Mapping e Big Data
Lezione 9

Mario Verdicchio
Università degli Studi di Bergamo
Anno Accademico 2019-2020

Data-driven geography (parte 2)

Harvey J. Miller & Michael F. Goodchild
The Ohio State University, Columbus, USA

GeoJournal 80 (2015)

Teoria nelle discipline data-driven: quelli a favore (1/4)

- La posizione generale in questa fazione è “il diluvio di dati rende obsoleto il metodo scientifico”
- Prendendo la **fisica** e la **biologia** come esempi, è “evidente” che teorie e modelli sono solo caricature di una realtà più profonda che non può essere facilmente e forse non sarà mai spiegata
- Tuttavia, la spiegazione non è necessaria per il progresso: la correlazione sostituisce la causalità e la scienza può avanzare anche senza modelli coerenti, teorie unificate o una spiegazione meccanicistica

Teoria nelle discipline data-driven: quelli a favore (2/4)

- Anche nelle **scienze sociali**, c'è chi afferma che volumi senza precedenti di dati sociali hanno il potenziale per rivoluzionare la nostra comprensione della società
- Questa comprensione però non sarà sotto forma di leggi generali delle scienze sociali o di relazioni sociali causa-effetto
- Non si proclama la fine della teoria, ma piuttosto un tipo più modesto di teoria che includa proposizioni generali (come ad esempio quali interventi funzionano per particolari problemi sociali) o ipotesi su come fatti sociali più comuni si compongano per generare risultati meno comuni

Teoria nelle discipline data-driven: quelli a favore (3/4)

- Il sociologo Robert Merton a metà del XX secolo le chiamò teorie di medio raggio (“mid-range”): teorie che affrontano fenomeni sociali identificabili anziché entità astratte come l'intero sistema sociale
- Le teorie di medio raggio sono empiriche: si basano su osservazioni e servono a ricavare ipotesi che possono essere investigate
- Tuttavia, non sono punti di arrivo: piuttosto sono delle milestone temporanee verso schemi concettuali generali che possono comprendere più teorie di medio raggio

Teoria nelle discipline data-driven: quelli a favore (4/4)

- La scienza basata sui dati sembra comportare uno spostamento dal generale verso lo specifico, lontano dai tentativi di trovare leggi universali
- Ci sono chiaramente alcuni vantaggi di questo cambiamento: nella **pianificazione urbana** nell'era degli Scarce Data ci si è concentrati su cambiamenti radicali e massicci a lungo termine delle città, con poca preoccupazione per i piccoli spazi e i movimenti locali
- La pianificazione urbana data-driven può correggere alcuni dei problemi urbani conseguenti da questa omissione consentendo una maggiore attenzione al locale e alla routine

Teoria nelle discipline data-driven: conseguenze a lungo termine

- Su periodi più lunghi e domini spaziali più ampi, il locale e la routine si fondono con il lungo termine
- Una sfida scientifica fondamentale è come i Big Data locali e a breve termine possano informare la nostra comprensione dei processi su orizzonti temporali e spaziali più lunghi
- In altre parole, il tradizionale problema scientifico della **generalizzazione** si ripresenta

E in geografia?

Brevissima storia della geografia
nella battaglia Generale vs. Particolare

Ricordiamo che…

- la conoscenza **nomotetica** è quella che deriva da attività scientifica impegnata nella ricerca di leggi generali
- la conoscenza **idiografica** è invece la conoscenza particolare che si ottiene cercando di fornire la descrizione di casi specifici

La geografia ai tempi di…

- Strabone (64/63 a.C.-24 d.C.) e Tolomeo (90-168 d.C.) comportava sia generalizzazioni sulla Terra che descrizioni intime di luoghi e regioni specifici
- Questi due aspetti erano considerati due lati della stessa medaglia

La geografia ai tempi di…

- …Bernhardus Varenius (1622-1650) era costituita da conoscenze generali (scientifiche) e speciali (regionali), sebbene egli considerasse quest'ultima una sussidiaria della prima

La geografia ai tempi di…

- …Alexander von Humboldt (1769–1859) e Carl Ritter (1779–1859), spesso considerati i fondatori della geografia moderna, puntava a ricavare leggi generali attraverso un’attenta misurazione dei fenomeni geografici in luoghi e tempi particolari

La geografia ai tempi di…

- dell'inizio del XX secolo vedeva un predominio della geografia nomotetica nelle vesti del determinismo ambientale nei primi anni del 1900…
- …seguito da un contraccolpo contro i suoi abusi e dal conseguente aumento della geografia idiografica sotto forma di differenziazione areale: Richard Hartshorne dichiarò in The Nature of Geography (1939) che l'unica legge della geografia è che “tutte le aree sono uniche”

La geografia ai tempi di…

- …metà XX secolo vede un altro contraccolpo
- Il dominio della geografia idiografica e la concomitante crisi nella geografia accademica americana (in particolare, la chiusura del programma geografico di Harvard nel 1948) portarono alla rivoluzione quantitativa degli anni ‘50 e ‘60
- Geografi come Fred Schaefer, William Bunge, Peter Haggett e Edward Ullman affermano che la geografia dovrebbe essere una scienza alla ricerca della legge che risponde alla domanda “perché?” piuttosto che costruire una raccolta di fatti che descrivono ciò che sta accadendo in determinate regioni

La geografia ai tempi di…

- …oggi: i geografi fisici si sono disimpegnati da questi dibattiti, ma la tensione tra approcci nomotetici e idiografici persiste nella geografia umana
- Tuttavia, i tentativi di conciliare la conoscenza nomotetica e idiografica non sono morti con Humboldt e Ritter
- Approcci come la geografia del tempo cercano di catturare il contesto e la storia e riconoscere i ruoli sia dell'agenzia che della struttura nel comportamento umano

La geografia ai tempi d'oggi

- Nell'analisi spaziale, la tendenza verso le statistiche locali, esemplificata dalla regressione geograficamente ponderata e dagli indicatori locali di associazione spaziale, rappresenta un compromesso in cui i principi generali della geografia nomotetica possono esprimersi in modo diverso attraverso lo spazio geografico
- Goodchild ha caratterizzato i GIS come l'unione di:
  - conoscenza nomotetico, nei loro software e algoritmi, e
  - conoscenza idiografica nei loro database

Geografia data-driven

- I percorsi di conoscenza geografica generati da approcci ad alta intensità di dati come la geografia del tempo, la disaggregazione delle statistiche spaziali e GIScience sono un ritorno ai primordi della geografia in cui né la ricerca della legge generale né la ricerca della descrizione del caso particolare erano privilegiate
- Le generalizzazioni e le leggi geografiche sono possibili ma lo spazio conta: la dipendenza spaziale e l'eterogeneità spaziale creano un contesto locale che modella i processi fisici e umani mentre si evolvono sulla superficie della Terra
- Questa convinzione è supportata anche da recenti scoperte nella teoria dei sistemi complessi, che suggerisce che i modelli di interazioni locali portano a comportamenti **emergenti** che non possono essere compresi isolatamente a livello locale o a livello globale
- Comprendere le interazioni tra agenti all'interno di un ambiente è la colla scientifica che lega il locale al globale.

Geografia data-driven

- La geografia basata sui dati non è necessariamente una rottura radicale con la tradizione geografica: la geografia ha una credenza di vecchia data nel valore della conoscenza idiografica presa da sola, così come nel suo ruolo nella costruzione della conoscenza nomotetica
- Sebbene questa convinzione sia stata a volte contestata, la geografia basata sui dati può fornire i percorsi tra conoscenza idiografica e nomotetica che i geografi hanno cercato per due millenni
- Tuttavia, mentre la teoria della complessità supporta questa prospettiva, suggerisce anche che questo tipo di conoscenza possa avere limiti intrinseci:
  il comportamento emergente è per definizione sorprendente

Approcci
alla geografia data-driven

Se funziona, funziona.

Premessa

- Se accettiamo la premessa - almeno fino a prova contraria - che i Big Data e la scienza basata sui dati si armonizzino con temi e credenze di lunga data nella geografia, la domanda che segue è:

  come possono gli approcci basati sui dati adattarsi alla ricerca geografica?

- Gli approcci basati sui dati possono supportare sia la scoperta della conoscenza geografica sia la modellazione spaziale, con le solite sfide e precauzioni che devono essere riconosciute

Knowledge Discovery

- La Knowledge Discovery geografica si riferisce alla fase iniziale del processo scientifico in cui lo sperimentatore forma la sua visione concettuale del sistema, sviluppa ipotesi da testare e costruisce le basi per supportare il processo di costruzione della conoscenza
- I dati geografici facilitano questa fase cruciale del processo scientifico sostenendo attività come la selezione e la ricognizione del sito di studio, l'etnografia, la progettazione sperimentale e la logistica.

Knowledge Discovery

- Forse l'impatto più trasformativo della scienza basata sui dati sulla scoperta della conoscenza geografica sarà attraverso l'esplorazione dei dati e la generazione di ipotesi
- Simile a un telescopio o un microscopio, i sistemi per l'acquisizione, l'archiviazione e l'elaborazione di enormi quantità di dati possono consentire agli investigatori di aumentare le loro percezioni della realtà e vedere cose che altrimenti sarebbero nascoste o troppo deboli per essere percepite
- Da questo punto di vista, la scienza basata sui dati non è necessariamente un approccio radicalmente nuovo, ma piuttosto un modo per migliorare l'inferenza per i processi di esplorazione e generazione di ipotesi di lunga data prima della costruzione della conoscenza attraverso analisi, modellizzazione e verifica

Ragionamento abduttivo

- La scoperta della conoscenza basata sui dati ha una base filosofica: il ragionamento **abduttivo**, una forma di inferenza articolata dall'astronomo e matematico C. S. Peirce (1894-1914)
- Il ragionamento abduttivo inizia con i dati che descrivono qualcosa e termina con un'ipotesi che spiega i dati
- È una forma più debole di inferenza rispetto al ragionamento **deduttivo** o **induttivo**:
  - il ragionamento deduttivo mostra che X deve essere vero
  - il ragionamento induttivo mostra che X è vero
  - mentre il ragionamento abduttivo mostra solo che X può essere vero
- Tuttavia, il ragionamento abduttivo è di fondamentale importanza nella scienza, in particolare nella fase di scoperta iniziale che precede l'uso di approcci deduttivi o induttivi alla costruzione della conoscenza.

Ragionamento abduttivo

- Il ragionamento abduttivo richiede quattro capacità:
  1. la capacità di sostenere nuovi frammenti di teoria
  2. un vasto insieme di conoscenze da cui attingere, che vanno dal buon senso alle competenze di dominio
  3. un mezzo per cercare in questa raccolta di conoscenze connessioni tra modelli di dati e possibili spiegazioni
  4. strategie complesse di risoluzione dei problemi come analogia, approssimazione e ipotesi

Geovisualizzazione

- Gli esseri umani hanno dimostrato di avere più successo delle macchine nell'esecuzione di questi compiti complessi, suggerendo che la scoperta della conoscenza basata sui dati dovrebbe cercare di sfruttare queste capacità umane attraverso metodi come la geovisualizzazione piuttosto che tentare di automatizzare il processo di scoperta
- Si può proporre un processo incentrato sull'uomo in cui la geovisualizzazione funge da quadro centrale per la creazione di catene di inferenza tra approcci abduttivi, induttivi e deduttivi nella scienza, consentendo più interazioni e sinergie tra questi approcci alla costruzione della conoscenza geografica

Esplorazione dei dati

- Uno dei problemi con i Big Data è la dimensione e la complessità dello spazio informazioni implicato da un enorme database multivariato
- Un buon sistema di esplorazione dei dati dovrebbe generare tutti i modelli interessanti in un database, ma **solo** quelli interessanti per evitare di sovraccaricare l'analista
- Due modi per gestire il gran numero di potenziali modelli sono le **conoscenze di base** e le **misure di interesse**

Conoscenze e misure

- La conoscenza di base guida la ricerca di schemi rappresentando le conoscenze accettate sul sistema per focalizzare la ricerca di nuovi schemi
- Al contrario, possiamo usare le misure di interesse a posteriori per filtrare i modelli spuri valutando ogni modello in base a dimensioni come semplicità, certezza, utilità e novità
- I pattern con rating al di sotto di una soglia specificata dall'utente vengono scartati o ignorati
- Entrambi questi approcci richiedono la formalizzazione della conoscenza geografica, una sfida discussa in precedenza

Modellazione basata sui dati

- Gli approcci **tradizionali** alla modellistica sono **deduttivi**: lo scienziato sviluppa (o modifica o prende in prestito) una teoria e deriva una rappresentazione formale che può essere manipolata per generare previsioni sul mondo reale che possono essere testate con i dati
- La modellazione **senza teoria**, invece, costruisce modelli basati sull'**induzione dai dati** piuttosto che sulla deduzione dalla teoria

Modellazione basata sui dati

- Nel campo dell'economia molti hanno lavorato con la modellazione basata sui dati sotto forma di modellazione "generale-specifico".
- In questa strategia, il ricercatore inizia con il modello più complesso possibile e lo riduce a un modello più elegante sulla base dei dati
- Il tutto è fondato sulla convinzione che, con una quantità sufficiente di dati, solo il modello "vero" sopravvivrà a una batteria rigorosa di test statistici progettati per tagliare via variabili dal modello

Modellazione basata sui dati

- Ciò contrasta con la tradizionale strategia da specifico a generale in cui si inizia con un modello preconfezionato basato sulla teoria e si costruisce in modo conservativo un modello più complesso.
- Tuttavia, l'approccio generale-specifico è controverso, con alcuni che sostengono che, dato l'enorme numero di potenziali modelli, si dovrebbe essere molto fortunati a riconoscere il vero modello all'interno del modello iniziale e complesso
- Pertanto, la performance **predittiva** è l'unico criterio rilevante: **la spiegazione è irrilevante**

GAM

- La geografia ha anche assistito a tentativi di modellazione senza teoria, non senza polemiche
- Stan Openshaw è un sostenitore particolarmente forte dell'uso della potenza dei computer per costruire modelli dai dati: esempi includono la Geographical Analysis Machine (GAM) per il raggruppamento spaziale di dati puntuali e sistemi automatizzati per la modellazione delle interazioni spaziali
- GAM utilizza una tecnica che genera cluster locali o “hot spot” senza richiedere una teoria a priori o conoscenze sulla distribuzione statistica sottostante
- GAM cerca i cluster espandendo sistematicamente la ricerca circolare da posizioni all'interno di un reticolo
- Il sistema salva i cerchi con conteggi osservati maggiori del previsto e quindi varia sistematicamente i raggi e la risoluzione del reticolo per ricominciare la ricerca

Senza teoria

- Il ricercatore non ha bisogno di ipotizzare o avere alcuna aspettativa riguardo alla distribuzione spaziale del fenomeno: il sistema fa una ricerca, in maniera bruta, di tutte le possibili (o ragionevoli, almeno) risoluzioni spaziali e relative zone
- Il sistema automatizzato utilizza la tecnica programmazione genetica (metaforica) per generare modelli di interazione spaziale da elementi di base come le variabili del modello, forme funzionali, parametrizzazioni e operatori usando il livello di adattamento come criterio di selezione (non dimenticate il denigratorio “curve fitting”)

Problemi problemi problemi (1)

- Una sfida nella modellazione priva di teoria è che toglie un potente meccanismo per migliorare l'efficacia di una ricerca di un modello esplicativo: la **teoria**
- La teoria ci dice dove cercare una spiegazione e (forse ancora più importante) dove non cercare
- Nel caso specifico della modellizzazione dell'interazione spaziale, ad esempio, la necessità che i modelli siano coerenti dimensionalmente può limitare le opzioni
- Lo spazio informativo implicito in un universo di potenziali modelli può essere enorme anche in un dominio limitato come l'interazione spaziale
- Computer potenti e tecniche di ricerca intelligenti possono sicuramente migliorare le nostre possibilità, ma all'aumentare del volume, della varietà e della velocità dei dati, aumenta anche la dimensione degli spazi informativi per i possibili modelli, portando a un tipo di corsa agli armamenti (arms race) senza forse un chiaro vincitore

Problemi problemi problemi (2)

- Una seconda sfida nella modellazione basata sui dati è che i dati guidano la **forma del modello**, il che significa che non esiste alcuna garanzia che lo stesso modello derivi da un set di dati diverso
- Anche dato lo stesso set di dati, potrebbero essere generati molti modelli diversi che si adattano ai dati, il che significa che lievi alterazioni del criterio di buon fit utilizzato per guidare la selezione dei modelli possono produrre modelli molto diversi
- Questo è essenzialmente il problema dell'**overfitting** statistico, un problema ben noto con tecniche induttive come le reti neurali artificiali e l'apprendimento automatico
- Tuttavia, nonostante i metodi e le strategie per evitare l'eccessivo adattamento, esso sembra essere endemico: alcuni stimano che i tre quarti degli articoli scientifici pubblicati sull'apprendimento automatico siano difettosi a causa dell'overfitting

Problemi problemi problemi (3)

- Una terza sfida nella modellazione senza teoria è la **complessità** dei modelli risultanti.
- La costruzione di modelli tradizionali nella scienza utilizza la parsimonia come principio guida: il modello migliore è quello che spiega di più con il minimo
- Questo viene talvolta chiamato “il rasoio di Occam”: dati due modelli con uguale validità, il modello più semplice è migliore
- L’interpretazione del modello è un test informale ma chiave: il modellista deve essere in grado di spiegare cosa dicono i risultati del modello sulla realtà
- I modelli ricavati in maniera computazionale dai dati e messi a punto in base al feedback delle previsioni possono generare **previsioni affidabili** da **processi troppo complessi** per il cervello umano
- Ad esempio, è noto che il sistema automatizzato di Openshaw per la riproduzione di modelli di interazione spaziale genera modelli molto complessi e non intuitivi, molti dei quali sono anche dimensionalmente incoerenti

$$T_{ij} = \frac{\left(\frac{\left(\arctan V_{ij}^{-0.8}\right) + 1}{\operatorname{cotan}\left(\operatorname{lgamma}\left(O_i\right)^{-0.4}\right)} + 0.19\ln\left(O_i D_j\right)\right)}{\operatorname{lgamma}\left(\cos\left(O_i D_j\right) + 0.04\right) - \operatorname{cotan}\left(\cosh A_{ij} - 0.9\right)}$$

$$T_{ij} = \left(\arctan\left(\tanh\left(V_{ij}\right)^{-1.05}\right) + 1\right)\operatorname{atan}\left(\operatorname{lgamma}\left(O_i\right)^{-0.82}\right) + \frac{\sinh\left(\tanh\left(O_i D_j\right) - 0.6\right)}{\operatorname{lgamma}\left(\cos\left(O_i D_j\right) + 0.4\right)} - \operatorname{cotan}\left(\tan\left(A_{ij}\right)^{2.3}\right)$$

$$T_{ij} = \frac{\left(\operatorname{atan}\left(V_{ij}^{-1.4}\right)\tanh\left(\operatorname{lgamma}\left(O_i\right)^{-0.9}\right) + \tanh\left(D_j\right)\right)}{\cos\left(5.1\tanh\left(O_i D_j\right)\right)}$$

Problemi problemi problemi (4)

- La conoscenza dei modelli basati sui dati può essere complessa e non comprimibile: i **dati** sono la spiegazione
- Ma se la spiegazione non è comprensibile, abbiamo davvero una spiegazione?
- Forse la natura della spiegazione si sta evolvendo
- Forse i computer sono fondamentali nella scienza basata sui dati non solo per la scoperta, ma anche per rappresentare schemi complessi che vanno oltre la comprensione umana

NO!

- Forse questo è un temporaneo divario fino a quando non raggiungeremo la convergenza tra intelligenza umana e macchina come prevedono alcuni (Kurzweil 1999)

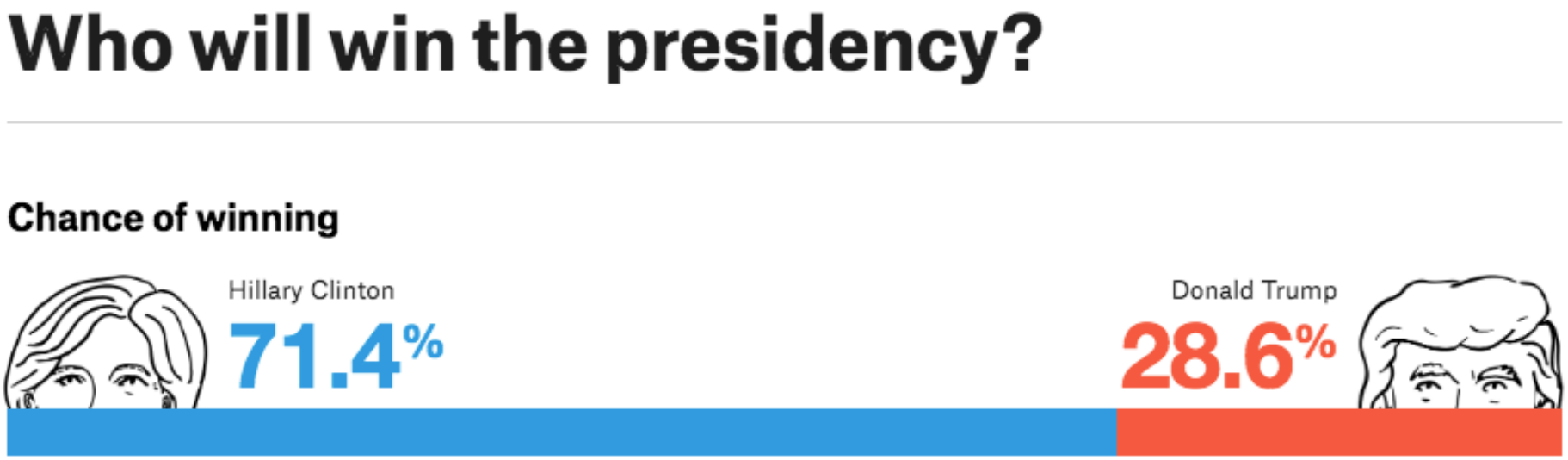
Inventore di indubbio talento, è stato pure assunto ai vertici di Google (suscitando non poche perplessità); esprime posizioni molto controverse su “macchine pensanti”; oggi vende anche vitamine per allungare la vita…

Avvertenza

- Sebbene non possiamo sperare di risolvere questa domanda (o le sue implicazioni filosofiche) a breve, possiamo aggiungere una nota cautelativa di Nate Silver:

  raccontare storie sui dati invece della realtà è pericoloso e può portare a confondere il rumore per segnale

Nate Silver nel 2016



Fonte: https://projects.fivethirtyeight.com/2016-election-forecast/

Problemi problemi problemi (5)

- Un'ultima sfida nella modellazione spaziale basata sui dati è lo **de-skilling**: una perdita di capacità di modellazione e analisi
- Mentre assegnare compiti banali ai computer rende gli esseri umani liberi di svolgere attività sofisticate, ci sono momenti in cui le abilità banali diventano cruciali
- Ad esempio, ci sono casi documentati di piloti di compagnie aeree, a causa della mancanza di esperienza di volo manuale, hanno reagito male in caso di emergenza quando l'autopilota si spegne
- Sebbene raramente pericoloso per la vita, si potrebbe argomentare in modo simile sulla costruzione automatica di modelli: se un processo di modellazione basato sui dati genera risultati anomali, l'analista sarà in grado di determinare se sono artefatti o autentici?
- Più i risultati sono anomali, più profondo è il pensiero richiesto, ma il pensiero umano può degradare a causa di un affidamento eccessivo sulle macchine (**automation bias**)

Automazione e geografia

- Rileggendo il saggio di Jerry Dobson sulla geografia automatizzata quasi 40 anni dopo ("Automated geography" su *The Professional Geographer*, 35, 135-143,1983), si rimane colpiti dal numero di attività geografiche che un tempo erano scrupolose ma ora sono ridotte alla pressione di un pulsante
- I geografi di una certa età si ricordano di corsi di cartografia di base e di produzione
- Quali abilità che consideriamo essenziali oggi saranno considerate obsolete domani?
- Che cosa perderemo?

Cautela, cautela, cautela

- Note cautelative sull'impatto della geografia basata sui dati sulla società più ampia:
  - dobbiamo essere consapevoli del luogo in cui si sta svolgendo questa ricerca, alla luce della ricerca accademica in cui è possibile la revisione tra peer e la riproducibilità, oppure dietro le porte chiuse di società del settore privato e agenzie governative, come prodotti proprietari senza revisione tra peer e senza piena riproducibilità?
  - la privacy è una preoccupazione vitale, non solo come un diritto umano ma anche come una potenziale fonte di contraccolpo che chiuderà la ricerca guidata dai dati
  - dobbiamo stare attenti a evitare i pre-crimini e le pre-punizioni: categorizzare e reagire a persone e luoghi in base a potenziali derivati da correlazioni piuttosto che dal comportamento reale
  - dobbiamo evitare una dittatura dei dati: la ricerca basata sui dati dovrebbe supportare, non sostituire, il processo decisionale di esseri umani intelligenti e scettici