

Strategic Data Science (SDS)

Introduction to Data Science

Karl Ho

School of Economic, Political and Policy Sciences
University of Texas at Dallas

"ipsa scientia potestas est"

"ipsa scientia potestas est"

"Knowledge itself is Power."

"ipsa scientia potestas est"

"Knowledge itself is Power."

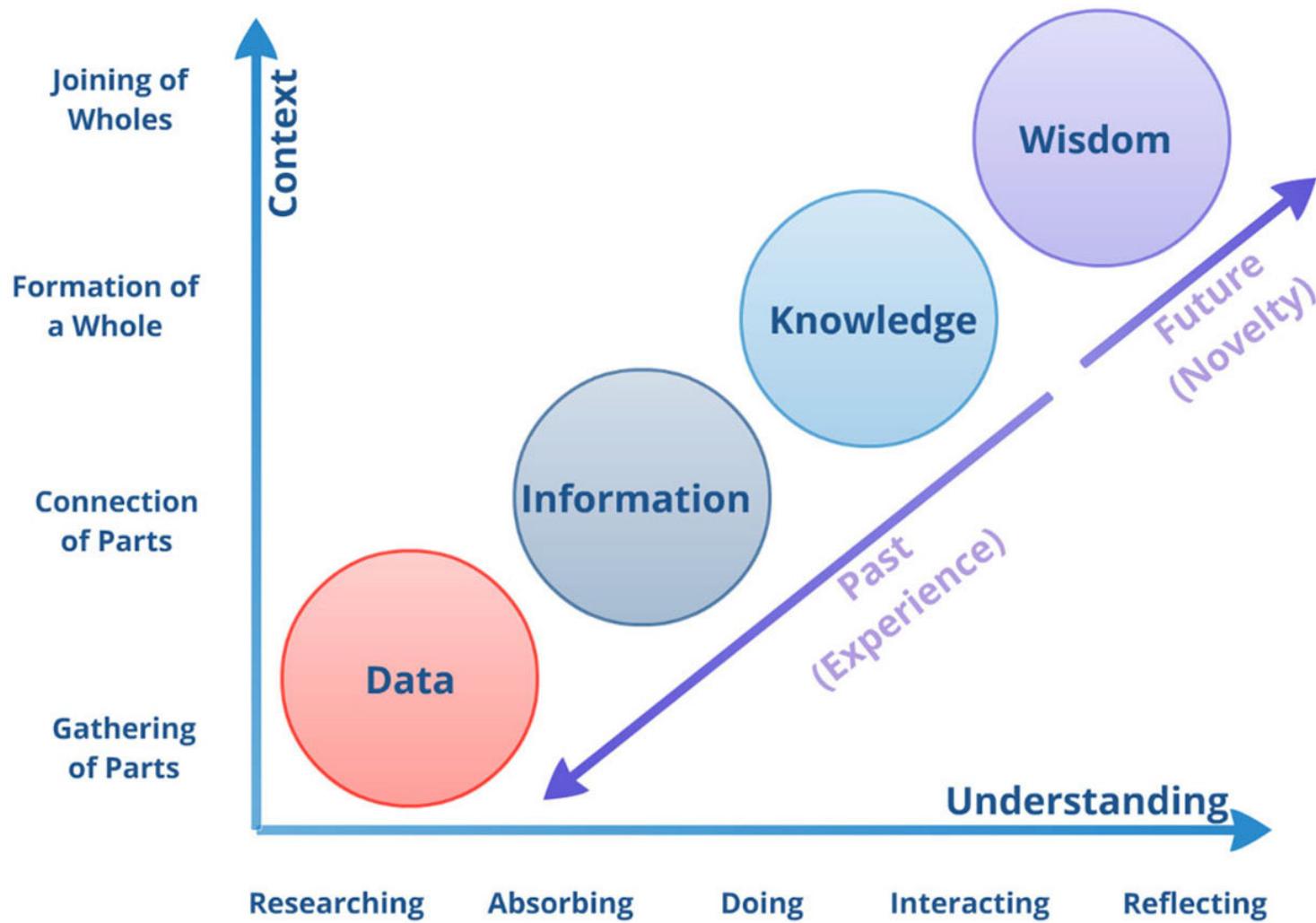
- Sir Francis Bacon

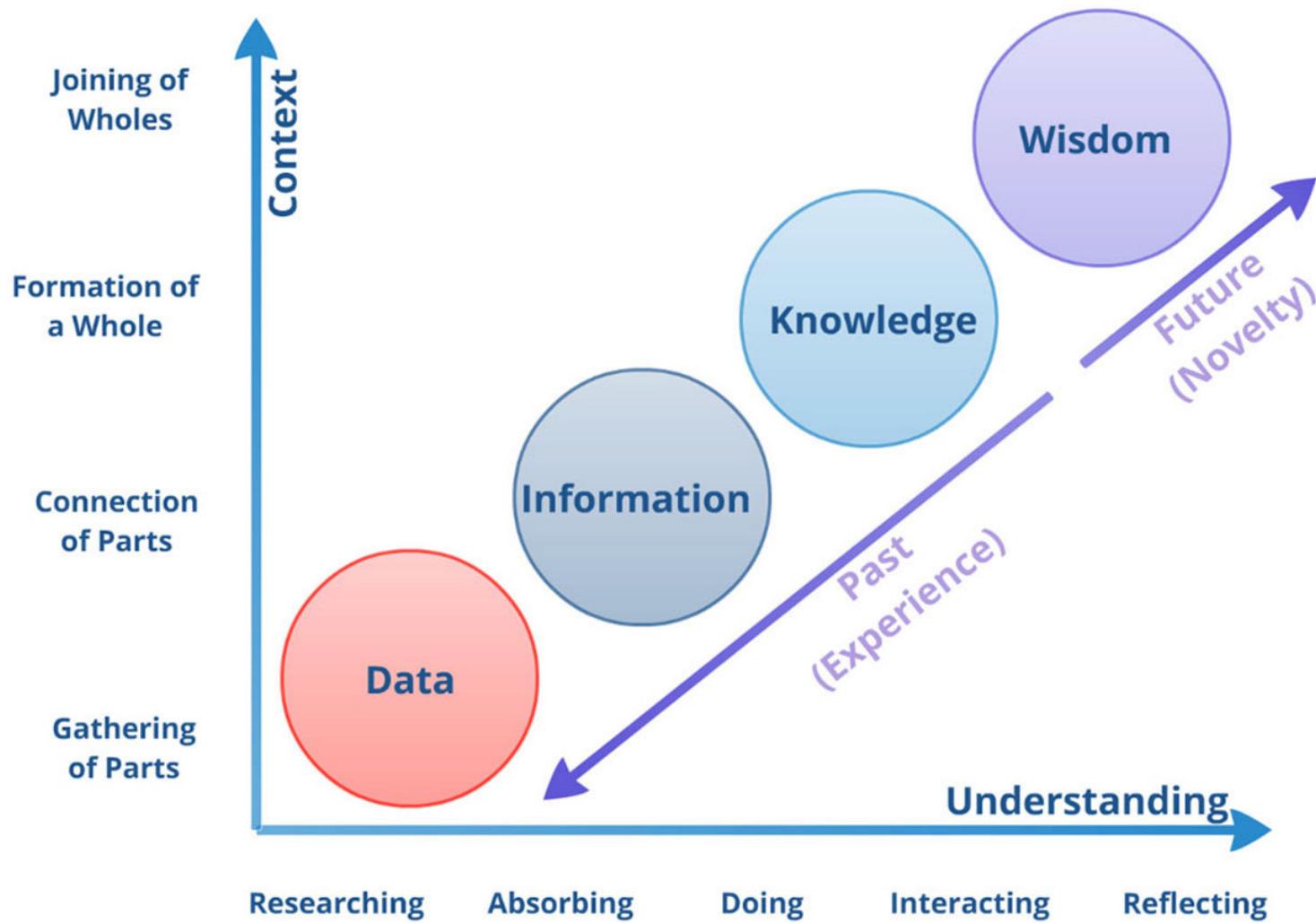
"ipsa scientia potestas est"

"Knowledge itself is Power."

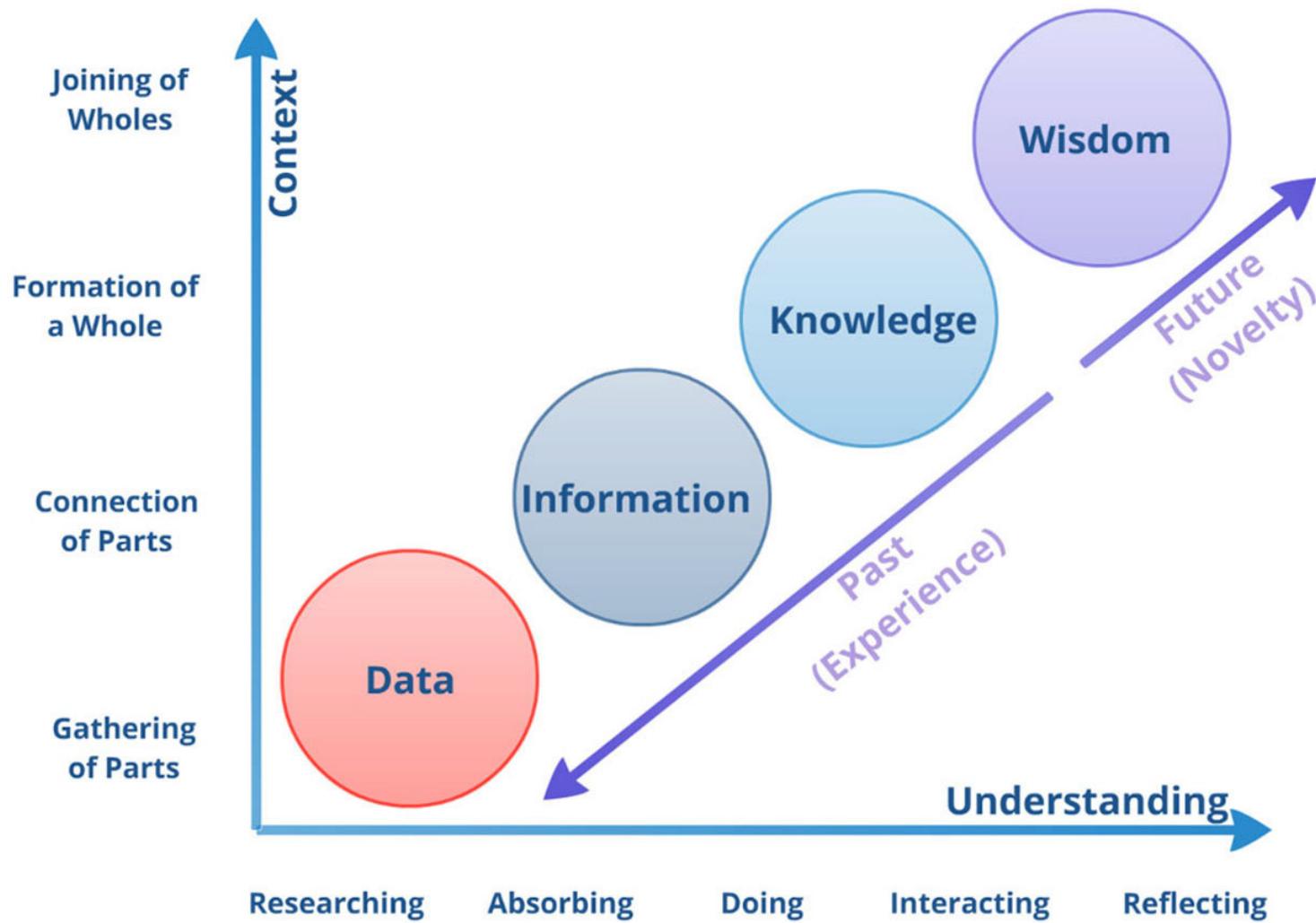
- Sir Francis Bacon

$$Power = f(Size_{Knowledge}, Veracity_{Knowledge}, Speed_{Knowledge})$$

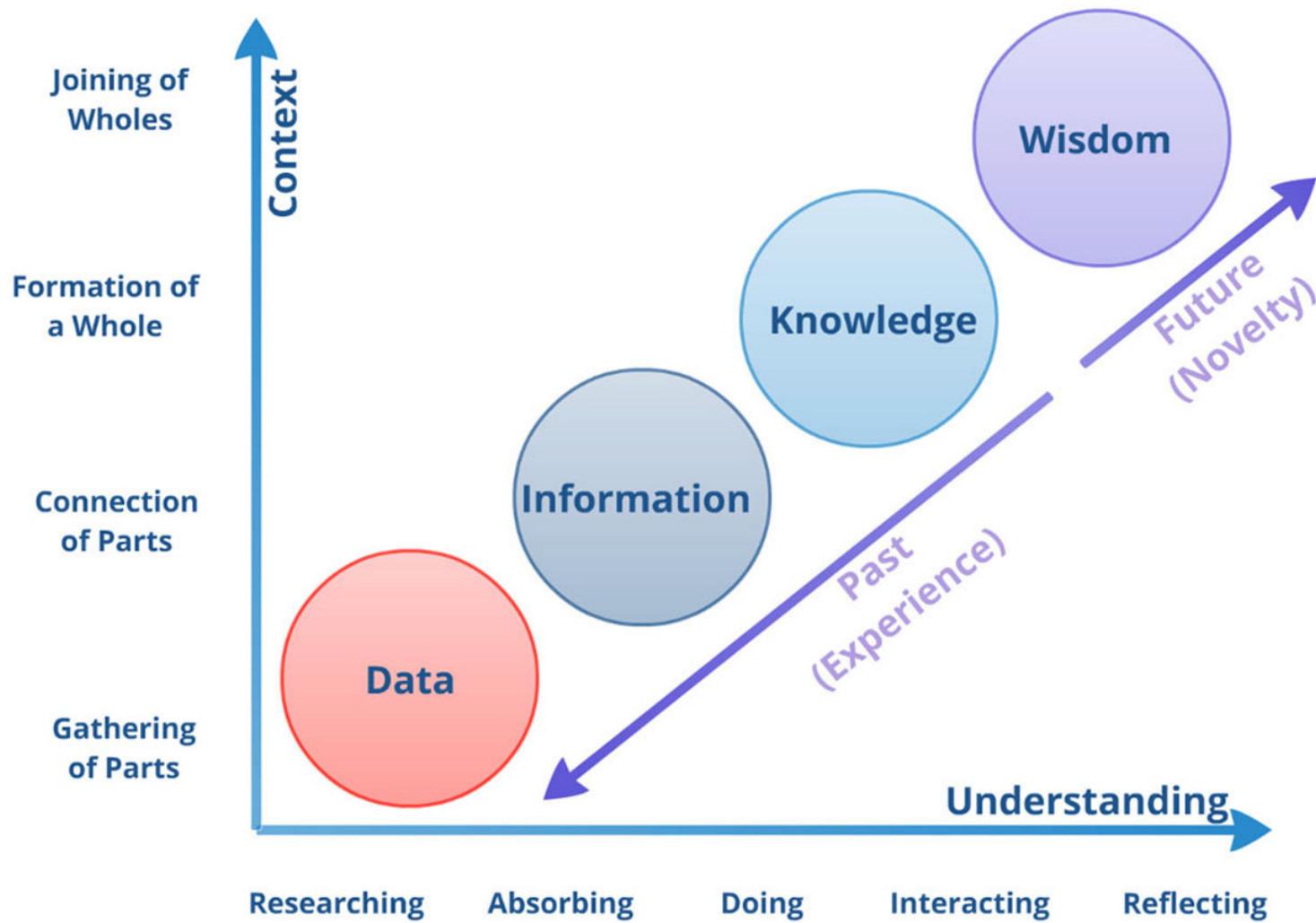




Ackoff, R.L., 1989. From data to wisdom. *Journal of applied systems analysis*, 16(1), pp.3-9.



Ackoff, R.L., 1989. From data to wisdom. *Journal of applied systems analysis*, 16(1), pp.3-9.



Ackoff, R.L., 1989. From data to wisdom. *Journal of applied systems analysis*, 16(1), pp.3-9.

"Knowledge is data."

"Knowledge is data."

"Data is power."

- 1700s Agricultural Revolution

- 1700s Agricultural Revolution
- 1780 Industrial Revolution

- 1700s Agricultural Revolution
- 1780 Industrial Revolution
- 1940 Information Revolution

- 1700s Agricultural Revolution
- 1780 Industrial Revolution
- 1940 Information Revolution
- 1950s Digital Revolution

- 1700s Agricultural Revolution
- 1780 Industrial Revolution
- 1940 Information Revolution
- 1950s Digital Revolution
- Knowledge Revolution

- 1700s Agricultural Revolution
- 1780 Industrial Revolution
- 1940 Information Revolution
- 1950s Digital Revolution
- Knowledge Revolution
- Data Revolution

What is Data?

What is Data?

Data is everything.

What is Data?

Data is everything.

- Data is ever growing.....**

What is Data?

Data is everything.

- Data is ever growing.....
 - Moore's Law

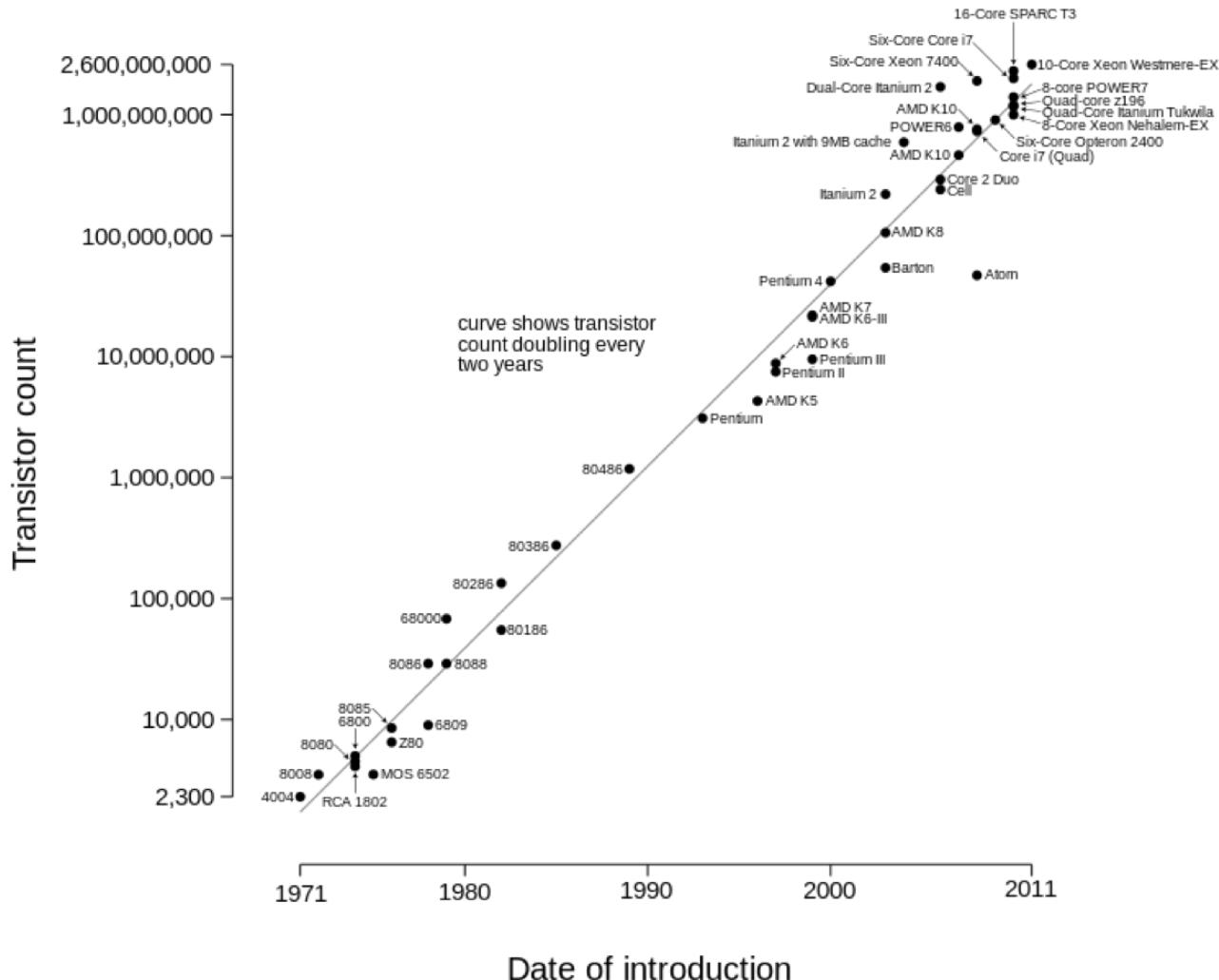
What is Data?

Data is everything.

- Data is ever growing.....
 - Moore's Law
 - Parkinson's Law

Moore's Law

Microprocessor Transistor Counts 1971-2011 & Moore's Law



Moore's Law

Moore's Law

Moore's Law

General-purpose computing capacity
grew at an annual rate of 58%.

Moore's Law

**General-purpose computing capacity
grew at an annual rate of 58%.**

**Computing power doubles every 18
months.**

Telecommunication

Telecommunication

The world's capacity for bidirectional telecommunication grew at 28% per year, closely followed by the increase in globally stored information (23%)

Telecommunication

The world's capacity for bidirectional telecommunication grew at 28% per year, closely followed by the increase in globally stored information (23%)

Hilbert, M. and López, P., 2011. The world's technological capacity to store, communicate, and compute information. *science*, p.1200970.

Digital Revolution

Humankind's capacity for unidirectional information diffusion through broadcasting channels has experienced comparatively modest annual growth (6%). Telecommunication has been dominated by digital technologies since 1990 (99.9% in digital format in 2007), and the majority of our technological memory has been in digital format since 2000s.

Parkinson's Law of Data

Parkinson's Law of Data

“Data expands to fill the space available for storage.”

Quick note about size

Bits: 8 bits = 1 byte

Bytes: 1024 bytes = 1 KB (1 to 3 digits)

Kilobytes: 1024 KB = 1 MB (4 to 6 digits)

Megabytes: 1024 MB = 1 GB (7 to 9 digits)

Gigabytes: 1024 GB = 1 TB (10 to 12 digits)

Terabytes: 1024 TB = 1 PB (13 to 15 digits)

Petabytes: 1024 PB = 1 EB (16 to 18 digits)

Exabytes: 1024 EB = 1 ZB (19 to 21 digits)

Zettabytes: 1024 ZB = 1 YB (22 to 24 digits)

Yottabytes: more than enough... (25 to 27 digits)

A Taxonomy of Data

A Taxonomy of Data

1. Numbers

A Taxonomy of Data

1. Numbers
2. Text

A Taxonomy of Data

1. Numbers
2. Text
3. Images

A Taxonomy of Data

1. Numbers
2. Text
3. Images
4. Audio

A Taxonomy of Data

1. Numbers
2. Text
3. Images
4. Audio
5. Video

A Taxonomy of Data

1. Numbers
2. Text
3. Images
4. Audio
5. Video
6. Signals

A Taxonomy of Data

1. Numbers
2. Text
3. Images
4. Audio
5. Video
6. Signals
7. Data of data: Metadata and Paradata

Categories of Data (by method)

Categories of Data (by method)

1. Survey

Categories of Data (by method)

- 1. Survey**
- 2. Experiments**

Categories of Data (by method)

- 1. Survey**
- 2. Experiments**
- 3. Qualitative Data**

Categories of Data (by method)

- 1. Survey**
- 2. Experiments**
- 3. Qualitative Data**
- 4. Text Data**

Categories of Data (by method)

- 1. Survey**
- 2. Experiments**
- 3. Qualitative Data**
- 4. Text Data**
- 5. Web Data**

Categories of Data (by method)

- 1. Survey**
- 2. Experiments**
- 3. Qualitative Data**
- 4. Text Data**
- 5. Web Data**
- 6. Complex Data**

Categories of Data (by method)

- 1. Survey**
- 2. Experiments**
- 3. Qualitative Data**
- 4. Text Data**
- 5. Web Data**
- 6. Complex Data**
 - 1. Network Data**

Categories of Data (by method)

- 1. Survey**
- 2. Experiments**
- 3. Qualitative Data**
- 4. Text Data**
- 5. Web Data**
- 6. Complex Data**
 - 1. Network Data**
 - 2. Multiple-source linked Data**

What is Big Data?

What is Big Data?

The Big data is about data that has huge volume, cannot be on one computer.

Has a lot of variety in data types, locations, formats and form. It is also getting created very very fast (velocity) (Doug Laney 2001).

What is Big Data?

The Big data is about data that has huge **volume**, cannot be on one computer.

Has a lot of **variety** in data types, locations, formats and form. It is also getting created very very fast (**velocity**) (Doug Laney 2001).

What is Big Data?

What is Big Data?

Burt Monroe (2012)

5Vs of Big data

What is Big Data?

Burt Monroe (2012)

5Vs of Big data

- Volume

What is Big Data?

Burt Monroe (2012)

5Vs of Big data

- Volume
- Variety

What is Big Data?

Burt Monroe (2012)

5Vs of Big data

- Volume
- Variety
- Velocity

What is Big Data?

Burt Monroe (2012)

5Vs of Big data

- Volume
- Variety
- Velocity
- Vinculation

What is Big Data?

Burt Monroe (2012)

5Vs of Big data

- Volume
- Variety
- Velocity
- Vinculation
- Validity

Big Data Research

Table 1. Number of NSF “Big Data” Projects, by Directorate, 2009–2013

| | 2009 | 2010 | 2011 | 2012 | 2013 | Total |
|-------|------|------|------|------|------|-------|
| CSE | 2 | 0 | 3 | 36 | 124 | 163 |
| ENG | 0 | 0 | 0 | 3 | 30 | 33 |
| SBE | 0 | 0 | 0 | 4 | 23 | 27 |
| MPS | 0 | 0 | 0 | 7 | 18 | 25 |
| Other | 0 | 0 | 0 | 9 | 19 | 28 |
| Total | 2 | 0 | 3 | 59 | 214 | 276 |

CSE – Computer and Information Science and Engineering

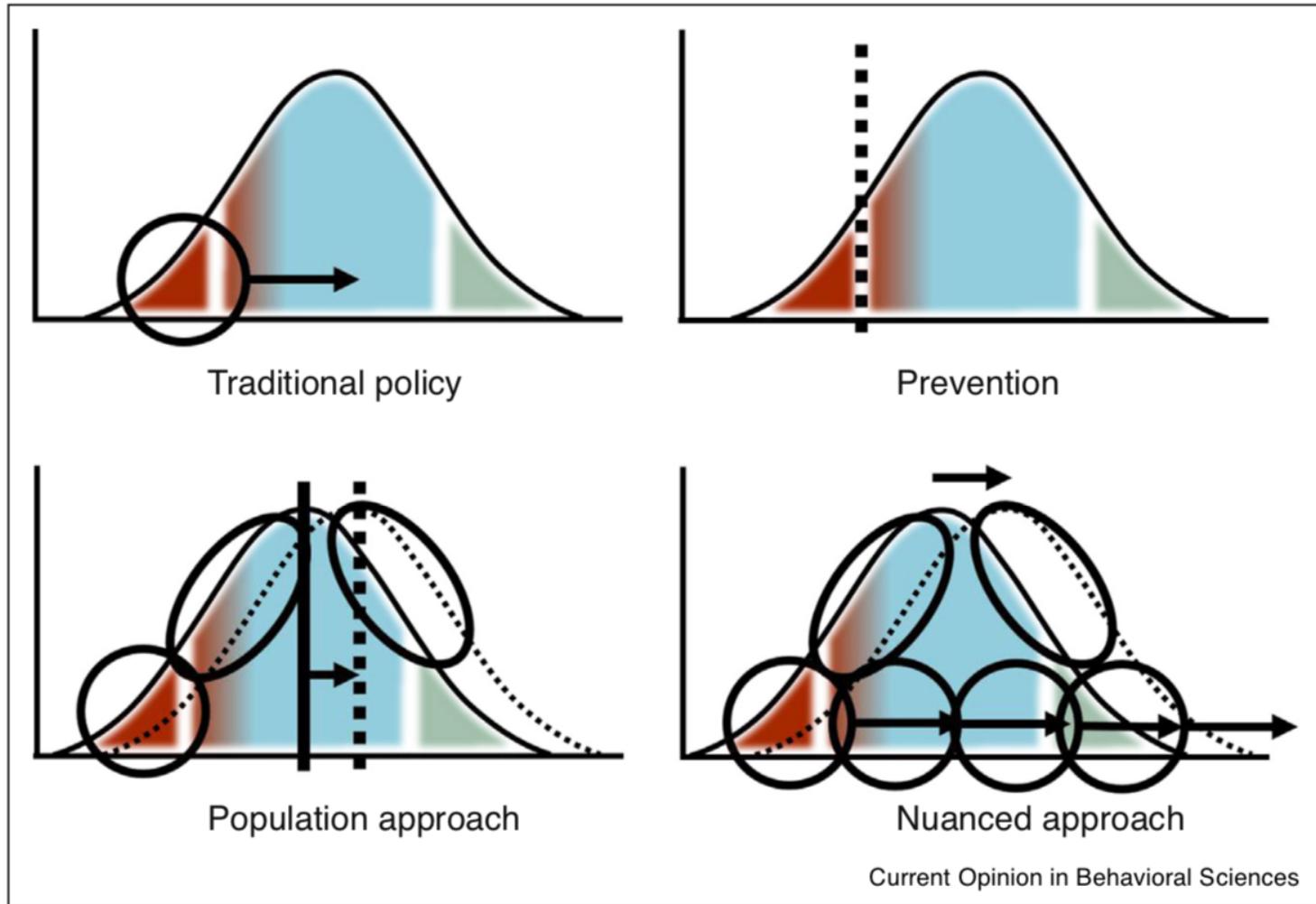
ENG – Engineering

SBE – Social Behavioral and Economic Sciences

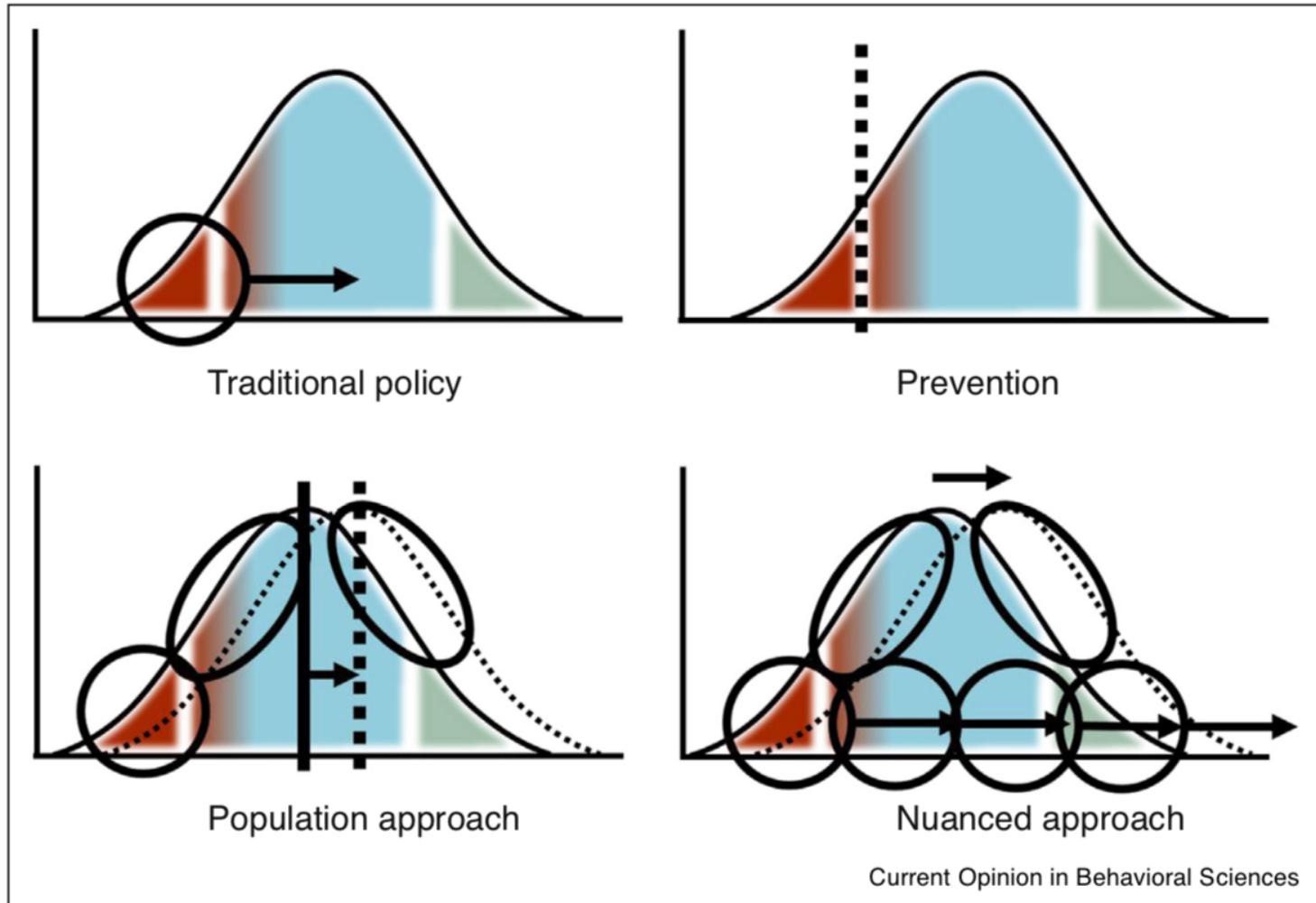
Mathematics and Physical Science

Public Policy and Big Data

Public Policy and Big Data



Public Policy and Big Data



Using big data to solve real problems through academic and industry partnerships

Academic Partner

Generate theories based on academic literature

Analyze data and publish papers

Inform academic theories



Industry Partner

Create program for industry purposes

Save data for academic partner

Inform product development

Inform real world problems

Current Opinion in Behavioral Sciences

Using big data to solve real problems through academic and industry partnerships

Academic Partner

Generate theories based on academic literature

Analyze data and publish papers

Inform academic theories



Industry Partner

Create program for industry purposes

Save data for academic partner

Inform product development

Current Opinion in Behavioral Sciences

Table 1**Summary of four key trade-offs and future integrative directions for the use of Big Data in social and behavioral science research**

| No. | Consideration | Trade-off | Directions |
|-----|-----------------------|--|--|
| 1 | Outcome focus | Prediction versus explanation | <ul style="list-style-type: none">- Use Big Data to predict relevant outcomes- Use Big Data approaches to inform explanatory research- Establish a productive cycle that creates new theoretical insights based on prediction efforts and that uses explanatory insights to build better prediction models |
| 2 | Epistemological focus | Induction versus deduction | <ul style="list-style-type: none">- Use Big Data for bottom-up, data-driven research- Use Big Data to inform top-down, theory-driven research- Integrate data-driven and theory-driven approaches, thereby providing a rich knowledge base that allows to investigate increasingly fine-grained questions and explanations |
| 3 | Data optimization | Bigness versus representativeness | <ul style="list-style-type: none">- Make use of available Big Data- Complement Big Data with more traditional dynamic, interactive and multi-modal laboratory-based and field-based data- Team-up and pool alternative rich data-sets across laboratories worldwide to a degree that they become Big Data |
| 4 | Data usage | Data access versus scientific independence | <ul style="list-style-type: none">- Make use and profit from the accessibility to Big Data- Work on Big Data solutions that assure scientific independence and quality standards (e.g. with regard to methodology and transparency)- Establish independent actors as Big Data players that maintain scientific independence and high standards of scientific quality |

Table 1**Summary of four key trade-offs and future integrative directions for the use of Big Data in social and behavioral science research**

| No. | Consideration | Trade-off | Directions |
|-----|-----------------------|--|--|
| 1 | Outcome focus | Prediction versus explanation | <ul style="list-style-type: none">- Use Big Data to predict relevant outcomes- Use Big Data approaches to inform explanatory research- Establish a productive cycle that creates new theoretical insights based on prediction efforts and that uses explanatory insights to build better prediction models |
| 2 | Epistemological focus | Induction versus deduction | <ul style="list-style-type: none">- Use Big Data for bottom-up, data-driven research- Use Big Data to inform top-down, theory-driven research- Integrate data-driven and theory-driven approaches, thereby providing a rich knowledge base that allows to investigate increasingly fine-grained questions and explanations |
| 3 | Data optimization | Bigness versus representativeness | <ul style="list-style-type: none">- Make use of available Big Data- Complement Big Data with more traditional dynamic, interactive and multi-modal laboratory-based and field-based data- Team-up and pool alternative rich data-sets across laboratories worldwide to a degree that they become Big Data |
| 4 | Data usage | Data access versus scientific independence | <ul style="list-style-type: none">- Make use and profit from the accessibility to Big Data- Work on Big Data solutions that assure scientific independence and quality standards (e.g. with regard to methodology and transparency)- Establish independent actors as Big Data players that maintain scientific independence and high standards of scientific quality |

Mahmoodi, J., Leckelt, M., van Zalk, M.W., Geukes, K. and Back, M.D., 2017. Big Data approaches in social and behavioral science: four key trade-offs and a call for integration. *Current Opinion in Behavioral Sciences*, 18, pp.57-62.

Table 1**Summary of four key trade-offs and future integrative directions for the use of Big Data**

| No. | Consideration | Trade-off | |
|-----|-----------------------|--|--|
| 1 | Outcome focus | Prediction versus explanation | <ul style="list-style-type: none">- Use Big Data to predict relevant outcomes- Use Big Data approaches to inform explanations- Establish a productive cycle that creates predictions and that uses explanatory insights to better inform predictions |
| 2 | Epistemological focus | Induction versus deduction | <ul style="list-style-type: none">- Use Big Data for bottom-up, data-driven theory development- Use Big Data to inform top-down, theoretical models- Integrate data-driven and theory-driven approaches that allows to investigate increasingly fine-grained phenomena |
| 3 | Data optimization | Bigness versus representativeness | <ul style="list-style-type: none">- Make use of available Big Data- Complement Big Data with more traditional laboratory-based and field-based data- Team-up and pool alternative rich data sources before they become Big Data |
| 4 | Data usage | Data access versus scientific independence | <ul style="list-style-type: none">- Make use and profit from the accessibility of Big Data- Work on Big Data solutions that assume scientific independence with regard to methodology and transparency- Establish independent actors as Big Data custodians |

- Prediction-explanation gap
- Induction-deduction gap
- Bigness-representativeness gap
- Data access gap

What is Data Science?

What is Data Science?

What is Data Science?

1. Science of Data

What is Data Science?

1. Science of Data
2. Understand Data Scientifically

**The key word in "Data
Science" is not Data....**

**The key word in "Data
Science" is not Data....
it is Science.**

**The key word in "Data
Science" is not Data....
it is Science.**

- Jeff Leek

The long term impact of
Data Science will be
measured by the scientific
questions we can answer
with the data.

The long term impact of
Data Science will be
measured by the scientific
questions we can answer
with the data.

- Jeff Leek

Data Science Keywords

Data Science Keywords

- Data management

Data Science Keywords

- Data management
- Data analytics

Data Science Keywords

- Data management
- Data analytics
- Data scientists

Data Science Keywords

- Data management
- Data analytics
- Data scientists
- Data curation

Data Science Keywords

- Data management
- Data analytics
- Data scientists
- Data curation
- Modeling

Data Science Keywords

- Data management
- Data analytics
- Data scientists
- Data curation
- Modeling
- CRMs

How Data are generated?

How Data are generated?

- Computers

How Data are generated?

- Computers
- Web

How Data are generated?

- Computers
- Web
- Mobile devices

How Data are generated?

- Computers
- Web
- Mobile devices
- IoT (Internet of Things)

How Data are generated?

- Computers
- Web
- Mobile devices
- IoT (Internet of Things)
- Further extension of human users (e.g. AI, avatars)

How Data are generated?

How Data are generated?

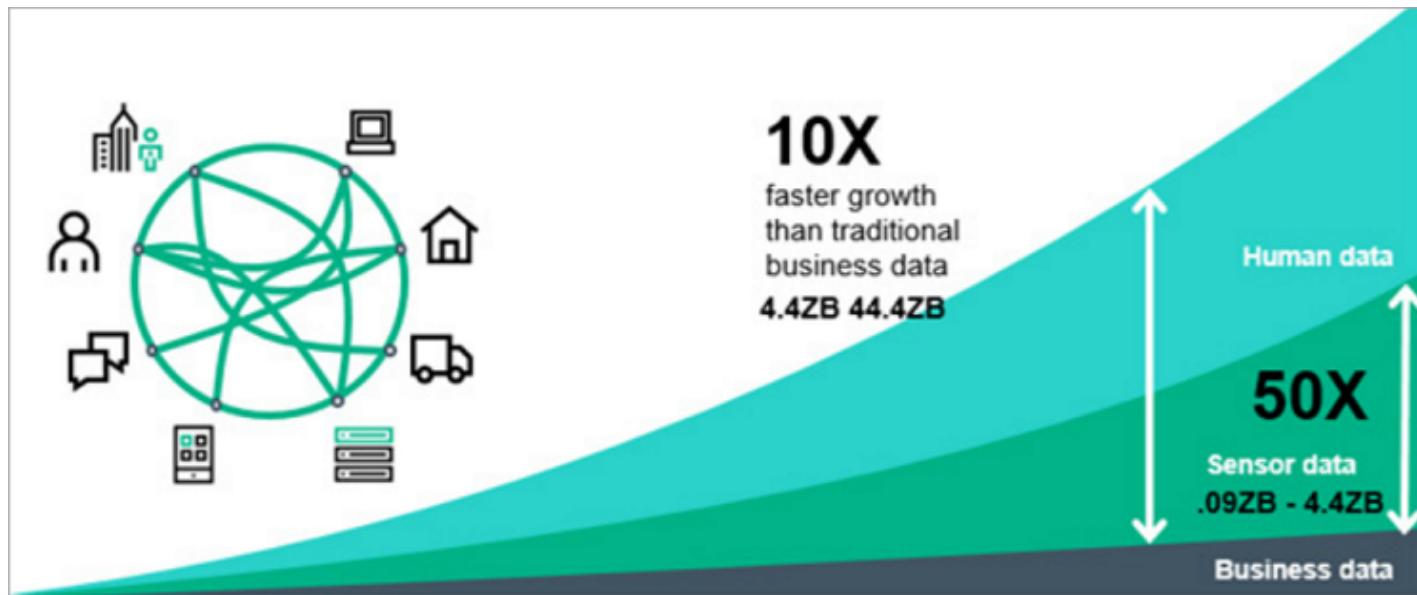
The size of the digital universe will double every two years at least.

- InsideBigdata.com

How Data are generated?

The size of the digital universe will double every two years at least.

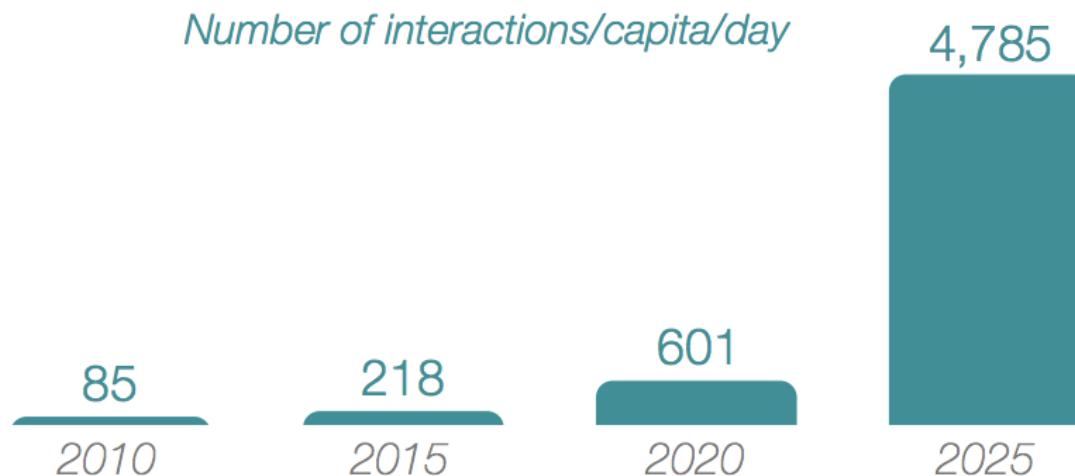
- InsideBigdata.com



How Data are generated?

How Data are generated?

Figure 8. | Interactions per Connected Person per Day



Source: IDC's Data Age 2025 study, sponsored by Seagate, April 2017

"Data Lake" Ubiquitous

"Data Lake" Ubiquitous

**Massive raw data repository in its rawest form
pending processing.**

Data Analytics vs. Data Analysis

Data Analytics vs. Data Analysis

Data analytics refers to generation, acquisition, management, modeling and visualization of data.

Data Analytics vs. Data Analysis

Data analytics refers to generation, acquisition, management, modeling and visualization of data.

Thomas Davenport and his colleagues (2007) emphasize the ability to "collect, analyze and act on data".

Data Analytics vs. Data Analysis

Data analytics refers to generation, acquisition, management, modeling and visualization of data.

Thomas Davenport and his colleagues (2007) emphasize the ability to "collect, analyze and act on data".

Davenport, Thomas H., and Jeanne G. Harris. 2007. *Competing on analytics: The new science of winning.*

Harvard Business Press.

Data Analytics vs. Data Analysis

Data Analytics vs. Data Analysis

Data analytics goes beyond only providing analysis of data but focuses on the action or decision making informed by data.

Social Data Analytics: A journey just set afoot

Social Data Analytics: A journey just set afoot

Social data was initially referred to data generated from social media. It is now not confined to that generation mode but is more generally data generated by people or users.

Social Data Analytics: A journey just set afoot

Social data was initially referred to data generated from social media. It is now not confined to that generation mode but is more generally data generated by people or users.

Social data analytics encompasses the generation, management, modeling and visualization of social data.

**.... social science is beginning to shape
the world of big data.**

.... social science is beginning to shape the world of big data.

Much of big data is social data.... It is the responsibility of social scientists to assume their central place in the world of big data, to shape the questions we ask of big data, and to characterize what does and does not make for a convincing answer.

.... social science is beginning to shape the world of big data.

Much of big data is social data.... It is the responsibility of social scientists to assume their central place in the world of big data, to shape the questions we ask of big data, and to characterize what does and does not make for a convincing answer.

- Monroe, Pan, Roberts, Sen and Sinclair 2015

.... social science is beginning to shape the world of big data.

Much of big data is social data.... It is the responsibility of social scientists to assume their central place in the world of big data, to shape the questions we ask of big data, and to characterize what does and does not make for a convincing answer.

- Monroe, Pan, Roberts, Sen and Sinclair 2015

Monroe, B.L., Pan, J., Roberts, M.E., Sen, M. and Sinclair, B., 2015. No! Formal theory, causal inference, and big data are not contradictory trends in political science. PS: Political Science & Politics, 48(1), pp.71-74.

Social (Data) Scientist's mission

Social (Data) Scientist's mission

Two major areas to which social scientists can contribute, based on decades of experience and work with end users, are:

Social (Data) Scientist's mission

Two major areas to which social scientists can contribute, based on decades of experience and work with end users, are:

1. Inference

Social (Data) Scientist's mission

Two major areas to which social scientists can contribute, based on decades of experience and work with end users, are:

1. Inference
2. Data quality.

Social (Data) Scientist's mission

Two major areas to which social scientists can contribute, based on decades of experience and work with end users, are:

1. Inference
2. Data quality.

- Foster *et al.* 2016

Social (Data) Scientist's mission

Social (Data) Scientist's mission

Compared to computer scientists and business analytics researchers, we are distinct in not only our familiarity with data, statistical models and inference.

Social (Data) Scientist's mission

Compared to computer scientists and business analytics researchers, we are distinct in not only our familiarity with data, statistical models and inference.

Social scientists pursue a good cause, something we can contribute: to make a difference, to bring public good and to shape a better society.

Social (Data) Scientist's mission

Social (Data) Scientist's mission

Grimmer, J., 2015. We are all social scientists now: how big data, machine learning, and causal inference work together. *PS: Political Science & Politics*, 48(1), pp.80-83.

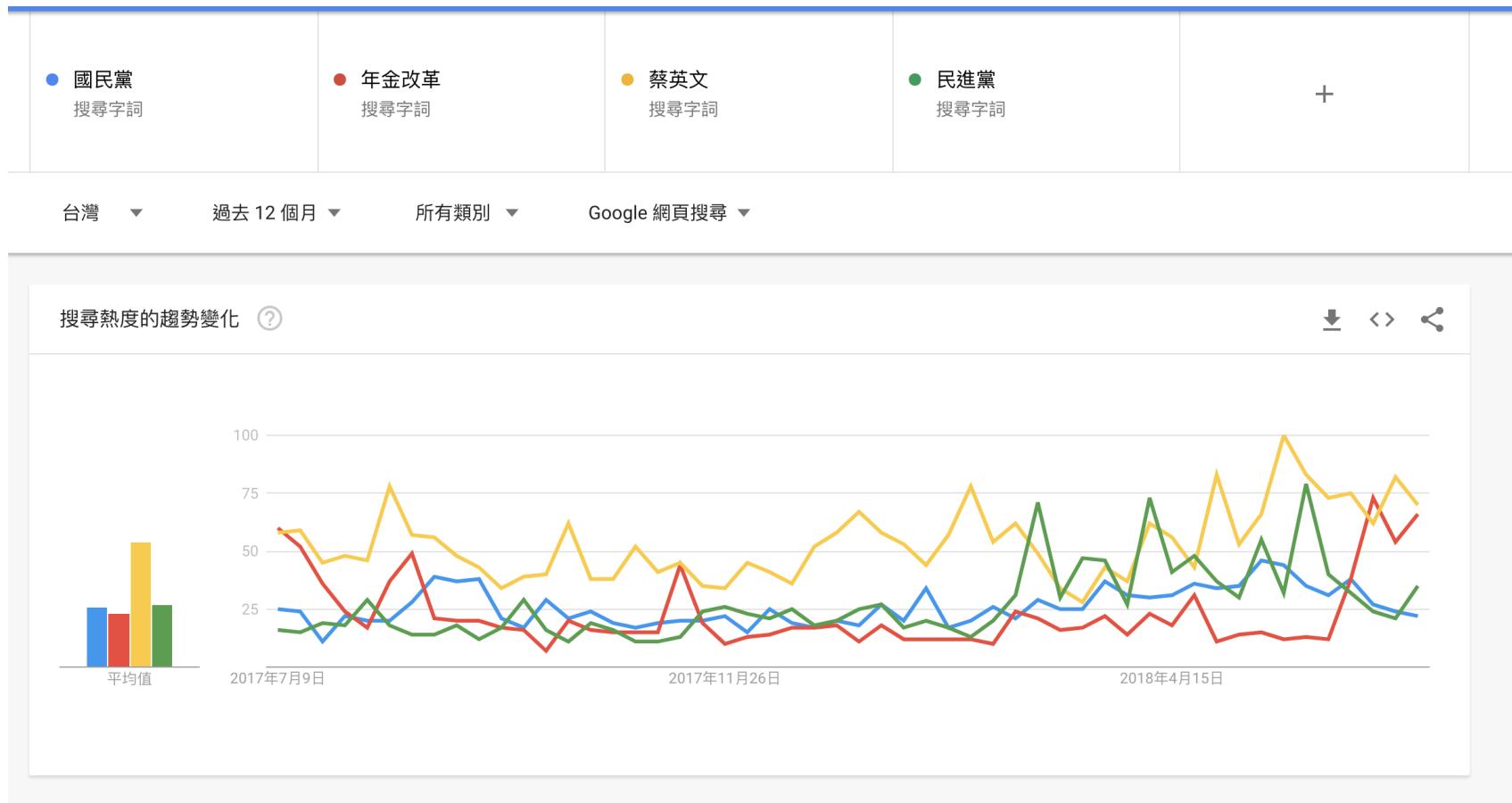
Social (Data) Scientist's mission

Social scientists know that large amounts of data will not overcome the selection problems that make causal inference so difficult.

Grimmer, J., 2015. We are all social scientists now: how big data, machine learning, and causal inference work together. *PS: Political Science & Politics*, 48(1), pp.80-83.

Google Trends

Google Trends



The story of Google Flu Trend

The story of Google Flu Trend

By using Big Data of search queries, Google Flu Trend (GFT) predicted the flu-like illness rate in a population.

The story of Google Flu Trend

By using Big Data of search queries, Google Flu Trend (GFT) predicted the flu-like illness rate in a population.

However, the journal *Nature* where GFT published the findings on figured the GFT overestimated as much as twice than the actual data. Two political scientists helped fix and address the problem.

The story of Google Flu Trend

By using Big Data of search queries, Google Flu Trend (GFT) predicted the flu-like illness rate in a population.

However, the journal *Nature* where GFT published the findings on figured the GFT overestimated as much as twice than the actual data. Two political scientists helped fix and address the problem.

Lesson we learn:

Political Science can save the world!

The story of Google Flu Trend

The story of Google Flu Trend

Lazer, Kennedy, King and Vespignani (2014)

Traditional “small data” often offer information that is not contained (or containable) in big data, and the very factors that have enabled big data are enabling more traditional data collection ([watch TED talk by Dr. Joel Selanikio](#)). The Internet has opened the way for improving standard surveys, experiments, and health reporting. (Lazer et al. 2014 *Science*)

A Theory of Data: Understanding Data Generation

Data Generation

Data Generation

R. Connelly et al. / Social Science Research 59 (2016) 1–12

| Made Data Experimental | Made Data Observational (e.g. Social Surveys) | Found Data Administrative Data | Found Data Other Types of Big Data |
|---|--|--|---|
| <ul style="list-style-type: none">• Data are collected to investigate a fixed hypothesis.• Usually relatively small in size.• Usually relatively uncomplex.• Highly systematic.• Known sample / population. | <ul style="list-style-type: none">• Data may be used to address multiple research questions.• Data may be very large and complex (but usually smaller than big data).• Highly systematic.• Known sample / population. | <ul style="list-style-type: none">• Data are not collected for research purposes.• May be large and complex.• Semi-systematic.• May be messy (i.e. may involve extensive data management to clean and organise the data).• Multidimensional (i.e. may involve multiple fragments of data which have to be brought together through data linkage).• Usually a known sample / population. | <ul style="list-style-type: none">• Data are not collected for research purposes.• May be very large and very complex.• Some sources will be very unsystematic (e.g. data from social media posts).• Very messy / chaotic.• Multidimensional (i.e. may involve multiple fragments of data which have to be brought together through data linkage).• Sample / population usually unknown. |

Fig. 1. Characteristics of quantitative social science data resources.

Data Generation

R. Connelly et al. / Social Science Research 59 (2016) 1–12

| Made Data Experimental | Made Data Observational (e.g. Social Surveys) | Found Data Administrative Data | Found Data Other Types of Big Data |
|---|--|--|---|
| <ul style="list-style-type: none">• Data are collected to investigate a fixed hypothesis.• Usually relatively small in size.• Usually relatively uncomplex.• Highly systematic.• Known sample / population. | <ul style="list-style-type: none">• Data may be used to address multiple research questions.• Data may be very large and complex (but usually smaller than big data).• Highly systematic.• Known sample / population. | <ul style="list-style-type: none">• Data are not collected for research purposes.• May be large and complex.• Semi-systematic.• May be messy (i.e. may involve extensive data management to clean and organise the data).• Multidimensional (i.e. may involve multiple fragments of data which have to be brought together through data linkage).• Usually a known sample / population. | <ul style="list-style-type: none">• Data are not collected for research purposes.• May be very large and very complex.• Some sources will be very unsystematic (e.g. data from social media posts).• Very messy / chaotic.• Multidimensional (i.e. may involve multiple fragments of data which have to be brought together through data linkage).• Sample / population usually unknown. |

Fig. 1. Characteristics of quantitative social science data resources.

Data Generation

R. Connelly et al. / Social Science Research 59 (2016) 1–12

| Made Data Experimental | Made Data Observational (e.g. Social Surveys) | Found Data Administrative Data | Found Data Other Types of Big Data |
|---|--|--|---|
| <ul style="list-style-type: none">• Data are collected to investigate a fixed hypothesis.• Usually relatively small in size.• Usually relatively uncomplex.• Highly systematic.• Known sample / population. | <ul style="list-style-type: none">• Data may be used to address multiple research questions.• Data may be very large and complex (but usually smaller than big data).• Highly systematic.• Known sample / population. | <ul style="list-style-type: none">• Data are not collected for research purposes.• May be large and complex.• Semi-systematic.• May be messy (i.e. may involve extensive data management to clean and organise the data).• Multidimensional (i.e. may involve multiple fragments of data which have to be brought together through data linkage).• Usually a known sample / population. | <ul style="list-style-type: none">• Data are not collected for research purposes.• May be very large and very complex.• Some sources will be very unsystematic (e.g. data from social media posts).• Very messy / chaotic.• Multidimensional (i.e. may involve multiple fragments of data which have to be brought together through data linkage).• Sample / population usually unknown. |

Fig. 1. Characteristics of quantitative social science data resources.

Administrative Data

Administrative data are defined as data which derive from the operation of administrative systems, typically by public sector agencies

- Connelly et al. 2016

Data Methods

Data Methods

1. Survey

Data Methods

1. Survey
2. Experiments

Data Methods

1. Survey
2. Experiments
3. Qualitative Data

Data Methods

1. Survey
2. Experiments
3. Qualitative Data
4. Text Data

Data Methods

1. Survey
2. Experiments
3. Qualitative Data
4. Text Data
5. Web Data

Data Methods

1. Survey
2. Experiments
3. Qualitative Data
4. Text Data
5. Web Data
6. Machine Data

Data Methods

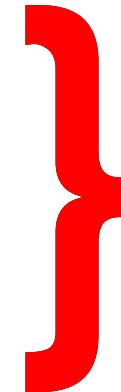
1. Survey
2. Experiments
3. Qualitative Data
4. Text Data
5. Web Data
6. Machine Data
7. Complex Data
 1. Network Data

Data Methods

1. Survey
2. Experiments
3. Qualitative Data
4. Text Data
5. Web Data
6. Machine Data
7. Complex Data
 1. Network Data
 2. Multiple-source linked Data

Data Methods

- 1. Survey
- 2. Experiments
- 3. Qualitative Data
- 4. Text Data
- 5. Web Data
- 6. Machine Data
- 7. Complex Data
 - 1. Network Data
 - 2. Multiple-source linked Data



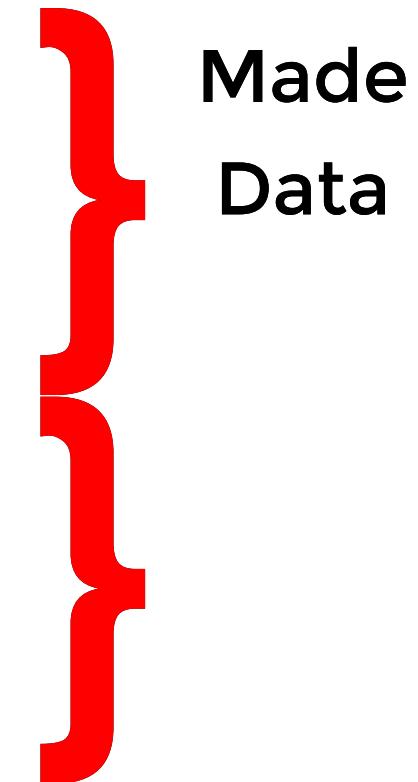
Data Methods

- 1. Survey
- 2. Experiments
- 3. Qualitative Data
- 4. Text Data
- 5. Web Data
- 6. Machine Data
- 7. Complex Data
 - 1. Network Data
 - 2. Multiple-source linked Data

} Made
Data

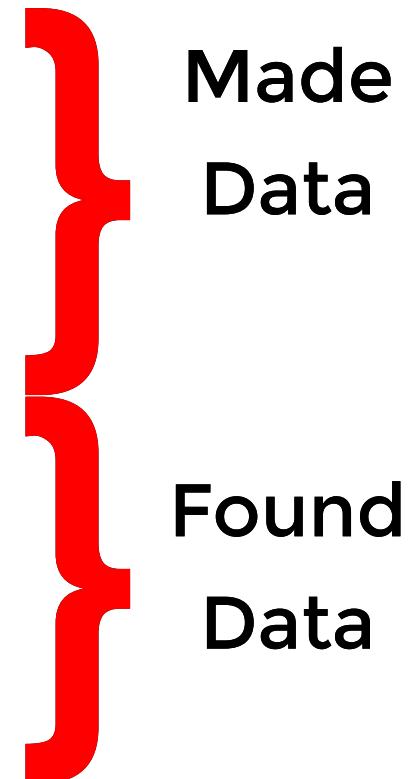
Data Methods

- 1. Survey
- 2. Experiments
- 3. Qualitative Data
- 4. Text Data
- 5. Web Data
- 6. Machine Data
- 7. Complex Data
 - 1. Network Data
 - 2. Multiple-source linked Data



Data Methods

- 1. Survey
- 2. Experiments
- 3. Qualitative Data
- 4. Text Data
- 5. Web Data
- 6. Machine Data
- 7. Complex Data
 - 1. Network Data
 - 2. Multiple-source linked Data



Statistical Modeling: The Two Cultures

Statistical Modeling: The Two Cultures

Leo Breiman 2001: *Statistical Science*

Statistical Modeling: The Two Cultures

Leo Breiman 2001: *Statistical Science*

One assumes that
the data are
generated by a
given stochastic
data model.

Statistical Modeling: The Two Cultures

Leo Breiman 2001: *Statistical Science*

**One assumes that
the data are
generated by a
given stochastic
data model.**

**The other uses
algorithmic models
and treats the data
mechanism as
unknown.**

Statistical Modeling: The Two Cultures

Leo Breiman 2001: *Statistical Science*

**One assumes that
the data are
generated by a
given stochastic
data model.**

**The other uses
algorithmic models
and treats the data
mechanism as
unknown.**

Data Model

Statistical Modeling: The Two Cultures

Leo Breiman 2001: *Statistical Science*

One assumes that
the data are
generated by a
given stochastic
data model.

Data Model

The other uses
algorithmic models
and treats the data
mechanism as
unknown.

Algorithmic Model

Statistical Modeling: The Two Cultures

Leo Breiman 2001: *Statistical Science*

One assumes that
the data are
generated by a
given stochastic
data model.

The other uses
algorithmic models
and treats the data
mechanism as
unknown.

Data Model

Small data

Algorithmic Model

Statistical Modeling: The Two Cultures

Leo Breiman 2001: *Statistical Science*

One assumes that
the data are
generated by a
given stochastic
data model.

The other uses
algorithmic models
and treats the data
mechanism as
unknown.

Data Model

Small data

Algorithmic Model

Complex, big data

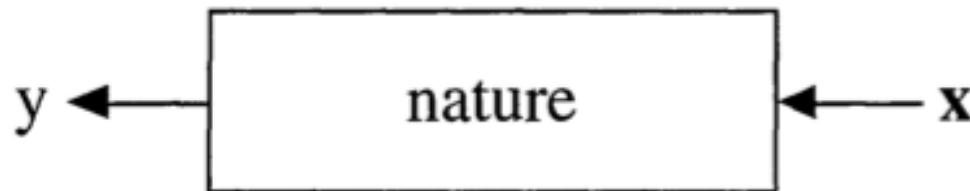
Theory: Data Generation Process

Theory: Data Generation Process

Data are generated in many fashions. Picture this: independent variable x goes in one side of the box-- we call it nature for now-- and dependent variable y come out from the other side.

Theory: Data Generation Process

Data are generated in many fashions. Picture this: independent variable x goes in one side of the box-- we call it nature for now-- and dependent variable y come out from the other side.



Theory: Data Generation Process

Theory: Data Generation Process

Data Model

Theory: Data Generation Process

Data Model

The analysis in this culture starts with assuming a stochastic data model for the inside of the black box. For example, a common data model is that data are generated by independent draws from response variables.

Theory: Data Generation Process

Data Model

The analysis in this culture starts with assuming a stochastic data model for the inside of the black box. For example, a common data model is that data are generated by independent draws from response variables.

Response Variable= $f(\text{Predictor variables, random noise, parameters})$

Theory: Data Generation Process

Data Model

The analysis in this culture starts with assuming a stochastic data model for the inside of the black box. For example, a common data model is that data are generated by independent draws from response variables.

Response Variable= $f(\text{Predictor variables, random noise, parameters})$

Reading the response variable is a function of a series of predictor/independent variables, plus random noise (normally distributed errors) and other parameters.

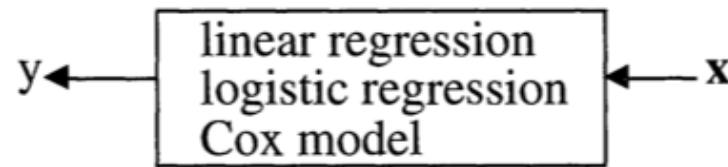
Theory: Data Generation Process

Theory: Data Generation Process

Data Model

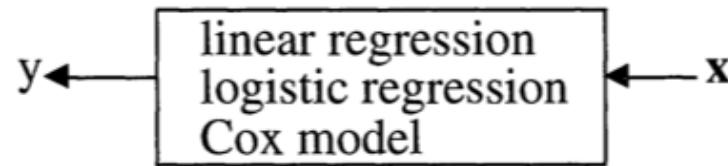
Theory: Data Generation Process

Data Model



Theory: Data Generation Process

Data Model



The values of the parameters are estimated from the data and the model then used for information and/or prediction.

Theory: Data Generation Process

Theory: Data Generation Process Algorithmic Modeling

Theory: Data Generation Process

Algorithmic Modeling

The analysis in this approach considers the inside of the box complex and unknown. Their approach is to find a function $f(x)$ -an algorithm that operates on x to predict the responses y .

Theory: Data Generation Process

Algorithmic Modeling

The analysis in this approach considers the inside of the box complex and unknown. Their approach is to find a function $f(x)$ -an algorithm that operates on x to predict the responses y .

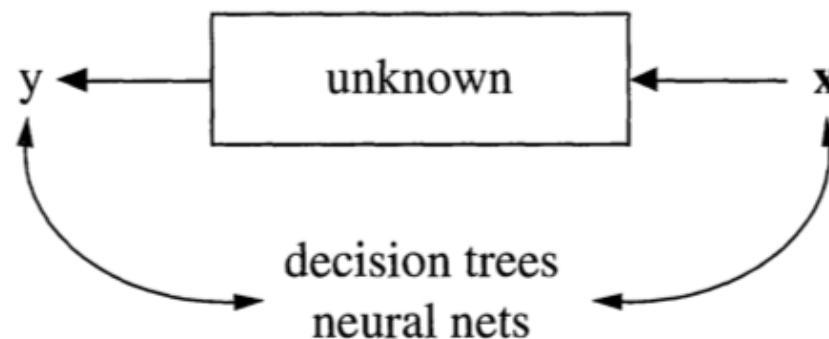
The goal is to find algorithm that accurately predicts y .

Theory: Data Generation Process

Algorithmic Modeling

The analysis in this approach considers the inside of the box complex and unknown. Their approach is to find a function $f(x)$ -an algorithm that operates on x to predict the responses y .

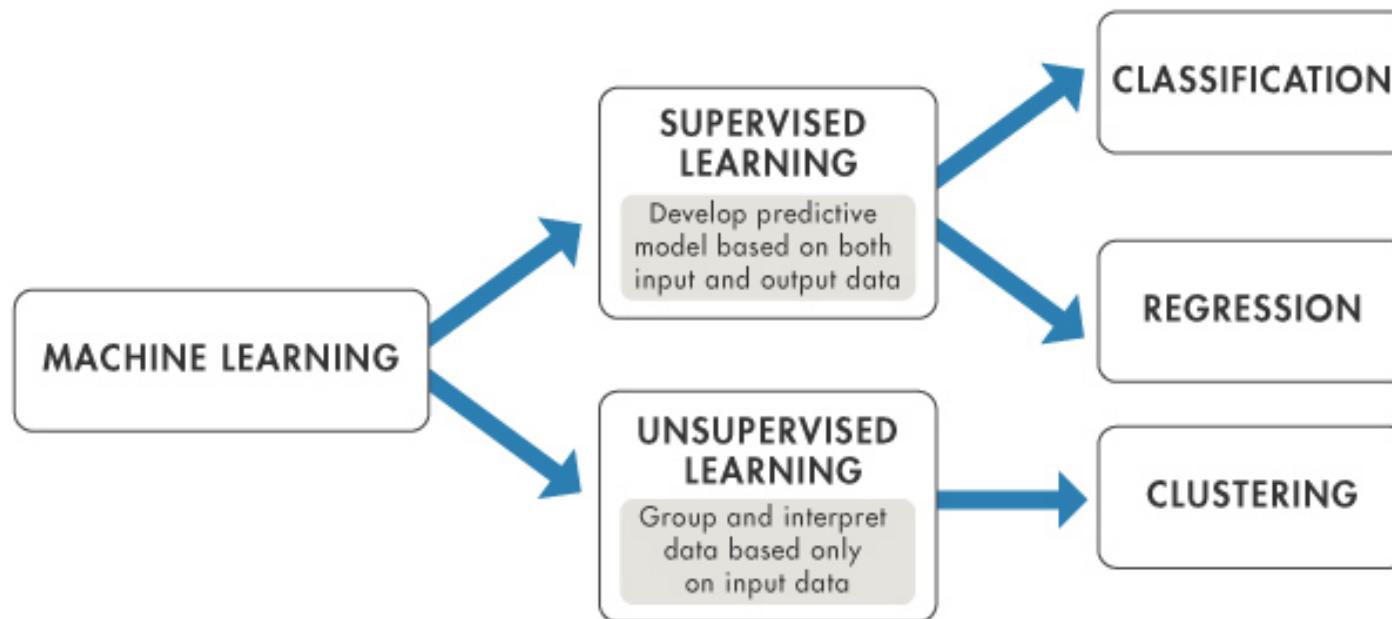
The goal is to find algorithm that accurately predicts y .



Theory: Data Generation Process Algorithmic Modeling

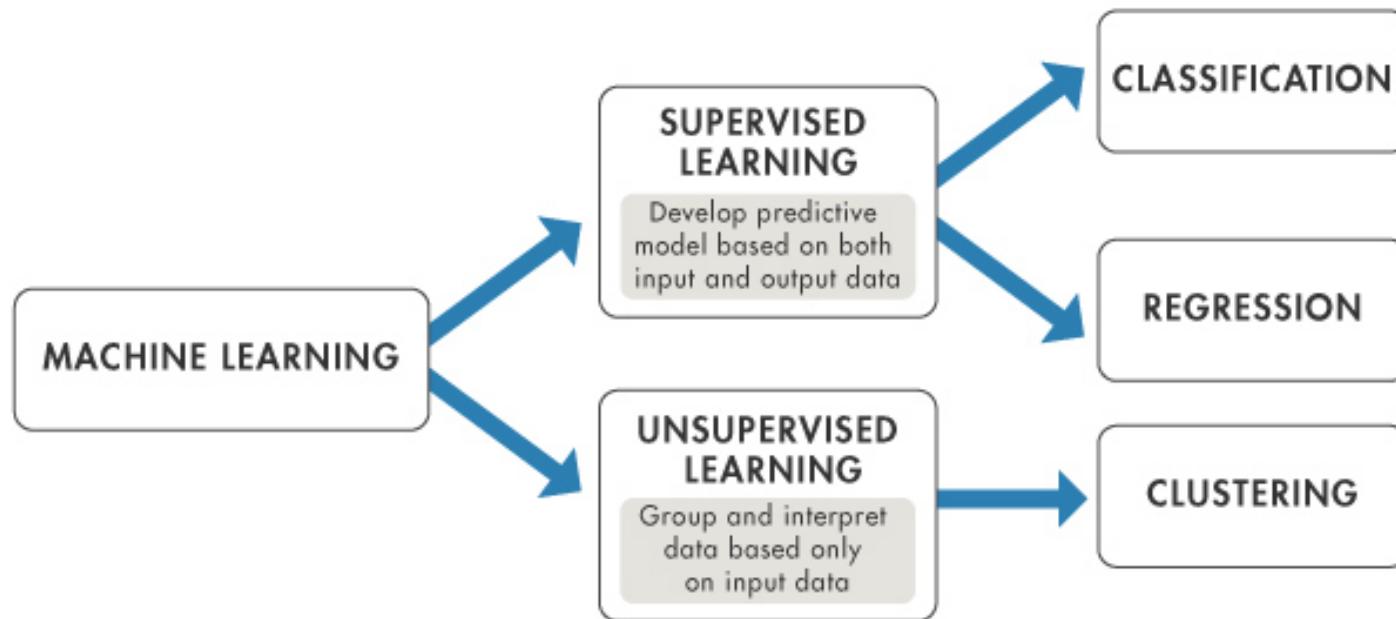
Theory: Data Generation Process

Algorithmic Modeling



Theory: Data Generation Process

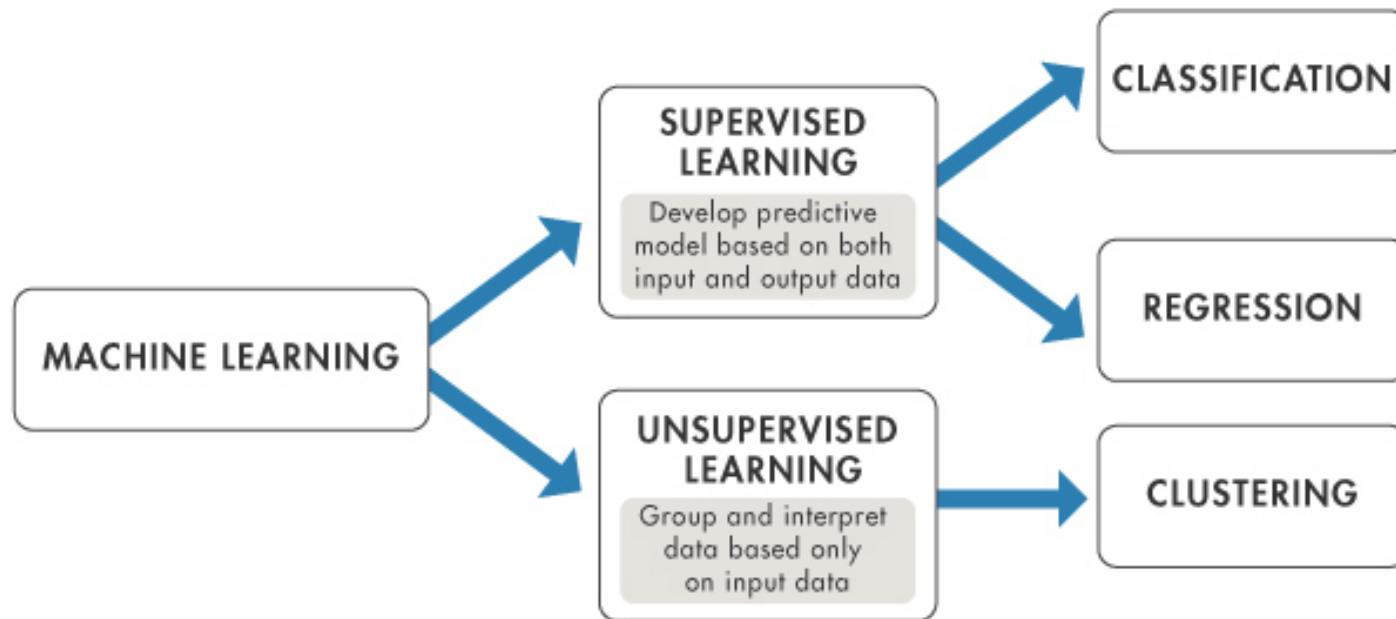
Algorithmic Modeling



Supervised Learning vs. Unsupervised Learning

Theory: Data Generation Process

Algorithmic Modeling



Supervised Learning vs. Unsupervised Learning

Source: <https://www.mathworks.com>

Algorithm and Inference

Very broadly speaking, algorithms are what statisticians do while inference says why they do them.

- Efron and Hastie 2017

Machine Learning

Machine Learning

Machine Learning

Machine can excel at frequent and high-volume task, at much faster rate and lower variance.

Machine Learning

Machine can excel at frequent and high-volume task, at much faster rate and lower variance.

Human can tackle noble situations.

Machine Learning

Machine Learning

Machine Learning

A Random Forest is an ensemble learning method that grows multivalued Decision Trees in different training sets.

Machine Learning

A Random Forest is an ensemble learning method that grows multivalued Decision Trees in different training sets.

To classify a new sample, input parameters are given to each tree in the forest to classify outcomes by taking the majority vote over all the trees in the forest.

Machine Learning

A Random Forest is an ensemble learning method that grows multivalued Decision Trees in different training sets.

To classify a new sample, input parameters are given to each tree in the forest to classify outcomes by taking the majority vote over all the trees in the forest.

The Random Forest is a typical learning model with the goal of reducing variance.

Illustration: HKES

Illustration: HKES

Research Question:

Illustration: HKES

Research Question:

What concerns Hong Kong people most?

Illustration: HKES

Research Question:

What concerns Hong Kong people most?

- Choice-based conjoint analysis**

Illustration: HKES

Research Question:

What concerns Hong Kong people most?

- Choice-based conjoint analysis
- Survey respondents were asked to choose between two social/political reform proposals

Illustration: HKES

Research Question:

What concerns Hong Kong people most?

- Choice-based conjoint analysis
- Survey respondents were asked to choose between two social/political reform proposals
- Each proposal consists of a set of reform items, representing three big concerns: (1) procedural democracy; (2) welfare benefits; (3) integration with

Illustration: HKES

Illustration: HKES

- Value of each reform item is randomly drawn

Illustration: HKES

- Value of each reform item is randomly drawn
- Advantages of the conjoint design

Illustration: HKES

- Value of each reform item is randomly drawn
- Advantages of the conjoint design
- Respondents need not report preferences for individual items, thereby lowering the risk of preference falsification

Illustration: HKES

- Value of each reform item is randomly drawn
- Advantages of the conjoint design
- Respondents need not report preferences for individual items, thereby lowering the risk of preference falsification
- Put all big theories in a single decision, so that we can rank order respondents' major concern

Illustration: HKES

Illustration: HKES

YouGov

Pair 1 of 4

| Reform Item | Proposal 1 | Proposal 2 |
|--|--|--|
| Constitutional Development | | |
| Percentage of Directly Elected Seats in the Legislative Council | Change from the current 50% to 100% | Change from the current 50% to 25% |
| Chief Executive Election: Nomination Method | Change the status quo by giving people outside of the Election Committee the power to nominate | Change the status quo by giving people outside of the Election Committee the power to nominate |
| Chief Executive Election: Election Method | Change the status quo by implementing universal suffrage | Change the status quo by implementing universal suffrage |
| Social Development | | |
| Mainland immigrants daily quota | Change from the current quota of 150 to 100 | Change from the current quota of 150 to 200 |
| National Education in Primary and Secondary Schools | Maintain the status quo, where there is no national education | Change the status quo by introducing national education |
| Public Housing (e.g. Public Rental Estates and Housing Ownership Scheme) as a Share of Total Housing | Change the current percentage at around 45% to 30% | Change the current percentage at around 45% to 60% |
| Operating Online and Printed Media | Change the status quo by requiring government approval | Maintain the status quo, where no government approval is required |
| Economic Benefits from Mainland China | Hong Kong is given a piece of land from the Guangdong province for economic development | Increase the number of mainland visitors |

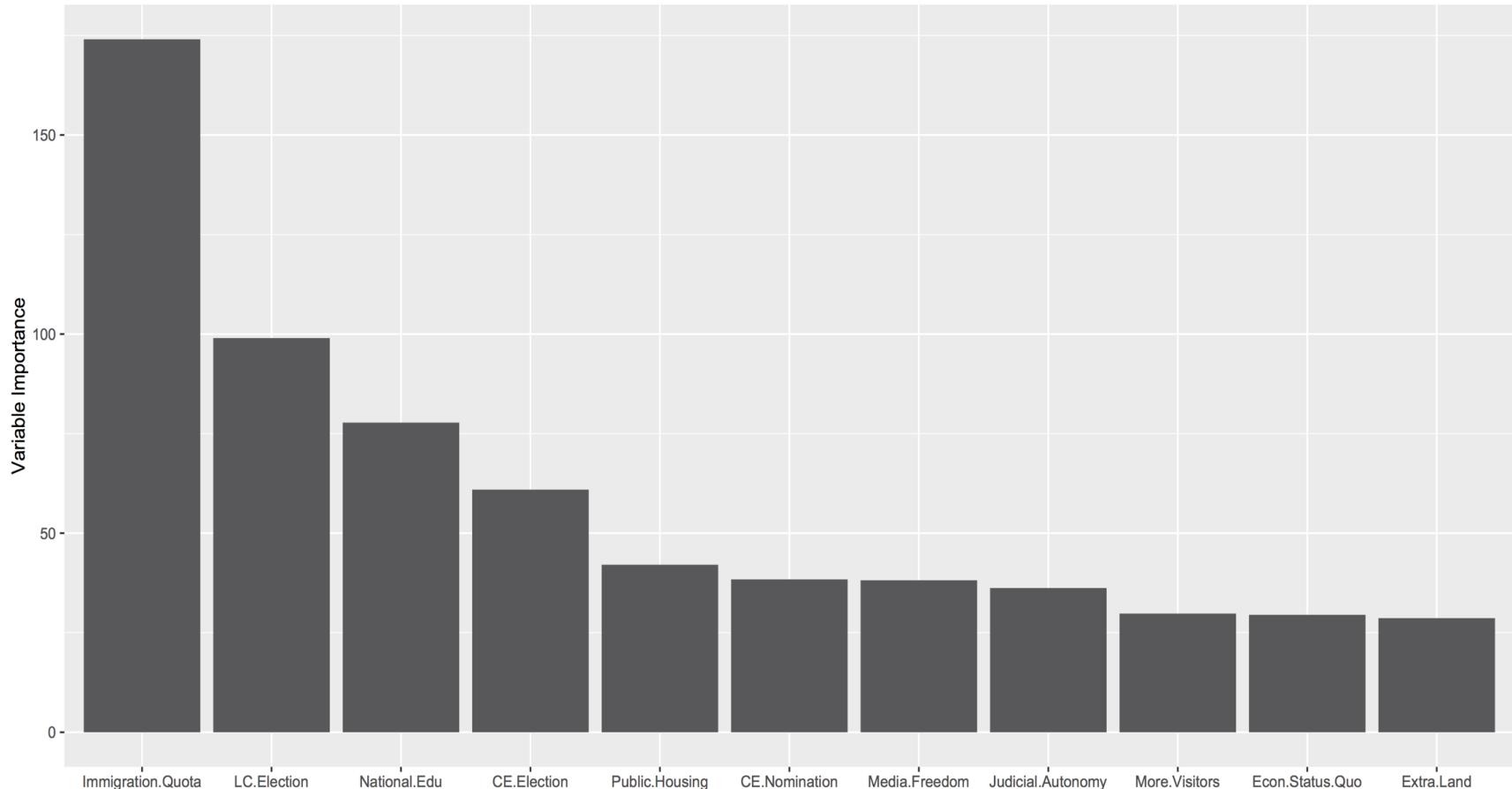
| <u>Reform Item</u> | <u>Proposal 1</u> | <u>Proposal 2</u> |
|---|--|--|
| Constitutional Development | | |
| Percentage of Directly Elected Seats in the Legislative Council | Change from the current 50% to 100% | Change from the current 50% to 25% |
| Chief Executive Election: Nomination Method | Change the status quo by giving people outside of the Election Committee the power to nominate | Change the status quo by giving people outside of the Election Committee the power to nominate |
| Chief Executive Election: Election Method | Change the status quo by implementing universal suffrage | Change the status quo by implementing universal suffrage |
| Social Development | | |
| Mainland immigrants daily quota | Change from the current quota of 150 to 100 | Change from the current quota of 150 to 200 |
| National Education in Primary and Secondary Schools | Maintain the status quo, where there is no national education | Change the status quo by introducing national education |
| Public Housing (e.g. Public Rental Estates and Housing Ownership Scheme) as a Share of Total Housing | Change the current percentage at around 45% to 30% | Change the current percentage at around 45% to 60% |
| Operating Online and Printed Media | Change the status quo by requiring government approval | Maintain the status quo, where no government approval is required |
| Economic Benefits from Mainland China | Hong Kong is given a piece of land from the Guangdong province for economic development | Increase the number of mainland visitors |
| Extradition of Hong Kong Citizens Who Committed Crimes in Mainland to | Change the status quo by introducing a mechanism for | Change the status quo by introducing a mechanism for |

| <u>Reform Item</u> | <u>Proposal 1</u> | <u>Proposal 2</u> |
|---|--|--|
| Constitutional Development | | |
| Percentage of Directly Elected Seats in the Legislative Council | Change from the current 50% to 100% | Change from the current 50% to 25% |
| Chief Executive Election: Nomination Method | Change the status quo by giving people outside of the Election Committee the power to nominate | Change the status quo by giving people outside of the Election Committee the power to nominate |
| Chief Executive Election: Election Method | Change the status quo by implementing universal suffrage | Change the status quo by implementing universal suffrage |
| Social Development | | |
| Mainland immigrants daily quota | Change from the current quota of 150 to 100 | Change from the current quota of 150 to 200 |
| National Education in Primary and Secondary Schools | Maintain the status quo, where there is no national education | Change the status quo by introducing national education |
| Public Housing (e.g. Public Rental Estates and Housing Ownership Scheme) as a Share of Total Housing | Change the current percentage at around 45% to 30% | Change the current percentage at around 45% to 60% |
| Operating Online and Printed Media | Change the status quo by requiring government approval | Maintain the status quo, where no government approval is required |
| Economic Benefits from Mainland China | Hong Kong is given a piece of land from the Guangdong province for economic development | Increase the number of mainland visitors |
| Extradition of Hong Kong Citizens Who Committed Crimes in Mainland to | Change the status quo by introducing a mechanism for | Change the status quo by introducing a mechanism for |

Illustration: HKES

Illustration: HKES

Variable Importance from Conjoint Analysis Using Random Forests



Machine Learning Caveat

Machine Learning Caveat

"when you present two systems to a company, a simple one with explanations that does ok, and a more complicated system that works better, every single time they will take the second. Every single time."

- Yann LeCun, Director of AI research at Facebook on Deep Learning

Machine Learning Caveat

"when you present two systems to a company, a simple one with explanations that does ok, and a more complicated system that works better, every single time they will take the second. Every single time."

- Yann LeCun, Director of AI research at Facebook on Deep Learning

Does the job, with no explanation or theory?

Let the dataset change your mindset.

Let the dataset change your mindset.

- Hans Rosling

Hans Rosli

Swedish physician and
statistician

Hans Roslin

Swedish physician and statistician

- Founded Gapminder Foundation

Hans Roslin

Swedish physician and statistician

- Founded Gapminder Foundation
- Visualize historical data on public health and poverty

Hal Varian

Hal Varian



Hal Varian

Chief Economist,
Google

Professor of
Economics,
University of
California, Berkeley.

**Big Data: New
Tricks for
Econometrics
Machine Learning
and Econometrics**



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.

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

“The Three Sexy Skills of Data Geeks”: “... with the Age of Data upon us, those who can model, munge, and visually communicate data...

“The Three Sexy Skills of Data Geeks”: “... with the Age of Data upon us, those who can model, munge, and visually communicate data...

- Mike Driscoll

Data Science Roadmap

Data Science Roadmap

1. Introduction - Data theory

Data Science Roadmap

1. Introduction - Data theory
2. Data methods

Data Science Roadmap

1. Introduction - Data theory
2. Data methods
3. Statistics

Data Science Roadmap

1. Introduction - Data theory
2. Data methods
3. Statistics
4. Programming

Data Science Roadmap

1. Introduction - Data theory
2. Data methods
3. Statistics
4. Programming
5. Data Visualization

Data Science Roadmap

1. Introduction - Data theory
2. Data methods
3. Statistics
4. Programming
5. Data Visualization
6. Information Management

Data Science Roadmap

1. Introduction - Data theory
2. Data methods
3. Statistics
4. Programming
5. Data Visualization
6. Information Management
7. Data Curation

Data Science Roadmap

1. Introduction - Data theory
2. Data methods
3. Statistics
4. Programming
5. Data Visualization
6. Information Management
7. Data Curation
8. Spatial Models and Methods

Data Science Roadmap

1. Introduction - Data theory
2. Data methods
3. Statistics
4. Programming
5. Data Visualization
6. Information Management
7. Data Curation
8. Spatial Models and Methods
9. Machine Learning

Data Science Roadmap

1. Introduction - Data theory
2. Data methods
3. Statistics
4. Programming
5. Data Visualization
6. Information Management
7. Data Curation
8. Spatial Models and Methods
9. Machine Learning
10. NLP/Text mining

Data Science Roadmap

1. Introduction - Data theory
 1. Fundamentals
 1. Data concepts
 2. Data Generation Process (DGP)
 2. Algorithm-based vs. Data-based approaches
 3. Taxonomy

Data Science Roadmap

2. Data methods

Data Science Roadmap

- 2. Data methods
- 1. Passive data

Data Science Roadmap

- 2. Data methods
 - 1. Passive data
 - 2. Data at will

Data Science Roadmap

- 2. Data methods
 - 1. Passive data
 - 2. Data at will
 - 3. Qualitative data

Data Science Roadmap

- 2. Data methods
 - 1. Passive data
 - 2. Data at will
 - 3. Qualitative data
 - 4. Complex data

Data Science Roadmap

- 2. Data methods
 - 1. Passive data
 - 2. Data at will
 - 3. Qualitative data
 - 4. Complex data
 - 5. Text data

Data Science Roadmap

3. Statistics

Data Science Roadmap

- 3. Statistics
 - 1. Sample and Population
 - 2. Inference
 - 3. Size and power
 - 4. Representation

Data Science Roadmap

4. Programming

Data Science Roadmap

4. Programming

1. R

Data Science Roadmap

4. Programming

1. R

2. Python

Data Science Roadmap

4. Programming

1. **R**

2. **Python**

3. **HTML**

Data Science Roadmap

4. Programming
 1. R
 2. Python
 3. HTML
 4. Java script

Data Science Roadmap

Data Science Roadmap

5. Data Visualization

1. Tableau

Data Science Roadmap

5. Data Visualization

1. Tableau
2. ggplot2

Data Science Roadmap

5. Data Visualization

1. **Tableau**
2. **ggplot2**
3. **Shiny**

Data Science Roadmap

5. Data Visualization

1. **Tableau**
2. **ggplot2**
3. **Shiny**
4. **D3.js**

Data Science Roadmap

5. Data Visualization

1. **Tableau**
2. **ggplot2**
3. **Shiny**
4. **D3.js**
5. **Animation**

Data Science Roadmap

Data Science Roadmap

6. Information Management

1. MapReduce
2. Hadoop
3. Cassandra
4. MongoDB
5. NoSQL

Data Science Roadmap

Data Science Roadmap

7. Data curation

1. Google OpenRefine
2. Sampling
3. Missing value concepts and management

Data Science Roadmap

8. Spatial Models and Methods

Data Science Roadmap

8. Spatial Models and Methods
 1. GIS

Data Science Roadmap

8. Spatial Models and Methods

1. **GIS**

2. **R/Leaflet**

Data Science Roadmap

8. Spatial Models and Methods

1. **GIS**
2. **R/Leaflet**
3. **Python Map**

Data Science Roadmap

8. Spatial Models and Methods

1. **GIS**
2. **R/Leaflet**
3. **Python Map**
4. **Remote Sensing**

Data Science Roadmap

9. Machine Learning

1. Supervised
2. Unsupervised
3. Regression methods
4. Neural Networks

Data Science Roadmap

10. NLP/Text Mining

Data Science Roadmap

10. NLP/Text Mining

1. Corpus

Data Science Roadmap

10. NLP/Text Mining

1. Corpus
2. Text Analysis

Data Science Roadmap

10. NLP/Text Mining

1. Corpus
2. Text Analysis
3. Sentiment Analysis

Data Science Roadmap

10. NLP/Text Mining

1. Corpus
2. Text Analysis
3. Sentiment Analysis
4. Natural Language Processing

Data Literacy

Data Literacy

1. Data generating process

Data Literacy

1. Data generating process
2. Graphic grammar

Data Literacy

1. Data generating process
2. Graphic grammar
3. Statistical judgement

Data Literacy

Data Literacy

1. Data generating process

Data Literacy

1. Data generating process
 1. How data are generated

Data Literacy

1. Data generating process
 1. How data are generated
 2. Distribution

Data Literacy

1. Data generating process
 1. How data are generated
 2. Distribution
 3. Missing values

Data Literacy

1. Data generating process
 1. How data are generated
 2. Distribution
 3. Missing values
 4. Wrong data

Data Literacy

Data Literacy

2. Graphic grammar

Data Literacy

2. Graphic grammar
 1. Bad charts deliver incorrect message

Data Literacy

- 2. Graphic grammar
 - 1. Bad charts deliver incorrect message
 - 2. Poor design

Data Literacy

2. Graphic grammar
 1. Bad charts deliver incorrect message
 2. Poor design
 3. Color

Data Literacy

2. Graphic grammar
 1. Bad charts deliver incorrect message
 2. Poor design
 3. Color
 4. Label

Data Literacy

2. Graphic grammar
 1. Bad charts deliver incorrect message
 2. Poor design
 3. Color
 4. Label
 5. Scale

Data Literacy

Data Literacy

3. Statistical understanding

Data Literacy

3. Statistical understanding
 1. Size does (not) matter

Data Literacy

3. Statistical understanding
 1. Size does (not) matter
 2. Representativeness does

Data Literacy

3. Statistical understanding
 1. Size does (not) matter
 2. Representativeness does
 3. Forecast/prediction minded

Data Literacy

3. Statistical understanding
 1. Size does (not) matter
 2. Representativeness does
 3. Forecast/prediction minded
 4. Explanation

Data Literacy

1. Why we need numeric data?
2. History of data

Darkest hour: Churchill and typist

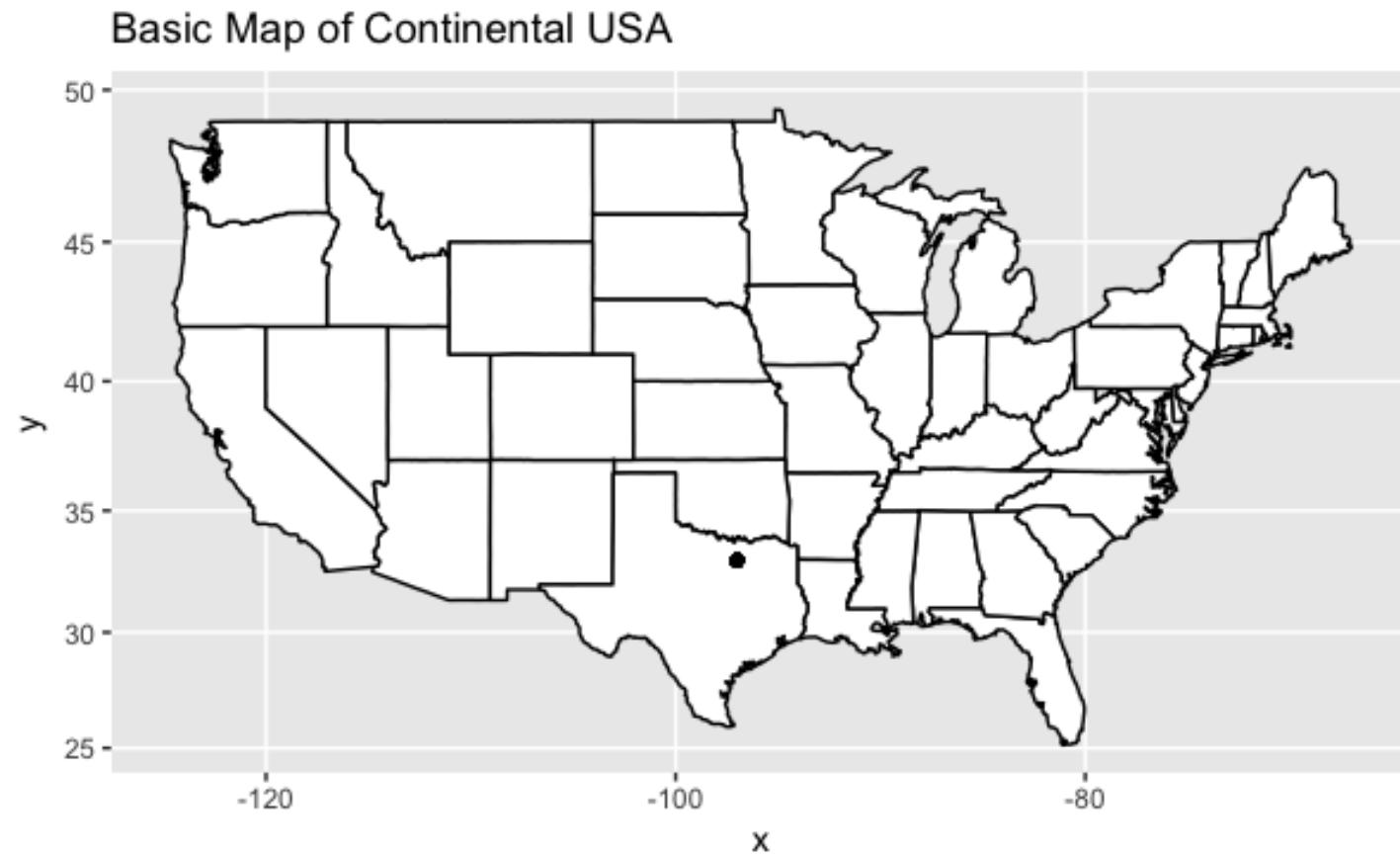
Darkest hour: Churchill and typist



- Data Thinking
- Multi-disciplinary Thinking
- Machine Thinking

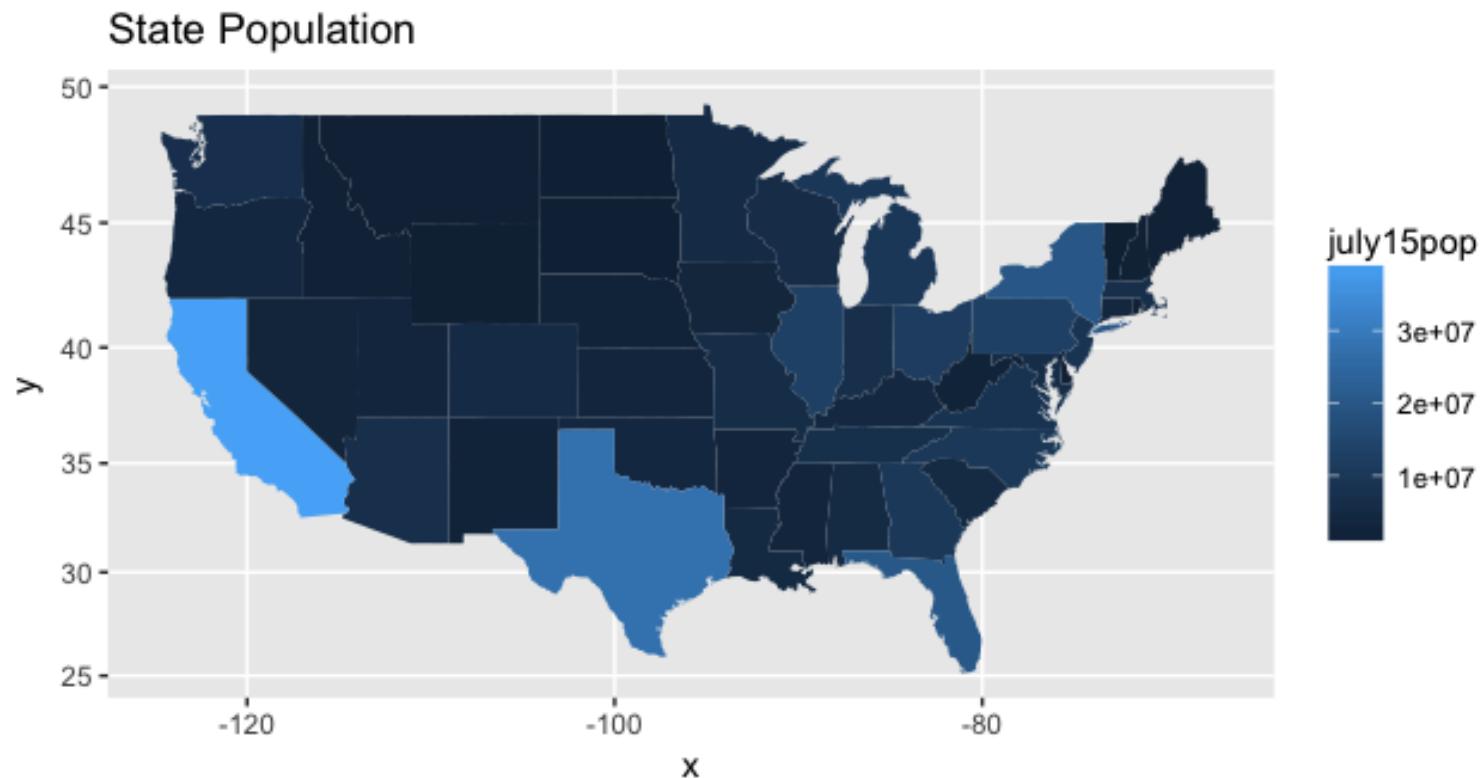
Spatial Data: United States

Spatial Data: United States



Spatial Data: United States

Spatial Data: United States



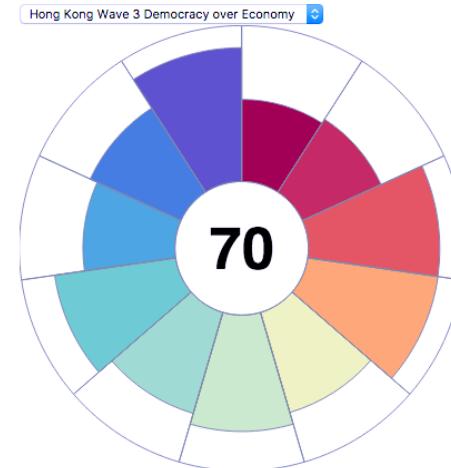
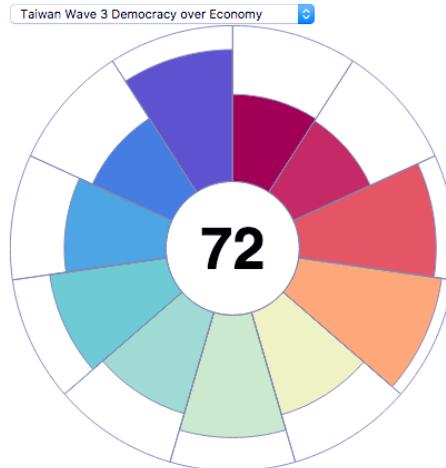
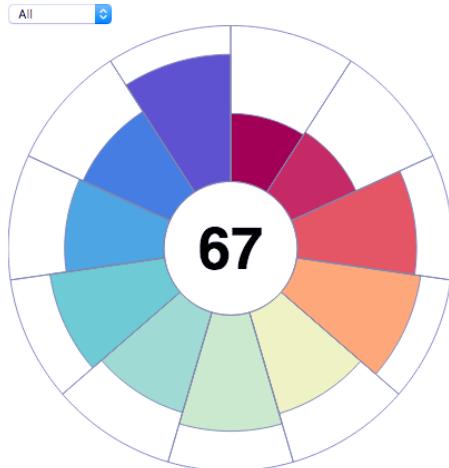
Java: D3 Library

Java: D3 Library

Latent Profile Models:

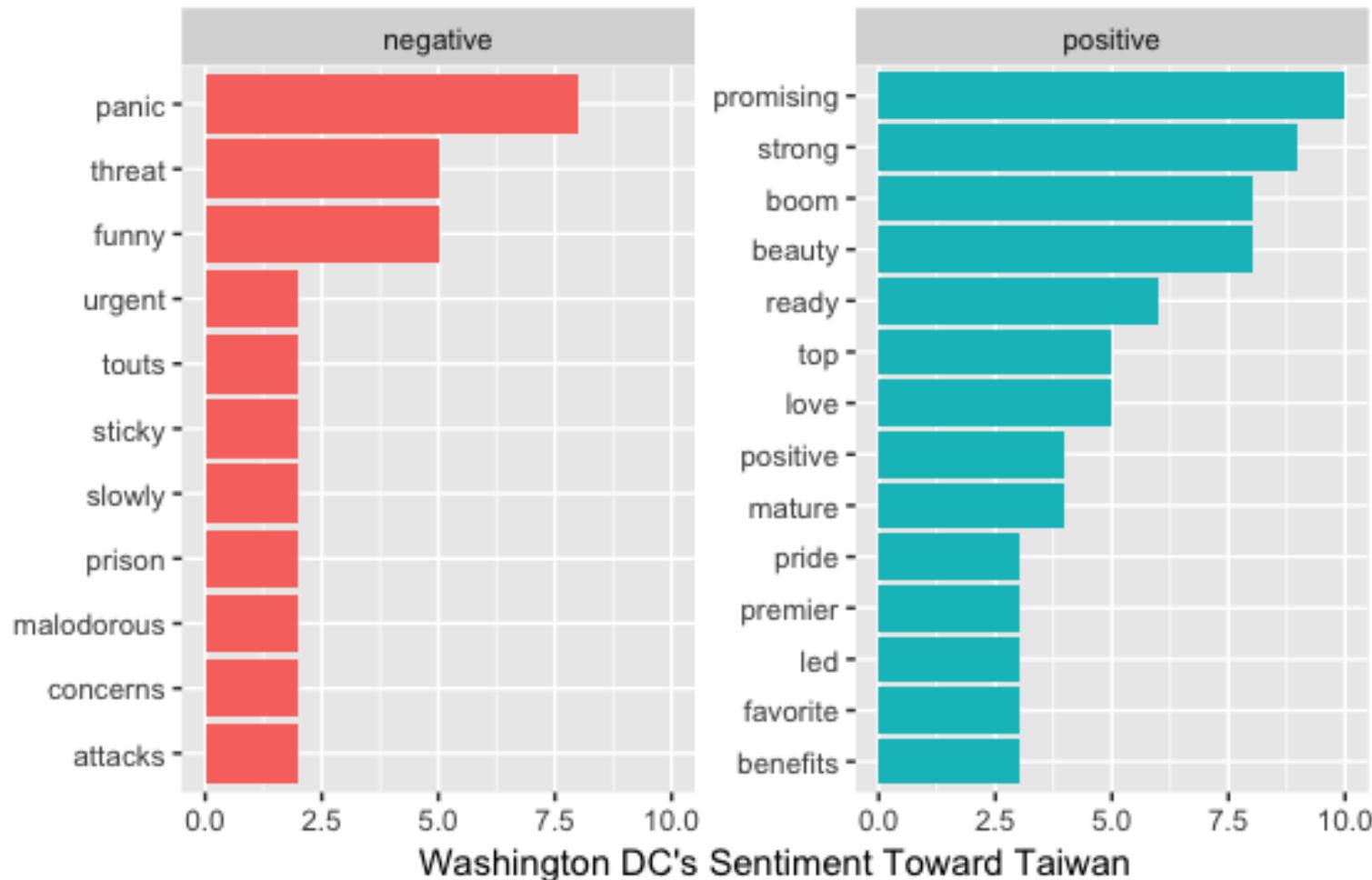
Hong Kong and Taiwan

Wave 3 (2010, 2012) | Wave 4 (2014, 2016)



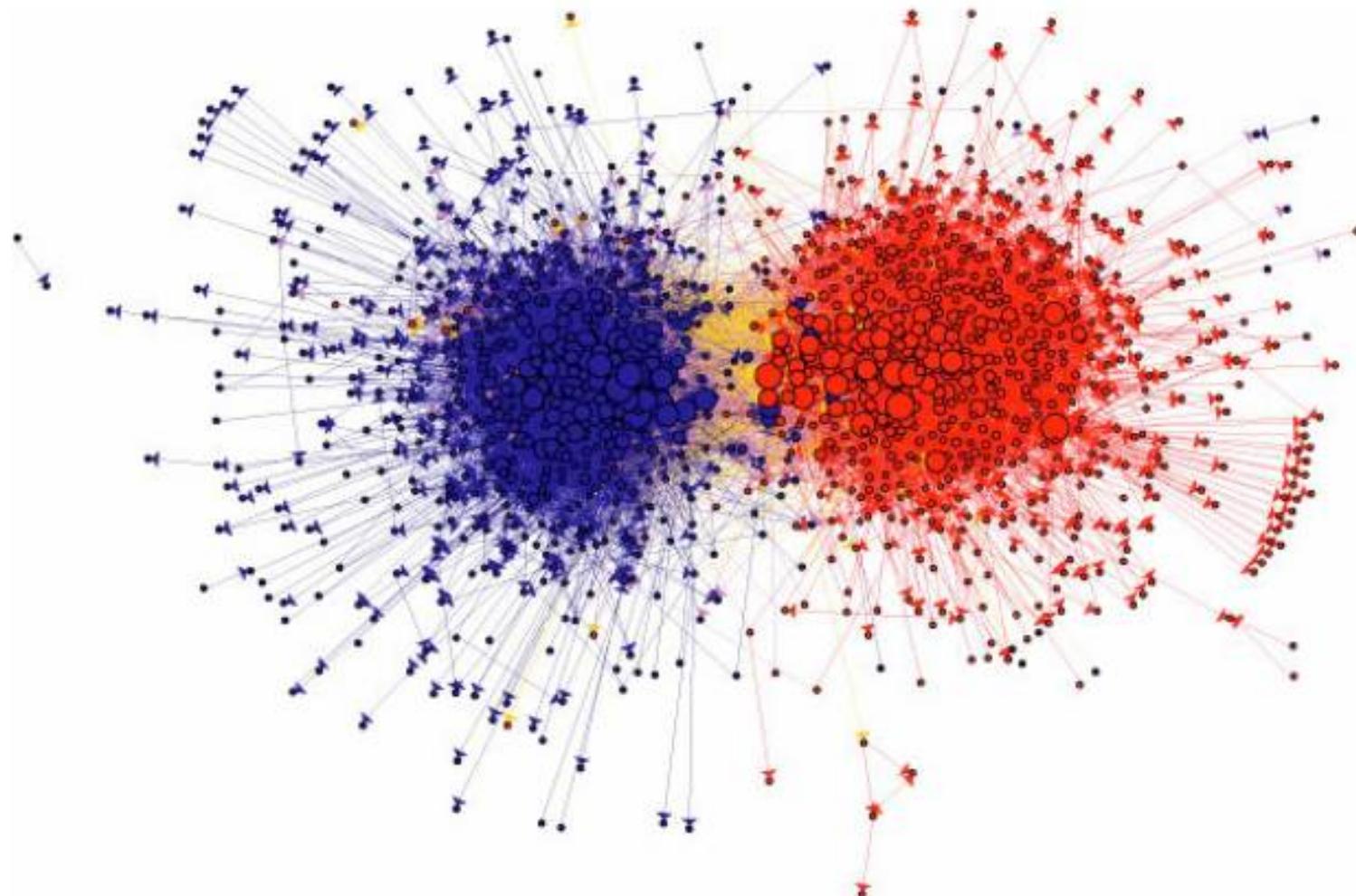
Sentiment Analysis

Sentiment Analysis

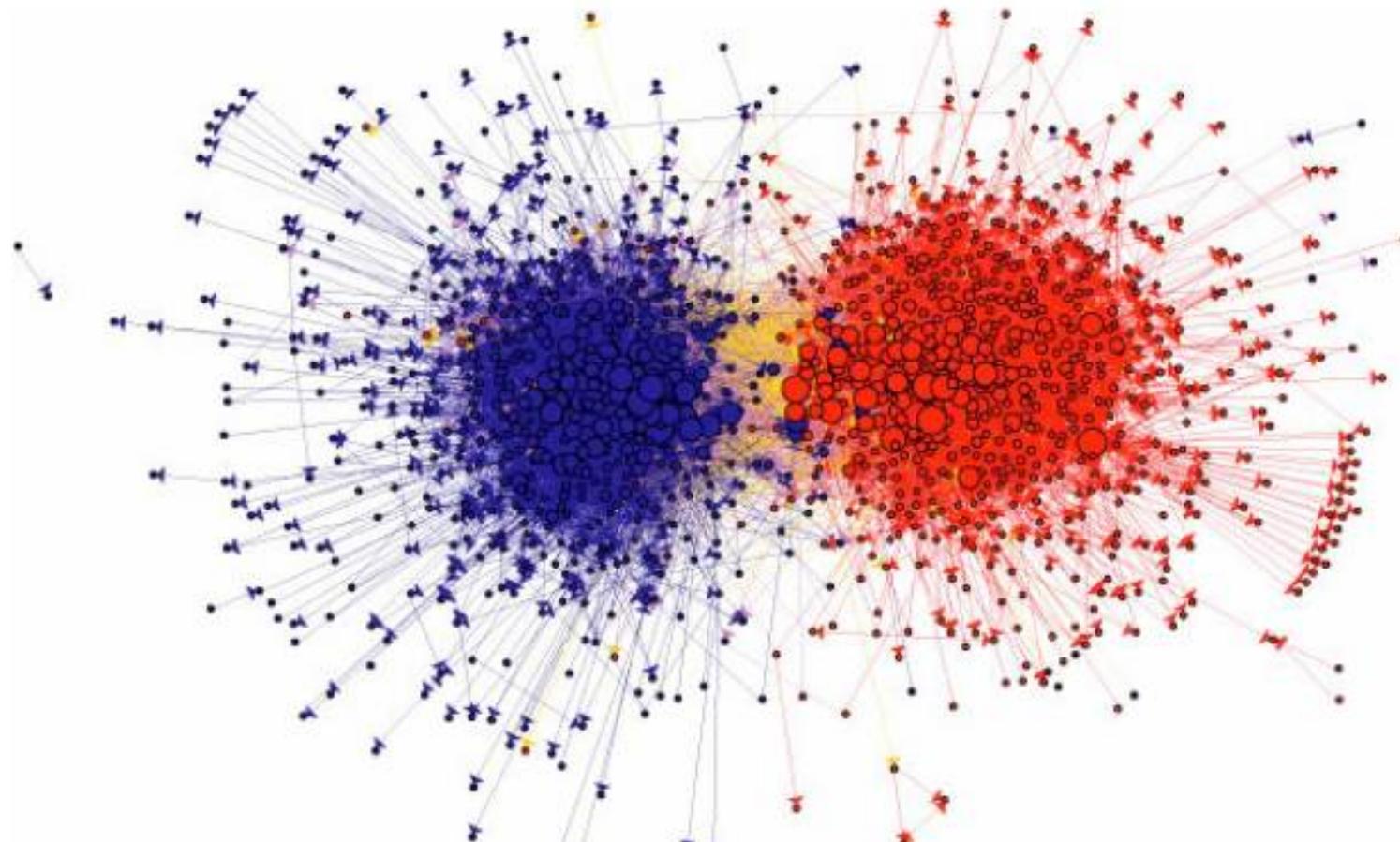


Lazer et al. 2009 Life in the network

Lazer et al. 2009 Life in the network



Lazer et al. 2009 Life in the network



This figure summarizes the link structure within a community of political blogs (from 2004), where red nodes indicate conservative blogs, and blue liberal. Orange links go from liberal to conservative, and purple ones from conservative to liberal. The size of each blog reflects the number of other blogs that link to it

Cumulated/Repeated Data

Cumulated/Repeated Data

