Ciència de dades i ciberseguretat

Laia Subirats Maté

Barcelona Activa

Ajuntament de Barcelona

Missió

Barcelona Activa vol contribuir a la millora de la qualitat de vida de les persones promovent la competitivitat econòmica de la ciutat i el reequilibri dels territoris, a través del foment de l'ocupació de qualitat, l'impuls a l'emprenedoria i el suport a un teixit empresarial plural, divers i sostenible, des d'una perspectiva econòmica, social i ambiental.

Visió

Les accions de l'entitat es regeixen per la voluntat de fer de Barcelona una ciutat de referència internacional per treballar, emprendre i viure amb valors socials i ambientals.

Valors

L'activitat de Barcelona Activa es desenvolupa al voltant d'aquests 4 valors:

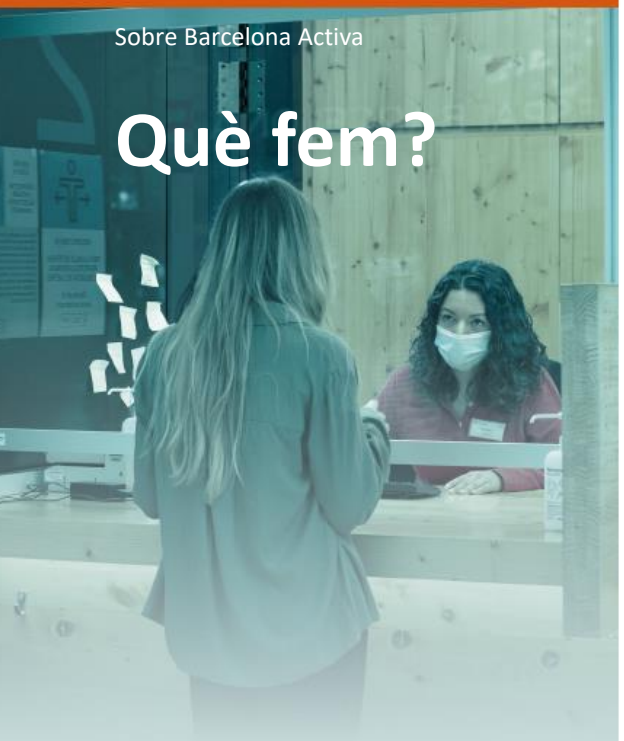
1. Igualtat d'oportunitats i progrés social

2. Cooperació dins l'organització i amb altres

3. Economia al servei de les persones: social i sostenible

4. Esperit de servei públic i ètica professional i personal

Què fem?



Acompanyem les persones que cerquen feina durant tot el procés

barcelonactiva.cat/treball

Donem suport a les persones emprenedores per fer realitat la seva idea de negoci

barcelonactiva.cat/emprenedoria

Ajudem les empreses i organitzacions a impulsar la seva activitat

barcelonactiva.cat/empreses

Oferim formació a les persones que cerquen feina, emprenedores i professionals.

barcelonactiva.cat/formacio

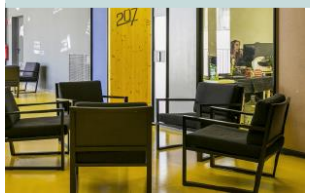
PROXIMITAT, INNOVACIÓ I IGUALTAT

Seu Central
de Barcelona Activa

Incubadora Glòries
Acull empreses de nova creació

Oficina d’Atenció a les Empreses
L’espai de referència per a les empreses i pimes de Barcelona

Porta22
2.000 m^2 per a l’ocupació i la carrera professional



Incubadora Almogàvers
Incubadora publicoprivada per a empreses innovadores

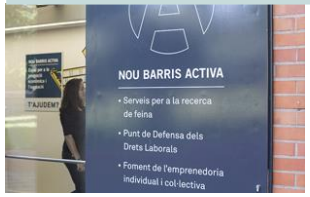
Parc Tecnològic
Clúster d’innovació i capacitació tecnològica al nord de la ciutat

Convent de Sant Agustí
Centre referent en ocupació per a joves a Ciutat Vella

Incubadora MediaTIC
Allotja empreses d’alt impacte tecnològic

Cibernàrium 22@
Centre de capacitació tecnològica i IT Academy

Ca n’Andalet
Espai per a la formació i l’ocupació a Horta-Guinardó



Nou Barris Activa
L’espai de referència per a la promoció econòmica i l’ocupació a Nou Barris

InnoBA
Centre municipal per a la Innovació socioeconòmica

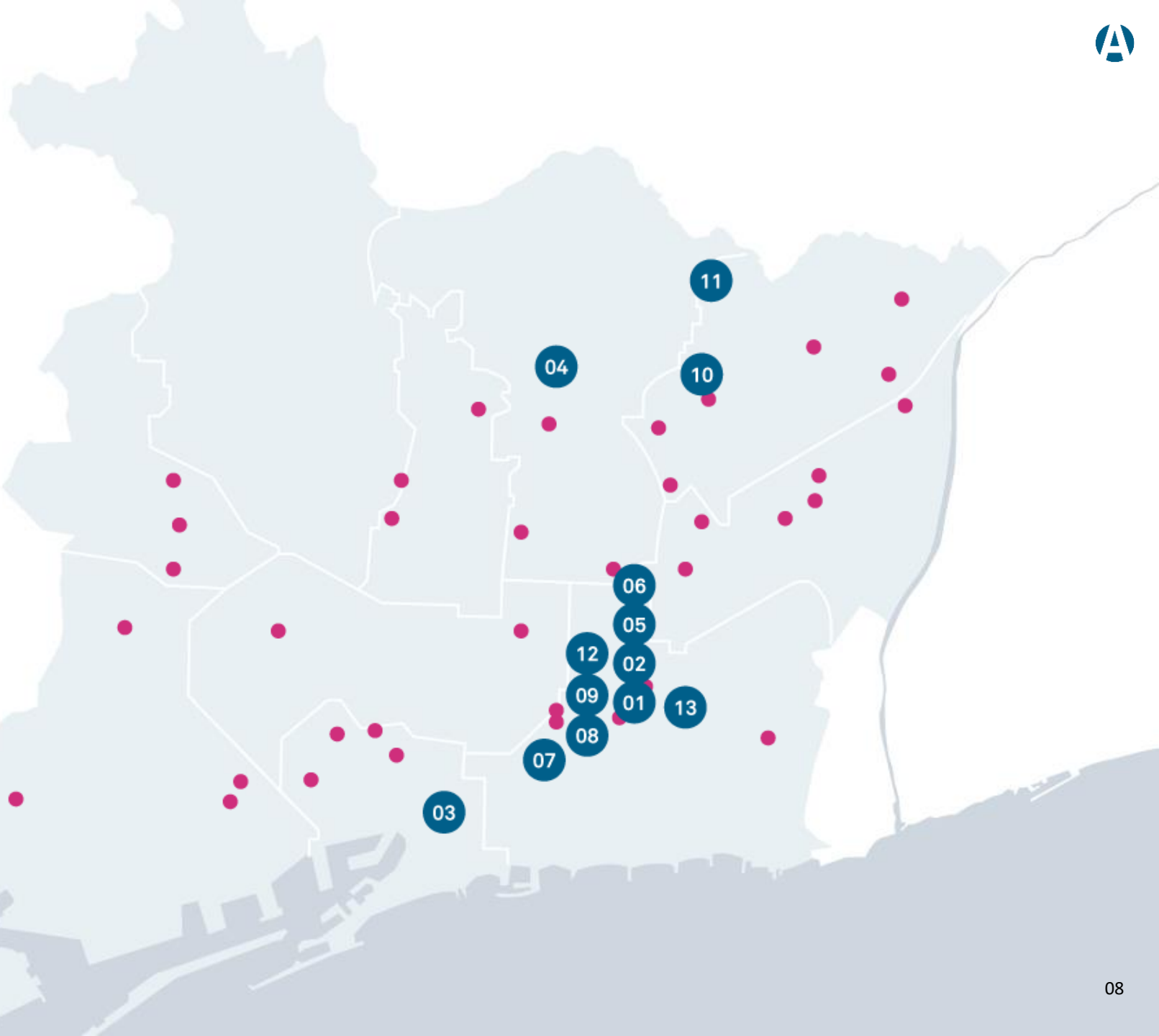
Centre per a la Iniciativa Emprenedora Glòries
L’equipament per emprendre a la ciutat

Cibernàrium Nou Barris
Centre de formació digital d’iniciació ubicat al Parc Tecnològic

Xarxa d’equipaments de Barcelona Activa

Barcelona Activa està present a **48** punts de la ciutat.

1. **Seu Central**
2. **Porta22**
3. **Convent de Sant Agustí**
4. **Ca n’Andalet**
5. **Centre per a la Iniciativa Emprenedora Glòries**
6. **Incubadora Glòries**
7. **Incubadora Almogàvers**
8. **Incubadora MediaTIC**
9. **Oficina d’Atenció a les Empreses**
10. **Parc Tecnològic i Cibernàrium Nou Barris**
11. **Nou Barris Activa**
12. **Cibernàrium 22@**
13. **InnoBA**



Punts d’atenció a la ciutat

Índex

0. Presentació ponent
1. Reptes de seguretat en projectes de ciència de dades
2. Criteris i procediments en l'emmagatzematge de dades
3. Gestió de la privadesa de les dades
4. Control d'accés i prevenció d'accessos no deguts
5. Introducció a l'aplicació de la ciència de dades en ciberseguretat

Laia Subirats

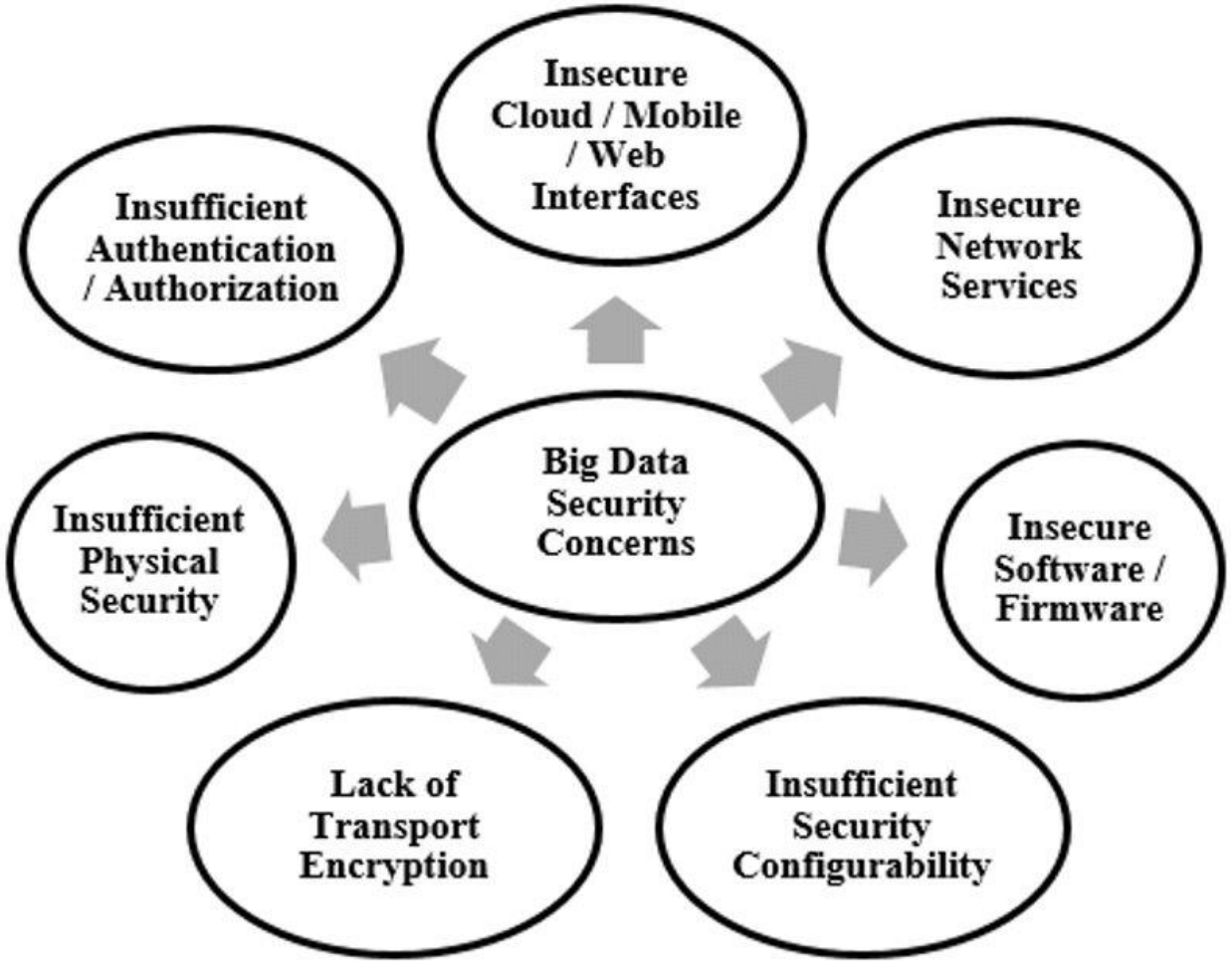
Experiència laboral:

- **Investigadora @ Eurecat (2009-Actualitat)**
- **Consultora @ Universitat Oberta de Catalunya (2016-Actualitat)**
- Consultora @ Universitat Pompeu Fabra (Estiu 2011, Estiu 2012)
- Estudiant @ Google Summer of Code (Estiu 2008)
- Estudiant d'estiu @ European Organization for Nuclear Research (CERN) (Estiu 2007)
- Becària @ Telefónica I+D (2006-2008)

Estudis:

- Doctorat en Informàtica @ Universitat Autònoma de Barcelona (2009-2015)
- Màster en Telemàtica @ Universitat Politècnica de Catalunya (2008-2009)
- Enginyeria de Telecomunicacions @ Universitat Pompeu Fabra (2003-2008)

1. Reptes de seguretat en projectes de ciència de dades

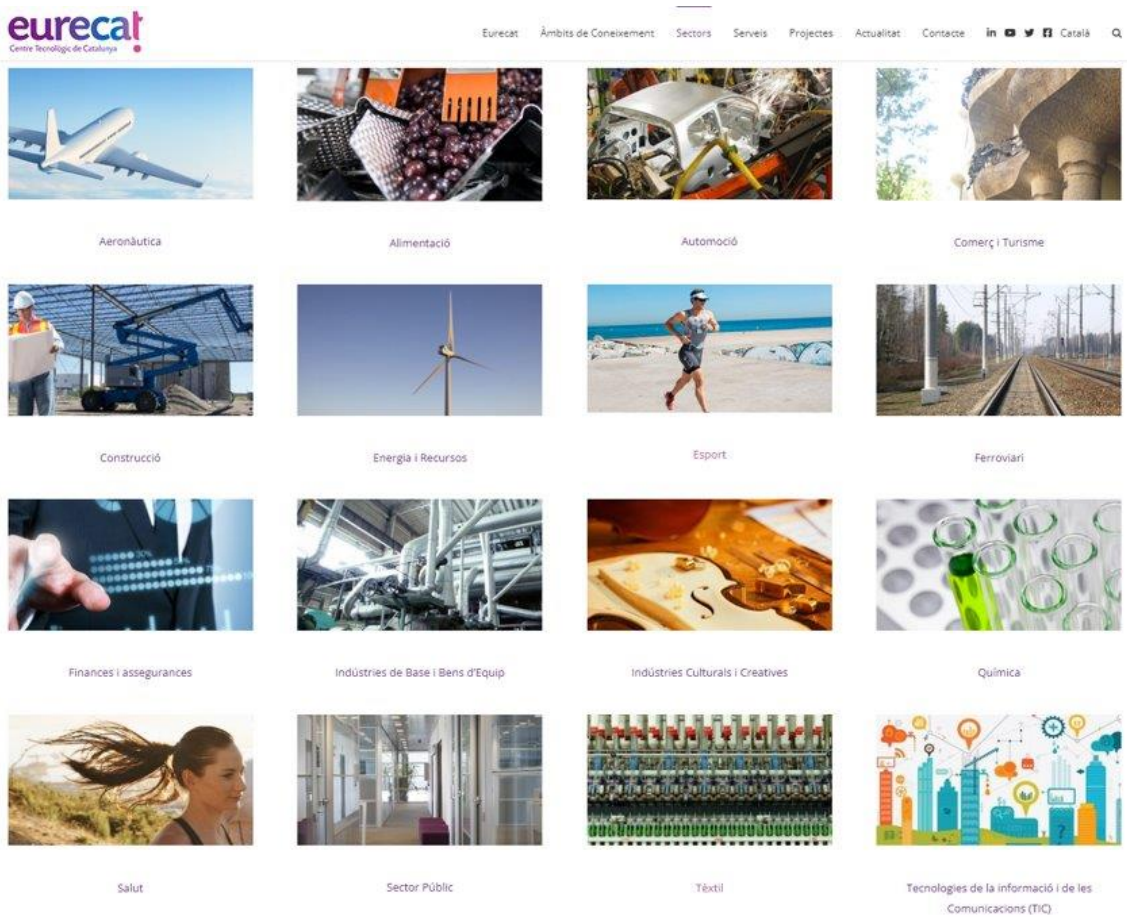


Data mining

- **Passos:**

1. **Plantejar-se una pregunta**
2. **Aconseguir dades**
3. **Processar les dades**
4. **Anàlisi de dades**
5. **Prendre una decisió i actuar**

Camps d’aplicació



Conjunts de dades

- Podeu buscar conjunts de dades a https://datasetsearch.research.google.com
- APIs
- Dades Obertes
- Web scraping

Perfils data science

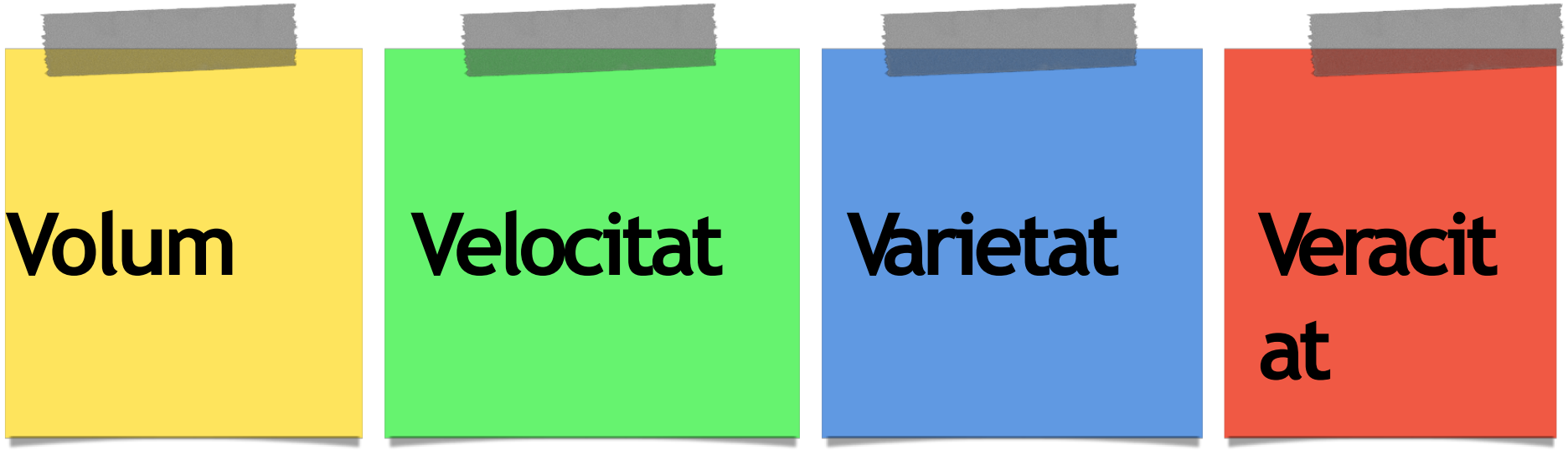
https://todobi.com/tipos-de-roles-en-analytics-business

- Harvard Business review
- Sexiest job of the 21st century

Què és Big Data?

Big data es més que mida/volum.

Es caracteriza, normalment, per les denominades 4 V's



Què és Big Data?

ARCHIVES
Archives of scanned documents, statements, insurance forms, medical record and customer correspondence, paper archives, and print stream files that contain original systems of record between organizations and their customers

DOCS
XLS, PDF, CSV, email, Word, PPT, HTML, HTML 5, plain text, XML, JSON, etc.

MEDIA
Images, videos, audio, Flash, live streams, podcasts, etc.

DATA STORAGE
SQL, NoSQL, Hadoop, doc repository, file systems, etc.

BUSINESS APPS
Project management, marketing automation, productivity, CRM, ERP content management systems, HR, storage, talent management, procurement, expense management, Google Docs, intranets, portals, etc.

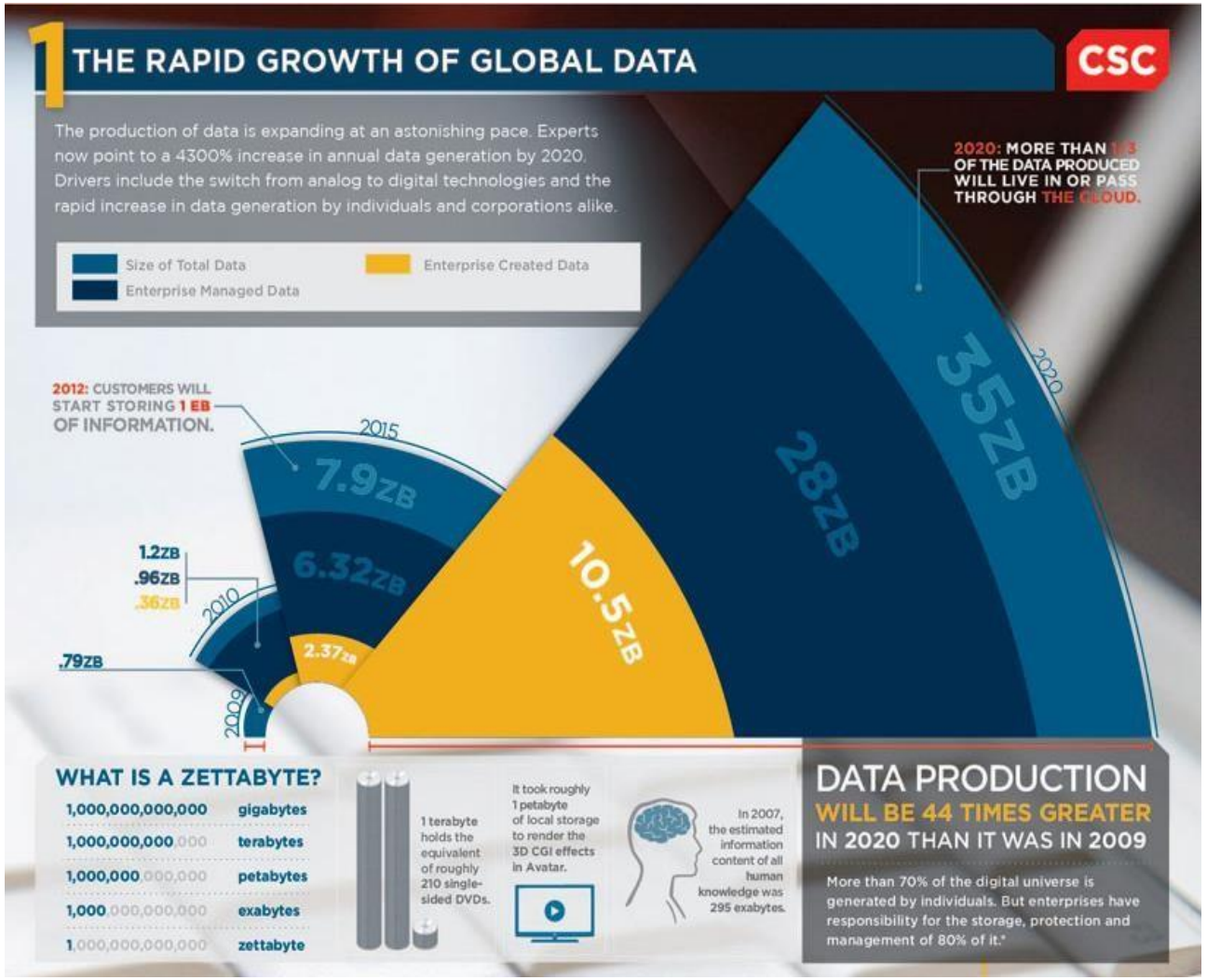
PUBLIC WEB
Government, weather, competitive, traffic, regulatory, compliance, health care services, economic, census, public finance, stock, OSINT, the World Bank, SEC/Edgar, Wikipedia, IMDb, and other Web services

SOCIAL MEDIA
Twitter, LinkedIn, Facebook, Tumblr, Blog, SlideShare, YouTube, Google+, Instagram, Flickr, Pinterest, Vimeo, Wordpress, IM, RSS, Review, Chatter, Jive, Yammer, etc.

MACHINE LOG DATA
Event logs, server data, application logs, business process logs, audit logs, call detail records (CDRs), mobile location, mobile app usage, clickstream data, etc.

SENSOR DATA
Medical devices, smart electric meters, car sensors, road cameras, satellites, traffic recording devices, processors found within vehicles, video games, cable boxes or household appliances, assembly lines, office buildings, cell towers and jet engines, air conditioning units, refrigerators, trucks, farm machinery, etc.

Què és Big Data?



Benefits of Big Data

Benefits of Big Data

40%

Increased Efficiency

50%

Better Business Decision Making

80%

Improved Customer Experience and Engagement

90%

Achieved Financial Savings

slidesalad

Here are some of the Big Data Security challenges that companies should mitigate:

- Big Data Security Issues: Data Storage
- Big Data Security Issues: Fake Data
- Big Data Security Issues: Data Privacy
- Big Data Security Issues: Data Management
- Big Data Security Issues: Data Access Control
- Big Data Security Issues: Data Poisoning
- Big Data Security Issues: Employee Theft

Tipus de seguretat en les dades

https://www.fortinet.com/resources/cyberglossary
https://www.fortinet.com/resources/cyberglossary/data-security

Encryption

Data Erasure: More secure than standard data wiping, data erasure uses software to completely overwrite data on any storage device. It verifies that the data is unrecoverable.

Data Masking: By masking data, organizations can allow teams to develop applications or train people using real data. It masks personally identifiable information (PII) where necessary so that development can occur in environments that are compliant.

Data Resiliency: Resiliency is determined by how well an organization endures or recovers from any type of failure

Best Practices for Ensuring Data Security and Privacy

Why is data security important? Primarily, it keeps your data secure and builds confidence among your customers. Here are some best practices that have been effective for other organizations:

1. **Secure your information:** This means managing who has access and encrypting your data. Only people who need it to perform essential functions should have access, and information should be encrypted as it goes back and forth between the database and their computer or device.
2. **Prepare ahead of time for threats:** You can get ready for a potential data security incident by testing your system, educating employees, devising an incident management plan, and creating a data recovery plan.
3. **Delete data you are not using:** You should get rid of both digital and physical copies of data you no longer need. In this way, you reduce the chances of a hacker discovering it and using it for profit.

Solucions

Data discovery and classification tools
Data and file activity monitoring
Vulnerability assessment and risk analysis tools
Automated compliance reporting

2. Criteris i procediments en l'emmagatzematge de dades

Estratègies:

Physical security of servers and user devices
Access management and controls
Application security and patching
Backups
Employee education
Network and endpoint security monitoring and controls

Tendències de seguretat en les dades

AI
Multicloud security
Quantum

Interacció

Achieving enterprise-grade data security
Data security and the cloud
Data security and BYOD (Bring your own Device)

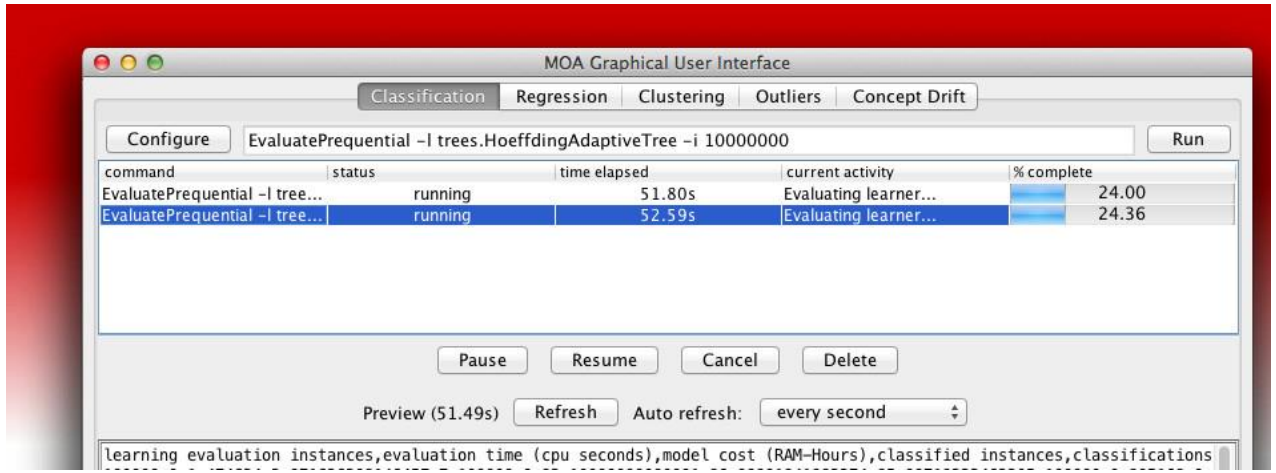
Data security, privacy and protection solutions

Data security solutions
Data security devices
Homomorphic encryption
Storage data backup (Github) and recovery
Data encryption solutions
Data privacy
Ransomware protection
Zero trust security
Flash storage solutions
GDPR

3. Gestió de la privadesa de les dades

1. Stream data mining
2. Biaixos
3. Publicar dades obertes

https://scikit-multiflow.github.io

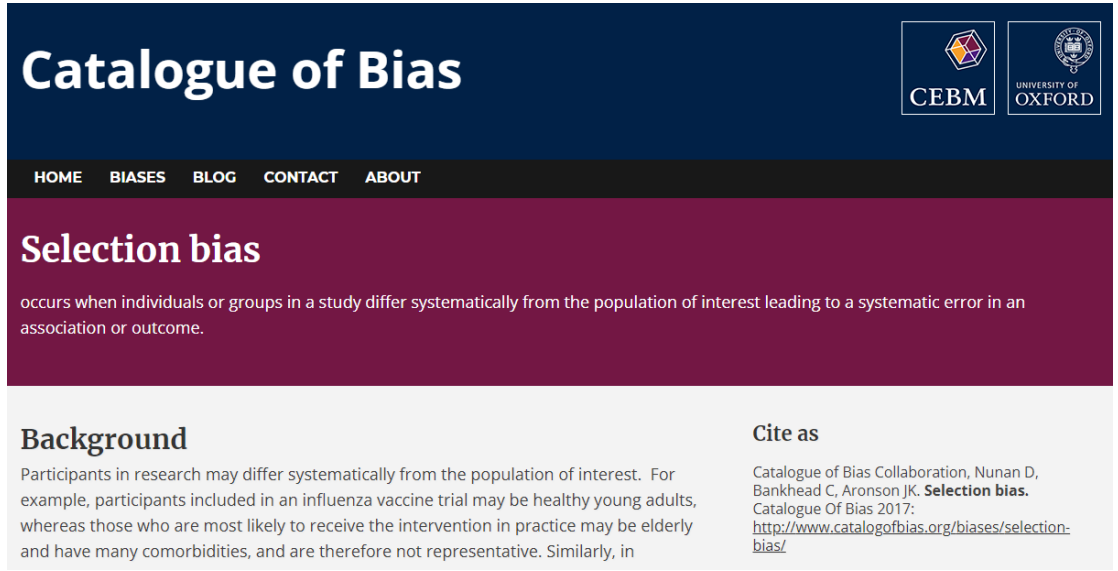


Catàleg de biaixos

Researchers have recognised long time ago that bias is an important problem:

The catalogue of bias group biases according to:

- Background
- Examples
- Impact
- Preventive steps
- Sources



https://catalogofbias.org

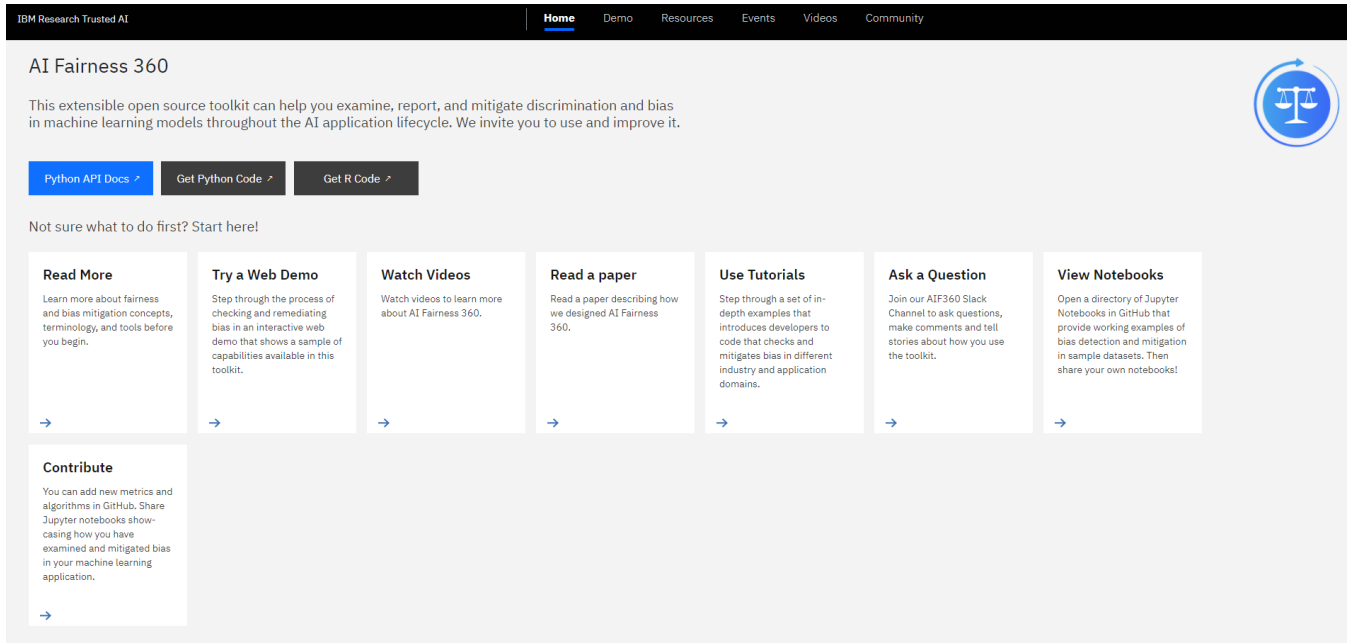
AI i sexe i gènere

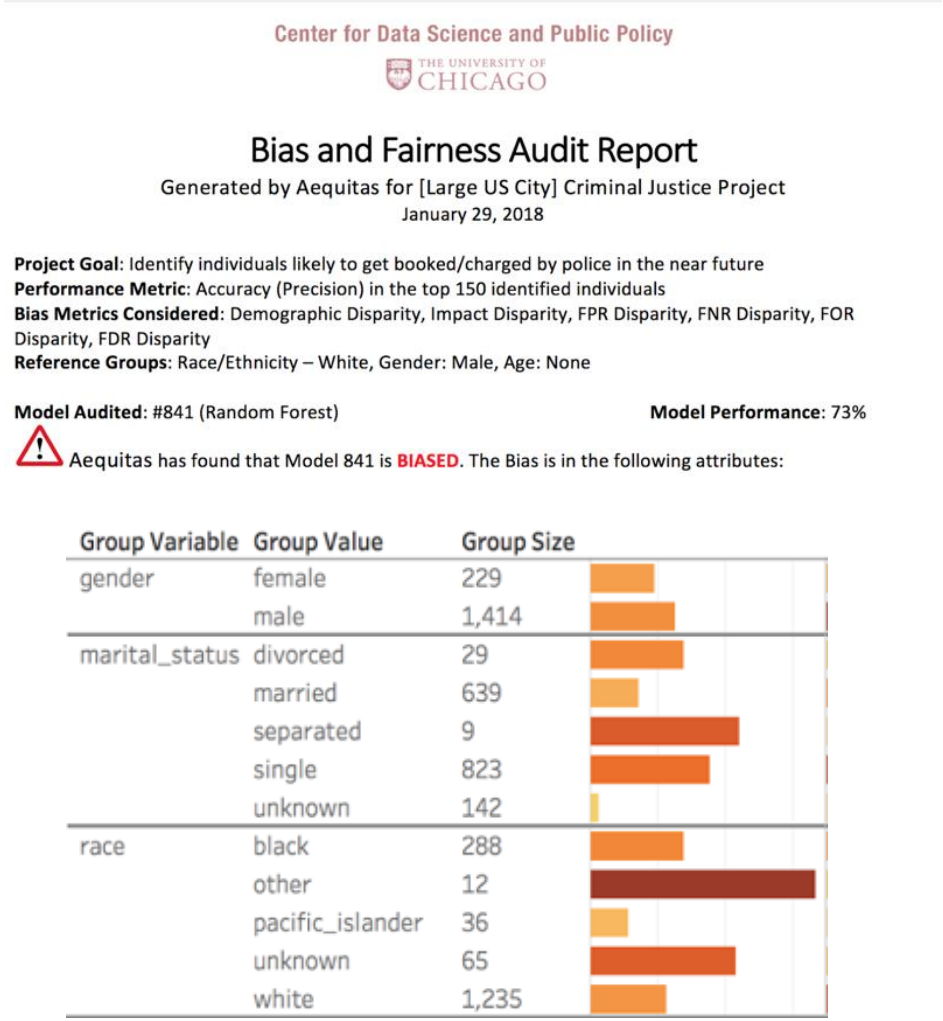
Iniciativa, URL	Año, País	Misión
Data2x, https://data2x.org	2012, US	To improve availability, quality and use of gender data.
Women in Big Data, https://www.womeninbigdata.org	2015, Global	To increase the diversity in the Big Data field and to attract, connect, engage and grow female talent.

Open source toolkits

IBM AI Fairness 360 (AIF360) aif360.mybluemix.net

Aequitas dsapp.uchicago.edu/projects/aequitas





Center for Data Science and Public Policy

THE UNIVERSITY OF CHICAGO

Bias and Fairness Audit Report

Generated by Aequitas for [Large US City] Criminal Justice Project

January 29, 2018

Project Goal: Identify individuals likely to get booked/charged by police in the near future
Performance Metric: Accuracy (Precision) in the top 150 identified individuals
Bias Metrics Considered: Demographic Disparity, Impact Disparity, FPR Disparity, FNR Disparity, FOR Disparity, FDR Disparity
Reference Groups: Race/Ethnicity – White, Gender: Male, Age: None

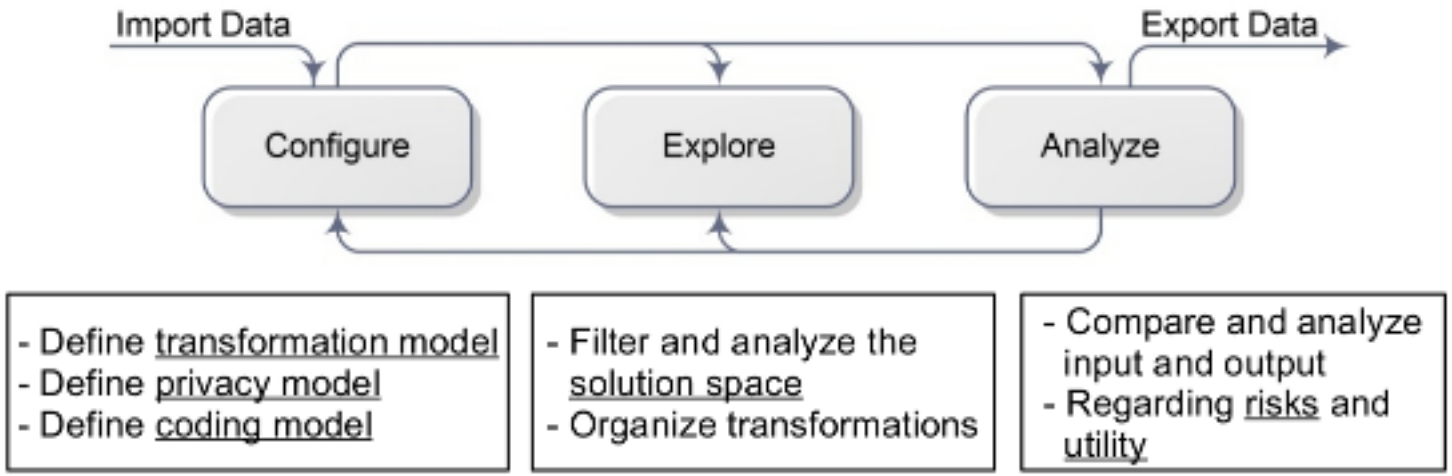
Model Audited: #841 (Random Forest) **Model Performance**: 73%

Aequitas has found that Model 841 is **BIASED**. The Bias is in the following attributes:

Group Variable	Group Value	Group Size
gender	female	229
	male	1,414
marital_status	divorced	29
	married	639
	separated	9
	single	823
	unknown	142
race	black	288
	other	12
	pacific_islander	36
	unknown	65
	white	1,235

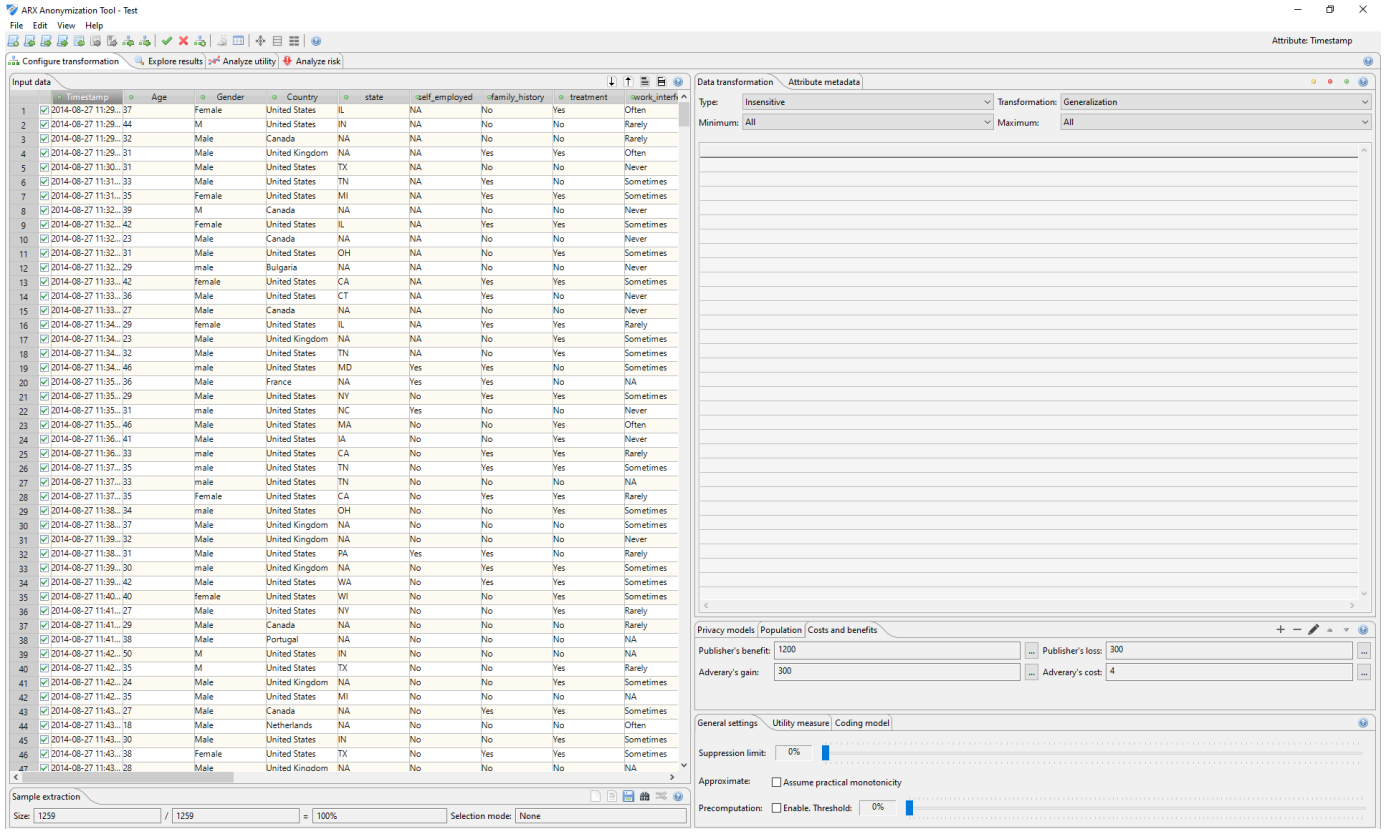
ARX Data anonymization tool

https://arx.deidentifier.org

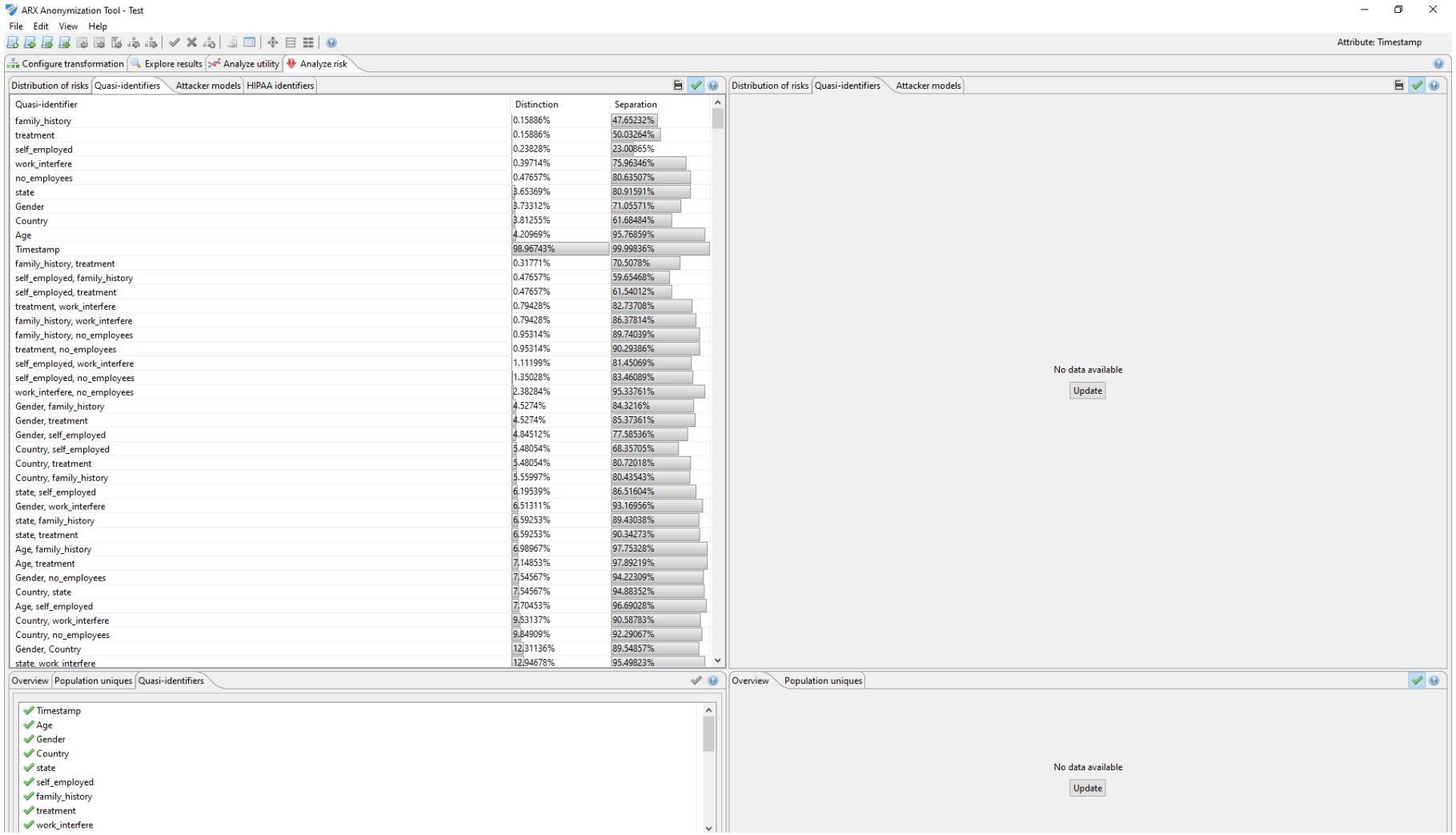


Exemple pràctic ARX anonymization tool

Carreguem dades de:
https://www.kaggle.com/datasets/osmi/mental-health-in-tech-survey



Exemple pràctic ARX anonymization tool



Challenges

Data challenges include:

- Poorly selected data
- Incomplete, incorrect or outdated data selection bias
- Unintentional perpetuation and promotion of historical biases

Algorithm challenges include:

- Poorly designed matchmaking systems
- Personalization and recommendation services that reduce rather than expand user options
- Decision-making systems that assume correlation necessarily imply causality
- Data sets that lack information or disproportionately represent specific populations

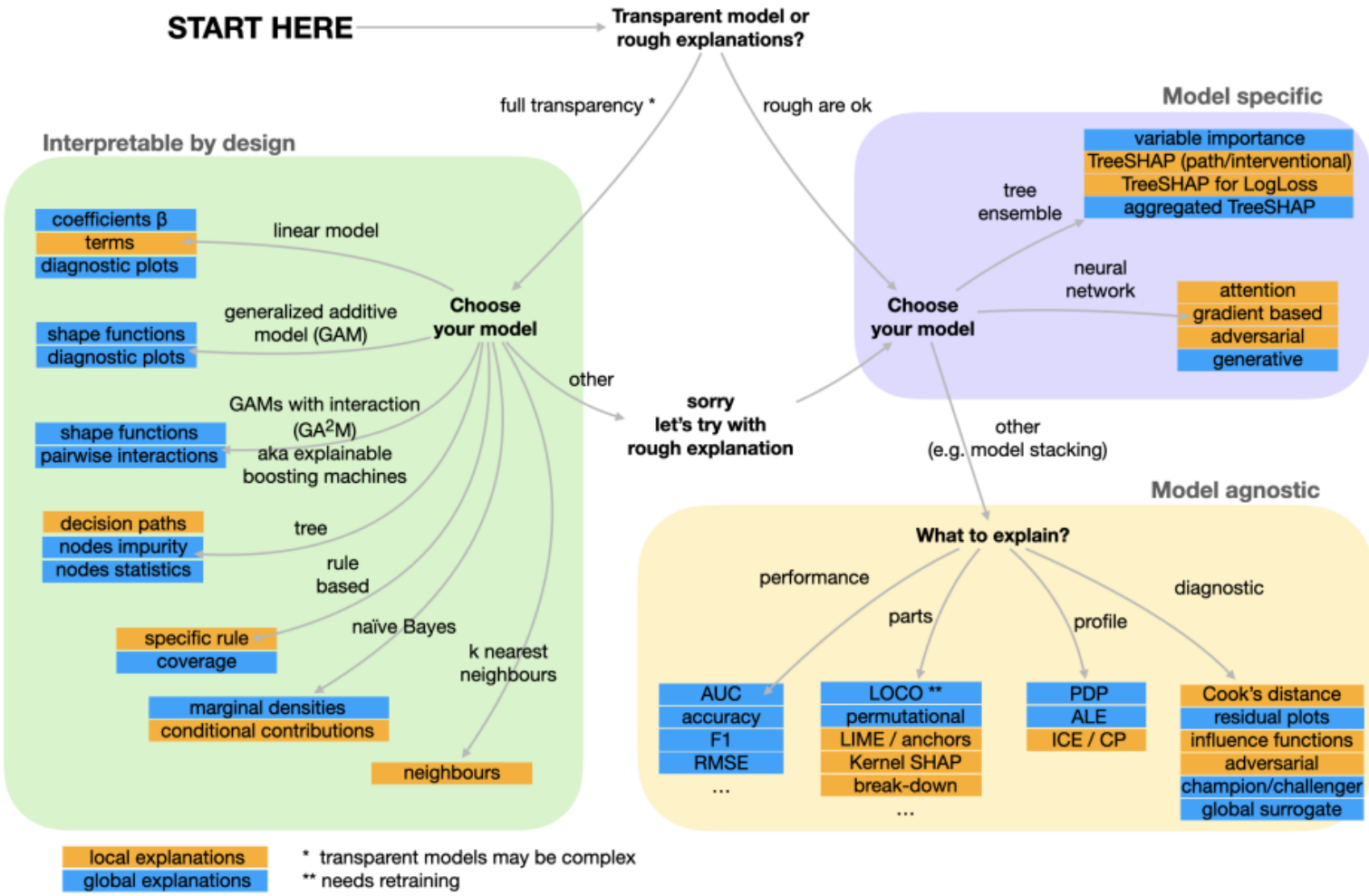
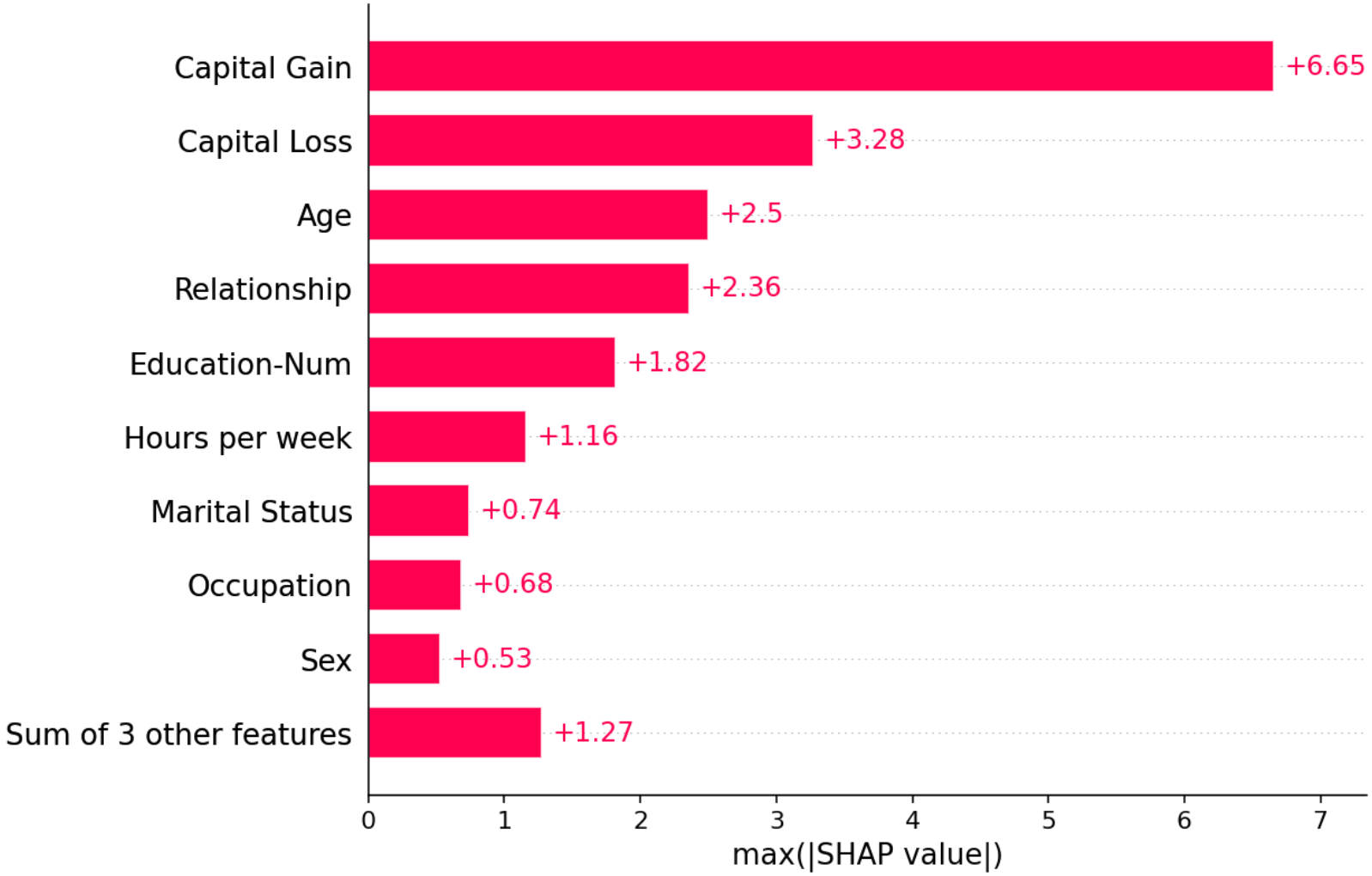


Figure 1: Model oriented taxonomy for XAI methods.

https://arxiv.org/pdf/2009.13248.pdf

SHAP **(SHapley Additive exPlanations)** is a game theoretic approach to explain the output of any machine learning model. It connects optimal credit allocation with local explanations using the classic Shapley values from game theory and their related extensions



https://shap.readthedocs.io/en/latest/index.html

Conclusions

- **Preventing bias** in Big Data analysis of education is essential to ensure **fairness and ethics**.
- Bias can arise from **data or algorithms**. **Fair design algorithms** are an opportunity to overcome algorithmic bias, but data bias could still be present. Taking sex and gender into account is essential to provide a more personalized prediction in the **diagnosis and prognosis**.
- More data and insights are needed, and increased **collaboration** between educational data and AI stakeholders will also be critical to ensure progress.

Publishing open data

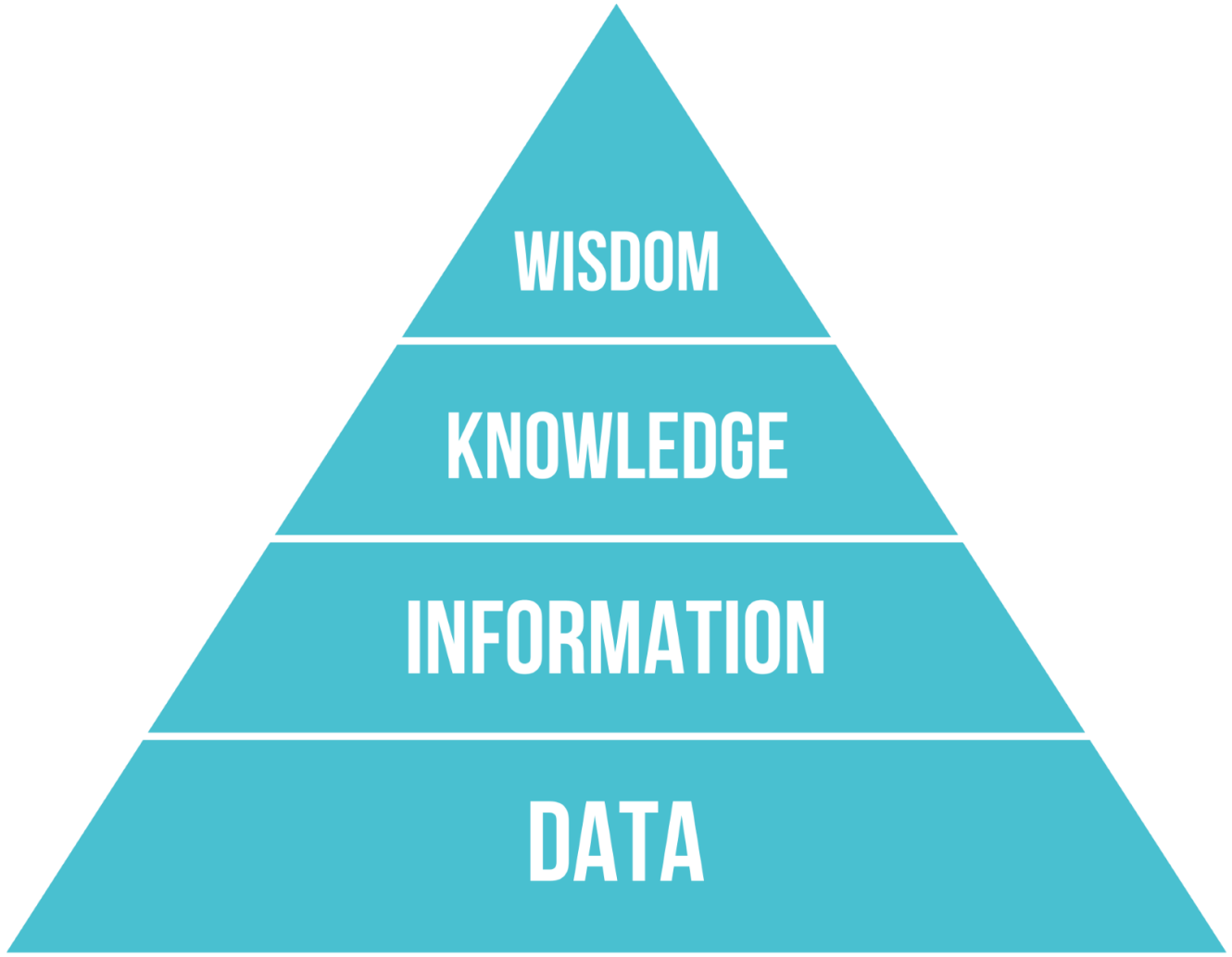
- Introduction to Open Data
- Steps for data publishing
- Publish in the best place: Creating a dataset with a doi through Zenodo or other journals / platforms
- Describe your data
- Use the best file format: 5-star Open Data
- Licenses to publish the data
- References

Publishing open data

Open data has many benefits, for example enables data driven decision making, performance measurement, reduction of government costs, transparency, economic development, increased citizen engagement and talent attraction / retention.

Figure 1. Data extracted from https://opendatabarometer.org

Publishing open data



Publishing open data

- Determine if you can publish your data: Before you decide to publish your data, check to ensure your legally and ethically allowed to.
- Publish in the best place: There are many places to publish data. It's up to you to decide if you want to publish in a discipline specific repository that contains data like yours, the University's institutional repository, or a general repository such as Figshare.
- Describe your data: Publishing a thorough description of your data along with the dataset means that others will be able to understand and reuse your data.
- Use the best file format: choosing an open or standard file format means that everyone will be able to open your dataset.
- License your data: Applying a licence to your data enables others to understand how your data can be used.
- Get a persistent identifier: Persistent identifiers, like DOIs and ORCiDs, are the easiest way of linking items together, such as a dataset to a related publication. They also make your data easier to cite and enable you to measure the reuse of your data through metrics.

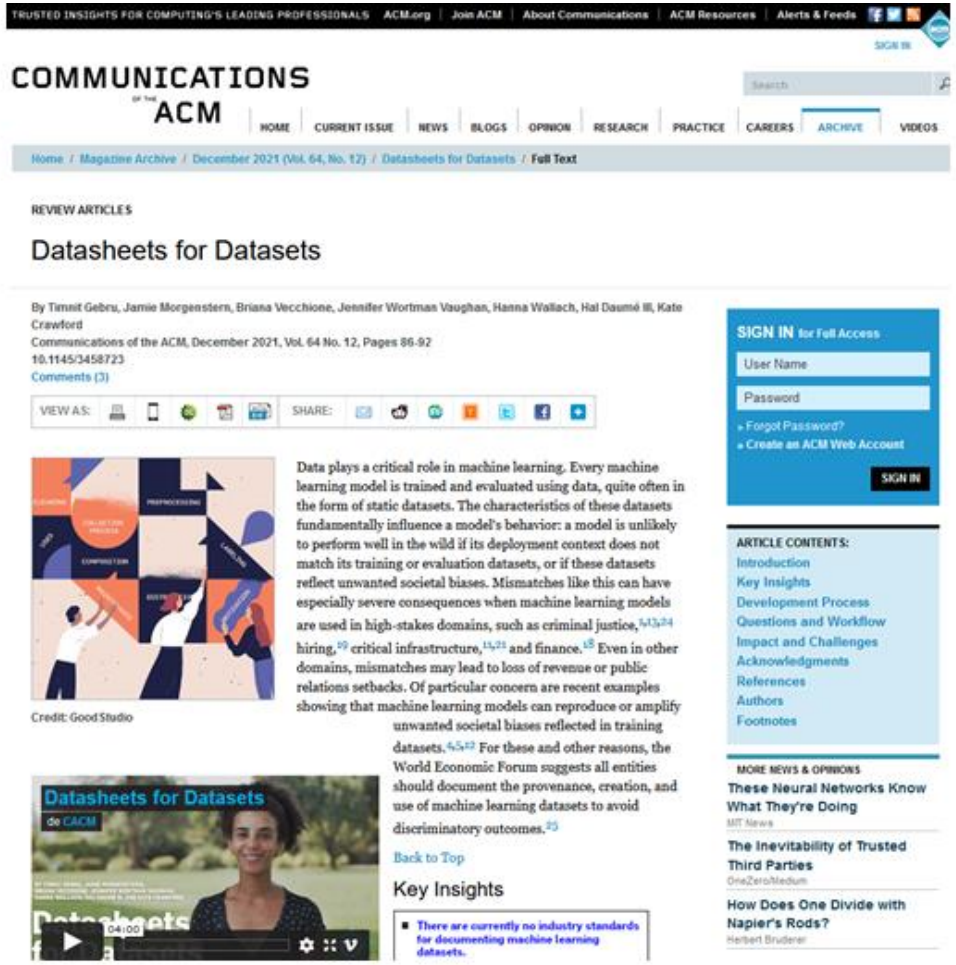
Publishing open data

The following are some journals where you can publish your data

- Zenodo
- Scientific Data (nature.com)
- Data in Brief - Journal - Elsevier
- Data | An Open Access Journal from MDPI
- Datasets Documentation | Kaggle

Publishing open data

Datasheets for datasets are intended to address the needs of two key stakeholder groups: dataset creators and dataset consumers.

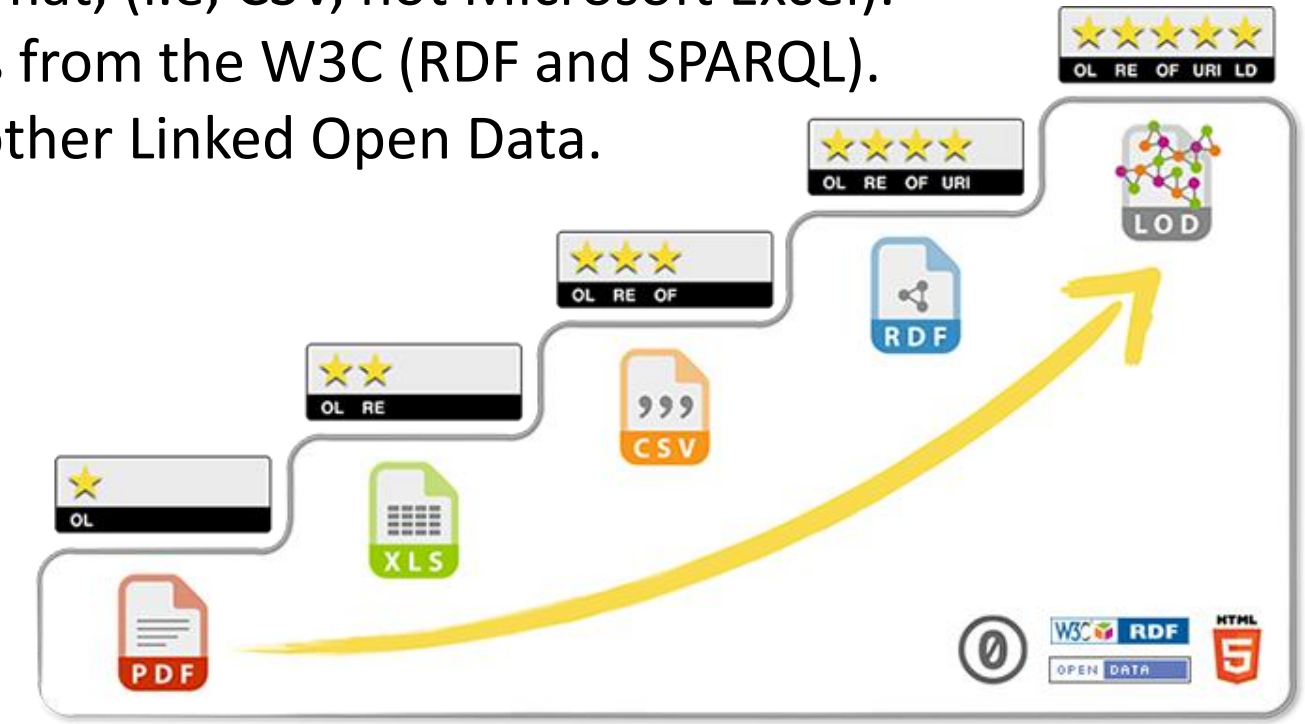


https://cacm.acm.org/magazines/2021/12/256932-datasheets-for-datasets/fulltext

Publishing open data

Tim Berners-Lee, the inventor of the Web and initiator of the Linked Data project, suggested a **5 star deployment scheme** for Linked Data:

- ☆ Data is available on the Web, in whatever format.
- ☆☆ Available as machine-readable structured data, (i.e., not a scanned image).
- ☆☆☆ Available in a non-proprietary format, (i.e, CSV, not Microsoft Excel).
- ☆☆☆☆ Published using open standards from the W3C (RDF and SPARQL).
- ☆☆☆☆☆ All of the above and links to other Linked Open Data.



Publishing open data

- More courses and references available at https://theodi.org/events/courses/
- Wiki about publishing open data https://www.wikidata.org/wiki/Wikidata:Open_data_publishing
- FAIR Tools https://www.fairsfair.eu/tools-software
- https://github.com/aina1997 --> Master class DS

4. Control d'accés i prevenció d'accessos no deguts

What Are the Components of Access Control?
How Does Access Control Work?
What Is the Difference Between Authentication and Authorization?
What Are the Different Types of Access Controls?

What Are the Components of Access Control?

Access control is managed through several components:

Authentication

Authentication is the initial process of establishing the identity of a user. For example, when a user signs in to their email service or online banking account with a username and password combination, their identity has been authenticated. However, authentication alone is not sufficient to protect organizations' data.

Authorization

Authorization adds an extra layer of security to the authentication process. It specifies access rights and privileges to resources to determine whether the user should be granted access to data or make a specific transaction.

For example, an email service or online bank account can require users to provide two-factor authentication (2FA), which is typically a combination of something they know (such as a password), something they possess (such as a token), or something they are (like a biometric verification). This information can also be verified through a 2FA mobile app or a thumbprint scan on a smartphone.

Access

Once a user has completed the authentication and authorization steps, their identity will be verified. This grants them access to the resource they are attempting to log in to.

Manage

Organizations can manage their access control system by adding and removing the authentication and authorization of their users and systems. Managing these systems can become complex in modern IT environments that comprise cloud services and on-premises systems.

Audit

Organizations can enforce the principle of least privilege through the access control audit process. This enables them to gather data around user activity and analyze that information to discover potential access violations.

How Does Access Control Work?

Access control is used to verify the identity of users attempting to log in to digital resources. But it is also used to grant access to physical buildings and physical devices.

Physical Access Control

Common examples of physical access controllers include:

Barroom Bouncers

Bouncers can establish an access control list to verify IDs and ensure people entering bars are of legal age.

Subway Turnstiles

Access control is used at subway turnstiles to only allow verified people to use subway systems. Subway users scan cards that immediately recognize the user and verify they have enough credit to use the service.

Keycard or Badge Scanners in Corporate Offices

Organizations can protect their offices by using scanners that provide mandatory access control. Employees need to scan a keycard or badge to verify their identity before they can access the building.

Logical/Information Access Control

Logical access control involves tools and protocols being used to identify, authenticate, and authorize users in computer systems. The access controller system enforces measures for data, processes, programs, and systems.

Signing Into a Laptop Using a Password

A common form of data loss is through devices being lost or stolen. Users can keep their personal and corporate data secure by using a password.

Unlocking a Smartphone With a Thumbprint Scan

Smartphones can also be protected with access controls that allow only the user to open the device. Users can secure their smartphones by using biometrics, such as a thumbprint scan, to prevent unauthorized access to their devices.

Remotely Accessing an Employer’s Internal Network Using a VPN

Smartphones can also be protected with access controls that allow only the user to open the device. Users can secure their smartphones by using biometrics, such as a thumbprint scan, to prevent unauthorized access to their devices.

What Is the Difference Between Authentication and Authorization?

Authentication and authorization are crucial to access control in security. Authentication is the process of logging in to a system, such as an email address, online banking service, or social media account. Authorization is the process of verifying the user's identity to provide an extra layer of security that the user is who they claim to be.

Importance of Access Control in Regulatory Compliance

Access control is crucial to helping organizations comply with various data privacy regulations. These include:

PCI DSS

The Payment Card Industry Data Security Standard (PCI DSS) is a security standard that protects the payment card ecosystem. An access control system is crucial to permitting or denying transactions and ensuring the identity of users.

HIPAA

The Health Insurance Portability and Accountability Act (HIPAA) was created to protect patient health data from being disclosed without their consent. Access control is vital to limiting access to authorized users, ensuring people cannot access data that is beyond their privilege level, and preventing data breaches.

SOC 2

Service Organization Control 2 (SOC 2) is an auditing procedure designed for service providers that store customer data in the cloud. It ensures that providers protect the privacy of their customers and requires organizations to implement and follow strict policies and procedures around customer data. Access control systems are crucial to enforcing these strict data security processes.

ISO 27001

The International Organization for Standardization (ISO) defines security standards that organizations across all industries need to comply with and demonstrate to their customers that they take security seriously. ISO 27001 is the ISO's gold standard of information security and compliance certification. Implementing access controls is crucial to complying with this security standard.

What Are the Different Types of Access Controls?

There are various types of access controls that organizations can implement to safeguard their data and users. These include:

Attribute-based Access Control (ABAC)

ABAC is a dynamic, context-based policy that defines access based on policies granted to users. The system is used in identity and access management (IAM) frameworks.

Discretionary Access Control (DAC)

DAC models allow the data owner to decide access control by assigning access rights to rules that users specify. When a user is granted access to a system, they can then provide access to other users as they see fit.

Mandatory Access Control (MAC)

MAC places strict policies on individual users and the data, resources, and systems they want to access. The policies are managed by an organization's administrator. Users are not able to alter, revoke, or set permissions.

Role-Based Access Control (RBAC)

RBAC creates permissions based on groups of users, roles that users hold, and actions that users take. Users are able to perform any action enabled to their role and cannot change the access control level they are assigned.

Break-glass Access Control

Break-glass access control involves the creation of an emergency account that bypasses regular permissions. In the event of a critical emergency, the user is given immediate access to a system or account they would not usually be authorized to use.

Rule-based Access Control

A rule-based approach sees a system admin define rules that govern access to corporate resources. These rules are typically built around conditions, such as the location or time of day that users access resources.

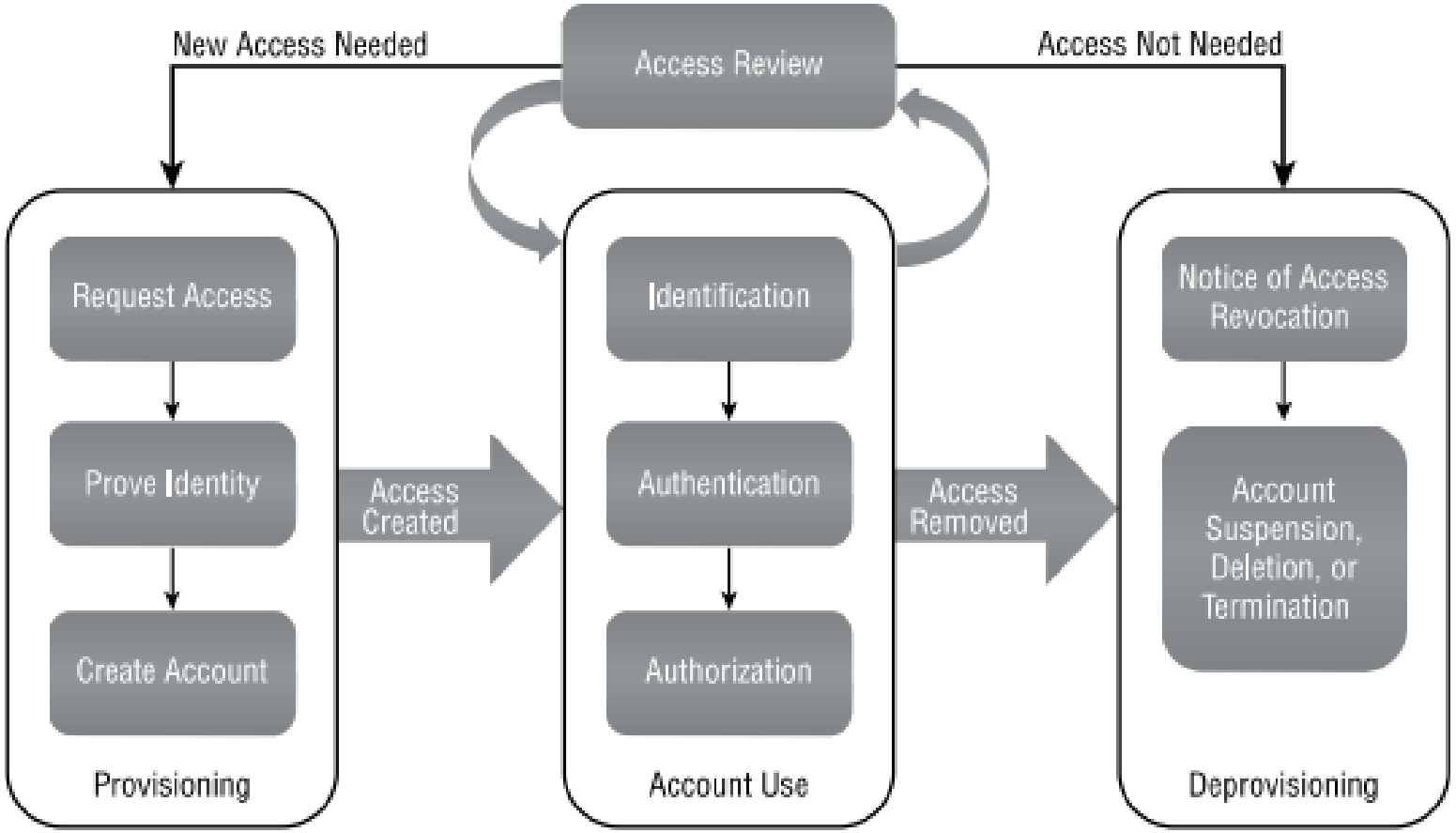


FIGURE 5.1 **The access management lifecycle**

https://es.wikipedia.org/wiki/OpenID

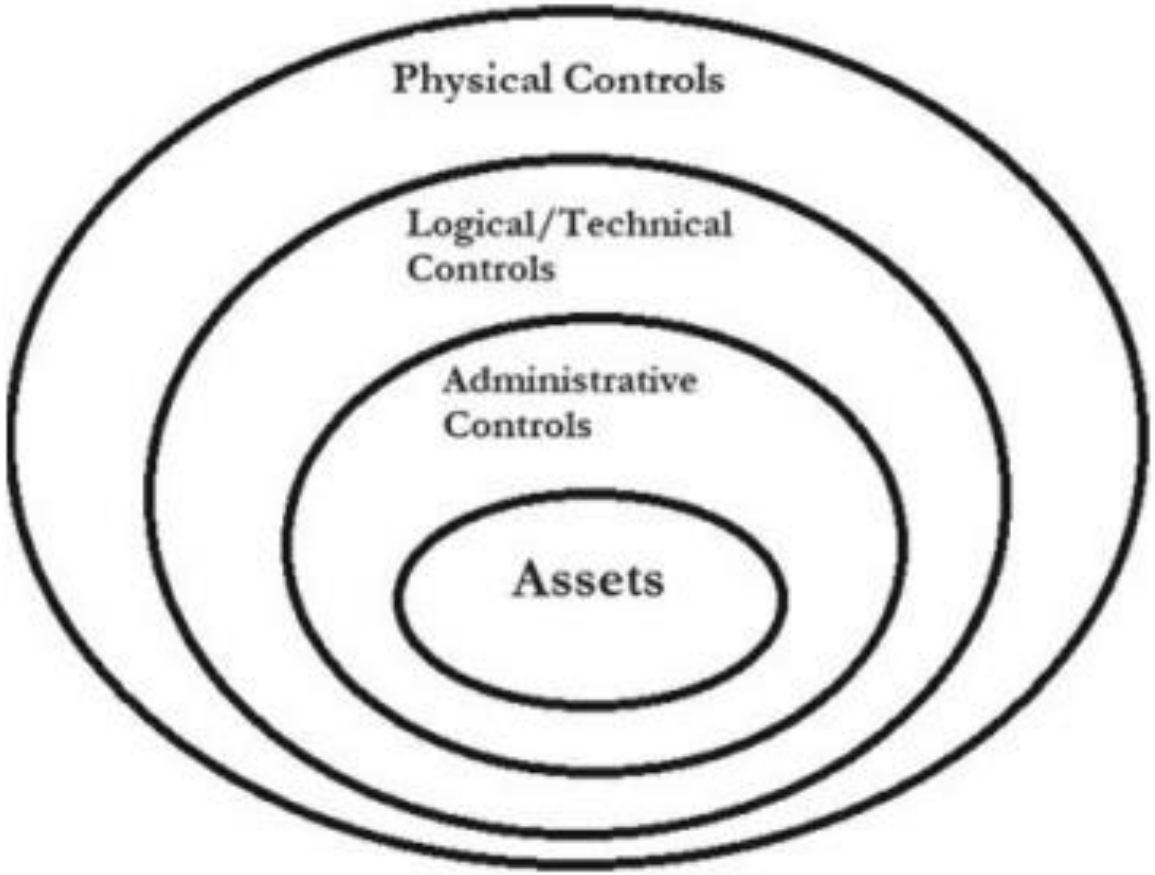


Fig. 9.1 Defense-in-depth implementation in security controls

CISSP®

Certified Information Systems Security Professional

An (ISC)² Certification

Certification **Exam Outline**

Effective Date: May 1, 2021

Figure 9: Placing the ECSF role profiles in the context of a management circle

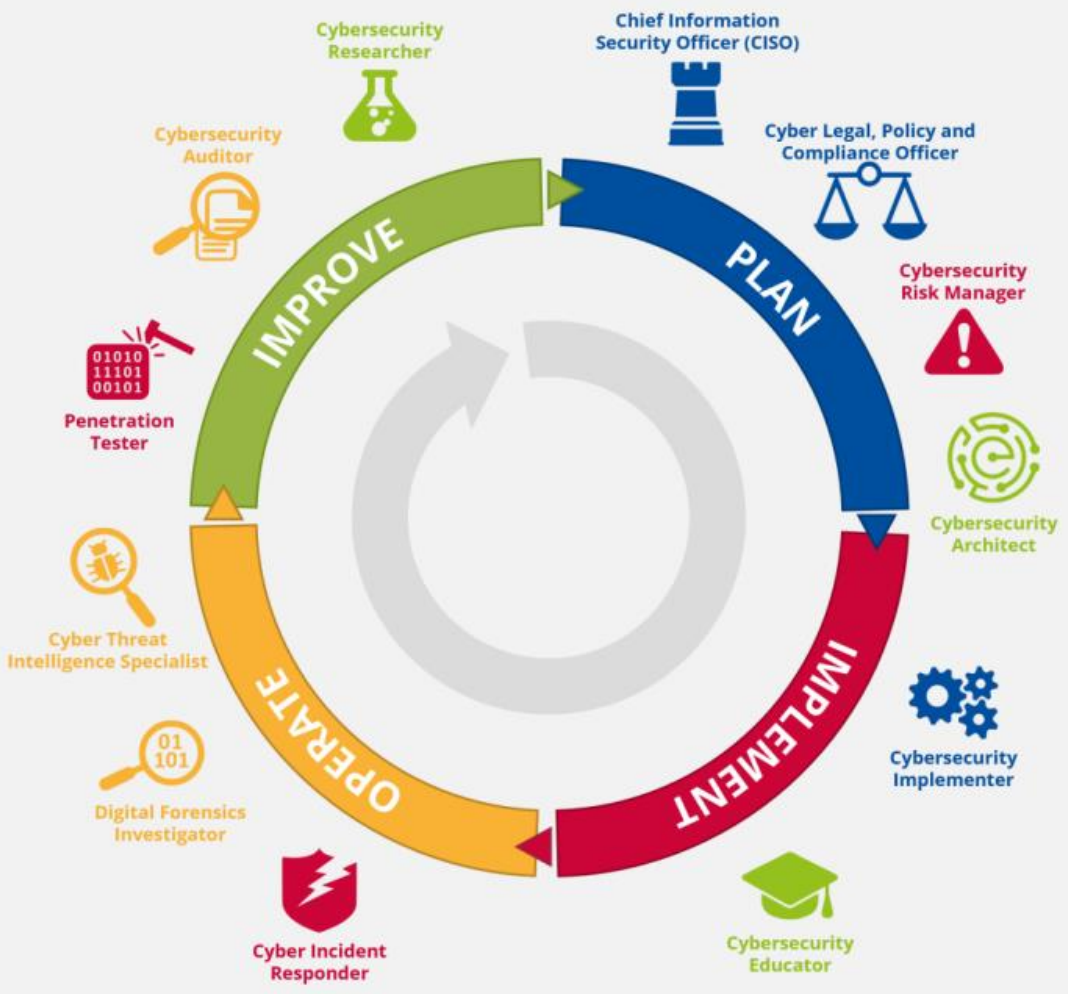


Table 5: The ESCO profiles and the ECSF profiles relationships

ESCO code	ESCO Occupation	Relationship	ECSF role profile
2149.2.8	Research engineer	might include	Cybersecurity Researcher
2310.1	Higher education lecturer	might include	Cybersecurity Educator
2356	Information technology trainer	might include	Cybersecurity Educator
2511.18	IT auditor	might include	Cybersecurity Auditor
2519.2	ICT auditor manager	might include	Cybersecurity Auditor
2529.1	Chief ICT security officer	is	Chief Information Security Officer (CISO)
2529.2	Digital forensic expert	is	Digital Forensics Investigator
2529.3	Embedded system security engineer	might be included	Cybersecurity Implementer
2529.4	Ethical hacker	is	Penetration Tester
2529.6	ICT Security administrator	might be included	Cybersecurity Implementer
2529.7	ICT security engineer	might be included	Cybersecurity Architect
2529.7	ICT security engineer	might be included	Cybersecurity Implementer
2619.4	Data protection officer	is	Cyber Legal, Policy & Compliance Officer

https://esco.ec.europa.eu/en/classification/occupation_main

Hacking for dummies

- Cracker
- White hat
- Black hat
- Gray hat
- Green hat
- Red hat
- Blue hat

Key loggers, SQL injection attack, distributed denial of service, ethical hacking

Seguretat dades

Tabla 2. Clasificación según el impacto de seguridad de los datos

Objetivo de seguridad	Impacto bajo	Impacto moderado	Impacto alto
Confidencialidad	La divulgación no autorizada de información puede tener un efecto adverso **limitado** en la organización.	La divulgación no autorizada de información puede tener un efecto adverso **grave** en la organización.	La divulgación no autorizada de información puede tener un efecto adverso grave o **catastrófico** en la organización.
Integridad	La modificación o destrucción no autorizada de la información puede tener un efecto adverso **limitado** en la organización.	La modificación o destrucción no autorizada de la información puede tener un efecto adverso **grave** en la organización.	La modificación o destrucción no autorizada de la información puede tener un efecto adverso grave o **catastrófico** en la organización.
Disponibilidad	La interrupción del acceso o uso de la información o un sistema de información puede tener un efecto adverso **limitado** en la organización.	La interrupción del acceso o uso de la información o un sistema de información puede tener un efecto adverso **grave** en la organización.	La interrupción del acceso o uso de la información o un sistema de información puede tener un efecto adverso grave o **catastrófico** en la organización.

Exercici

- El propietari d'una perruqueria perd accidentalment una carpeta que contenia les factures dels seus proveïdors dels tres últims mesos.

- Un empleat d'una assessoria legal obre accidentalment un correu que conté *ransomware* al seu ordinador portàtil, que està connectat exclusivament a Internet.

- Similar a el cas anterior, però aquesta vegada l'empleat està utilitzant el PC de la seva oficina que té accés a Internet i a la intranet corporativa.

Qualitat de les dades

- Exactitut
- Completitut
- Consistència
- Puntualitat
- Unicitat
- Validesa

Springer Link

Search Log in

Textbook | © 2023

Cybersecurity and Identity Access Management

Authors: Bharat S. Rawal, Gunasekaran Manogaran, Alexender Peter

Discusses the different third-party attacks and hacking processe

Maximizes reader insights into effective cyber security, malware, firewalls and cryptography

Simplifies a complicated picture into hacking and penetration testing

1933 Accesses

Sections

Table of contents
About this book
Keywords
Authors and Affiliations
About the authors
Bibliographic Information

This is a preview of subscription content, access via your institution.

Access via your institution →

eBook EUR 85.59

Price includes VAT (Spain)

- ISBN: 978-981-19-2658-7
- Instant PDF download
- Readable on all devices
- Own it forever
- Exclusive offer for individuals only
- Tax calculation will be finalised during checkout

Buy eBook

Hardcover Book EUR 103.99

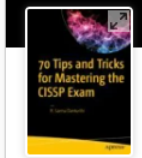
Learn about institutional subscriptions

https://link.springer.com/book/10.1007/978-981-19-2658-7

Springer Link

Search Log in

Introducing new learning courses and educational videos from Apress. Start watching



Book | © 2020

70 Tips and Tricks for Mastering the CISSP Exam

Apress®

Authors: R. Sarma Danturthi

Shows you how to read an exam question and correctly pick an answer from a set of solutions

Provides tips and tricks to tackle the verbiage given in an exam question, quickly eliminate wording that is not required, get the gist of the question, and find the solution without spending too much time

Teaches you how words such as MOST, BEST, FIRST, and LAST are used in exam questions and how to discern the correct answer to a question when multiple options look close as possible solutions

3931 Accesses

Sections

Table of contents
About this book
Keywords
Authors and Affiliations

Access via your institution →

eBook EUR 36.99

Price includes VAT (Spain)

- ISBN: 978-1-4842-6225-2
- Instant PDF download
- Readable on all devices
- Own it forever
- Exclusive offer for individuals only
- Tax calculation will be finalised during checkout

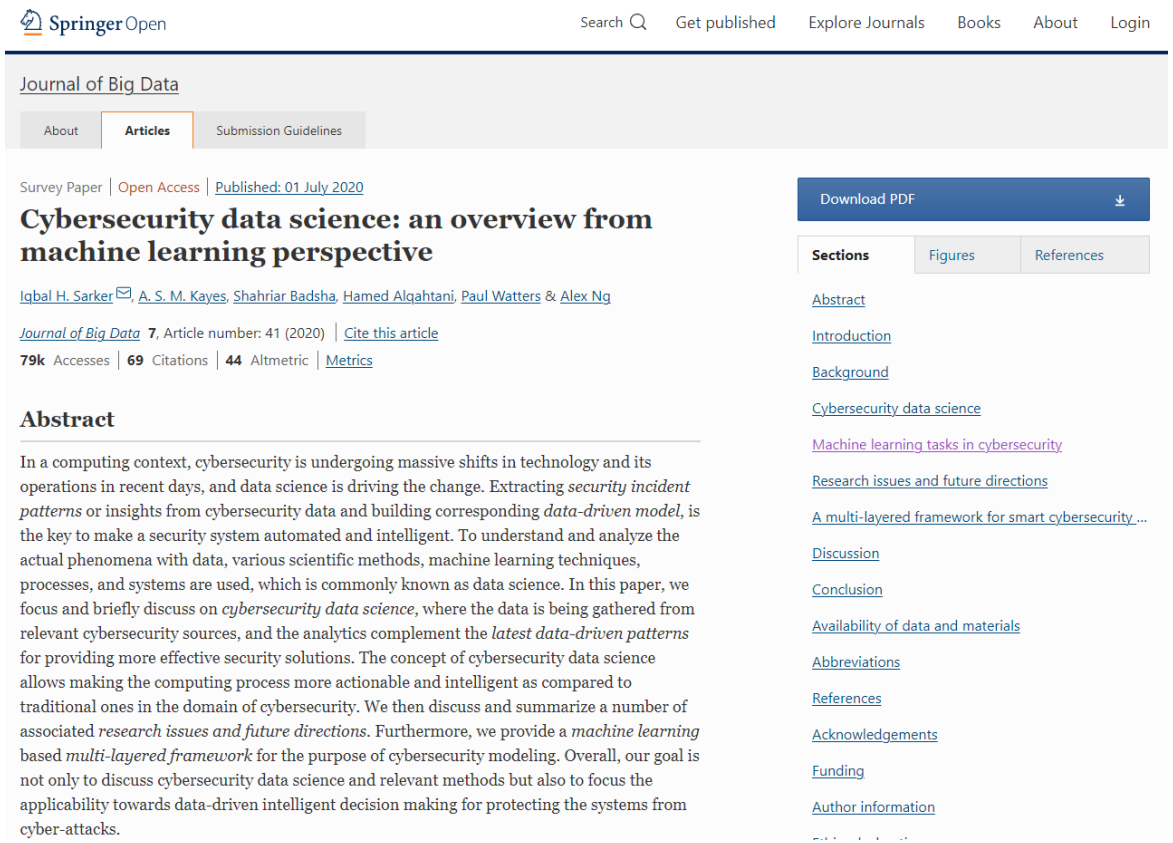
Buy eBook

Softcover Book EUR 46.79

Learn about institutional subscriptions

https://link.springer.com/book/10.1007/978-1-4842-6225-2

5. Introducció a l'aplicació de ciència de dades en ciberseguretat



SpringerOpen Search Get published Explore Journals Books About Login

Journal of Big Data

About | Articles | Submission Guidelines

Survey Paper | Open Access | Published: 01 July 2020

Cybersecurity data science: an overview from machine learning perspective

Iqbal H. Sarker, A. S. M. Kayes, Shahriar Badsha, Hamed Alqahtani, Paul Watters & Alex Ng

Journal of Big Data **7**, Article number: 41 (2020) | Cite this article

79k Accesses | **69** Citations | **44** Altmetric | Metrics

Abstract

In a computing context, cybersecurity is undergoing massive shifts in technology and its operations in recent days, and data science is driving the change. Extracting *security incident patterns* or insights from cybersecurity data and building corresponding *data-driven model*, is the key to make a security system automated and intelligent. To understand and analyze the actual phenomena with data, various scientific methods, machine learning techniques, processes, and systems are used, which is commonly known as data science. In this paper, we focus and briefly discuss on *cybersecurity data science*, where the data is being gathered from relevant cybersecurity sources, and the analytics complement the *latest data-driven patterns* for providing more effective security solutions. The concept of cybersecurity data science allows making the computing process more actionable and intelligent as compared to traditional ones in the domain of cybersecurity. We then discuss and summarize a number of associated *research issues and future directions*. Furthermore, we provide a *machine learning* based *multi-layered framework* for the purpose of cybersecurity modeling. Overall, our goal is not only to discuss cybersecurity data science and relevant methods but also to focus the applicability towards data-driven intelligent decision making for protecting the systems from cyber-attacks.

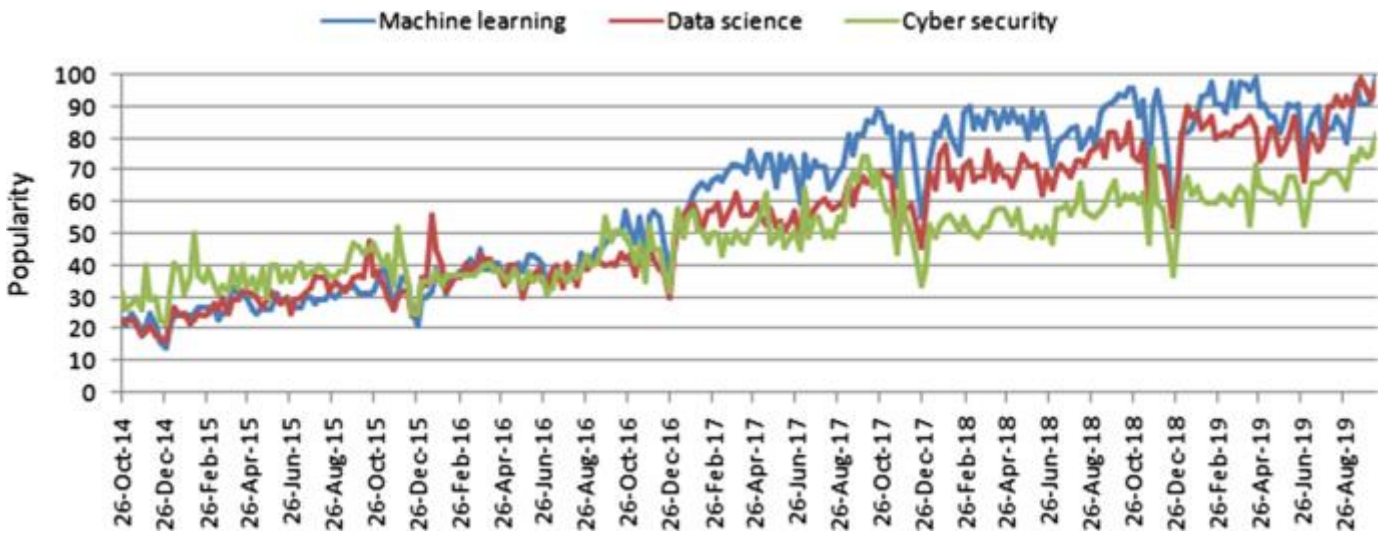
Download PDF

Sections | Figures | References

Abstract
Introduction
Background
Cybersecurity data science
Machine learning tasks in cybersecurity
Research issues and future directions
A multi-layered framework for smart cybersecurity ...
Discussion
Conclusion
Availability of data and materials
Abbreviations
References
Acknowledgements
Funding
Author information

https://journalofbigdata.springeropen.com/articles/10.1186/s40537-020-00318-5#Sec10

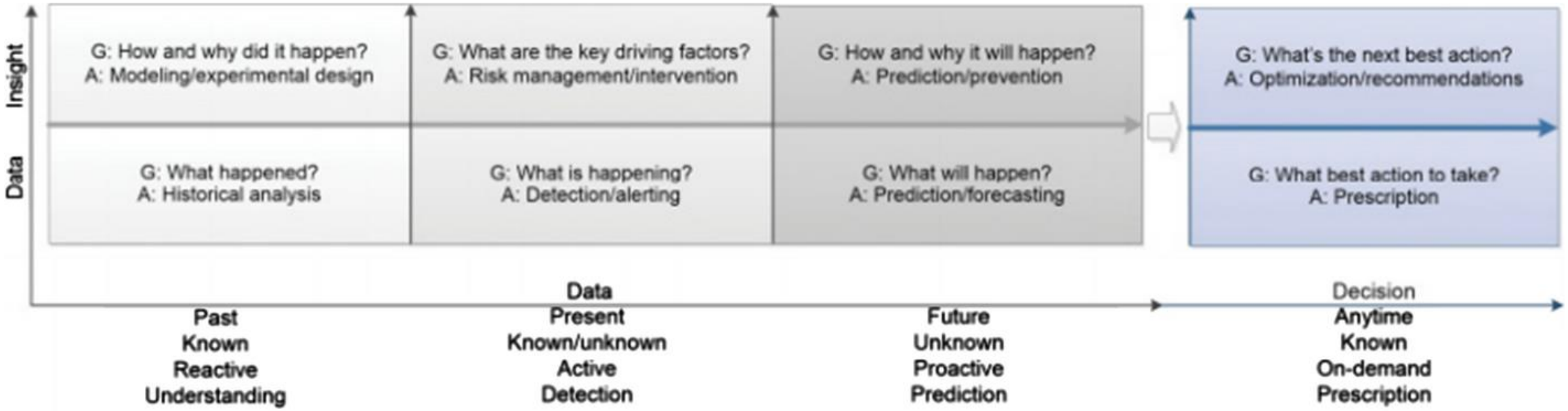
Popularity trends



Introducció regressió, classificació i clustering

- https://edarxiv.org/qba8w
- https://www.mdpi.com/2076-3417/9/22/4960
- https://archaeologytutor.herokuapp.com

Data analytic stages in data science



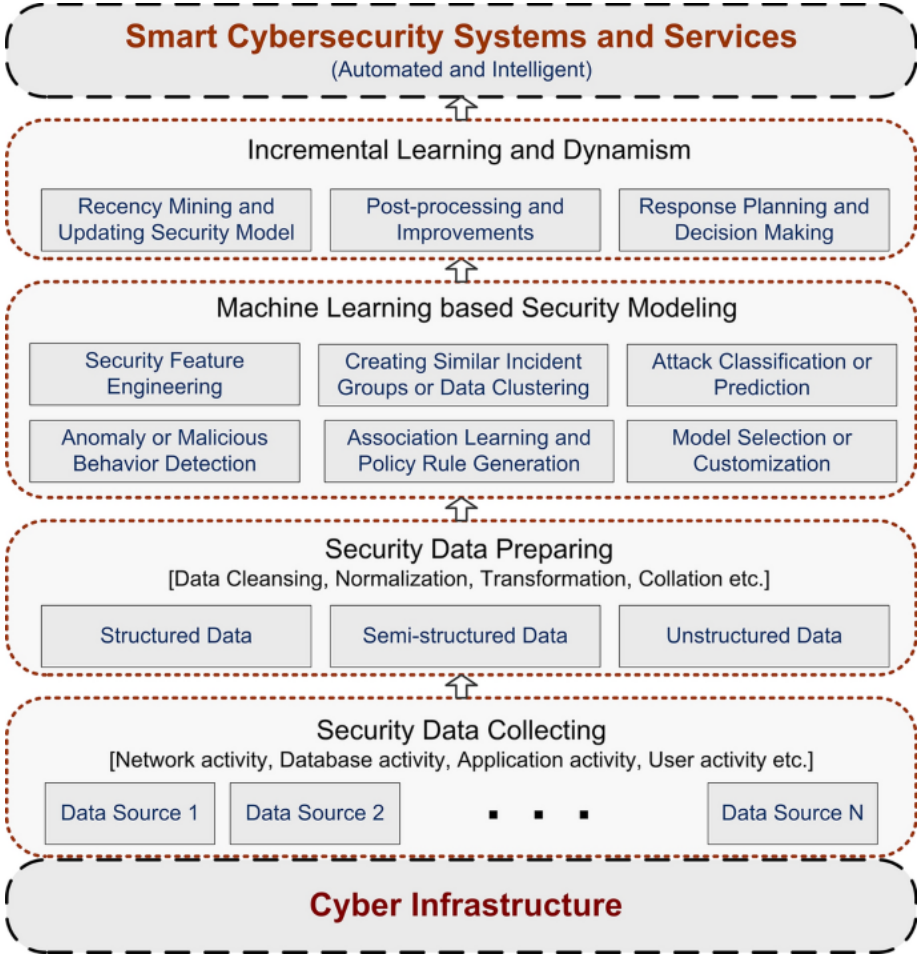
Data science and cybersecurity

Used Technique	Purpose	References
SVM	To classify various attacks such as DoS, Probe, U2R, and R2L	Kotpalliwar et al. [85]
SVM	Feature selection, intrusion detection and classification	Pervez et al. [86], Yan et al. [87], Li et al. [88], Raman et al. [89]
SVM	DDoS detection and analysis in SDN-based environment	Kokila et al. [90]
SVM	Evaluating host-based anomaly detection systems	Xie et al. [91]
SVM-PSO	To build intrusion detection system	Saxena et al. [92]
FCM clustering, ANN and SVM	To build network intrusion detection system	Chandrasekhar et al. [93]
KNN	Network intrusion detection system	Shapoorifard et al. [94], Vishwakarma et al. [95]
KNN	To reduce the false alarm rate	Meng et al. [96]
SVM and KNN	To build intrusion detection system	Dada et al. [97]
K-means and KNN	To build intrusion detection system	Sharifi et al. [98]
KNN and Clustering	To build intrusion detection system	Lin et al. [99]
Naive Bayes	To build an intrusion detection system for multi-class classification.	Koc et al. [100]
Decision Tree	To detect the malicious code's behavior information by running malicious code on the virtual machine and analyze the behavior information for intrusion detection.	Moon et al. [101]
Decision Tree	Feature selection and to build an effective network intrusion detection system	Ingre et al. [102], Malik et al. [103], Relan et al. [104], Rai et al. [105], Sarker et al. [106], Puthran et al. [107]
Decision Tree and KNN	Anomaly intrusion detection system	Balogun et al. [108]
Genetic Algorithm and Decision Tree	To solve the problem of small disjunct in the decision tree based intrusion detection system	Azad et al. [109]
Decision Tree and ANN	To measure the performance of intrusion detection system	Jo et al. [110]
Random Forests	To build network intrusion detection systems	Zhang et al. [111]
Association Rule	To build network intrusion detection systems	Tajbakhsh et al. [112]
Behavior Rule	To build intrusion detection system for safety critical medical cyber physical systems	Mitchell et al. [113]
Supervised	For malware detection and analysis	Alazab et al. [114], Alazab et al. [4]
Semi-supervised Adaboost	For network anomaly detection	Yuan et al. [115]
Hidden Markov Models	To build an intrusion detection system	Ariu et al. [116], Aarnes et al. [117]
Genetic Algorithm	For prevention of cyberterrorism through dynamic and evolving intrusion detection	Hansen et al. [118], Aslahi et al. [119]
Deep Learning Recurrent, RNN, LSTM	To build anomaly intrusion detection system and attack classification	Alrawashdeh et al. [120], Yin et al. [121], Kim et al. [122], Almiani et al. [123]
Deep Learning Convolutional	Malware traffic classification system	Kolosnjaji et al. [124], Wang et al. [125]
Deep and Reinforcement Learning	Malicious activities and intrusion detection system	Alauthman et al. [126], Blanco et al. [127], Lopez et al. [128]

Future directions

- Cibersecurity datasets
- Handling quality problems in cybersecurity datasets
- Security policy rule generation
- Hybrid learning method
- Protecting the valuable security information
- Context-awareness in cybersecurity
- Feature engineering in cybersecurity
- Remarkable security alert generation and prioritizing
- Recency analysis in cybersecurity solutions

A generic multi-layered framework based on machine learning techniques for smart cybersecurity services

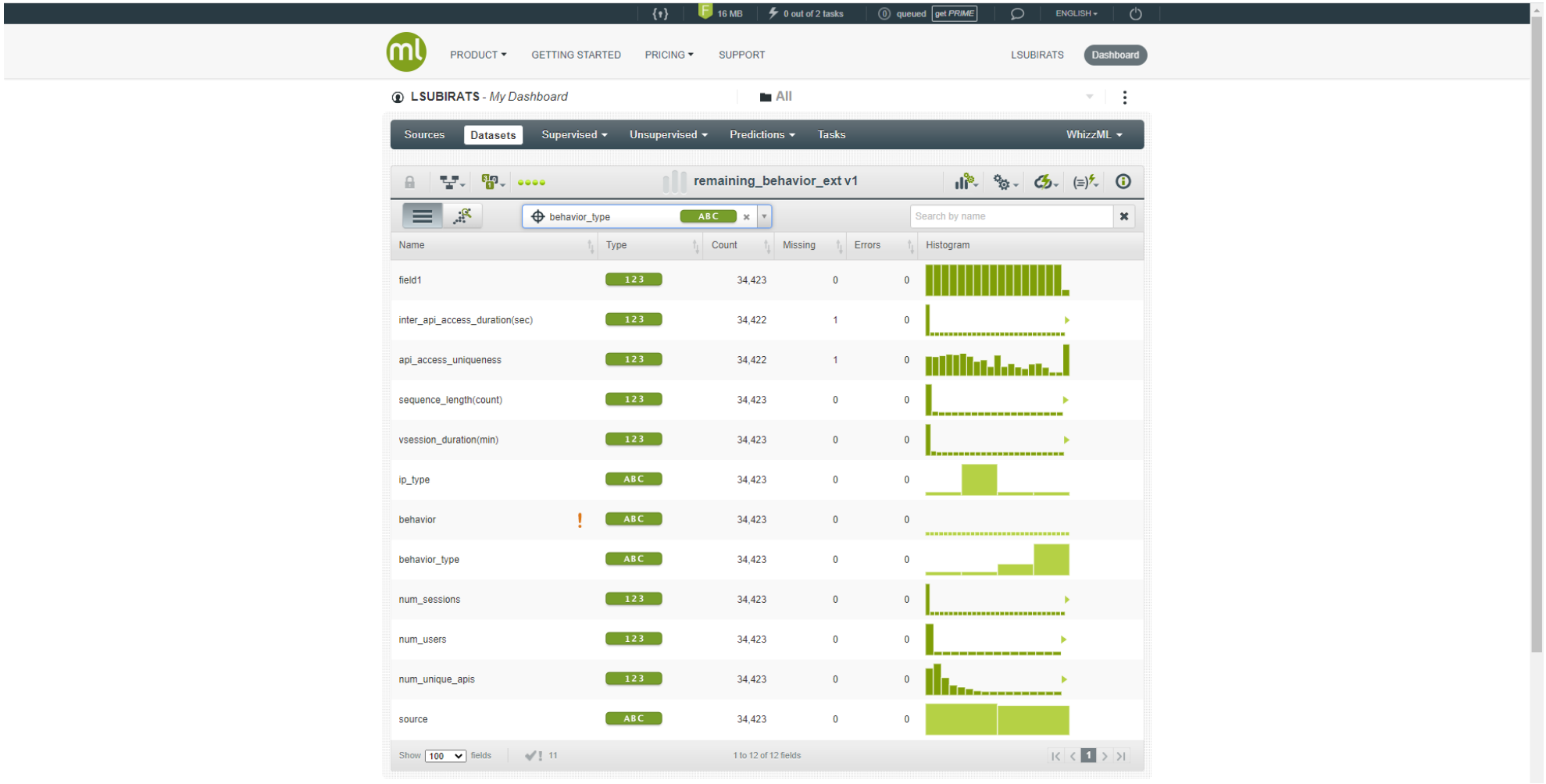


Enllaços útils

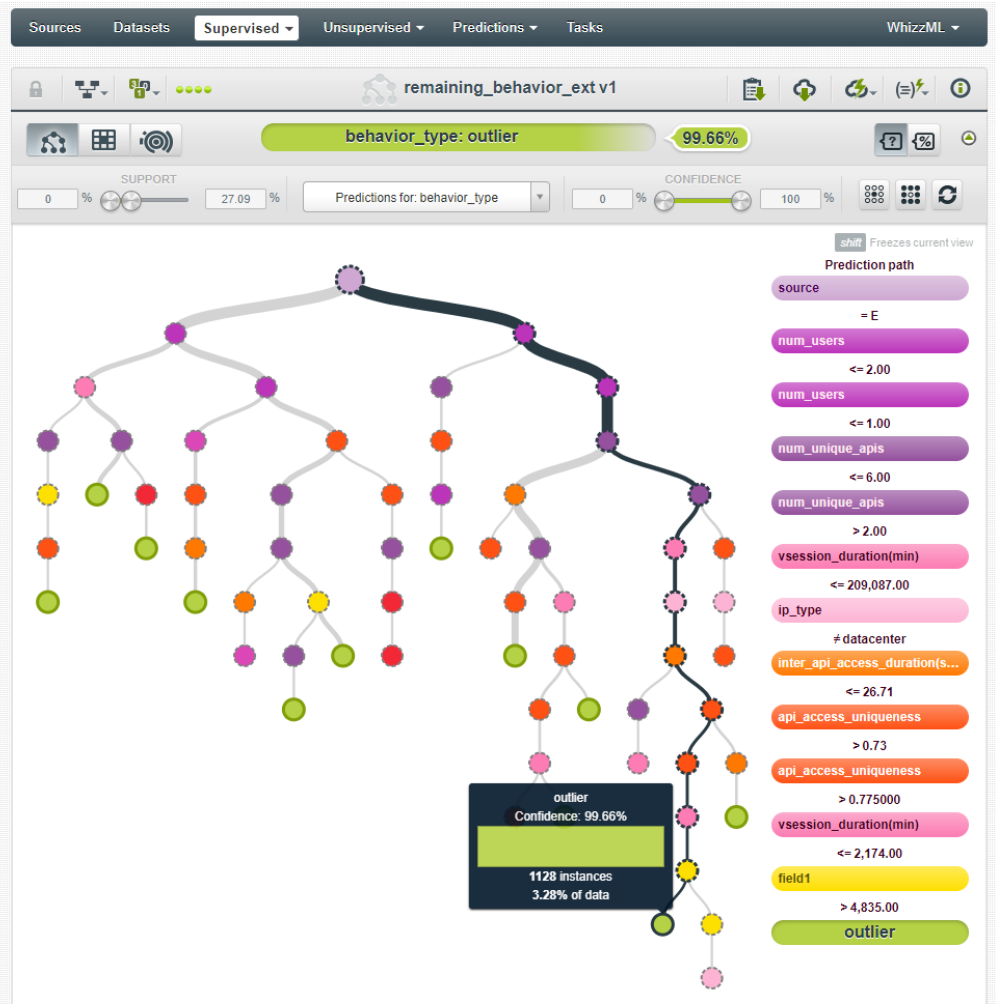
- https://bigml.com/

API security: Access behavior anomaly dataset | Kaggle

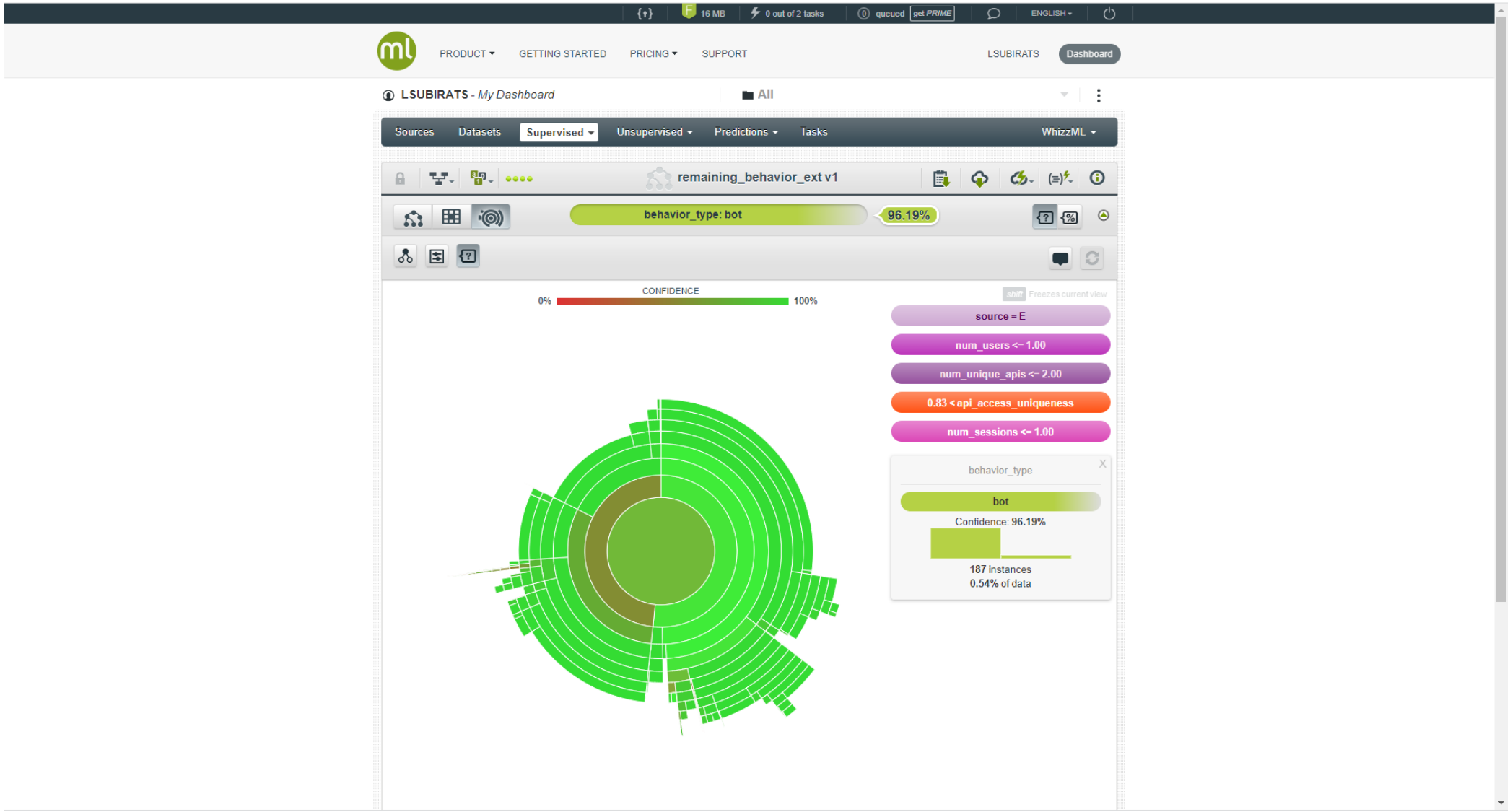
Exemple BigML i ciberseguretat



Exemple BigML i ciberseguretat



Exemple BigML i ciberseguretat



Ajuntament de Barcelona

Barcelona Activa

barcelonactiva.cat