



BIG DATA EUROPE

Empowering Communities
with Data Technologies

On the need for intelligent access to big data in life sciences



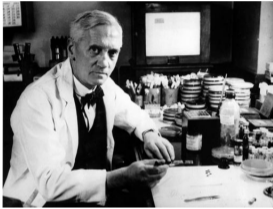
The **BioA2Q** experience

21-05-2015

Georgios Paliouras, NCSR "Demokritos"

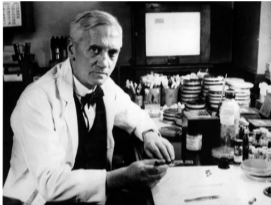


Life sciences then and now





Life sciences then and now






These data are too big for me!



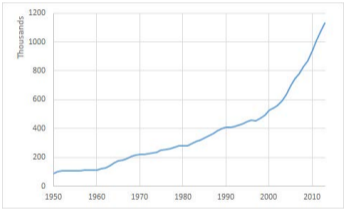
"After careful consideration of all 437 charts, graphs, and metrics, I've decided to throw up my hands, hit the liquor store, and get snocked. Who's with me?!"

By Anderson for 

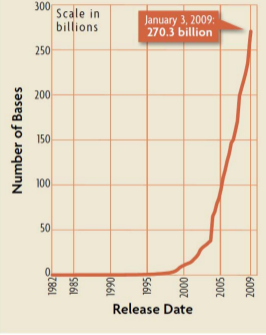


Data growth is exponential

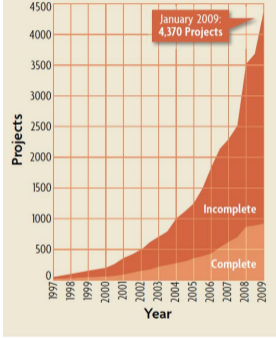
Number of articles indexed by MEDLINE (PUBMED) per year



Growth Rate of EMBL-Bank



Genome Sequencing Projects on GOLD

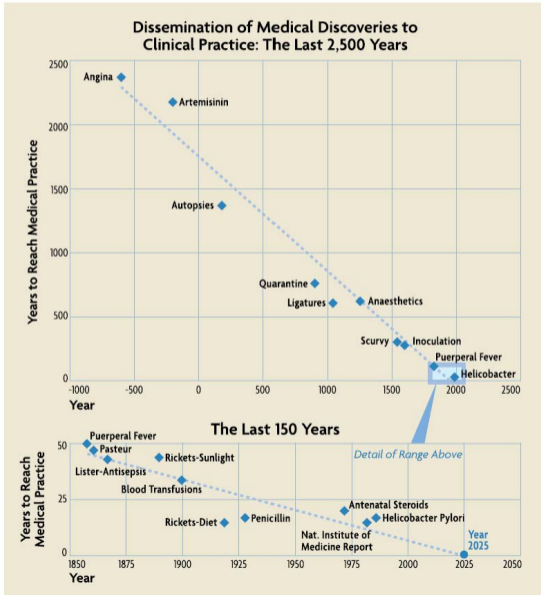


<http://dan.corlan.net/medline-trend.html>

Southan and Cameron, *Beyond the tsunami: developing the infrastructure to deal with life sciences data*, The fourth Paradigm: Data-Intensive Scientific Discovery, Microsoft Corp., 2009.



Data is source of knowledge



Gillam et al., *The healthcare singularity and the age of semantic medicine*, The fourth Paradigm: Data-Intensive Scientific Discovery, Microsoft Corp., 2009.



One just needs to make the connection

The Swanson case: Fish oil and Raynaud's syndrome

- Public knowledge since 1975: Raynaud's syndrome is associated with high blood viscosity, platelet aggregability, vasoconstriction.
- Public knowledge since 1984: Fish oil leads to reductions in blood lipids, platelet aggregability, blood viscosity, and vascular reactivity.
- Swanson puts the two together in 1986: Can dietary fish oil ameliorate or prevent Raynaud's syndrome? He supports his evidence with relevant literature.
- DiGiacomo confirms the hypothesis in 1989.

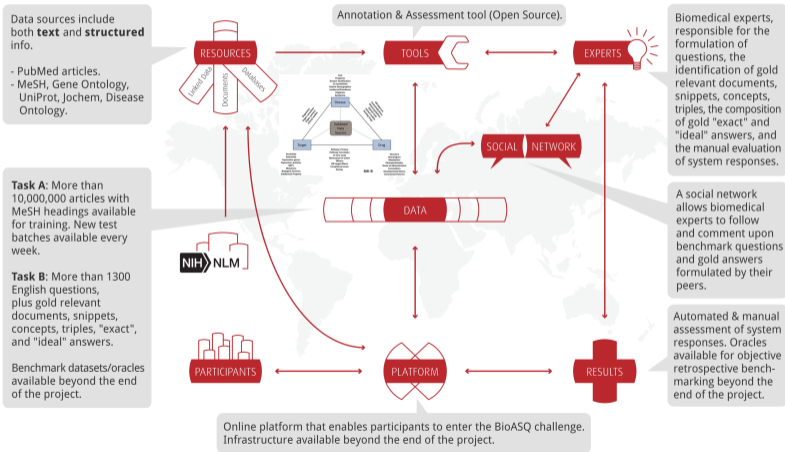
Vision: Create big data machinery that help produce and support more such cases.



- 2 articles published in biomedical journals **every minute!**
- Make sure this knowledge is used to the benefit of patients
- Need to make it accessible to biomedical experts
- Search is not effective enough
- Push research in automated answering of questions
- A challenge for such systems can achieve a multiplying effect



BioASQ ecosystem



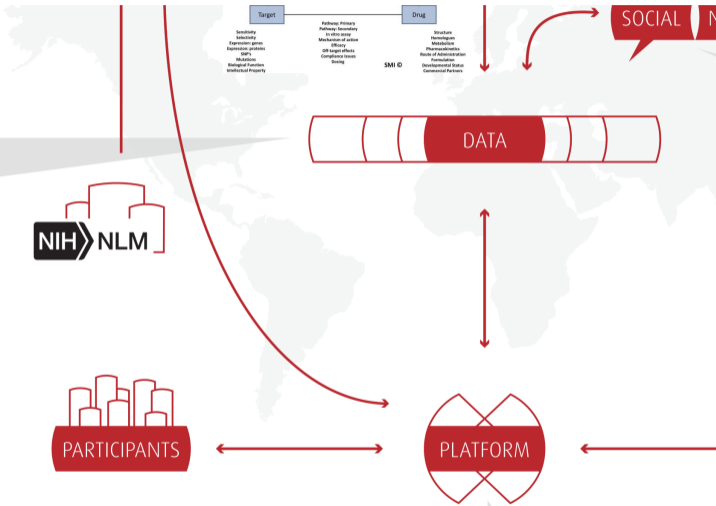


BioASQ ecosystem

Task A: More than 10,000,000 articles with MeSH headings available for training. New test batches available every week.

Task B: More than 1300 English questions, plus gold relevant documents, snippets, concepts, triples, "exact", and "ideal" answers.

Benchmark datasets/oracles available beyond the end of the project.



Online platform that enables participants to enter the BioASQ c
 Infrastructure available beyond the end of the project.



Data sources include both **text** and **structured** info.

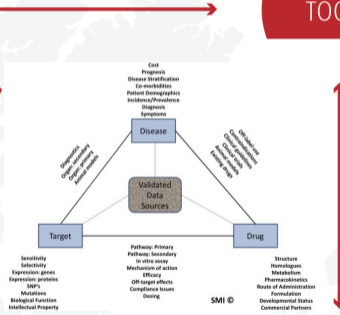
- PubMed articles.
- MeSH, Gene Ontology, UniProt, Jochem, Disease Ontology.

Task A: More than 10,000,000 articles with MeSH headings available for training. New test



Annotation & As

TOO

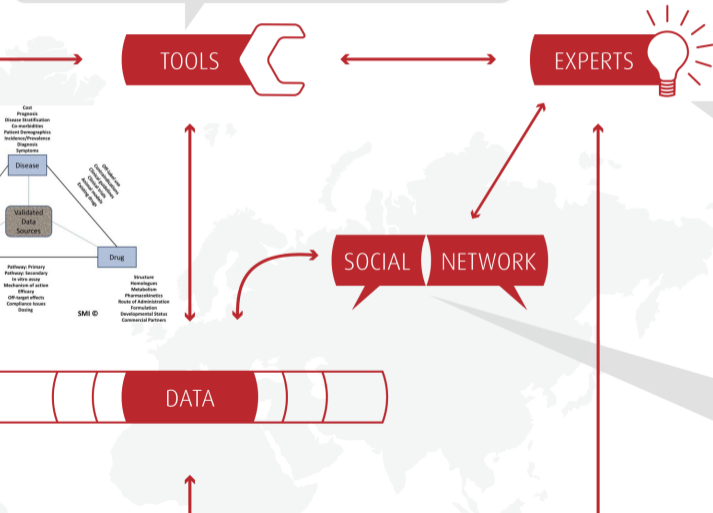


DATA



BioASQ ecosystem

Annotation & Assessment tool (Open Source).



Biomedical experts, responsible for the formulation of questions, the identification of gold relevant documents, snippets, concepts, triples, the composition of gold "exact" and "ideal" answers, and the manual evaluation of system responses.

A social network allows biomedical experts to follow and comment upon benchmark questions and gold answers



Talking to BioASQ experts

“as I’m growing older . . . I spend more time in front of the computer but I learn less. . . . the complexity has increased, the variety has increased and my time has been reduced.”

“When I do research I use IT stuff all the time, I’m looking for papers and data...I’m also doing statistical analysis”



Talking to BioASQ experts

"as I'm growing older . . . I spend more time in front of the computer but I learn less. . . . the complexity has increased, the variety has increased and my time has been reduced."

"When I do research I use IT stuff all the time, I'm looking for papers and data...I'm also doing statistical analysis"

"PubMed and all this of course, we really depend on that. We cannot work if we don't search in those."

"The bulk of information, that's the main problem. . . .if someone has some extra time and starts reading the results of a search then this might never end!"

"Sometimes you get irrelevant results. That's the main problem."



Talking to BioASQ experts

"as I'm growing older . . . I spend more time in front of the computer but I learn less. . . . the complexity has increased, the variety has increased and my time has been reduced."

"When I do research I use IT stuff all the time, I'm looking for papers and data...I'm also doing statistical analysis"

"PubMed and all this of course, we really depend on that. We cannot work if we don't search in those."

"The bulk of information, that's the main problem. . . .if someone has some extra time and starts reading the results of a search then this might never end!"

"Sometimes you get irrelevant results. That's the main problem."

"There is abundance of structured information . . . Unfortunately not all structured databases are included into one."

"I am looking at least into twenty different places for the same protein."

". . . since I use a number of different programs I forget them by the time I want to use them again and I have to remember them once more."



(Vision) Information systems that act like peers to human experts:



- understand the information need of the expert
- represent the need in machine-readable format
- match it to the information and data available in various sources
- provide comprehensive and comprehensible response, with supporting material

(Big data added value) Integration of information from many sources and large-scale semantic indexing.

(Outlook) Long way ahead but the impact of even marginal progress on public health can be very significant!



Where do we stand?

- Big data is getting linked
- We have a range of tools for analysing and indexing such data
-  BDE is set to bring the pieces together
- Challenges, such as  push research further; NLM has improved their MeSH indexing engine by 5%, in the first year of BioASQ!
- IBM Watson to be put in use by 14 US cancer research institutes
- Robotic science assistants making their appearance; “Adam” generating functional genomics hypotheses about the yeast *Saccharomyces cerevisiae*



Further information

www.big-data-europe.eu

www.bioasq.org, participants-area.bioasq.org

Follow @BigData_Europe, @BioASQ

