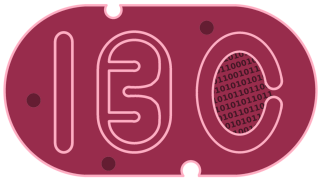
Data Science & Science Data

Challenges and opportunities

Patrick Valduriez



Institut de biologie computationnelle

Data versus Information

- Data
  - Elementary definition of a fact
    - E.g. temperature, exam grade, account balance, message, photo, transaction, etc.
  - Can be complex
    - E.g. a satellite image
  - Can also be very simple, and taken in isolation, not very useful

Les données en question

PAR
Stéphane Grumbach
Patrick Valduriez

NIVEAU DE LECTURE
Facile

PUBLIÉ LE
31/03/2016

Au cœur de la connaissance et de l'information, les données ont peu à peu pris une importance qui nous dépasse. Mais qu'entend-on exactement par données ? Quels sont les enjeux autour de leur gestion ou de leur analyse ? Quels impacts sur la société ?

© Fotolia - ptnphotof

Une donnée est la description élémentaire d'une réalité ou d'un fait, comme par exemple un

https://interstices.info

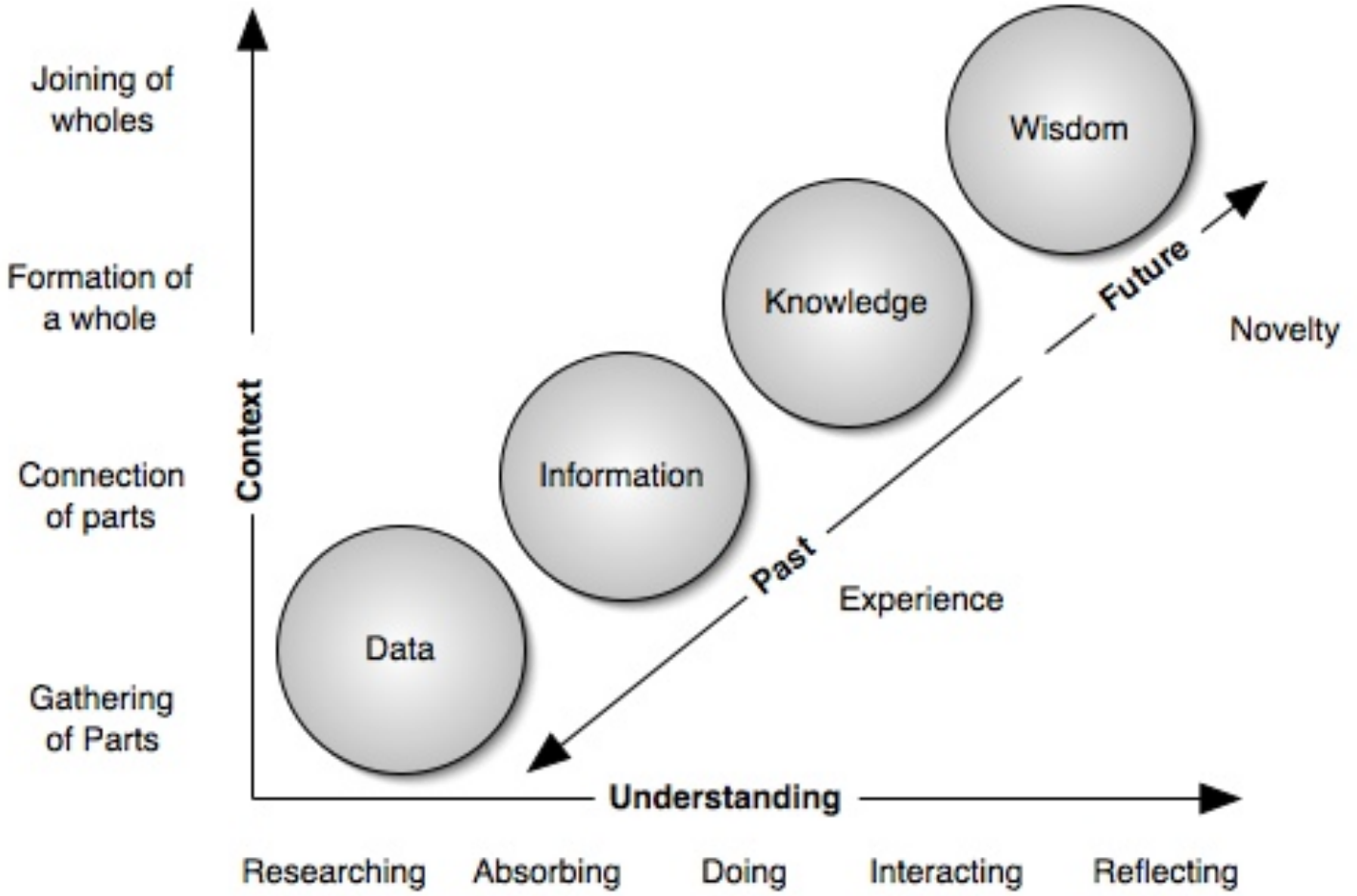
- Information
  - Obtained by integration, interpretation and analysis of data to yield sense in a given *context*
  - Can be very useful to understand the world
    - E.g. climate evolution, ranking of a student, etc.

Data and Algorithm

> "Content without method leads to fantasy,
> method without content to empty sophistry."
>
> Johann Wolfgang von Goethe (Maxims and Reflections, 1892)

- The better the datasets, the better the machine learning algorithms
- Milestones
  - 1997: IBM Deep Blue defeats Chess world champion Garry Kasparov
    - Negascout planning algorithm (1983)
    - Dataset of 700 thousands of chess games (1991)
  - 2016: Google Alphago defeats Go master Lee Sedol (4-1)
    - Monte Carlo method based algorithm (from the 1940's) and neural network
    - Dataset of 30 millions of go moves

The Continuum of Understanding



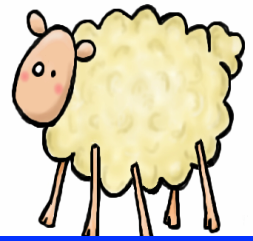
Outline

1. Data science
2. The good, the bad and the ugly
3. Science data
4. Challenges
5. Opportunities

Data Science

Data Science: definition

Used to be hard to find data scientists



New training programs all over the world

But many "fake" data scientists on the market

- Data scientist
  - Not to be confused with data analyst
  - Strong technical skills
    - Statistics, data analysis and CS (programming, data mgt, machine learning)
  - AND good knowledge of the business domain, to interpret the analysis results and draw meaningful conclusions

Big Data: what is it?

- A buzz word!
  - It depends on your perspective
    - E.g. 10 terabytes is big for an OLTP system, but small for a web search engine
- A definition (Wikipedia)
  - Consists of data sets that grow so *large* that they become awkward to work with
  - *But size is only one dimension of the problem*
  - Dimensions (Vs): volume, velocity, variety, veracity, validity
- How *big* is big?
  - Moving target: terabyte (10^{12} bytes), petabyte (10^{15} bytes), exabyte (10^{18}), zetabyte (10^{21})
  - Landmarks in DBMS products
    - 1980: Teradata database machine
    - 2010: Oracle Exadata database machine

Big Data Analytics (BDA)

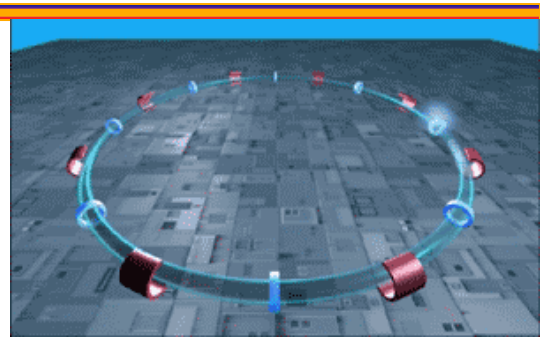
- Objective: find useful information and discover knowledge in data
  - Using data mining, data analysis, machine learning ...
- Why is this hard?
  - Low information density (unlike in corporate data)
    - Like searching for needles in a haystack
  - External data from various sources
    - Hard to verify and assess, hard to integrate
  - Different kinds of data
    - Structured data: transaction, decision-support, scientific
    - Unstructured: web document, social network, open data, IoT
    - Hard to integrate
  - Simple machine learning models don't work
    - See next: "When big data goes bad" stories

Some BDA Killer Apps

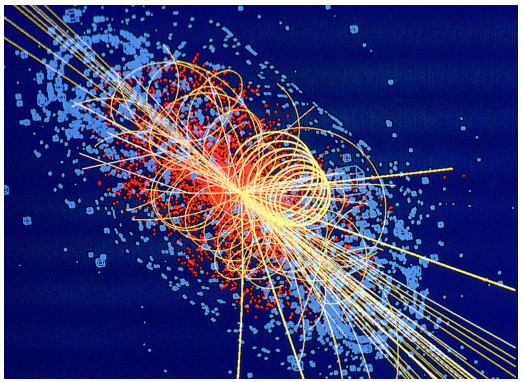
- 360° view of customers
  - Marketing, recommendation
    - Requires combining corporate (structured) data with external (unstructured) data (web, social networks, phone recordings, …)
- Online fraud detection across massive databases
  - E-commerce, banking, telephony, etc.
- National security
  - Signal intelligence, cyber analytics
- Medical science
  - Personalized medicine, with major investment from the GAFAM

Data Science
the good, the bad and the ugly

The Good: Higgs Boson @ CERN



- LHC (Large Hadron Collider)
  - Instrument to study the properties of fundamental particules in physics
  - Produces 15 petabytes / year
  - Made available through the LHC Computing Grid to several computing centers
    - E.g. CC-IN2P3, Lyon
  - Up to 200,000 simultaneous analyses



- High Boson discovery
  - 2012: CERN announces that it had discovered a particle that was probably a Higgs boson particle as predicted by the Standard Model of particle physics
  - 2014: CERN confirms the discovery

The Bad

amazon.fr Prime — Toutes nos catégories — patrick valduriez — SOLDES Jusqu'à -70% & Bons Plans

Votre adresse de livraison: Montpellier 34000 — Parcourir les catégories — Chez vous — Soldes & Ventes Flash — Chèques-cadeaux — Vendre — Aide — Bonjour. Identifiez-vous Votre compte — Testez Prime — Vos Listes — 0 Panier

Amazon.fr — Ventes Flash — Outlet — Meilleures ventes — Offres reconditionnées — Nos idées cadeaux — Services Amazon — Amazon Assistant — Coupons

1-16 sur 18 résultats pour "patrick valduriez" — Trier par Dernières nouveautés

Afficher les résultats pour

Livres
- Entreprise et Bourse
- Informatique & Internet
- Art, Musique & Cinéma

Voir les 3 catégories

Affiner par

Option d'expédition (Qu'est-ce?)
- prime
- Livraison gratuite

Formats livres
Broché

Auteurs
Georges Gardarin
Christian Soutou

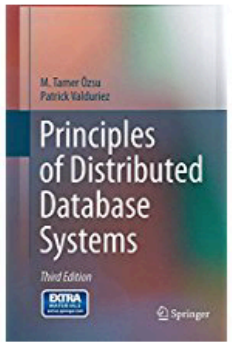
Livraison internationale (Qu'est-ce?)
- Livraison internationale disponible

Advances in Databases and Information Systems: 19th East European Conference, ADBIS 2015, Poitiers, France, September 8-11, 2015, Proceedings 14 août 2015
de Morzy Tadeusz et Patrick Valduriez

Format Kindle
47,26 €

Broché
67,52 € prime
En stock

Plus de choix d'achat
67,52 € (8 d'occasion & neufs)



Principles of Distributed Database Systems 23 août 2016
de M. Tamer OEzsu et Patrick Valduriez

Broché
105,06 €

Plus de choix d'achat
105,06 € (8 d'occasion & neufs)

Relié
183,90 €
Plus que 1 ex. Commandez vite !

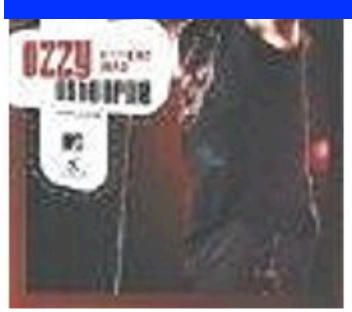
Plus de choix d'achat
96,32 € (4 d'occasion & neufs)

The Bad

Ozzy Osbourne - Talking. 31 août 2003
de Patrick Valduriez

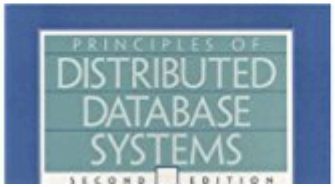
Broché
26,56 € (2 d'occasion & neufs)

Problem: how do I get this fixed?



Plus que 1 ex. Commandez vite !

Plus de choix d'achat
9,99 € (5 d'occasion & neufs)



Principles of Distributed Database Systems: United States Edition 19 janvier 1999
de M. Tamer Ozsu et Patrick Valduriez

Broché
60,25 € prime

19

The Bad at Wall Street

by CNNMoney Staff @CNNMoneyInvest

April 23, 2013: 7:06 PM ET

The tweet that cost $139 BILLION: Researchers analyse impact of hacked 2013 message claiming President Obama had been injured by White House explosion

- **Tweet was sent in 2013 after Syrian hackers accessed AP account**
- **Dow Jones Industrial Average dropped 143.5 points**
- **Standard & Poor's 500 Index lost more than $136 billion**
- **Researchers warned that algorithms were to blame for massive drop**

By MARK PRIGG FOR DAILYMAIL.COM

PUBLISHED: 00:07 GMT, 21 May 2015 | **UPDATED:** 00:10 GMT, 21 May 2015

The Ugly

A tiny company in Worcester, Mass., has paid the ultimate price for posting offensive T-shirts for sale online.

Fierce public backlash brought down Solid Gold Bomb, which made headlines in March for offering shirts that said "Keep Calm and Rape a Lot." The company closed its doors last week and let go its remaining three employees.

The Ugly

- Excerpts (from when big data goes bad)

Solid Gold Bomb, the company that made the shirt, wasn't necessarily aware that it was even selling it. Solid Gold Bomb's business isn't in artfully designing T-shirts. Instead, it writes code that takes libraries of words that

Problem: context-independent model, but context does matter!

template of a T-shirt and automatically get posted as an Amazon item for sale.

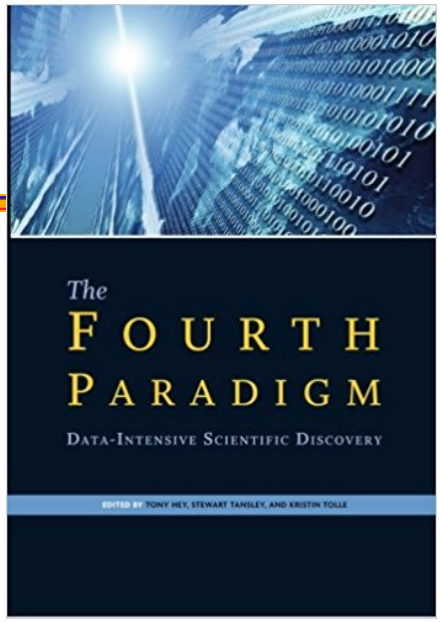
Their mistake was overlooking a single word in a list of 4,000 or so others.

Data Science Best Practices

- Have a data strategy
  - Business objectives, data requirements, data quality, data privacy, data governance
- Invest in a strong data team
  - Data-driven culture
  - Identify roles and skills
    - Data scientist, data analyst, data engineer, etc.
- Select the right technologies and tools
  - Choose between using off-the-shelf BI tools or develop code with Hadoop
- Design data processes and architectures
  - Must be evolutive and reusable across business lines

Science Data

Context: data-intensive science



- Modern science such as astronomy, biology and computational engineering must deal with overwhelming amounts of data
  - Generated by sensors, simulation programs, scientific instruments (LHC, gene sequencing tools, etc.)
- Increasingly, scientific breakthroughs will be powered by advanced computing capabilities that help researchers manipulate and explore these massive datasets

Requirements for Science Data

- As identified in recent interdisciplinary conferences (e.g. XLDB)
  - Support for result reproducibility, data sharing and collaboration
  - Scalability to hundreds of petabytes
  - Efficient metadata management to help integrating data coming from heterogeneous data sources
  - Open source software to insure data independence from proprietary systems

Science Data Management

The problem

"*Scientists are spending most of their time manipulating, organizing, finding and moving data, instead of researching. And it's going to get worse*"

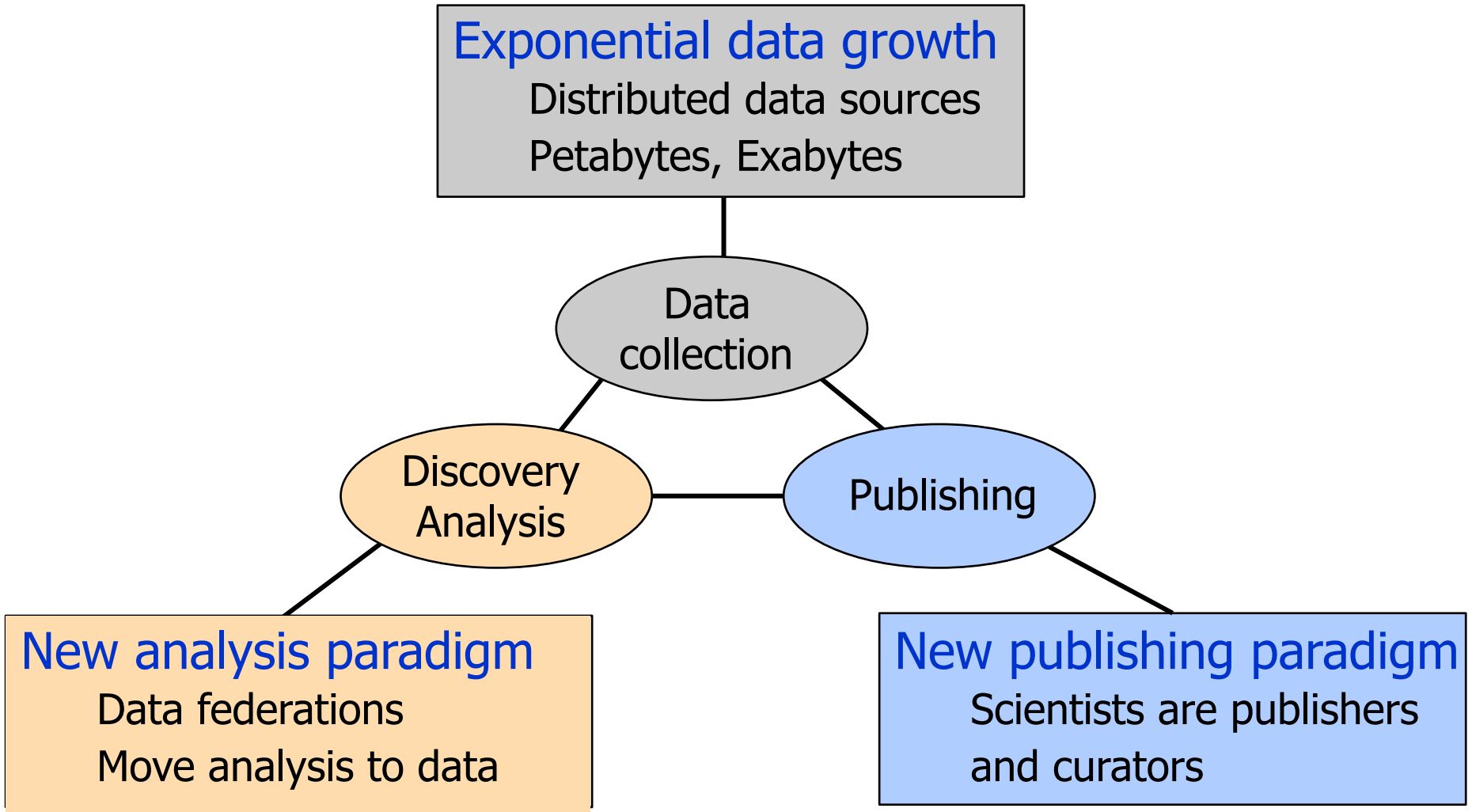
The Office Science Data Management Challenge (USA DoE 2004)

In bioinformatics, the time to deal with data can be well above 50% (IBC annual review 2017)

Challenges


scientific
method
research
hypothesis
science
experiment
test
concept
methodology
analysis
data
theory
biology
chemistry
physics
oxygen
sterile
medical
abstract
molecular
genetic
technical
methods
researcher
subject
futuristic
scientist
organic
invention
results
proof
measurement
examination
infrastructure
inquiry
genetically
transparency
consistent
laboratory
tools
equipment
observation
document
representation
reasoning
chemist
knowledge
study
conclusion
organism
engineering
imagination
discovery
energy
answer
computer
artificial
doctor
teacher
culture
analyzing
experimental
established
instrument
education
validation
future
exam
technology
integrity
system
history
contemporary
analyze
question
chemical
various
experimentation
interpretations
conceptual
determination
inspiration
lab
idea
professor
researchers
fields
thinking
innovation
information
context
atom
solution
unbiased
medicine
define
compound
investigation
success
definition

New Paradigms for Scientists



Source: Jim Gray (ACM Turing Award)

Science Data Sharing

- Scientific databases and web portals
  - Astronomy (SkyServer), Biology (GenBank), etc.
- Web portals
  - HAL, GoogleScholar, DBLP, data.gouv.fr, AgroPortal, …
- Data storage & computing platforms
  - BOINC (Seti@Home project), LHC Computing Grid, Grid5000, PlanetLab, etc.
- Towards open science
  - Data papers
  - Overlay journals, e.g. episcience.org
  - Crowdsourcing platforms, e.g. GalaxyZoo, Telabotanica

Impact on Scientific Practice

- Example in climate change
  - 97% of the papers in climate change research conclude that global warming is real
  - But what about those 3% of papers that reach contrary conclusions?
- A fascinating story: Learning from Mistakes in Climate Research
  - R.E. Benestad, et al. Theoretical and Applied Climatology 126: 699 (2016)
  - By the team of Katharine Hayhoe, director of the Climate Science Center at Texas Tech University

The Story

- The team tried to replicate the results of those 3% of sceptic papers
  - Looked at the 38 papers published in peer-reviewed journals in the last decade that denied global warming
- Findings
  - Every single one had an error, in their assumptions, methodology, or analysis
  - Many had cherry-picked the results that conveniently supported their conclusion, while ignoring other data
- Conclusion
  - Using the papers' data and after corrections of the errors, the results not only contradict the original papers but do support that global warming is real

The Ideal Scientific Publication

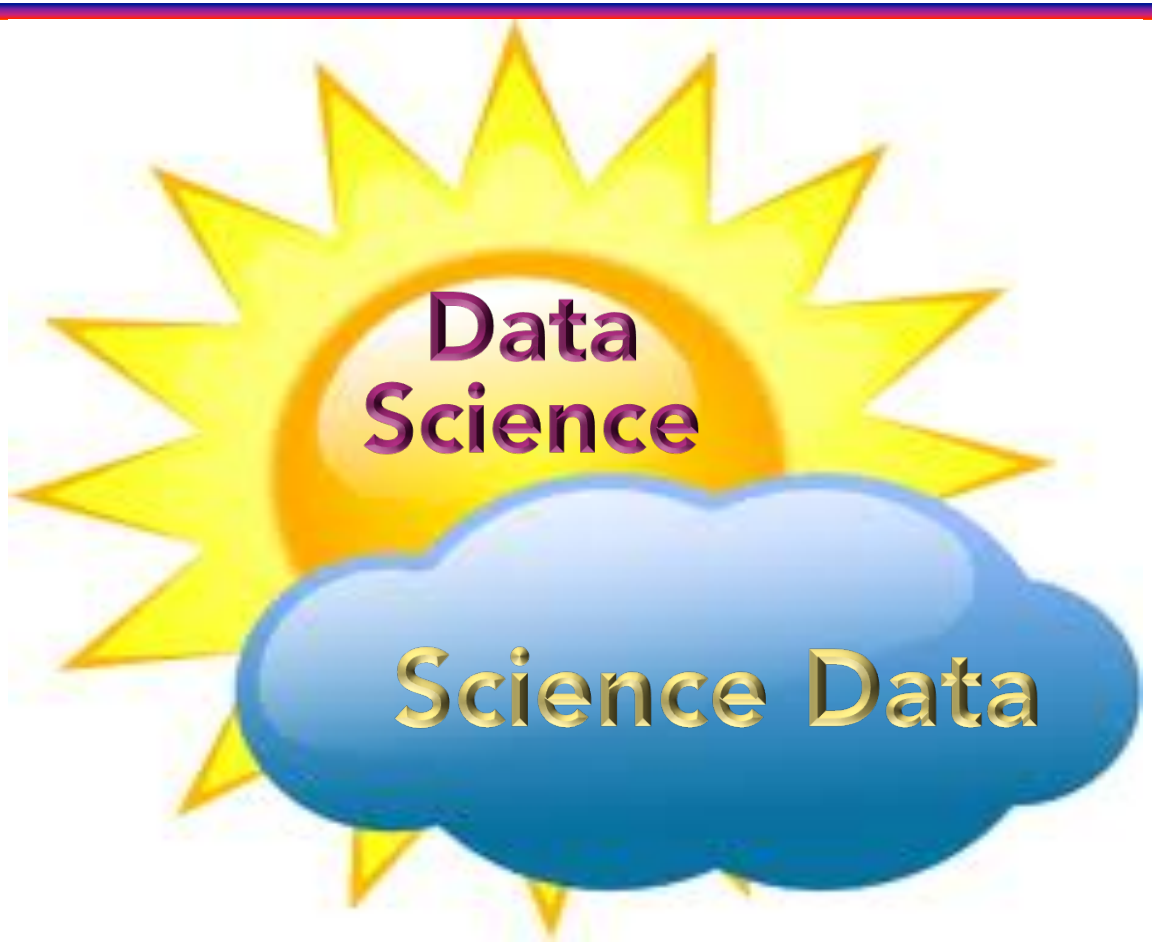
- Should foster reproducible results
  - This requires access to raw data sets, intermediate data, algorithms, programs and scripts
- Data papers: a first step forward
  - But looses the relationship with algorithms & code
- A solution by Jens Dittrich & Patrick Bender
  - Janiform Intra-Document Analytics for Reproducible Research. PVLDB 8(12): 1972-1975 (2015)
  - A janiform document, e.g. both static pdf and dynamic HTML, that embeds all data & code necessary to understand and reproduce the results

Opportunities

Towards Open, Responsible Science

- Better scientific quality & integrity
  - Requires implementing data science best practices - investing in strong data teams
  - And educating scientists in data science (continuum of understanding) and its impact on the world (values)
- Opening science to people
  - Data crowdsourcing
  - Open data / services platforms
- Better, faster knowledge discovery
  - Virtuous circle between machine learning and big data
  - Requires scientists' control to avoid bad stories

Thanks



Data Science Landscape

dave@vcdave.com

blogs.forbes.com/davefeinleib