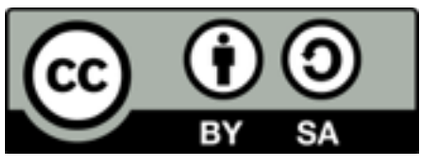
The Scientific Method in the era of Big (Open) Data

Antonio Vetrò
Director of Research - Nexa Center for Internet & Society

Topics in Internet & Society Interdisciplinary Studies,
Politecnico di Torino, 28 June 2016



Credits

Some of the slides have been reused (with permission) by works of Paolo Silos Labini, Angelo Vulpiani, Daniel Méndez Fernández, Manfred Broy

Outline

- Introduction

- Postulates

- Pars destruens

- Pars costruens

Impact of Internet on Science

- Wider reach out of scientific results
- (Potential) engagement of new actors in the scientific process
- Large availability of data and technologies to analyse them

From: timbl@info .cern.ch (Tim Berners-Lee)
Newsgroups: alt.hypertext
Subject: WorldWideWeb: Summary
Keywords: heterogeneous hypertext, web, source, protocol, index, information retrieval
Message-ID: <6487@cernvax.cern.ch>
Date: 6 Aug 91 16:00:12 GMT
References: <6484@cernvax.cern.ch>
Sender: news@cernvax.cern.ch
Lines: 84

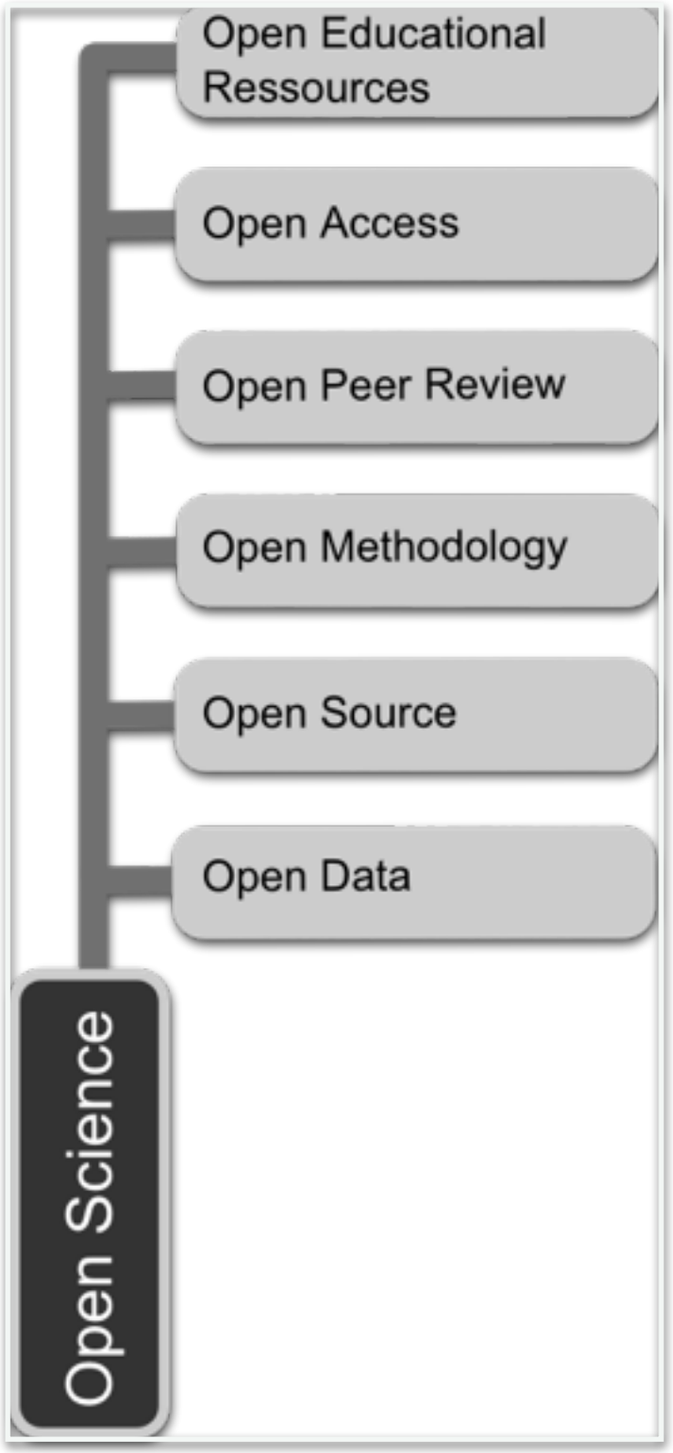
In article <6484@cernvax.cern.ch> I promised to post a short summary of the
WorldWideWeb project. Mail me with any queries.

WorldWideWeb - Executive Summary

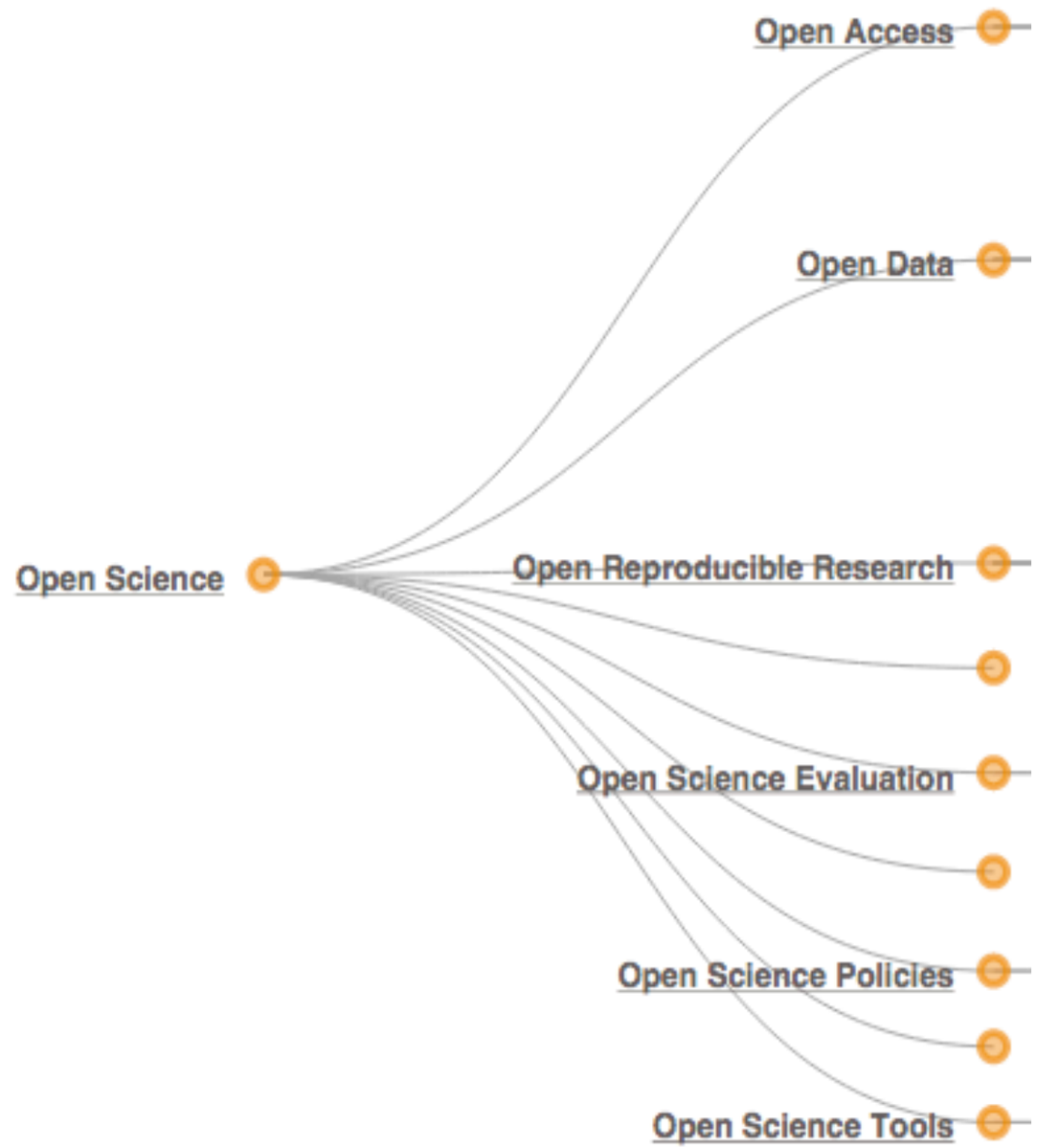
The WWW project merges the techniques of information retrieval and hypertext to
make an easy but powerful global information system.

The project started with the philosophy that much academic information should
be freely available to anyone. It aims to allow information sharing within
internationally dispersed teams, and the dissemination of information by
support groups.

Open (Digital) Science



source : https://en.wikipedia.org/wiki/Open_science



source: Pontika : et al. ; "Fostering Open Science to Research using a Taxonomy and an eLearning Portal" (2015)

Impact of Internet on Science

- Wider reach out of scientific results
- (Potential) engagement of new actors in the scientific process
- Large availability of data and technologies to analyse them

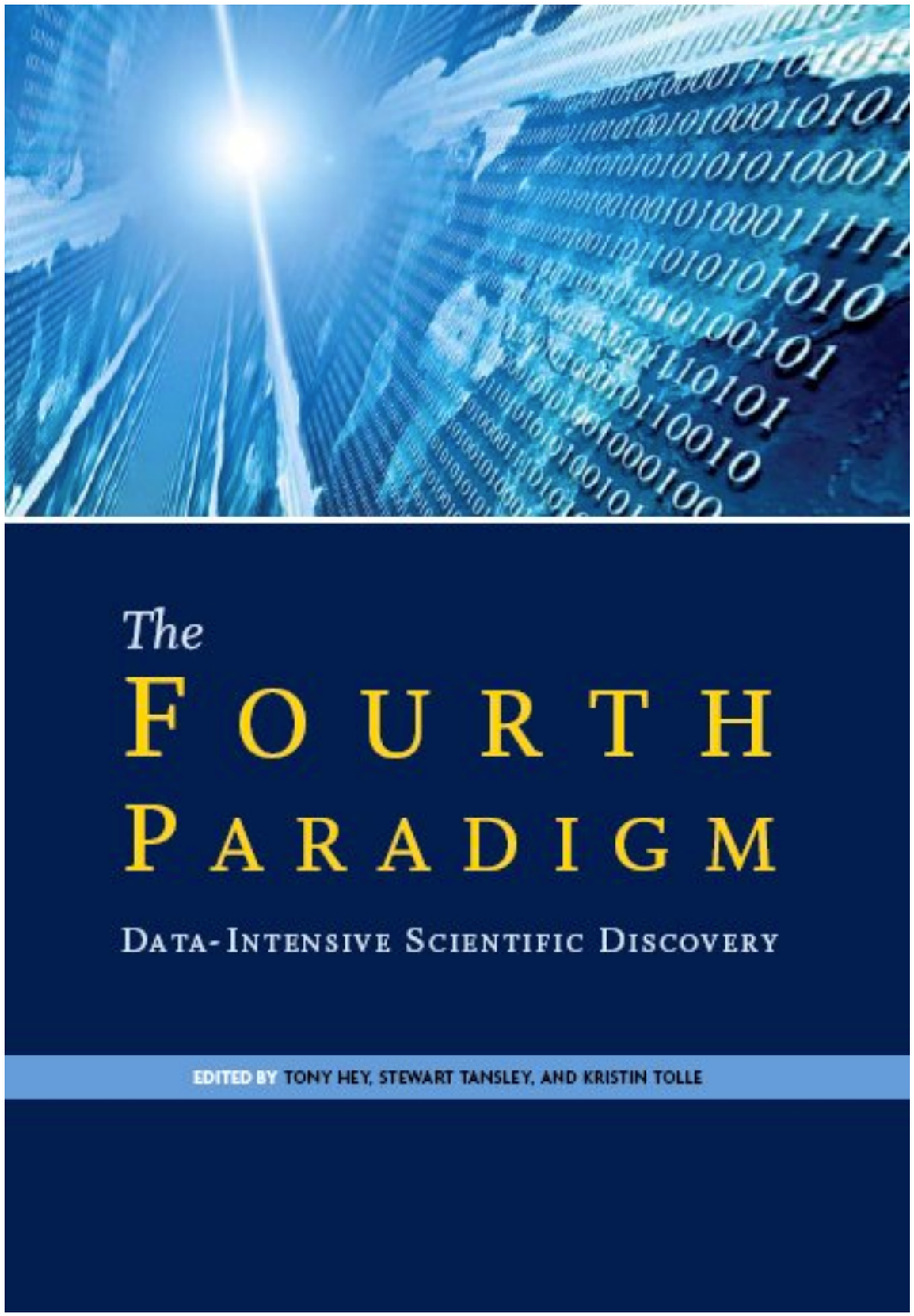
CHRIS ANDERSON SCIENCE 06.23.08 12:00 PM

THE END OF THEORY: THE DATA DELUGE MAKES THE SCIENTIFIC METHOD OBSOLETE

Illustration: Marian Bantjes

http://www.wired.com/2008/06/pb-theory/

"*The new availability of huge amounts of data, along with the statistical tool to crunch these number, offers a whole new way of understanding the world. Correlation supersedes causation, and science can advance even without coherent model, unified theories, or really any mechanistic explanation at all*."



"In the 21st century, much of the vast volume of scientific data captured by new instruments on a 24/7 basis, along with information generated in the artificial world of computer models, is likely to reside forever in a live, substantially publicly accessible, curated state for the purpose of continued analysis. This analysis will result in the development of many new theories !"

extract from the Foreword

JOURNAL OF ONCOLOGY PRACTICE

Screening for Pancreatic Adenocarcinoma Using Signals From Web Search Logs: Feasibility Study and Results

John Paparrizos, MSc, Ryen W. White, PhD and Eric Horvitz, MD, PhD

Author Affiliations

Corresponding author: Ryen W. White, PhD, Microsoft Research, One Microsoft Way, Redmond, WA 98052; e-mail: ryenw@microsoft.com.

Abstract

Introduction: People's online activities can yield clues about their emerging health conditions. We performed an intensive study to explore the feasibility of using anonymized Web query logs to screen for the emergence of pancreatic adenocarcinoma. The methods used statistical analyses of large-scale anonymized search logs considering the symptom queries from millions of people, with the potential application of warning individual searchers about the value of seeking attention from health care professionals.

Methods: We identified searchers in logs of online search activity who issued special queries that are suggestive of a recent diagnosis of pancreatic adenocarcinoma. We then went back many months before these landmark queries were made, to examine patterns of symptoms, which were expressed as searches about concerning symptoms. We built statistical classifiers that predicted the future appearance of the landmark queries based on patterns of signals seen in search logs.

Results: We found that signals about patterns of queries in search logs can predict the future appearance of queries that are highly suggestive of a diagnosis of pancreatic adenocarcinoma. We showed specifically that we can identify 5% to 15% of cases, while preserving extremely low false-positive rates (0.00001 to 0.0001).

Conclusion: Signals in search logs show the possibilities of predicting a forthcoming diagnosis of pancreatic adenocarcinoma from combinations of subtle temporal signals revealed in the queries of searchers.

Proceedings of the National Academy of Sciences of the United States of America

> Current Issue > vol. 111 no. 24 > Adam D. I. Kramer, 8788–8790, doi: 10.1073/pnas.1320040111

Experimental evidence of massive-scale emotional contagion through social networks

Adam D. I. Kramer[a,1], Jamie E. Guillory[b,2], and Jeffrey T. Hancock[b,c]

Author Affiliations

Edited by Susan T. Fiske, Princeton University, Princeton, NJ, and approved March 25, 2014 (received for review October 23, 2013)

A correction has been published

A correction has been published

Abstract Full Text Authors & Info Figures Metrics Related Content PDF

Significance

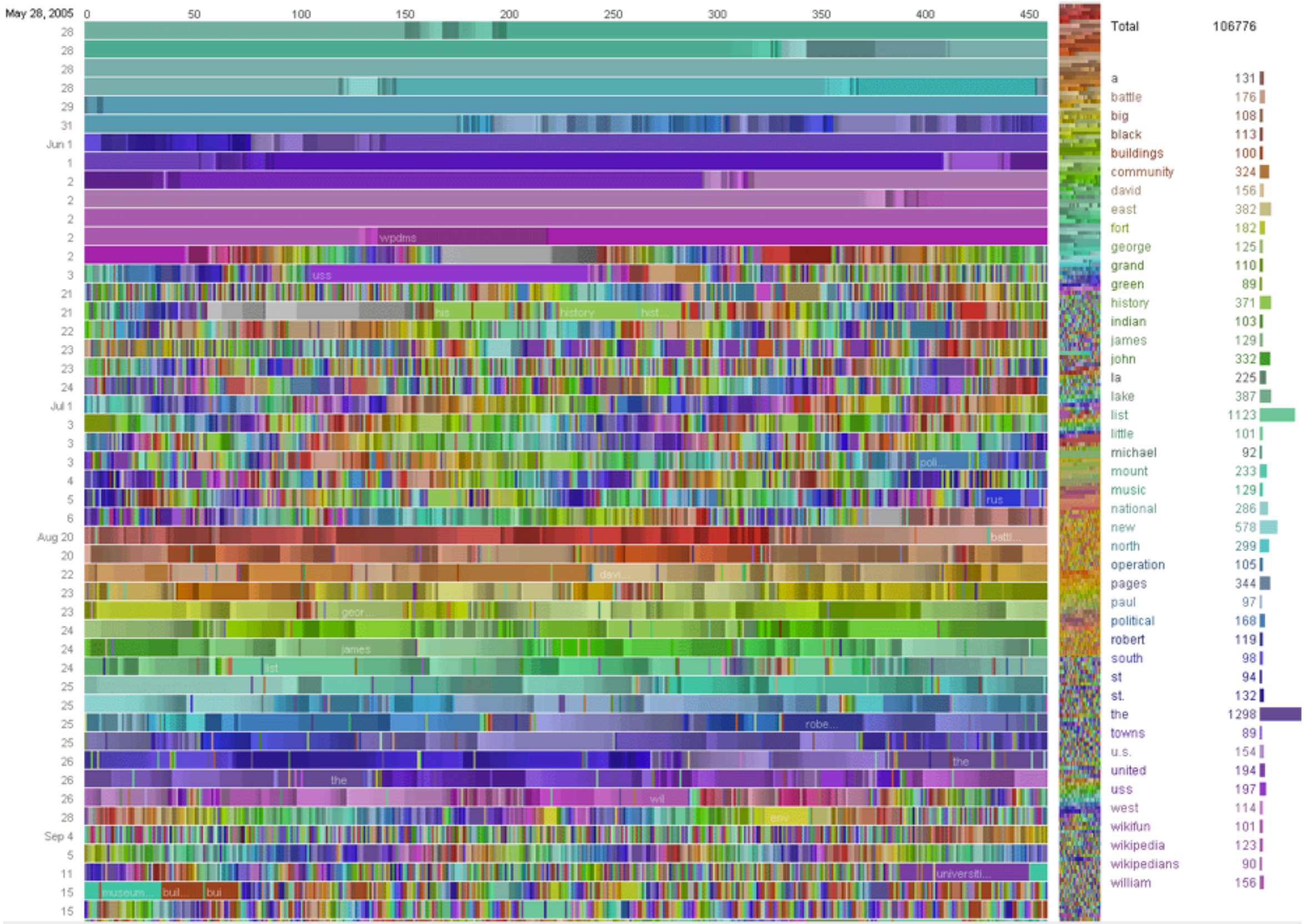
We show, via a massive (N = 689,003) experiment on Facebook, that emotional states can be transferred to others via emotional contagion, leading people to experience the same emotions without their awareness. We provide experimental evidence that emotional contagion occurs without direct interaction between people (exposure to a friend expressing an emotion is sufficient), and in the complete absence of nonverbal cues.

Postulates

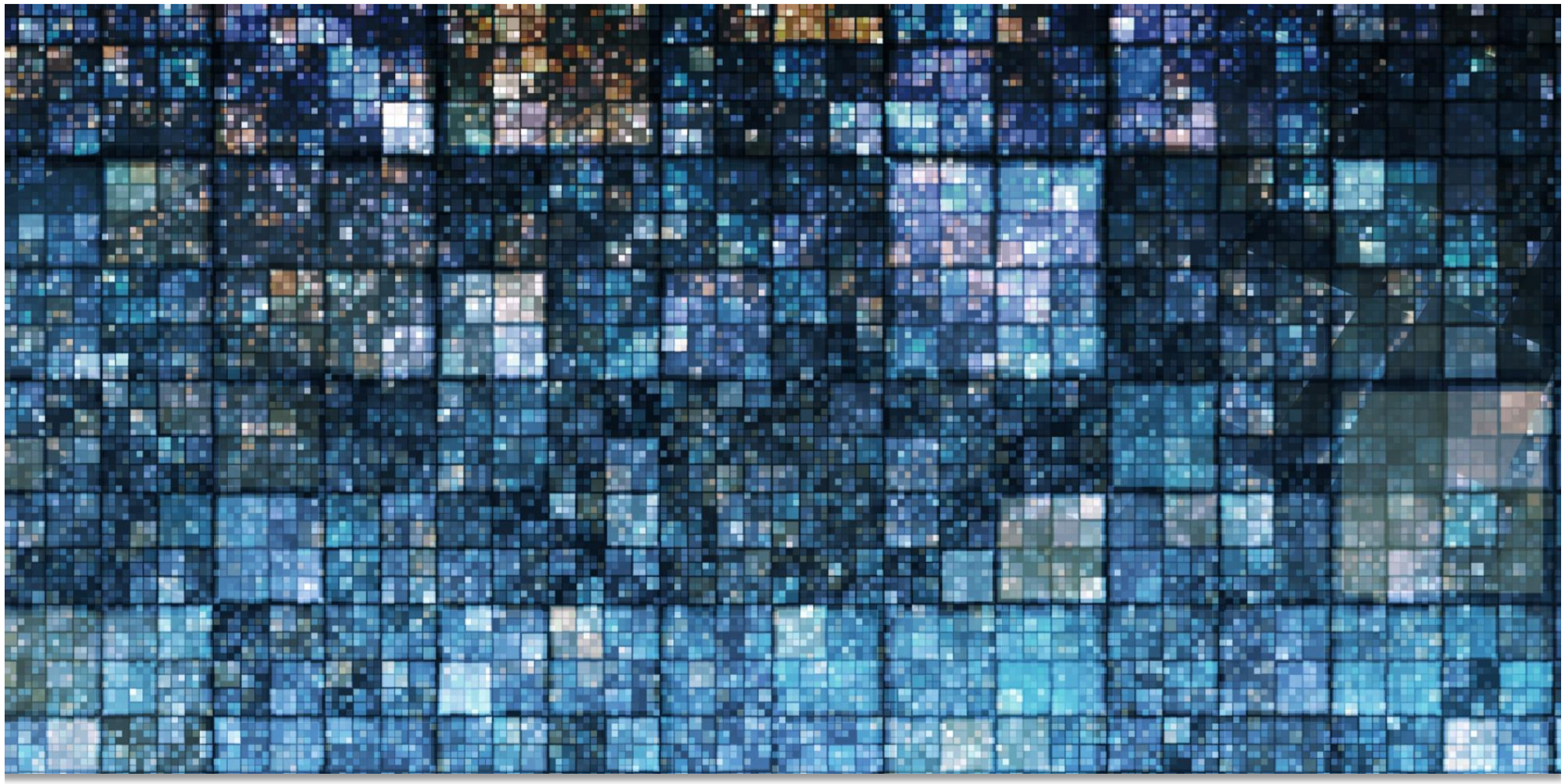
Big Data

Big data is a term for data sets that are so large or complex that traditional data processing applications are inadequate. Challenges include analysis, capture, data curation, search, sharing, storage, transfer, visualization, querying, updating and information privacy. The term often refers simply to the use of predictive analytics or certain other advanced methods to extract value from data, and seldom to a particular size of data set.

Source: https://en.wikipedia.org/wiki/Big_data



Visualization of daily Wikipedia edits created by IBM



Every day 2.5 quintillion bytes of data are created

About 75% of data is unstructured, coming from sources such as text, voice and video

Source: Giuseppe Futia, Nexa Center for Internet & Society

Big Data in Science: example

At the Large Hadron Collider (LHC), protons collide some 1 billion times per second, and the CERN data centre store more than 30 petabytes of data per year from the LHC experiments.

Excerpt from : http://home.cern/scientists/updates/2015/09/big-data-takes-root

Theory

A framework to

Describe - Explain - Predict

phenomena

Swanson, Richard A. Theory Building in Applied Disciplines. San Francisco, CA: Berrett-Koehler Publishers 2018.

Science

Science: Systematically and objectively gaining (and preserving), documenting, and disseminating knowledge

- In principle, science tries to be *objective* by aspiring knowledge based on "facts" (independent of subjective judgment!)

However:

- Accepting scientific results is a social process (documentation, communication, following rules).
- Some elements of science (mathematics, logics) seem to be unbiased – but nevertheless rely on acceptance by the peers and capabilities to apply the theories.
- One could also say: "In the end, it is also a matter of beliefs, capability, and individual and social judgment"
  (following some basic principles, rules, and codes)

Science evolves, too

- Aristoteles (384-324 BC)
  - Search for truth
  - Search for laws and reasoning for phenomena
  - Understanding the nature of phenomena
- Francis Bacon (1561-1626)
  - Progress of knowledge of nature (reality)
  - Scientific knowledge based on inductive and careful observation of events in nature
- Era of (French) Enlightenment (Voltaire (1694-1778), Diderot (1713-1784))
  - Emancipation from god and beliefs
- Kant (1724-1804)
  - System of Epistemology
- Constructivism (Förster (1911-2002), N. Luhmann (1927-1998))
  - Subjective construction
- Popper : empirical falsification (1934 - 1970 ca)
- Other important references: Kuhn, Lakatos, Feyerabend

Adapted from: Orkunoglu, 2010

Science and Philosophy

Ontology

object - subject relation

Is there a world independent of subjectivity?

- Idealism
- Realism
- Solipsism
- …

Epistemology

From where do discoveries result? From experiences?

- Rationalism
- Empiricism
- Scepticism
- …

Ethics

From where does ethics result? Does there exist something like universal ethics?

- Normative Ethics
- Descriptive Ethics
- Everyday Ethics
- …

From: Orkunoglu, 2010

What is the notion of truth ?

- We speak about truth, if no subjective interpretation and distortion is possible
- We could also say: "Whenever I repeat my treatment to a certain population, it will always lead to the same observation"
- If we have "universal truth", we can call our results "generalisable" ("externally valid")

Challenges: Obtaining truth

- Can we obtain something as "universal truth"?
- Can we do so in a life time? Or even within a PhD?
- What if my observations/interpretations/analyses are dependent on human factors?
- Often things can be true for certain <u>contexts</u>

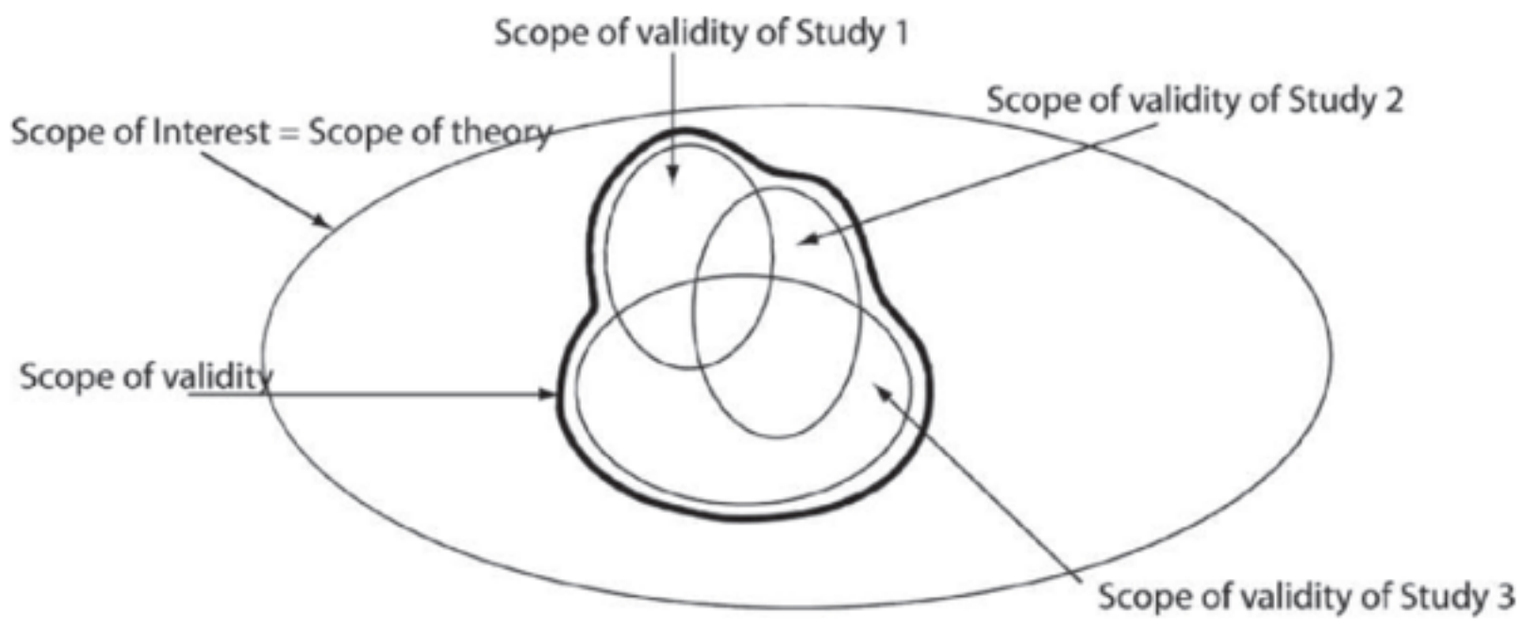


Image: Sjøberg, 2011

Science and methods

Philosophy of science

Principle ways of working

Methods and tools

Fundamental theories

You are (usually) here

Example

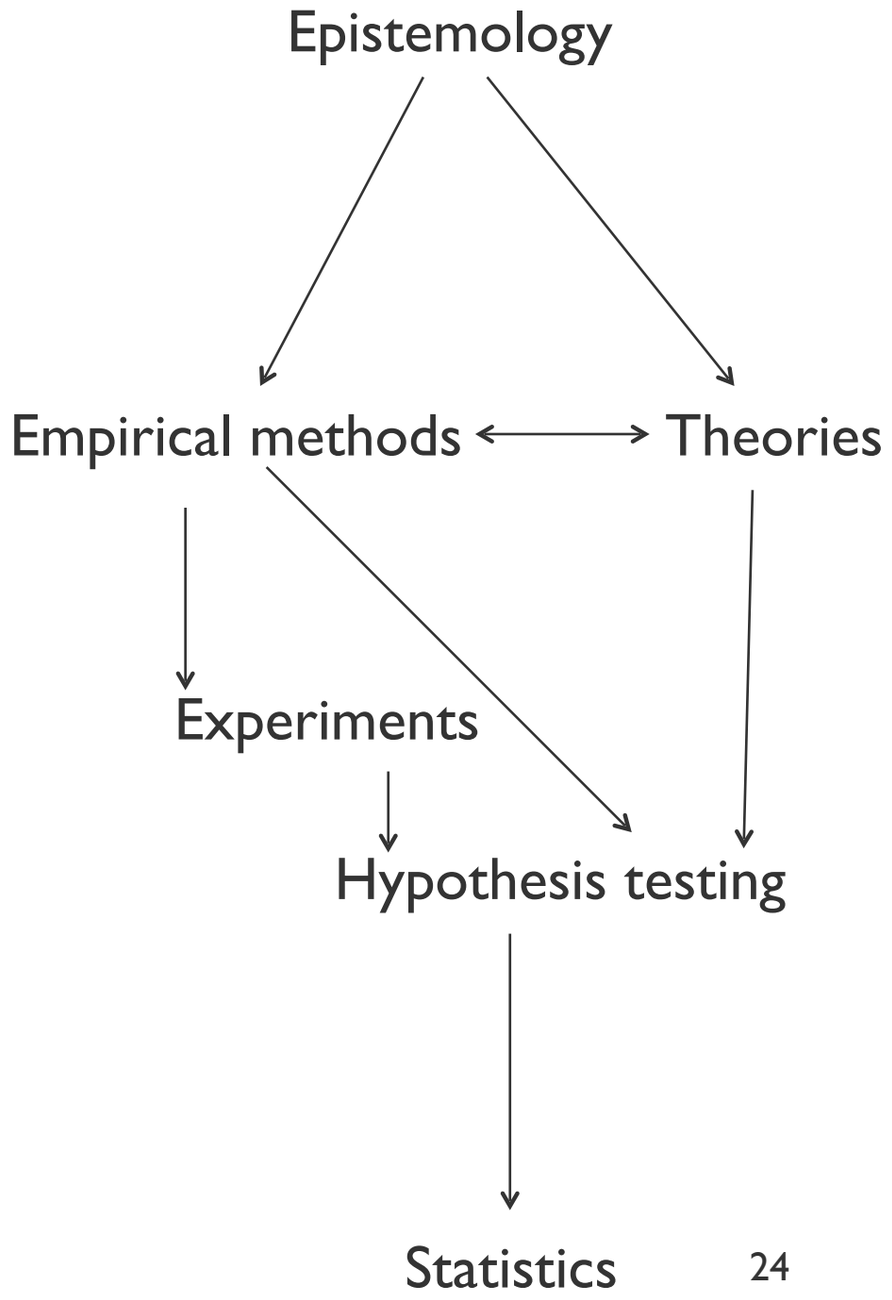
Philosophy of science

Principle ways of working

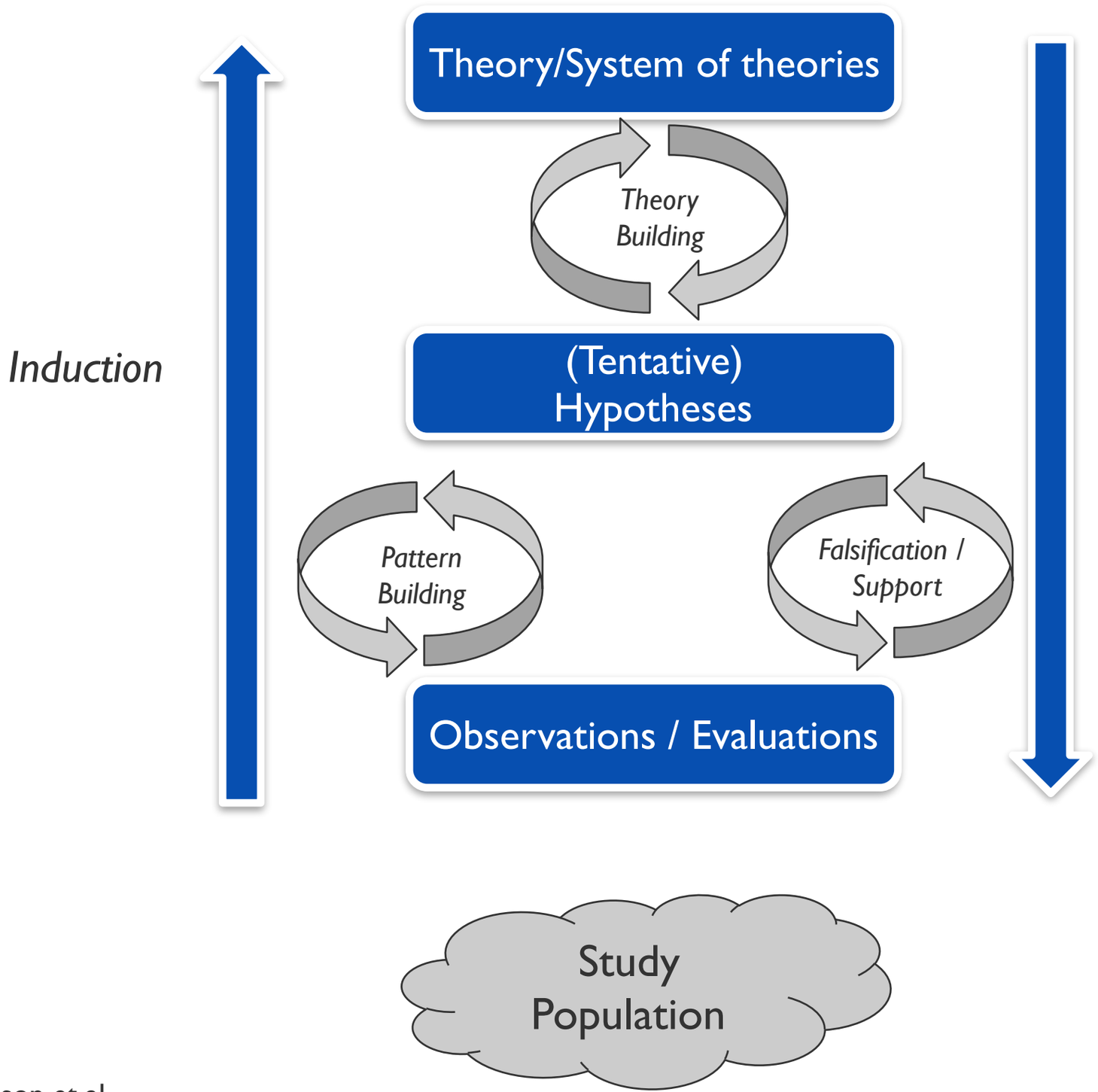
Methods and tools

Fundamental theories

Example



Scientific methods



Further reading: Runeson et al.
Case Study Research in Software Engineering: Guidelines and Experiments

Scientific methods

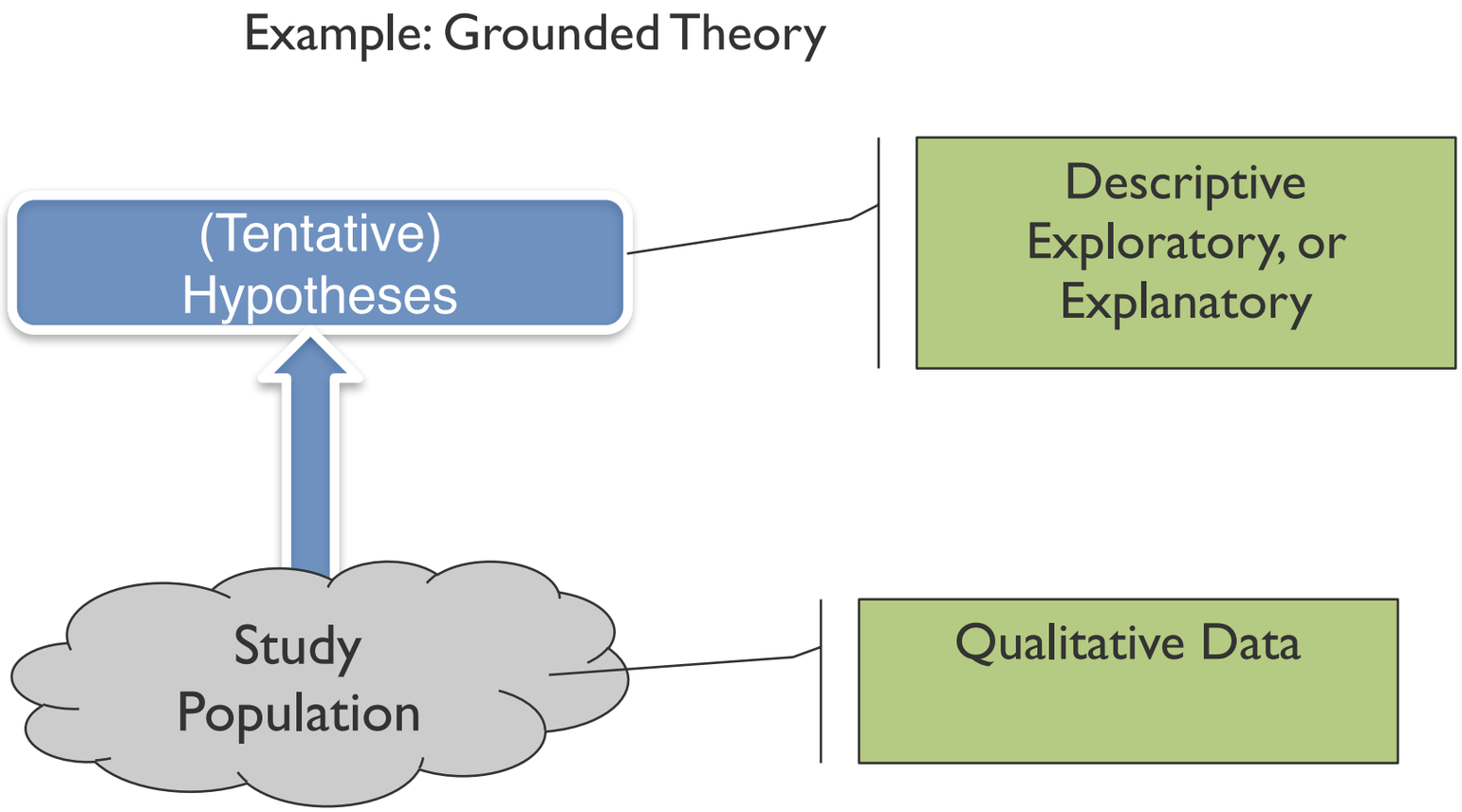
- Each method I can apply…
  - Has a specific **purpose**
  - Relies on a specific **data type**

Purposes

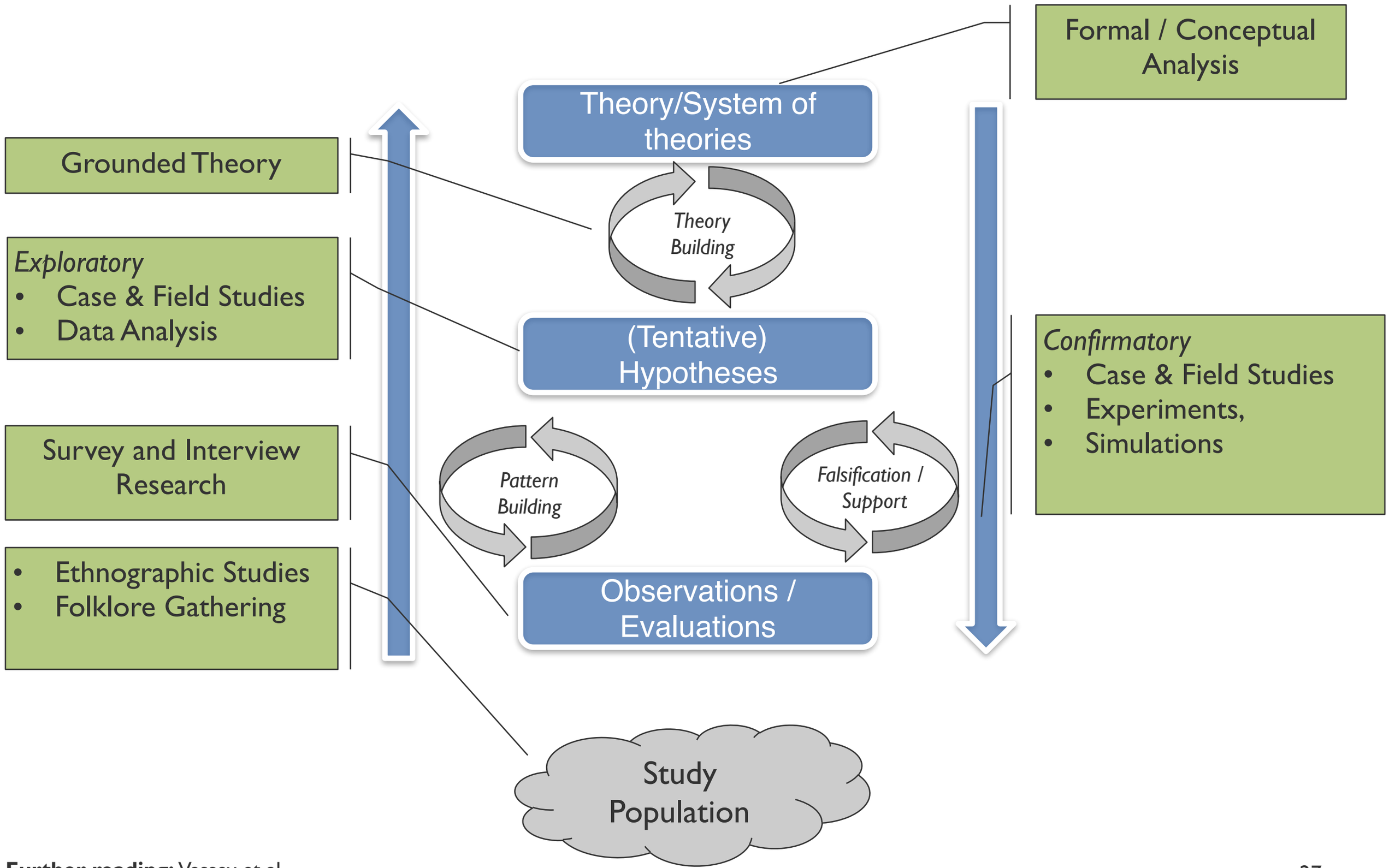
- Exploratory
- Descriptive
- Explanatory
- Improving

Data Types

- Qualitative
- Quantitative



Scientific methods



Further reading: Vessey et al
A unified classification system for research in the computing disciplines

Scientific method: inductive **and** deductive

- Hypothesising
- Modeling
- Testing

BIG (OPEN) DATA:

The end of traditional scientific method ?

CHRIS ANDERSON SCIENCE 06.23.08 12:00 PM

THE END OF THEORY: THE DATA DELUGE MAKES THE SCIENTIFIC METHOD OBSOLETE

Illustration: Marian Bantjes

Pars destruens

Limitations

- Statistical tests are a powerful method to give scientific foundations and provide evidence to theories with the observations it is possible to get

- However they have some limitations
  - not all assumptions often hold (e.g., normality in parametric tests)
  - often we don't have enough data to draw proper conclusions

Limitations: pval

Misleading p-values showing up more often in biomedical journal articles

Date: March

Source: Stanfor

Summary: A reviev misund uncerta

Taylor & Francis
Taylor & Francis Group

The American Statistician

ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: http://amstat.tandfonline.com/loi/utas20

The ASA's Statement on p-Values: Context, Process, and Purpose

NATURE | RESEARCH HIGHLIGHTS: SOCIAL SELECTION

Psychology journal bans *P* values

Test for reliability of results 'too easy to pass', say editors.

Chris Woolston

26 February 2015 | Clarified: 09 March 2015

r

ole A. Lazar (2016) The ASA's Statement on nerican Statistician, 70:2, 129-133, DOI:

l/00031305.2016.1154108

Opportunity of Big Data

- Nowadays large availability of data allows the application of better techniques to reduce the error in drawing conclusions from statistical tests

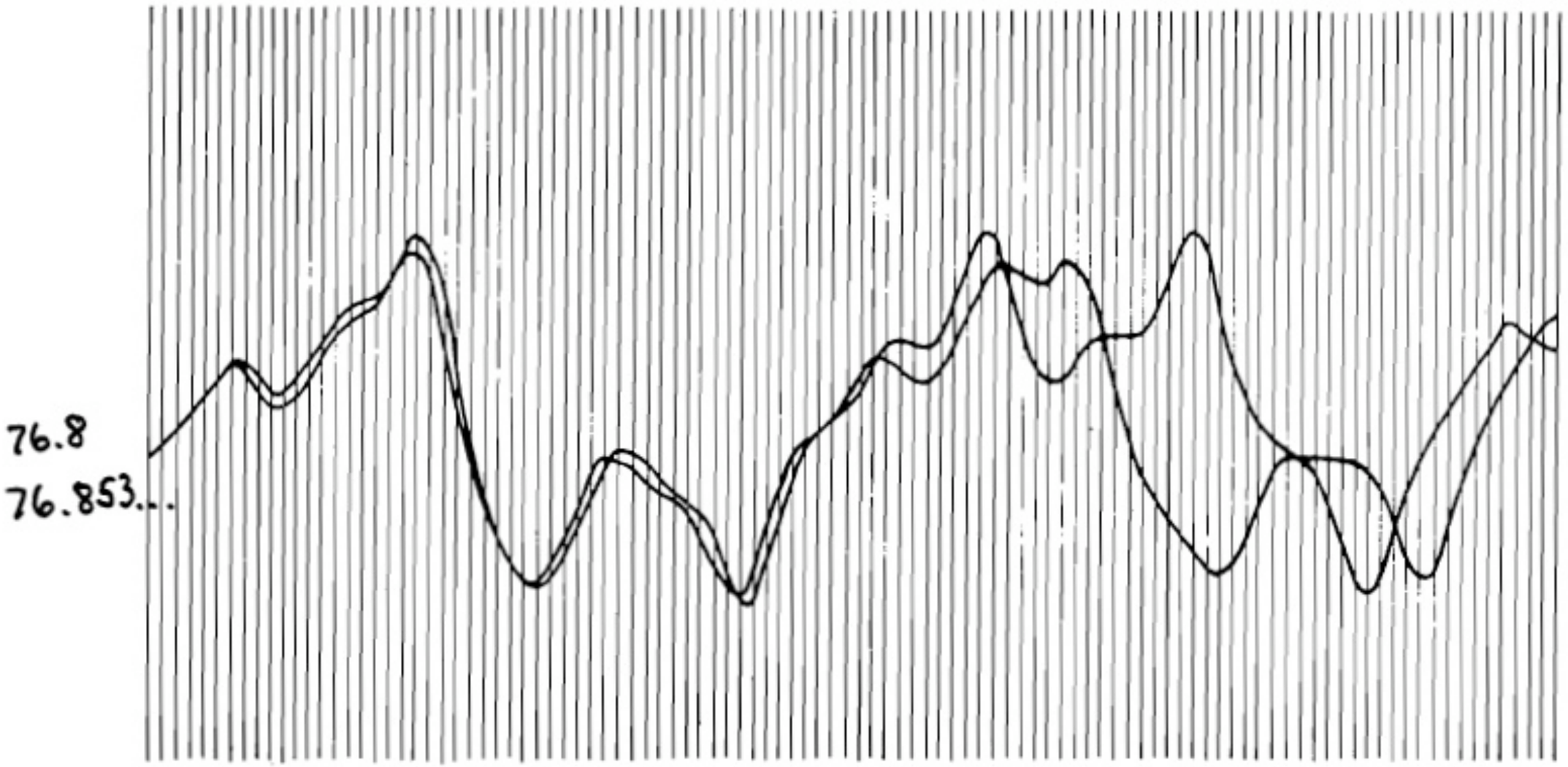
- Example: control on false discovery rate rather than type I error (p val)

- Paradigm: "Collect data first, ask question later"
- It translates in "inference from data" with no a priori questions
  - E.g.: regression , patterns and hypothesis building from historical data
- Often applied for prediction purposes

Not so straightforward

Case1:
deterministic approach

Deterministic approach
The problem of chaos (Poincaré work)



How two weather patterns diverge. From nearly the same starting point, Edward Lorenz saw his computer weather produce patterns that grew farther and farther apart until all resemblance disappeared. (From Lorenz's 1961 printouts.)

The problem of chaos (Poincaré work): an example from physics

Let's take a simple prediction function (Logistic map):

$$x(t+1) = 4x(t)(1-x(t))$$

Its prediction error doubles at every step:

$$\delta x(t) \sim 2^t \delta x(0)$$

In a deterministic chaotic system we have that:
(Lyapunov exponent $\lambda>0$):

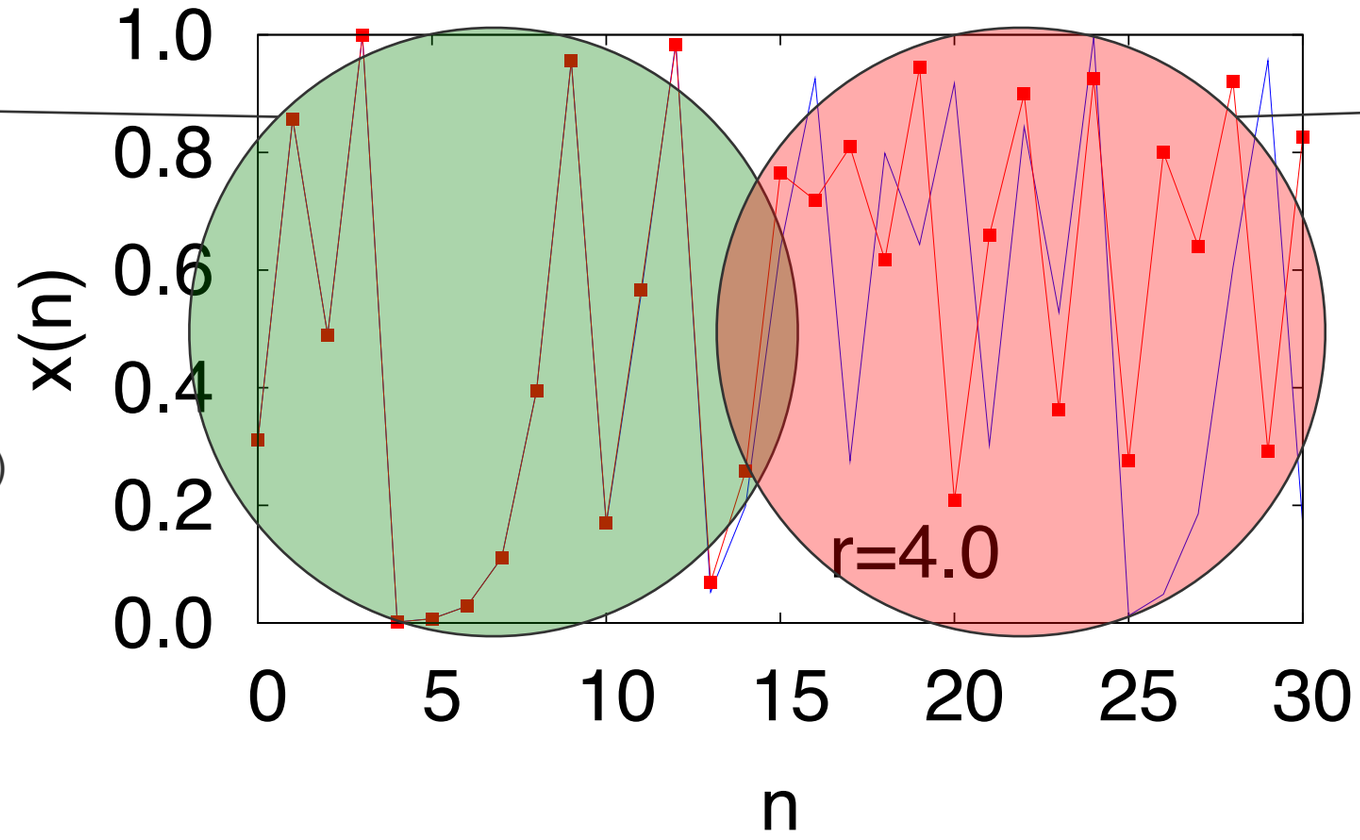
$$|\delta x(t)| \sim e^{\lambda t} |\delta x(0)|$$

That means: a system can be predicted with a tolerance Δ **only within a certain time** which is dependent on λ

$$T_p \sim \frac{1}{\lambda} \ln\left(\frac{\Delta}{|\delta x(0)|}\right)$$

Stability, high predictability

Chaos, low predictability



Example with logistic map:
Despite the very similar initial conditions ($|x(0) - x'(0)| = 4 \times 10^{-6}$), after t=16 the two trajectories are completely different ("Butterfly effect")

$\lambda = \ln 2 \simeq 0.693$

Source: Sylos Labini

Prediction window

$$T_p \sim \frac{1}{\lambda} \ln\left(\frac{\Delta}{|\delta x(0)|}\right)$$

- Predicting eclipses and tides is easier because those phenomena are less chaotic, i.e. Lyapunov exponent λ is lower (and so the predictability window is large). That's why ancient populations (such as Maya) could understand the periodicity of the planets' movement, without having a physical reference model

- The atmosphere is much more chaotic system, and the predictability window is quite short (i.e. λ is higher) → see Lorentz efforts

Source: Sylos Labini

Case2:
probabilistic approach

Probabilistic approach : method of the analogs

- Most predictions algorithm work under the following basic idea
  - We know the past , i.e. a series ($x_1, x_2, ..., x_M$) where $x_j = x(j\Delta t)$
  - We want to forecast the future, i.e. x_{M+t}
  - We look back in the past to find a situation similar to the present (time M), i.e. a vector x_k with k<M and $|x_k - x_M| < \varepsilon$
  - Predict at time M + t

$$\widehat{\boldsymbol{x}}_{M+T} = \boldsymbol{x}_{k+T}$$

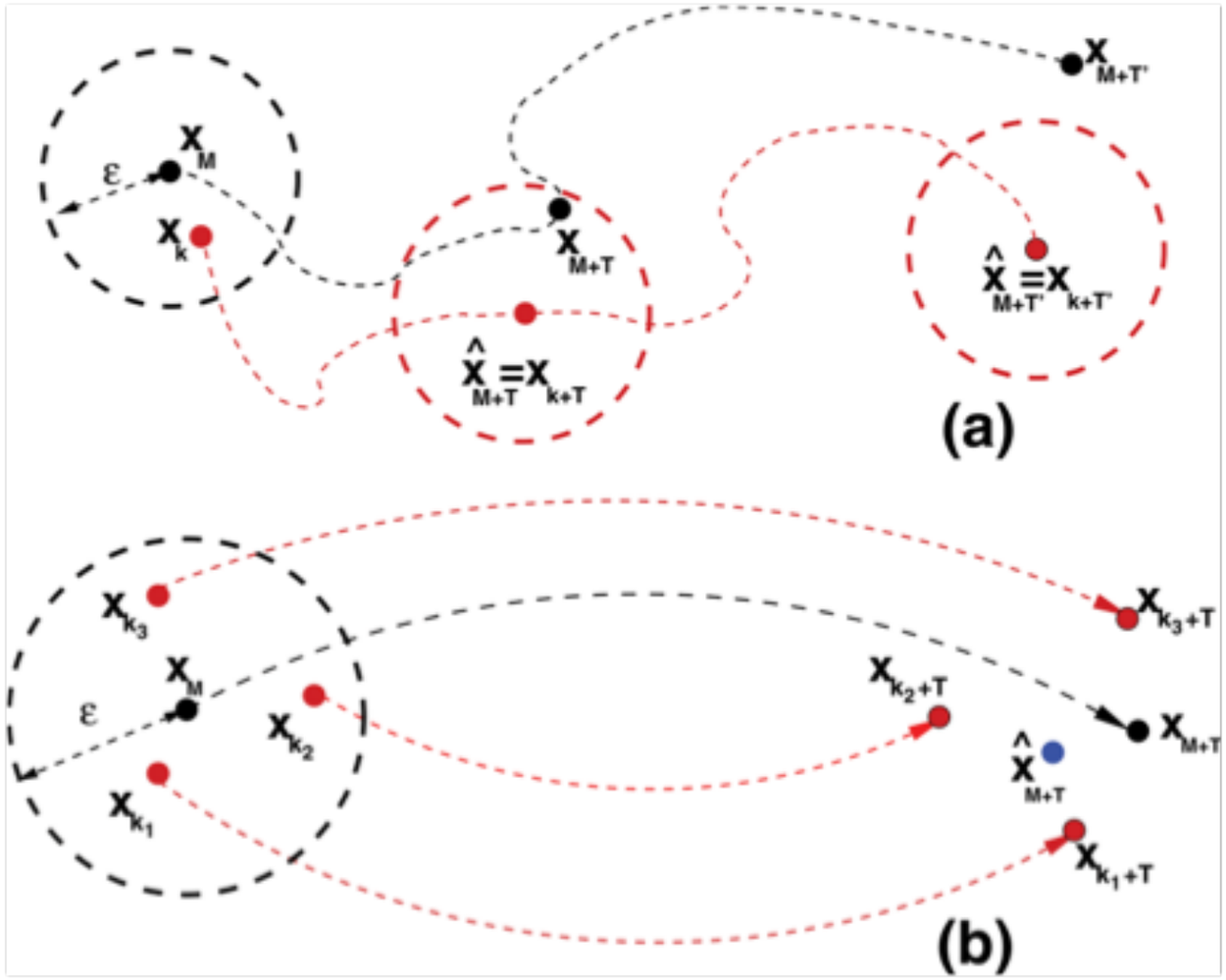


Fig. 1. (Color online) Sketch of the method of analogs: (a) illustration of Eq. (1) and of the error growth; (b) generalization of the method to more than one analog. In particular, if N_a analogs $\{\boldsymbol{x}_{k_a}\}_{a=1}^{N_a}$ are found, Eq. (1) can be replaced by $\widehat{\boldsymbol{x}}_{M+T} = \sum_{a=1}^{N_a} \mathbb{E}_a \boldsymbol{x}_{k_a+T}$ where the matrices $\mathbb{E}_a$ can be computed by suitable interpolations.

Predicting the future from the past: An old problem from a modern perspective
Cecconi, F. and Cencini, M. and Falcioni, M. and Vulpiani, A., American Journal of Physics, 80, 1001-1008 (2012), DOI:http://dx.doi.org/10.1119/1.4746070

Considerations

In an ergodic system, the average return time of a set A is proportional to a system's characteristic time τ_0 and inverse proportional to the probability of A (*Lak's lemma*):

$$\langle \tau(A) \rangle = \frac{\tau_0}{P(A)}$$

Which in a system of linear dimensions O(ε) is inverse proportional to the number of variables involved*

$$P(A) \sim \left(\frac{\epsilon}{L}\right)^D$$

Good news: to find an analog in the past with precision ε , we must go back in time of

$$\langle \tau(A) \rangle \sim \tau_0 \left(\frac{L}{\epsilon}\right)^D$$

Bad news: to find an analog, the minimum length of the series should be of the same order

(eg precision 5%, t = 6×10^7)

$$\tau_0 \left(\frac{L}{\epsilon}\right)^D$$

$$M_{min} \sim \frac{\tau_0}{\Delta t}\left(\frac{L}{\epsilon}\right)^D$$

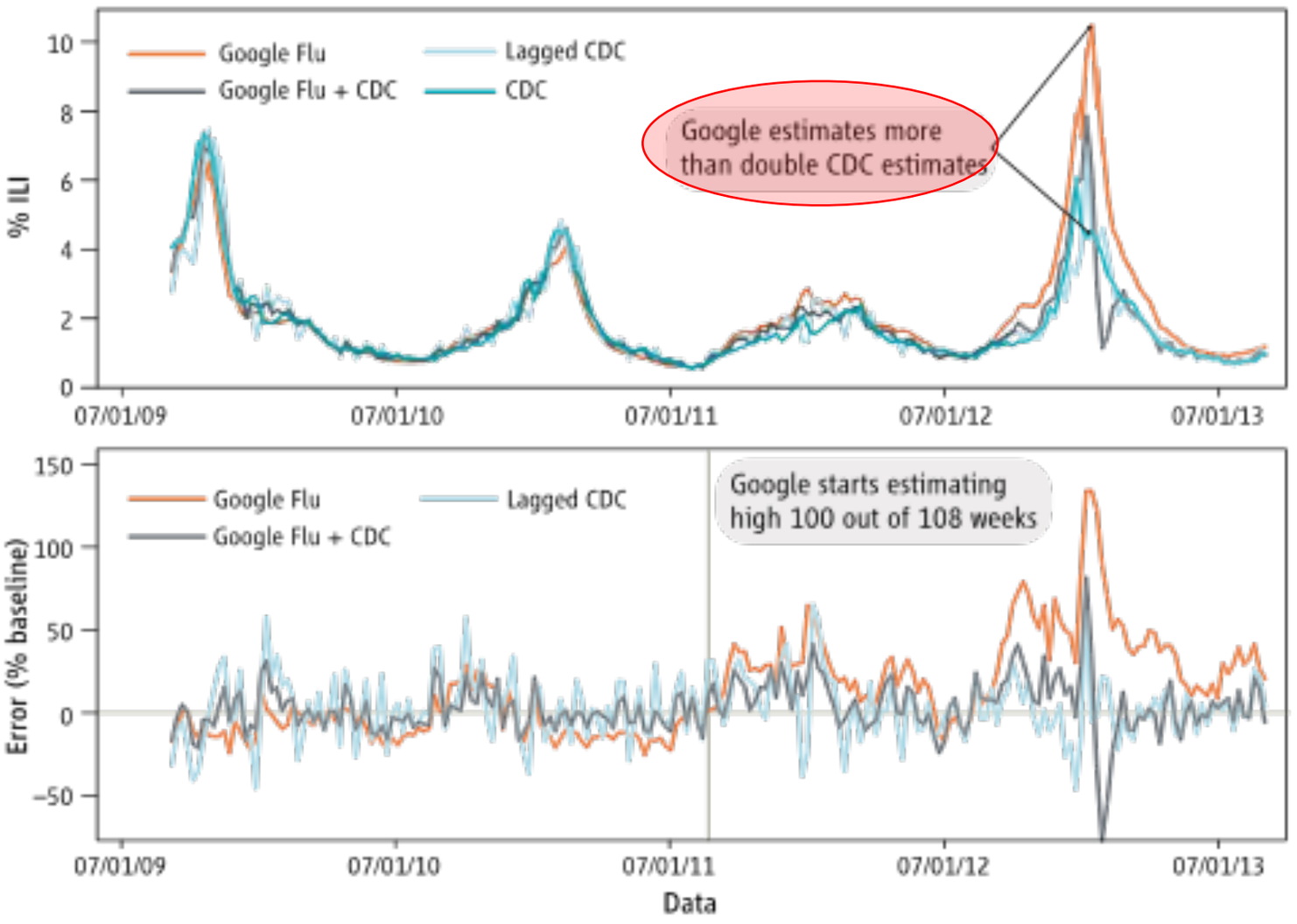
* For details see Cecconi, F. and Cencini, M. and Falcioni, M. and Vulpiani, A. DOI:http://dx.doi.org/10.1119/1.4746070

Considerations (continued)

$$M_{min} \sim \frac{\tau_0}{\Delta t}\left(\frac{L}{\epsilon}\right)^D$$

- In very complex systems (e.g. earthquakes) the state vector is not known a priori, and data are not enough to get appreciable precision in predictions (because it is almost impossible to find an analogue back in the past)

- Example:
  - the case of Google Flu parabole



GFT overestimation. GFT overestimated the prevalence of flu in the 2012–2013 season and overshot the actual level in 2011–2012 by more than 50%. From 21 August 2011 to 1 September 2013, GFT reported overly high flu prevalence 100 out of 108 weeks. (**Top**) Estimates of doctor visits for ILI. "Lagged CDC" incorporates 52-week seasonality variables with lagged CDC data. "Google Flu + CDC" combines GFT, lagged CDC estimates, lagged error of GFT estimates, and 52-week seasonality variables. (**Bottom**) Error [as a percentage {[Non-CDC estmate) − (CDC estimate)]/(CDC) estimate)}. Both alternative models have much less error than GFT alone. Mean absolute error (MAE) during the out-of-sample period is 0.486 for GFT, 0.311 for lagged CDC, and 0.232 for combined GFT and CDC. All of these differences are statistically significant at $P < 0.05$. See SM.

The Parable of Google Flu: Traps in Big Data Analysis

David Lazer, Ryan Kennedy, Gary King, Alessandro Vespignani

SCIENCE, Vol. 343, 14 March 2014

ILI : influenza-like illness
CDC : Centers for Disease Control and Prevention, which bases its estimates on surveillance reports from laboratories across the United States

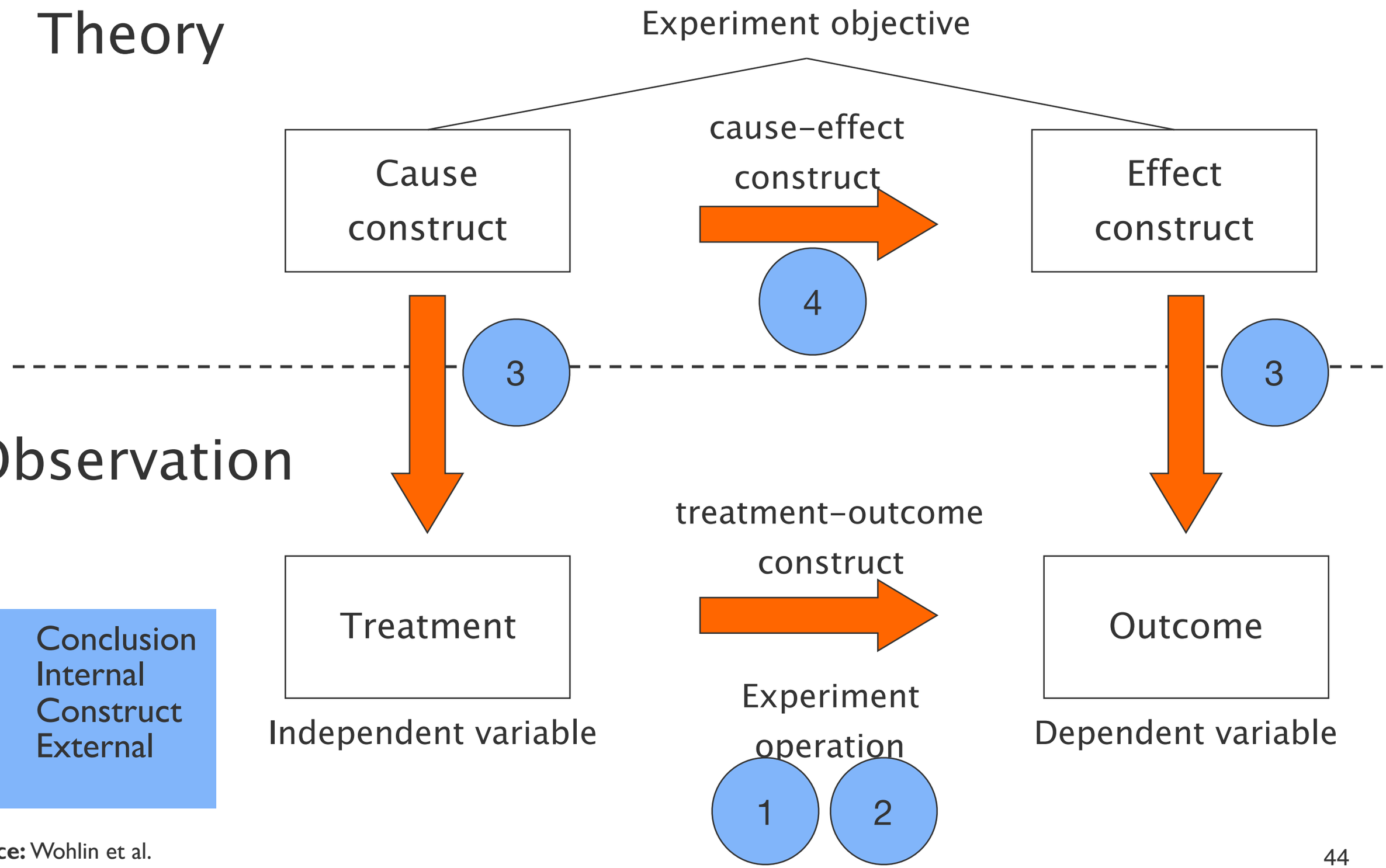
Considerations

- Important questions for prediction:
  - Which are the relevant variables ?
  - What kind of laws regulates the system ?
  - What kind of perspective do we take: deterministic or probabilistic ?

- Different situations
  - Evolution laws in the system exist and are known
  - We don't know whether the system has some laws

- Common problems
  - Not always the equations of the phenomena are known (do they exist?)
  - Often we even don't have a set of variables which describe the phenomenon
  - Which are the confounding factors ?

Main types of validity (for experiment design)

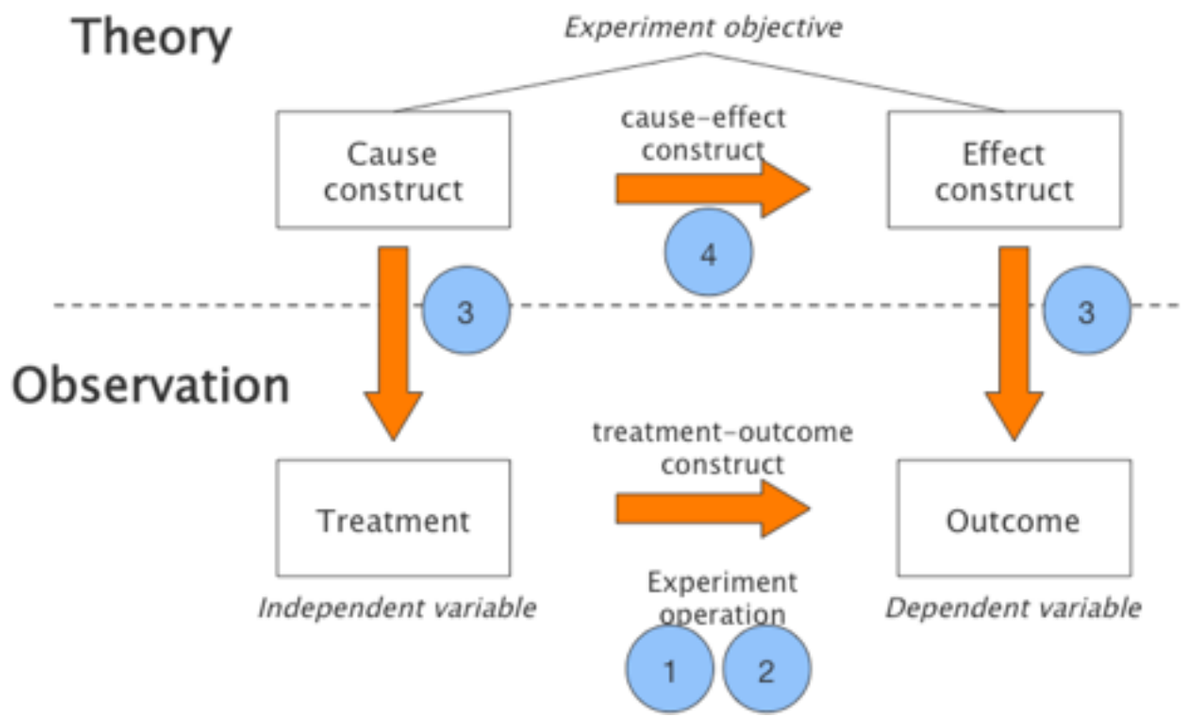


1. Conclusion
2. Internal
3. Construct
4. External

Source: Wohlin et al.
Experimentation in Software Engineering: An Introduction.

Main types of validity (for experiment design)

- Following classification scheme has been established for empirical SE:
  1. Conclusion validity:
     *"In this study, is there a **relationship** between treatment and outcome ?*
  2. Internal Validity:
     *"Assuming there is a relationship in this study, is the relationship a **causal** one?"*
  3. Construct Validity:
     *Assuming that there is a causal relationship in this study, can we claim that the treatment **reflects** well our cause construct and that our measure **reflects** well our idea of the construct of the measure ?*
  4. External Validity:
     *"Assuming that there is a causal relationship in this study between the cause and the effect, can we **generalize** this effect to other persons, places or times ?*



The root cause of our problem is a philosophy of scientific inference that is supported by the statistical methodology in dominant use. This philosophy might best be described as a form of "naïve inductivism," a belief that all scientists seeing the same data should come to the same conclusions.

Goodman, S. N. *Epidemiology* **12,** 295–297 (2001).

check out also:
http://fivethirtyeight.com/features/science-isnt-broken/

What about Ethics ?

What about Ethics ?

JOURNAL OF ONCOLOGY PRACTICE

Screening for Pancreatic Adenocarcinoma Using Signals From Web Search Logs: Feasibility Study and Results

John Paparrizos, MSc, Ryen W. White, PhD and Eric Horvitz, MD, PhD

Author Affiliations

Corresponding author: Ryen W. White, PhD, Microsoft Research, One Microsoft Way, Redmond, WA 98052; e-mail: ryenw@microsoft.com.

Abstract

Introduction: People's online activities can yield clues about their emerging health conditions. We performed an intensive study to explore the feasibility of using anonymized Web query logs to screen for the emergence of pancreatic adenocarcinoma. The methods used statistical analyses of large-scale anonymized search logs considering the symptom queries from millions of people, with the potential application of warning individual searchers about the value of seeking attention from health care professionals.

Methods: We identified searchers in logs of online search activity who issued special queries that are suggestive of a recent diagnosis of pancreatic adenocarcinoma. We then went back many months before these landmark queries were made, to examine patterns of symptoms, which were expressed as searches about concerning symptoms. We built statistical classifiers that predicted the future appearance of the landmark queries based on patterns of signals seen in search logs.

Results: We found that signals about patterns of queries in search logs can predict the future appearance of queries that are highly suggestive of a diagnosis of pancreatic adenocarcinoma. We showed specifically that we can identify 5% to 15% of cases, while preserving extremely low false-positive rates (0.00001 to 0.0001).

Conclusion: Signals in search logs show the possibilities of predicting a forthcoming diagnosis of pancreatic adenocarcinoma from combinations of subtle temporal signals revealed in the queries of searchers.

Proceedings of the National Academy of Sciences of the United States of America

CrossMark click for updates

Experimental evidence of massive-scale emotional contagion through social networks

Adam D. I. Kramer[a,1], Jamie E. Guillory[b,2], and Jeffrey T. Hancock[b,c]

Author Affiliations

Edited by Susan T. Fiske, Princeton University, Princeton, NJ, and approved March 25, 2014 (received for review October 23, 2013)

A correction has been published

A correction has been published

Significance

We show, via a massive (N = 689,003) experiment on Facebook, that emotional states can be transferred to others via emotional contagion, leading people to experience the same emotions without their awareness. We provide experimental evidence that emotional contagion occurs without direct interaction between people (exposure to a friend expressing an emotion is sufficient), and in the complete absence of nonverbal cues.

Rule of Ethics

- Requirements defined in Common Rule
- Approval by Institutional Review Boards
- —> In the example studies, no process for respecting ethical requirements was needed

PARS COSTRUENS

ethics

MEMBERS & MISSION
DATA PROTECTION
PRESS & NEWS
PUBLICATIONS
EVENTS
HUMAN RESOURCES
DATA PROTECTION OFFICER
PUBLIC PROCUREMENT
CONTACT
IPEN
LEGAL NOTICE
ETHICS ADVISORY GROUP

HIGHLIGHTS

BLOG
EU DATA PROTECTION REFORM
STRATEGY 2015-2019

ETHICS ADVISORY GROUP

EDPS starts work on a New Digital Ethics - press release

Towards a new Digital Ethics - video

The European Data Protection Supervisor has set up an **Ethics Advisory Group (*Advisory Group*)** on the ethical dimensions of data protection by his decision of 3 December 2015.

The overall objective of the Advisory Group is to explore the relationships between human rights, technology, markets and business models in the 21st century from an ethical perspective, with particular attention to the implications for the rights to privacy and data protection in the digital environment.

The Group will be active between February 2016 and January 2018.

The Terms of Reference explain the role of the Advisory Group. The Members of the Advisory Group will be chosen following a Call for Interest, published on this website and in The Economist.

More background on the EDPS' work on digital ethics can be found in the Strategy adopted by the EDPS for the period 2015-2019 and in the EDPS Opinion of 11 September 2015 - Towards a new digital ethics: Data, Dignity and Technology.

The Report on the selection of the members of the Ethics Advisory Group is available here.

https://secure.edps.europa.eu/EDPSWEB/edps/EDPS/Ethics

Areas for further research (I)

- Should human data science be regarded as human-subjects research?

- What are the quantifiable risks posed by correlative and/or predictive data research?

- Similarly, how should we account for the risk of sharing datasets when we cannot know what auxiliary datasets they will be combined (munged) with in the future?

Source: Perspectives on Big Data, Ethics, and Society

Areas for further research (II)

- How is big data redefining both when and how the public benefits from research? And what are more precise ways to assess public benefit or justice considerations in big data research?

- How should data privacy and security scientists approach illicitly gained, publicly available data?

- What are the options for self-regulation in data science?

Source: Perspectives on Big Data, Ethics, and Society

Areas for further research (III)

- What resources are needed in the university context to encourage engagement with data ethics issues, particularly outside of the IRB?

- How can integrative approaches to data ethics be fostered in classroom environments? What pedagogical resources are needed?

- How can integrative approaches to data ethics be fostered in classroom environments? What pedagogical resources are needed?

Source: Perspectives on Big Data, Ethics, and Society

Areas for further research (IV)

- What are the ecological and environmental impacts of a rise in big data research and industry?

- How can ethical issues be integrated into core technical research?

- What motivates data scientists—and their colleagues and employers—in industry to establish ethics processes? Which ethics review structures do and do not work inside industry?

- What is the proper purview of “research ethics” as a topic in the age of big data?

method

How can I support validity in general?

In general, we have 2 possibilities:

1. Support the validity *by construction* (often referred to as "validity procedures")
2. Increase the validity *after the fact*

Constructively supporting validity

Conclusion Validity

- Capture and critically discuss statistical assumptions and estimate probability of making errors
- Draw baselines to compare representatives of samples

Internal Validity

- Minimise side-effects and confounding factors
- Be unbiased!
- Refer to method and subject triangulation

Construct Validity

- Reproducibly define research questions and methods before any analysis (e.g. by using GQM)

External Validity

- Observe and explain objects and subjects → Qualitative studies
- Refer to data triangulation
- Refer to independent replication studies!

Further Tips

- Define and report the study according to available guidelines
- Be patient, be flexible
- Recognise the positive value of checking the threats to validity!

Example

- **Comparing four approaches for technical debt identification**,

  Nico Zazworka, Antonio Vetro', Clemente Izurieta, Sunny Wong, Yuanfang Cai, Carolyn Seaman & Forrest Shull,

  Software Quality Journal, 21(2), 2013

- Large correlational analyses (~ 100.000 data points) on 13 releases of Hadoop open source software to discover relationship between quality structural metrics (at code, design and architectural level) and rework indicators (defect proneness and change proneness)

Threat	Type	Control strategy
Choice of statistical significance thresholds	Conclusion	Literature-based choice of thresholds
Data transformation [0,N] → [0,1]	Conclusion	Distribution check
Metrics not normalized by classes size	Conclusion	Correlation check
Correlations found are incidental	Internal	Effect measured on two outcomes
Classes size measured by nr of methods	Construct	Correlation check
Defect proneness measured by nr of bug fixes	Construct	Checked with three different computation methods
Findings generalizability	External	Aggregation on 13 different releases

Increasing the validity after the fact

Independent Confirmation

- Case study /experimental research of theories by researchers not involved in development of theory
- Replication of experiments or case studies until reaching saturation (or getting retired)

Challenges

- What can we expect from a PhD thesis?

Reproducibility vs replicability

- Reproducibility (requires change): *ceteris paribus*
- Replicability (avoids change) :"poor substitute for reproducibility" * ?

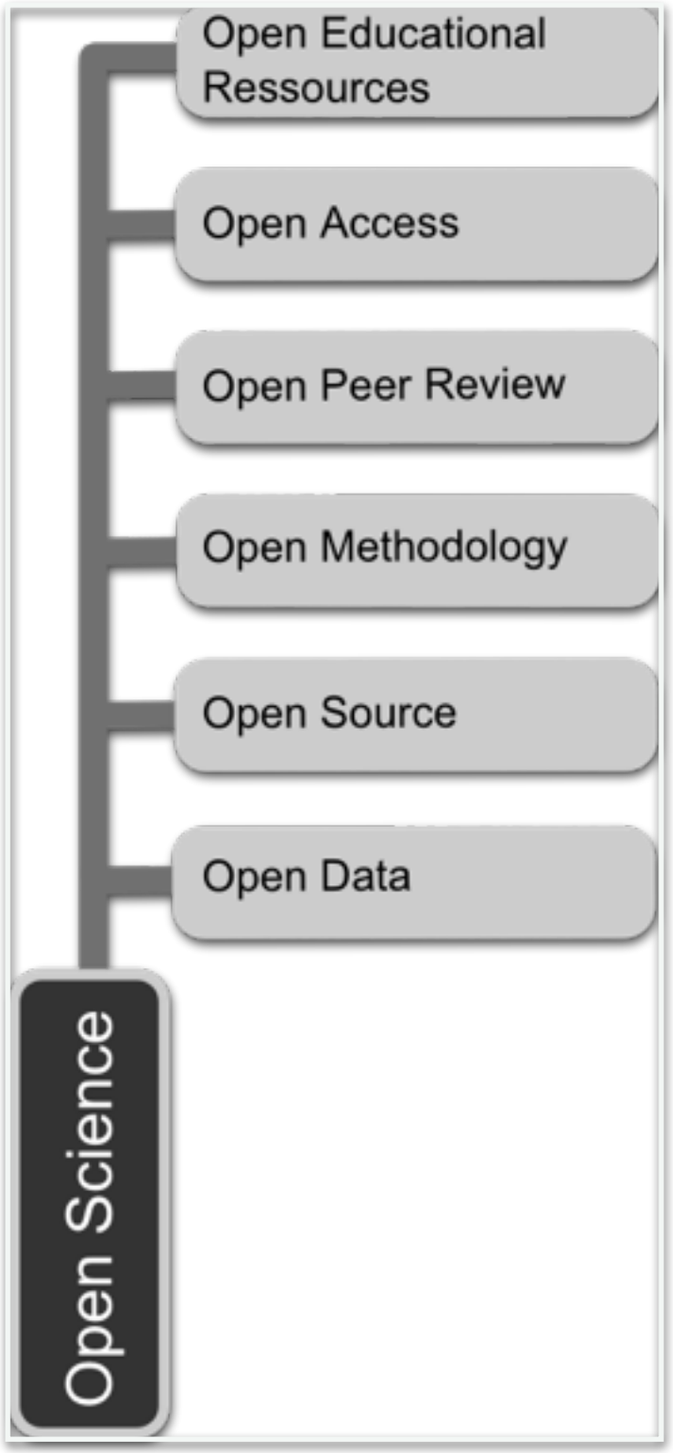
- Nature initiative:
  - no space limitations on Methods sections
  - statisticians help review papers and measures
  - encourage raw data online
  - checklist for life science submissions

- Other ongoing initiatives:
  - The Recomputation Manifesto
  - ARRIVE – Animal Research: Reporting In Vivo Studies
  - National Institutes of Health of the United States (NIH)

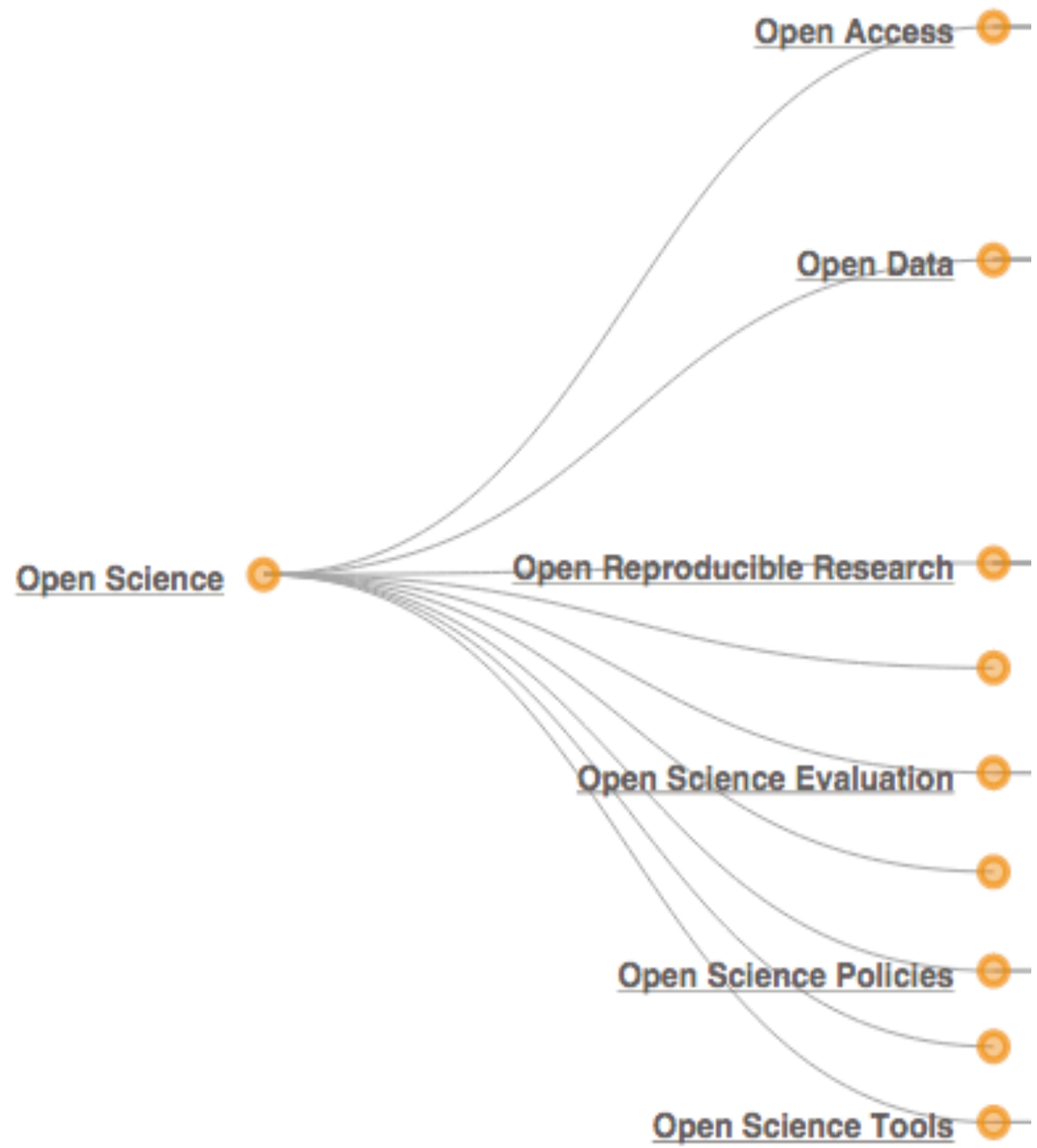
*http://cogprints.org/7691/7/ICMLws09.pdf

infrastructures & norms

Open (Digital) Science



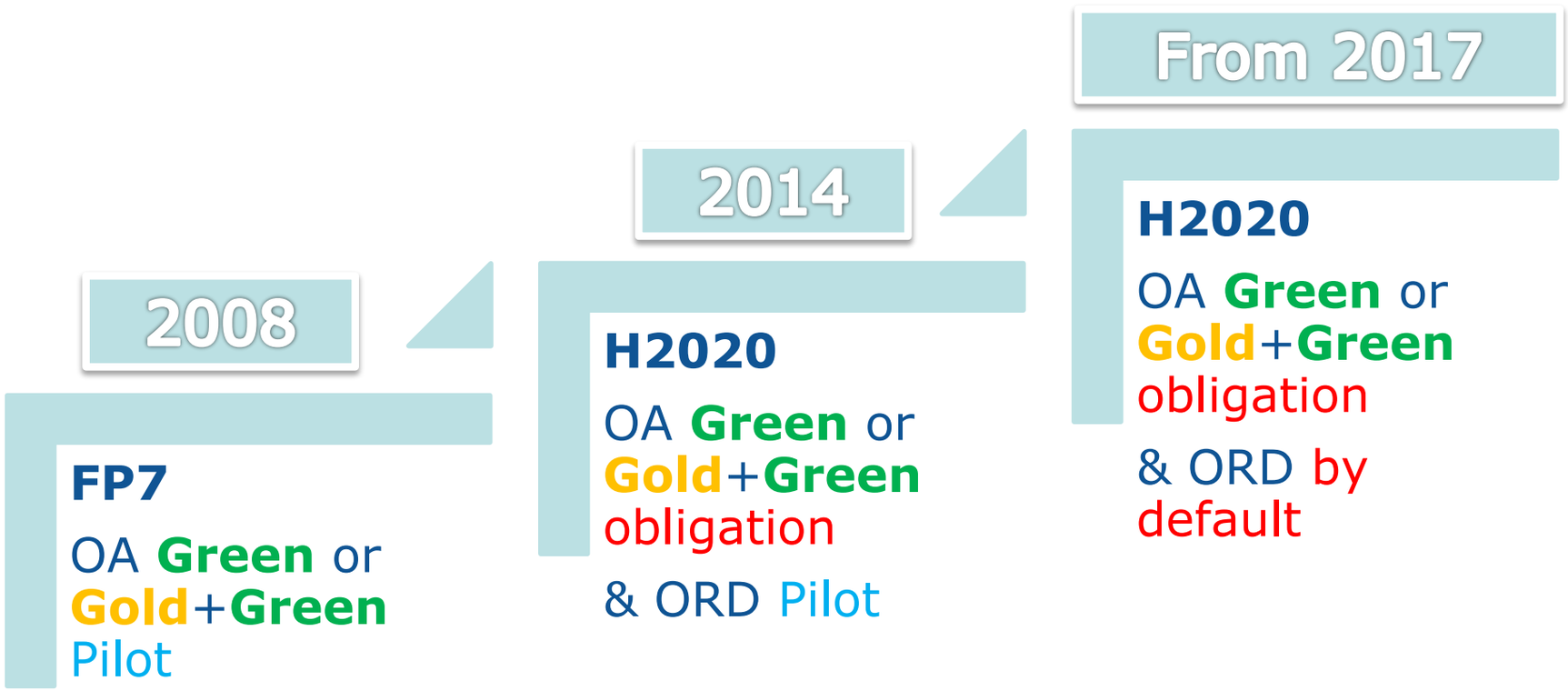
source : https://en.wikipedia.org/wiki/Open_science



source: Pontika : et al. ; "Fostering Open Science to Research using a Taxonomy and an eLearning Portal" (2015)

Open Research Data

OA and EC Framework Programmes



source: http://www.pasteur4oa.eu/sites/pasteur4oa/files/generic/DECHAMP_PASTEUR4OA_rev.pdf

European Open Science Cloud

20 June 2016 – first draft report from the High Level Expert Group

On 20 June 2016, the Commission High Level Expert Group on the European Open Science Cloud (HLEG EOSC) made the draft of their first report available: **A Cloud on the 2020 Horizon. Realising the European Open Science Cloud: first report and recommendations** 565 KB .

The Commission will review the draft report shortly and publish the final version in the course of the summer.

Enquiries in relation to the draft report can be made directly to members of the HLEG EOSC and/or to the Commission at **RTD-EOSC@ec.europa.eu**.

http://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud

Commons based on scientific data

"a federated environment for scientific data sharing and re-use"

see https://www.youtube.com/watch?v=SC4-O8BmI4I

GREEN LIGHT FOR OPEN ACCESS

17-18 MAY 2016 AMSTERDAM

#5

WHAT'S NEXT?

PART 1

POLICIES

THINGS DO CHANGE...

OPEN MONITORING!

GOAL POLICIES: SUPPORT THIS MASSIVE CULTURAL SHIFT!

policies → working habits!

SUPPORT SERVICES FOR USERS!

"ADVOCATE THE BENEFITS FOR THE USERS!"

SUSTAINABILITY!

COORDINATION!

- RISKS
- GAPS
- IMPROVEMENT

GOVERNANCE OF

INFRASTRUCTURE

connect to related infrastructures →

how can we avoid "occupation" by parties we don't trust?

WE NEED KNOWLEDGE ABOUT

- TECHNICS
- USERS
- POLITICS

800 UNIVERSITIES

EUROPEAN UNIVERSITY ASSOCIATION (EUA)

"use & reuse of research results is crucial!"

democratic society

PLATFORM

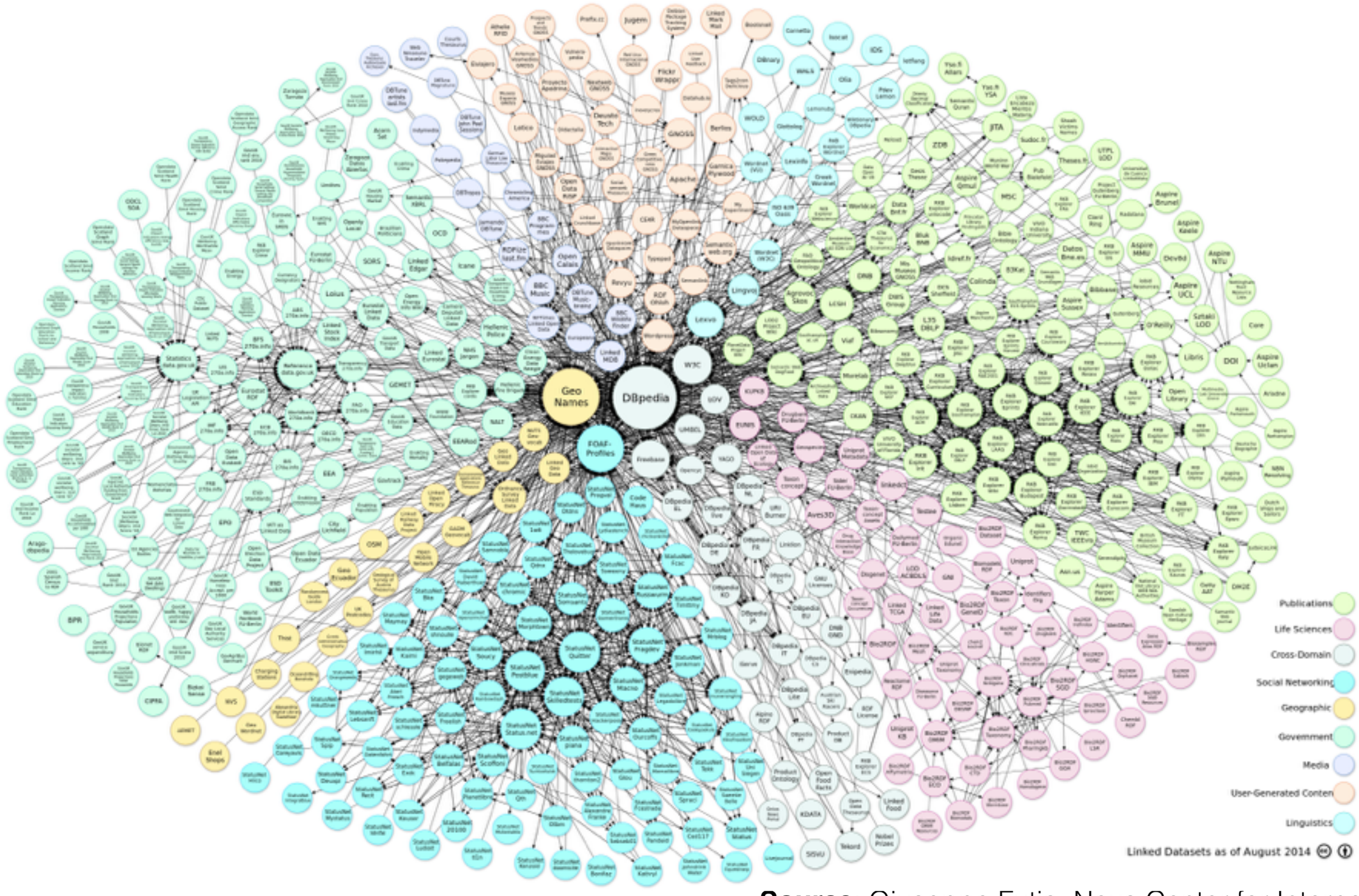
DIALOG WITH ALL STAKEHOLDERS

GET & SHARE INFORMATION

PASTEUR4OA

DENKSCHETS.NL

Linked Data Cloud



Source: Giuseppe Futia, Nexa Center for Internet & Society

Big Data and Development Incubator Symposium, "Big Data and Human Development"
Date: September 15-16, 2016
Location: Said Business School, Oxford, UK

This symposium will serve as a bridge between methodological knowledge about big data, critical academic research on the topic, and the desires of stakeholders and practitioners to achieve key developmental outcomes and goals.

Keynote Speakers include Professor Bitange Ndemo, Professor Alex (Sandy) Pentland, and Dr. Linnet Taylor

We welcome the submission of abstracts (of max 250 words) for talks, panels, and sessions at the workshop. Submit them before July 15, 2016, to christopher.dobyns@oii.ox.ac.uk

Please contact Mark Graham (mark.graham@oii.ox.ac.uk) with any questions.

Science does not rest upon solid bedrock. The bold structure of its theories rises, as it were, above a swamp. It is like a building erected on piles. The piles are driven down from above into the swamp, but not down to any natural or 'given' base; and when we cease our attempts to drive our piles into a deeper layer, it is not because we have reached firm ground. We simply stop when we are satisfied that they are firm enough to carry the structure, at least for the time being. (1959)

— Karl Raimund Popper

Useful links and readings

- **Readings on Big Data and application to Science:**

  - Viktor Mayer-Schnberger. 2013. Big Data: A Revolution that will Transform how We Live, Work and Think. Viktor Mayer-Schnberger and Kenneth Cukier. John Murray Publishers, , UK.

  - The Fourth Paradigm: Data-Intensive Scientific Discovery In The Fourth Paradigm: Data-Intensive Scientific Discovery (2009) by Anthony J. G. Hey, Stewart Tansley, Kristin M. Tolle

- **Perspective from Philosophy and Ethics:**

  - http://www.recode.net/2016/6/14/11923286/facebook-emotional-contagion-controversy-data-research-review-policy-ethics

  - http://www.theverge.com/2014/12/9/7360441/facebook-screwing-with-user-emotions-was-2014s-most-shared-scientific

  - Perspectives on Big Data, Ethics, and Society, THE COUNCIL FOR BIG DATA, ETHICS, AND SOCIETY, http://bdes.datasociety.net/wp-content/uploads/2016/05/Perspectives-on-Big-Data.pdf

  - An Introduction to Philosophy of Science, Kent W. Staley, Cambridge University Press

Useful links and readings

- **Perspective from Statistics:**

- R. Foygel Barber and E. J. Candès. Controlling the false discovery rate via knockoffs.

- Sheldon M. Ross, Introduction to probability and statistic for engineers and scientists, ELSEVIER

- Ronald E. Walpole, Raymond H. Myers, Sharon L. Myers, Keying Ye, Probability & Statistics for Engineers & Scientists

- Bradley Efron, Large-Scale Inference , Empirical Bayes Methods for Estimation, Testing, and Prediction, ISBN: 9781107619678, Jan 2013

- **Perspective from Physics**

- Cecconi, F. and Cencini, M. and Falcioni, M. and Vulpiani, Predicting the future from the past: An old problem from a modern perspective, A., American Journal of Physics, 80, 1001-1008 (2012), DOI:http://dx.doi.org/10.1119/1.4746070

- Francesco Sylos Labini, Big Data Complexity and Scientific Method

- Chris Anderson, The End of Theory

- L.F. Richardson, Weather Prediction by Numerical Process (Cambridge University Press, 1922)