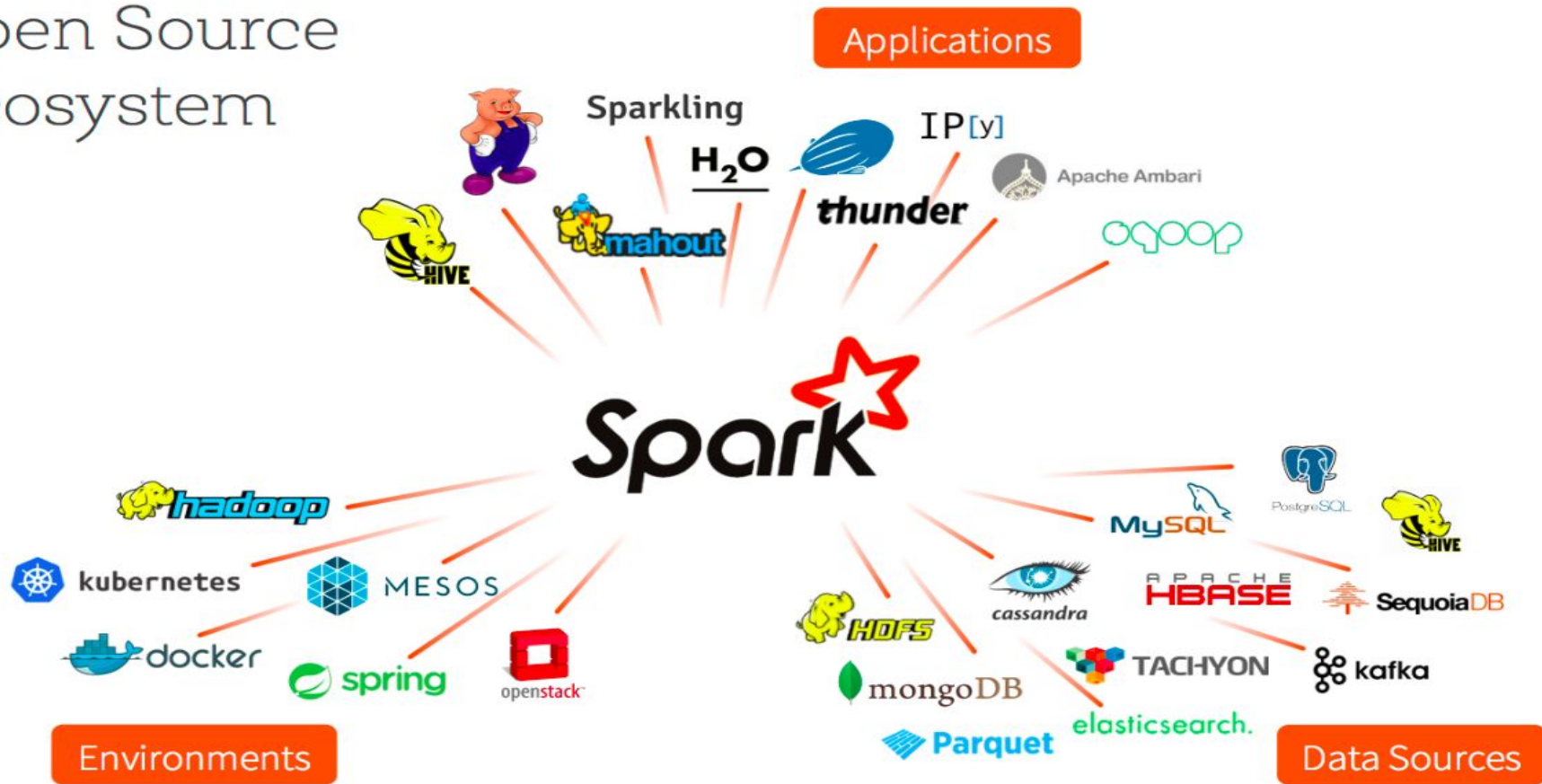
- A framework
- Birth for large-scale data processing
- Generalize programming model for data processing [more than MR]
- Provides high-level APIs : Scala, Python, R, Java
- Arm to teeth : SQL, Streaming, Machine Learning, GraphX
- Compatible with previous ecology : hadoop, mesos, hdfs, cassandra, hbase, s3 ...

Spark ! When, What, *Why* ?

- General
- Fast in develop
 - REPL explore
 - RDD operations
 - Less code
- Fast in processing
- Compatible
- Packages and 3-party packages
- ***Memory, cheaper and cheaper***
- ***Company who accepts Spark***

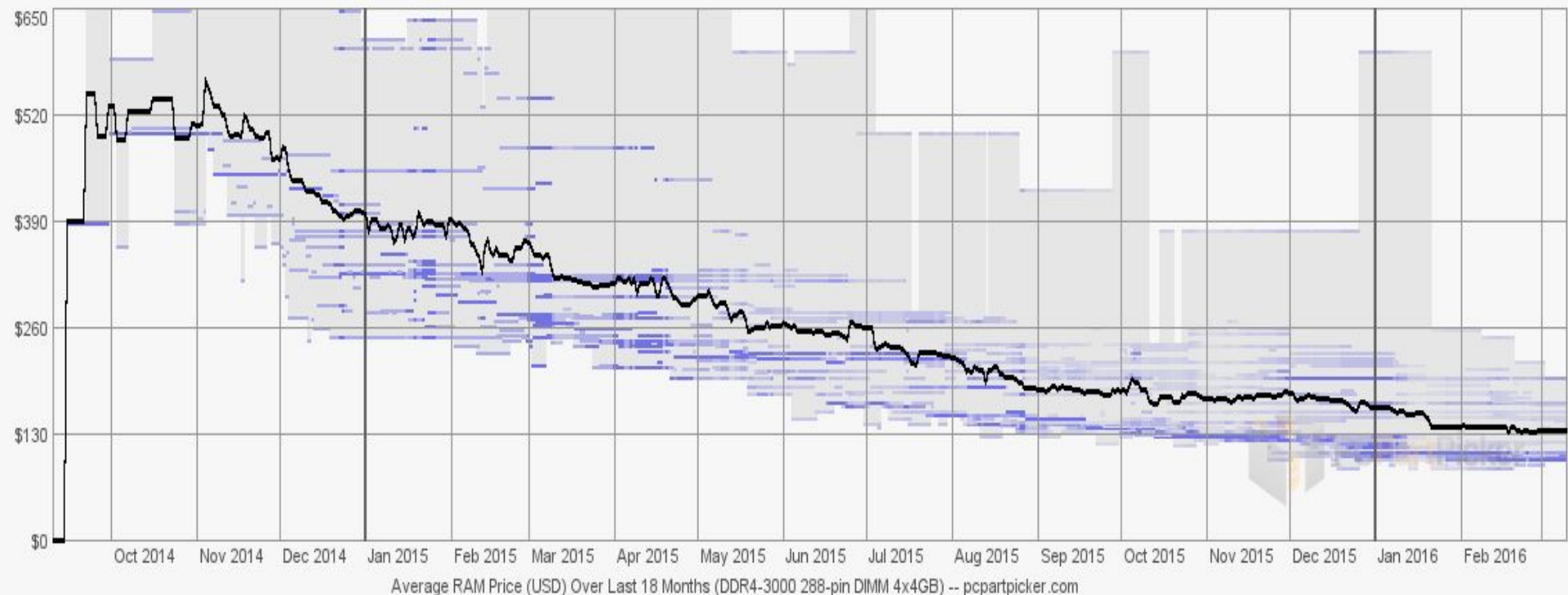
Spark ! When, What, *Why* ?

Open Source
Ecosystem

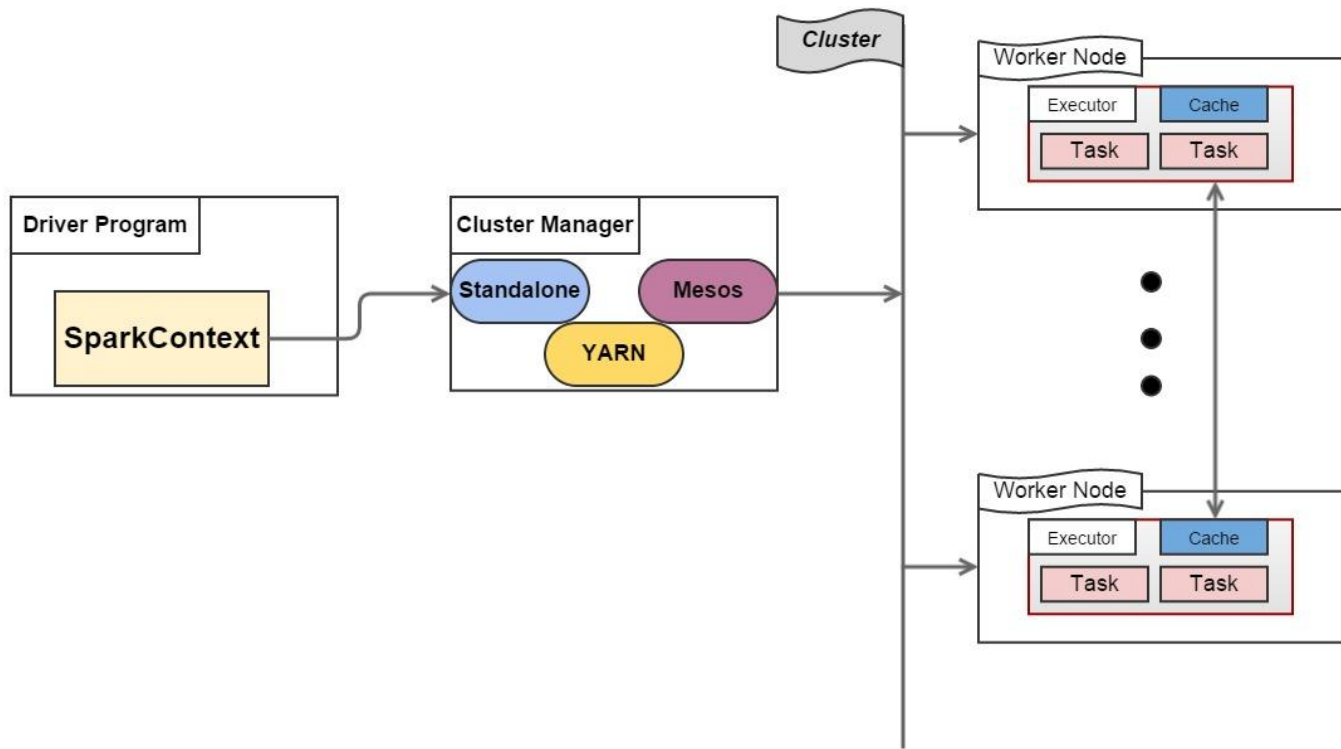


Spark ! When, What, *Why* ?

DDR4-3000 288-pin DIMM 4x4GB Price Trend



Basic Concepts in Spark

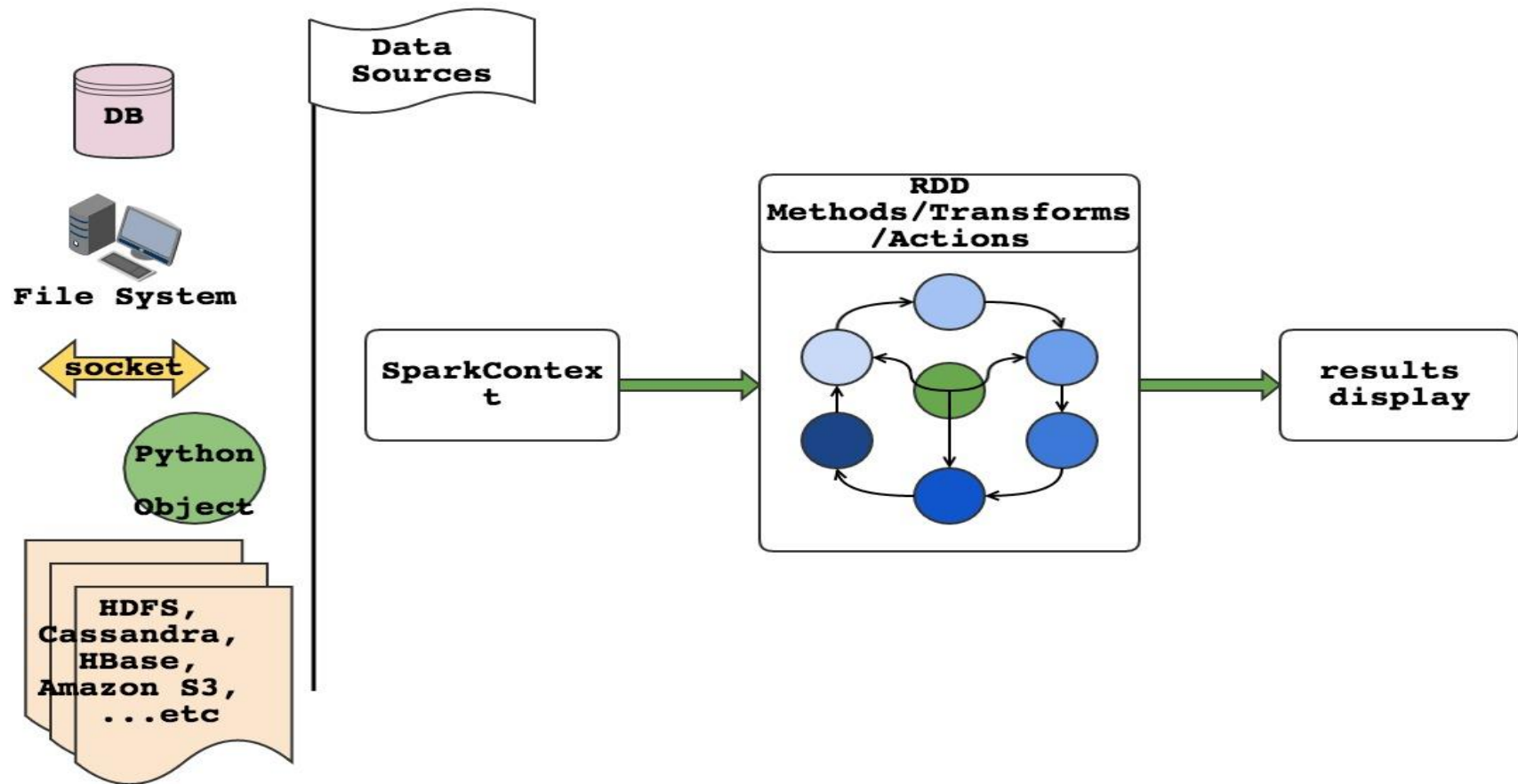


Basic Concepts in Spark

- Driver, Master, Worker, Executor
- Application
- SparkContext, i.e : sc
- RDD
- Transform & Action in RDD

need more ? check : [『 Spark 』2. spark 基本概念解析](#)

Programming Model in Spark



Programming Model in Spark

Three basic steps to build a Spark Application

- load dataset
 - static dataset
 - dynamic dataset
- Processing
 - RDD operation
 - UDF
 - Cache
- Output Display
 - collect
 - store in database, file system ...

Demo, Demo, Demo

- Monte Carlo in Spark
-
- Spark in finance : index similarity calculating
-
- Spark in finance : distributed backtesting strategy

Demo, Demo, Demo

Q & A

