

Dati, GDPR e Data Ethics

- **Cambiamento radicale** del mondo in cui viviamo dal punto di vista della raccolta e conservazione dei dati
- Dibattito filosofico e legale su **giusto** e **legittimo** uso dei dati
- Come creatori, produttori e utilizzatori di dati dovremmo interrogarci su considerazioni etiche

- Bilanciamento tra rischi e benefici . . . zero rischi, zero benefici
- **Data Ethics**: principi che portano a concludere che cosa è giusto fare
- **Accettabilità** e **valore sociale** del Data Science

- Non è il dato in sè a comportare interrogativi etici, **ma l'uso che se ne fa**
- Benefici previsti: salute, sicurezza, educazione, giustizia
- Rischi potenziali: sicurezza, mancato rispetto della privacy, arbitrarietà nelle decisioni
- **Etica**: ciò che una persona **dovrebbe** fare
- **Legge**: ciò che una persona **deve** fare

Ethics is knowing the difference between
what you have the right to do ... and what is right to do

- Pervasività delle 'data technologies' le rende una **infrastruttura** della società
- Interconnessione tra dati
- Natura dinamica dei dati
- Analisi real-time e online (aggiornamento continuo)
- Capacità di usare i dati per obiettivi imprevisti e rivelare informazioni inattese (data mining)
- Rischi di intrusività e abusi

I dati sono spesso descritti come il nuovo petrolio, ma con una importante differenza: i dati non vengono consumati, possono essere utilizzati nuovamente e per scopi diversi

- L'era dei Big Data ha reso disponibile una enorme mole di dati a livello individuale, relativamente al comportamento delle persone e alle loro connessioni
- Si parla in questo senso di **Behavioural Big Data**, BBD
- Si è reso quindi sempre più necessario sviluppare **principi etici** e linee guida allo scopo di bilanciare i potenziali benefici scientifici e sociali dei BBD con una maggiore esigenza di rispetto della privacy

- La **European General Data Protection Regulation** è il risultato di una lunga evoluzione e maturazione di un insieme di norme che estende il tradizionale concetto di privacy
- La **protezione dei dati personali** è considerata dalle istituzioni Europee come un diritto fondamentale
- La GDPR è particolarmente importante perchè si tratta di un **regolamento**, che trascende i processi legislativi nazionali ed ha applicazione immediata

- La **European General Data Protection Regulation** è stata progettata per porre limiti e restrizioni all'uso dei dati personali da parte di compagnie e organizzazioni che operano nella UE e al di fuori dell'UE, nella misura in cui queste organizzazioni *monitorano il comportamento e offrono beni e servizi* a soggetti residenti nella UE (art. 24)
- La GDPR può quindi interessare qualsiasi azienda o organizzazione che utilizzi dati di soggetti residenti nella UE, a prescindere da dove questo avvenga
- L'utilizzo di dati personali a livello industriale è sempre stato meno regolamentato, rispetto alla ricerca accademica

Nonostante vi sia un crescente interesse per la GDPR e il suo impatto, fare una sintesi del suo contenuto non è semplice

Si cercherà quindi di rispondere a queste due domande:

- 1 Quali sono i **principali concetti** e principi della GDPR che un data scientist dovrebbe conoscere?
- 2 Come questi principi potrebbero essere incorporati in un **processo di data science**?

Del resto, la non osservanza della GDPR può risultare in sanzioni finanziarie fino a 20 milioni di euro, o il 4% dei ricavi dell'anno precedente!

6 principi generali della GDPR (art. 5)

- 1 Legalità, correttezza, trasparenza
- 2 Limitazione dello scopo: scopo specificato, esplicito e legittimo
- 3 Minimizzazione dei dati: dati adeguati, rilevanti e limitati a ciò che è necessario (possibility driven vs. purpose driven)
- 4 Accuratezza: dati accurati e aggiornati
- 5 Limitazione del data storage
- 6 Sicurezza dei dati

Obiettivo della raccolta dei dati

- Aziende e organizzazioni raccolgono e usano dati personali per una notevole varietà di scopi, ad esempio sviluppare, aggiornare o migliorare un servizio; pianificare nuovi servizi, anche personalizzati; identificare frodi, abusi o rischi alla sicurezza
- Gli obiettivi sono stabiliti dal **data controller**
- Il data controller può essere un'azienda, un'università, o qualsiasi entità che possieda dati relativi a persone.
- La definizione degli obiettivi è l'elemento che distingue il data controller dal **data processor**
- Ovviamente un'azienda può essere sia data controller che data processor

Principio di limitazione dello scopo (purpose limitation):

- La GDPR richiede che sia esplicitato dalle aziende quali dati intendono raccogliere e per quali scopi
- Solo dopo aver ricevuto il **consenso informato** da parte degli utenti i dati possono essere raccolti e utilizzati
- Molti siti di aziende contengono sezioni '*che tipo di informazione raccogliamo?*' e '*come usiamo questa informazione?*'
- Il consenso informato richiede che venga chiarito a **cosa** servono i miei dati e **come** vengono usati ...

Necessità contrattuale e Limitazione di scopo

- Secondo il principio di **limitazione di scopo**, se una banca ottiene il consenso da parte dei propri clienti a raccogliere e processare dati personali, ad esempio per l'apertura di un conto **-necessità contrattuale-**, gli stessi dati non possono essere utilizzati per finalità di marketing
- La 'necessità contrattuale' viene citata come base legale per il processamento dei dati personali

Interesse legittimo

- L'utilizzo di dati raccolti per altre finalità, definito **repurposing**, può rappresentare un **interesse legittimo** del data controller, ad esempio nel caso di prevenzione di frodi
- Il repurposing per finalità di marketing risulta invece più ambiguo, e ricade nel caso dell'interesse legittimo quando esiste una appropriata e rilevante relazione tra l'istituzione e il cliente

Esistono alcuni obiettivi speciali per cui sono previsti delle eccezioni nella GDPR

- ricerca scientifica
- scopi statistici
- interesse pubblico
- scopi storici

- Per 'ricerca scientifica' si intende un ampio spettro di attività, ad esempio 'sviluppo tecnologico', 'dimostrazione', 'ricerca di base', 'ricerca applicata', 'ricerca finanziata pubblicamente'
- Con il termine 'scopi statistici' si intende la produzione di 'sondaggi statistici' e 'risultato statistici'
Nonostante la terminologia sia vaga, i concetti si riferiscono a risultati aggregati. Quindi le operazioni di aggregazione e calcolo di statistiche di sintesi ricadono all'interno del concetto di 'scopi statistici'
- 'Interesse pubblico e archiviazione' sono un concetto non esplicitamente definito dalla GDPR
Tuttavia esempi di 'ragioni di interesse pubblico' includono casi di 'scambio di dati tra autorità competenti in materia fiscale o di sanità pubblica'
- Gli 'scopi storici' si intende soprattutto ricerca di tipo genealogico.

Data subject

Nella GDPR la principale entità di interesse è il **data subject**, con un focus sulle variabili definite come dati personali e dati sensibili

Un data scientist deve essere a conoscenza dei seguenti concetti chiave:

- **soggetto dei dati**
- **dati personali**
- **dati sensibili**
- **dati pseudonomizzati**
- **dati statistici**

Dati personali

Si riferiscono a un ampio spettro di informazioni relative a una persona sulla sua identità fisica, genetica, mentale, economica, culturale e sociale. Da notare che anche indirizzi IP e cookies possono essere considerati come dati personali, nel senso che, se combinati con altre informazioni, possono essere utilizzati per costruire profili personali.

Dati sensibili

Si tratta di categorie di dati che rivelano l'appartenenza di un individuo a una categoria speciale

La lista delle categorie viene formulata all'art.9 (ad esempio origine etnica, orientamento sessuale, condizione di salute ...)

Il processamento dei dati sensibili è **proibito** dalla GDPR, a meno di un consenso esplicito da parte degli utenti

Dati pseudonomizzati

Un sottoinsieme dei dati personali dai quali sono stati rimossi gli identificativi individuali, in modo tale che non sia possibile riconoscere una specifica persona

La GDPR stabilisce ripetutamente che l'operazione di 'pseudonomizzazione' dovrebbe essere la base di qualsiasi operazione di raccolta dei dati

Analisi

Con il termine **analisi** la GDPR si riferisce a qualsiasi operazione di **data processing**

Ovviamente il processamento dei dati può avvenire in molti modi, più o meno sofisticati, manuali o automatizzati

Il **data processor** è la persona o l'organizzazione che concretamente realizza l'analisi dei dati

Utilità

L'utilità di un'analisi statistica è una funzione obiettivo che un data scientist usa per valutarne la performance, ad esempio numero di clicks, ROI, tasso di churn

La GDPR non discute esplicitamente metriche o misure di performance, ma specifica tre importanti aspetti collegati:

- diritto alla privacy
- principio di proporzionalità
- interesse legittimo

Diritto alla privacy

Lo scopo della GDPR è quello di **assicurare un consistente ed elevato livello di protezione delle persone**

È interessante notare la grande differenza con gli USA dove *'collecting and processing personal data is allowed unless it causes harm or is expressly limited by US law'* (opt-out model)...

mentre in UE *'processing of personal data is prohibited unless there is an explicit legal basis that allows it'* (opt-in model)

Principio di proporzionalità

Stabilisce che la protezione dei dati personali non è un diritto assoluto

Principio di bilanciamento rispetto ad altri diritti fondamentali

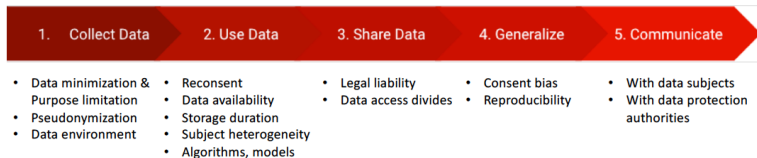
Interesse legittimo

Alla base di questo concetto, c'è l'idea che il beneficio economico derivante dall'uso dei dati sia superiore a qualsiasi danno arrecato al soggetto dei dati

→ verifica di bilanciamento

Impatto della GDPR sul data science

Esempio di Data Science workflow



Impatto della GDPR sul data science

Raccolta dei dati

- Pre-raccolta: Rispetto dei principi di minimizzazione dei dati e di limitazione dello scopo
- Dati strettamente rilevanti per il raggiungimento dell'obiettivo.
- Questo comporta che la pianificazione di progetti di Business Analytics sia sempre più orientata verso obiettivi che possono essere raggiunti **senza utilizzare dati personali**
- Il principio della minimizzazione introduce un trade-off nella fase di raccolta: campione abbastanza grande per evitare identificazione, ma senza raccogliere più dati del necessario!

Raccolta dei dati

- Post-raccolta: un aspetto chiave della GDPR sta nella pratica di pseudonomizzazione.
Si tratta di un metodo per ridurre la possibilità che un certo dato venga attribuito a una persona senza l'uso di ulteriore informazione (che deve essere tenuta separata)
- Da notare che è pratica differente dall'anonimizzazione
- Conseguenze della pseudonomizzazione: rimozione totale di identificativi individuali rende difficile la previsione a livello individuale (sistemi di raccomandazione e previsioni personalizzate)

Impatto della GDPR sul data science

Utilizzo dei dati

Un problema riguarda l'utilizzo dei dati raccolti prima della GDPR

Molti siti e piattaforme online hanno chiesto agli utenti un nuovo consenso sull'utilizzo di quei dati

Aggiornamento di policy da parte di alcune compagnie, ad esempio Kaggle:
'possiamo usare dati aggregati e anonimizzati che abbiamo derivato da vostra informazione personale prima della sua cancellazione, ma non in modo da rendere possibile il riconoscimento di una persona'

Impatto della GDPR sul data science

Utilizzo dei dati

Inoltre:

- Data storage limitation
- Data accuracy

Dopo un certo periodo di tempo i dati devono essere cancellati per rispettare questi due principi

Vale poi il principio del *diritto di rettifica* da parte del soggetto dei dati

Impatto della GDPR sul data science

Utilizzo dei dati

Uno degli aspetti più importanti relativamente all'uso dei dati riguarda la scelta di modelli e algoritmi

Il principio di legalità, correttezza e trasparenza favorisce modelli più facilmente interpretabili

Impatto della GDPR sul data science

Condivisione dei dati e generalizzazione

Si sottolineano alcuni aspetti critici:

- Dal momento che secondo la GDPR, un'azienda può essere ritenuta responsabile dei danni arrecati da un uso illegale dei dati personali, c'è molta riluttanza alla condivisione
- Problema del **consent bias**

Impatto della GDPR sul data science

Comunicazione

L'aspetto della comunicazione è molto importante e riguarda:

- Soggetti dei dati
- Autorità preposte alla protezione dei dati (compliance)

Communication skills sempre più importanti nel futuro

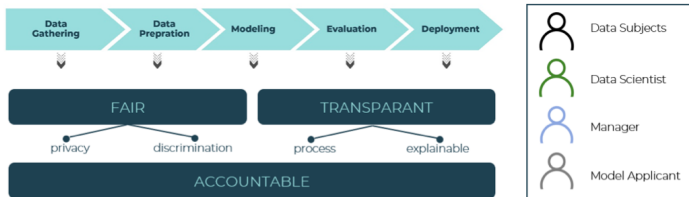
- La mancata comprensione di come le decisioni vengono prese ha importanti **risvolti etici** ...
- Perché è importante considerare una dimensione etica nel data science?
- **Etica**: obiettivo esistenziale di per sé ... ma non solo
- Rischio di perdita di **reputazione** per aziende grandi e piccole

- **Valore del pensiero etico** → miglioramenti nella qualità dei dati e dei modelli, con previsioni più accurate, migliore accettazione dei modelli e migliore immagine del brand
- Generazione Z (nati tra 1995 e 2010): particolarmente attenti a giustizia sociale e etica

- Ruolo dell'individuo
- Fase del processo di analisi
- Criterio di valutazione

Data Ethics

Data Science + Etica: 3 dimensioni



Un processo di data science etico dovrebbe essere

- **Corretto** → non discrimina gruppi sensibili, non trae vantaggio da possibili violazioni della sfera privata
- **Trasparente** → criterio che ha implicazioni su correttezza e responsabilità, dipende dal soggetto considerato (manager, data scientist, data subject ...)
- **Responsabile** → dalla teoria alla pratica ... i buoni propositi non bastano

Data Science Etica: checklist

- Avere consapevolezza di pratiche istituzionali e procedure
- Avere consapevolezza di leggi e regolamenti nazionali
- Mantenere traccia di come i dati vengono modificati e manipolati
- Conoscere la provenienza dei dati
- Conservare i dati in modo sicuro
- Determinare per quanto a lungo i dati devono essere conservati
- Garantire l'uso dei metodi e modelli appropriati
- Essere chiari sui benefici delle analisi e chi dovrebbe riceverli

- Greene, T., Shmueli, G., Ray, S., & Fell, J. (2019). Adjusting to the GDPR: The impact on data scientists and behavioral researchers. *Big data*, 7(3), 140-162.
- Martens, D. (2022). *Data Science Ethics: Concepts, Techniques, and Cautionary Tales*. Oxford University Press.
- Hand, D. J. (2018). Aspects of data ethics in a changing world: Where are we now?. *Big data*, 6(3), 176-190.