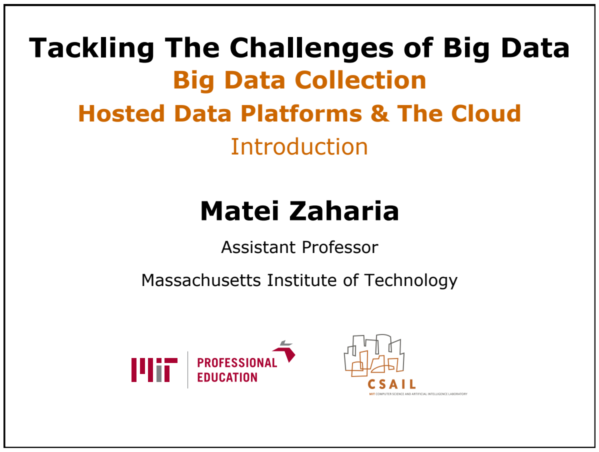
Tackling The Challenges of Big Data

Big Data Collection

Matei Zaharia

Assistant Professor

Massachusetts Institute of Technology



Tackling The Challenges of Big Data

Big Data Collection

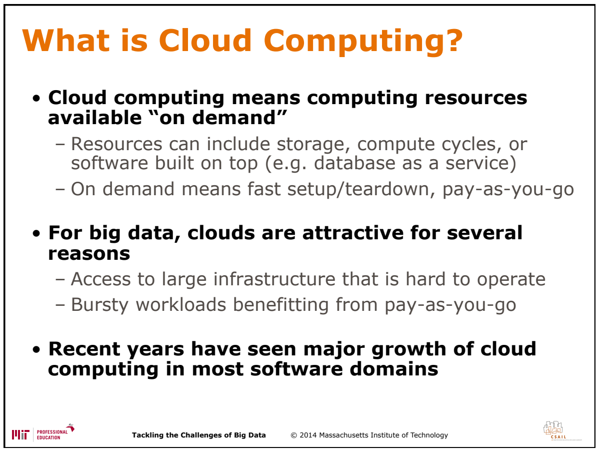
Hosted Data Platforms & The Cloud

Introduction

Matei Zaharia

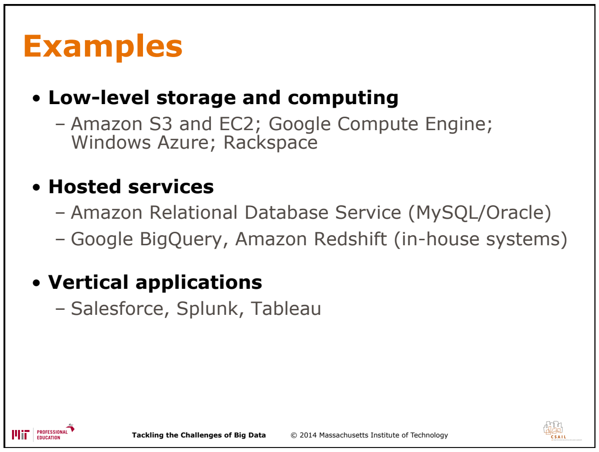
Assistant Professor

Massachusetts Institute of Technology



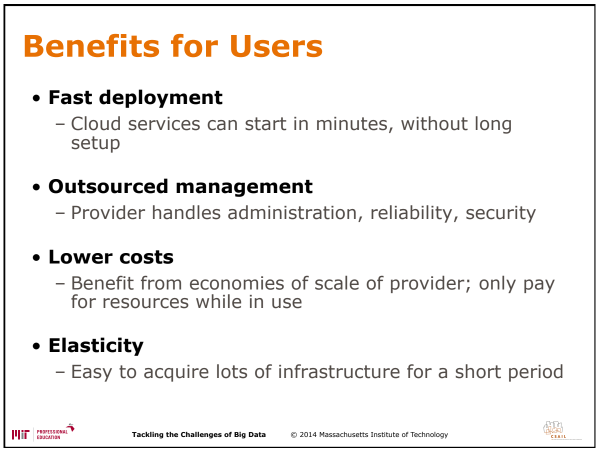
What is Cloud Computing?

- **Cloud computing means computing resources available "on demand"**
  - – Resources can include storage, compute cycles, or software built on top (e.g. database as a service)
  - – On demand means fast setup/teardown, pay-as-you-go
- **For big data, clouds are attractive for several reasons**
  - – Access to large infrastructure that is hard to operate
  - – Bursty workloads benefitting from pay-as-you-go
- **Recent years have seen major growth of cloud computing in most software domains**



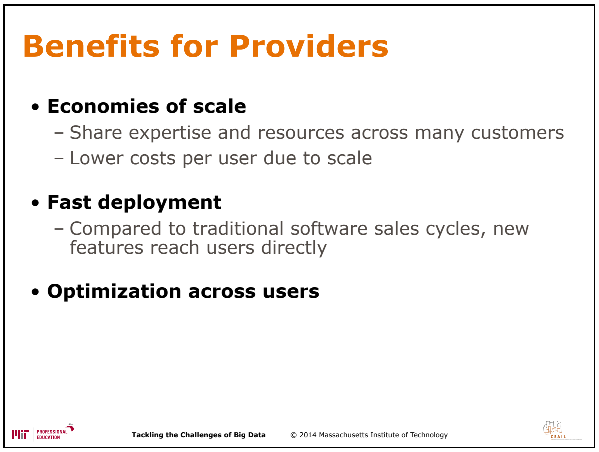
Examples

- **Low-level storage and computing**
  - Amazon S3 and EC2; Google Compute Engine; Windows Azure; Rackspace
- **Hosted services**
  - Amazon Relational Database Service (MySQL/Oracle)
  - Google BigQuery, Amazon Redshift (in-house systems)
- **Vertical applications**
  - Salesforce, Splunk, Tableau



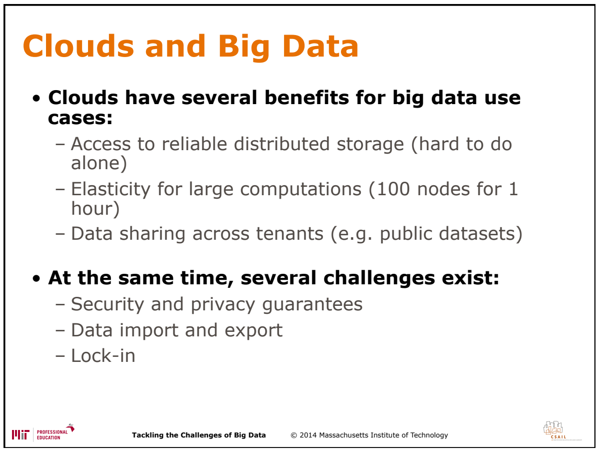
Benefits for Users

- **Fast deployment**
  - Cloud services can start in minutes, without long setup
- **Outsourced management**
  - Provider handles administration, reliability, security
- **Lower costs**
  - Benefit from economies of scale of provider; only pay for resources while in use
- **Elasticity**
  - Easy to acquire lots of infrastructure for a short period



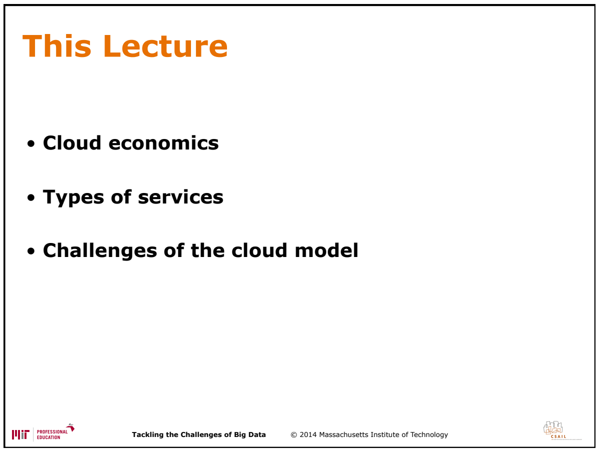
Benefits for Providers

- **Economies of scale**
  - Share expertise and resources across many customers
  - Lower costs per user due to scale
- **Fast deployment**
  - Compared to traditional software sales cycles, new features reach users directly
- **Optimization across users**



Clouds and Big Data

- **Clouds have several benefits for big data use cases:**
  - Access to reliable distributed storage (hard to do alone)
  - Elasticity for large computations (100 nodes for 1 hour)
  - Data sharing across tenants (e.g. public datasets)
- **At the same time, several challenges exist:**
  - Security and privacy guarantees
  - Data import and export
  - Lock-in



This Lecture

- **Cloud economics**
- **Types of services**
- **Challenges of the cloud model**

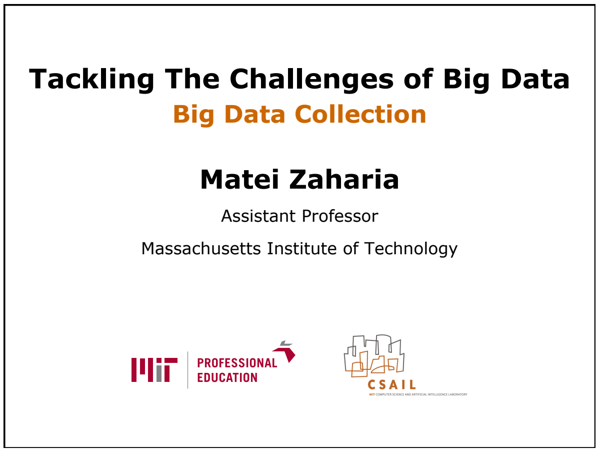
Tackling The Challenges of Big Data

Big Data Collection

Hosted Data Platforms & The Cloud

Introduction

THANK YOU



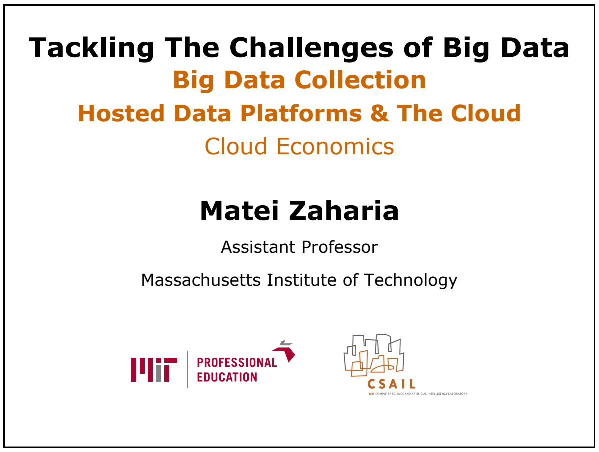
Tackling The Challenges of Big Data

Big Data Collection

Matei Zaharia

Assistant Professor

Massachusetts Institute of Technology



Tackling The Challenges of Big Data

Big Data Collection

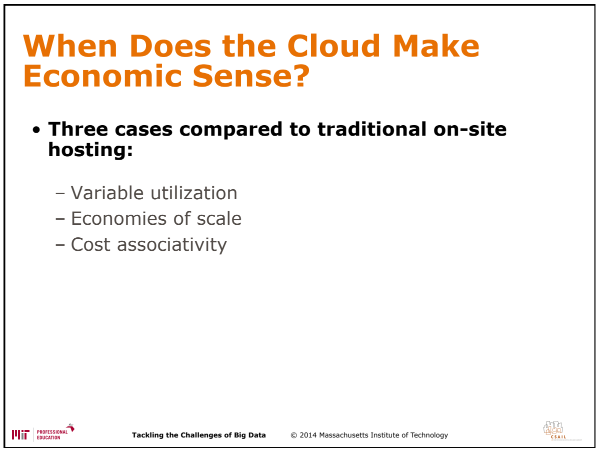
Hosted Data Platforms & The Cloud

Cloud Economics

Matei Zaharia

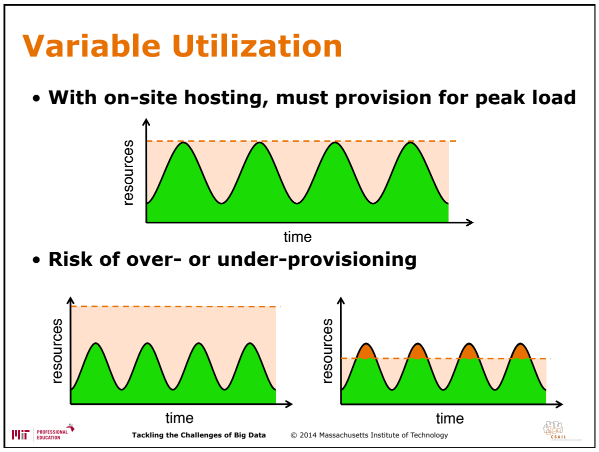
Assistant Professor

Massachusetts Institute of Technology



When Does the Cloud Make Economic Sense?

- **Three cases compared to traditional on-site hosting:**
  - Variable utilization
  - Economies of scale
  - Cost associativity



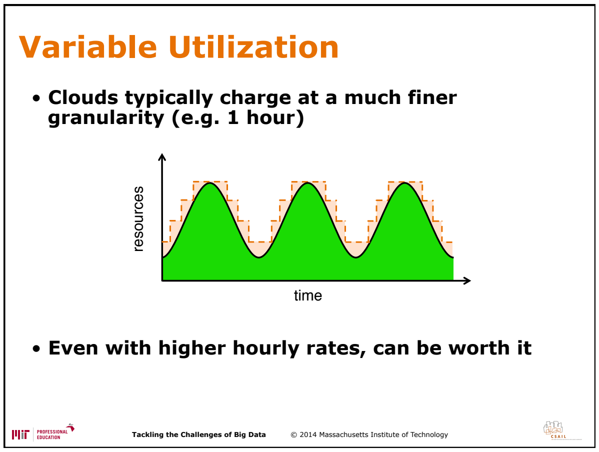
Variable Utilization

- With on-site hosting, must provision for peak load

resources
time

- Risk of over- or under-provisioning

resources
time
resources
time

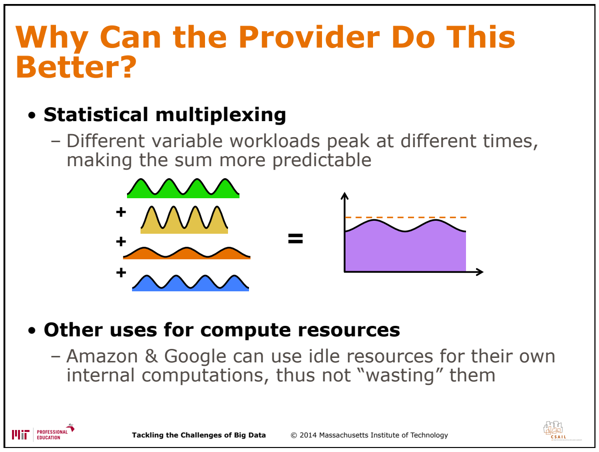


Variable Utilization

- Clouds typically charge at a much finer granularity (e.g. 1 hour)

resources
time

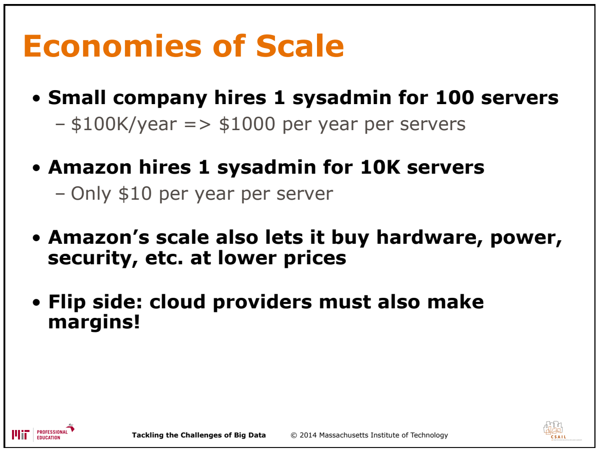
- Even with higher hourly rates, can be worth it



Why Can the Provider Do This Better?

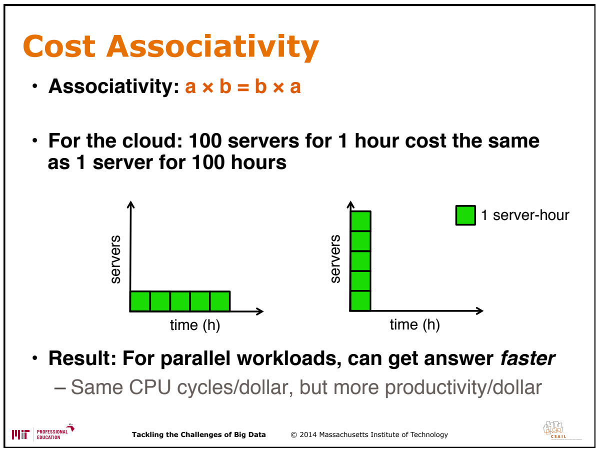
- Statistical multiplexing
  - Different variable workloads peak at different times, making the sum more predictable

- Other uses for compute resources
  - Amazon & Google can use idle resources for their own internal computations, thus not “wasting” them



Economies of Scale

- **Small company hires 1 sysadmin for 100 servers**
  - $100K/year => $1000 per year per servers
- **Amazon hires 1 sysadmin for 10K servers**
  - Only $10 per year per server
- **Amazon's scale also lets it buy hardware, power, security, etc. at lower prices**
- **Flip side: cloud providers must also make margins!**

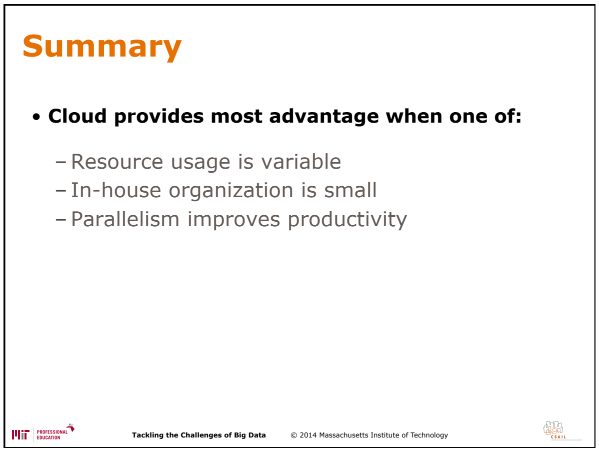


Cost Associativity

- **Associativity:** a × b = b × a
- **For the cloud: 100 servers for 1 hour cost the same as 1 server for 100 hours**

servers
time (h)
servers
time (h)
🟩 1 server-hour

- **Result: For parallel workloads, can get answer *faster***
  - Same CPU cycles/dollar, but more productivity/dollar



Summary

- **Cloud provides most advantage when one of:**
  - Resource usage is variable
  - In-house organization is small
  - Parallelism improves productivity

Tackling The Challenges of Big Data

Big Data Collection

Hosted Data Platforms & The Cloud

Cloud Economics

THANK YOU

PROFESSIONAL EDUCATION

CSAIL

Tackling The Challenges of Big Data

Big Data Collection

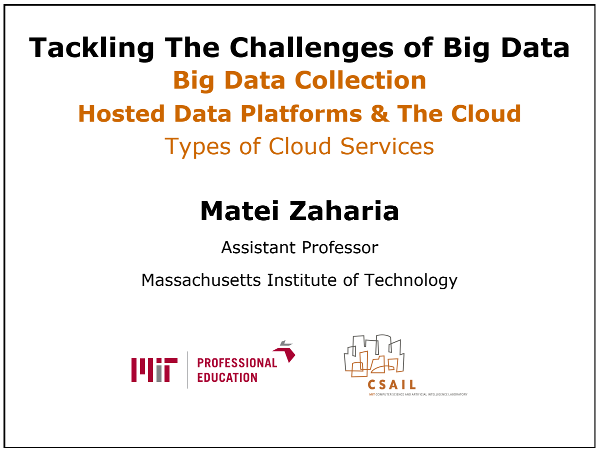
Matei Zaharia

Assistant Professor

Massachusetts Institute of Technology

PROFESSIONAL EDUCATION

CSAIL



Tackling The Challenges of Big Data

Big Data Collection

Hosted Data Platforms & The Cloud

Types of Cloud Services

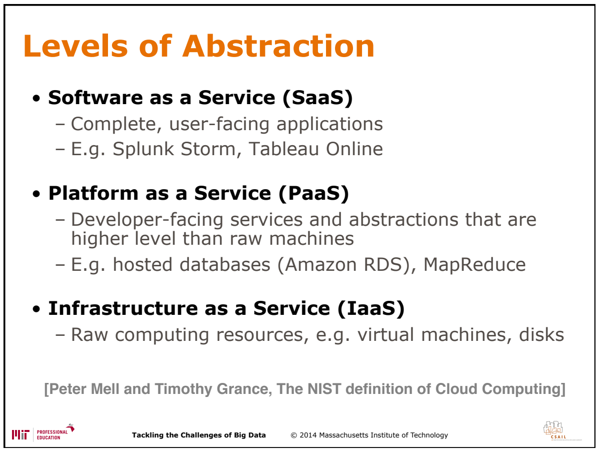
Matei Zaharia

Assistant Professor

Massachusetts Institute of Technology

PROFESSIONAL EDUCATION

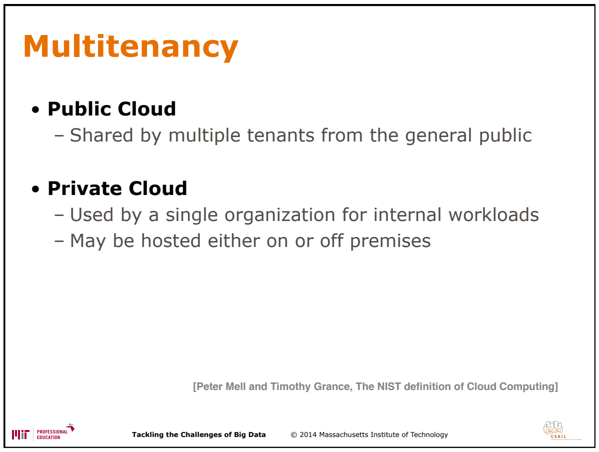
CSAIL



Levels of Abstraction

- **Software as a Service (SaaS)**
  - Complete, user-facing applications
  - E.g. Splunk Storm, Tableau Online
- **Platform as a Service (PaaS)**
  - Developer-facing services and abstractions that are higher level than raw machines
  - E.g. hosted databases (Amazon RDS), MapReduce
- **Infrastructure as a Service (IaaS)**
  - Raw computing resources, e.g. virtual machines, disks

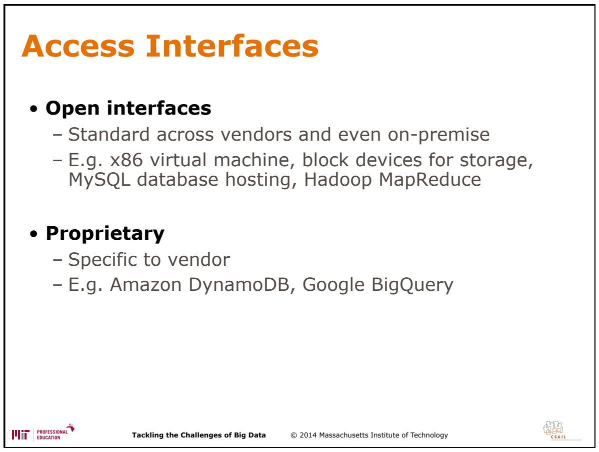
[Peter Mell and Timothy Grance, The NIST definition of Cloud Computing]



Multitenancy

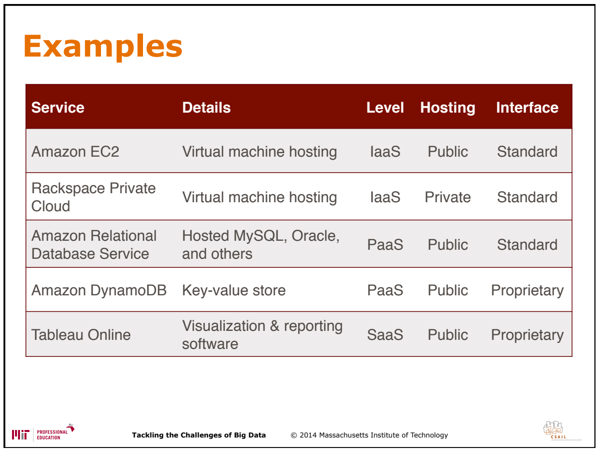
- **Public Cloud**
  - Shared by multiple tenants from the general public
- **Private Cloud**
  - Used by a single organization for internal workloads
  - May be hosted either on or off premises

[Peter Mell and Timothy Grance, The NIST definition of Cloud Computing]



Access Interfaces

- **Open interfaces**
  - Standard across vendors and even on-premise
  - E.g. x86 virtual machine, block devices for storage, MySQL database hosting, Hadoop MapReduce
- **Proprietary**
  - Specific to vendor
  - E.g. Amazon DynamoDB, Google BigQuery



Examples

Service	Details	Level	Hosting	Interface
Amazon EC2	Virtual machine hosting	IaaS	Public	Standard
Rackspace Private Cloud	Virtual machine hosting	IaaS	Private	Standard
Amazon Relational Database Service	Hosted MySQL, Oracle, and others	PaaS	Public	Standard
Amazon DynamoDB	Key-value store	PaaS	Public	Proprietary
Tableau Online	Visualization & reporting software	SaaS	Public	Proprietary

Tackling the Challenges of Big Data © 2014 Massachusetts Institute of Technology

Tackling The Challenges of Big Data

Big Data Collection

Hosted Data Platforms & The Cloud

Types of Cloud Services

THANK YOU

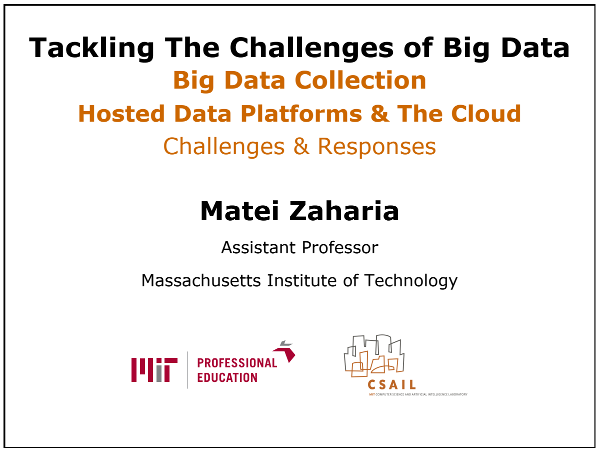
Tackling The Challenges of Big Data

Big Data Collection

Matei Zaharia

Assistant Professor

Massachusetts Institute of Technology



Tackling The Challenges of Big Data

Big Data Collection

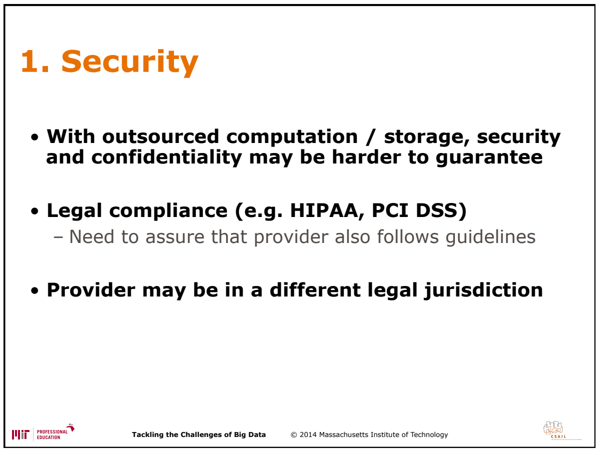
Hosted Data Platforms & The Cloud

Challenges & Responses

Matei Zaharia

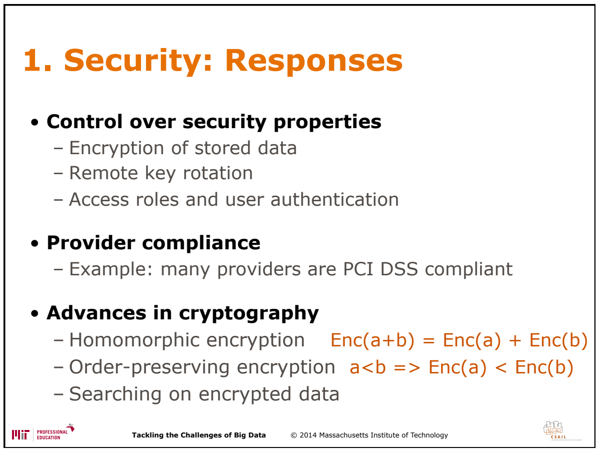
Assistant Professor

Massachusetts Institute of Technology



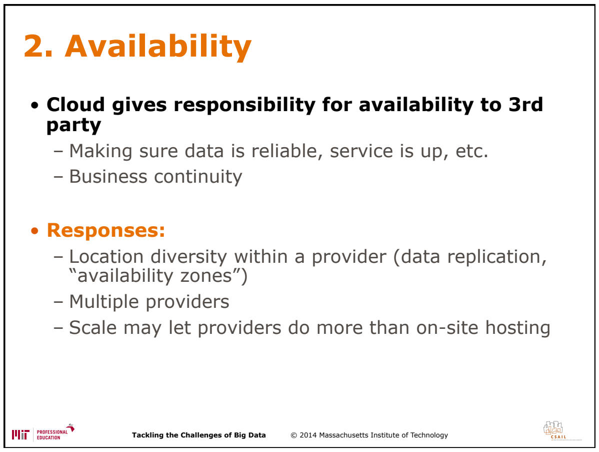
1. Security

- **With outsourced computation / storage, security and confidentiality may be harder to guarantee**
- **Legal compliance (e.g. HIPAA, PCI DSS)**
  - Need to assure that provider also follows guidelines
- **Provider may be in a different legal jurisdiction**



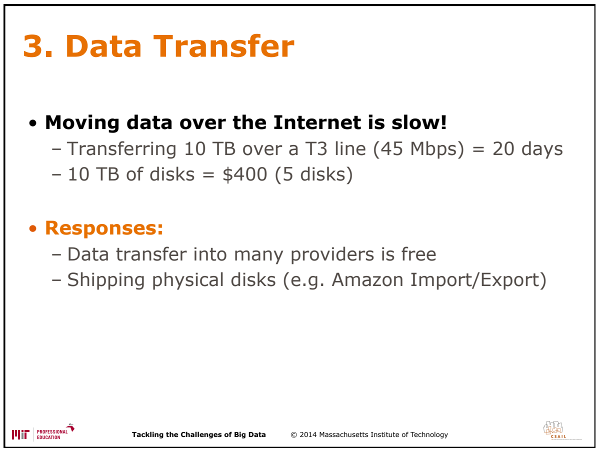
1. Security: Responses

- **Control over security properties**
  - Encryption of stored data
  - Remote key rotation
  - Access roles and user authentication
- **Provider compliance**
  - Example: many providers are PCI DSS compliant
- **Advances in cryptography**
  - Homomorphic encryption Enc(a+b) = Enc(a) + Enc(b)
  - Order-preserving encryption a<b => Enc(a) < Enc(b)
  - Searching on encrypted data



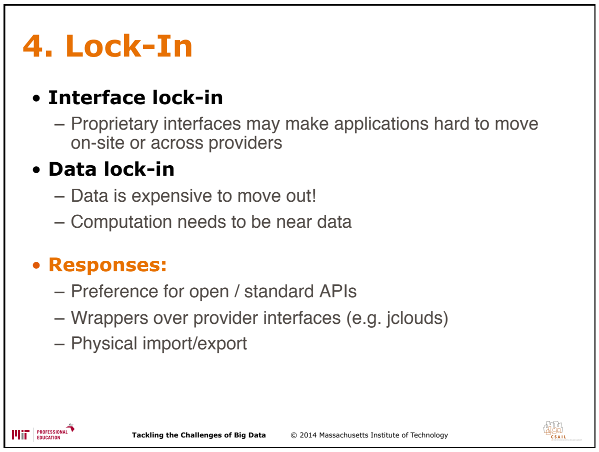
2. Availability

- **Cloud gives responsibility for availability to 3rd party**
  - Making sure data is reliable, service is up, etc.
  - Business continuity
- **Responses:**
  - Location diversity within a provider (data replication, “availability zones”)
  - Multiple providers
  - Scale may let providers do more than on-site hosting



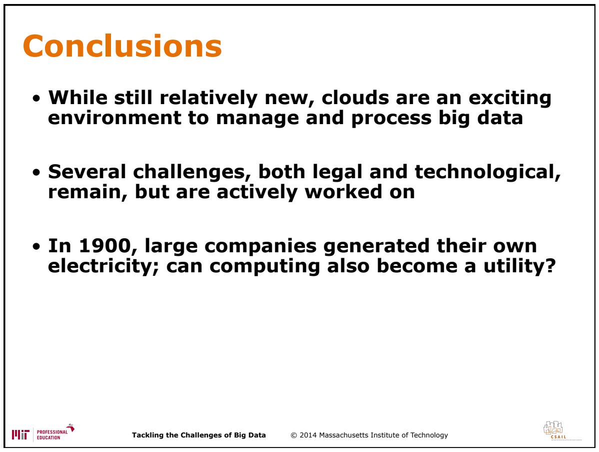
3. Data Transfer

- **Moving data over the Internet is slow!**
  - Transferring 10 TB over a T3 line (45 Mbps) = 20 days
  - 10 TB of disks = $400 (5 disks)
- **Responses:**
  - Data transfer into many providers is free
  - Shipping physical disks (e.g. Amazon Import/Export)



4. Lock-In

- **Interface lock-in**
  - Proprietary interfaces may make applications hard to move on-site or across providers
- **Data lock-in**
  - Data is expensive to move out!
  - Computation needs to be near data
- **Responses:**
  - Preference for open / standard APIs
  - Wrappers over provider interfaces (e.g. jclouds)
  - Physical import/export



Conclusions

- While still relatively new, clouds are an exciting environment to manage and process big data
- Several challenges, both legal and technological, remain, but are actively worked on
- In 1900, large companies generated their own electricity; can computing also become a utility?

Tackling the Challenges of Big Data © 2014 Massachusetts Institute of Technology

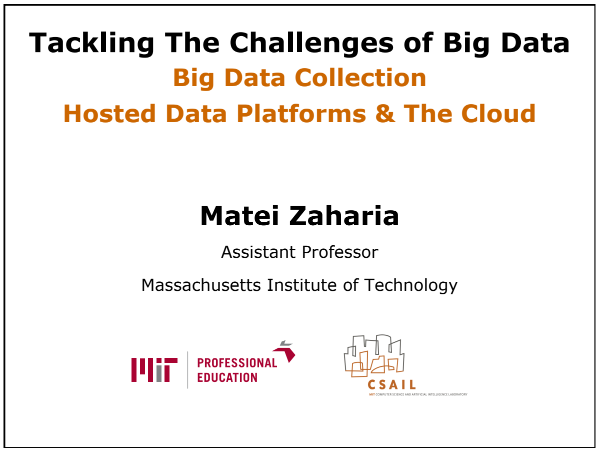
Tackling The Challenges of Big Data

Big Data Collection

Hosted Data Platforms & The Cloud

Challenges & Responses

THANK YOU



Tackling The Challenges of Big Data

Big Data Collection

Hosted Data Platforms & The Cloud

Matei Zaharia

Assistant Professor

Massachusetts Institute of Technology