

Fourth anniversary  
Gold Open Access  
in the field of AI

[Open for submissions >](#)



**Thomas Hartung & team**

***Big Data/AI and their  
role in regulatory  
decision-making***

**Slides available:**

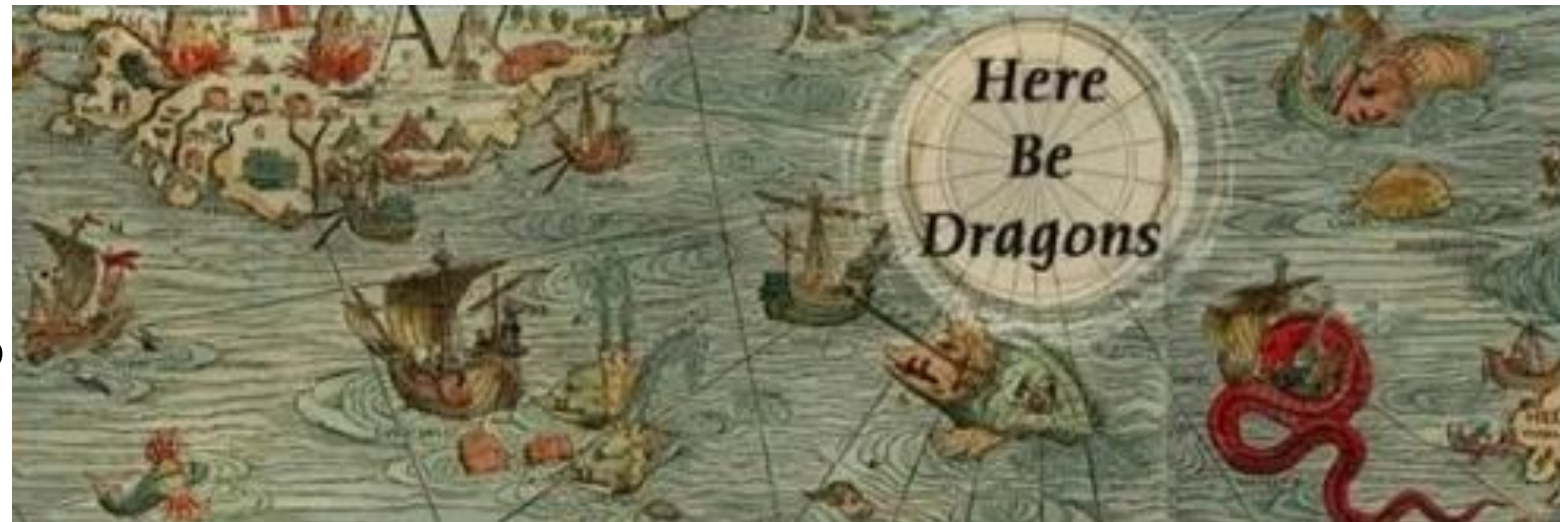




# The toxicological world

**Large parts unknown**

**A few known and  
imagined monsters**



# A.I. = Making big sense of

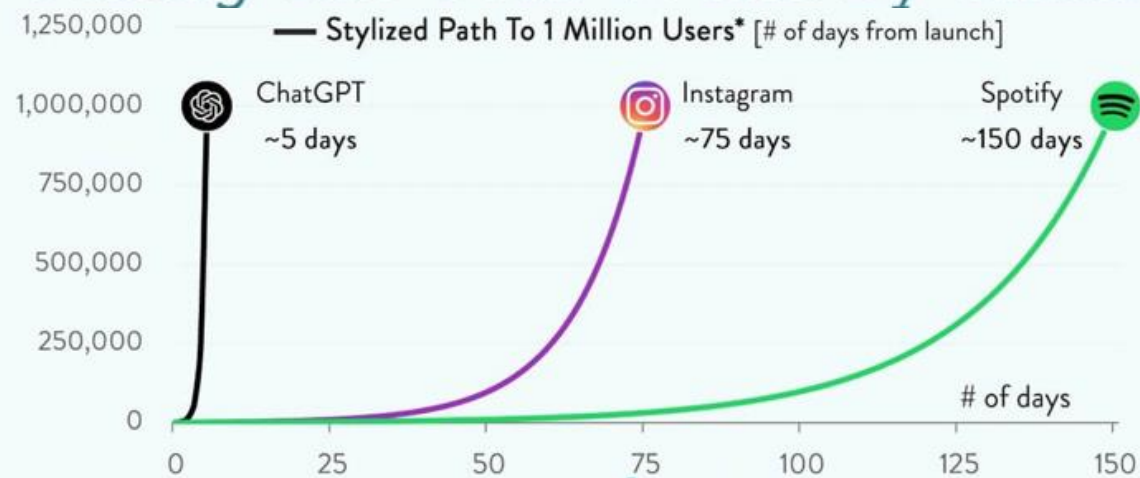


If data is the  
new oil (Clive  
Humby, 2006),  
AI is the new  
combustion  
engine!

<https://theamericangenius.com/editorials/big-data-is-watching-you-some-will-panic-others-will-rejoice/>

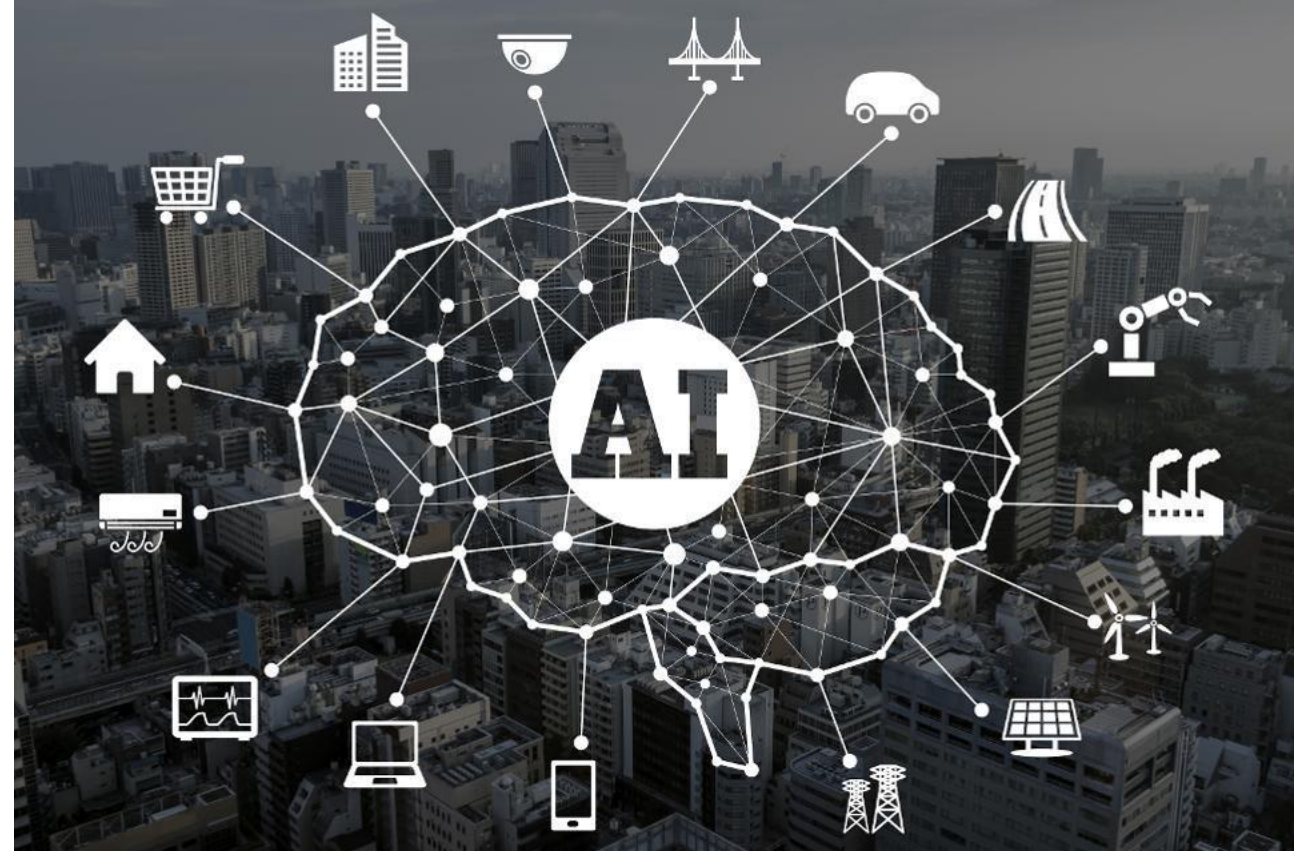
# GPT-3 fastest technology roll-out in history

## ChatGPT From OpenAI Is A Bot Taking The Tech World By Storm



Sources: Google, Subredditstats, Media Reports

charttr \*Path is stylized to the 1m milestone

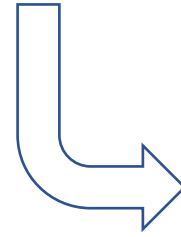


**AI industry expected to increase by over 13x over the next 7 years (CAGR of 38%).**

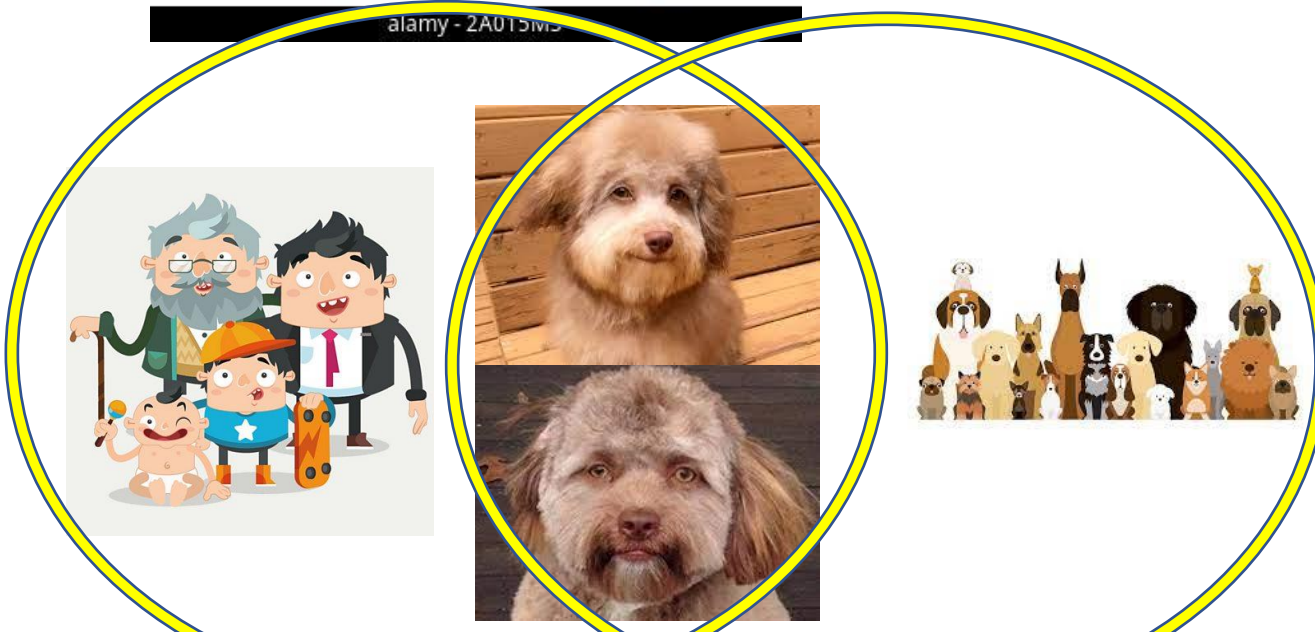
# Unsupervised vs. supervised AI



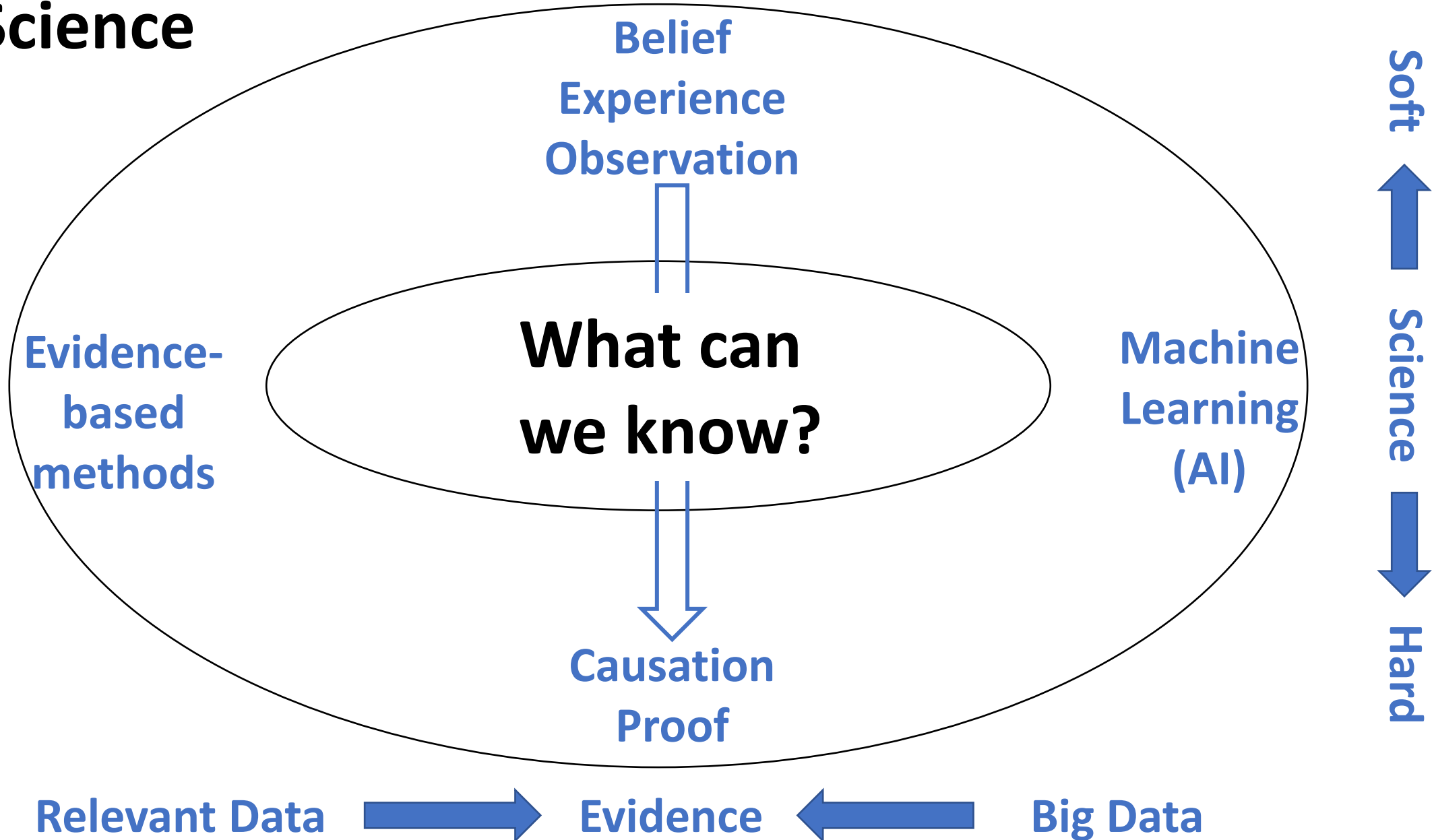
Autonomous  
clustering



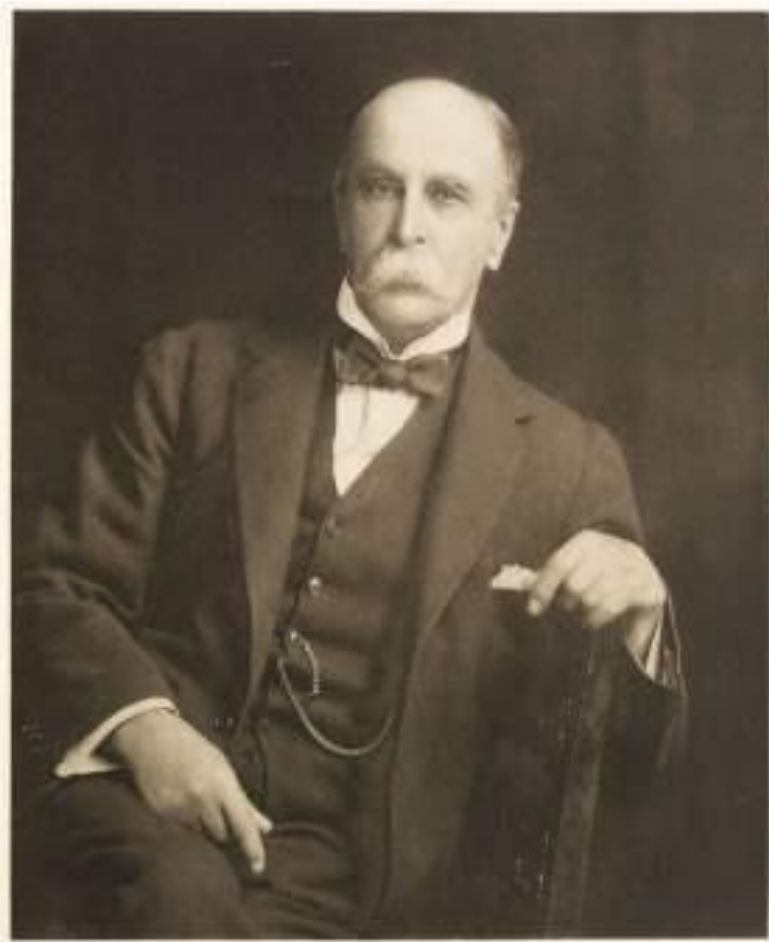
Human  $\leftrightarrow$  Dog  
Toxic  $\leftrightarrow$  Non-toxic



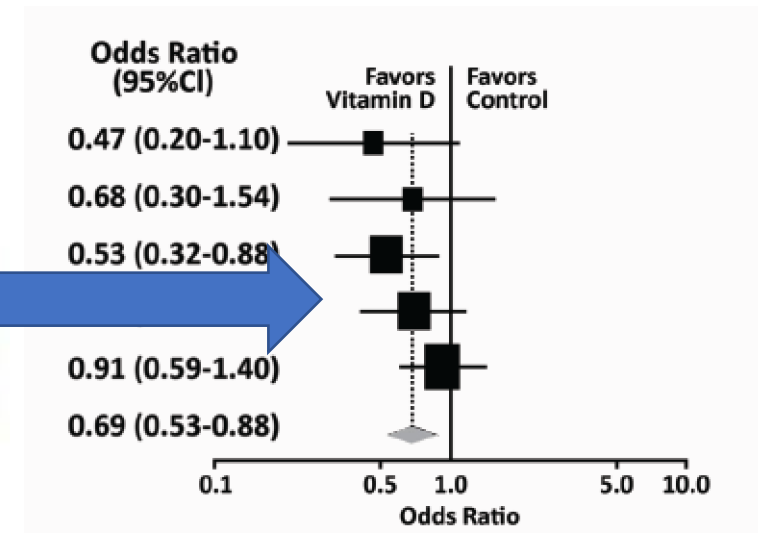
# Science



*“Medicine is a science of uncertainty  
and an art of probability”*  
**Sir William Osler (1849-1919)**

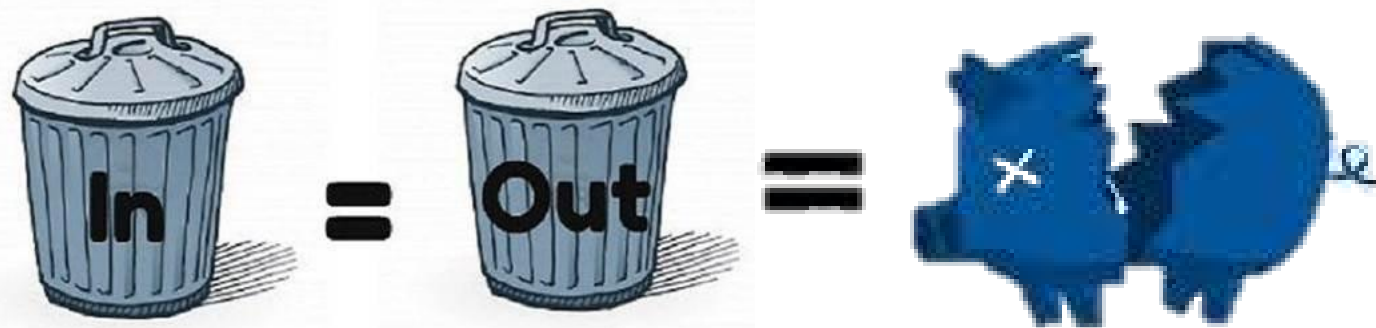


*“Evidence-based medicine: A science of  
uncertainty and an art of probability”*



**Forrest Plot of  
probabilities**

A  
B



**The question is only  
who sorts the garbage**

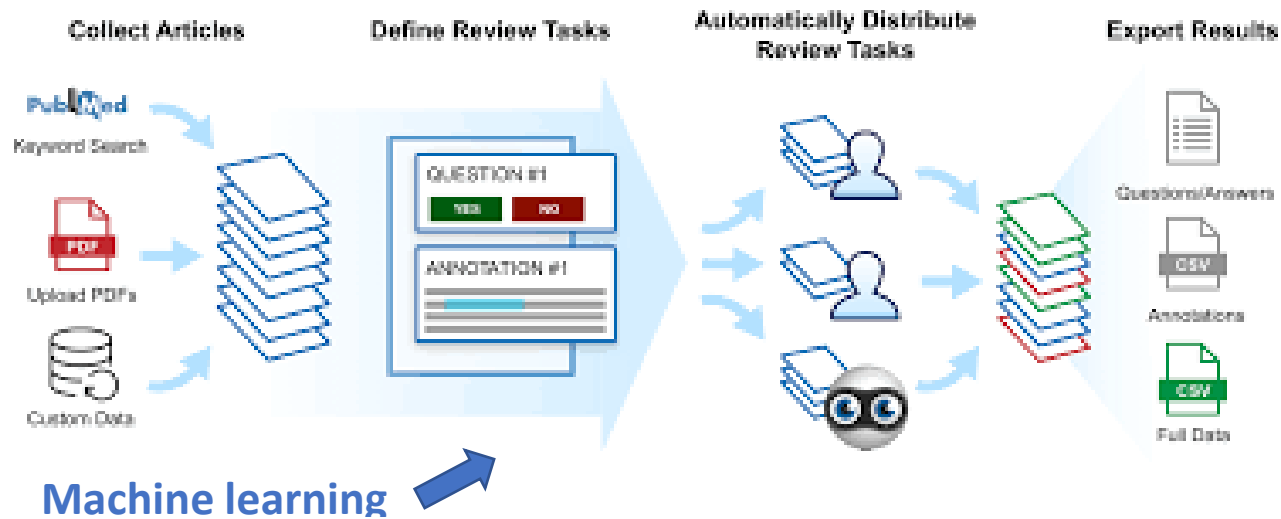




# Data extraction and mining as key part of ONTOX

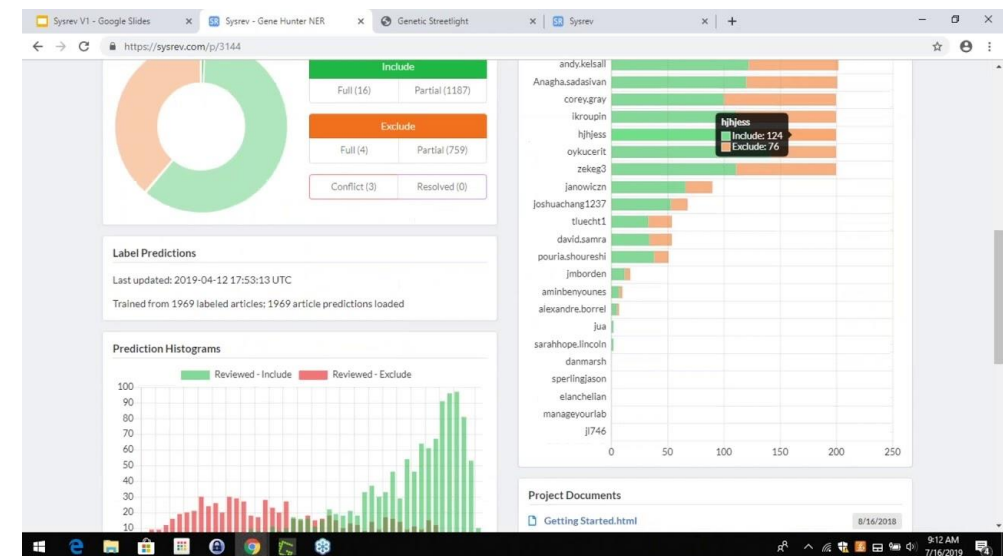


# ToxTrack



## Semi-automated systematic review:

- Auto-extract /annotate papers
- Auto-analyze clustering of papers
- Learn from manual inclusion / exclusion
- Automated inclusion / exclusion suggestions
- Natural Entity Recognition & Causal Relationship Extraction
- Feed into ontologies and AI
- chatGPT -> bioGPT -> toxGPT (?)



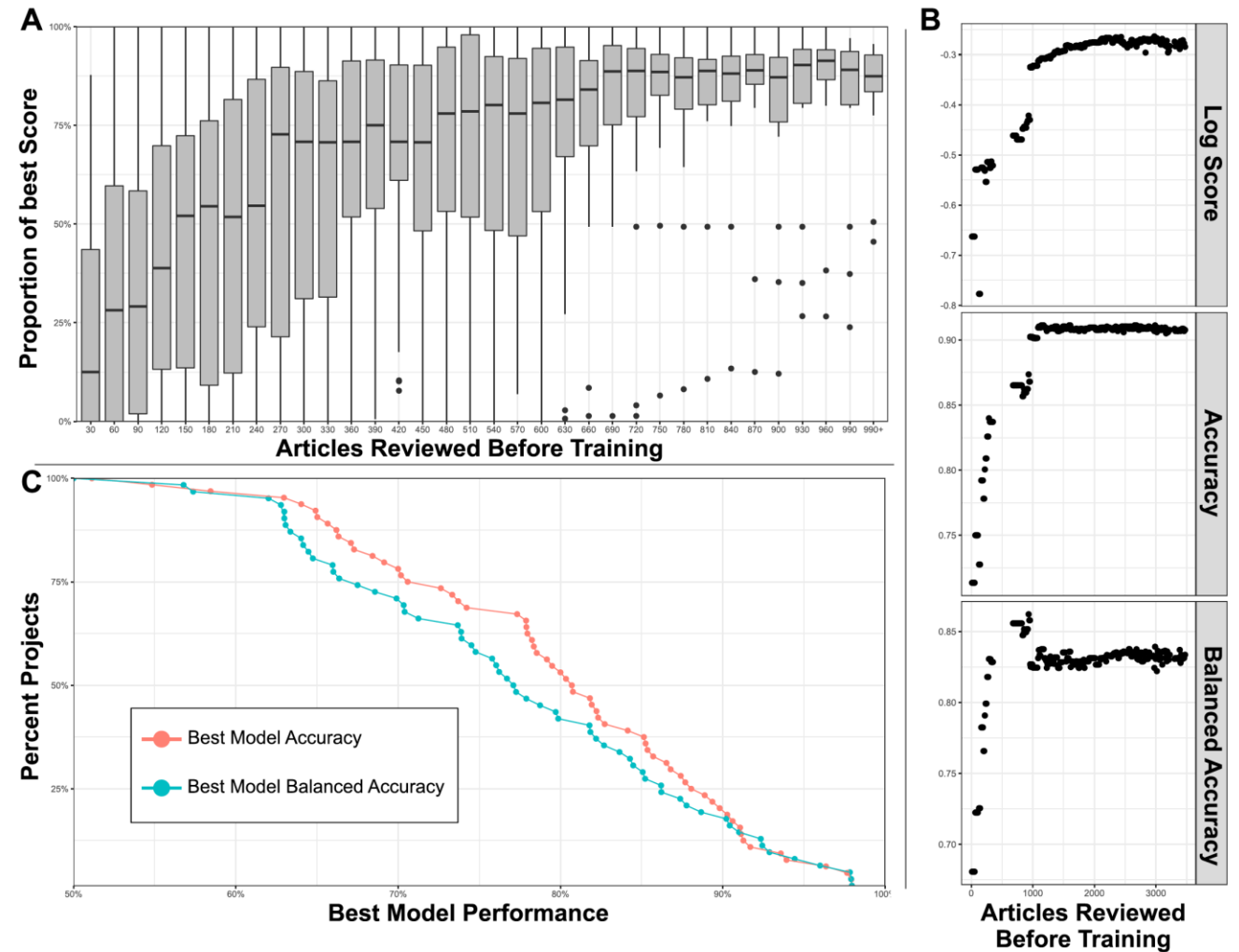
<https://www.youtube.com/channel/UCoUbMAvxBSZpOlqKjOkxNzQ/videos>



## Sysrev: A FAIR Platform for Data Curation and Systematic Evidence Review

Thomas Bozada Jr.<sup>1</sup>, James Borden<sup>1</sup>, Jeffrey Workman<sup>1</sup>, Mardo Del Cid<sup>1</sup>, Jennifer Malinowski<sup>2</sup> and Thomas Luechtefeld<sup>1,3\*</sup>

Models improve rapidly until 300 articles have been reviewed.



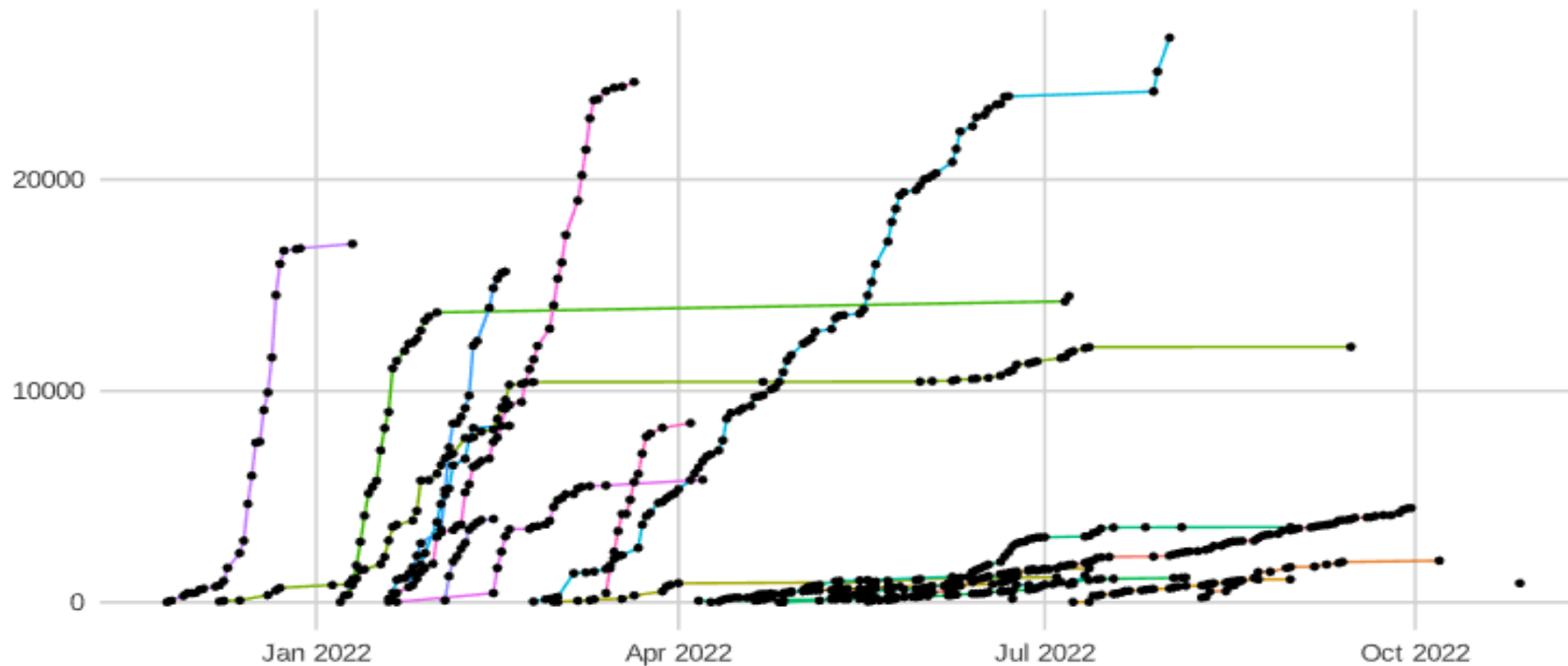
**FIGURE 7 |** (A) Box plots indicate distribution of model performance relative to the worst and best model in a given model's project. Models are bucketed according to the number of articles labeled before model training. Models improve rapidly until 300 articles have been reviewed. (B) Accuracy metrics for a large Sysrev reviewing insect population changes. Model performance is charted as a function of number of articles used in training, across 3 performance metrics, and evaluated on a consistent holdout set. (C) Best model accuracy and balanced accuracy evaluated in 64 Sysrevs.

# Data Mining for Ontologies & AI

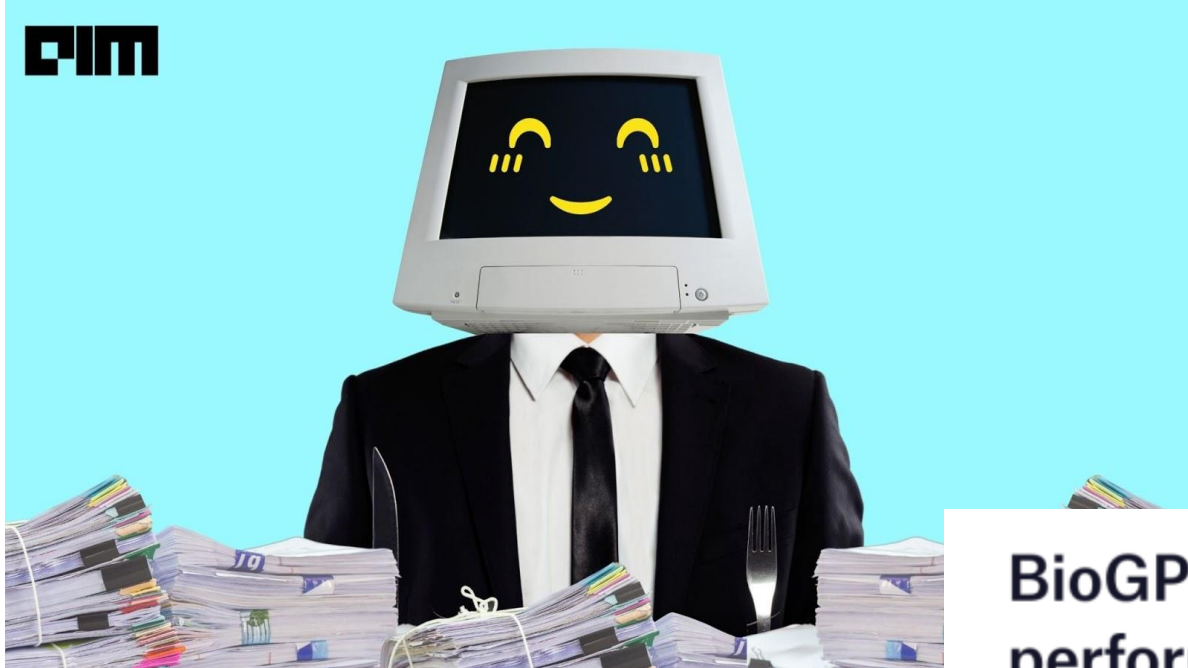


## Project activity

A simple ggplot can now track project activity over time:

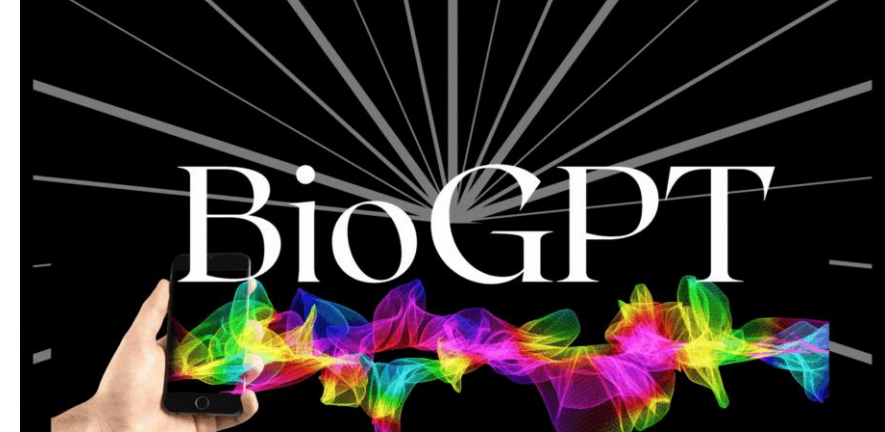


Progress Dashboard Visualization: Number of *labels* in each ONTOX project over time



<https://analyticsindiamag.com/10-useful-resources-to-access-ai-ml-research/>

**Pre-trained Large Language Models outperform human annotators of scientific papers...**



## BioGPT and human annotator have comparable performance in biomedical research test

Selected performances on PubMedQA, which tests biomedical language processing

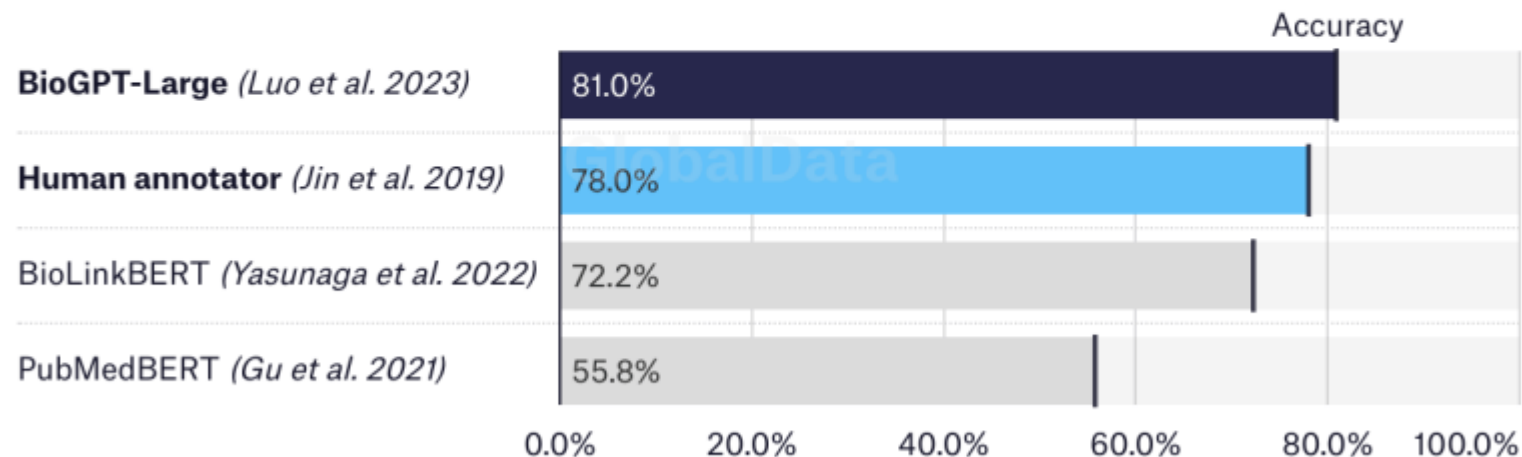
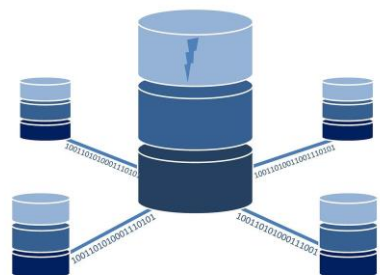


Chart: GlobalData • Source: PubMedQA



Literature



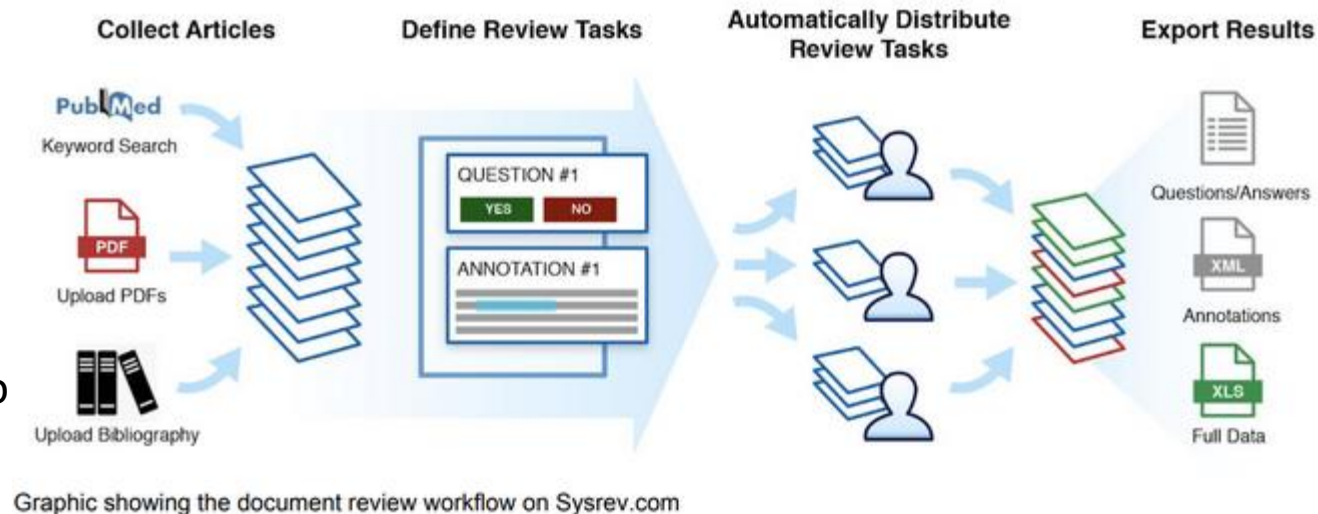
Databases



Internet



<https://www.youtube.com/c/SysRev?app=desktop>



biobricks-ai/  
**bricktools**

a set of tools for auditing bricks

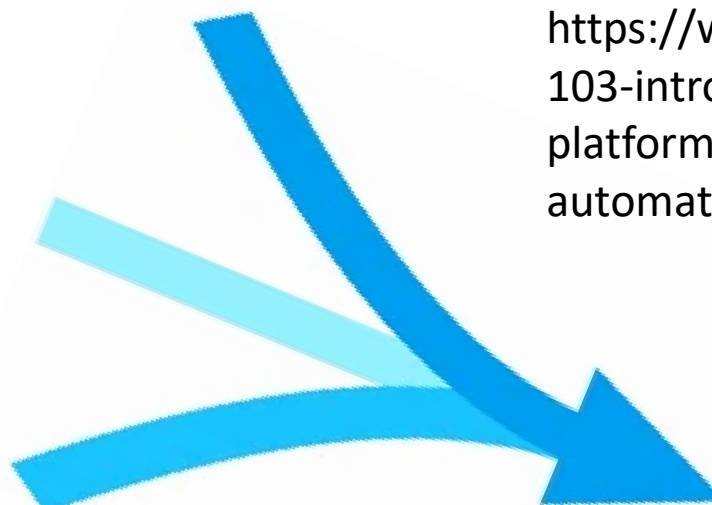
3 Contributors 1 Issue 1 Star 0 Forks



<https://www.biopharmatrend.com/post/103-introducing-sysrev-the-intelligent-platform-for-document-review-and-automated-data-extraction/>

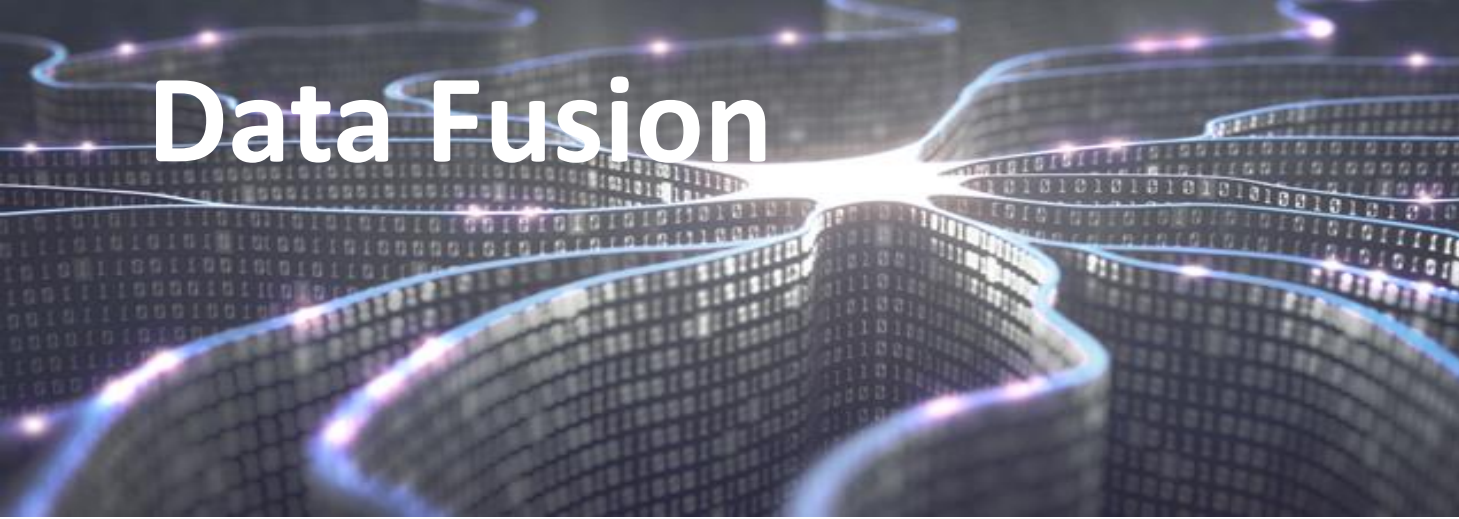
 **Chemchart**

<http://chemchart.com>



**DATA**

# Data Fusion



When looking for eye irritants,  
look also for neighbors' information on skin  
irritation, chemical reactivity etc.

**Output:  
probabilities**



**The different  
way of  
evidence  
integration**

# Ongoing RASAR developments

**79% (n=131) and 80% (n=375) accuracy in predicting  
HUMAN skin sensitization** (Golden et al., ALTEX, 2020)

**38,250 predictions for 4,729 food-relevant substances  
83% accurate (n=139) (Fu et al., 2022)**

Present Knowledge in  
**Food Safety**  
A Risk-Based Approach Through The Food Chain



Edited by  
Michael E. Knowles, Lucia E. Anelich,  
Alan R. Boobis and Bert Popping

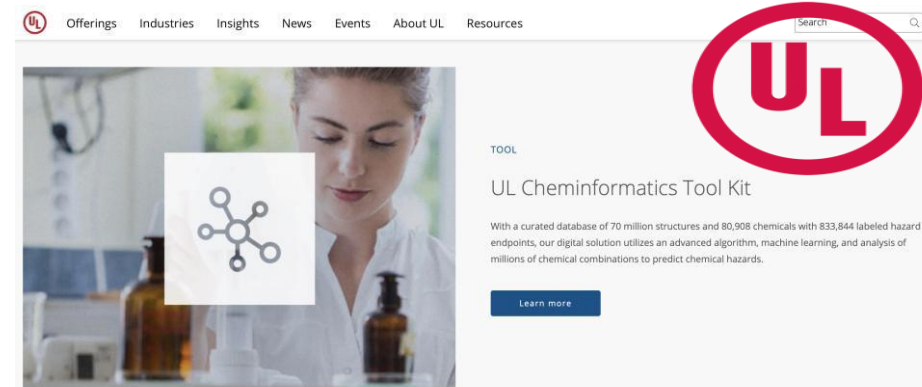


The use of artificial intelligence and big data for the safety evaluation of US food-relevant chemicals

Yuqi Fu<sup>1</sup>, Thomas Luechtefeld<sup>1,2</sup>, Agnes Karmaus<sup>3</sup> and Thomas Hartung<sup>1,4</sup>



**= 38,000 animal studies  
at \$250+ million**



**Accepted for Australian Industrial  
Chemical Legislation 2020**



<https://www.dreamstime.com/photos-images/sky-limit.html>

**TABLE 39.1** Five simplified manual curation categorization inventories for 1215 chemicals.

Manual curation categorization	Category number	Chemical count	Percentage
Direct food additive	1	502	39%
Pesticide/residue	3	329	25%
Indirect food additive	2	284	22%
Nonfood	4	106	8%
Not included in manual curation	5	76	6%
		1297 <sup>a</sup>	100%

1215 of 4729 chemicals had use categories

**TABLE 39.4** Percentage of positive and negative r

Positivity rate and negativity rate					
		Direct food additive			
		Positive rate	Negative rate		
TAI	Acute oral toxicity	28%	72%		
	Acute dermal irritation	57%	43%		
	Acute dermal toxicity	20%	80%		
	Acute aquatic toxicity	16%	84%		
	Acute inhalation toxicity	24%	76%		
	Chronic aquatic toxicity	23%	77%		
	Eye irritation	61%	39%		
	Mutagenicity	12%	88%		
	Skin sensitization	34%	66%		
Category and toxicity endpoint.					
Category	Toxicity endpoint	Pesticide/residue		Not included in manual curation	
		Positive rate	Negative rate	Positive rate	Negative rate
Direct food additive	Acute oral toxicity	54%	46%	37%	63%
	Acute dermal irritation	40%	60%	66%	34%
Indirect food additive	Acute oral toxicity	43%	57%	33%	67%
	Acute dermal irritation	59%	41%	11%	89%
Pesticide/residue	Acute oral toxicity	47%	53%	34%	66%
	Acute dermal irritation	66%	34%	16%	84%
Nonfood	Acute oral toxicity	38%	62%	73%	27%
	Acute dermal irritation	25%	75%	18%	82%
Not included in manual curation	Acute oral toxicity	51%	49%	42%	58%
	Acute dermal irritation				

Substantial number of hazards

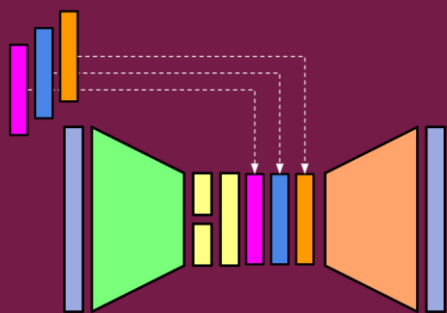
**TABLE 39.5** Comparison of RASAR predictions and complementary experimental data classification for high-confidence chemicals.

Chemical	CASRNs	Manual Curation Categorization	Acute				Skin			Chronic	
			Acute Ora Toxicity	Inhalation Toxicity	Acute Derma Toxicity	Acute Dermal Irritation	Eye Irritation	Sensitization	Acute Aquatic Toxicity	Aquatic Toxicity	Mutagenicity
Allyl cyclohexanepropionate	2705-87-5	Direct food additive	(0.965) + ✓-✓	(0.922) + ✓-✓	(0.860) + ✓-✓	(0.399) - /-✓	(0.252) - x-/-	(0.928) + ✓-✓	(0.996) + ✓-✓	(0.960) + ✓-✓	(0.059) - x-x
α-phellandrene	99-83-2	Direct food additive	(0.025) - x-✓	(0.093) - x-✓	(0.039) - x-/-	(0.127) - x-✓	(0.048) - x-✓	(0.905) + ✓-✓	(0.994) + ✓-/-	(0.829) + ✓-✓	(0.060) - x-/-
Methyl butyrate	623-42-7	Direct food additive	(0.034) - x-x	(0.185) - x-x	(0.140) - x-/-	(0.991) + ✓-✓	(0.993) + ✓-✓	(0.132) - x-x	(0.076) - x-/-	(0.130) - x-/-	(0.161) - x-/-
3-(Methylthio)propyl isothiocyanate	505-79-3	Direct food additive	(0.950) + ✓-✓	(0.818) + ✓-✓	(0.767) + ✓-✓	(0.959) + ✓-✓	(0.941) + ✓-✓	(0.907) + ✓-/-	(0.992) + ✓-✓	(0.919) + ✓-✓	(0.209) - x-/-
Pentachloropyridine	2176-62-7	Indirect food additive	(0.822) + ✓-✓	(0.980) + ✓-✓	(0.969) + ✓-✓	(0.969) + ✓-✓	(0.945) + ✓-✓	(0.924) + ✓-✓	(0.991) + ✓-✓	(0.902) + ✓-✓	(0.174) - x-x
Furfural	98-01-1	Indirect food additive	(0.882) + ✓-✓	(0.949) + ✓-✓	(0.941) + ✓-✓	(0.988) + ✓-✓	(0.991) + ✓-✓	(0.966) + ✓-✓	(0.617) + /-✓	(0.473) - /-✓	(0.872) + ✓-✓
2,4-Diaminotoluene	95-80-7	Indirect food additive	(0.934) + ✓-✓	(0.937) + ✓-✓	(0.990) + ✓-✓	(0.901) + ✓-✓	(0.822) + ✓-✓	(0.983) + ✓-✓	(0.824) + ✓-✓	(0.810) + ✓-✓	(0.987) + ✓-✓
Dichlorobenzene	106-46-7	Indirect food additive Pesticides/residues	(0.758) + ✓-✓	(0.796) + ✓-✓	(0.325) + ✓-x	(0.989) + ✓-✓	(0.974) + ✓-✓	(0.713) + ✓-✓	(0.539) + /-✓	(0.290) - x-✓	(0.903) + ✓-x
Coumaphos	56-72-4	Non-food	(0.827) + ✓-✓	(0.937) + ✓-✓	(0.892) + ✓-✓	(0.354) - x-✓	(0.278) - x-✓	(0.811) + ✓-✓	(0.996) + ✓-✓	(0.857) + ✓-✓	(0.515) + /-x
Coumatetralyl	5836-29-3	Non-food	(0.963) + ✓-✓	(0.846) + ✓-✓	(0.837) + ✓-✓	(0.859) + ✓-✓	(0.824) + ✓-✓	(0.938) + ✓-x	(0.995) + ✓-✓	(0.970) + ✓-✓	(0.518) + /-✓
sulfotep	3689-24-5	Non-food	(0.812) + ✓-✓	(0.780) + ✓-✓	(0.803) + ✓-✓	(0.493) + /-✓	(0.274) - x-✓	(0.577) + /-✓	(0.993) + ✓-✓	(0.907) + ✓-✓	(0.234) - x-x
2,4-D-1-butyl ester	94-80-4	Non-food	(0.973) + ✓-✓	(0.984) + ✓-/-	(0.971) + ✓-/-	(0.703) + ✓-✓	(0.590) + /-✓	(0.976) + ✓-✓	(0.984) + ✓-✓	(0.902) + ✓-✓	(0.245) - x-x
Terbufos	13071-79-9	Pesticides/residues	(0.648) + /-✓	(0.626) + /-✓	(0.557) + /-✓	(0.490) + /-✓	(0.225) - x-✓	(0.617) + /-/-	(0.998) + ✓-✓	(0.922) + ✓-✓	(0.299) - x-x
Tefluthrin	79538-32-2	Pesticides/residues	(0.644) + /-✓	(0.822) + ✓-✓	(0.691) + ✓-✓	(0.031) - x-✓	(0.015) - x-✓	(0.065) - x-x	(0.997) + ✓-✓	(0.971) + ✓-✓	(0.063) - x-x
Deltamethrin	52918-63-5	Pesticides/residues	(0.831) + ✓-✓	(0.715) + ✓-✓	(0.658) + /-/-	(0.763) + ✓-✓	(0.587) + ✓-✓	(0.940) + ✓-✓	(0.997) + ✓-✓	(0.907) + ✓-✓	(0.346) - /-/-
Cypermethrin	52315-97-8	Pesticides/residues	(0.556) + /-✓	(0.454) - /-✓	(0.284) - x-/-	(0.009) - x-✓	(0.039) - x-✓	(0.310) - x-✓	(0.996) + ✓-✓	(0.939) + ✓-✓	(0.043) - x-x
Fenvalerate	51630-58-1	Pesticides/residues	(0.466) - /-✓	(0.039) - x-/-	(0.020) - x-/-	(0.022) - x-✓	(0.016) - x-✓	(0.157) - x-✓	(0.994) + ✓-✓	(0.948) + ✓-✓	(0.021) - x-x
2,5-Dimethylfuran	625-86-5	Not included in manual curation	(0.933) + ✓-✓	(0.979) + ✓-✓	(0.975) + ✓-/-	(0.996) + ✓-✓	(0.996) + ✓-✓	(0.975) + ✓-✓	(0.595) + /-/-	(0.514) + /-/-	(0.821) + ✓-x

Small subset compared to  
literature findings:

83% correct

## Additive CVAE



### Why Should I Care?

1. Property Specific embedding function
2. Allows independent training of posteriors
3. Can be scaled
4. Can be interpreted probabilistically
5. Can generate compounds
6. Updates have a geometric interpretation
7. High Accuracy, Interpretable Property Prediction

**Conditional variational autoencoder**

**Limited at this moment by computational power:**

- Used only 1.2 of 200 million datapoints
- Only 152 properties considered



# Biological similarity

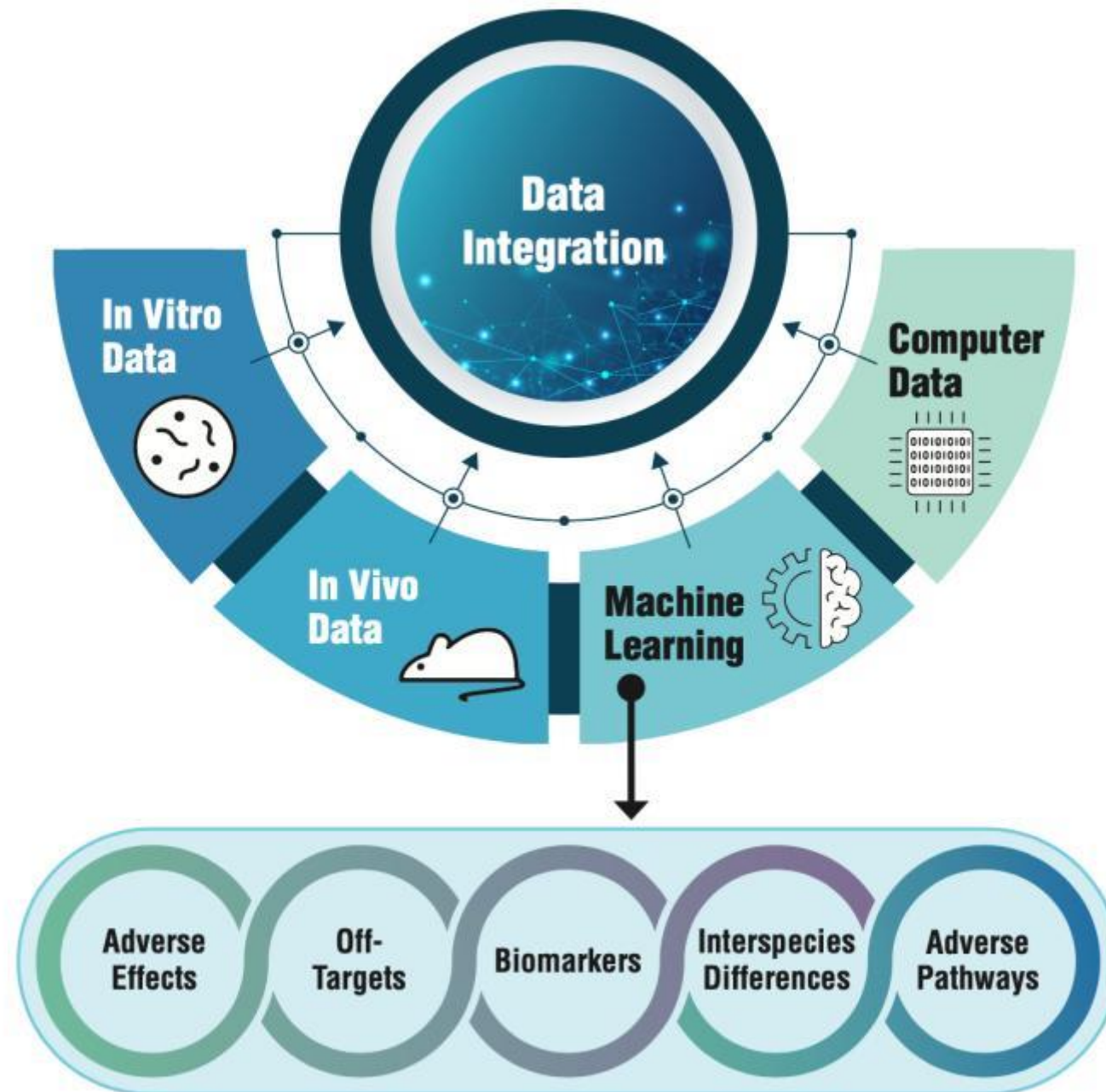
**Create property-specific additive modifiers that update embeddings for each of the 152 properties**

**~ 80% AUC**

**Property specific embeddings for combinations of properties  
= additive CVAE**

**Can create compounds based on the 152 properties!**

# Goals Data Integration Toxicology





***For every complex problem there  
is an answer that is clear, simple,  
and wrong.***

**Henry L. Mencken  
(Baltimore, 1880-1956)**

**The best description of traditional toxicology I know**

# The problem

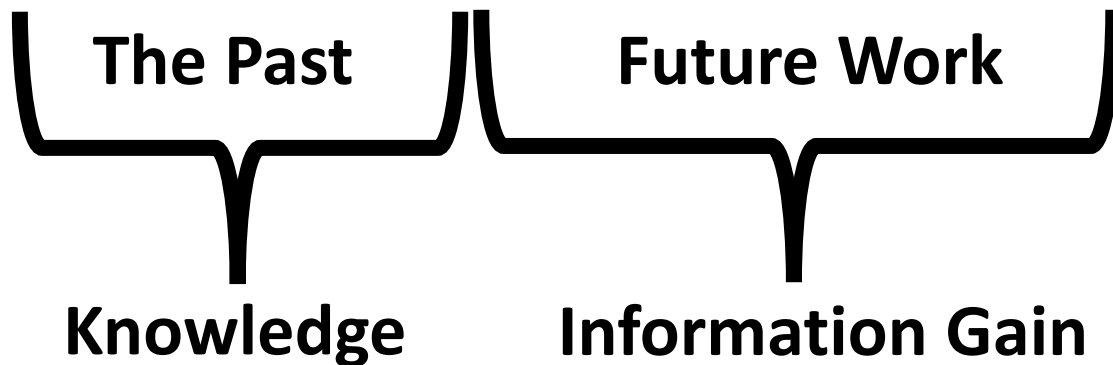
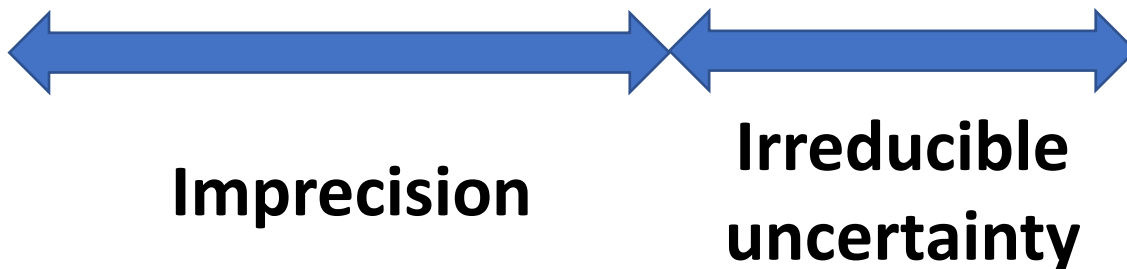
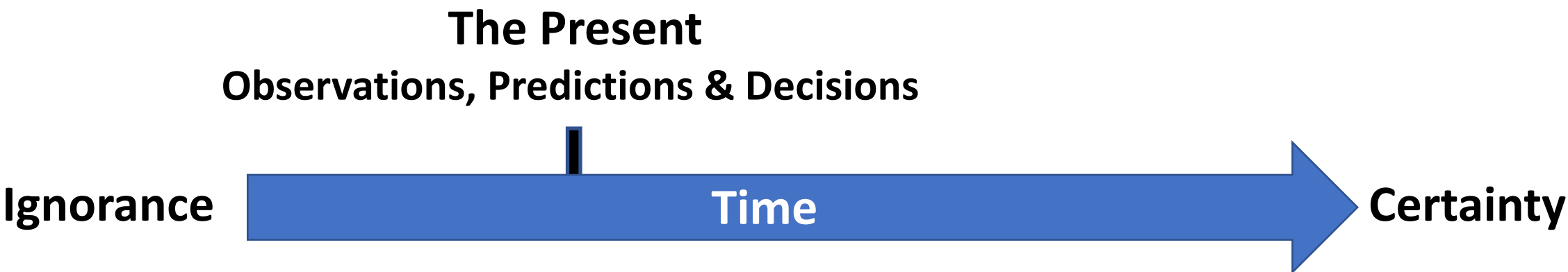


**toxic**

**non-toxic**



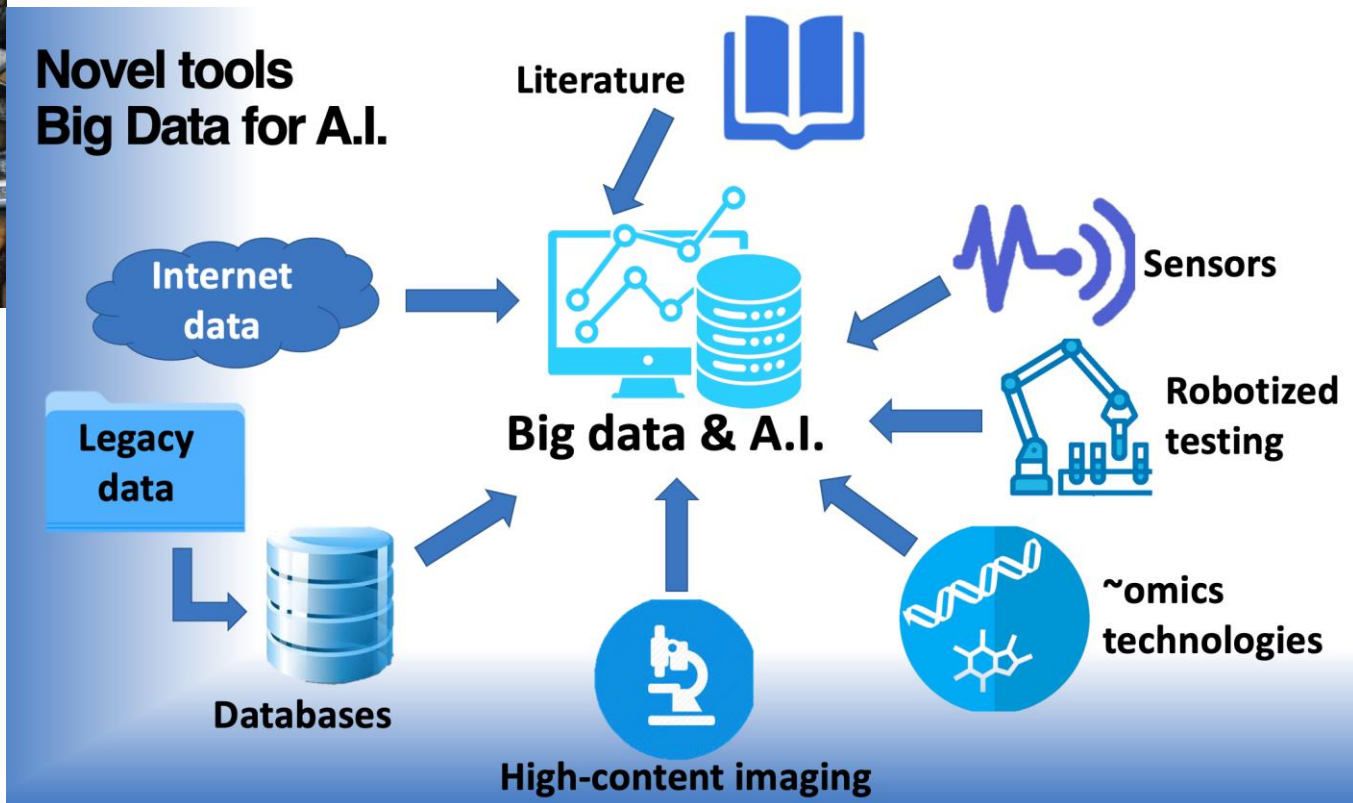
**reality**  
**= uncertainty**



*"Probability is expectation founded upon partial knowledge. A perfect acquaintance with all the circumstances affecting the occurrence of an event would change expectation into certainty, and leave neither room nor demand for a theory of probabilities" George Boole (1815-1864)*



**Novel tools in toxicology  
require evidence integration**  
**Probability-changer**



# **Deterministic vs. Probabilistic**

**QSAR**

**RASAR**

**Descriptive statistics**

**Big Data mining**

**Dynamic modeling**

**Machine Learning, AI**

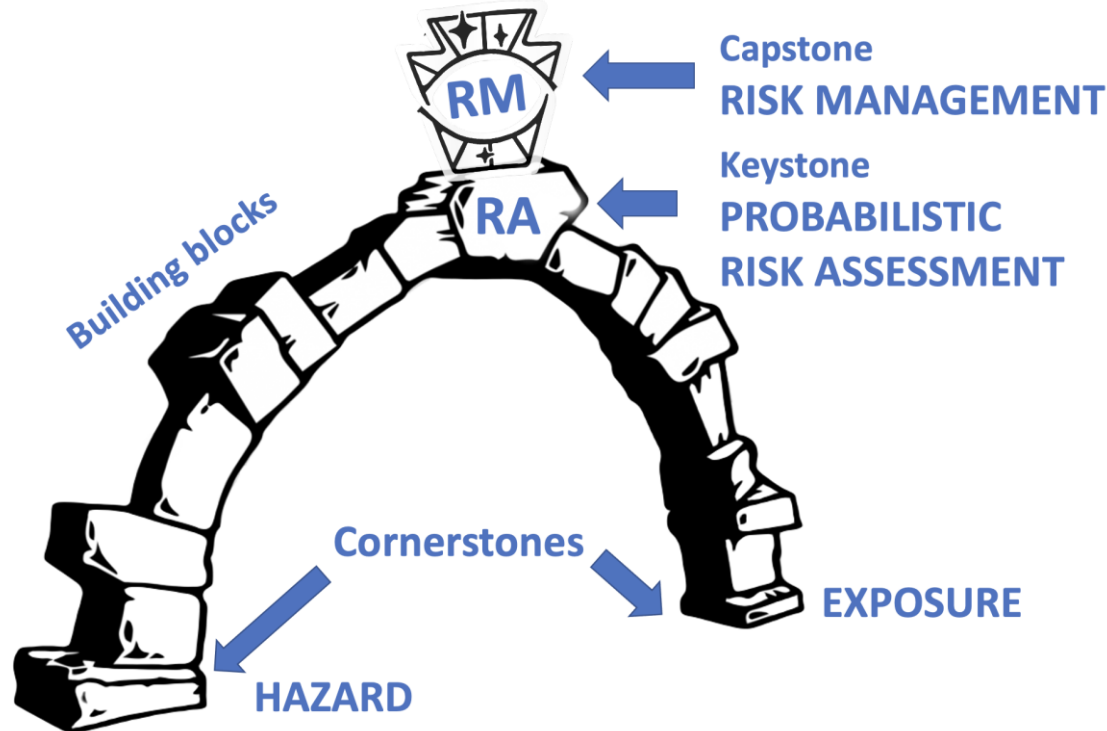
**Systems Toxicology**

**??? Systems ToxAicology**

Food for Thought ...

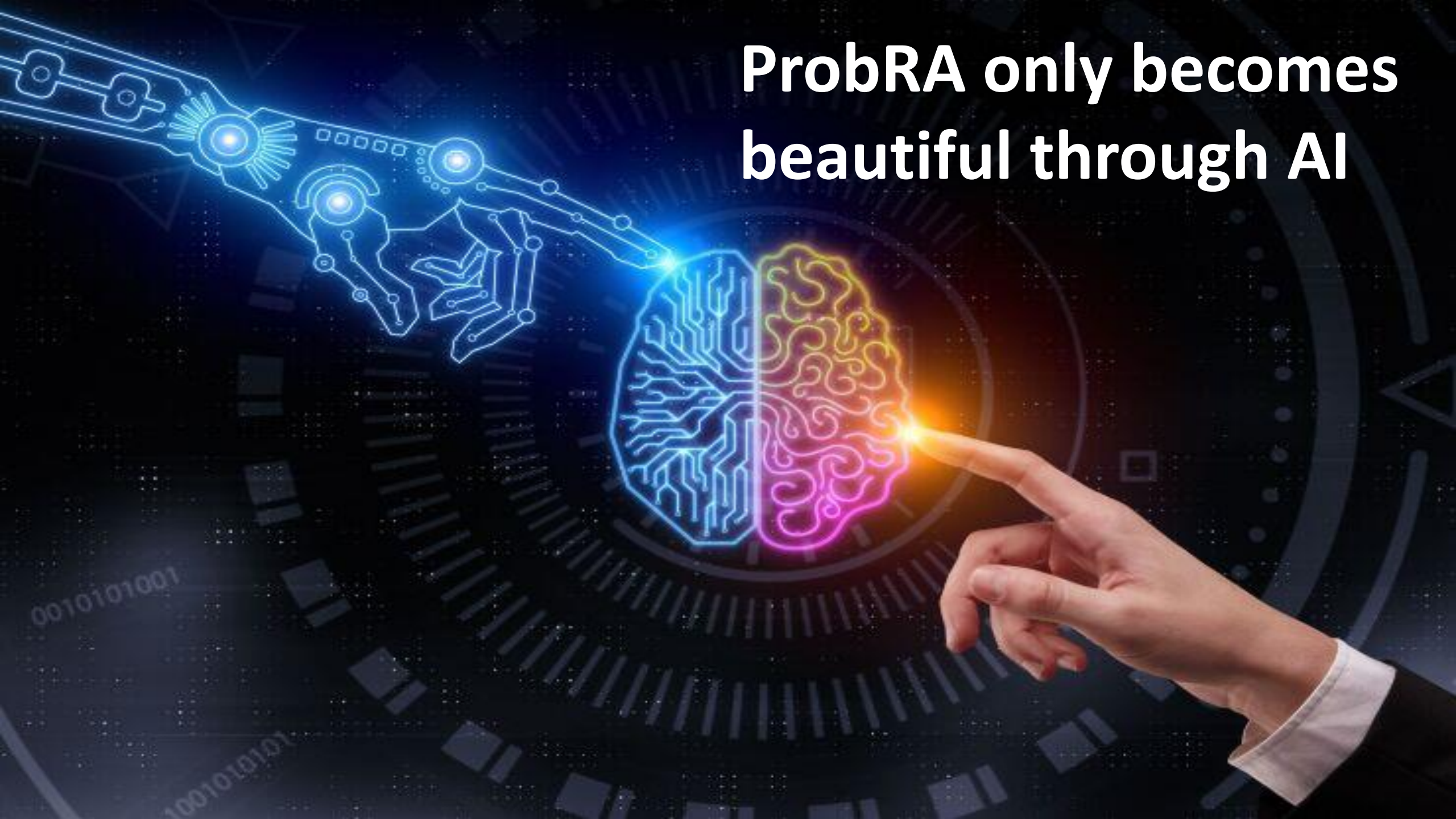
## Probabilistic Risk Assessment – the Keystone for the Future of Toxicology

Alexandra Maertens<sup>1</sup>, Emily Golden<sup>1</sup>, Thomas H. Luechtefeld<sup>1,2</sup>, Sebastian Hoffmann<sup>1,3</sup>,  
Katya Tsaoun<sup>1</sup> and Thomas Hartung<sup>1,4</sup>



2 Workshops 2022 & 2023  
Ranco, Italy

**ProbRA only becomes  
beautiful through AI**



# DATA

# PMDEP App

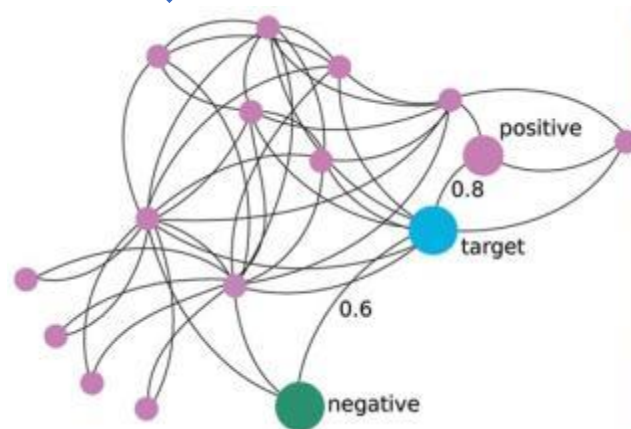
<https://youtu.be/YG0gjm&GD5K>

Marc Teunis



Systems Biology  
Markup Language

CellDesigner™



From perturbation of physiology

# Probability of hazard

RASAR  
+ QSAR

From chemical structure and properties  
including biological ones (!)

Do we really want an A.I. tool predicting (animal) toxicology?

- Ends a \$4 billion industry
- Carcinogenicity testing has 52% pos. (n ~3,500, Ames & Gold 2000, Gold et al. 2005)

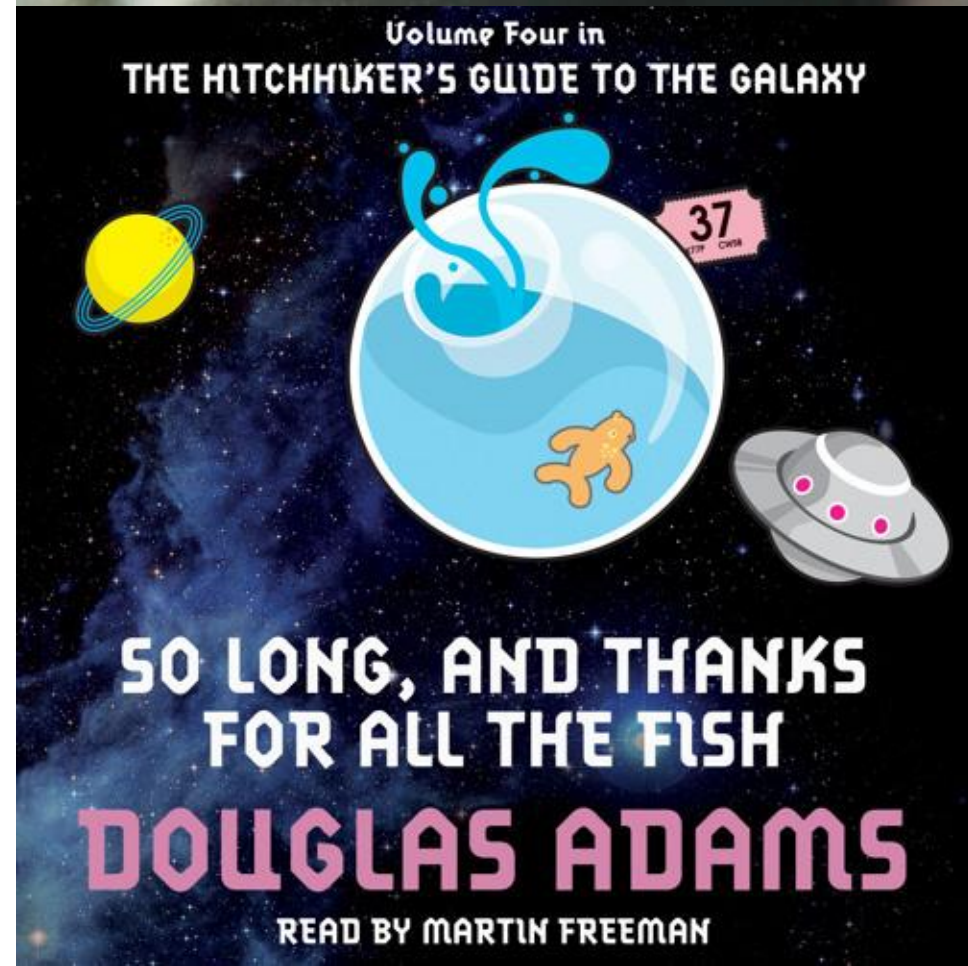
*Drugs on market: 44% of 282 (FDA), 49% of 241 (PDR)*

*Ingredients of coffee: 72% of 32*



**Chemophobia**  
**Legal liabilities**

So long and thanks for all the feed



AUTONOMOUS  
WEAPONS

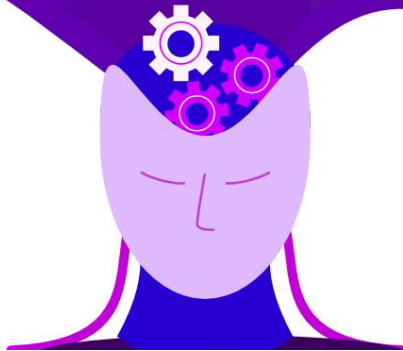
LOSS OF HUMAN  
JOBS

ARTIFICIAL  
INTELLIGENCE  
BIAS

# ARTIFICIAL INTELLIGENCE DANGERS

ARTIFICIAL  
INTELLIGENCE  
TERRORISM

INVASION OF  
PRIVACY



JOHNS HOPKINS  
BLOOMBERG SCHOOL  
of PUBLIC HEALTH



[Home](#) [About Us](#) [Programs/Activities](#) [Publications](#) [Resources](#) [Media Center](#) [Contact Us](#) [Make a Gift](#) [Search Site](#)

CAATwalk Newsletter

Slides available:



Co-pilot not  
autonomous – we  
need the human-in-  
the-loop!

