Lesson 1.1: Why Distributed Computing?

DISTRIBUTED COMPUTING WITH SPARK SQL

Why Distributed Computing?

Brooke Wenig/Conor Murphy
Machine Learning Practice Lead/
Data Scientist, Databricks

UCDAVIS
Continuing and Professional Education

Copyright ©The Regents of the University of California

Slide 2: Learning Objectives

Learning Objectives

Motivate the business need for processing big data

Identify key concepts related to distributed computing

Slide 3: Qualities of Big Data

Qualities of Big Data

Volume

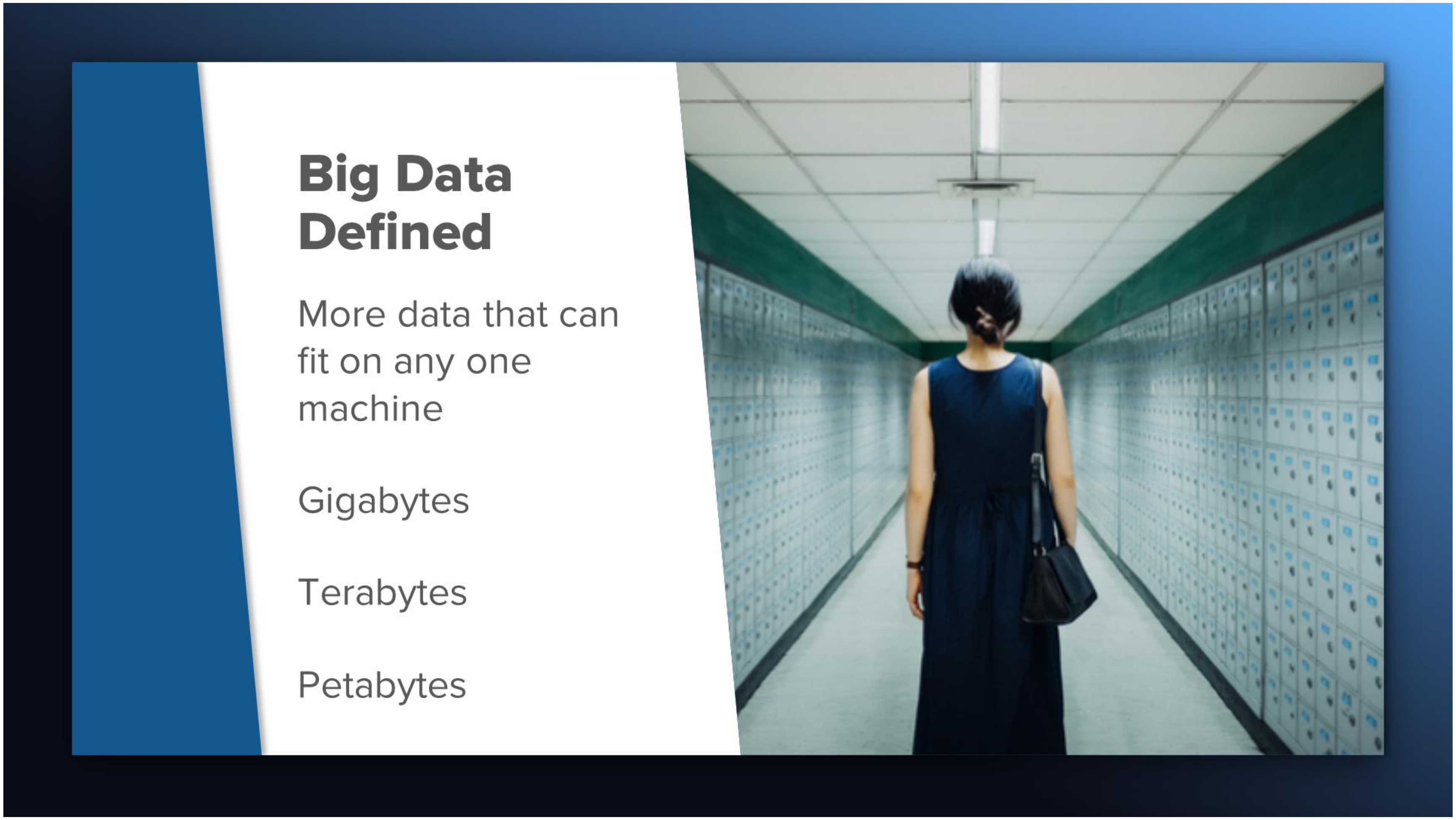
Velocity

Variety

Veracity

Copyright ©The Regents of the University of California

Slide 4: Big Data Defined



Big Data Defined

More data that can fit on any one machine

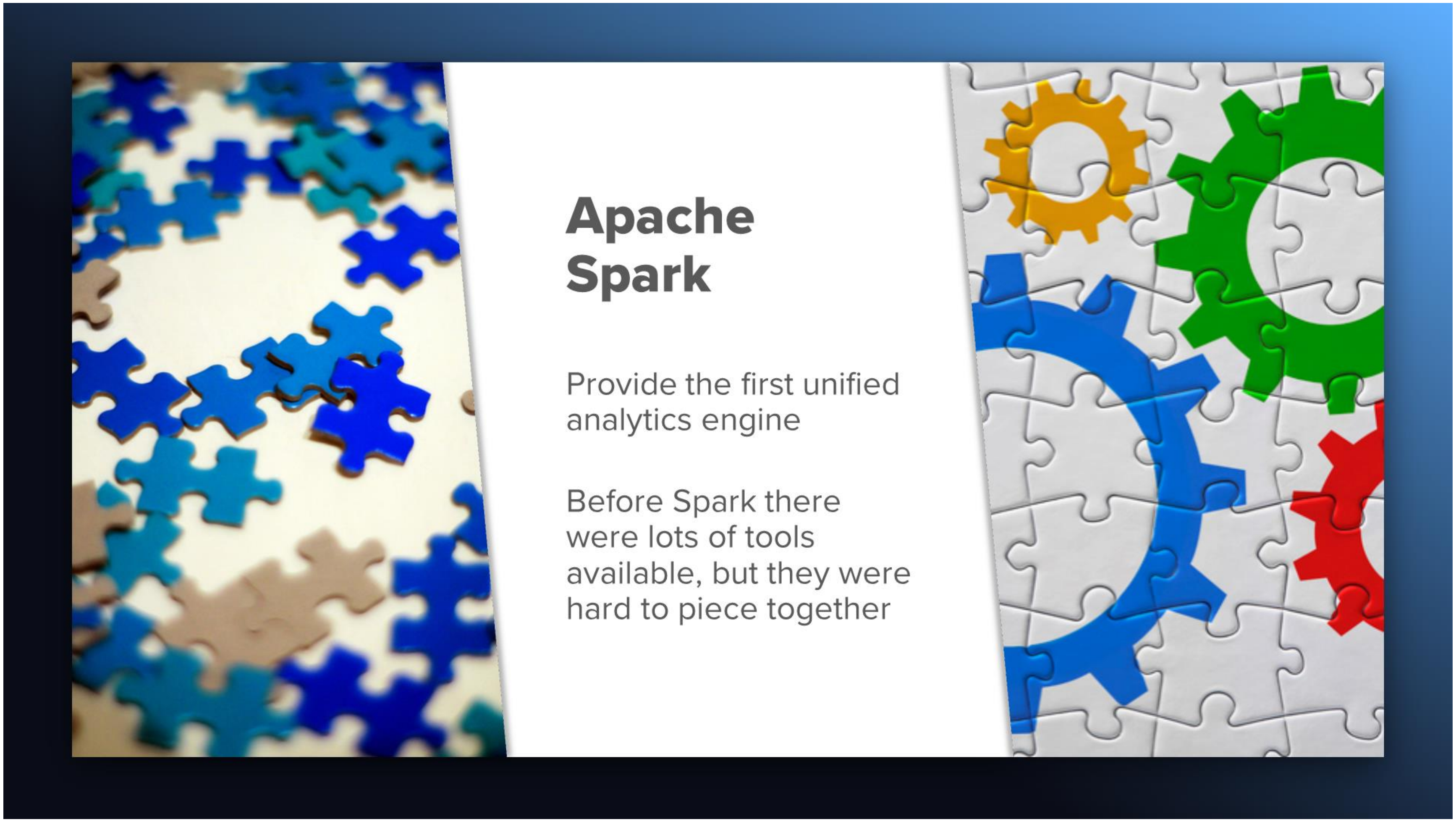
Gigabytes

Terabytes

Petabytes

Copyright ©The Regents of the University of California

Slide 5: Apache Spark



Apache Spark

Provide the first unified analytics engine

Before Spark there were lots of tools available, but they were hard to piece together

Copyright ©The Regents of the University of California

Slide 6: Spark is Multilingual and Supports Many Languages

Spark Supports Many Languages

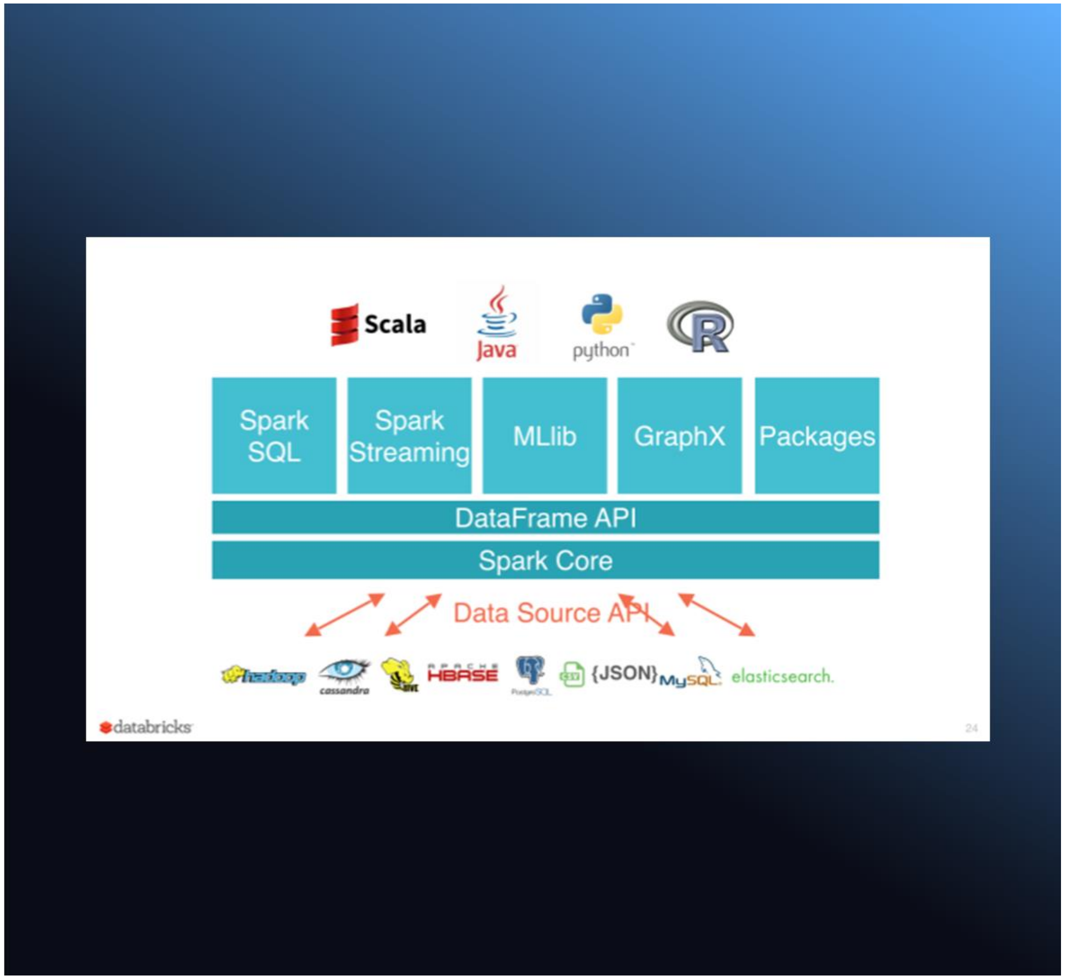
SQL

Python

Scala

Java

R



Copyright ©The Regents of the University of California

Slide 7: Why Spark is Popular

Why Spark is Popular

Reads & processes data from many sources

Works with many file types

Solves many data problems faced by analysts

Copyright ©The Regents of the University of California

Slide 8: Apache Spark: Origin Story

Apache Spark: Origin Story

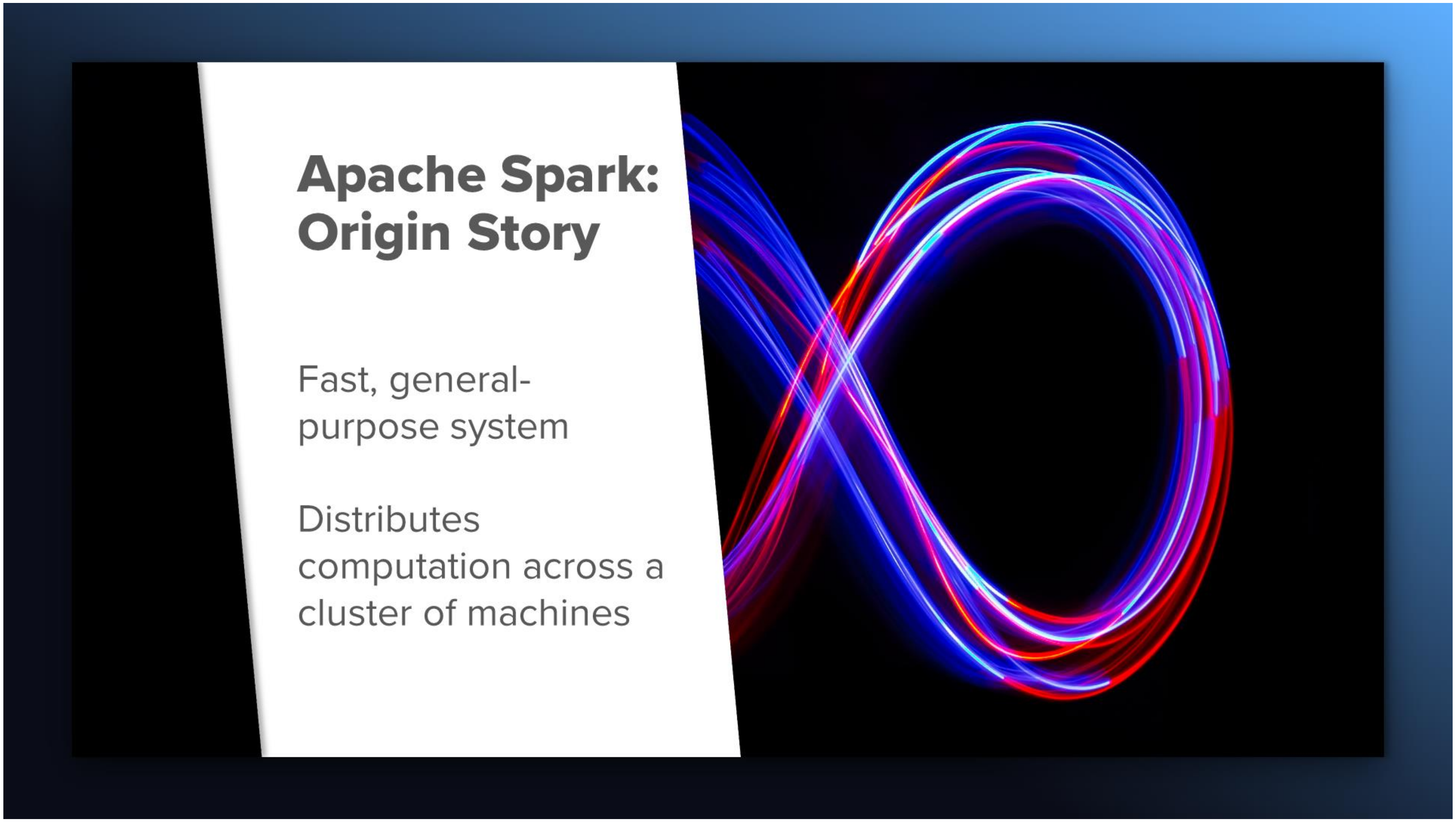
UC Berkeley, 2009

Matei Zaharia, co-founder of Databricks

Collaboration between professors and graduate students

Copyright ©The Regents of the University of California

Slide 9: Apache Spark: Origin Story



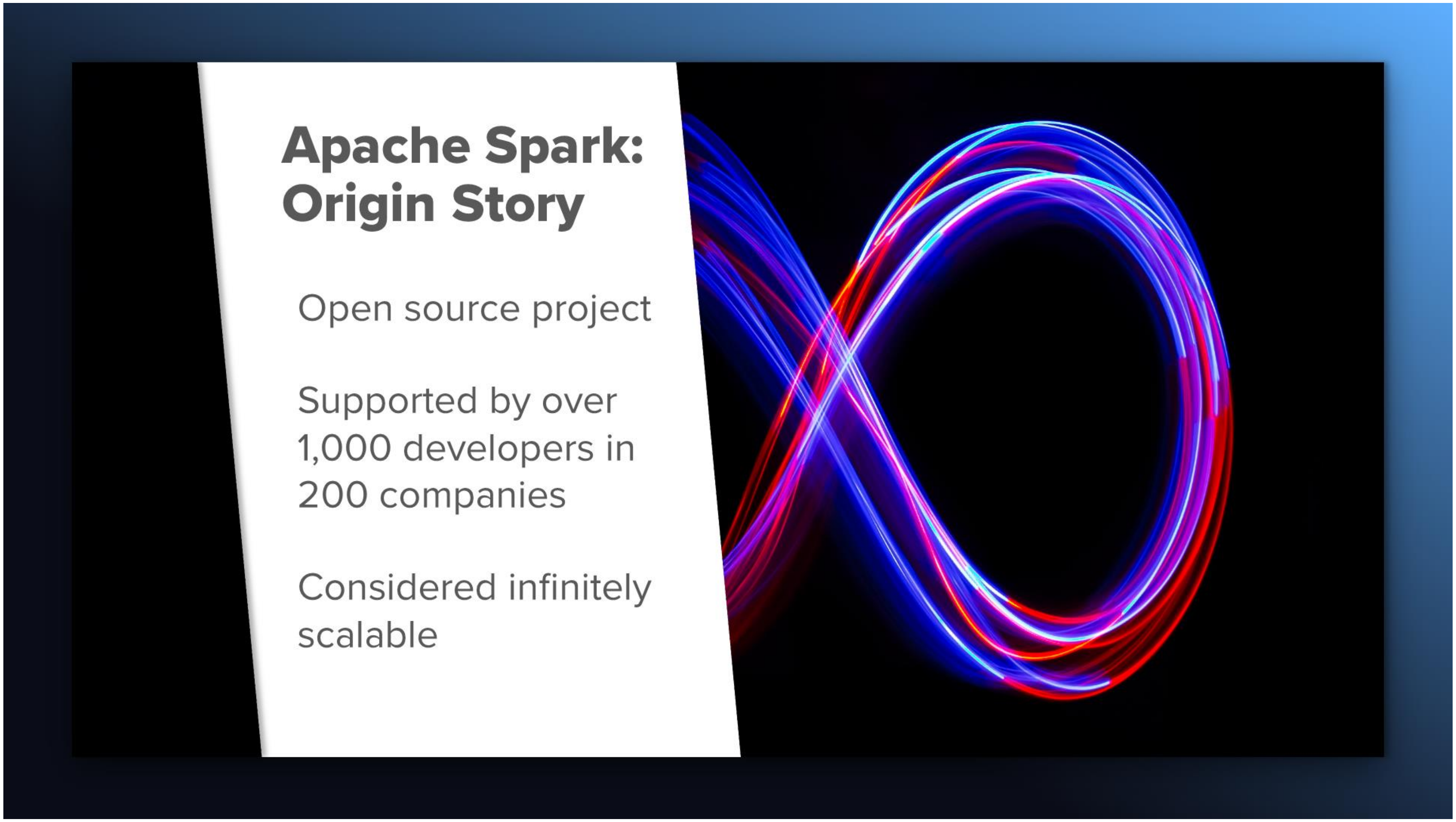
Apache Spark: Origin Story

Fast, general-purpose system

Distributes computation across a cluster of machines

Copyright ©The Regents of the University of California

Slide 10: Apache Spark: Origin Story



Apache Spark: Origin Story

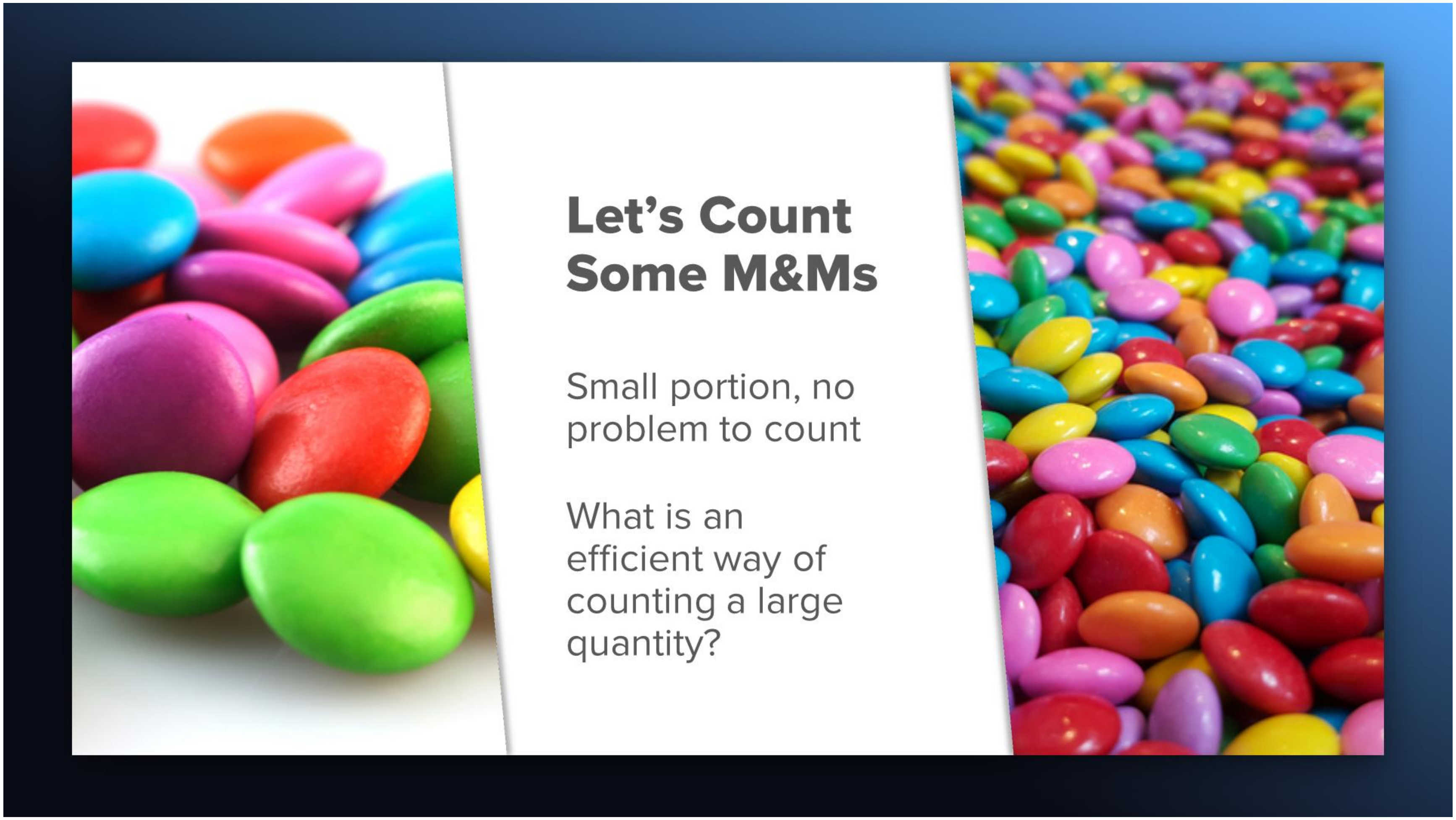
Open source project

Supported by over 1,000 developers in 200 companies

Considered infinitely scalable

Copyright ©The Regents of the University of California

Slide 11: Let’s Count Some M&Ms



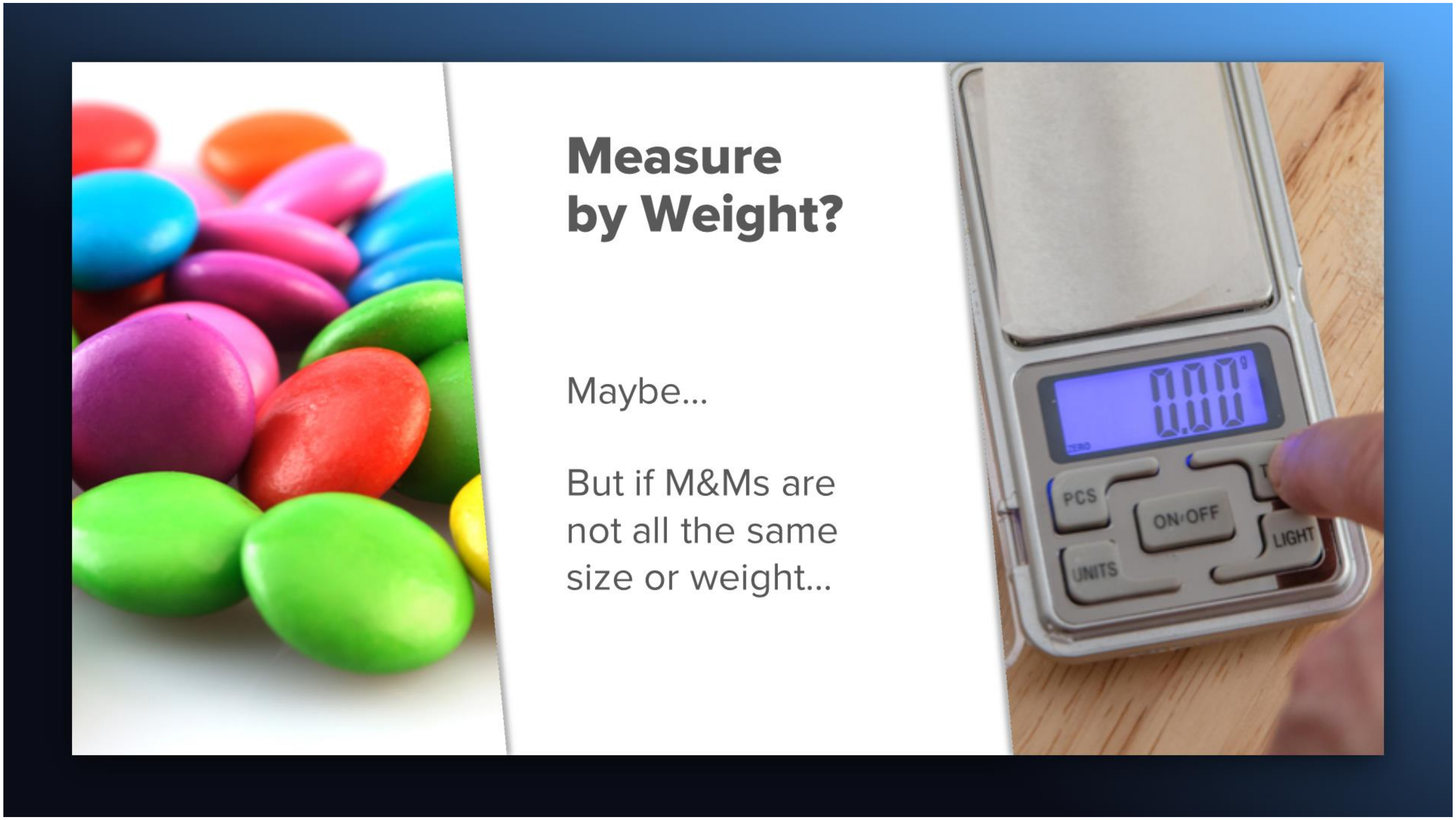
Let’s Count Some M&Ms

Small portion, no problem to count

What is an efficient way of counting a large quantity?

Copyright ©The Regents of the University of California

Slide 12: Measure by Weight?



0.00 g
PCS
ON/OFF
UNITS
LIGHT

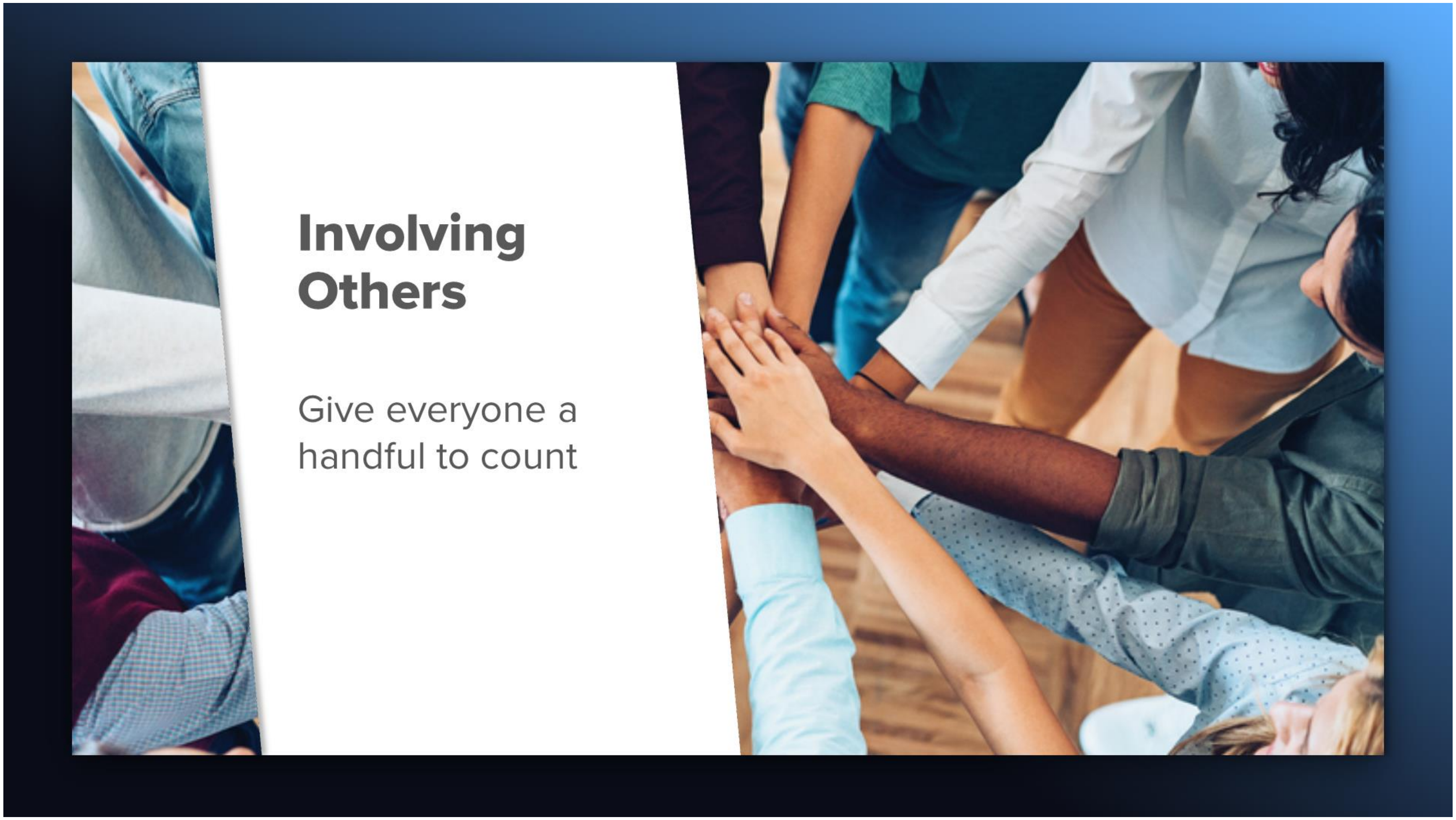
Measure by Weight?

Maybe...

But if M&Ms are not all the same size or weight...

Copyright ©The Regents of the University of California

Slide 13: Involving Others

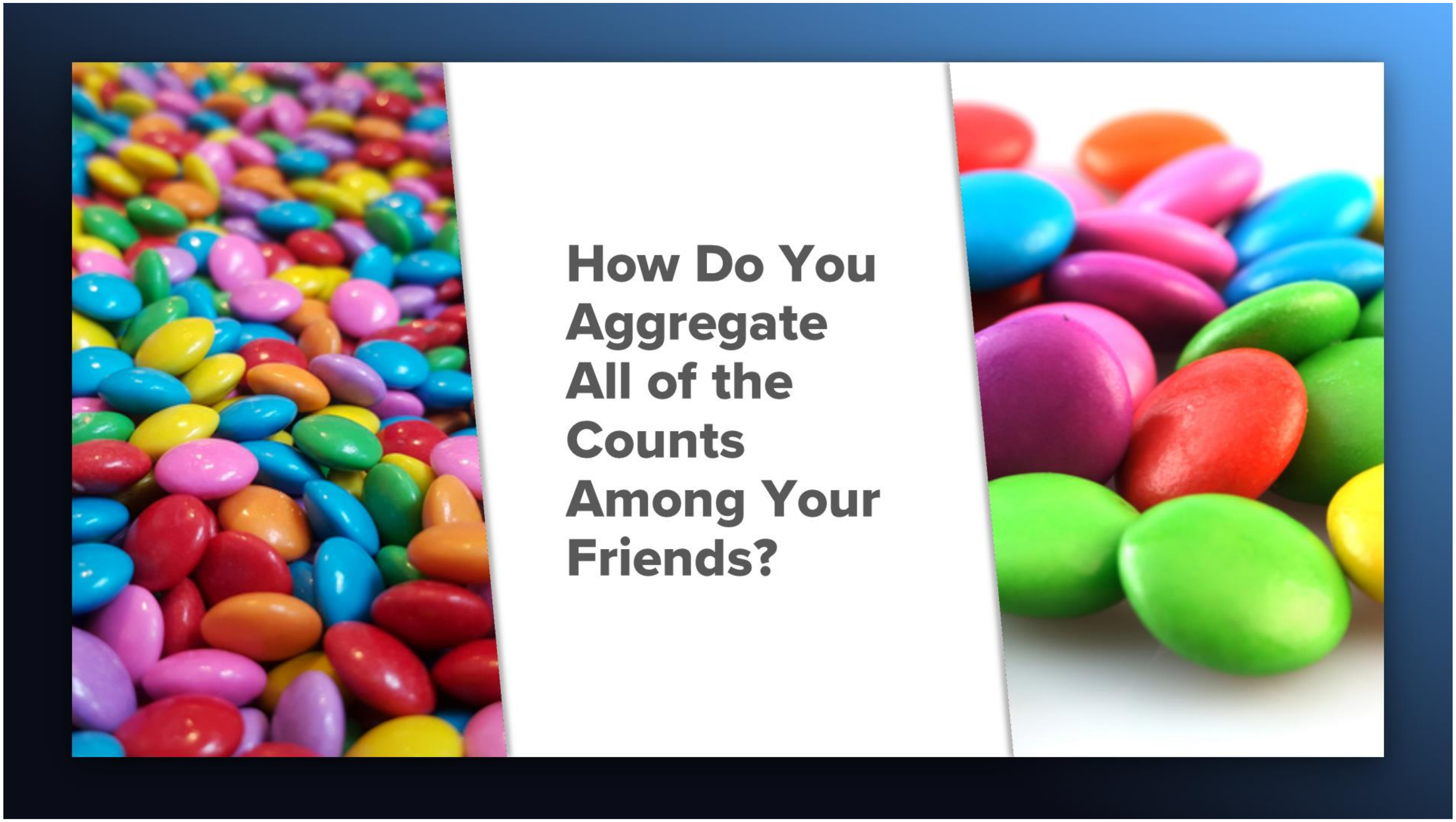


Involving Others

Give everyone a handful to count

Copyright ©The Regents of the University of California

Slide 14: How Do You Aggregate All of the Counts Among Your Friends?



How Do You Aggregate All of the Counts Among Your Friends?

Copyright ©The Regents of the University of California

Slide 15: Drivers and Executors

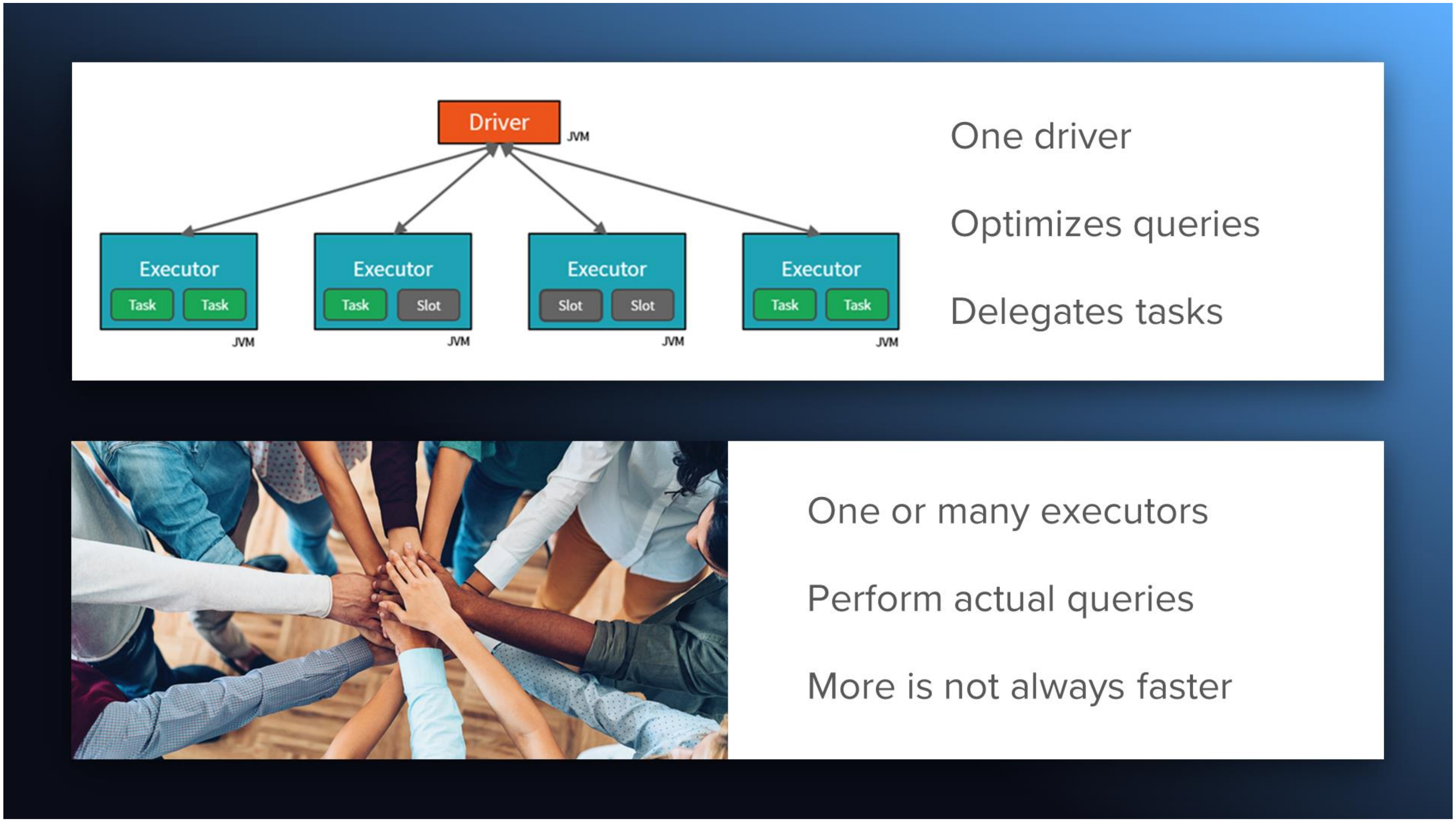
Driver and Executors

You are the driver

Friends are executors

Copyright ©The Regents of the University of California

Slide 16: Drivers and Executors



One driver

Optimizes queries

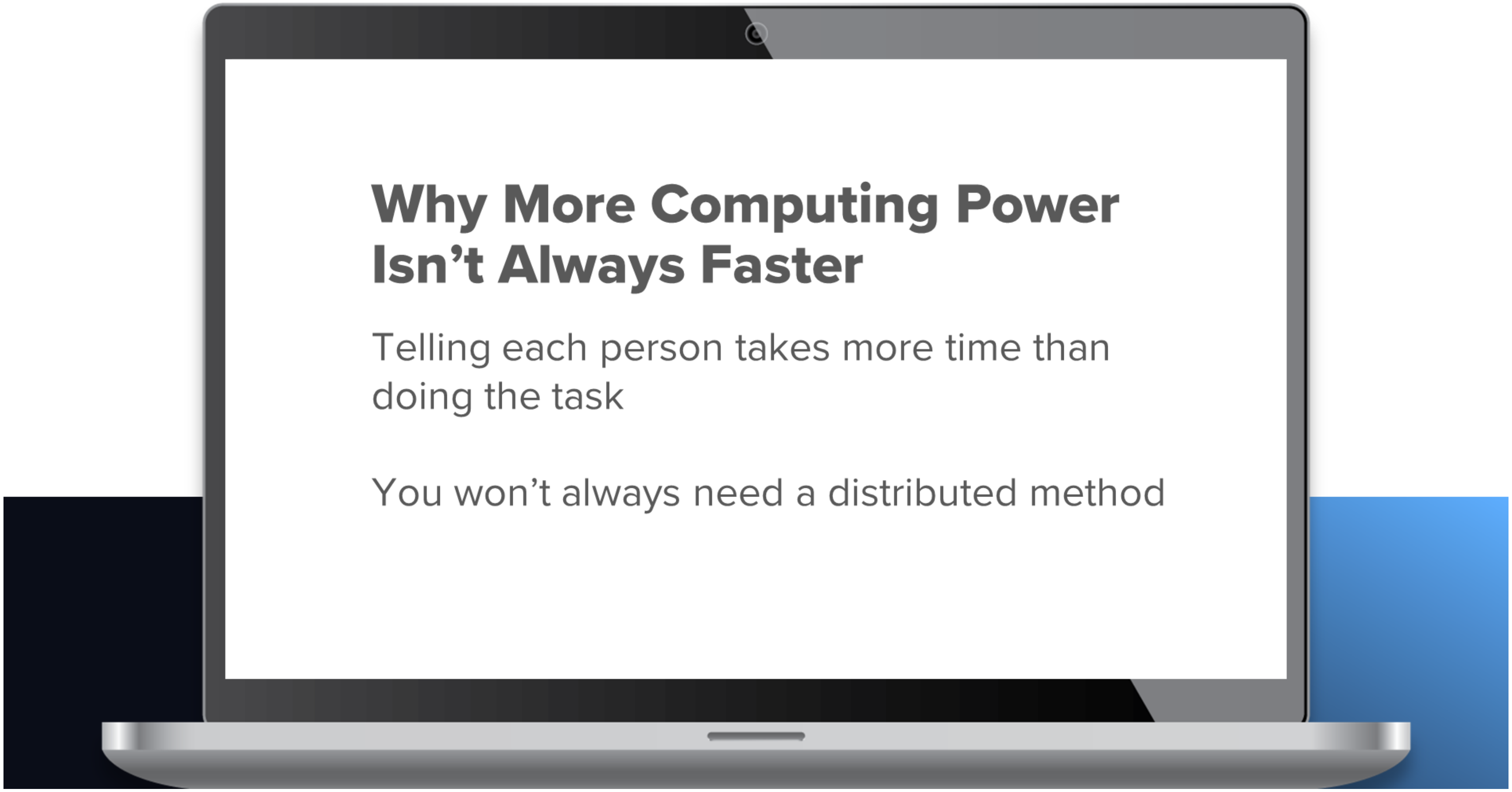
Delegates tasks

One or many executors

Perform actual queries

More is not always faster

Slide 17: Why More Computing Power Isn't Always Faster



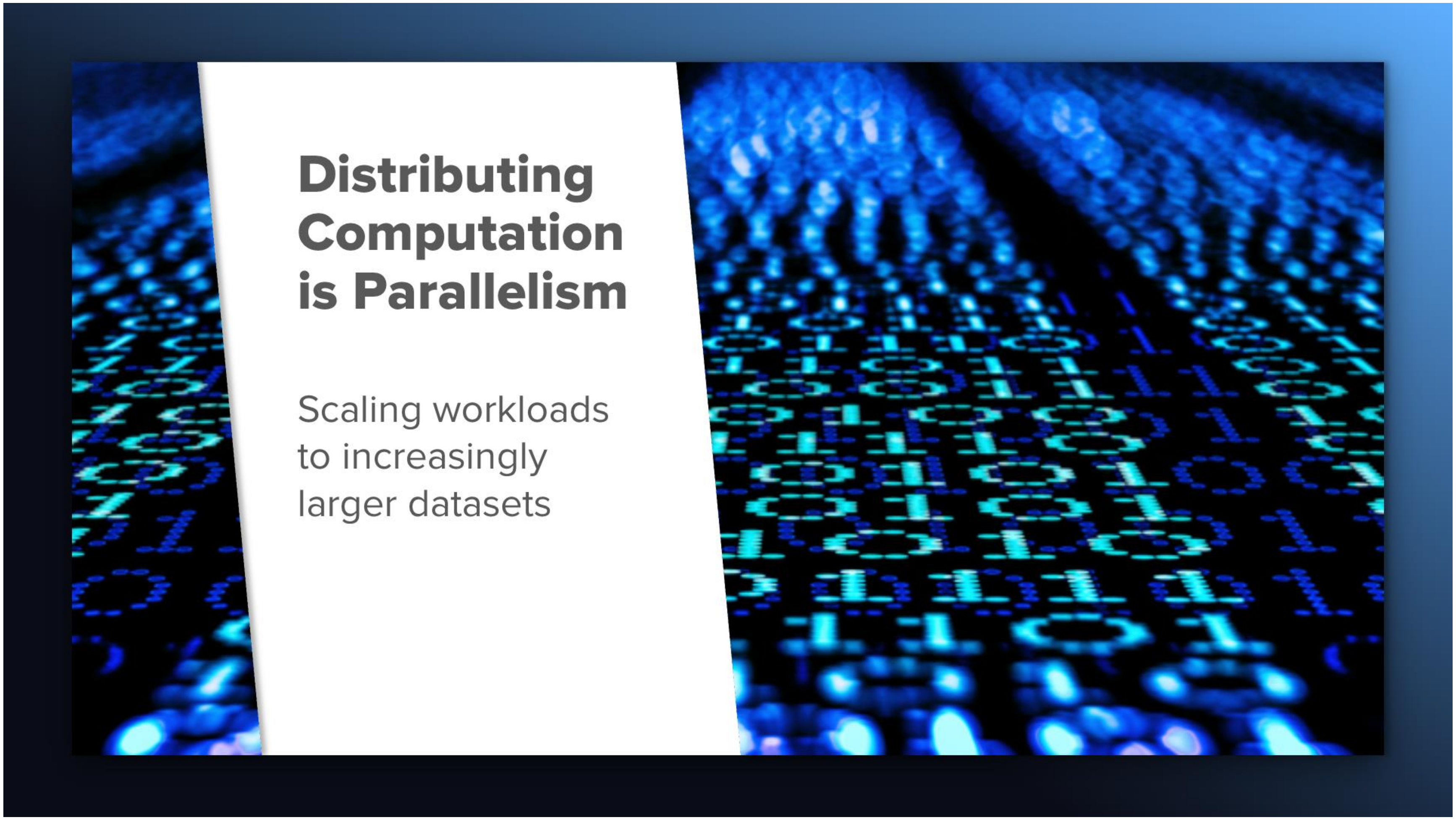
Why More Computing Power Isn't Always Faster

Telling each person takes more time than doing the task

You won't always need a distributed method

Copyright ©The Regents of the University of California

Slide 18: Distributing Computation is Parallelism



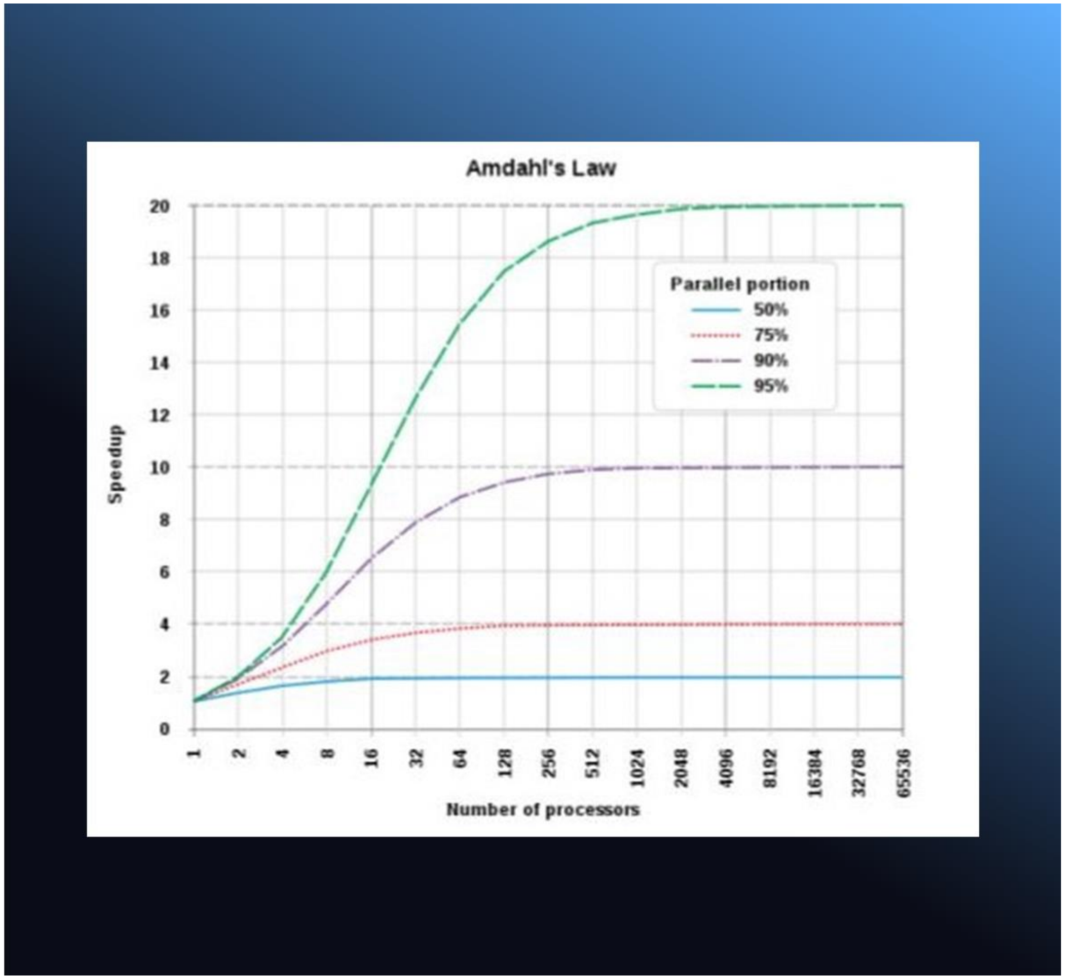
Distributing Computation is Parallelism

Scaling workloads to increasingly larger datasets

Slide 19: Amdahl’s Law

Amdahl’s Law

The amount of acceleration we would see from parallelizing a task is a function of what portion of the task can be completed in parallel



Copyright ©The Regents of the University of California

Slide 20: Linear Scalability

Linear Scalability

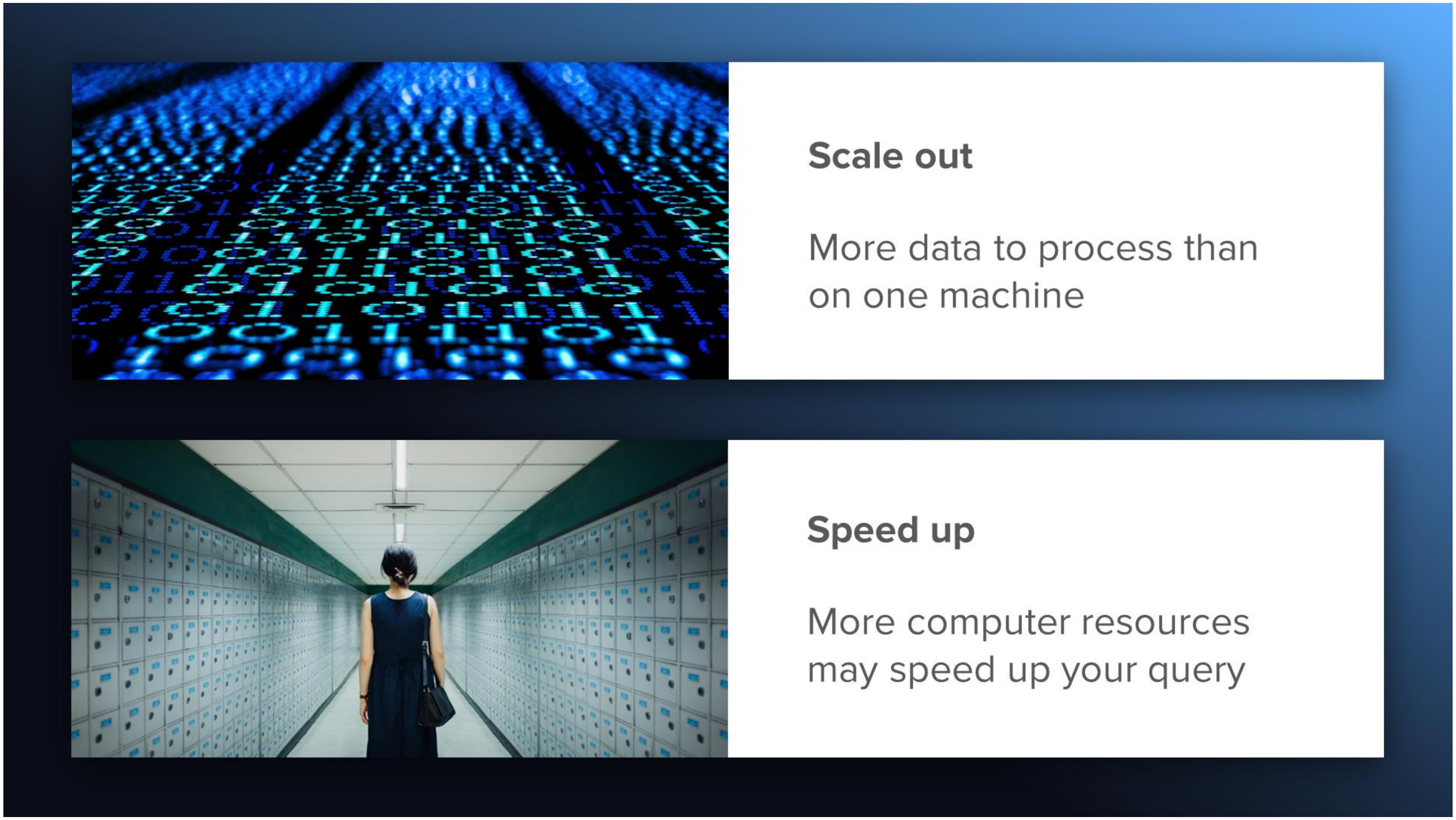
Dividing tasks across a cluster of machines

We see improvements up to thousands of machines

Spark is on par with best distributed computing solutions on the market

Copyright ©The Regents of the University of California

Slide 21: Scalability



Scale out

More data to process than on one machine

Speed up

More computer resources may speed up your query

Slide 22: Coming Up

Coming Up

Learn about core Spark concepts

Copyright ©The Regents of the University of California