

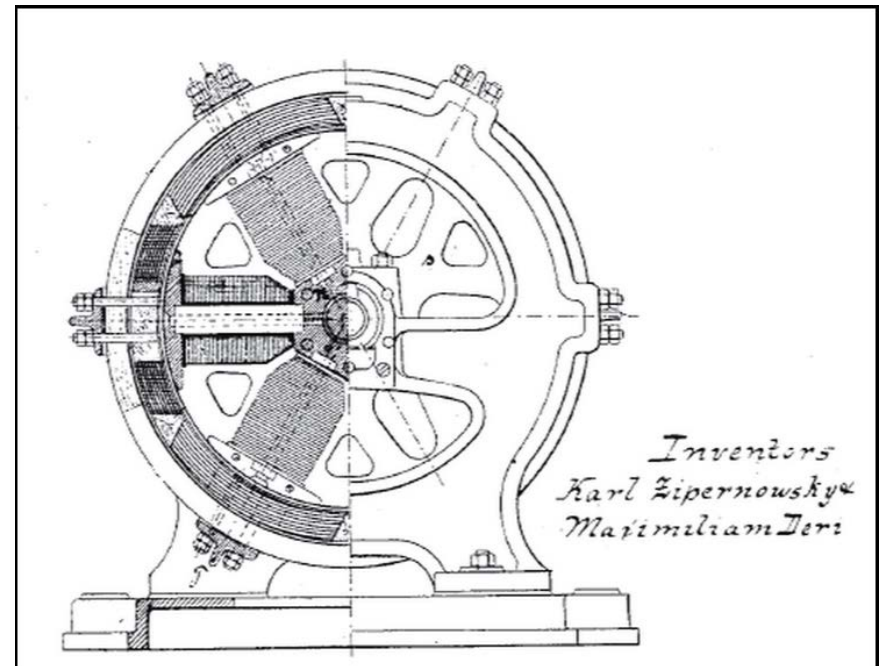


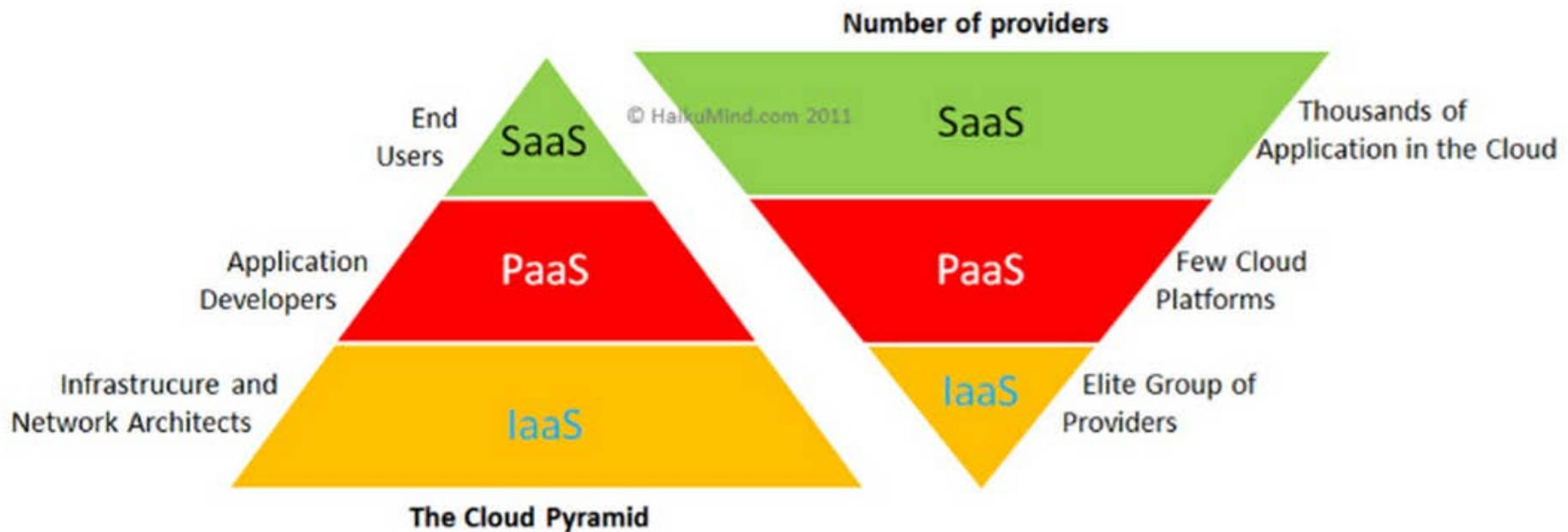
Data Science in the Cloud

Andy Catlin

"Cloud computing today is what the electric grid was 100 years ago."

-- Nicholas Carr, *The Big Switch*





Source: <http://www.haikumind.com/cloud-computing-acronyms-iaas-paas-and-saas/>

Why do our data science work in the cloud?

Why do our data science work in the cloud?

- R
- Machine Learning
- Hadoop
- Operationalizing Processes
- Other?

- **R**

- Memory bound

- **Machine Learning**

- <http://www.computing.co.uk/ctg/news/2408008/microsoft-versus-amazon-microsoft-has-the-edge-in-machine-learning-argues-forresters-mike-gualtieri>

- **Hadoop**

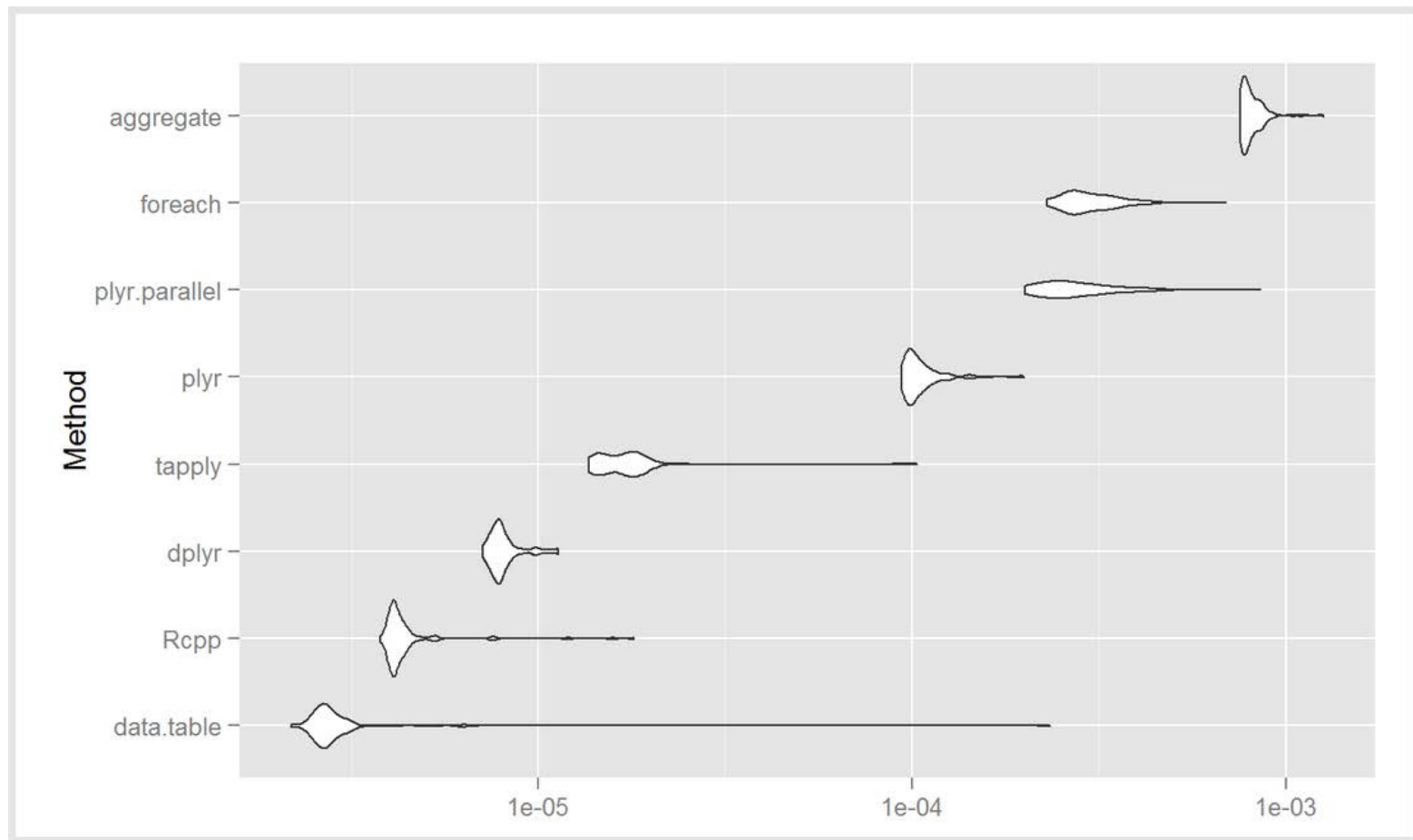
- Distributed Storage

- Distributed Processing

- “Code to Data” instead of data to code

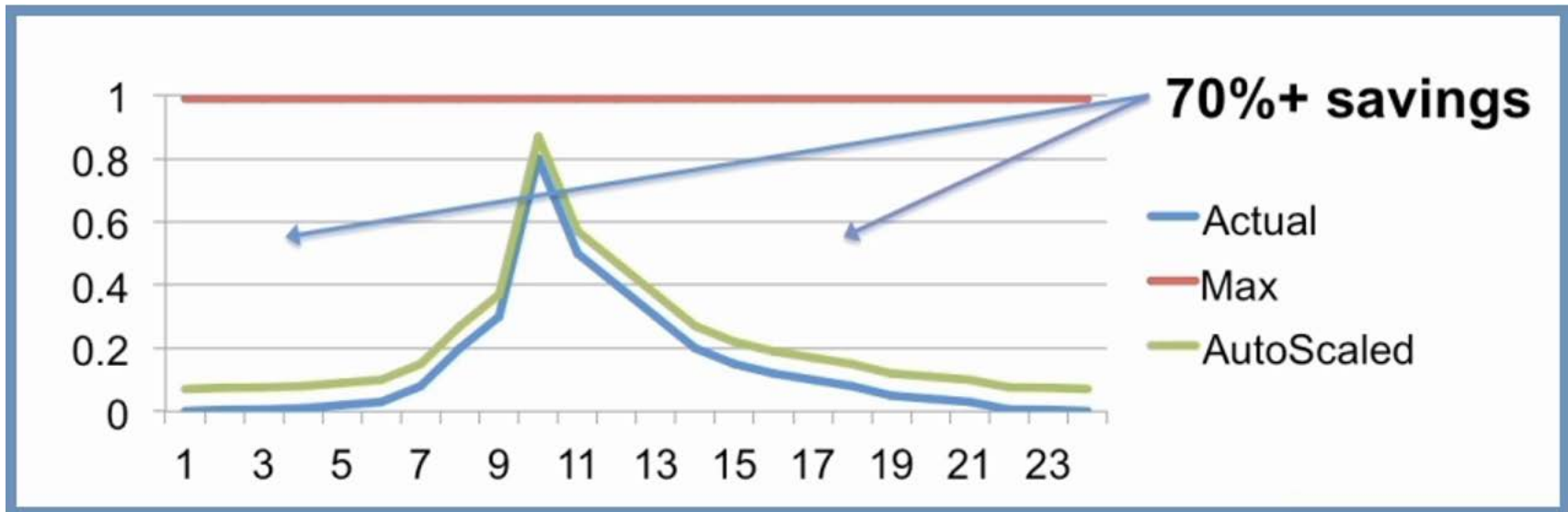
- **Operationalizing Processes**

• R Performance



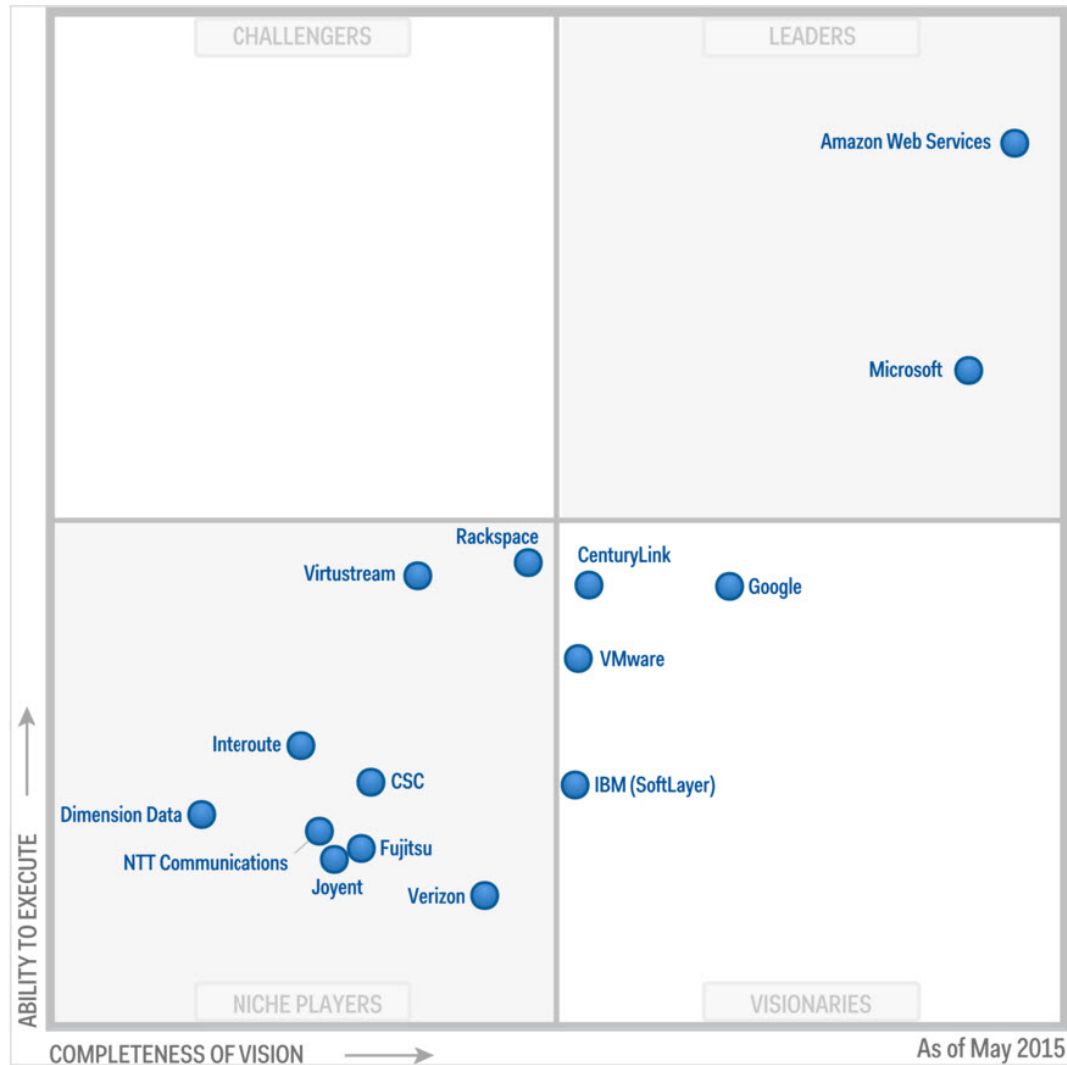
Source: <http://jaredlander.com/content/2015/04/MakingRGoFasterAndBigger.html#1>,
Jared Lander, New York 2015.

- Operationalizing Processes
 - Server Administrator efficiencies
 - Smoothing Demand curve



Source: "Cloud Computing with AWS," Rich Morrow, Infinite Skills, Jul 24, 2013.

Which is the Best Cloud Service Provider?



Source: Gartner Magic Quadrant for Cloud Service Providers, May 2015.

<https://aws.amazon.com/resources/gartner-2015-mq-learn-more/>

Demonstration: Setting up RStudio on Amazon AWS

- "RStudio Server Amazon Machine Image (AMI)," http://www.louisaslett.com/RStudio_AMI/
- "RStudio in the cloud, for dummies," Ken Kleinman, <http://www.r-bloggers.com/rstudio-in-the-cloud-for-dummies/>. Feb 13, 2012.
- "Connecting to Your Linux Instance from Windows Using PuTTY," <https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/putty.html>

- Best practice? Beefy local developer's workstation + Linux server in Cloud
- Discussion