Deuxièmes rencontres R

Karim Chine

R and the Cloud

Cloud Era Ltd 27 June 2013 karim.chine@cloudera.co.uk

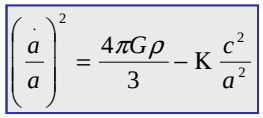
Outline

- Introduction
- Rethinking virtual research and teaching
- Elastic-R: Towards a universal platform for data science
- Elastic-R: Design and technologies overview
- Elastic-R: The scriptability framework
- Demo
- Conclusion

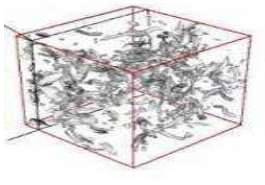
Introduction

- Science and the 4th paradigm

Experimental Science



Theoretical Science



Computational Science

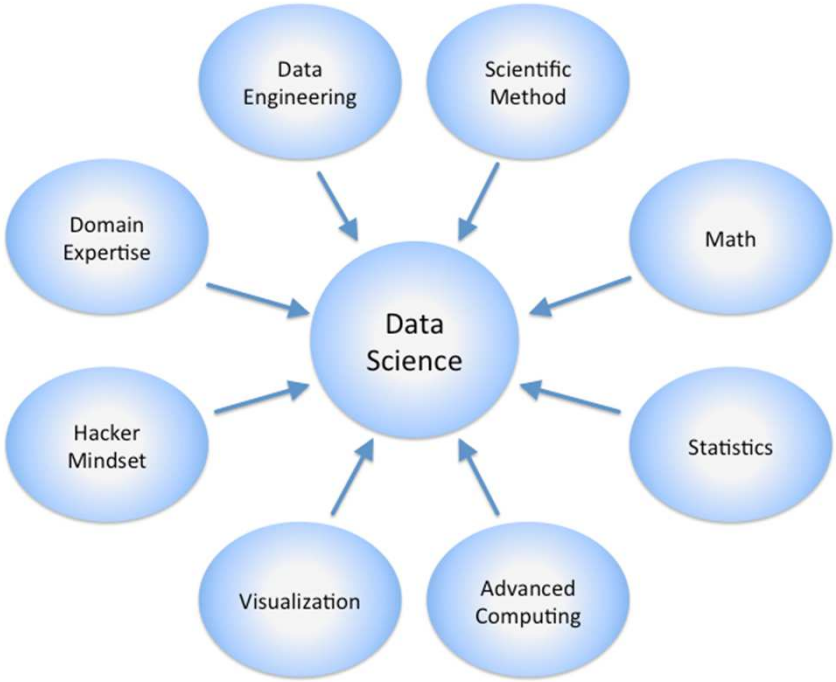
e-Science / Data-intensive Science

- The e-Research dancing bears

The townspeople gather to see the wondrous sight as the massive, lumbering beast shambles and shuffles from paw to paw. The bear is really a terrible dancer, and the wonder isn't that the bear dances well but that the bear dances at all.

Introduction

The rise of data science



Data science incorporates varying elements and builds on techniques and theories from many fields, including math, statistics, data engineering, pattern recognition and learning, advanced computing, visualization, uncertainty modeling, data warehousing, and high performance computing with the goal of extracting meaning from data and creating data products

* http://en.wikipedia.org/wiki/Data_science

Introduction

Fragmentation and friction in the data science arena

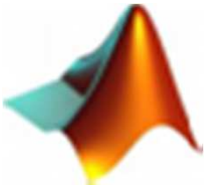
www.python.org

www.scipy.org

www.scilab.org

www.sagemath.org

www.wolfram.com



www.mathworks.com

office.microsoft.com

www.spss.com

http://root.cern.ch

F#

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**

Introduction

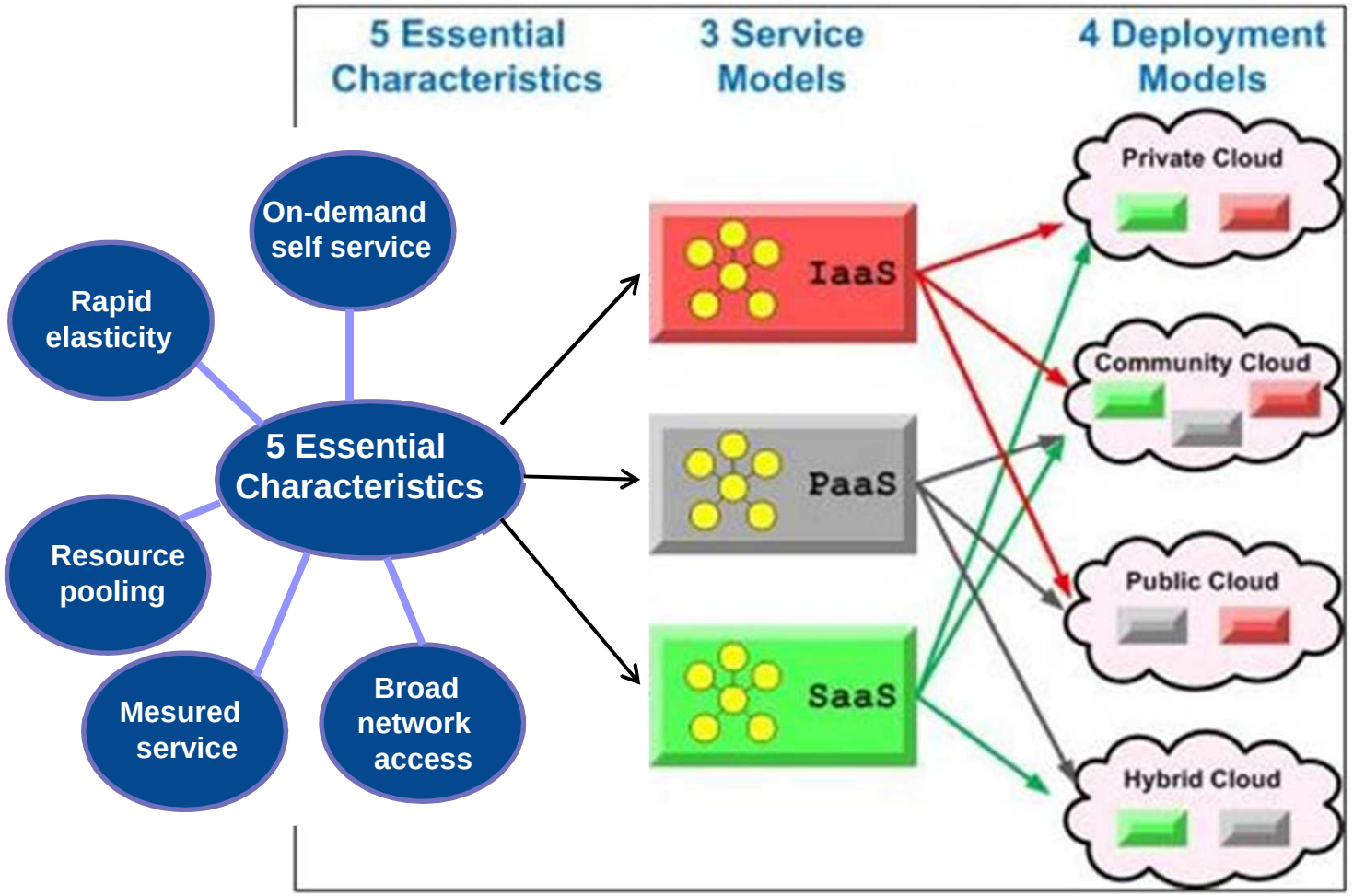
The Next Generation Data Science Platform

Arduino / Raspberry pi
Democratizing electronics

Elastic-R
Democratizing data science

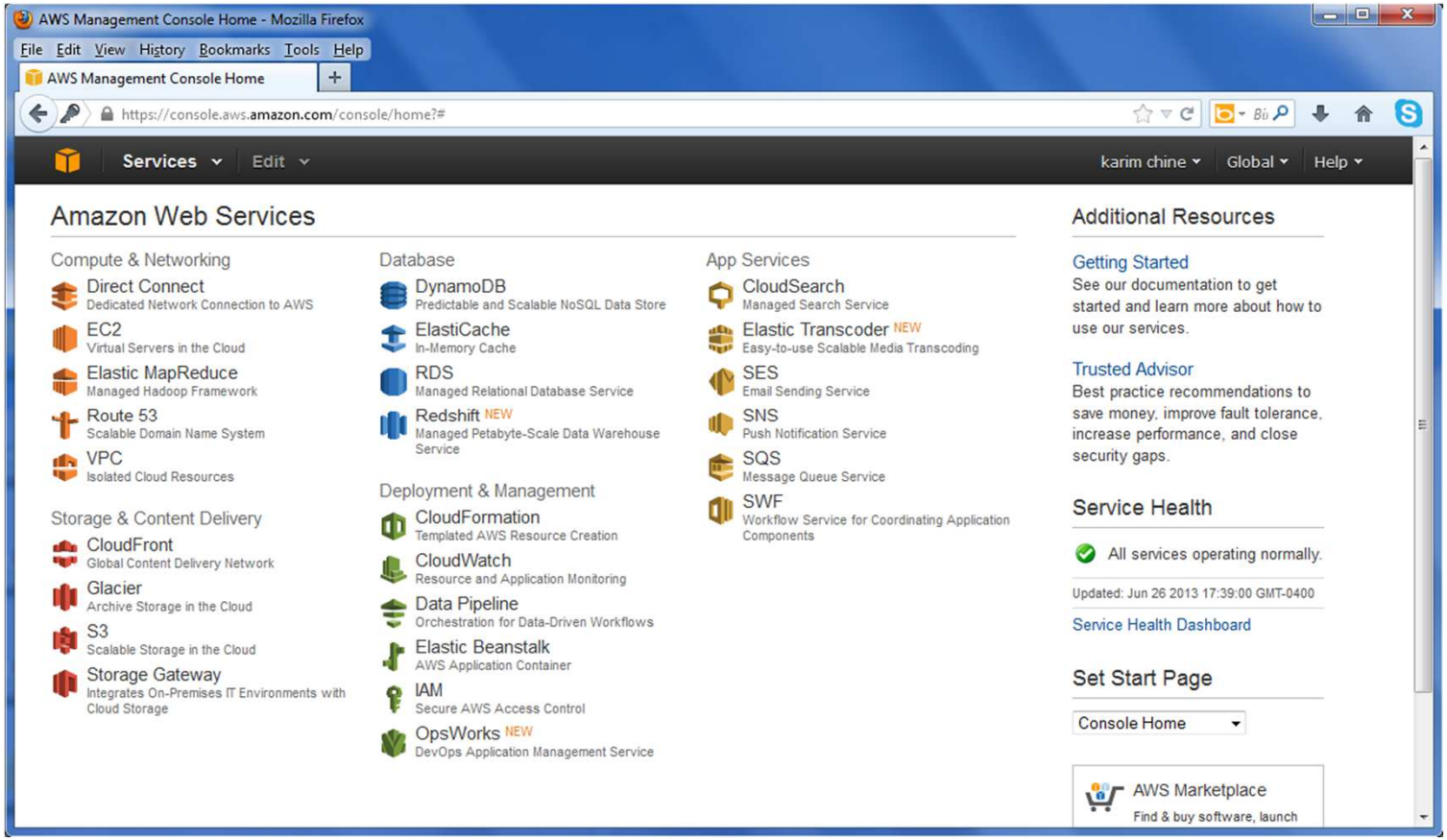
Introduction

The Cloud and its capabilities



Introduction

The cloud and its capabilities



Introduction

The cloud and its capabilities

- 3D printers are becoming a common place:
  - Creating three-dimensional solid object of virtually any shape from a digital model.
  - Scripting the physical world
  - Sharing physical reality on Facebook is now as easy as sharing your holiday pictures

Rethinking virtual research and teaching

- Free researchers from their IT services dictatorship. Give them self-service access to the IT resources they need
- Make real-time collaboration a free, reliable and ubiquitous service
- Allow Researchers to share without restrictions. Make the cloud become an ecosystem for Open Science where all research artifacts can be produced, discovered and reused
- Allow researchers to produce and publish to the web advanced applications/services without recourse to developers/admins

Rethinking virtual research and teaching

- Provide platforms for Science-as-a-Service
- Allow Researchers to « sell » the software/models/algorithms/techniques they invent seamlessly: Create a market place for data science artifacts and application
- Provide capabilities for making data analysis and computational research traceable and reproducible
- Bridge the gap between the different computational research tools: interconnect SCEs, workflow workbenches, Documents editors…

Rethinking virtual research and teaching

- Provide affordable and reliable tools for remote education
  - High-quality voice and video chat for a large number of users
  - Self-service collaboration tools: Editors, White boards, IDEs, etc.
  - Modules for Traceability/reproducibilty
- Extend existing on-line courses platforms to include capabilities such as:
  - Companion software environments in SaaS mode
  - Collaborative problem solving tools
  - Interactive courses
  - Tokens for Ready-to-run e-Learning applications
  - E-Learning environments' visual designers

Elastic-R: Towards a universal platform for data science

Computational Components

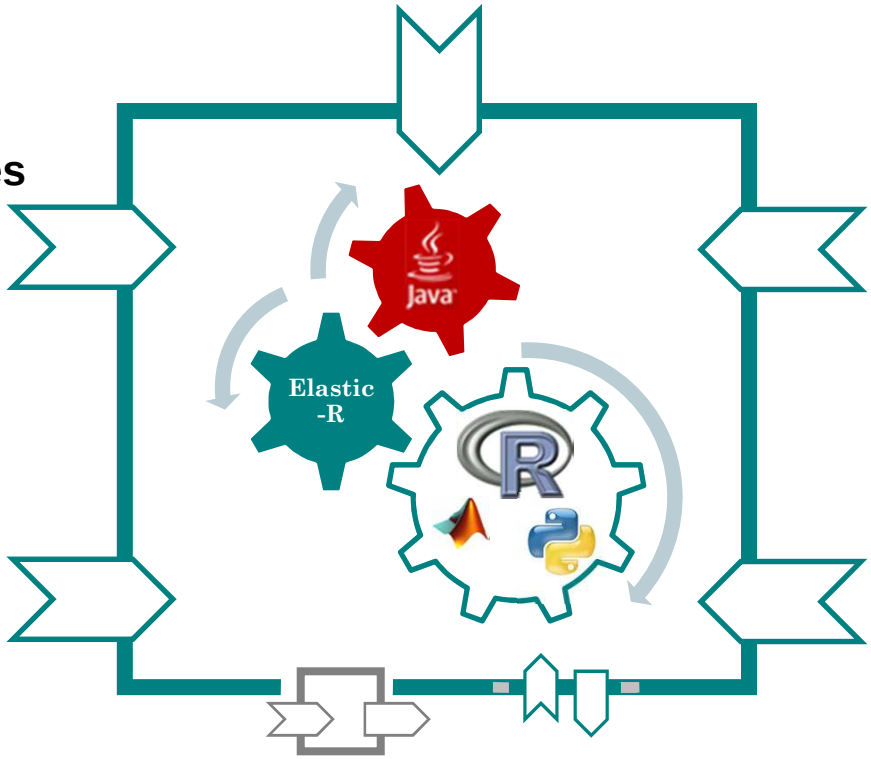
R packages, Wrapped C,C++,Fortran code, Python modules, Matlab Toolkits…

Open source or commercial

Computational Resources

Clusters, grids, private or public clouds

Free or pay-per-use



Computational GUIs

HTML5 and Desktop Workbench

Built-in views /Plugins /Collaborative views

Open source or commercial

Computational Storage

Local, NFS, FTP, Amazon S3, EBS

Computational Scripts

R / Python / Matlab / Groovy

Generated Computational Web Services

Stateful or stateless, mapping of R objects/functions

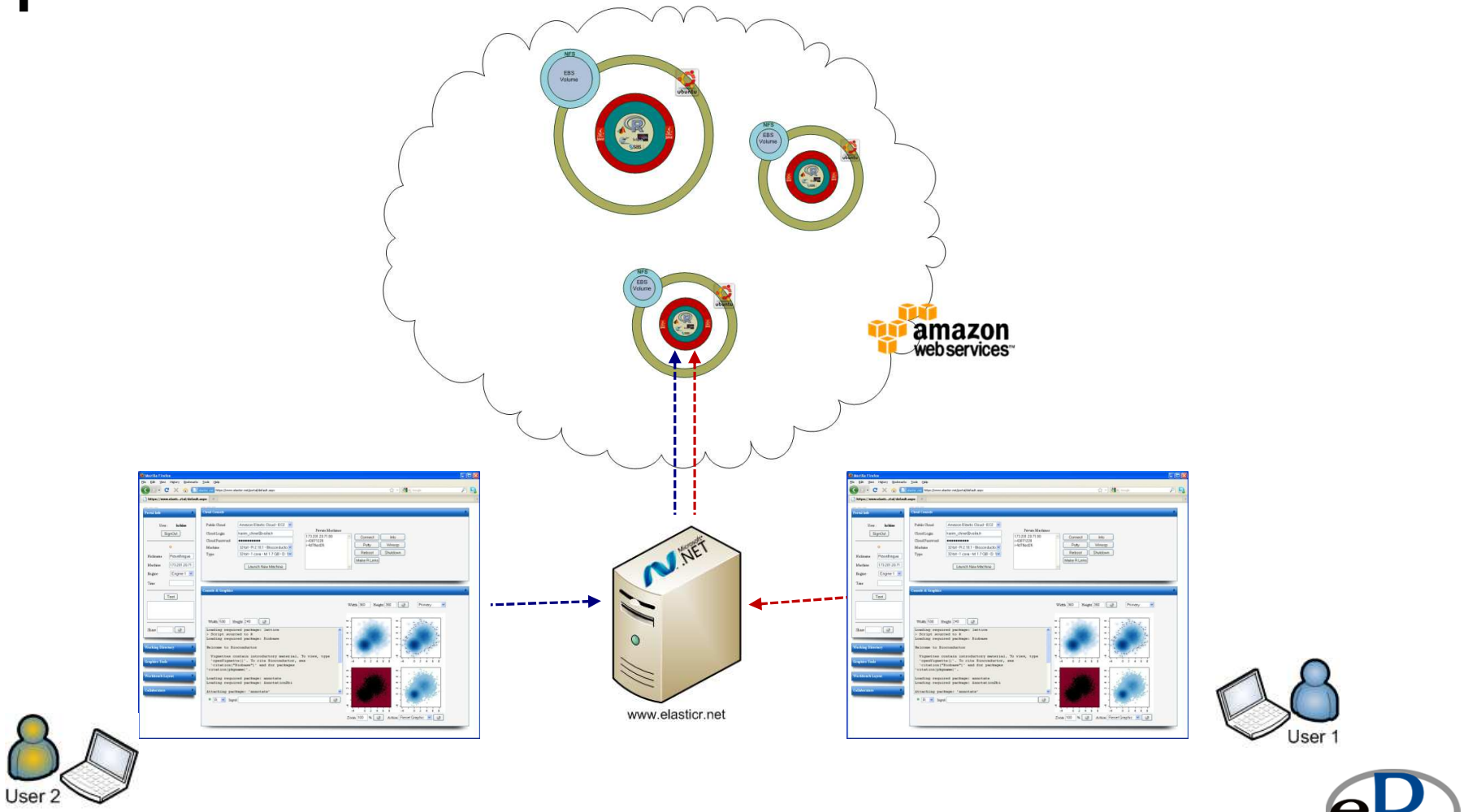
Computational APIs

Java / SOAP / REST, Stateless and stateful

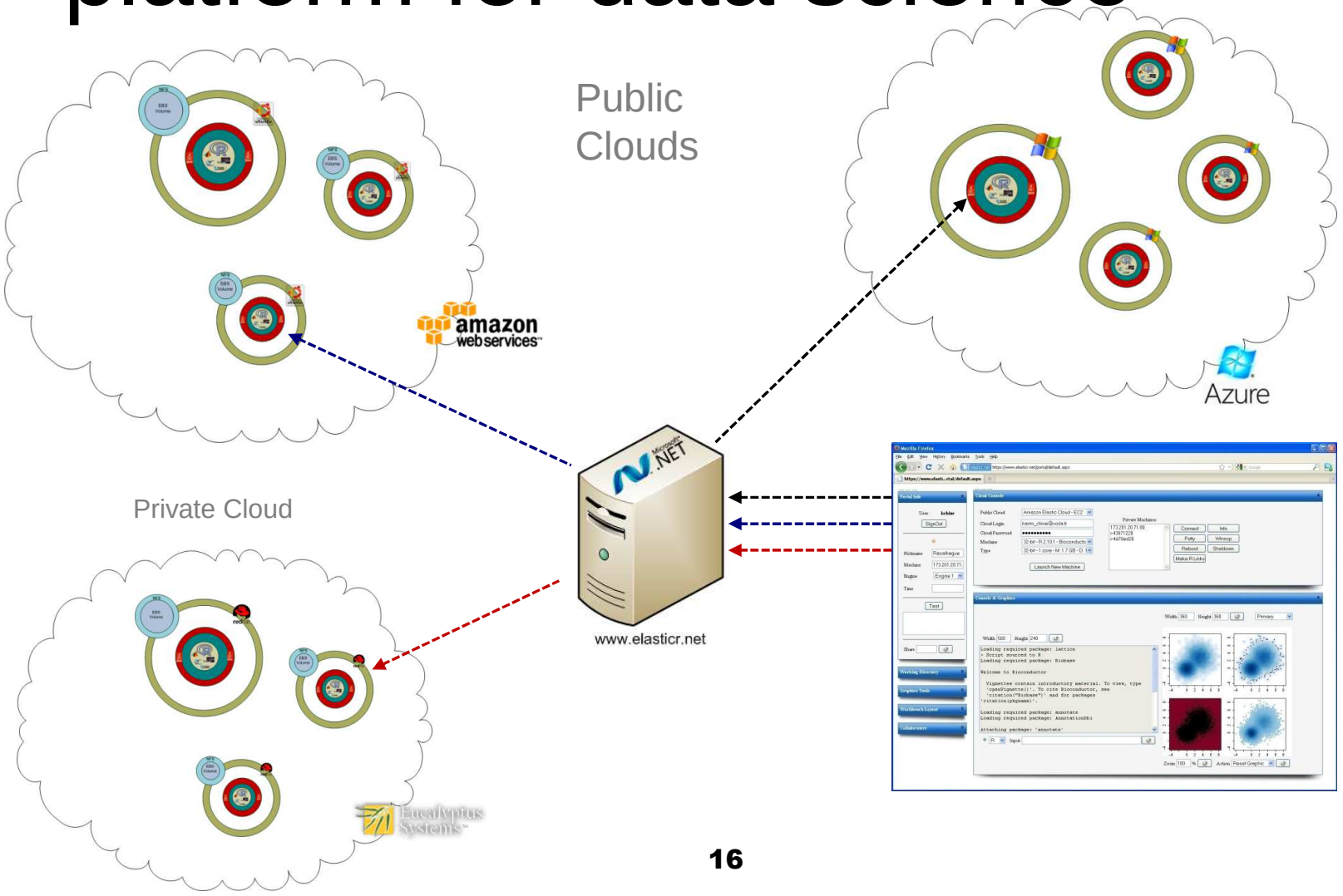
Elastic-R: Towards a universal platform for data science

Robot submarine dives to the deepest part of the ocean controlled by a 7-mile cable as thin as single human hair

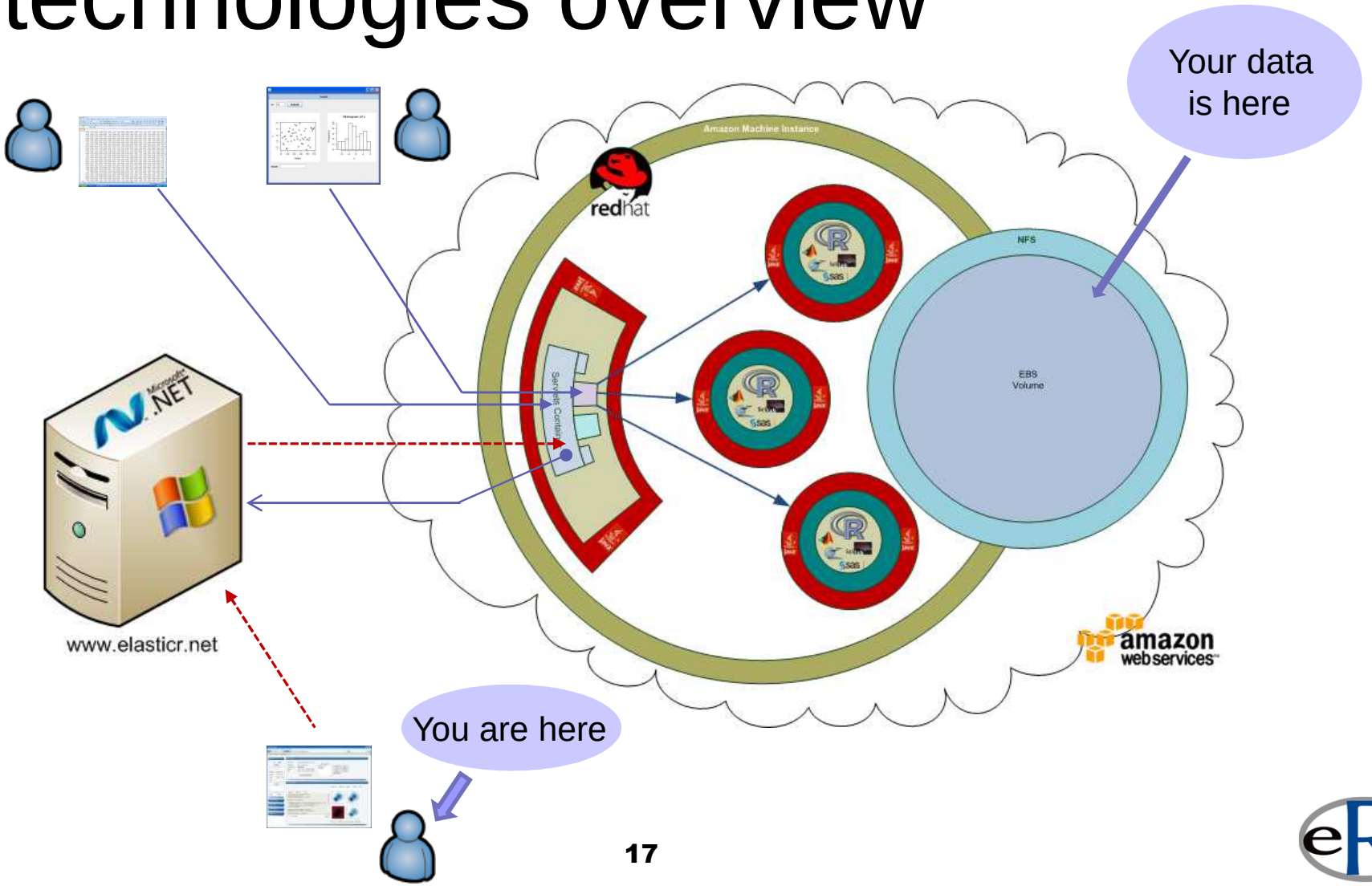
Elastic-R: Towards a universal platform for data science



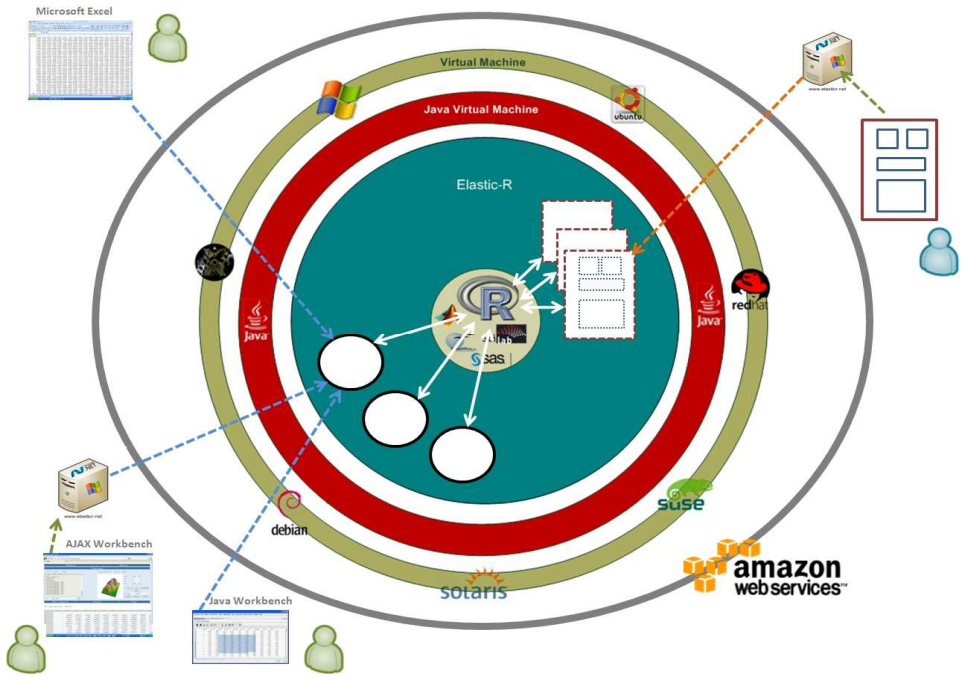
Elastic-R: Towards a universal platform for data science



Elastic-R: Design and technologies overview

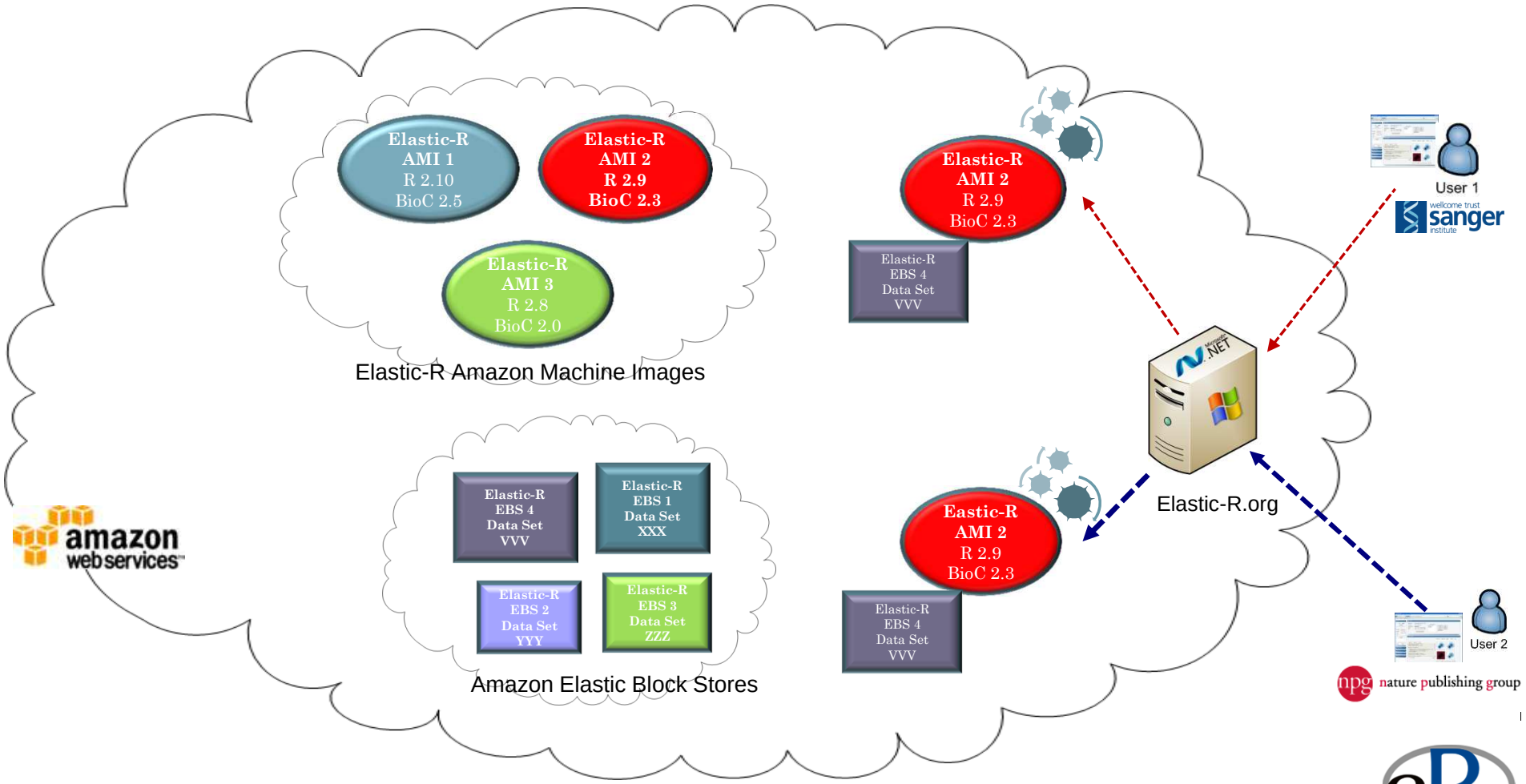


Elastic-R: Design and technologies overview



- Remote Java/R Processes
  - Events-driven Remote Objects/Engines
  - R, Python, Mathematica, Matlab, Scilab, ...
  - Collaborative Spreadsheets
  - Collaborative Scientific Graphics Canvas
  - Collaborative Dashboard with collaborative widgets

Elastic-R: Design and technologies overview

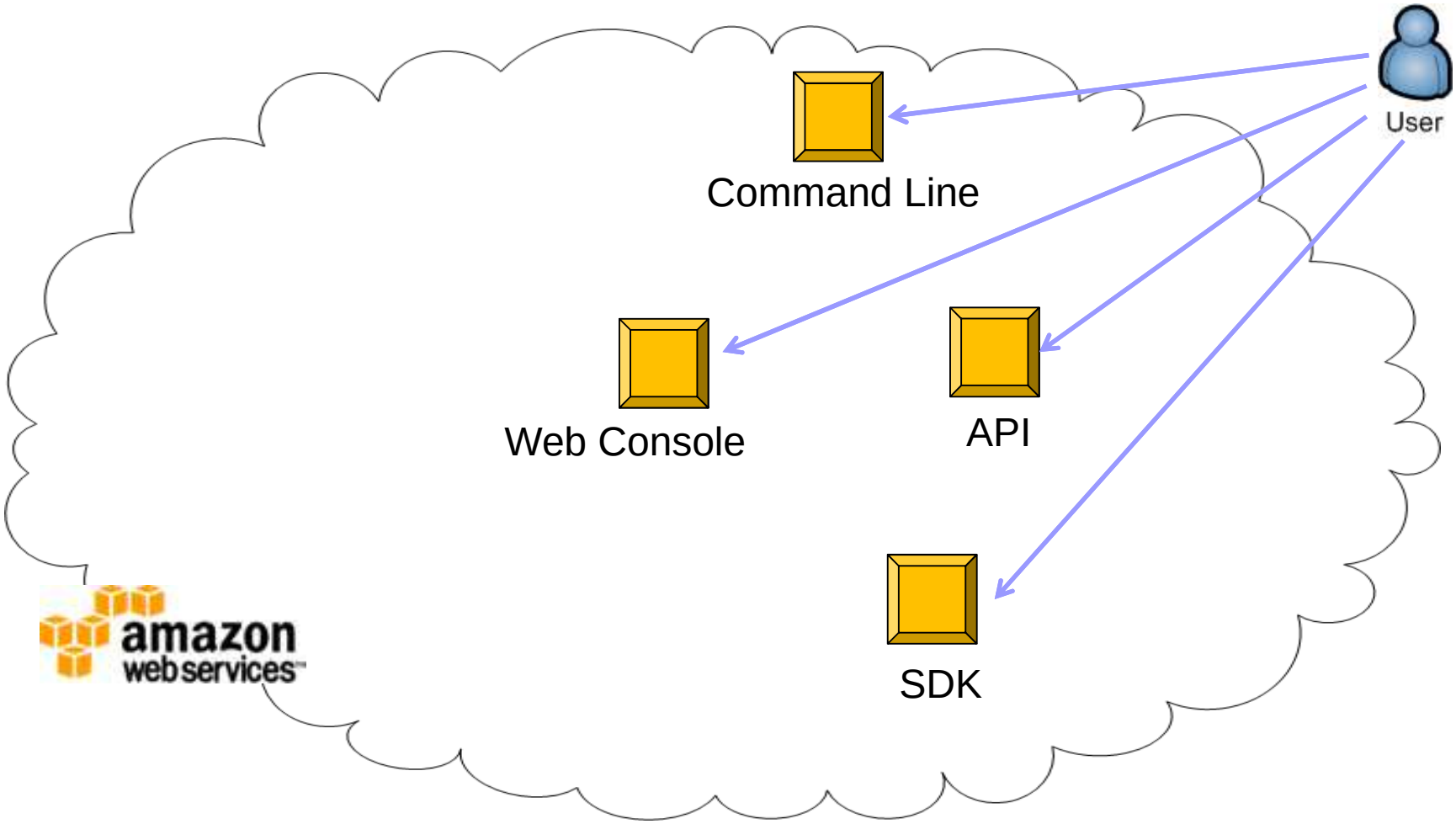


Elastic-R: Design and technologies overview

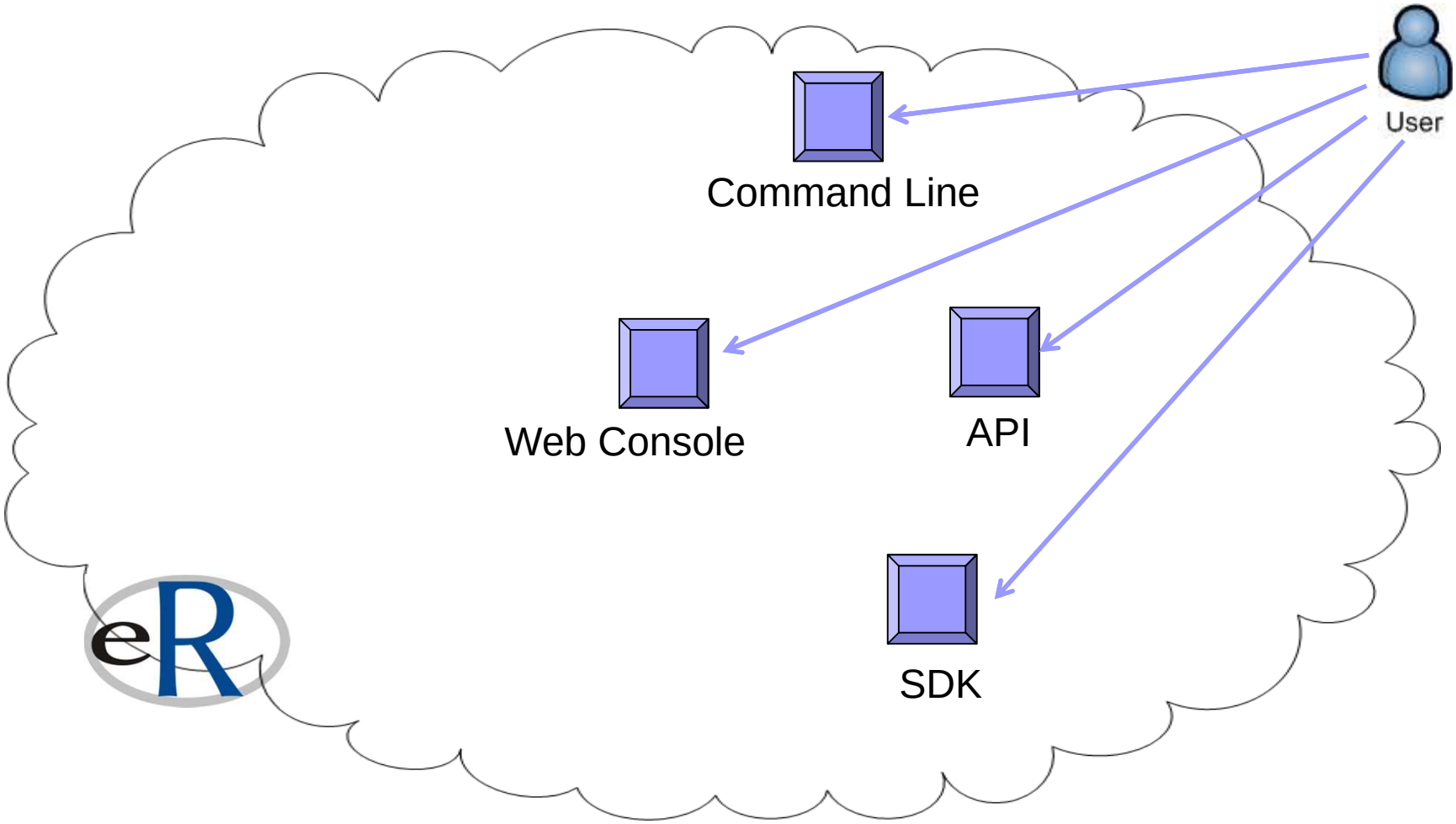
Modes of access

- Individuals with AWS accounts
  - Standard AMIs (Amazon Machine Images): paid-per-use to Amazon. For Academic use and trial purposes
  - Paid AMI: Paid-per-use software model. For business users
- Individuals without AWS accounts
  - Trial tokens, purchased tokens, tokens granted by other users.
  - Resources (data science engines) shared by other users
  - Individual subscribtions
- Companies/Educational & Research Institutions
  - Dedicated Platform and AMIs on an Amazon VPC (Virtual Private Cloud). Paid via subscribtion

Elastic-R: The scriptability framework



Elastic-R: The scriptability framework



Elastic-R: The scriptability framework

Script / globals.r

square ← function(x) {return(x^2) }
typeInfo(square) ← SimultaneousTypeSpecification(TypedSignature(x = "numeric"), returnType = "numeric")

Script / rjmap.xml

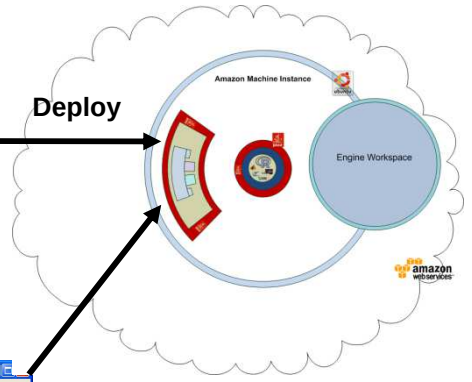
```
<rj>
  <publish>
    <functions> <function name="square" forWeb="true"/> </functions>
  </publish>
  <scripts>  <initScript name="globals.r" embed="true"/> </scripts>
</rj>
```

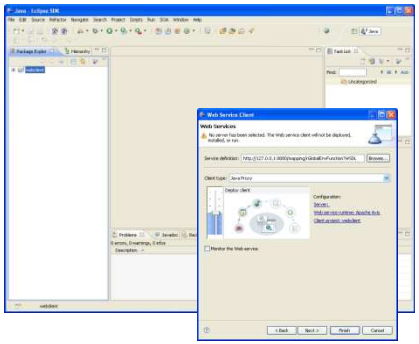
WS generator

rws.war

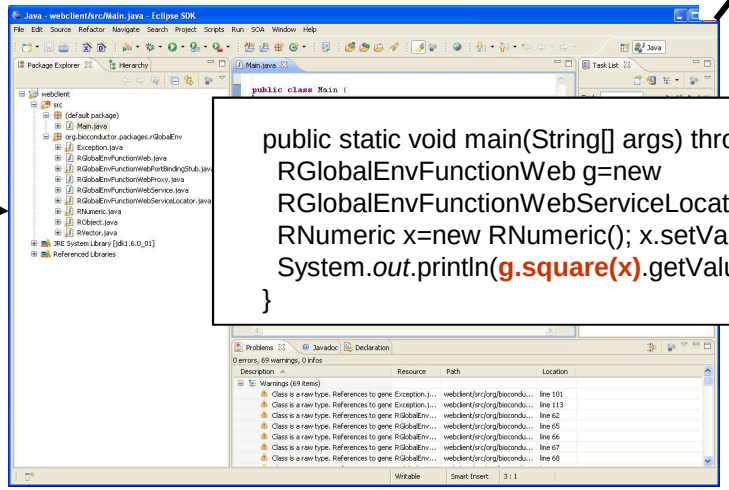
+ mapping.jar
+ pooling framework
+ R Java Bridge
+ JAX-WS
  - Servlets
  - Generated artifacts

Deploy





Eclipse Web Service Client Generator



```
public static void main(String[] args) throws Exception {
  RGlobalEnvFunctionWeb g=new
  RGlobalEnvFunctionWebServiceLocator().getrGlobalEnvFunctionWebPort();
  RNumeric x=new RNumeric(); x.setValue(new Double[]{6.0});
  System.out.println(g.square(x).getValue()[0]);
}
```

Elastic-R: The scriptability framework

- API: Soap and Restful Web Services
  - Data analysis engines control
  - Data analysis engines management (life cycle, etc.)
  - Virtual appliances/artifacts management
  - Platform administration
  - Generated web services from R functions
- SDKs: Java, R, Microsoft Office (Vba)
- Command line: elasticR package
- Html5 Workbench
  - Elastic-R artifacts management interface
  - Engines control interface

Demo

- Register to Elastic-R academic and trial portal (www.elastic-r.org)
- Create data science engines using trial tokens
- Work with R, Python and scientific Spreadsheets in the browser
- Share Data Science Engine and Collaborate
- Use The Visual and Collaborative Scientific Applications designer to create and publish to the web an interactive dashboard
- Connect to the remote Data Science Engine from withing a local R session, push and pull data, execute commands and show impact on the dashboard

Conclusion

- Elastic -R unlocks the potential of the cloud for Data scientists and educators
- With Elastic-R, the cloud becomes a cyberspace for collaborative research and sharing and an eco-system suited for open Science, open innovation and open education
- Elastic-R improves dramatically the productivity of the data scientists: The entire data science factory chain, from resources acquisition to services and applications publishing, becomes under their direct control
- Elastic-R provides Analytics-as-a-Service platform that can extend any existing portal or application

What to do Next

- Register to Elastic-R and try the HTML 5 Workbench and the collaboration
- Download the R package elasticR and use it to access the cloud from local R sessions
- Download the Java SDK and try to create your first Analytical application using AWS and the most advanced tools for programming with data.
- Get in touch with me to explore potential collaborations

Contact details

- Karim Chine
- karim.chine@cloudera.co.uk