



UNIVERSITÀ
DEGLI STUDI
DI TORINO



LA SCIENZA DEI BIG DATA E LO STUDIO DEI SISTEMI SOCIALI

Relatore: Dott. Fabio Gatto

SCHEMA DELLA PRESENTAZIONE:

① Teoria dei Network

- a) Cenni storici
- b) Sistemi complessi: definizione ed esempi
- c) Regole e caratteristiche principali
- d) Impatto nei vari settori
- e) Grafi

② Machine learning

- a) Esempi di applicazione del machine learning
- b) Tracce digitali e informazioni implicite
- c) Costruzione dell'algoritmo per l'apprendimento dei dati
- d) Applicazioni nei sistemi sociali
- e) Panoramica delle materie legate a questo tipo di analisi

I BIG DATA SONO DAVVERO BIG?

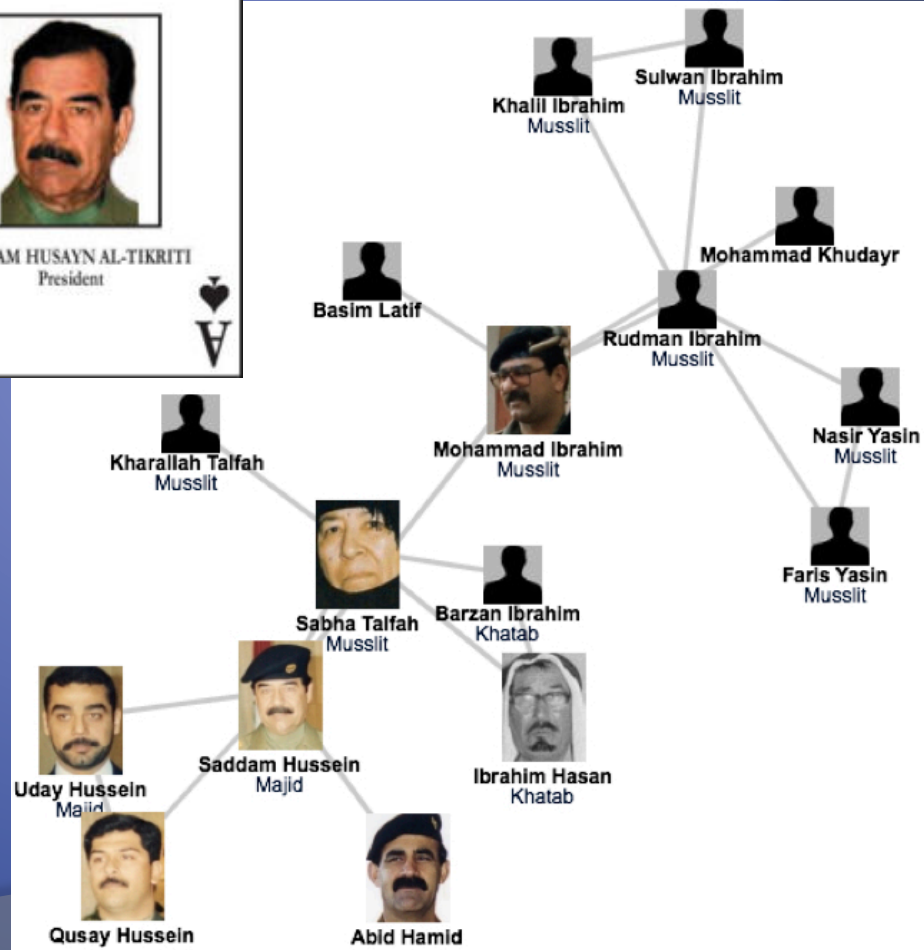
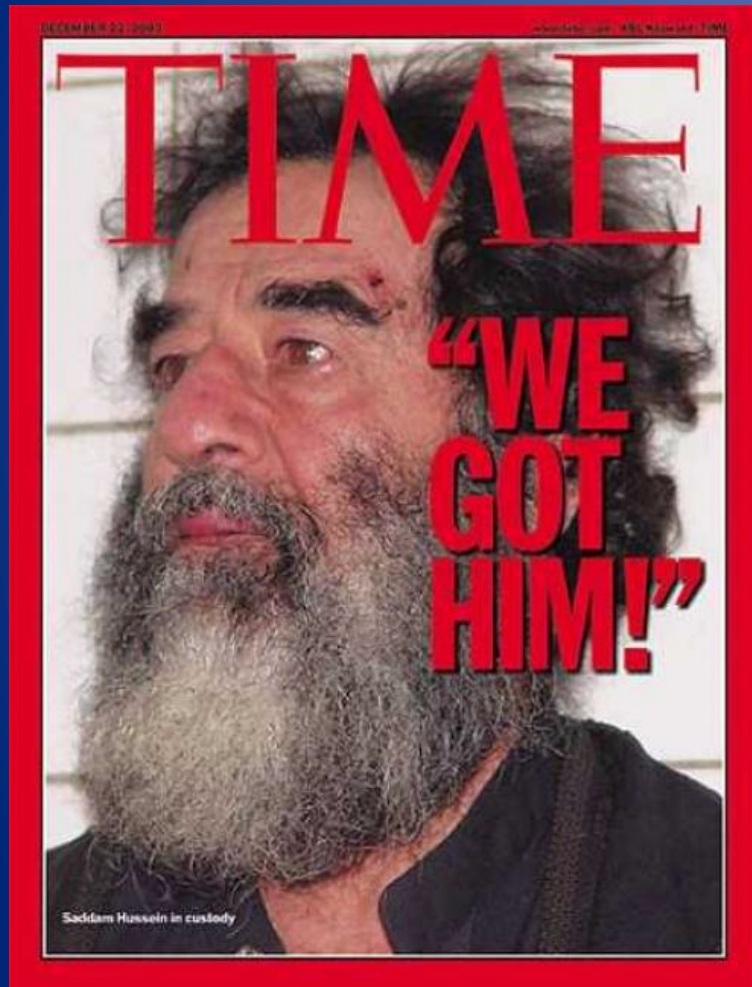
Se ci limitiamo alle comunicazioni digitali, nel mondo, ogni giorno, viene prodotto un numero di byte maggiore del numero di Avogadro.

È come se stessimo ricoprendo il pianeta di una “sostanza” intangibile e virtuale che ha più byte di quanti atomi abbia un frammento macroscopico di materia; mistura di preziosa informazione e di caotico rumore che chiama i “data scientist” a un compito difficile, ovvero estrarre dai dati informazione intelligente.

cit. Mario Rasetti

LA TEORIA DEI NETWORK

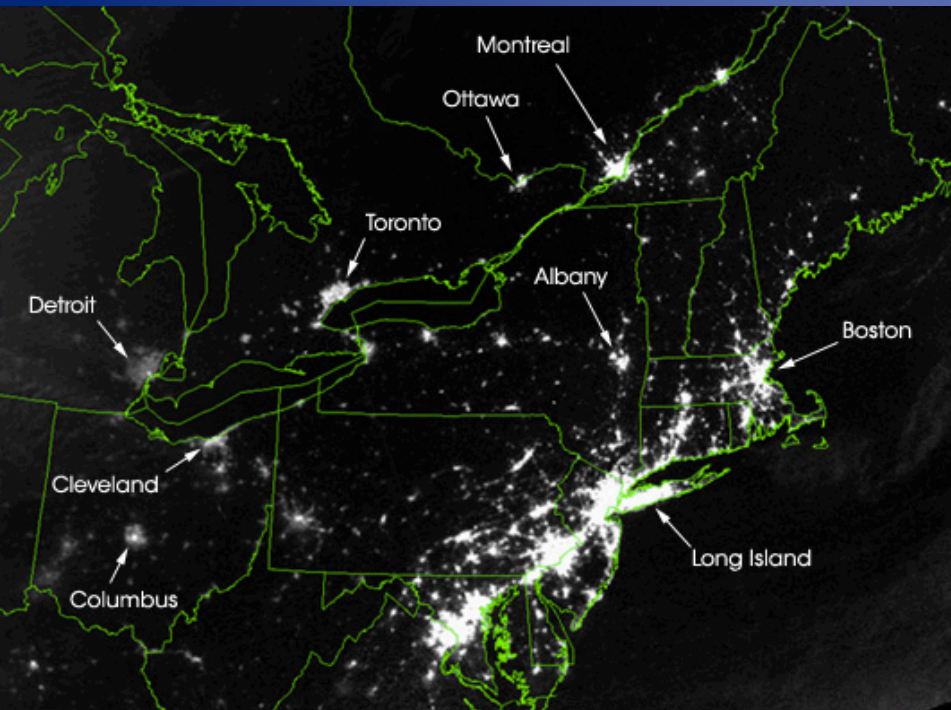
La cattura di Saddam e la scienza dei network



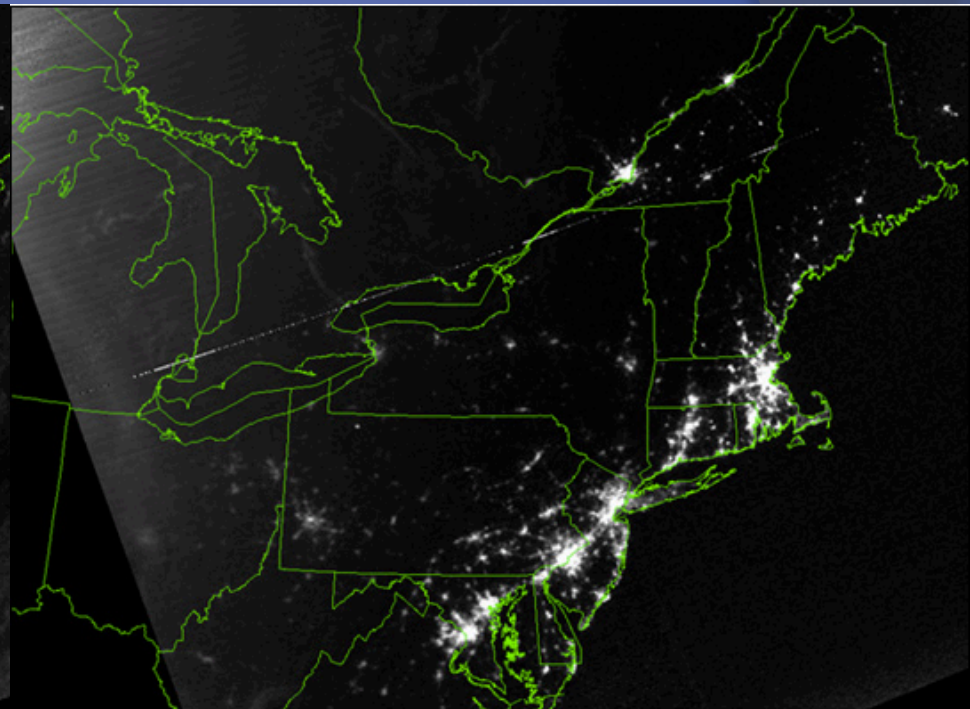
LA CATTURA DI SADDAM HUSSEIN:

- Mostra il forte potere predittivo della teoria dei network.
- Dimostra l'importanza di riuscire ad ottenere una mappa molto accurata della rete di interazioni che vogliamo studiare, ma in certi casi questa mappatura richiede delle azioni eroiche da parte delle persone che cercano le informazioni.
- Evidenzia quanto questo tipo di rete è stabile nel tempo, perché la cattura di Hussein da parte dell'intelligence non è stata basata su informazioni recenti ma sulle interazioni sociali che aveva prima dell'invasione in particolare quelle relative ai familiari.
- Dimostra che la scelta della rete sulla quale decidiamo di concentrarci fa una differenza enorme. Infatti l'albero gerarchico che era stato studiato in precedenza non aveva portato a nessun risultato utile.

Vulnerabilità dell'interconnettività



August 14, 2003: 9:29pm EDT
20 hours before



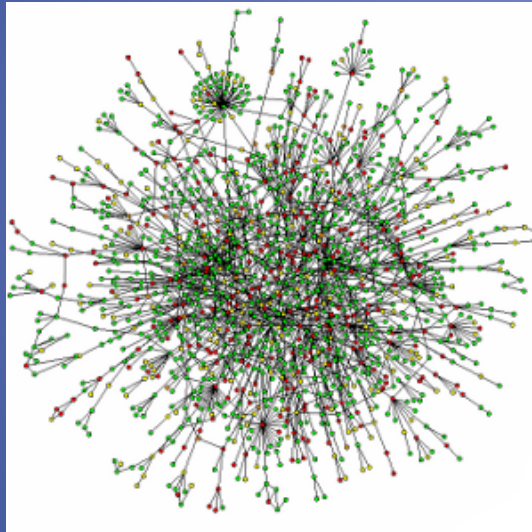
August 15, 2003: 9:14pm EDT
7 hours after

TEMI DA AFFRONTARE IN QUESTO TIPO DI CLASSI :

- Bisogna capire come le strutture delle reti influenzano la robustezza del sistema complesso.
- È necessario sviluppare strumenti quantitativi per valutare l'interazione tra la struttura delle reti e il loro impatto sui fallimenti.
- È possibile dimostrare che i fallimenti seguono delle leggi riproducibili, pertanto tramite degli studi statistici sulla scienza delle reti possiamo quantificare i fallimenti e anche prevederli.

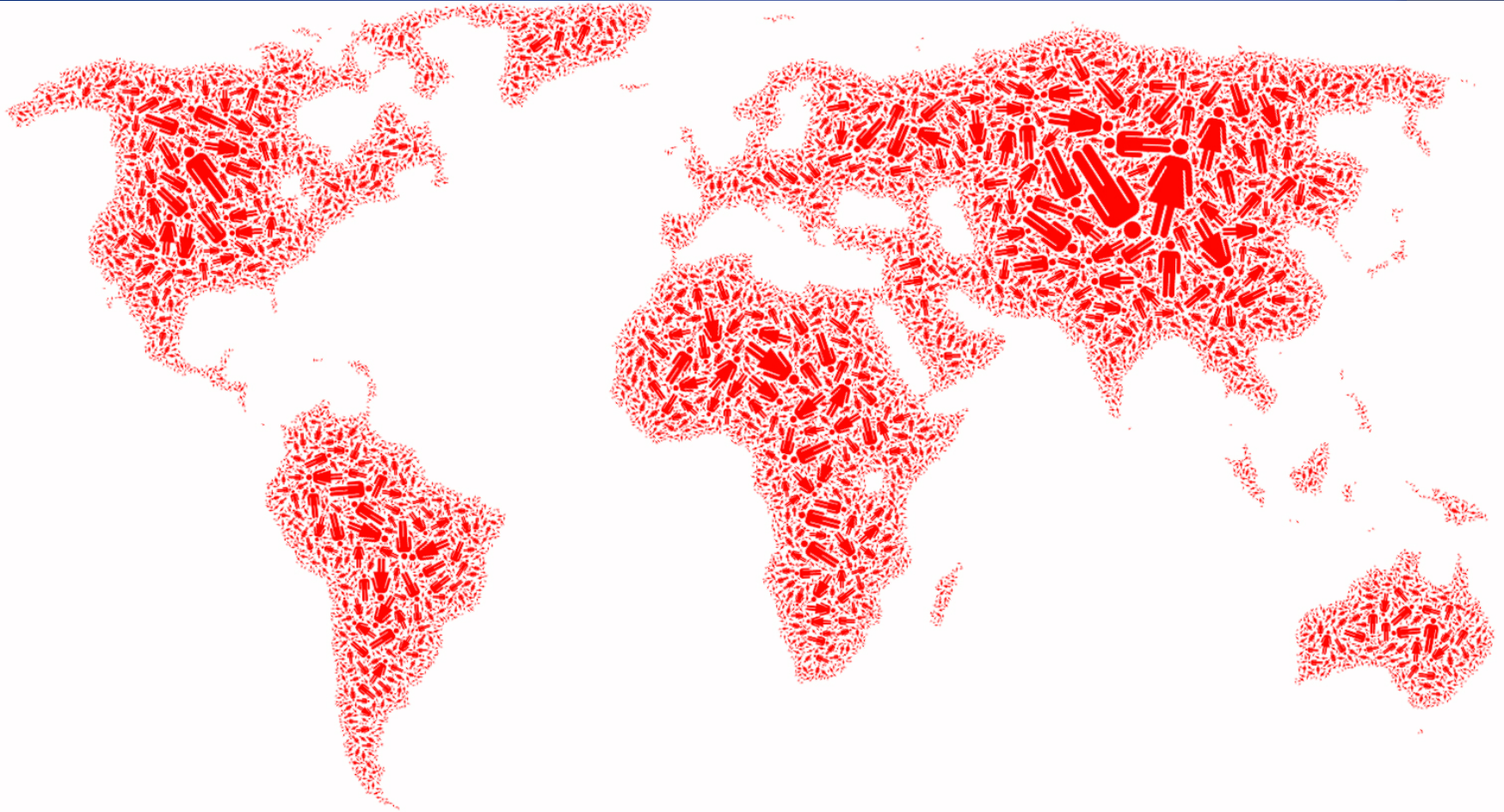
Sistemi Complessi

La complessità è una teoria scientifica che afferma l'esistenza di alcuni sistemi di visualizzazione di fenomeni comportamentali totalmente inspiegabili dalle analisi di singole componenti di tali sistemi.



Questi fenomeni, comunemente indicati come comportamenti emergenti, sembrano verificarsi in molti sistemi che vengono pertanto definiti complessi, come i sistemi viventi e sociali.

LA SOCIETA'



The "Day of 7 Billion" has been targeted by the United States Census Bureau to be in July 2012.

http://en.wikipedia.org/wiki/World_population

LE REGOLE DEI NETWORK

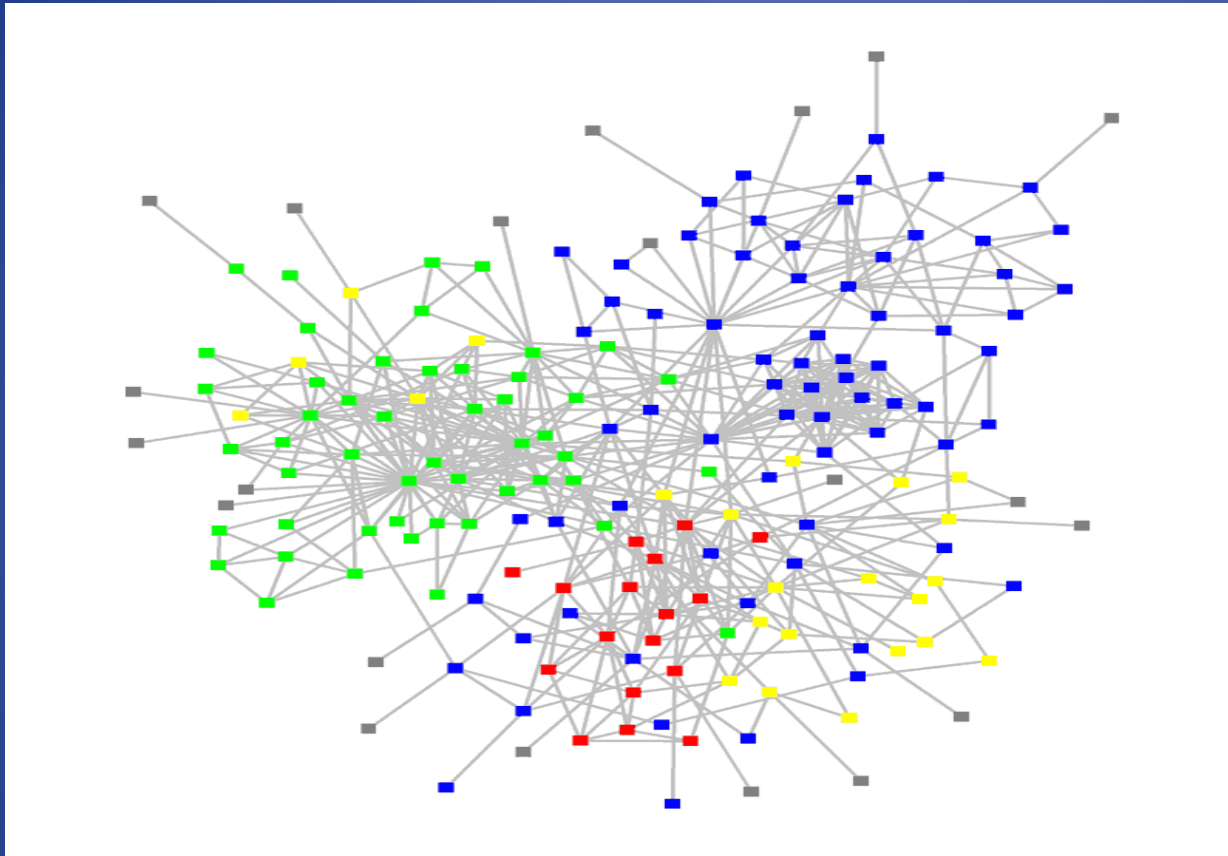
Dietro ogni sistema complesso c'è una rete, che definisce le interazioni tra le varie componenti del sistema.

Facebook

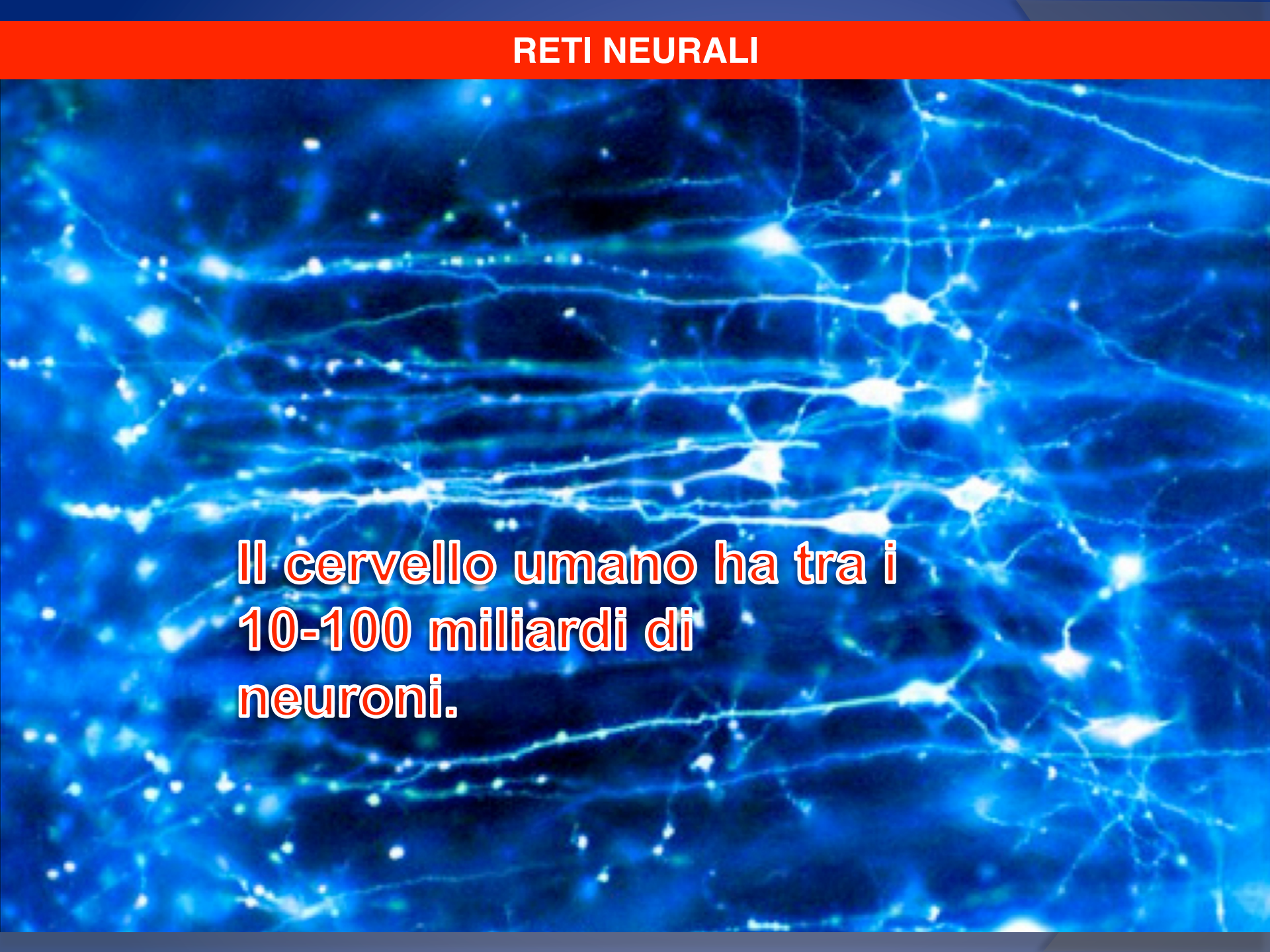


The “Social Graph” behind Facebook

Struttura di un'organizzazione



-    : departments
-  : consultants
-  : external experts



**Il cervello umano ha tra i
10-100 miliardi di
neuroni.**

BUSINESS TIES IN US BIOTECH-INDUSTRY

Nodes:

Companies

Investment

Pharma

Research Labs

Public

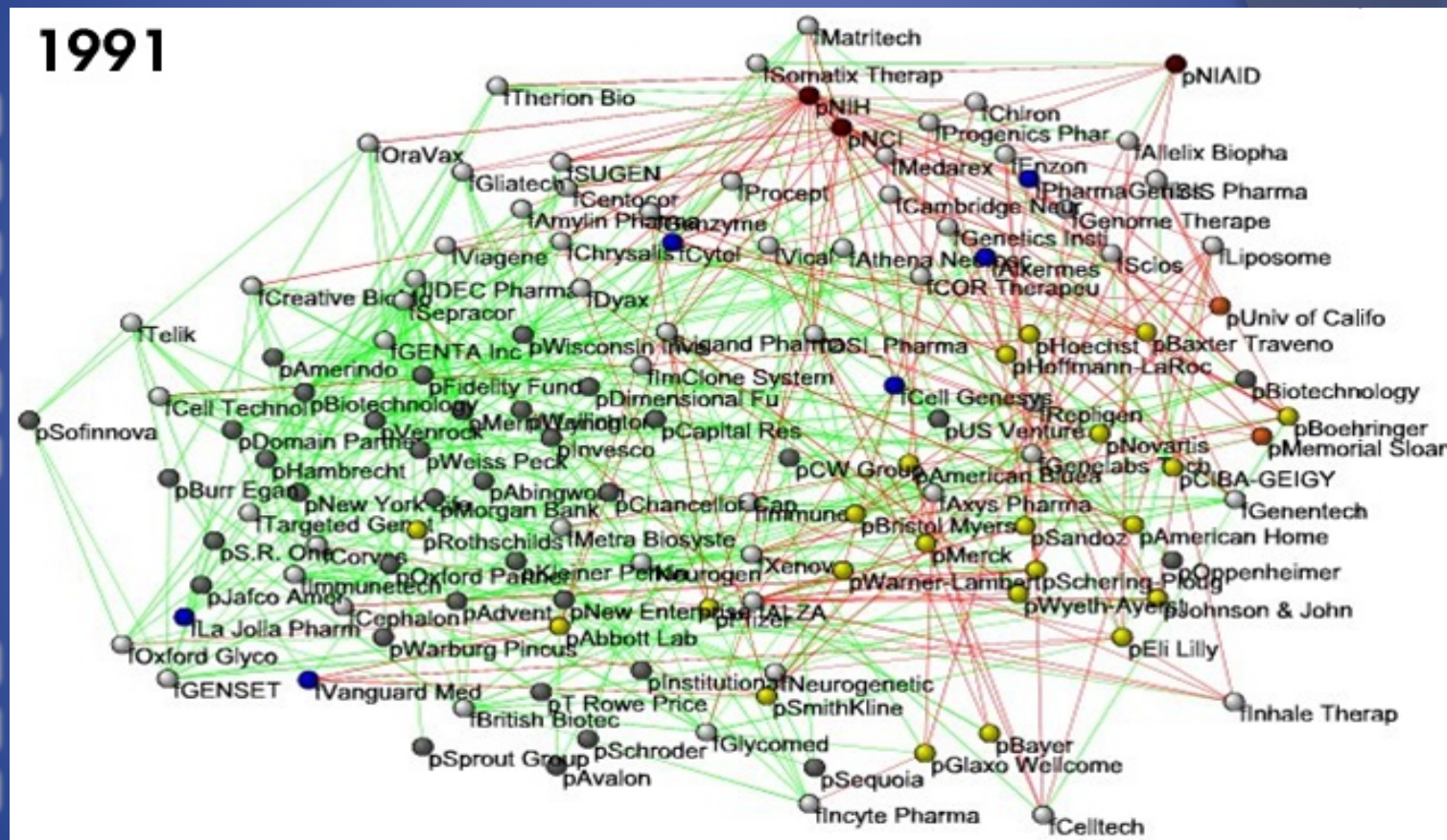
Biotechnology

Links:

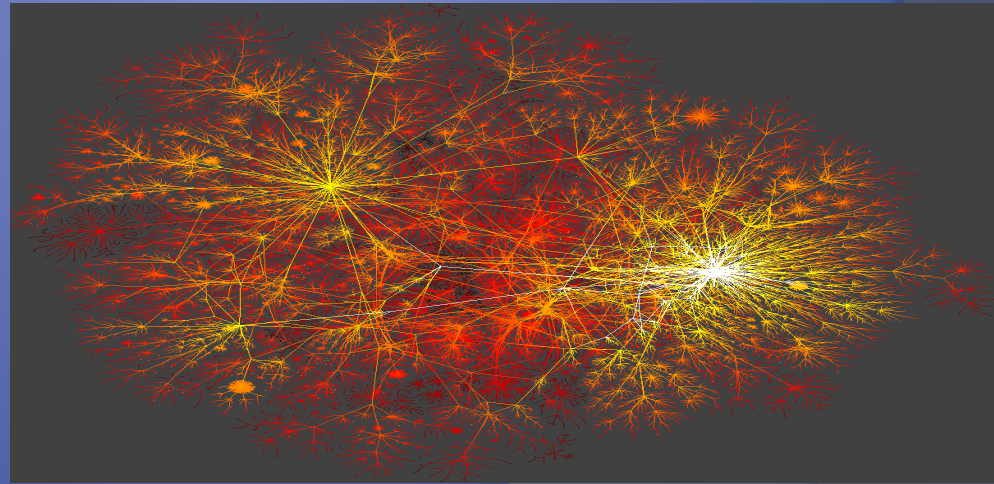
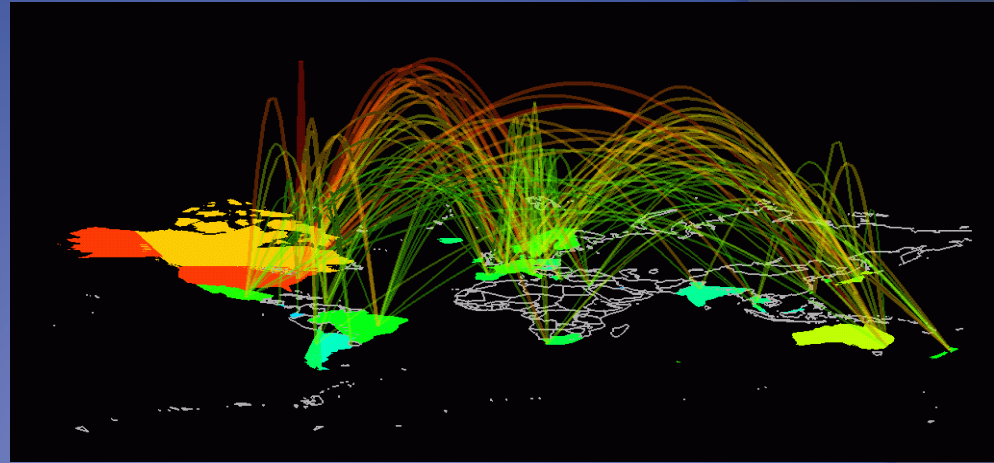
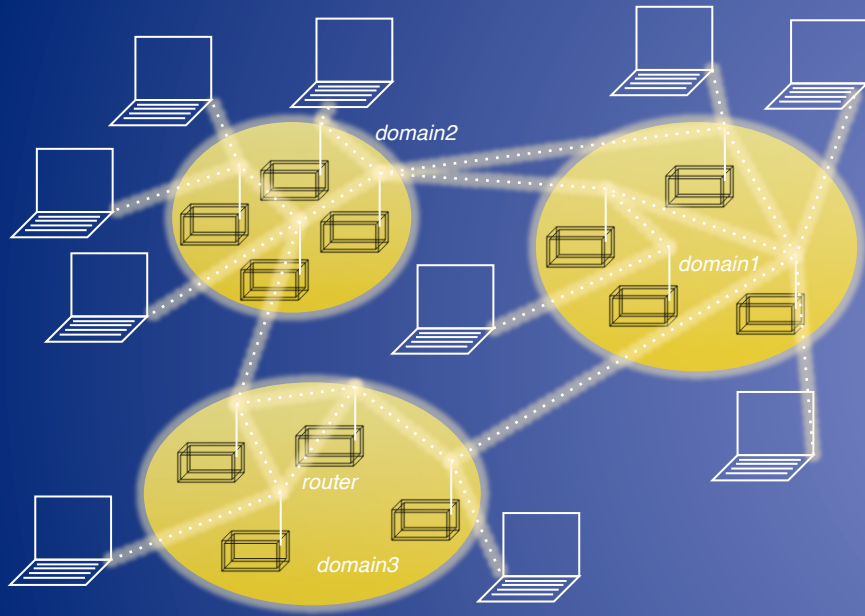
Collaborations

Financial

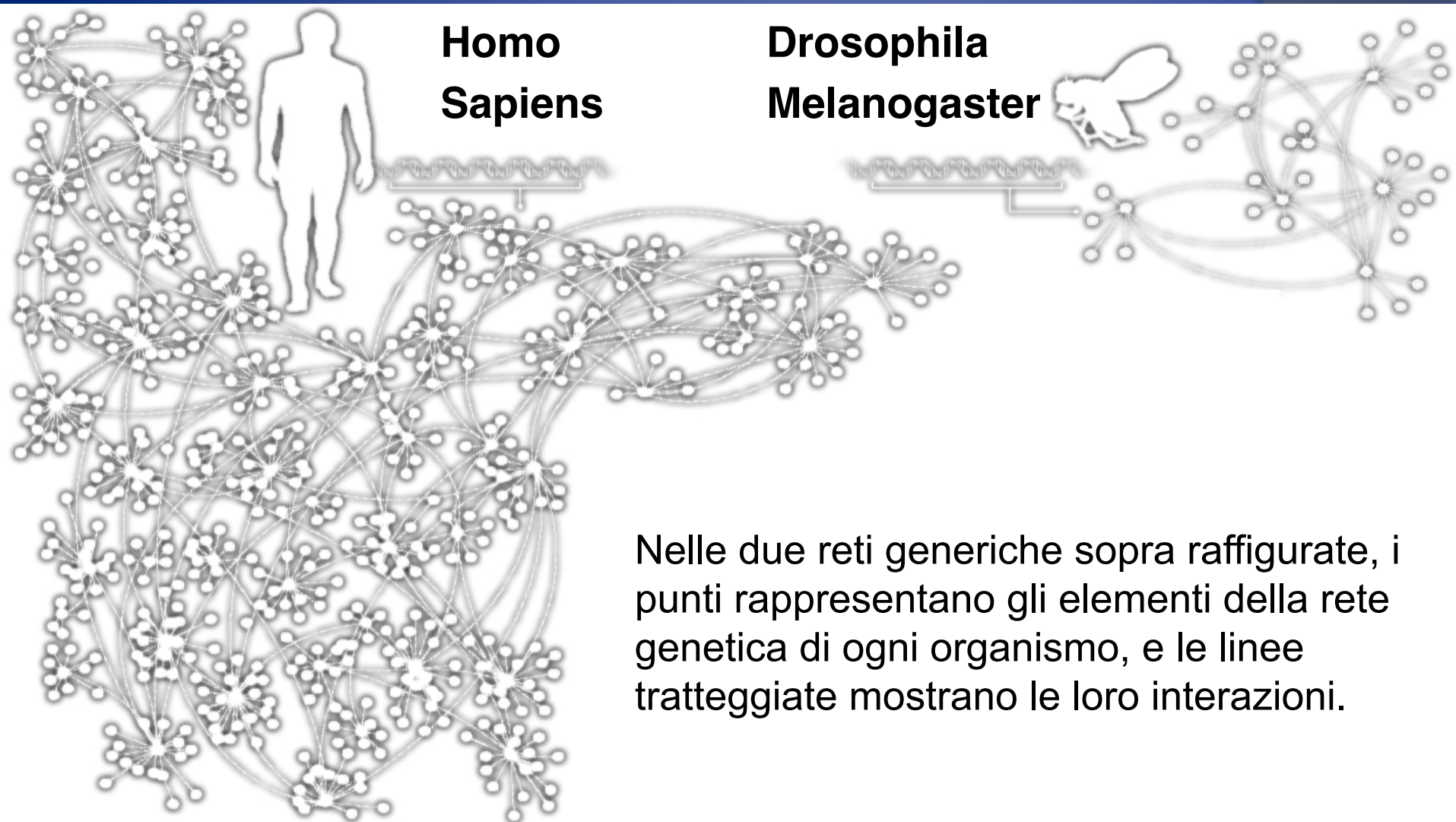
R&D



INTERNET



GENI UMANI



Homo
Sapiens

Drosophila
Melanogaster



I Sistemi Complessi

Sono costituiti da molti **elementi** non identici collegati tra loro da diverse **interazioni**.



NETWORK

LE REGOLE DEI NETWORK

Dietro ogni sistema studiato con la teoria della complessità c'è un diagramma di cablaggio intricato, o di una rete, che definisce le interazioni tra i componenti.

Non è possibile capire un sistema complesso senza prima mapparlo, in modo tale da poter studiare la rete di interazioni che lo costituisce.

LE CARATTERISTICHE DELLA SCIENZA DEI NETWORK

Interdisciplinare

Empirica

Quantitativa e Matematica

Computazionale

Interdisciplinare

Empirica

Quantitativa e Matematica

Computazionale

Interdisciplinare

Empirica

Quantitativa e Matematica

Computazionale

Interdisciplinare

Empirica

Quantitativa e Matematica

Computazionale

L'IMPATTO DELLA SCIENZA DEI NETWORK

IMPATTO ECONOMICO

Google

Market Cap(2010 Jan 1):
\$189 billion

Cisco Systems

networking gear Market
cap (Jan 1, 2010):
\$112 billion

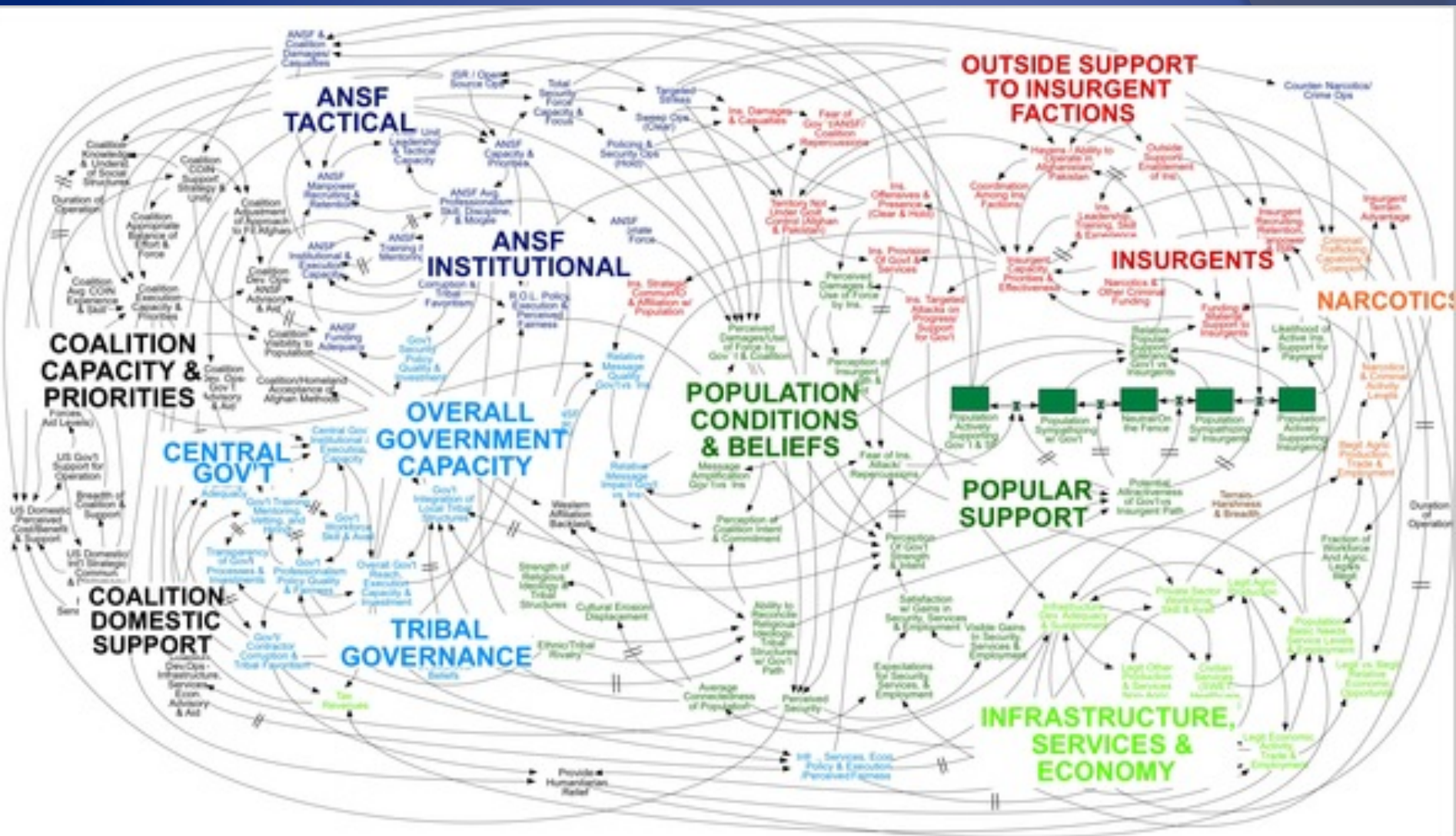
Facebook

market cap:
\$50 billion

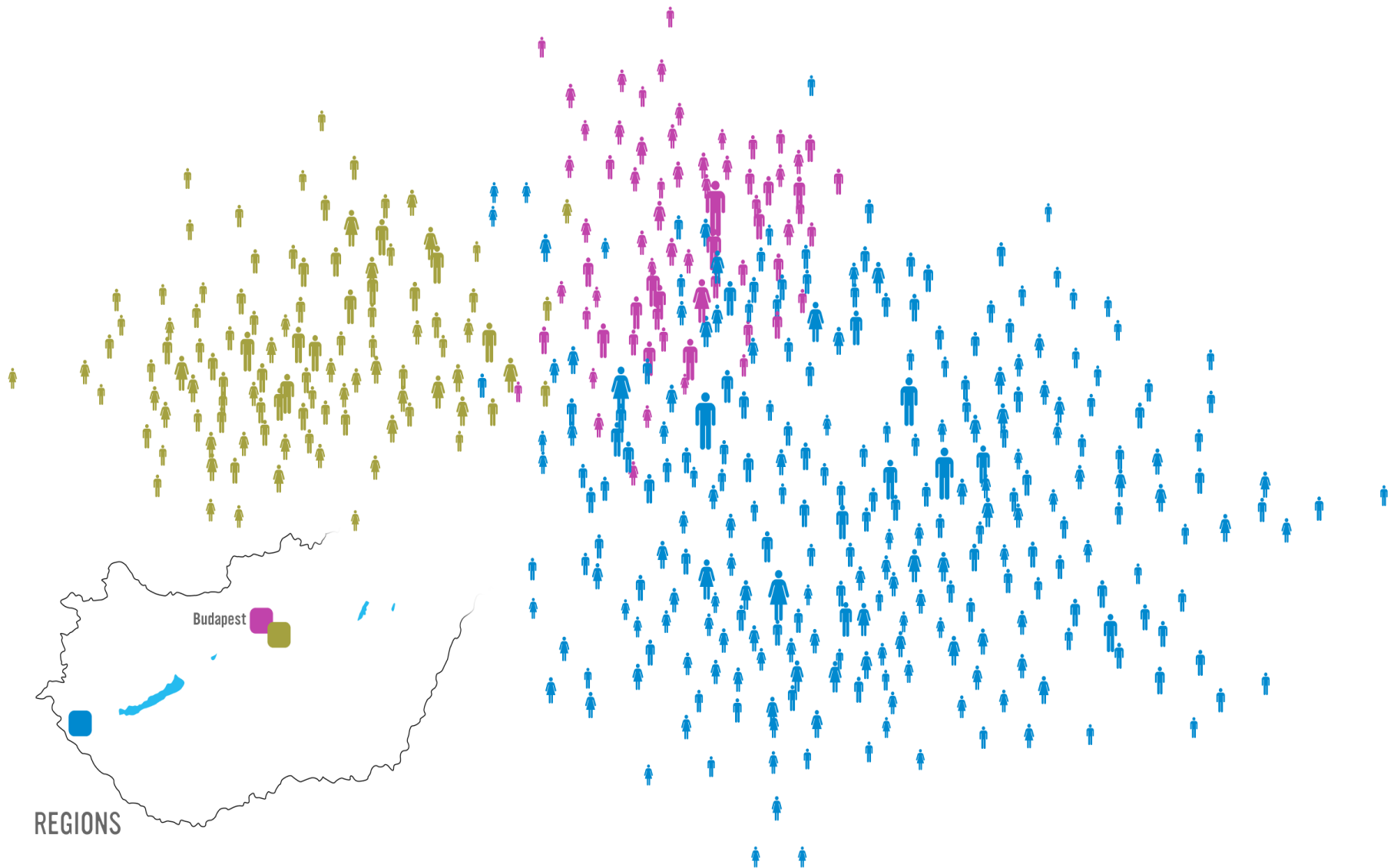
www.bizjournals.com/austin/news/2010/11/15/facebooks... - Cached



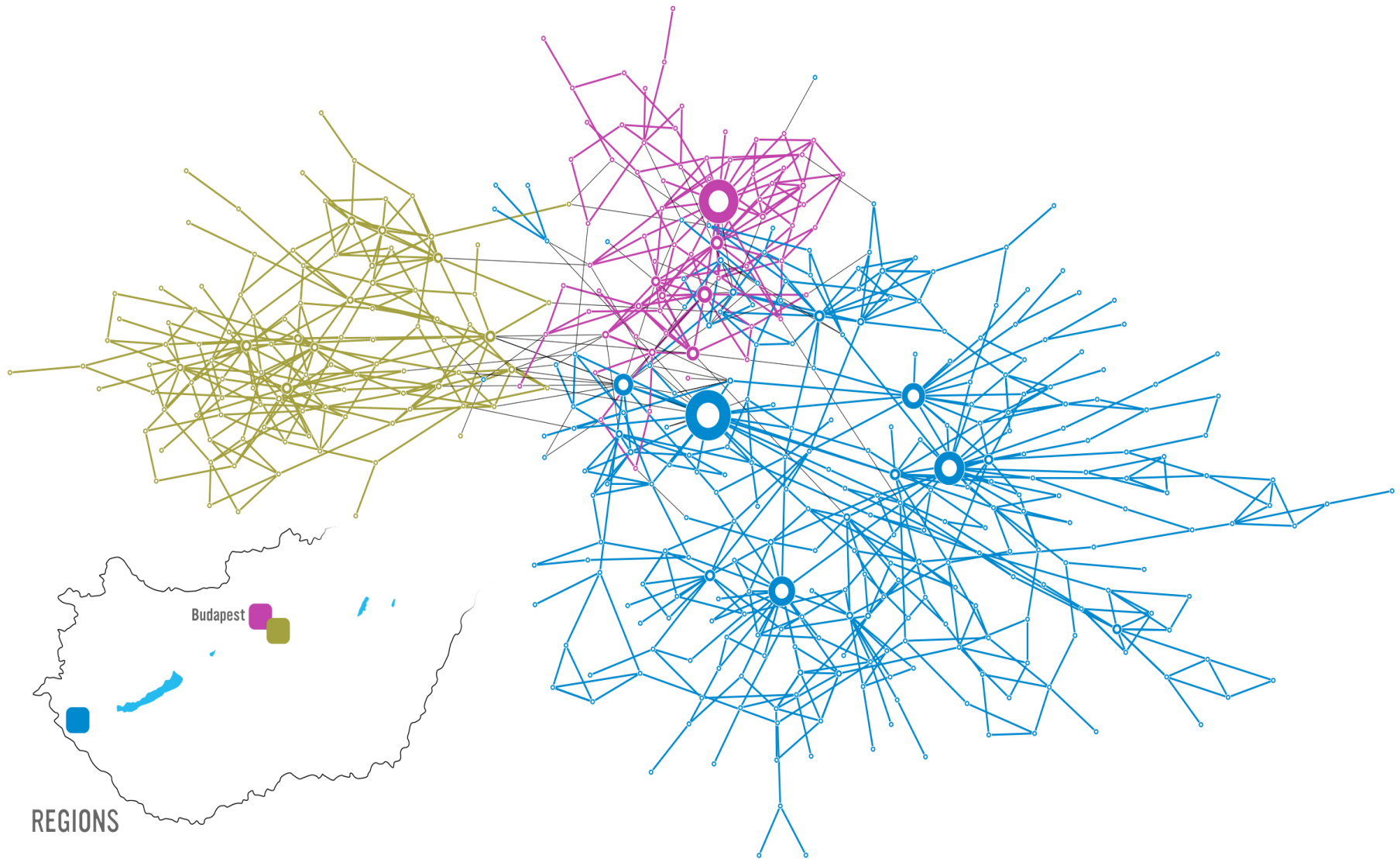
Reti caratterizzanti un impegno militare



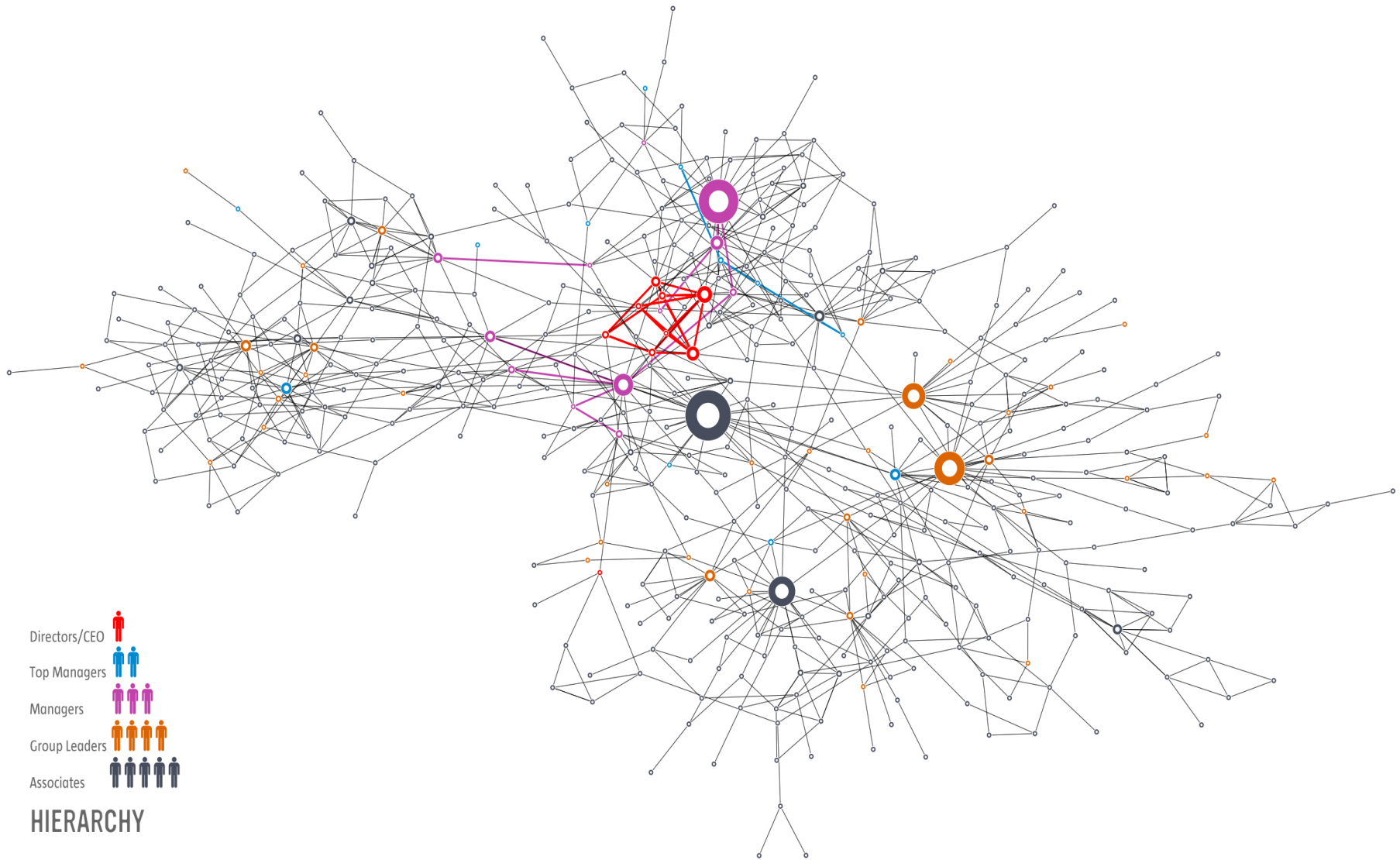
GESTIONE AZIENDE



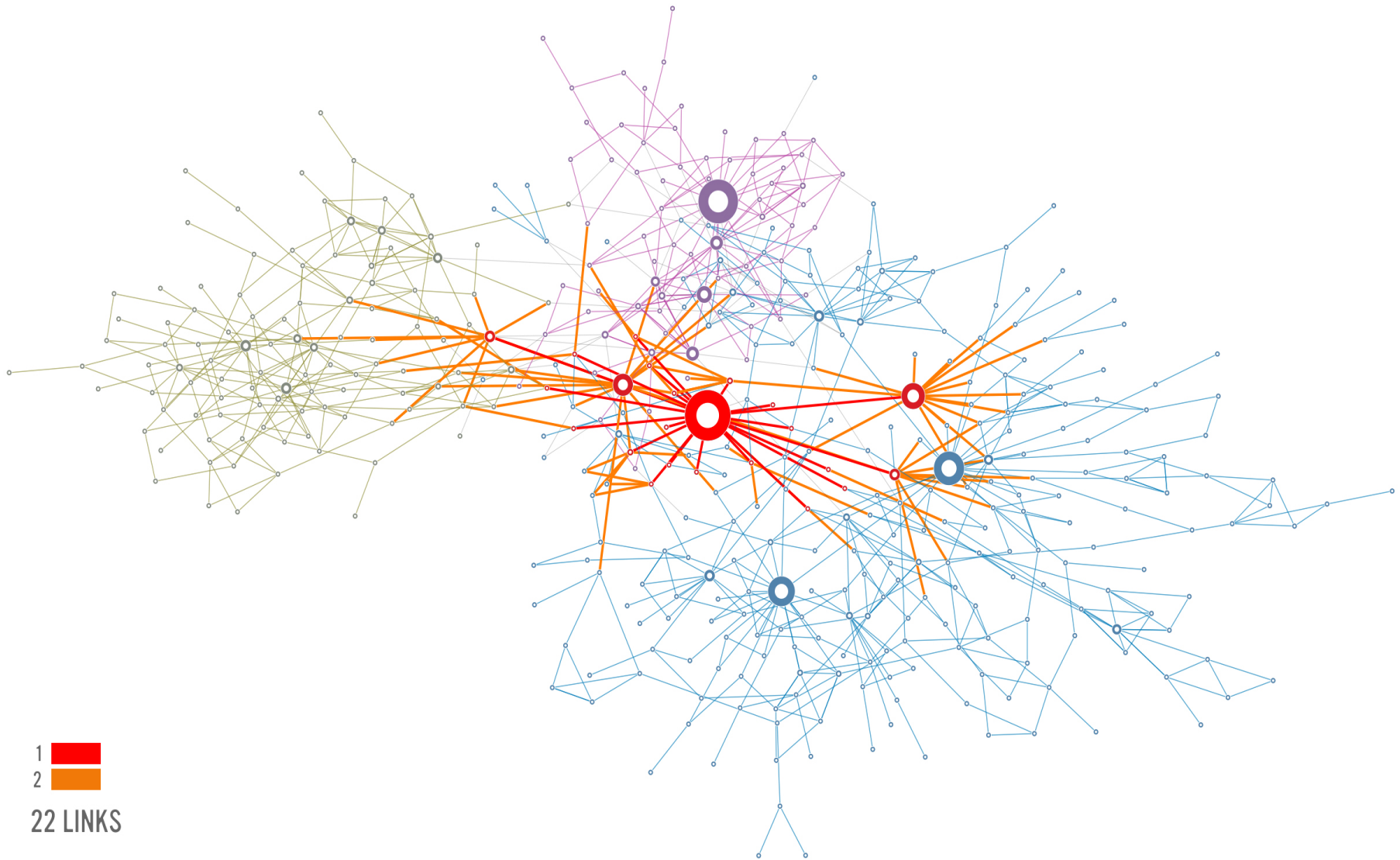
GESTIONE AZIENDE



GESTIONE AZIENDE

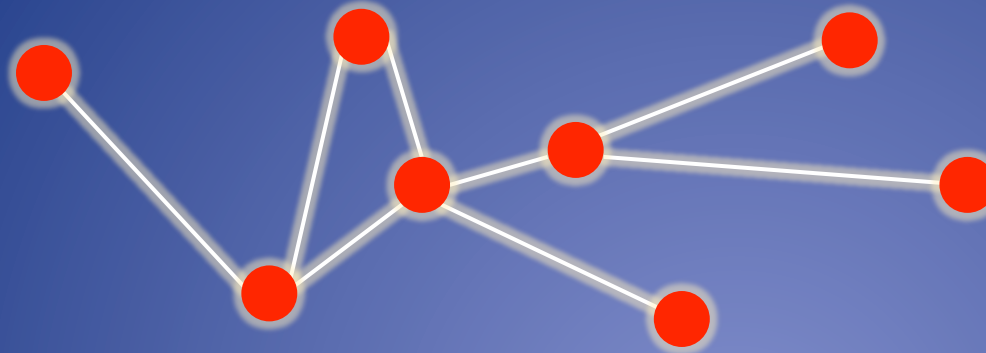


GESTIONE AZIENDE



NETWORK & GRAFI

ELEMENTI DI UN SISTEMA COMPLESSO



- ◆ **Elementi:** nodi, vertice $\rightarrow N$
- ◆ **Interazioni:** links, spigoli $\rightarrow L$
- ◆ **Sistemi:** network, grafi $\rightarrow (N,L)$

NETWORK O GRAFI ?

I network sono in prevalenza riferiti a sistemi reali

- www,
- social network
- metabolic network.

Linguaggio: (Network, node, link)

I grafi: rappresentazione matematica dei Network

- web graph,
- social graph (a Facebook term)

Linguaggio: (Graph, vertex, edge)

Nonostante l'esistenza di questa differenza I due termini vengono usati praticamente come sinonimi

SCELTA DELL'OPPORTUNA RAPPRESENTAZIONE

La scelta di come rappresentare la rete determina la nostra capacità di utilizzare la teoria dei network nel modo corretto permettendoci di ottenere il maggior numero di informazioni possibili.

In alcuni casi la rappresentazione può essere unica, ma per la maggior parte dei sistemi reali le rappresentazioni possibili sono diverse e per ognuna di esse si ottengono delle informazioni differenti.

Per esempio il modo in cui decidiamo di assegnare il legame tra individui determina la natura delle domande che possiamo andare a studiare.

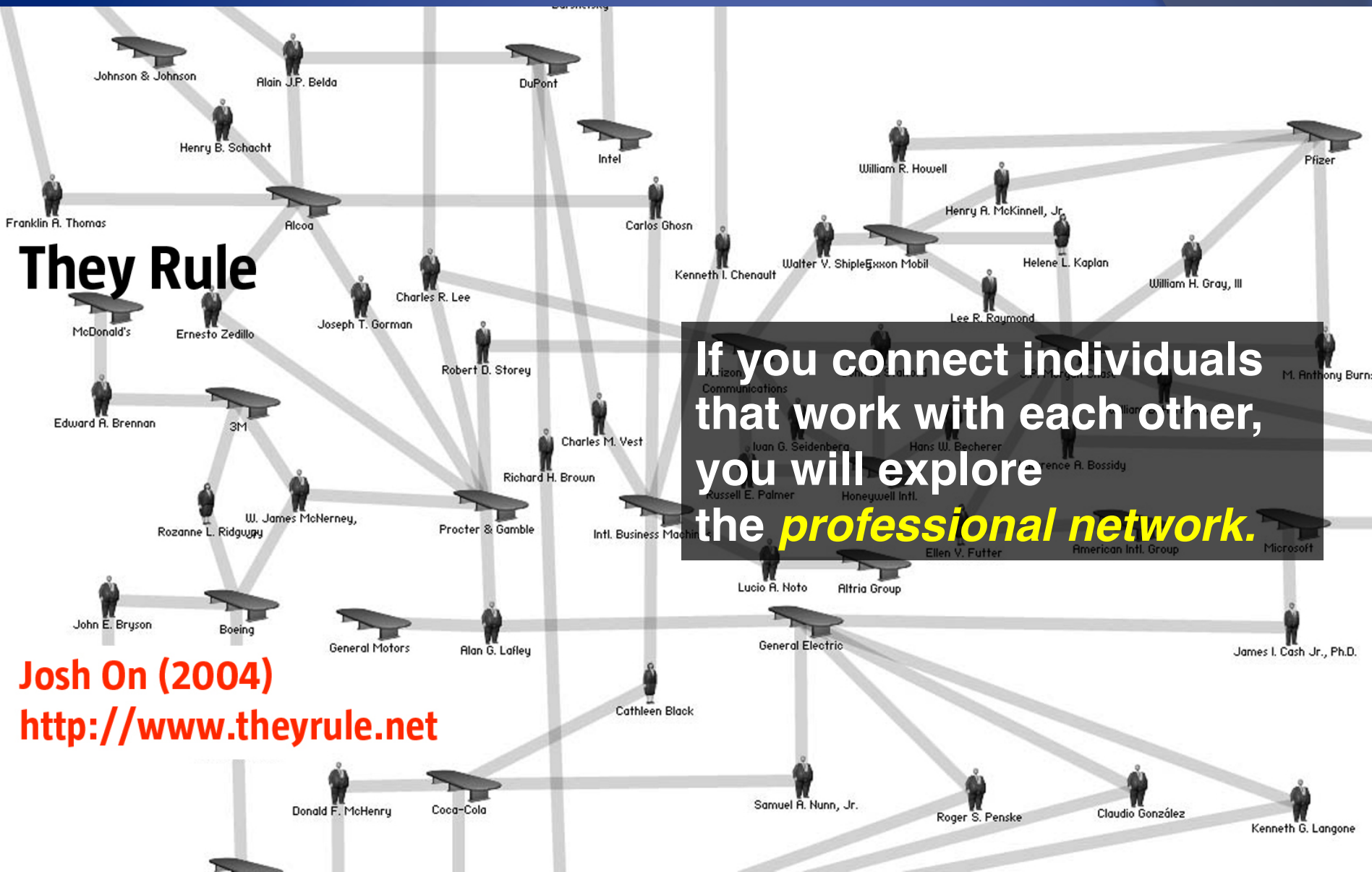
SCELTA DELL'OPPORTUNA RAPPRESENTAZIONE

They Rule

If you connect individuals
that work with each other,
you will explore
the **professional network**.

Josh On (2004)

<http://www.theyrule.net>



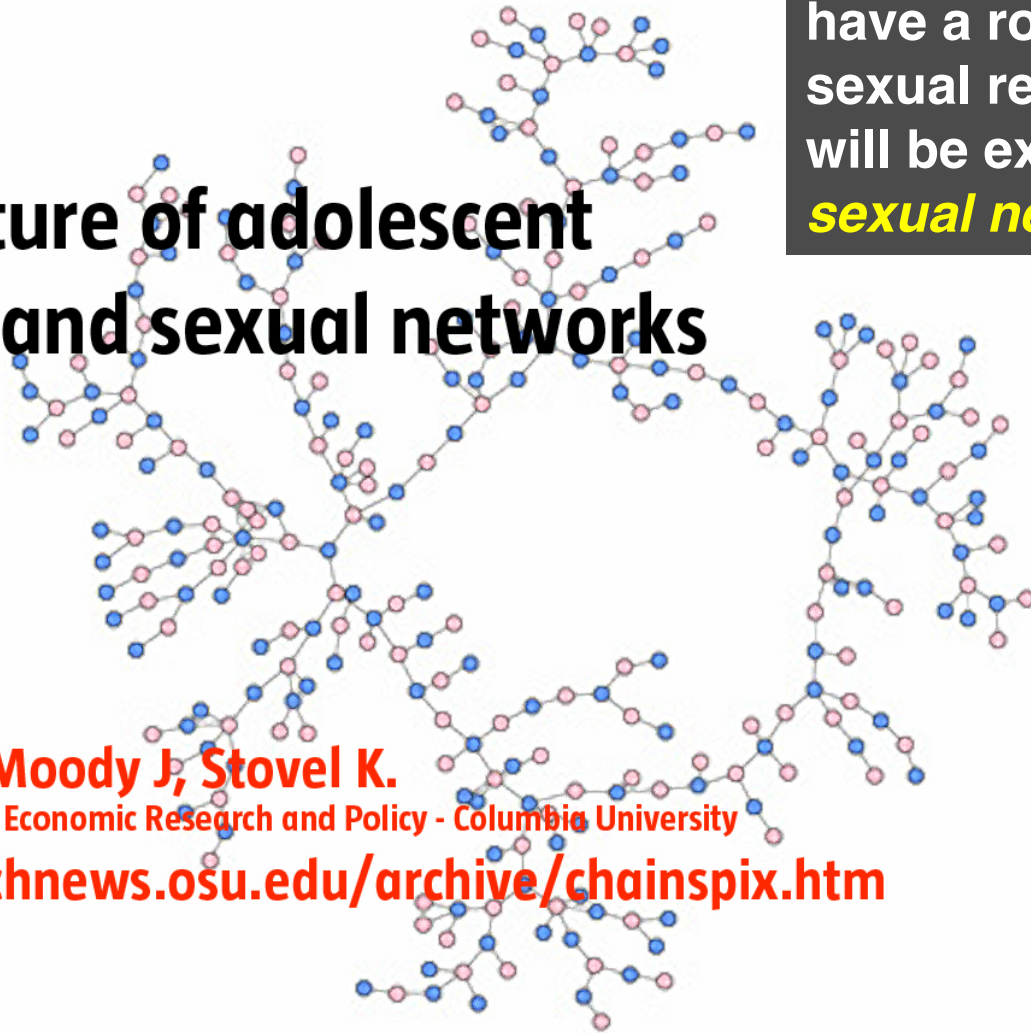
The structure of adolescent romantic and sexual networks

If you connect those that have a romantic and sexual relationship, you will be exploring the *sexual networks*.

Bearman PS, Moody J, Stovel K.

Institute for Social and Economic Research and Policy - Columbia University

<http://researchnews.osu.edu/archive/chainspix.htm>



Possiamo anche decidere di connettere gli individui basandosi sul loro nome (*ad esempio collegare tutti i Mario tra di loro*), però in questo caso cosa stiamo studiando?

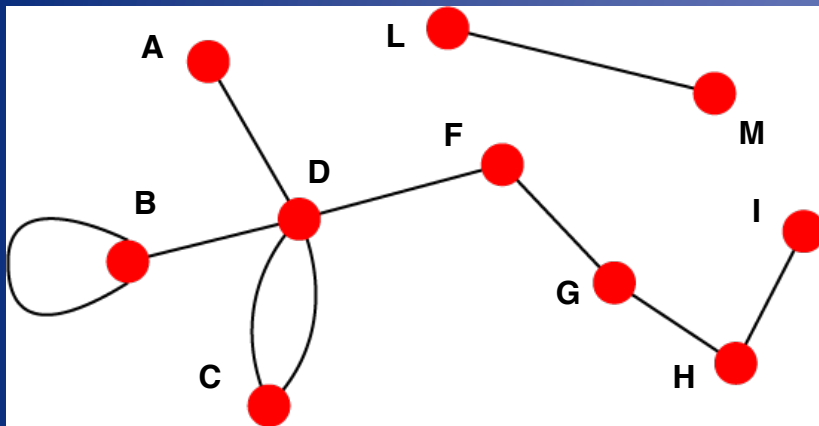
Questo è sicuramente un network, tuttavia.....

NETWORK INDIRETTI VS. NETWORK DIRETTI

INDIRETTI

Link: Indiretti (*sono simmetrici*)

Grafi:



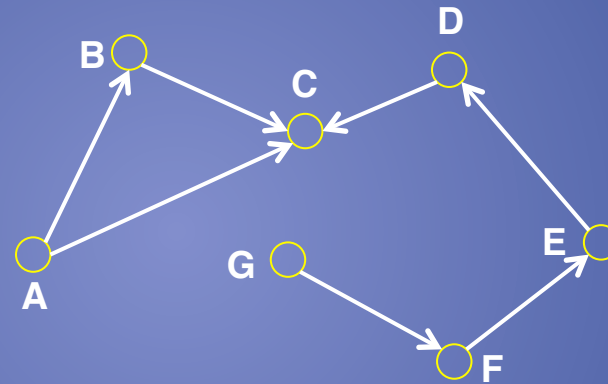
Esempi di link indiretti:

Coauthorship links
Actor network
Protein interactions

DIRETTI

Link: Diretti (*archi*).

Digraph = grafi diritti:



Un link indiretto è la sovrapposizione di due link diretti opposti.

Esempi di link diretti:

URLs on the www
Phone calls
Metabolic reactions

REFERENCE NETWORKS

NETWORK	NODES	LINKS	DIRECTED UNDIRECTED	N	L
Internet	Routers	Internet connections	Undirected	192,244	609,066
WWW	Webpages	Links	Directed	325,729	1,497,134
Power Grid	Power plants, transformers	Cables	Undirected	4,941	6,594
Mobile Phone Calls	Subscribers	Calls	Directed	36,595	91,826
Email	Email addresses	Emails	Directed	57,194	103,731
Science Collaboration	Scientists	Co-authorship	Undirected	23,133	93,439
Actor Network	Actors	Co-acting	Undirected	702,388	29,397,908
Citation Network	Paper	Citations	Directed	449,673	4,689,479
E. Coli Metabolism	Metabolites	Chemical reactions	Directed	1,039	5,802
Protein Interactions	Proteins	Binding interactions	Undirected	2,018	2,930

IL MACHINE LEARNING

ESEMPI DI ALGORITMI DI APPRENDIMENTO DATI



Cerca con Google

Mi sento fortunato

ESEMPI DI ALGORITMI DI APPRENDIMENTO DATI



amazon.it

traduttore - Cerca con Google

Amazon.it libri in italiano: narrativa, romanzi, attualità, libri per ragazzi, twiilight saga, Harry Potter e tanto altro...

In primo piano

- Libri Civil War: Scopri la selezione
- LA STAMPA: Nuovo percorso di lettura
- Adelphi al -25%: Oltre 1.000 titoli
- Tea Sisters al -25%: Scopri tutte le avventure
- Sperling e Pickwick al -25%: Il benessere non finisce mai

Consigliati per te

Visualizza altri consigli d'acquisto

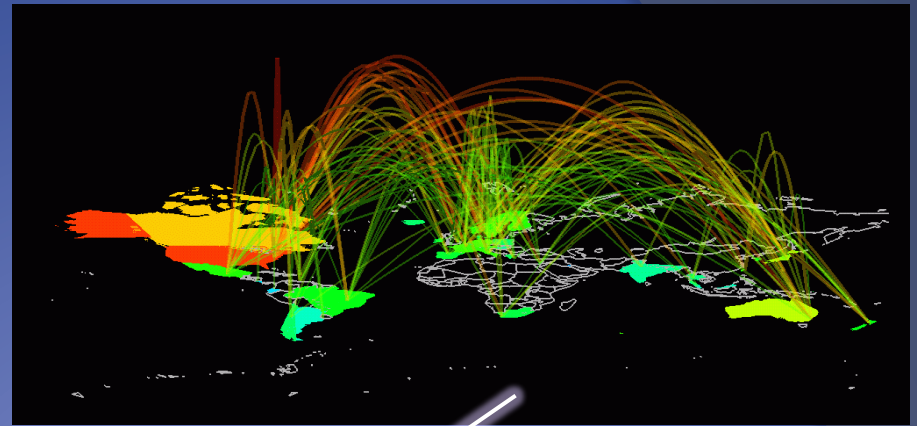
Book cover	Title	Author	Format	Rating	Price	Prime	Recommendation
	La fisica del diavolo	Maxwell...	Copertina rigida	★★★★☆ (46)	EUR 17,00	✓	Migliora i nostri consigli d'acquisto per te
	Alice nel paese dei quant	Robert Gilmore, P. D. Napolitani	Copertina flessibile	★★★★★ (7)	EUR 17,00	✓	Migliora i nostri consigli d'acquisto per te
	Il libro che cambierà per sempre le nostre idee sulla Bibbia	Mauro Biglino	Copertina flessibile	★★★★★ (49)	EUR 14,37	✓	Migliora i nostri consigli d'acquisto per te

Novità e prossime uscite consigliate dagli editori

Tutte le novità e prossime uscite consigliate dagli editori

- Svetlana Aleksievich: Gli ultimi testimoni
- Lilli Gruber: Prigionieri dell'Islam
- Cristina Pignatelli: Guida

L'algoritmo di Amazon decide, più di qualsiasi altra persona, quali libri vengono letti nel modo.

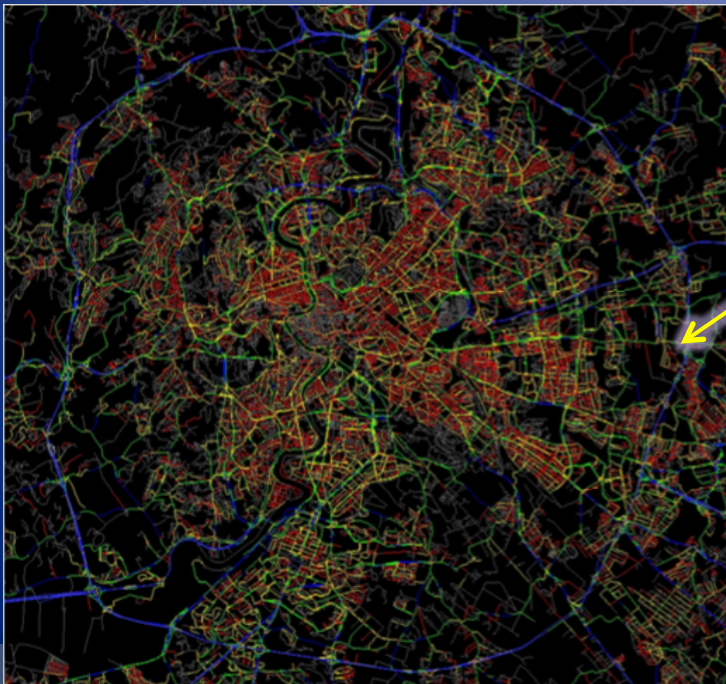


Gli algoritmi dell'NSA decidono se siete dei potenziali terroristi.

I modelli di scelta dei titoli azionari governano l'economia più di qualunque essere umano, o quasi.



Black Boxes



tracce digitali

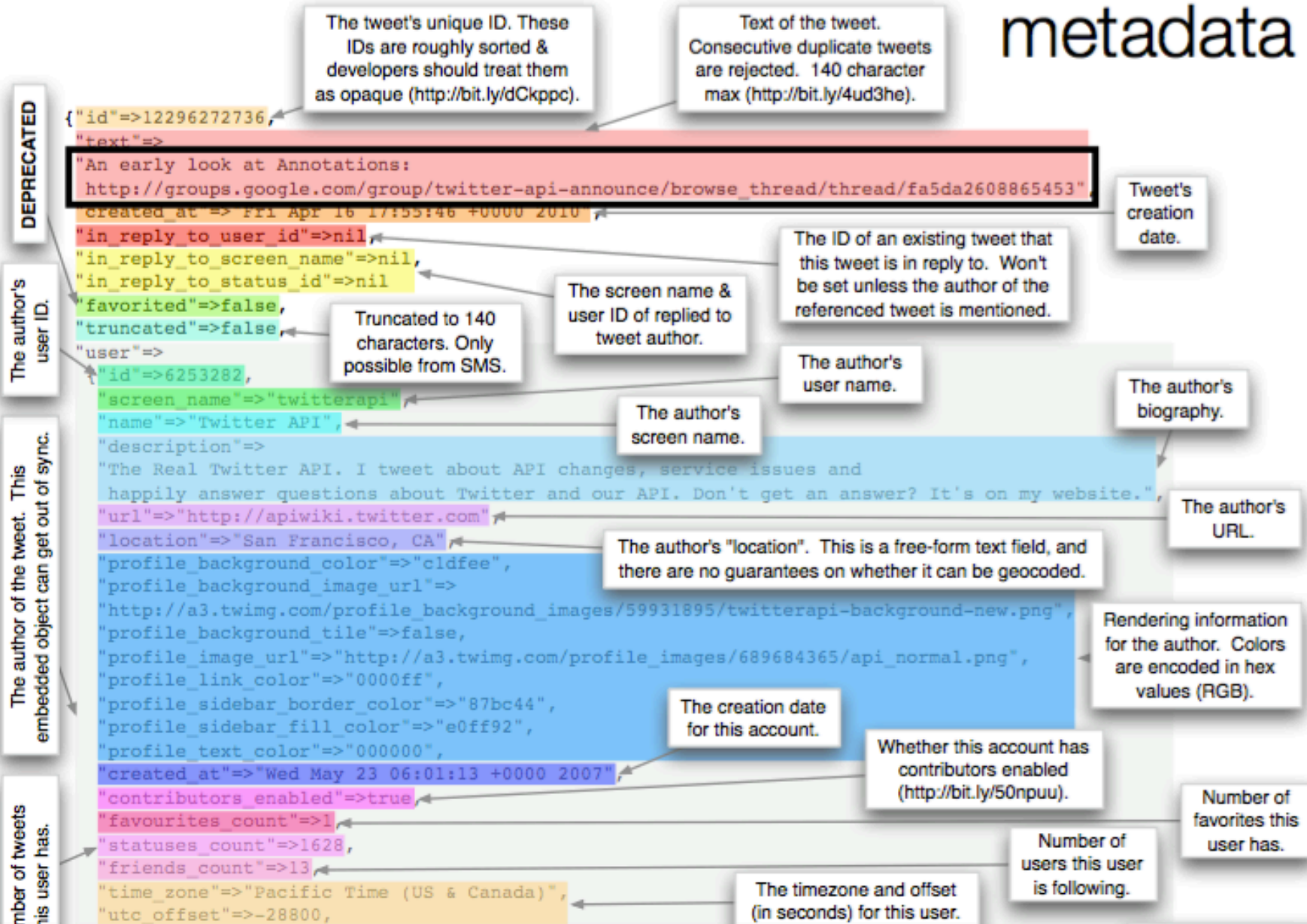
The background of the slide is a photograph of a person walking on a large sand dune under a clear blue sky. A series of footprints are visible in the sand, leading from the foreground towards the person in the distance. A red arrow points from the text 'tracce digitali' down to the list of challenges, and a blue arrow points from the same text towards the text on the right.

Quadro storico
Orizzonte temporale
Riproducibilità limitata
Contesto limitato
Sfide per la protezione
dei dati

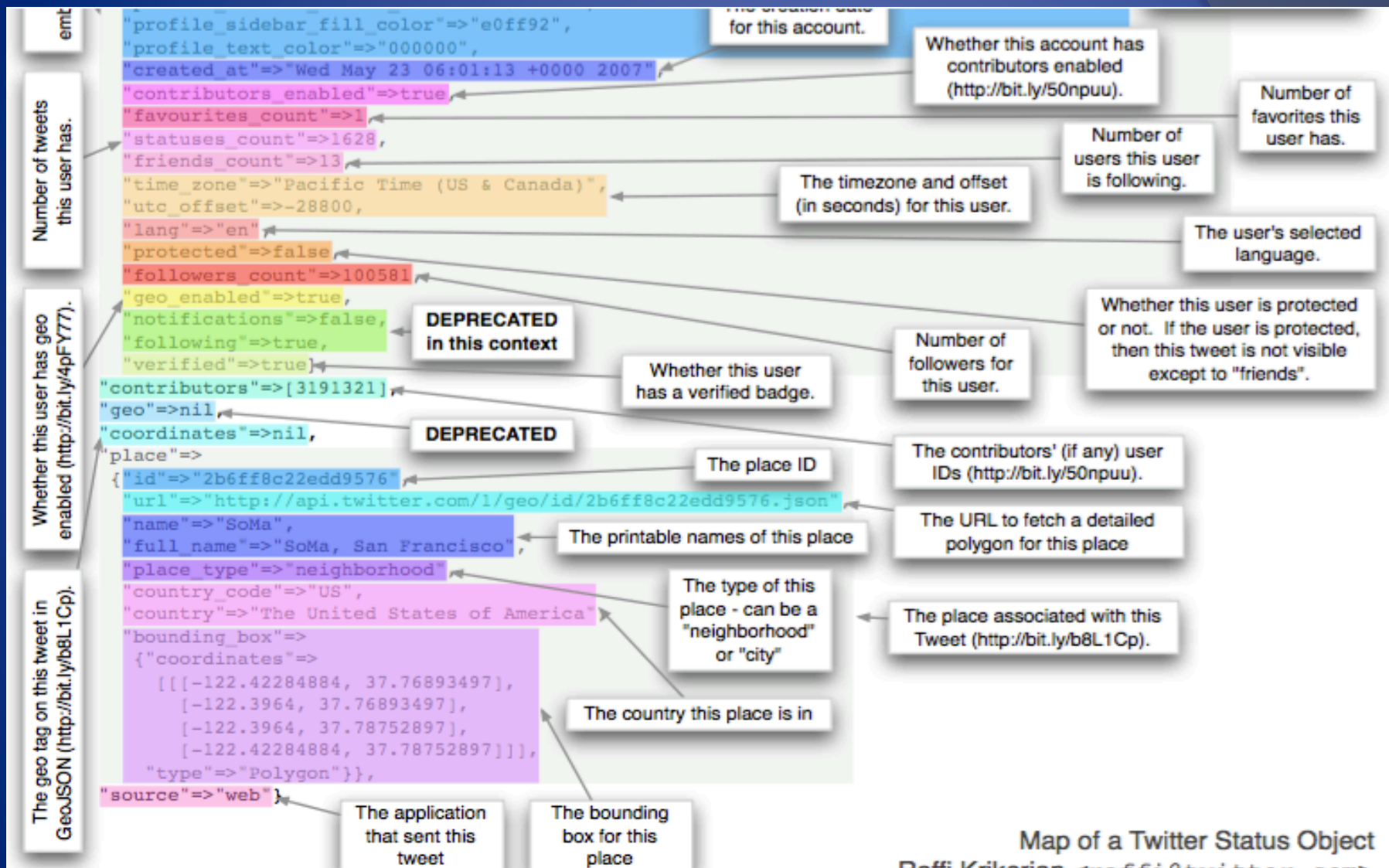
Scoperta la possibilità di utilizzare analisi di segnali digitali in tantissimi campi, con la possibilità di creare modelli ad ampia copertura.

INFORMAZIONI IMPLICITE

metadata



INFORMAZIONI IMPLICITE



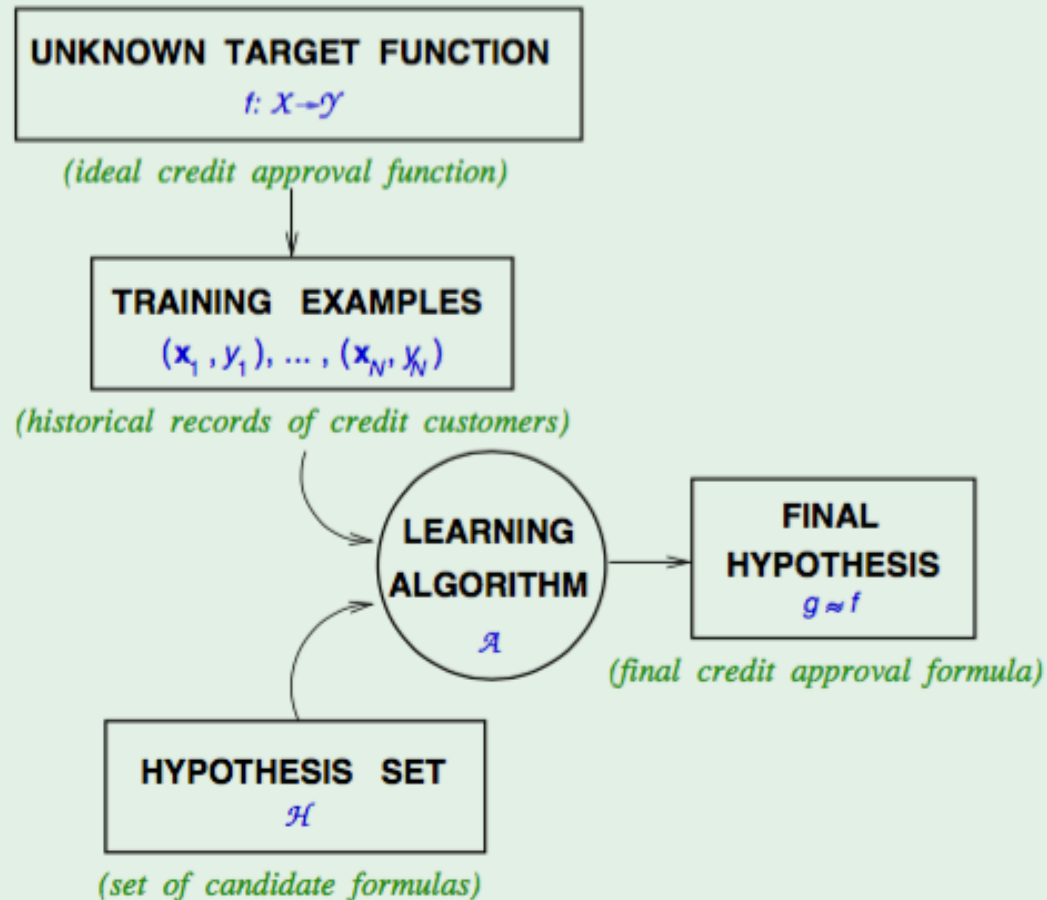
Map of a Twitter Status Object
 Raffi Krikorian <raffi@twitter.com>
 18 April 2010

Components of learning

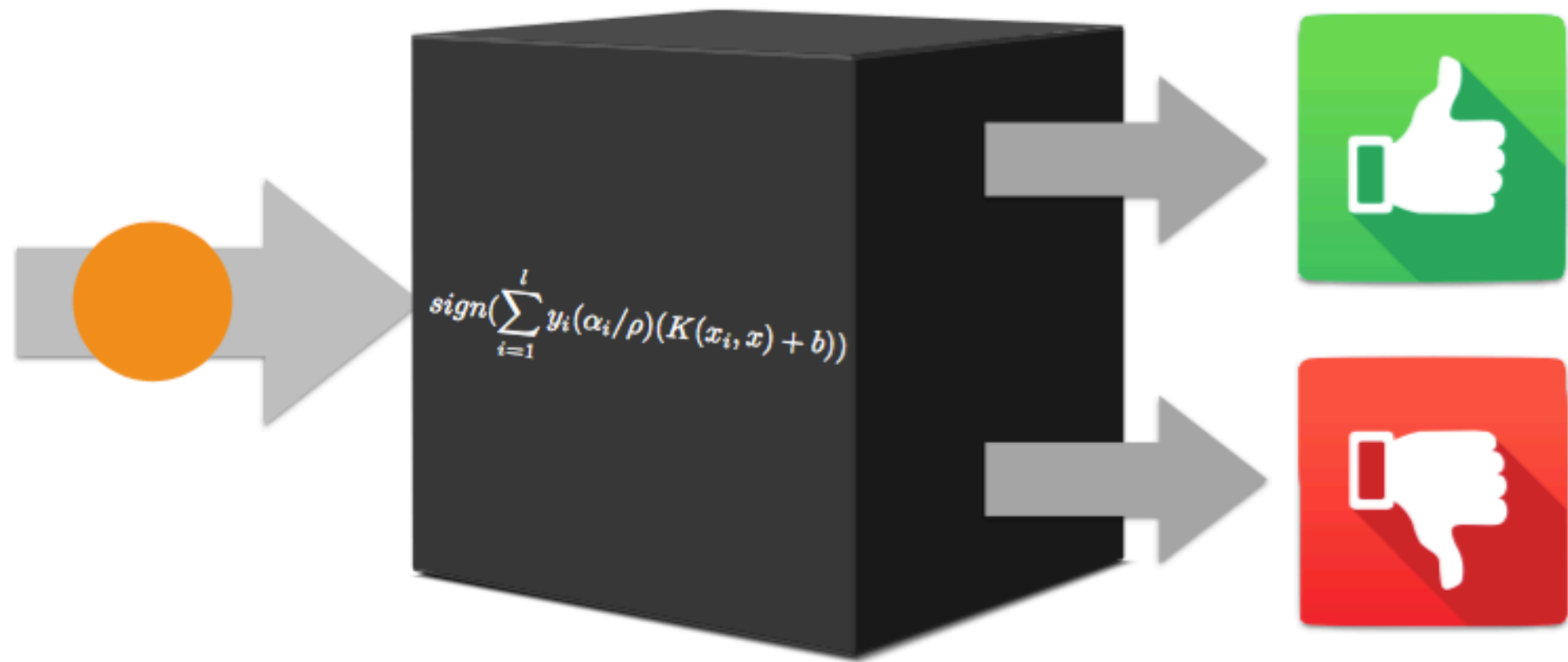
Formalization:

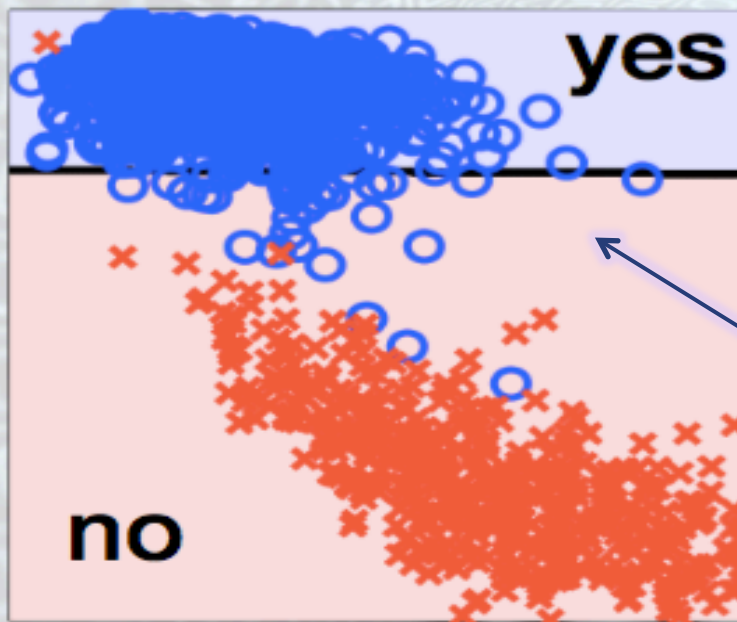
- Input: $\mathbf{x}$ (*customer application*)
 - Output: y (*good/bad customer?*)
 - Target function: $f : \mathcal{X} \rightarrow \mathcal{Y}$ (*ideal credit approval formula*)
 - Data: $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_N, y_N)$ (*historical records*)
- ↓ ↓ ↓
- Hypothesis: $g : \mathcal{X} \rightarrow \mathcal{Y}$ (*formula to be used*)

Struttura di funzionameto dell'Algoritmo

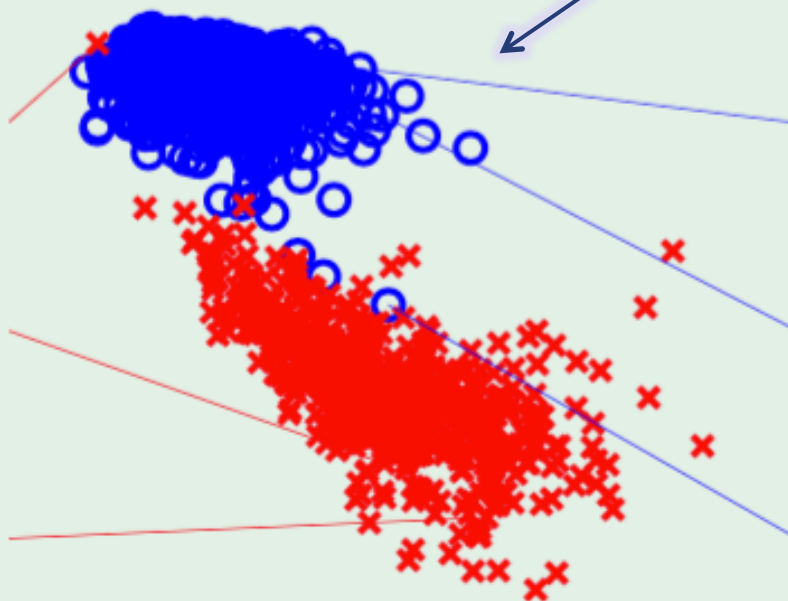
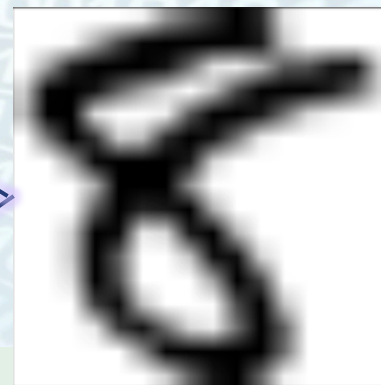


DECISIONI DELL'ALGORITMO





decision boundary



SEGMENTAZIONE & RAGGRUPPAMENTO



SEGMENTAZIONE & RAGGRUPPAMENTO

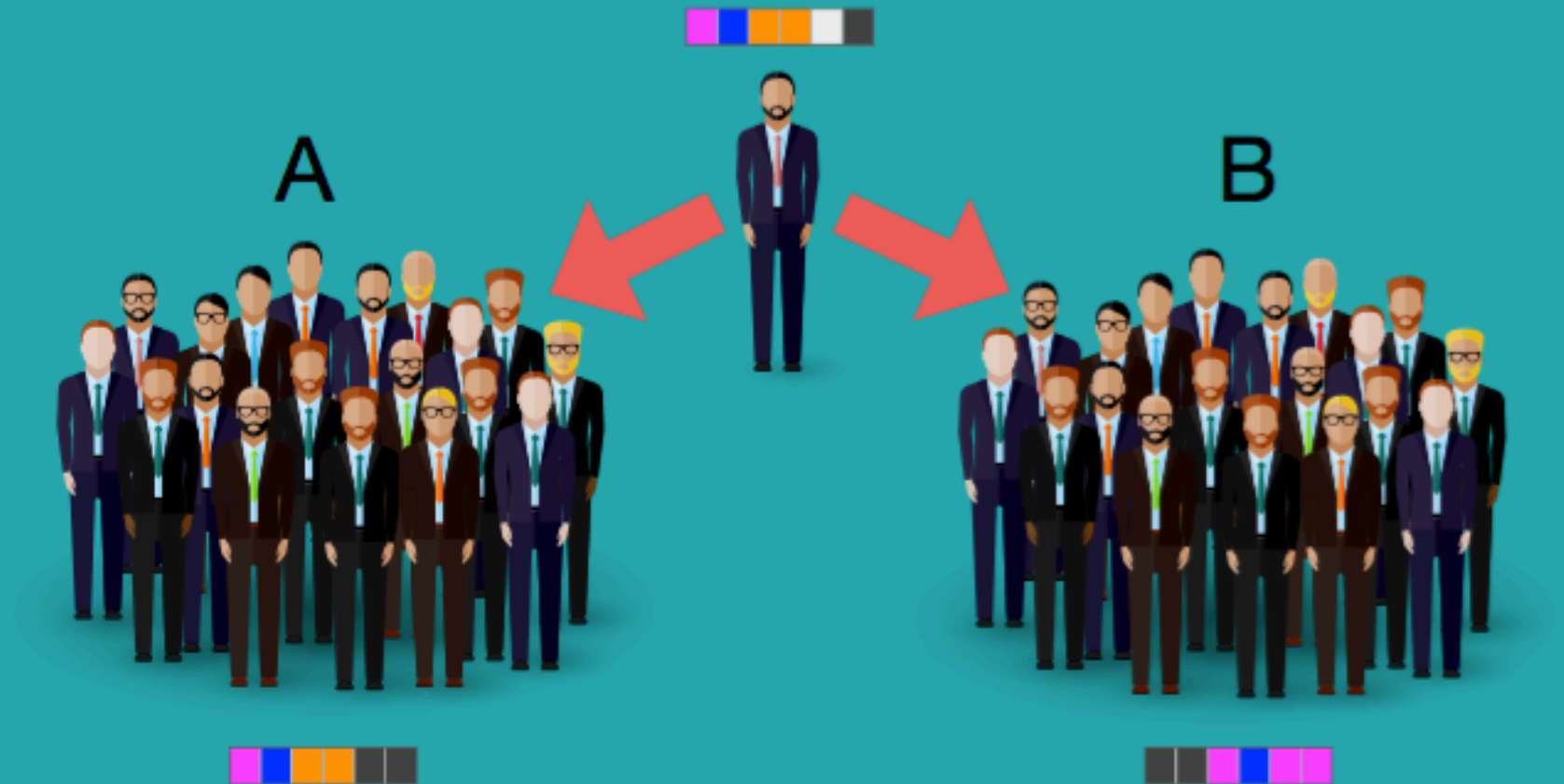
A



B



SEGMENTAZIONE & RAGGRUPPAMENTO



INTERVENTI MIRATI



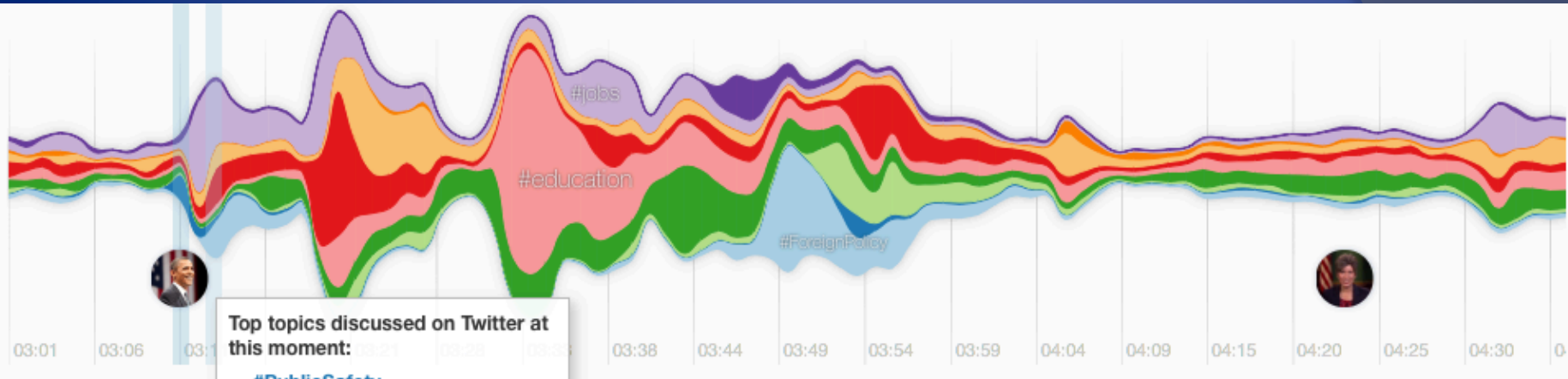
A



B



RISPOSTE A DISCORSI DI PROPAGANDA POLITICA

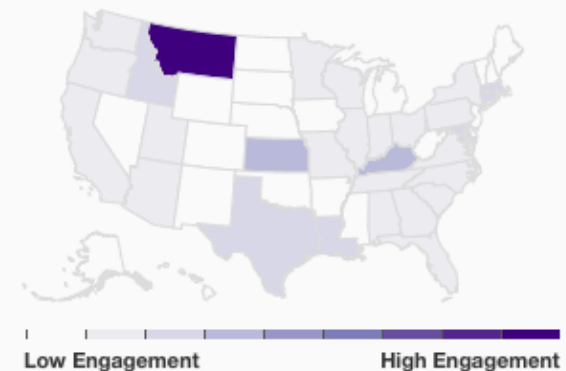


now lower than it was before the financial crisis. More of our kids are graduating than ever before; more of our people are insured than ever before; we are as free from the grip of foreign oil as we've been in almost 30 years.

Tonight, for the first time since 9/11, our combat mission in Afghanistan is over. Six years ago, nearly 180,000 American troops served in Iraq and Afghanistan. Today, fewer than 15,000 remain. And we salute the courage and sacrifice of every man and woman in this 9/11 Generation who has served to keep us safe. We are humbled and grateful for your service.

America, for all that we've endured; for all the grit and hard work required to come back; for all the tasks that lie ahead, know this:

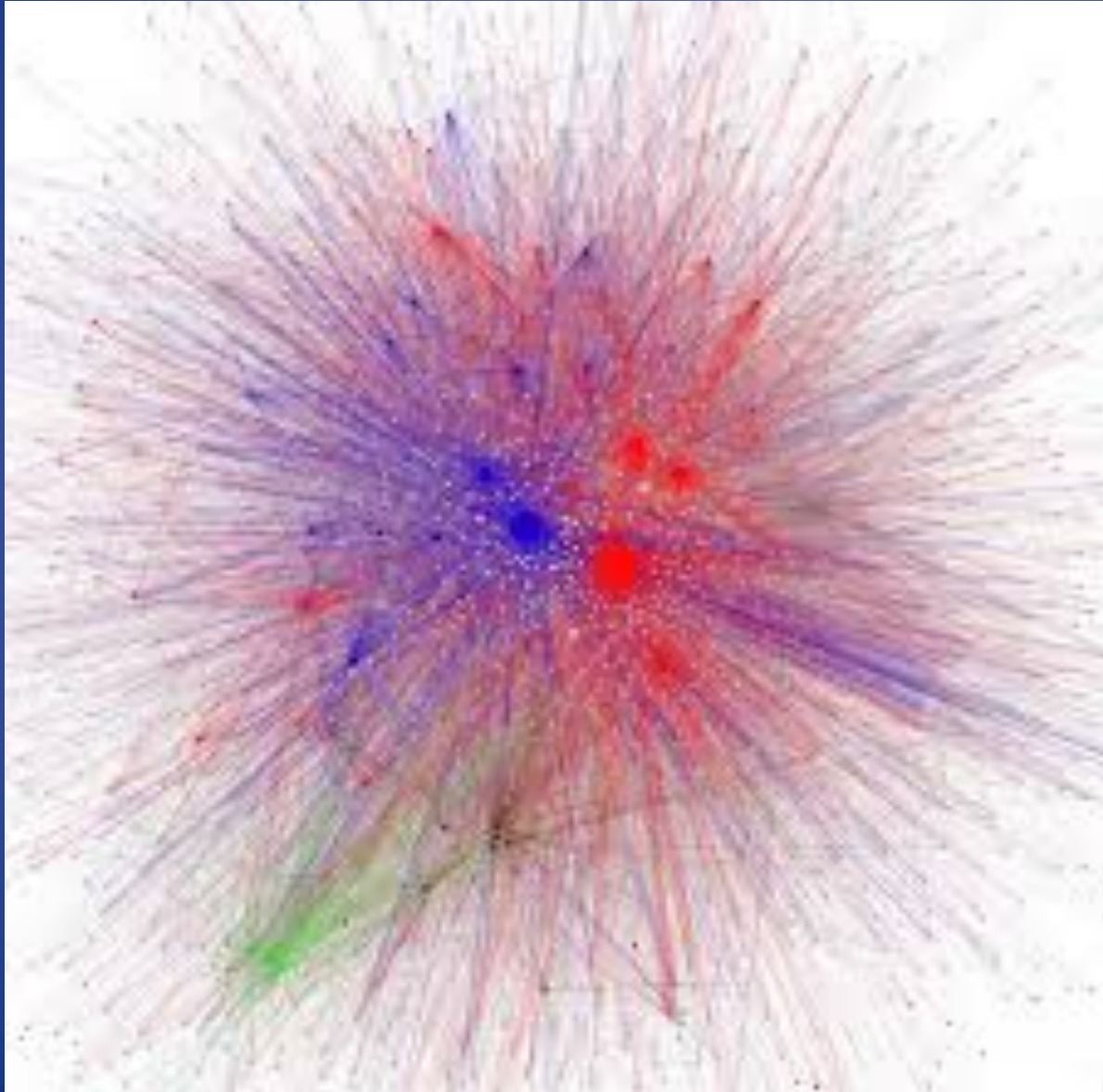
Real-time engagement distribution on Twitter for this paragraph



Map for #jobs



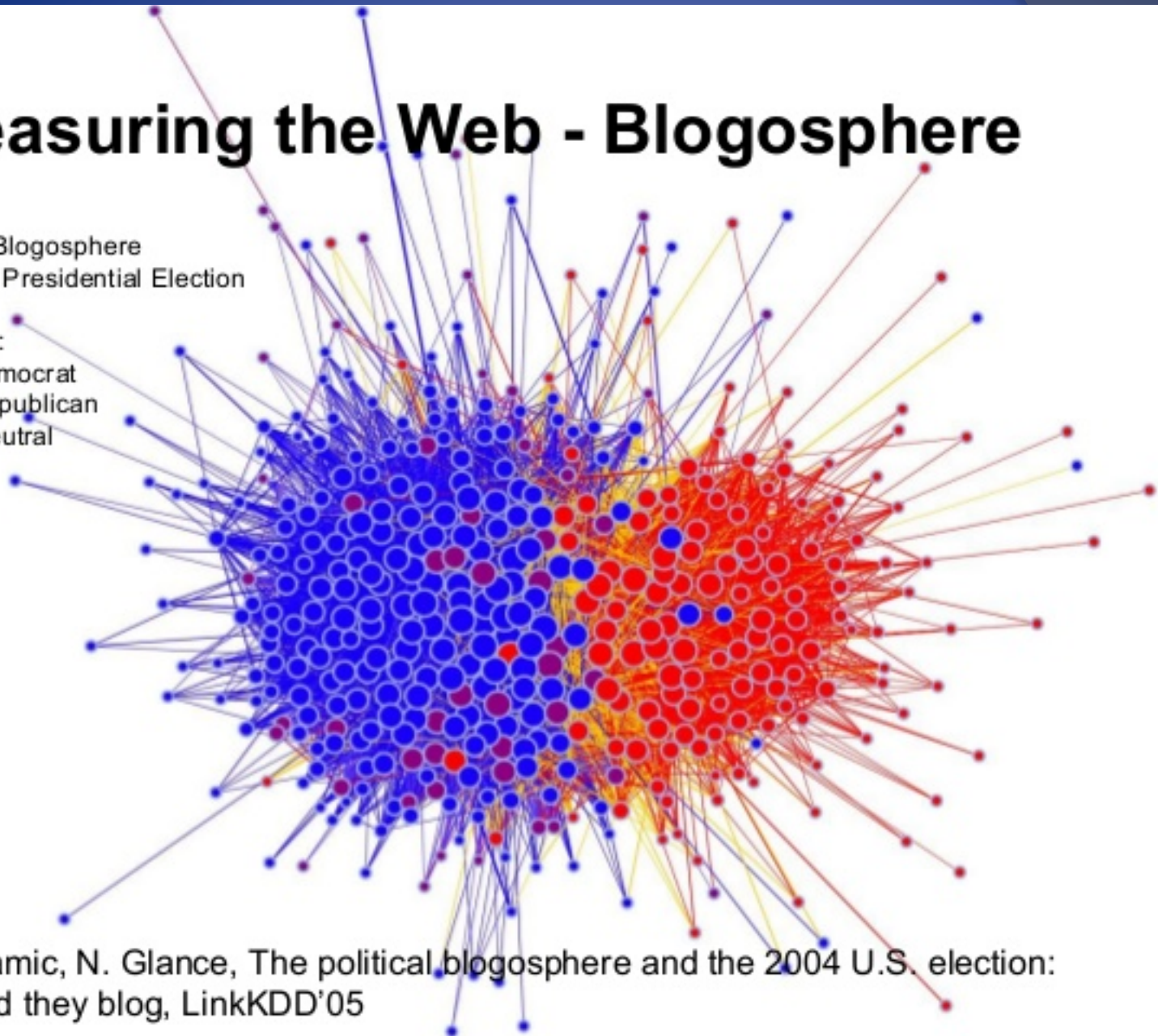
RETE DISTRIBUZIONE POLITICA DAI SOCIAL-NETWORK



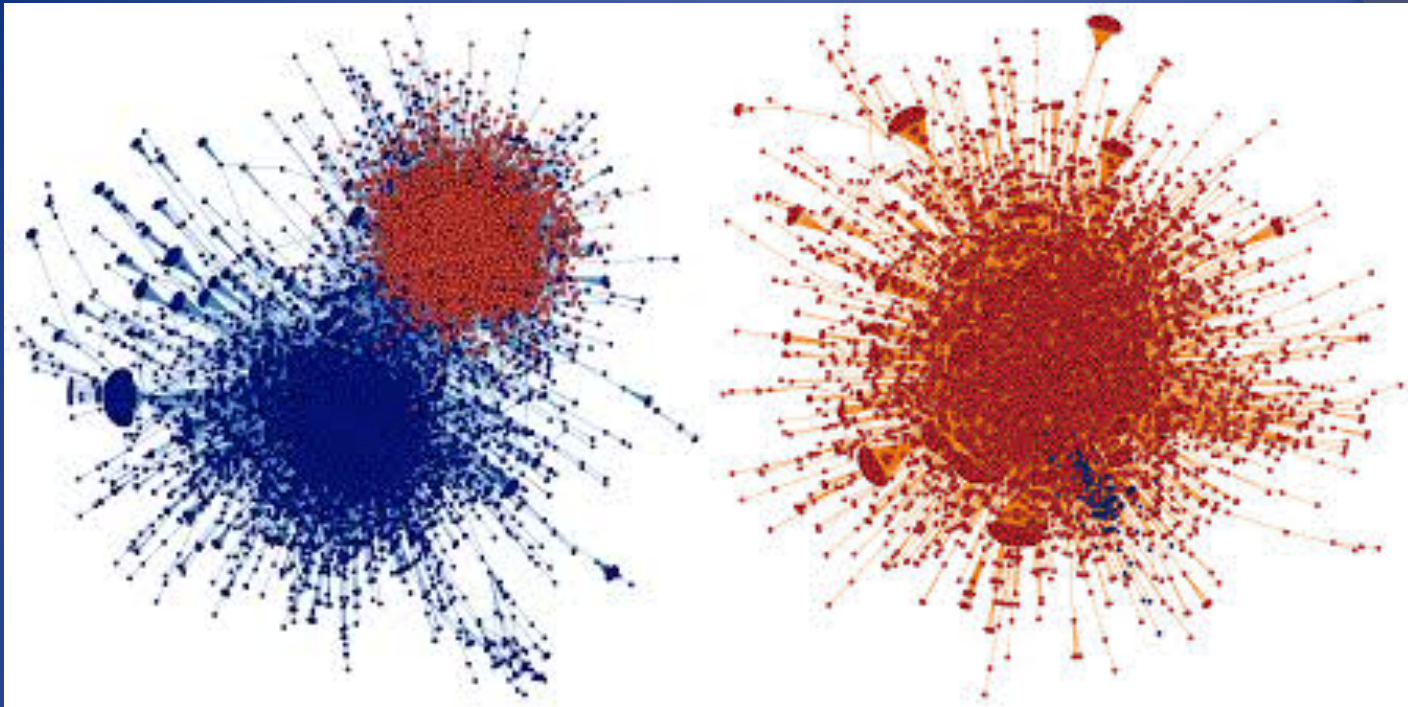
Measuring the Web - Blogosphere

Political Blogosphere
2004 US Presidential Election

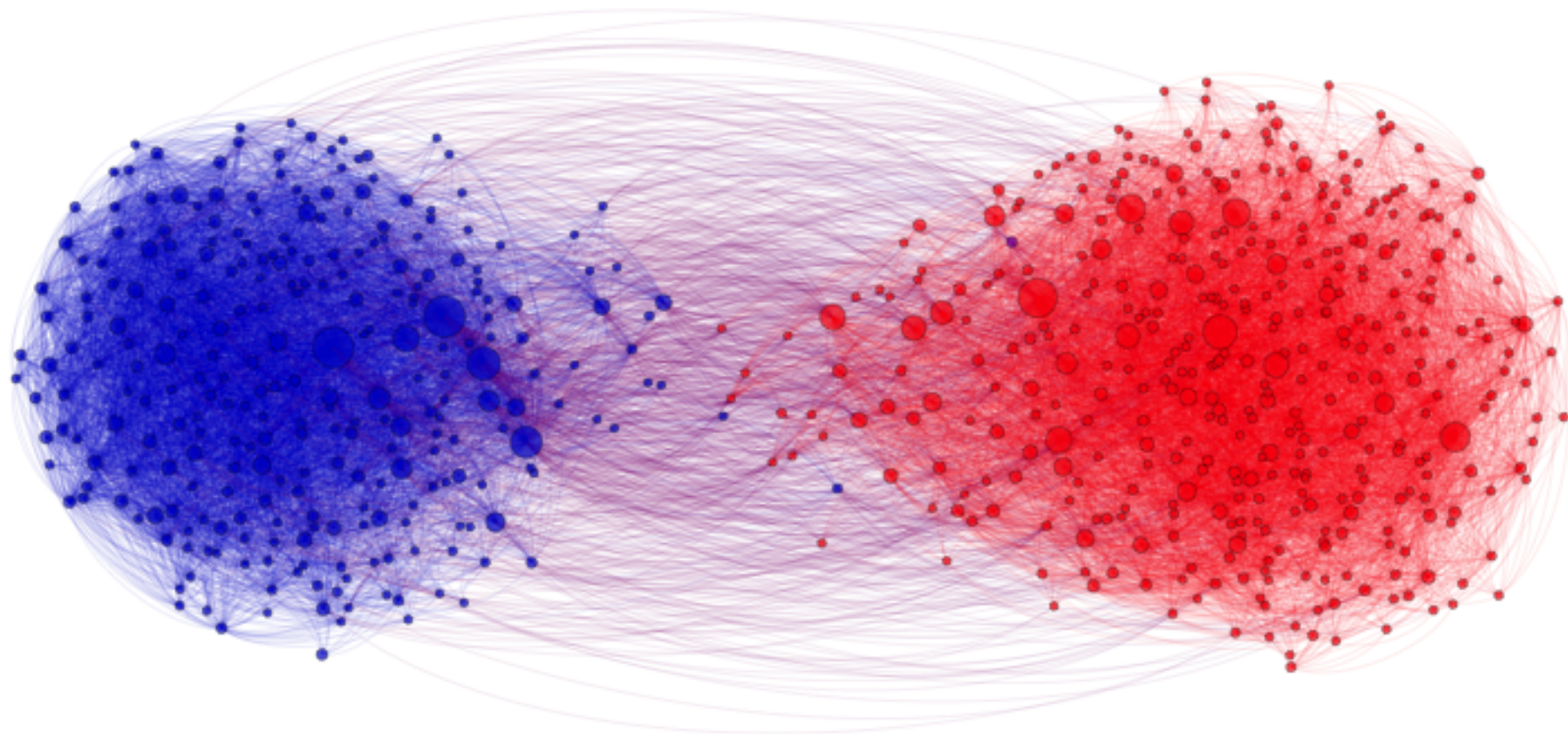
Bloggers:
Blue - Democrat
Red - Republican
Pink - Neutral



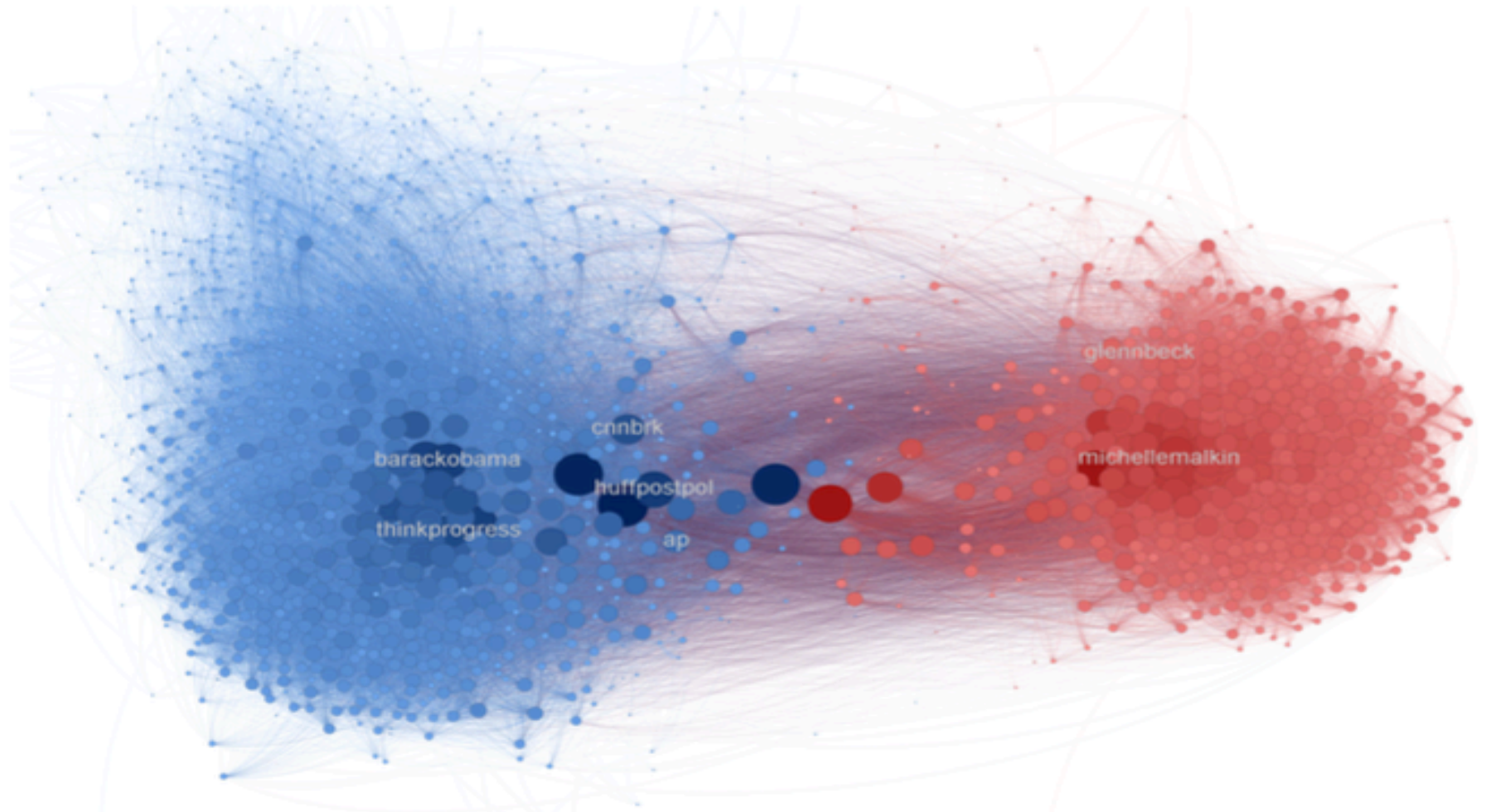
L. Adamic, N. Glance, The political blogosphere and the 2004 U.S. election:
divided they blog, LinkKDD'05



A sinistra troviamo la rete dei Retweet, mentre a destra la rete delle menzioni. Queste informazioni sono state ottenute mediante l'algoritmo force-directed. I nodi colorati riflettono l'assegnazione a grappolo. Risulta evidente la formazione di una struttura comunitaria nella rete Retweet, mentre nell'altra risulta molto meno marcata. Tramite studi statistici si può dimostrare che il grappolo rosso nella rete Retweet è composto dell'80% di utenti di sinistra.



La rete visualizza più di 1500 account che hanno co-retweettato i durante il secondo dibattito. Analizzando questi dati tramite un algoritmo di rilevamento, si scopre che i nodi di colore blu appartengono ad account di persone con un orientamento politico liberale mentre quelli rossi appartengono a utenti di orientamento conservatore. Con un'ulteriore analisi, si evidenziano dei nodi particolarmente connessi che appartengono ad organizzazioni di notizie riguardanti la politica.



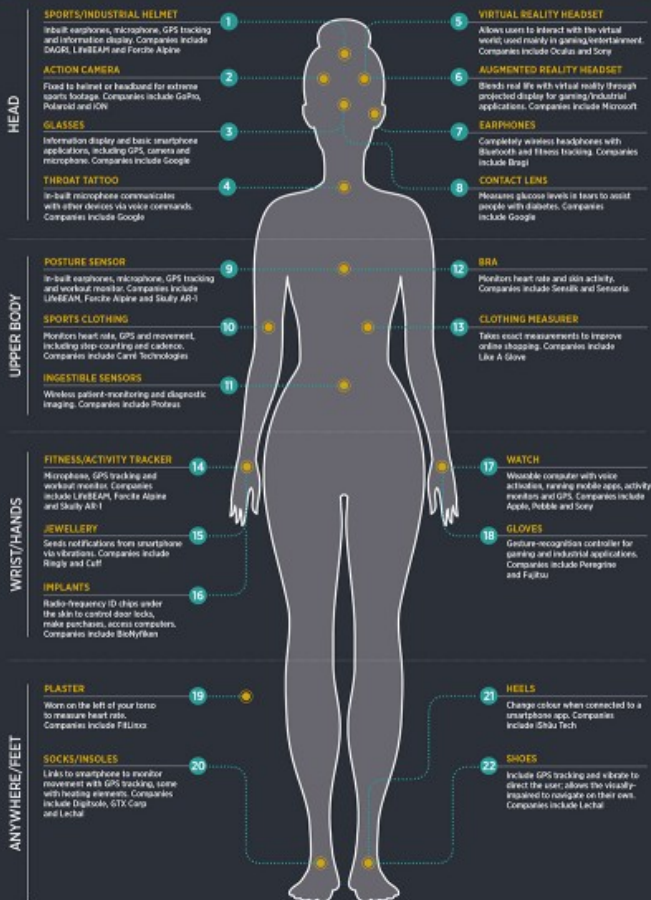
La rete visualizza più di 1500 account che hanno co-retweettato i durante il secondo dibattito. Analizzando questi dati tramite un algoritmo di rilevamento, si scopre che i nodi di colore blu appartengono ad account di persone con un orientamento politico liberale mentre quelli rossi appartengono a utenti di orientamento conservatore. Con un'ulteriore analisi, si evidenziano dei nodi particolarmente connessi che appartengono ad organizzazioni di notizie riguardanti la politica.

Incremento produzione ed utilizzo di dispositivi wireless.

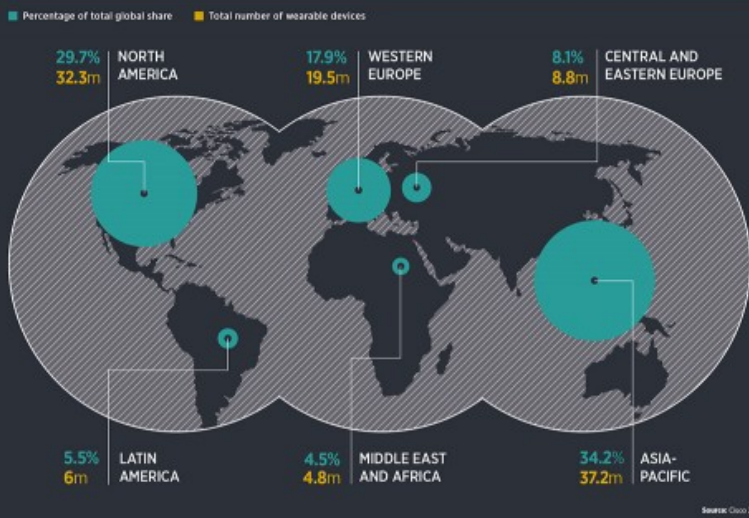
Growing trends in wearable technology



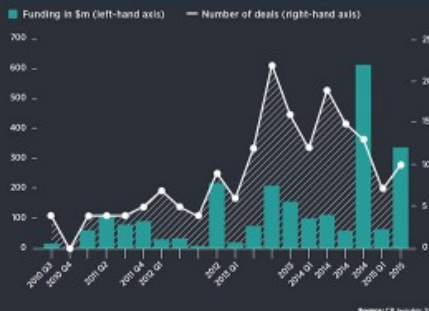
ANALYSIS AND APPLICATION OF WEARABLE DEVICES



WEARABLES REGIONAL MARKET SHARE



QUARTERLY PRIVATE WEARABLE TECHNOLOGY FUNDING 2010-2015



GLOBAL ENTERPRISE AND INDUSTRIAL WEARABLE REVENUE SHARE BY DEVICE



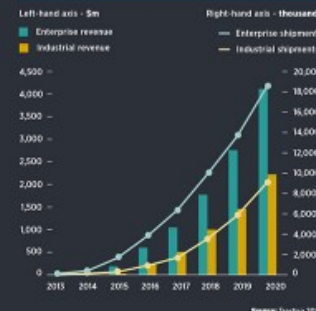
TOP BENEFITS OF USING A WEARABLE DEVICE IN AN ENTERPRISE



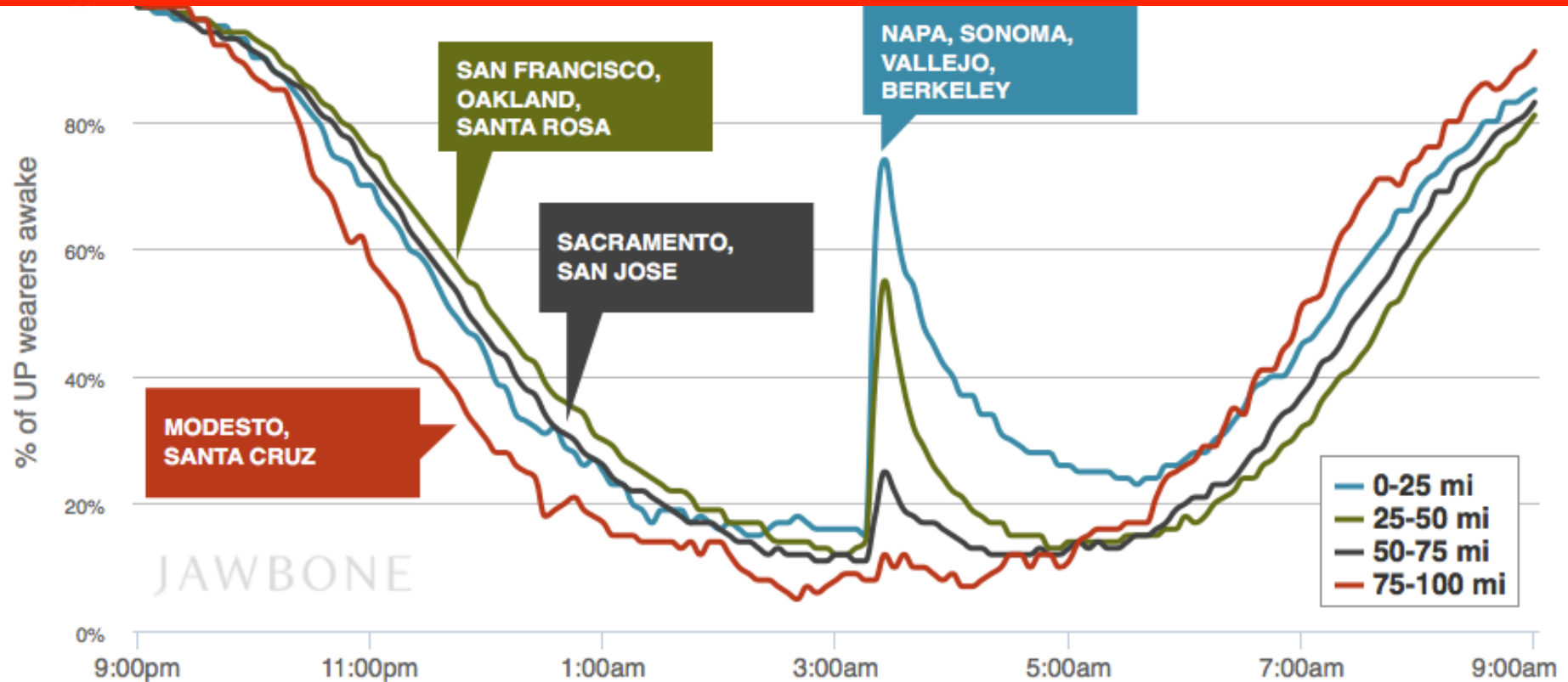
TOTAL ESTIMATED SHIPMENT VOLUMES OF GLOBAL WEARABLES (IN UNIT SALES)



GLOBAL ENTERPRISE AND INDUSTRIAL WEARABLE SHIPMENT AND REVENUE



DISPOSITIVI FITNESS



Wikipedia Usage Estimates Prevalence of Influenza-Like Illness in the United States in Near Real-Time

David J. McIver*, John S. Brownstein

Boston Children's Hospital, Harvard Medical School, Boston, Massachusetts

Abstract

Circulating levels of both seasonal and pandemic influenza in the population. While up-to-date information is typically available to two weeks. We introduce a novel method to estimate the prevalence of influenza-like illness (ILI) in the United States (US) by monitoring the rate of page views to certain influenza- or health-related Wikipedia pages. We compared these data to official ILI data from the CDC (CDC). We developed a Poisson model that accurately predicted ILI two weeks ahead of the CDC, with an absolute error of 1.0. Wikipedia-derived ILI models performed well during the 2009 H1N1 pandemic (as well as unusually low usage accurately estimated the week of peak activity). Wikipedia usage accurately estimated the week of peak activity and was accurate in its measure of ILI intensity. With frequent monitoring of ILI activity in the US and to provide near real-time

Citation: McIver DJ, Brownstein JS (2014) Wikipedia Usage Estimates Prevalence of Influenza-Like Illness in the United States in Near Real-Time. *PLOS Computational Biology* 10(4): e1003581. doi:10.1371/journal.pcbi.1003581

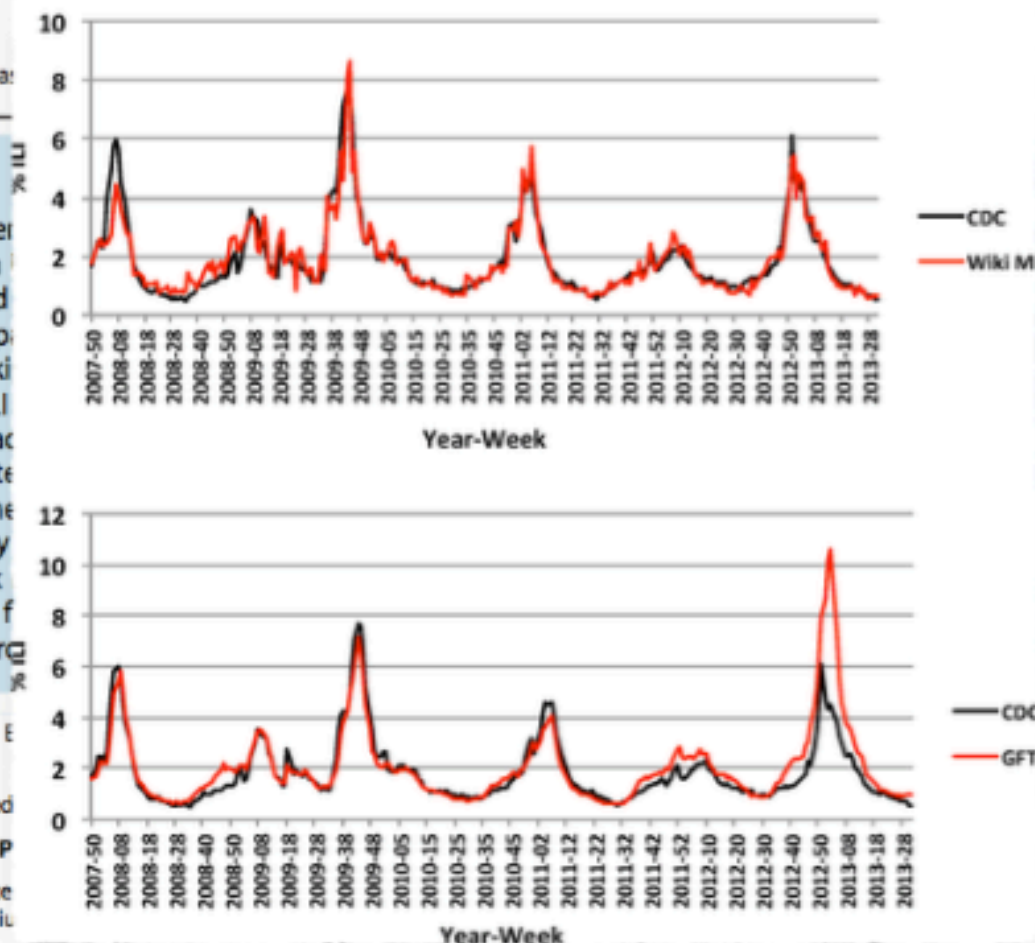
Editor: Marcel Salathé, Pennsylvania State University, UNITED STATES

Received: December 20, 2013; **Accepted:** March 11, 2014; **Published:** April 10, 2014

Copyright: © 2014 McIver, Brownstein. This is an open-access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was funded by the National Institutes of Health and National Library of Medicine 1R01LM010812-03. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.



ALGORITMI di LEARNING AVANZATI

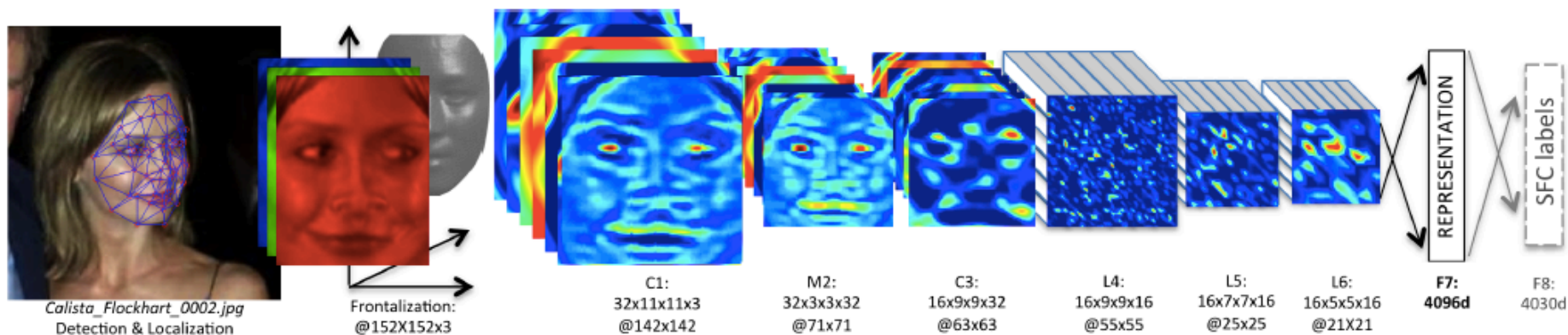


Figure 2. Outline of the DeepFace architecture. A front-end of a single convolution-pooling-convolution filtering on the rectified input, followed by three locally-connected layers and two fully-connected layers. Colors illustrate outputs for each layer. The net includes more than 120 million parameters, where more than 95% come from the local and fully connected layers.

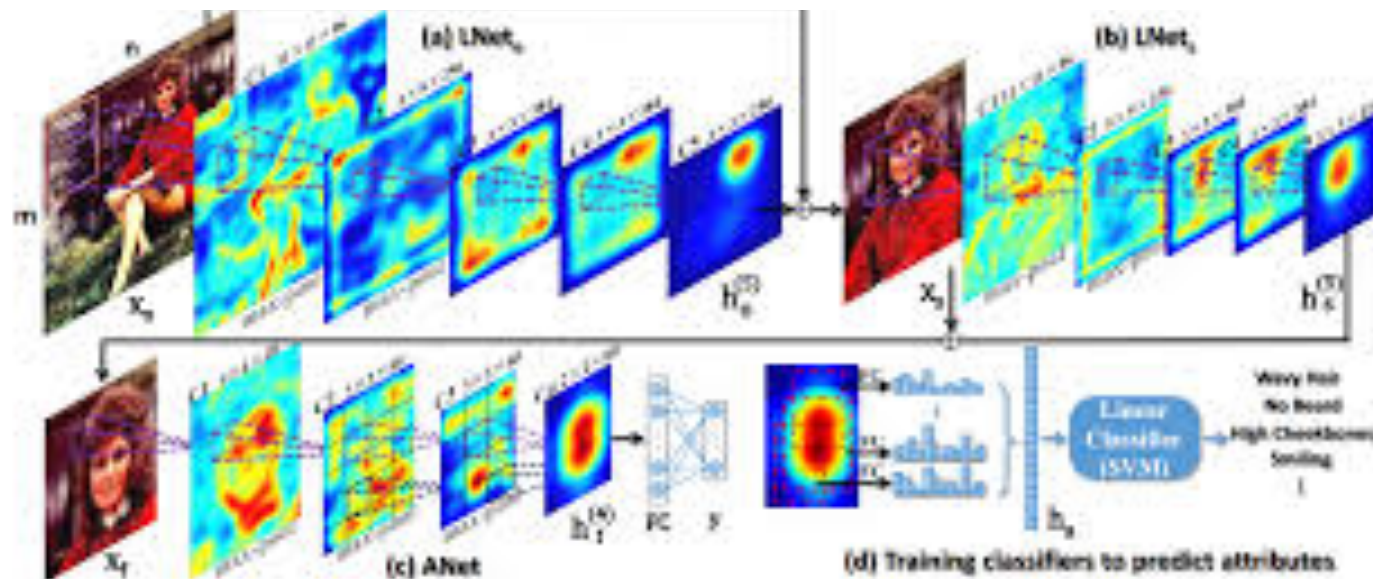
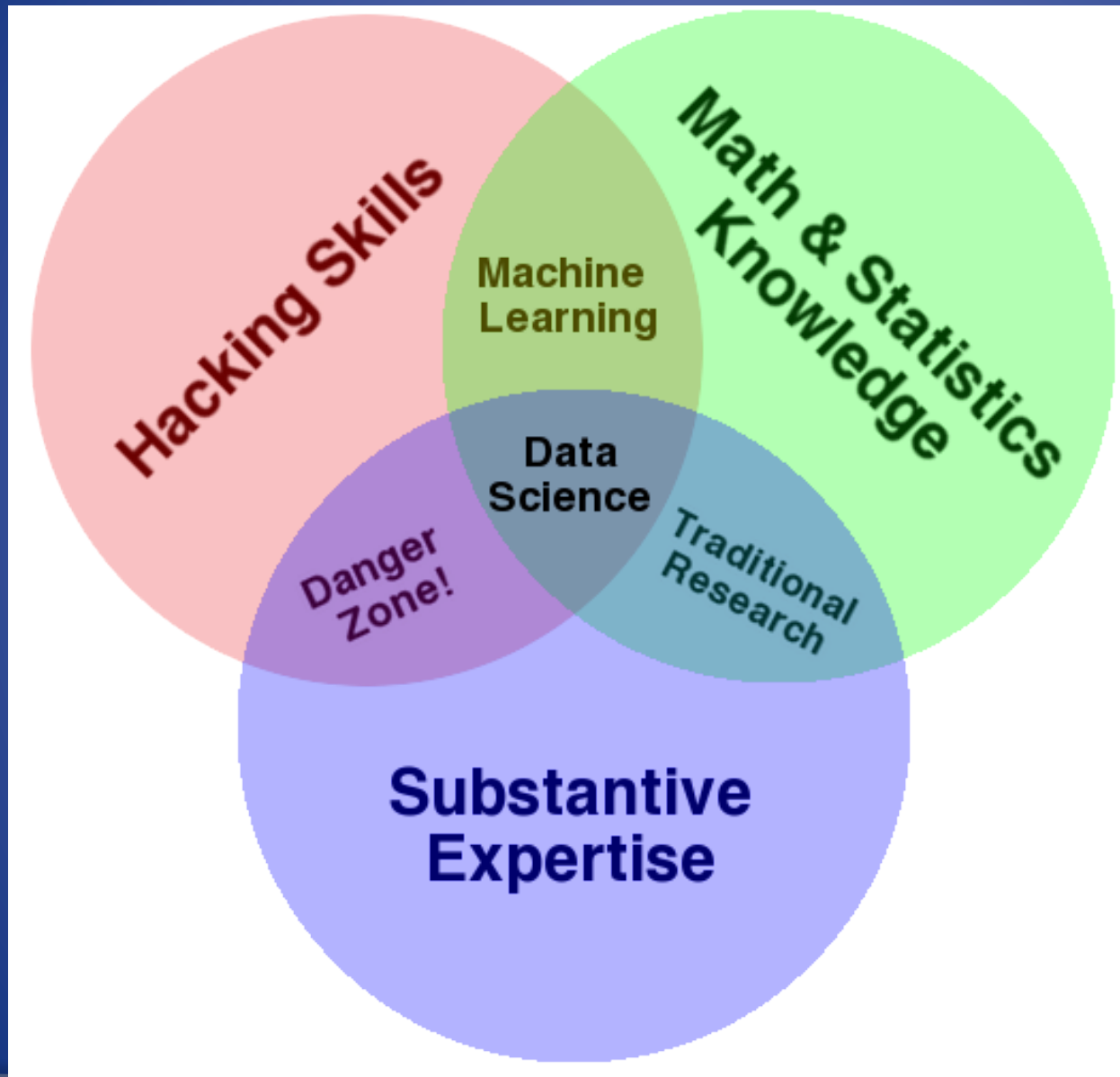


Figure 3. The proposed pipeline of attribute inference.

CARATTERISTICHE MACHINE LEARNING



LA FRONTIERA DELLA RICERCA

L'approccio convenzionale, via intelligenza artificiale (con il “data mining” che ricorre alla possibilità che la macchina impari a cercare tra i dati) e scienza delle reti complesse, sta mostrando i suoi limiti nel far fronte allo tsunami di dati.



I dati possono essere organizzati in modo naturale in strutture geometriche discrete, che si prestano a essere rappresentate come complessi simpliciali. Pertanto nello spazio dei dati l'informazione induce la topologia. La “teoria topologica dei campi sullo spazio dei dati” vuole estrarre informazioni dai dati in modo diverso e più efficace, globale anziché locale, e capace di ottenere maggiori informazioni nascoste nei dati e dai dati.



I primi risultati sono già straordinari: applicata, in forma approssimata, a dati relativi a immagini di cervelli umani ottenute con Risonanza magnetica funzionale, la teoria è riuscita infatti a distinguere soggetti “normali” da altri trattati con psicofarmaci, rivelando che nei dati era nascosta un'informazione che la stessa macchina con cui erano stati acquisiti non era stata addestrata a “vedere”.



*“I think the next century
will be the century
of complexity.”*

Stephen Hawking
January 23, 2000

REFERENZE

- **Network Science** by Albert-László Barabási
- **Linked: The New Science of Networks** by Albert-László Barabási
<http://barabasi.com/book/network-science>
- **LEARNING FROM DATA** by Yaser S. Abu-Mosta, Malik Magdon-Ismael
<http://amlbook.com/>
- **L'algoritmo definitivo. La macchina che impara da sola e il futuro del nostro mondo** by Pedro Domingos
- <http://snap.stanford.edu/data/>