Big Data & Career Paths

Conference Paper · June 2014

CITATIONS
0

READS
1,881

1 author:

Marcos Colebrook
Universidad de La Laguna
32 PUBLICATIONS **375** CITATIONS

Big Data & Career Paths

Marcos Colebrook
Univ. de La Laguna
@MColebrook

Contents

- Big Data facts
- Definition of Big Data
- Techs & Tools
- Data Science: skills and career paths
- Conclusions

Big Data everywhere!!

FAD?

BUZZWORD?

TREND?

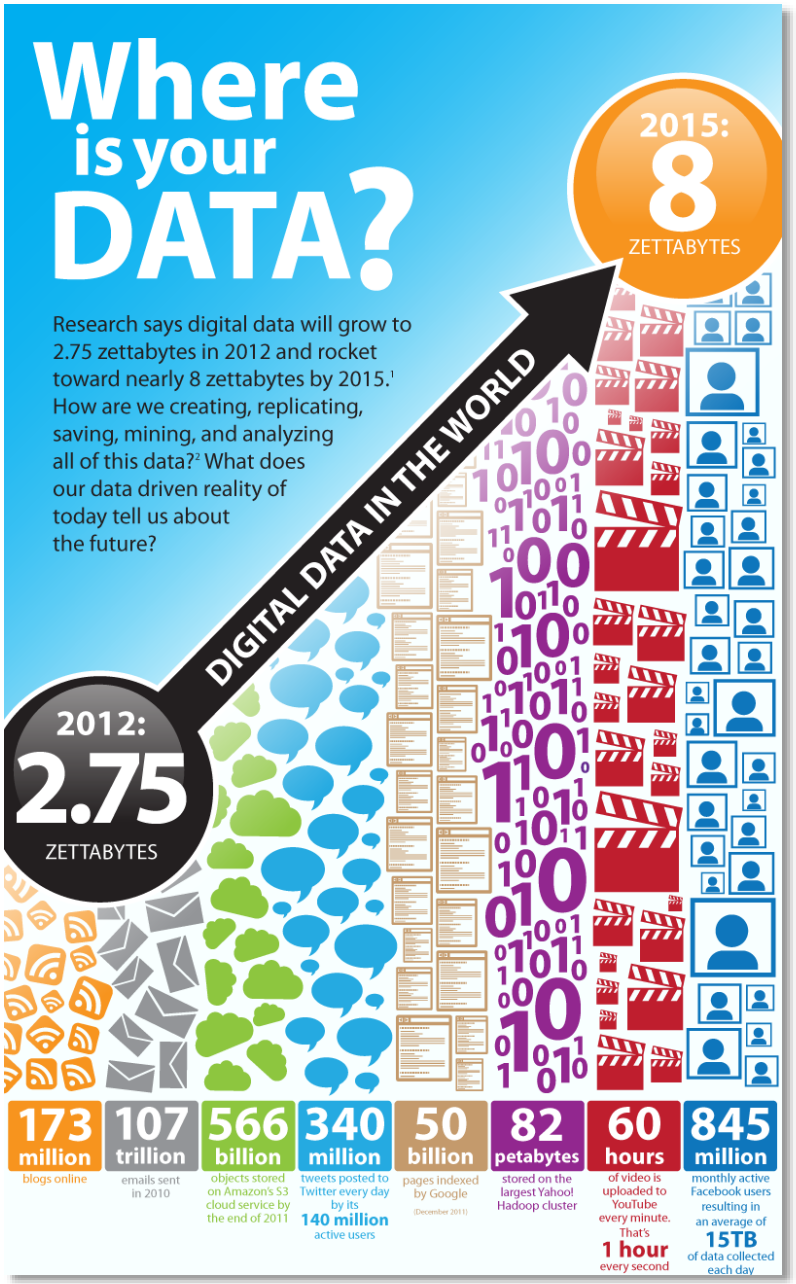
HYPE?

FUSS?

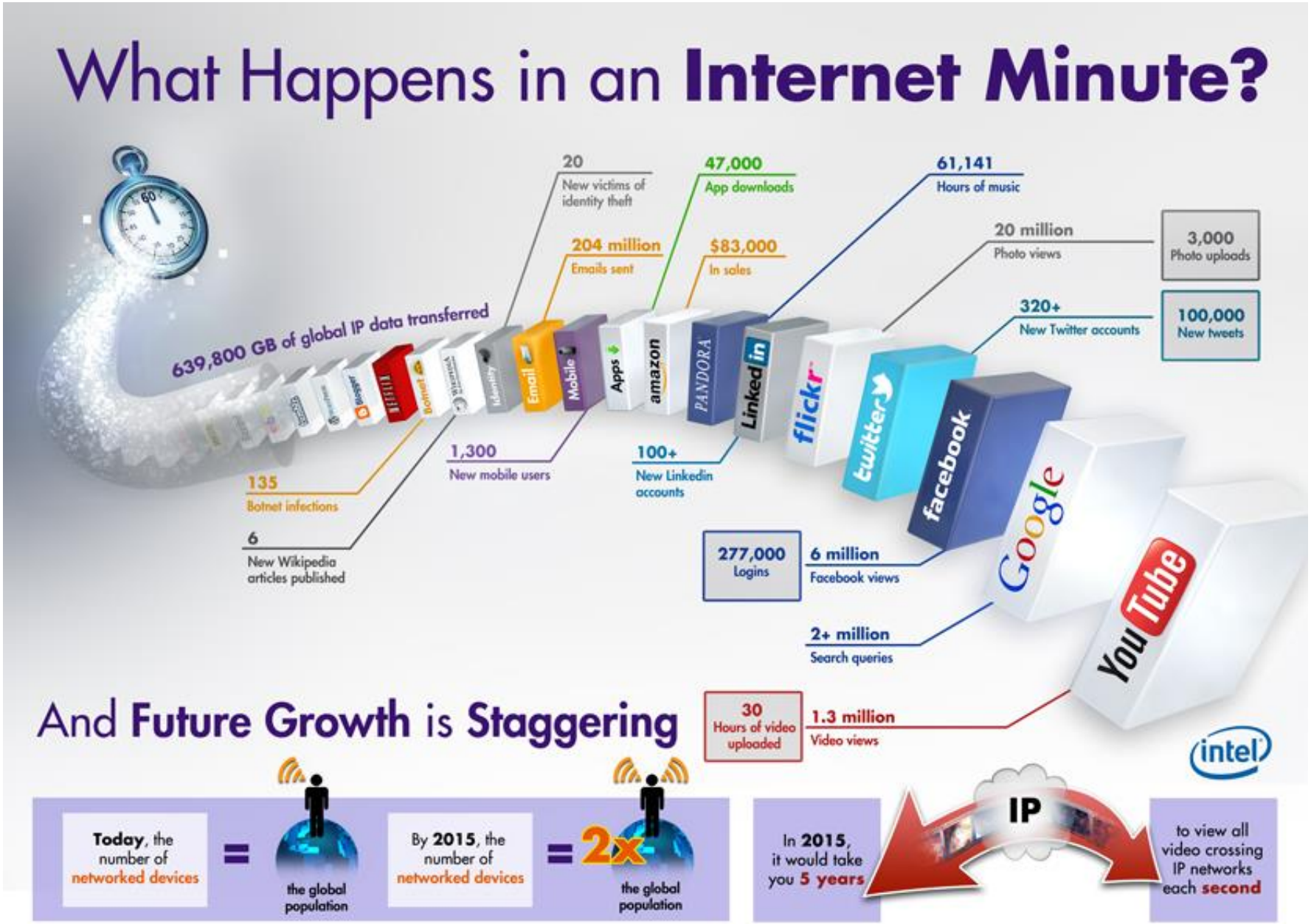
Data vs. God

*"In God we trust, all others bring **data**."*

– W.E. Deming



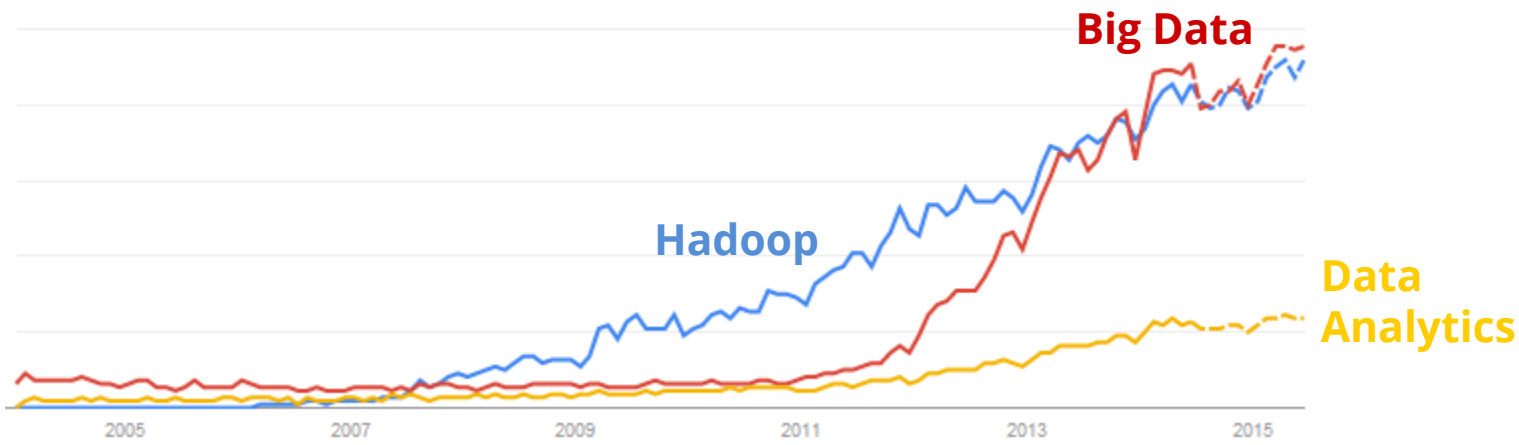
Source: M. Deutscher, When Will the World Reach 8 Zetabytes of Stored Data? (2012).

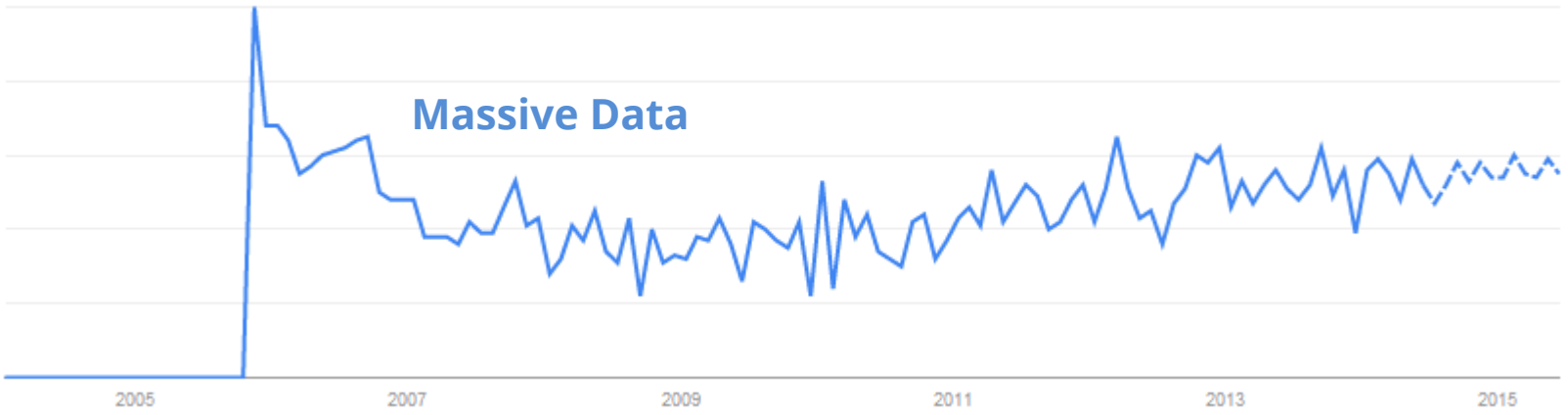


Source: Intel (2014), What Happens In An Internet Minute?

Big Data in Facebook

Google trends on Big Data





Father to the ‘Big Data’ term

John R. Mashey
Chief Scientist at Silicon Graphics

Source: S. Lohr (2013), The Origins of ‘Big Data’: An Etymological Detective Story.

Big Data: *think-tank* Policy Exchange

- **Big Data**: datasets that are too awkward to work with using traditional, hands-on database management tools.
- **Big Data Analytics**: the process of examining and **interrogating** big data assets to derive **insights** of value for decision making.

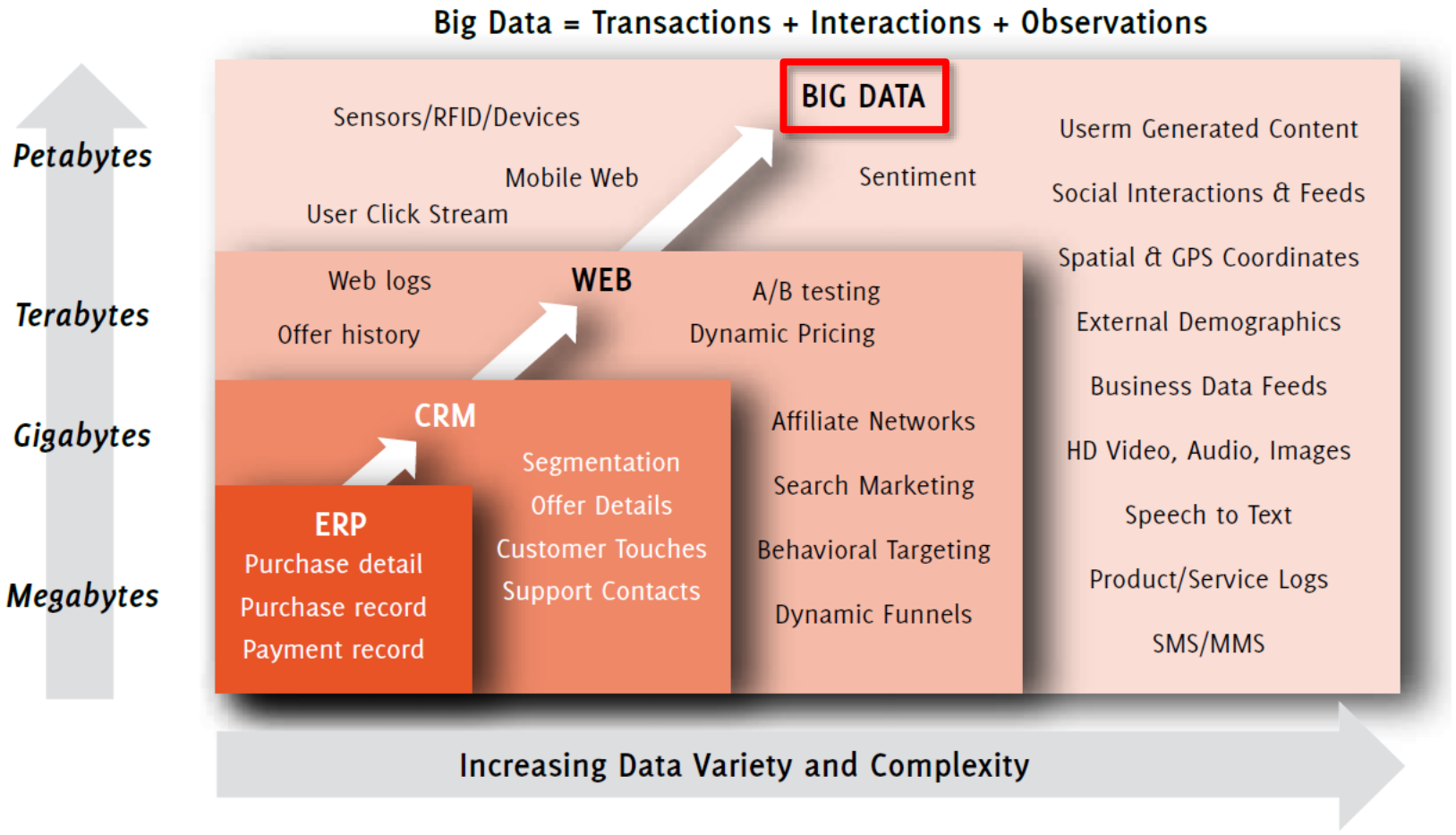
Source: C. Yiu (2012), The Big Data Opportunity.

What is Big Data?

Big Data is a term that describes **large volumes** of **high velocity**, **complex** and **variable data** that require advanced techniques and technologies to enable the capture, storage, distribution, management, and analysis of the information.

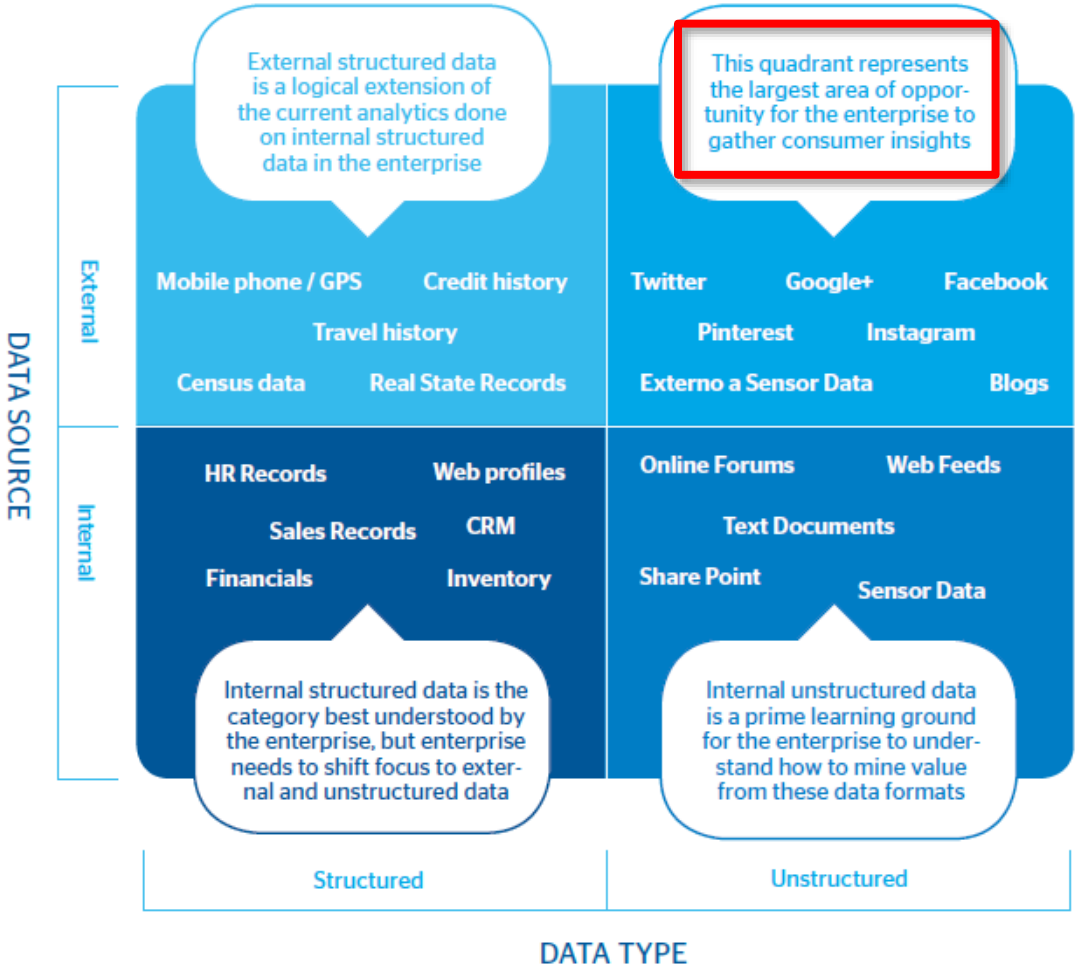
Source: Demystifying Big Data (2012), TechAmerica Foundation.

Big Data



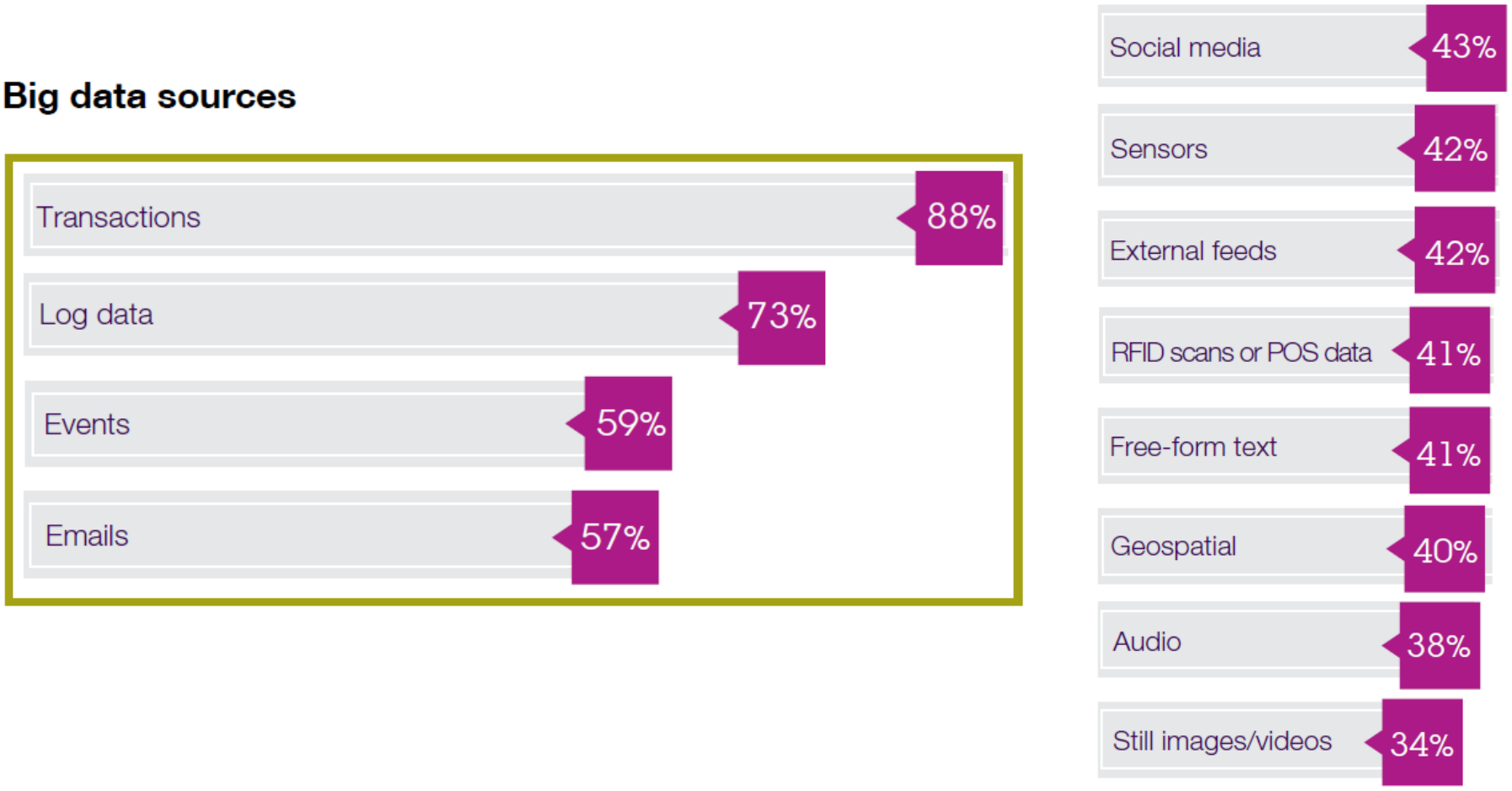
Source: J. Bloem *et al*. (2012), VINT Research Report 1: Creating Clarity with Big Data.

Sources & types of data



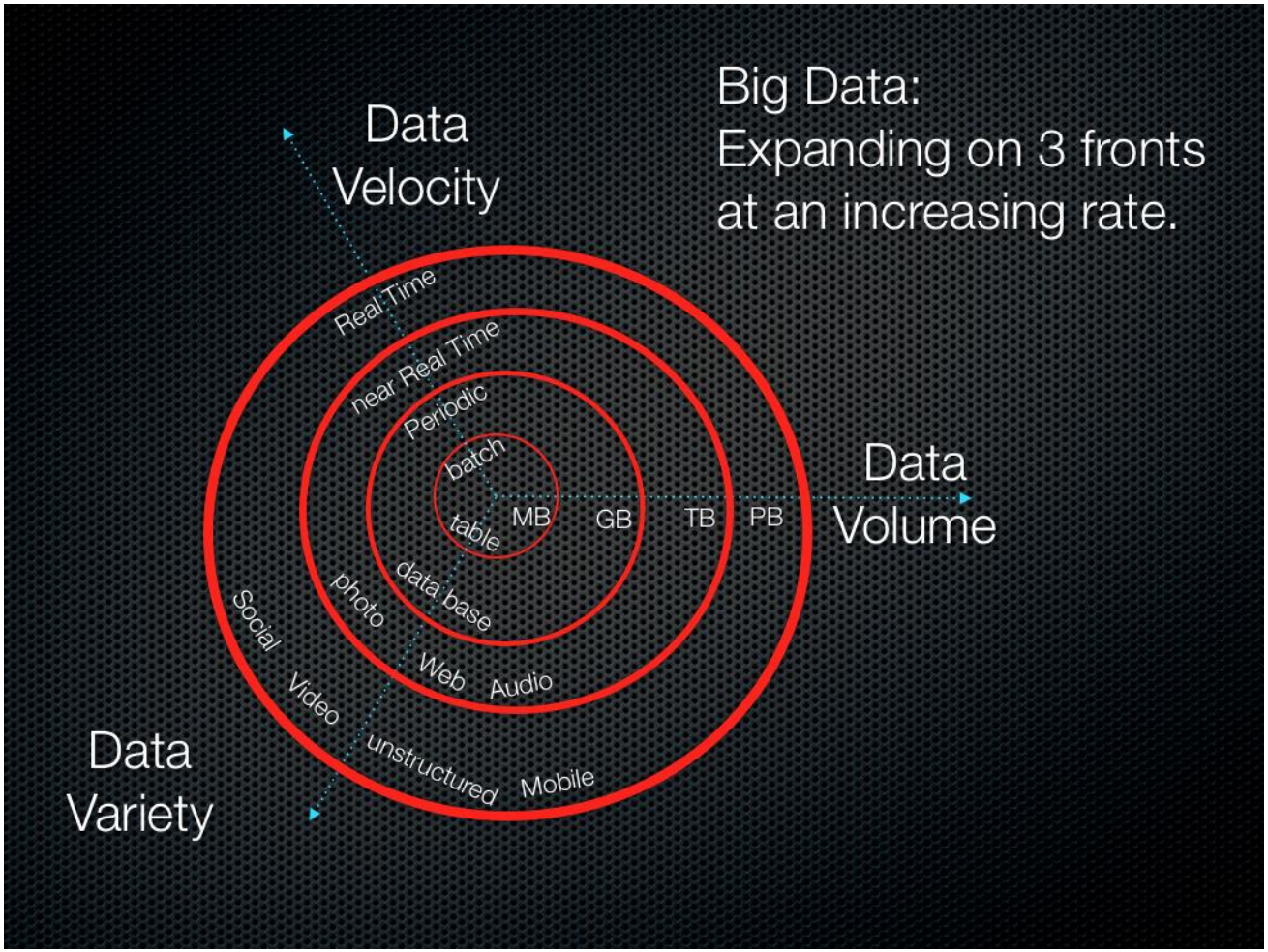
Source: Big Data, BBVA Innovation Edge 2013 (from Booz & Company "Benefitting from Big Data: Leveraging Unstructured Data Capabilities for Competitive Advantage")

Big Data sources



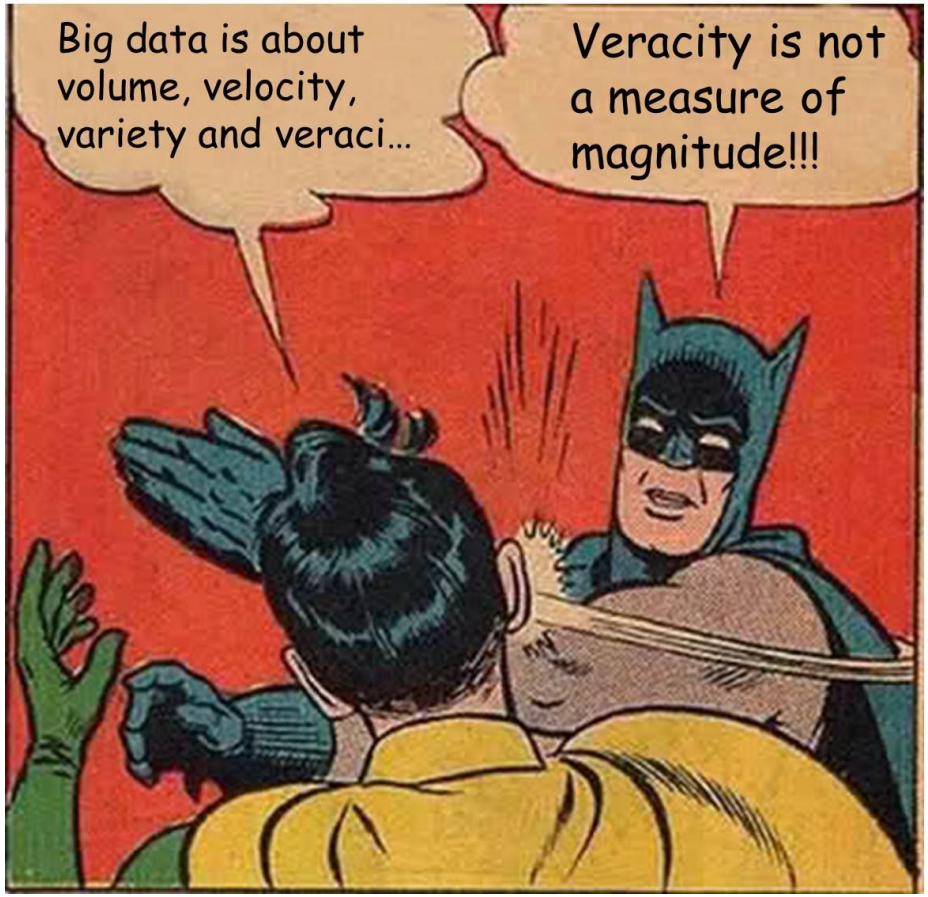
Source: M. Schroeck *et al*. (2012), Analytics: The Real-World Use of Big Data.

The three Vs of Big Data



Source: D. Soubra (2012), The 3Vs that define Big Data.

The other "Vs" in Big Data



*" 'Vs' like **veracity**, **validity**, **value**, **viability**, etc. are aspirational **qualities of all data**, not definitional qualities of Big Data."*

– Doug Laney

Source: D. Laney (2013), Batman on Big Data.

What is really important in Big Data?

*"The **Big** in Big Data relates to **importance** not size"*

– Rafael Irizarry

Source: R. Irizarry (2014), The Big in Big Data relates to importance not size.

My best “V”

Is Big Data a marketing campaign?

"If you're like me, the mere mention of Big Data now turns your stomach.

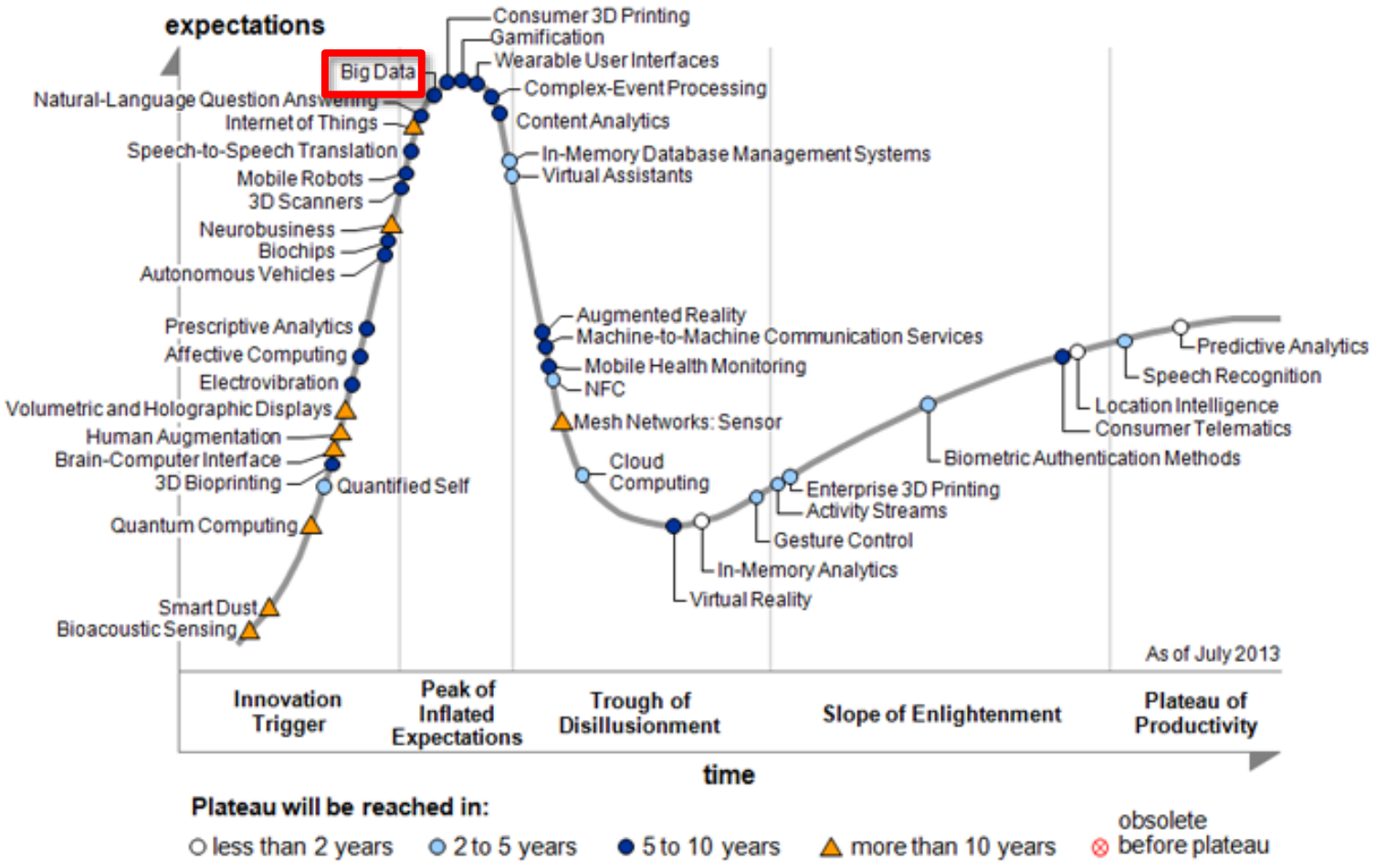
Nearly every business intelligence (BI) vendor, publication, and event has ***Big Data flashing*** *in neon colors in Times Square dimensions.*

Never before have I seen an idea in the BI space elicit this much obsession. Why all the fuss? Why, indeed.

Essentially, ***Big Data is a marketing campaign****, pure and simple."*

– Stephen Few

Gartner's 2013 Hype Cycle



Source: Gartner's 2013 Hype Cycle for Emerging Technologies

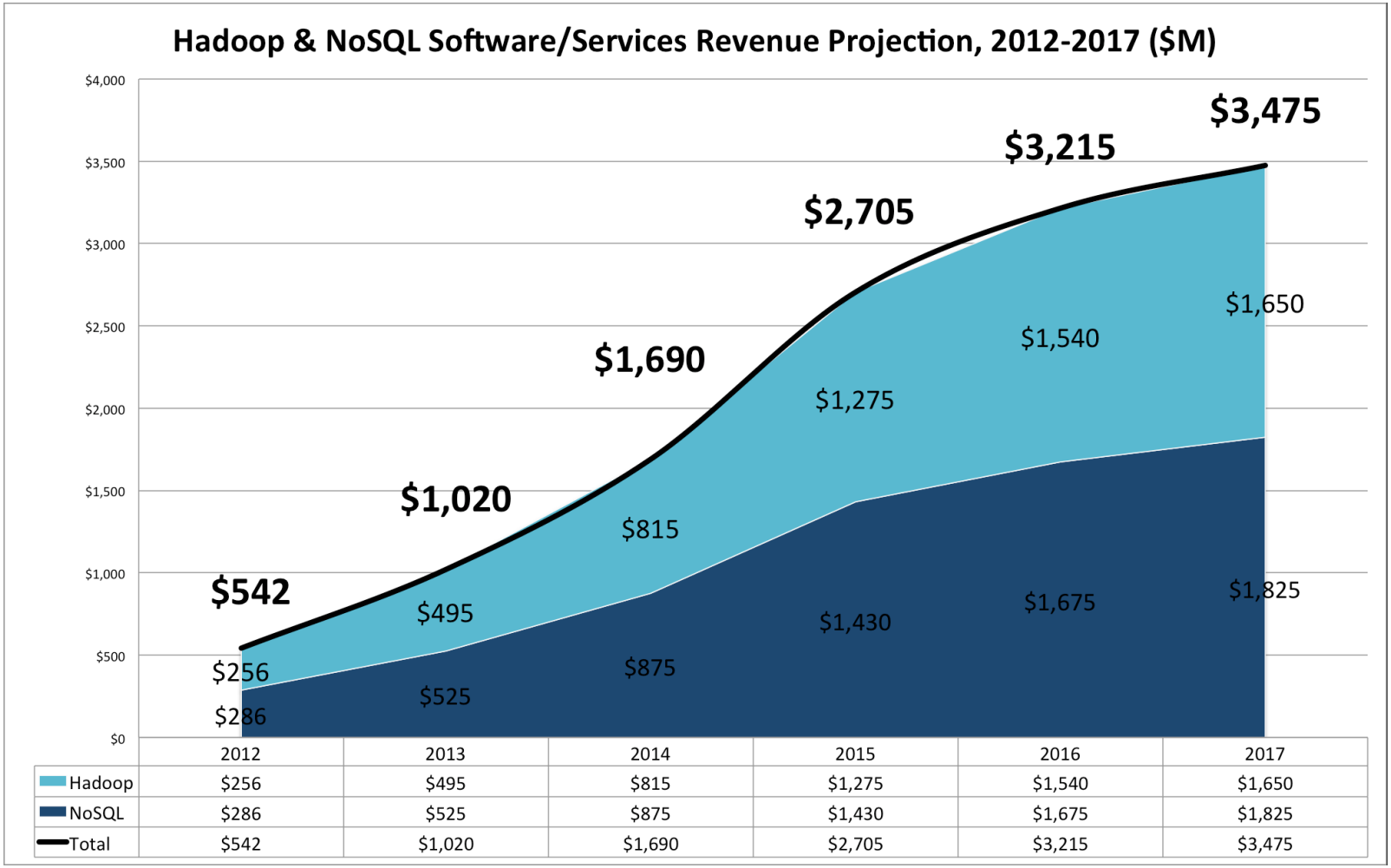
Big Data: McKinsey Report

- 140.000 – 190.000 more deep **analytical talent positions**, and 1.5 million **data savvy managers** needed to take full advantage of Big Data in the USA.
- **Techniques**: data mining (cluster analysis, classification, regression, etc), (un)supervised learning, ML, neural networks, optimization, predictive modeling, statistics, simulation, etc.
- **Technologies**: BI, Cassandra, DW, ETL, Hadoop, HBase, Map/Reduce, R, RDBMS, etc.
- Potential of Big Data in **five domains**:
  - Healthcare
  - Public Sector
  - Retail
  - Manufacturing
  - Telecommunications.

Source: J. Manyika, et al. (2012), Big Data: The Next Frontier for Innovation, Competition and Productivity.


hadoop

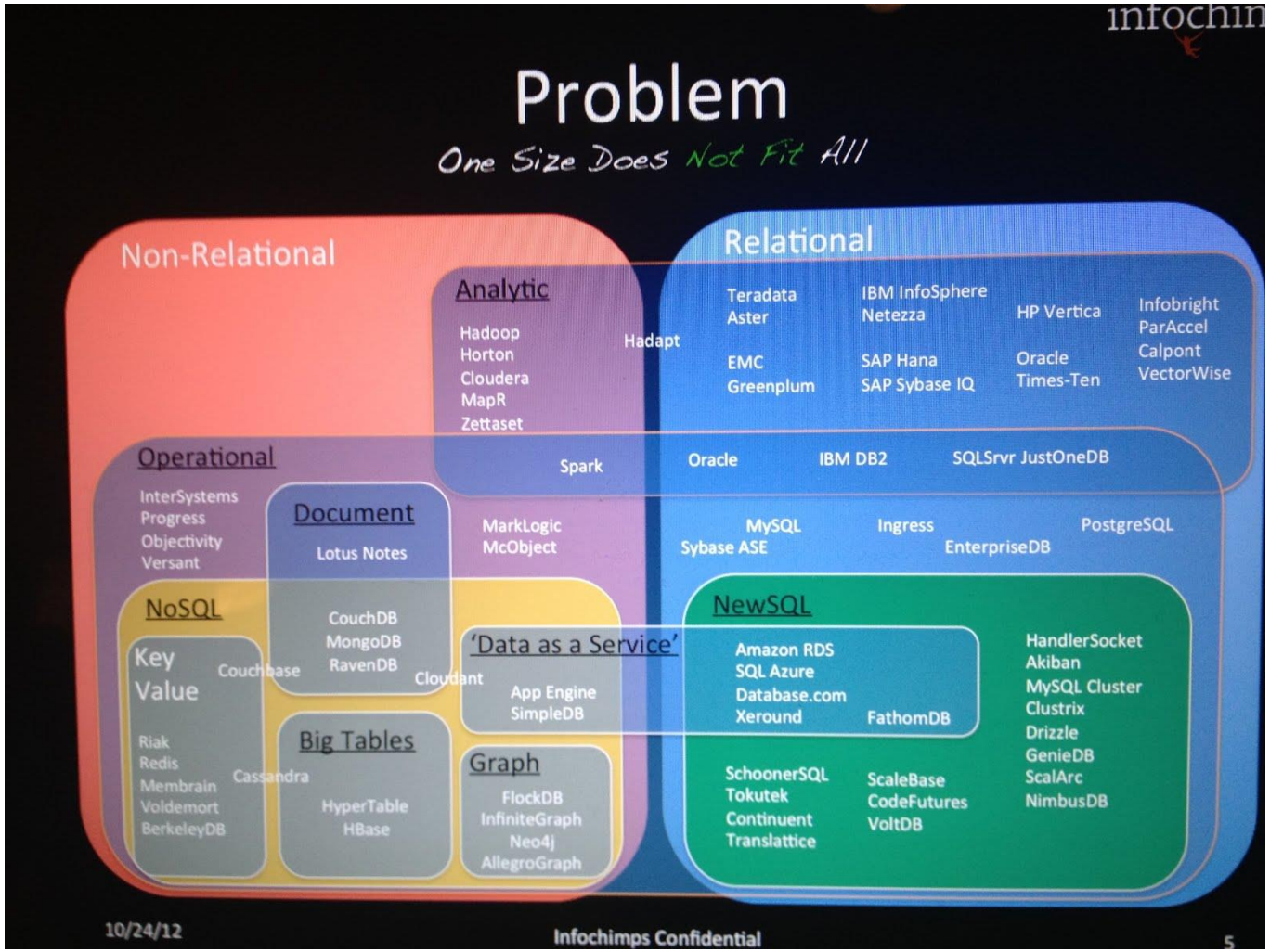
Hadoop-NoSQL Market Forecast 2012-2017



	2012	2013	2014	2015	2016	2017
Hadoop	$256	$495	$815	$1,275	$1,540	$1,650
NoSQL	$286	$525	$875	$1,430	$1,675	$1,825
Total	$542	$1,020	$1,690	$2,705	$3,215	$3,475

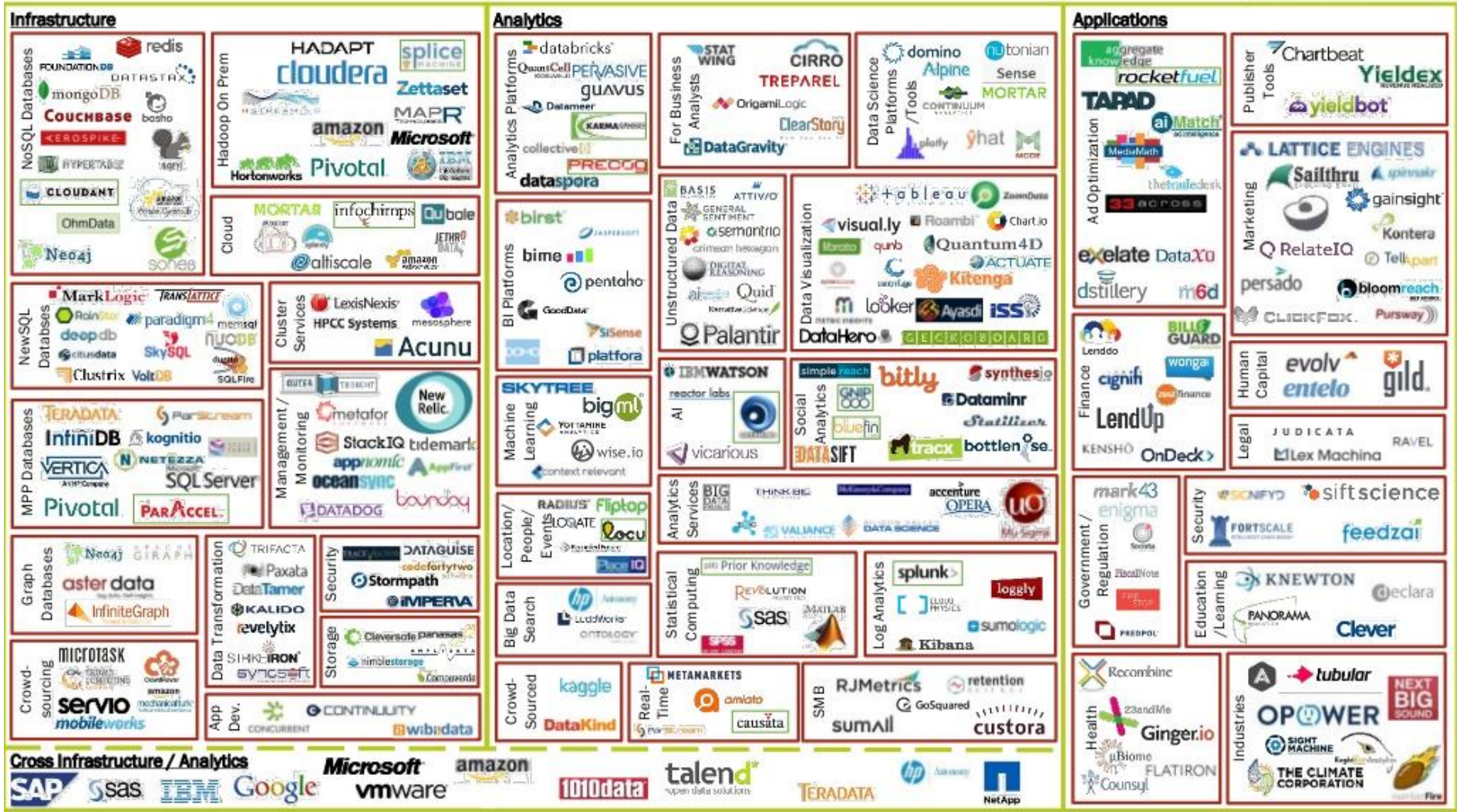
Source: J. Kelly (2013), Hadoop-NoSQL Software And Services Market Forecast 2012-2017.

Big Data Techs

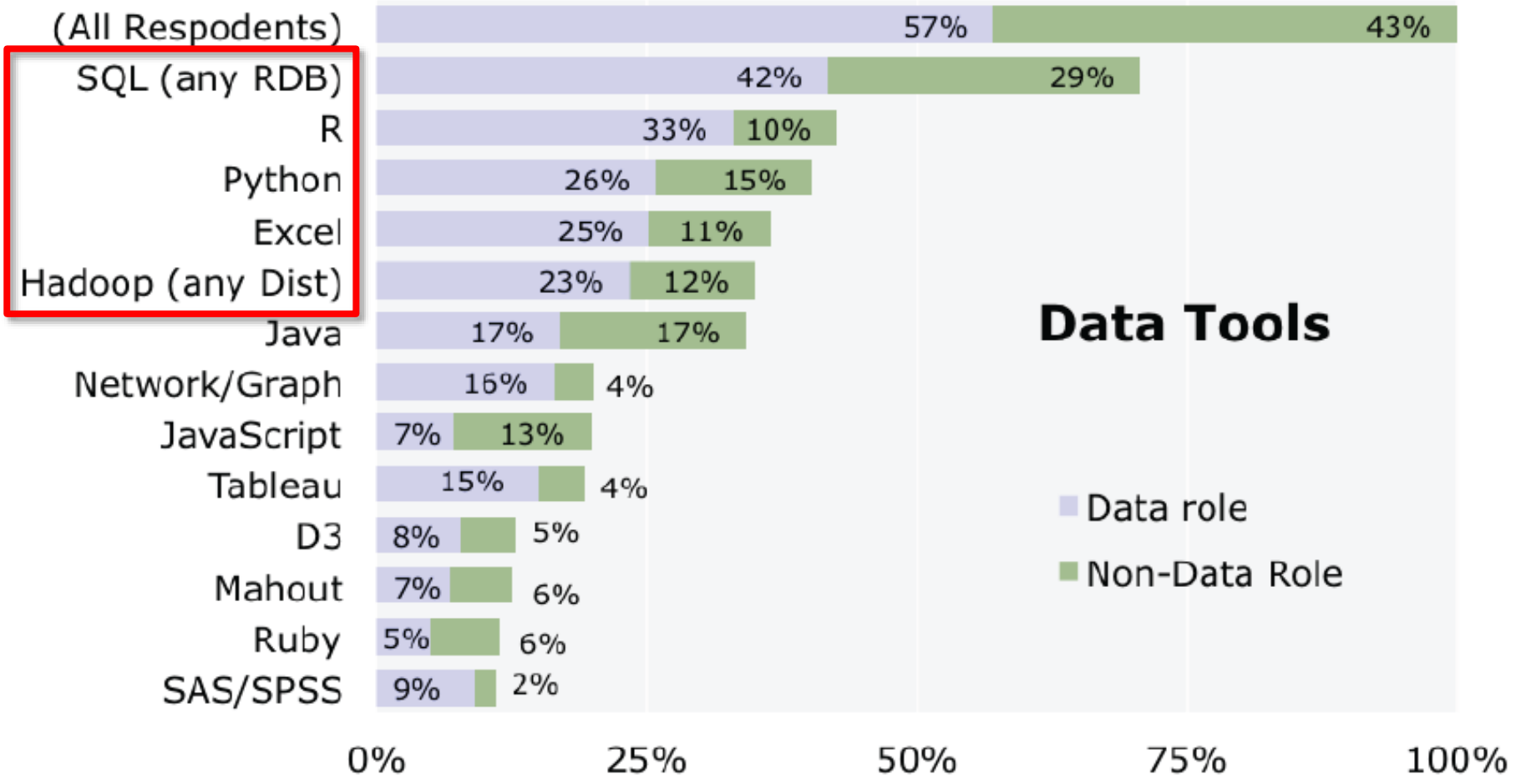


BIG DATA LANDSCAPE, VERSION 3.0

Exited: Acquisition or IPO

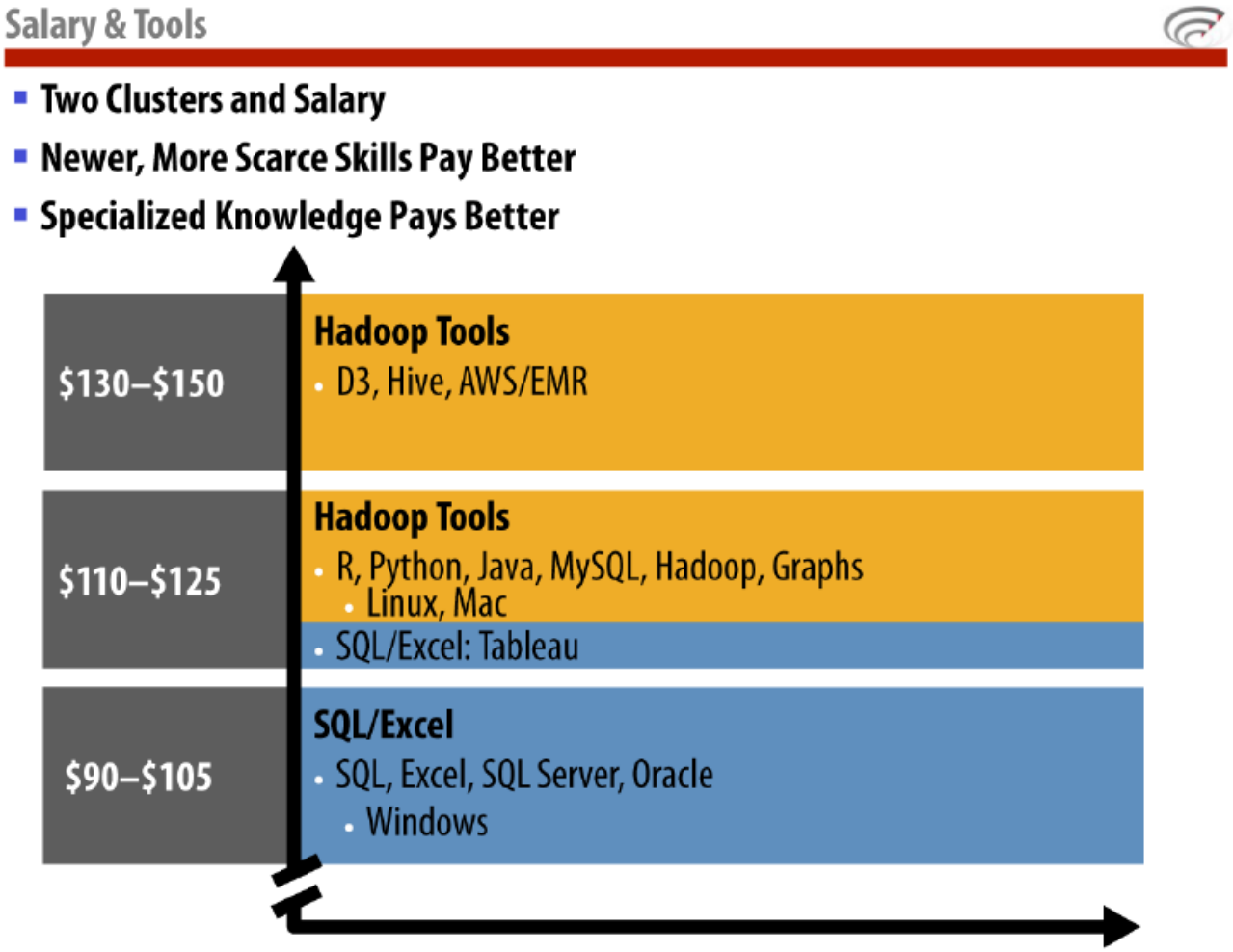


Data Tools



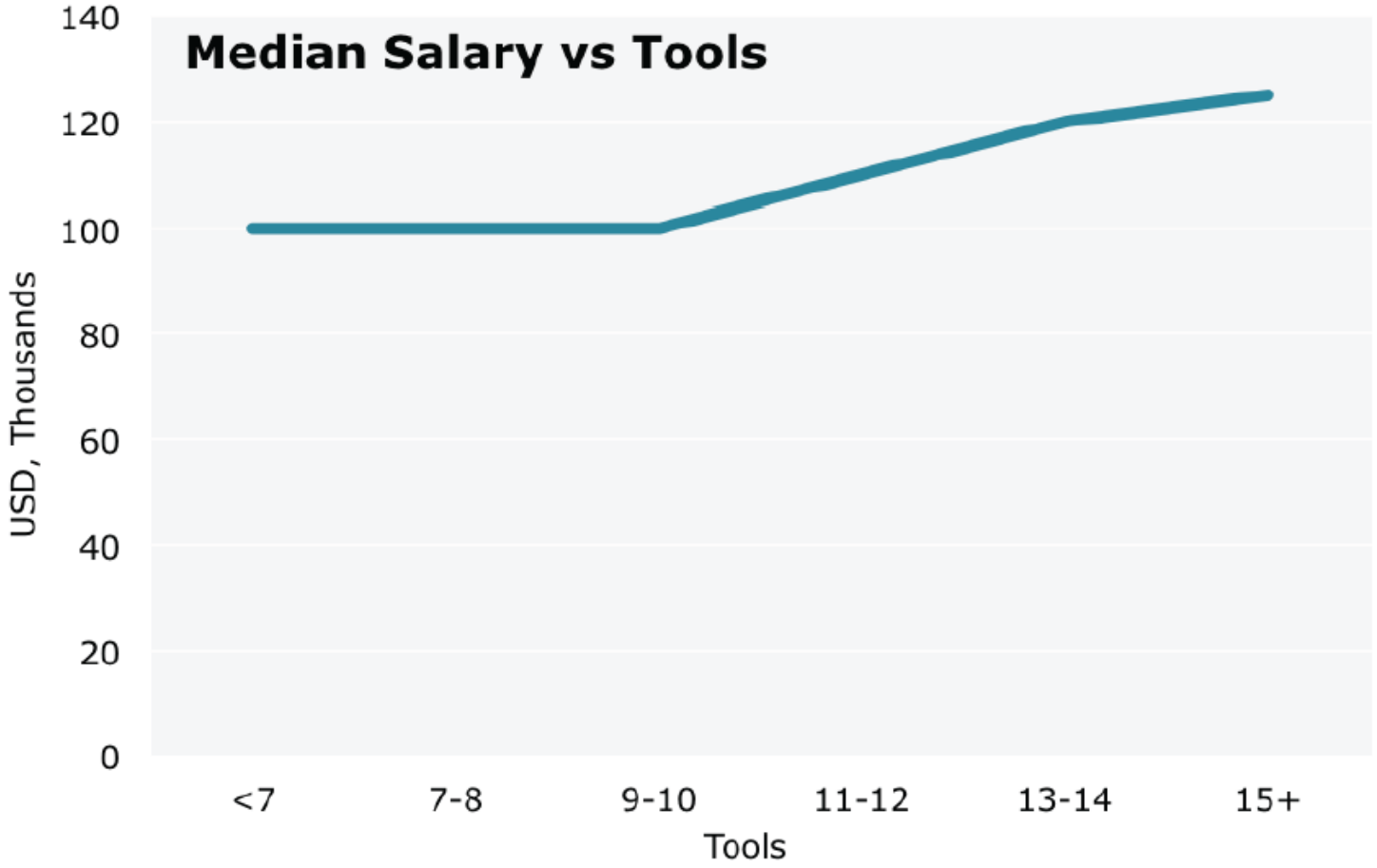
Source: J. King, R. Magoulas (2013), Data Science Salary Survey.

Salary vs. Data Tools



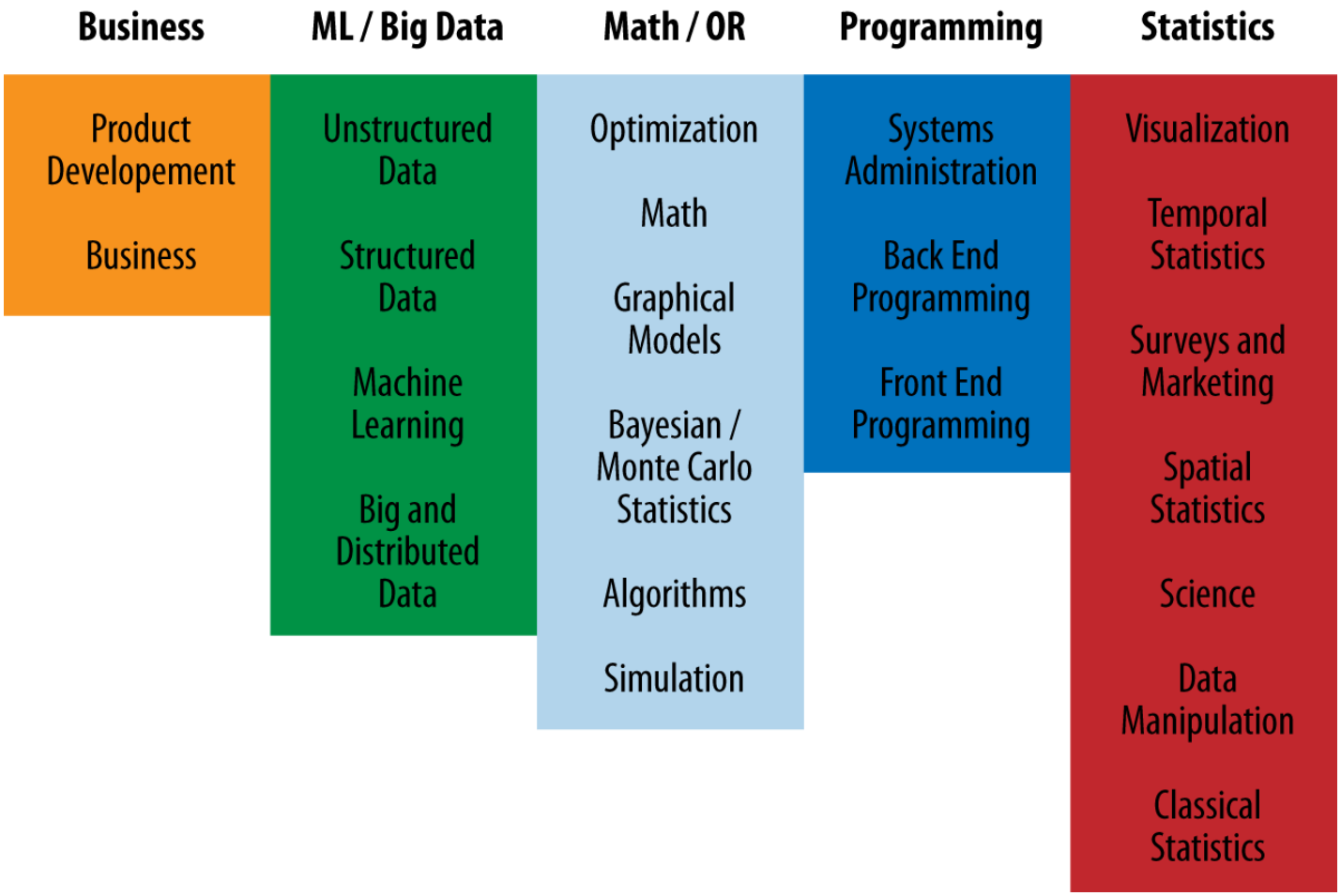
Source: J. King, R. Magoulas (2013), Data Science Salary Survey.

Median Salary vs. #Tools



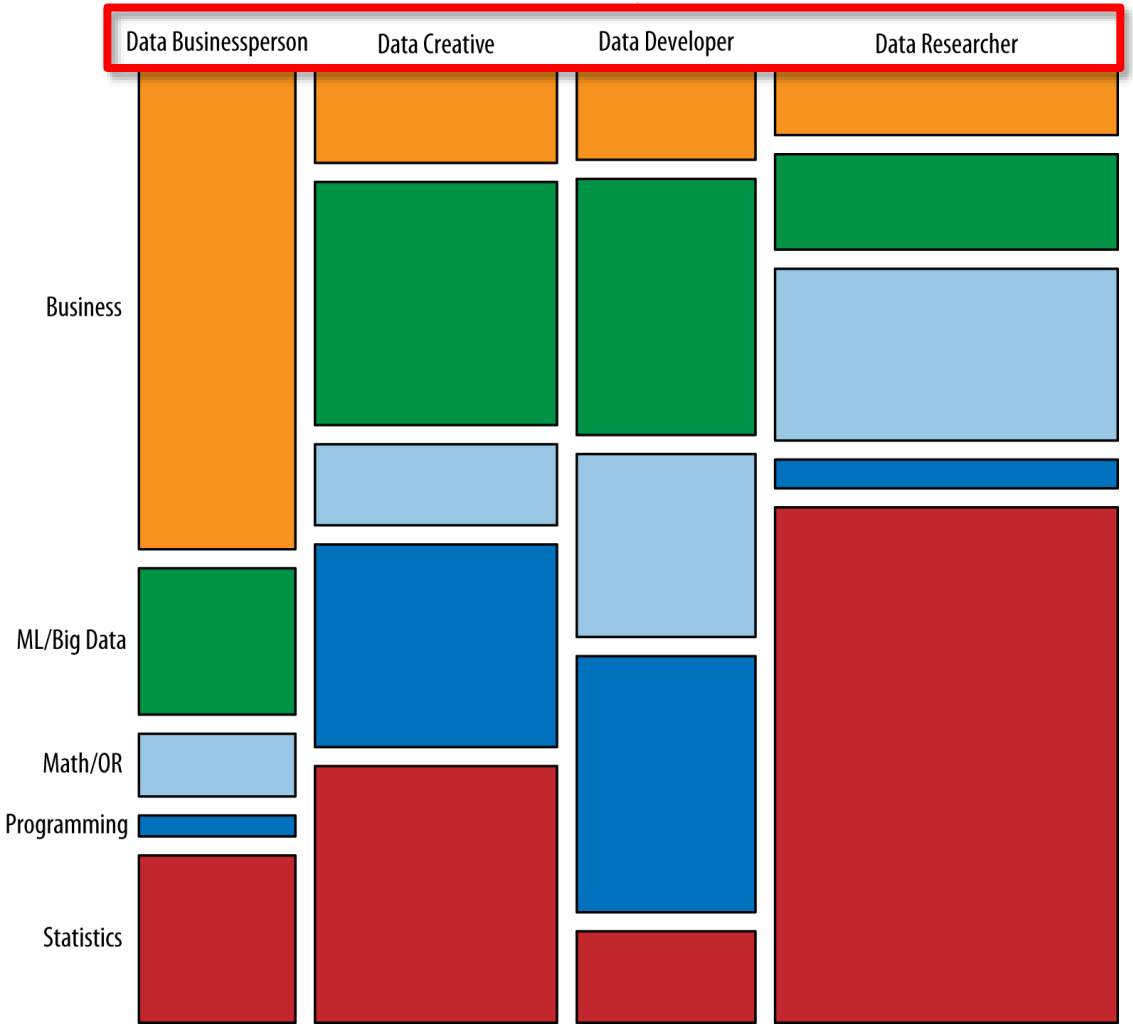
Source: J. King, R. Magoulas (2013), Data Science Salary Survey.

Data Skills



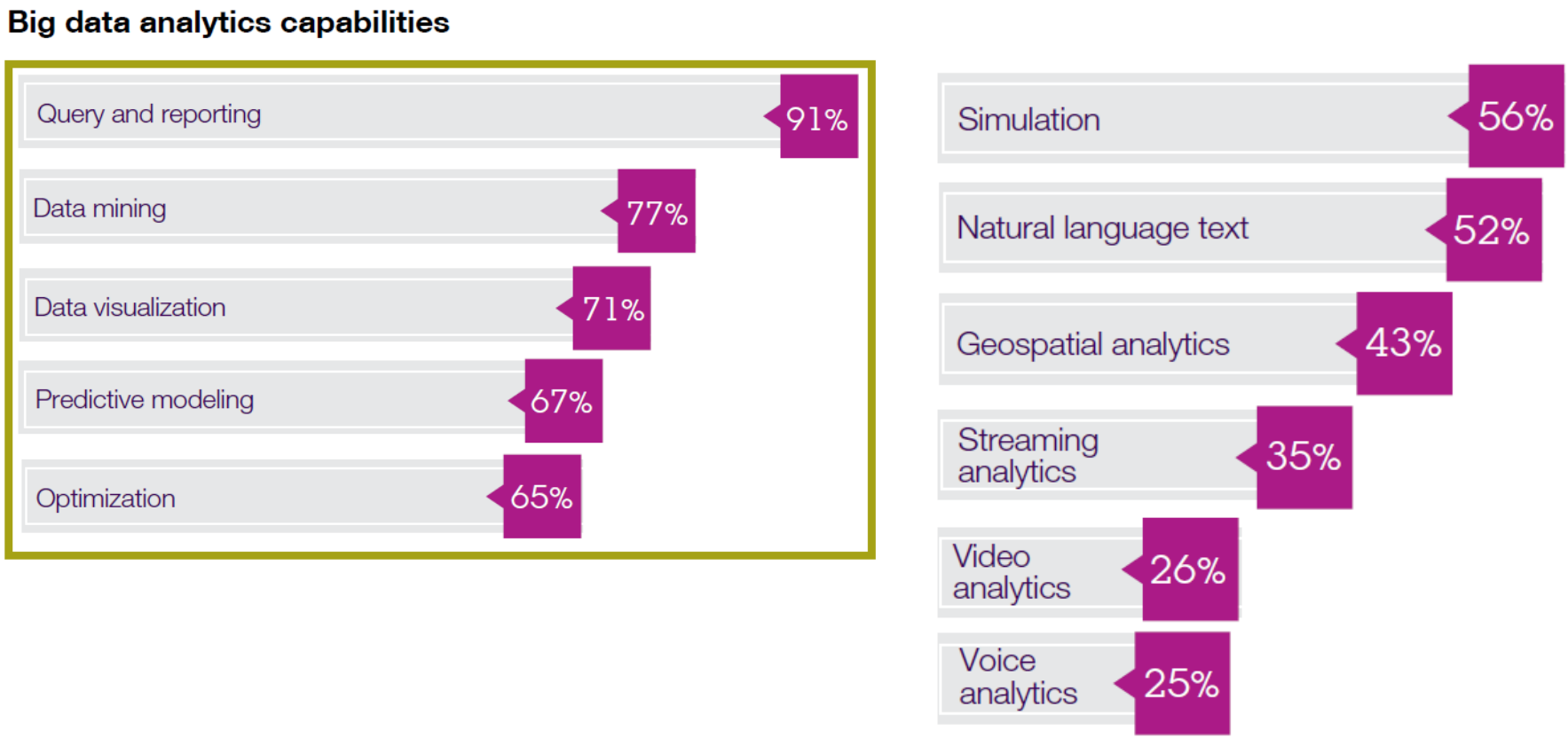
Source: H.D. Harris *et al.* (2013), Analyzing the Analyzers

Data Role vs. Data Skills



Source: H.D. Harris *et al.* (2013), Analyzing the Analyzers

Big Data capabilities



Source: M. Schroeck *et al.* (2012), Analytics: The Real-World Use of Big Dat.

Market & jobs opportunity

- The demand for **Big Data services** spending projected to reach **$132,300M** in 2015.
- By 2015, **Big Data demand** will reach **4.4 million jobs** globally, but **only one-third** of those jobs will be filled.
- The demand for services will generate **550,000 external services jobs** in the next **3 years**.
- Another **40,000 jobs** will be created at software vendors in the next 3 years.

Source: Big Data, BBVA Innovation Edge 2013 (from Gartner's "Top Technology Predictions for 2013 and Beyond")

Statiscian: a sexy job

*"I keep saying the **sexy job** in the next ten years will be **statisticians**.*

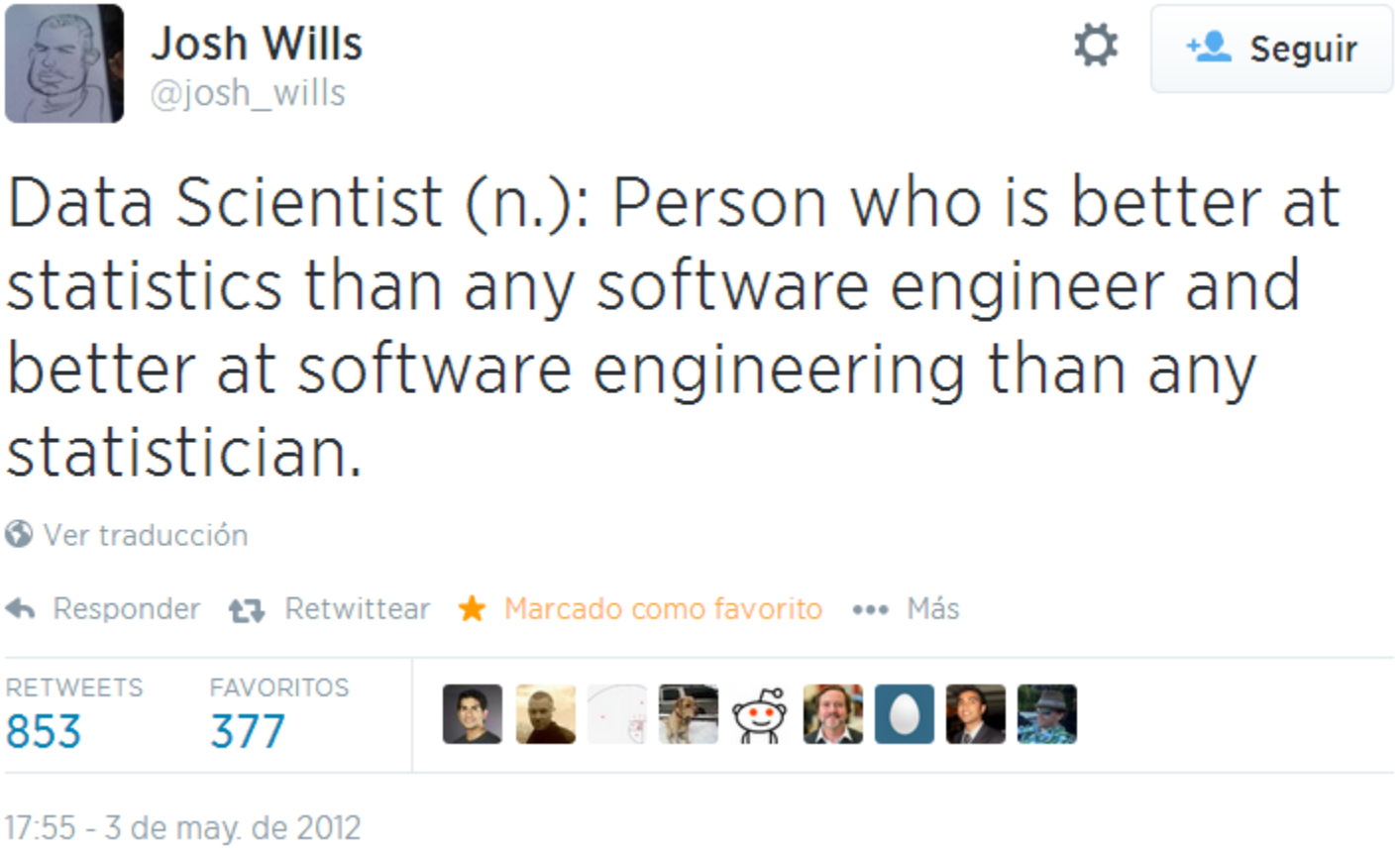
*People think I'm joking, but who would've guessed that **computer engineers** would've been the sexy job of the 1990s?*

*The ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it—that's going to be a **hugely important skill** in the next decades [...]"*

– Hal Varian
Google's Chief Economist

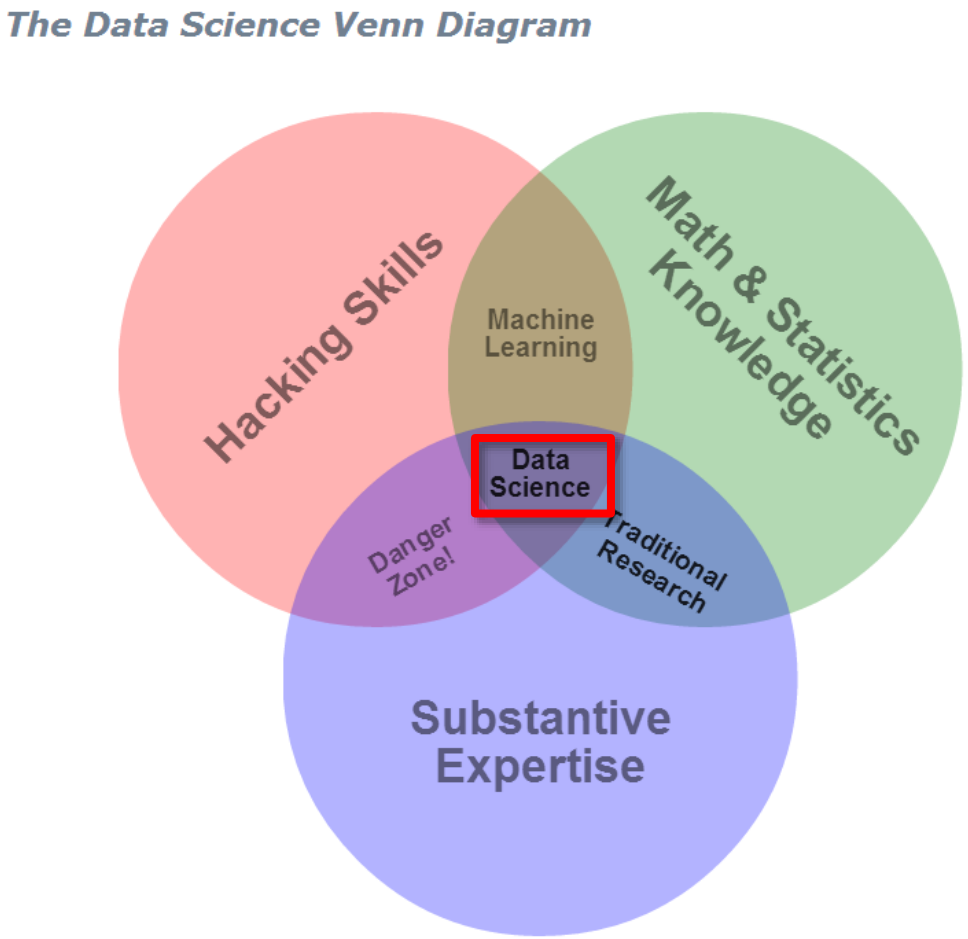
Source: Hal Varian on how the Web challenges managers, McKinsey & Co. 2009.

Data Scientist



Source: Josh Wills (2012).

Data Science Venn Diagram



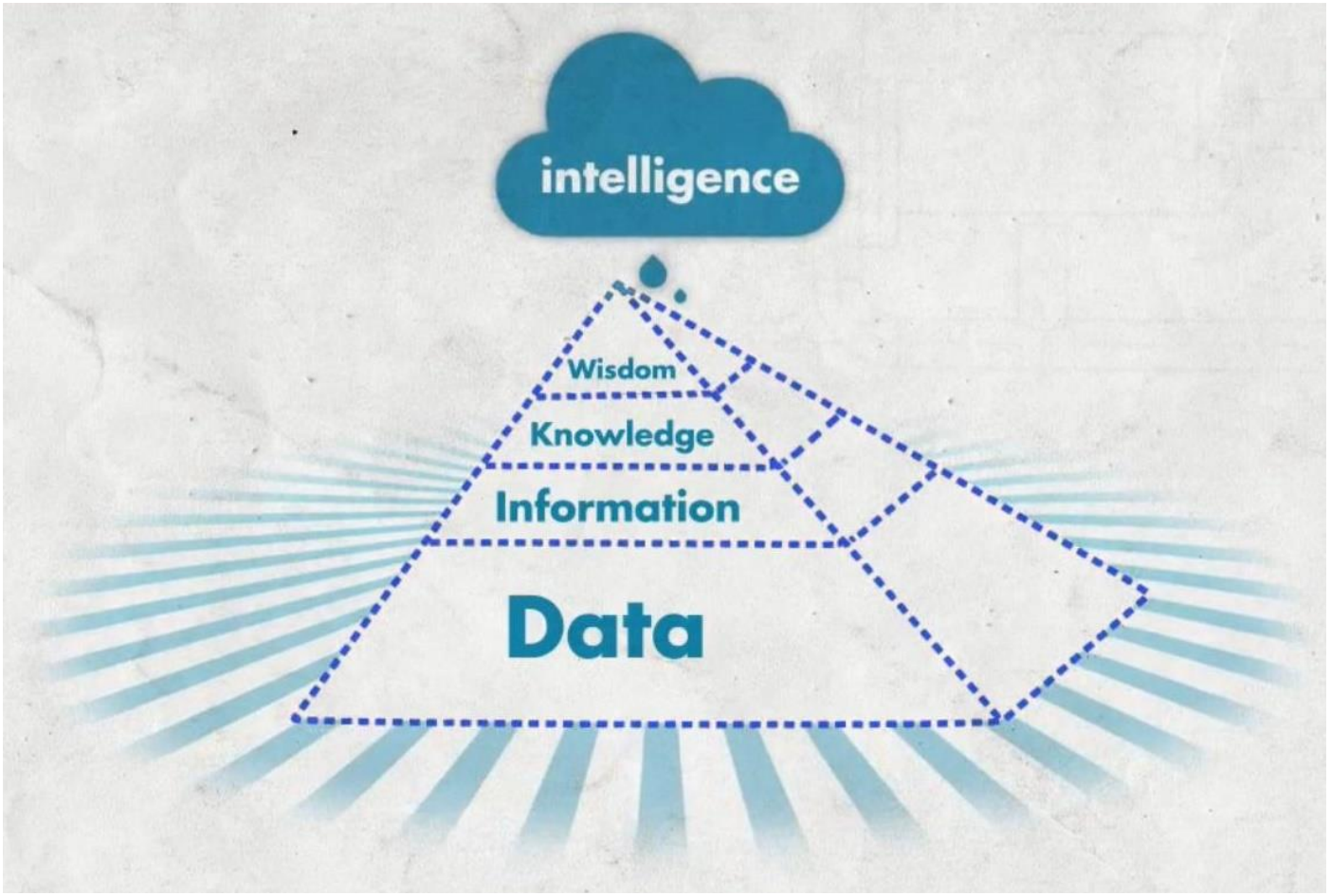
Source: Drew Conway (2010).

Data Scientist skill set: ACM

A data scientist requires an integrated skill set spanning **mathematics**, **machine learning**, **artificial intelligence**, **statistics**, **databases**, and **optimization**, along with a deep understanding of the craft of **problem formulation** to engineer effective solutions.

Source: V. Dhar (2013), Data Science and Prediction, Comm. of the ACM.

Intelligence over DIKW



Source: The Internet of Things 2010 at YouTube (1:40).

Data→Info→Knowledge→Understanding →Wisdom!!

*"There are **known knowns**. These are things we know that we know.*

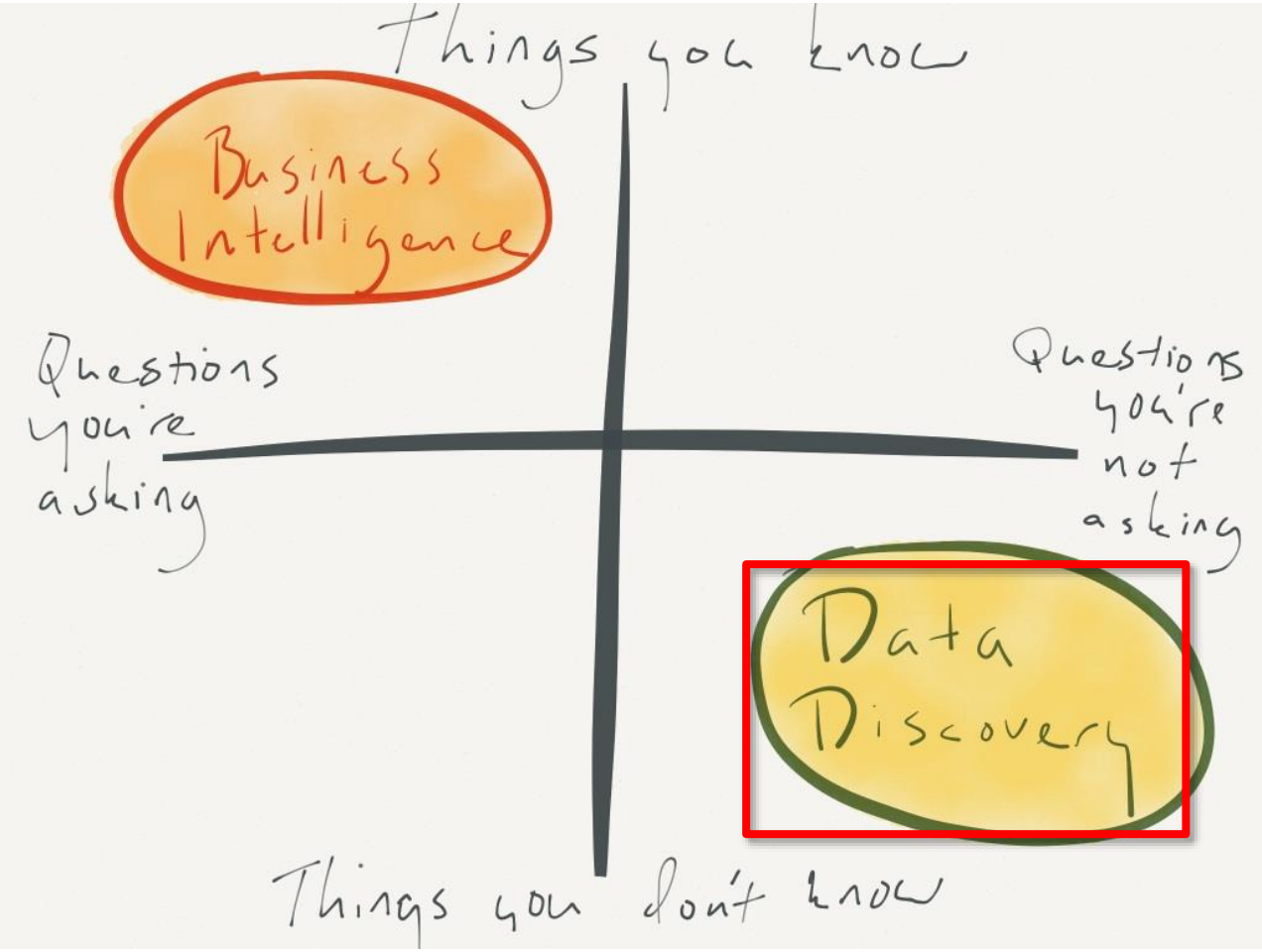
*There are **known unknowns**. That is to say, there are things that we know we don't know.*

*But there are also **unknown unknowns**. There are things we don't know we don't know."*

– Donald Rumsfeld

Source: C. Somohano (2013), Big Data [sorry] & Data Science: What Does a Data Scientist Do?

BI vs. Data Discovery



Source: J. Kolb (2010), The New Reality for Business Intelligence and Big Data.

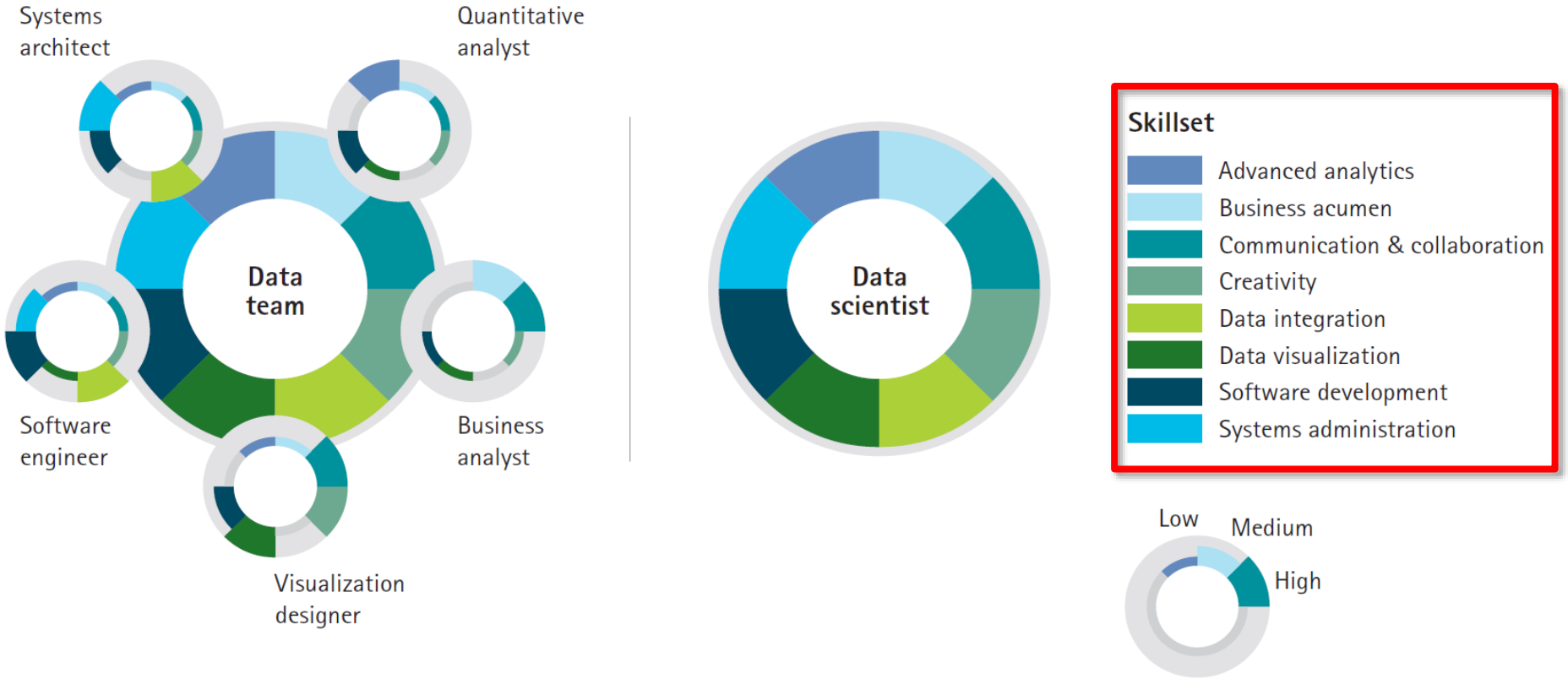
Data Science Teams

Data scientists as having the following qualities:

- **Technical expertise**: the best data scientists typically have deep expertise in some scientific discipline.
- **Curiosity**: a desire to go beneath the surface and discover and distill a problem down into a very clear set of hypotheses that can be tested.
- **Storytelling**: the ability to use data to tell a story and to be able to communicate it effectively.
- **Cleverness**: the ability to look at a problem in different, creative ways.

Source: D.J. Patil (2011), Building Data Science Team.

Data Science skills: Accenture



Source: J.G. Harris *et al.* (2013), The Team Solution to the Data Scientist Shortage.

Insight Data Science Fellow Program

- **6 week**, full-time, **postdoctoral** data science training fellowship in Silicon Valley or New York City.
- Self-directed, project-based learning (**no classes!**).
- **Software Engineering Best Practices**: Python, Git, Flask, Javascript.
- **Storing and Retrieving Data**: MySQL, Hadoop, Hive.
- **Statistical Analysis & Machine Learning**: NumPy & SciPy, Pandas, scikit-learn, R.
- **Visualizing and Communicating Results**: D3 Javascript library, visualization and presentation best practices.

Insight Data Engineering Fellow Program

- **6 week**, full-time, **professional** data engineering training fellowship in Silicon Valley, California.
- Self-directed, project-based learning (**no classes!**).
- **Big Data Infrastructure**.
- **Extracting** data.
- **Transforming** data.
- **Loading / Storing** data.
- Building **visualizations** and **dashboards**.

Conclusions

- Big Data is still an **emerging topic** that gathers a lot of new technologies, and needs some time to mature.
- But, on the other hand, it has a **true market** opportunity.
- Data Science / Engineering skills to acquire:
  - Math/Statistics and business knowledge.
  - Technical expertise: R, Python, Hadoop, Spark/Storm, D3, Java/Javascript, ...
  - Curiosity and cleverness.
  - Storytelling: ability to communicate results.
- Trends:
  - **Data Visualization**
  - Predictive Modelling
  - Social Analytics
  - Data Mining / Machine Learning
  - Forensic Computer Science
  - Spark / Storm vs. Hadoop MapReduce

References (1/3)

1. Big Data (2013), **BBVA Innovation Edge** (31 pp).
2. Demystifying Big Data: A Practical Guide To Transforming The Business of Government (2012), **TechAmerica Foundation** (40 pp).
3. Gartner's 2013 Hype Cycle for Emerging Technologies Maps Out Evolving Relationship Between Humans and Machines (2013), **Gartner**.
4. Hal Varian on How the Web Challenges Managers (2009), **McKinsey & Co.**
5. Insight Data Engineering Fellows Program (2014).
6. Insight Data Science Fellows Program (2014).
7. The Internet of Things (2010), **IBM Social Media**.
8. What Happens In An Internet Minute? (2014), **Intel**.

References (2/3)

9. J. Bloem, M. van Doorn, S. Duivestein, T. van Manen, E. van Ommeren (2012), VINT Research Report 1: Creating Clarity with Big Data, **SOGETI**.
10. D. Conway (2010), The Data Science Venn Diagram.
11. M. Deutscher, When Will the World Reach 8 Zetabytes of Stored Data? (2012), Silicon Angle (blog).
12. V. Dhar (2013), Data Science and Prediction, **Communications of the ACM** 56 (12), pp. 64-73.
13. S. Few (2012), Big Data, Big Ruse, **Perceptual Edge - Visual Business Intelligence Newsletter** (blog, 8 pp).
14. H.D. Harris, S.P. Murphy, M. Vaisman (2013), Analyzing the Analyzers, **O'Reilly Media** (40 pp).
15. J.G. Harris, N. Shetterley, A.E. Alter, K. Schnell (2013), The Team Solution to the Data Scientist Shortage, **Accenture Institute for High Performance**.
16. R. Irizarry (2014), The Big in Big Data Relates to Importance Not Size, **Simply Statistics** (blog).
17. J. King, R. Magoulas (2013), Data Science Salary Survey, **O'Reilly Media** (23 pp).
18. J. Kelly (2013), Hadoop-NoSQL Software and Services Market Forecast 2012-2017, **Wikibon** (blog).
19. J. Kolb (2010), The New Reality for Business Intelligence and Big Data, **Applied Data Labs** (blog).
20. D. Laney (2013), Batman on Big Data, **Gartner**.

References (3/3)

21. D. Laney (2013), Batman on Big Data, **Gartner**.
22. S. Lohr (2013), The Origins of 'Big Data': An Etymological Detective Story, **The New York Times**.
23. J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, A.H. Byers (2012), Big Data: The Next Frontier for Innovation, Competition and Productivity, **McKinsey Global Institute** (156 pp).
24. R. Nair, A. Narayanan (2012), Benefitting from Big Data: Leveraging Unstructured Data Capabilities for Competitive Advantage, **Booz & Company** (16 pp).
25. D.J. Patil (2011), Building Data Science Teams, **O'Reilly Media** (26 pp).
26. G. Piatetsky (2014), Big Data Landscape v3.0 Analyzed, **KDnuggets** (blog).
27. J. Podesta, P. Pritzker, E.J. Moniz, J. Holdren, J. Zients (2014), Big Data: Seizing Opportunities, Preserving Values, **The White House** (79 pp).
28. M. Schroeck, R. Shockley, J. Smart, D. Romero-Morales, P. Tufano (2012), Analytics: The Real-World Use of Big Data, **IBM Global Services**.
29. C. Somohano (2013), Big Data [sorry] & Data Science: What Does a Data Scientist Do?, **Data Science London** (55 pp).
30. D. Soubra (2012), The 3Vs that define Big Data, **Data Science Central** (blog).
31. C. Yiu, The Big Data Opportunity (2012), **Policy Exchange** (36 pp).
32. P. Zikopoulos, C. Eaton, D. deRoos, T. Deutsch, G. Lapis (2012), Understanding Big Data, McGraw-Hill.

Datos de contacto y cuestiones

¡¡Gracias!!

¿Preguntas?

- Datos de contacto:
  - Marcos Colebrook
  - Email: mcolesan@ull.edu.es
  - Twitter: @MColebrook
  - SlideShare: www.slideshare.net/MarcosColebrookSantamaria