Tools for Data Science

Vadim Y. Bichutskiy
@vybstat

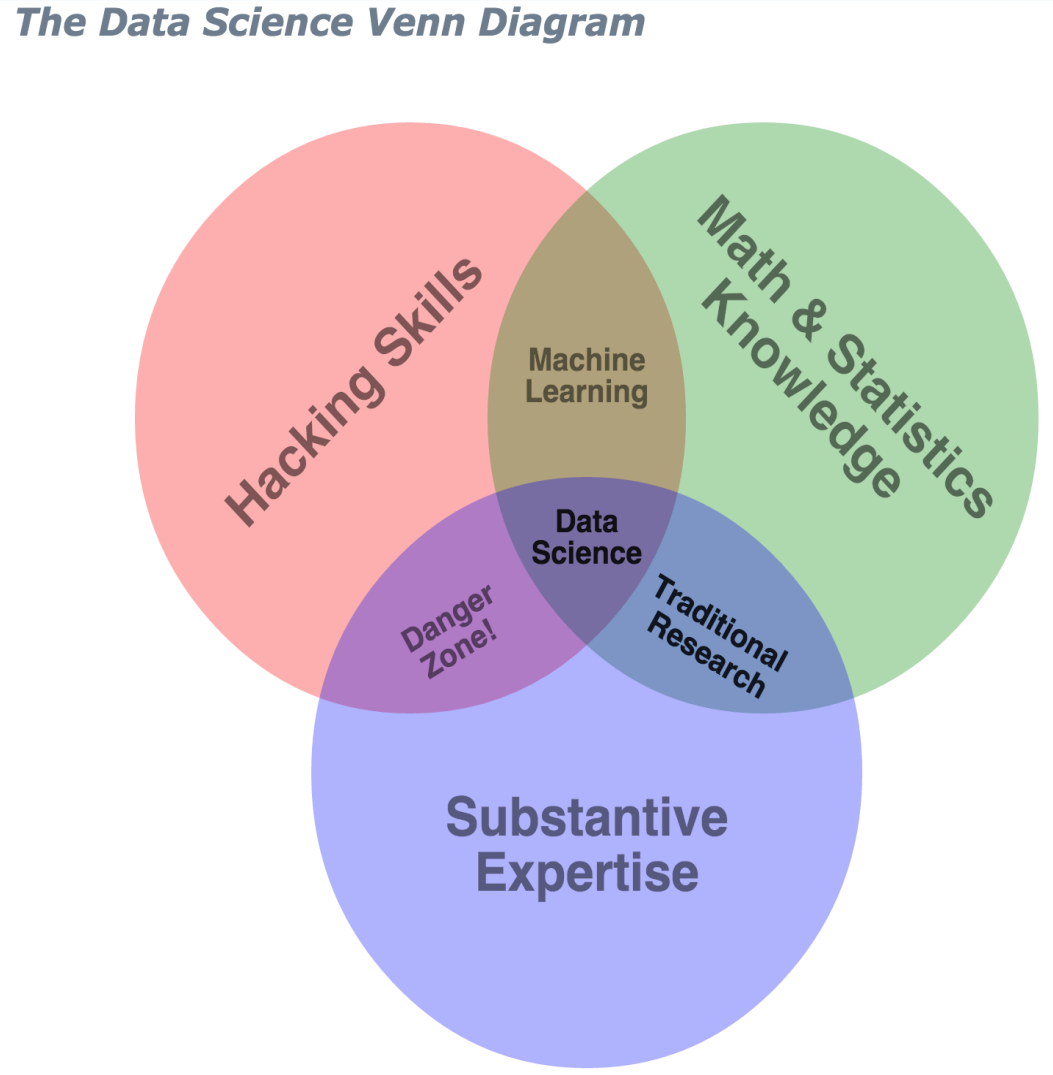
Data Science Seminar, GMU
April 10, 2015

So you want to be a data scientist?

- Good news
  - Data is everywhere
  - “Big Data”, “Analytics”, “Data Science” is changing the world
  - Hot and sexy
  - Lots of opportunity to get creative and innovate
  - Many open problems
  - Fun!
  - Demand is off the charts / low supply
  - High salaries
- Bad news
  - Requires lots of education: PhD is NOT enough
  - Can be overwhelming and stressful
  - Theory, practical tools, experience
  - Long working hours
  - Not enough sleep
  - Bad for health?
  - Versatile, flexible, curious
  - Continuous training

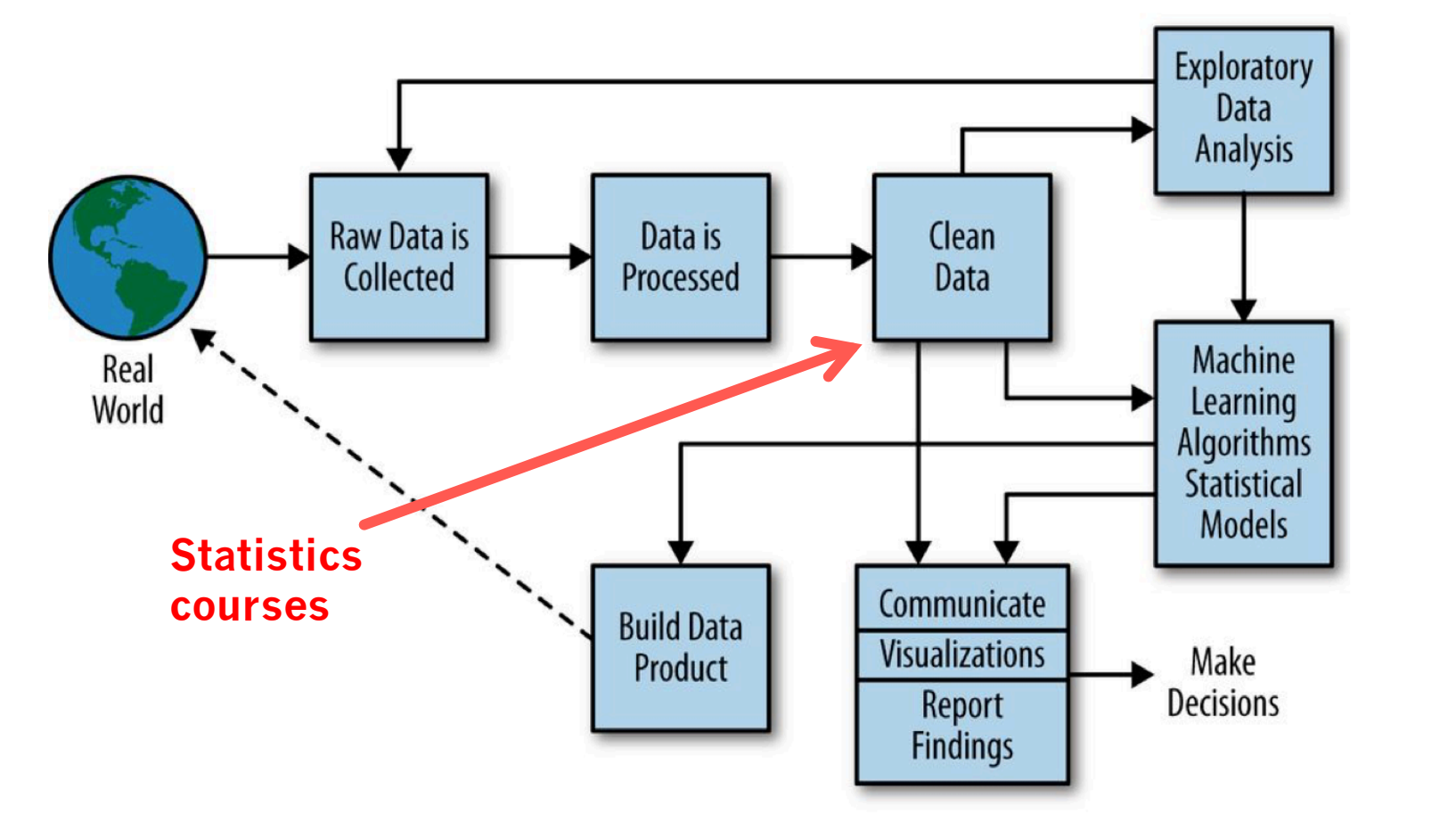
https://www.whitehouse.gov/blog/2015/02/18/white-house-names-dr-dj-patil-first-us-chief-data-scientist

What's Data Science?



http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

Data scientists: "Create order from chaos"



O'Neil, Cathy and Schutt, Rachel, *Doing Data Science: Straight Talk from the Frontline*, O'Reilly, 2014

Data collection, processing, cleaning is 80% of the effort

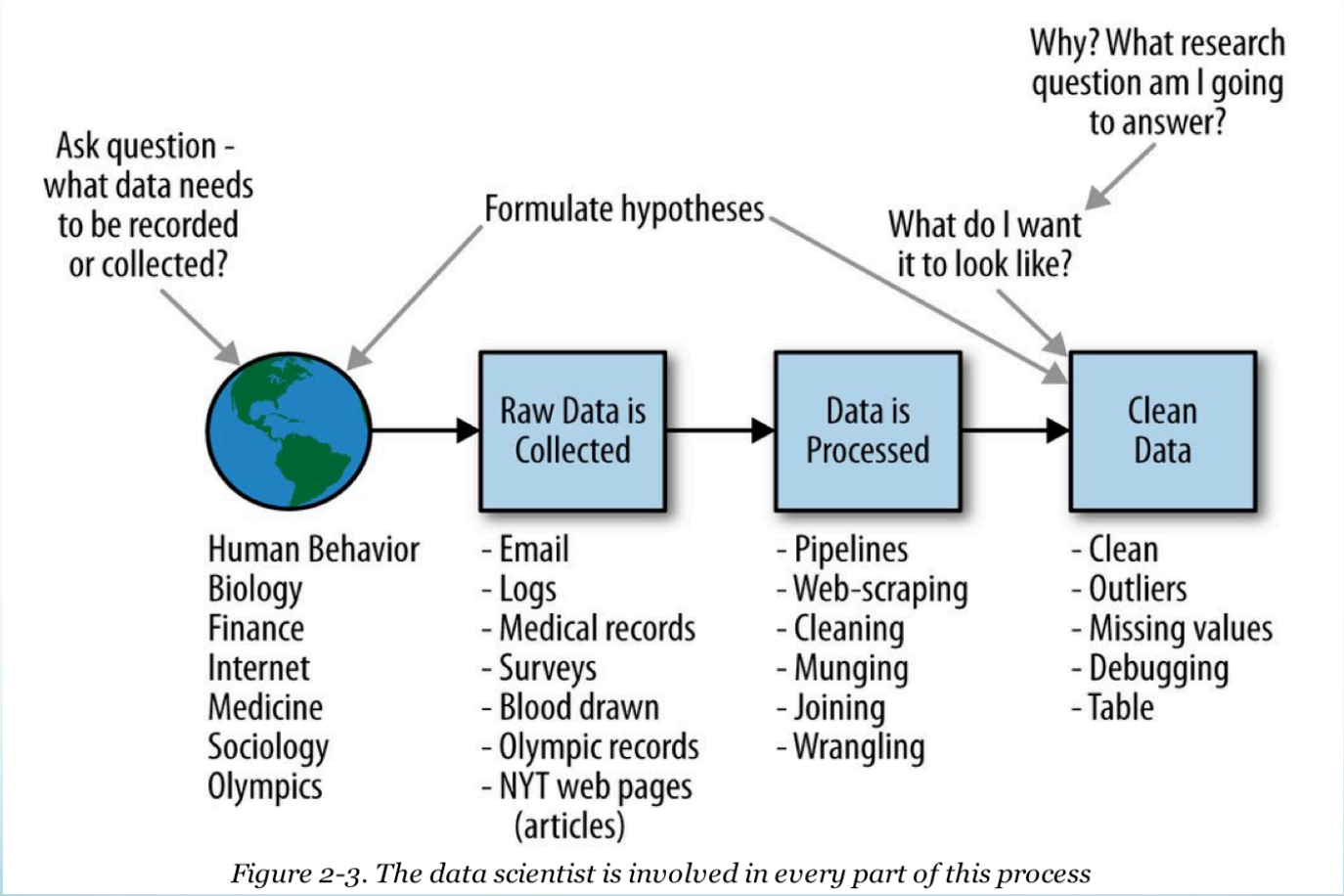
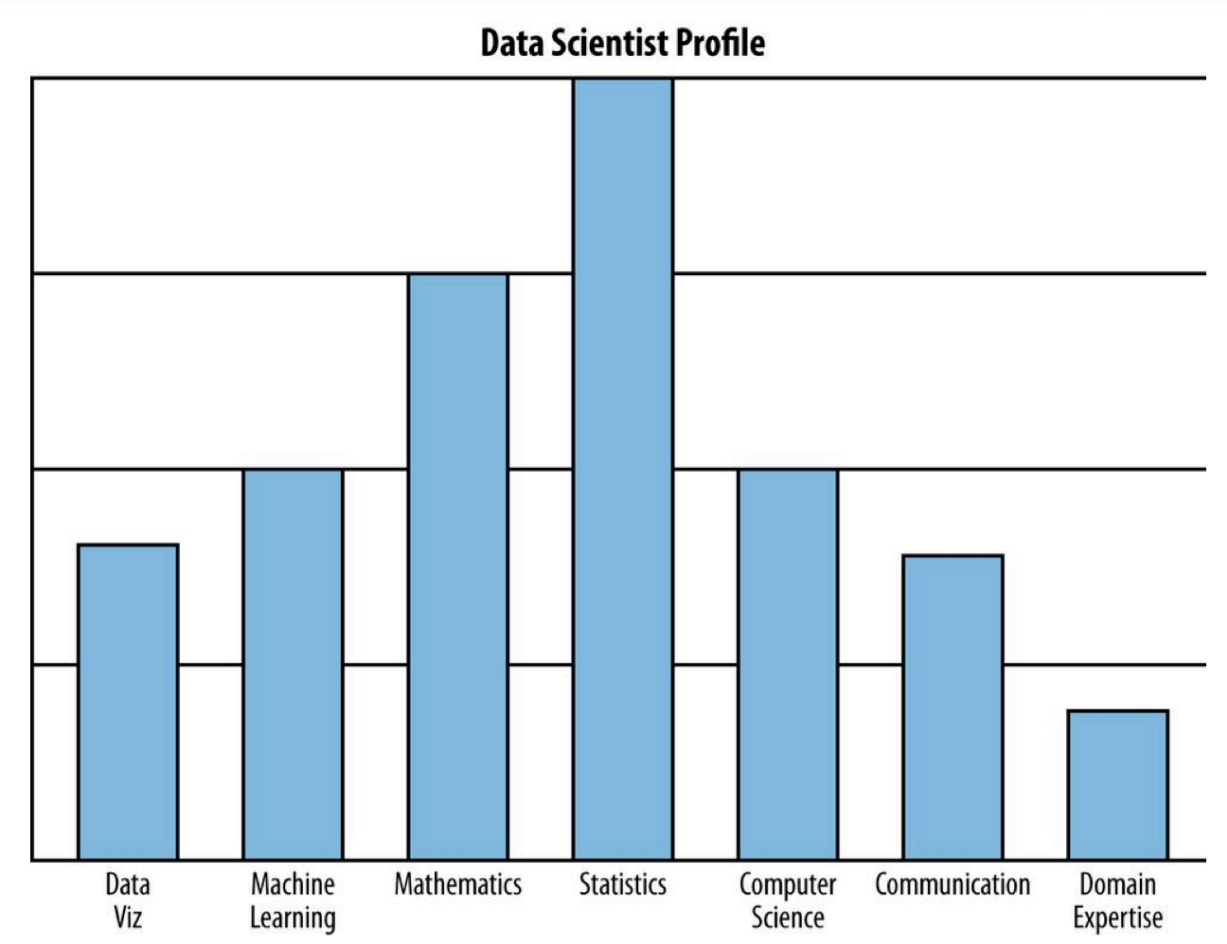


Figure 2-3. The data scientist is involved in every part of this process

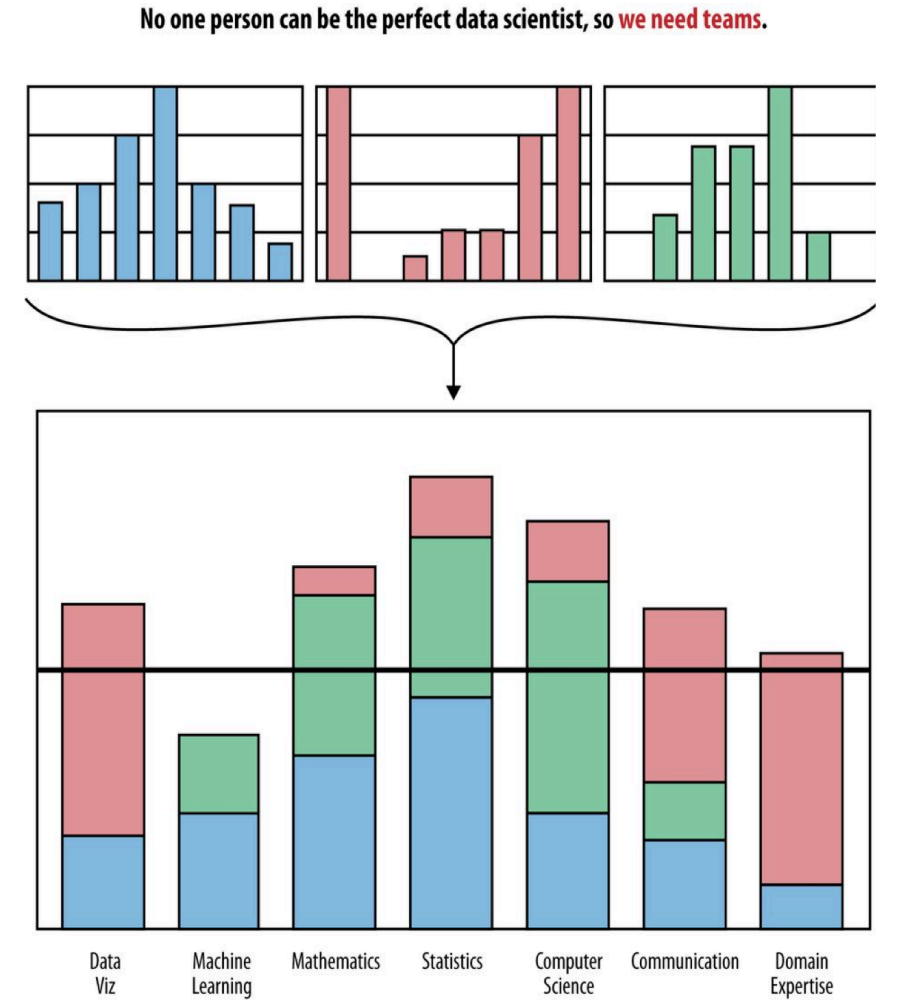
O'Neil, Cathy and Schutt, Rachel, *Doing Data Science: Straight Talk from the Frontline*, O'Reilly, 2014

Stats/CSI PhD



O'Neil, Cathy and Schutt, Rachel, *Doing Data Science: Straight Talk from the Frontline*, O'Reilly, 2014

“Data science is a team sport” --DJ Patil



O'Neil, Cathy and Schutt, Rachel, *Doing Data Science: Straight Talk from the Frontline*, O’Reilly, 2014

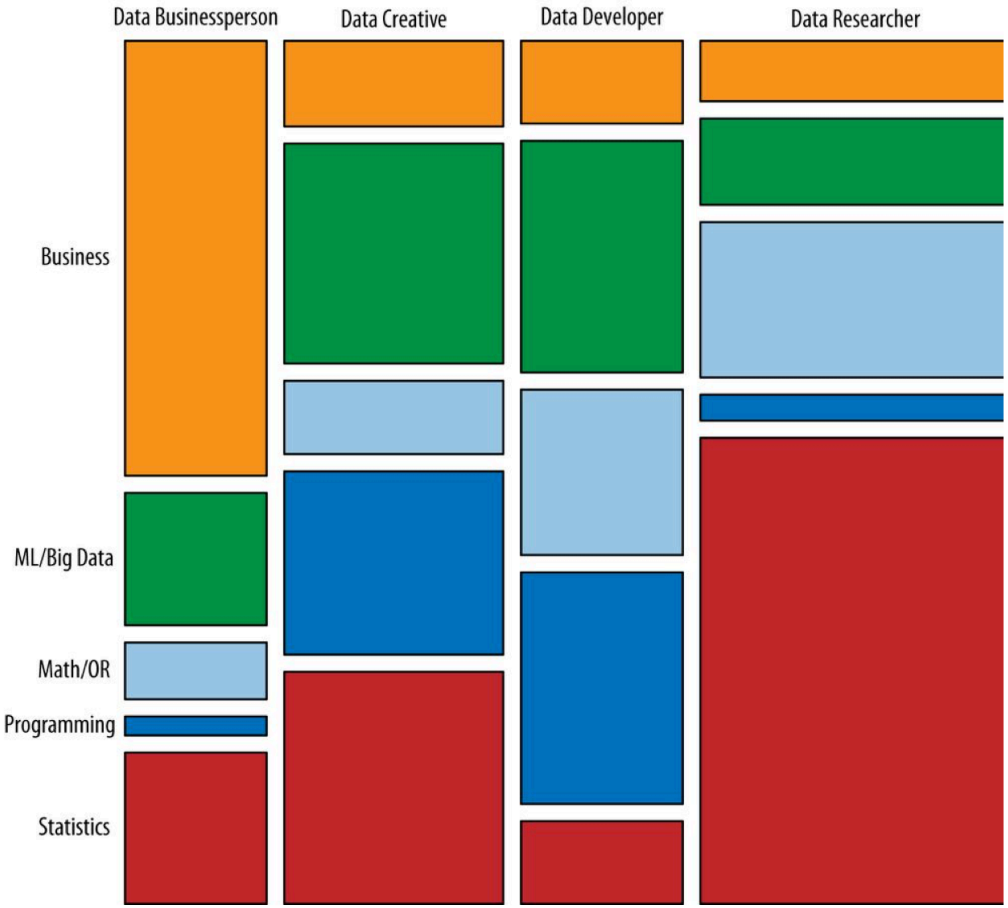
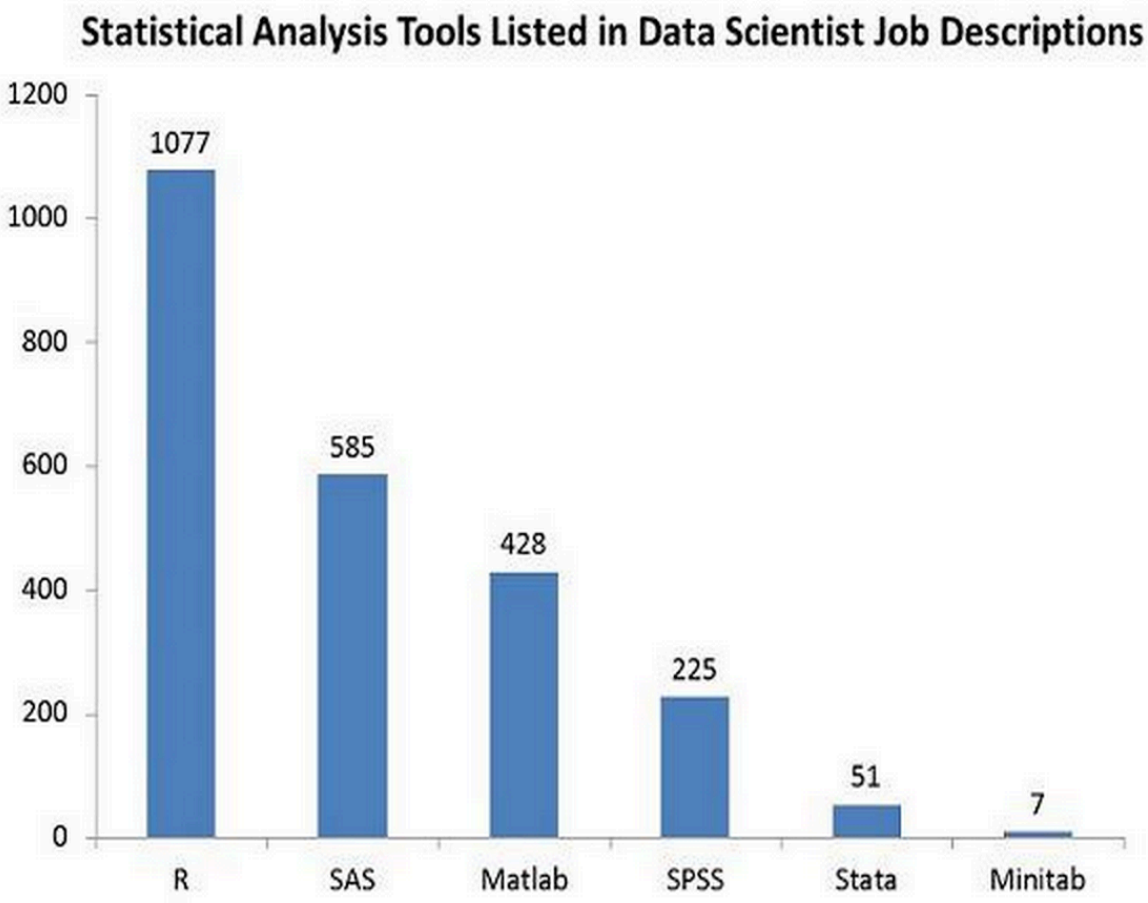


Figure 1-4. Harlan Harris's clustering and visualization of subfields of data science from Analyzing the Analyzers (O'Reilly) by Harlan Harris, Sean Murphy, and Marck Vaisman based on a survey of several hundred data science practitioners in mid-2012

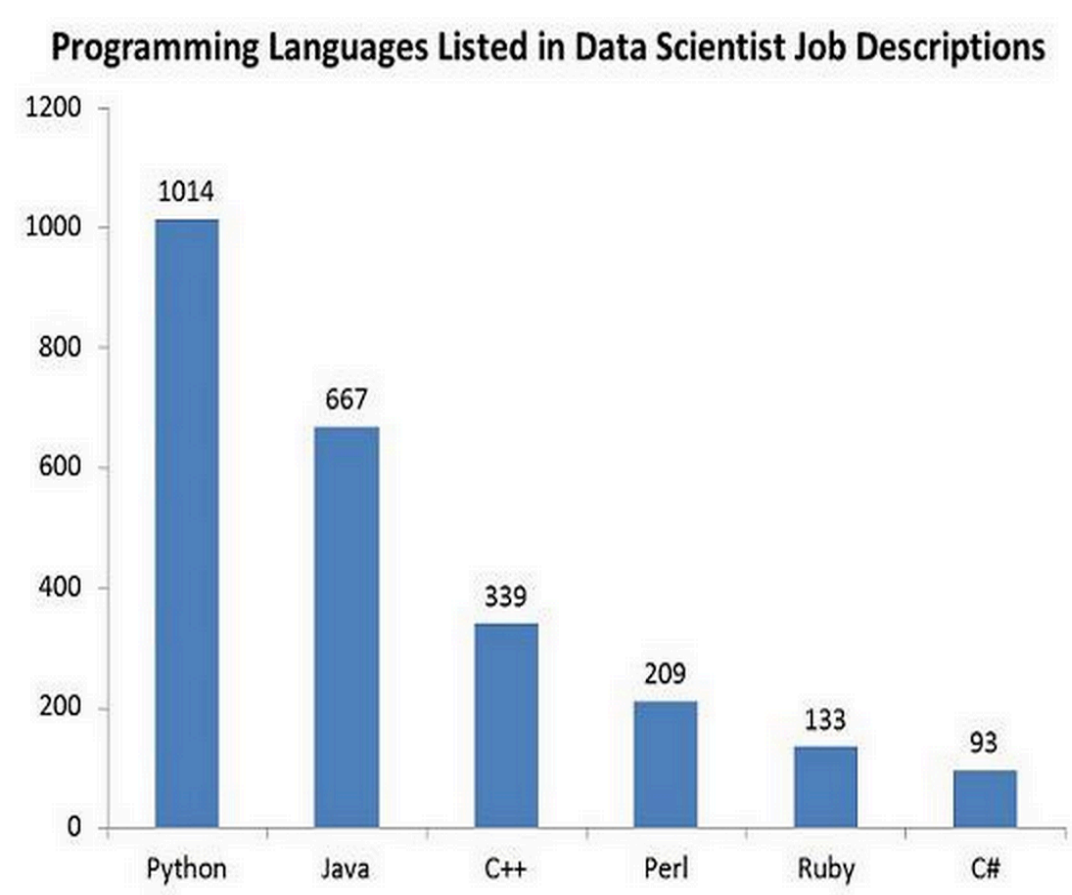


Kirk Borne @KirkDBorne · Apr 3

Tech Skills that #DataScientist Jobs Require: bit.ly/1DHBGbe #abdsc #BigData #Python #Rstats @DataScienceCtrl

72 54

http://www.datasciencecentral.com/profiles/blogs/what-technology-tool-skills-do-data-scientists-jobs-require



Kirk Borne @KirkDBorne · Apr 3

Tech Skills that #DataScientist Jobs Require: bit.ly/1DHBGbe #abdsc #BigData #Python #Rstats @DataScienceCtrl

72 54

http://www.datasciencecentral.com/profiles/blogs/what-technology-tool-skills-do-data-scientists-jobs-require

Data Science Skills

- Core
  - R
  - Python
  - SQL/NoSQL/database concepts
  - Unix command line
  - Statistics/machine learning/CS
  - Graph Theory/Networks
  - Data visualization/dashboards: Tableau, D3
  - Data representation: JSON, XML
  - Communication, domain expertise
- Project/position dependent
  - Java/C++
  - Amazon Web Services/Cloud computing
  - Hadoop
  - JavaScript/PHP/Web frameworks
- Emerging
  - Scala
  - Swift
  - Spark/Cluster computing
  - Real-time Analytics
  - Docker, Vagrant

Tools Usage

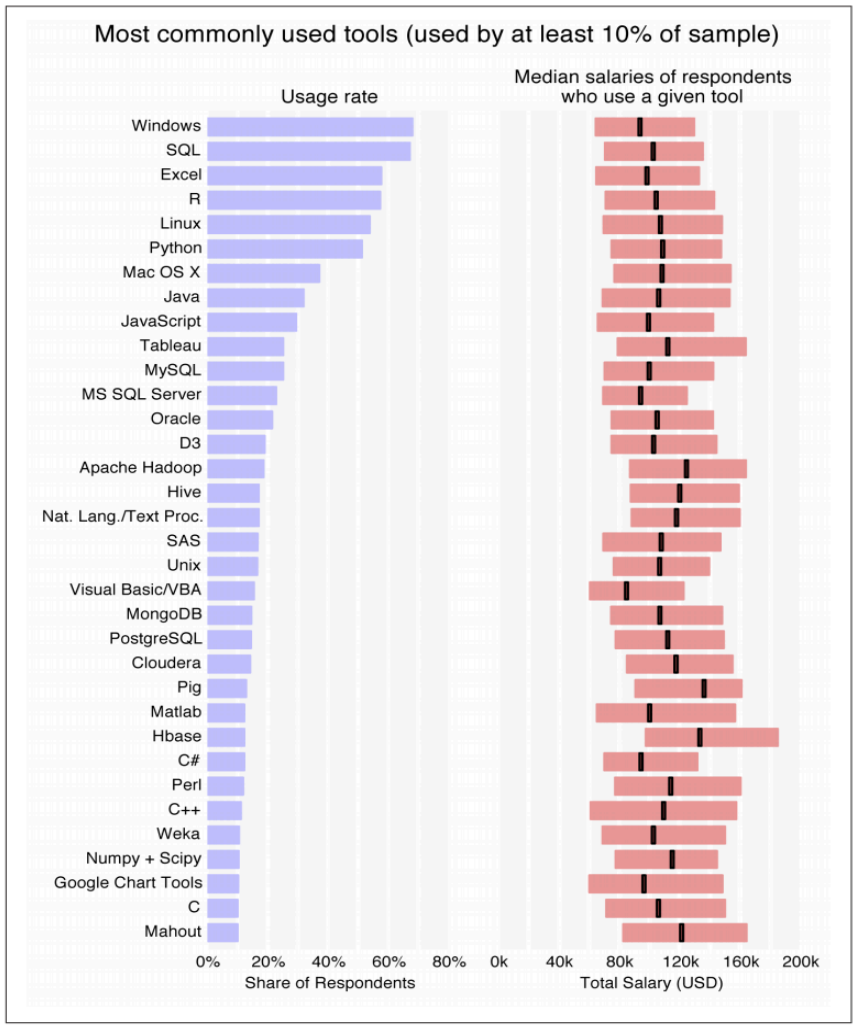


Figure 1-10. Most commonly used tools

http://www.oreilly.com/data/free/2014-data-science-salary-survey.csp

Tool Salaries

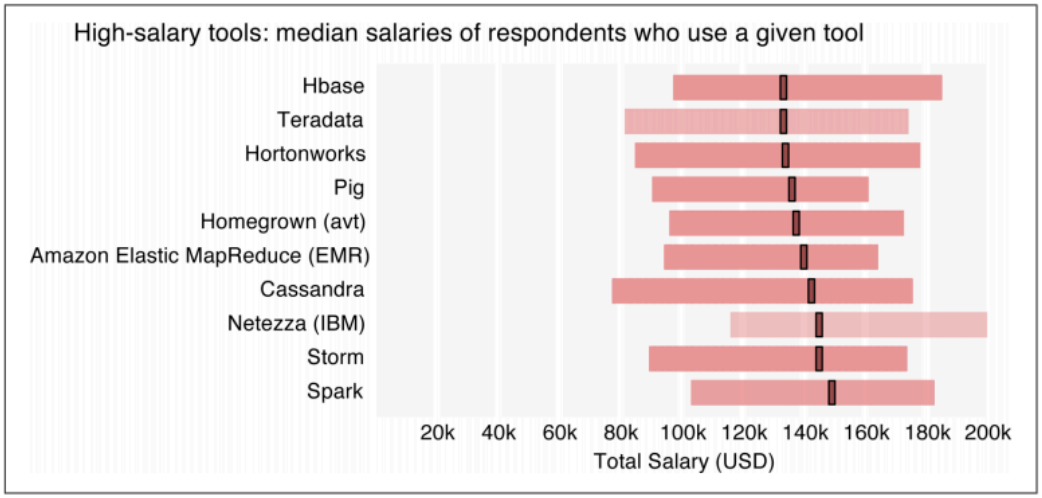


Figure 1-11. High-salary tools: median salaries of respondents who use a given tool

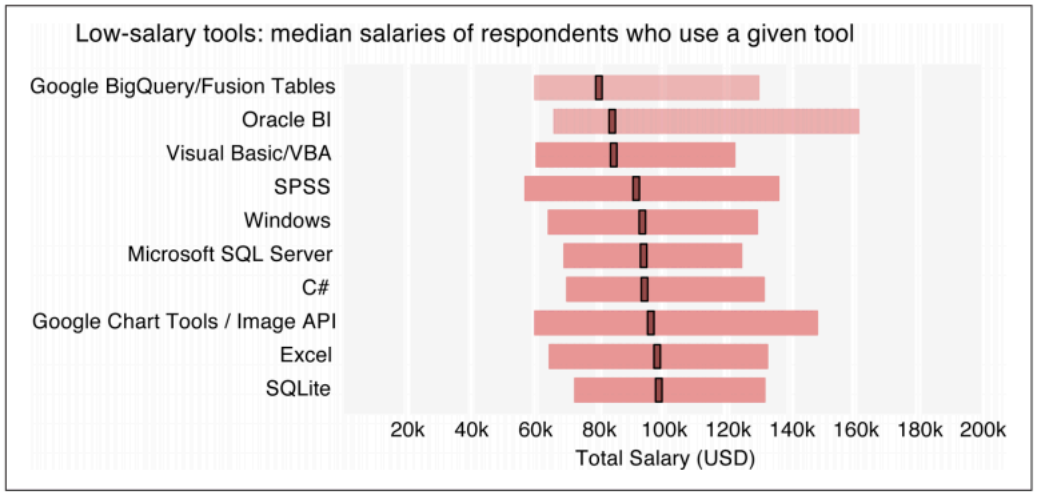
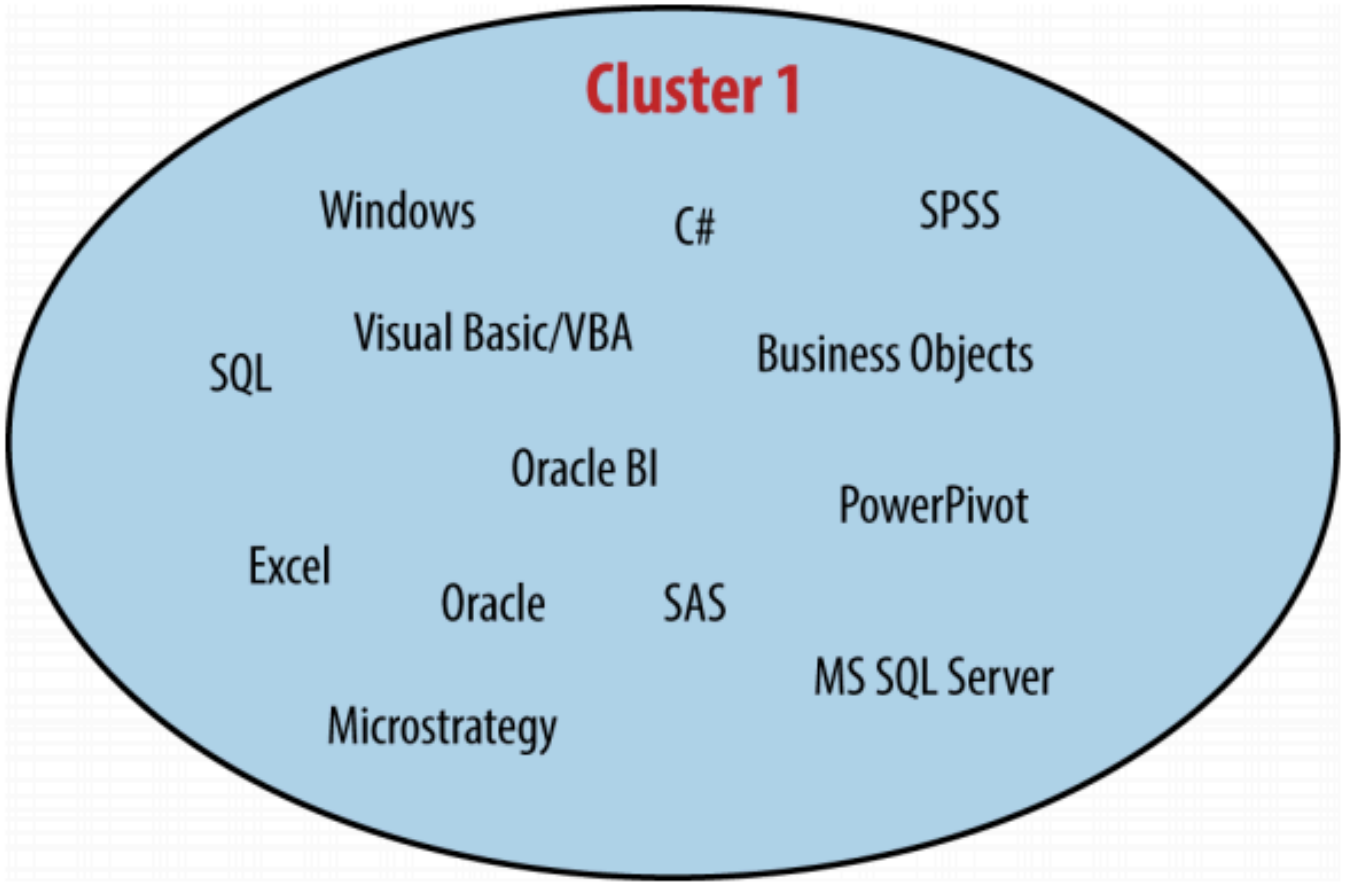


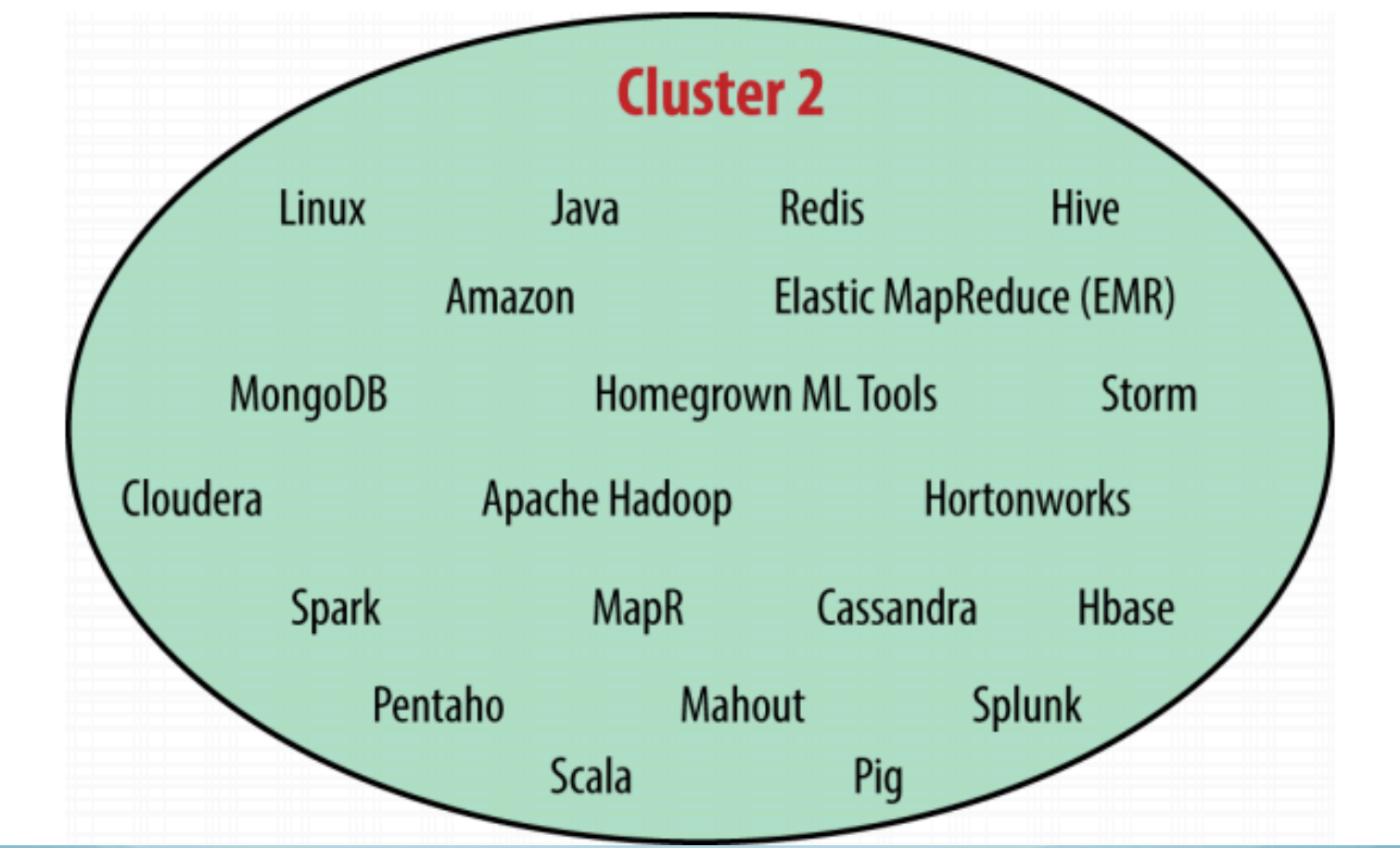
Figure 1-12. Low-salary tools: median salaries of respondents who use a given tool

http://www.oreilly.com/data/free/2014-data-science-salary-survey.csp

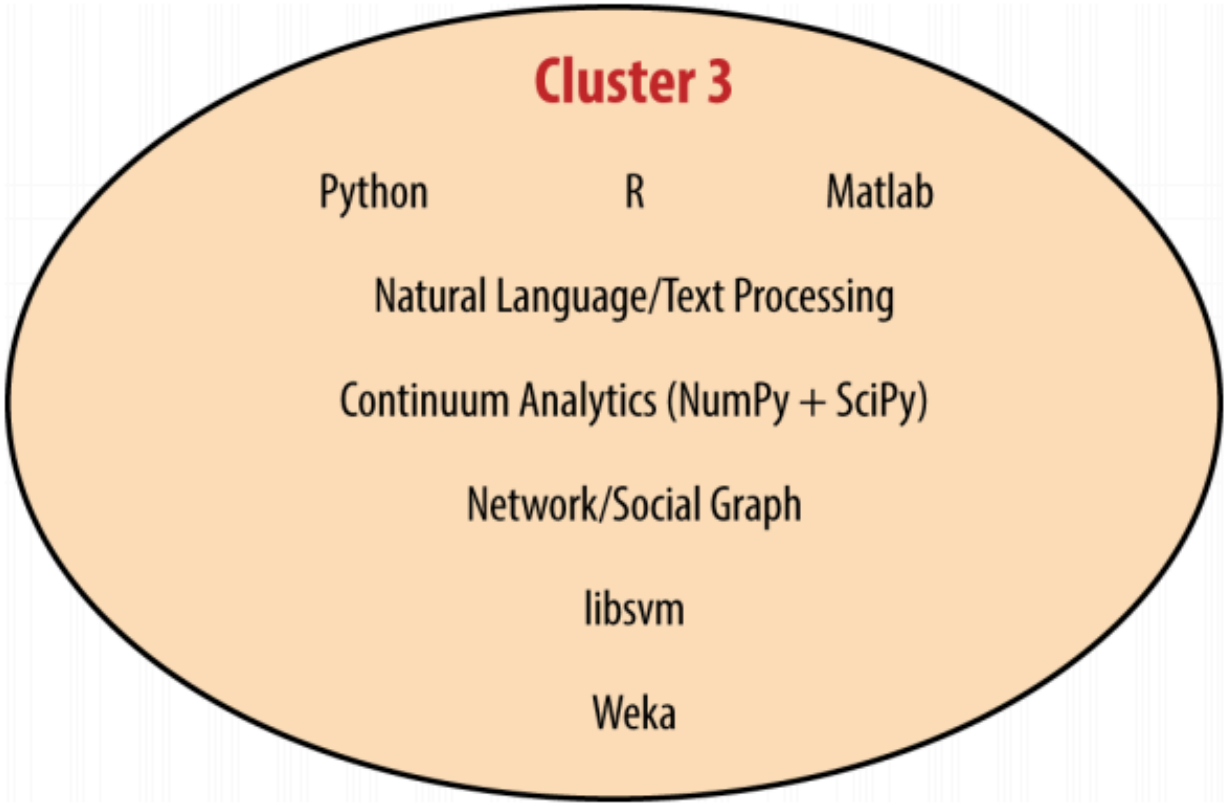
“Microsoft-Excel-SQL”



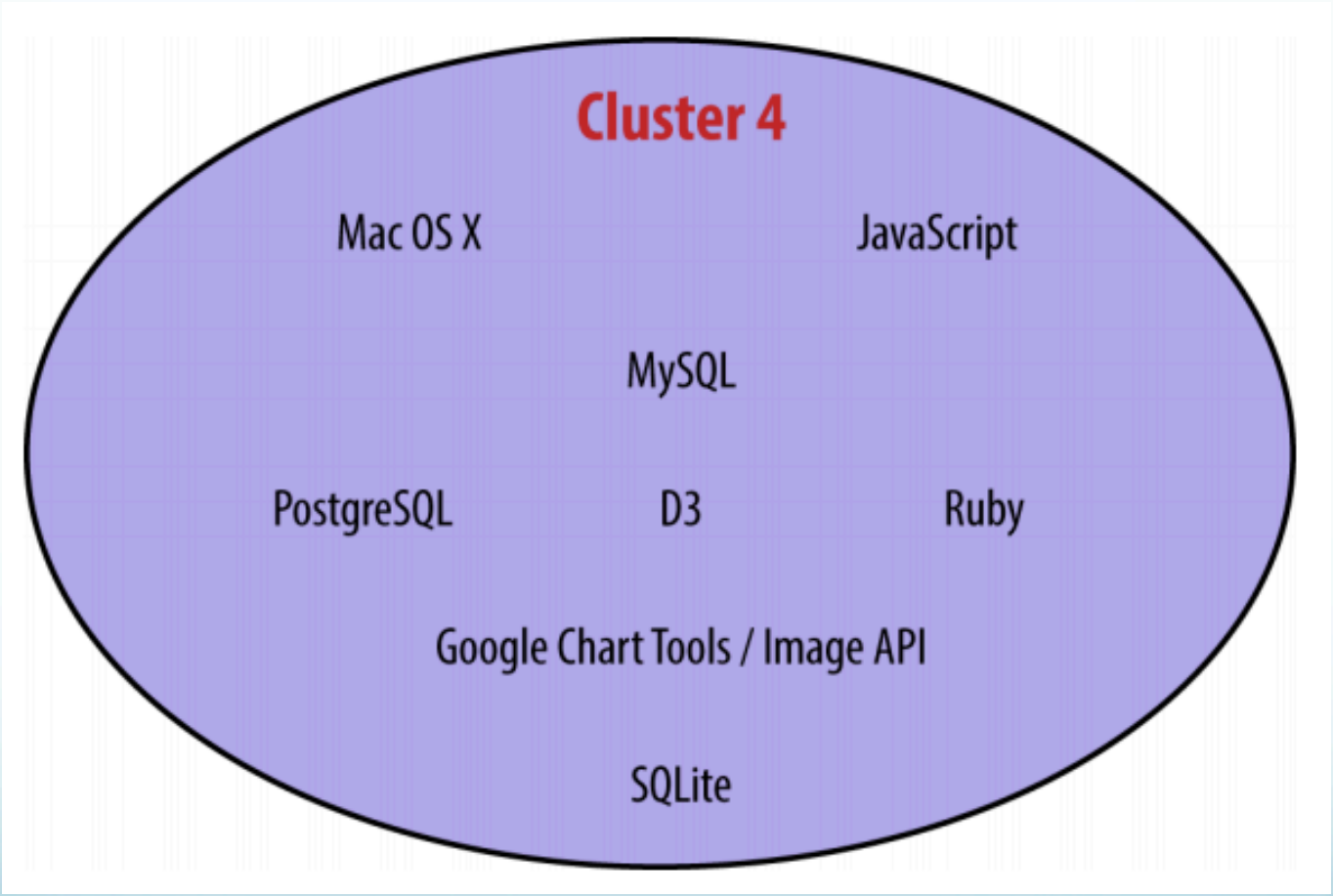
“Hadoop-Java-Cloud Computing”



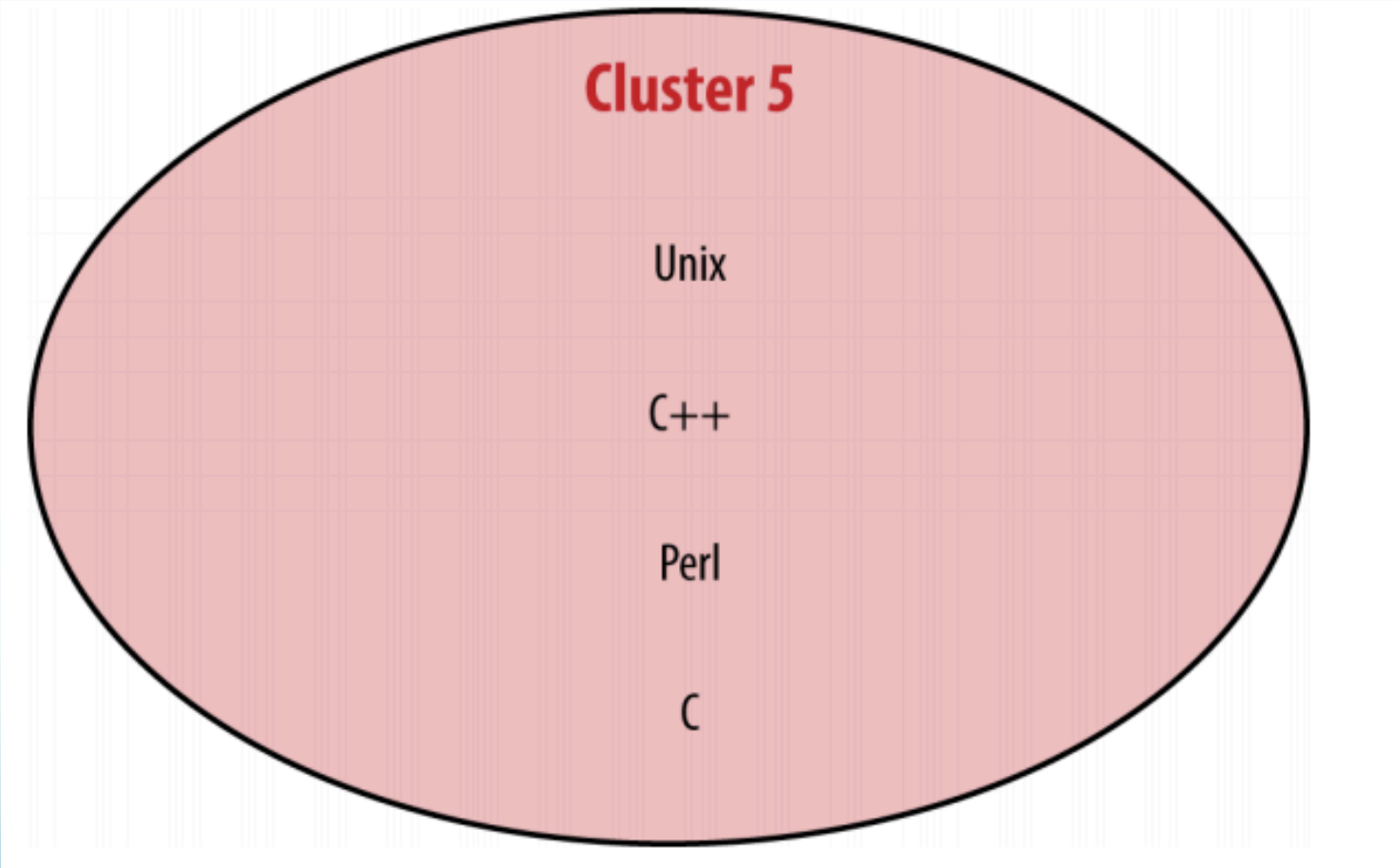
“R-Python-Analytics”



“MySQL-D3-JavaScript”



“Old tools”



Amazon MLaaS

AWS Official Blog

Amazon Machine Learning – Make Data-Driven Decisions at Scale

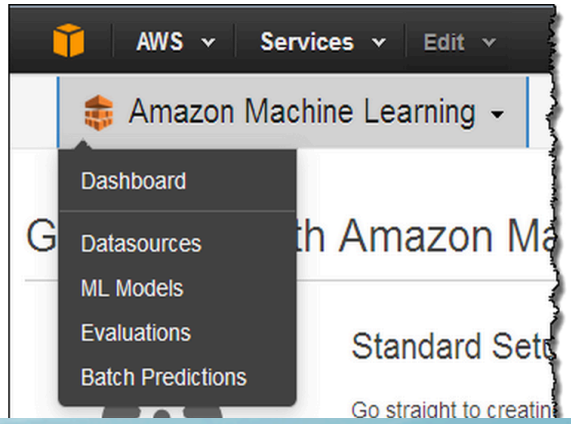
by Jeff Barr | on 09 APR 2015 | in Amazon Machine Learning | Permalink

Today, it is relatively straightforward and inexpensive to observe and collect vast amounts of operational data about a system, product, or process. Not surprisingly, there can be tremendous amounts of information buried within gigabytes of customer purchase data, web site navigation trails, or responses to email campaigns. The good news is that all of this data can, when properly analyzed, lead to statistically significant results that can be used to make high-quality decisions. The bad news is that you need to find data scientists with relevant expertise in machine learning, hope that your infrastructure is able to support their chosen tool set, and hope (again) that the tool set is sufficiently reliable and scalable for production use.

The science of Machine Learning (often abbreviated as ML) provides the mathematical underpinnings needed to run the analysis and to make sense of the results. It can help you to turn all of that data into high-quality predictions by finding and codifying patterns and relationships within it. Properly used, Machine Learning can serve as the basis for systems that perform fraud detection (is this transaction legitimate or not?), demand forecasting (how many widgets can we expect to sell?), ad targeting (which ads should be shown to which users?), and so forth.

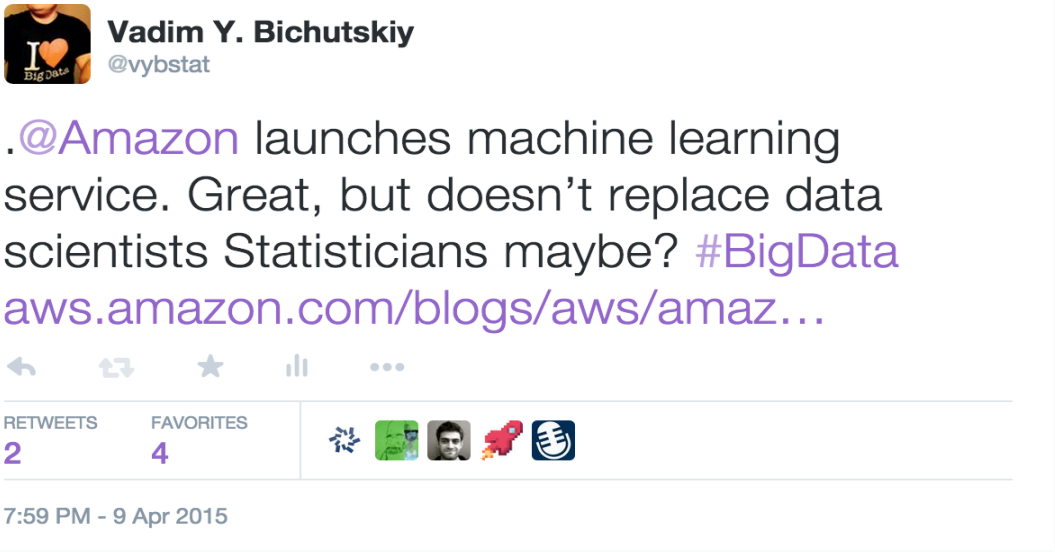
Introducing Amazon Machine Learning

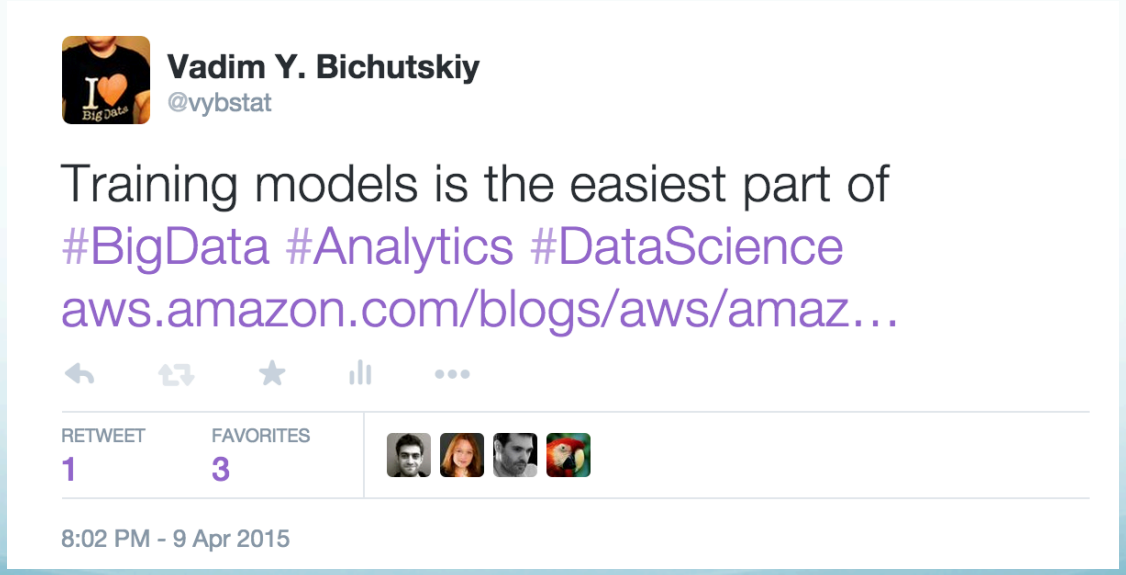
Today we are introducing Amazon Machine Learning. This new AWS service helps you to use all of that data you've been collecting to improve the quality of your decisions. You can build and fine-tune predictive models using large amounts of data, and then use Amazon Machine Learning to make predictions (in batch mode or in real-time) at scale. You can benefit from machine learning even if you don't have an advanced degree in statistics or the desire to setup, run, and maintain your own processing and storage infrastructure.



I'll get to the details in just a minute. Before I do so, I'd like to review some of the terminology and the concepts that you need to know in order to fully understand what machine learning does and how you can take advantage

http://aws.amazon.com/blogs/aws/amazon-machine-learning-make-data-driven-decisions-at-scale/


Vadim Y. Bichutskiy
@vybstat
.@Amazon launches machine learning service. Great, but doesn't replace data scientists Statisticians maybe? #BigData aws.amazon.com/blogs/aws/amaz…
RETWEETS
2
FAVORITES
4
7:59 PM - 9 Apr 2015


Vadim Y. Bichutskiy
@vybstat
Training models is the easiest part of #BigData #Analytics #DataScience aws.amazon.com/blogs/aws/amaz…
RETWEET
1
FAVORITES
3
8:02 PM - 9 Apr 2015

Resources (1)

- R
  - http://www.r-project.org/
  - http://www.rstudio.com/
- Python
  - https://www.python.org/
- JSON
  - http://json.org/
- Amazon Web Services
  - http://aws.amazon.com/
  - http://aws.amazon.com/blogs/aws/
- Hadoop
  - https://hadoop.apache.org/

Resources (2)

- Scala
  - http://www.scala-lang.org/
- Spark
  - https://spark.apache.org/
  - https://spark.apache.org/docs/latest/
- Docker
  - https://www.docker.com/
- Vagrant
  - https://www.vagrantup.com/
- Swift
  - apple.co/1CAAKQA