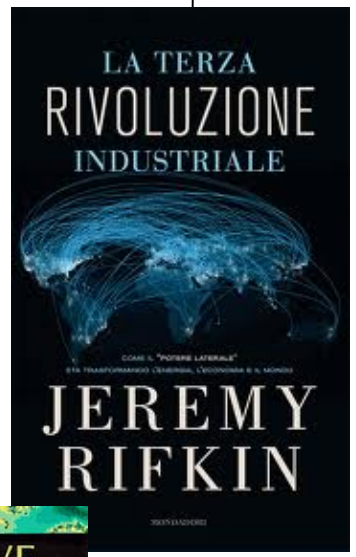
Prospettive dei dati aperti in Italia e in Europa: il punto sulla protezione dalle discriminazioni e l'accesso civico ai dati

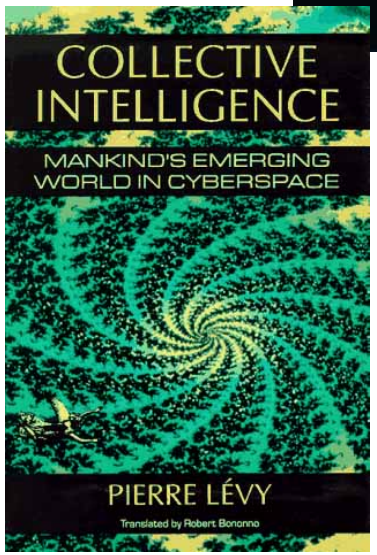
Monica Palmirani
CIRSFID, Università di Bologna
3 Marzo 2018, Ferrara

Internet e la “Società della conoscenza” condivisa

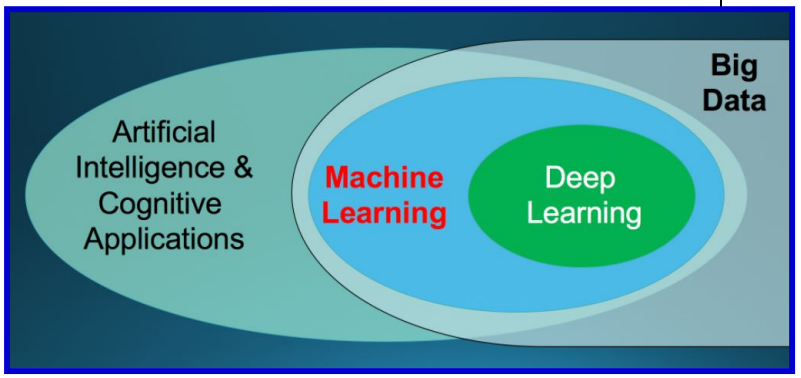
Dati/informazioni + Esperienza = *Conoscenza*



Open Source INTelligence

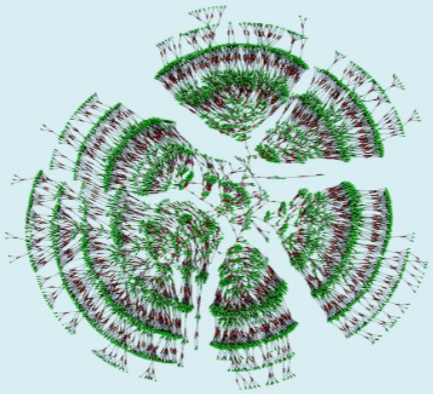


Esempi di applicazione di AI e big open data



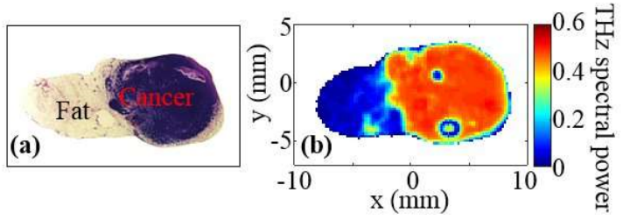
Examples

An example of a (small) FamiLinx pedigree of 6,000 people that spans over 7 generations:
Green nodes denote individuals and red nodes denote marriages



Researchers move closer to improved method of detecting breast cancer

March 1, 2018, University of Arkansas



Pathology image, left, and corresponding terahertz image, right, of excised tissue from mouse breast tumor. Credit: University of Arkansas

Engineering researchers at the University of Arkansas have moved closer to developing an alternative method of detecting and possibly treating breast cancer.

- https://youtu.be/NCu0O9Buf0o

Health

Followers 2

Datasets 7.6k

Follow

Countries

- United Kingdom (3349)
- Germany (1220)
- Spain (815)
- Ireland (619)
- France (421)
- Netherlands (328)
- Moldova, Republic of (196)
- Italy (195)
- Austria (190)
- Belgium (85)

PORTALE EUROPEO DEI DATI

- **805.060 dataset totali**
- **7.694 dataset sulla salute**
- **0,95%**

Cerca set di dati per categoria di dati

Agricoltura, pesca, silvicoltura e prodotti alimentari

Energia

Regioni e città

Trasporti

Economia e finanza

Affari internazionali

Governo e settore pubblico

Giustizia, sistema legale e pubblica sicurezza

Ambiente

Istruzione, cultura & sport

Salute

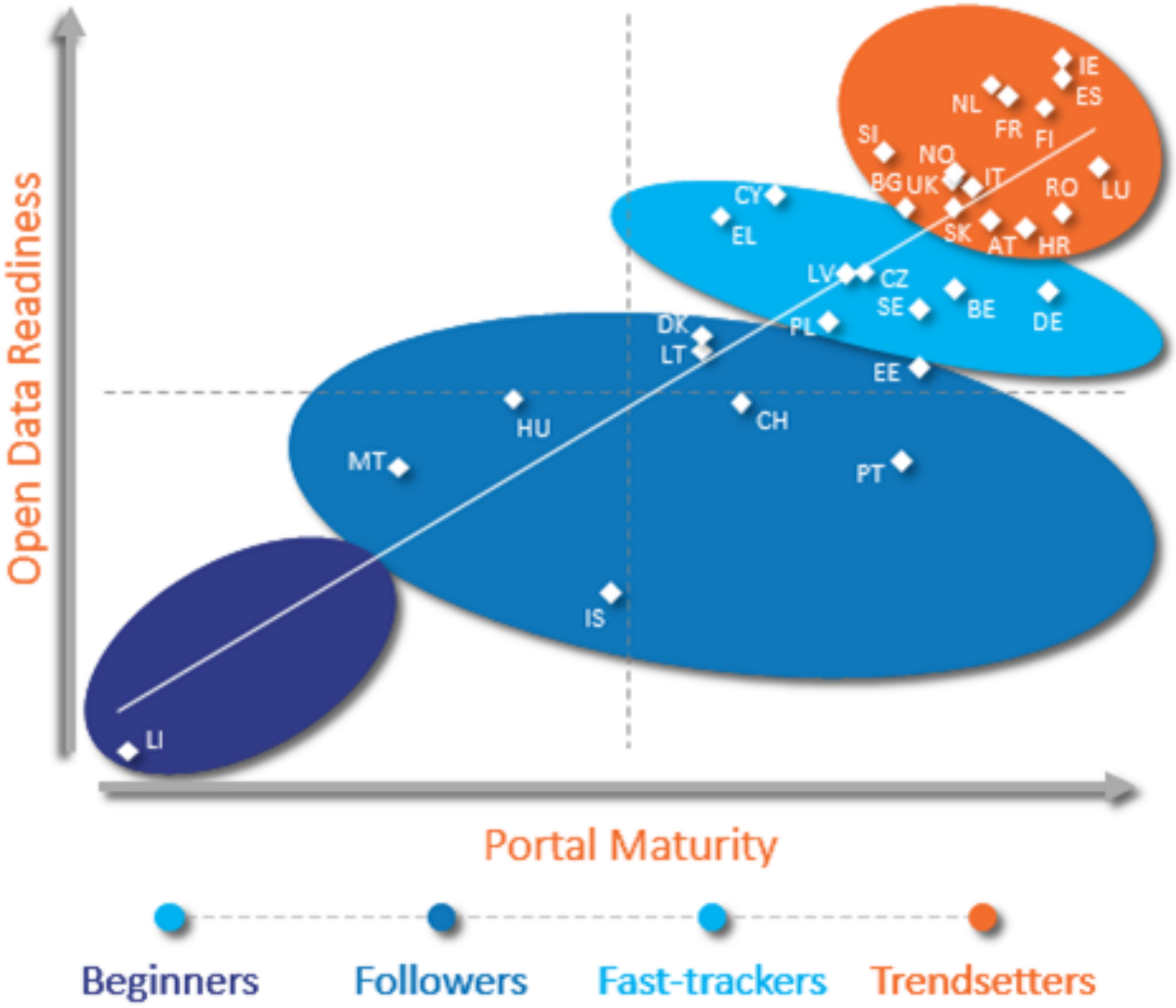
Popolazione e società

Scienza e tecnologia

Cataloghi

Tutti i dati

Maturità 2017



Barometro degli open data 2017

Italy – Overview

EUROPEAN DATA PORTAL

Portal

- http://www.dati.gov.it/ launched in 2011
- A large number of local and regional portals exist

Ranking

- Open Data Barometer Rank: #20
- Open Data Index Rank: #32

Transposition PSI Directive

✓ Yes

Revised PSI Directive

Open Data Policy

- ✓ Open Data policy in place
- ✓ More than 8 events held annually
- ✓ Priority domains identified
- ✓ Pre-defined approach to ensure data sets are up-to-date
- ✓ National 5 year strategy
- ✓ 90-99% data uploaded automatically

91%

Licensing Norms

- ✓ 100% data free of charge
- ✓ 100% data open licensed
- ✓ Policy encourages to use the CC-BY 4.0 international license

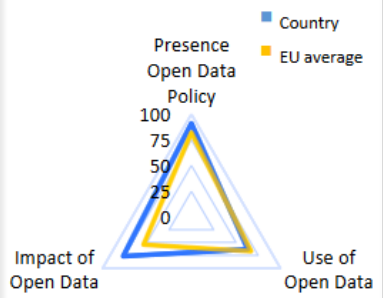
100%

National coordination

- ✓ National guidelines on publication
- ✓ Many regional data initiatives
- ✓ Many regional portals integrated
- X Not all regional initiatives coordinated at national level

86%

Score Italy



Usability of portal

- ✓ API accessible
- ✓ Search datasets
- ✓ Download datasets
- ✓ Possibility to give feedback
- ✓ ≥90% of the datasets are machine readable
- X Contribute to portal

Use & Reusability of Open Data

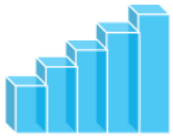
- 82,000 (0,014% of inhabitants in Italy) unique visitors on average per month
- 18% of the visitors are foreign
- 79,4% of traffic generated by humans
- Typical profile of visitors is mostly public sector
- Most datasets are available in CSV format
- Data published by other administrations is not re-used

17645 datasets available

70%

Top data set & domains

- Most downloaded data set: Race-goods-services.

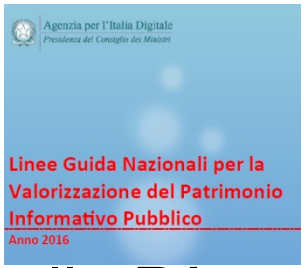


Top 5 most consulted domains:

1. Environment
2. Population and society
3. Education, culture and sport
4. Government and public sector
5. Economy and Finance

Piano Triennale 2017-2019

- Open Data pilastro dell'Agenda Digitale Italiana
- http://www.agid.gov.it/monitoraggio



- Punto del Piano triennale per l'informatica nella PA 2017-2019

Definizione dal piano triennale 2017

4.1.2 Open data

Gli *open data* sono definiti "dati di tipo aperto" nell'art. 68 del CAD e sono considerati elementi fondanti nel recepimento della Direttiva europea sull'informazione nel settore pubblico[49].

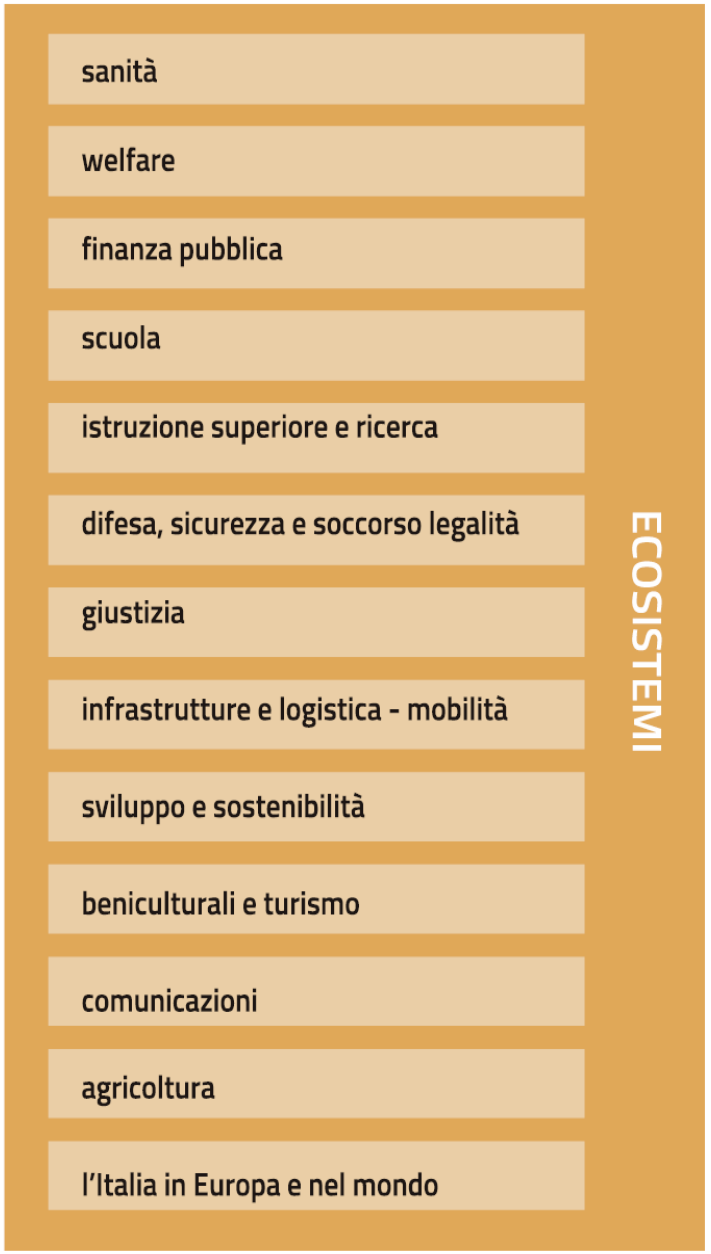
I dati pubblici sono aperti se:

- non sono riferibili a singole persone;
- sono resi disponibili in formato aperto, ovvero non proprietario, corredati dei relativi metadati;
- hanno associata una licenza che ne consente a chiunque il più ampio riutilizzo. Sono ammessi al massimo due vincoli: indicare la fonte di provenienza dei dati, riutilizzarli secondo gli stessi termini per cui sono stati licenziati originariamente;
- sono resi disponibili gratuitamente o ai soli costi marginali per la loro riproduzione e divulgazione, salvo casi eccezionali che siano trasparentemente e chiaramente identificati dalle amministrazioni titolari dei dati insieme ad AgID.

Ecosistemi

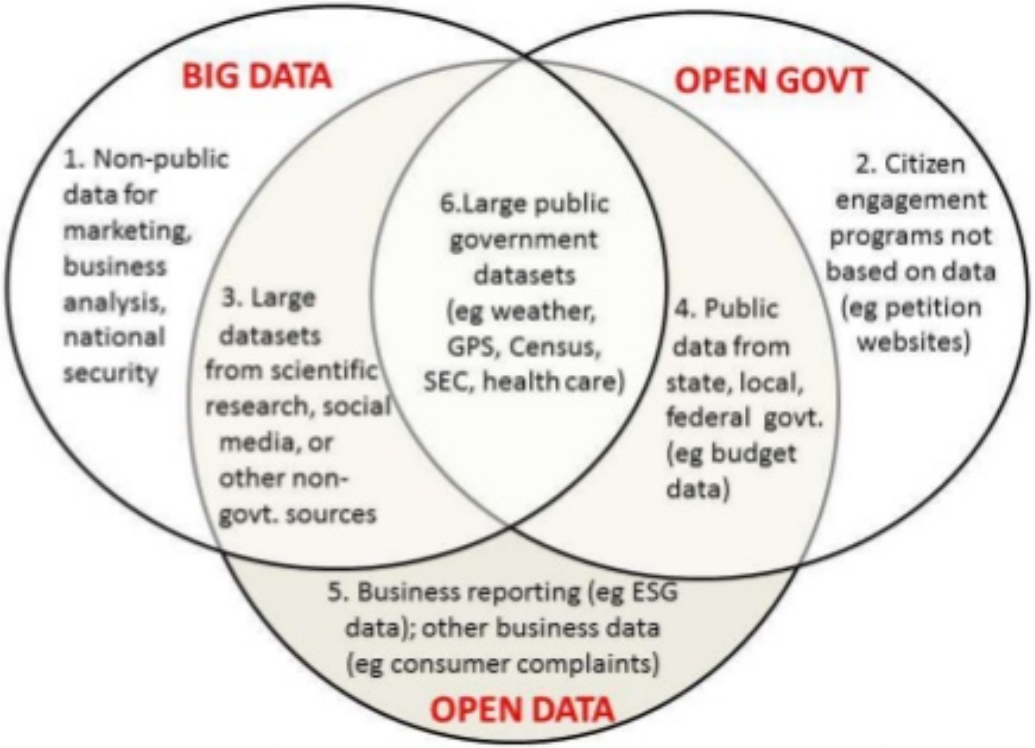
- Cultura, Turismo,
- Ambiente, Sanità

Figura 7 - Ecosistemi



Open Data e Big Data: un confine sempre più sottile

Buzzword Bingo 1/3: Open Data vs. Big Data vs. Open Government



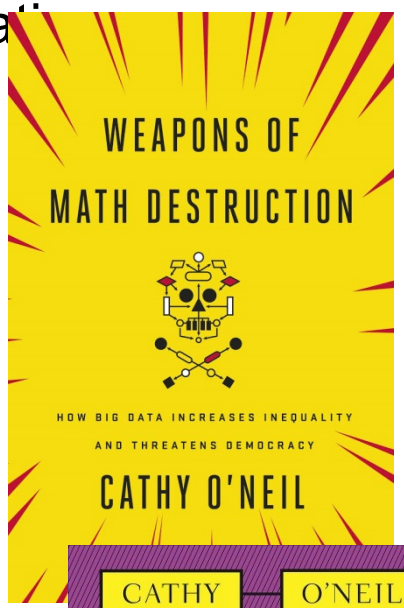
- http://www.opendatanow.com/2013/11/new-big-data-vs-open-data-mapping-it-out/ [1]

Big data/open data come arma distruttiva di massa ?

Possibili rischi dell'aggregazione dei dati

- Discriminazione
- Steriotipizzazione
- Deduzioni approssimate
- Predizione non corrette

- Non dobbiamo però cedere a due tentazioni:
  - Criminalizzare i dati (tutti i dati sono male)
  - Adottare una atteggiamento semplicistico all'apertura dei dati senza analisi di qualità (pubblico e non mi preoccupo)

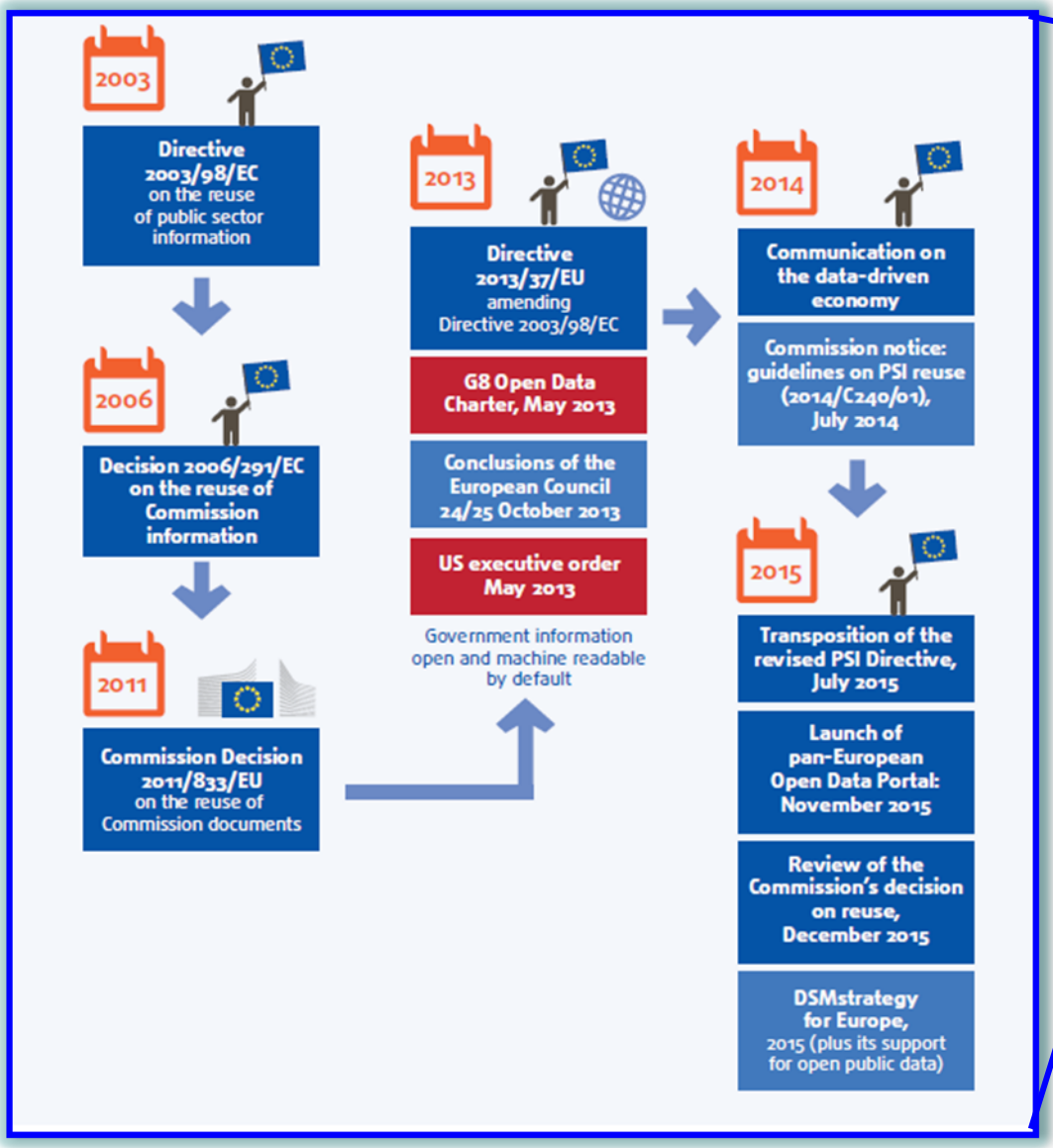


UNITED NATIONS HUMAN RIGHTS
OFFICE OF THE HIGH COMMISSIONER

A-72-43103_EN

- 73. Big Data challenges these principles while posing ethical issues and social dilemmas arising from the poorly considered use of algorithms. Rather than solving public policy problems, there is a risk of unintended consequences that undermine human rights such as freedom from all forms of discrimination, including against women, persons with disabilities and others.

Norme in supporto all'Open Data



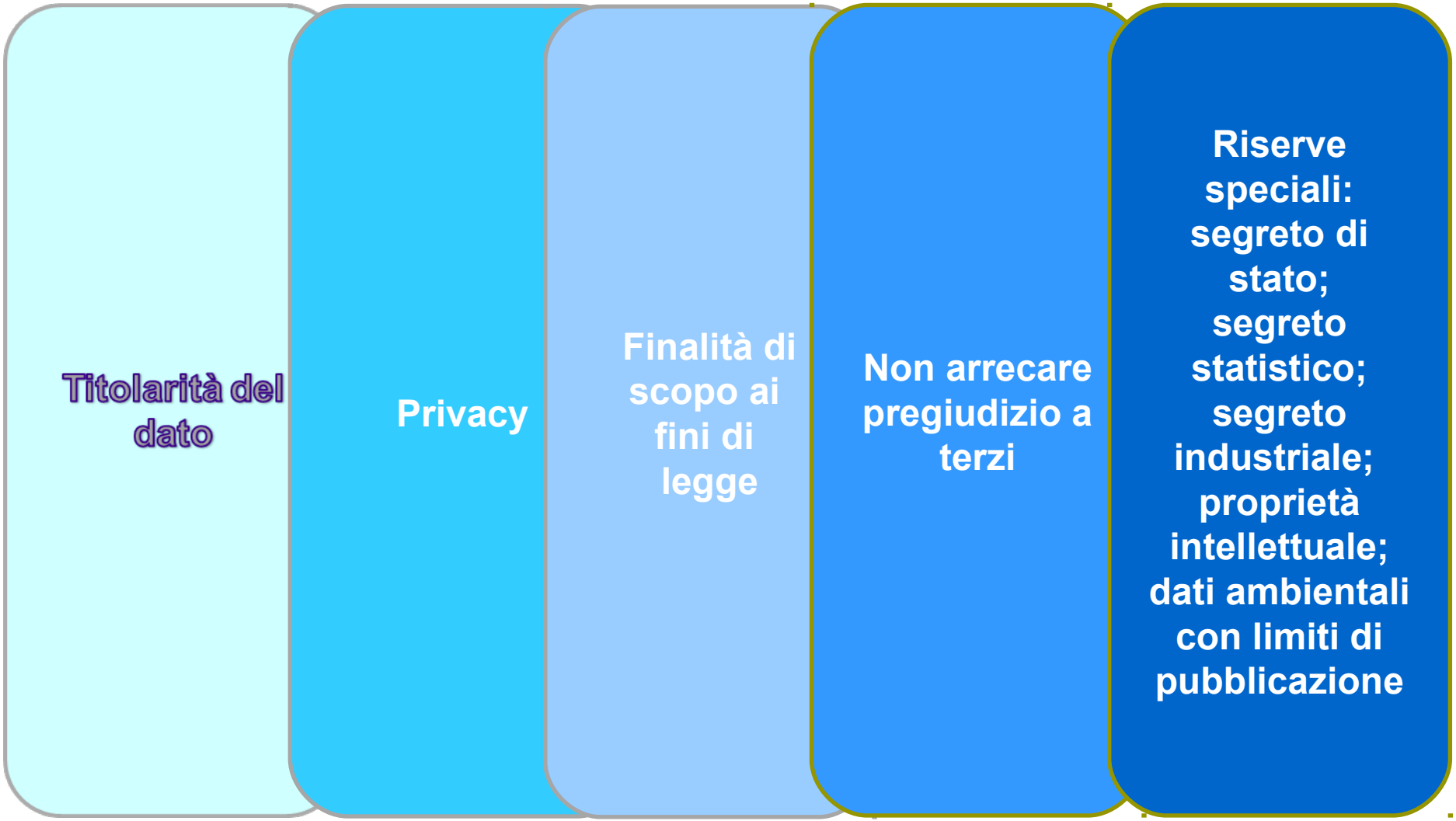
DIRETTIVA EUROPEA
PSI
2003
2013

CAD
Art. 52
2012

FOIA4ITALY
FOIA
2013
2016

GDPR
2016/697

Principi generali



Rischio De-anonimizzazione

- Il dato open di natura pubblica non deve contenere elementi - ***quasi-identificatori*** - tali da poter facilmente risalire all'identità degli individui o delle imprese a cui questi dati si riferiscono – **de-anonimizzazione o de-identificazione**

- **Art. 29 Working party – Opinion Adopted on 10 April 2014**
- **http://opengovfoundation.org/setting-the-standards-for-open-government-data/**

Health Data in an Open World

A report on re-identifying patients in the MBS/PBS dataset and the implications for future releases of Australian government data.

Chris Culnane, Benjamin Rubinstein and Vanessa Teague[1],
School of Computing and Information Systems
The University of Melbourne, 18 Dec 2017
{christopher.culnane, benjamin.rubinstein, vjteague}@unimelb.edu.au

Science. 2013 Jan 18;339(6117):321-4. doi: 10.1126/science.1229566.

Identifying personal genomes by surname inference.

Gymrek M[1], McGuire AL, Golan D, Halperin E, Erlich Y.

Author information

Abstract

Sharing sequencing data sets without identifiers has become a common practice in genomics. Here, we report that surnames can be recovered from personal genomes by profiling short tandem repeats on the Y chromosome (Y-STRs) and querying recreational genetic genealogy databases. We show that a combination of a surname with other types of metadata, such as age and state, can be used to triangulate the identity of the target. A key feature of this technique is that it entirely relies on free, publicly accessible Internet resources. We quantitatively analyze the probability of identification for U.S. males. We further demonstrate the feasibility of this technique by tracing back with high probability the identities of multiple participants in public sequencing projects.

Comment in

Genomic privacy in the information age. [Clin Chem. 2013]
Data re-identification: societal safeguards. [Science. 2013]

PMID: 23329047 DOI: 10.1126/science.1229566

[Indexed for MEDLINE] **Free full text**

- Privacy differenziale, Crittografia omomorfica e il multiparty computation
- Consenso dinamico

Pseudonimizzazione - GDPR

GDPR: dati personali, dati anonimi, dati pseudoanonimi

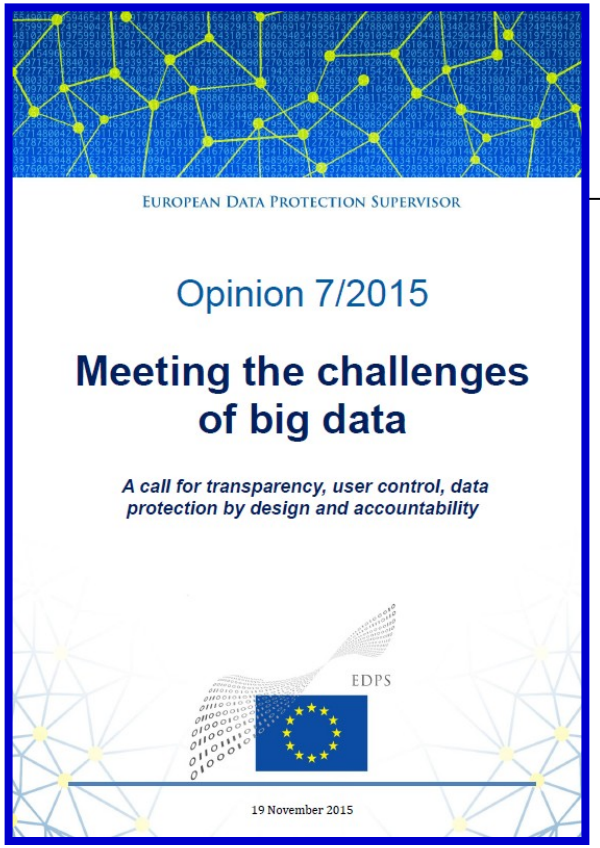
La “pseudonimizzazione”

«5) il trattamento dei dati personali in modo tale che i dati personali non possano più essere attribuiti a un interessato

specifico senza l'utilizzo di informazioni aggiuntive, a condizione che tali informazioni aggiuntive siano conservate separatamente e soggette a misure tecniche e organizzative intese a garantire che tali dati personali non siano attribuiti a una persona fisica identificata o identificabile» (art. 4, punto 5 del GDPR)

(28) L'applicazione della pseudonimizzazione ai dati personali può **ridurre i rischi per gli interessati e aiutare i titolari del trattamento e i responsabili del trattamento a rispettare i loro obblighi di protezione dei dati**. L'introduzione esplicita della «pseudonimizzazione» nel presente regolamento non è quindi intesa a precludere altre misure di protezione dei dati.

Opinion 7/2015



One result is the emergence of a revenue model for Internet companies relying on tracking online activity. Such 'big data' should be considered personal even where anonymisation techniques have been applied: it is becoming and will be ever easier to infer a person's identity by combining allegedly 'anonymous' data with publicly available information such as on social media. Furthermore, with the advent of the 'Internet of Things', much of the data collected and communicated by the increasing number of personal and other devices and sensors will be personal data: the data collected by them can be easily related to the users of these devices whose behaviour they will monitor. These may include highly sensitive data including health information and information relating to our thinking patterns and psychological make-up.

Opinion 7/2015

EUROPEAN DATA PROTECTION SUPERVISOR

- **Mancanza di trasparenza**
- **Asimmetria informativa**
- **Discriminazione**
- **Perdita di potenziale innovativo**

Opinion 7/2015

Meeting the challenges of big data

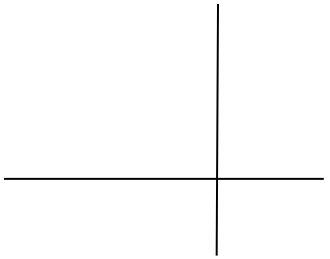
A call for transparency, user control, data protection by design and accountability

The expected benefits of statistics based prediction may further increase overconfidence in its capabilities. **Big data applications may find spurious correlations in data**, even in cases where there is no direct cause and effect between two phenomena that show a close correlation. In these cases there is a risk of drawing inaccurate but also – when applied at the individual level – **potentially unfair and discriminatory conclusions**.

These and other characteristics of big data, extensive use of automated decisions and predictive analysis may also lead to broader undesirable changes in the development of our societies. Importantly, they may lead to discrimination, re-enforcement of existing stereotypes, **social and cultural segregation and exclusion**[12].

The accumulation of massive personal data sets which feed big data analytics is possible because of the constant, invisible tracking of online activity. This surveillance may also have a **chilling effect on creativity and innovation**.

Tecniche di anonimizzazione e pseudonimizzazione



- randomization
  - Noise addition
  - Permutation
  - Differential privacy
- generalization
  - Aggregation and K-anonymity
  - L-diversity/T-closeness
- pseudonymisation (GDPR)
  - Encryption
  - Hash
  - Token
  - Crittografia omomorfica

Processo decisionale automatizzato relativo alle persone fisiche, compresa la profilazione - GDPR

«1. L'interessato ha il diritto di non essere sottoposto a una decisione basata **unicamente sul trattamento automatizzato**, compresa la profilazione, che produca effetti giuridici che lo riguardano o che incida in modo analogo significativamente sulla sua persona.

3. Nei casi di cui al paragrafo 2, lettere a) e c), il titolare del trattamento attua misure appropriate per tutelare i diritti, le libertà e i legittimi interessi dell'interessato, almeno il diritto di ottenere **l'intervento umano** da parte del titolare del trattamento, di esprimere la propria opinione e di contestare la decisione.» (art. 22, comma 1 e 3 – GDPR)

ARTICLE 29 DATA PROTECTION WORKING PARTY

17/EN

WP 251

Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679

Adopted on 3 October 2017

Profiling may be unfair and create discrimination, for example by denying people access to employment opportunities, credit or insurance, or targeting them with excessively risky or costly financial products. The following example illustrates how unfair profiling can lead to some consumers being offered less attractive deals than others.

Guidelines for De-identification of Personal Data

- Guide for De-identification Standards and Support/Management System -

Office for Government Policy Coordination | Ministry of Interior
Korea Communications Commission | Financial Services Commission
Ministry of Science, ICT and Future Planning | Ministry of Health and Welfare

JKMS Journal of Korean Medical Science jkms.org

J Korean Med Sci. 2018 Jan 29; 33(5): e41. PMCID: PMC5773854
Published online 2017 Dec 26. doi: 10.3346/jkms.2018.33.e41

Issues and Solutions of Healthcare Data De-identification: the Case of South Korea

Soo-Yong Shin

Author information ▶ Article notes ▶ Copyright and License information ▶

<Example> Attribute Values

Individual Characteristics	○ Gender, age, nationality, hometown, address, postal code, military service record, marital status, religion, hobbies, club/group affiliation, etc. ○ Smoking, drinking, vegetarian diet, matter of interests, etc.
Physical Characteristics	○ Blood type, height, weight, waist, blood pressure, eye color, etc. ○ Physical examination results, disability type, disability grade, etc. ○ Disease name or code, administration code, medical record, etc.
Credit Characteristics	○ Tax payment, credit rate, donation, etc. ○ Health insurance payment, income level, medical service recipient, etc.
Career Characteristics	○ Name of school, major, grade, academic performance, and academic background, etc. ○ Professional experiences, occupation, job category, company, department, position, previous jobs, etc.
Electronic Characteristics	○ Cookie data, date and time of login, date and time of visit, record of using services, access log, etc. ○ Internet access record, record of using mobile phones, GPS data, etc.
Family Characteristics	○ Family data (spouse, children, parents, and siblings), legal representative data, etc.

Crime maps – rischio di identificare persone fisiche

SFGOV

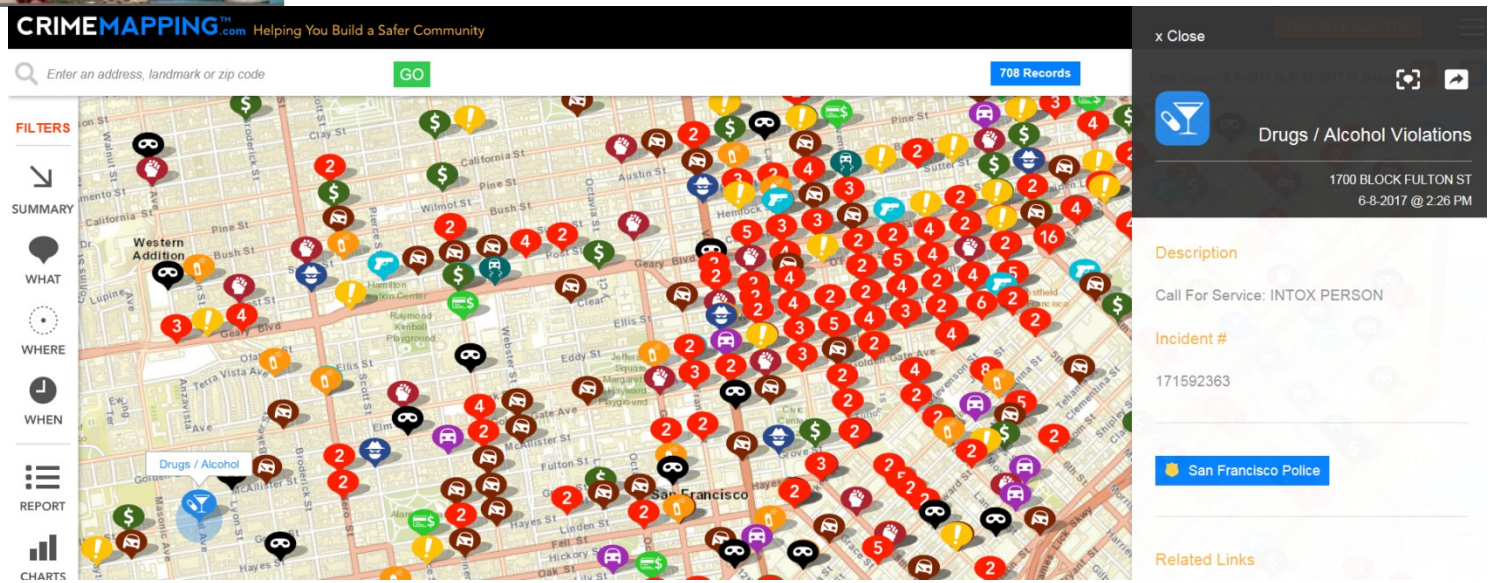
RESIDENTS BUSINESS OPEN GOV VISITORS ONLINE SERVICES HELP

Home > SF Crime Maps

SF Crime Maps

Access a map providing details about the locations where crimes are being committed within the City and County of San Francisco territory.

Link:
http://www.crimemapping.com/map/ca/sanfrancisco



23andMe®

These San Diego Scientists Can Predict How You Look Using Only Your Anonymous DNA

Monday, September 4, 2017

By David Wagner

CREDIT: PNAS

Investimento 150 milioni di dollari

- http://www.privacyitalia.eu/privacy-dati-sanitari-degli-italiani-regalati-ibm/2431/

Ma il Garante della privacy, **Antonello Soro**, già il 22 febbraio 2017 ha inviato una lettera a Governo e Regione Lombardia per avere chiarimenti sulla vicenda. Il Garante ha ribadito due concetti fondamentali per il rispetto della privacy dei cittadini italiani:

- I trattamenti di dati sanitari per fini di ricerca medica, biomedica ed epidemiologica possono prescindere dal consenso dell'interessato **solo quando la ricerca sia prevista da un'espressa disposizione di legge.**
- Possono trattare dati personali solo *"soggetti appositamente designati responsabili esterni del trattamento, individuati tra soggetti che, per esperienza, capacità e affidabilità, forniscano idonea garanzia del pieno rispetto delle vigenti disposizioni".*

FINANCIAL TIMES

WORLD US COMPANIES MARKETS OPINION WORK & CAREERS LIFE & ARTS

National Health Service + Add to myFT

NHS to trial artificial intelligence app in place of 111 helpline

London experiment will perform triage for urgent but non-life-threatening conditions

e app, like the one already built by Babylon, rather than pany; NHS

The Guardian view on Google's NHS grab: legally inappropriate

Editorial

FINANCIAL TIMES

HOME WORLD US COMPANIES MARKETS OPINION WORK & CAREERS LIFE & ARTS

Data protection + Add to myFT

DeepMind given 'legally inappropriate' access to NHS data

Records were shared with the Google-owned AI company to trial Streams health

Check list Vademecum

- Privacy
- Proprietà intellettuale della sorgente
- Licenza di rilascio
- Limiti alla pubblicazione
- Segretezza
- Condizioni economiche
- Temporalizzazione

Appendix A: Check-list

Privacy		Si/No
	sono i dati liberi da ogni informazione personale che possa identificare in modo diretto l'individuo (nome, cognome, indirizzo, codice fiscale, patente, telefono, email, foto, descrizione fisica, etc.)?	
	sono i dati liberi da ogni informazione indiretta che possa identificare l'individuo (caratteristiche personali che possono identificare facilmente il soggetto)? In caso negativo queste informazioni sono autorizzate per legge?	
	sono i dati liberi da ogni informazione sensibile che può essere ricondotta all'individuo? In caso negativo queste informazioni sono autorizzate per legge	
	sono i dati liberi da ogni informazione relativa al soggetto che incrociata con dati comunemente reperibili nel web (e.g. google maps, linked data, etc.) possa identificare l'individuo? In caso negativo queste informazioni sono autorizzate per legge?	
	sono i dati liberi da ogni record relativo a profughi, protetti di giustizia, vittime di violenze o in ogni caso categorie protette?	
	hai usato un tool per calcolare il rischio di de-anonimizzazione del tuo dataset prima di pubblicarlo?	

The Data Ethics Canvas user guide

September 2017

Data Ethics Canvas

What are your data sources?

Name and describe key data sources used in your project, whether you're collecting them yourself or getting access from third parties.

Who has rights over your data sources?

Where did you get the data from? e.g. is it data produced by an organisation or data collected directly from individuals?

Do you have permission or another basis on which you're allowed to use this data? What ongoing rights will the data source have?

What's your core purpose for using this data?

What is your primary use case, your business model?

Are you collecting more data than is needed for your purpose?

Who could be negatively affected?

Could the manner in which this data is collected, shared, used cause harm?

- be used to target, profile, prejudice people
- unfairly restrict access (eg exclusive arrangements)

Could people "perceive" it to be harmful?

Are you communicating potential risks/issues, if any?

How are limitations and risks being communicated to people affected by your project, and organisations using data?

What channels are you using?

Are there any limitations in your data sources?

Which might influence the outcomes of your project, like:

- bias in data collection, inclusion, algorithm
- gaps, omissions
- other sensitivities

What policies/laws shape your use of this data?

Data protection legislation, IP and database rights legislation, sector specific data sharing policies/regulation (e.g. health, employment, taxation)

Sector specific ethics legislation?

Do people understand your purpose?

If this is a project/use that could impact on people or more broadly shape/impact society, do people understand your purpose?

Has this been clearly communicated to them?

How are you minimising negative impact?

What steps can you take to minimise harm? Are there measures you could take to reduce limitations in your data sources? Could you monitor potential negative impact to support mitigating activities? What benefits will these actions add to your project?

When is your next review?

When will this Data Ethics Canvas be reviewed?

How will ongoing issues be monitored?

Are you going to be sharing this data with other organisations?

If so, who?

Who will be positively affected by this project?

What individuals, demographics, organisations?

How will they be positively affected?

Do they know and understand how they are positively affected?

How can people engage with you?

Can people affected appeal or request changes to the service? To what extent?

Are the appeal mechanisms reasonable?

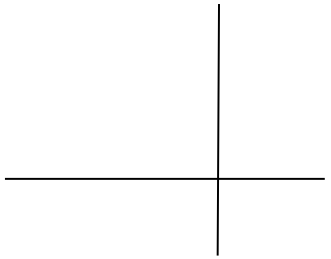
What are your actions?

What steps are you going to take prior to moving forward with this project?

Un bilanciamento di fenomeni mediante principi etici

Tutele	Diritto di Accesso
• privacy • dati statistici • licenza con limitazioni di riuso • a pagamento • metadati locali	• trasparenza • dati grezzi • cc0, senza limitazioni di riuso • gratuito • metadati condivisi


KEEP
CALM
AND
OPEN
DATA



- Licenze

Attenzione alla titolarità

- Se il titolare del dato impedisce il processo di open data un altro soggetto non può aprire i dati con licenza aperta
- Esempio del Catasto e Agenzia delle Entrate
  - D.L. 2 marzo 2012 n. 16, dal 1° ottobre 2012

Linee guida AGID 2016

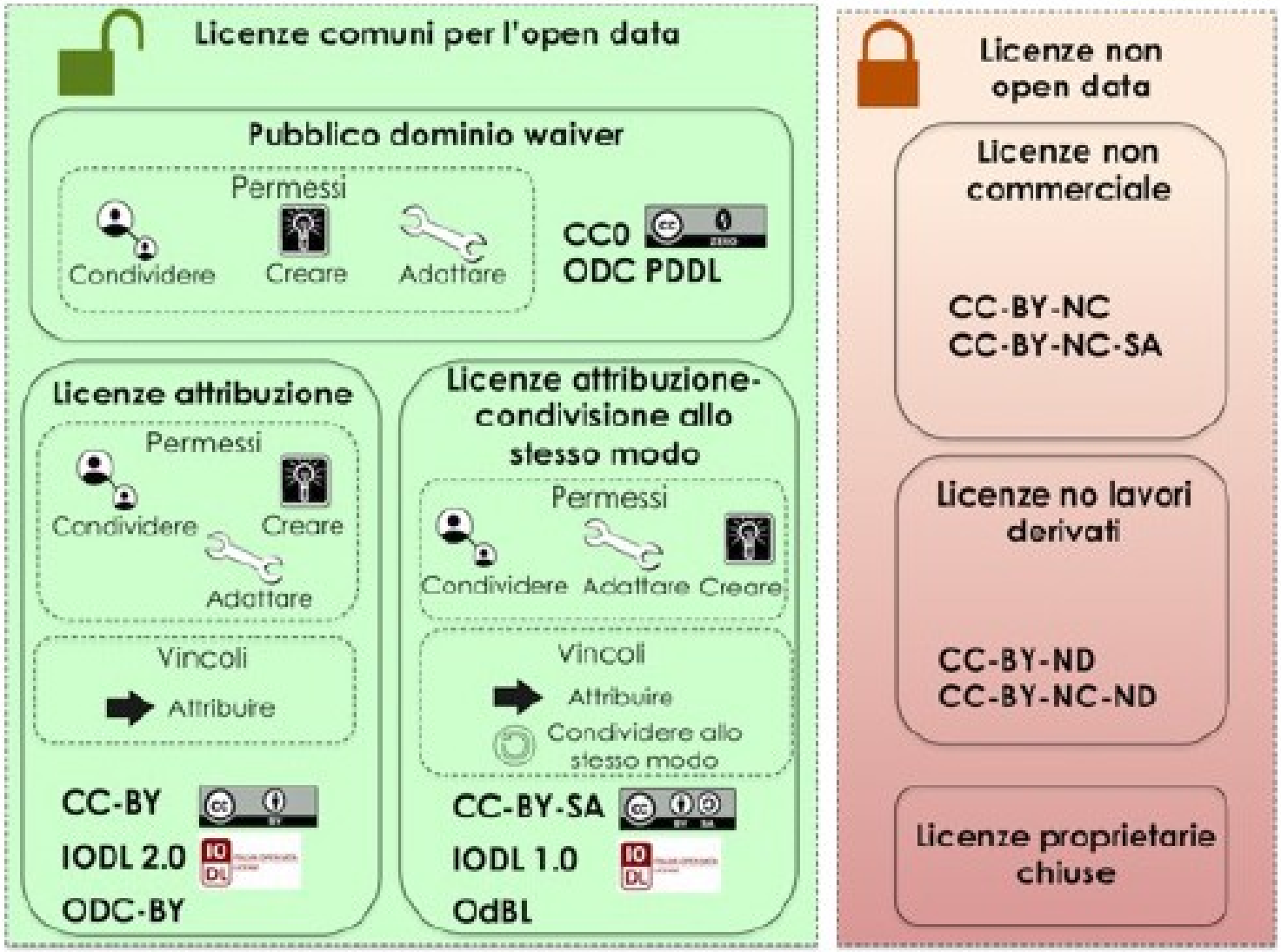


Figura 7: Licenze aperte e non aperte per i dataset

Licenze compatibilità

COMPATIBILITÀ TRA LICENZE

Un'indicazione di compatibilità tra le licenze Open Data indicate in Figura 7 è riportata nella seguente tabella[13]:

Licenza opera derivata / Licenza opera originaria	CC0	CC-BY	CC-BY-SA	IODL v. 2.0	IODL v. 1.0	ODbL
CC0						
CC-BY						
CC-BY-SA						
IODL v. 2.0						
IODL v. 1.0						
ODbL						

Tabella 1: Compatibilità tra licenze

La creazione di un'opera derivata e la sua pubblicazione è possibile

La creazione di un'opera derivata potrebbe essere possibile ma vi è incertezza (ad esempio sui diritti licenziati) circa l'effettiva compatibilità o altri problemi (problema di stratificazione delle attribuzioni), oppure sul tipo di prodotto derivato (e.s. per la ODbL le modifiche dei dati sono rilasciabili solo con ODbL mentre i prodotti derivati come le mappe con ogni altra licenza).

La creazione di un'opera derivata sotto la licenza proposta è impossibile

Creative Commons

https://creativecommons.org/licenses/by/4.0/legalcode

	Can someone use it commercially?	Can someone create new versions of it?
Attribution		
Share Alike		Yup, AND they must license the new work under a Share Alike license.
No Derivatives		
Non-Commercial		Yup, AND the new work must be non-commercial, but it can be under any non-commercial license.
Non-Commercial Share Alike		Yup, AND they must license the new work under a Non-Commercial Share Alike license.
Non-Commercial No Derivatives		

- Bologna –CC-by 3.0 it
- Barcelona – CC BY 3.0

- London – licenza non esclusiva
  - Creative Commons BY-NC-SA
  - Crown Copyright
  - Open Government License
  - Open Database License
  - Open Data License