Data Science
the business point of view

Marius Marcu, November 2016

mariusmarcu@global.t-bird.edu

Outline

- Big data – the new natural resource
- The data scientists – the new, modern gold miners
- How you can make most of your Data Science opportunity

mariusmarcu@global.t-bird.edu

Big Data – the new natural resource

- **Observations** captured one by one and entered manually on paper or in the computer
- **Human activity** on the web leaves traces captured by various entities
- **Internet of things** – streams of data automatically captured from sensors or human activity into databases or sophisticated graphs

- Many “mountains” of data
  - Cost of storage low, # of devices/sensors higher every year
  - Reasons to store: financial vs other (competitive advantage)
  - Creating Data + Metadata (data about the data)

90% of all of the world’s data generated in the last two years

Source: Accenture white paper 2013

40

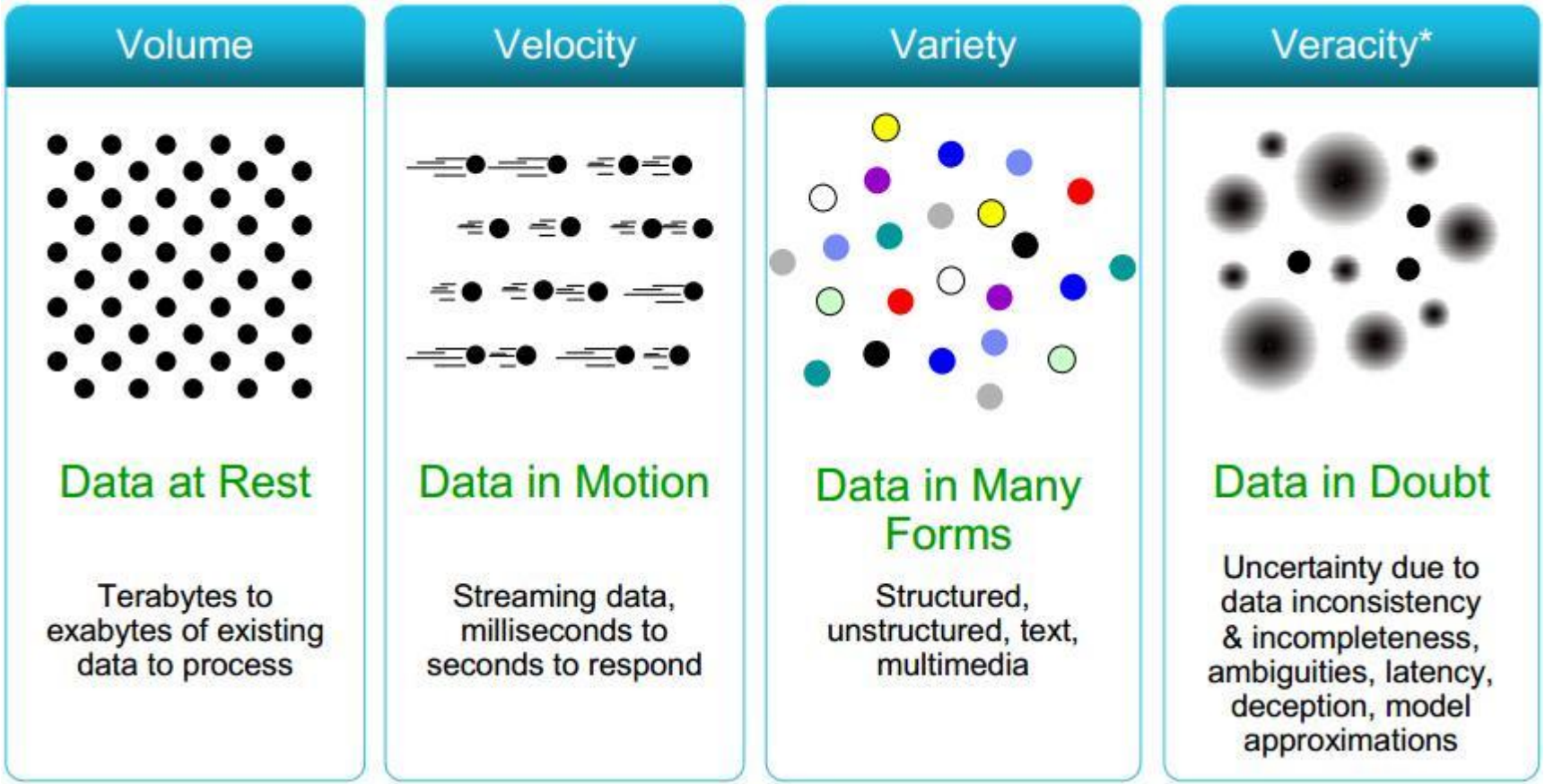
TRILLION GIGABYTES

Size of digital universe by 2020, up from 130 billion in 2005.

Source: IDC Digital Universe study, April 2014

mariusmarcu@global.t-bird.edu

Big Data characteristics

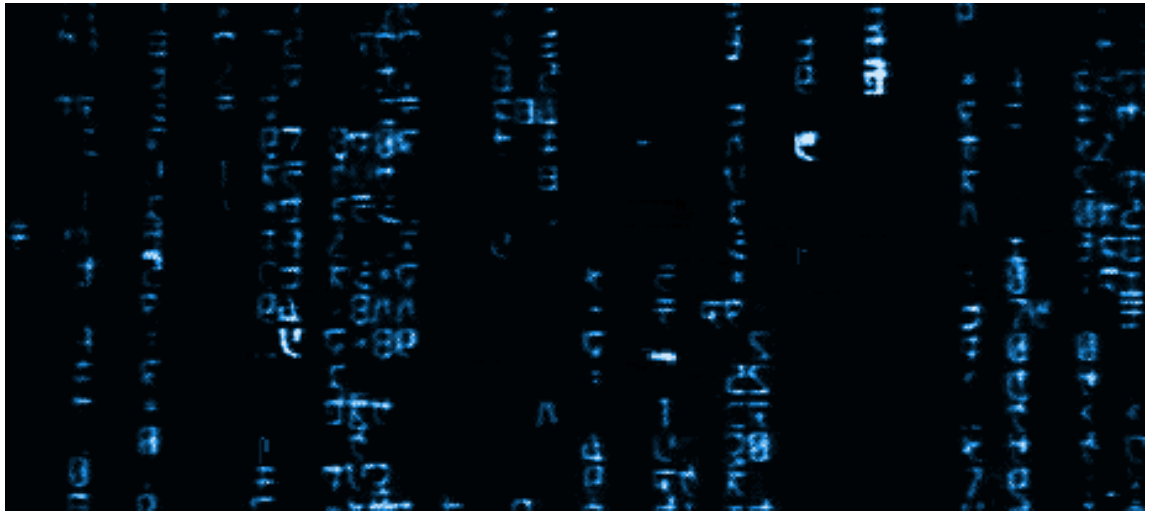


Source: http://www.datasciencecentral.com/profiles/blogs/data-veracity

mariusmarcu@global.t-bird.edu

Who produces/collects data and for what purpose?

- **Users/individuals** – they like to keep/share what they produce. “Sentimental value” vs “make money”. Data mostly shared.
- **Businesses** – optimize/grow the business – make more money. Data mostly NOT shared except when the main business is collection & sale of data and or insights.
- **Government** – Govern/protect/serve citizens. Mostly NOT shared, except sometimes for information purposes and public transparency. Sometimes they try to make a buck to balance gov operating costs.
- **Education institutions** – research purposes. Data shared to the extent allowed by research and community scrutiny.



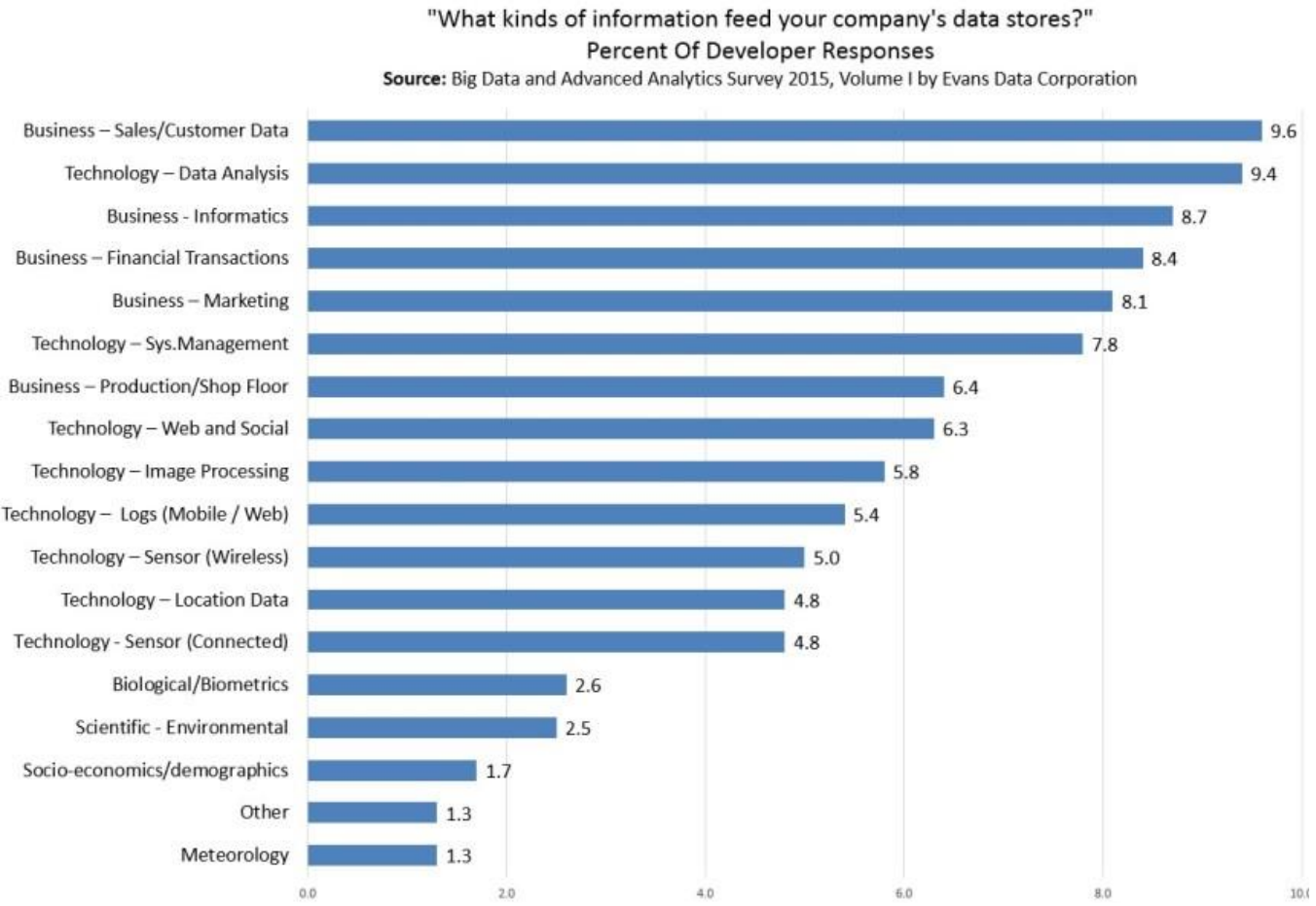
mariusmarcu@global.t-bird.edu

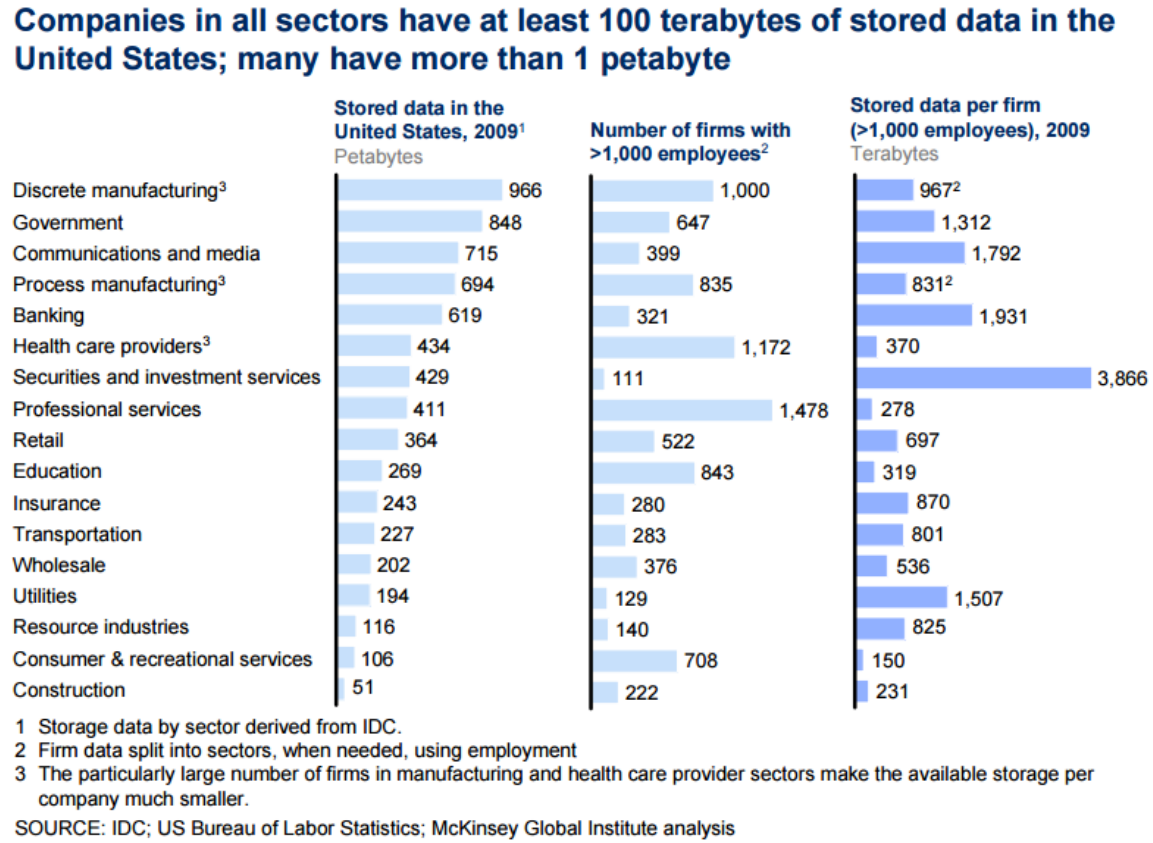
The landscape of big data and the tools to mine it

- Mountains of data collected today but businesses are interested only in the golden nuggets : metrics vs insights vs predictive analytics
- Few bridges between mountains
  - Data sets in diff frameworks/storage models that don't necessarily talk with each other
- Even fewer data miners
  - Few know how to apply the Scientific Method to big data sets
- Nascent data mining within most orgs (BRONZE AGE?)
  - **High hopes** for what they can get out of advanced analytics.
  - **Organization** don't have (YET) the capabilities they need to exploit big data
  - **Lack of alignment** on key issues for people inside individual organizations.

mariusmarcu@global.t-bird.edu

What data gets collected and who's analyzing it?

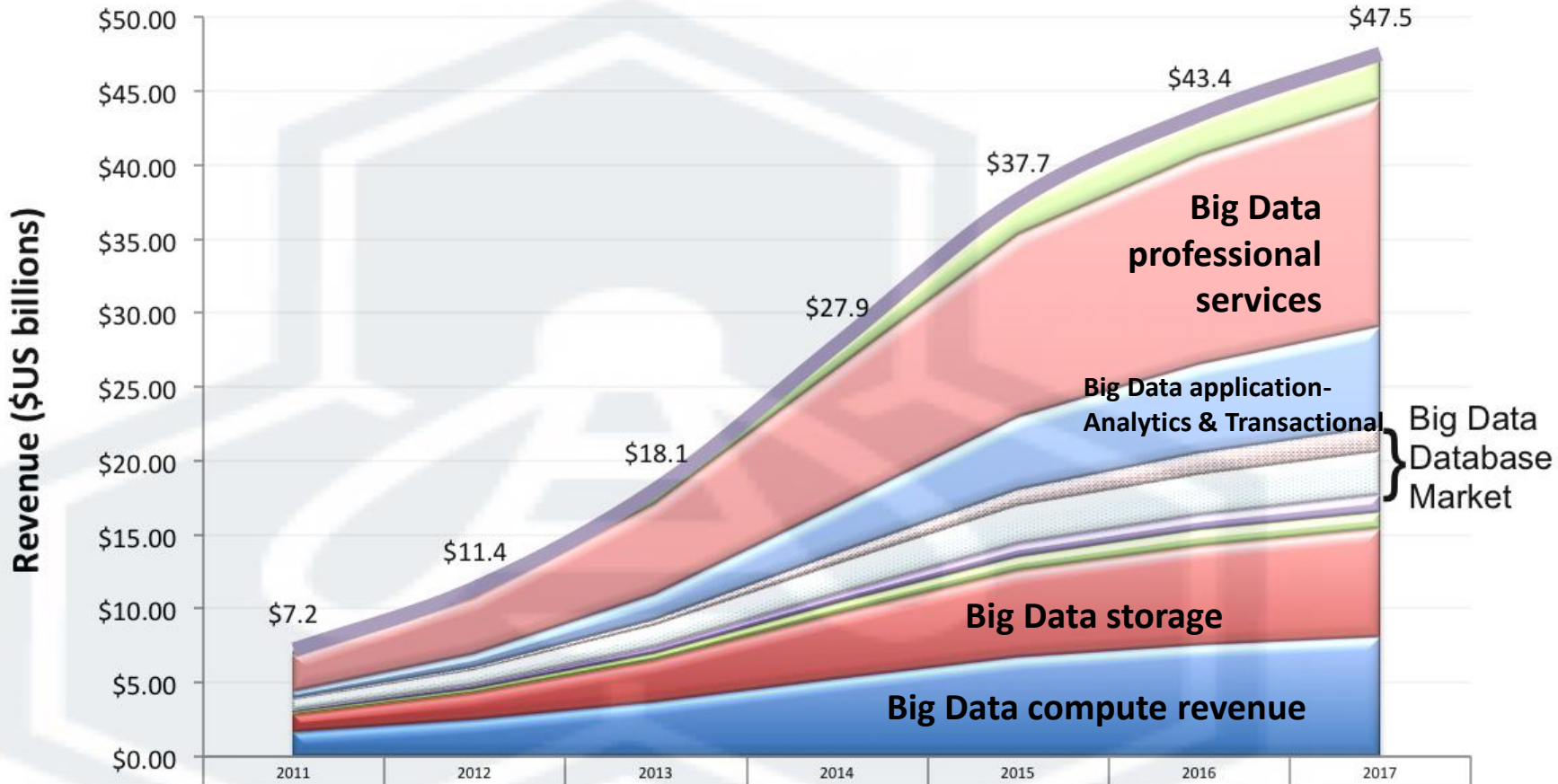




Lots of departments interested get lots of insights from lots of data.

Source: http://www.business2community.com/big-data/2015-big-data-market-update-01355960?utm_source=Sailthru&utm_medium=email&utm_campaign=Issue:%202015-10-20%20CIO%20Dive&utm_term=CJO%20Dive#TLWhJGgeLlh21X66.97

Big Data Market Forecast by Component, 2011-2017 ($US billions)



	2011	2012	2013	2014	2015	2016	2017
Big Data XaaS Revenue	$0.34	$0.60	$1.03	$1.71	$2.43	$2.87	$3.19
Big Data Professional Services Revenue	$2.43	$3.85	$6.07	$9.24	$12.31	$14.06	$15.30
Big Data Application (Analytic and Transactional) Revenue	$0.48	$0.93	$1.77	$3.24	$4.94	$6.05	$6.89
Big Data NoSQL Database Revenue	$0.10	$0.19	$0.39	$0.73	$1.14	$1.41	$1.62
Big Data SQL Database Revenue	$0.72	$1.02	$1.45	$2.00	$2.48	$2.74	$2.91
Big Data Infrastructure Revenue	$0.15	$0.25	$0.42	$0.67	$0.93	$1.08	$1.19
Big Data Networking Revenue	$0.18	$0.28	$0.44	$0.67	$0.89	$1.02	$1.11
Big Data Storage Revenue	$1.16	$1.83	$2.88	$4.39	$5.85	$6.68	$7.27
Big Data Compute Revenue	$1.64	$2.45	$3.64	$5.23	$6.70	$7.50	$8.06
Total Big Data Revenue	$7.2	$11.4	$18.1	$27.9	$37.7	$43.4	$47.5
Database as % of Total Big Data Market	11.4%	10.7%	10.2%	9.8%	9.6%	9.6%	9.5%

Source: © Wikibon Big Data Model 2011-2017

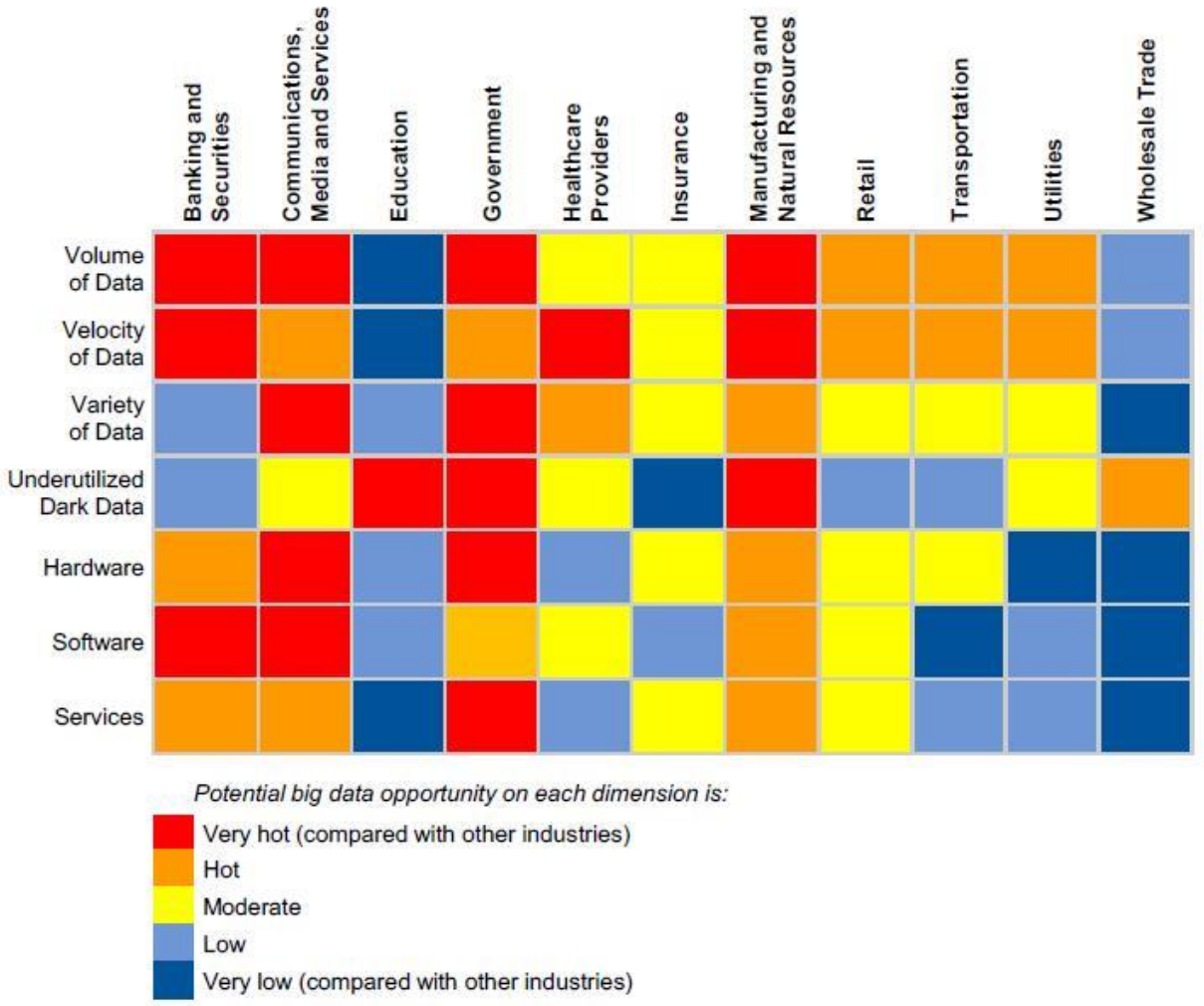
Big data market to reach almost $50bn by 2017

mariusmarcu@global.t-bird.edu

Greatest potential opportunities for Big Data (from a volume of data perspective):

- Banking and securities
- **Communications**
- **Media and Services**
- Government
- **Manufacturing**
- Natural Resources

Figure 2. Big Data Opportunity Heat Map by Industry



Source: Gartner (July 2012)

mariusmarcu@global.t-bird.edu

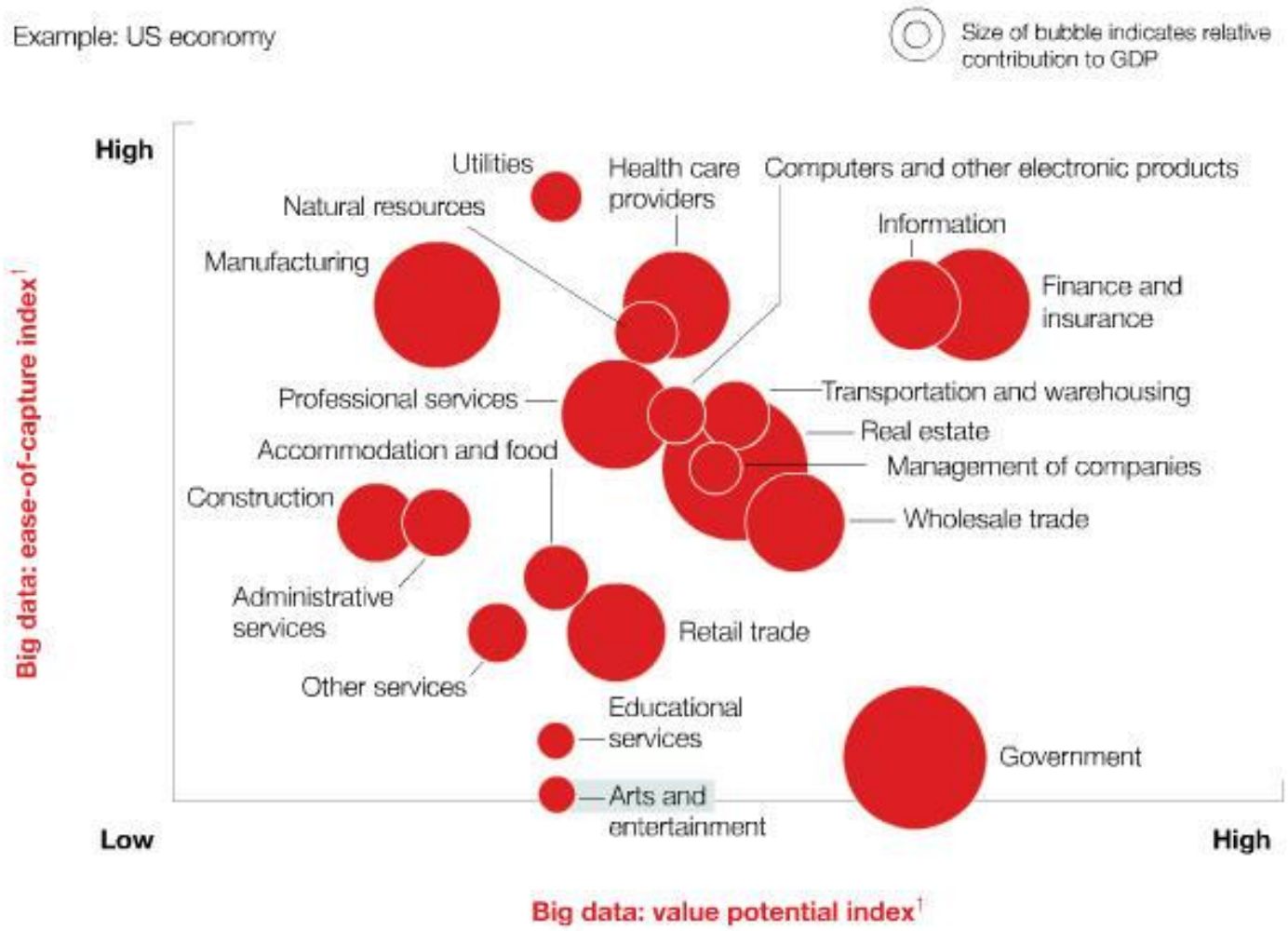
Finance and insurance data is EASY to capture and has HIGH value potential

Construction data NOT EASY to capture and has LOW value potential

What about ….?

- Healthcare
- Professional Services

The ease of capturing big data's value, and the magnitude of its potential, vary across sectors.



[1] For detailed explication of metrics, see appendix in McKinsey Global Institute full report *Big data: The next frontier for innovation, competition, and productivity*, available free of charge online at mckinsey.com/mgi.

Source: US Bureau of Labor Statistics; McKinsey Global Institute analysis

mariusmarcu@global.t-bird.edu

Plenty of room to compete and win with analytics solutions

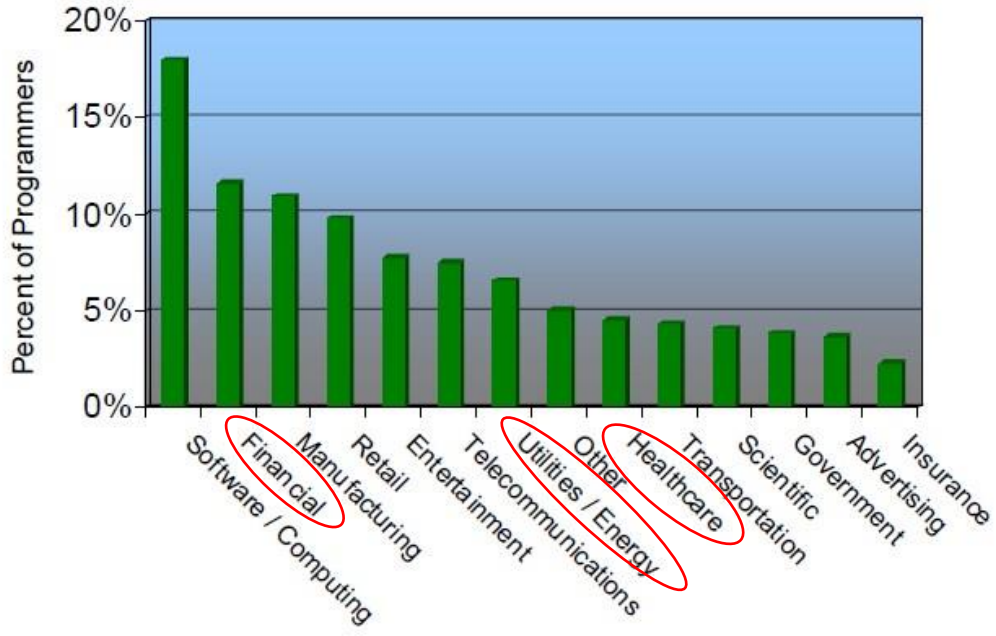
Large industries have significant untapped value in big data adoption.

Complex large-market industries, including financial services and healthcare, are underdeveloped when considering the potential big data adoption has for significant disruption and value creation. Big data strategies in these sectors have been slowed by difficulty of data capture and level of regulation.

U.S. market size vs. SVB Big Data Maturity Index[10]

Industry value ($B)

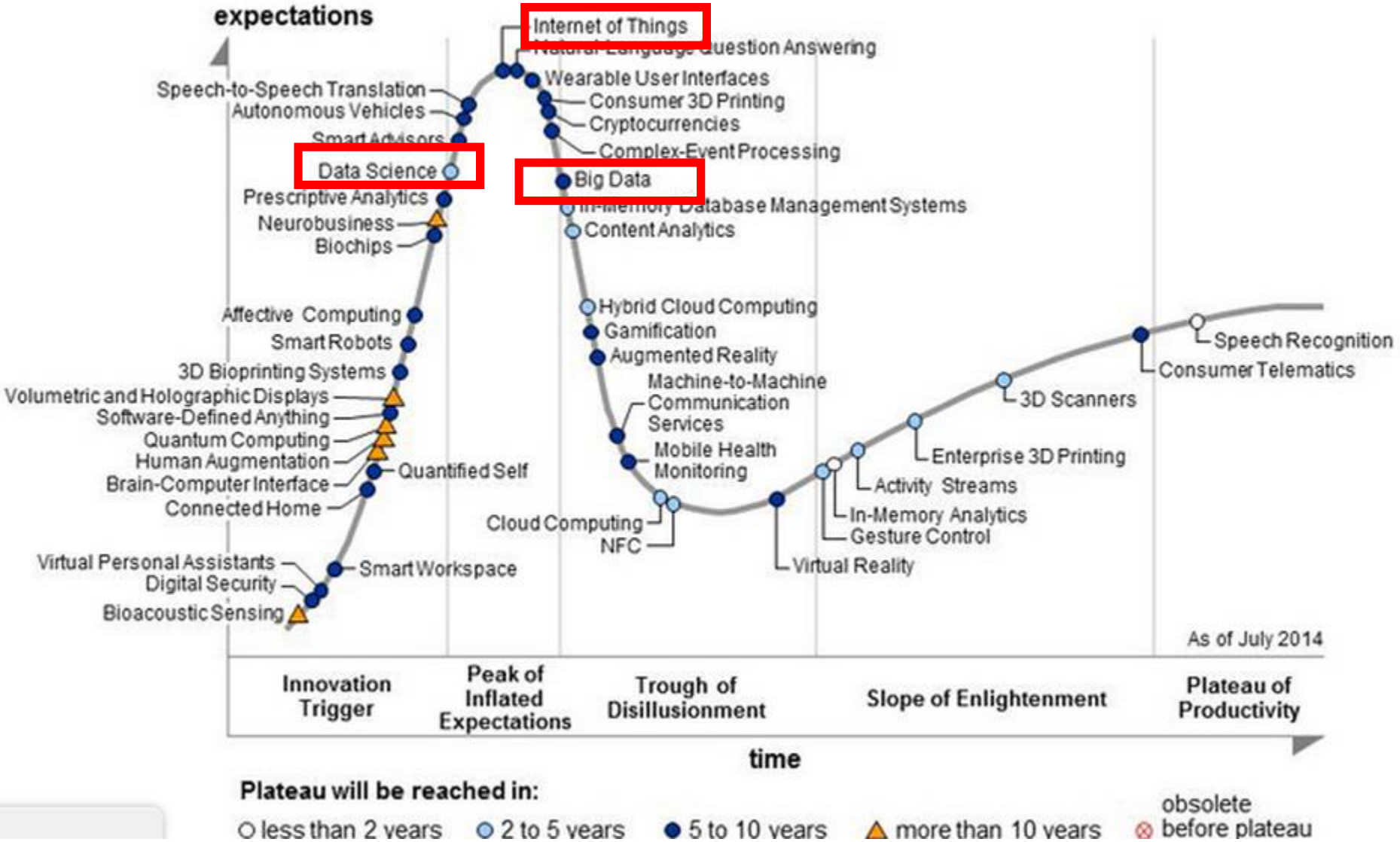
What industry does your data and analytics solutions address?



Big Data & Advanced Analytics Survey, Vol. I © 2015 Evans Data Corp.

mariusmarcu@global.t-bird.edu

Gartner's 2014 Hype Cycle for Emerging Technologies



We got (big) data.
What do we do with it?
Data Science.

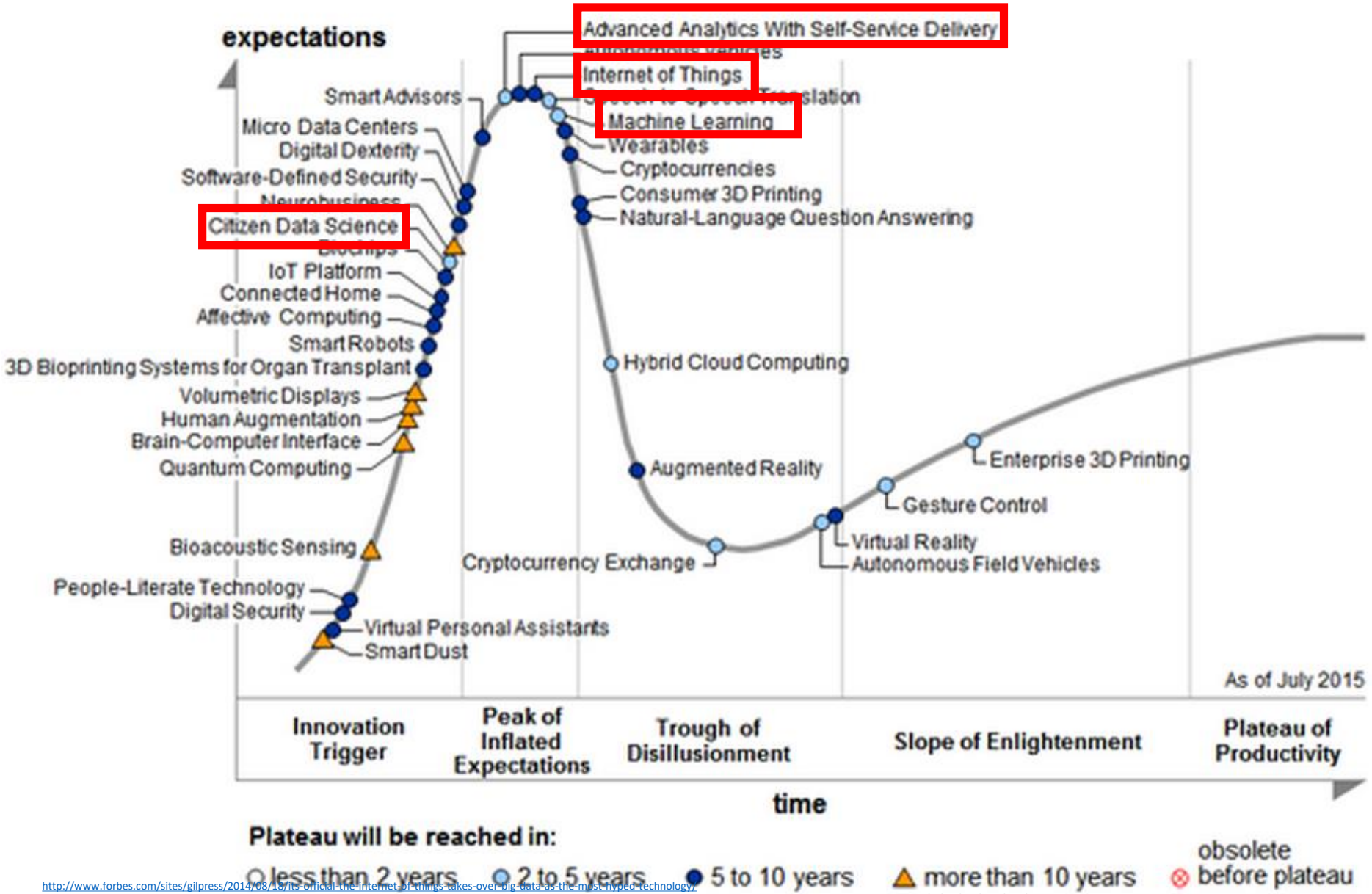
And the IOT
is going to give us
unlimited data supply!

Setting		Description
	Human	Devices attached to or inside the human body
	Home	Buildings where people live
	Retail environments	Spaces where consumers engage in commerce
	Offices	Spaces where knowledge workers work
	Factories	Standardized production environments
	Worksites	Custom production environments
	Vehicles	Systems inside moving vehicles
	Cities	Urban environments
	Outside	Between urban environments (and outside other settings)

SOURCE: McKinsey Global Institute analysis

http://www.forbes.com/sites/gilpress/2014/08/18/its-official-the-internet-of-things-takes-over-big-data-as-the-most-hyped-technology/

Gartner's 2015 Hype Cycle- Things change ...fast!



Big data seems to be the new normal, now found in data lakes

Democratization of data science at the horizon?

Machine learning still hot

http://www.forbes.com/sites/gilpress/2014/08/18/its-official-the-internet-of-things-takes-over-big-data-as-the-most-hyped-technology/

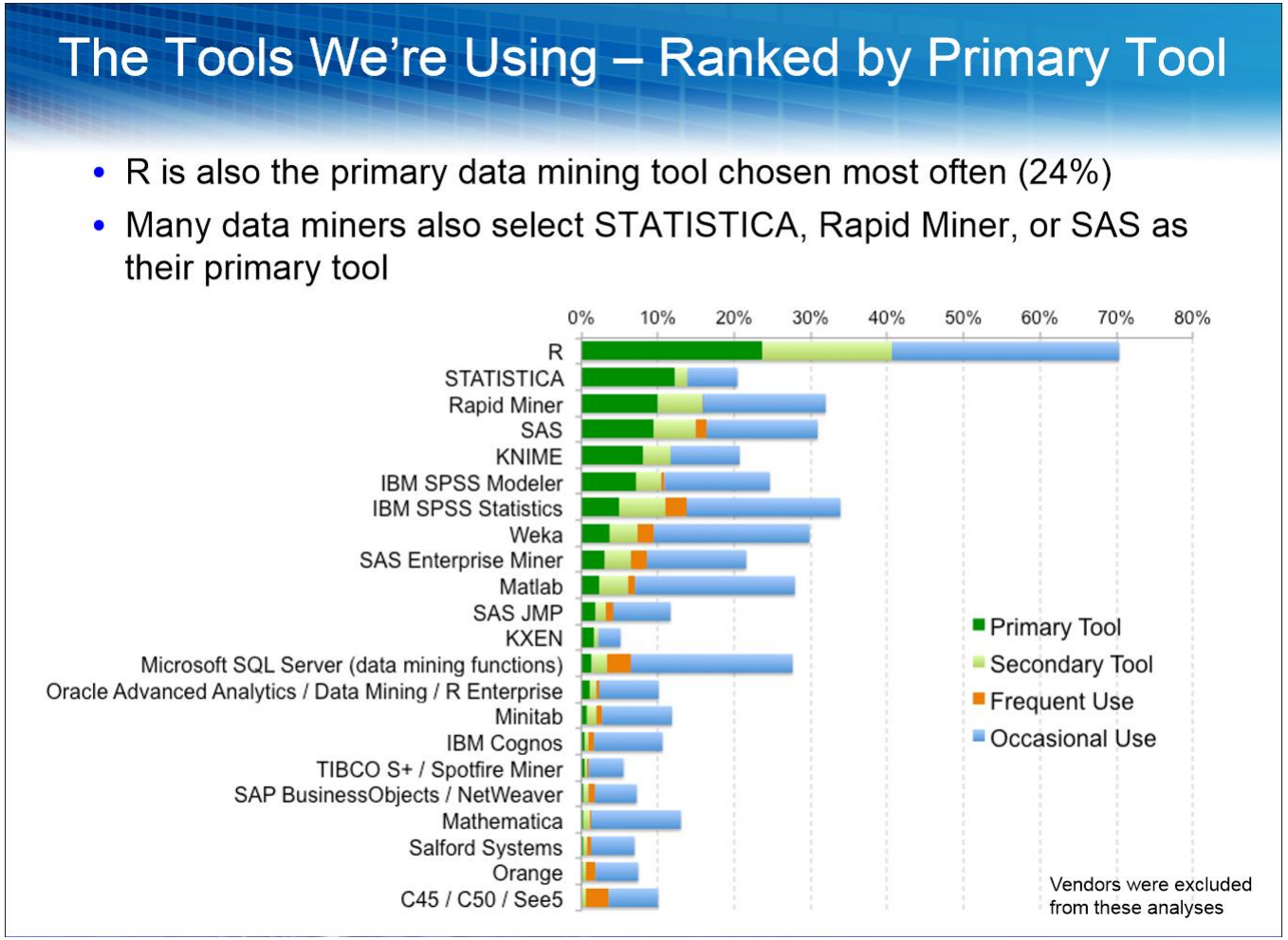
mariusmarcu@global.t-bird.edu

Data Scientists –
the new, modern, gold miners

mariusmarcu@global.t-bird.edu

Data scientists, the modern gold miners

- Must know how to construct intelligent hypotheses
- Understand the principles of experimental testing and design*
- Able to evaluate the validity of data analyses.



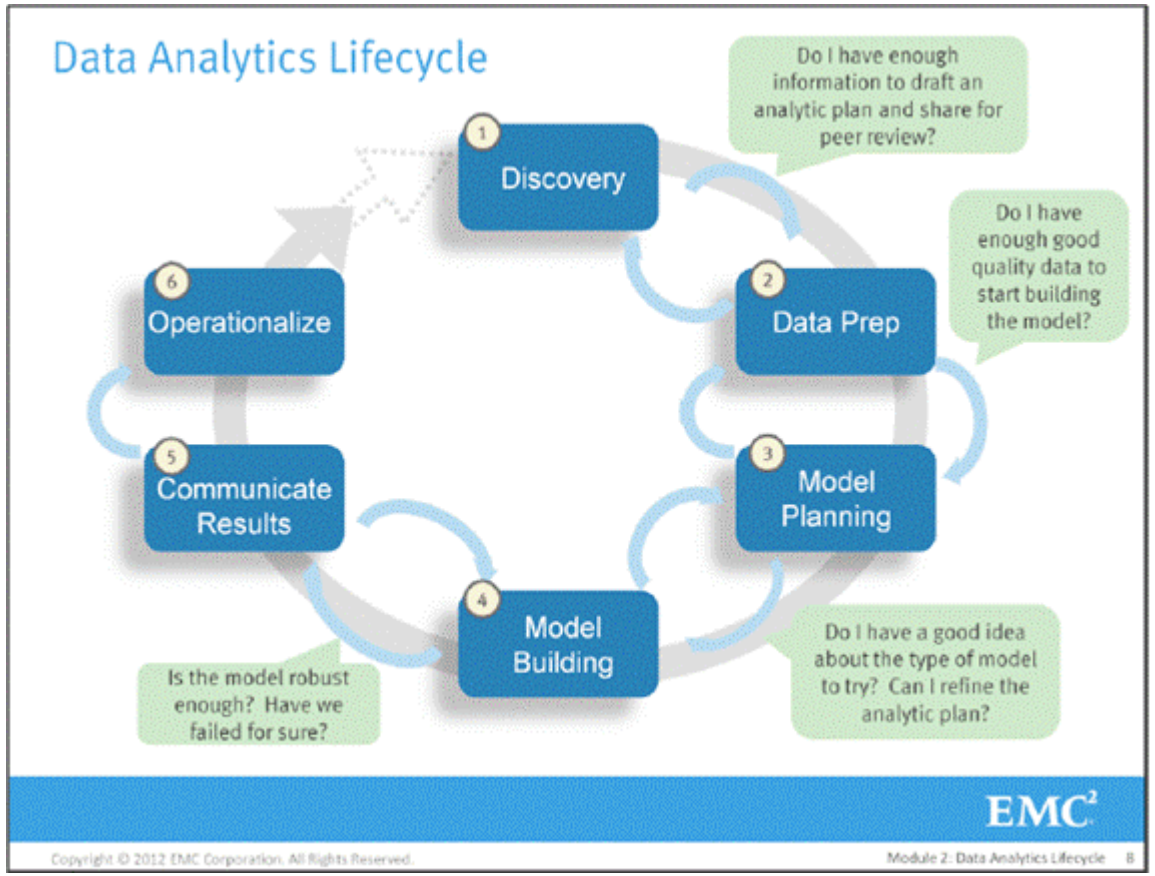
*A background in scientific experimental design will be particularly valued as randomized testing and experimentation becomes more commonplace

mariusmarcu@global.t-bird.edu

Data analytics everywhere… forever? Forever ever?

- Data collection –make vs buy
- Data preparation – 60% of project time!!!
- Analysis & modelling based on (clear) objectives from org
- Insights, predictions
- Recommend optimizations to business decision makers
- Management actions
- Results

Rinse and repeat

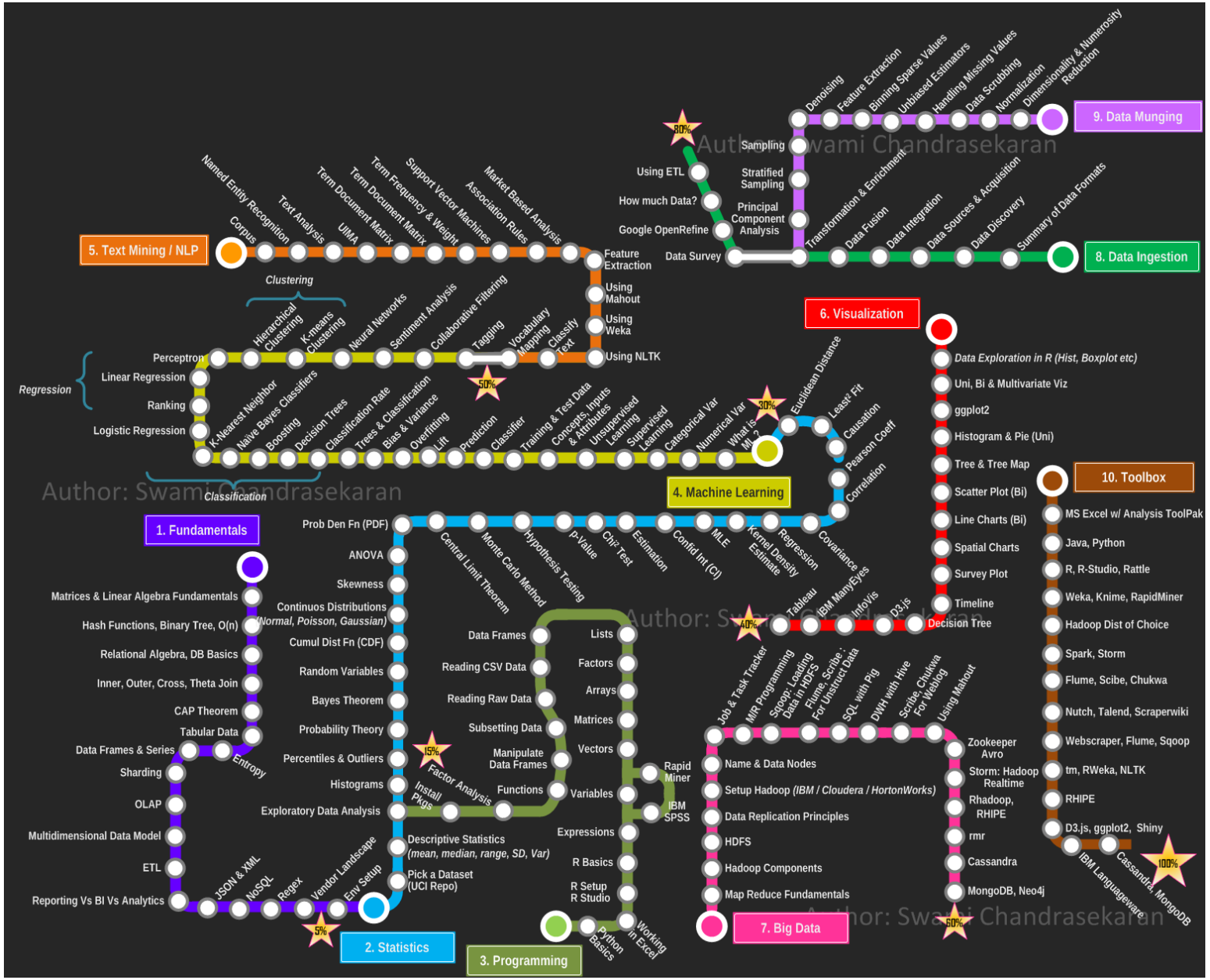


mariusmarcu@global.t-bird.edu

Looking for the (perfectly) skilled Data Scientist?

…keep looking…

in Data Science, specialization may be key



mariusmarcu@global.t-bird.edu

Categories of data scientists

1. Those strong in **statistics**: they sometimes develop new statistical theories for big data, that even traditional statisticians are not aware of. They are expert in statistical modeling, experimental design, sampling, clustering, data reduction, confidence intervals, testing, modeling, predictive modeling and other related techniques.
2. Those strong in **mathematics**: NSA (national security agency) or defense/military people working on big data, astronomers, and *operations research* people doing analytic business optimization (inventory management and forecasting, pricing optimization, supply chain, quality control, yield optimization) as they collect, analyze and extract value out of data.
3. Those strong in **data engineering**, Hadoop, database/memory/file systems optimization and architecture, API's, Analytics as a Service, optimization of data flows, data plumbing.
4. Those strong in **machine learning** / computer science (algorithms, computational complexity)
5. Those strong in **business**, ROI optimization, decision sciences, involved in some of the tasks traditionally performed by business analysts in bigger companies (dashboards design, metric mix selection and metric definitions, ROI optimization, high-level database design)
6. Those strong in production code development, **software engineering** (they know a few programming languages)
7. Those strong in **visualization**
8. Those strong in GIS, **spatial data**, data modeled by graphs, graph databases
9. **Those strong in a few of the above**. After 20 years of experience across many industries, big and small companies (and lots of training), I'm strong both in stats, machine learning, business, mathematics and more than just familiar with visualization and data engineering. This could happen to you as well over time, as you build experience. I mention this because so many people still think that it is not possible to develop a strong knowledge base across multiple domains that are traditionally perceived as separated (the *silo* mentality). Indeed, that's the very reason why *data science* was created.

Which one will you grow into?

http://www.datasciencecentral.com/profiles/blogs/six-categories-of-data-scientists

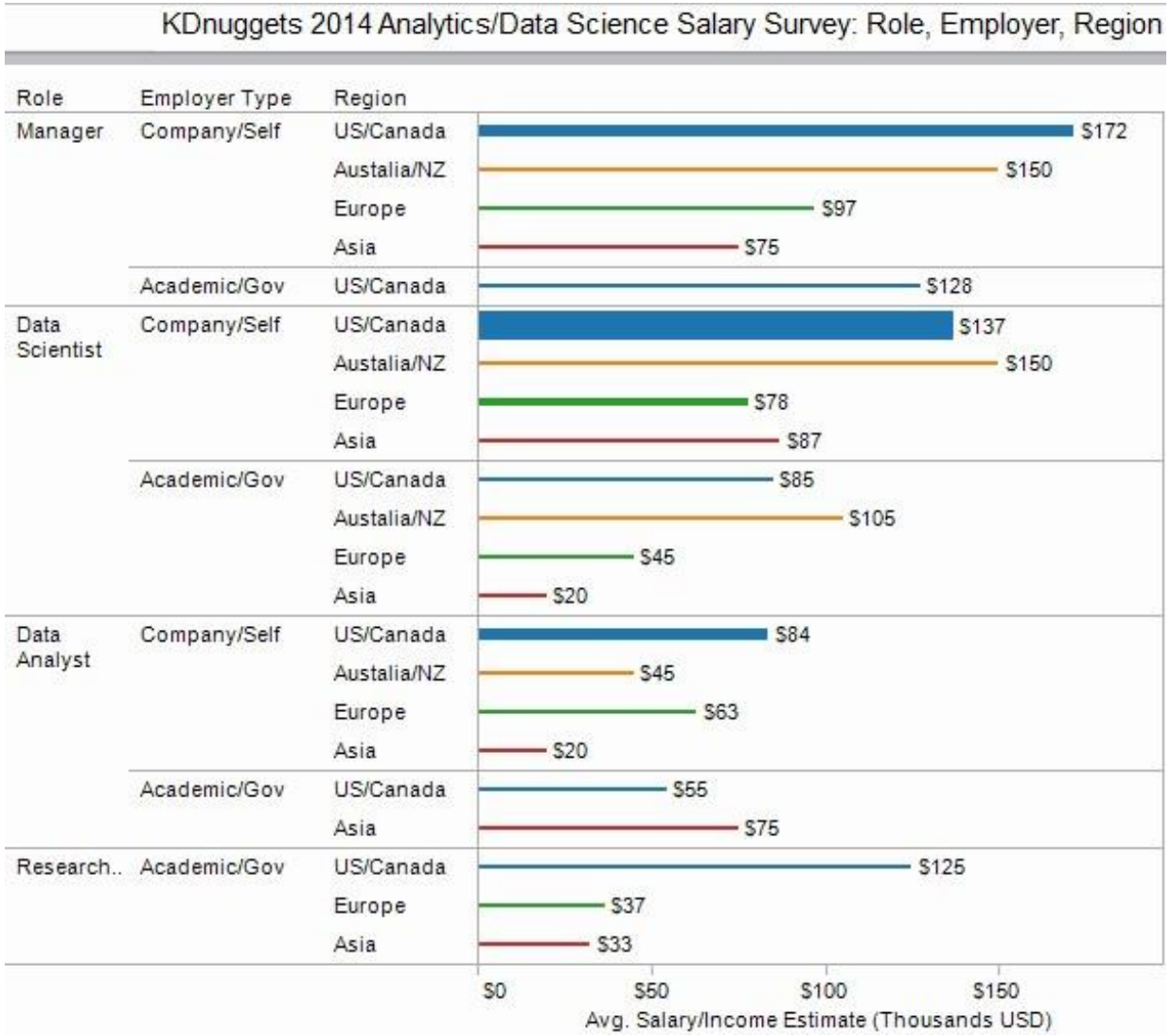
mariusmarcu@global.t-bird.edu

Pay rates for data scientists around the world

US/Canada – life is good

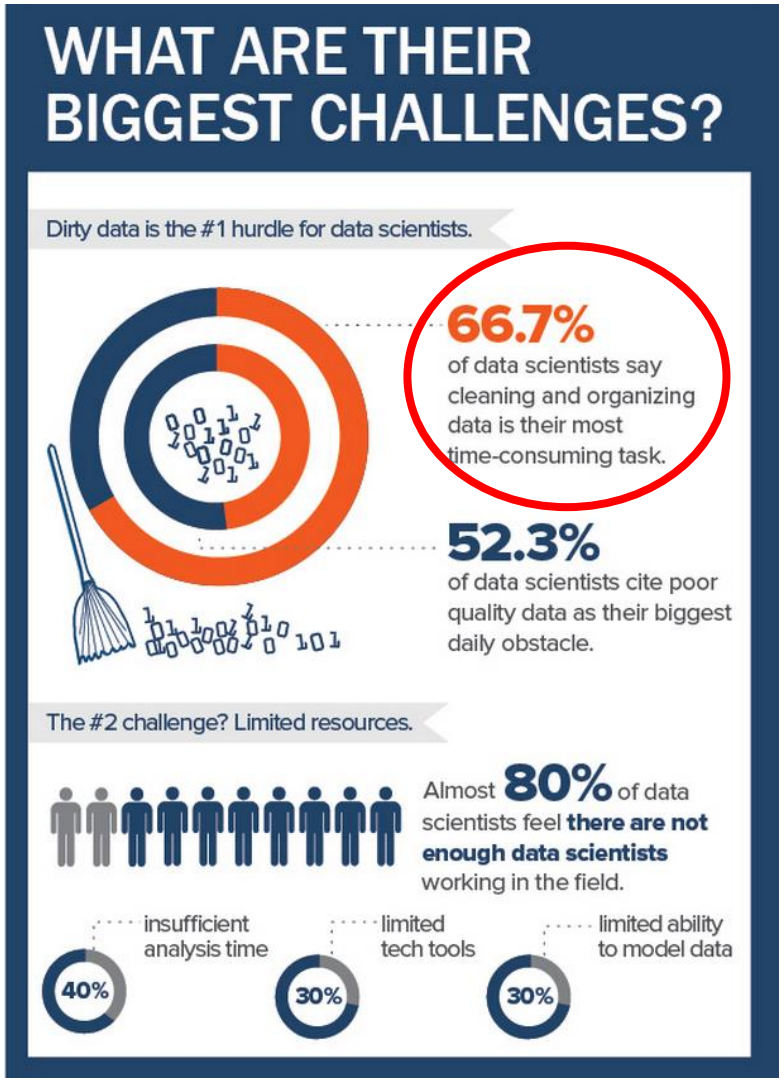
Australia/NZ - thought leaders

Outsource oppts : Asia and Europe



mariusmarcu@global.t-bird.edu

Working hard and dreaming about a better future



Source: Crowdflower Data Scientist report 2014

mariusmarcu@global.t-bird.edu

The rise of Citizen Data Scientists

- Citizen Data Scientist -"a person who creates or generates models that leverage predictive or prescriptive analytics but whose primary job function is outside of the field of statistics and analytics."

- # of citizen data scientists will grow 5x faster through 2017 than the number of data scientists.

- The Citizen Data Scientist plays games with data:
  - Visualizations,
  - New kinds of KPIs

- Speed vs depth of insight

- Complements the Data Scientist, sources of ideas and support

Citizen data scientists could be your biggest fans!

Smart Data Discovery tools and Capabilities

	Prepare Data	Find Patterns in Data		Share Findings
Sample Vendors	**Smart Self-Service Data Preparation**	**Smart Visualizations**	**Automated Pattern detection**	**NLG of Findings**
IBM Watson Analytics	R	✓	✓	
SAP Lumira	✓			✓
SAS Visual Analytics	R	✓	✓	
BeyondCore	✓	✓	✓	✓
DataRPM	✓	✓	✓	✓
Saffron			✓	
Ayasdi			✓	
Paxata	✓			
Trifacta	✓			
Tamr	✓			
ClearStory	✓	✓		
Automated Insights			✓	✓
Narrative Science			✓	✓
Datameer	✓		✓	
Platfora	✓	R	R	
Yseop			✓	✓
Qlik	R	✓		
Tableau	R	R		
	R = On the roadmap		✓ Have	
NLG = natural-language generation				

Source: Gartner (June 2015)

mariusmarcu@global.t-bird.edu

Food for thought: Data Forensics not Data Science could be the skill shortage

- 2011- McKinsey : US alone short of 140,000 and 190,000 people with deep analytical skills
- 2014 Capgemini - the biggest challenge in big data is often the provenance of the data.
  - "you only get out what you put in"

- From enterprise data (largely controlled) to lots of disparate sources
- 100s of small variations in the way business is conducted = "finer adjustments" that need more accuracy not less.
- Knowing more about your data sources can better inform your modelling

3 key dimensions to asses data:

1. **Provenance** – do you trust that source, what level of quality can we expect in the data? Adjust models accordingly

2. **Legality** important to understand what is and isn't allowed

3. **Sensitivity** - breaching some ethical boundaries?
   - brand reputation and image can make or break companies,
   - impact of people knowing about your use of their data (NSA revelations; Target "targeting" pregnant women)

Understanding your data sources could be the real skill in turning big data into value

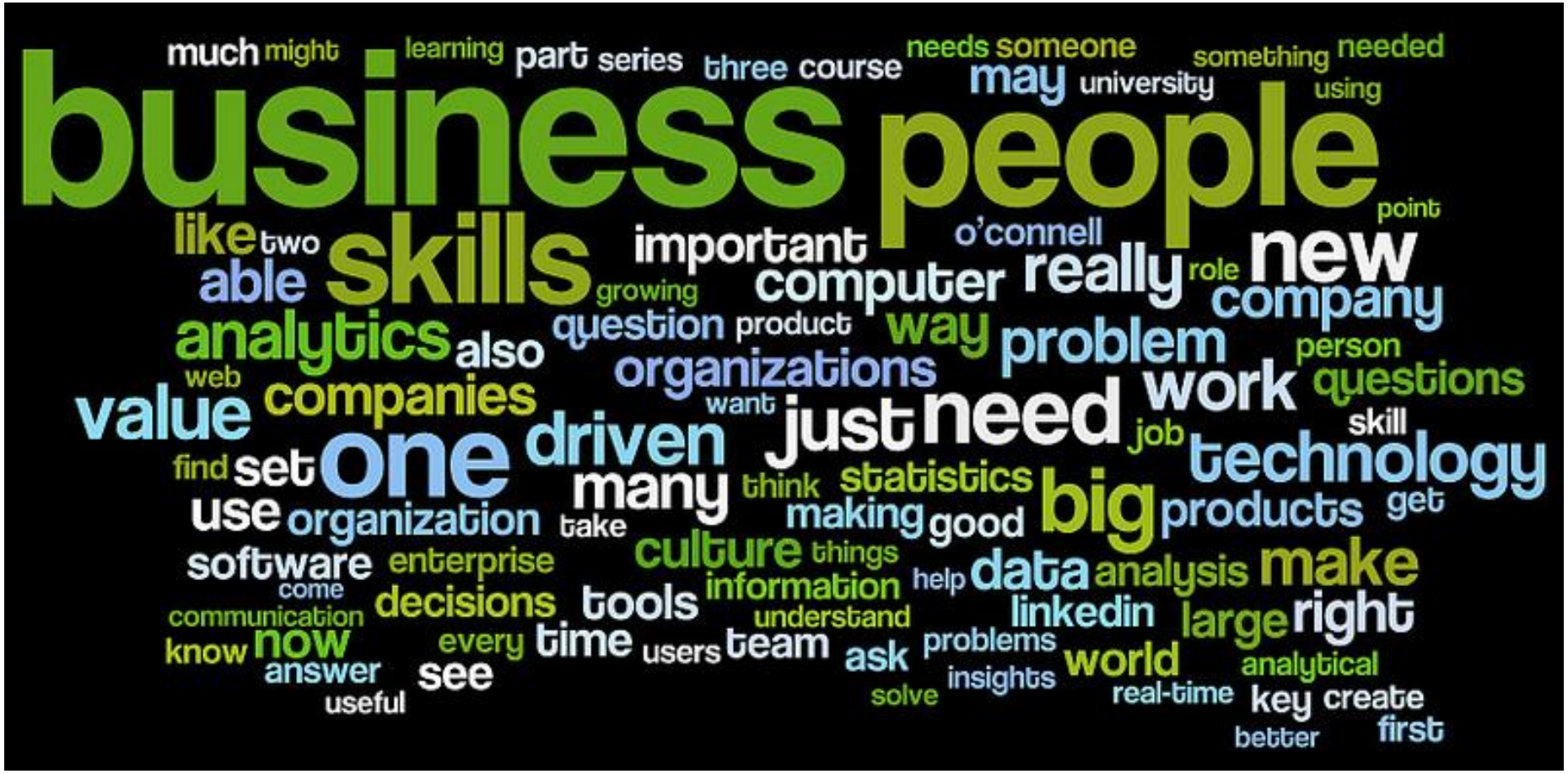
http://www.capgemini.com/blog/capping-it-off/2014/02/data-forensics-not-data-science-will-be-the-skill-shortage-for-the-big

mariusmarcu@global.t-bird.edu

How you can make most of your Data Science opportunity

mariusmarcu@global.t-bird.edu

Data Science – mostly about business ?



Source: Forbes "what is a data scientist" series word cloud

Data Scientists - helping change business and the world

Which job will you help replace?

- Front-line Military Personnel Will Be Replaced With Robots
- Private Bankers and Wealth Managers Will Be Replaced With Algorithms
- Lawyers, Accountants, Actuaries, and Consulting Engineers Will Be Replaced With Artificial Intelligence

What are the jobs that will be in demand in this brave new world only a decade away?

- Personal Worker Brand Coaches And Managers, **Professional Triber***, **Freelance Professors**, Urban Farmers, End-Of-Life Planner, Senior Carer, **Remote Health Care Specialist**, Neuro-Implant Technicians, Smart-Home Handyperson, **Virtual Reality Experience Designer**, Sex Worker Coach (?!)

*a freelance professional manager that specializes in putting teams together for very specific projects
http://www.fastcompany.com/3046277/the-new-rules-of-work/the-top-jobs-in-10-years-might-not-be-what-you-expect

mariusmarcu@global.t-bird.edu

Do no evil

People's understanding of current technology is following Moore's law, sadly in the reverse direction.

“Cloud" – the 1st Tsunami people hardly understand the concept and how it impacts their life

"Big Data“ – the 2nd Tsunami, looks like Big Brother to many, keeping an eye on everything you do, and it is almost impossible to hide anything from it, no matter how personal

“Internet of things” – the 3rd Tsunami? Sensors will be everywhere!

“Data Science” -> Machine Intelligence -> AI

Making sense of it all. Helps understand, predict, influence and drive behavior of the user/consumer/target

http://www.kdnuggets.com/2014/11/big-data-comic-privacy.html

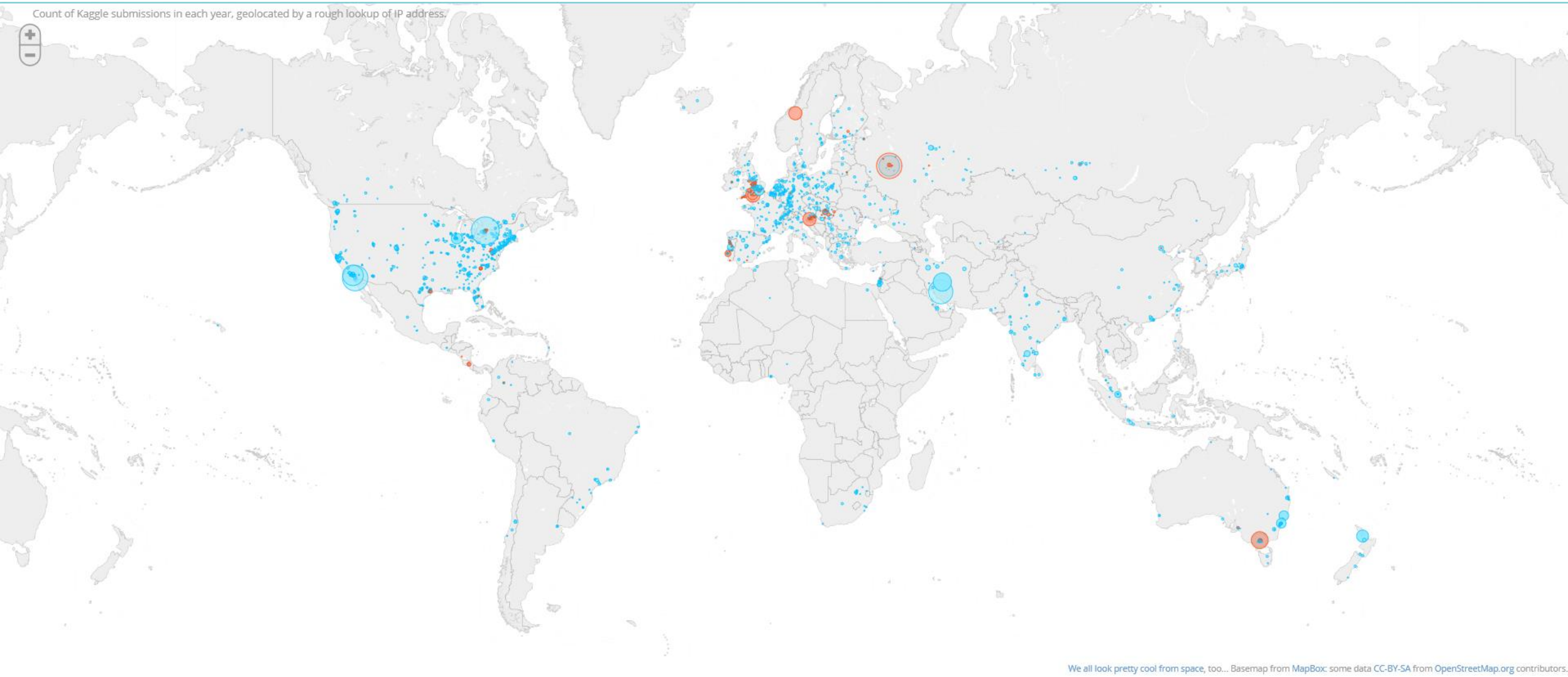
mariusmarcu@global.t-bird.edu

Data Science is : Science & Art

- **Where do you fit?** – where you can add most value today
- **Where do you want to be?** – where you can get paid best and work the least (or where you like it most)
- **Gravitate** towards companies with big data sophistication and thought leaders in big data (high tech, cloud companies, pharma… etc)
- **Start in science and move into art**
- **Packaging & delivery** of insights can be as important as content
- **Connect** with fellow data scientists
- **Kaggle** can help you grow - Data Hero?

mariusmarcu@global.t-bird.edu

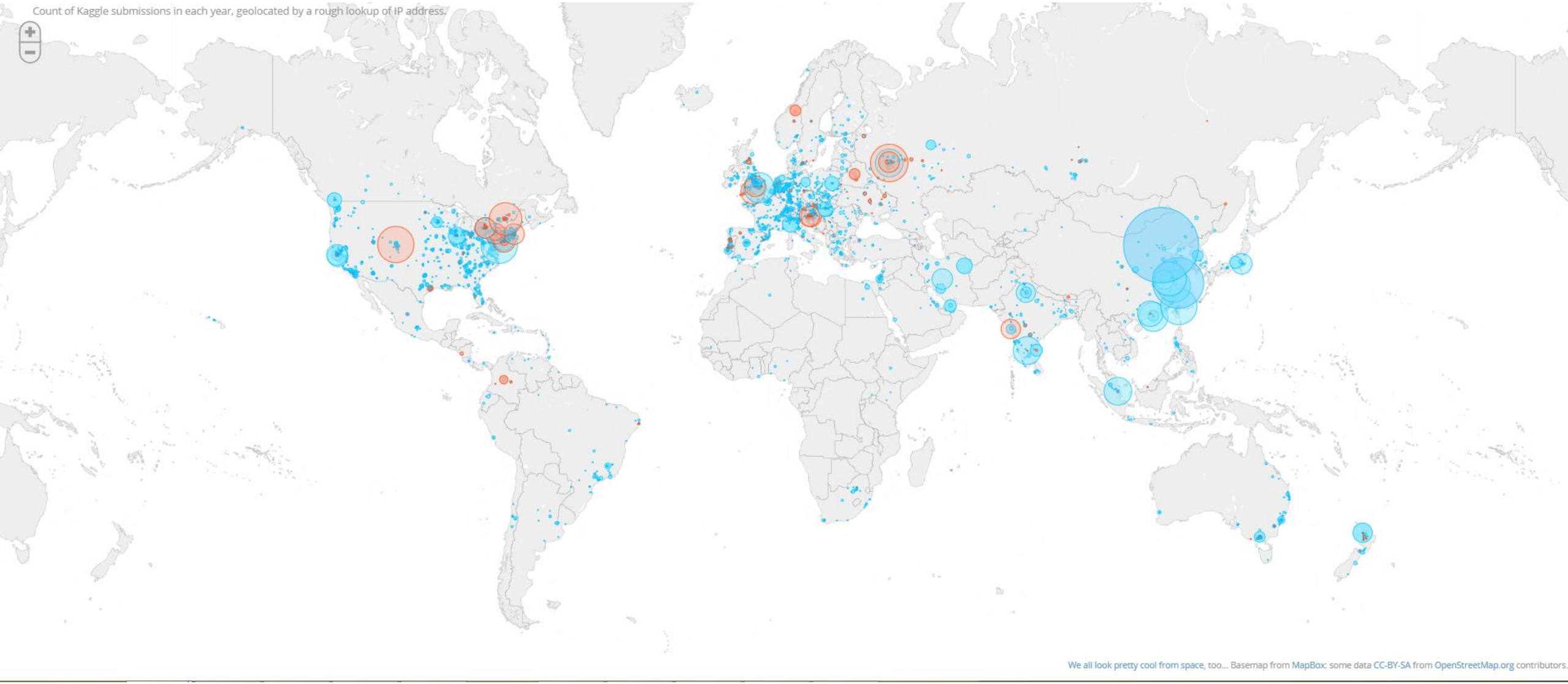
Kaggle submissions by IP location 2011



Source: http://garimeme.com/kaggle/grown.html#2.50/18.70/7.00

mariusmarcu@global.t-bird.edu

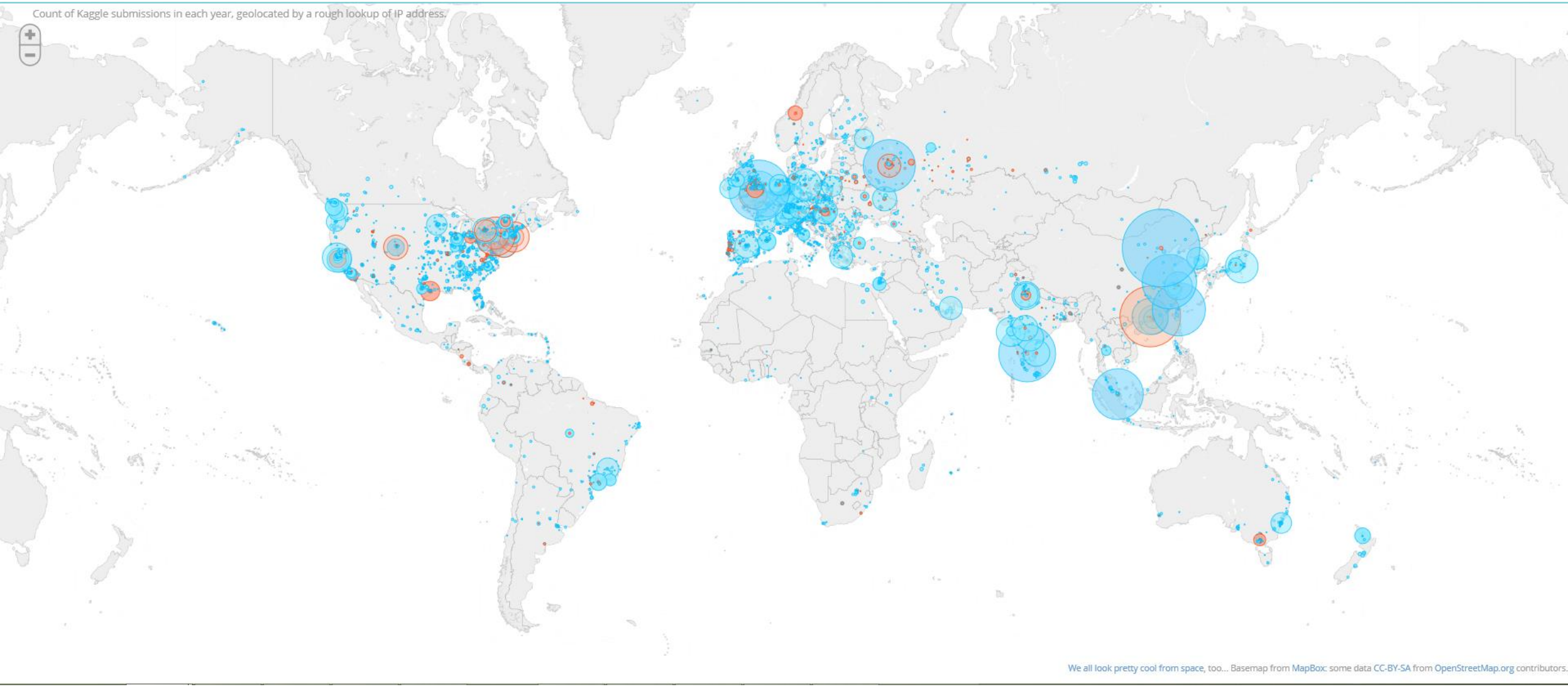
Kaggle submissions by IP location 2012



Source: http://garimeme.com/kaggle/grown.html#2.50/18.70/7.00

mariusmarcu@global.t-bird.edu

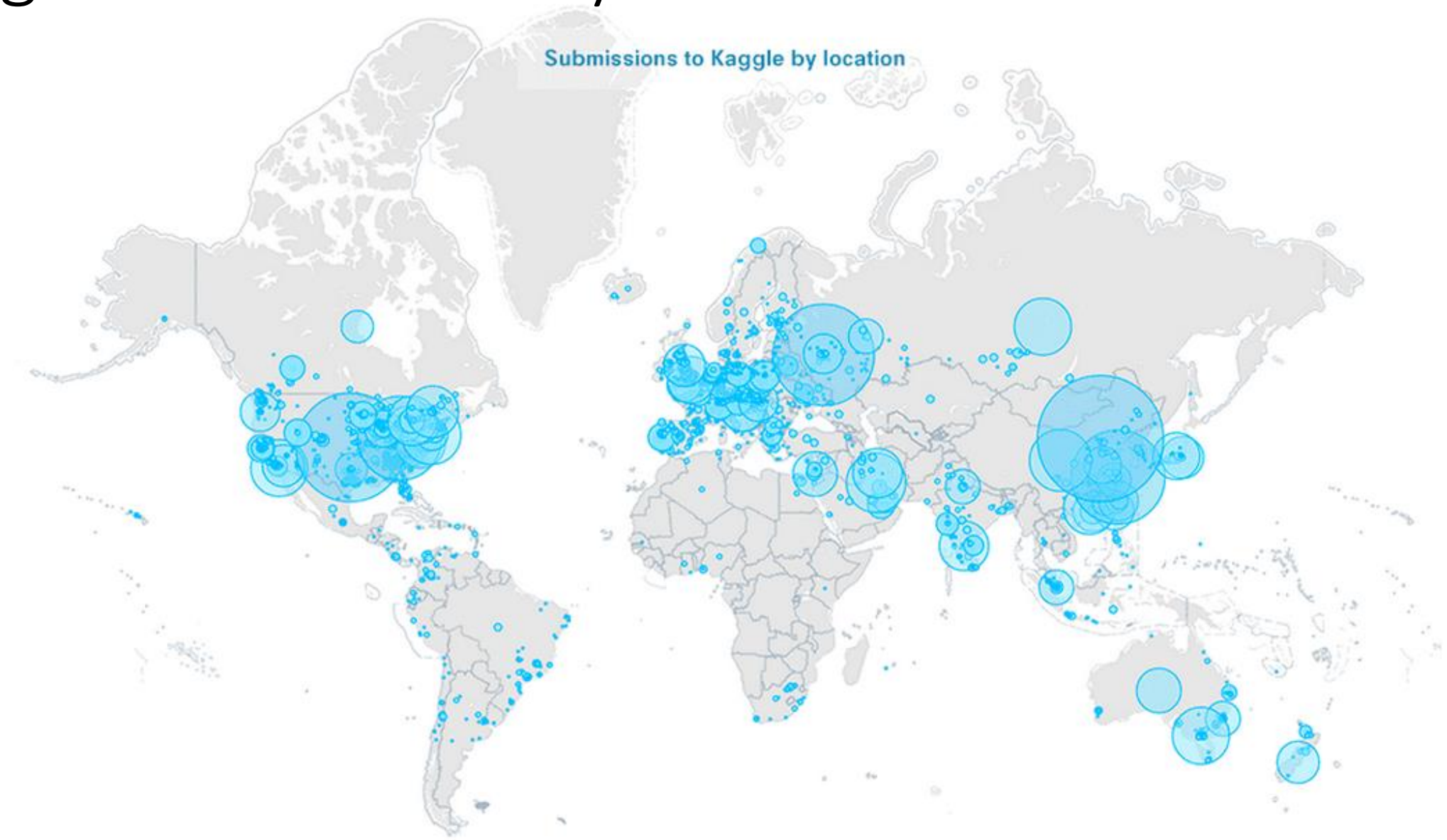
Kaggle submissions by IP location 2013



Source: http://garimeme.com/kaggle/grown.html#2.50/18.70/7.00

mariusmarcu@global.t-bird.edu

Kaggle submissions by IP location 2014



Source: http://garimeme.com/kaggle/grown.html#2.50/18.70/7.00

mariusmarcu@global.t-bird.edu

Summary

- Big Data – the new natural resource
- Everyone is looking for the gold nuggets
- Data Scientists – the new, modern gold miners
- Data Science tools - bronze age
- Data science is here to stay and you are in the right boat
- Data Science = Science & Art
- Find the start point that is right for you and….

…. Get to rule Kaggle!

mariusmarcu@global.t-bird.edu

Q & A

…Marius plans to make his first $1M in the next 5yrs
with a project/company empowered by Data Science…

What is your plan?

Let’s connect & chat

mariusmarcu@global.t-bird.edu