“Big Data” módszerek, eredmények és (félre-)értelmezésük az élelmiszer- és táplálkozástudományban

Baranyi József, University of Debrecen, Hungary

What is "Big Data"?

Volume
SCALE OF DATA

40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005

It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day

6 BILLION PEOPLE
have cell phones

Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored

Variety
DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]

30 BILLION PIECES OF CONTENT
are shared on Facebook every month

By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month

400 MILLION TWEETS
are sent per day by about 200 million monthly active users

Velocity
ANALYSIS OF STREAMING DATA

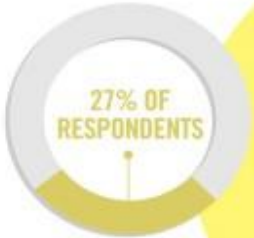
The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session

Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections per person on earth

Veracity
UNCERTAINTY OF DATA

1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions

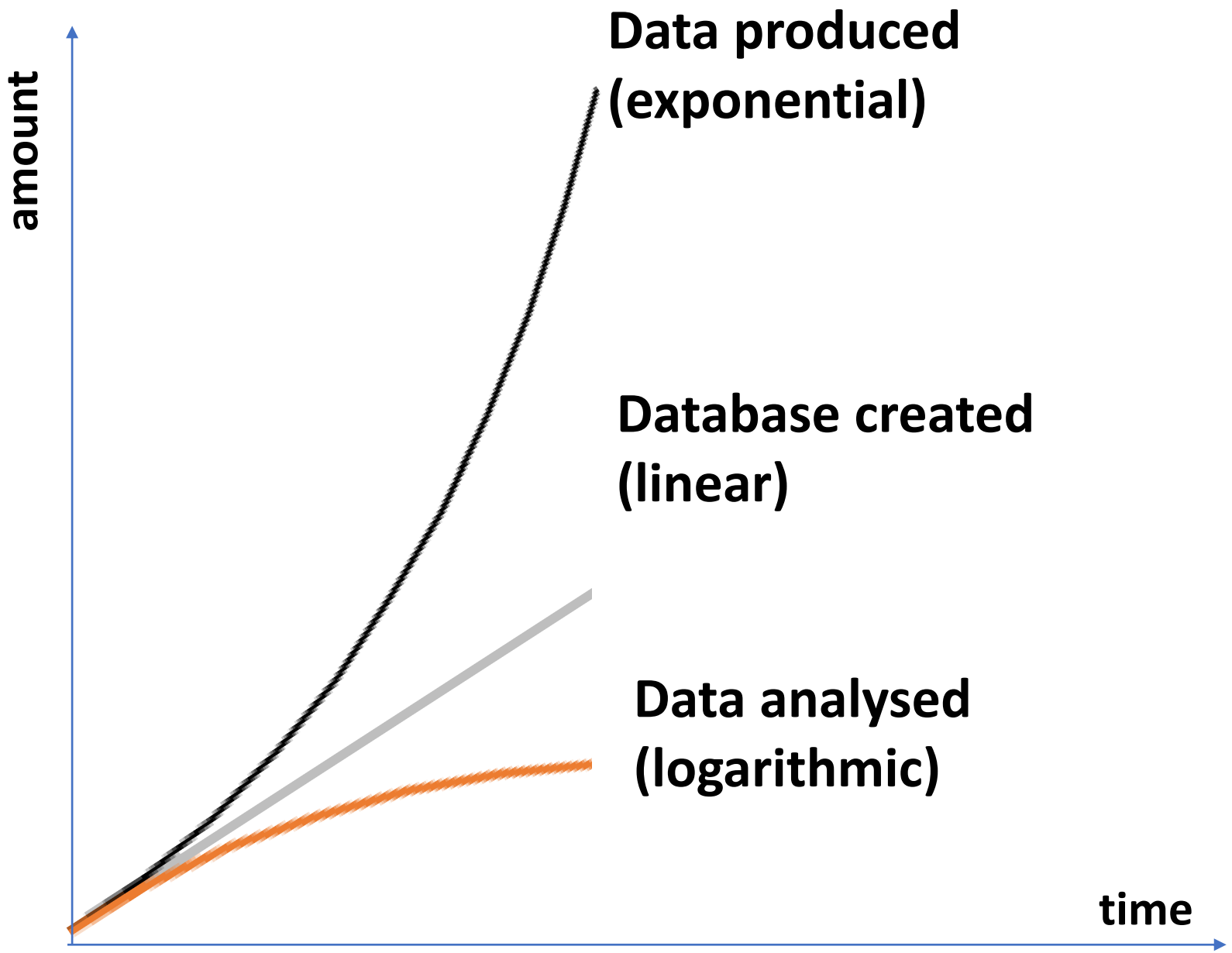


in one survey were unsure of how much of their data was inaccurate

Poor data quality costs the US economy around
$3.1 TRILLION A YEAR

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

IBM



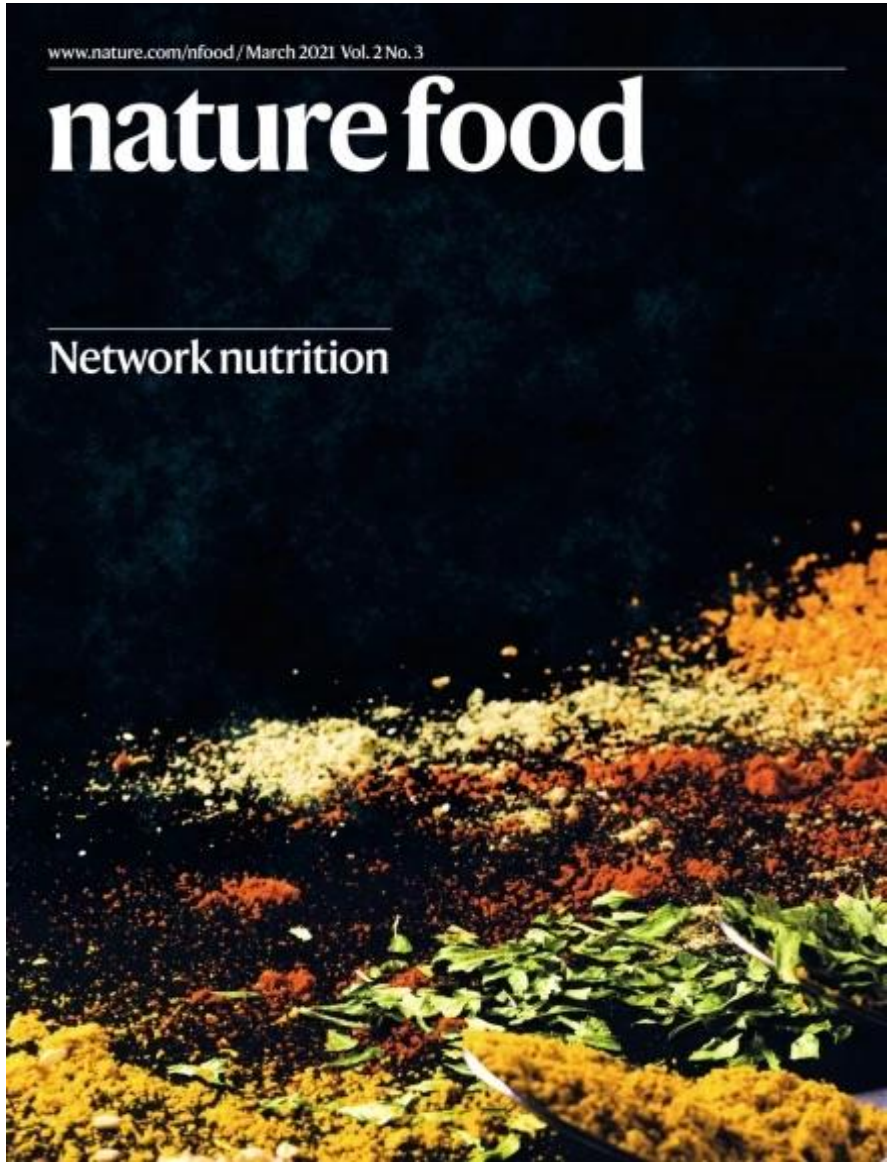
Based on P. Wolfe:
Making sense of
big data
PNAS 110 / 45

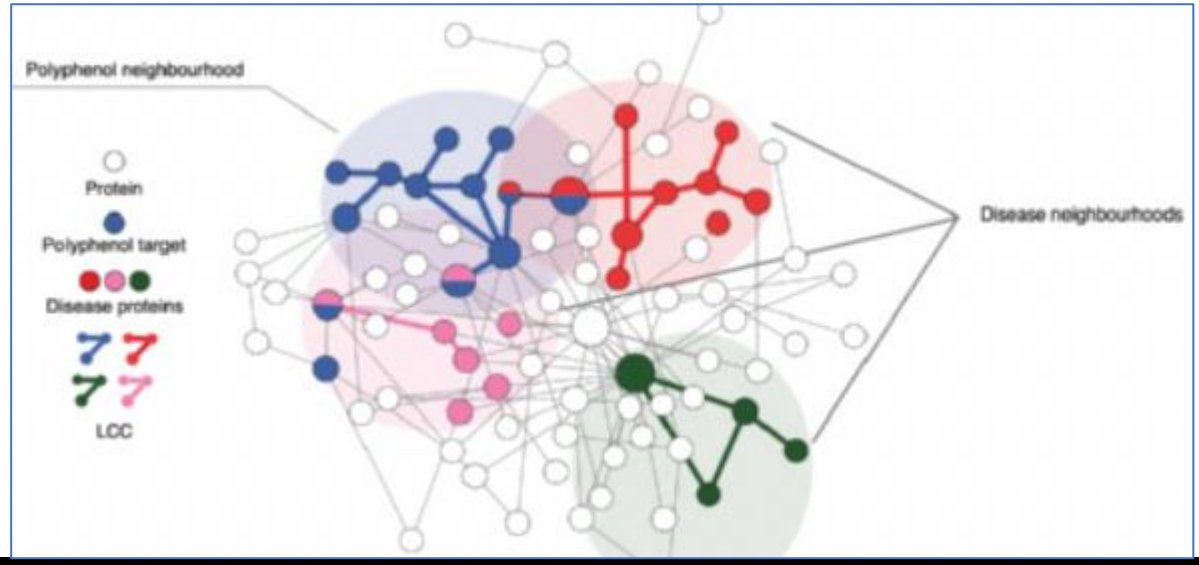
Top 20 Food-related databases

Based on Crit Rev Food Sci Nutr 57:11, 2286

Top 20 Food-related databases	Database type	Country	Organisation	Link/source
GEMS/food	Monitoring data	Global	WHO	https://extranet.who.int/gemsfood/
JECFA Evaluations Database	Hazard evaluations	Global	JECFA	http://apps.who.int/food-additives-contaminants-jecfa-database
RASFF	Alerts/notifications	European Union	European Commission	https://webgate.ec.europa.eu/rasff-window/portal
FDA Recent Recalls, Market Withdrawals, & Safety Alerts	Alerts/notifications	USA	USFDA	http://www.fda.gov/Safety/Recalls/default.htm
FDA Archive Recalls, Market Withdrawals, & Safety Alerts	Alerts/notifications	USA	USFDA	http://google2.fda.gov/search
Codex Alimentarius	Standards	Global	WHO/FAO	http://www.codexalimentarius.org/standards/list-of-standards/en
EU pesticides database	Pesticide approval	EU	European Commission	http://ec.europa.eu/sanco_pesticides/public/index.cfm
FSANS Food standards code	Food (safety) standards codes	Australia & New Zealand	FSANZ	http://www.foodstandards.gov.au/code/Pages/default.aspx
The EFSA Comprehensive European Food Consumption Database	Consumption data	EU	EFSA	http://www.efsa.europa.eu/en/datexfoodcdb/datexfooddb.htm
JECFA Specifications for Flavourings	Chemical/biological specifications	Global	JECFA	http://www.fao.org/food/food-safety-quality/scientific-advice/jecfa/jecfa-flav/en/
Foodborne Diseases Active Surveillance Network (FoodNet)	Outbreak surveillance	USA	CDC	http://www.cdc.gov/foodnet/index.html
Foodborne Outbreak Online Database (FOOD)	Outbreak surveillance	USA	CDC	http://wwwn.cdc.gov/foodborneoutbreaks/
ComBase	Quantitative microbiology	USA	USDA-ARS	http://www.combase.cc/index.php/en/
Global G.A.P.	Supplier information	Global	GLOBALG.A.P.	http://www.globalgap.org/uk_en/buyers/Sourcing-Certified-Products/index.html
International Food Additive Database	Maximum levels	USA	USDA; GMA; USDEC; BCI	http://www.foodadditivedatabase.com/
USDA Production, Supply and Distribution Online	Production/supply	USA	USDA-PSD	http://apps.fas.usda.gov/psdonline/psdHome.aspx
USDA Global Agricultural Trade System (GATS)	Import/export	USA	USDA-FAS	http://apps.fas.usda.gov/gats/default.aspx
AllergenOnline	Chemical information	USA	University of Nebraska-Lincoln	http://www.allergenonline.org/
SDAP - Structural Database of Allergenic Proteins	Chemical information	USA	UTMB-Health	http://fermi.utmb.edu/SDAP/
USDA National Nutrient Database for Standard Reference	Food product information	USA	USDA-NAL	http://ndb.nal.usda.gov/

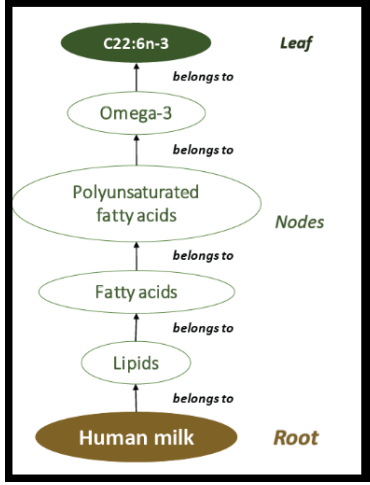
“FOODOME”





BARABASI LAB SCIENCE

PUBLICATIONS PROJECTS BOOKS ABOUT



MILKYBASE, A DATABASE OF HUMAN MILK COMPOSITION AS A FUNCTION OF MATERNAL-, INFANT- AND MEASUREMENT CONDITIONS

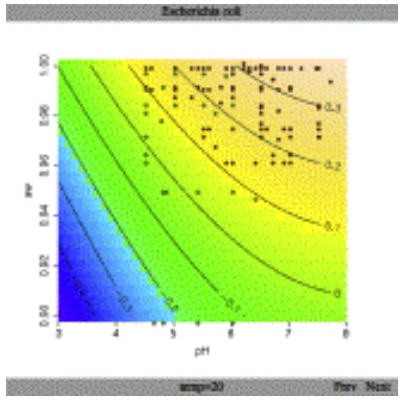
TÜNDE PACZA, MAYARA L. MARTINS, MAHA ROCKAYA, KATALIN MÜLLER, AYAN CHATTERJEE, ALBERT-LÁSZLÓ BARABÁSI & JÓZSEF BARANYI

Scientific Data volume 9, Article number: 557 (2022)

ABSTRACT FIGURES ↓ PDF ↓

From data to predictions and decisions

Predictive software tool to *AID* decision making



Artificial Intelligence: a tool to *MAKE decisions*

Data, Database of observations

Visualization, Statistics

Mathematical model

Predictive model implemented in a software tool

Decision

Model: in a cognitive fiction space

observations ➔ descriptions ➔ models

In some way, every prediction is an extrapolation

World around us

Observations ➔ descriptions

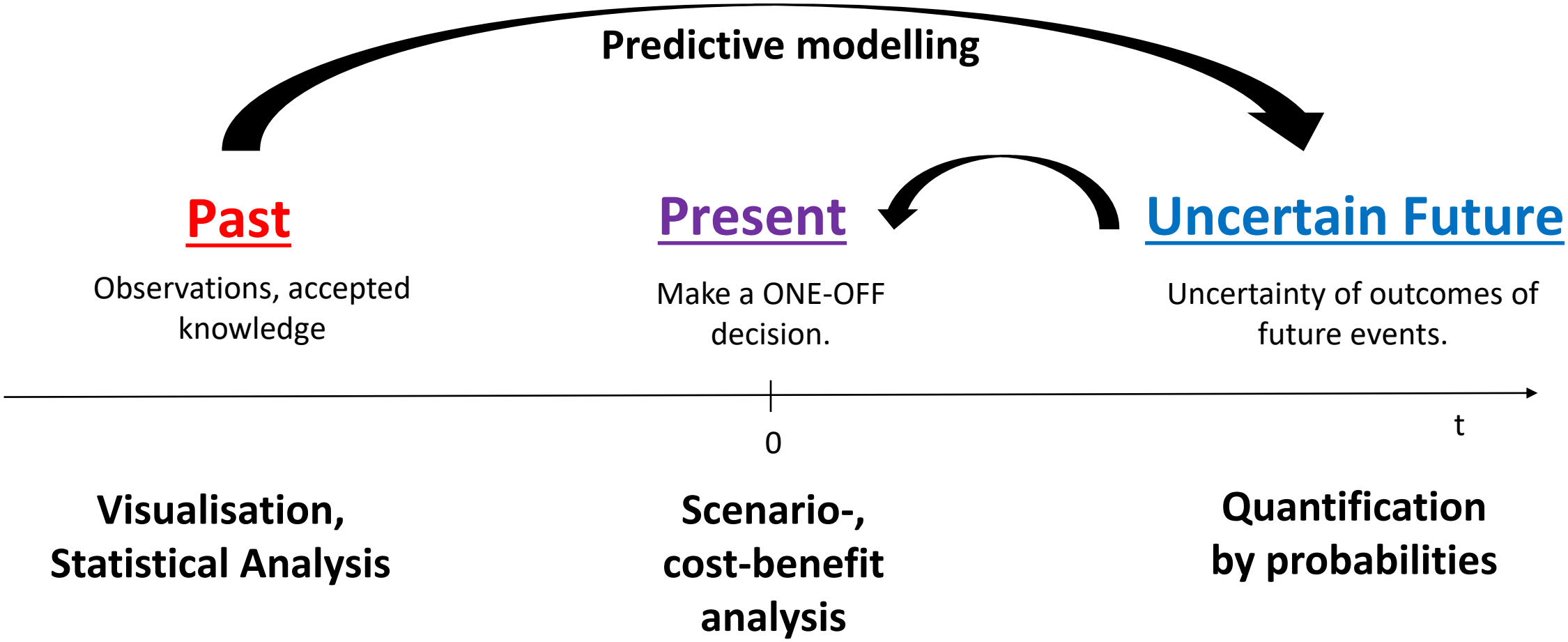
Descriptive science

Predictions

PREDICTIVE science

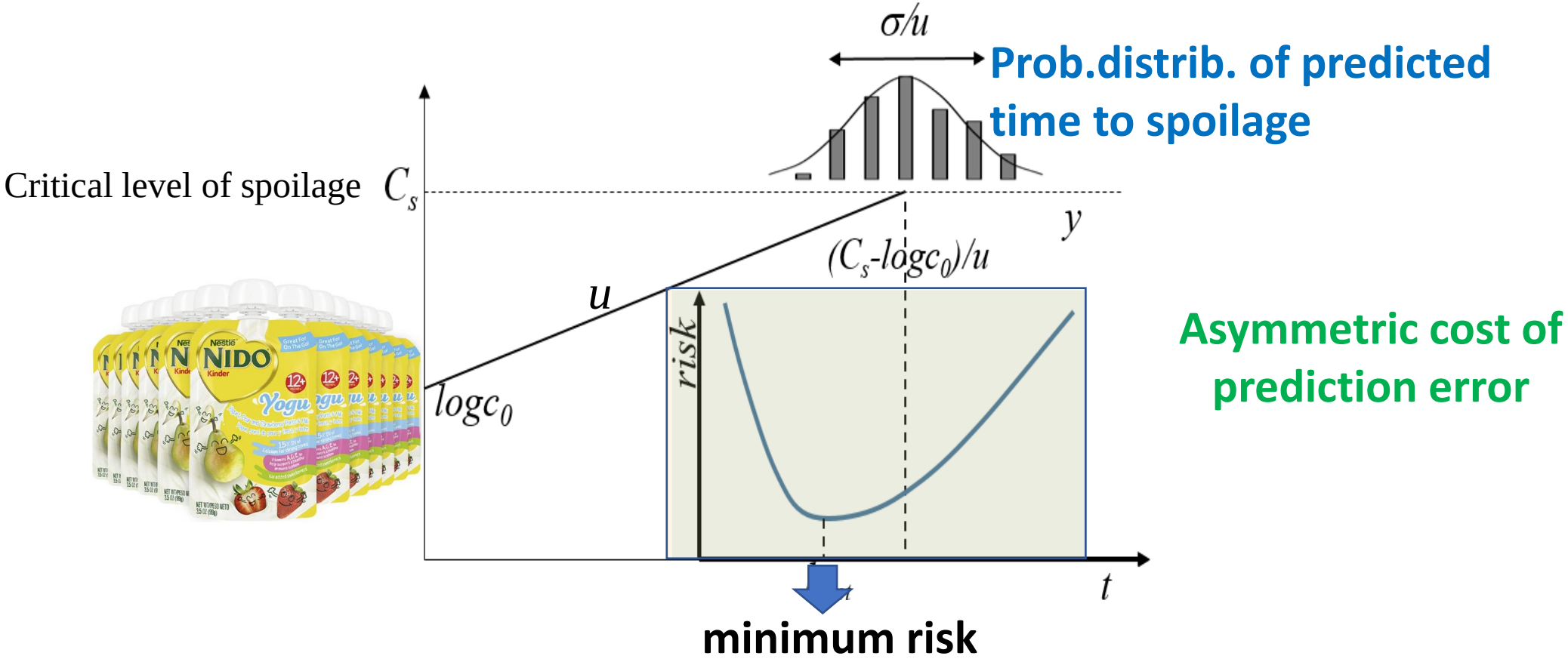
An homage to Yuval Harari

A risk-minimizing strategy to make decisions



Basis of decision: minimize the mean cost of the discrepancy between expected outcome and reality

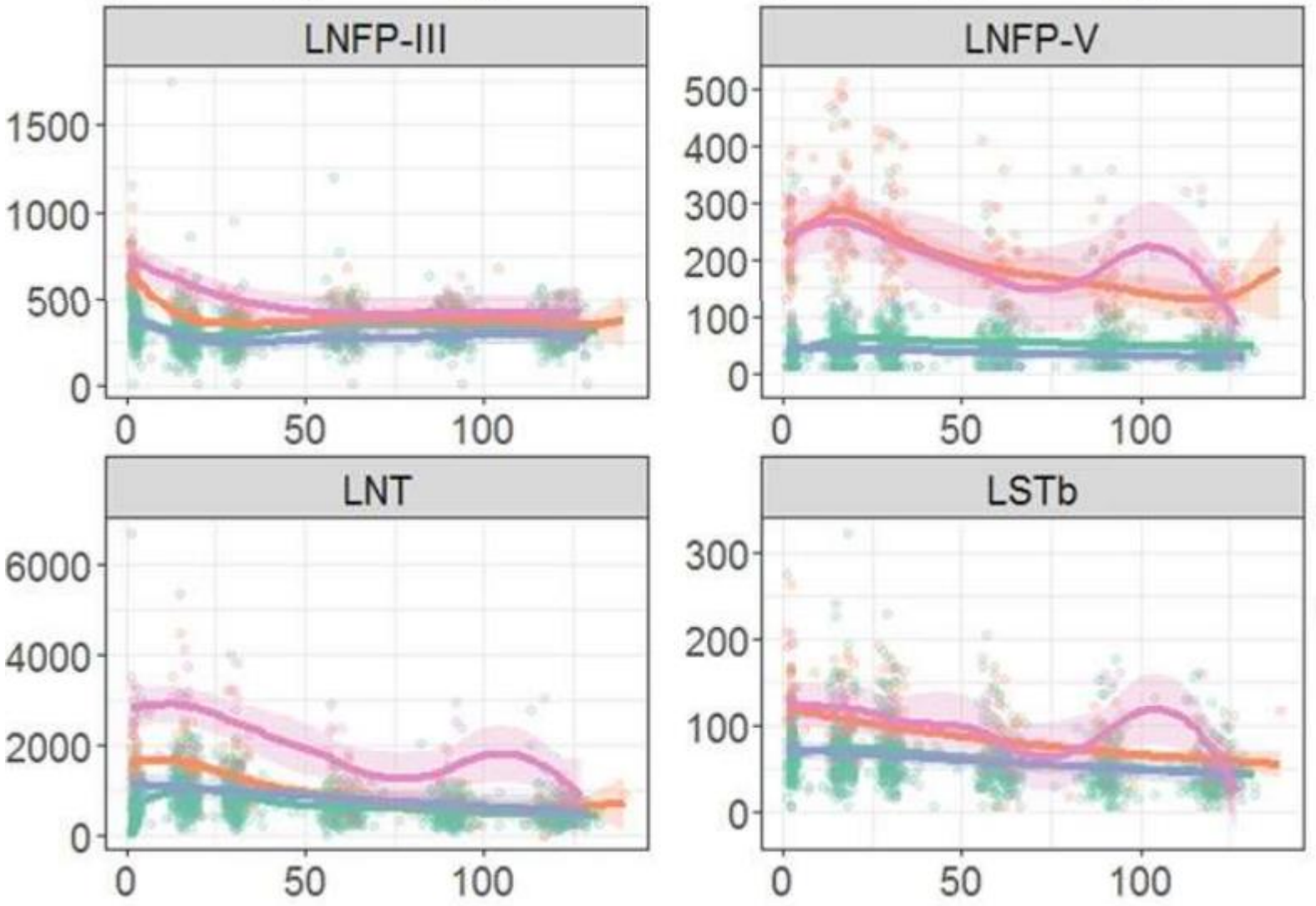
Make a decision whether a product should be taken off the shelf



The cost is different for under- and over-estimations of time to spoilage

Descriptive statistics is frequently inadequate for prediction.
Even worse, if empirical intra- and extrapolations are presented with seemingly scientific words (e.g. “statistically significant”) but interpreted incorrectly to make biased decisions.

Trajectories of HMO concentrations during the first 4 months of lactation separated by milk group. The solid lines represent the smoothing curves via local polynomial regression (LOESS – Locally Weighted Scatter-plot Smoother) and the shaded area represents the 95% confidence interval. (Details on statistical differences between milk groups can be found in Supplementary Table 6).



Samuel et al: Impact of maternal characteristics on human milk oligosaccharide on composition over the first 4 months of lactation in a cohort of healthy European mothers. Nat. Sci.Rep. (2022)

Összefoglalás

- Arra, hogy adatokból tudást nyerjünk, nagyságrendekkel kevesebb energiát fordítunk mint adatok generálására

- Adat-tisztítás, strukturálás, statisztika és prediktív modellezés segít az adatok értelmezésében.

- A túlzott törekvés az eredmények “egyszerű” nyelven történő elmagyarázására gyakran azok félre-értelmezéshez vezet.

- Morálisan elfogadhatatlan, ha ez a félre-értelmezés tudatosan, üzleti vagy politikai megfontolásokból történik

Köszönöm a figyelmet