Analysis Pipeline Introduction

David Levine

August 9, 2017

Session Outline

- Computing environments
- Analysis pipeline architecture
- Demo on local cluster
- Demo on AWS Batch
- Wrap up

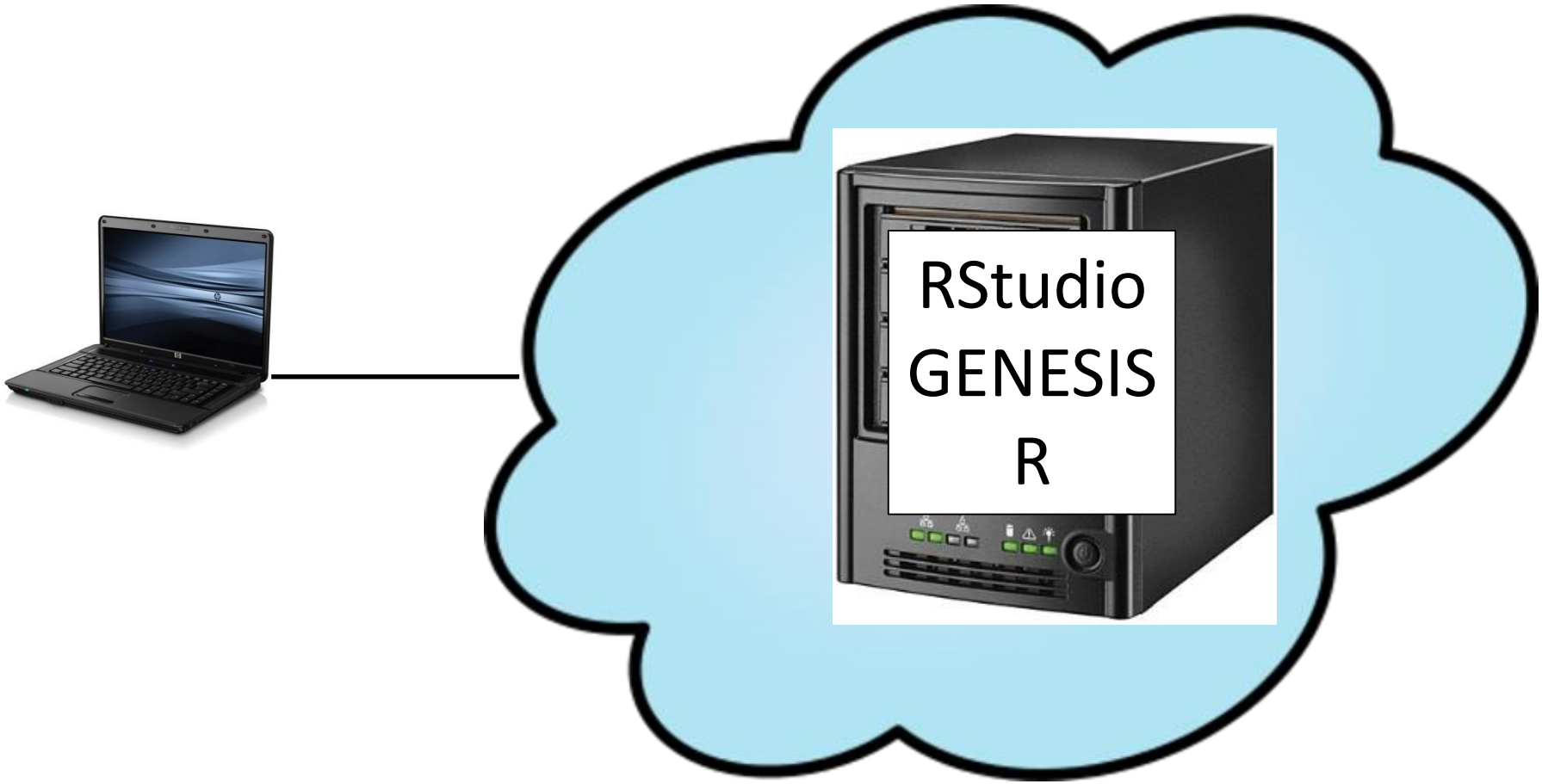
Generally not powerful enough

Data Set	Samples	Variants
workshop	1,126	25,760
freeze.1c	2,643	112,275,224
freeze.2a	9,109	140,980,783
freeze.3a	16,558	185,970,832
freeze.3a.phased	18,258	200,750,986
freeze.4	18,526	219,154,455
freeze.5	60,000	470,000,000

- Memory
- CPU
- Disk space

Single server OK for small data sets

Workshop: Server access via AWS*



* Amazon Web Services

TOPMed data sets belong on a cluster

- Local Cluster?
- Cloud Cluster?

GENESIS on a cluster

Local cluster SGE	AWS CfnCluster	AWS AWS Batch	Google Dataflow	Azure Batch
Python GENESIS R	Python GENESIS R	Python GENESIS R	Python GENESIS R	Python GENESIS R
current			future	

Calling GENESIS in other applications