Lecture on the Discussion Document:

A New Framework for Statistical Thinking in Times of Big Data and Digital Economy

Dr. Orlando GONZÁLEZ
Graduate School of Human Sciences
Assumption University, Thailand

Dr. Masami ISODA
CRICED / Faculty of Human Sciences
University of Tsukuba, Japan

Dr. Roberto ARAYA
Center for Advanced Research in Education (CIAE)
The University of Chile, Chile

Dr. Maitree INPRASITHA
IRDTP / Faculty of Education
Khon Kaen University, Thailand

APEC-Tsukuba International Conference XIII
University of Tsukuba, Tokyo Campus
Tokyo, Japan – February 7, 2019

Outline

This lecture will last 40 minutes, organized as follows:

- Understanding Big Data in the Digital Era
- Statistical Thinking as the Practice of Statistics
- A New Framework for Statistical Thinking
- Discussion of an Exemplar Application of the Framework: Aging Population Issues in APEC Countries
- Q & A Session

Understanding Big Data in the Digital Era

Big banks turn to big data to track clients

BIG PICTURE Thousands of data points such as personal calls, email and social media interaction combine to give a more complete picture

Understanding Big Data in the Digital Era

But… What is “Big Data”?

CISCO

The Zettabyte Era Officially Begins (How Much is That?)

Thomas Ba… Se… …its

For mor… …decade, Cisco has tracked and projecte… …al Int… traffic gr… …h … …te… …etw…king trends through t… …ual Networking Index (VNI). From day on… …has been … …rk … …t ur … …ysts have targeted a … …ajor networking milestone.

A Zetta…yte is a w…ole lot of traffic (to put it mildly). So, since we're counting, here's how … Ze…

- A zettabyte is a measure of storage capacity and is … …ed … …10^{21} (1,000,000,000,000,000,000,000 bytes) or 1 sextill… by…
- One Zetta…yte is approximately equal to a thousand Exabytes, a billion Terabytes, or a trillion Gigabytes.

WHAT DO YOU THINK?

Understanding Big Data in the Digital Era

- As many of you mentioned, data in this digital era usually meet the following five criteria:

Volume

Variety

Velocity

Veracity

Value

(Claverie-Berge, 2012; Storey & Song, 2017)

Understanding Big Data in the Digital Era

- **Volume**: This refers to data at rest. Datasets with sizes in the order of terabytes, petabytes and zettabytes.
- **Variety**: This refers to the many forms (e.g., text, images, videos, audio files, emails) and sources (e.g., spreadsheets, databases, social media and monitoring devices) of data.
- **Velocity**: This refers to data in motion: the speed at which data flows in from sources (e.g., real time streaming).
- **Veracity**: This refers to data in doubt. Depending on its origin, processing technologies, and collection methods, data can have biases and inaccuracies attached, which need to be identified and accounted for.
- **Value**: This refers to turning data into profit. Just having big data is of no use unless we can turn it into value.

Statistical Thinking as the Practice of Statistics

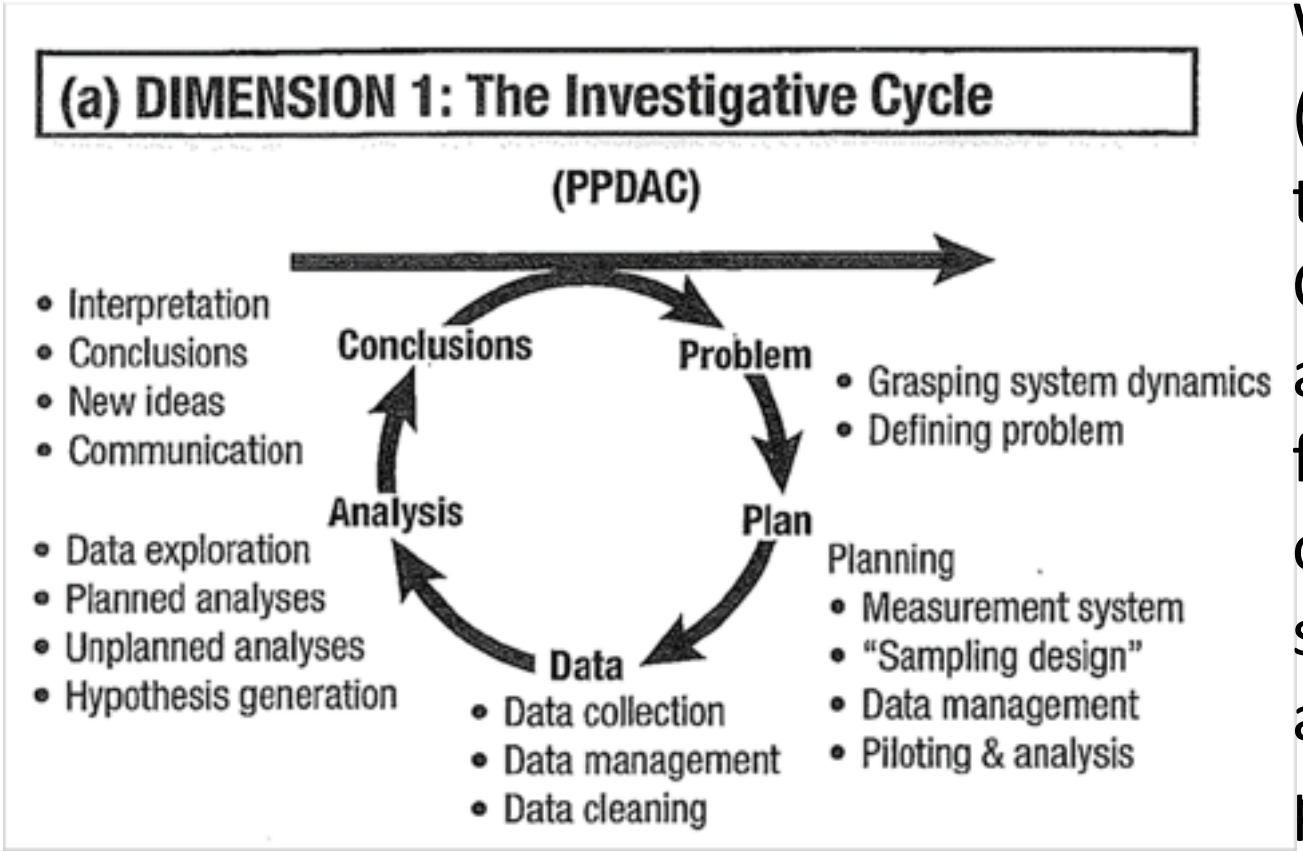
BASIC LITERACY	REASONING	THINKING
IDENTIFY DESCRIBE REPHRASE TRANSLATE INTERPRET READ	WHY? HOW? EXPLAIN (THE PROCESS)	APPLY CRITIQUE EVALUATE GENERALIZE (delMas, 2002)

- Many researchers (e.g., del Mas, 2002; Watson et al., 2018) consider statistical thinking as the practice of statistics through the enactment of the different thought processes involved in statistical problem solving and statistical investigations, in such a way that it leads to the most reliable, logical, and trustworthy conclusions.

Statistical Thinking as the Practice of Statistics

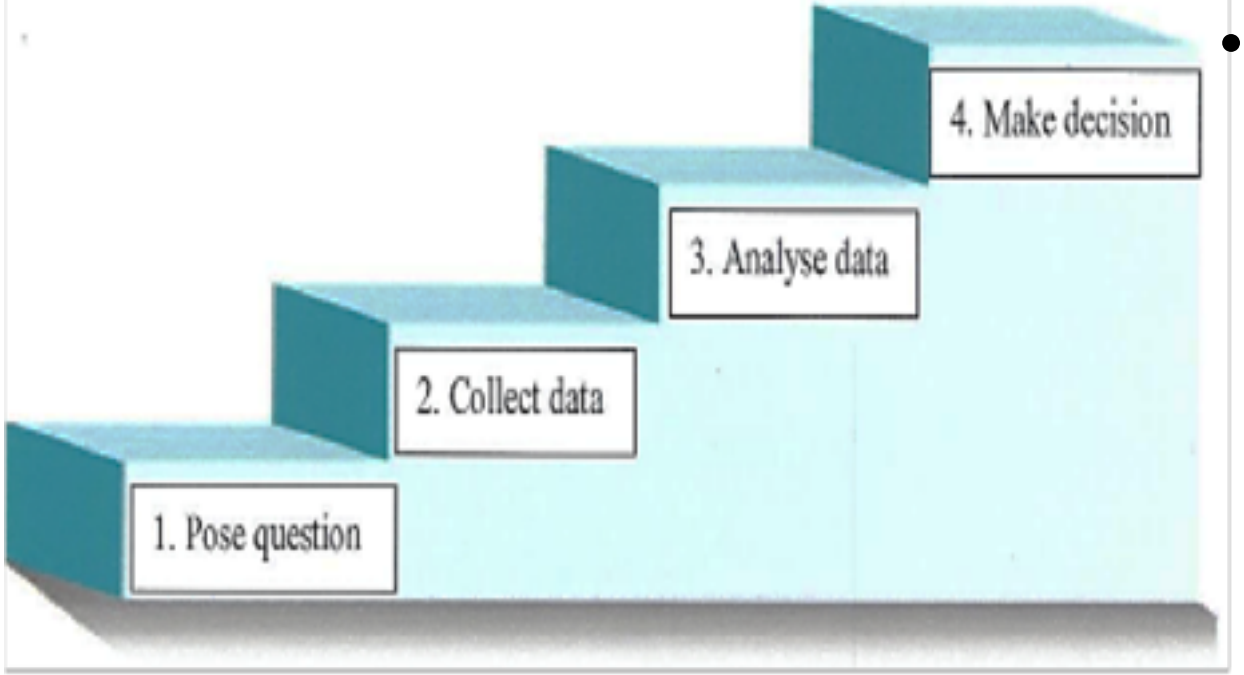
- The school curriculum should provide both
  - experience in the practice of statistics, and
  - opportunities to critically evaluate data from varied sources (e.g., other students, science reports, media extracts, big data platforms).
- But, how do we give students the opportunity to experience the "Practice of Statistics" with Big Data in the classroom?
- Before discussing this, let us revise some of the current curriculum frameworks of statistical thinking.

The Practice of Statistics: The PPDAC Model



Wild and Pfannkuch (1999) suggested the Investigative Cycle (better known as PPDAC Model) from their analysis of the work of students of statistics and practicing professional statisticians.

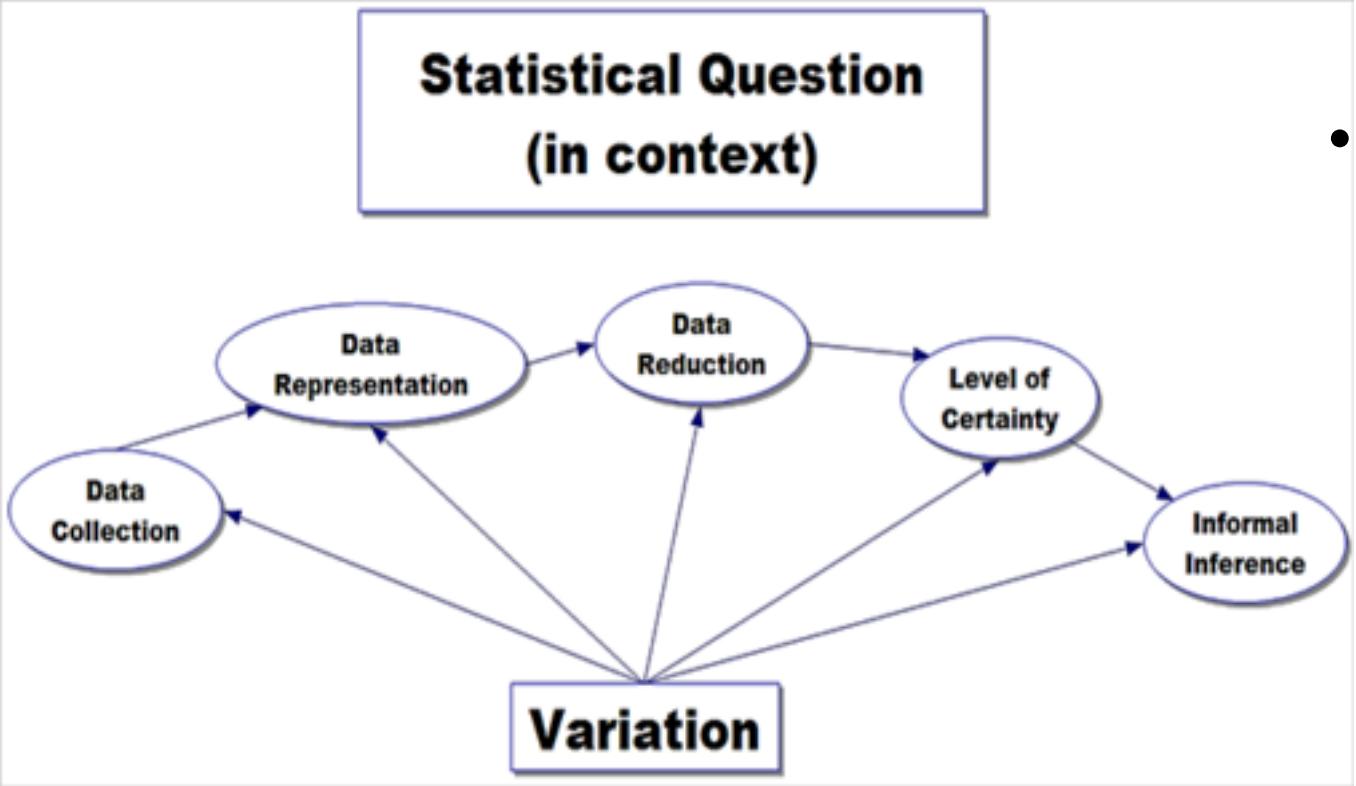
The Practice of Statistics: Franklin et al. (2007) Model



A typical example of the traditional question-then-answer method. From the start, we ask questions like "How many tourists visited Tokyo in the last decade?"

- In the GAISE report of the American Statistical Association, Franklin et al. (2007) suggested the practice of statistics through engaging in four-stage statistical investigations.

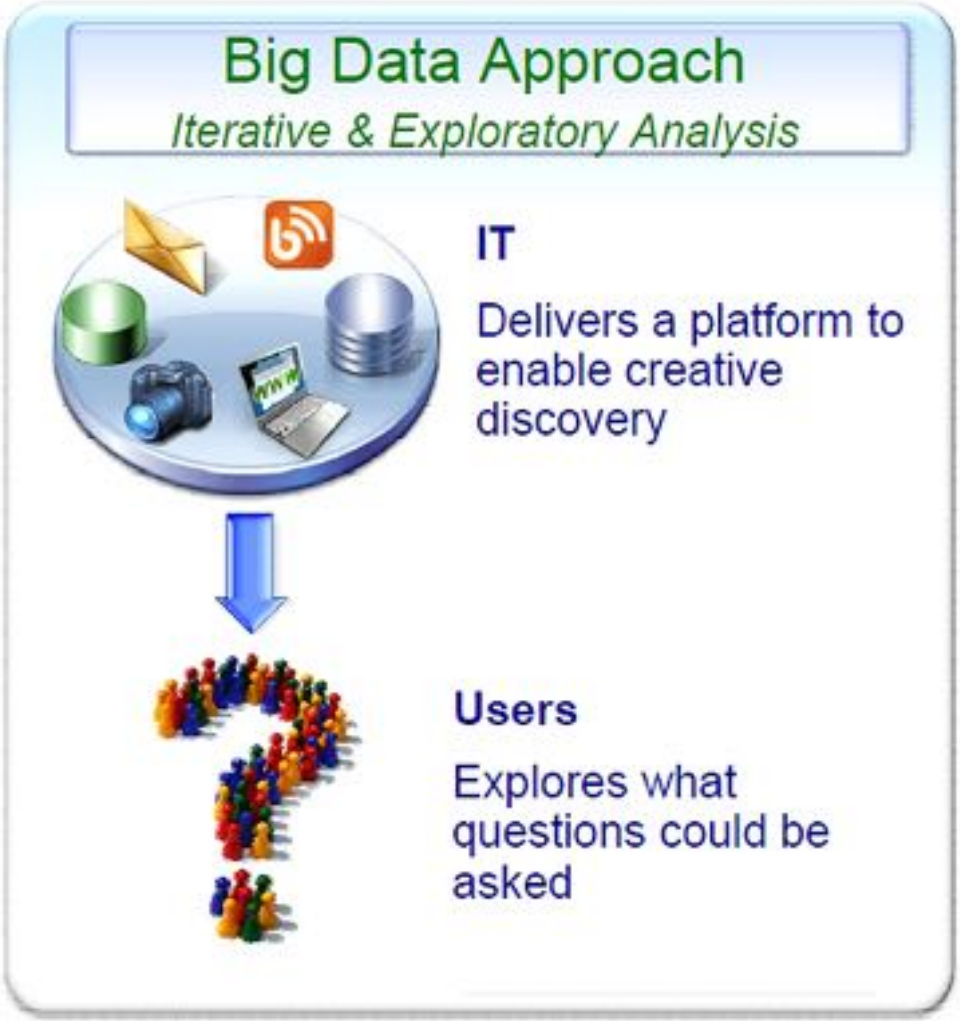
The Practice of Statistics: Watson (2009) Model



- Watson (2009) suggested a way to engage students in the practice of statistics through statistical investigations.

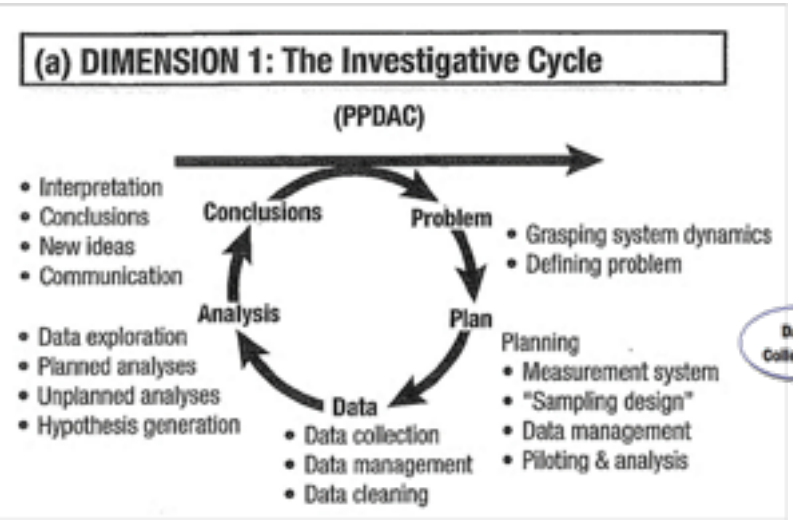
Another typical example of the traditional question-then-answer method.

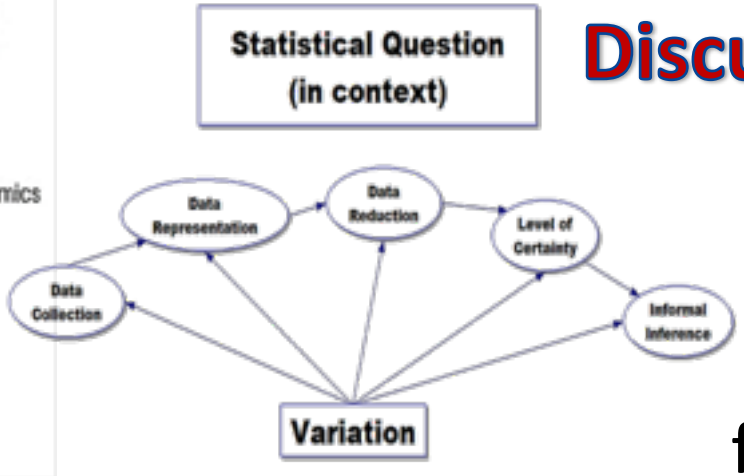
The Practice of Statistics: Big Data Approach (IBM, 2012)

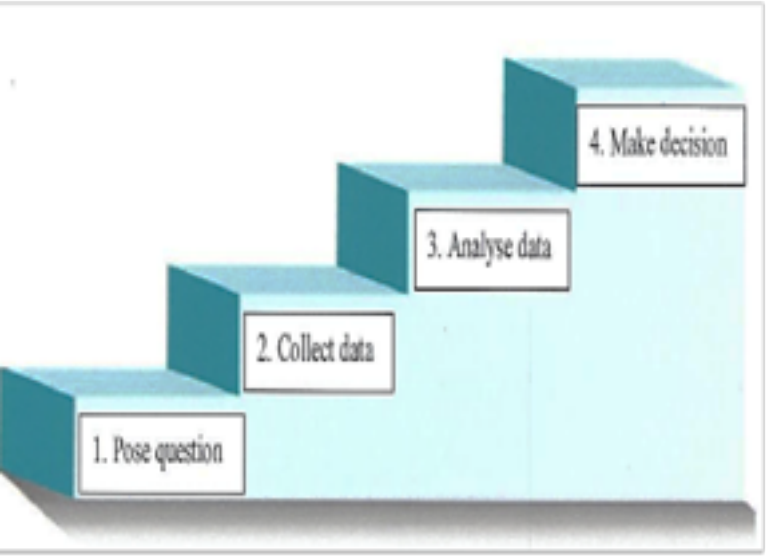


- In order to start the practice of statistics with big data, an IT platform that enables creative discovery is needed.
- Then, users look for connections and significant relationships within the available data.
- This will result in data-first answers, and then users will work backward to find the questions that should have been asked.

Statistical Thinking as the Practice of Statistics







Discussion time!

How can we revise current curriculum frameworks of statistical thinking to incorporate big data for the digital economy?

Let's hear your ideas!

A New Framework for Statistical Thinking

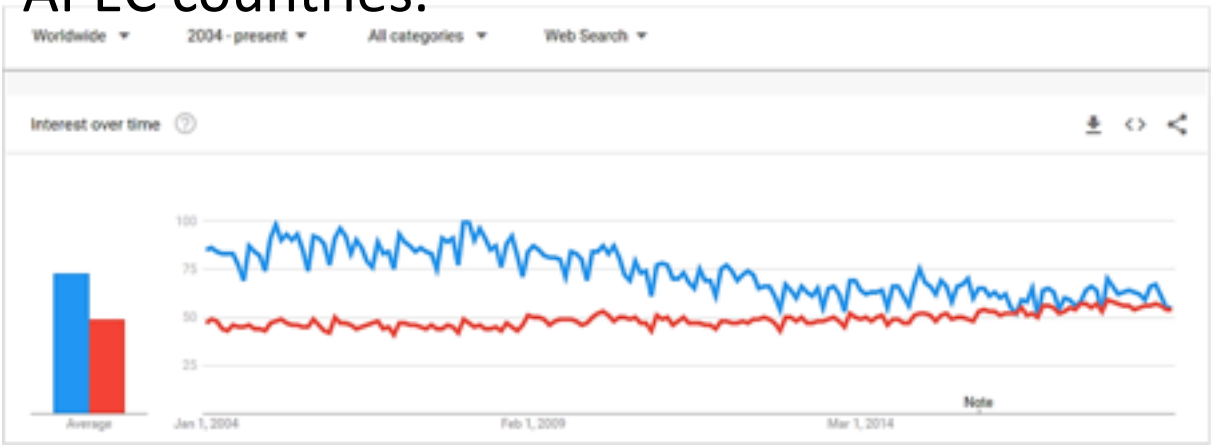
- We consider that any up-to-date framework for statistical thinking must be designed giving consideration to the criticisms to previous models, and to the way big data analytics is carried out.
- By doing so, we came up with a framework that understands statistical thinking as a cognitive process comprised of the following five phases:
  - *Patterns and relationships from data*.
  - *Questions*.
  - *Objectives*.
  - *Data mining*.
  - *Understanding and/or designing*.

Exemplar Application of the Framework: Aging Population Issues in APEC Countries

- Let us suppose that we are interested in exploring issues related to population ageing, focusing on APEC countries.
- To do so, we checked the worldwide trend of web searches for the terms “social security” and “nursing home”, using the website “Google Trends” (https://trends.google.com).

Patterns and Relationships

Using Google Trends platform, we look for patterns (trends) and relationships within the data, based on our particular interest: online search terms for “social security” and “nursing home” in APEC countries.



Worldwide search trends in the last 15 years

Exemplar Application of the Framework: Aging Population Issues in APEC Countries

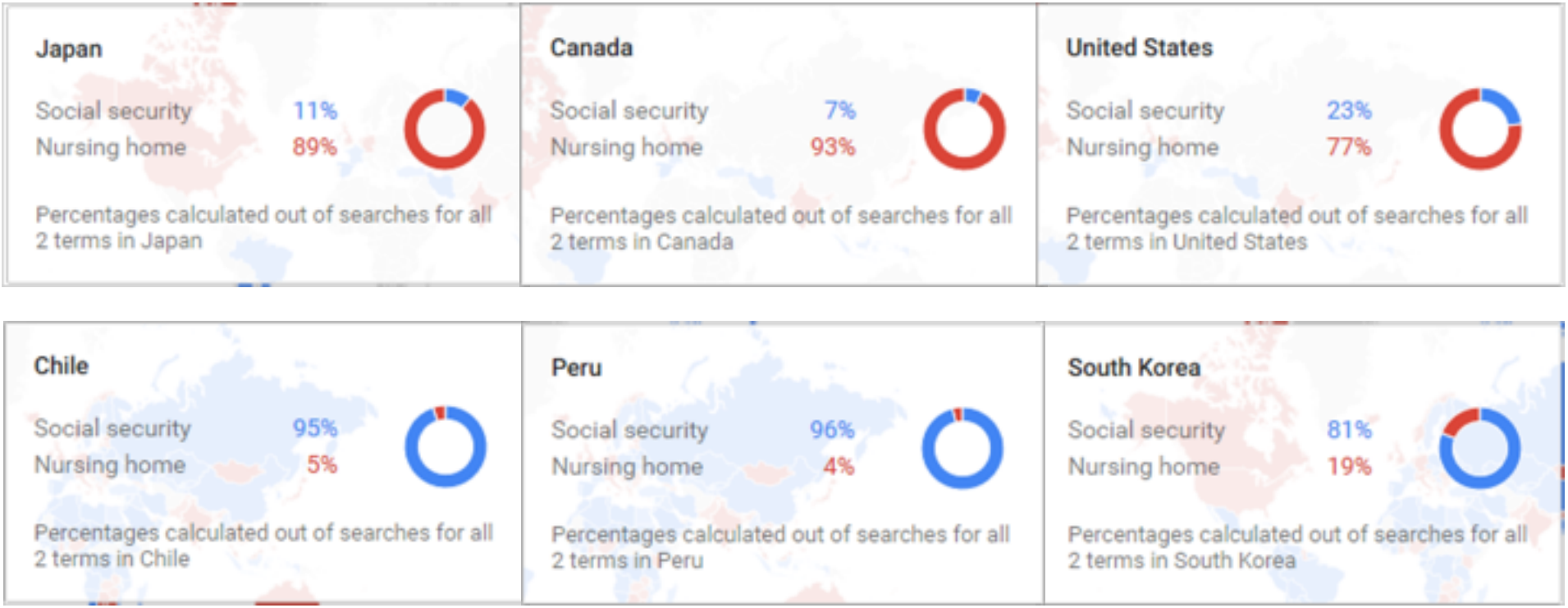


Figure 2: Percentage comparison of Google searches for the terms “social security” (in blue) and “nursing home” (in red) in six APEC countries in 2018.

Questions

From the obtained visual representations, we can pose questions about to find explanations to the patterns and relationships found.

What questions would you pose? Let's hear them!

Some questions that might be posed for this example are:

- Why do some countries (such as Japan, Canada and USA) seem to show considerably more interest on "nursing home" than on "social security"?
- Why do some countries (such as Chile, Peru and South Korea) seem to show considerably more interest on "social security" than on "nursing home"?
- What could be the behavior of these trends for individual APEC countries in the next decade?

Objectives

From the posed questions, we can set clear objectives to address.

Some objectives stemming from the questions we posed earlier are the following:

1. To look for and identify the reasons why some APEC countries seem to show considerably more or less interest on “nursing home” than on “social security”.
2. To determine the behavior for individual APEC countries regarding individual queries, such as the mentioned above.
3. To predict the trends of web searches for the terms “nursing home” and “social security” in APEC countries in the next decade.

Data mining

Addressing the previous objectives will lead to a re-examination of the data, from which new insights and knowledge discovery will emerge from three types of data mining:

- *Big data oriented data mining*.
- *Explanation-oriented data mining*.
- *Future-oriented data mining*.

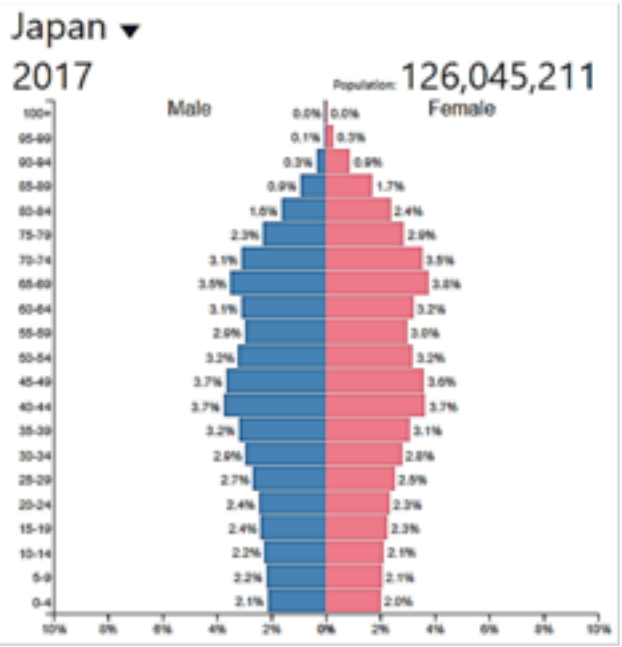
For the purpose of exemplifying this phase, let us address the Objective 1 (i.e., to look for and identify the reasons why some APEC countries seem to show considerably more or less interest on "nursing home" than on "social security").

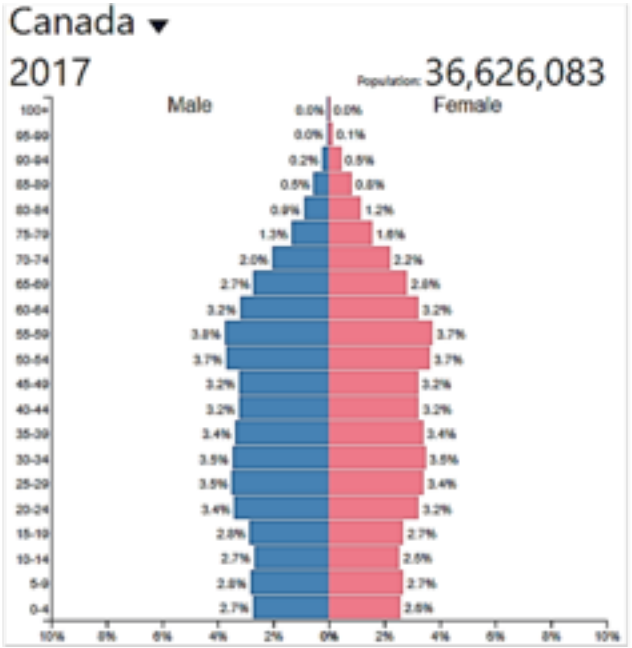
Data mining

- For APEC countries showing considerably more interest on searching for the term "nursing home", the main reason could be the current structure of their population pyramids.
- This is ***explanation-oriented data mining***: providing reasonable explanations for the new knowledge (patterns and relationships) extracted from big data.
- To do this, old and new sets of raw big data will be explored, looking for identifying significant correlations and new connections, in order to arrive at insights and extract new knowledge (patterns and relationships).
- This is ***big data-oriented data mining***.
- This is not the traditional data exploration (i.e., digging into data searching for specific metrics). Here, identifying new patterns and relationships is the goal.

Exemplar Application of the Framework: Aging Population Issues in APEC Countries

Data mining





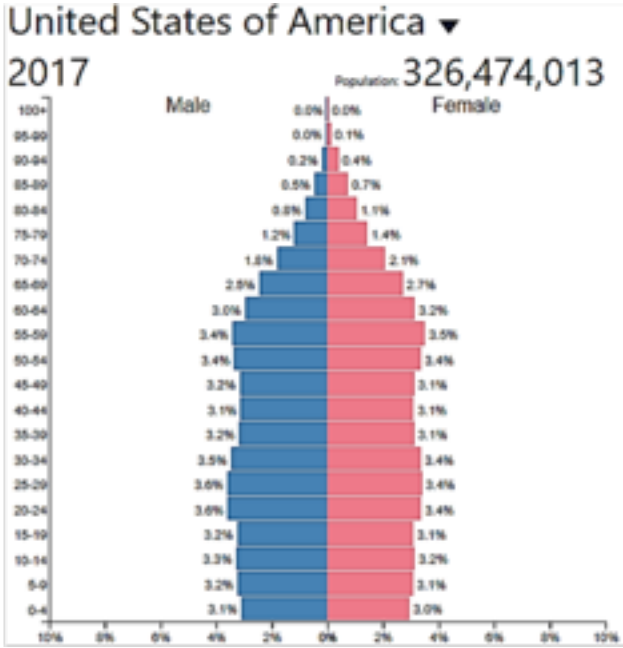


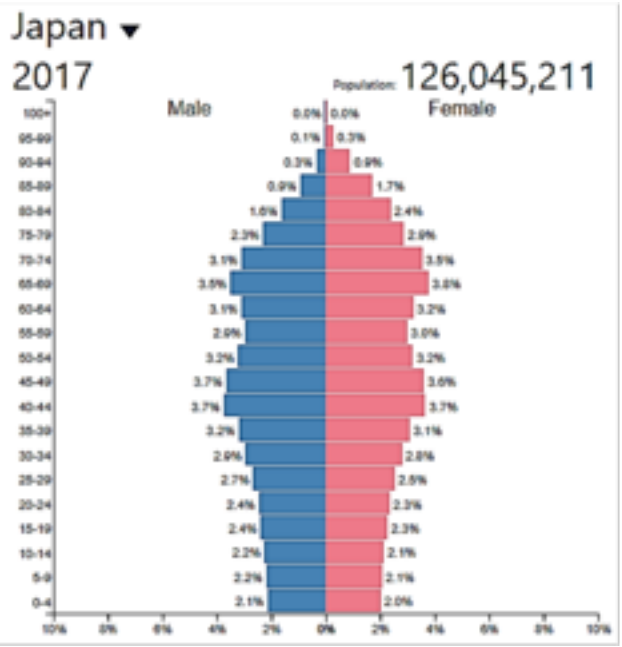
Figure 3: Population pyramids in 2017 for Japan, Canada and the USA.

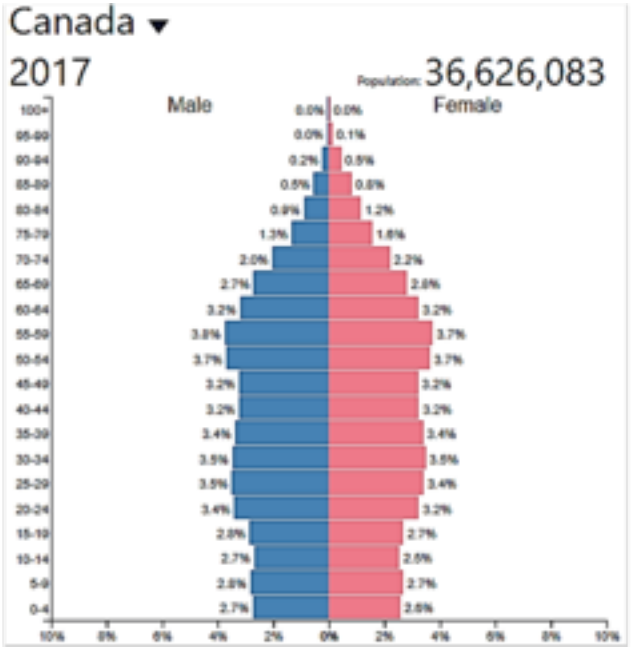
From the big data-oriented data-mining:

- In Japan, a large proportion of the population close to or over retirement age.
- In Canada, a large proportion of the population close to retirement age.
- In the US, a large proportion of the population is below 50 years-old.

Exemplar Application of the Framework: Aging Population Issues in APEC Countries

Data mining





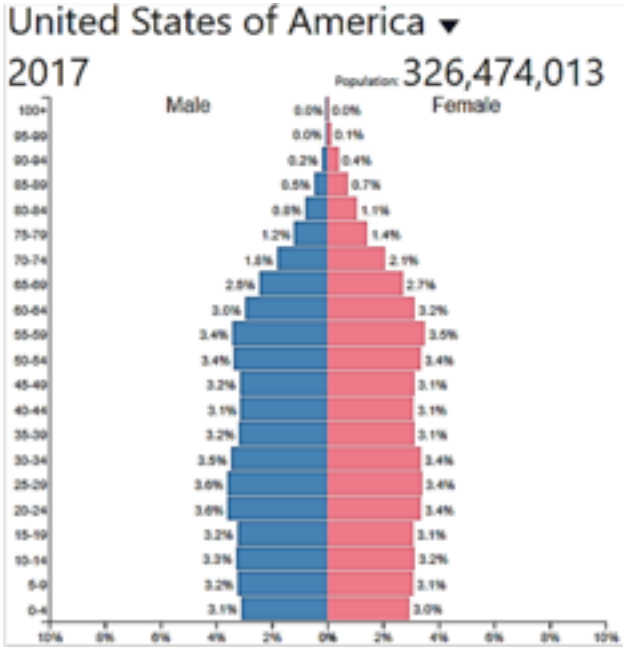


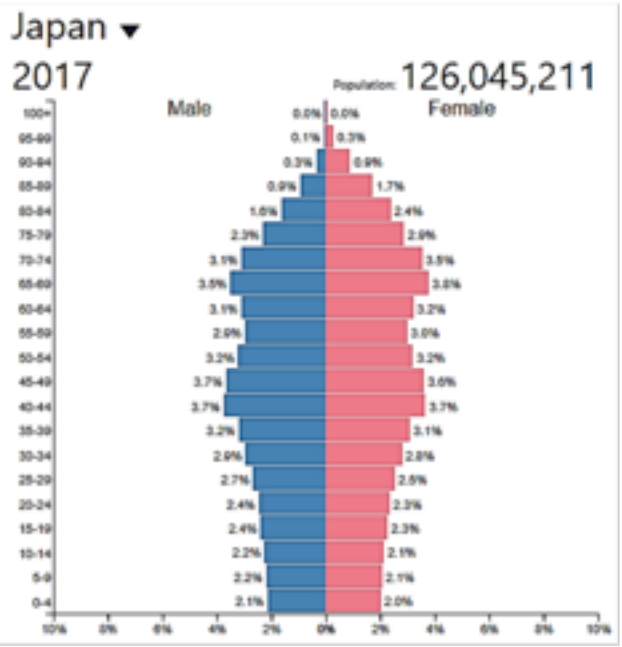
Figure 3: Population pyramids in 2017 for Japan, Canada and the USA.

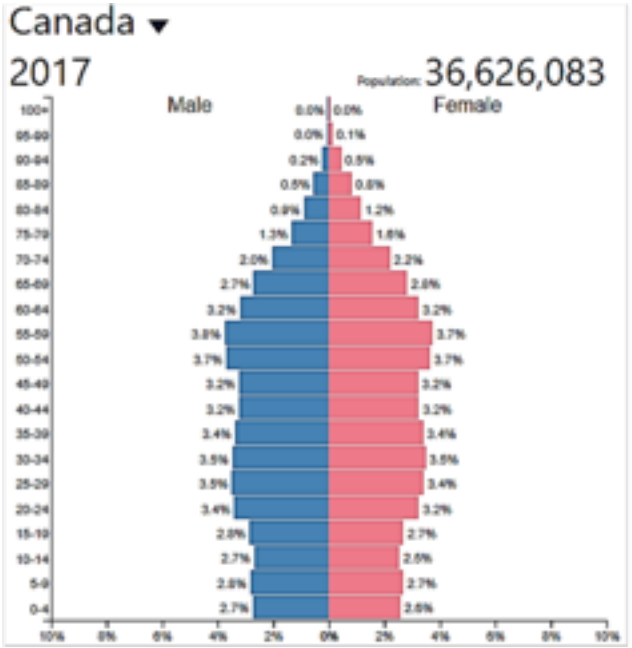
Now, let's hear your explanations why some APEC countries seem to show considerably more interest on, for example, "nursing home" (explanation-oriented data mining)?

Remember: Explanations might not be found in the data initially examined!

Exemplar Application of the Framework: Aging Population Issues in APEC Countries

Data mining





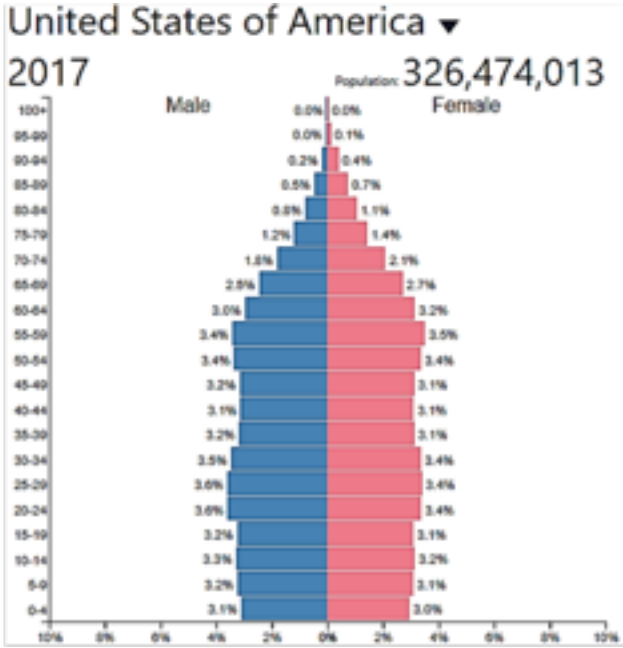
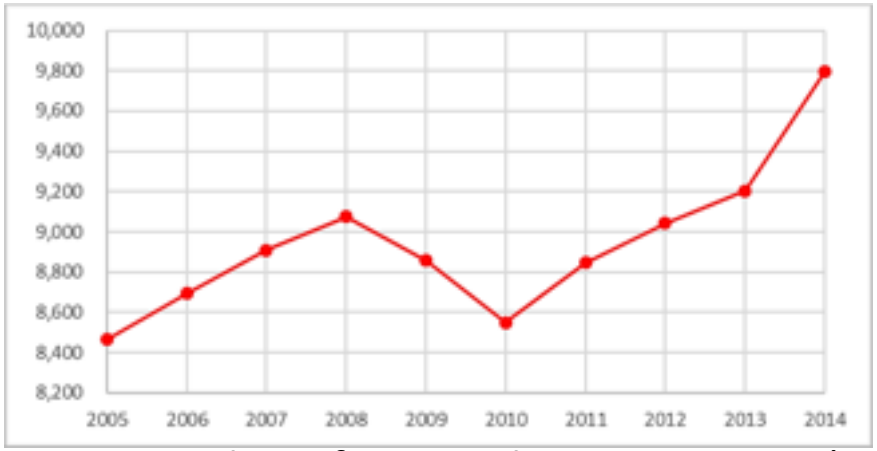


Figure 3: Population pyramids in 2017 for Japan, Canada and the USA.

Confirming your explanations will require more big-data-oriented data mining, possibly using new big data sources.

Data mining

From our hypothetical explanations, we might make inferences, imagining the future of "nursing homes" in Japan, Canada and the US (*future-oriented data mining*), using big data.



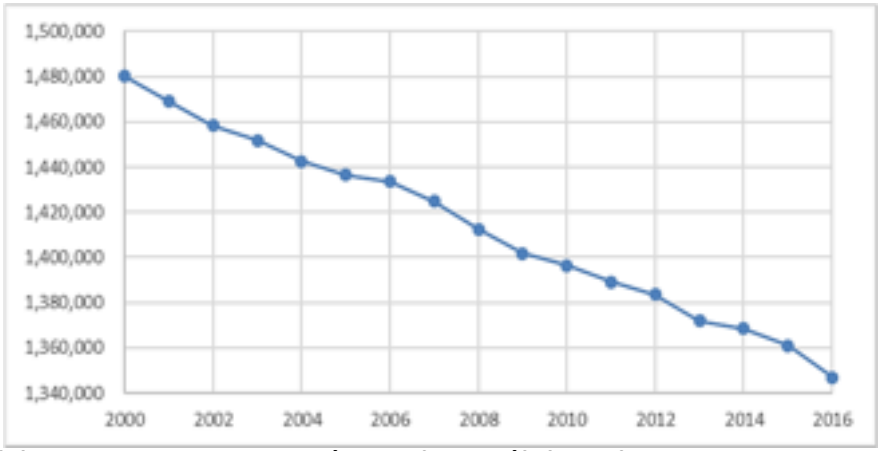


Figure 4: Number of nursing homes in Japan (red line, 2005–2014) and US (blue line, 2000–2016)

- In which country nurses for elderly care will be a good career option in the near future?
- In which country will be the establishment of nursing homes a profitable business?

Any more questions about the future? Let's hear them!

Exemplar Application of the Framework: Aging Population Issues in APEC Countries

Understanding and/or designing

From the new knowledge, inferences and plausible explanations generated in the previous phase, we have gained valuable understanding about the topics of interest.

This understanding provides us with ideas to develop new activities related to the topics or variables of interest, aiming to turning them into value of some sort.

Any more ideas about how to turn this knowledge into value? Let's hear them!

For example, from understanding the rising need for nursing homes and nurses in Japan, someone could design business plans targeting senior citizens: in-home care services, senior citizen transportation services, e-commerce store for the elderly, wheelchair manufacturing, foreign nurse recruitment agency, etc.

Understanding and/or designing

In the USA, most of these ideas (e.g., a foreign nurse recruitment agency to provide care service for the elderly) will not work everywhere, but could be successful in states like Oregon and Virginia.

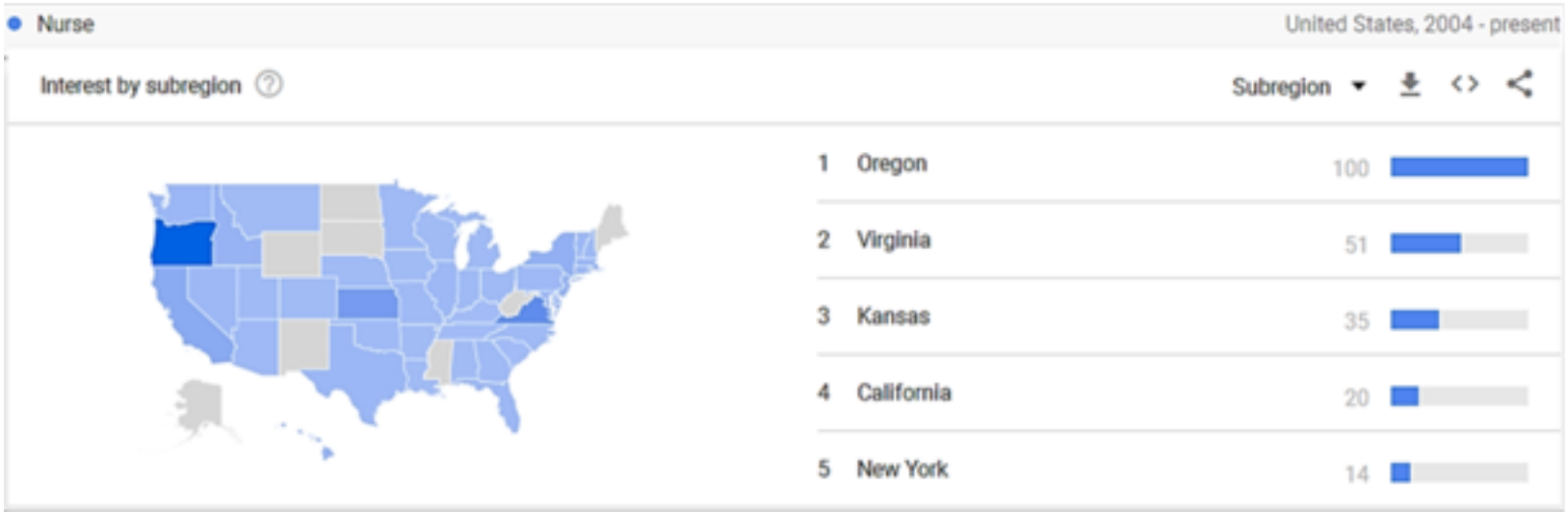


Figure 5: Interest in Google searches for the term “nurse” in United States in the last 15 years

After this presentation, what have we learned about the following questions?

1. In order to function effectively in a society driven by big data and digital economy, what are the necessary processes of statistical thinking required to handle big data?
2. How can we revise current curriculum frameworks of statistical thinking to incorporate big data for the digital economy?
3. How can we incorporate core ideas of big data for the digital economy into the high school curriculum?
4. What are plausible instructional activities (exemplar applications) for teaching the fundamental ideas of statistics while fostering statistical thinking for big data and the digital economy?

Questions and Comments are Welcomed

Thank You for your Attention!