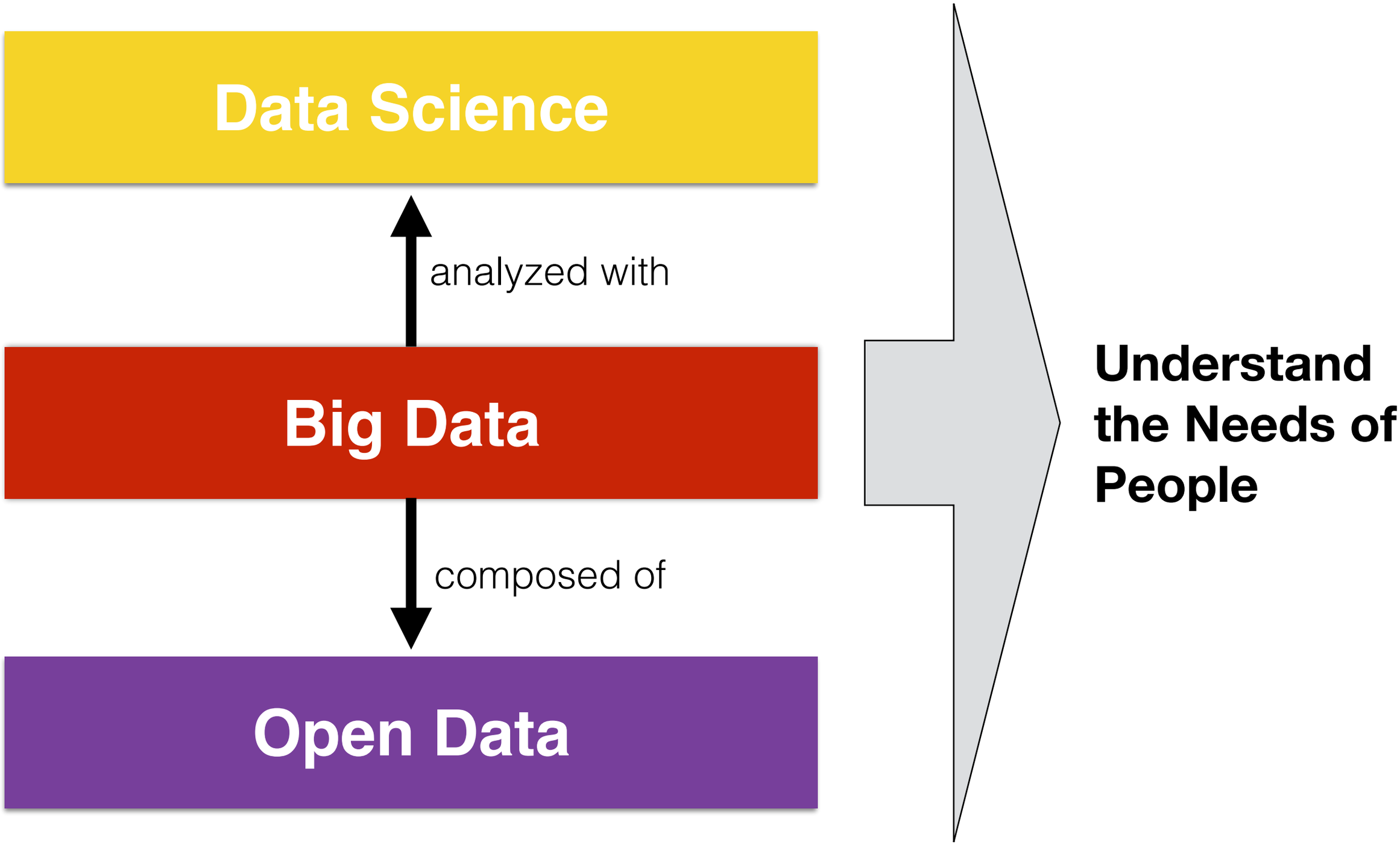
Big Data & Data Science Serving the Mission

Harvey Alférez, Ph.D.
Global Software Lab,
School of Engineering and Technology,
Universidad de Montemorelos, Mexico
www.harveyalferez.com

"the Savior mingled with men as one who desired their good. He showed His sympathy for them, ***ministered to their <u>needs</u>***, and won their confidence. Then He bade them, 'Follow Me.'" [1]

1. E. G. White, The Ministry of Healing (Review & Herald, 1905), p. 143.

My current interest…



The Digital Universe is Huge

- The digital universe is **doubling in size every two years**.
- By **2020** it will reach **44 zettabytes**, or **44 trillion gigabytes** [1].
- These facts have motivated **companies** and **scientists** in the last years to find new ways to understand **big data** in the digital universe.

1. EMC Corporation. (2014). The digital universe of opportunities: Rich data and the increasing value of the internet of things. URL: http://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm

- **Big data** is a termed that can be used to describe data sets so **large** and **complex** that they become difficult to work with using standard techniques [2].

- **Big data is the next big thing**. **<u>The new oil</u>** [3].

2. Snijders, C., Matzat, U., and Reips, U.-D. (2012). “Big data”: Big gaps of knowledge in the field of Internet science. International Journal of Internet Science, 1(1):1–5.

3. Rotella, P. (2012). Is data the new oil? URL: http://www.forbes.com/sites/perryrotella/2012/04/02/is-data-the-new-oil/

Big Data Innovation Summit

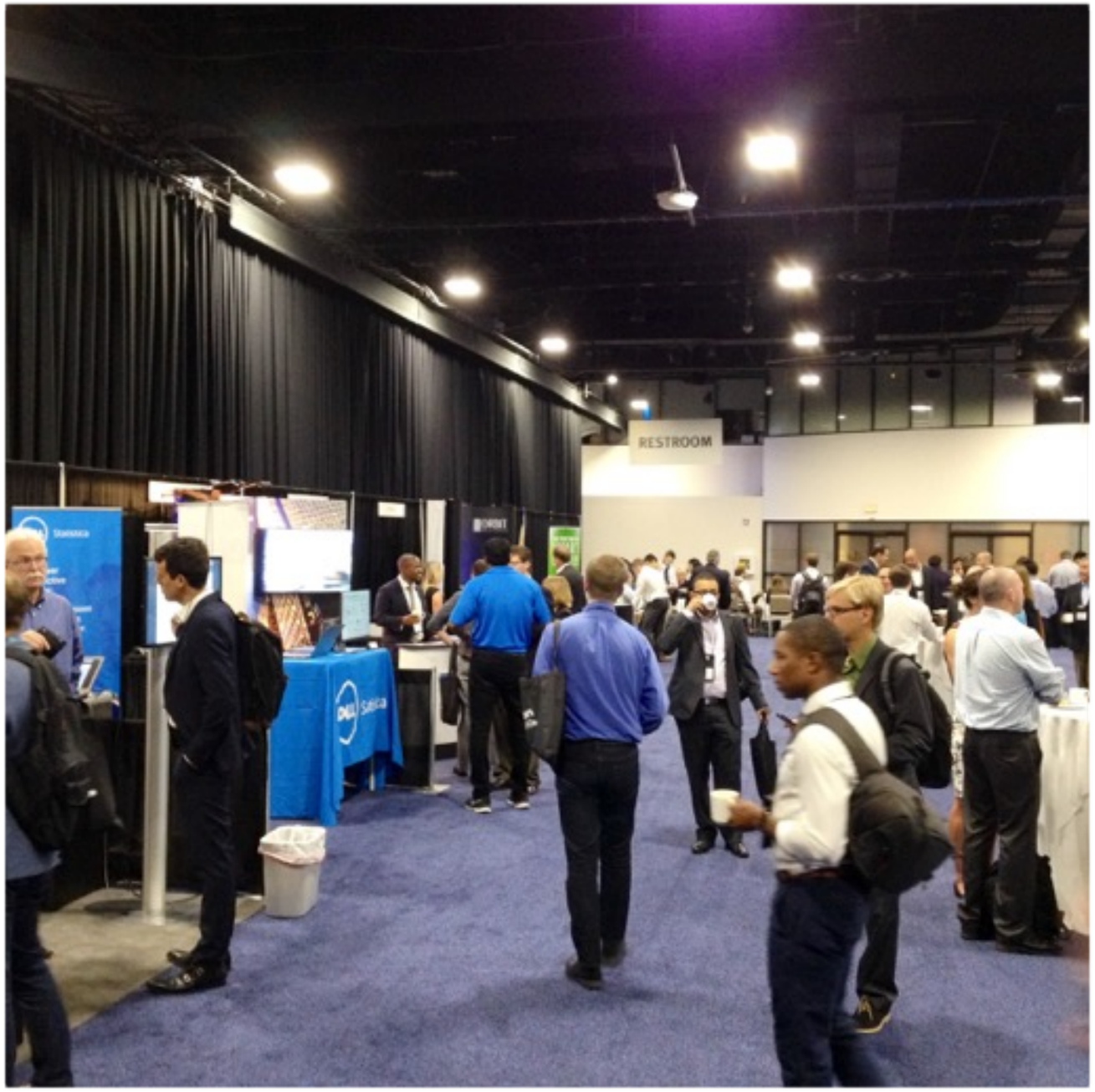
"Strategize Your Data Capabilities To Maximize Business Performance"

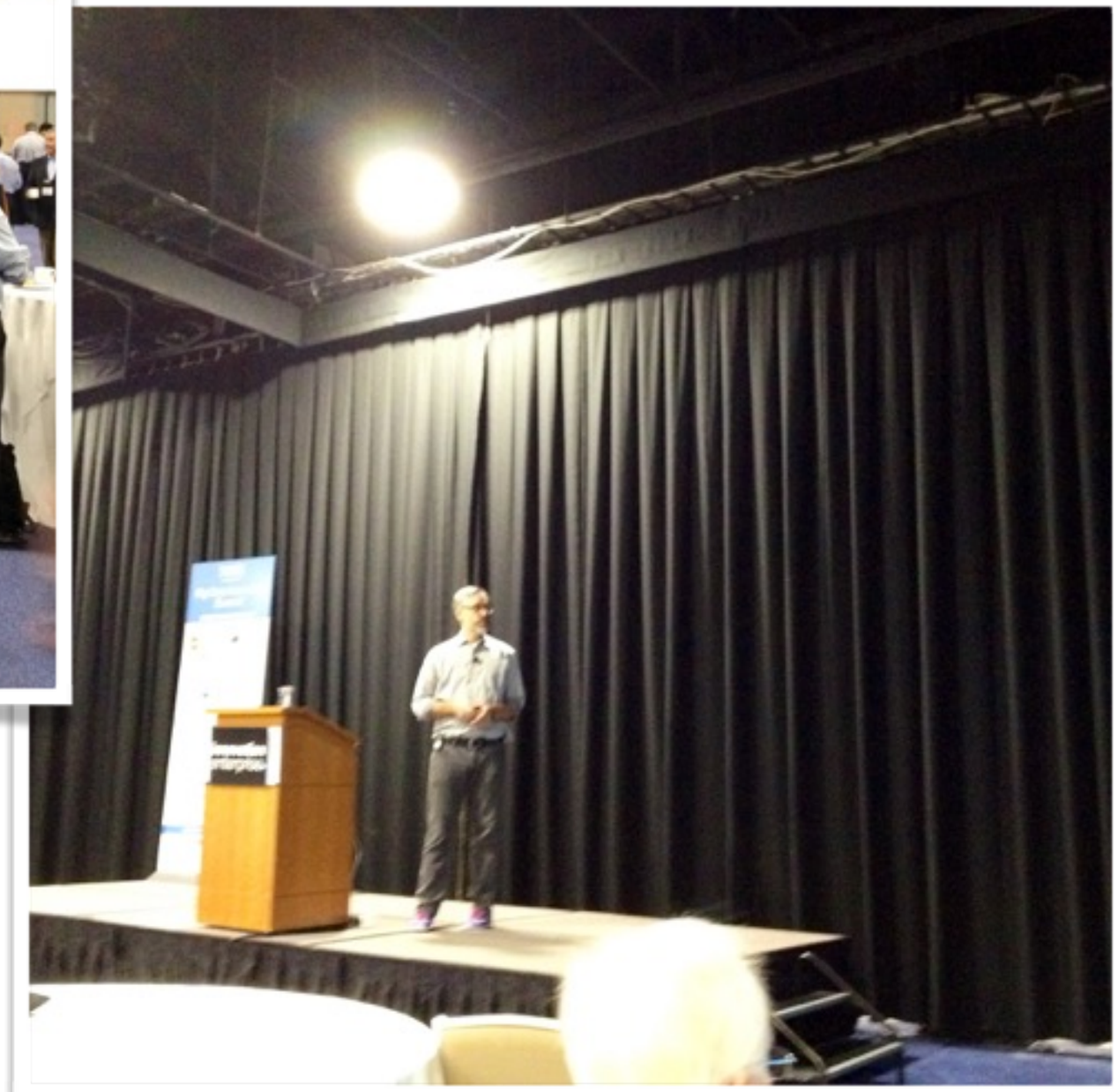
Request an Invitation
Download Brochure

Speakers Include:

Schneider Electric	facebook	The Bronx Defenders	MIT	tripadvisor
Engineering Fellow **Schneider Electric**	Head of Analytics, Instagram **Facebook**	Chief Scientist **The Bronx Defenders**	Digital Capability Leader **MIT**	Principal Software Engineer **TripAdvisor**

More


RESTROOM



Data Science can be defined as the study of the generalizable extraction of knowledge from data [1].

1. V. Dhar, “Data science and prediction,” *Commun. ACM , 56* (12, 2013), pp. 64-73.

- **Data science:**
  - A multi-disciplinary approach
  - *Not* only restricted to big data
  - A data scientists makes discoveries while swimming in data

Why do we need a new term like **data science** when we have had statistics for centuries?

1. The raw material, the "**data**" part of data science, is increasingly **heterogeneous** and **unstructured**.

2. Traditional database methods are *not* suited for **knowledge discovery**.

   Unlike database querying, which asks "What data satisfies this pattern (query)?"

   discovery asks "What **interesting** and **robust patterns satisfy** this **data**?"

Open data is **data** that **anyone** can **access**, **use** or **share**. Simple as that.

open
data
institute

My Way Towards Research on Big Data & Data Science

2014 **2015** **2016** **2017**

Understanding Data

Software (IJSC, SERP 2014)
Geoscience (ICAI 2015)
Smart Cities (ICAI 2015)

Full references are available on
www.harveyalferez.com

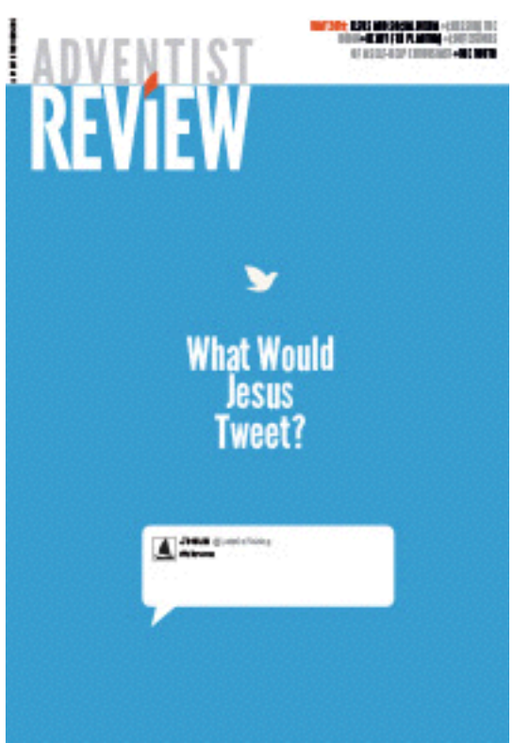
My Way Towards Research on Big Data & Data Science

2014 **2015** **2016** **2017**

Is it possible to use big data & data science to understand the needs of people?

My Way Towards Research on Big Data & Data Science

2014 **2015** **2016** **2017**



Big Data for Reaching a Big World

Andrews Research Conference

My Way Towards Research on Big Data & Data Science

2014 **2015** **2016** **2017**

Tweeting in New York City, Data Science Can Teach Us to Sympathize

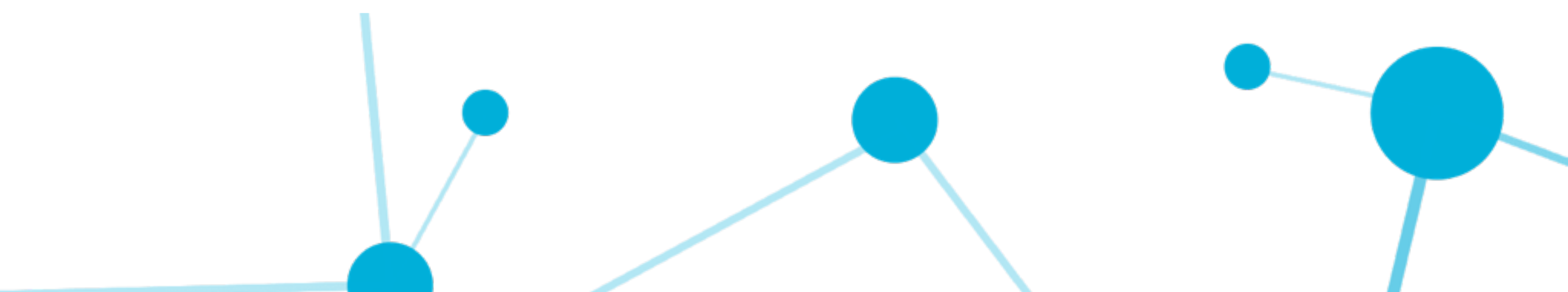
GAiN

My Way Towards Research on Big Data & Data Science

2014 **2015** **2016** **2017**

What are we doing now at Montemorelos?

Big Data for Reaching a Big World



Use **big data analysis** to try to understand how **culture** perceives our **fundamental beliefs.**

Big Data and Our Church

- In this study, the computational data analysis was based on **culturomics**.
  - The application of high-throughput data collection and analysis to the study of human culture [4].

Andrews Research Conference

4. Michel, J.-B. et al. (2011). Quantitative analysis of culture using millions of digitized books. Science, 331(6014):176–182.

Big Data and Our Church

- The full **data set** used in the experiments is available for download at:

  https://books.google.com/ngrams

  - This data set is composed of digitized texts containing about **4% of all books ever printed** between **1800** and **2008** (5,195,769 books).

  - **Books** in **English** (361 billion words) and in **Spanish** (45 billion words)

Big Data and Our Church

- The corpus **cannot be read by a human** [4]:
  - If you try to read only English-language entries from the year 2000 alone, at the reasonable pace of 200 words/min, without interruptions for food or sleep, it would take **80 years**.

4. Michel, J.-B. et al. (2011). Quantitative analysis of culture using millions of digitized books. Science, 331(6014):176–182.

Big Data and Our Church

- The **Google Ngram Viewer** was used to visualize the results.
  - A **1-gram** is a string of characters uninterrupted by a space. This includes words (“car”, “MICHIGAN”) but also numbers (“3.14”) and typos (“excesss”).
  - An **n-gram** is a sequence of 1-grams, such as the phrases “stock market” (a 2-gram) and “the United States of America” (a 5-gram) [4].

4. Michel, J.-B. et al. (2011). Quantitative analysis of culture using millions of digitized books. Science, 331(6014):176–182.

Church, Religion and Bible

N-gram Frequency (Corpus of English Books)

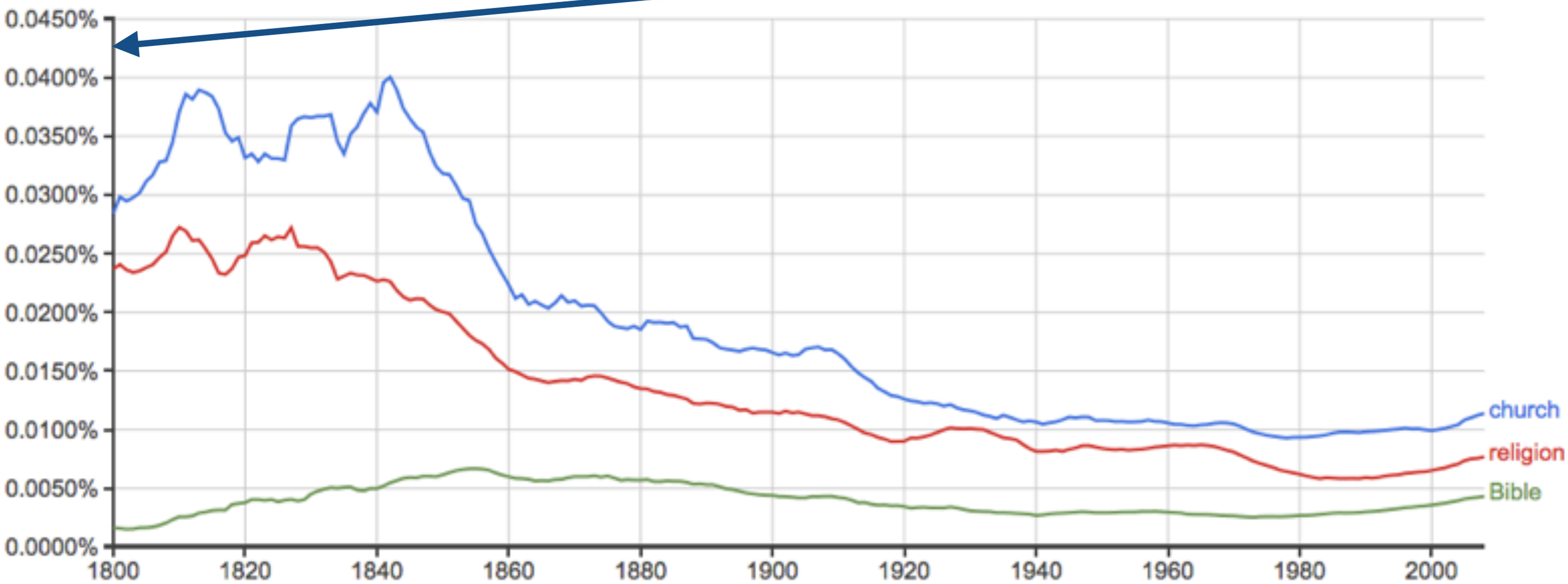


Figure 1. Church, Religion, and Bible Trends

Years

Secularization

Figure 2. Secularization Trend

God

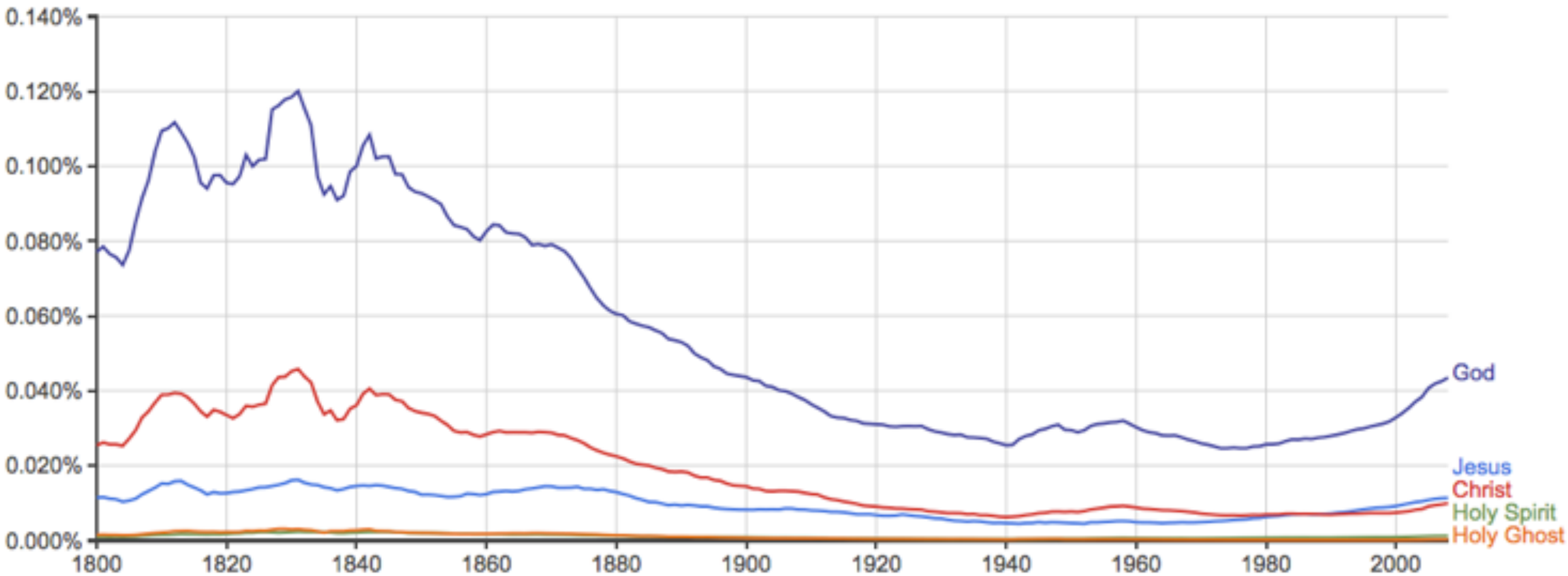


Figure 3. God-Related Trend

Creation

Figure 4. Creationism vs. Theory of Evolution (Case Insensitive Search)

The Sabbath

Figure 5. Sabbath Trend

Nature of Man

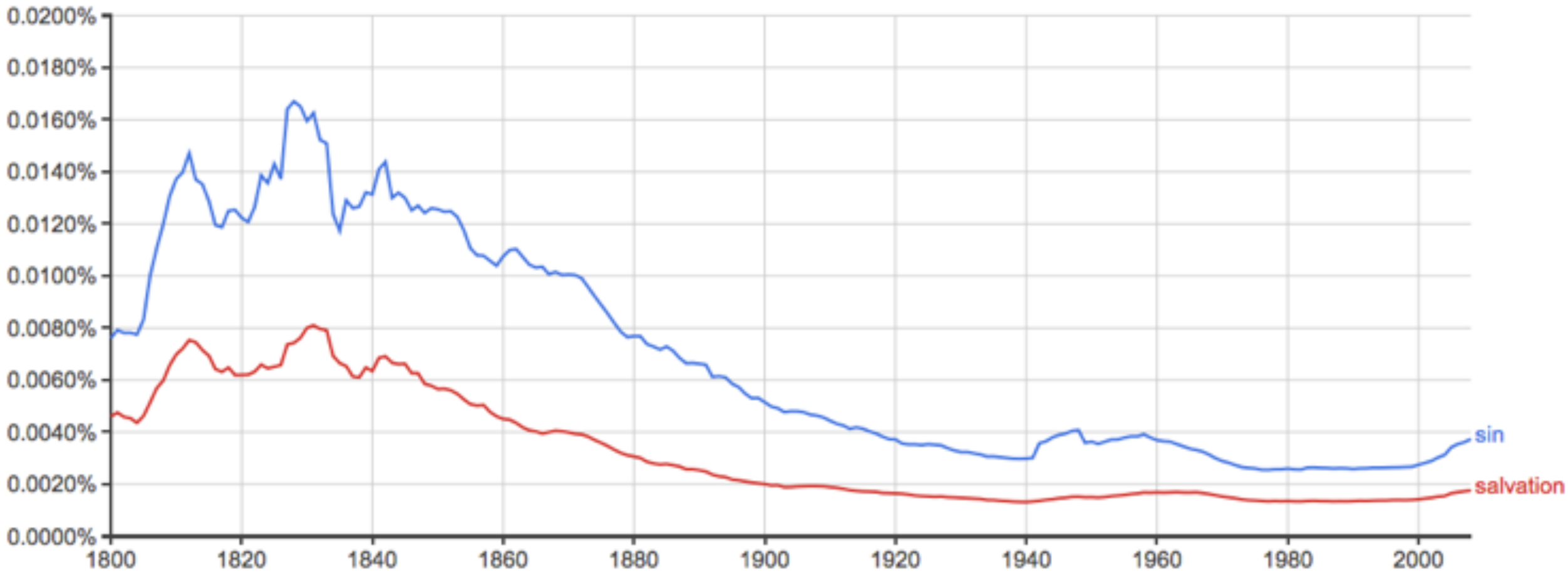


Figure 6. Sin and Salvation Trends

The Law of God and Justification by Faith

Figure 7. Ten Commandments and Justification by Faith Trends
(Case Insensitive Search)

The Second Coming of Christ

Figure 8. Second Coming of Christ, English Vs. Spanish Trends (Case Insensitive Search)

Healthy Living

Figure 9. Increasing Interest in Healthy Living and Vegetarianism

The Role of Big Data in Our Church

- Our **Church** can do something **valuable** with **big data**.
  - For instance, big data can help us to **make our beliefs relevant in a postmodern culture**.
- **Computational approaches** can be used to understand large pools of data, discover patterns, and make **“data-driven” decisions**.

The Role of Big Data in the Church

"The analysis of **big data** is not only a matter of solving computational problems… For the analysis of big data to truly yield answers to society's [Church's] biggest problems, **we must recognize that it is as much about social science as it is about computer science**" [5].

- Justin Grimmer, Stanford University

5. Grimmer, J. (2015). We Are All Social Scientists Now: How Big Data, Machine Learning, and Causal Inference Work Together. Political Science & Politics. 48(01): 80-83.

The Role of Big Data in the Church

- The applications of big data in the social sciences have not been well documented: **we know very little of how big data is actually being used in the social sciences** [6].

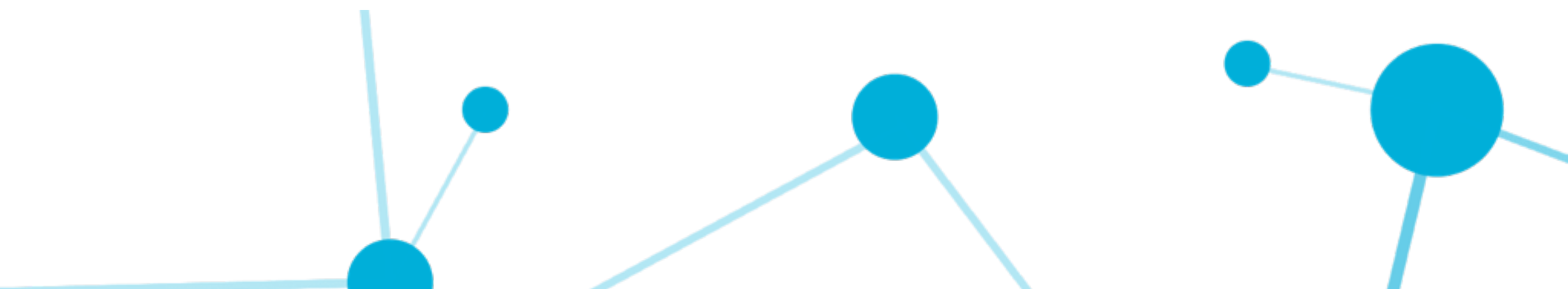
- **Emerging Field:** Computational Social Science.

  - Use large-scale demographic, behavioral and network data to investigate human activity and relationships [7].

6. University of Oxford. (2014). Accessing and Using Big Data to Advance Social Science Knowledge. URL: http://www.oii.ox.ac.uk/research/projects/?id=98

7. Microsoft Research (n.a.). Computational Social Science. URL: http://research.microsoft.com/en-us/projects/css/

Understanding the Needs of People in Big Cities through Data Science



Cities are growing fast

- **66%** of the world's population will live in urban areas by **2050** [1].

- There are more than **500** cities with a population of **1 million or more people**. However, these cities have an average of **1 Adventist congregation** for every **89,000 people**! [2].

1. Department of Economic and Social Affairs, United Nations, "World's Population Increasingly Urban with More than Half Living in Urban Areas," *United Nations* (July 10, 2014) https://www.un.org/development/desa/en/news/population/world-urbanization-prospects.html; retrieved November 10, 2015.

2. A. Oliver, "Adventist Church Implements Assessment Plan for Urban Mission," *Adventist News Network* (October 25, 2013) http://news.adventist.org/en/all-news/news/go/2013-10-25/adventist-church-implements-assessment-plan-for-urban-mission/; retrieved November 11, 2015.

“**The work in the cities is the essential work for this time.** When the cities are worked as God would have them, the result will be the setting in operation of **a mighty movement such as we have not yet witnessed**” [1].

1. E. G. White, Medical Ministry (Pacific Press Pub, 1963), p. 304.

Use **data science** to understand the **needs** of people in **New York City**.

Which *data* to use to understand the needs of people in big cities?

Twitter is the largest searchable archive of human thought, that's public, that's ever existed [1].

1. T. Simonite, "Twitter Boasts of What It Can Do with Your Data," *MIT Technology Review* (October 21, 2015) http://www.technologyreview.com/news/542711/twitter-boasts-of-what-it-can-do-with-your-data/; retrieved November 10, 2015.

Reaching People's Tweets

Sentiment analysis was used to discover the **needs** of people from tweets.

The **computational study** of **opinions**, **sentiments**, and **emotions** expressed in text [1].

Sentiment analysis has been **satisfactory** used to classify users' sentiments in tweets [2].

1. B. Ling, "Sentiment Analysis and Subjectivity," in N. Indurkhya, & F. J. Damerau, *Handbook of Natural Language Processing*, 2nd ed., (Boca Raton, Fl: Chapman & Hall, 2010), pp. 627-665.

2. A. Tumasjan, T. O. Sprenger, & P. G., Sa. "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment," *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media.* AAAI, (2010), pp. 178-185.

Reaching People's Tweets

- **Tweets** are **classified**
  - as ***positive*** when they communicate a positive sentiment, such as happiness;
  - as ***negative*** when a negative sentiment is attached to them (e.g. sadness);
  - and as ***neutral*** when no emotions are implied.

Reaching People's Tweets

Machine learning [1] was used as a tool to differentiate tweets with *positive*, *negative*, and *neutral* sentiments.

Machine learning explores the study and construction of **algorithms** that can **learn from** and **make predictions on data**.

1. A. Go, R. Bhayani, & L. Huang, *Twitter Sentiment Classification using Distant Supervision* (Stanford University, 2009)

Listening Closely to the Birds

Over a period of six weeks (September 22 to November 3, 2015), we collected 2,084 tweets from New York City, 1,633 of them bearing positive sentiments and 451 expressing negative sentiments. Tweets with neutral sentiments were not collected.

Listening Closely to the Birds

30 specified keywords:

Adventist, addiction, Bible, children, Christ, church, contamination, divorce, education, elderly, exercise, family, God, health, Jesus, obesity, peace, poverty, religion, rest, safety, salvation, Savior, stress, teenagers, teens, terrorism, vegetarian, violence, youth

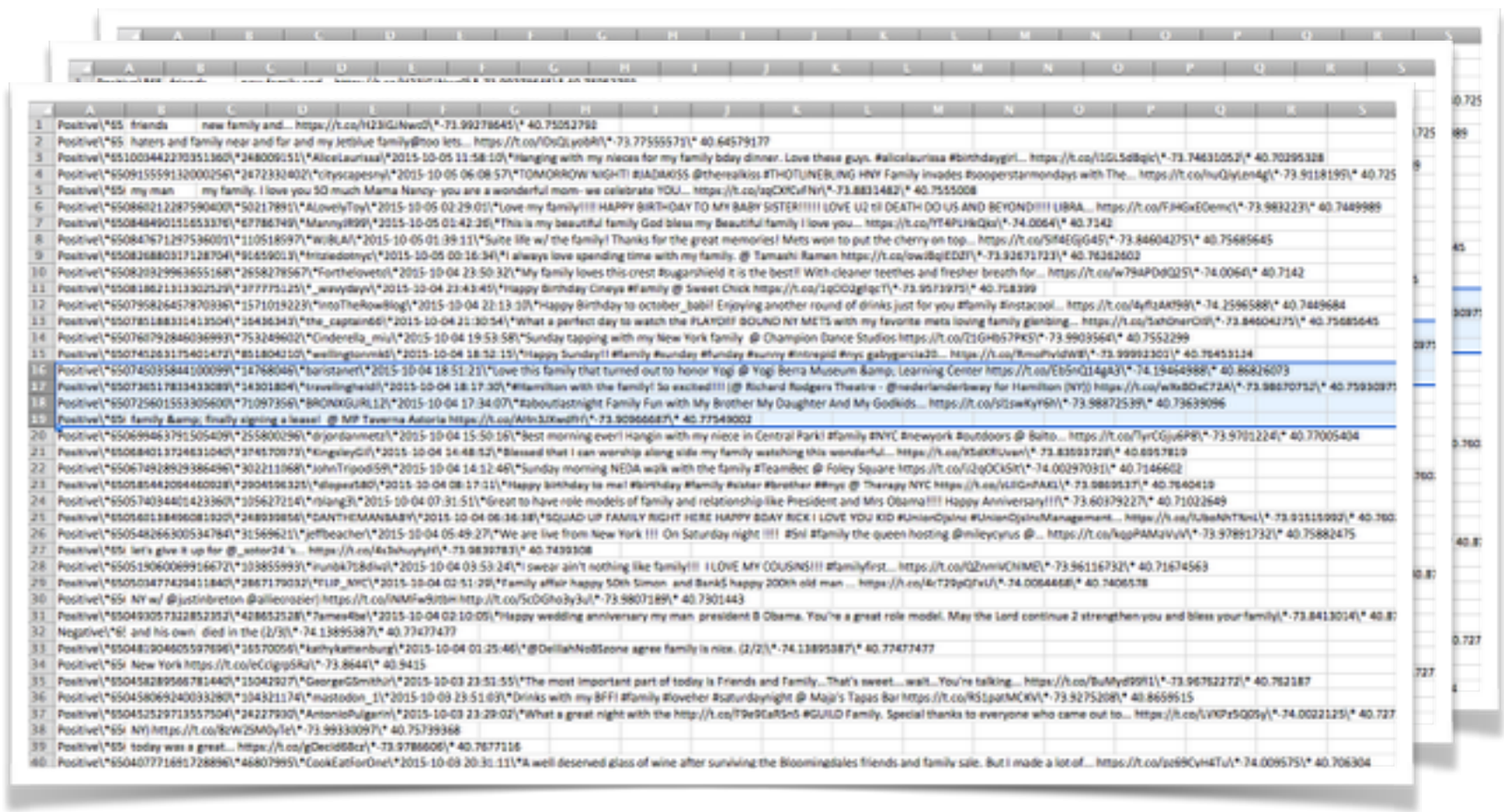
1. Collects

python

2. Classifies

+ and -

3. Stores



Positive Tweet about Vegetarian Food

- Positive
- her*
- 2015/10/02 02:08:16
- I want to be vegetarian. I really do. @arrogantswine @ East Williamsburg Brooklyn https://t.co/rpatPGyhXw\
- -73.939 (longitude)
- 40.714 (latitude)

Negative Tweet about Family

- Negative
- And*
- 11/10/15 18:48
- My ex has made them hate me, but I still see the children in my dreams.
- -73.74663446 (longitude)
- 40.69729011 (latitude)

Listening Closely to the Birds

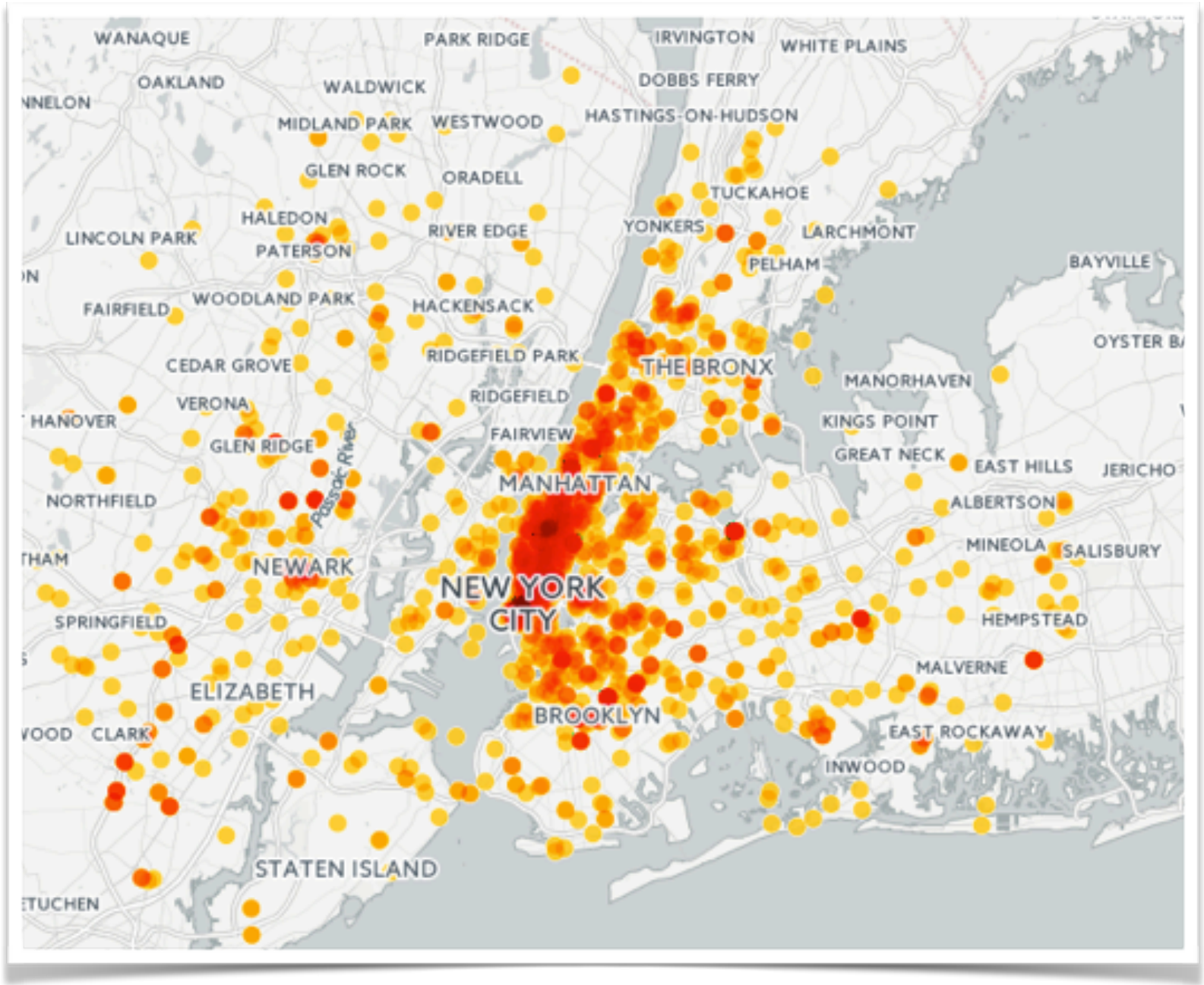


Figure 10. Intensity of tweets in New York City

Listening Closely to the Birds

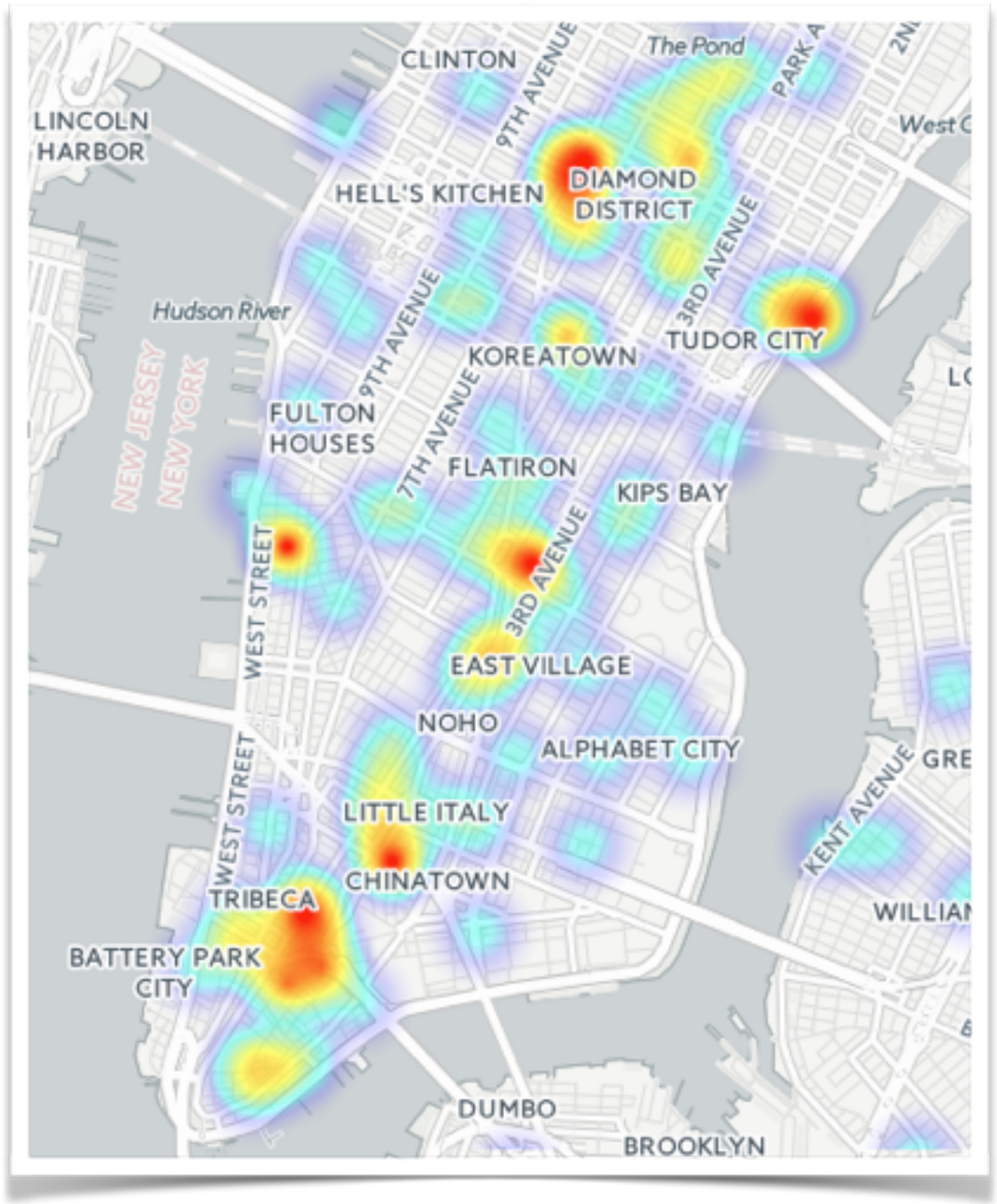


Figure 11. Areas with negative tweets in Manhattan

Upbeat and Downbeat

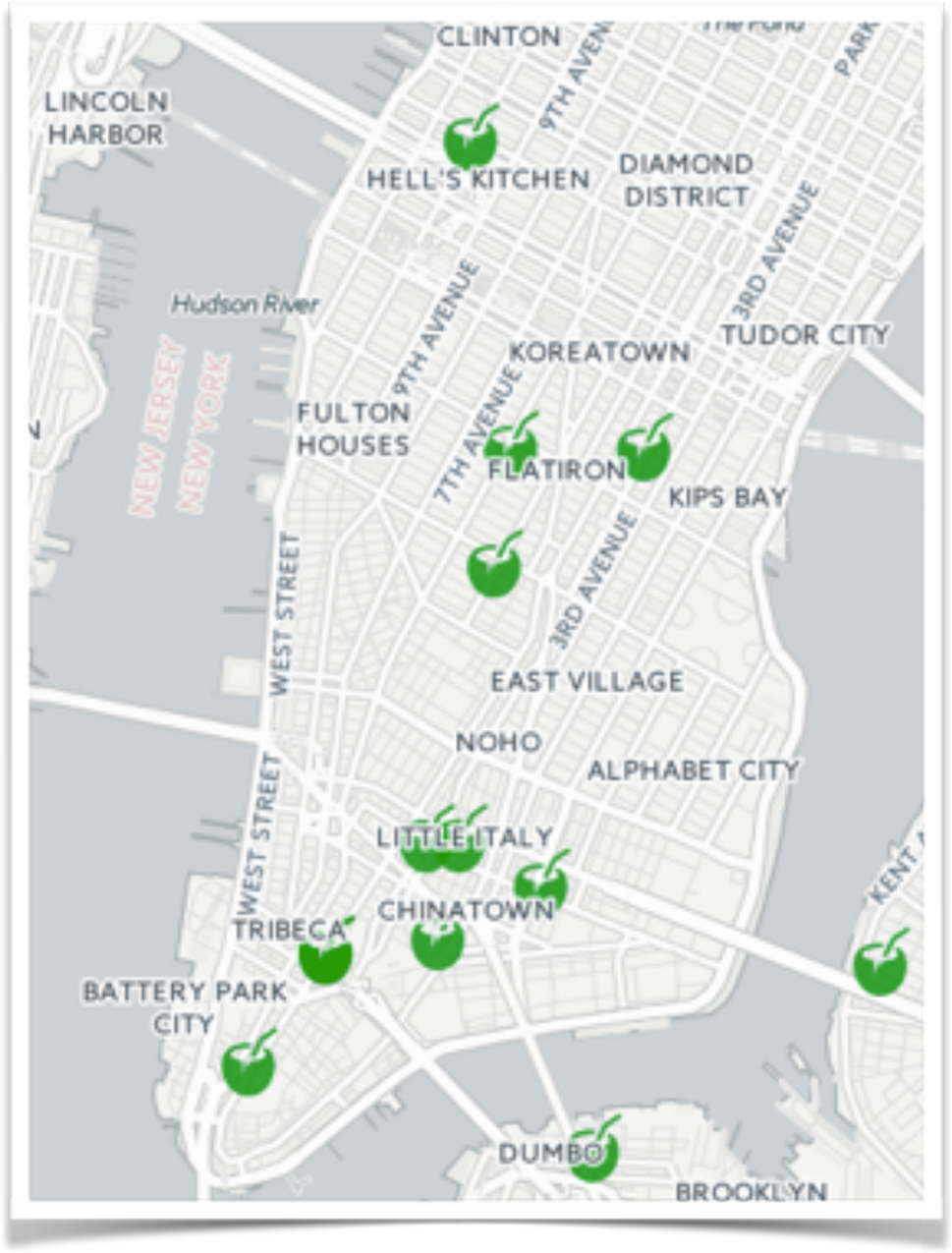


Figure 12. Positive tweets about vegetarian food in Manhattan

Upbeat and Downbeat

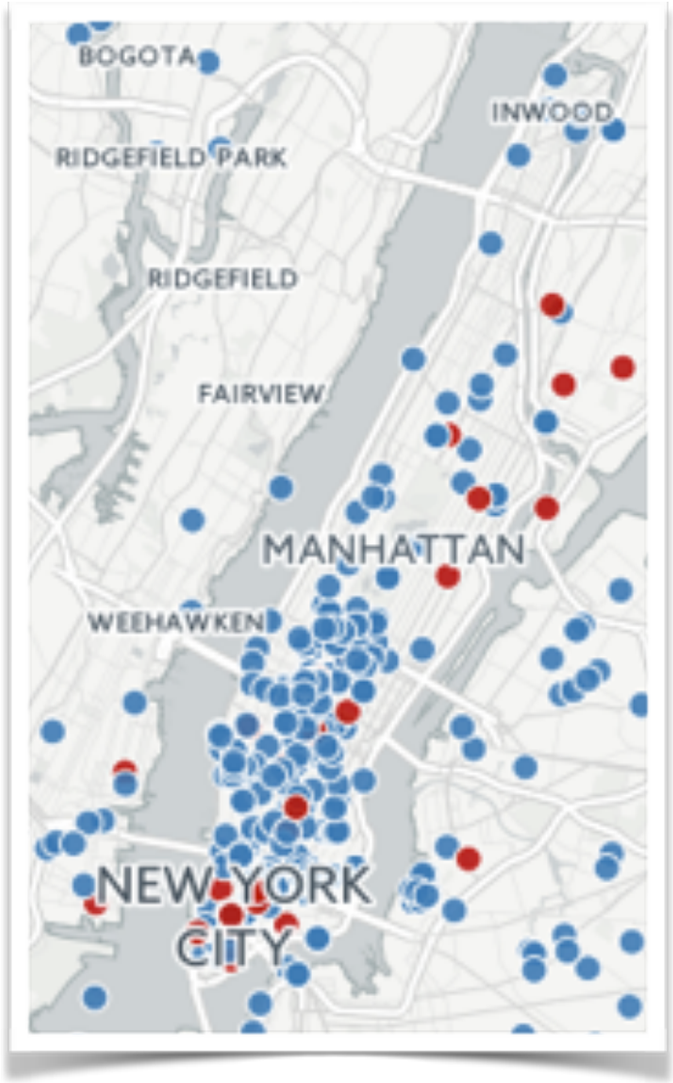
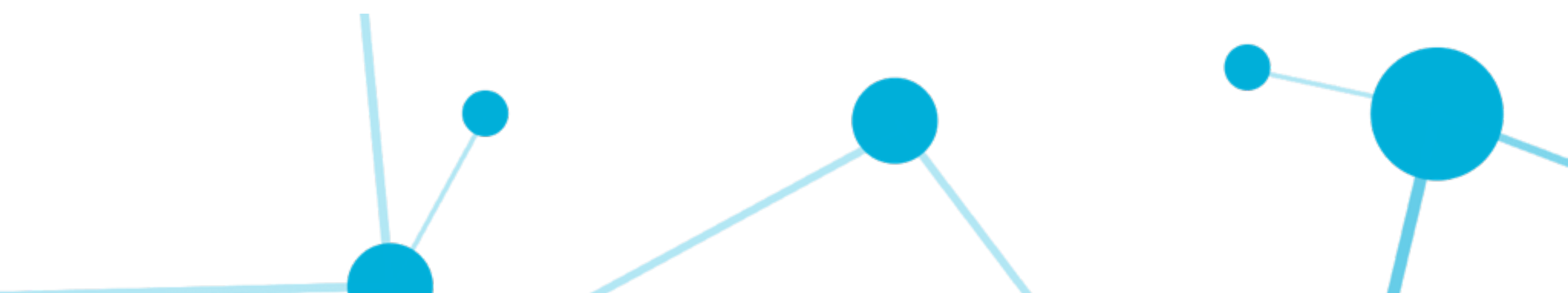


Figure 13. Positive [blue] and negative [red] tweets about family in Manhattan.

Data science has the potential to help us understand the **needs of people in big cities** in an **unprecedented way**.

What are we doing Now at Montemorelos?



Software to Guide the Application of Data Science in the 10/40 Window. Case Study: Data Analysis in the Middle East and North Africa Union

Merari González (MCC)

Using Data Science to Understand Segments of Individuals Who Have been Removed from Membership in the Inter-Oceanic Mexican Union Conference from 2005 to 2013

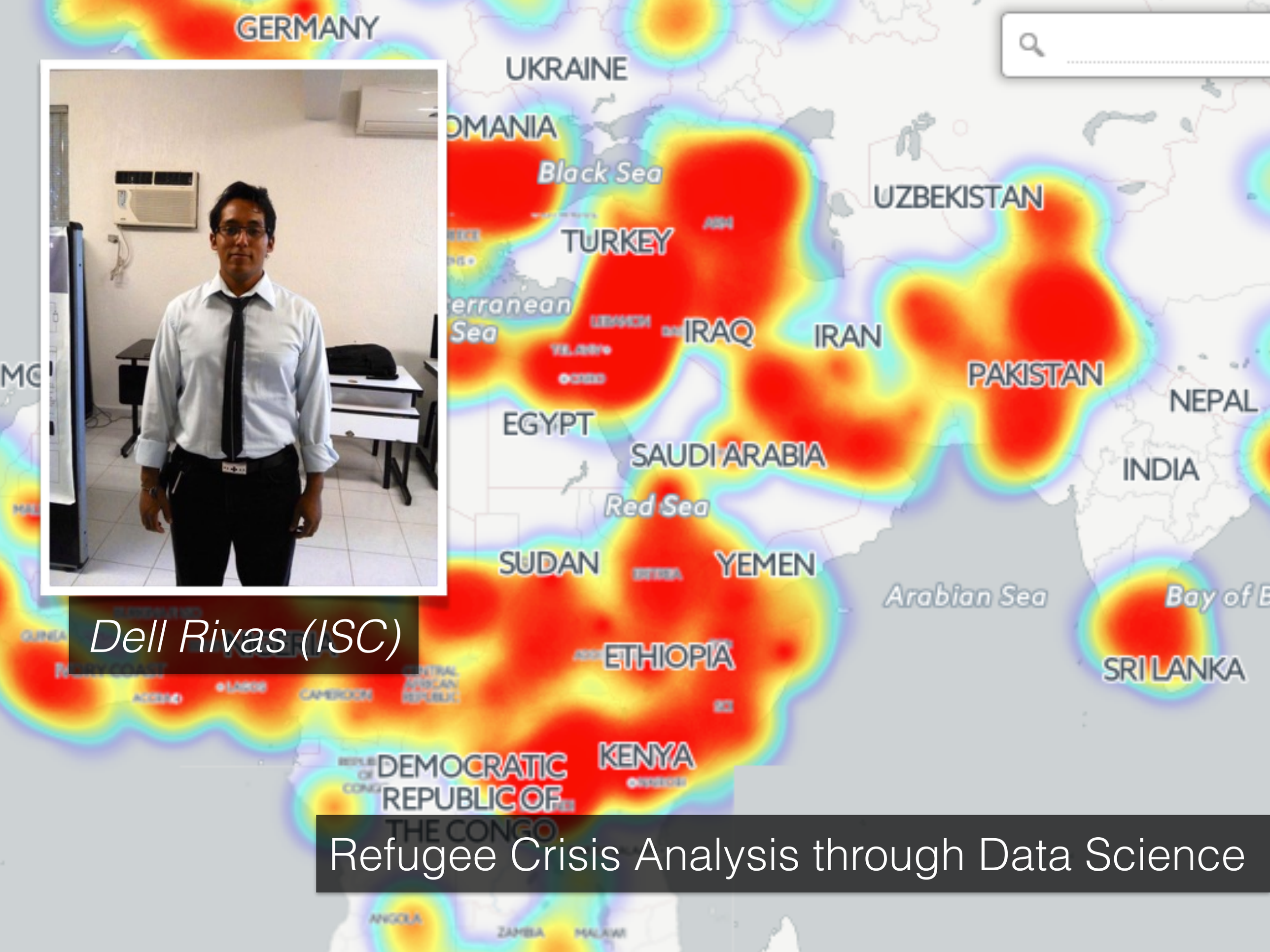
Dr. Germán H. Alférez, *Universidad de Montemorelos,* Erón Zebadúa, *Inter-Oceanic Mexican Union Conference,* and Enoc Cruz, *Universidad Linda Vista*

Technical Report June 23, 2016. Global Software Lab, School of Engineering and Technology, Universidad de Montemorelos

Abstract—Removing individuals from membership in the Seventh-day Adventist Church is the ultimate discipline that the church can administer. Our contribution is to present how we have applied state-of-the-art data science techniques to identify the segments of individuals who have been baptized from 2005 to 2013 and also been removed from membership in the same period of time at the Inter-Oceanic Mexican Union Conference. The dataset that was analyzed is composed of 14,388 records of members who have been removed. The results can guide further church decisions to prevent membership lost, specially among youth and among people who are baptized after evangelistic campaigns. Our data-science approach could be easily extrapolated to other divisions and conferences.

Erón Zebadúa (MCC)

Enoc Cruz (MCC)



Dell Rivas (ISC)

Refugee Crisis Analysis through Data Science

Alférez, G.H., Jiménez, J., Hernández-Navarro, H., González, M., Domínguez, R., Briones, A., Hernández-Villalvazo, H. (2016). **Application of Data Science to Discover the Relationship between Dental Caries and Diabetes in Dental Records.** Proceedings of the 2016 International Conference on Health Informatics and Medical Systems (HIMS 2016), Las Vegas, NV, USA.

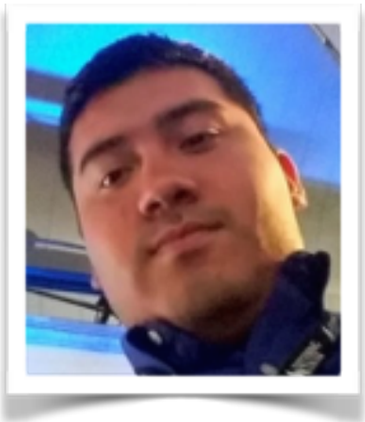
Marín, C., Alférez, G.H. Córdova, J., & González, V. (2015). **Detection of Melanoma Through Image Recognition and Artificial Neural Networks.** Proceedings of the 2015 IUPESM World Congress on Medical Physics and Biomedical Engineering (WC 2015), Toronto, Canada.

Automatic Detection of Retinopathy in Mexican Patients through Deep Learning

Anthony McCulloch (ISC)

Glaucoma Detection through Deep Learning

Marco Espinoza (ISC)



Software to Classify Patients through Machine Learning and Open Maternal Mortality Data in Mexico

Rusbel Domínguez (MCC)

SEVENTH-DAY ADVENTIST® CHURCH
NORTH AMERICAN DIVISION

Big Data & Data Science Serving the Mission

Harvey Alférez, Ph.D.
Global Software Lab,
School of Engineering and Technology,
Universidad de Montemorelos, Mexico
www.harveyalferez.com