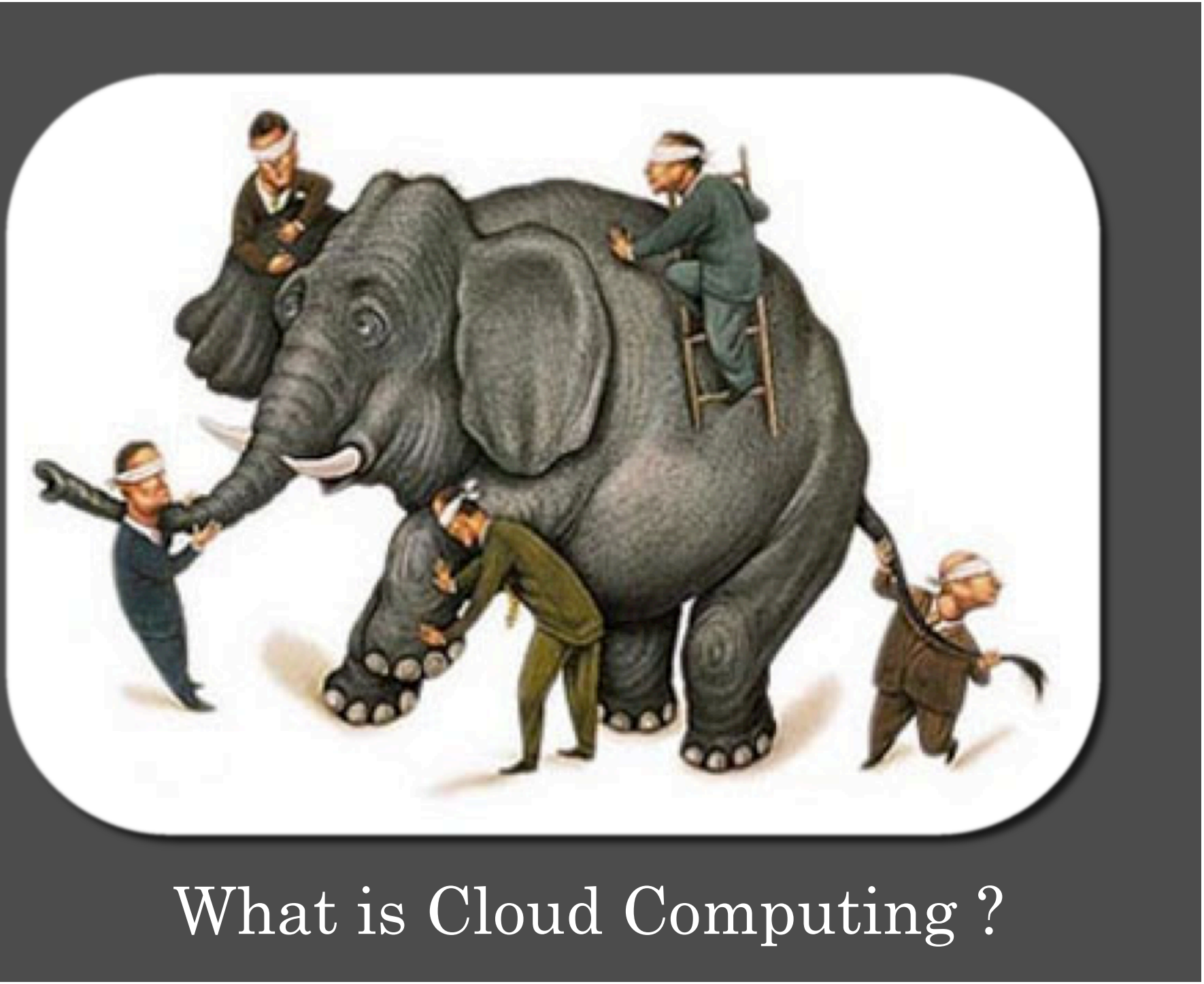
Elastic-R: SaaS for Scientific Computing

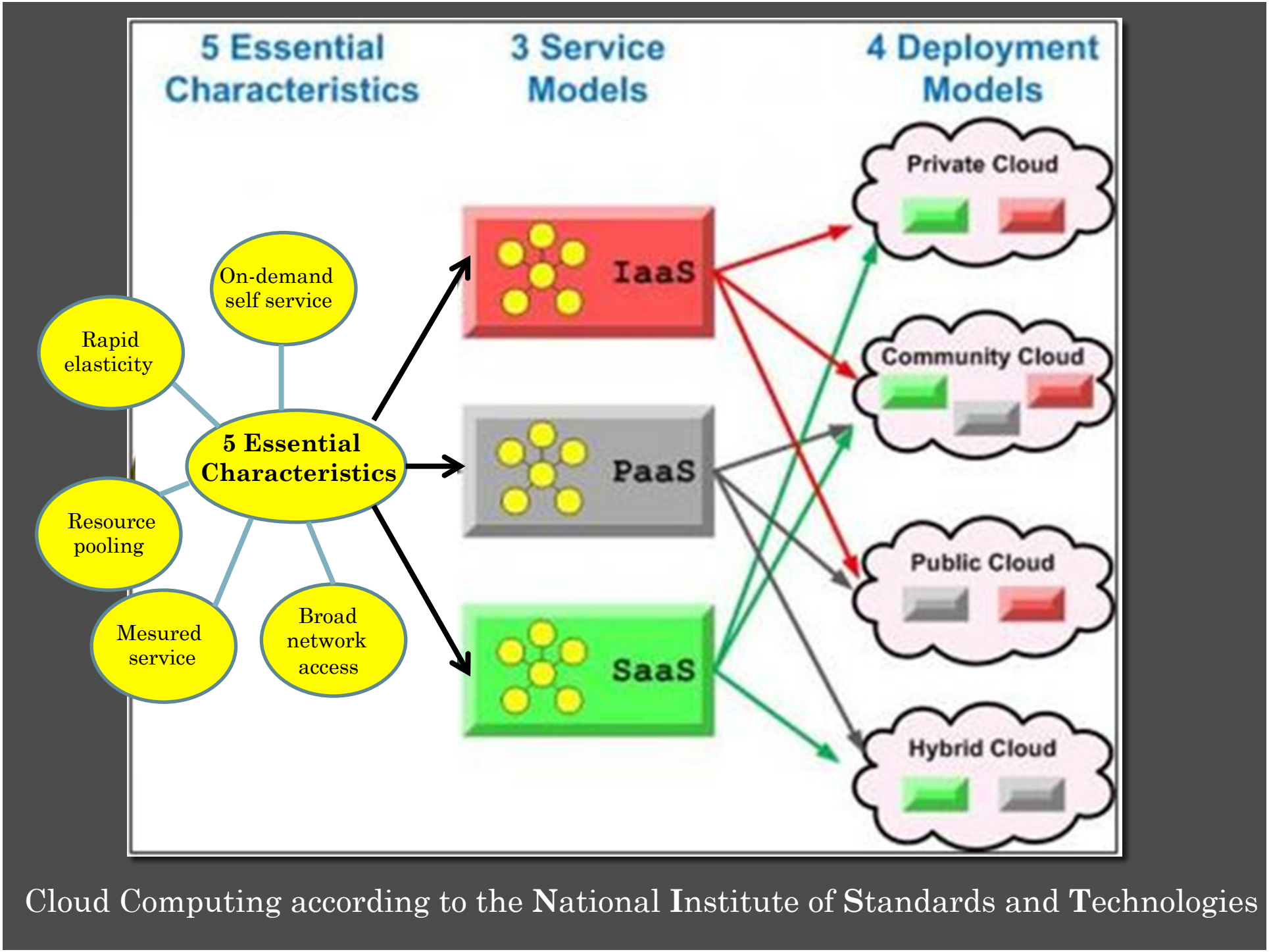
Fourth DAAD Summer School
on Current Trends in Distributed Systems

Sousse, 6 September 2012

Karim Chine
karim.chine@cloudera.co.uk



What is Cloud Computing ?



Cloud Computing according to the National Institute of Standards and Technologies

Cloud = public IaaS = Amazon Web Services

Firefox | Windows Azure Offerings Are Expand... | Privacy & Data Practices | AddThis

cloudcomputing.sys-con.com/node/2292662

Most Visited | Getting Started | Latest Headlines | elasticfox.xpi | Bookmarks

Welcome! Register | Sign-in Search Jump to a Magazine

CLOUD COMPUTING JOURNAL

Home | Subscribe | Advertise | Authors | Topics | Videos | Events | Webcasts

Cloud Expo: Email Alerts | Newsletters | Get Cloud Expo: Homepage | Mobile

.NET | AJAX | CLOUD | ECLIPSE | FLEX | OPEN WEB | IPHONE | JAVA | LINUX | OPEN SOURCE | ORACLE | POWERBUILDER | SECURITY | SOA | VIRTUALIZATION | WEB 2.0

BIG DATA EXPO
NEW YORK CITY • JUNE 10-13, 2013

Cloud Expo Authors: Steve Weisfeldt, Liz McMillan, Ajay Budhraja, Jnan Dash, Scott Kinka
Related Topics: Cloud Expo, SOA & WOA, .NET, Virtualization

Cloud Expo: Blog Post

Windows Azure Offerings Are Expanded with New IaaS Offerings

New Windows Azure Virtual Machine and Virtual Networking bolster Hybrid scenarios for your Private Cloud

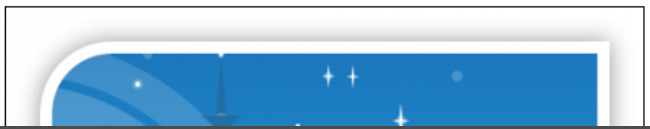
BY KEITH MAYER ARTICLE RATING: ☆☆☆☆☆

JUNE 9, 2012 11:00 AM EDT READS: 1,460

RELATED | PRINT | EMAIL | FEEDBACK | ADD THIS | BLOG THIS

On June 7th, Microsoft announced several new infrastructure-as-a-service offerings within the Windows Azure cloud that are now available as a customer preview: Windows Azure Virtual Machines, Virtual Network, Web Sites and Media Services. A full release of these features is anticipated later this calendar year.

In particular, the new Windows Azure Virtual Machine and Virtual Network capabilities offer increased flexibility in planning private cloud deployments through the introduction of "hybrid" scenarios that leverage both private and public cloud resources. Private Clouds, typically associated with on-premises managed fabrics of virtual resource pools, can be extended into the Windows Azure cloud securely and cost-effectively using these new features.



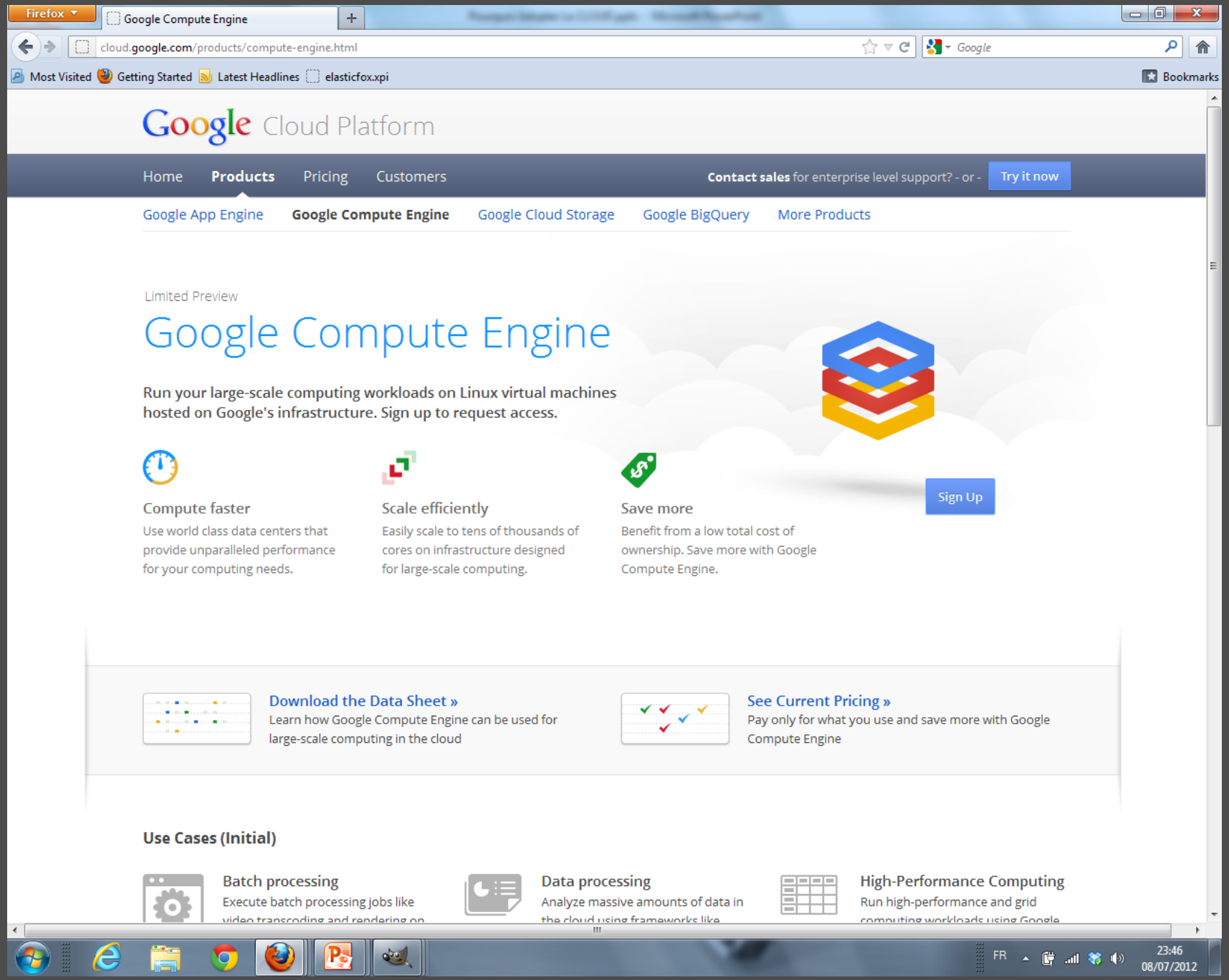
Comments

Perhaps I Haven't Made Myself Clear...

By Bruce Armstrong

bruce.armstrong wrote: Somebody just said it better than I did, and with more chops to say it: Open Letter to Mark Zuckerberg, Sheryl Sandberg & Facebook Mobile

May. 8, 2012 02:11 PM


Firefox
Google Compute Engine
cloud.google.com/products/compute-engine.html
Most Visited
Getting Started
Latest Headlines
elasticfox.xpi
Bookmarks
Google Cloud Platform
Home
Products
Pricing
Customers
Contact sales for enterprise level support? - or -
Try it now
Google App Engine
Google Compute Engine
Google Cloud Storage
Google BigQuery
More Products
Limited Preview
Google Compute Engine
Run your large-scale computing workloads on Linux virtual machines hosted on Google's infrastructure. Sign up to request access.
Sign Up
Compute faster
Use world class data centers that provide unparalleled performance for your computing needs.
Scale efficiently
Easily scale to tens of thousands of cores on infrastructure designed for large-scale computing.
Save more
Benefit from a low total cost of ownership. Save more with Google Compute Engine.
Download the Data Sheet »
Learn how Google Compute Engine can be used for large-scale computing in the cloud
See Current Pricing »
Pay only for what you use and save more with Google Compute Engine
Use Cases (Initial)
Batch processing
Execute batch processing jobs like
Data processing
Analyze massive amounts of data in
High-Performance Computing
Run high-performance and grid
FR
23:46
08/07/2012

Firefox | NASA Embraces Amazon Cloud, Leaves ...

www.wired.com/wiredenterprise/2012/06/nasa-web-services-openstack/

Most Visited | Getting Started | Latest Headlines | elasticfox.xpi | Bookmarks

ADVERTISEMENT

PSST! HEY WIRED READERS!

LOOKING FOR A TECH JOB? SEARCH WIRED JOBS TODAY.

WIRED

SEARCH JOBS

subscribe to WIRED

PRINT AND DIGITAL ACCESS

SUBSCRIBE | RENEW | GIVE A GIFT | INTERNATIONAL

WIRED ENTERPRISE

IT HAPPENS

PREVIOUS POST | NEXT POST

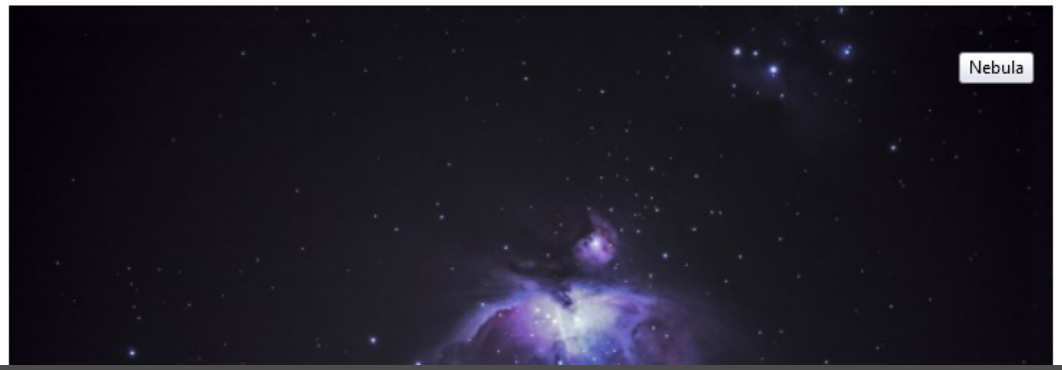
NASA Embraces Amazon Cloud, Leaves OpenStack Behind

322 Tweet | 21 +1 | 76 Share

By Caleb Garling | June 12, 2012 | 7:55 pm | Categories: Announcements, Government, Infrastructure as a Service

Follow @calebgarling

Like | Send | 46 people like this. Be the first of your friends.



SUBSCRIBE TO WIRED MAGAZINE

subscribe to WIRED

PRINT AND DIGITAL ACCESS

- Subscribe to WIRED
- Renew
- Give a gift
- International Orders

EDITORIAL TEAM

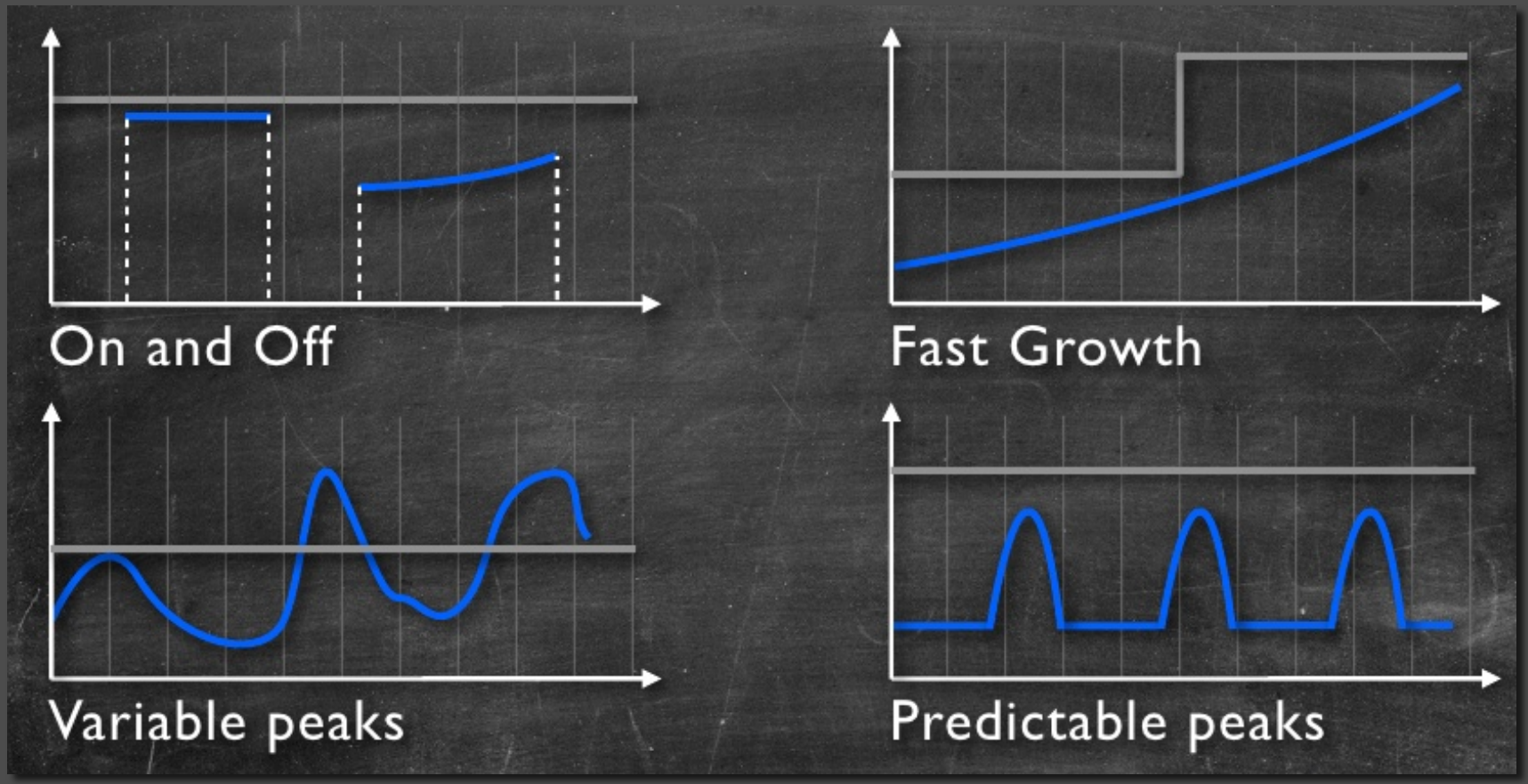
Editor: Cade Metz | E-mail | Twitter

Staff Writer: Caleb Garling | E-mail | Twitter

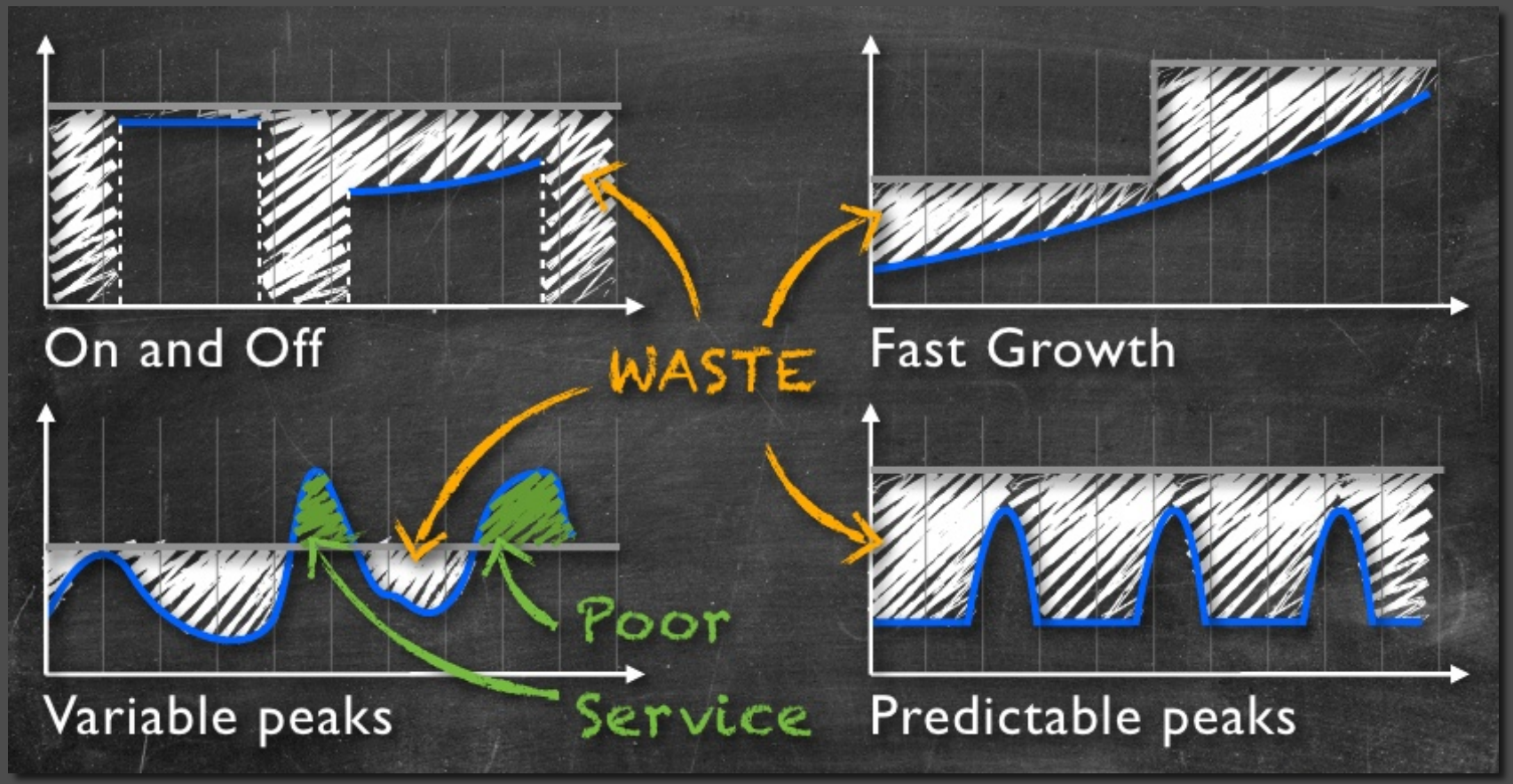
Staff Writer: Robert McMillan | E-mail | Twitter

Why adopt **PUBLIC IaaS ?**

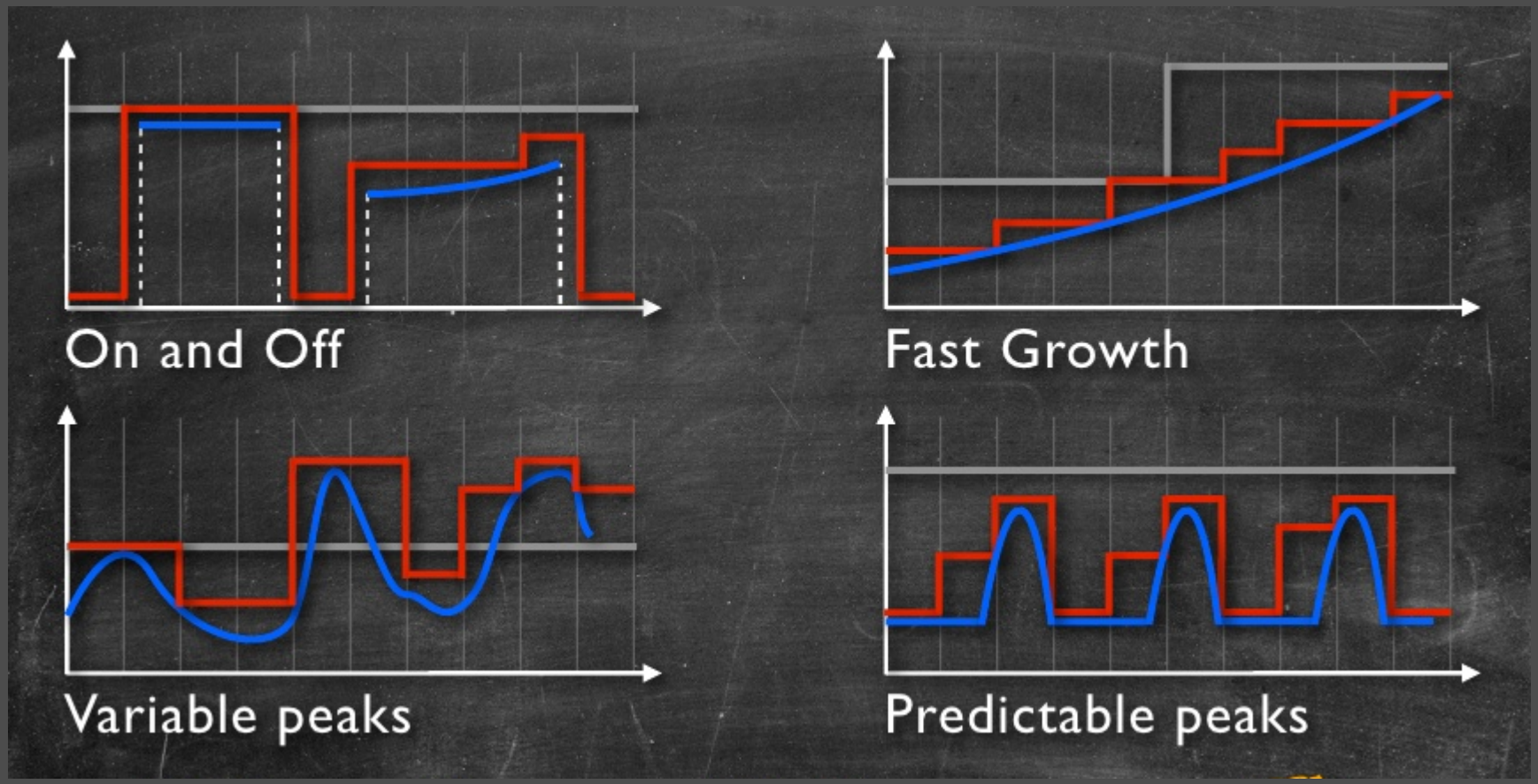
1.Elastic Capacity,
cost effectiveness



Infrastructure Usage Patterns



Classic IT Infratructures



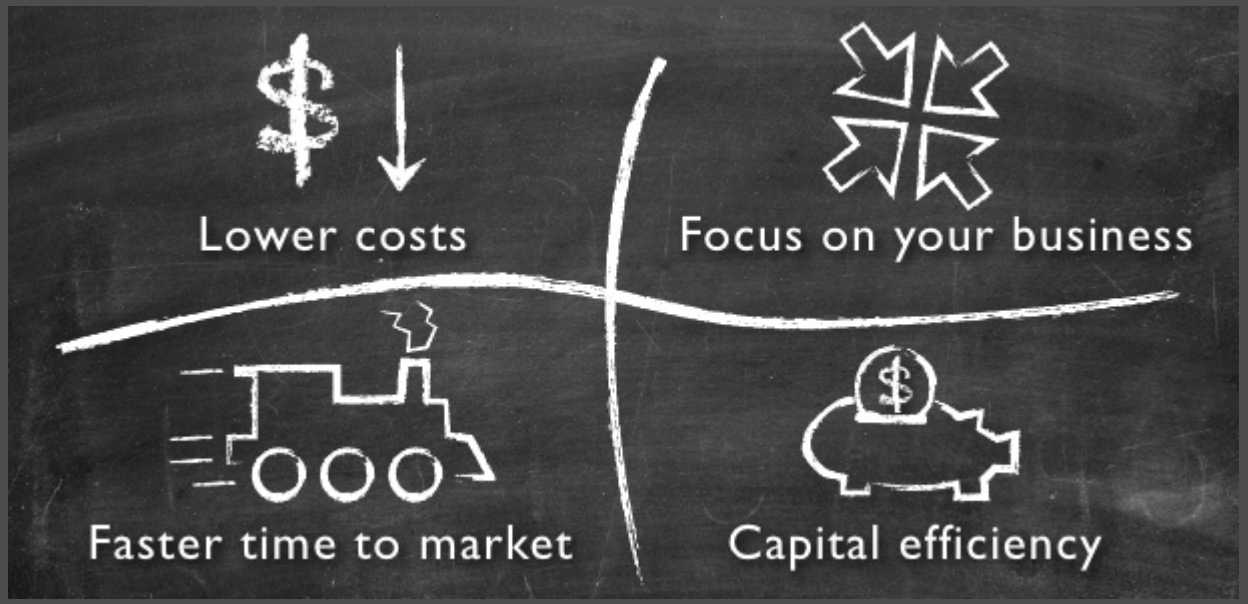
Cloud infrastructures

2. **Technological Agility**

Rapid and easy services deployment

3. **Business Agility**

No Capex, no initial investments



4. **No Committment**, pay-per-use model

5. Integral automation,

Universal « LEGO » of IT/Applications reusable components

6. Maximal Security

Security expertise/resourcs mutualisation/critical mass

7. mySuperComputer,

A Super-computer for everyone

8. **myExperimentationLab**

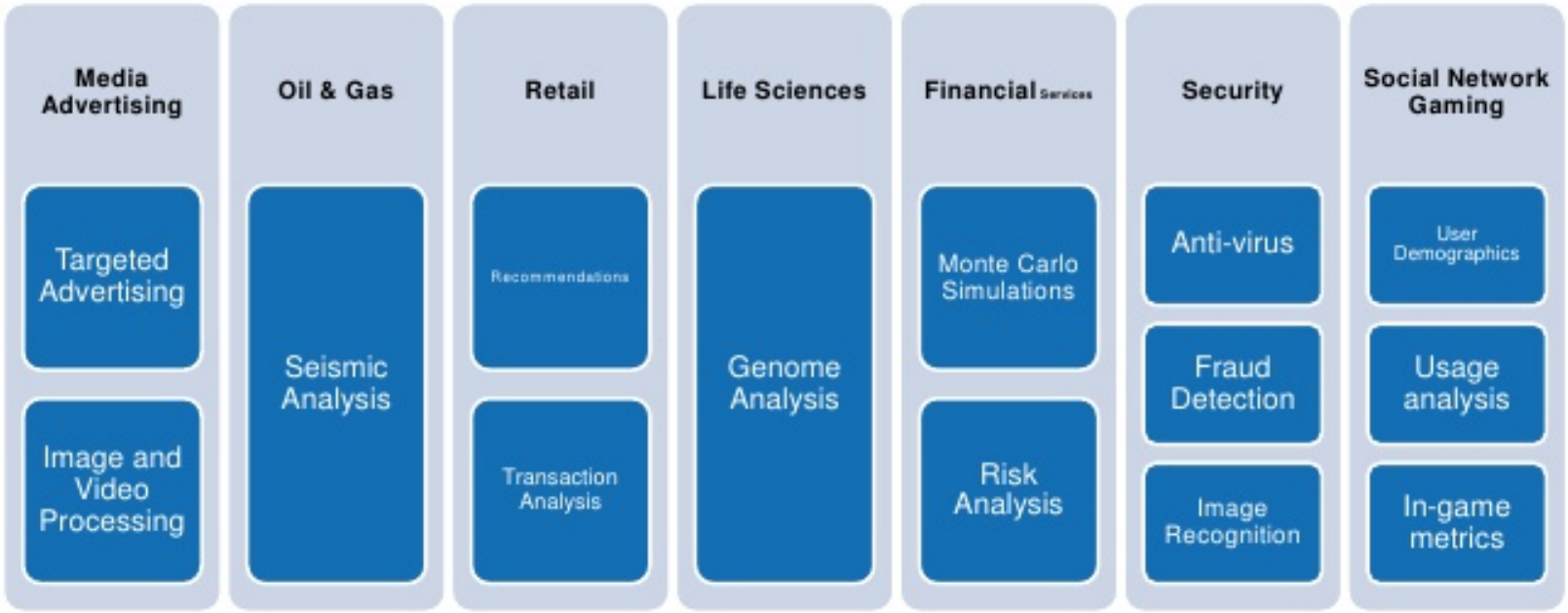
Fail quickly, fail cheaply, fail often!!

Want to increase innovation?
Lower the cost of failure.

- Joi Ito

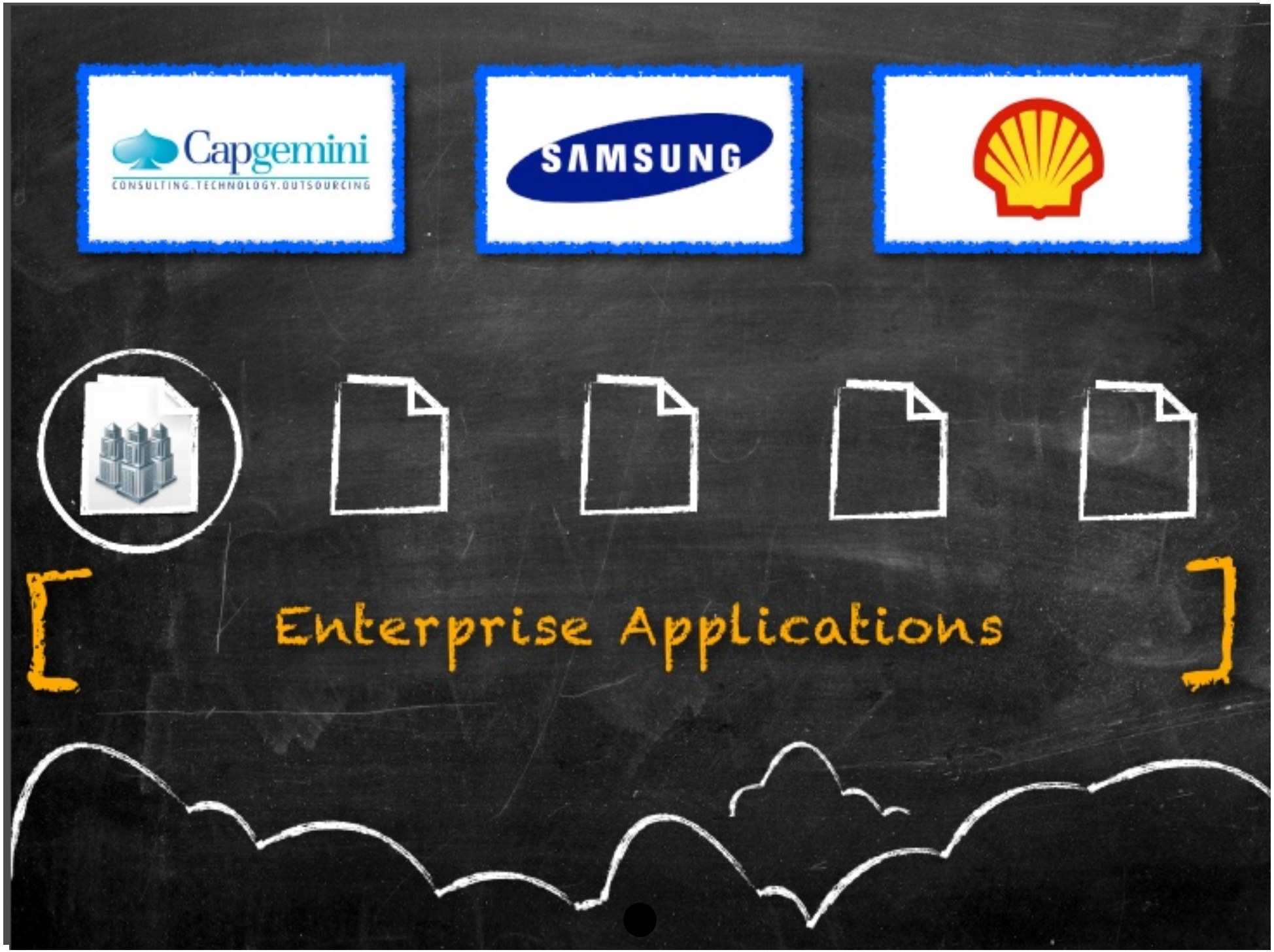
9. myBigData,
Big Data for everyone

Big Data Verticals



To be in the cloud or not to be ?

(http://www.slideshare.net/AmazonWebServices/aws-101-cloud-computing-seminar-2012)


Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING
SAMSUNG
Enterprise Applications

NETFLIX
ANIMOTO
PBS
www
Media and Web Applications

_ 26 Million subscribers

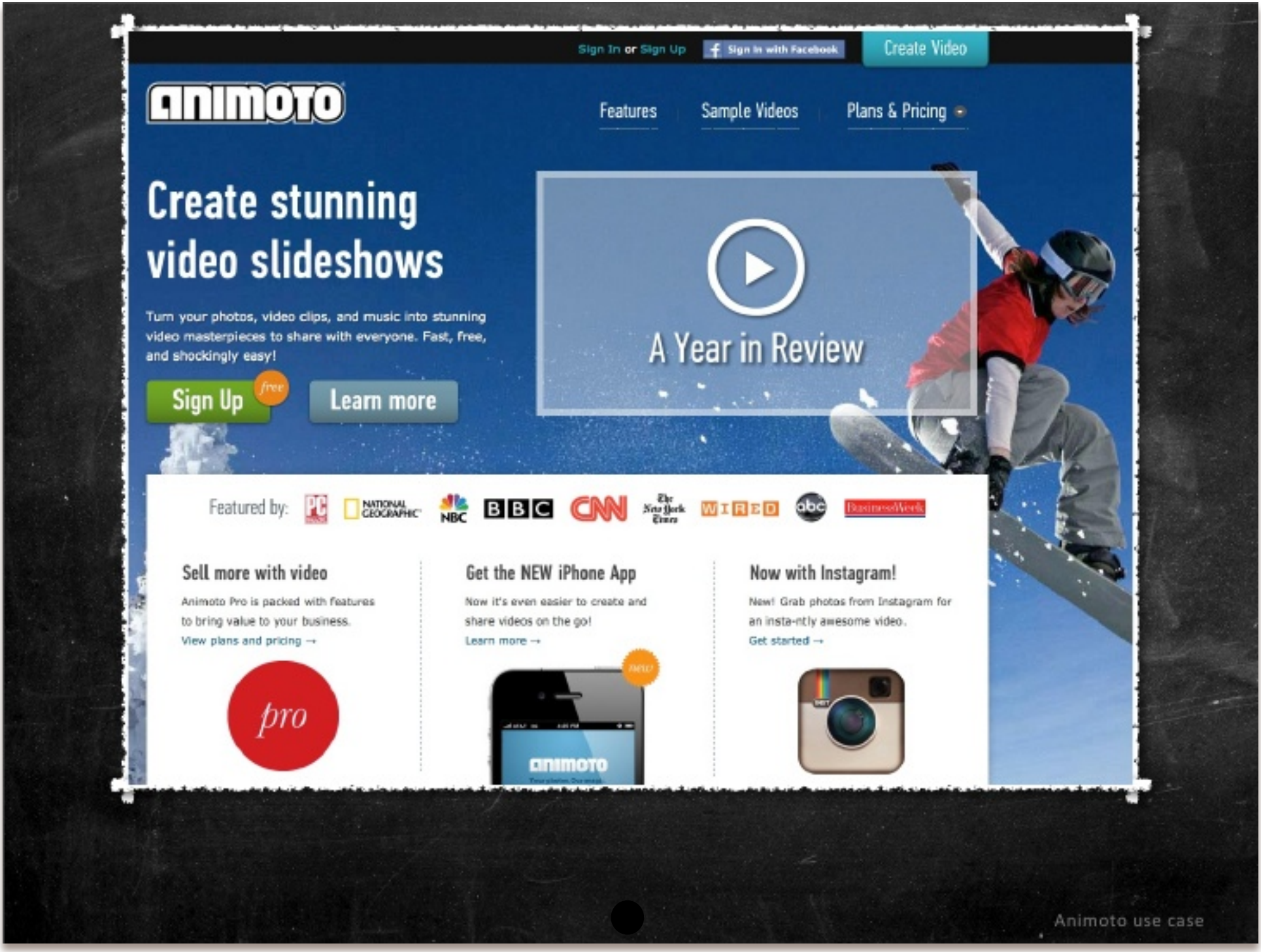
_ Netflix is ~25% of US Internet traffic

_ They are ~100% on AWS

_ Adrian Cockcroft, Cloud Architect:

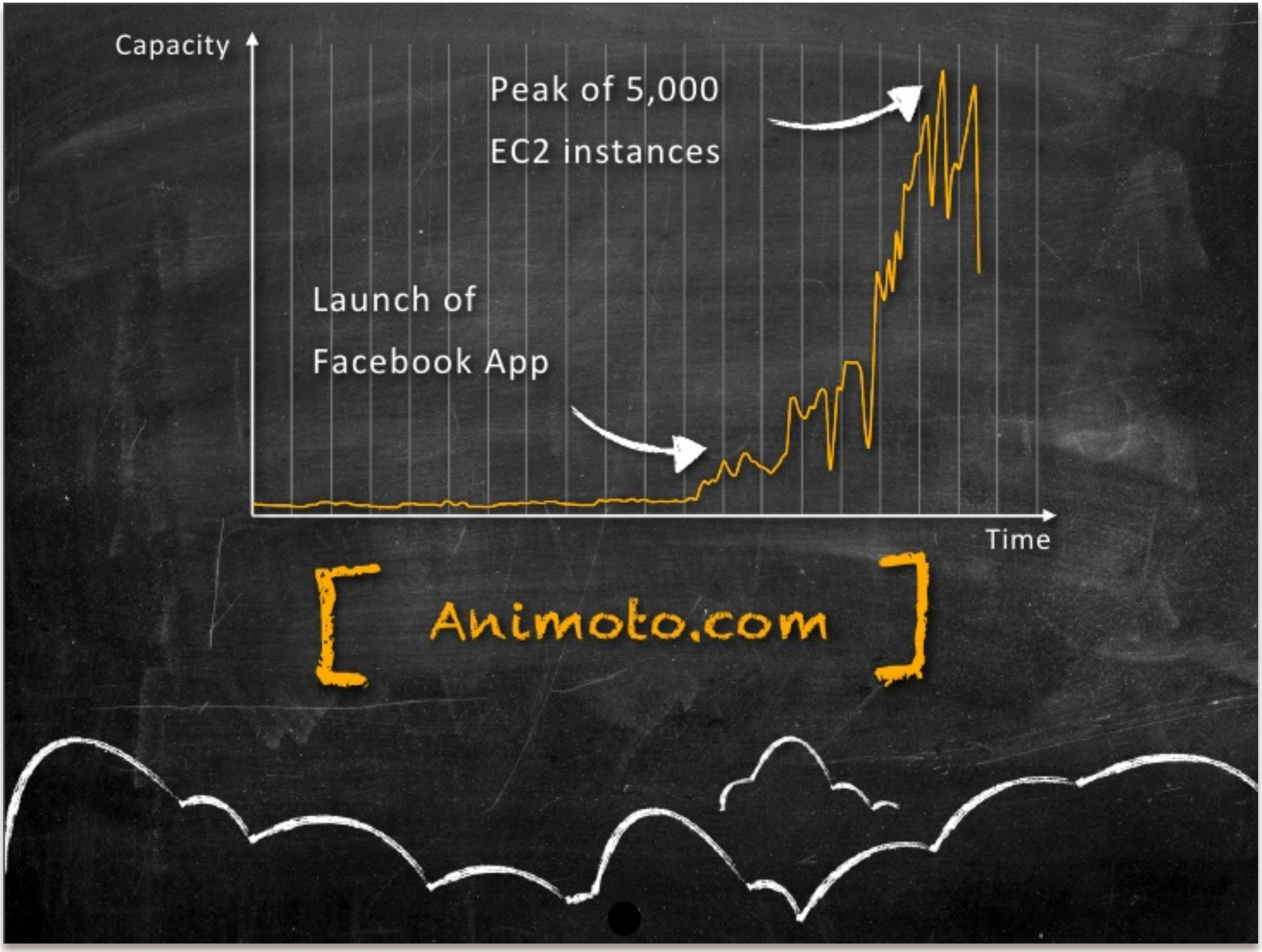
http://www.slideshare.net/adrianco

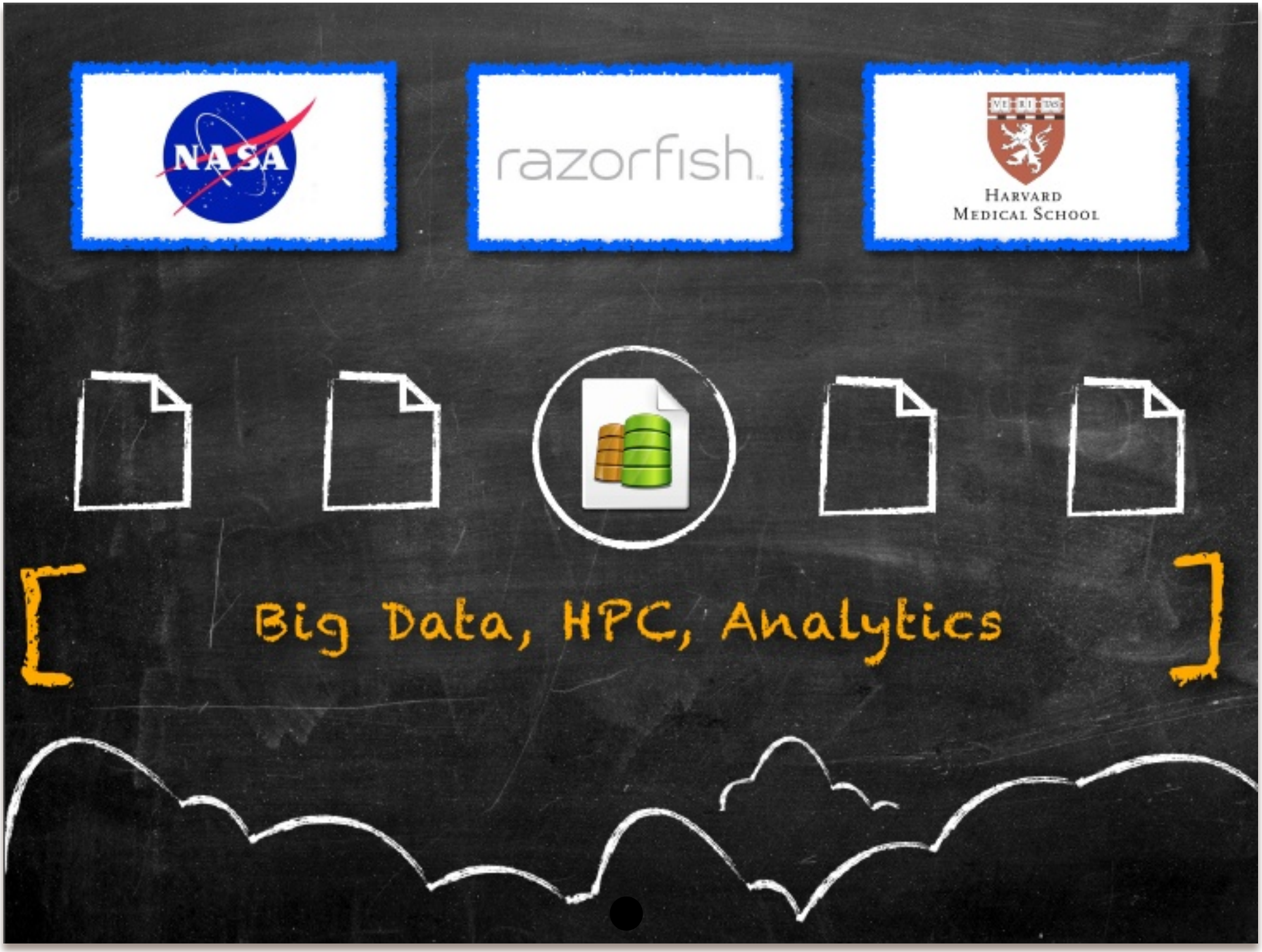
[Netflix.com]

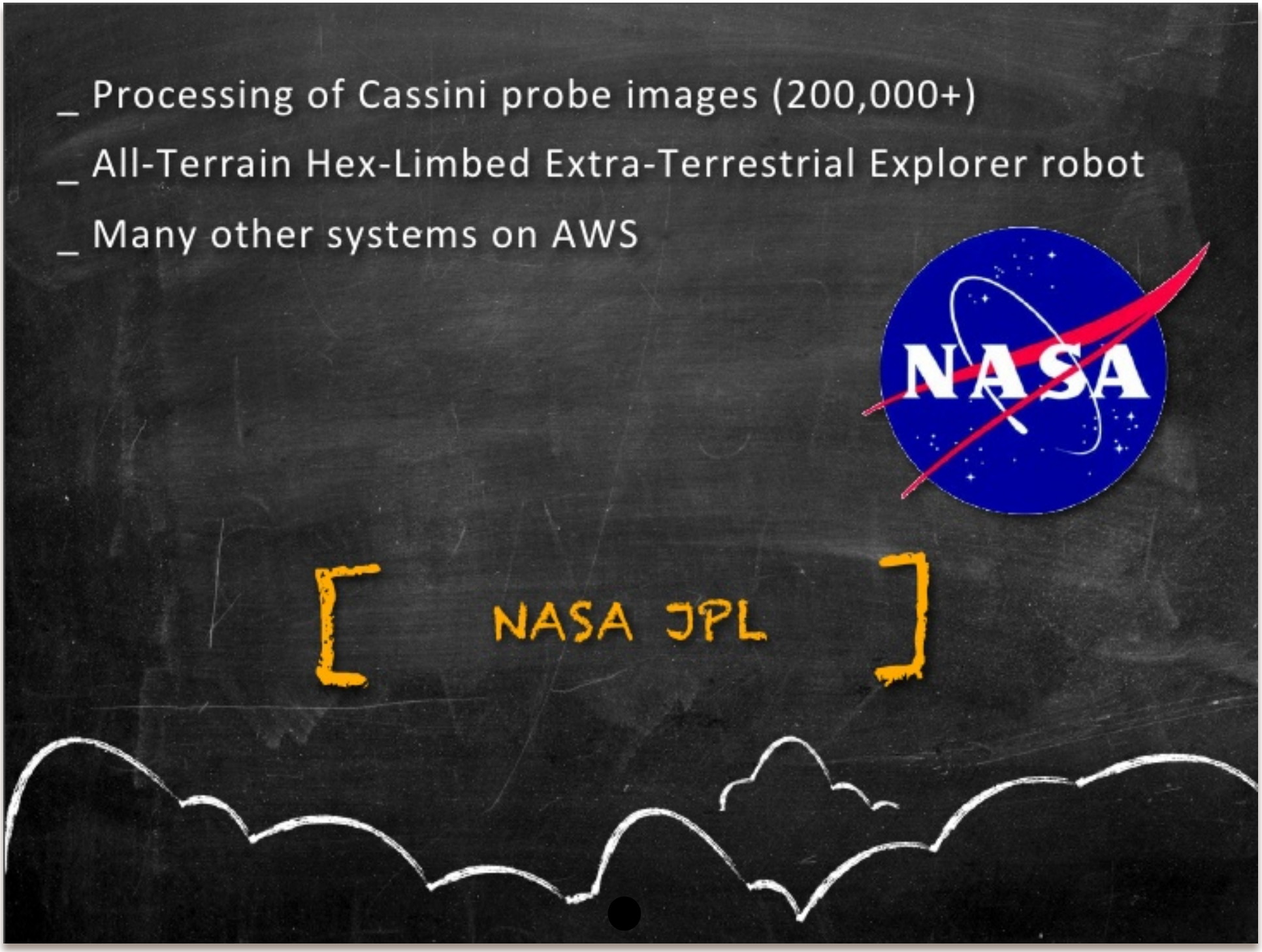


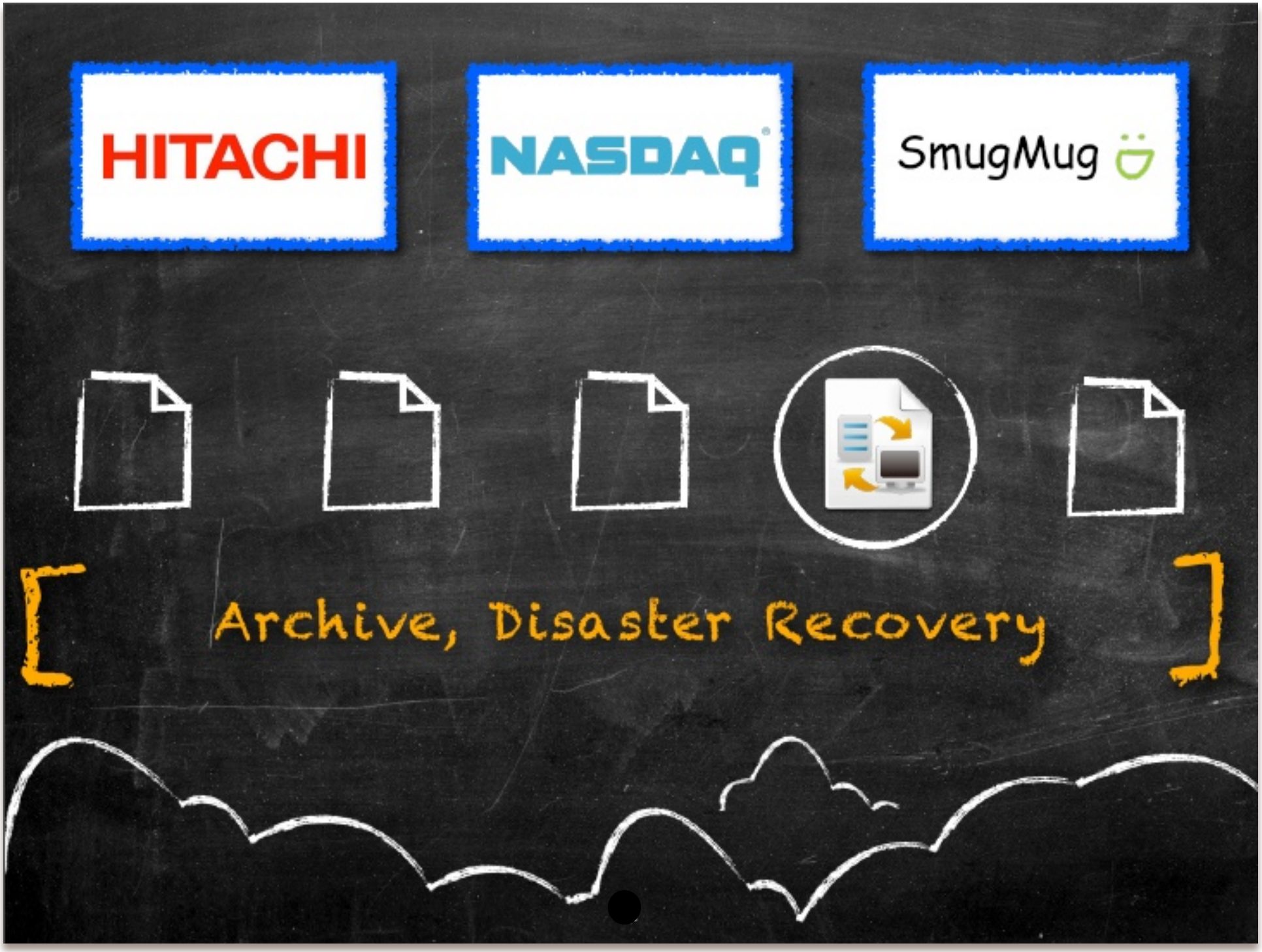
Animoto use case

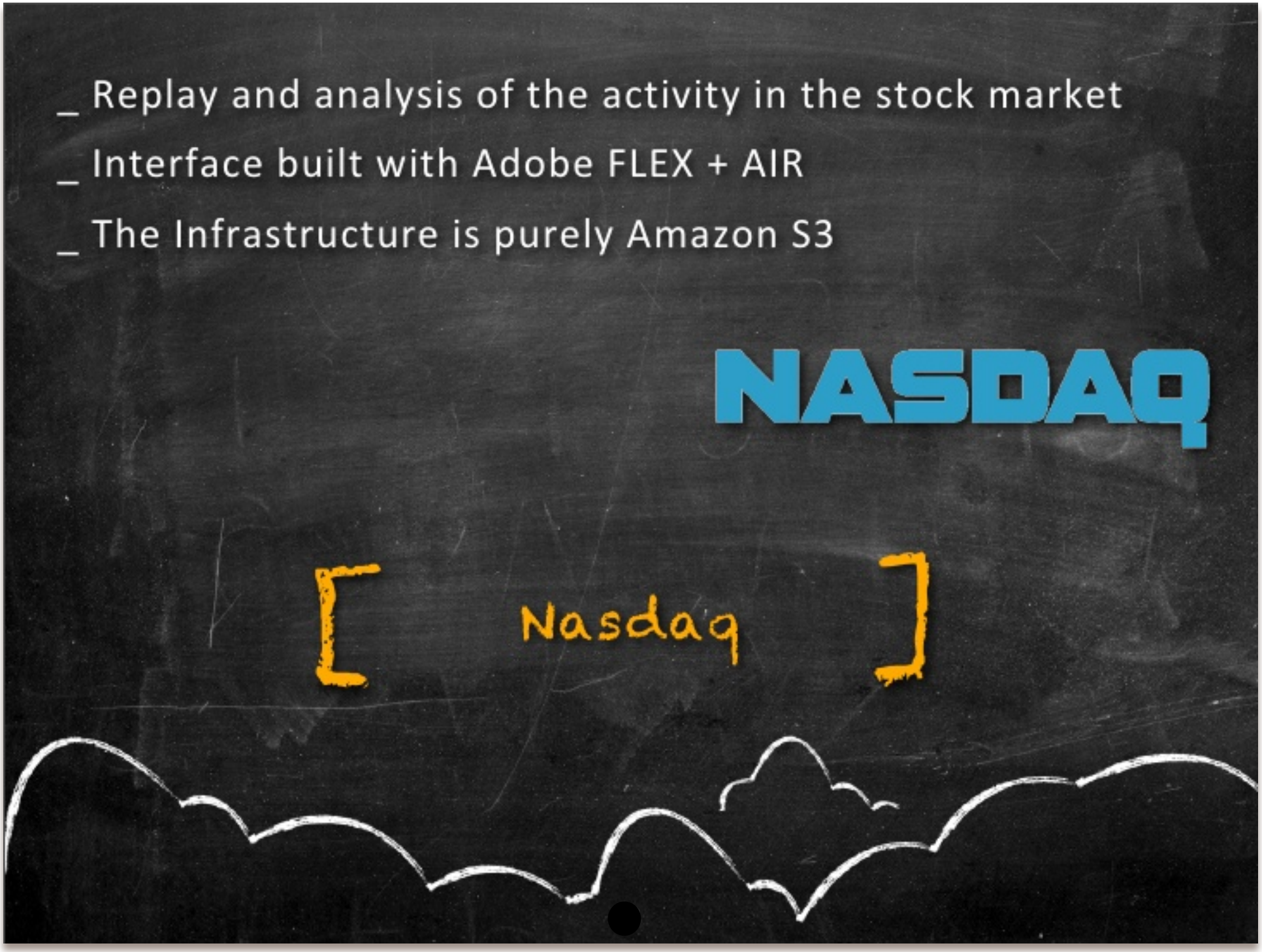
_ Simple way to create videos with your pictures
_ Facebook App: From 50 to 5,000 servers in 3 days
_ Later optimization with Cluster GPU Instances (EC2)
_ Freemium model
animoto
[Animoto.com]


Capacity
Peak of 5,000
EC2 instances
Launch of
Facebook App
Time
Animoto.com


NASA
razorfish
Harvard
Medical School
Big Data, HPC, Analytics


_ Processing of Cassini probe images (200,000+)
_ All-Terrain Hex-Limbed Extra-Terrestrial Explorer robot
_ Many other systems on AWS
NASA
[NASA JPL]


HITACHI
NASDAQ
SmugMug
Archive, Disaster Recovery


_ Replay and analysis of the activity in the stock market
_ Interface built with Adobe FLEX + AIR
_ The Infrastructure is purely Amazon S3
NASDAQ
[Nasdaq]

Nasdaq on Amazon S3

gumi
playfish
zynga
Mobile and games

_ Two founders in Perth, Australia

_ “Discovr Apps” for iPhone: 250,000 downloads in 4 days

_ They are #1 in 28+ Countries

[Filter Squad]

_ Based in Beijing

_ Among the top 7 game developers on Facebook

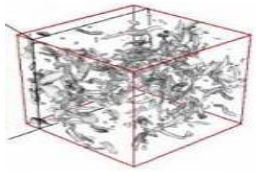
_ 3 Million daily active users globally

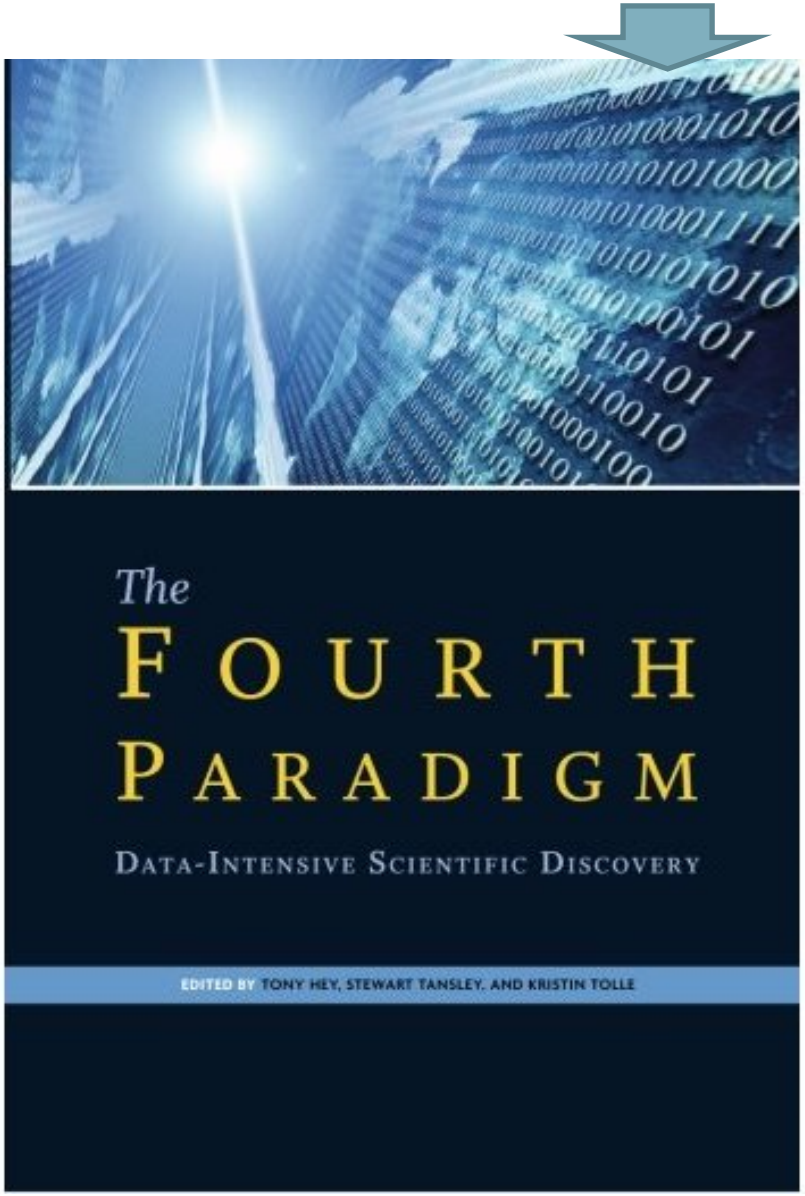
[FunPlus Game]

"With our social game growing from 1 million to 3 million in 3 months, we can spin out more than 300 servers quickly, only managed by ONE engineer."
Chen Qi, COO, Fun+ Game
FUN+GAME
[FunPlus Game]

Elastic-R

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - \mathrm{K}\frac{c^2}{a^2}$$





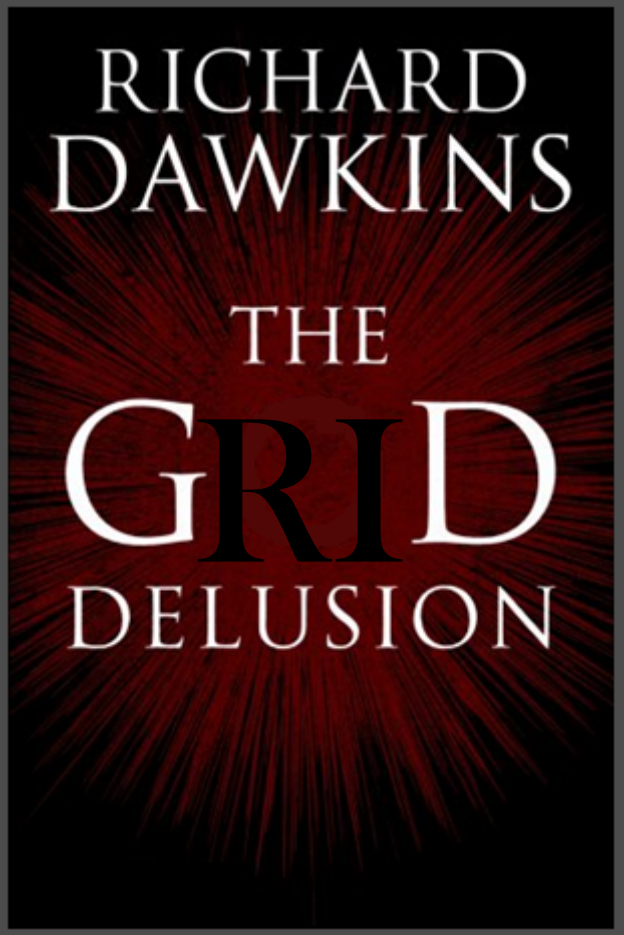
Jim Gray with his colleagues Gianfranco Putzulo and Irving Traiger in the late '70' / early '80s when they did roundbreaking work on concurrency control for databases (image courtesy of Heather Gray)

e-Research

e-Research refers to the development of, and the support for, information and computing technologies to facilitate all phases of research processes. The term e-Research originates from the term e-Science but expands its remit to all research domains not just the sciences. It's concerned with technologies that support all the processes involved in research including creating and sustaining research collaborations and discovering, analysing, processing, publishing, storing and sharing research data and information. Typical technologies in this domain include: Virtual Research Environments, Grid computing, visualisation services, and text and data mining services.

Cyberinfrastructure

Technological solution to the problem of efficiently connecting data, computers, and people with the goal of enabling derivation of novel scientific theories and knowledge



What's wrong with the GRID ?

"the abstractions that Grids expose – to the end-user, to the deployers and to application developers – are inappropriate and they need to be higher level" (Jha, Merzky, & Fox, 2009)

Suppose [a person] had a basket full of apples and, being worried that some of the apples were rotten, wanted to take out the rotten ones to prevent the rot spreading. How would he proceed? Would he not begin by tipping the whole lot out of the basket? And would not the next step be to cast his eye over each apple in turn, and pick up and put back in the basket only those he saw to be sound, leaving the others? In just the same way, those who have never philosophized correctly have various opinions in their minds which they have begun to store up since childhood, and which they therefore have reason to believe may in many cases be false. They then attempt to separate the false beliefs from the others, so as to prevent their contaminating the rest and making the whole lot uncertain. Now the best way they can accomplish this is to reject all their beliefs together in one go, as if they were all uncertain and false. They can then go over each belief in turn and re-adopt only those which they recognize to be true and indubitable. (Replies 7, AT 7:481)

Data analysis and statistical computing in the cloud

www.scipy.org

www.python.org

www.sagemath.org

www.scilab.org

F#

www.wolfram.com

www.mathworks.com

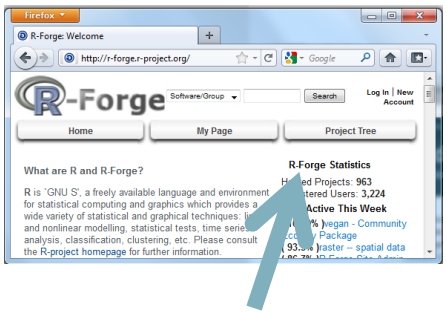
office.microsoft.com

www.spss.com

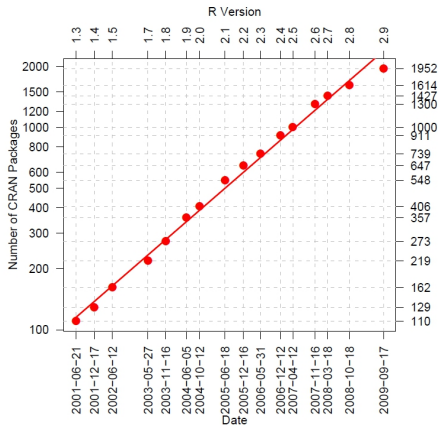
http://root.cern.ch

www.sas.com

- **Open-source (GPL) software environment for statistical computing and graphics**
- ***Lingua franca* of data analysis.**
- **Repositories of contributed R packages related to a variety of problem domains in life sciences, social sciences, finance, econometrics, chemo metrics, *etc.* are growing at an exponential rate.**
- **R is Super Glue**

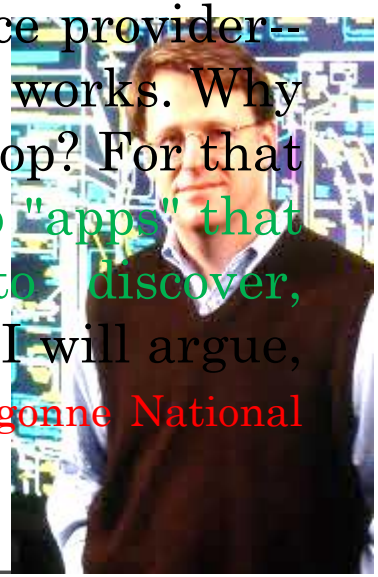


Hosted Projects: **963**
Registered Users: **3,224**

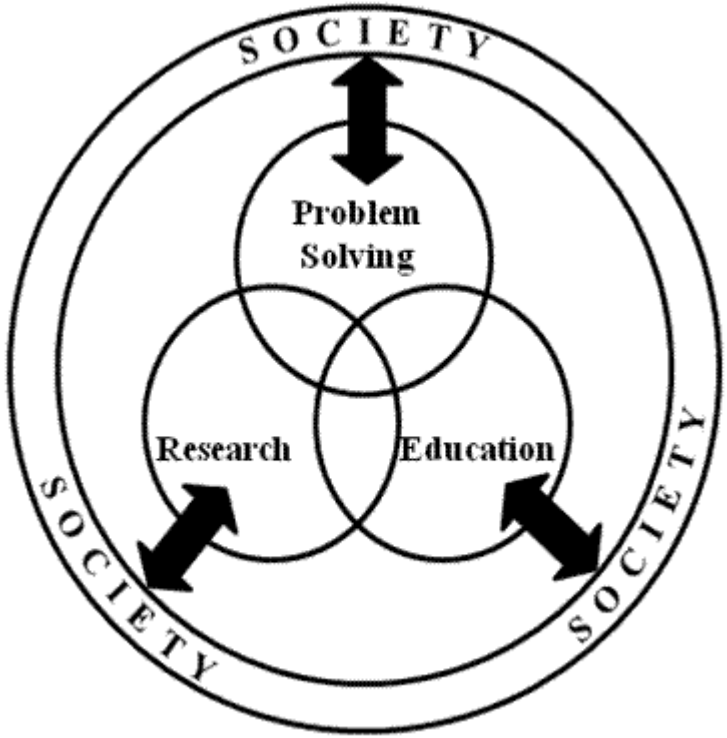


Evolution of the CRAN Packages number

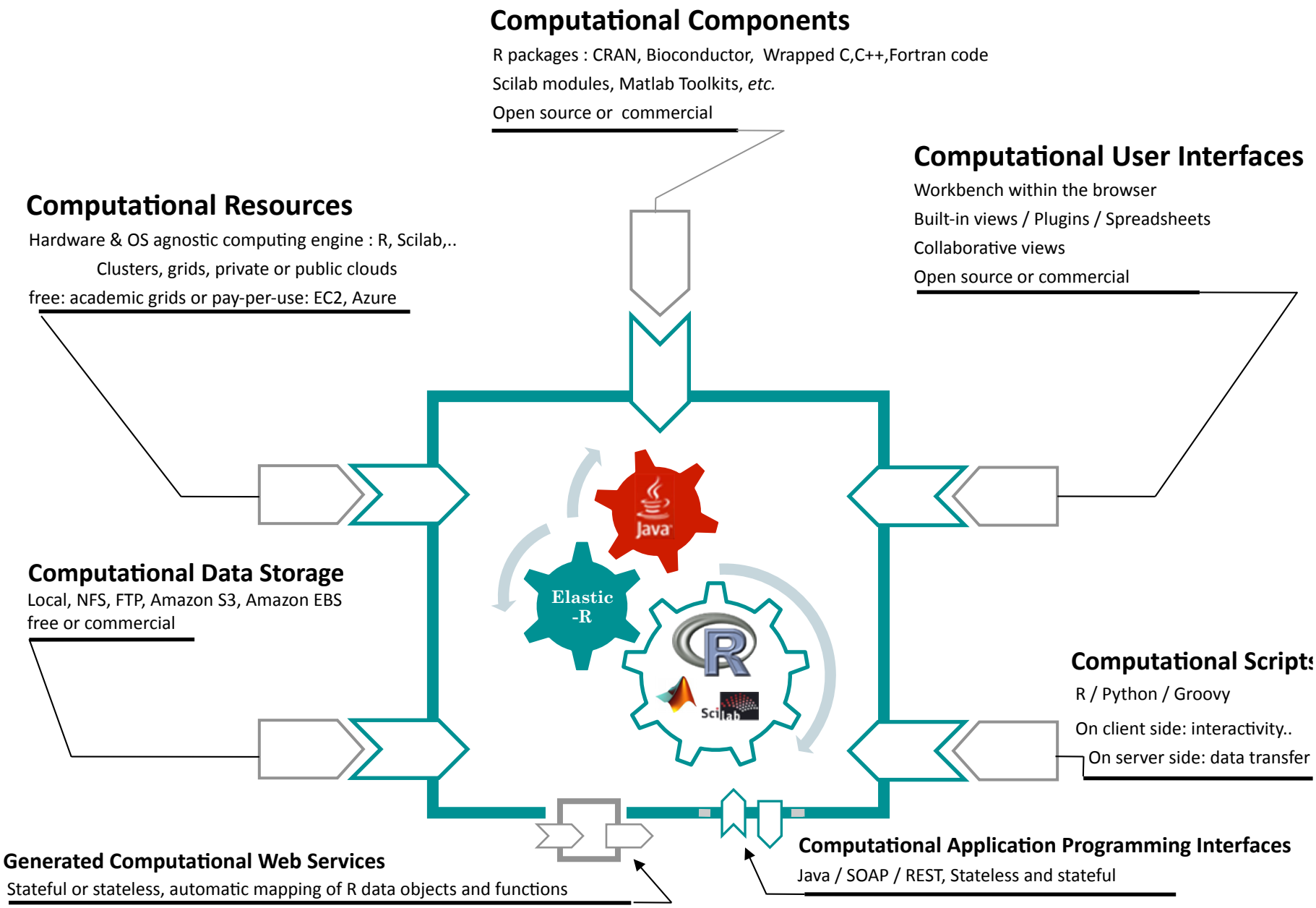
We've all heard about how on-demand computing and storage will transform scientific practice. But by focusing on resources alone, we're missing the real benefit of the large-scale outsourcing and consequent economies of scale that cloud is about. The biggest IT challenge facing science today is not volume but complexity. Sure, terabytes demand new storage and computing solutions. But they're cheap. It is establishing and operating the processes required to collect, manage, analyze, share, archive, etc., that data that is taking all of our time and killing creativity. And that's where outsourcing can be transformative. An entrepreneur can run a small business from a coffee shop, outsourcing essentially every business function to a software-as-a-service provider--accounting, payroll, customer relationship management, the works. Why can't a young researcher run a research lab from a coffee shop? For that to happen, we need to make it easy for providers to develop "apps" that encapsulate useful capabilities and for researchers to discover, customize, and apply these "apps" in their work. The effect, I will argue, will be a dramatic acceleration of discovery. Ian Foster, Argonne National Laboratory



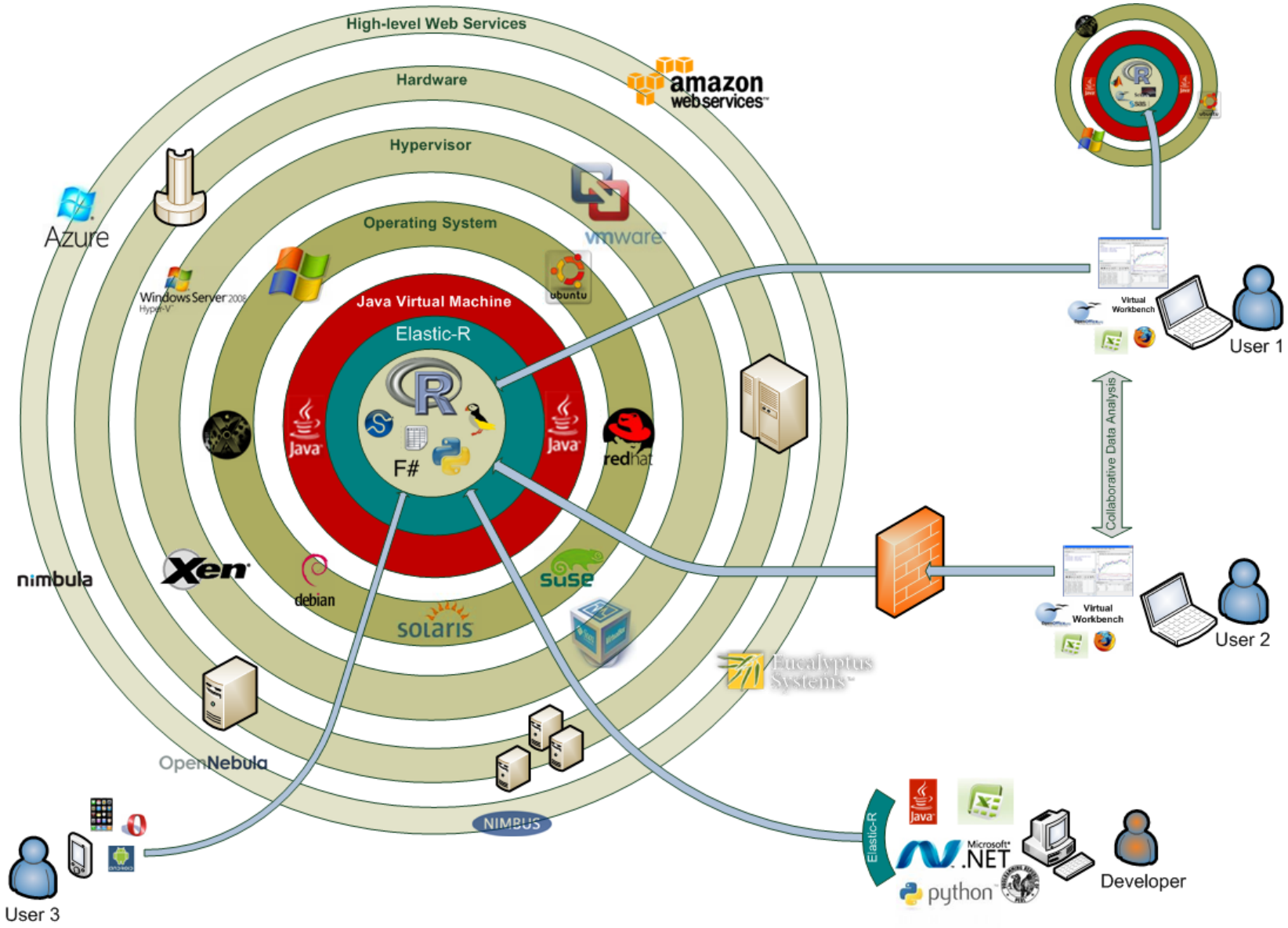
Integrating Research, Education, and Problem Solving

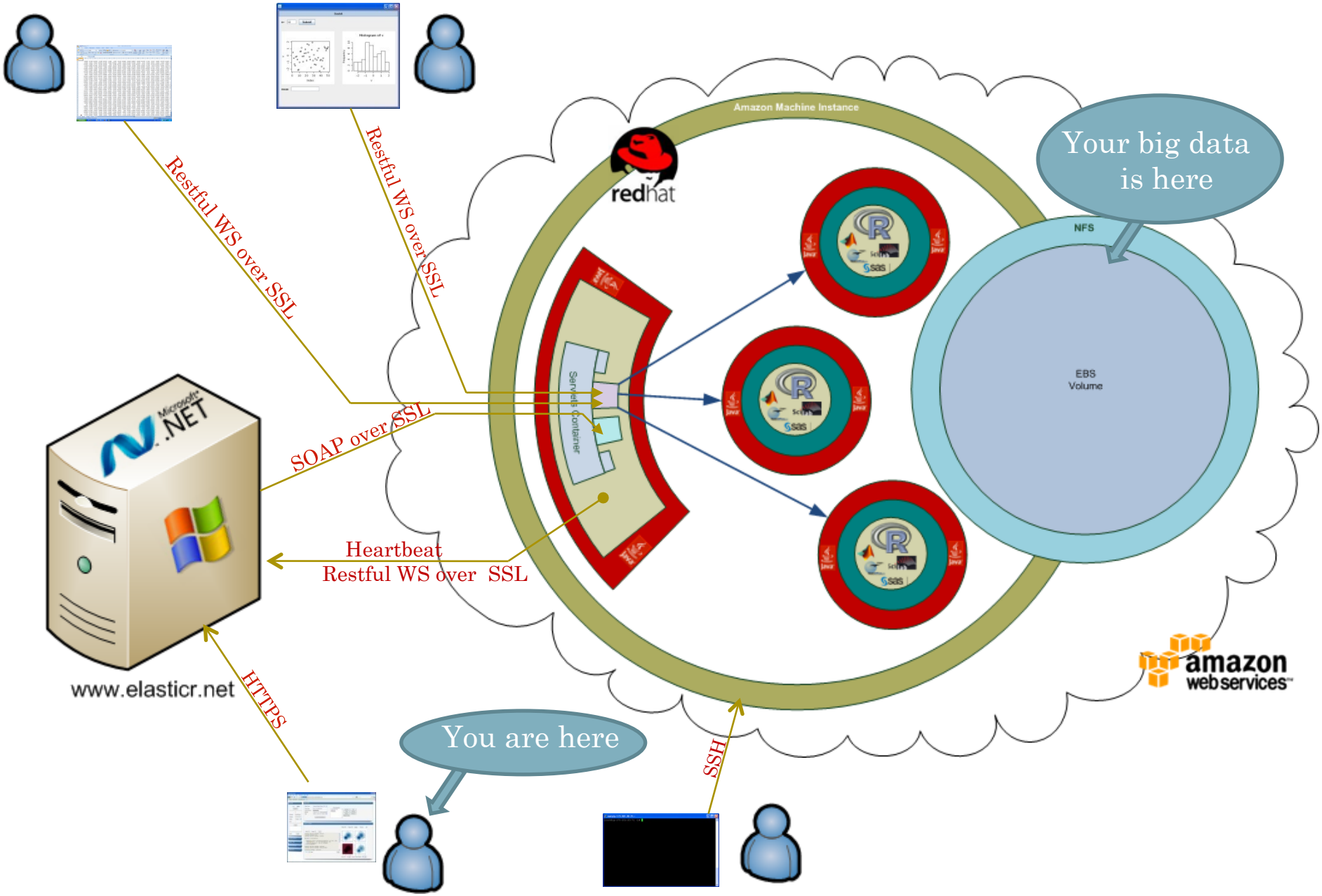


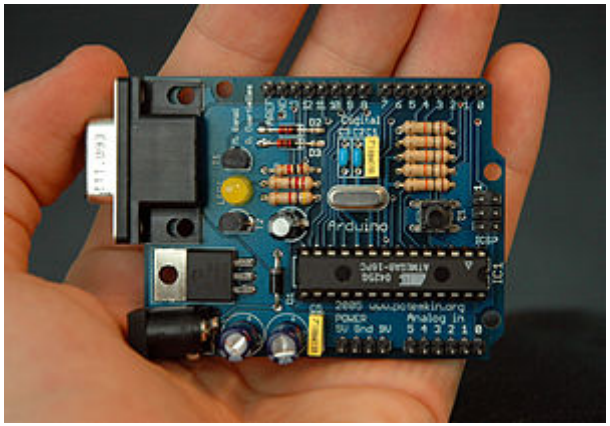
ELASTIC-R: PLUG-AND-PLAY SCIENTIFIC AND STATISTICAL COMPUTING

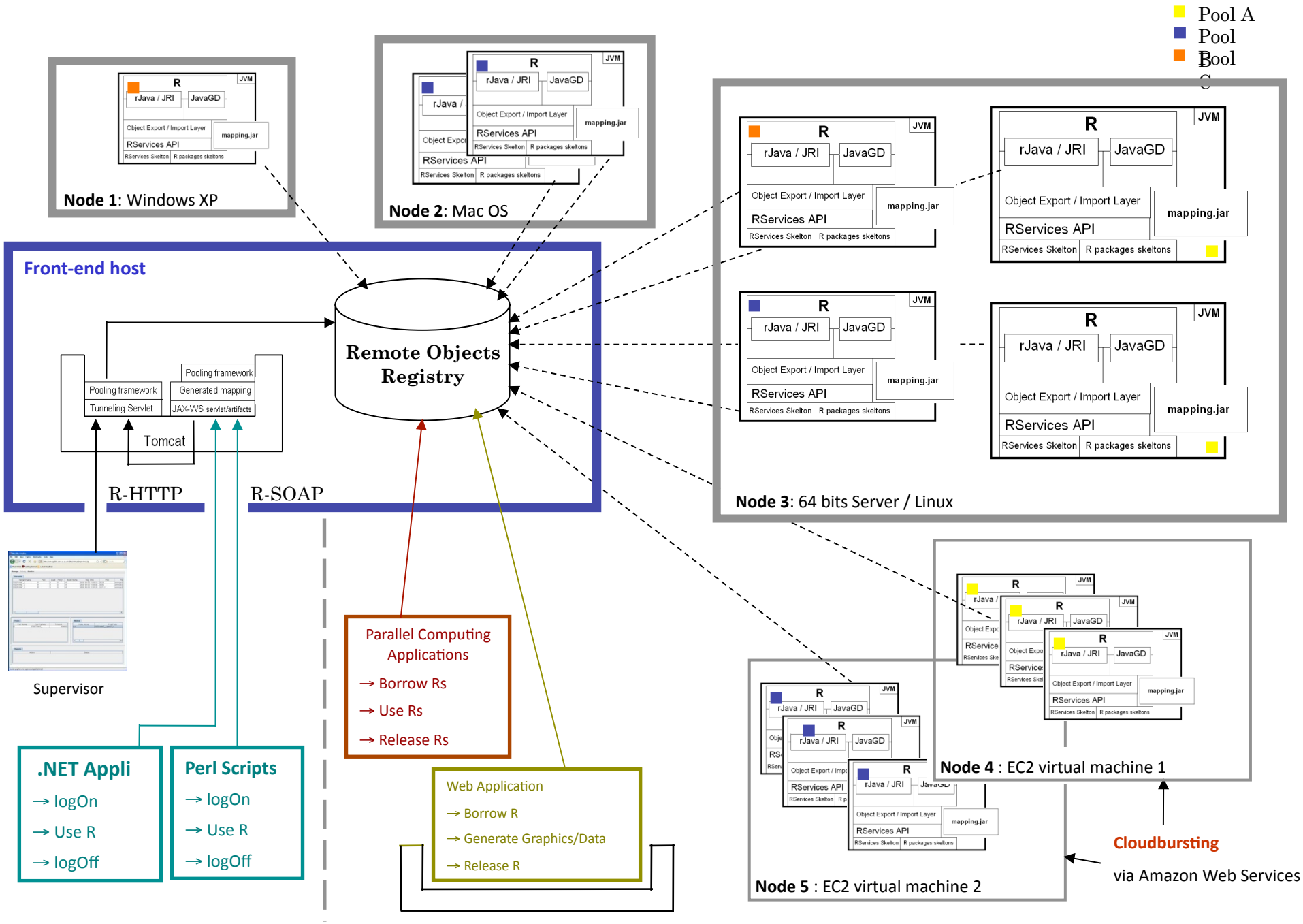


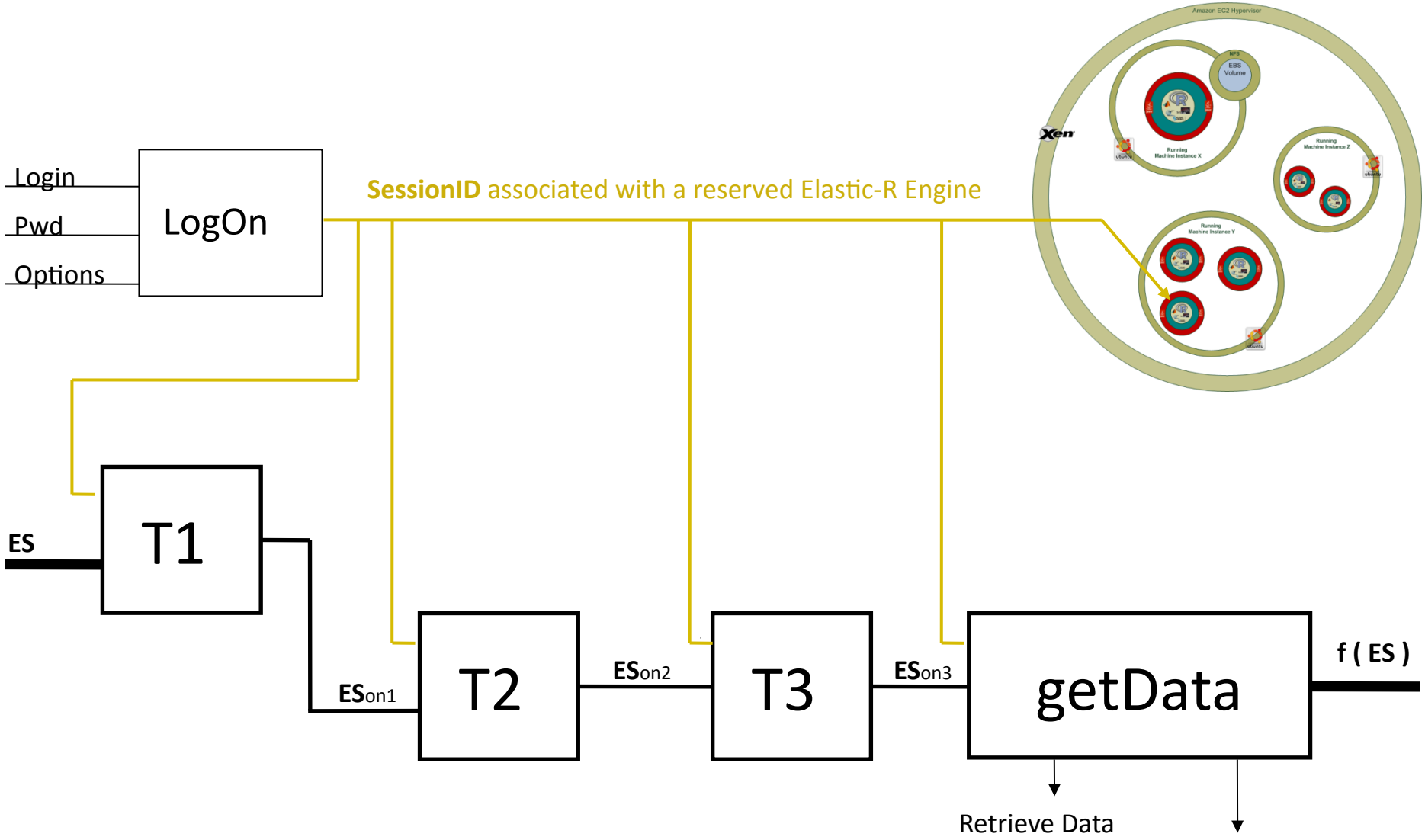
Elastic-R : a platform for scientific computing on Infrastructure-as-a-Service style clouds




Amazon Machine Instance
redhat
Your big data is here
NFS
EBS Volume
Servelt Container
Restful WS over SSL
Restful WS over SSL
SOAP over SSL
Heartbeat
Restful WS over SSL
HTTPS
SSH
www.elasticr.net
You are here
amazon web services


Digital
Arduino
2005 www.potemkin.org
POWER
Analog in


Pool A
Pool B
Pool C
Node 1: Windows XP
Node 2: Mac OS
Front-end host
Remote Objects Registry
Pooling framework
Tunneling Servlet
Pooling framework
Generated mapping
JAX-WS servlet/artifacts
Tomcat
R-HTTP
R-SOAP
Node 3: 64 bits Server / Linux
Node 4 : EC2 virtual machine 1
Node 5 : EC2 virtual machine 2
Cloudbursting
via Amazon Web Services
Supervisor
Parallel Computing Applications
→ Borrow Rs
→ Use Rs
→ Release Rs
.NET Appli
→ logOn
→ Use R
→ logOff
Perl Scripts
→ logOn
→ Use R
→ logOff
Web Application
→ Borrow R
→ Generate Graphics/Data
→ Release R
R
rJava / JRI
JavaGD
JVM
Object Export / Import Layer
RServices API
RServices Skelton
R packages skeltons
mapping.jar

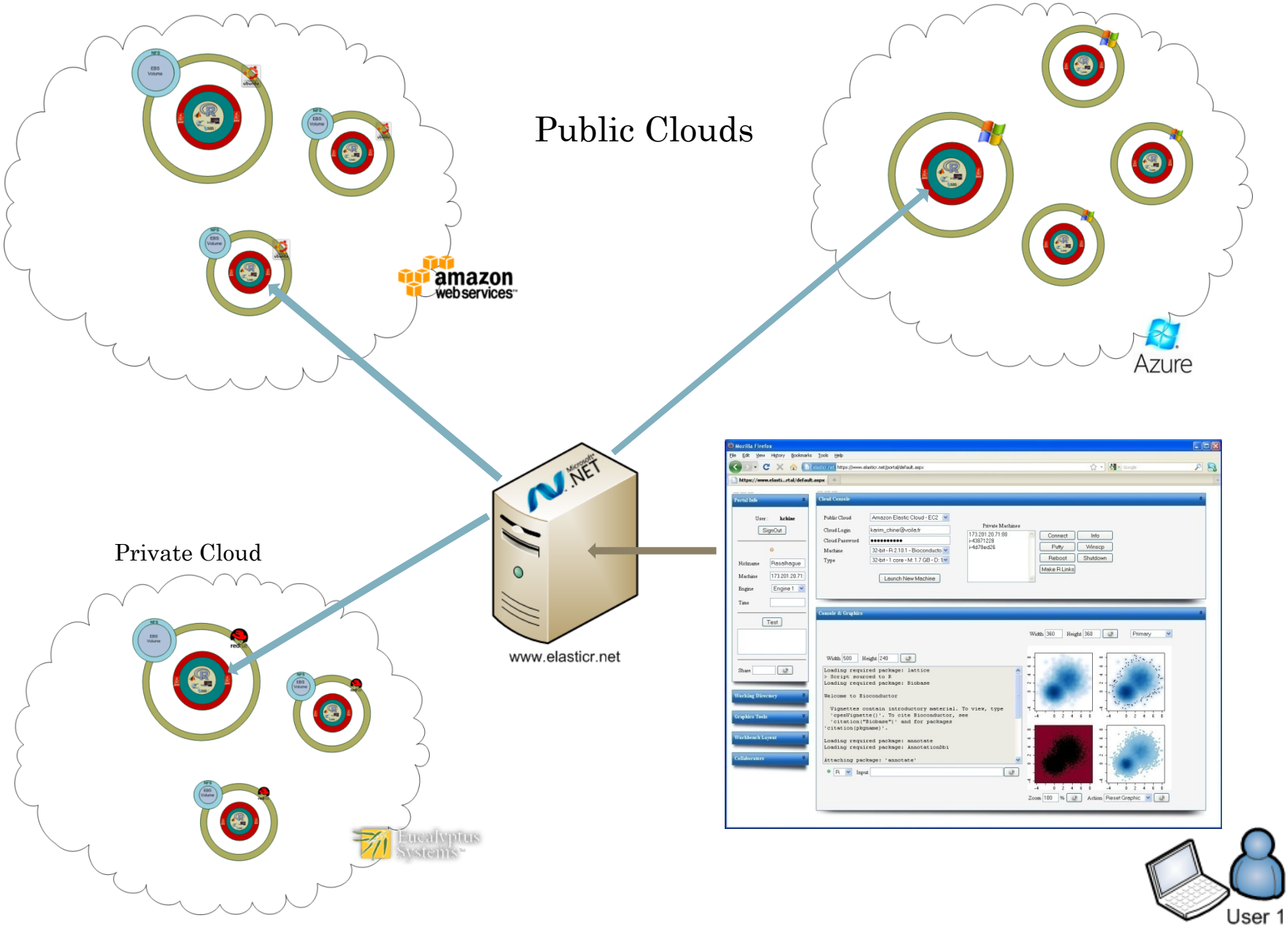


T1,**T2**,**T3** : Generated Stateful Web Services for R functions T1,T2 & T3
LogOn, **getData** : R-SOAP methods

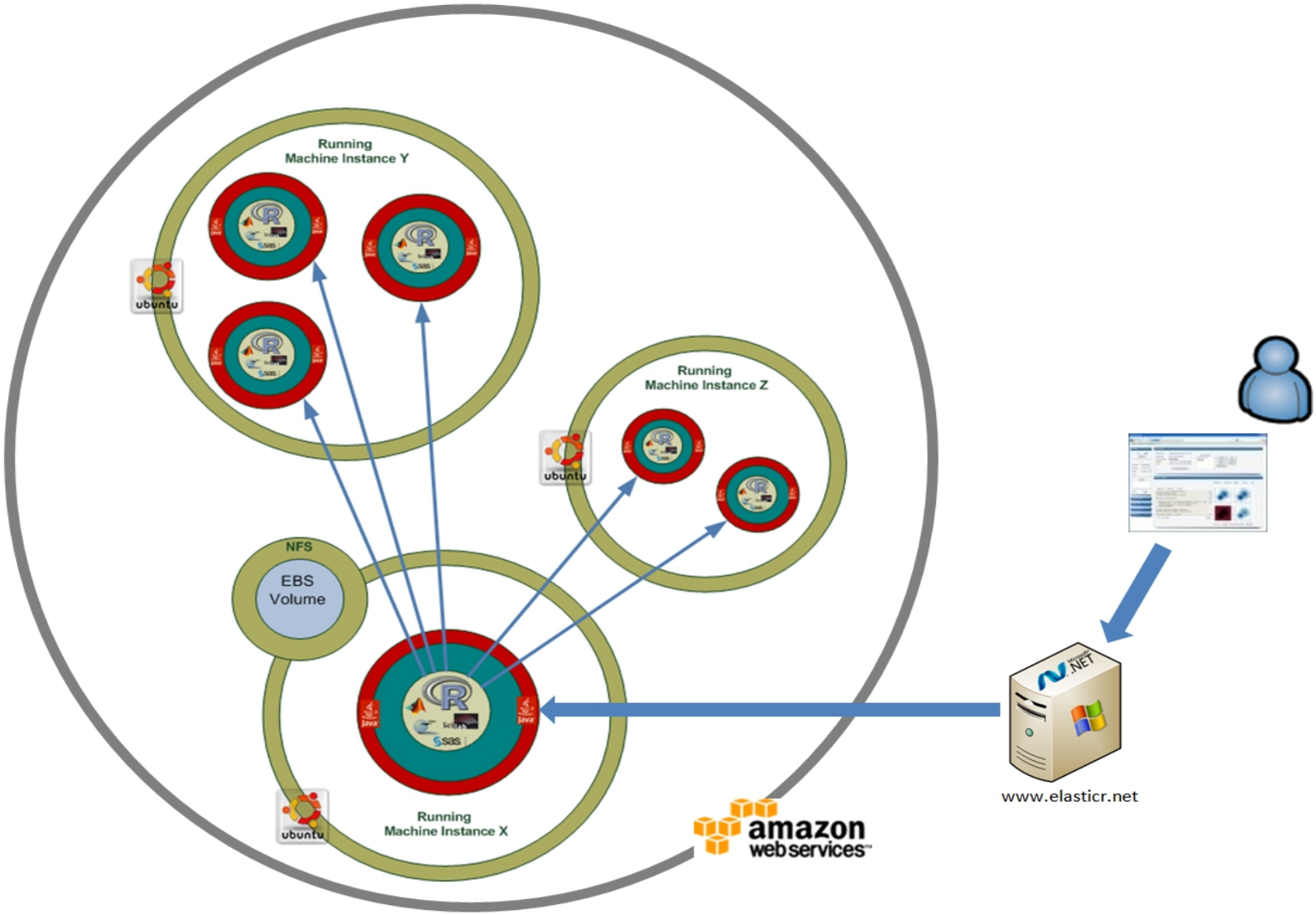
ES : ExpressionSet
ESon1, **ES**on2, **ES**on3 : ExpressionSet Object Names

f = T3 o T2 o T1

Elastic-R portal: single facade to public/private clouds

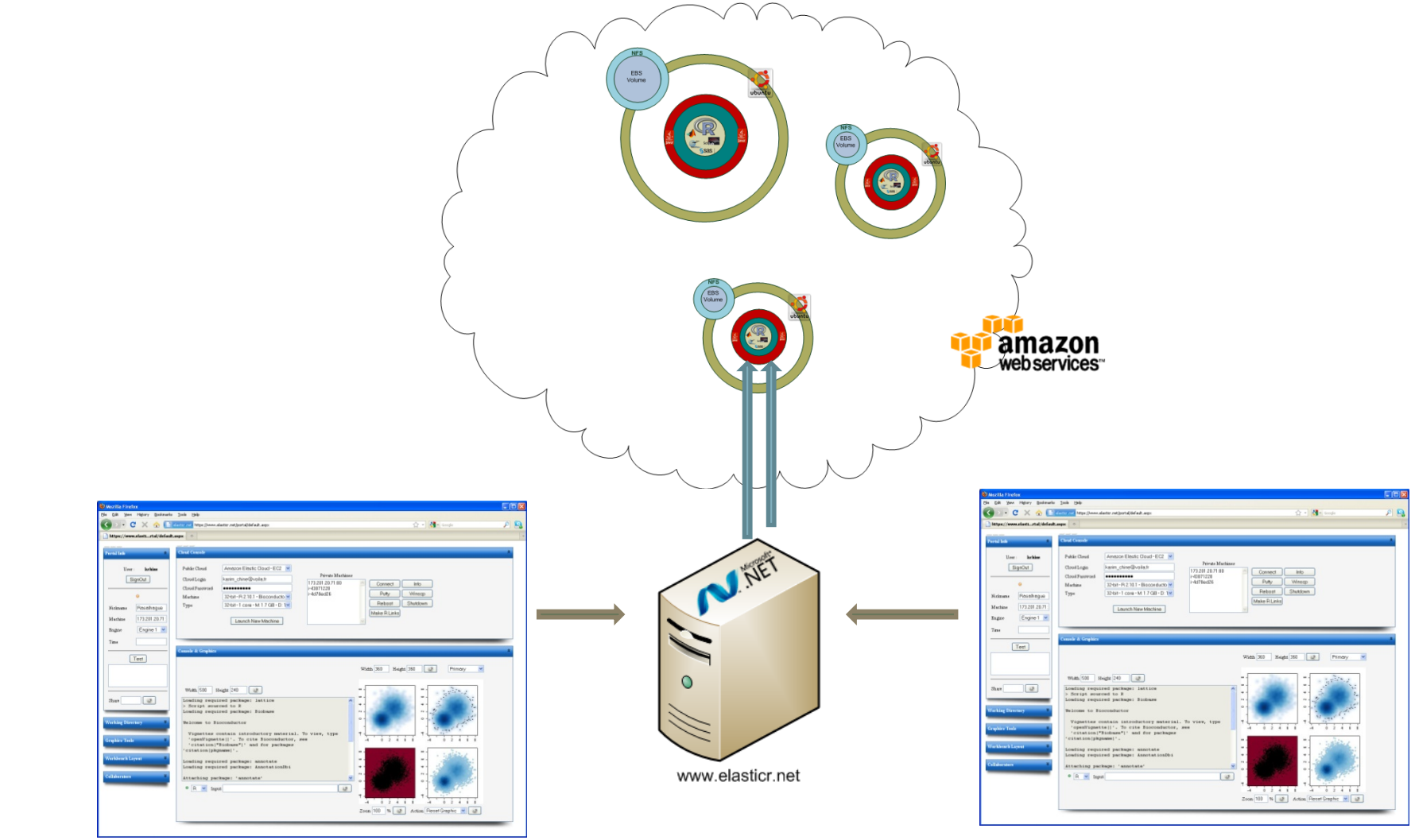


Scientist can control in parallel any number of stateful R/Python engines from within an R/Python session on the cloud or on a local machine

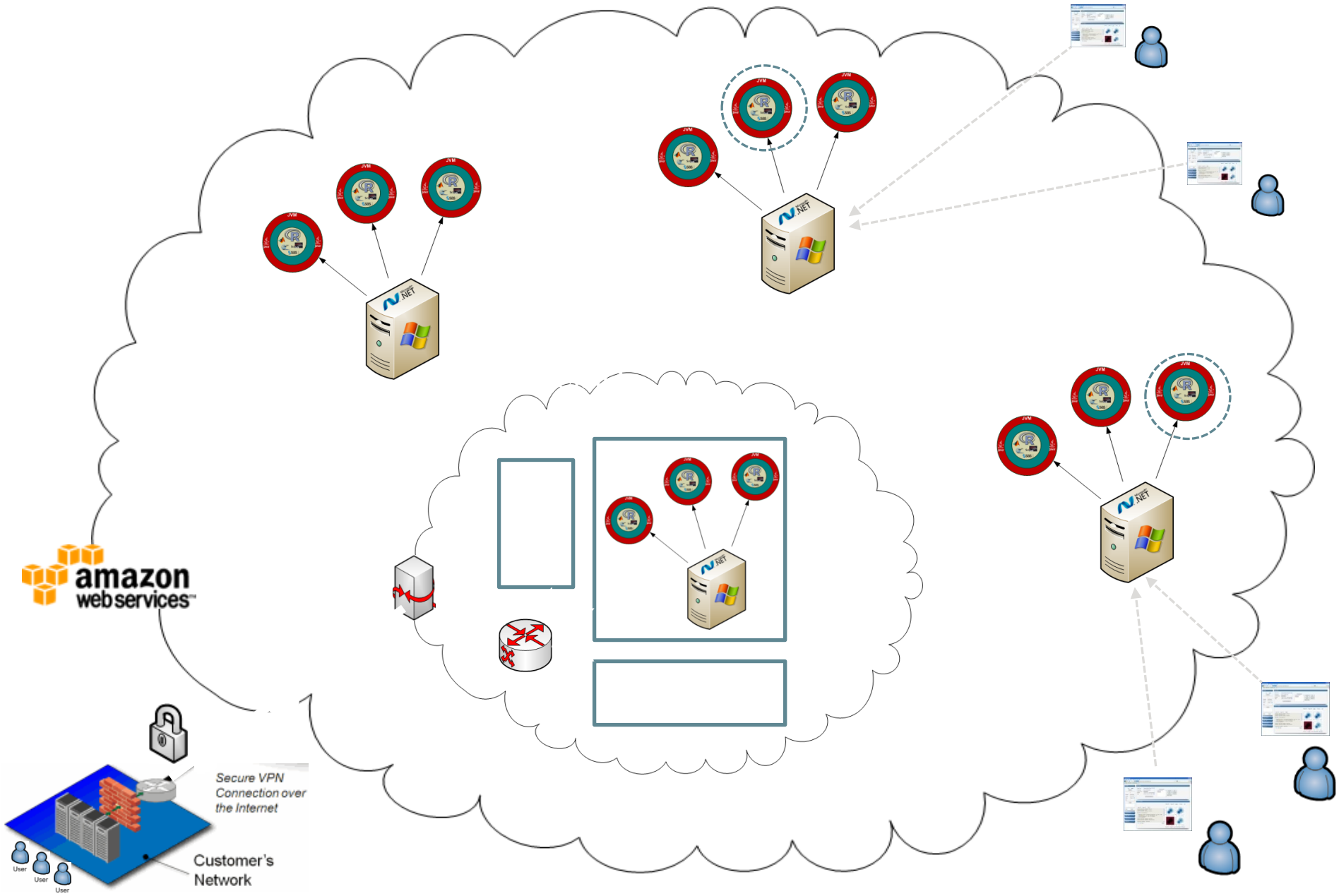


Collaborative e-Science and e-Learning in the cloud

Scientists, educators and students can share their machine instances, computing engines, data, spreadsheets, GUIs, etc. and collaborate on real-time

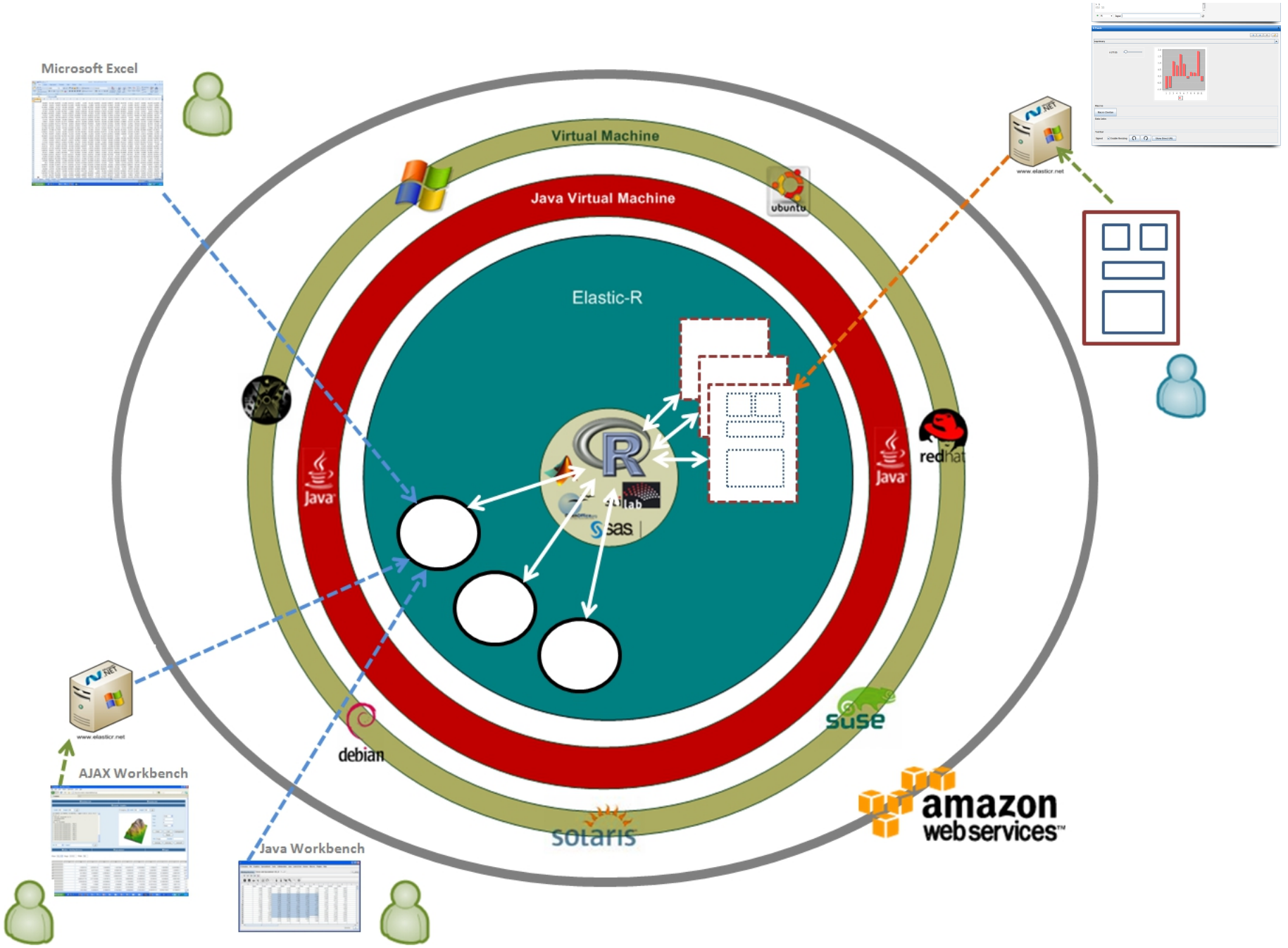


Dedicated portals for decentralized and private collaboration

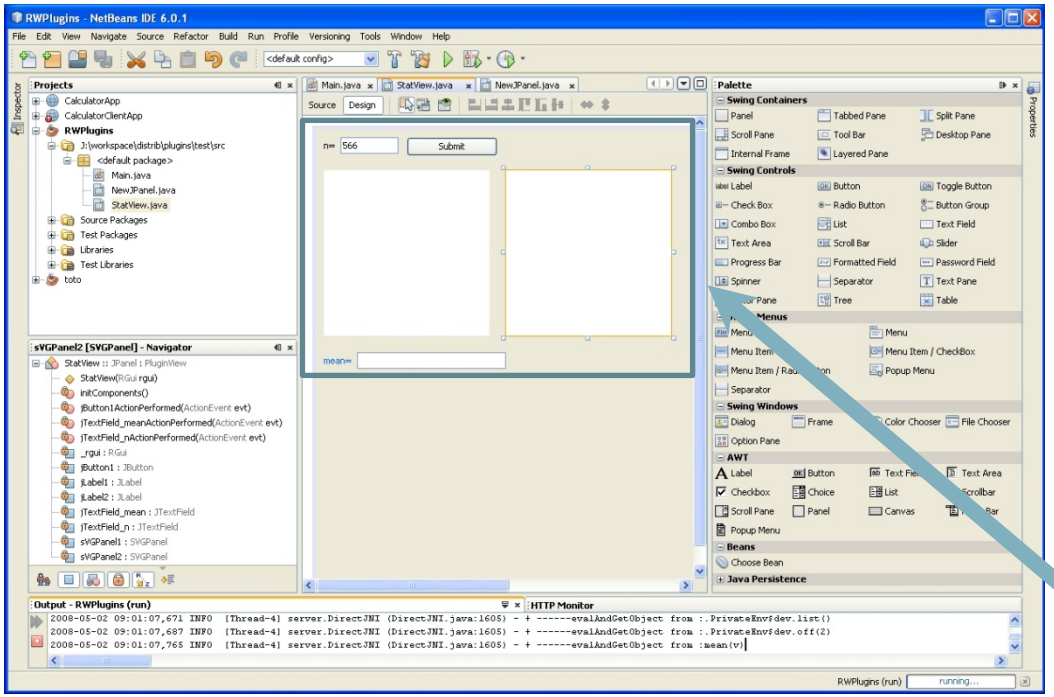


Software convergence in the cloud

Rapid e-Science/e-Learning applications development in the cloud


Microsoft Excel
Virtual Machine
Java Virtual Machine
Elastic-R
AJAX Workbench
Java Workbench
debian
suse
solaris
amazon
web services

Netbeans 6 – Visual GUI builder



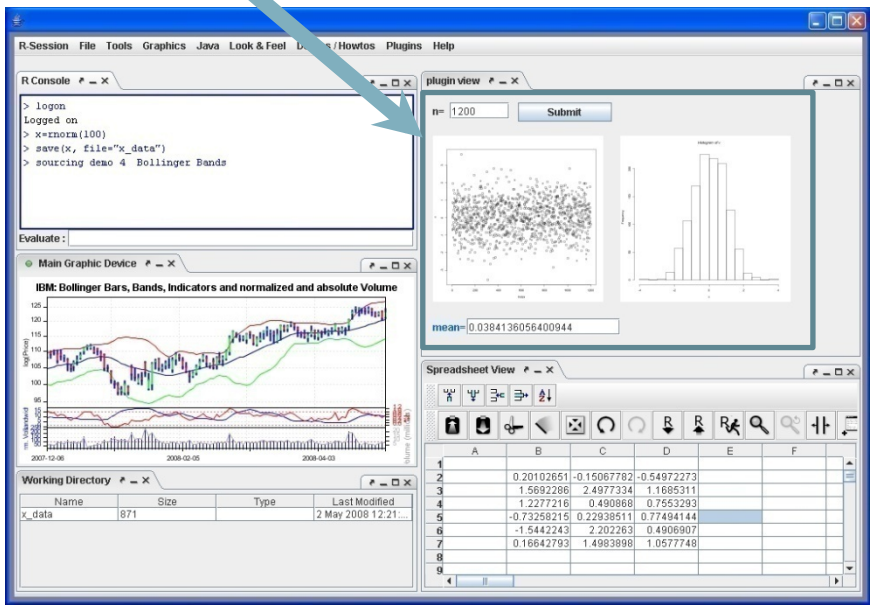
Build

myPlugin.jar

- myView1
- myView2
- descriptor.xml

Import Plugin

Elastic-R Workbench



Upload plugin

Plugins Repository

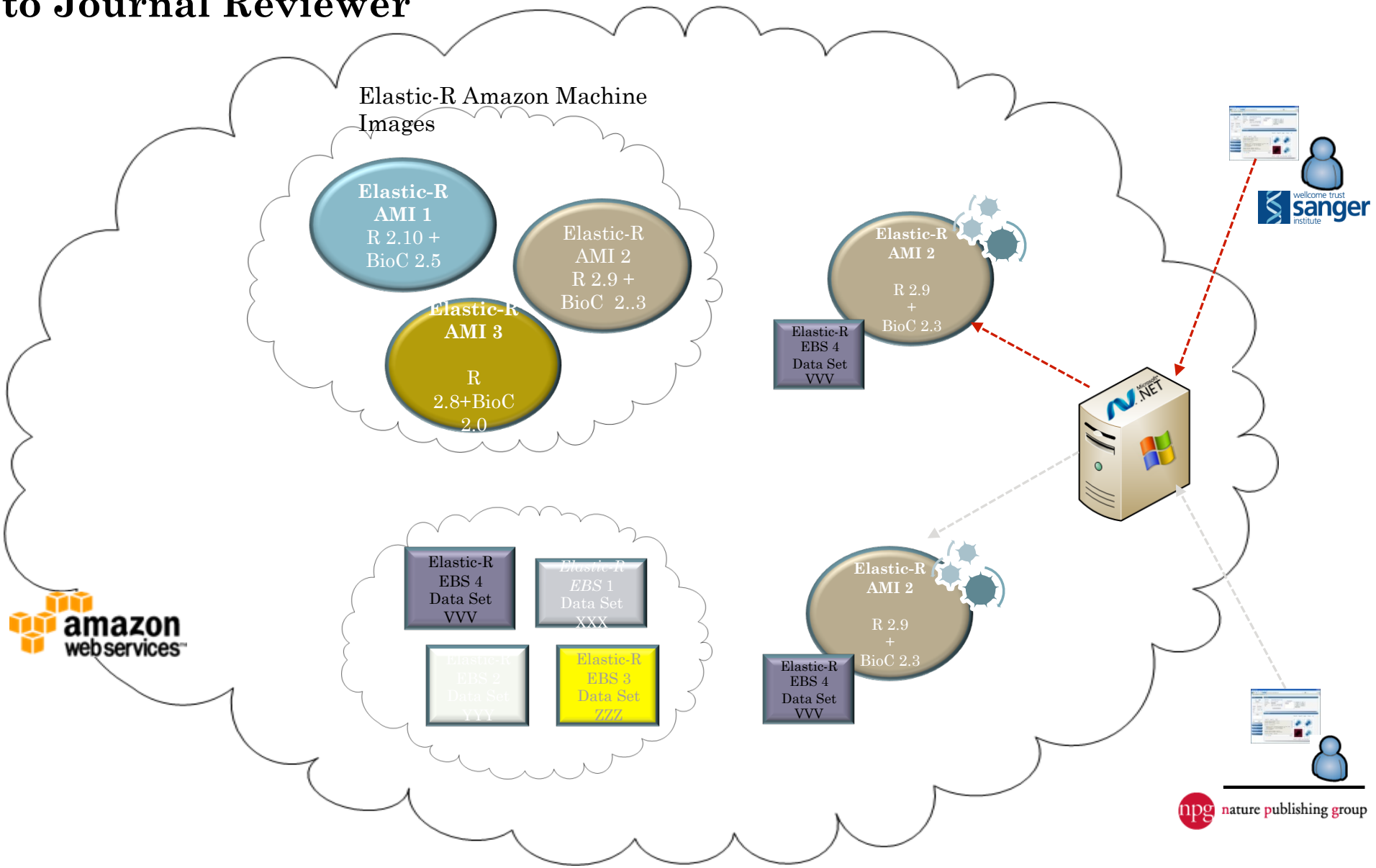
- myPlugin
- myDashboard

Browse Repository

Download Plugin

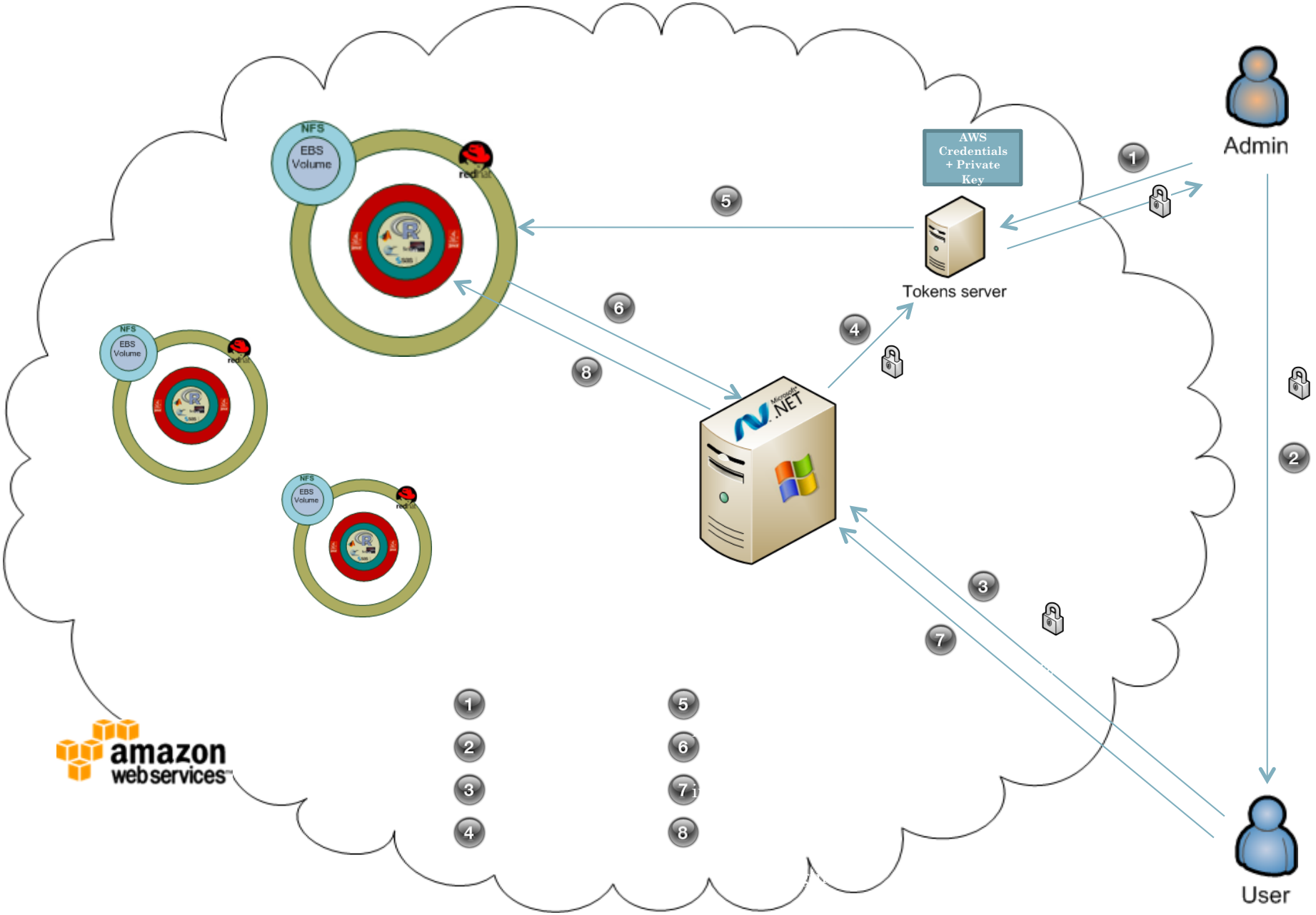
The cloud as an open ecosystem for science and education

Author form the Sanger Institute communicating details about the virtual appliances used to produce a paper's computational results to Journal Reviewer



Reducing the digital divide, the cloud for developing countries

Compute resources sharing: Elastic-R's digital tokens



Acknowledgments

ACS: Madi Nassiri **Amazon**: Simone Brunozzi, Deepak Singh **AT&T Research Labs**: Simon Urbanek **Auckland Centre for eResearch**: Nick Jones **Banca d'Italia**: Giuseppe Bruno **Bio-IT World**: Kevin Davies **BNP Paribas**: Ousseynou Nakoulima **Cambridge Healthtech Institute**: Cindy Crowninshield, Deborah Shear **City University of New York**: Mario Morales, Makram Talih **Columbia University**: Omar Besbes **Dassault Systèmes**: Omri Ben Ayoun, Patrick Johnson **Dataspora**: Michael E. Driscoll **EDF**: Alejandro Ribes **EBI**: Alvis Brazma, Wolfgang Huber, Kimmo Kallio, Misha Kapushesky, Michael Kleen, Alberto Labarga, Philippe Rocca-Serra, Ugis Sarkans, Kirsten Williams, Eamonn Maguire **EPFL**: Darlene Goldstein **ESPRIT**: Farouk Kammoun, Tahar. Benlakhdar **e-Taalim**: Nadhir Douma **ETH Zürich**: Yohan Chalabi, Diethelm Würtz, Martin Mächler **European Commission**: Konstantinos Glinos, Enric Mitjana, Monika Kacik, Ioannis Sagias **FHCRC**: Martin Morgan, Nianhua Li, Seth Falcon **Google**: Olivier Bosquet **FVG LLC**: Lisa Wood **Harvard University**: Tim Clark, Sudeshna Das, Douglas Burke,Paolo Ciccarese **IBM**: Jean-Louis Bernaudin, Pascal Sempe, Loic Simon, Lea A Deleris, Alex Fleischer, Alain Chabrier **Imperial College London**: Asif Akram, Vasa Curcin, John Darlington, Brian Fuchs **Indiana University**:Michael Grobe **INRIA**: David Monteau, Christian Saguez, Claude Gomez, Sylvestre Ledru **JISC**: John Wood, David Flanders **Johnson & Johnson - Janssen Pharmaceutica**: Patrick Marichal **KXEN**: Eric Marcade **Lancaster University**: Robert Crouchley, Daniel Grose **Leibniz Universität Hannover**: Kornelius Rohmeier **LIAMA:** Baogang Hue, Kang Cai **Limagrain**: Zivan Karaman **Mekentosj**: Alexander Griekspoor, Matt Wood **Microsoft**: Eric Le Marois, Tony Hey **Mubadala**: Ghazi Ben Amor **Nature Publishing Group**: Ian Mulvany, Steve Scott **NCeSS**: Peter Halfpenny, Rob Procter, Marzieh Asgari-Targhi, Alex Voss, YuWei Lin, Mercedes Argüello Casteleiro, Wei Jie, Meik Poschen, Katy Middlebrough, Pascal Ekin, June Finch, Farzana Latif, Elisa Pieri, Frank O'Donnell **New York Java User Group**: Frank D Greco **OeRC**: Dimitrina Spencer, Matteo Turilli, David Wallom, Steven Young **OMII-UK**: Neil Chue Hong, Steve Brewer **OpenAnalytics**: Tobias Verbeke **Oracle**: Dominique van Deth, Andrew Bond **OSS Watch**: Ross Gardler **Platform Computing**: Christopher Smith **Royal Society**: James Wilsdon **San Diego Supercomputer Center**: Nancy R. Wilkins-Diehr **Sanger Institute**: Lars Jorgensen, Phil Butcher **Shell**: Wayne.W.Jones, Nigel Smith **Société Générale**: Anis Maktouf **Stanford University**: John Chambers, Balasubramanian Narasimhan, Gunter Walther **SYSTEM@TIC**: Karim Azoum **Technische Universität Dortmund**: Uwe Ligges, Bernd Bischl **Technoforge**: Pierre-Antoine Durgeat **Tekiano**: Samy Ben Naceur **Télécom-ParisTech**: Isabelle Demeure, Georges Hebrail, Nesrine Gabsi **The Generations Network**: Jim Porzak **Total**: Yannick Perigois **Tunisian Ministry of Communication Technologies**: Naceur Ammar, Lamia Chaffai-Sghaier, Mohamed Saïd Ouerghi, Syrine Tlili **Tunisian Ecole Polytechnique**: Riadh Robbana **UC Berkeley**: Noureddine El Karoui, Terry Speed **UC Davis**: Rudy Beran, Debashis Paul, Duncan Temple Lang **UCL**: Daniel Jeffares **UCLA**: Ivo Dinov, Jeroen Ooms **UC San Diego**: Anthony Gamst **UCSF**: Tena Sakai **Université Catholique de Louvain**: Christian Ritter **University of Cambridge**: Ian Roberts, Robert MacInnis Peter Murray-Rust, Jim Downing, Michael Simmons, Mark Hayes **University of Manchester**: Carole Goble, Len Gill, Simon Peters, Richard D Pearson, Iain Buchan, John Ainsworth **University of Plymouth**: Paul Hewson **University of Split**: Ivica Puljak **UTK**: Ajay Ohri **World Bank Group-IFC**: Oualid Ammar **Yahoo**: Laurent Mirguet, Rob Weltman

Merci!

-> www.elastic-r.org

karim.chine@cloudera.co.uk

..