Dark Data y los nuevos desafíos para Open Data y Open Science

Juan M. Durán

Delft University of Technology

The Netherlands

TUDelft

Outline

1. Dark Data
    - Contexto institucional
    - Definición y algunas consecuencias legales
    - El Scientific Data Officer (SDO), Open Data y Open Science
    - Consequencias éticas
2. - Big Data como política de estado

Contexto Institucional

Dark Data and the SDO

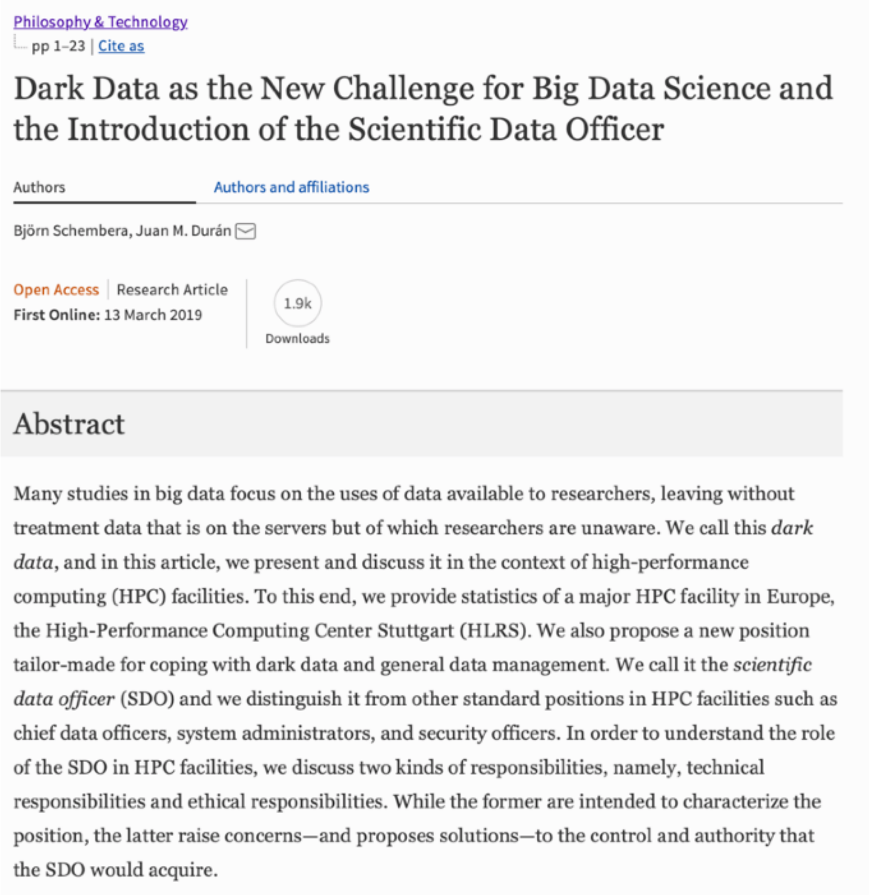
VOLUME 26 NO. 3 2013

PHILOSOPHY & TECHNOLOGY

Editor-in-Chief: Luciano Floridi

Springer

ISSN 2210-5433



Philosophy & Technology
pp 1–23 | Cite as

Dark Data as the New Challenge for Big Data Science and the Introduction of the Scientific Data Officer

Authors | Authors and affiliations

Björn Schembera, Juan M. Durán

Open Access | Research Article
First Online: 13 March 2019

1.9k
Downloads

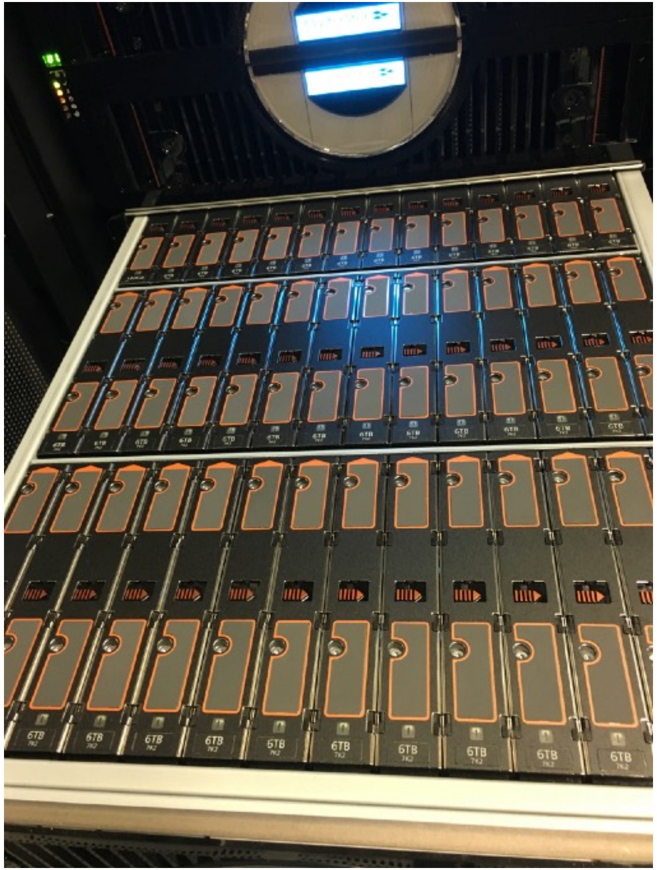
Abstract

Many studies in big data focus on the uses of data available to researchers, leaving without treatment data that is on the servers but of which researchers are unaware. We call this *dark data*, and in this article, we present and discuss it in the context of high-performance computing (HPC) facilities. To this end, we provide statistics of a major HPC facility in Europe, the High-Performance Computing Center Stuttgart (HLRS). We also propose a new position tailor-made for coping with dark data and general data management. We call it the *scientific data officer* (SDO) and we distinguish it from other standard positions in HPC facilities such as chief data officers, system administrators, and security officers. In order to understand the role of the SDO in HPC facilities, we discuss two kinds of responsibilities, namely, technical responsibilities and ethical responsibilities. While the former are intended to characterize the position, the latter raise concerns—and proposes solutions—to the control and authority that the SDO would acquire.

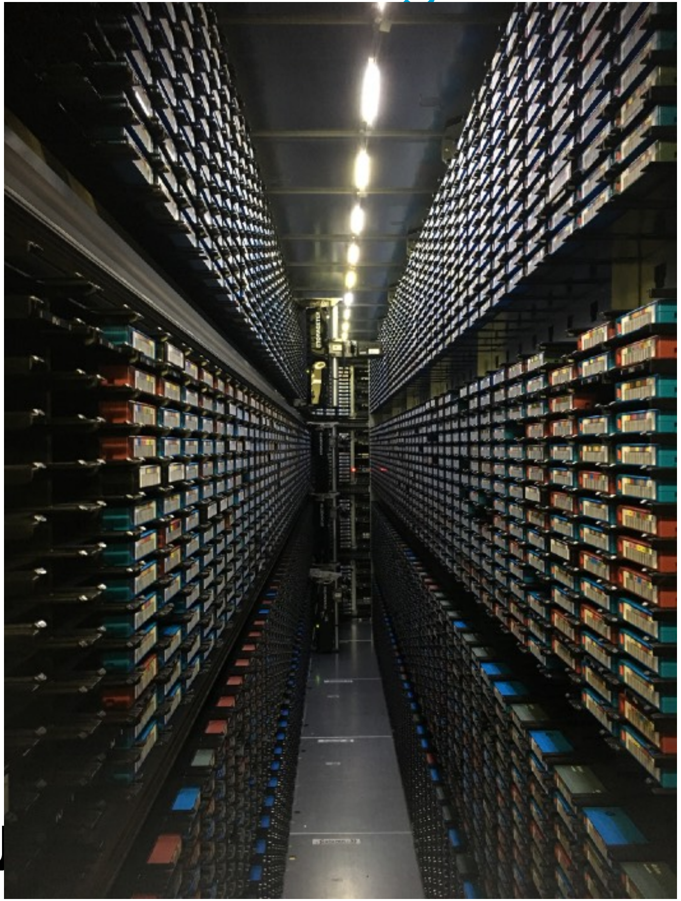
Höchstleistungsrechenzentrum Stuttgart (HLRS)

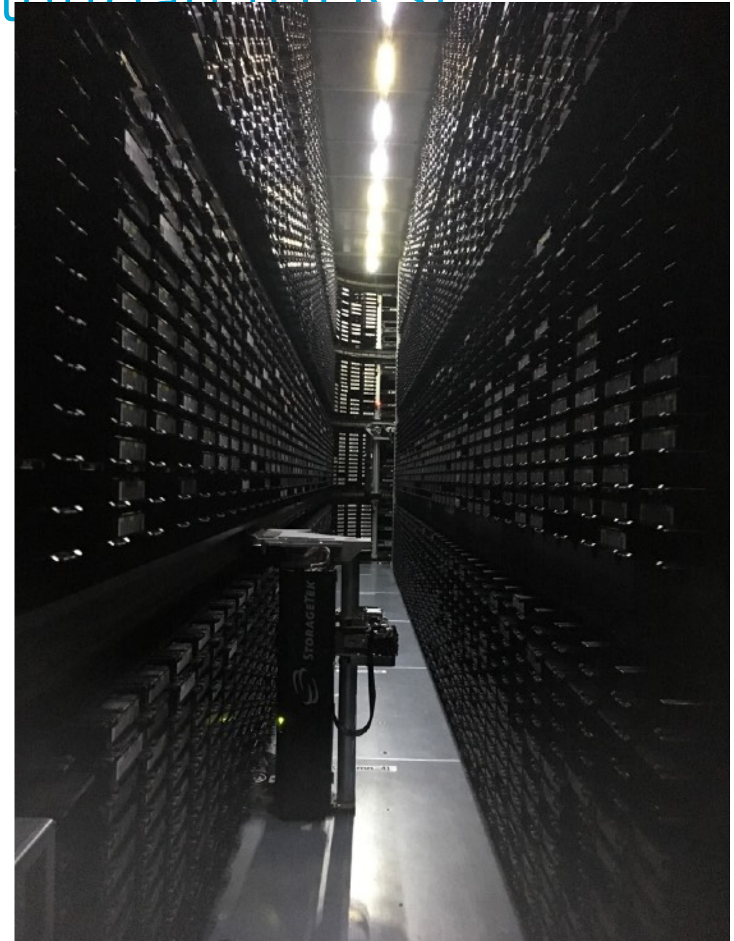
Cray XC40 Hazel-Hen

Höchstleistungsrechenzentrum Stuttgart (HLRS)



Höchstleistungsrechenzentrum Stuttgart (HLRS)

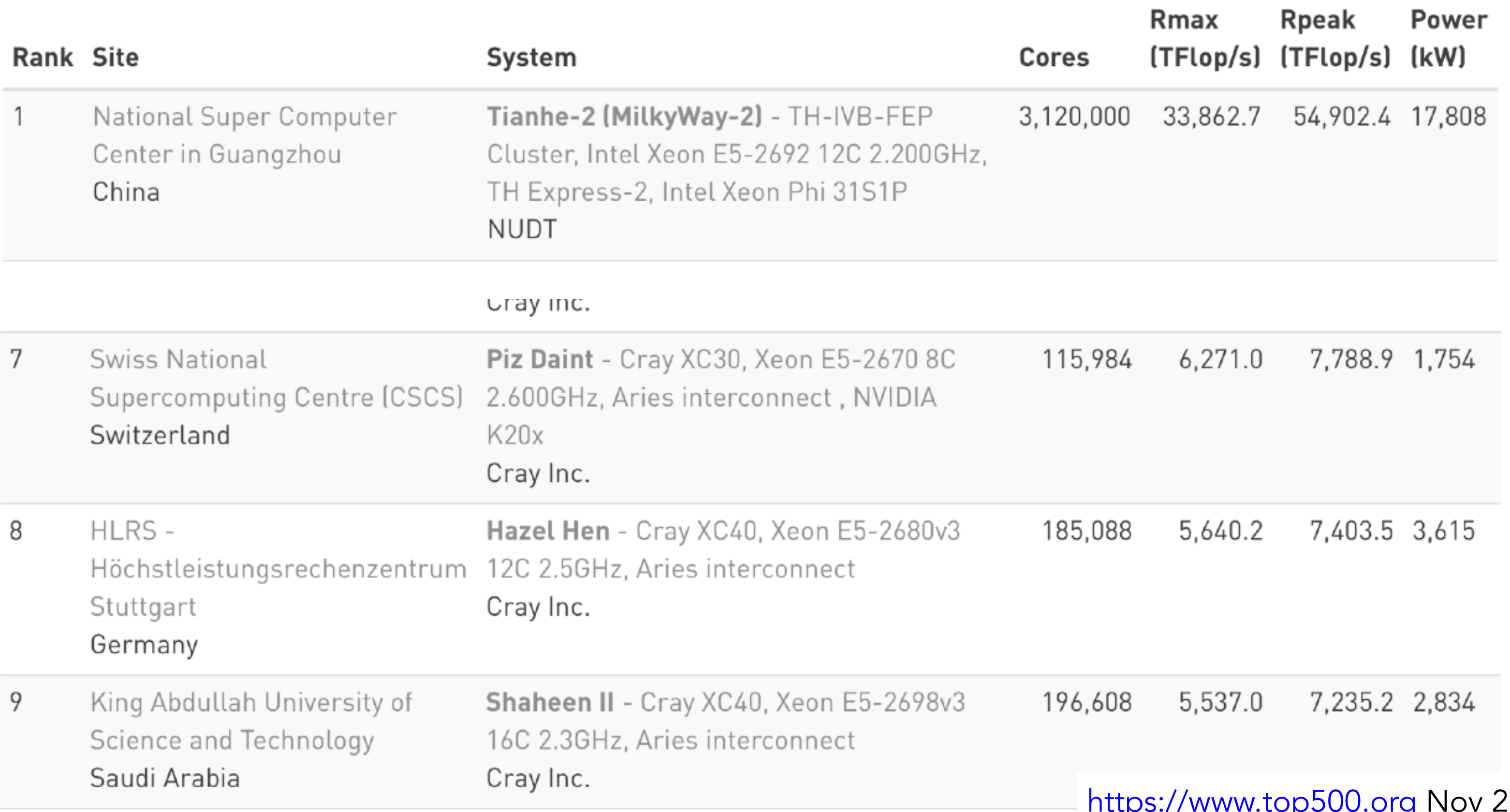




Höchstleistungsrechenzentrum Stuttgart (HLRS)

- Octubre 2015
- 5,6 Peta-FLOPS per second (peak performance = 7,4 Peta-FLOPS)
- 185,088 cores
- Noviembre 2017 (3ra computadora más rápida en la UE)
- Noviembre 2015 8va más rápida del mundo. Julio 2019 cae al puesto 34.
- En 2019 tendrán una nueva “Hawk” con un costo ca. 38M€

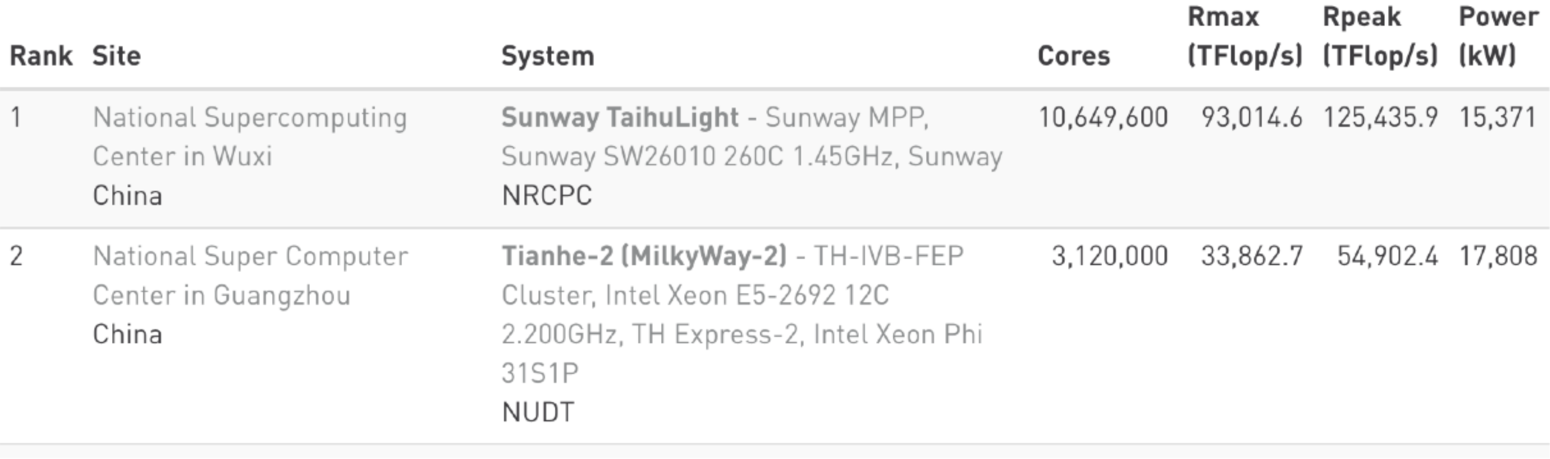
Höchstleistungsrechenzentrum Stuttgart (HLRS)

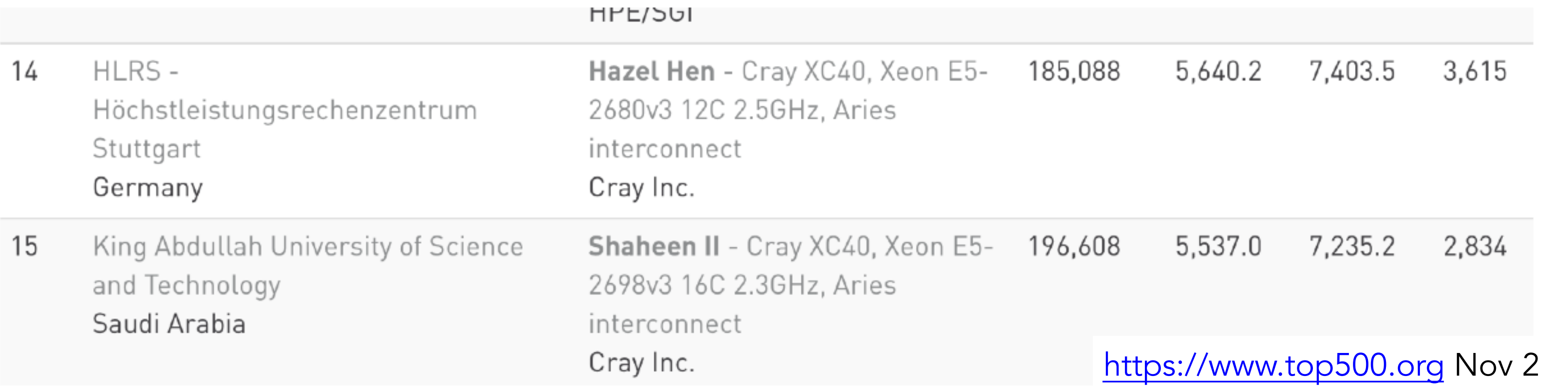


Rank	Site	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	National Super Computer Center in Guangzhou China	**Tianhe-2 (MilkyWay-2)** - TH-IVB-FEP Cluster, Intel Xeon E5-2692 12C 2.200GHz, TH Express-2, Intel Xeon Phi 31S1P NUDT	3,120,000	33,862.7	54,902.4	17,808
		Cray Inc.				
7	Swiss National Supercomputing Centre (CSCS) Switzerland	**Piz Daint** - Cray XC30, Xeon E5-2670 8C 2.600GHz, Aries interconnect , NVIDIA K20x Cray Inc.	115,984	6,271.0	7,788.9	1,754
8	HLRS - Höchstleistungsrechenzentrum Stuttgart Germany	**Hazel Hen** - Cray XC40, Xeon E5-2680v3 12C 2.5GHz, Aries interconnect Cray Inc.	185,088	5,640.2	7,403.5	3,615
9	King Abdullah University of Science and Technology Saudi Arabia	**Shaheen II** - Cray XC40, Xeon E5-2698v3 16C 2.3GHz, Aries interconnect Cray Inc.	196,608	5,537.0	7,235.2	2,834

https://www.top500.org Nov 2015

Höchstleistungsrechenzentrum Stuttgart (HLRS)

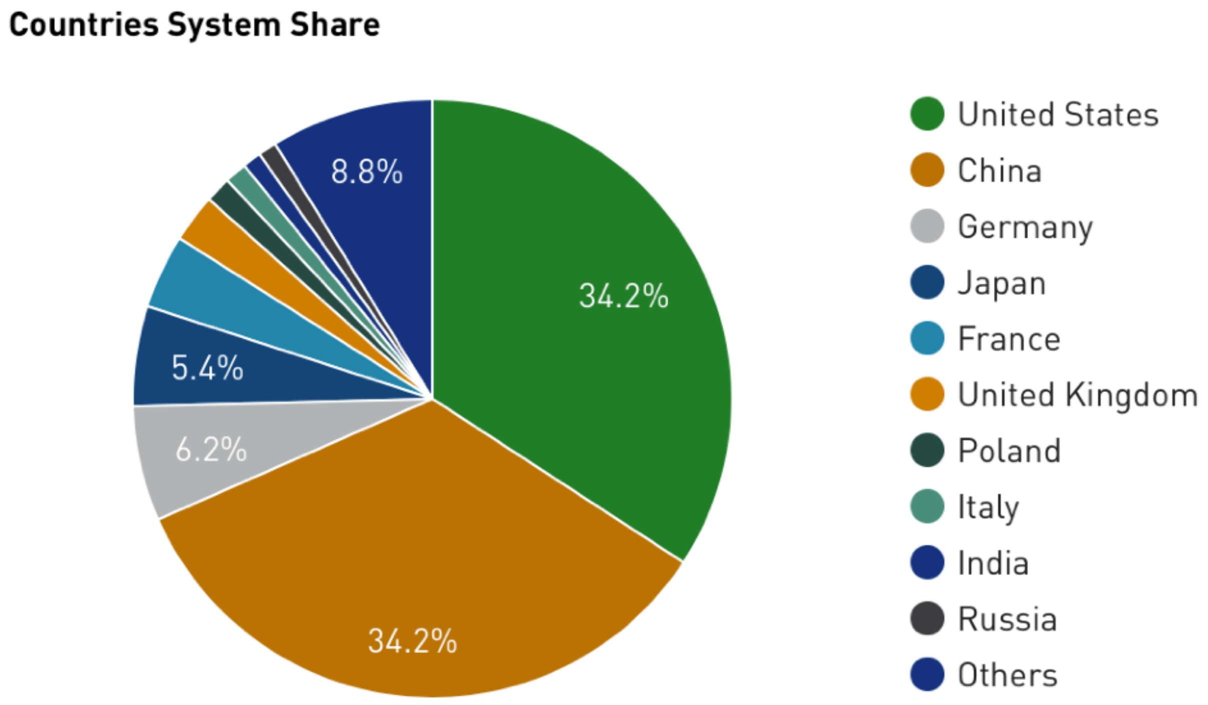




Rank	Site	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	National Supercomputing Center in Wuxi China	**Sunway TaihuLight** - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway NRCPC	10,649,600	93,014.6	125,435.9	15,371
2	National Super Computer Center in Guangzhou China	**Tianhe-2 (MilkyWay-2)** - TH-IVB-FEP Cluster, Intel Xeon E5-2692 12C 2.200GHz, TH Express-2, Intel Xeon Phi 31S1P NUDT	3,120,000	33,862.7	54,902.4	17,808
		HPE/SGI				
14	HLRS - Höchstleistungsrechenzentrum Stuttgart Germany	**Hazel Hen** - Cray XC40, Xeon E5-2680v3 12C 2.5GHz, Aries interconnect Cray Inc.	185,088	5,640.2	7,403.5	3,615
15	King Abdullah University of Science and Technology Saudi Arabia	**Shaheen II** - Cray XC40, Xeon E5-2698v3 16C 2.3GHz, Aries interconnect Cray Inc.	196,608	5,537.0	7,235.2	2,834

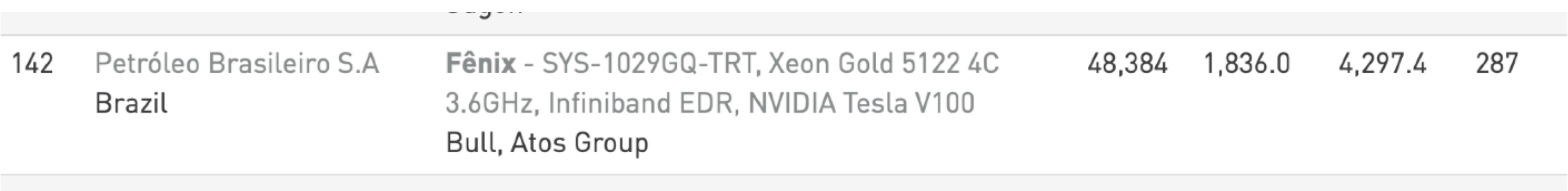
https://www.top500.org Nov 2016

Distribución de HPC

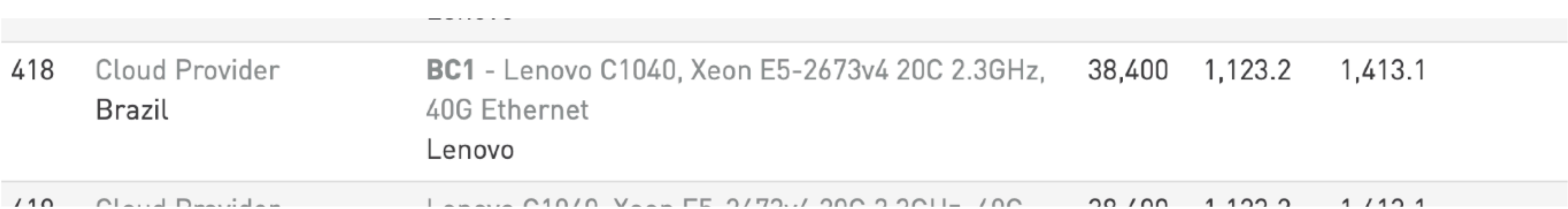


https://www.top500.org Nov 2016

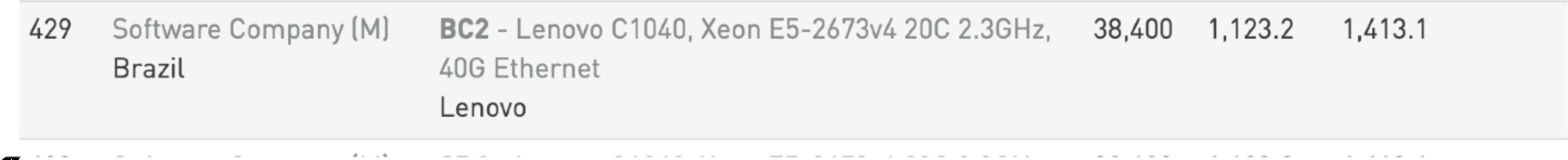
AméricaLatina (bueno, Brasil…)



142	Petróleo Brasileiro S.A Brazil	**Fênix** - SYS-1029GQ-TRT, Xeon Gold 5122 4C 3.6GHz, Infiniband EDR, NVIDIA Tesla V100 Bull, Atos Group	48,384	1,836.0	4,297.4	287



418	Cloud Provider Brazil	**BC1** - Lenovo C1040, Xeon E5-2673v4 20C 2.3GHz, 40G Ethernet Lenovo	38,400	1,123.2	1,413.1
419	Cloud Provider	Lenovo C1040, Xeon E5-2673v4 20C 2.3GHz, 40G	38,400	1,123.2	1,413.1



429	Software Company (M) Brazil	**BC2** - Lenovo C1040, Xeon E5-2673v4 20C 2.3GHz, 40G Ethernet Lenovo	38,400	1,123.2	1,413.1

https://www.top500.org Oct 2019

Dark Data

Dark Data - una definición posible

“Dark data” is [data] not carefully indexed and stored so it becomes nearly invisible to scientists and other potential users and therefore is more likely to remain under-utilised and eventually lost [..] the type of data that exists only in the bottom left-hand desk drawer of scientists on some media that is quickly ageing and soon will be unreadable by commonly available devices” (Heidorn 2008, 281)

Dark Data - una definición posible

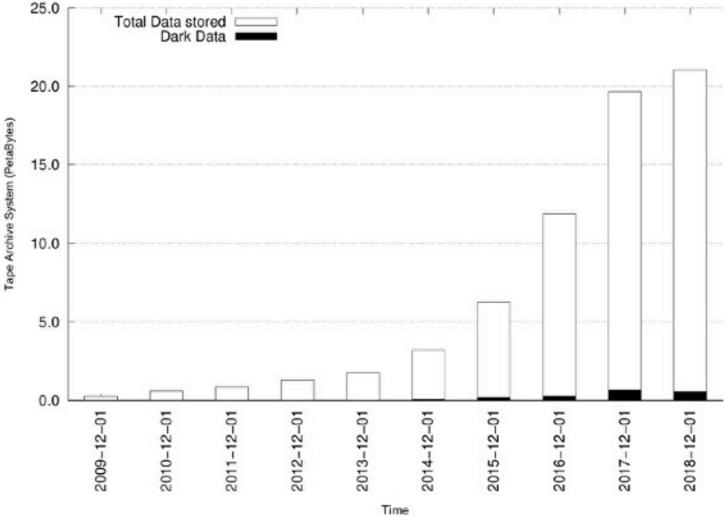
“Dark data” is [data] not carefully **indexed** and stored so it becomes nearly **invisible** to scientists and other potential users and therefore is more likely to remain **under-utilised** and eventually **lost** [..] the type of data that exists only in the bottom left-hand desk drawer of scientists on some media that is quickly ageing and soon will be **unreadable** by commonly available devices” (Heidorn 2008, 281)

Dark Data - Metadata

Para nosotros“dark data” también son:

- Datos que no tienen meta-datos (i.e., no han sido etiquetados), que están mal organizados (tal vez por limitaciones en el filesystem), que están dispersos.
    - Son difícil de contabilizar: definiciones y mediciones son necesarias

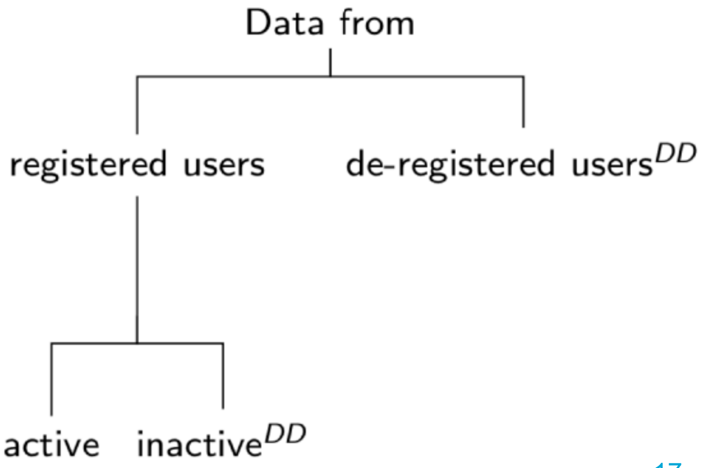
Dec-2017	668.03 TB / 19,663 TB (3,40 % DD)
Mar-2019	1,021.40 TB / 21,240 TB (4,81 % DD)



Dark Data - Usuarios

Para nosotros “dark data” también un origen en los usuarios del sistema:

- Datos guardados que persisten en el sistema de archivos por usuarios inactivos o des-registrados
  - Cómo determinar que un usuario no está más activo?
    - Tiempo inactivo?
    - Cantidad de datos transferidos?



Por qué hay presencia de Dark Data?

- Mantener los datos y los metadatos no motiva, ni profesionalmente, ni económicamente, ni socialmente. En otras palabras, por qué utilizar “tiempo precioso” para organizar mis datos (más allá de lo necesario para que yo pueda acceder a esos datos)
- Centros de HPC adoptan nueva infraestructura, nuevos in-house sistemas de archivos, contribuye a los dark data.

Implicaciones del Dark Data

- Recursos:
  - Dark Data desperdicia recursos (almacenamiento, tiempo de cómputo, mantenimiento)
  - 500TB de Dark Data cuestan ca. $ 10,000 dependiendo de la tecnología de almacenamiento de datos.
  - Se invierte dinero en educar a un ingeniero/científico.

- En varios sentidos el Dark Data viola regulaciones del GDPR
  - e.g., cuando los datos contienen información personal
  - e.g., cuando los datos tienen derechos de propiedad

- Dark Data no es FAIR: **F**indable, **A**ccessible, **I**nter-operable, **R**eusable (por definición!)

Scientific Data Officer (SDO)

Scientific Data Officer: necesidad & responsabilidades

- Etiquetar (o dar soporte en el etiquetado) de datos con algún formato estandarizado de metadatos de acuerdo con FAIR (inexistente!)
- Estandarizar los datos y metadatos para todos los proyectos en un centro de HPC
- Reduce Dark Data
- Chequeos periódicos del inventario de datos (y de Dark Data)
- Administración del sistema de datos
- Autoridad sobre el problema de decidir sobre usuarios inactivos/des-registrados.

Scientific Data Officer: responsabilidades

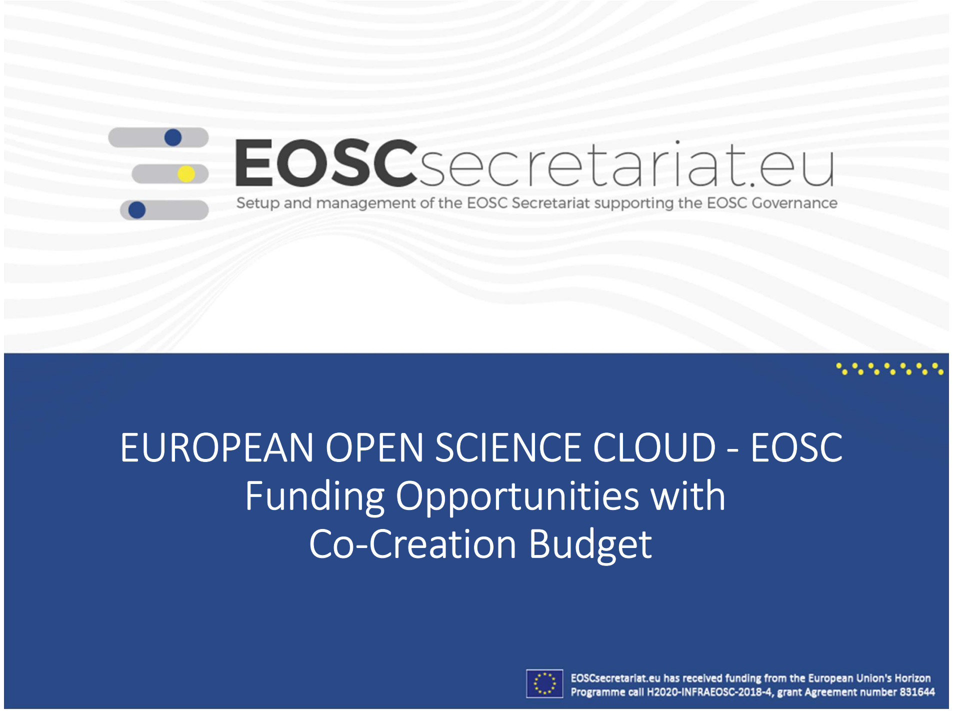
- Stewardship/curadora de datos
- Multiplicadora de conocimiento y su transferencia (no se necesita más tener "el contacto" o conocer "al juez")
- Rol mediadora (e.g., comparte información y datos intra e inter-institucional)
    - Una red de SDOs?
- Dar apertura a los datos (cumplir con los ideales de OpenScience and OpenData) y cumplir con FAIR
- La SDO debe observar y hacer respetar reglas de buena práctica científica (incluido la autoridad de sancionar y accionar legalmente?).

Scientific Data Officer: Problemas éticos

- Se presentan varias consideraciones sobre el comportamiento de las SDOs:
  - Malversación de los resultados de una investigación (embezzlement/ missappropriation)
    - Manipulation y selección de datos
  - Interferencia en la producción de resultados
  - Obstrucción en la investigación científica
    - Priorizar el acceso al sistema de archivos a ciertos grupos en desventaja de otros
  - No cumplimientos con códigos de conducta y éticos
  - Gatekeeping: determina quién continua con una pertenencia a la institución

Big Data como política de estado

Making Dark Data FAIR



EOSCsecretariat.eu

Setup and management of the EOSC Secretariat supporting the EOSC Governance

EUROPEAN OPEN SCIENCE CLOUD - EOSC
Funding Opportunities with
Co-Creation Budget

EOSCsecretariat.eu has received funding from the European Union's Horizon Programme call H2020-INFRAEOSC-2018-4, grant Agreement number 831644

SoBigDataPlusPlus

Next Events in Bruxelles

Conference on Artificial Intelligence:
The European Approach for Citizens' Wellbeing

November 19th 2019, 08:30 to 12:00
Room ASP 5E2, European Parliament

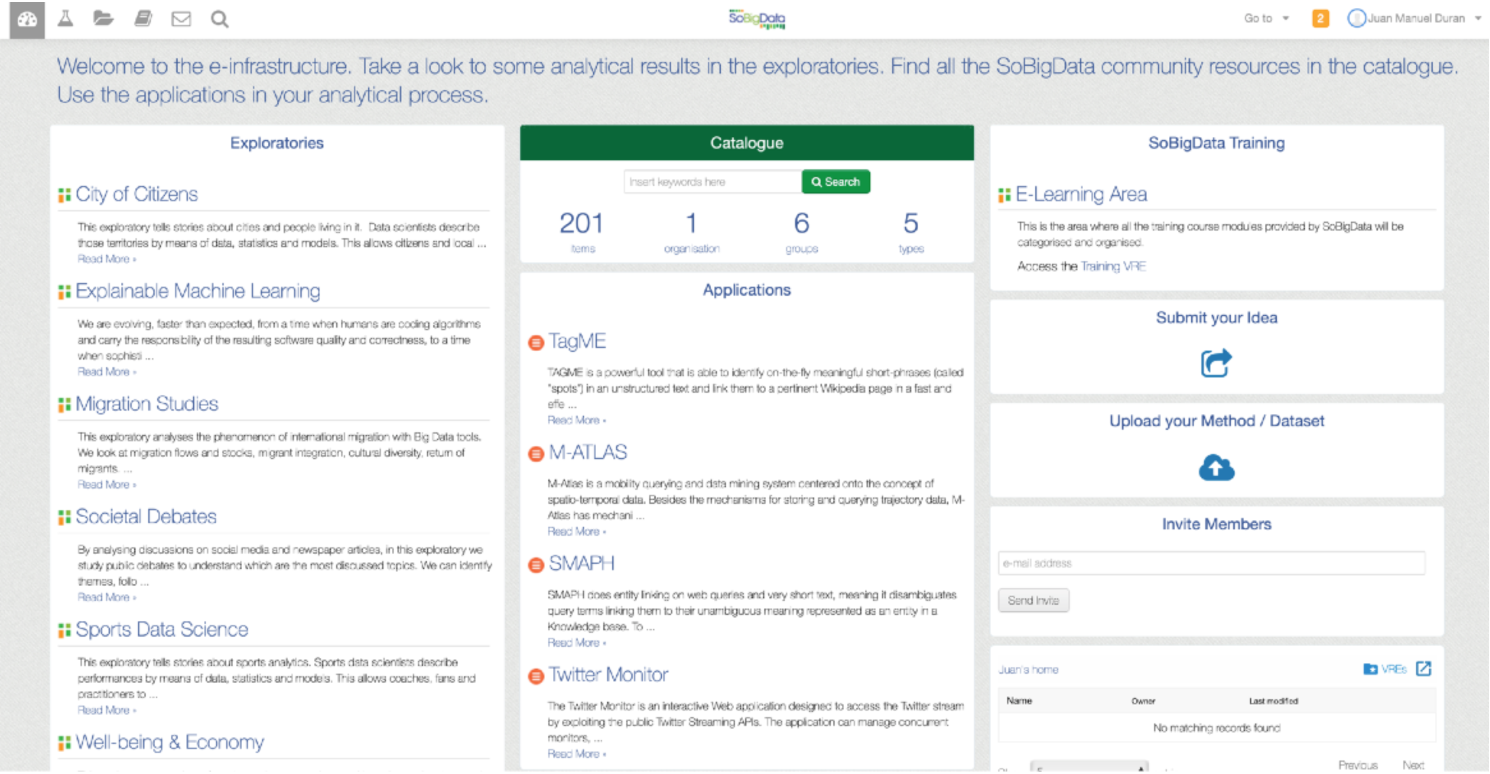
From SoBigData to SoBigData++:
Toward an Advanced Community of Data Scientists

November 19th 2019, 14:00 to 18:00
Meeting Room, Regione Toscana
Rond Point Schuman 14

SoBigData is the European Research Infrastructure for **Big Data** and Social Mining.
From data to knowledge, investigating stories **ethically**, paying attention to citizens privacy.

http://www.sobigdata.eu/index

SoBigDataPlusPlus



Welcome to the e-infrastructure. Take a look to some analytical results in the exploratories. Find all the SoBigData community resources in the catalogue. Use the applications in your analytical process.

Exploratories

City of Citizens

This exploratory tells stories about cities and people living in it. Data scientists describe those territories by means of data, statistics and models. This allows citizens and local ...
Read More »

Explainable Machine Learning

We are evolving, faster than expected, from a time when humans are coding algorithms and carry the responsibility of the resulting software quality and correctness, to a time when sophisti ...
Read More »

Migration Studies

This exploratory analyses the phenomenon of international migration with Big Data tools. We look at migration flows and stocks, migrant integration, cultural diversity, return of migrants. ...
Read More »

Societal Debates

By analysing discussions on social media and newspaper articles, in this exploratory we study public debates to understand which are the most discussed topics. We can identify themes, follo ...
Read More »

Sports Data Science

This exploratory tells stories about sports analytics. Sports data scientists describe performances by means of data, statistics and models. This allows coaches, fans and practitioners to ...
Read More »

Well-being & Economy

Catalogue

Insert keywords here | Search

201	1	6	5
items	organisation	groups	types

Applications

TagME

TAGME is a powerful tool that is able to identify on-the-fly meaningful short-phrases (called "spots") in an unstructured text and link them to a pertinent Wikipedia page in a fast and effe ...
Read More »

M-ATLAS

M-Atlas is a mobility querying and data mining system centered onto the concept of spatio-temporal data. Besides the mechanisms for storing and querying trajectory data, M-Atlas has mechani ...
Read More »

SMAPH

SMAPH does entity linking on web queries and very short text, meaning it disambiguates query terms linking them to their unambiguous meaning represented as an entity in a Knowledge base. To ...
Read More »

Twitter Monitor

The Twitter Monitor is an interactive Web application designed to access the Twitter stream by exploiting the public Twitter Streaming APIs. The application can manage concurrent monitors, ...
Read More »

SoBigData Training

E-Learning Area

This is the area where all the training course modules provided by SoBigData will be categorised and organised.

Access the Training VRE

Submit your Idea

Upload your Method / Dataset

Invite Members

e-mail address

Send Invite

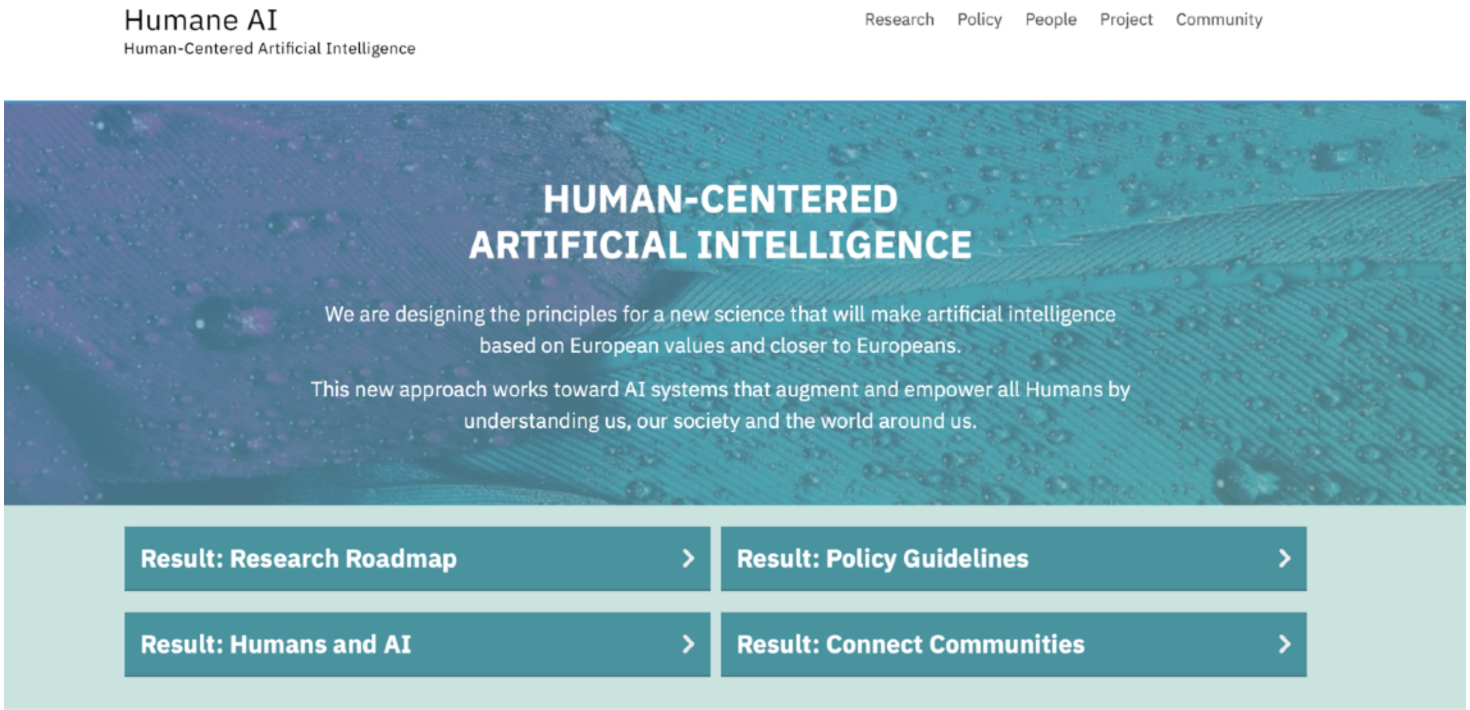
Juan's home | VREs

Name	Owner	Last modified
No matching records found		

Previous Next

https://sobigdata.d4science.org/group/sobigdata-gateway

HumaineAI



Humane AI

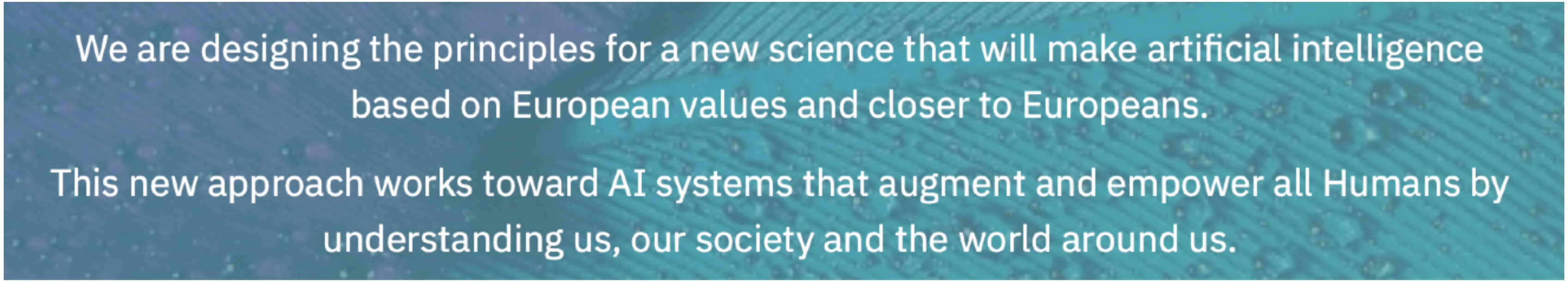
Human-Centered Artificial Intelligence

HUMAN-CENTERED ARTIFICIAL INTELLIGENCE

We are designing the principles for a new science that will make artificial intelligence based on European values and closer to Europeans.

This new approach works toward AI systems that augment and empower all Humans by understanding us, our society and the world around us.

HumaineAI



We are designing the principles for a new science that will make artificial intelligence based on European values and closer to Europeans.

This new approach works toward AI systems that augment and empower all Humans by understanding us, our society and the world around us.

HumaineAI

- Valores Europeos?
- China y EEUU:
  - China: Estado policíaco de vigilancia (surveillance state). Para 2018 China ya tenía una red de vigilancia de más de 170 millions de cámaras CCTV, con aprox. 400 million de nuevas cámaras a instalarse en los próximos tres años (el uso es el “facial recognition technology”) - Financiamieno es “virtualmene infinito”

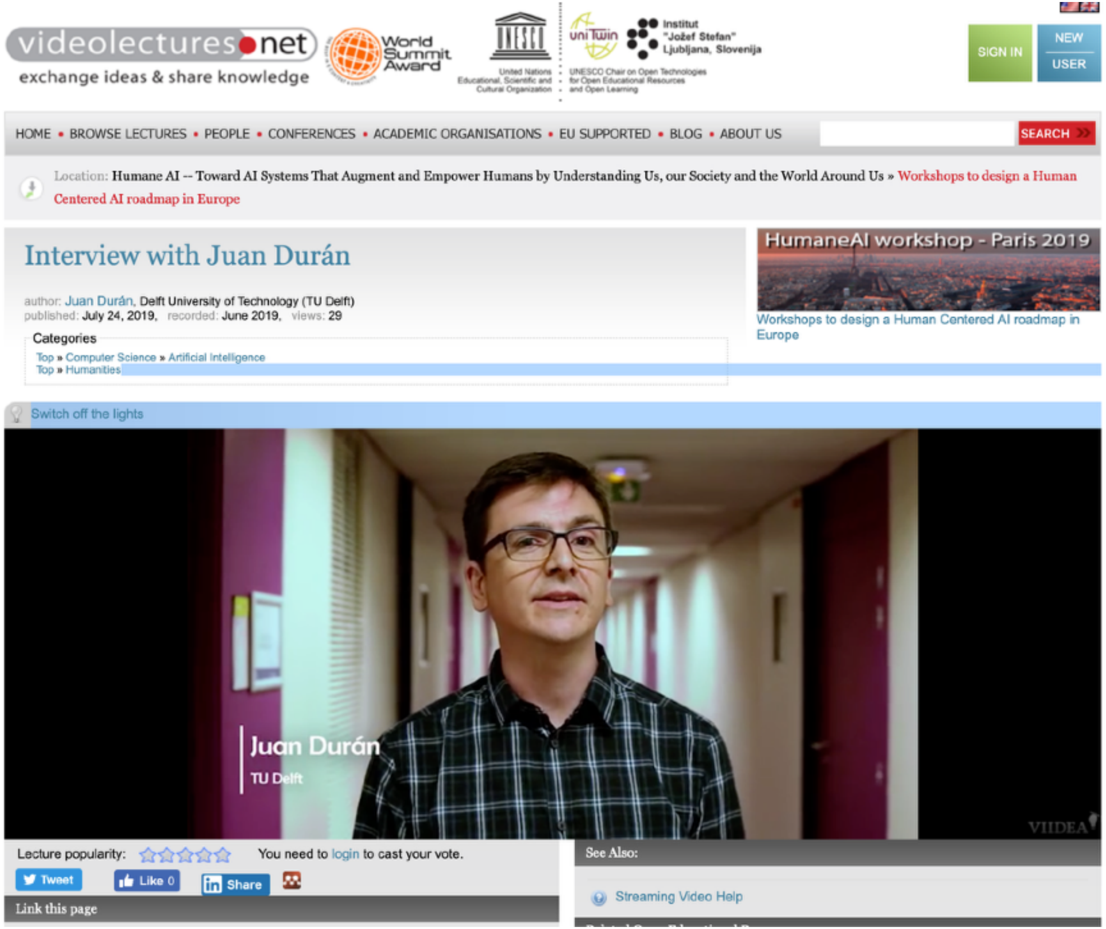
HumaineAI

- Valores Europeos?
- China y EEUU:
  - EEUU: El gran motor en Big Data es la industria, con sus valores capitalistas y con fines comerciales y obtención de ganancias. Esto, en principio, podría afectar libertades y derechos adquiridos, como el derecho a la privacidad, a la protección del estado, al acceso a tecnologías, etc.

HumaineAI

- Valores Europeos
    - Estado presente regulando la industria pero a su vez monitoreado por ONG y agencias independientes que garanticen los derechos de todos los ciudadanos.
    - GDPR

HumaineAI


videolectures.net
exchange ideas & share knowledge
World Summit Award
United Nations Educational, Scientific and Cultural Organization
UNESCO Chair on Open Technologies for Open Educational Resources and Open Learning
Institut "Jožef Stefan" Ljubljana, Slovenija
SIGN IN
NEW USER
HOME • BROWSE LECTURES • PEOPLE • CONFERENCES • ACADEMIC ORGANISATIONS • EU SUPPORTED • BLOG • ABOUT US
SEARCH
Switch off the lights
Juan Durán
TU Delft
VIIDEA
Tweet
Like 0
Share
Link this page
See Also:
Streaming Video Help

Location: Humane AI -- Toward AI Systems That Augment and Empower Humans by Understanding Us, our Society and the World Around Us » Workshops to design a Human Centered AI roadmap in Europe

Interview with Juan Durán

author: Juan Durán, Delft University of Technology (TU Delft)
published: July 24, 2019, recorded: June 2019, views: 29

Categories

Top » Computer Science » Artificial Intelligence
Top » Humanities

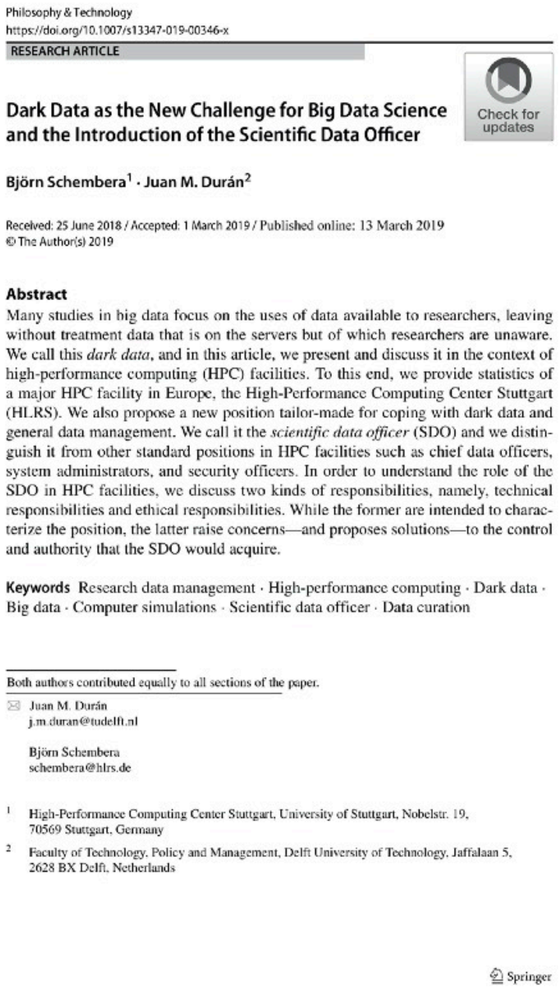
HumaneAI workshop - Paris 2019

Workshops to design a Human Centered AI roadmap in Europe

Lecture popularity: ☆☆☆☆☆ You need to login to cast your vote.

http://videolectures.net/humaneai_duran_interview/

Más información →



Philosophy & Technology
https://doi.org/10.1007/s13347-019-00346-x
RESEARCH ARTICLE

Check for updates

Dark Data as the New Challenge for Big Data Science and the Introduction of the Scientific Data Officer

Björn Schembera[1] · Juan M. Durán[2]

Received: 25 June 2018 / Accepted: 1 March 2019 / Published online: 13 March 2019

Abstract

Many studies in big data focus on the uses of data available to researchers, leaving without treatment data that is on the servers but of which researchers are unaware. We call this *dark data*, and in this article, we present and discuss it in the context of high-performance computing (HPC) facilities. To this end, we provide statistics of a major HPC facility in Europe, the High-Performance Computing Center Stuttgart (HLRS). We also propose a new position tailor-made for coping with dark data and general data management. We call it the *scientific data officer* (SDO) and we distinguish it from other standard positions in HPC facilities such as chief data officers, system administrators, and security officers. In order to understand the role of the SDO in HPC facilities, we discuss two kinds of responsibilities, namely, technical responsibilities and ethical responsibilities. While the former are intended to characterize the position, the latter raise concerns—and proposes solutions—to the control and authority that the SDO would acquire.

Keywords Research data management · High-performance computing · Dark data · Big data · Computer simulations · Scientific data officer · Data curation

Both authors contributed equally to all sections of the paper.

✉ Juan M. Durán
j.m.duran@tudelft.nl

Björn Schembera
schembera@hlrs.de

[1] High-Performance Computing Center Stuttgart, University of Stuttgart, Nobelstr. 19, 70569 Stuttgart, Germany

[2] Faculty of Technology, Policy and Management, Delft University of Technology, Jaffalaan 5, 2628 BX Delft, Netherlands

Springer

Muchas Gracias

Juan M. Durán - TU Delft

j.m.duran@tudelft.nl

juanmduran.net

TU Delft