19.06.2019

Irene Gittardi

Il database di casi giudiziari del Center for Science and Law

APRI ALLEGATO

#big_data #database #David_Eagleman #recidiva #reo #rischio #USA

Abstract. *Segnaliamo qui un ambizioso progetto di ricerca, sviluppato dal* team *di ricercatori del Center for Science and Law (SciLaw), diretto dal Prof. David Eagleman, finalizzato a raccogliere milioni di casi giudiziari celebrati dinanzi alle diverse corti nazionali statunitensi all'interno di un unico* database, *denominato* Criminal Record Database *– CRD.*

L'intento alla base della creazione del CRD – che oggi conta più di 28 milioni di casi, in parte già elaborati e in parte attualmente in fase di elaborazione, ed è liberamente consultabile online *–, come descritto dai suoi ideatori*[1]*, è quello di orientare la politica sociale secondo un modello* evidence-based, *che consenta di ridurre i tassi di detenzione e offrire nuove strade per risolvere il problema della criminalità.*

In passato, infatti, la gestione del fenomeno criminale è stata spesso guidata dall'intuizione e da considerazioni politiche piuttosto che dall'analisi oggettiva dei dati. In ragione dei costi sociali e della stigmatizzazione connessi ad alcuni tipi di reati, le decisioni assunte dai giudici e in ambito di policy *sono spesso il frutto di una risposta emotiva, e non tengono conto delle caratteristiche della persona e delle singole circostanze del caso. Ecco che, allora, grazie all'analisi di milioni di casi giudiziari raccolti nei diversi Stati, diventa possibile dare risposta ad alcune domande cruciali (quali scelte politiche negli ultimi decenni si sono rivelate efficaci nel ridurre la criminalità? In relazione a quali tipi di reato? Esistono particolari categorie di reato che inducono l'autore a commettere nuovi crimini in futuro? In quali casi la condanna risulta efficace nella prevenzione di condotte recidivanti? Qual è il legame tra età giovanile e criminalità?) e migliorare così la complessiva gestione della giustizia penale statunitense.*

SOMMARIO: 1. Introduzione e premessa. – 2. I metodi. – 2.1. L'acquisizione dei dati. – 2.2. L'elaborazione dei dati. 2.3. – La progettazione del sistema di catalogazione. – 2.4. Le decisioni. – 2.5. Voci aggiuntive. – 3. Discussione.

Per leggere l'articolo pubblicato nel 2015 sulla rivista The Journal of Science and Law*, oggetto della presente Riflessione, clicca su "apri allegato".*

1. Introduzione e premessa.

Negli Stati Uniti, il principale strumento che consente di condurre un'analisi, sotto il profilo quantitativo, del fenomeno criminale, è rappresentato da un *database* sviluppato dall'FBI, denominato *Uniform Crime Reports* – UCR, all'interno del quale le singole **Law Enforcement Agencies** di tutto il Paese inseriscono i dati di base, che vengono poi rielaborati direttamente dall'FBI.

Sebbene i *report* pubblicati dall'FBI su base annua costituiscano una delle raccolte più complete dei crimini contro la persona e contro il patrimonio perpetrati sul territorio statunitense, l'UCR presenta comunque alcuni **limiti rilevanti**, specie con riguardo ai possibili percorsi di ricerca realizzabile a partire da quei dati.

In particolare, come sottolineano i ricercatori del SciLaw di Houston diretto dal Prof. David Eagleman, i principali problemi sono i seguenti[2]:

1. non essendo disponibili identificatori univoci degli autori dei reati, è **impossibile individuare i soggetti recidivi**;
2. manca la descrizione di dettaglio dei singoli reati e dei relativi **epiloghi processuali**: il sistema presenta infatti solo un quadro di sintesi, che tralascia una serie di informazioni rilevanti relative al singolo caso quali, ad esempio, il numero delle imputazioni, gli esiti del procedimento, il tipo di pena inflitta, ecc.;
3. nei *report* finali dell'FBI viene inoltre riportata solo l'imputazione ritenuta più grave tra quelle contestate nei diversi casi;
4. infine, l'inserimento dei dati da parte di una pluralità di agenzie locali comporta un alto livello di **variabilità** e di **imprecisione** nel sistema poiché, negli Stati Uniti, ciascuno Stato dispone di un'ampia autonomia nella definizione dei singoli reati e nell'articolazione del sistema processuale.

Ecco che, per ovviare a tali limiti, i ricercatori del SciLaw hanno deciso di sviluppare un proprio *database* che raccoglie decine di milioni di singoli casi giudiziari provenienti da più corti statunitensi. Attualmente, l'archivio contiene **oltre 28 milioni di casi, celebrati tra il 1971 e il 2013** nell'ambito di cinque diverse zone del Paese (la contea di Harris, in Texas, quella di Miami-Dade, in Florida, la citta di New York e gli Stati dell'Alabama e del New Mexico – in quest'ultimo caso, l'analisi dei procedimenti, attualmente in fase di elaborazione, è protratta fino all'anno 2018).

> **Il CRD ha come scopo quello di trovare la risposta a un interrogativo di fondo, relativo ai meccanismi di funzionamento della condotta umana: cosa influenza i modi in cui gli autori dei reati compiono le proprie scelte?**

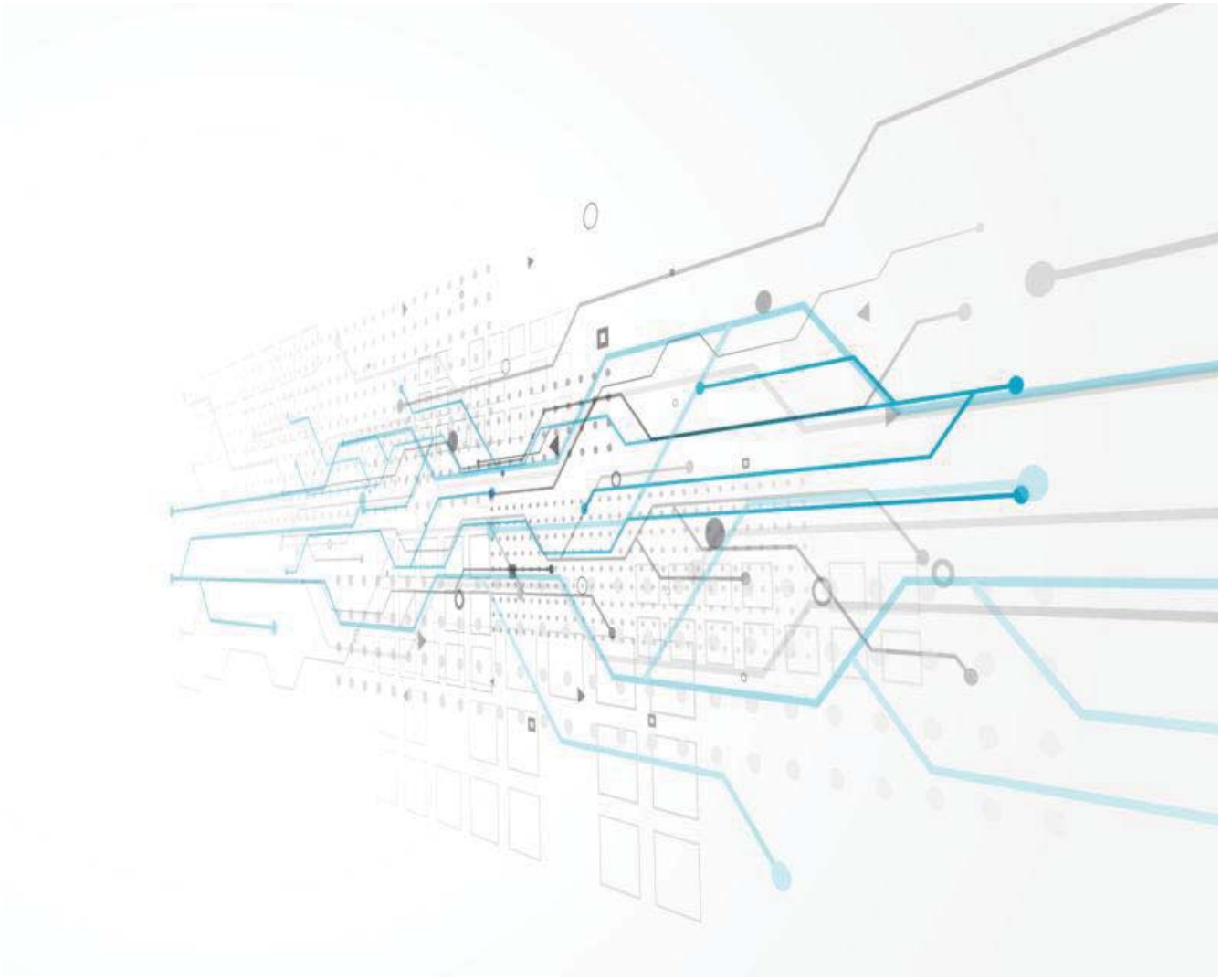
I **vantaggi** del *database*, secondo il team coordinato dal Prof. Eagleman, sono molteplici:

1. la presenza di identificatori individuali, associati a ciascun autore di reato, che rendono possibile (pur garantendo l'anonimato)[3] **analizzare il fenomeno della recidiva,** quantomeno con riferimento ai delitti commessi nell'ambito della medesima giurisdizione;
2. il fatto che, per ciascun caso, sono disponibili i dati relativi a **tutte le imputazioni**, il che permette di effettuare **analisi accurate per tutte le fattispecie** di reato previste dai diversi ordinamenti;
3. l'inserimento di un **numero di variabili** – relative all'autore del reato e alla decisione finale – **più ampio** di quello contemplato dai sistemi dell'FBI, così da consentire una valutazione più dettagliata, specie con riguardo alle variazioni nel *quantum* di pena;
4. la natura dal CRD come sistema in **continuo sviluppo**, posto che tra gli utenti del CRD non figurano solo i giudici e i pubblici ministeri, ma anche i membri della **comunità scientifica**, che metteranno a disposizione le proprie competenze per migliorare e integrare costantemente il *database*.

Il CRD ha come scopo quello di trovare la risposta a un interrogativo di fondo, relativo ai meccanismi di funzionamento della condotta umana: **cosa influenza i modi in cui gli autori dei reati compiono le proprie scelte?**

Consentendo un'analisi dei **fattori esterni** – come le politiche legislative o l'inclusione sociale – suscettibili di influire sulla decisione del singolo di commettere un reato, i ricercatori mirano a rendere possibile l'**effettuazione di ricerche criminologiche** in grado di impattare significativamente non solo sul mondo del diritto, ma anche sulla società nel suo complesso.

La finalità ultima del CRD, infatti, è quella di promuovere e incoraggiare politiche sociali che abbiano basi scientifiche, garantendo la possibilità di compiere un'attività di analisi fondata su dati e gratuita.



Credits to Freepik.com

2. I metodi.

2.1. L'acquisizione dei dati.

Per ottenere i dati poi confluiti nel *database*, gli sviluppatori del CRD si sono innanzitutto rivolti ai sistemi di gestione delle informazioni giudiziarie della città di New York, delle contee di Harris e di Miami-Dade e dello Stato del New Mexico per ottenere la documentazione relativa ai casi processuali ivi conservati[4].

2.2. L'elaborazione dei dati.

Il primo passo da compiere in sede di elaborazione dei dati è stato quello della **standardizzazione** (delle lettere maiuscole, dello *spelling* delle diverse voci di testo, all'eliminazione di errori di digitazione e della duplicazione delle informazioni); in particolare, dal momento che, negli Stati Uniti, le diverse giurisdizioni usano migliaia di e **denominazioni diverse** (talvolta discordanti) per indicare i medesimi delitti, i ricercatori del *team* hanno lavorato fianco a fianco con i programmatori per omogeneizzare la terminologia di volta in volta impiegata, in modo da conferire chiarezza e uniformità all'intero sistema.

2.3. La progettazione del sistema di catalogazione.

Per risolvere il problema causato dai **diversi titoli di reato** utilizzati dalle corti, tre avvocati hanno inoltre creato e implementato un particolare schema di classificazione, operando, anche in questo caso, a stretto contatto con i programmatori.

Così, i ricercatori del SciLaw hanno ideato un **sistema di classificazione** che fosse al contempo abbastanza ampio da superare le piccole divergenze nella descrizione degli elementi di un reato da parte delle varie giurisdizioni e abbastanza dettagliato da consentire una capillare comprensione di come i diversi tipi di reato si siano evoluti nel tempo.

In particolare, sono stati ideati due livelli di classificazione delle fattispecie di reato: uno "ampio" (*Broad*) e uno "dettagliato" (*Detailed*), contenenti rispettivamente **32 classificazioni** (dal furto, all'omicidio, ai crimini contro la pubblica amministrazione, ecc.) e in **152 classificazioni** (a grana più fine, che vanno dalla recidiva per guida in stato di ebbrezza[5], alla truffa ai danni dei servizi sociali, al possesso di un ridotto quantitativo di marijuana[6], ecc.).

Tutti i delitti oggetto di contestazione nei casi analizzati sono stati inseriti in entrambi gli schemi di classificazione.

A questo punto, si trattava di convertire i dati grezzi in un sistema uniforme: in questa fase, i ricercatori hanno dovuto fare i conti con tre tipi di **ostacoli**:

1. le difficoltà derivanti, ad esempio, dall'utilizzo di diverse forme di *spelling* o di abbreviazione per lo stesso titolo reato, o dall'utilizzo di riferimenti alle singole disposizioni legislative: la soluzione è consistita nell'identificare manualmente, mediante il supporto della squadra di avvocati, i singoli dati, valutarli e infine assegnarli alla categoria corretta;
2. il problema connesso all'estrema variabilità, nel tempo e nello spazio, dei titoli di reato (*labels*). Un esempio è quello dei reati sessuali: la condotta di "stupro" in alcuni Stati è definito "*rape*", in altri "*sexual assault*". Inoltre, a fronte della medesima denominazione, la fattispecie può richiedere la presenza di elementi oggettivi diversi
3. inoltre, occorre considerare che anche le definizioni degli elementi extrapenali delle varie fattispecie tendono a modificarsi nel corso del tempo. Così, ad esempio, la disciplina relativa all'individuazione e catalogazione delle sostanze stupefacenti ha imposto agli sviluppatori un complesso lavoro di allineamento delle "vecchie" fattispecie con i loro corrispettivi attuali per consentire la realizzazione di analisi intertemporali.

2.4. Le decisioni.

Sono poi stati armonizzati anche i diversi **epiloghi dei singoli casi**, poiché le giurisdizioni esaminate utilizzano ben **250 diverse denominazioni** per descrivere le possibili conclusioni di un procedimento penale.

Anche in questo caso, il team ha valutato manualmente ogni singola decisione, assegnandola a una delle seguenti sei categorie: *Dismissed* (letteralmente "rigetto")[7], *Acquittal* (assoluzione), *Guilty* (condanna), *Guilty by plea* (condanna in seguito ad ammissione di colpevolezza), *Conditional dismissal* (sospensione condizionale), *No action* (archiviazione).

2.5. Voci aggiuntive.

Nonostante il CRD sia composto unicamente da dati pubblici, si è cercato di minimizzare le potenziali lesioni della *privacy* attraverso un processo di de-identificazione dei protagonisti delle singole vicende giudiziarie. In particolare, il processo di **anonimizzazione** prescelto genera, secondo Eagleman e colleghi, un risultato finale che rende quasi impossibile individuare il dato originale alla sua base.

Al pubblico sono messi a disposizione esclusivamente i dati anonimi, privi di identificatori individuali. Internamente, invece, viene conservato il file che collega i dati anonimi al nome del soggetto e ai suoi identificatori.

I dati raccolti grazie al materiale fornito dalle singole corti sono stati inoltre arricchiti dai ricercatori in sede di elaborazione, così consentendo di individuare una serie di informazioni aggiuntive quali l'etnia di appartenenza, o il genere, dell'autore di reato[8].

3. Discussione.

Il CRD – con i suoi oltre **28 milioni di casi** giudiziari sviluppati nell'arco di 36 anni, **tra il 1977 e il 2013** – viene definito come **il** *database* **più grande e completo**, su base anonima e ad accesso gratuito.

Credits to Freepik.com

Questa risorsa apre le porte a una **varietà di temi di ricerca** – rendendo possibile, ad esempio, un'analisi e il confronto delle conseguenze per l'imputato che patteggia rispetto a colui che non lo fa, o della diversa influenza delle pene detentive e pecuniarie sui tassi di recidiva.

Inoltre, come sottolineano i suoi sviluppatori[9], il CRD permette di operare un **confronto inter-giurisdizionale** delle procedure di arresto: così, ad esempio, è stato valutato l'impatto, sui casi di arresto, della necessità (non prevista in tutti gli Stati) di ottenere un'autorizzazione preventiva da parte del pubblico ministero.

Come già anticipato, il sistema implementato dal team del SciLaw consente altresì **una più profonda comprensione della recidiva**, e ciò permette di **sviluppare politiche criminali** basate su prove e finalizzate a **prevenire e a controllare il crimine**. Ciò potrà portare benefici alla società in generale, consentendo al legislatore di ancorare le proprie decisioni in materia di *law enforcement* sulla valutazione delle reali tendenze criminali.

> **Il CRD – con i suoi oltre 28 milioni di casi giudiziari sviluppati nell'arco di 36 anni, tra il 1977 e il 2013 – viene definito come il database più grande e completo, su base anonima e ad accesso gratuito**

Il CRD però presenta anche alcuni **limiti**, di qui i suoi stessi creatori danno ampiamente conto (indicando anche, ove possibile, le relative soluzioni)[10].

Più in dettaglio:

1. il *database* **non contiene** i dati relativi al **diritto penale minorile**, perché non ottenibili con la richiesta prevista dal *Freedom of Information act*;
2. il *database* **non comprende i casi secretati** (*sealed*) o cancellati (*expunged*), che vengono di norma rimossi dai *database* locali su cui il CRD si fonda. È probabile che questa circostanza falsi le proporzioni di alcuni tipi di reati (ad esempio, quelli in materia di circolazione stradale);
3. il CRD **non contiene i dati relativi alle vittime**, impedendo così l'analisi, per esempio, di come l'appartenenza etnica o l'età della persona offesa possano aver influito sulla condanna;
4. il CRD **non contiene i dati relativi all'esecuzione della pena** (*correction records*), dal momento che essi non sono pubblici, secondo la legislazione di molti Stati. Di conseguenza, ad esempio, mentre è possibile conoscere il dato relativo alla condanna inflitta al reo all'esito del processo, non è dato sapere, attraverso il CRD, per quanto tempo questi sia effettivamente rimasto in carcere[11];
5. il sistema soffre poi di una serie di **limitazioni dovute all'approvvigionamento dei dati** e collegate alle scelte originarie delle singole agenzie locali in materia di selezione delle informazioni rilevanti o in tema di *privacy* (ad esempio, la severità della disciplina in materia di riservatezza fa sì che il CRD non contenga alcun dato con riferimento al Northwest e al Midwest);
6. l'analisi della recidiva che il CRD permette di condurre è inoltre limitata alle sole iscrizioni avvenute nella stessa giurisdizione (così evidentemente falsando l'individuazione dei **reali tassi di recidiva)**;

il CRD, infine, contiene dati **a partire dal momento dell'arresto**, mentre non prende in considerazione i precedenti stadi nel processo di *law enforcement* (contrariamente allo UCR, il *database* dell'FBI, comprensivo anche di tutti i *reports* di *law enforcement*).

---

[1] Si vedano, in particolare, P. A. Ormachea, G. Haarsma, S. Davenport, D.M. Eagleman, *A new criminal records database for large-scale analysis of policy and behavior*, in *Journal of Science and Law*, 1(1), 2015, pp. 1 ss., unitamente alla sezione del sito SciLaw dedicata al progetto, dai quali sono tratte le principali informazioni riportate nella presente Riflessione.

[2] P.A. Ormachea et al., *A new criminal records database*, cit., p. 1.

[3] In particolare, il *database* include la presenza di c.d. *anonymized identifiers* ("identificatori resi anonimi" – cfr. par. 2.5. dell'articolo in commento) degli autori dei reati, specificamente finalizzati a consentire l'analisi del fenomeno della recidiva pur garantendo l'anonimato dell'autore.

[4] Trattandosi di dati pubblici, le informazioni sono state ottenute con la richiesta regolata dal Freedom of Information Act.

[5] «Second time DWI» (driving while intoxicated).

[6] Nel *Supplementary Material* allegato all'articolo in commento sono riportati i dettagli di entrambi gli schemi di classificazione.

[7] L'espressione si riferisce ai casi in cui il giudice, prima dell'inizio del processo o comunque prima della sentenza, chiude il caso, per così dire, preliminarmente – ad esempio, per violazione del diritto a uno «*speedy trial*» o per mancanza di prove – senza però giungere a una pronuncia classificabile come *Acquittal*.

[8] Ad esempio, i dati grezzi di alcune giurisdizioni identificavano solo sommariamente l'etnia del reo, attraverso la dicitura "bianco" o "nero". Poiché è verosimile che molti ispanici siano stati impropriamente classificati come appartenenti a una di queste razze, è stato seguito il metodo dello US Census Bureau, fondato sull'esame del cognome,

per valutare l'appartenenza all'etnia ispanica. L'articolo segnala che si tratta di un metodo non perfetto ma con pochi falsi positivi. L'algoritmo elaborato dal CRD consente, peraltro, di inserire una nuova voce accanto a quella originaria: ciò significa che nel CRD, con riferimento alla razza, c'è sia la variabile originaria sia una aggiuntiva e inferenziale. Allo stesso modo, sono state utilizzate le tavole prese dallo US Census anche per individuare il genere dell'autore del reato. Qui, tuttavia, si è aggiunto il genere solo dove era originariamente mancante.

[9] P.A. Ormachea et al., *A new criminal records database*, cit., p. 6.

[10] *Ibidem.*

[11] Gli autori stanno cercando di ovviare a questo problema inserendo nel CRD dei *corrections record* ottenuti in modo "indipendente".