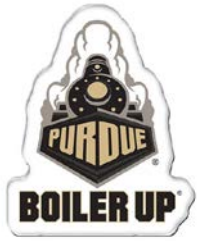


BIG DATA CONCEPTS

Sang Eun Woo

Purdue University

CARMA Webcast Lecture

November 8, 2019

PURDUE
UNIVERSITY

Background

- **Openness to New Experience (e.g., ideas, methods, changing nature of work)**

- 6th Purdue Symposium on Psychological Sciences on Big Data for Psychological Sciences (May 17–19, 2018)

→ An edited volume for APA Books (Woo, Tay, & Proctor): "**Big Data in Psychological Research**" (*forthcoming – 2020*)

Goals

Review contemporary uses of the "big data" concept

- Big data *phenomenon*
- Big data *applications*
- Big data ***method***
- Big data *science*

- **Show that "big data" is a topic worthy of scientific endeavors**
- **Identify key areas of opportunities and challenges for future big data research method in psychology (and related fields)**

Big data as a *phenomenon*

- **Mid-1990's**
  - John Mashey at Silicon Graphics

The New York Times

Bits
Business, Innovation, Technology, Society

The Origins of 'Big Data': An Etymological Detective Story

By Steve Lohr February 1, 2013 9:10 am

The term Big Data is so generic that the hunt for its origin was not just an effort to find an early reference to those two words being used together. Instead, the goal was the early use of the term that suggests its present connotation — that is, not just a lot of data, but different types of data handled in new ways.

When I called Mr. Mashey recently, he said that Big Data is such a simple term, it's not much a claim to fame. His role, if any, he said, was to popularize the term within a portion of the high-tech community in the 1990s. "I was using one label for a range of issues, and I wanted the simplest, shortest phrase to convey that the boundaries of computing keep advancing," said Mr. Mashey, a consultant to tech companies and a trustee of the Computer History Museum in Mountain View, Calif.

Lohr (2013)

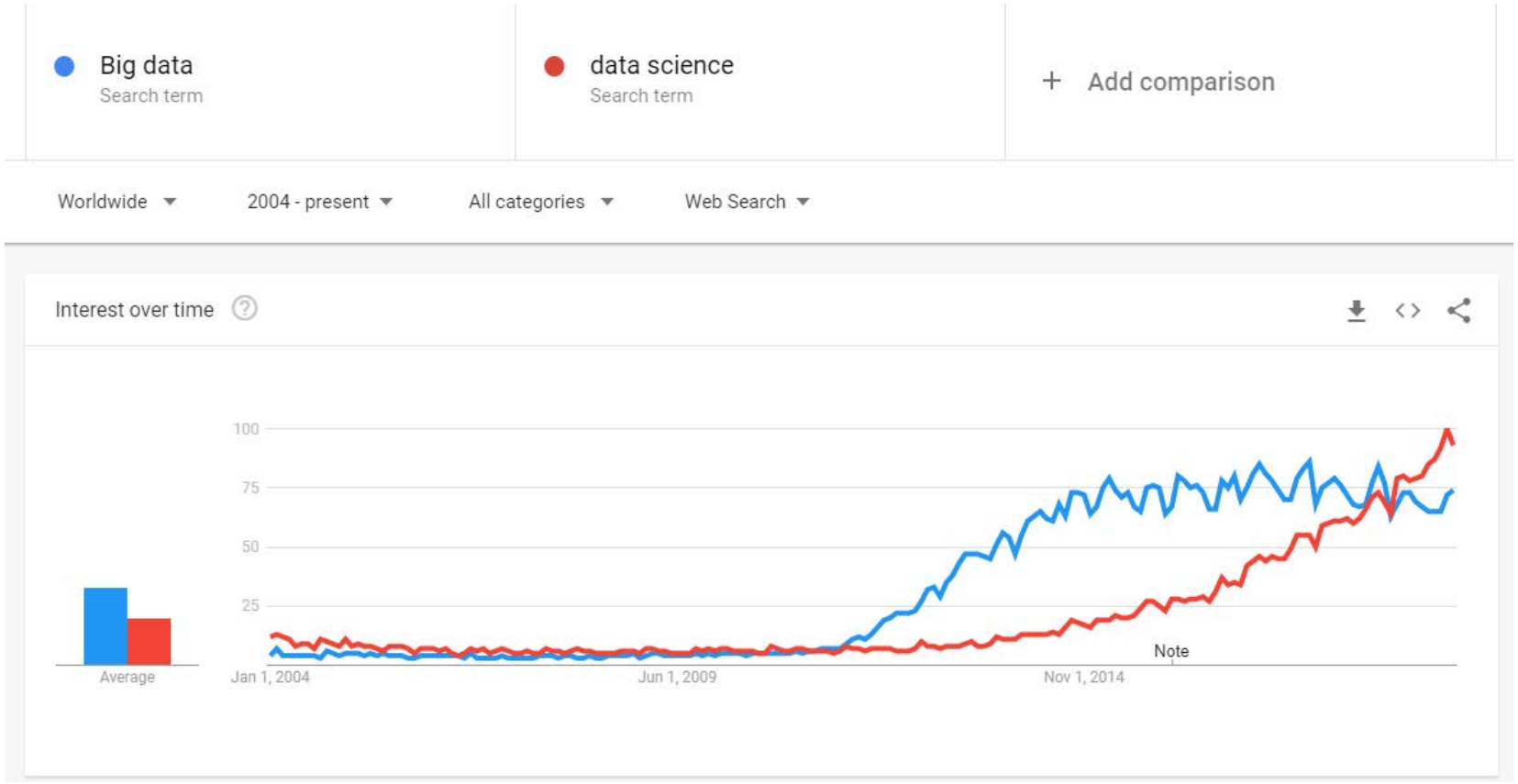
Big data as a *phenomenon*

- **Early 2000's**
  - Doug Laney (3 V's: **Volume, Variety, Velocity**)
  - Another V: *Veracity* (let's come back to this later)
  - More V-words added later on: value, variability, exhaustivity, versatility, volatility, vitality, virtuosity, visionary … virility… valueless, vampire-like, venomous, vulgar…. (Uprichard, 2013)
  - And then P-words… portentous, perverse, personal, productive, …. provocative… playful (Lupton, 2015)

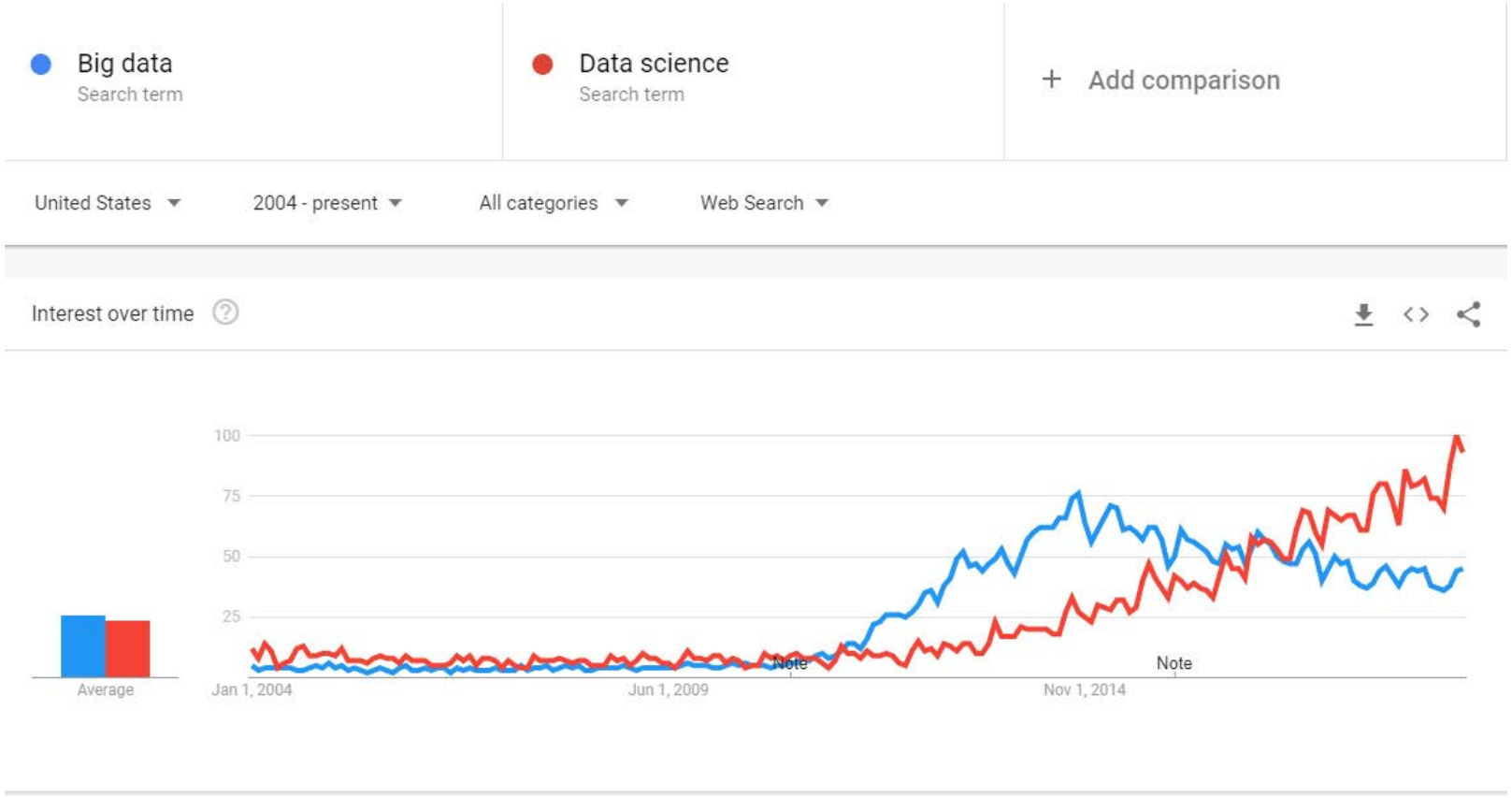
- Kitchin's (2013) 7 ontological traits of big data
  - Volume, Velocity, Variety, Exhaustivity, Resolution & Indexicality, Relationality, Extensionality & Scalability
  - Do these traits apply to all types of big data (e.g., mobile phone data, web searches, social media, digital CCTV, supermarket scanner and sales, credit card, flight movement, credit card, stock market trades, house price register)? Not really.

- **Where are we now?**

Big data as a *phenomenon*



Big data as a *phenomenon*



Article

The Double-Edged Sword of Big Data in Organizational and Management Research: A Review of Opportunities and Risks

Ramon Wenzel[1] and Niels Van Quaquebeke[2]

Organizational Research Methods
2018, Vol. 21(3) 548-591

DOI: 10.1177/1094428117718627
journals.sagepub.com/home/orm
SAGE

Industrial and Organizational Psychology, 8(4), pp 491–508 December 2015.
 doi:10.1017/iop.2015.40

Focal Article

Big Data Recommendations for Industrial–Organizational Psychology

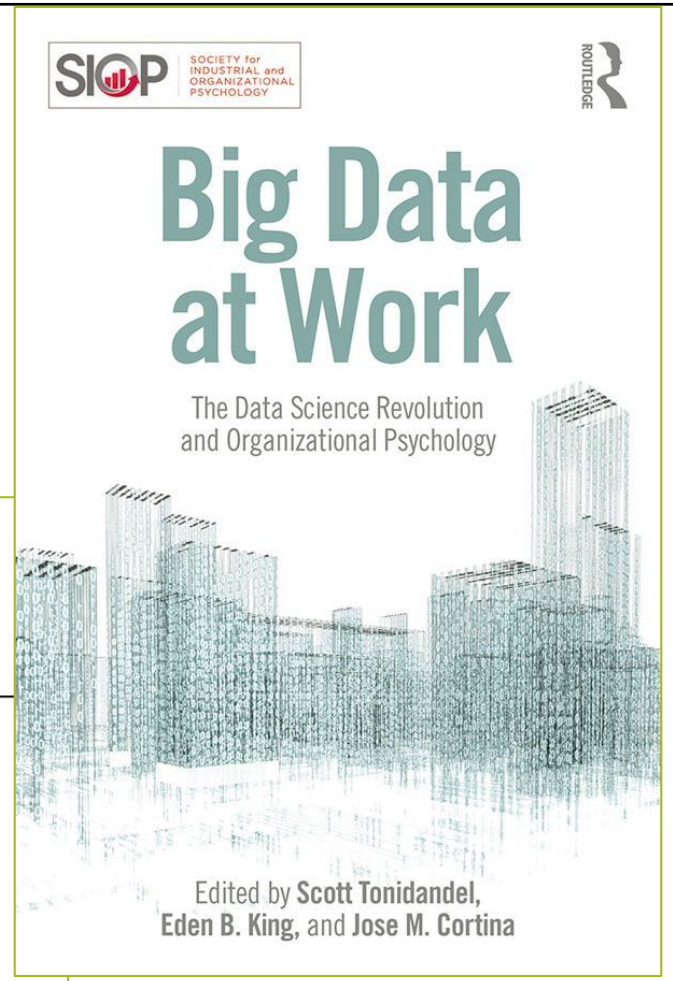
Richard A. Guzzo
Mercer, Washington, DC

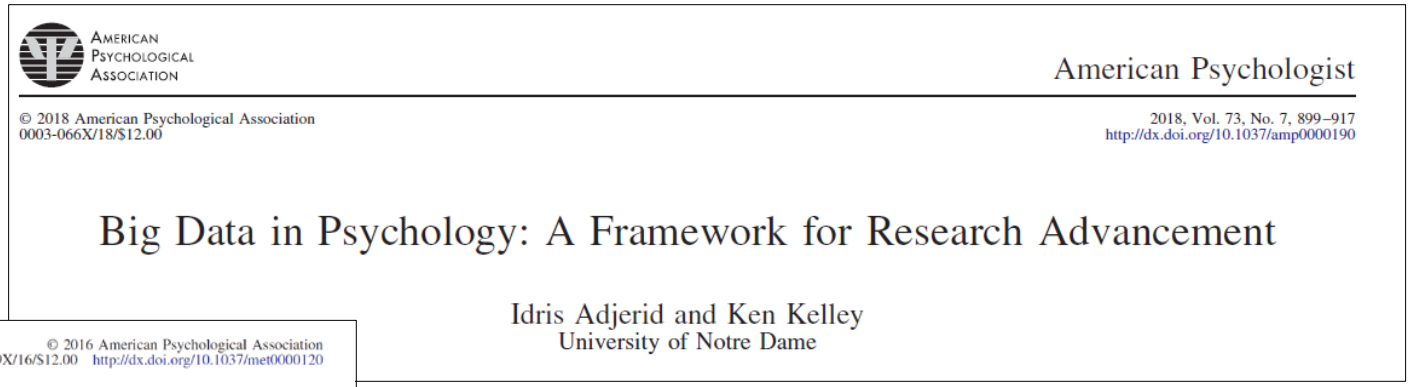
Alexis A. Fink
Intel, Portland, Oregon

Eden King
George Mason University

Scott Tonidandel
Davidson College

Ronald S. Landis
Illinois Institute of Technology





American Psychological Association

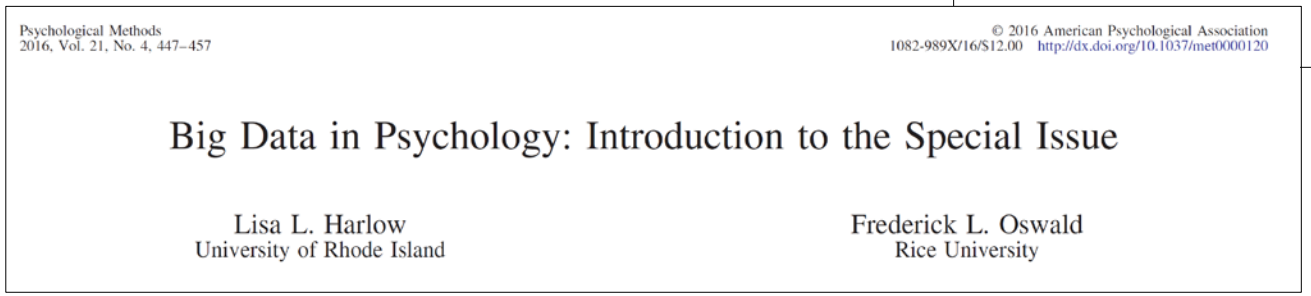
American Psychologist

© 2018 American Psychological Association
0003-066X/18/$12.00

2018, Vol. 73, No. 7, 899–917
http://dx.doi.org/10.1037/amp0000190

Big Data in Psychology: A Framework for Research Advancement

Idris Adjerid and Ken Kelley
University of Notre Dame



Psychological Methods
2016, Vol. 21, No. 4, 447–457

© 2016 American Psychological Association
1082-989X/16/$12.00 http://dx.doi.org/10.1037/met0000120

Big Data in Psychology: Introduction to the Special Issue

Lisa L. Harlow
University of Rhode Island

Frederick L. Oswald
Rice University

"**A Practical Guide to Big Data Research in Psychology**" by Eric Evan Chen and Sean P. Wojcik

"**A Primer on Theory-Driven Web Scraping: Automatic Extraction of Big Data from the Internet for Use in Psychological Research**" by Richard N. Landers, Robert C. Brusso, Katelyn J. Cavanaugh, and Andrew B. Collmus

"**Mining Big Data to Extract Patterns and Predict Real-Life Outcomes**" by Michal Kosinski, Yilun Wang, Himabindu Lakkaraju, and Jure Leskovec

"**Gaining Insights from Social Media Language: Methodologies and Challenges**" by Margaret L. Kern, Gregory Park, Johannes C. Eichstaedt, H. Andrew Schwartz, Maarten Sap, Laura K. Smith, and Lyle H. Ungar

"**Tweeting Negative Emotion: An Investigation of Twitter Data in the Aftermath of Violence on College Campuses**" by Nickolas M. Jones, Sean P. Wojcik, Josiah Sweeting, and Roxane Cohen Silver

"**Comparing Vector-Based and Bayesian Memory Models Using Large-Scale Datasets: User-Generated Hashtag and Tag Prediction on Twitter and Stack Overflow**" by Clayton Stanley and Michael D. Byrne

"**Theory-Guided Exploration with Structural Equation Model Forests**" by Andreas M. Brandmaier, John J. Prindle, John J. McArdle, and Ulman Lindenberger

"**Statistical Learning Theory for High Dimensional Prediction: Application to Criterion-Keyed Scale Development**" by Benjamin P. Chapman, Alexander Weiss, and Paul Duberstein

"**Finding Structure in Data Using Multivariate Tree Boosting**" by Patrick J. Miller, Gitta H. Lubke, Daniel B. McArtor, and C. S. Bergeman

"**Partial Least Squares Correspondence Analysis: A Framework to Simultaneously Analyze Behavioral and Genetic Data**" by Darek Beaton, Joseph Dunlop, and Hervé Abdi

Big Data in Psychological Research

Woo, Tay, & Proctor, forthcoming (hopefully by Summer 2020)

Table of Content

Big data as *applications* (people/talent analytics)

Promising directions

- Team assessment and training/intervention using wearable sensors
- Job analysis using natural language processing
- Data visualization → Human-guided visual analytics

Exciting but tricky directions

- Targeted recruitment using online job seeker data
- Automatic (algorithm-based) scoring of selection interviews
- Electronic performance monitoring
- Pulse surveys to track employee well-being and health outcomes
- Turnover management and intervention using attrition modeling

Practical & ethical concerns:

- Privacy
- Fairness/biases

Bauer et al. (2020); Guzzo et al. (2015); Tonidandel, King, & Cortina (2016), Song et al. (2020)

Big data *methods* (for psychology)

- "big data" for psychological research [...] broadly refers to *multiplying multiform data (e.g., structured, unstructured) and its supporting technological infrastructure (i.e., capture, storage, processing) and analytic techniques* that can enhance psychological research.

  (Woo, Tay, & Proctor, 2020)

- ***Data types (sources):*** Mobile communication, websites, social media/crowdsourcing, sensors, network cameras, etc.
- ***Technological infrastructure:*** Cloud-based technology, etc.
- ***Analytical approaches:*** Machine learning algorithms, visual analytics, etc.

Requires careful interdisciplinary collaborations

Big data *methods* (for psychology)

Adjerid & Kelley (2018). "Big data in psychology: A framework for research advancement." *American Psychologist*

- ***sample size (n), variables (v), time (t)***
  - *N, v, T (traditional research expanded)*
  - *n, V, T (small sample big data research)*
  - *N, V, t (small snapshot big data research)*
  - *N, V, T (idealized big data research)*

Big data *methods*

- **Two big issues**

1. Concerns and uncertainties about scientific value (*where is theory?*)
2. Measurement validity (*what are we really measuring?*)

"This is a world where massive amounts of data and applied mathematics replace every other tool that might be brought to bear. Out with every theory of human behavior, from linguistics to sociology. Forget taxonomy, ontology, and psychology. Who knows why people do what they do? The point is they do it, and we can track and measure it with unprecedented fidelity. With enough data, the numbers speak for themselves."

Chris Anderson, Editor-in-Chief of Wired (2008)

"Quantity of data does not mean that one can ignore foundational issues of measurement and construct validity and reliability and dependencies among data. The core challenge is that most big data that have received popular attention are not the output of instruments designed to produce valid and reliable data amenable for scientific analysis."

David Lazer and colleagues, Science (2014)

Is big data *atheoretical?*

- Some big data applications may be deemed atheoretical, if they are not informed by prior theory and/or do not lead to a theoretical explanation for the phenomenon….. **"dustbowl empiricism"**
- But, **big data as a research method is always (intentionally) theoretical!**
  - Informed, guided, or driven by theory – a lot of flexibility here
  - Used for subsequent theoretical advancement – this should always happen

Theory-informed vs. theory-informing big data research

- **Data Collection**: theory *informs*, *guides*, or *determines* what types of data to collect

- **Data Processing**: theoretical insights and interests are often incorporated into wrangling, munging, tidying, transforming & mapping, and visualizing data

- **Analysis and interpretation**:

Two cultures in statistics: data modeling vs. algorithmic modeling (Breiman, 2001)

  - The former focuses on theory-based model/parameter specification and evaluation (traditional inferential statistics within psych); the latter focuses on maximizing prediction accuracy via cross-validation of algorithms (originated from computer science – ML and data mining)

→ "modern algorithmic methods can be used alongside traditional data modeling approaches in a reciprocally informative manner" (Putka & Oswald 2016)

How exactly does this work?

Three methods of science

Deductive (Theory-Driven) Research

Deduction: Reaching a logical conclusion based on true premises

Start with a theory → develop testable hypotheses → data to disprove or support the theory (cannot ever "verify" the truth)

Inductive (Data-Driven) Research

Induction: Generalizing results beyond the observations at hand

Finding patterns in data from one sample → replicated in other data/samples → phenomenon detection

Abductive (Explanatory) Research

Abduction: Finding a feasible (or best) explanation for a phenomenon

Phenomenon detection → theory construction (generation – development – appraisal)

Haig (2020), Woo, O'Boyle, & Spector (2017)

Deductive research using big data

- **Data Collection**: theory should determine what types of data to collect to test a set of specific hypotheses
- **Data Processing**: the goal should be to derive theoretically meaningful variables for further analysis. This goal determines choices of wrangling, munging, tidying, transforming & mapping, and visualizing data
- **Analysis and interpretation**: use traditional inferential statistics to analyze the data, and interpret accordingly; when using algorithmic approach, evaluate the findings against the a priori hypotheses derived from theory

Highly restrictive in the way data are collected and used; the main focus is on testing existing theory, not on constructing a new one

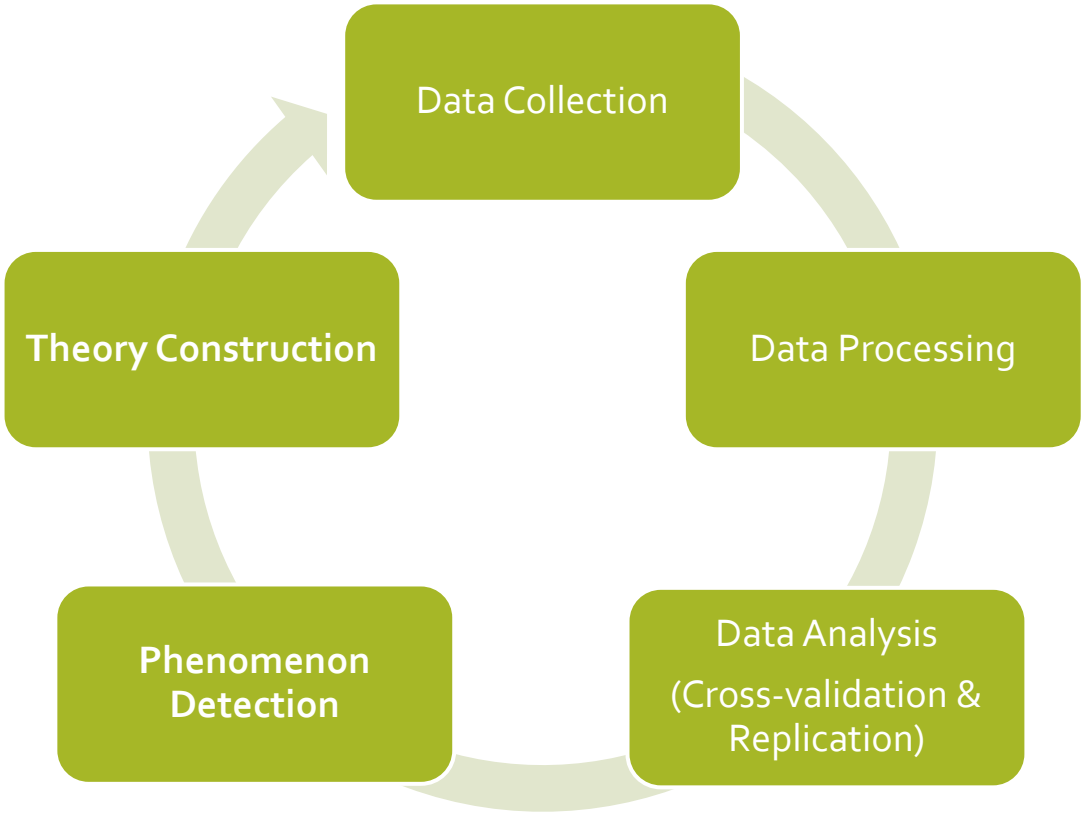
Inductive research using big data

- **Data Collection**: theory can inform what types of data to collect, but try to keep it loose and seek variety
- **Data Processing**: try out different methods of wrangling, munging, tidying, transforming & mapping, and visualizing data
- **Analysis and interpretation**: use algorithmic modeling approach to analyze the data and cross-validate the results – try different analytical methods to see if findings are replicated across methods

The main focus is on phenomenon detection, not on coming up with an explanation (i.e., theory)

McAbee, Landis, & Burke (2017); Woo, O'Boyle, & Spector (2017)

Abductive research using big data



The abductive sequence: phenomenon detection → explanation (i.e., theory)

Effectuation vs. Causation

Terminologies borrowed from the entrepreneurship literature…

- **Effectuation**: "Start with your means" "bird-in-hand" principle
- **Causation**: starting with a clear purpose, goals, or ends in mind

An Illustrative Example: Employee Turnover

Forthcoming IOP articles on attrition modeling

- Theories of turnover should inform practitioner's decisions in attrition modeling (Speer et al., in press)
  - Science-Practitioner model ("praxis") for SIOP
- Big data method such as attrition modeling can help better understand turnover-related phenomena (Woo, in press)
  - Big data *application* → big data *research*

What might big data turnover research look like?

- Example 1: **Profiles of Leavers and Stayers** (Woo & Allen, 2014)
  - Inductively derived four latent profiles of leavers and stayers
  - Two 'decent-sized' samples (Ns= 582 and 536)
- Example 2: **Resignation Styles** (Klotz & Bolino, 2016)
  - Seven resignation styles identified through grounded theory approach (qualitative)
- Example 3: **Pre-quitting Behaviors** (Gardner, Van Iddekinge, & Hom, 2018)
  - Introduced an initial set of observable behaviors shown by soon-to-be leavers

Incorporate bigger *n*, *v*, and *t* for further establishing the phenomenon and/or refining the new theory

Woo (in press)

Current gaps in turnover research that attrition modeling could address

- Gap 1: Contextualizing existing theories
- Gap 2: Internal job transfer and relocation
- Gap 3: Clarifying the interconnection of withdrawal behaviors

Woo (in press)

Big data and psychological measurement

- **One of the lesser-known V's of big data…. Veracity (or "validity")**

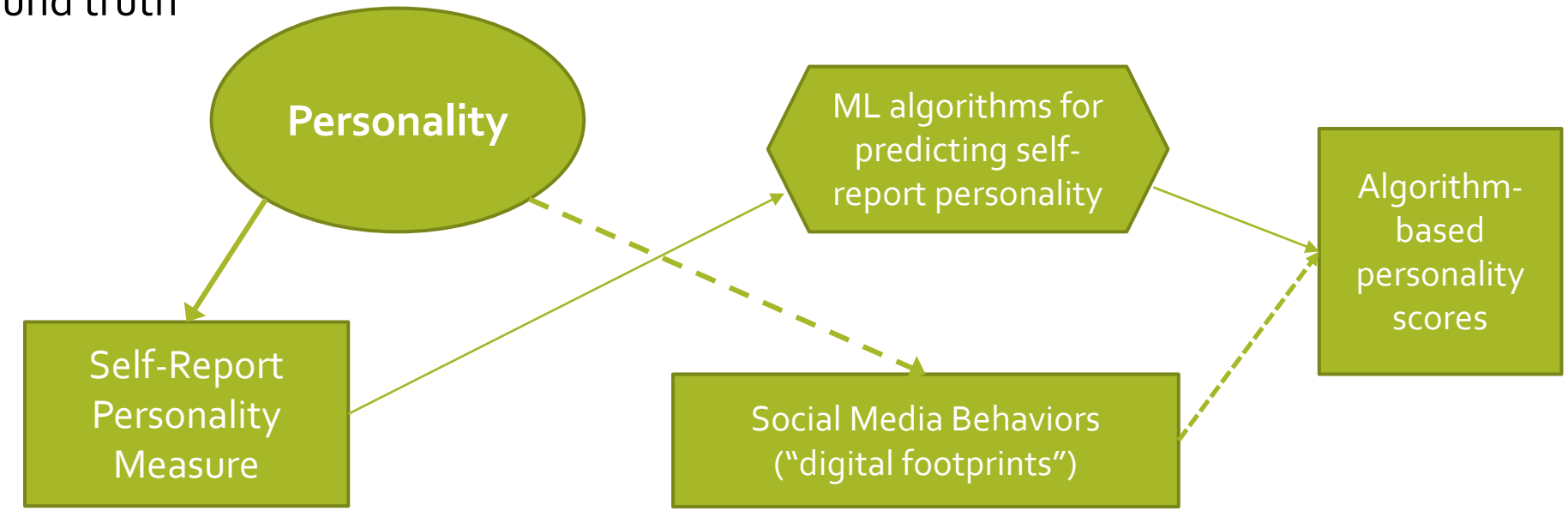
What are we measuring with big data? (Woo et al., 2020)

	Affective & Attitudinal States	Personality Traits	Interpersonal Relationships
Social media	√	√	√
Wearable sensors	√		√
Internet activities	√		
Public network camera	√		√
Smartphones	√	√	√

Top-down (theory-driven) evaluations of measurement validity are not readily available in the literature (with the exception of social media)… why?

Problems with top-down measurement validity claims

- You must assume (1) existence of the construct, and (2) a direct causal link between the construct and the responses/indicators (Borsboom, Mellenbergh, & Heerden, 2004)
- Big data research usually relies on machine-learning algorithms (data-driven analytic techniques) to maximize the prediction accuracy for another established measure as "ground truth"



Woo, Tay, Hickman, & Saef, under review

Big data and psychological measurement

- **An abductive approach to evaluating big-data measurement validity:**
  - ❑ Establish predictive accuracy without a strong validity claim or immediate application
  - ❑ Search for theoretical explanation (focusing on response processes)
  - ❑ Delineate the direct causal effects of known psychological constructs (e.g., personality) vs. their indirect effects via related constructs (e.g., interests, attitudes, goals) vs. contextual factors…
- **What does a bottom-up measurement validation process look like?**
- **Also consider the distinction between intentional vs. incidental measurements (Oswald, 2020)**

Big data *science* (data-intensive science)

- Rob Kitchin (2014) "Big Data, new epistemologies and paradigm shifts"

Table 1. Four paradigms of science.

Paradigm	Nature	Form	When
First	Experimental science	Empiricism; describing natural phenomena	pre-Renaissance
Second	Theoretical science	Modelling and generalization	pre-computers
Third	Computational science	Simulation of complex phenomena	pre-Big Data
Fourth	Exploratory science	Data-intensive; statistical exploration and data mining	Now

Compiled from Hey et al. (2009).

Big data *science* (data-intensive science)

- ***Paradigm shift** or **paradigm expansion?***

- Brian Haig's (2020) chapter: "Big data science: A philosophy of science perspective"
  - Abductive theory of scientific method
  - Challenges the conventional notion of "theory" and "causation" within psychology. Worth a read!

Concluding thoughts

- Validity of algorithm-based measurement: Need more scholarly attention as well as practical guidelines
- The data revolution is changing (or at least diversifying) the way we think about science, theory, and causation. One major source of disagreement on the scientific value of big data may be due to the differences in the conceptualizations of theory and causation.
- **Big data can move our field toward:**
  - **a more diversified, multi-mode paradigm of scientific inquiry**
  - **Triangulation via multiple methods**
  - **Continual cycle of inductive-abductive-deductive reasoning (or "abductive method of science" as a whole)**

THANK YOU!

Acknowledgement:
Louis Tay
Robert Proctor
Fred Oswald
Louis Hickman
Andrew Jebb
Scott Parrigon
Paul Spector
Ernest O'Boyle

For questions and full references,
contact Sang: sewoo@purdue.edu