INSTITUTE
FOR SCIENTIFIC INTERCHANGE
FOUNDATION

Big Data,

between science and society

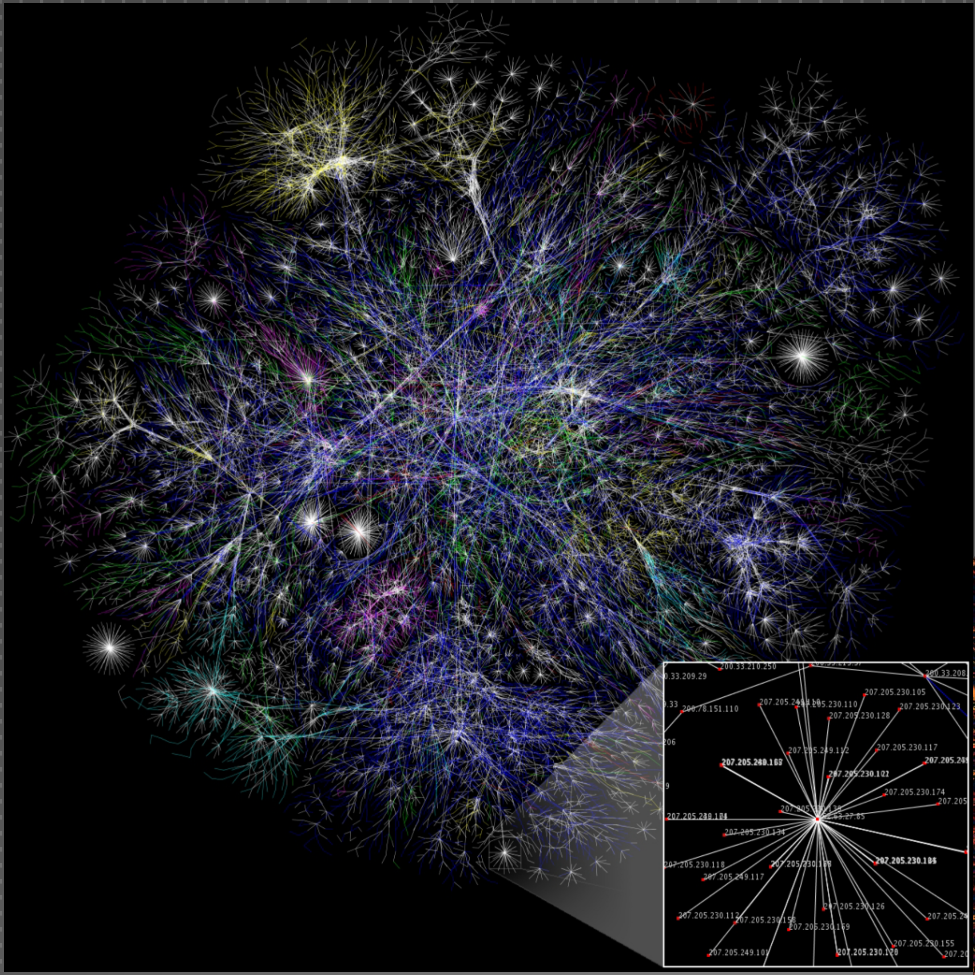
Mario Rasetti

ISI Foundation, Torino – New York

COMPLEX SYSTEMS

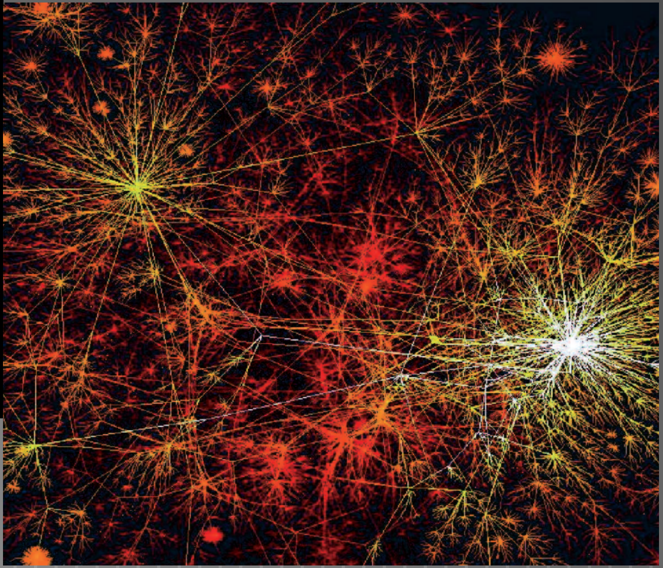
Bevys of Starlings

Multi-agent – Multi-scale – Emergent effects



Network representable

The Internet
Clustering (handshake)



- ✓ large number of components
- ✓ interactions between components
- ✓ multi-scale hierarchical structures
- ✓ coupling between scales
- ✓ self-organization (*no blueprint*)
- ✓ emergent properties
- ✓ “complex” is more than “complicated”
- ✓ **more is different** (P.W. Anderson)
    - ★ the end of linear thinking
    - ★ interdependence and systemic approach
    - ★ the problem of causal inference
    - ★ universality
    - ★ (BIG) DATA representable

BIG DATA : how many ?

We live in a world where:

- every day 320 billions e-messages and 25 billions SMS are exchanged ; and over 500 millions photographs are posted and shared on Facebook ;
- more than 4.5 billion people (> ½ world population!) have a mobile phone (or similar device)
- the total quantity of information created and exchanged has reached, in 2013, 4 *zettabyte*s [a zettabyte means 10^{21} bytes (characters) : the 1250 pages of Tolstoy's *War and Peace* could be stored 323 billion times in a zettabyte, and the whole content of the Library of Congress 4 million times] and grows 40% per year (will reach in 4 years a *yottabyte*, 10^{24})

- A world in which every year 1 billion cars take the road, over 2 billions people fly on airplanes; population and urbanization, commercial exchanges and migrations grow tumultuosly - entangled with a parallel universe of more and more sophisticated technologies, giving life to a unique, interconnected socio-technical system.
- The growing complexity of this scenario hides enormous opportunities as well as potential risks, generated just by the vulnerability that lies in the interdipendencies between different systems.
- For this society requires an increased capacity of making predictions, to anticipate, evaluate and correlate risks and understand the complexity of the novel, different, world that new technologies are giving life to.
- Big Data 2.0 will be sensors

Thanks to Data, digital imaging is tracking this world more and more closely

- this allows us to use computation to **extract patterns and establish causal inferences** using tools, e.g., from A.I. : **data mining**, **machine learning**, **statistics** → **MINING ALGORITHMS** (Google)
- mathematical modeling and forecast may then take place on a **data-rich landscape**, fed by **data streams** from multiple sources : (BI)SIMULATION
- thus we can assess at unprecedented speed and scale our **worldview against reality**, and feed it back to models, and help – through virtual scenarios – decision makers & policy makers to behave in a more rational way : PREDICTIONS

The great challenge :
understand, govern and control the process

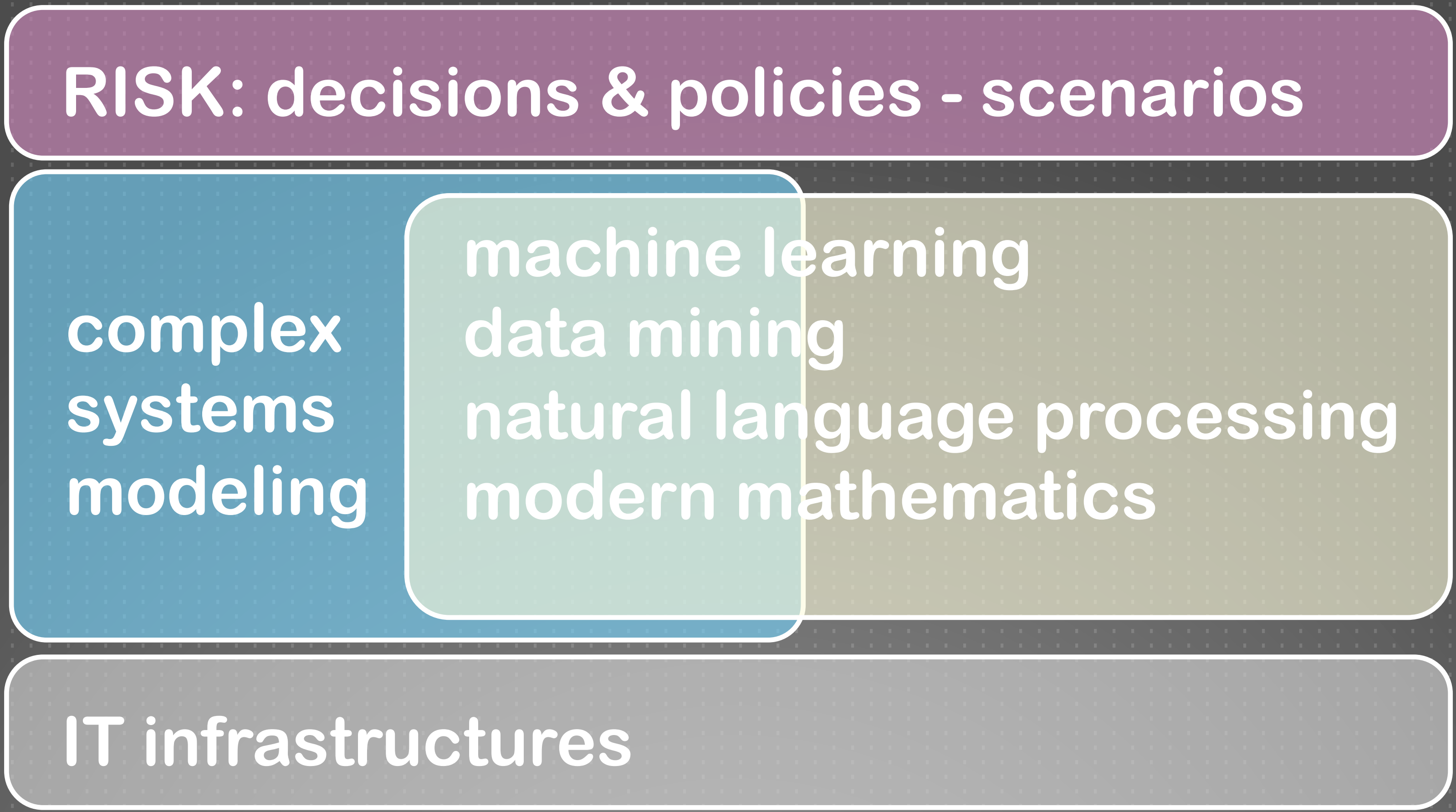
Data → Information → Knowledge → Wisdom

Data → Information → Data Mining

Information → Knowledge → Patterns & Correlations

Knowledge → Wisdom → A ‘Field Theory of Data’

The methodology



RISK: decisions & policies - scenarios

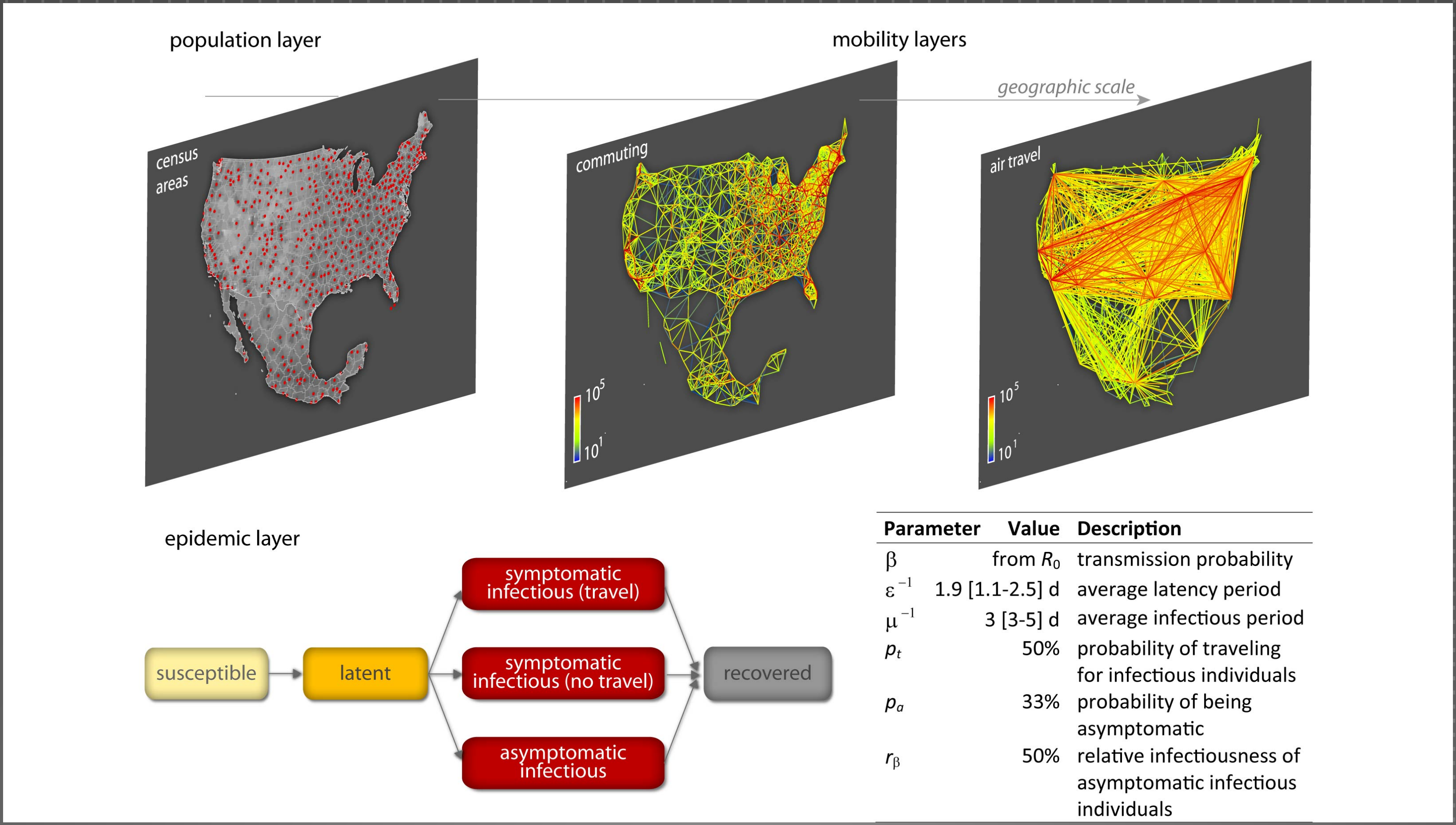
complex systems modeling

machine learning
data mining
natural language processing
modern mathematics

IT infrastructures

Data → Virtual Reality → (Simulation) → **Predictions**

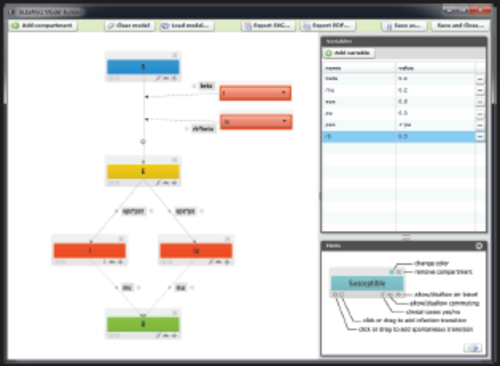
Predicting Pandemics

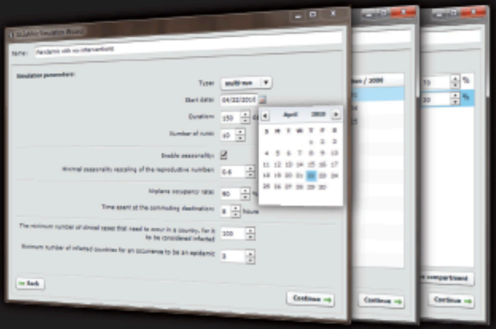


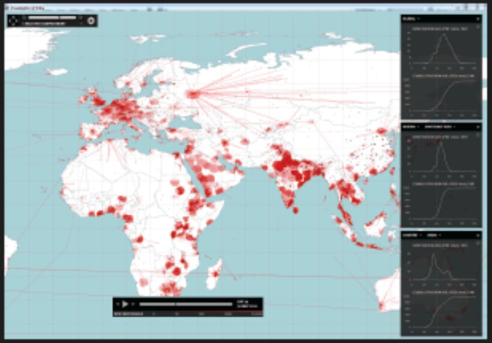
epidemic layer

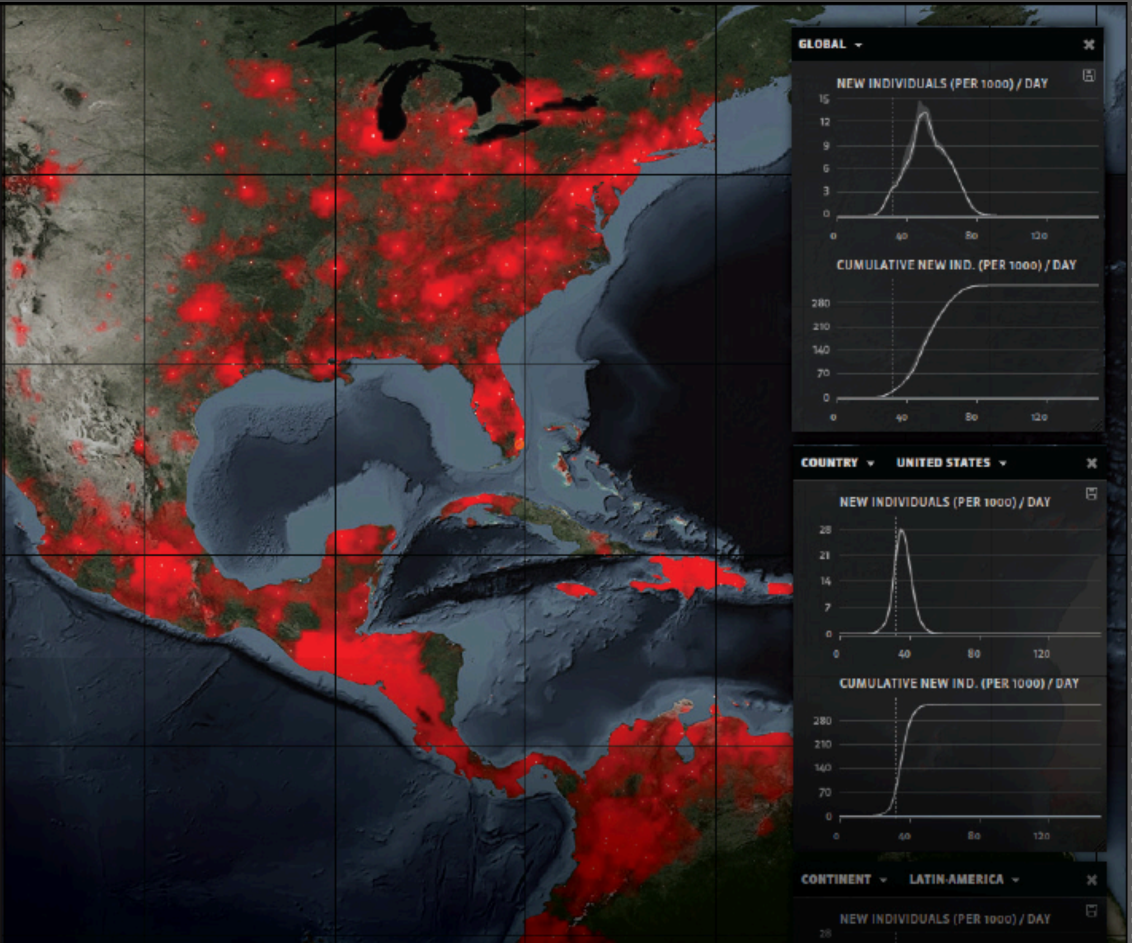
susceptible → latent → symptomatic infectious (travel) / symptomatic infectious (no travel) / asymptomatic infectious → recovered

Parameter	Value	Description
β	from R_0	transmission probability
ε^{-1}	1.9 [1.1-2.5] d	average latency period
μ^{-1}	3 [3-5] d	average infectious period
p_t	50%	probability of traveling for infectious individuals
p_a	33%	probability of being asymptomatic
r_β	50%	relative infectiousness of asymptomatic infectious individuals







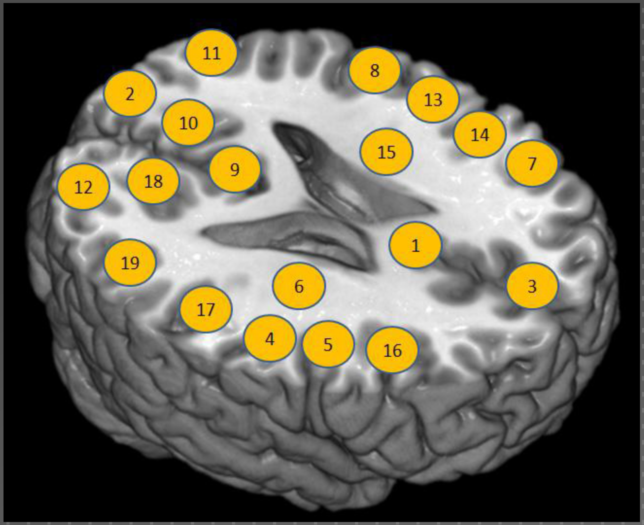

GLOBAL
NEW INDIVIDUALS (PER 1000) / DAY
CUMULATIVE NEW IND. (PER 1000) / DAY
COUNTRY
UNITED STATES
NEW INDIVIDUALS (PER 1000) / DAY
CUMULATIVE NEW IND. (PER 1000) / DAY
CONTINENT
LATIN-AMERICA
NEW INDIVIDUALS (PER 1000) / DAY

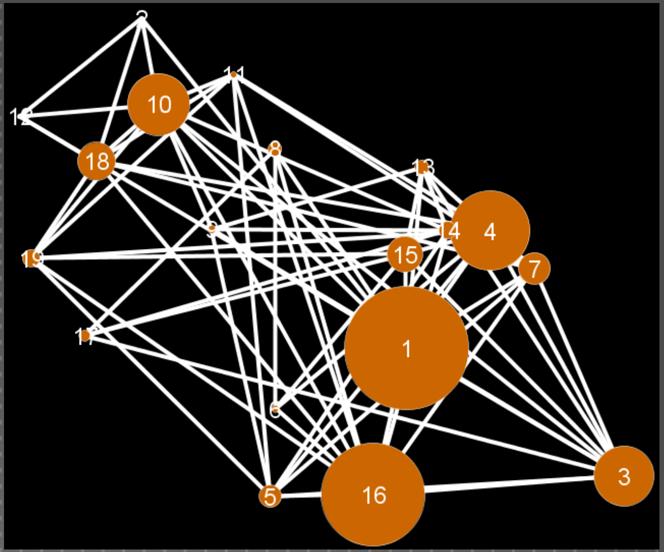
GLEaMviz is a multiplatform application that allows to interactively program and generate data with the GLEaM computational model.

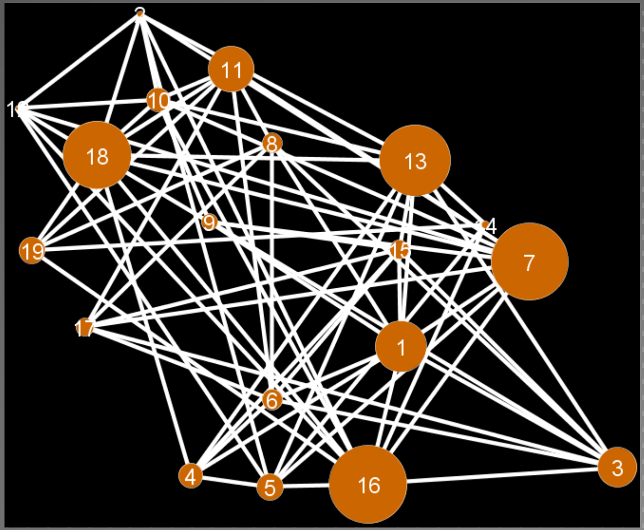
With the GLEaMviz application you can:

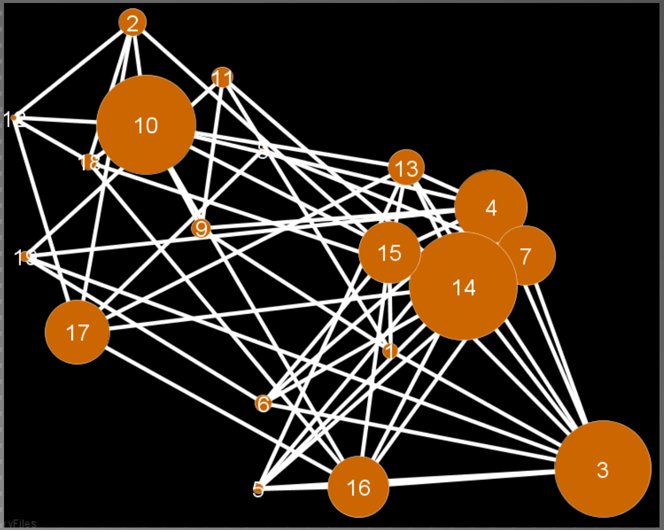
- Configure the epidemic model by setting the disease natural history.
- Define the simulation scenario by including environmental effects and the initial conditions of the outbreak.
- Define mitigation and containment policies and analyze their effectiveness.
- Explore simulation results through dynamic plots and maps.
- Download and share the generated data through a user-friendly interface.

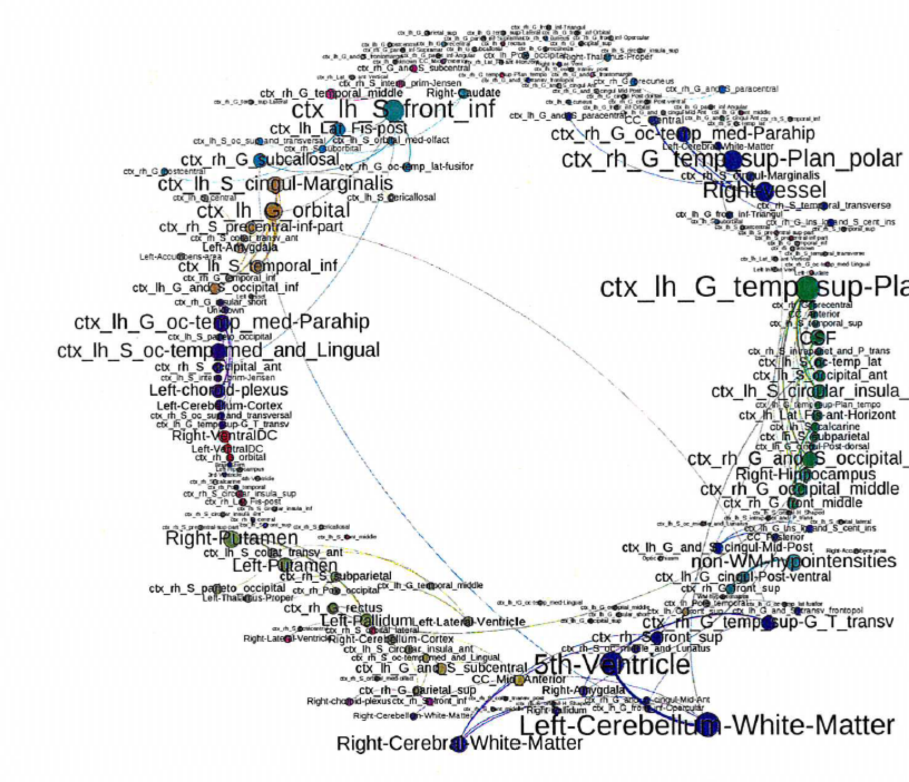
The Human Brain





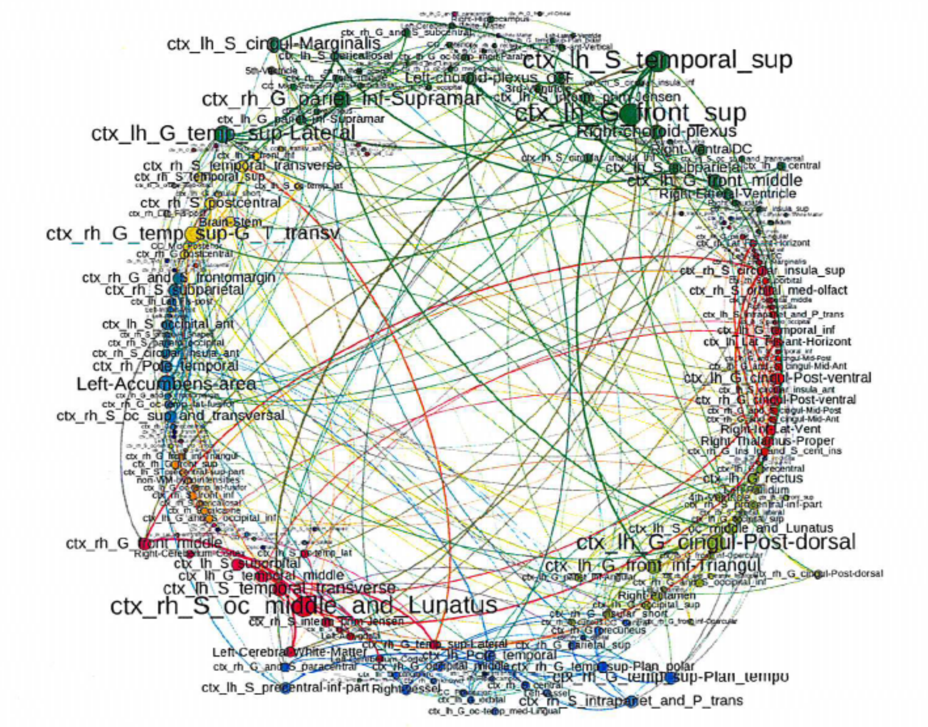




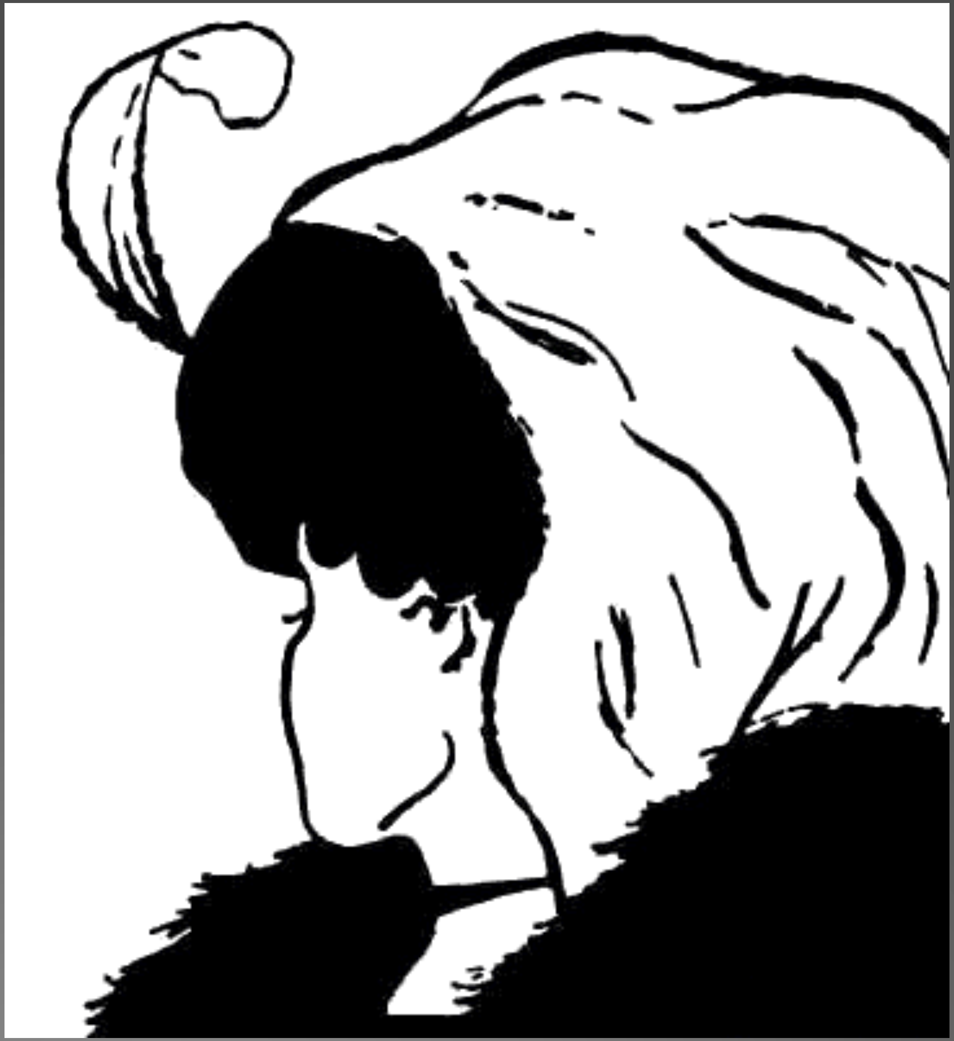

ctx_lh_S_front_inf
ctx_rh_G_temporal_middle
Right-Caudate
ctx_lh_Lat_Fis-post
ctx_rh_G_subcallosal
ctx_lh_S_cingul-Marginalis
ctx_lh_G_orbital
ctx_rh_S_precentral-inf-part
Left-Amygdala
ctx_lh_S_temporal_inf
ctx_lh_G_and_S_occipital_inf
ctx_lh_G_oc-temp_med-Parahip
ctx_lh_S_oc-temp_med_and_Lingual
ctx_rh_S_occipital_ant
Left-choroid-plexus
Left-Cerebellum-Cortex
Right-VentralDC
Left-VentralDC
Right-Putamen
ctx_lh_S_collat_transv_ant
Left-Putamen
ctx_rh_S_subparietal
ctx_rh_S_parieto_occipital
Left-Thalamus-Proper
ctx_rh_G_rectus
Left-Pallidum
Left-Lateral-Ventricle
Right-Cerebellum-Cortex
ctx_lh_S_circular_insula_ant
ctx_lh_G_and_S_subcentral
CC_Mid_Anterior
ctx_rh_G_parietal_sup
Right-Amygdala
5th-Ventricle
Left-Cerebellum-White-Matter
Right-Cerebral-White-Matter
ctx_rh_G_oc-temp_med-Parahip
ctx_rh_G_temp_sup-Plan_polar
Right-vessel
ctx_lh_G_temp_sup-Plan_polar
CSF
ctx_lh_S_circular_insula_inf
ctx_lh_Lat_Fis-ant-Horizont
ctx_rh_G_and_S_occipital_inf
Right-Hippocampus
ctx_rh_G_occipital_middle
ctx_rh_G_front_middle
ctx_lh_G_and_S_cingul-Mid-Post
non-WM-hypointensities
ctx_lh_G_cingul-Post-ventral
ctx_rh_G_front_sup
ctx_rh_G_temp_sup-G_T_transv
ctx_rh_S_front_sup

"normal subjects"

Psychoactive drug treated subjects

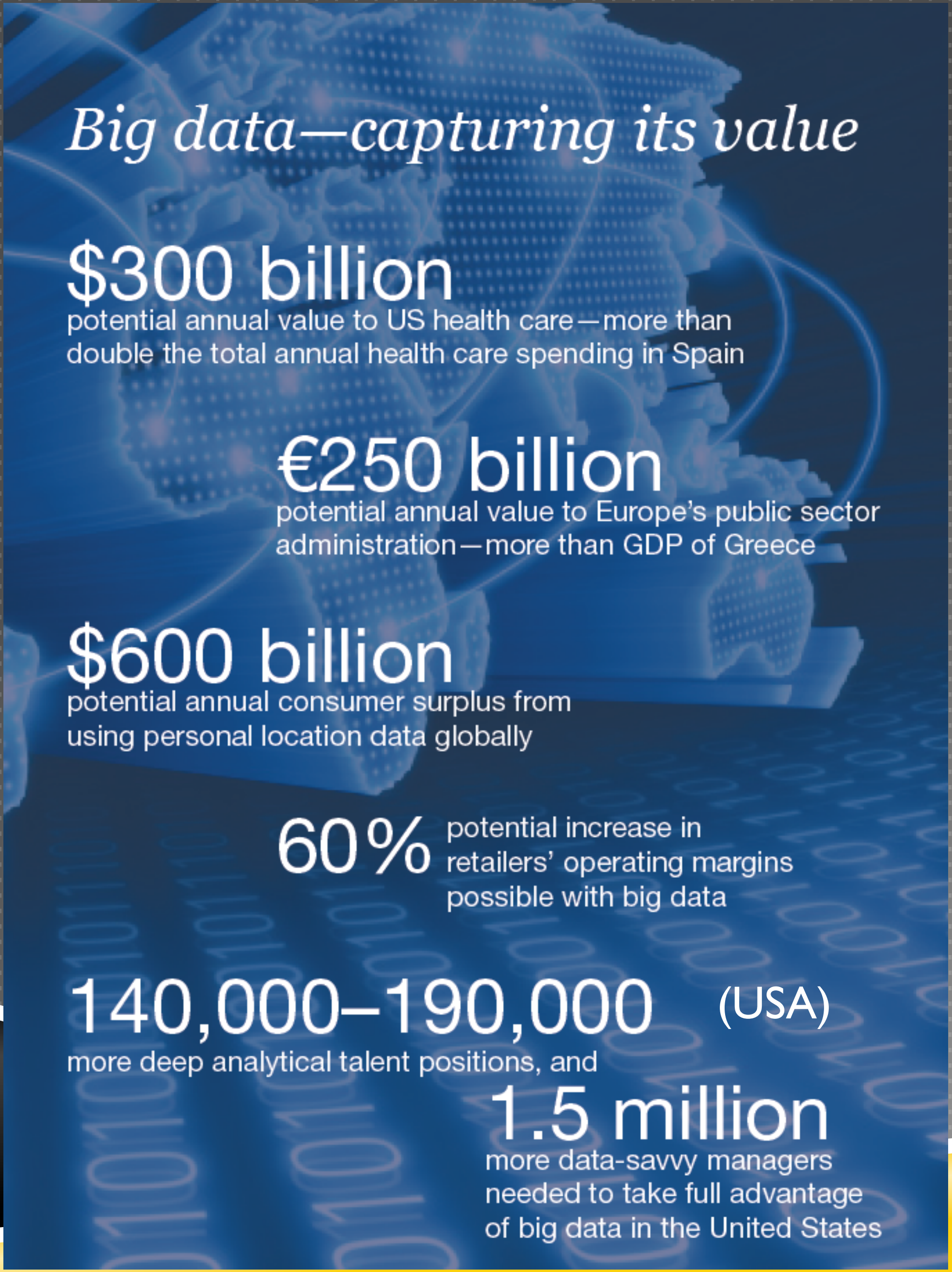

ctx_lh_S_cingul-Marginalis
ctx_lh_S_temporal_sup
ctx_rh_G_pariet_inf-Supramar
ctx_lh_G_front_sup
ctx_lh_G_temp_sup-Lateral
Right-choroid-plexus
Left-choroid-plexus
Right-VentralDC
ctx_lh_G_front_middle
ctx_rh_S_temporal_transverse
ctx_rh_S_postcentral
Brain-Stem
ctx_rh_G_temp_sup-G_T_transv
ctx_rh_G_and_S_frontomargin
ctx_rh_S_subparietal
ctx_lh_S_occipital_ant
ctx_rh_Pole_temporal
Left-Accumbens-area
ctx_rh_S_oc_sup_and_transversal
ctx_rh_S_circular_insula_sup
ctx_rh_S_orbital_med-olfact
ctx_lh_G_temporal_inf
ctx_lh_Lat_Fis-ant-Horizont
ctx_lh_G_cingul-Post-ventral
ctx_rh_G_cingul-Post-ventral
Right-Inf-Lat-Vent
Right-Thalamus-Proper
ctx_lh_G_rectus
ctx_rh_G_front_middle
ctx_lh_S_suborbital
ctx_lh_G_temporal_middle
ctx_lh_S_temporal_transverse
ctx_rh_S_oc_middle_and_Lunatus
ctx_lh_G_cingul-Post-dorsal
ctx_lh_G_front_inf-Triangul
ctx_lh_S_oc_middle_and_Lunatus
Right-Putamen
Left-Cerebral-White-Matter
ctx_lh_Pole_temporal
ctx_rh_G_temp_sup-Plan_polar
ctx_rh_G_temp_sup-Plan_tempo
ctx_lh_S_precentral-inf-part
Right-vessel
ctx_rh_S_intrapariet_and_P_trans

Francisco Varela



Ambiguous pictures

MacKinsey BD Report (2012)



Big data—capturing its value

$300 billion
potential annual value to US health care—more than double the total annual health care spending in Spain

€250 billion
potential annual value to Europe's public sector administration—more than GDP of Greece

$600 billion
potential annual consumer surplus from using personal location data globally

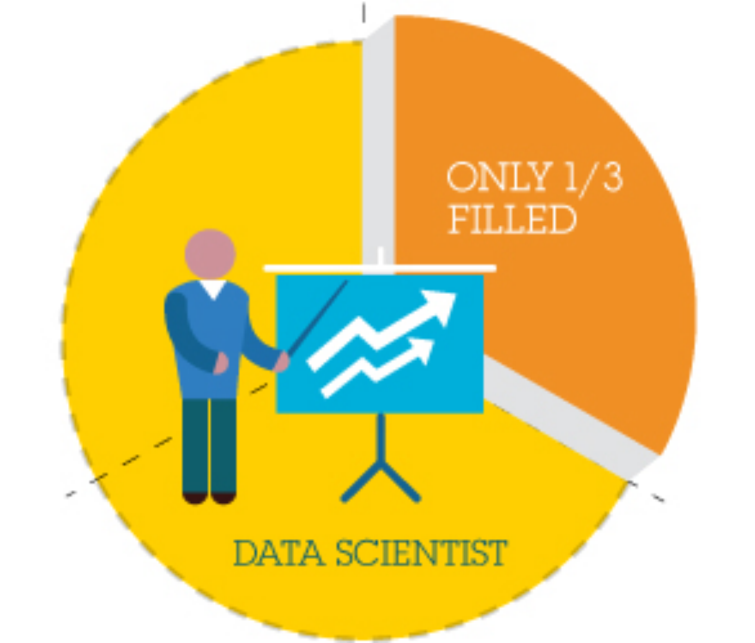
60% potential increase in retailers' operating margins possible with big data

140,000–190,000 (USA)
more deep analytical talent positions, and

1.5 million
more data-savvy managers needed to take full advantage of big data in the United States

IBM estimate (2014) - worldwide

4.4MILLION
data scientists
needed by 2015



The hardware side

What do we need to achieve?

Data comes today from different sources, in different formats and at different times.

Easy access requires:

- Scalability
- Fast & efficient data storage and access
- Tiering and compression
- Friendly visualization

What infrastructure features do we need?

Big data analytics are useful only if the insights are easily available to people and processes that need them.

This requires an infrastructure with:

- Self-healing capability
- High-grade continuous management
- Single- and multiple-site solutions

What speed do we need?

Also very efficient big data analytics tools are useless if insights require infinite time.

This requires an infrastructure capable of extracting information from big data and performing data analytics in real time:

- Low latency capability
- Ability to scale up, in and out fast
- Analytics customized for the system of record

- Over 50% business leaders don’t have access to the data they need
- Only 38% of organizations are prepared to deal with the onslaught of big data
- Infrastructures are needed able to handle up to 20 Petabytes of data
- Presently available storage systems can be improved to store in the same space up to 5 times the amount of data now stored

- In 2020 over 30 billion devices will be in use
- By 2017, 60% of cars will have connected services
- Infrastructures providing data analytics with essentially > 95% fewer outages and performance problems than present commodity platforms will be necessary
- Big data and data analytics infrastructures should deliver up to 99.999% uptime

- 40% of executives need real-time information accessible within the process time
- Estimated minimum need is capability of processing complex queries thousands of times faster
- Infrastructures providing data analytics results with essentially > 95% higher speed than present commodity platforms will be necessary
- Capacity of storing data to be processed enabling delivery insights at least 20 times larger