On the Big Impact of Big Computer Science

Stefano Ceri
Politecnico di Milano

The «Big Approach» in the pharma sector

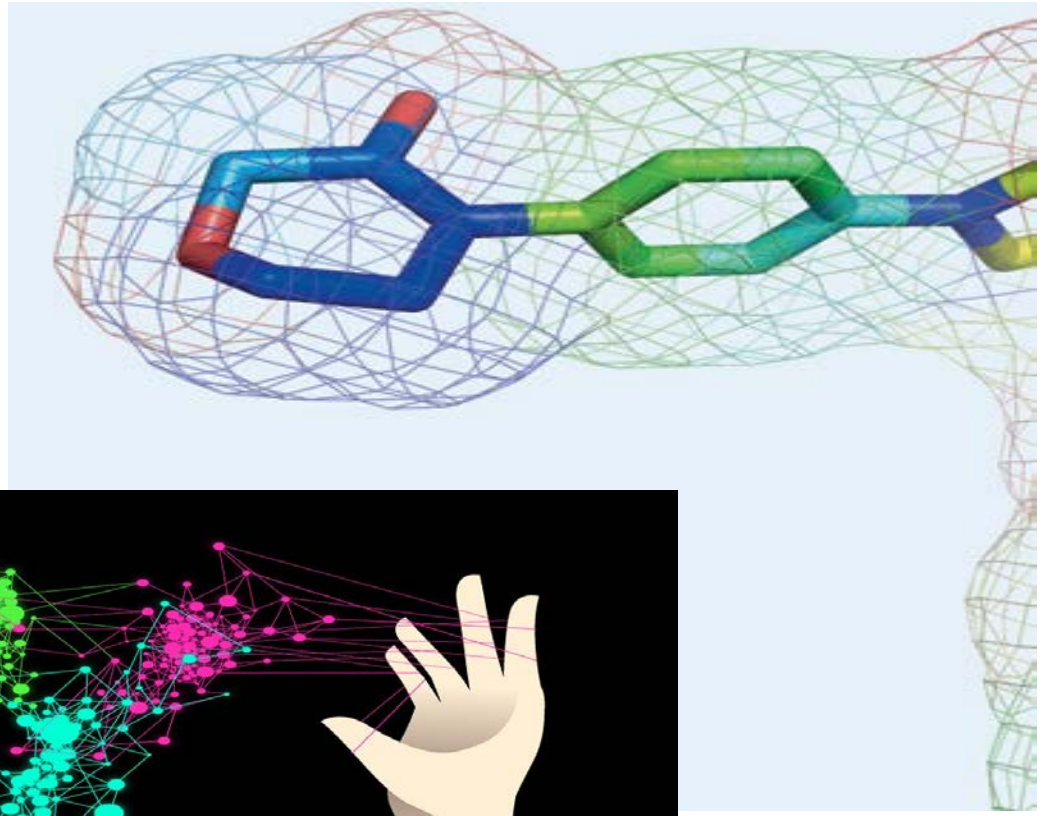
Bayer, From Molecules to Medicine,

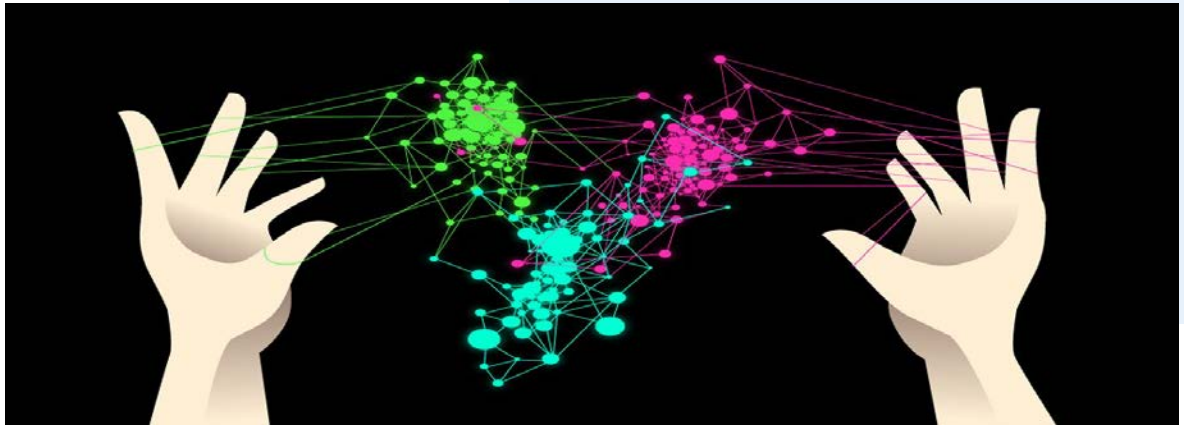
http://pharma.bayer.com/en/research-and-development/technologies/small-and-large-molecules/index.php, retrieved July 15, 2015.

1. DNA TESTING for TARGET DISCOVERY

2. High-Throughput Screening

3. STRUCTURAL BIOLOGY / COMPUTATIONAL CHEMISTRY



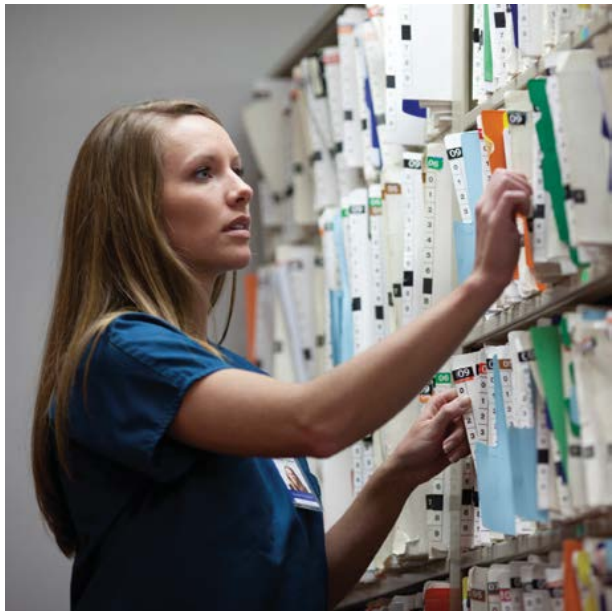


Then it is a long way to the production of medicines…

4. Finding the optimum: Medicinal Chemistry
5. Understanding effects: Pharmacology and Toxicology
6. Packaging the active ingredient: Galenics
7. Testing tolerability: Phase I
8. Confirming efficacy: Phases II and III
9. Predicting effects on individuals: Pharmacogenomics
10. Putting it all together: Regulatory Affairs

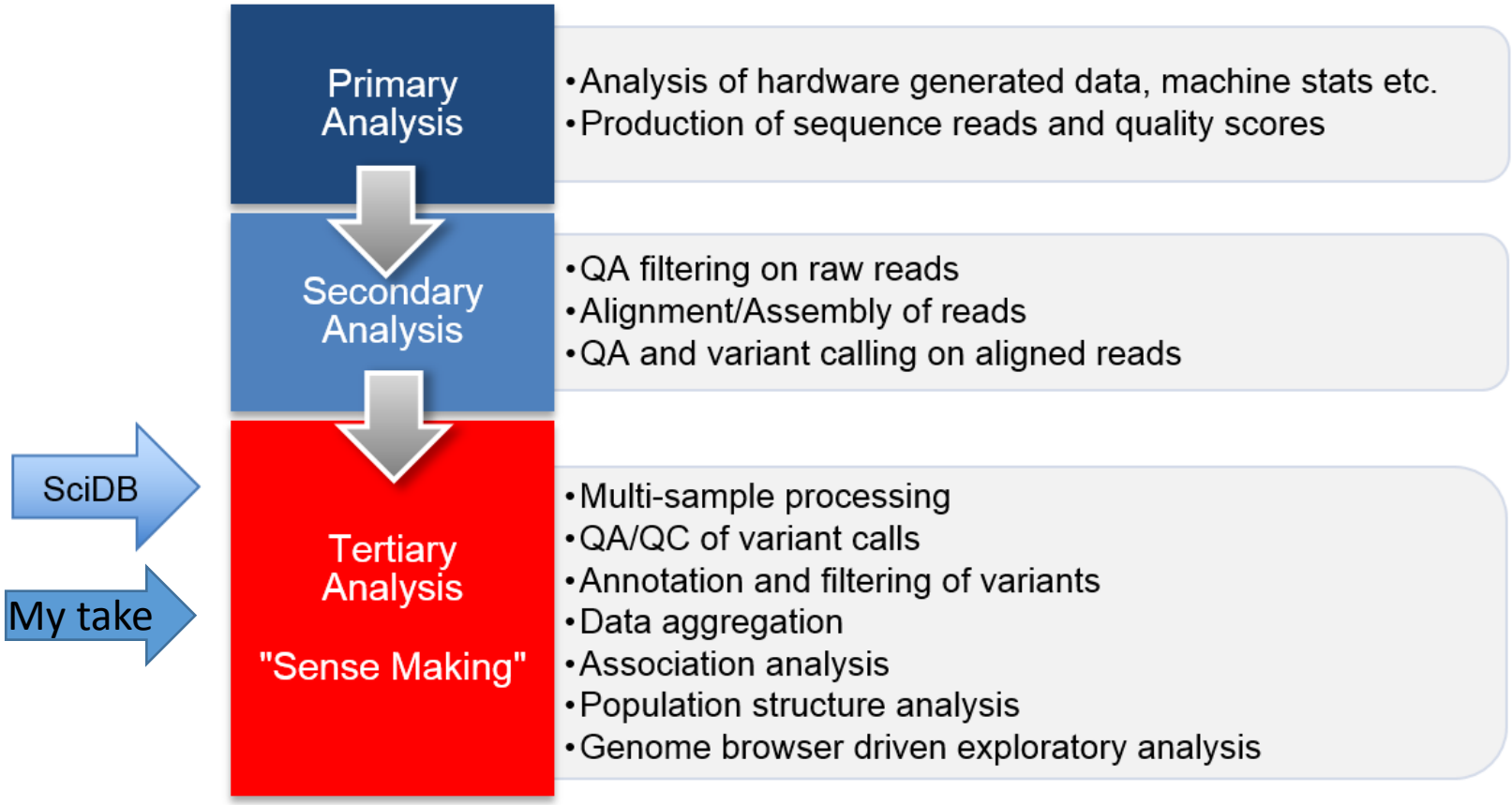
On the relevance of «Regulatory affairs»

The documentation submitted to a regulatory agency by the pharmaceutical company contains all the data generated during the development and test phases. This dossier with the results from chemical-pharmaceutical, toxicological and clinical trials may sometimes amount to capacities of more than 13GB or 500.000 pages. The regulatory agency reviews the documentation to see whether it provides sufficient evidence to prove the efficacy, safety and quality of the drug for the proposed indication.



My catch of todays' big science in biology

Big Data Analysis with Next Generation Sequencing (NGS)



Source: http://blog.goldenhelix.com/grudy/a-hitchhiker%E2%80%99s-guide-to-next-generation-sequencing-part-2/

Public Data

- **1000 Genomes: Deep Catalog of Human Genetic Variation**

The goal of the 1000 Genomes Project is to find most genetic variants that have frequencies of at least 1% in the populations.

- **The Cancer Genome Atlas (TCGA)**

Each cancer undergoes comprehensive genomic characterization and analysis. Generated data are freely available and widely used by the cancer community through the TCGA Data Portal.

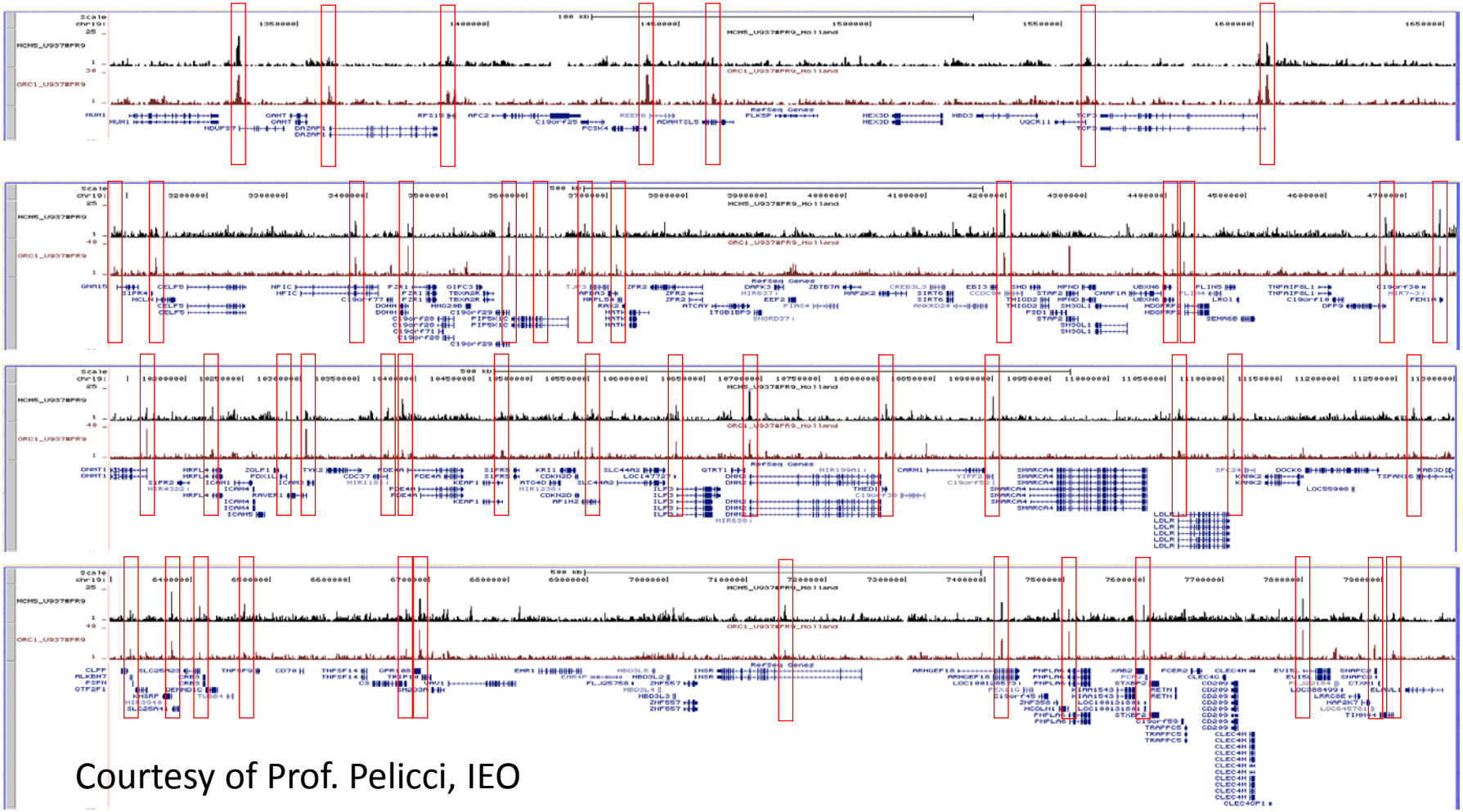
- **100,000 Genomes Project**

This UK project will sequence 100,000 genomes from around 70,000 people. Participants are NHS patients with a rare disease, plus their families, and patients with cancer.

- **ENCODE: Encyclopedia of DNA Elements**

The ENCODE (Encyclopedia of DNA Elements) Consortium is an international collaboration of research groups with the goal to build a comprehensive parts list of functional elements in the human genome, including elements that act at the protein and RNA levels, and regulatory elements that control cells and circumstances in which a gene is active.

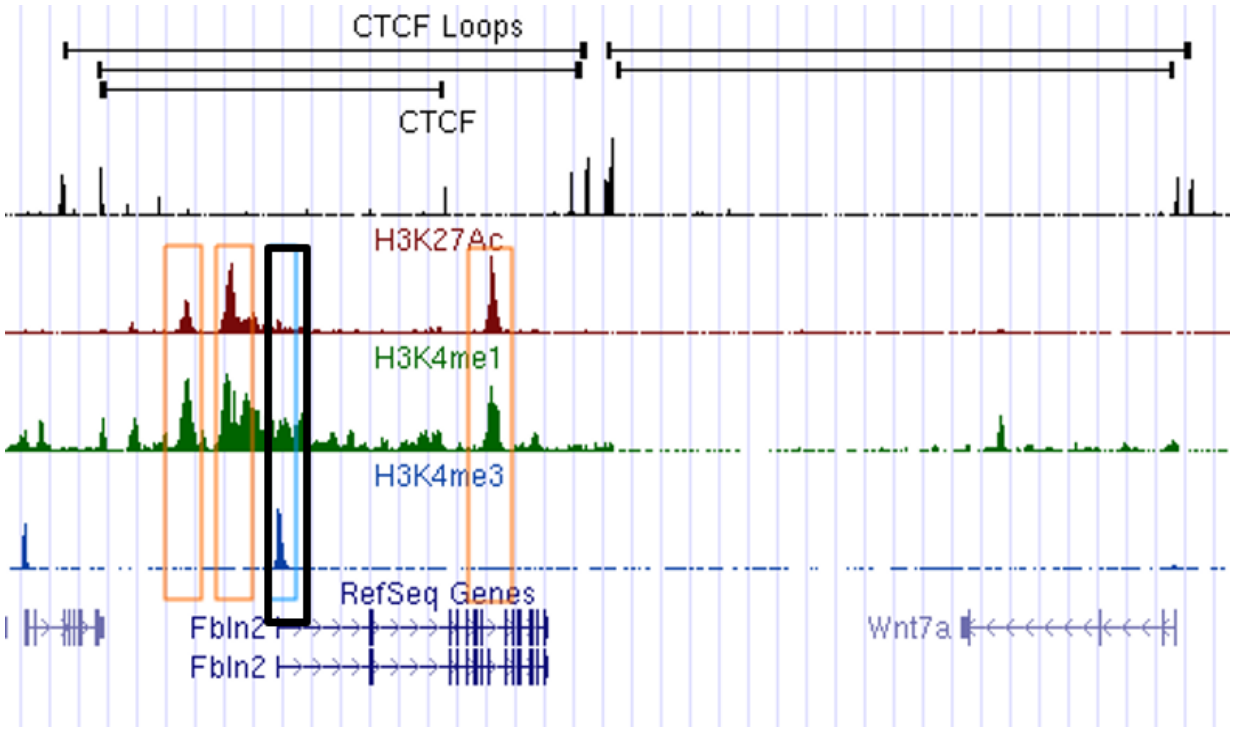
The needle and the haystack



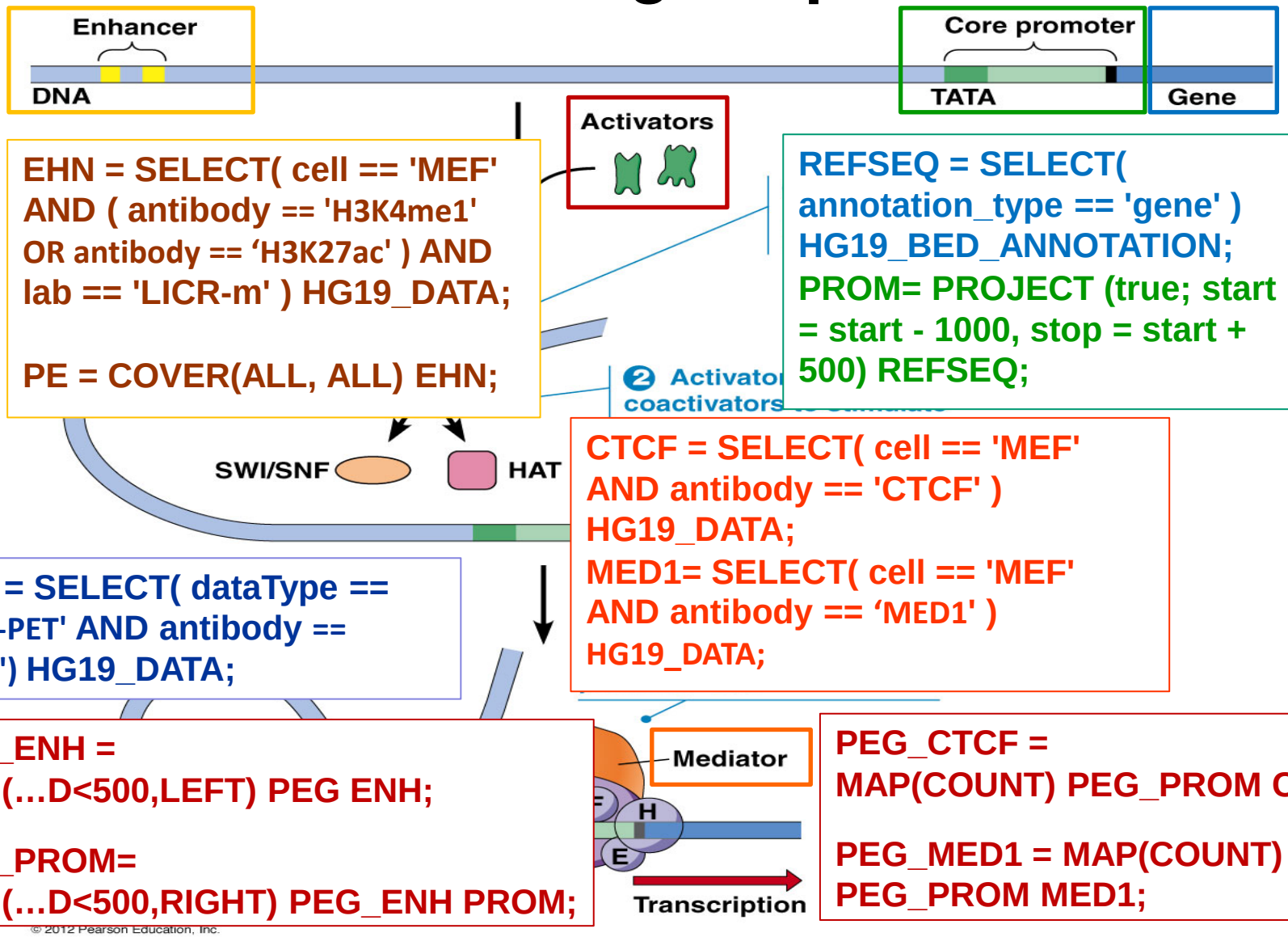
Courtesy of Prof. Pelicci, IEO

Search for patterns within small 3D loops of CTCF

- Yellow area: enhancers
- Blue area: promoters
- Black lines: CTCF loops



GenoMetric Query Language: Abstraction of biological phenomena



GQML operations

Classic relational operations – with genomic extensions

- SELECT, PROJECT, GROUP, ORDER/TOP, UNION, DIFFERENCE, MERGE

Domain-specific genomic operations:

- COVER, GENOMETRIC JOIN, MAP

GQML implementation

Cloud Computing

- VERSION 1: Translation to **PIG** under Hadoop
- VERSION 2: Optimized mapping to **SPARK** and **FLINK** engines

Storing public data from **ENCODE, TCGA, Epigenomic Roadmap**

For interested readers

- M. Masseroli, P. Pinoli, F. Venco, A. Kaitoua, V. Jalili, F. Paluzzi, H. Muller, S. Ceri. **GenoMetric Query Language: A novel approach to large-scale genomic data management**, *Bioinformatics*, 12(4):837-843, 2015.
- M. Bertoni, S. Ceri, A. Kaitoua, P. Pinoli. **Evaluating cloud frameworks on genomic applications**, *IEEE Conference on Big Data Management*, Santa Clara, Nov. 2015.

http://www.bioinformatics.deib.polimi.it/genomic_computing/ (GMQL on Google, - GMQL/)

Back to the topic

Small Science or The “formal”/“complete” approach

- The scientific method is built around testable hypotheses. These models, for the most part, are systems visualized in the minds of scientists. The models are then tested, and experiments confirm or falsify theoretical models of how the world works.
- Scientists are trained to recognize that correlation is not causation, that no conclusions should be drawn simply on the basis of correlation between X and Y (it could just be a coincidence), that “data without a model is just noise.”

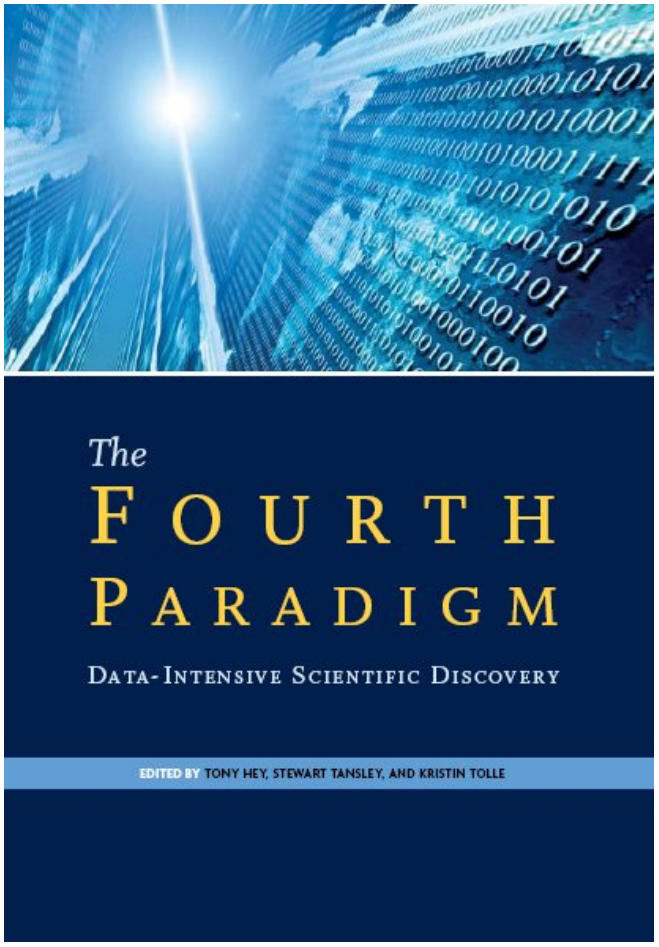
Big Science or Data driven approach

- Faced with massive data, the classic approach to science — hypothesize, model, test — is becoming obsolete. Petabytes allow us to say: "Correlation is enough."
- "**We can stop looking for models.** We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot." (Chris Anderson, Wired Ed. In Chief)

The big dilemma: can data-driven science stop looking for models?

- Moshe's talk: "the data-driven approach does not replace the formal-model approach"; in his two experiences, "the data-driven approach stands on the shoulders of the formal-model approach."
  - But: how many experiences like that? How many Moshe Vardi are around us?

Where it all started: The Fourth Paradigm



A tribute to Jim Gray in our youths

Why Four?

- First: empirical science and observations
- Second: theoretical science and mathematically-driven insights
- Third: computational science and simulation-driven insights
- Fourth: data-driven insights of modern scientific research.

The data perspective: Jim Gray's words

When people use the word **database**, fundamentally what they are saying is that the data should be **self-describing** and it should have a **schema**. That's really all the word database means.

So if I give you a particular collection of information, you can look at this information and say, "I want all the genes that have this property" or "I want all of the stars that have this property" or "I want all of the galaxies that have this property."

But if I give you just a bunch of files, you can't even use the concept of a galaxy and you have to hunt around and figure out for yourself what is the effective schema for the data in that file.

If you have a schema for things, you can index the data, you can aggregate the data, you can use parallel search on the data, you can have ad hoc queries on the data, and it is much easier to build some generic visualization tools.

My take on data design for «big data»

- Along Jim: even «big data» need some «structure» and a minimal level of data design, by assessing:
  - that data are self-described with a schema
  - that data are of «sufficient quality»
- But «big data» studies are bottom-up (data exists before being designed), therefore:
  - the best conceptual models which are built top-down usually don't fit – and nobody understands them
  - they need data integration which is a «lost war» of data management community
- In other words: data theory+abstractions are loosing ground but aren't totally dead.

Big science and education

An educational model of big science is emerging

- Pushing math-stats, data mining, machine learning.
- Problem-driven
- Traditional CS models used when/if needed but no longer the key foundational aspect of the curriculum.

Harvard: Master of Science in Computational Science and Engineering (CSE)

"What should a graduate of our CSE program be able to do?"

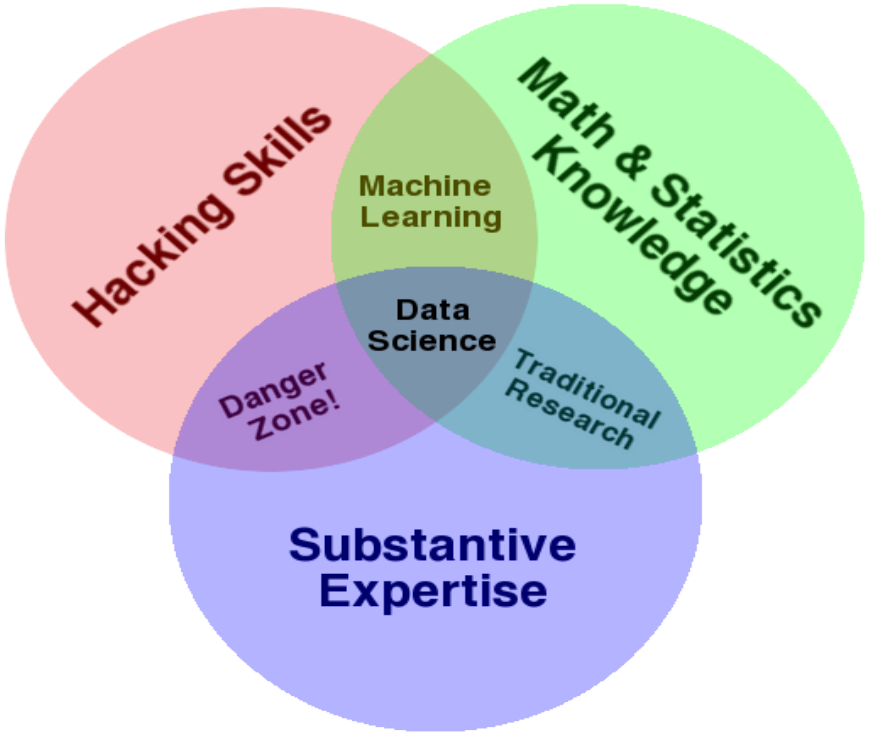
- Frame a real-world problem such that it can be addressed computationally
- Evaluate multiple computational approaches to a problem and choose the most appropriate one
- Produce a computational solution to a problem that can be comprehended and used by others
- Communicate across disciplines
- Collaborate within teams
- Model systems appropriately with consideration of efficiency, cost, and the available data
- Use computation for reproducible data analysis
- Leverage parallel and distributed computing
- Build software and computational artifacts that are robust, reliable, and maintainable
- Enable a breakthrough in a domain of inquiry

Many other one-year masters' in «big data» (e.g. PoliMi, Pisa, Bologna, …)

- Emphasis on:
- Problem-driven approach – first frame the problem, then choose the method
- Computational aspects (machine learning) and statistical methods (correlation/significance) highlighted
- «Business orientation»: where is the enterprise value
- «Story telling»: how to present (e.g. visualize) data

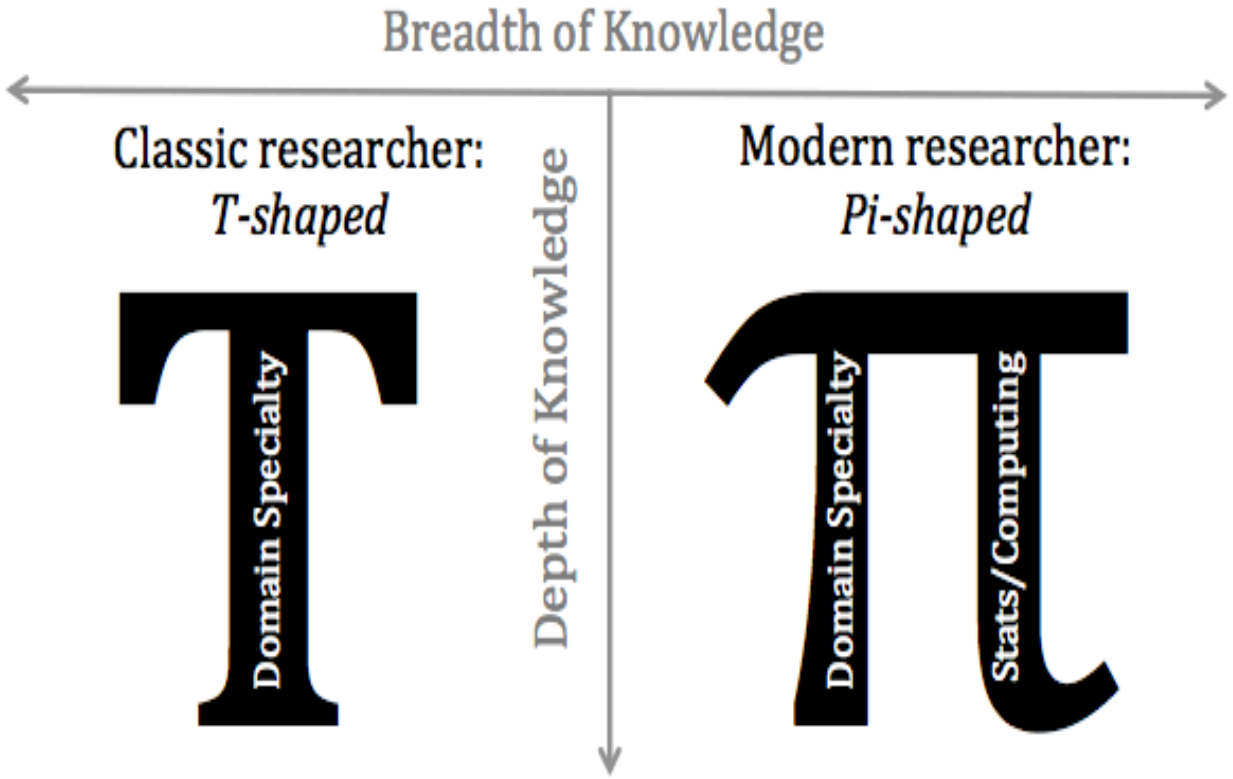
(my take: in one-year program there is little room for «models»)

Big Science and education: A visual view of Big Data Skills



- From: DrewConway.com, retrieved 25/6/2015

Data Scientists: T-shaped and Pi-shaped



Data Science and Academic Recognition

Why Data Science may not fit in Academia

- Scientific research more and more dependent on the careful analysis of large datasets, requiring a skill-set as broad as it is deep: scientists must be experts not only in their own domain, but in statistics, computing, algorithm building, and software design.
- Academia's reward structure is not well-poised to reward the value of this type of work.
- Time spent developing high-quality reusable software tools translates to less time writing and publishing, which under the current system translates to little hope for academic career advancement.
- Jake Vanderplas - Oct 26, 2013

Why industry is a better fit for data scientist

- Salary
- Stability / Opportunity for Advancement
- Respect of Peers
- Opportunity to work on open source software projects
- Flexibility to work on interesting projects
- Opportunity to travel & attend conferences
- Opportunity to publish / freedom from the burden of publishing
- Opportunity to teach / freedom from the burden of teaching
- Opportunity to mentor students / freedom from the burden of mentoring students

Fixing the value system to defend the data scientists' career

- press the importance of reproducibility in academic publication
- push for a new standard for tenure-track evaluation criteria
- create and fund positions which emphasize and reward the development of open, cross-disciplinary scientific software tools
- increase the pay of post-doctoral scientific research positions

Where Data Science should be housed by Academia

- each department should have its own training to data sciences
- part of applied computer science
- a consulting service within universities
- the natural location of interdisciplinary studies
- a new role for data curation (instead of libraries)

Conclusion: my take on Computer Science Education (and beyond)

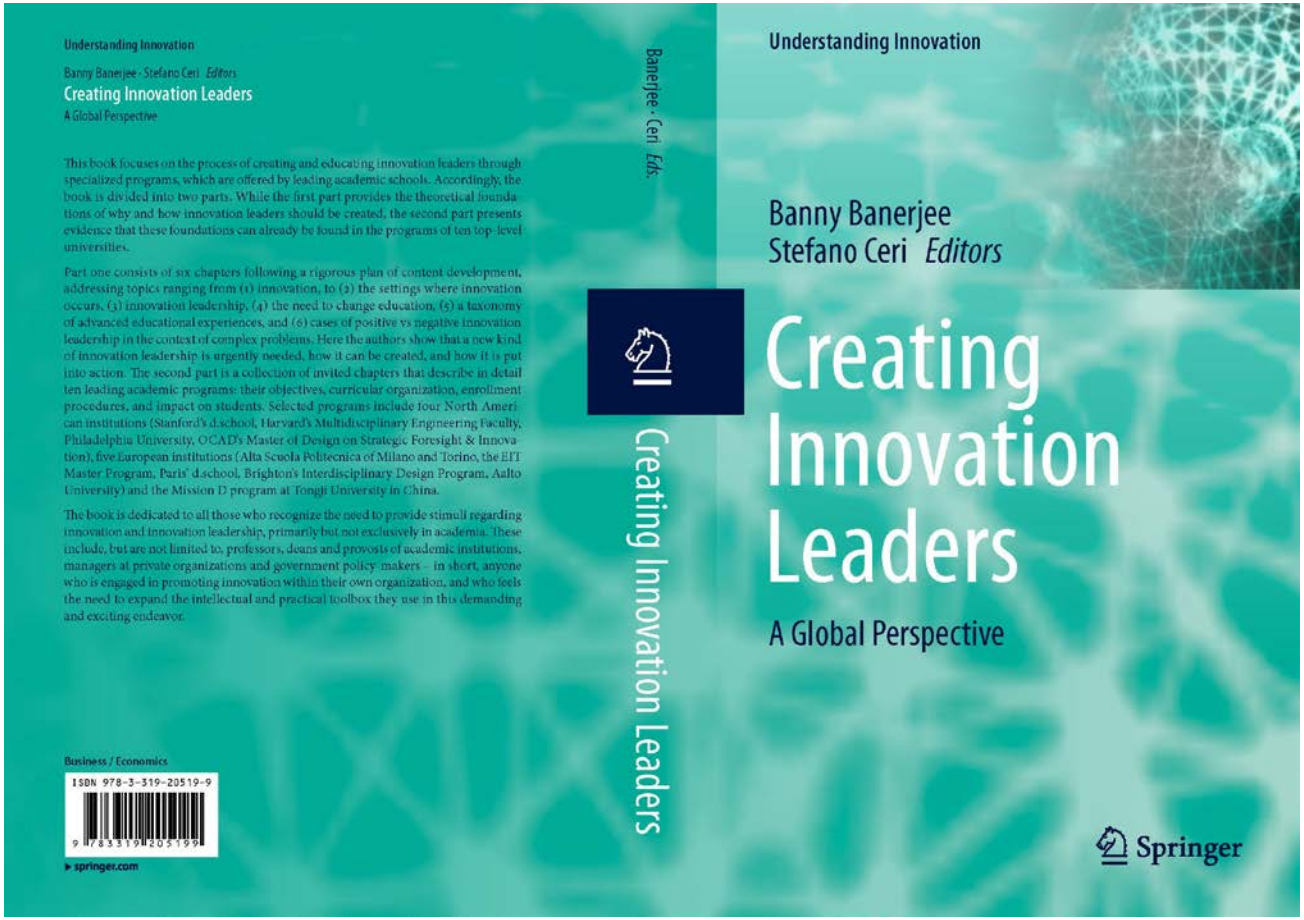
Where we stand:

- A growing discipline, beyond the ten-year-ago decline:
  - lots of undergraduate and graduate students
  - lots of jobs and lots of postions-to-be-filled
  - Still not too attractive: «nerds» are not popular within the brighests high-school graduates
- We can be self-referential
  - Enough models , methods and technologies belong to computer science

So why changing?

- CS's current challenge: multidisciplinarity
  - After >50 years of CS disciplinarity, we can finally face the challenge
  - But: standing instead of leaning…
- CS's unique feature: can be radically problem-driven.
  - In the small: when teaching just 2 hrs/semester of programming 101 to general students
  - In the large: when you have the breadth of a 1-2 year curriculum

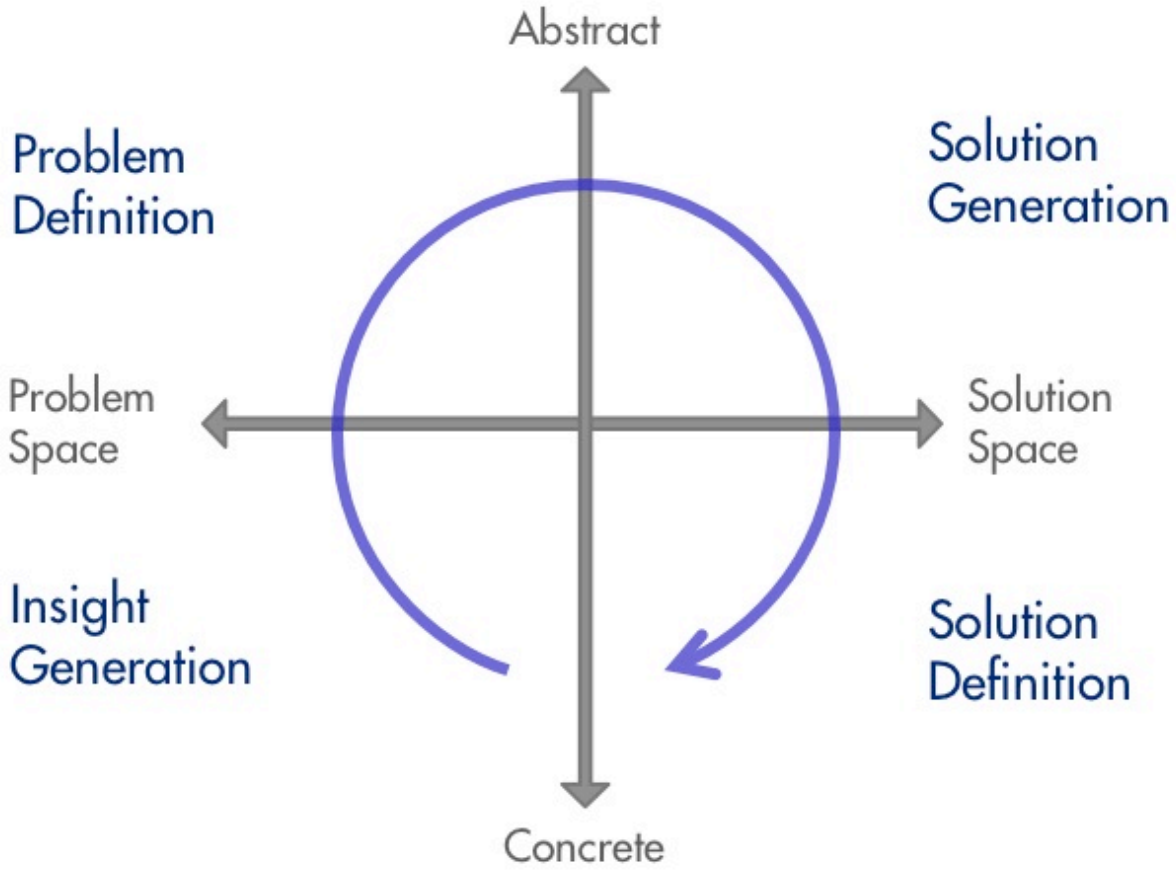
Change direction in education: creating innovation leadership



Innovation Leadership?

- Contrasting conventional (hierarchical) leadership
- A modality that involves fulfilling certain functions in the context of an organizational, institutional or project context.

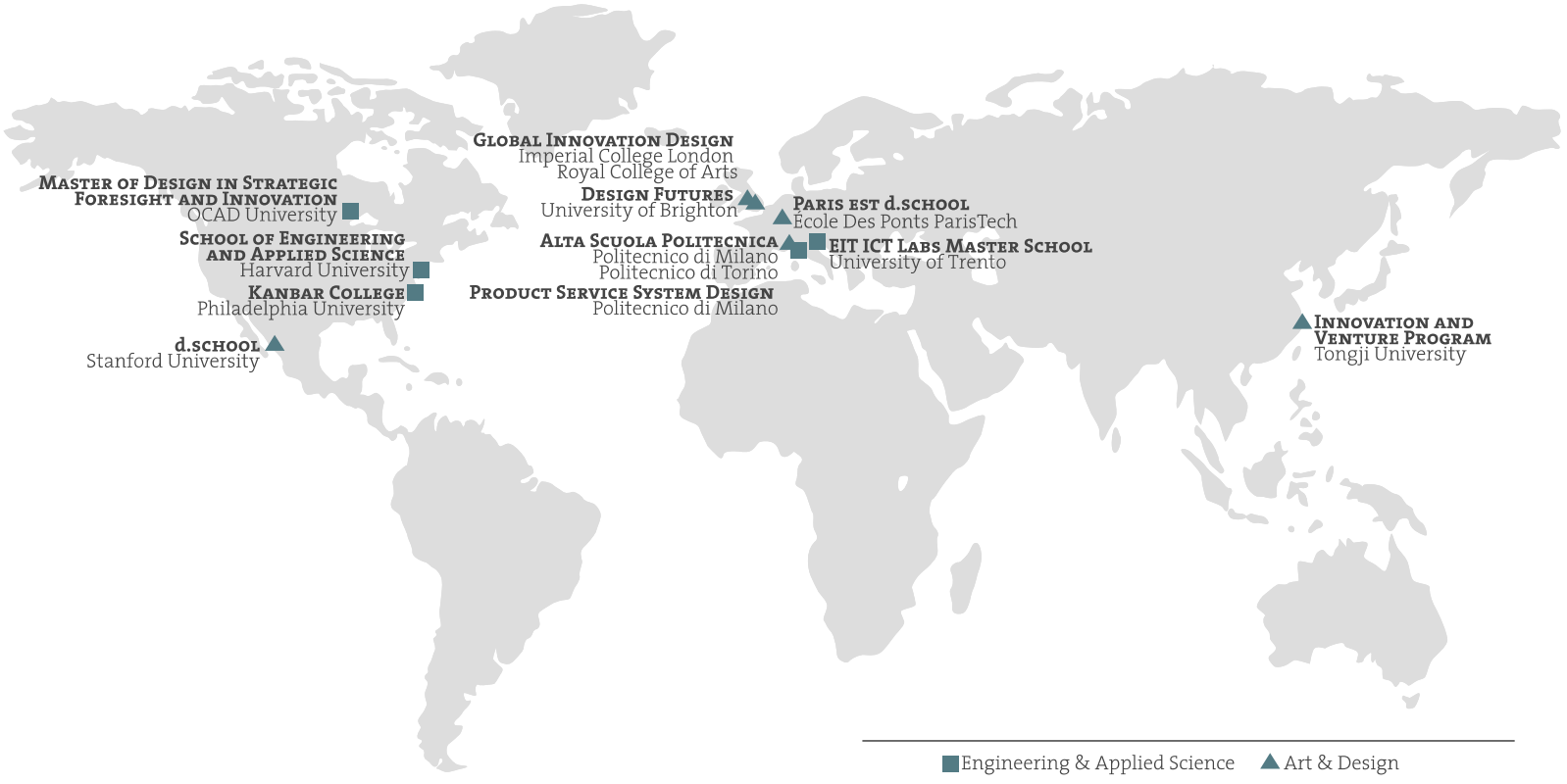
The innovation space



Implication on education

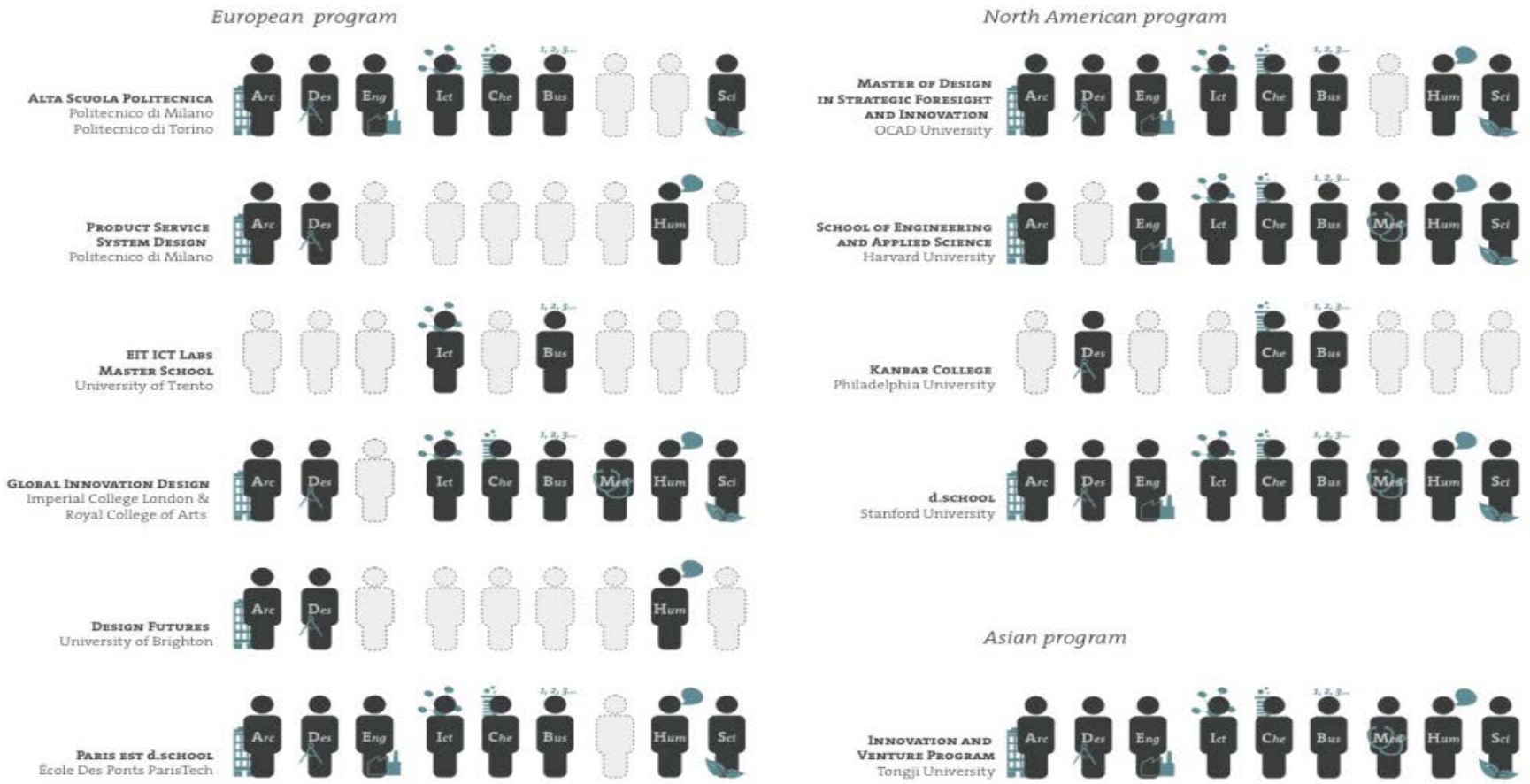
- The university should be learning centric.
- Disciplinary expertise should be valued, but equally important is the student's ability to find and solve problems by actively integrating many kinds of knowledge from disparate sources.
- Learning should be cooperative. The instructor guides the student during the knowledge integration process.
- Key ingredients include soft knowledge, such as collaboration and teamwork, decision-making, and leadership.

Several similar schools around the world emphasize these values

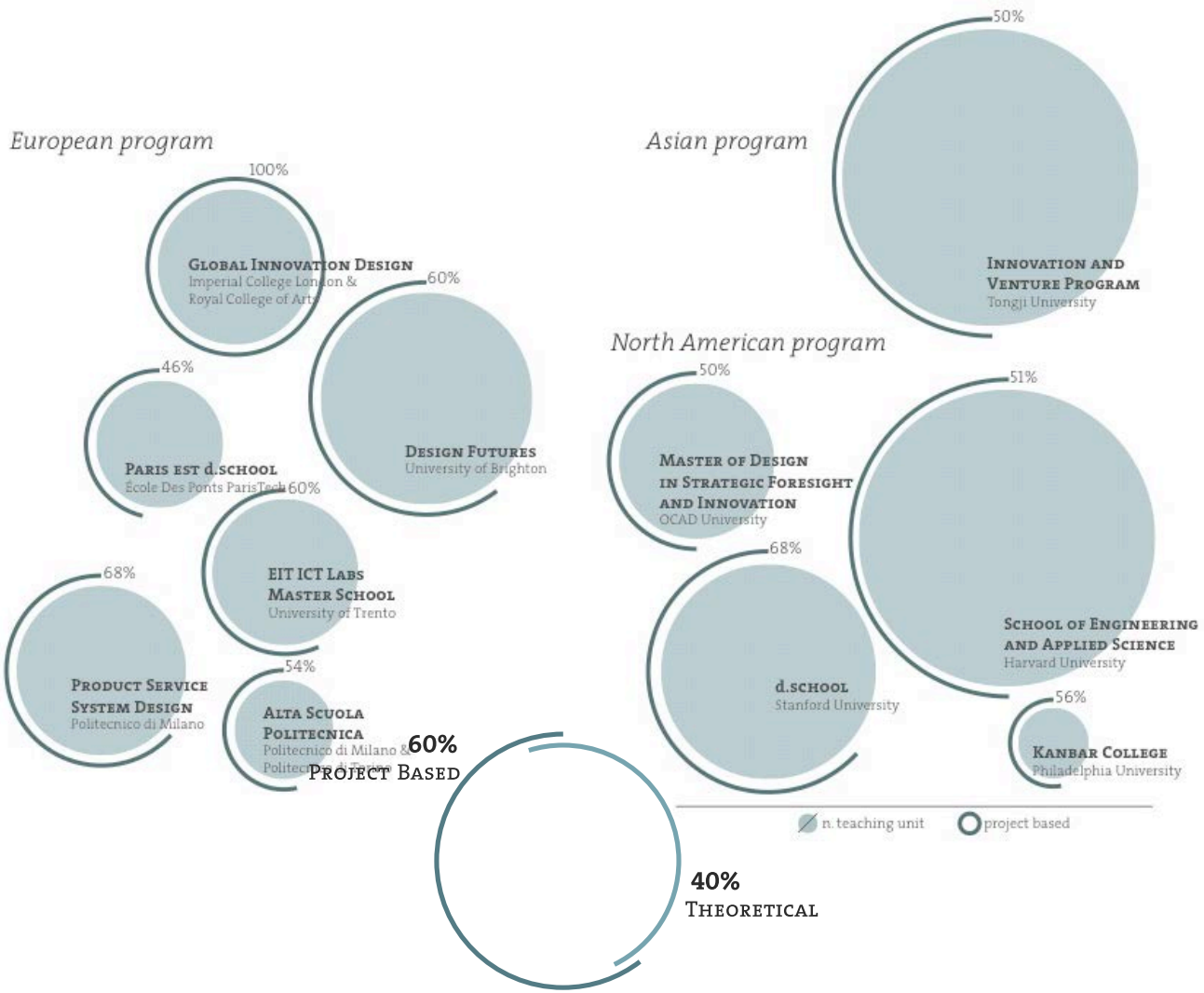


… share common drivers…

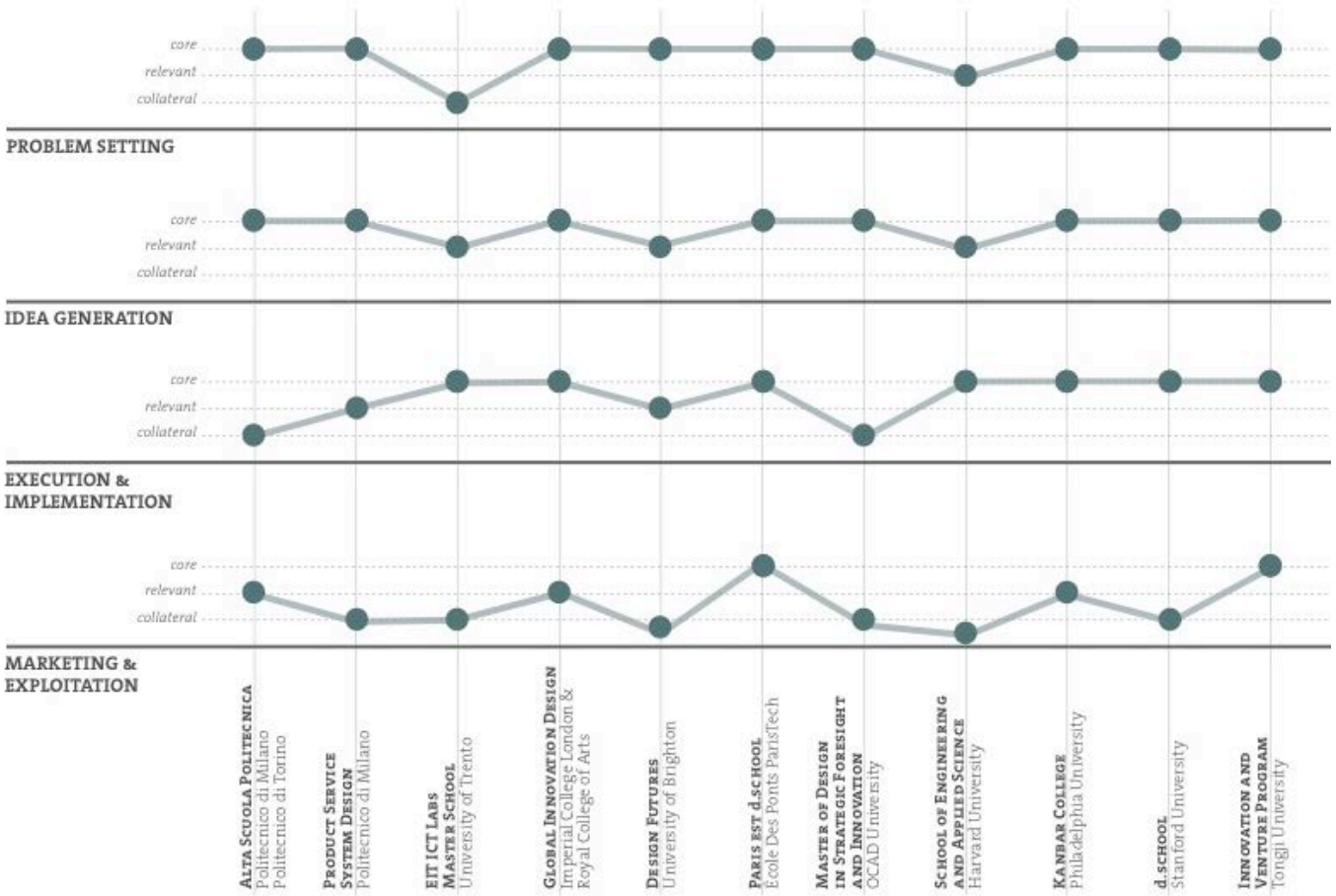
.. with emphasis on interdisciplinarity...



… and a strong problem-based approach…

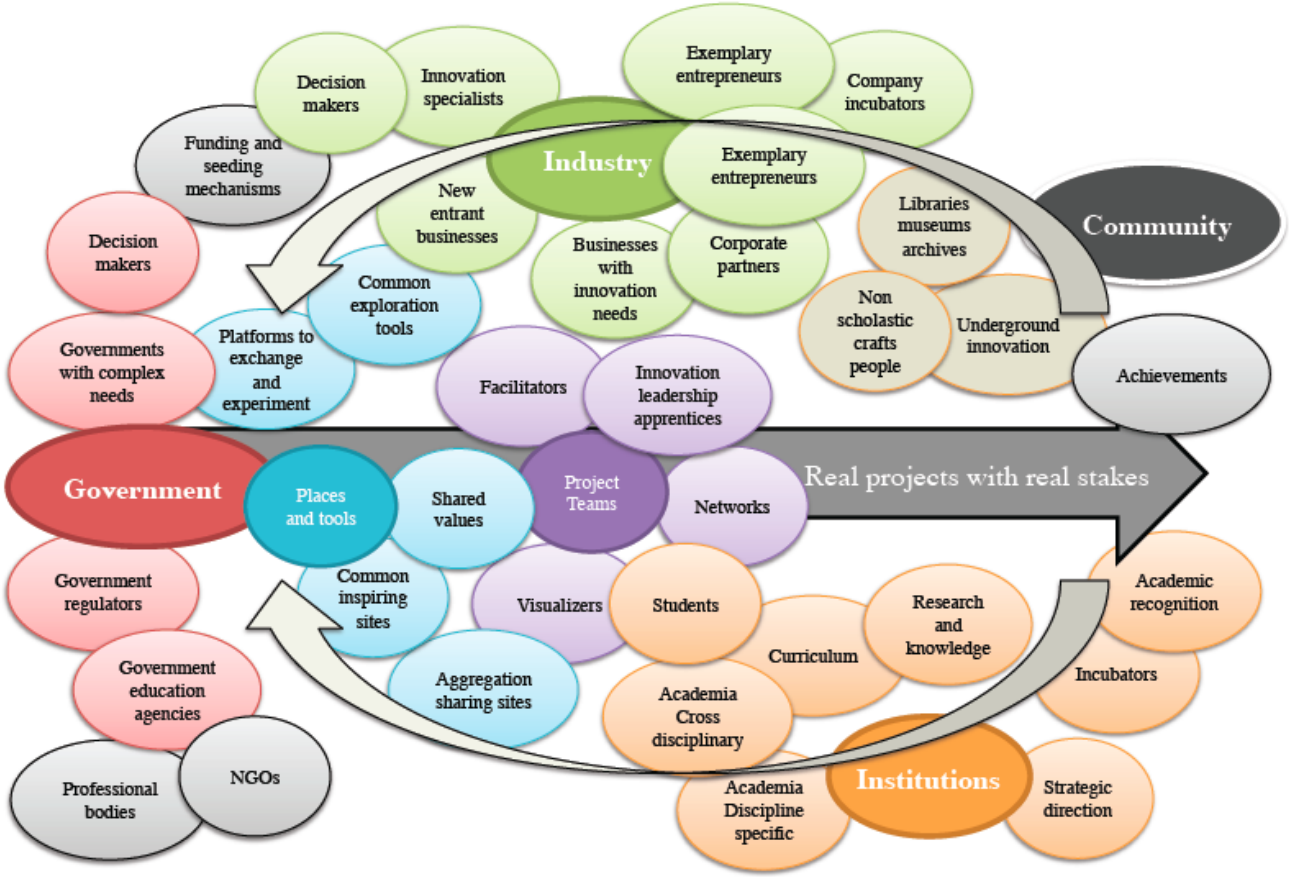


…with more emphasis on problem setting and idea generation…



..calling for a new model of leadership…

..placed within a complex ecosystem



(where humans are the essential ingredient)