Mapping e Big Data
Lezione 7

Mario Verdicchio
Università degli Studi di Bergamo
Anno Accademico 2022-2023

Data-driven geography

Harvey J. Miller & Michael F. Goodchild
The Ohio State University, Columbus, USA
GeoJournal 80 (2015)

Sommario 1/3

Il contesto della **ricerca** in geografia si è spostato da un ambiente **scarso** di dati a un ambiente **ricco** di **dati**, in cui i cambiamenti fondamentali non sono solo il **volume** di dati, ma la **varietà** e la **velocità** con cui possiamo acquisire dati **georeferenziati**; tendenze spesso associate al concetto di **Big Data**.

Una **geografia basata sui dati** potrebbe emergere in risposta alla ricchezza di dati georeferenziati che fluiscono da **sensori** e **persone** nell'ambiente. Sebbene ciò possa sembrare **rivoluzionario**, in realtà potrebbe essere meglio descritto come **evolutivo**.

Sommario 2/3

Alcuni dei **problemi** sollevati dalla geografia basata sui dati sono stati in effetti problemi **di lunga data** nella ricerca geografica, vale a dire **grandi volumi** di dati, gestione di popolazioni e dati **disordinati** e **tensioni** tra conoscenza idiografica (casi unici) e nomotetica (leggi generali).

La convinzione che il contesto **spaziale** sia importante è un tema importante nel pensiero geografico e una delle principali motivazioni alla base di approcci come la **time geography**, **statistiche spaziali disaggregate** e **GIScience**.

Sommario 3/3

È possibile utilizzare i **Big Data** per supportare sia la **scoperta** di conoscenza geografica sia la **modellazione** spaziale. Tuttavia, ci sono **sfide**, come ad esempio come **formalizzare** le conoscenze geografiche per **pulire** i dati e **ignorare** modelli spuri e come **costruire modelli** basati sui dati che siano sia **veri** che **comprensibili**.

Domande 1/2

Nonostante l'entusiasmo per i Big Data e i metodi basati sui dati, il **ruolo** che possono svolgere nella ricerca accademica, e in particolare la **ricerca** in geografia, potrebbe non essere immediatamente evidente.

La **teoria** e le **spiegazioni** sono oramai **obsolete** se possiamo misurare e descrivere così tanto, così rapidamente?

La **velocità** dei dati conta davvero nella ricerca, con le sue tradizioni di **attenta riflessione**?

Gli ovvi problemi associati alla varietà - mancanza di **controllo** di **qualità**, mancanza di un **rigoroso** disegno di **campionamento** - possono essere superati?

Domande 2/2

Possiamo fare **generalizzazioni** valide da una fortuita raccolta di dati in corso (anziché attentamente progettata e strumentata)?

Big Data e i metodi basati sui dati possono portare a **scoperte** significative nella ricerca geografica?

O la comunità di ricerca continuerà a fare affidamento su ciò che ai fini di questo documento chiameremo **Scarce Data**: i prodotti di programmi statistici del **settore pubblico** che a lungo sono stati il principale contributo alla ricerca in geografia umana quantitativa?


Metafora 1:
bere da un idrante

Bere da un idrante

Una metafora che rappresenta le potenziali difficoltà legate all'uso dei Big Data: ne arrivano troppi e troppo velocemente perché se ne possa fare un buon uso.

Non è una novità

Vale la pena riconoscere immediatamente, tuttavia, che questa metafora ha una storia relativamente lunga in geografia e che la disciplina non è affatto nuova all'abbondanza di dati voluminosi.

Il programma di telerilevamento satellitare basato su Landsat iniziò nei primi anni '70 acquisendo dati a velocità che erano ben al di sopra delle capacità analitiche dei sistemi computazionali dell'epoca.

I successivi miglioramenti nella risoluzione dei sensori e la proliferazione di satelliti militari e civili hanno fatto sì che quattro decenni più tardi i volumi di dati continuino a sfidare anche i più potenti sistemi computazionali.

Metafora 2: il cigno nero

Il cigno nero…

…è una metafora resa famosa da Nassim Taleb (“Il cigno nero. Come l'improbabile governa la nostra vita”, Il Saggiatore, 2014) per indicare eventi che sono improbabili ma significativi, opposti a eventi probabili e poco significativi.

Con i Big Data…

…non serve distinguere i cigni neri da quelli bianchi prima di raccogliere i dati: si raccolgono e si misurano tutti i cigni, e poi dopo si possono classificare quelli neri.

A volte un cigno nero può essere dato una combinazione inusuale di cigni bianchi.

Sembrerebbe tutto facile e migliorato dai Big Data, ma…

Data-driven Geography

Le sfide

Sfida 1: Popolazioni, non campioni

Quando l'analisi veniva eseguita in gran parte a mano piuttosto che dalle macchine, non era pratico gestire grandi volumi di dati.

Invece, i ricercatori usavano metodi per la raccolta di campioni rappresentativi e per far generalizzazioni sulla popolazione da cui tali campioni sono stati estratti.

Il **campionamento** casuale era quindi una strategia per affrontare il sovraccarico di informazioni. In casi come il censimento della popolazione era anche un modo per limitare i costi.

Il campionamento casuale funziona bene, ma è fragile: funziona solo fino a quando il campionamento è rappresentativo.

Un tasso di campionamento di uno su sei (il tasso precedentemente utilizzato dal Bureau of Census degli Stati Uniti per la versione lunga del censimento) può essere adeguato per alcuni scopi, ma diventa sempre più problematico quando l'analisi si concentra su sottocategorie relativamente rare.

Il campionamento casuale richiede anche un processo di enumerazione e poi di selezione nella popolazione (un frame di campionamento), il che è problematico se l'enumerazione è incompleta.

I campioni presentano inoltre una mancanza di estensibilità per usi secondari.

Poiché la casualità è così critica, è necessario pianificare attentamente il campionamento e potrebbe essere difficile riesaminare i dati per scopi diversi da quelli per i quali sono stati raccolti.

Al contrario, molte delle nuove fonti di dati sono costituite da **popolazioni**, non da campioni: la facilità di raccolta, archiviazione ed elaborazione dei dati digitali significa che invece di trattare con una piccola rappresentazione della popolazione possiamo lavorare con l'intera popolazione e quindi sfuggire a uno dei vincoli del passato.

Ma…

…un problema con le popolazioni è che sono spesso **auto-selezionate** piuttosto che campionate: per esempio,

- tutte le persone che si sono iscritte a Facebook
- tutte le persone che hanno uno smartphone
- tutte le auto che hanno viaggiato nella città di Londra tra le 8:00 e le 11:00 del 2 settembre 2013

I tweet geolocalizzati sono un'interessante fonte di informazioni sulle tendenze attuali, ma solo una piccola parte dei tweet viene geolocalizzata accuratamente utilizzando il GPS.

Poiché non conosciamo le caratteristiche demografiche di nessuno di questi gruppi, è impossibile generalizzare da essi a popolazioni più ampie da cui potrebbero essere state tratte.

I geografi hanno dovuto affrontare a lungo i problemi associati ai campioni e alle loro popolazioni madri.

Si consideri, ad esempio, un'analisi della relazione tra persone di età superiore ai 65 anni e persone registrate come repubblicane, il caso studiato da Openshaw e Taylor nel 1979.

Le 99 contee dell'Iowa (la loro fonte di dati) sono **tutte** le contee esistenti nello Iowa. Non sono quindi un campione casuale di contee dello Iowa, né un campione rappresentativo di contee degli Stati Uniti, quindi i metodi di statistiche inferenziali che presuppongono campionamenti casuali e indipendenti non sono applicabili.

Altro esempio: nel telerilevamento è comune analizzare **tutti** i pixel in una determinata scena. Ancora una volta, questi non sono un campione casuale di popolazione più ampia.

In questi esempi particolari possiamo essere certi che l'**intera** popolazione di interesse è inclusa: siamo interessati a tutta la copertura del suolo in una scena, o a tutte le persone con più di 65 anni e repubblicane nello Iowa.

Questo spesso non è vero con molte nuove fonti di dati: non abbiamo la certezza che l'intera popolazione di interesse sia inclusa.

Una sfida è come identificare le nicchie a cui i dati della popolazione monitorata possono essere applicati con ragionevole generalità.

Ciò **inverte** il classico problema di campionamento in cui identifichiamo una domanda e raccogliamo dati per rispondere a quella domanda.

Invece, raccogliamo i dati e determiniamo a quali domande possiamo rispondere.

Non più:

- È vero che x? Campioniamo la popolazione per verificarlo.

Bensì:

- Abbiamo questi dati. A che domanda rispondono in maniera generalizzabile? Quale domanda li rende un campione rappresentativo della popolazione?

Un altro problema riguarda ciò che le persone fanno quando offrono volontariamente informazioni geografiche e di altro tipo.

I social media come Facebook possono avere alti tassi di penetrazione rispetto alla popolazione, ma non hanno necessariamente alti tassi di penetrazione nella vita delle persone.

Il check-in a un concerto o una lezione di orchestra fornisce un'immagine nobile che una persona vorrebbe promuovere, mentre il check-in in un negozio di alcolici alle 10 del mattino è un'immagine che una persona potrebbe essere meno desiderosa di condividere.

Metafora 3: il teatro

Nel classico testo sociologico “The Presentation of Self in Everyday Life”, Erving Goffman (1959) usa il teatro come metafora e distingue tra comportamenti da palcoscenico e da dietro le quinte, con comportamenti da palcoscenico coerenti con il ruolo che le persone desiderano svolgere nella vita pubblica e comportamenti da dietro le quinte come azioni private che la gente desidera mantenere private.

Da queste osservazioni seguono diverse domande geografiche.

Qual è la geografia del palcoscenico vs. dietro le quinte in una città o regione?

Questa distribuzione varia in base all'età, al sesso, allo stato socio-economico o alla cultura?

Che cosa implicano questi fattori per ciò che possiamo sapere sul comportamento spaziale umano?

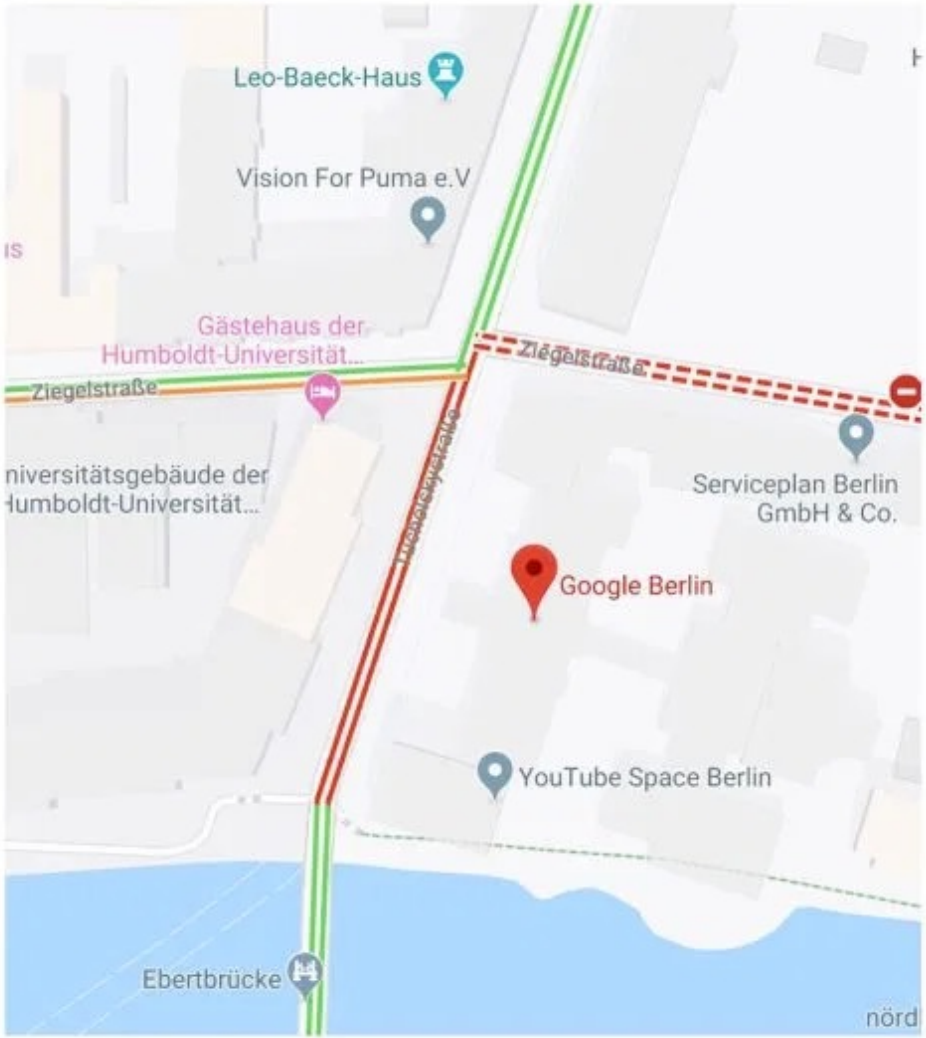
Oltre al volontariato selettivo di informazioni sulla propria vita, potrebbero esserci anche pregiudizi nella selezione delle informazioni che le persone si offrono volontariamente sugli ambienti.

Open Street Map (OSM) è spesso identificato come un progetto di mappatura crowdsourcing di successo: molte città del mondo sono state mappate da persone su base volontaria con un notevole grado di accuratezza.

Tuttavia, alcune regioni vengono mappate più rapidamente di altre, come località turistiche, aree ricreative e quartieri ricchi, mentre le località di minore interesse per coloro che partecipano all'OSM (come i quartieri più poveri) ricevono meno attenzione.

Mentre i pregiudizi esistono nelle mappe amministrative ufficiali (ad esempio, i governi dei paesi in via di sviluppo spesso non mappano insediamenti informali come le favelas), i pregiudizi nelle mappe di crowdsourcing sono probabilmente più sottili.

Allo stesso modo, l'ascesa dell'hacking civico in cui i cittadini generano dati, mappe e strumenti per risolvere i problemi sociali tende a concentrarsi sui problemi che i cittadini con laptop, connessioni Internet veloci, competenze tecniche e tempo disponibile considerano importanti.



Simon Weckert, Berlino, febbraio 2020

Sfida 2: Disordine, sporcizia

Le nuove fonti di dati sono spesso **disordinate**, costituite da dati non strutturati, raccolti senza controllo di qualità e spesso non accompagnati da documentazione né metadati. Esistono almeno due modi per gestire tale confusione.

Da un lato, possiamo limitare il nostro uso dei dati ad attività che non tentano di generalizzare o fare ipotesi sulla qualità. I dati disordinati possono essere utili in quelle che si potrebbero definire le aree più **soft** della scienza: esplorazione iniziale delle aree di studio o generazione di ipotesi.

L'etnografia, la ricerca **qualitativa** e le indagini della Grounded Theory spesso si concentrano sull'uso di interviste, testi e altre fonti per rivelare ciò che altrimenti non era noto o riconosciuto, e in tali contesti i tipi di rigorosi campionamenti e documentazione associati agli Scarse Data sono in gran parte inutili.

D'altra parte, possiamo tentare di **pulire** e **verificare** i dati, rimuovendo il più possibile il disordine, per l'uso nella costruzione di conoscenze scientifiche tradizionali.

Goodchild e Li discutono di questo approccio nel contesto delle informazioni geografiche di crowdsourcing («Assuring the quality of volunteered geographic information", *Spatial Statistics 1*, 2012). Notano che la produzione tradizionale di informazioni geografiche si è basata su più fonti e sull'esperienza di cartografi e scienziati per assemblare un quadro integrato del paesaggio.

Ad esempio, le informazioni sul terreno possono essere compilate da fotogrammetria, misurazioni del punto di elevazione e fonti storiche; come risultato di questo processo di sintesi, il risultato pubblicato potrebbe benissimo essere più accurato di una qualsiasi delle fonti originali.

Goodchild e Li sostengono che quel tradizionale processo di **sintesi**, che è in gran parte nascosto al pubblico e non appare nel risultato finale, diventerà esplicito e di fondamentale importanza nel nuovo mondo dei Big Data.

Identificano tre strategie per la pulizia e la verifica dei dati disordinati:

1. la soluzione **crowd**
2. la soluzione **sociale**
3. la soluzione della **conoscenza**

La soluzione crowd 1/2

La soluzione crowd si basa sulla “legge di Linus”, chiamata così in onore dello sviluppatore di Linux, Linus Torvalds:

“Con abbastanza occhi, tutti i bug sono in superficie”.

In altre parole, **più persone** possono accedere e rivedere il codice, maggiore è l'**accuratezza** del prodotto finale.

I fatti geografici che possono essere sintetizzati da più rapporti originali sono probabilmente più precisi dei singoli rapporti.

Questa è la strategia utilizzata da Wikipedia e dai suoi analoghi: i contributi aperti e l'editing aperto sono evidentemente in grado di produrre risultati ragionevolmente accurati se assistiti da varie procedure di editing automatizzato.

La soluzione crowd 2/2

Nel caso della geografia, tuttavia, sorgono diversi problemi che limitano il successo della soluzione crowd.

I rapporti sugli eventi in alcune località possono essere difficili da confrontare se i mezzi utilizzati per specificare la posizione (nomi dei luoghi, indirizzo, GPS) sono **incerti** e se i mezzi utilizzati per descrivere l'evento sono **ambigui**.

I fatti geografici possono essere oscuri, come i nomi delle montagne in parti remote del mondo, e il pubblico può quindi avere **scarso interesse** o capacità di modificare gli errori.

La soluzione sociale

Goodchild e Li descrivono la soluzione sociale come implementazione di una struttura gerarchica di **moderatori** e gatekeeper volontari.

Gli individui sono nominati ai ruoli nella gerarchia in base alla loro esperienza e all'accuratezza dei loro contributi.

I fatti riportati su base volontaria che sembrano discutibili o contestabili sono **riferiti** alla gerarchia, per essere accettati, interrogati o respinti a seconda dei casi.

Schemi come questo sono stati implementati da molti progetti, tra cui OSM e Wikipedia.

Il loro principale svantaggio è la **velocità**: poiché sono coinvolti gli esseri umani, la soluzione è più adatta alle applicazioni in cui il tempo non è critico.

La soluzione della conoscenza

La soluzione della conoscenza chiede come si possa sapere se un fatto presunto è falso o è probabile che sia falso.

Gli errori di ortografia e gli errori di sintassi sono semplici indicatori che tutti noi utilizziamo per valutare le e-mail dannose.

Nel caso geografico, ci si può chiedere se un fatto presunto sia coerente con ciò che è già noto sul mondo geografico, sia in termini di fatti che di teorie.

Inoltre, tali controlli di coerenza possono potenzialmente essere automatizzati, consentendo il triage in tempo quasi reale.

Questo approccio è stato implementato, sebbene su una base in qualche modo non strutturata, dalle aziende che ricevono quotidianamente migliaia di correzioni volontarie ai loro database geografici.

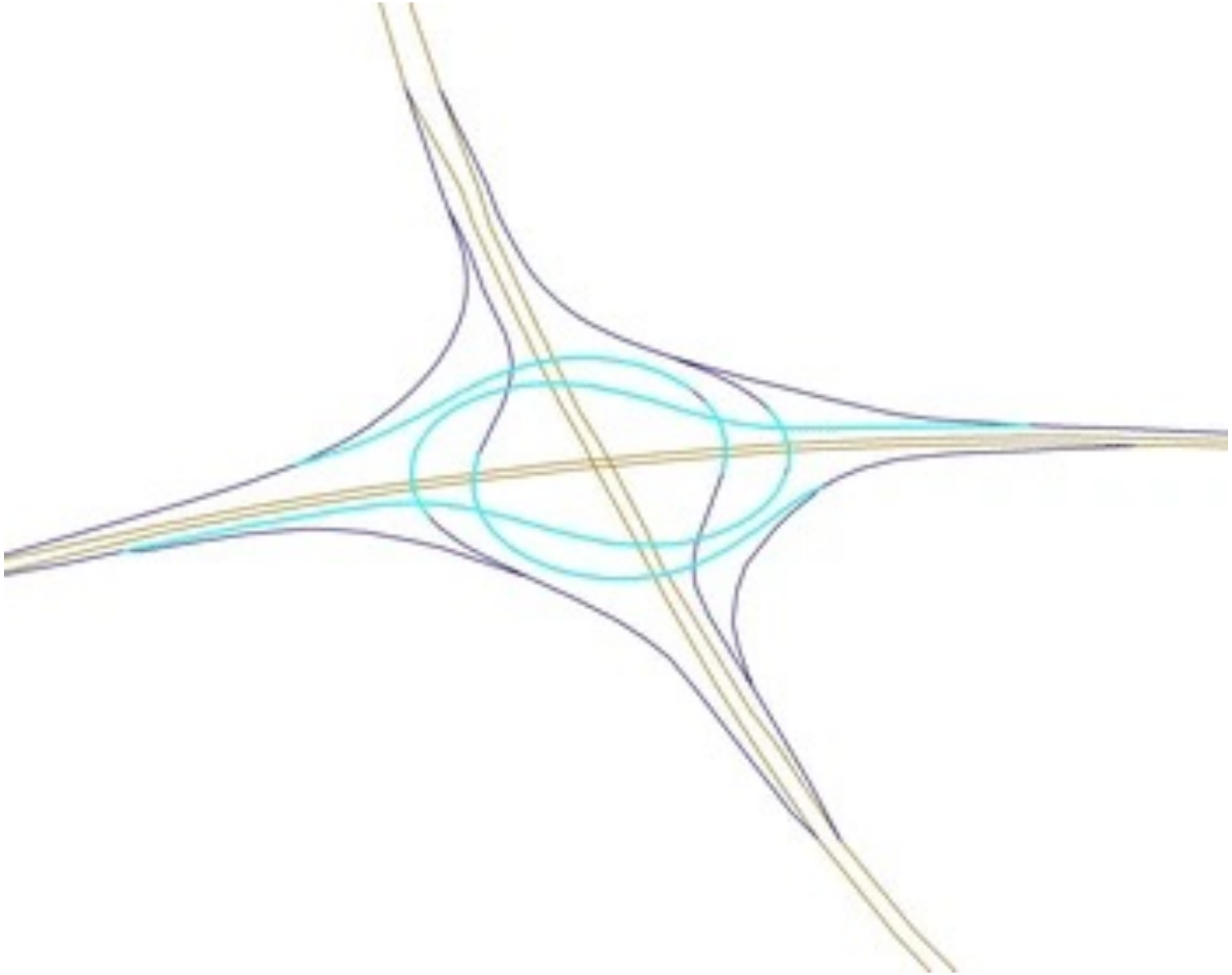
Fatti presunti

Un fatto presunto può discostarsi dalla conoscenza geografica consolidata nella sintassi o nella semantica, o in entrambi.

La **sintassi** si riferisce alle regole con cui il mondo è costruito, mentre la **semantica** si riferisce al significato di quei fatti.

La conoscenza sintattica è spesso più facile da controllare rispetto alla conoscenza semantica.

Ad esempio, la seguente figura illustra un esempio di conoscenza geografica sintattica.



Sappiamo dalle specifiche tecniche che una rampa può intersecare un'autostrada con un piccolo angolo (in genere 30 gradi o meno).

Se un database della rete stradale sembra avere intersezioni su rampa con angoli maggiori di 30 gradi, sappiamo che è probabile che i dati siano errati.

Nel caso in figura, molte delle intersezioni apparenti dei segmenti azzurri hanno maggiori probabilità di essere cavalcavia o sottopassi.

Tali errori sono stati definiti errori di **coerenza logica** nella letteratura della scienza dell'informazione geografica.

Al contrario, la seguente figura illustra conoscenze geografiche semantiche: una fotografia di un lago che è stata collegata alla mappa di Google Earth del campus della Ohio State University.

Tuttavia, questa fotografia sembra essere posizionata in modo errato: riconosciamo la scena come Mirror Lake, un'icona del campus a sud-est della presunta posizione indicata sulla mappa.

La posizione presunta deve essere errata, ma possiamo esserne certi?

Forse l'università ha spostato Mirror Lake per far posto a un nuovo edificio di geourbanistica?

O forse Mirror Lake era così popolare che l'università ha creato un Mirror Lake per gestire l'eccesso di folla?

Non possiamo immediatamente e con assoluta sicurezza respingere questo fatto empirico senza ulteriori indagini poiché non viola alcuna regola nota con cui il mondo è costruito: non c'è nulla che impedisca al Mirror Lake di essere spostato o duplicato.

Certo, ci sono alcuni fatti semantici che possono essere subito liquidati come assurdi: non ci si aspetterebbe di vedere il Mirror Lake sulla cima del monte Everest o nel deserto del Sahara.

Tuttavia, non esiste una linea di demarcazione forte tra **fatti semantici chiaramente assurdi** e **non assurdi**. Per esempio non ci si aspetterebbe di vedere Venezia nel deserto del Mojave, eppure…

Las Vegas esiste davvero.

Un compito importante per la soluzione della conoscenza è la formalizzazione della conoscenza per supportare la selezione automatizzata di fatti asseriti e la sintesi automatizzata di dati.

La conoscenza può essere derivata empiricamente o come predizione da teorie, modelli e simulazioni.

In quest'ultimo caso, potremmo ricercare dati in contrasto con le previsioni nell'ambito dei processi di scoperta e costruzione della conoscenza.

Ci sono almeno due grandi sfide nella formalizzazione della conoscenza geografica.

In primo luogo, concetti geografici come “quartiere”, “regione”, “il Midwest” e “nazioni in via di sviluppo” possono essere vaghi, fluidi e contestati.

Una seconda sfida è lo sviluppo di rappresentazioni esplicite, formali e calcolabili della conoscenza geografica.

Molta conoscenza geografica è sepolta in teorie, modelli ed equazioni formali che devono essere risolte o è espressa in un linguaggio informale che deve essere interpretato.

Al contrario, le tecniche di knowledge-discovery dei Big Data richiedono rappresentazioni esplicite come regole, gerarchie e reti di concetti a cui è possibile accedere direttamente senza elaborazione.

Sfida 3: Correlazione, non causalità

Tradizionalmente, la ricerca accademica si preoccupa di sapere **perché** succede qualcosa.

Le **correlazioni** da sole non sono sufficienti, perché l'esistenza della correlazione non implica che il cambiamento in una delle variabili **causi** un cambiamento nell'altra.

Nella correlazione esplorata da Openshaw e Taylor citata in precedenza, l'esistenza di una correlazione tra il numero di repubblicani registrati in una contea e il numero di persone di età pari o superiore a 65 anni non implica che uno dei due abbia un effetto causale dall'altra.

Nel corso degli anni, la scienza ha adottato frasi peggiorative per descrivere la ricerca che cerca correlazioni senza preoccuparsi della causalità o della spiegazione, come "*correlation is not causation*" o "*curve-fitting*".

Tuttavia le correlazioni possono essere utili per la previsione, specialmente se si è disposti a supporre che una correlazione osservata possa essere generalizzata al di là delle circostanze specifiche in cui è osservata.

Inoltre, sebbene possano essere condizioni sufficienti, la spiegazione e la causalità non sono condizioni necessarie per la ricerca scientifica: molta ricerca, specialmente in settori come l'analisi spaziale, riguarda l'avanzamento della metodologia, sia che il suo eventuale utilizzo sia per cercare spiegazioni o per fare previsioni.

La letteratura della scienza dell'informazione geografica è piena di strumenti che sono stati progettati non per trovare spiegazioni ma per attività più semplici come rilevare schemi o manipolare dati per la visualizzazione.

Tali strumenti sono chiaramente preziosi in un’era della scienza basata sui dati, in cui le domande sul “perché” potrebbero non essere così importanti.