AI LAB 2023

1.1 - INTRODUCTION

FRANCESCA M. BUFFA

LAB STRUCTURE

- INTRODUCTION
- THE DATA
- THE AI-LAB CHALLENGE – PART 1
- PART 1 - SHARING AND DISCUSSION

- UNSUPERVISED LEARNING EXAMPLES
- THE AI-LAB CHALLENGE – PART 2
- PART 2 - SHARING AND DISCUSSION

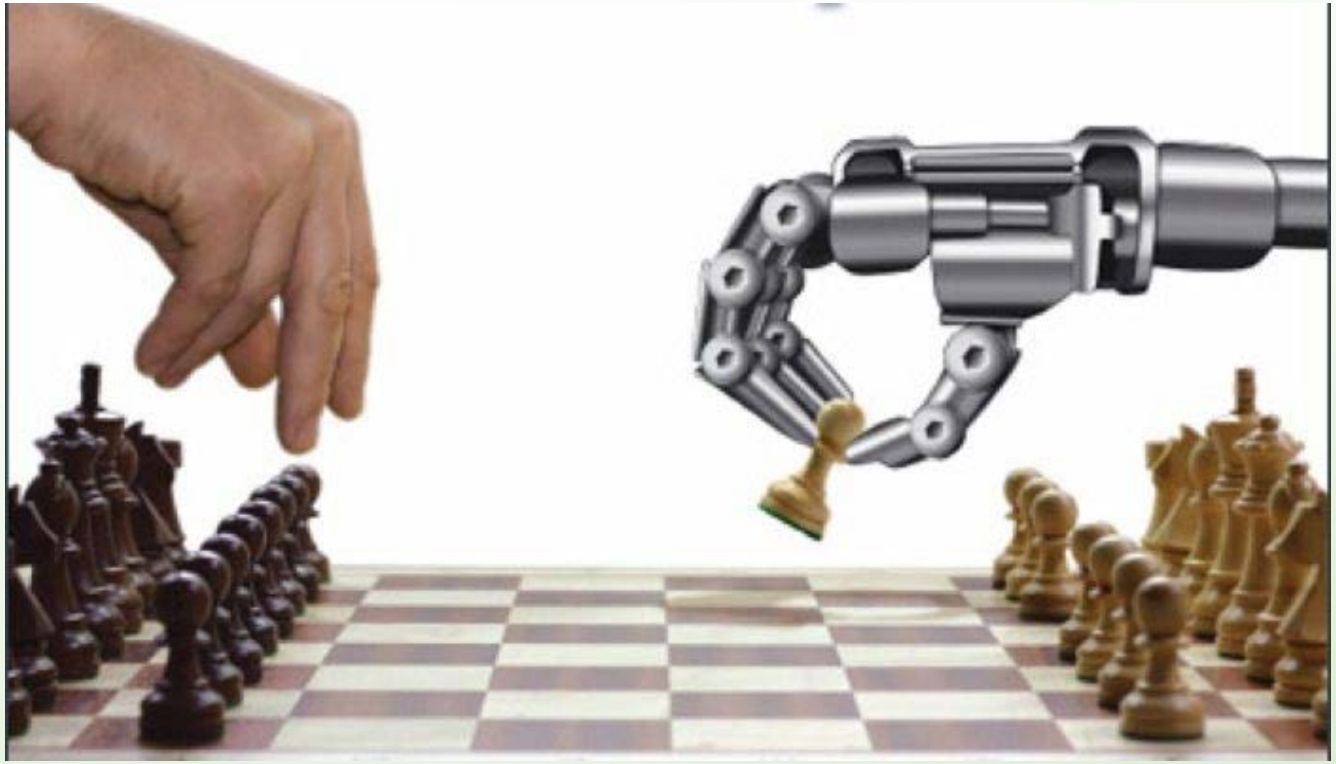
- SUPERVISED LEARNING EXAMPLES
- THE AI-LAB CHALLENGE – PART 3
- LARGE PROJECTS AND DATABASES

- THE AI-LAB CHALLENGE PARTS 1-3, SHARING AND DISCUSSION
- DATA INTERPRETATION
- DISCUSS AND PREPARE WORKSHOP PRESENTATIONS

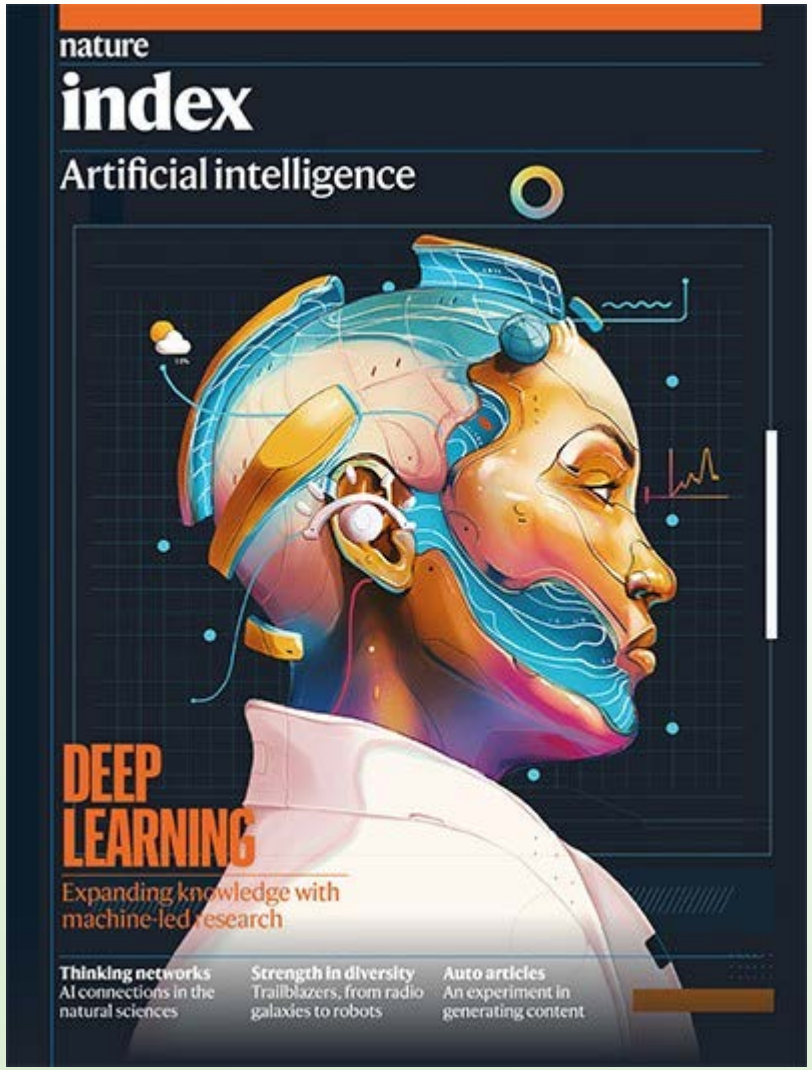
AI

“NARROW” AI

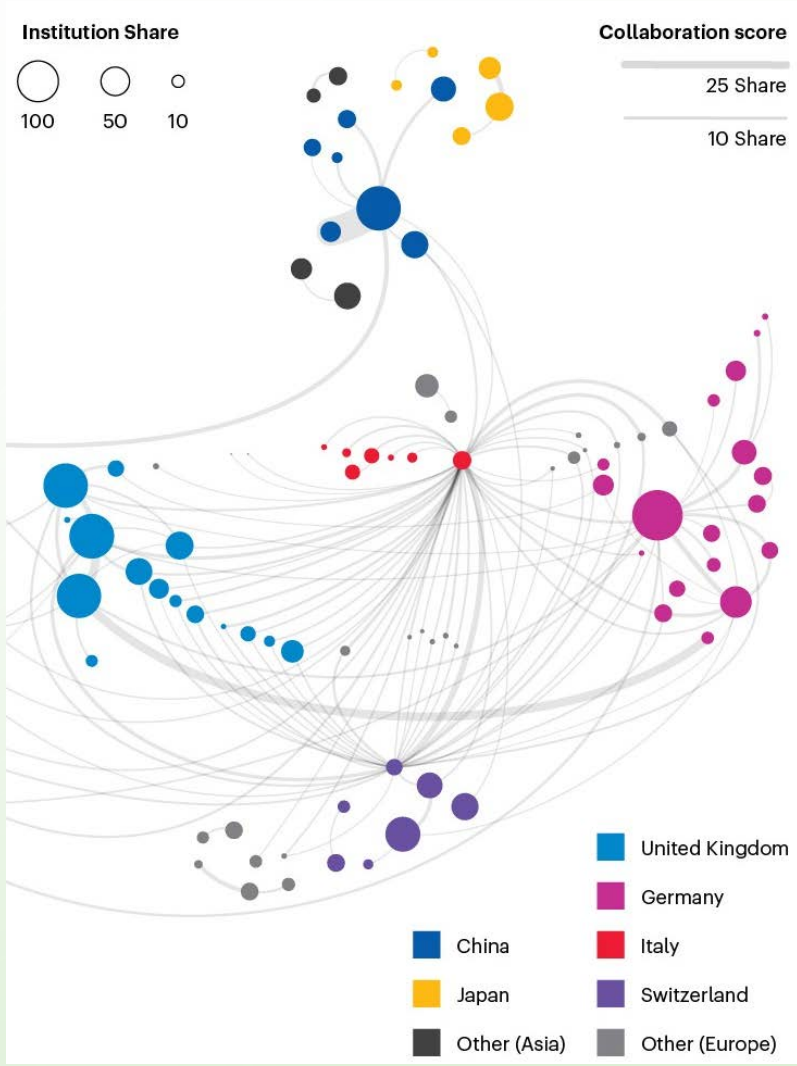
DEVELOPED FOR ONE SPECIFIC TASK



AI: THE INTERNATIONAL LANDSCAPE



Top 200 collaborations (lines) among 146 institutions (circles) between 2015 and 2019. Circles sized according to institution's share in AI. The thickness of the lines corresponds to a particular collaboration's total share between two institutions.



https://www.nature.com/articles/d41586-020-03409-8

https://www.nature.com/collections/hchdibjchj

TASKS THAT ARE HARDER FOR AI:

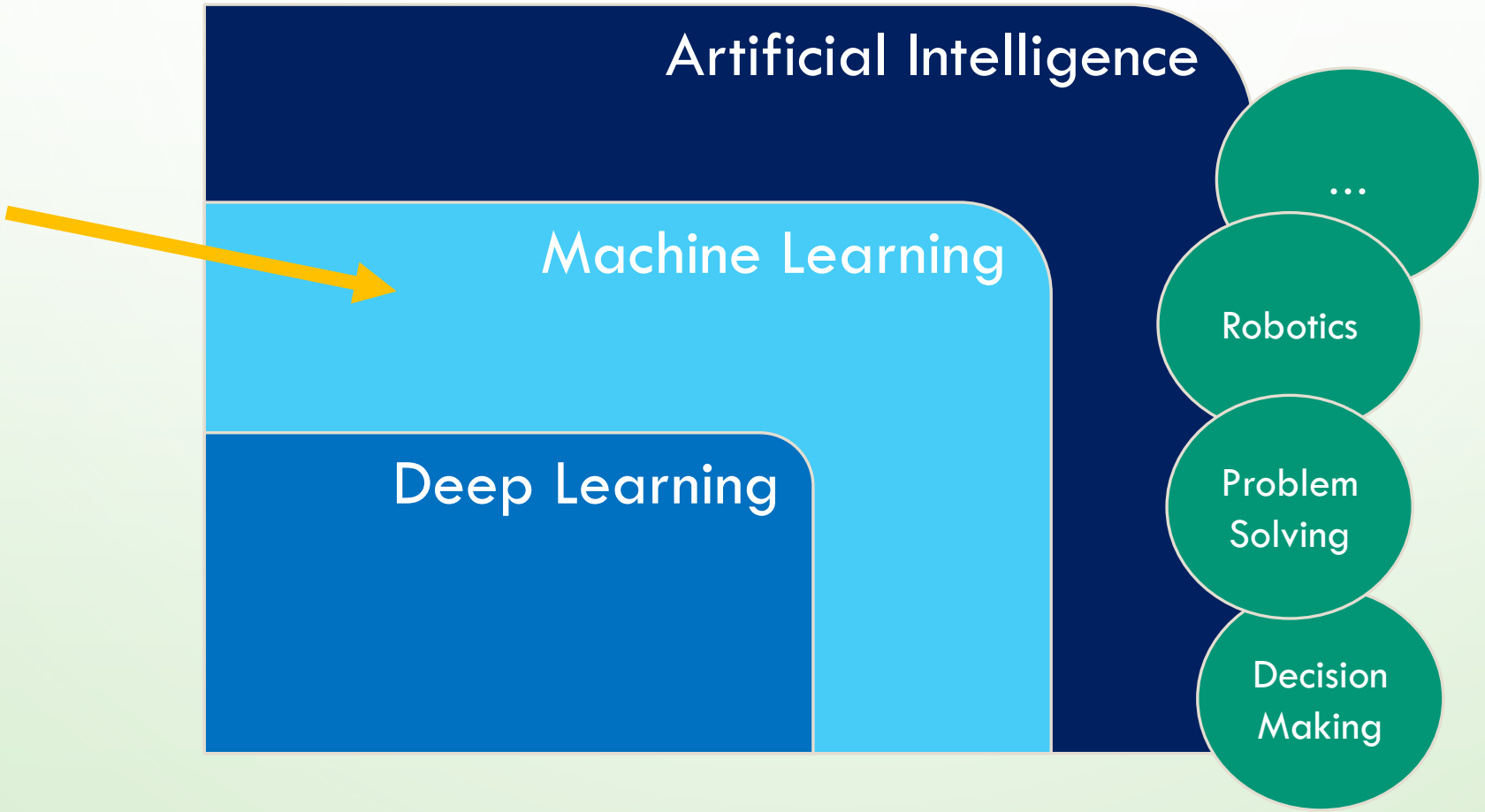
- UNSTRUCTURED DATA
  - NEED STRUCTURED DATA TO OPERATE EFFECTIVELY (EG. NUMERICAL DATA IN A DATABASE).
  - IMAGES, VIDEOS, AND TEXT DOCUMENTS ARE HARDER
- AMBIGUITY AND UNCERTAINTY
  - INCOMPLETE, LIMITED OR CONFLICTING INFORMATION CAN MAKE IT DIFFICULT FOR AI TO MAKE ACCURATE PREDICTIONS
- CREATIVITY AND INTUITION
  - NEW IDEAS AND SOLUTIONS ARE POSSIBLE BUT LIMITED AS AI RELIES ON PRE-EXISTING DATA AND PATTERNS
- CONTEXTUAL UNDERSTANDING
  - SOCIAL, CULTURAL, AND HISTORICAL FACTORS THAT SHAPE HUMAN BEHAVIOUR
  - HARD INTERPRET HUMAN LANGUAGE, BEHAVIOUR, AND INTERACTIONS.
- LEARNING FROM LIMITED DATA
  - NEED LARGE AMOUNTS OF DATA TO TRAIN AI MODELS
  - PERFORMANCE NOT EASY TO IMPROVE WHEN DATA ARE LIMITED OR INCOMPLETE
  - INACCURATE OR BIASED PREDICTIONS

TASKS WHERE AI CAN OUTPERFORM HUMANS:

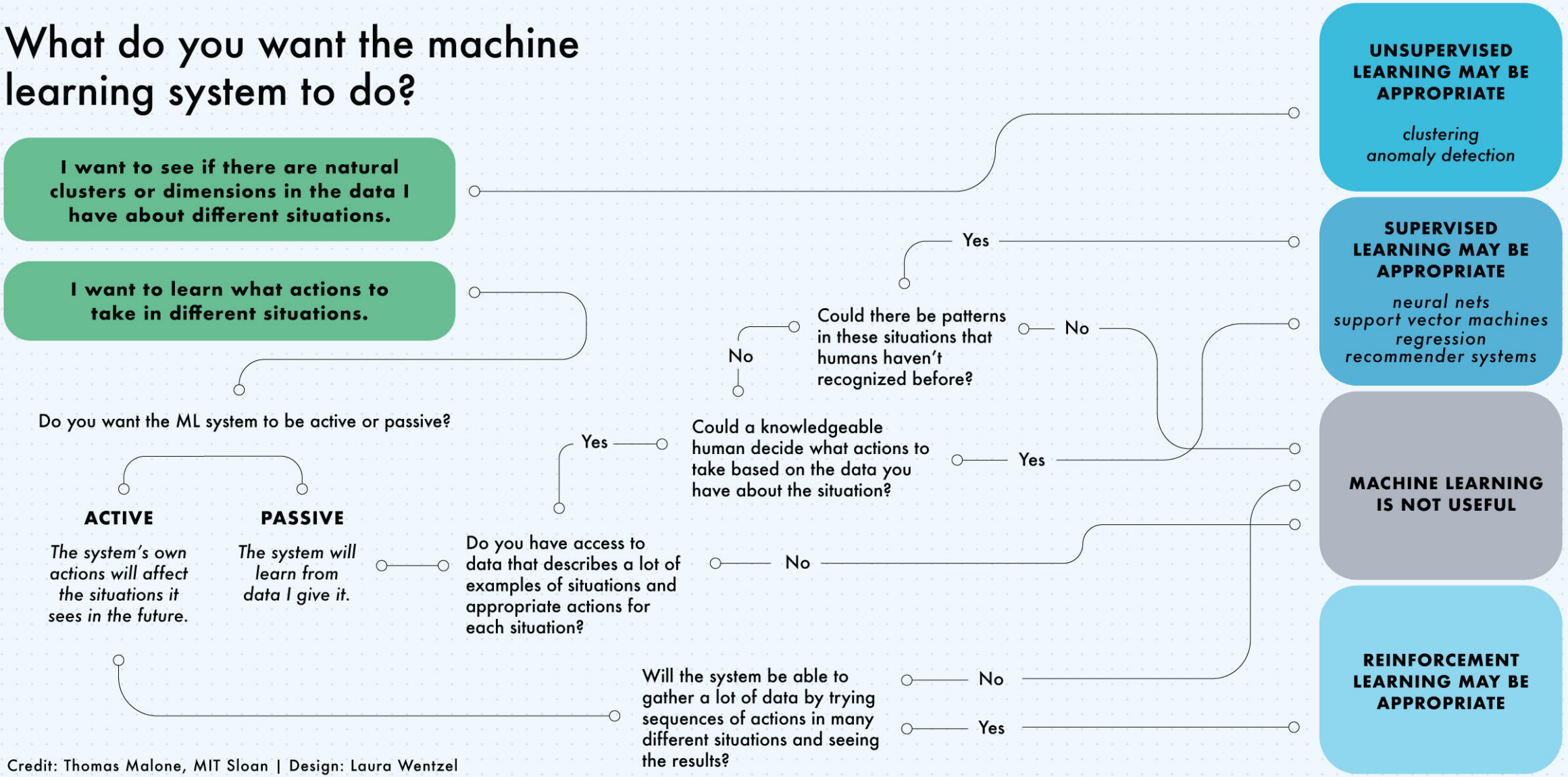
- COMPLEX CALCULATIONS
  - AI IS FASTER AND MORE ACCURATELY THAN HUMANS
- DATA ANALYSIS
  - AI CAN PROCESS LARGE AMOUNTS OF DATA QUICKLY AND ACCURATELY
- PATTERN RECOGNITION
  - AI CAN IDENTIFY PATTERNS IN DATA OR IMAGES THAT MAY BE DIFFICULT FOR HUMANS TO DISCERN, OR TOO RARE
- REPETITIVE TASKS
  - ASSEMBLY DATA ENTRY WITHOUT GETTING BORED OR MAKING MISTAKES
- OBJECT TRACKING
  - TRACK OBJECTS AND MOVEMENTS ACROSS LARGE AREAS, SECURITY SURVEILLANCE OR TRAFFIC MONITORING
- PLAYING GAMES
  - ABLE TO BEAT HUMAN PLAYERS AT COMPLEX GAMES SUCH AS CHESS, GO, AND POKER.

OVERALL, TASKS THAT REQUIRE: PROCESSING LARGE AMOUNTS OF DATA, PERFORMING COMPLEX CALCULATIONS OR SIMULATIONS, IDENTIFYING PATTERNS RARE OR DIFFICULT TO DETECT PATTERNS.

AI, MACHINE LEARNING



APPLICATION OF MACHINE LEARNING



Thomas Malone | MIT Sloan. https://bit.ly/3gvRho2

TYPICAL STEPS IN MACHINE LEARNING APPLICATIONS

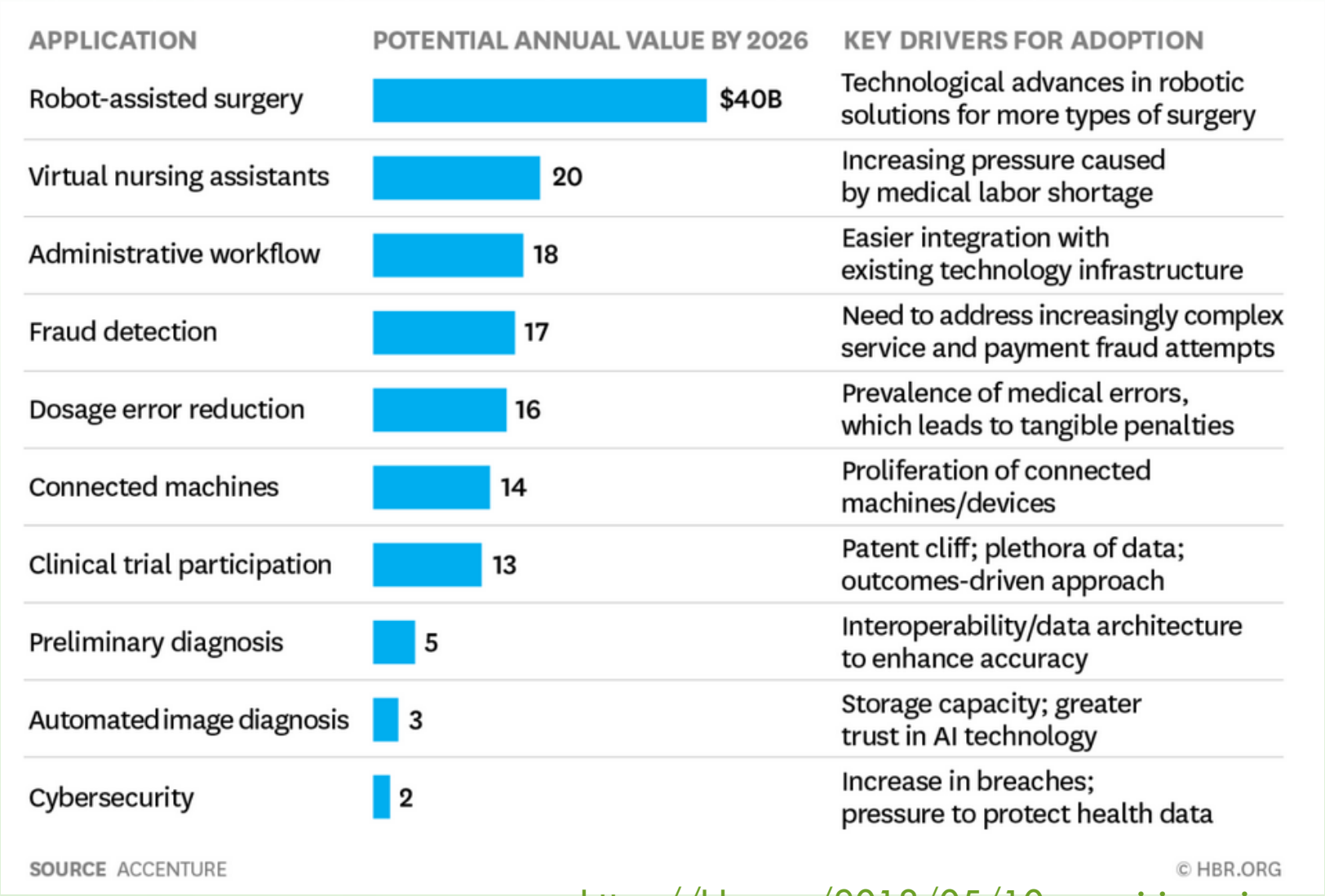
- UNDERSTAND THE APPLICATION CONTEXT
- ASK A QUESTION
- GENERATE OR COLLECT THE DATA
- LOOK AT THE DATA
- UNDERSTAND THE DATA
- CLEAN AND PROCESS THE DATA
- SELECT ALGORITHM OR ALGORITHMS
- PLAN THE TRAINING APPROACH
- TRAIN AND VALIDATE YOUR MODEL
- FINALIZE YOUR MODEL
- TEST YOUR MODEL BY PREDICTING ON NEW DATA

UNDERSTAND THE CONTEXT OF APPLICATION

- YOUR CONTEXT WILL BE LIFE SCIENCES AND HEALTH
- WHY DID WE CHOOSE THIS CONTEXT?
- INTRODUCTION TO THE BASIC CONCEPTS
- EXAMPLES OF PREVIOUS AI APPLICATIONS

HEALTH MAJOR APPLICATION FOR AI

Harvard Business Review

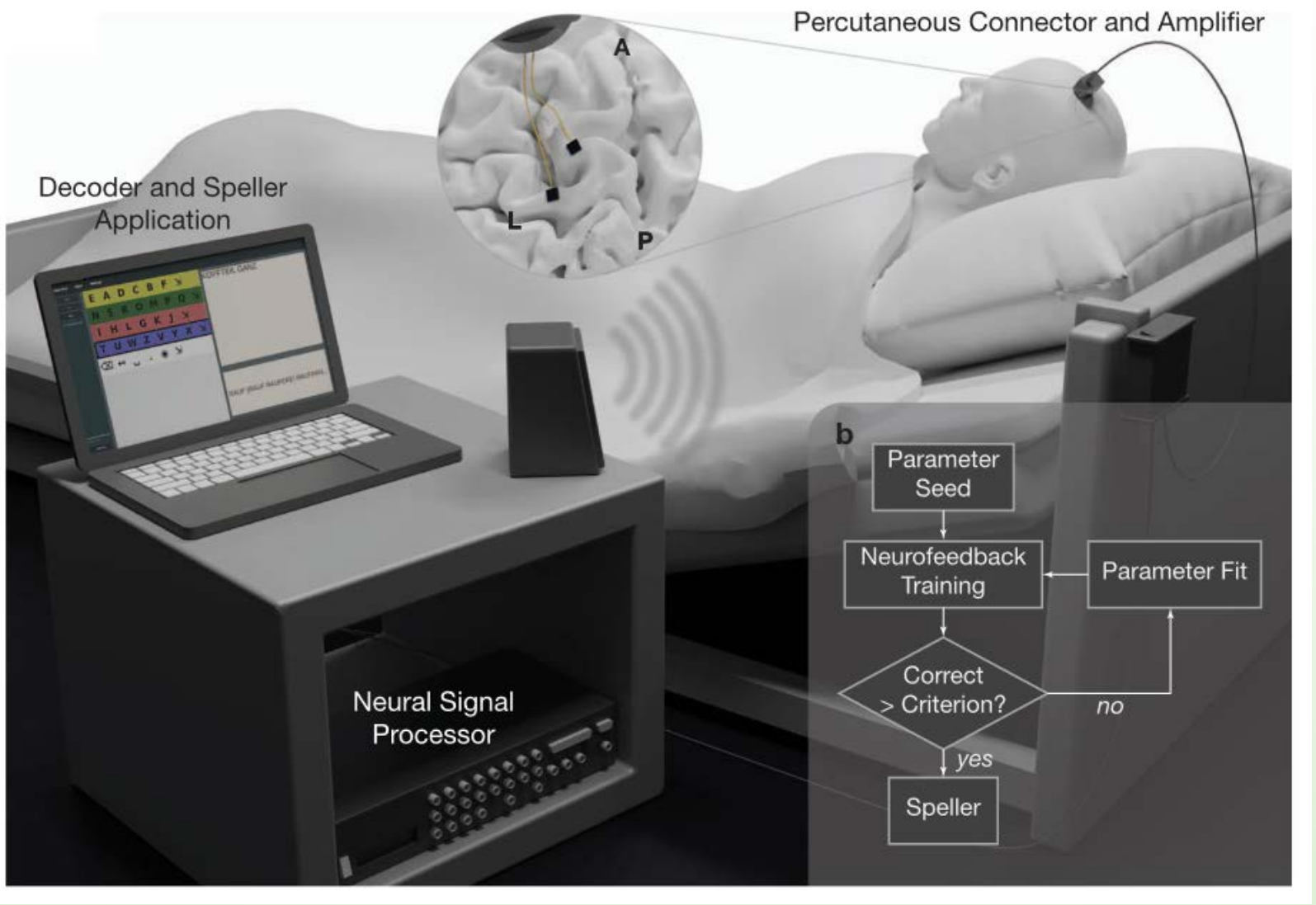


APPLICATION	POTENTIAL ANNUAL VALUE BY 2026	KEY DRIVERS FOR ADOPTION
Robot-assisted surgery	$40B	Technological advances in robotic solutions for more types of surgery
Virtual nursing assistants	20	Increasing pressure caused by medical labor shortage
Administrative workflow	18	Easier integration with existing technology infrastructure
Fraud detection	17	Need to address increasingly complex service and payment fraud attempts
Dosage error reduction	16	Prevalence of medical errors, which leads to tangible penalties
Connected machines	14	Proliferation of connected machines/devices
Clinical trial participation	13	Patent cliff; plethora of data; outcomes-driven approach
Preliminary diagnosis	5	Interoperability/data architecture to enhance accuracy
Automated image diagnosis	3	Storage capacity; greater trust in AI technology
Cybersecurity	2	Increase in breaches; pressure to protect health data

SOURCE ACCENTURE

https://hbr.org/2018/05/10-promising-ai-applications-in-health-care

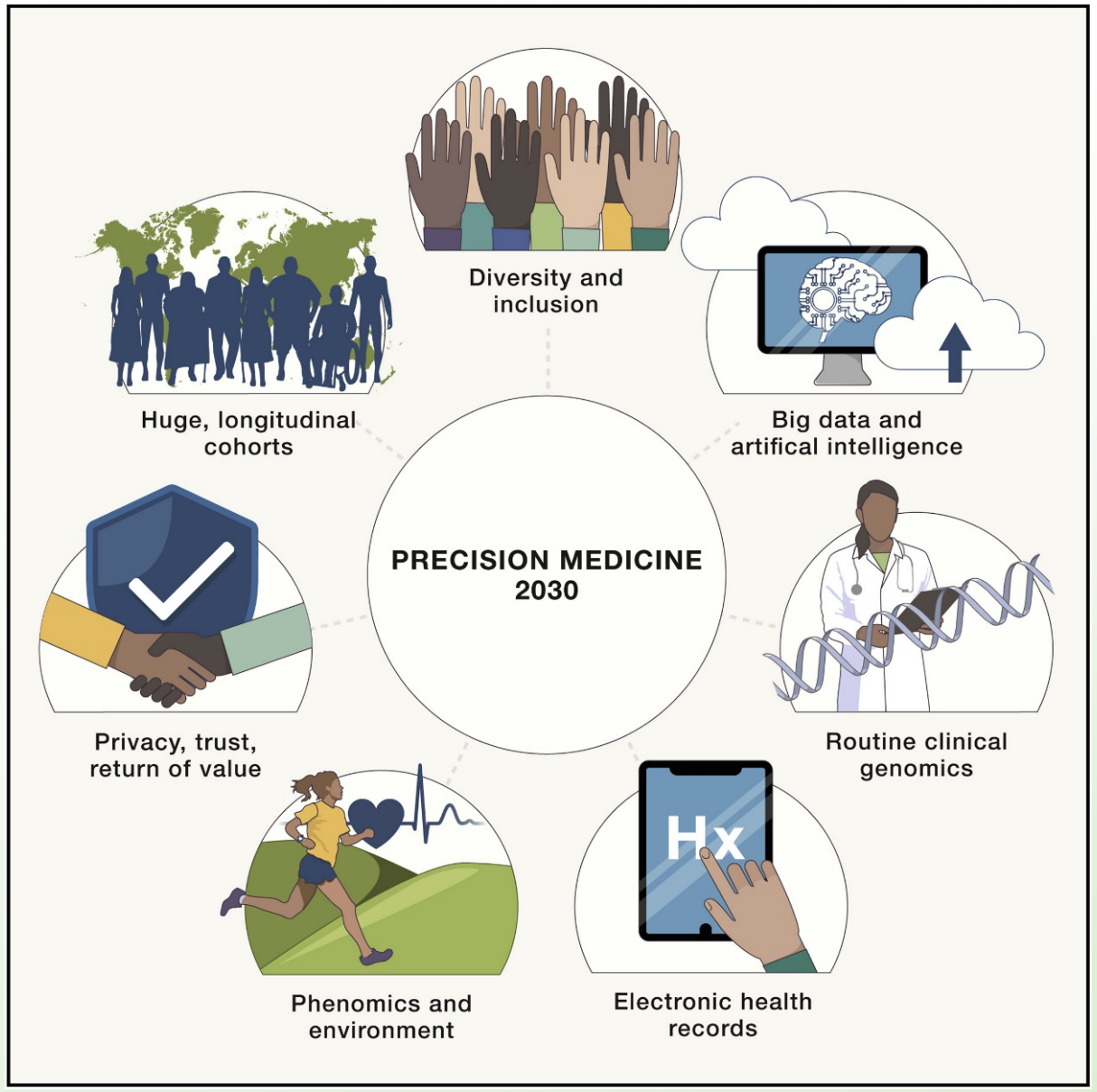
SPELLING INTERFACE USING INTRACORTICAL SIGNALS ENABLED VIA AUDITORY NEUROFEEDBACK TRAINING



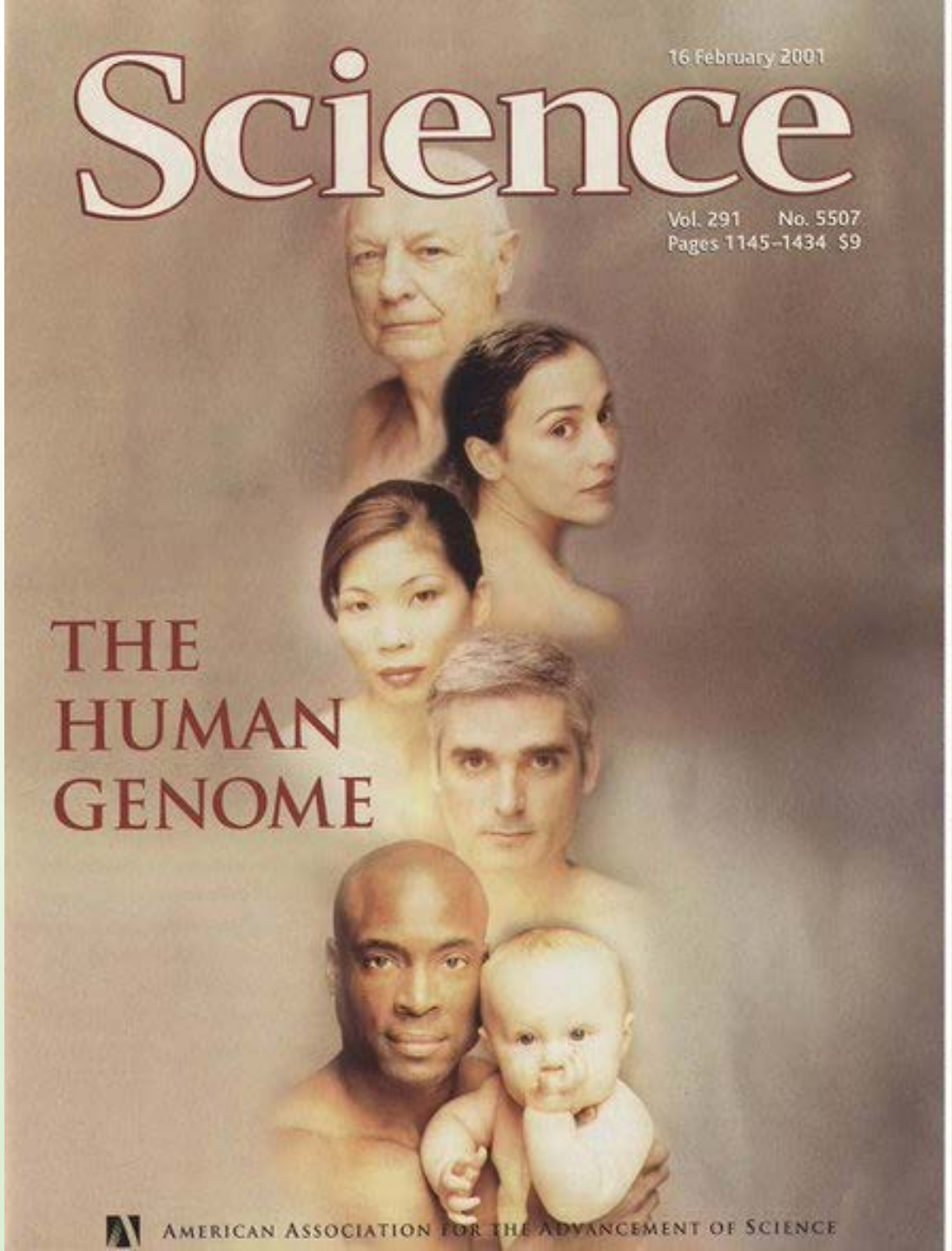
https://www.nature.com/articles/s41467-022-28859-8

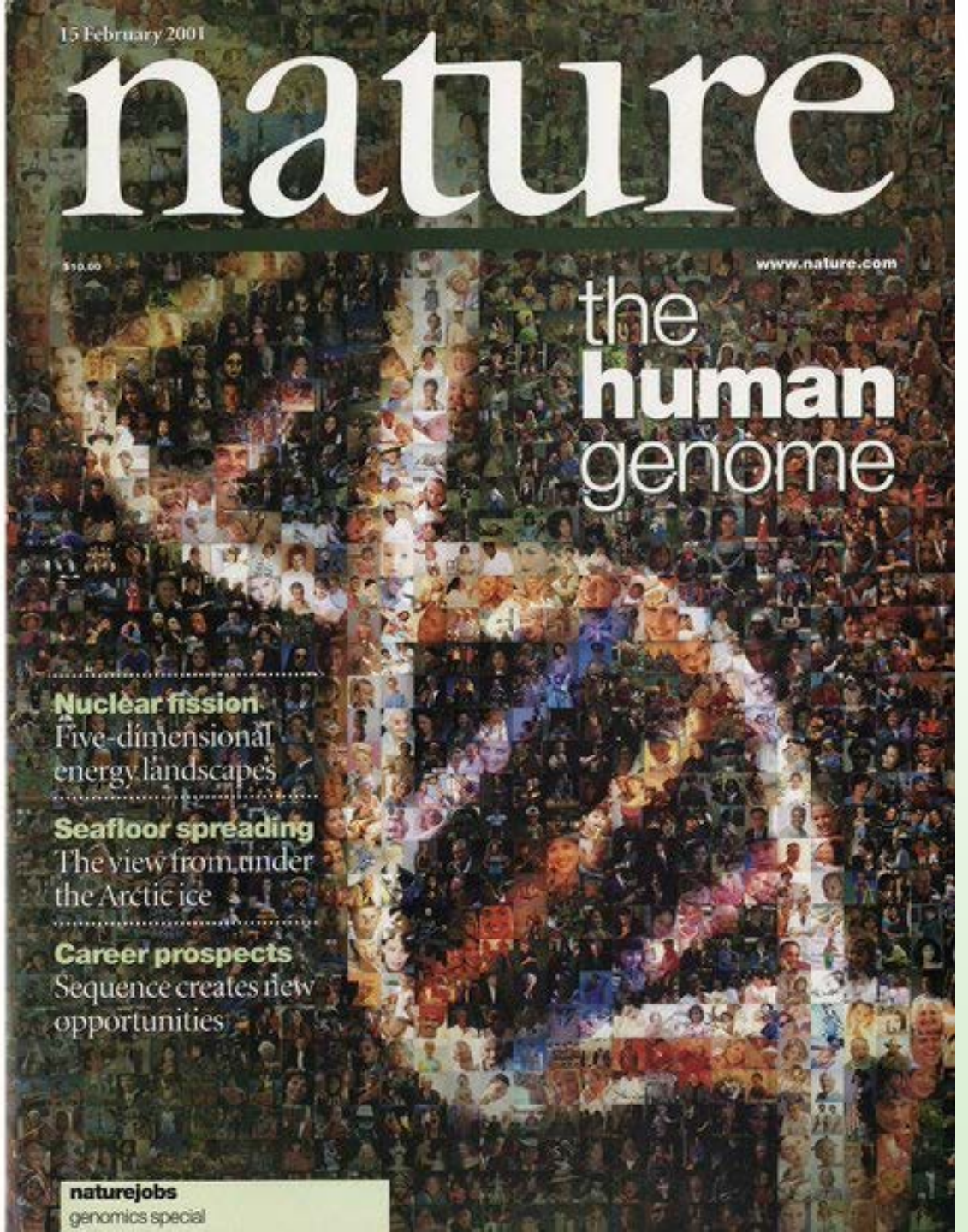
AI IN HEALTH

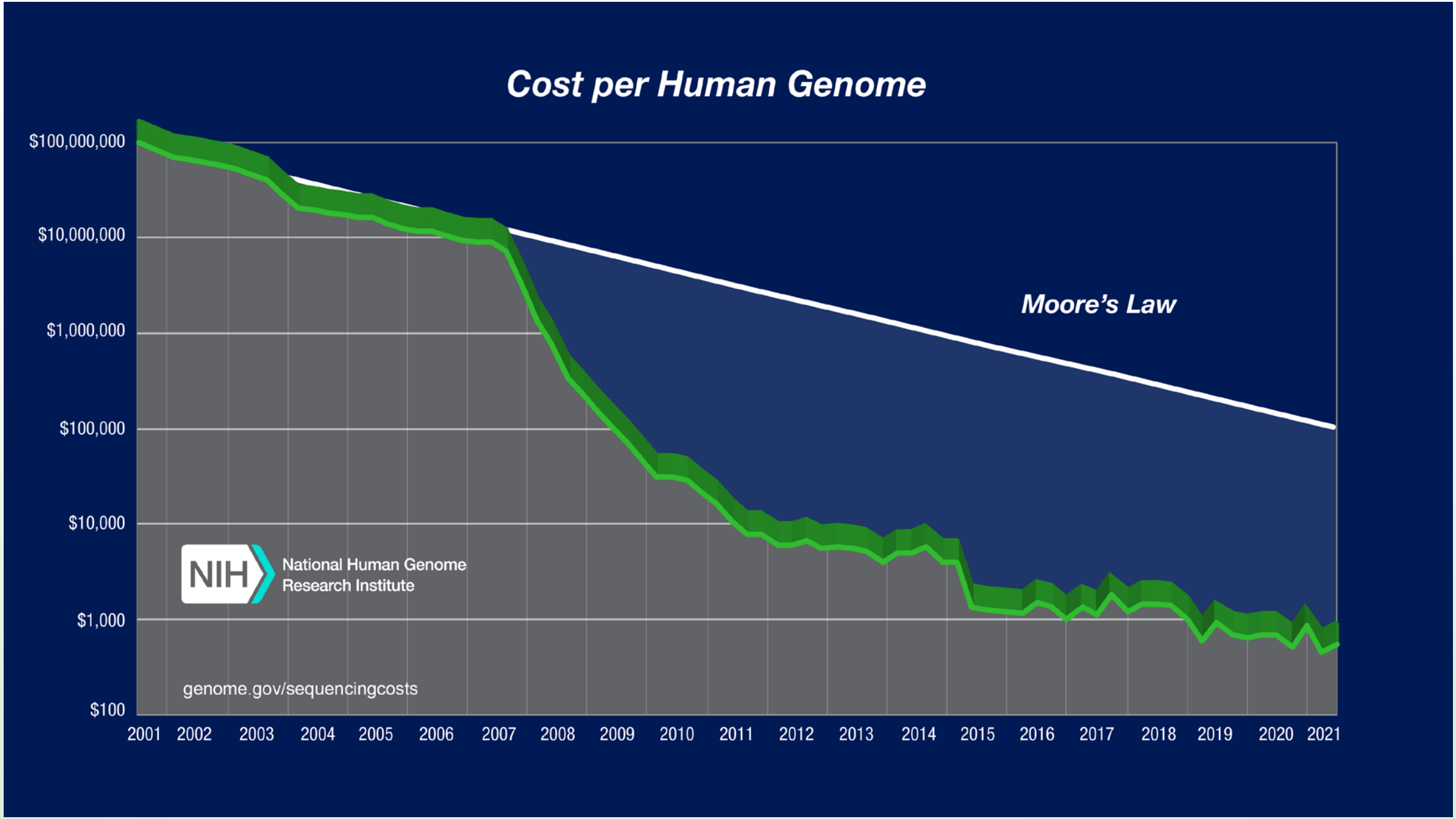
- New technologies enable to obtain and store large amount of biological and medical data.
- Both the quantity and type of data collected has changed dramatically, and is very diverse.
- To achieve impact on health we need to integrate and transform such data in information.
- This requires collaboration between disciplines: medicine, biology, genomics, engineering, ethics, law, big data, maths, stats, AI.



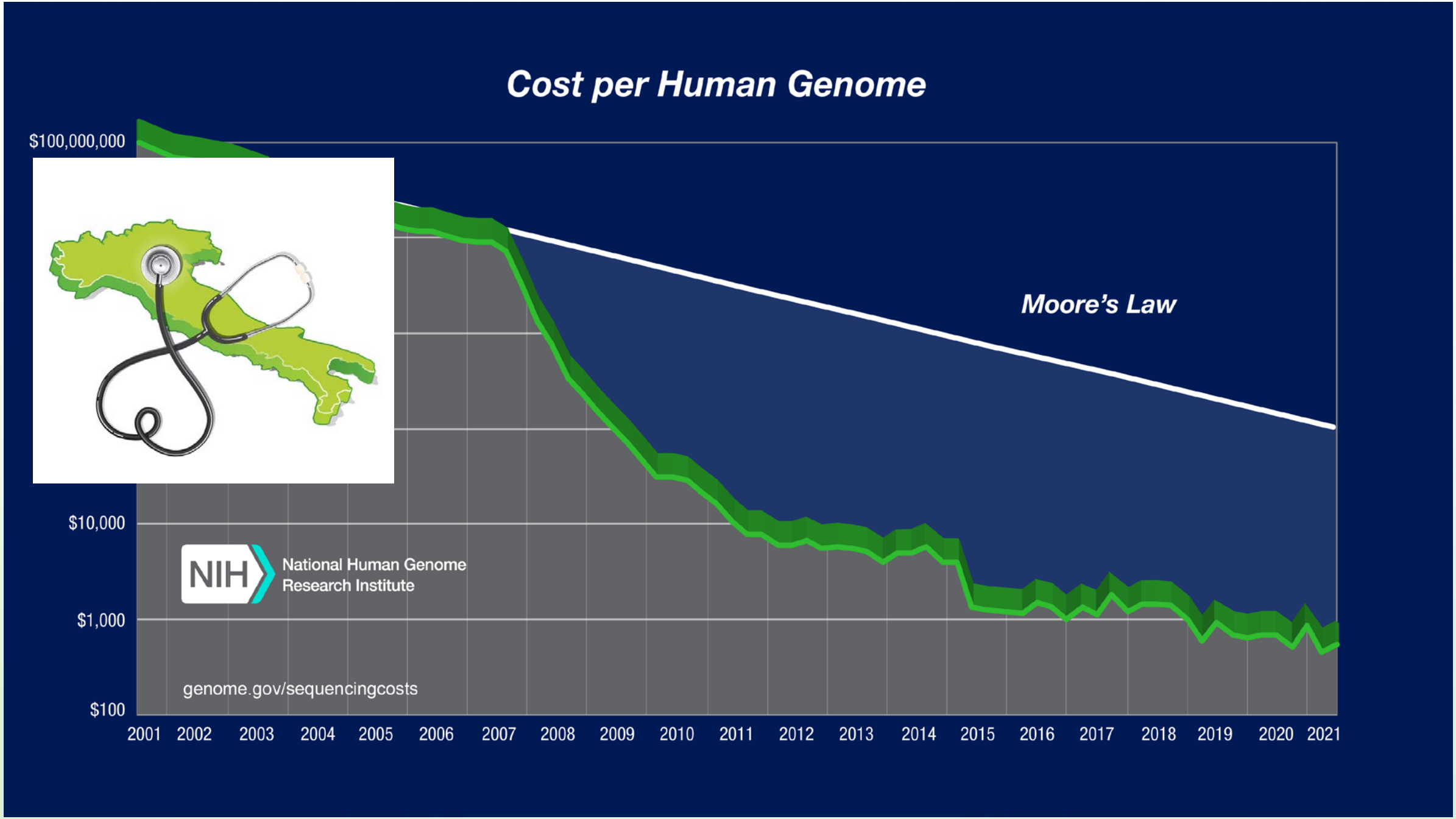
Denny & Collins, Cell, March 2021


16 February 2001
Science
Vol. 291 No. 5507
Pages 1145–1434 $9
THE HUMAN GENOME
AMERICAN ASSOCIATION FOR THE ADVANCEMENT OF SCIENCE

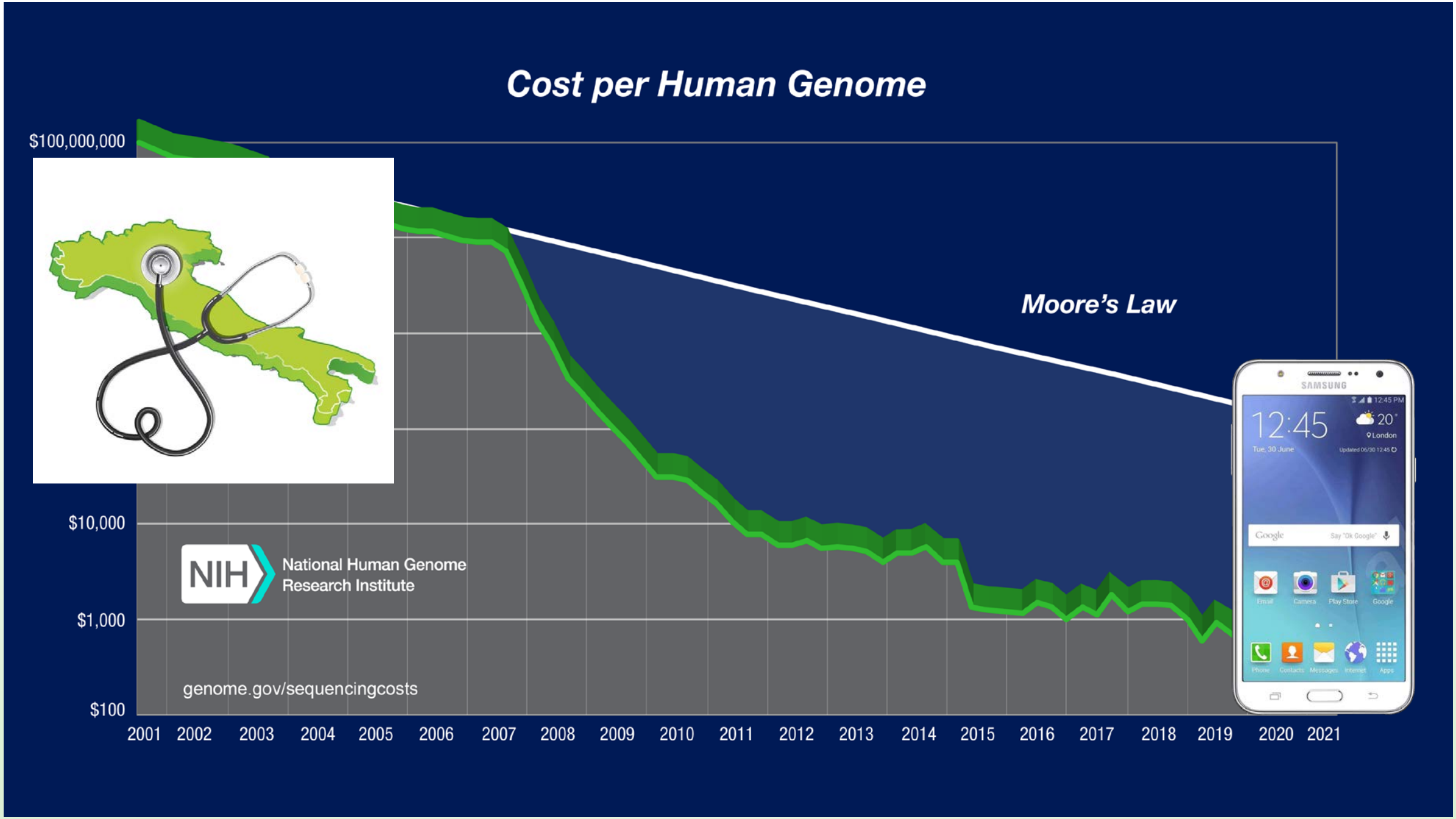

15 February 2001
nature
$10.00
www.nature.com
the human genome
Nuclear fission
Five-dimensional energy landscapes
Seafloor spreading
The view from under the Arctic ice
Career prospects
Sequence creates new opportunities
naturejobs
genomics special



https://www.genome.gov/about-genomics/fact-sheets/Genomic-Data-Science

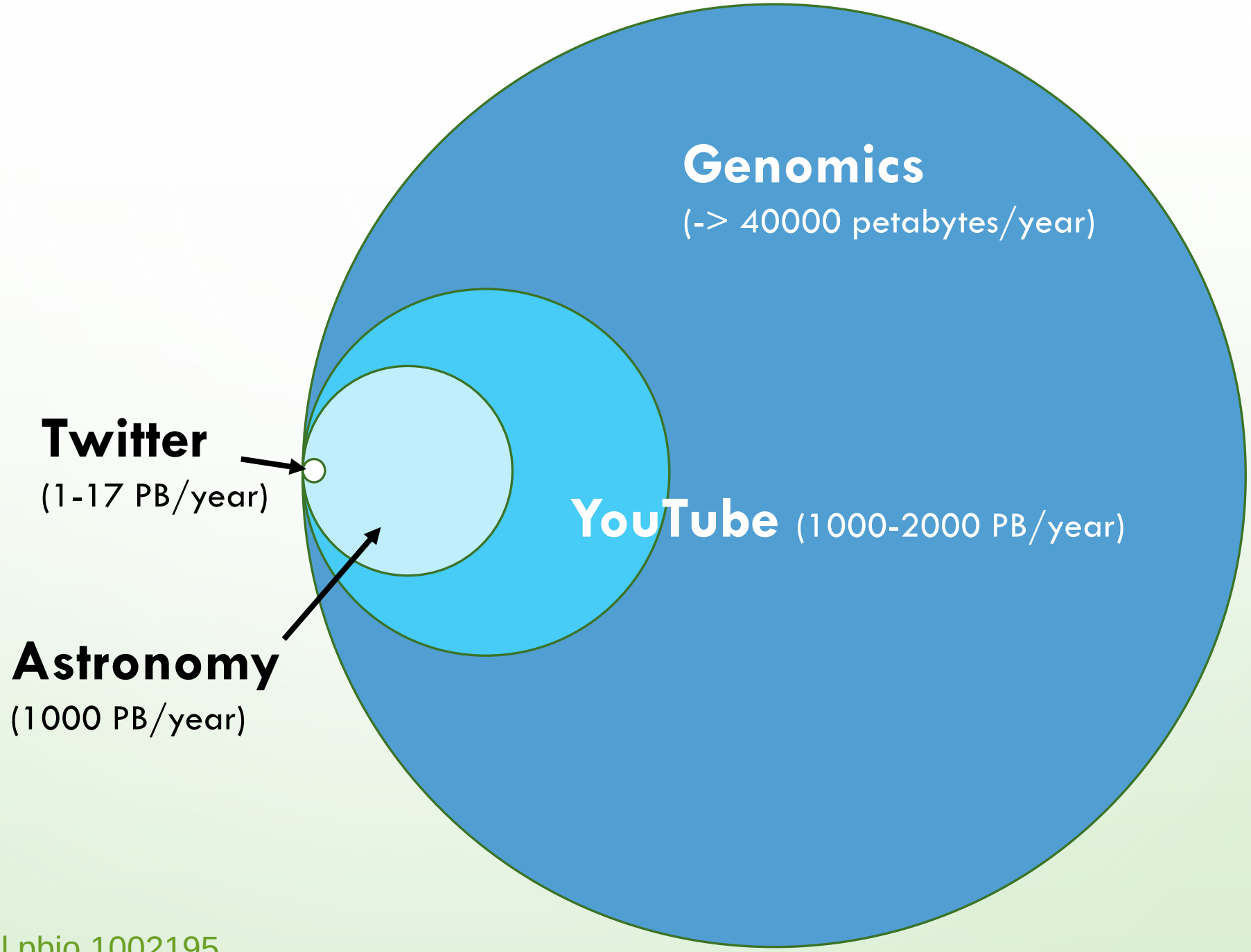


https://www.genome.gov/about-genomics/fact-sheets/Genomic-Data-Science



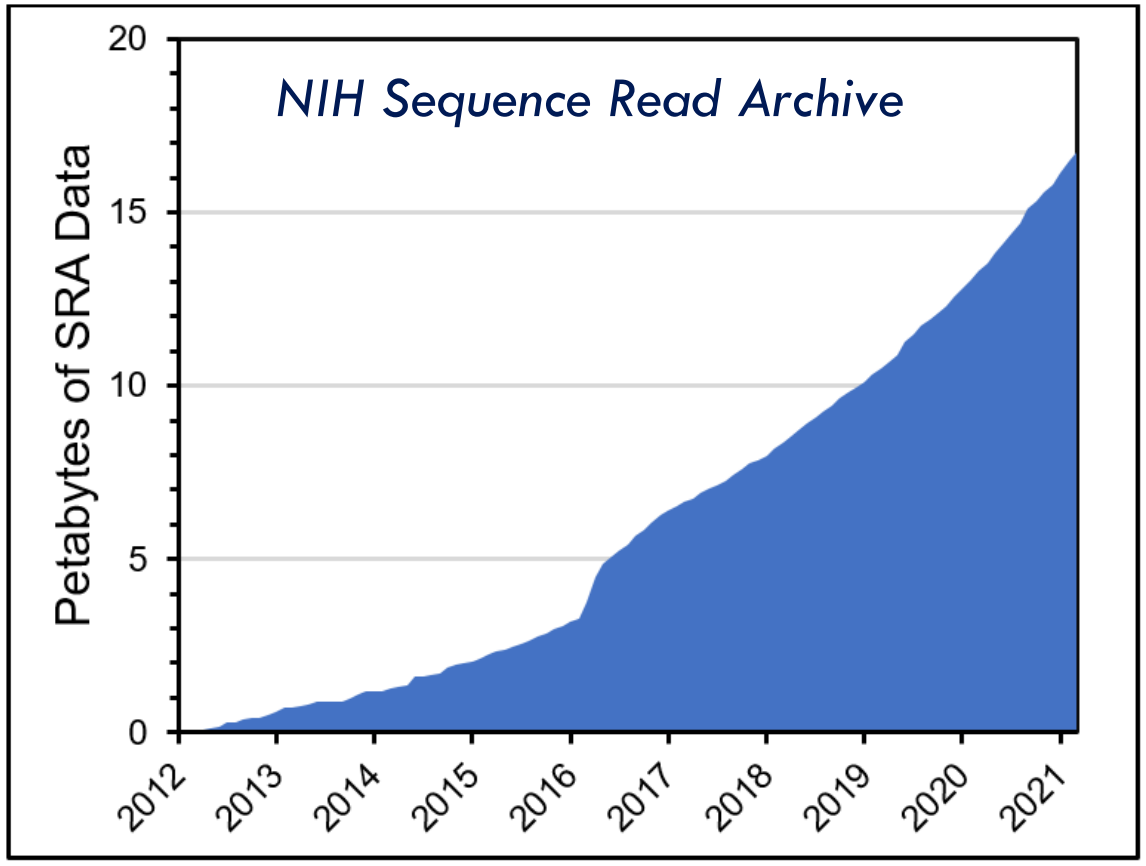
https://www.genome.gov/about-genomics/fact-sheets/Genomic-Data-Science

Big data

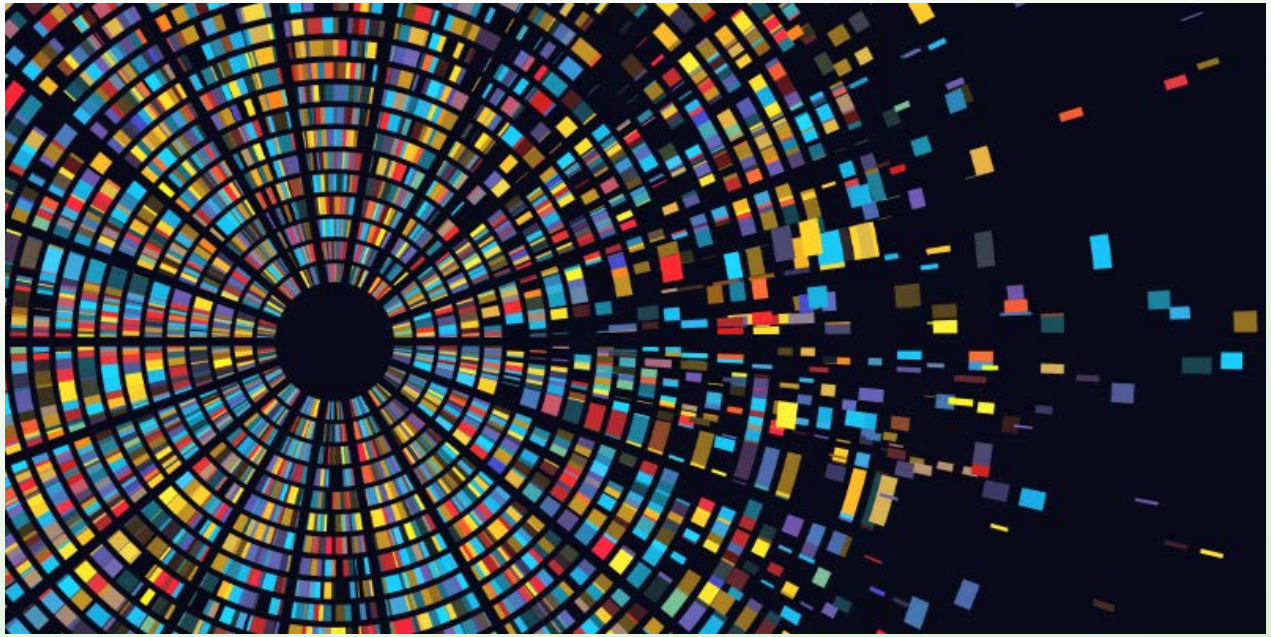


https://doi.org/10.1371/journal.pbio.1002195

TACKLING PETABYTE SCALE SEQUENCE SEARCH CHALLENGES



Dr. Susan Gregurick, NIH Associate Director for Data Science and ODSS Director, said: "We all share a common problem and a need to develop, enhance, and implement methods that streamline data access, search or findability, and ultimately data reuse."



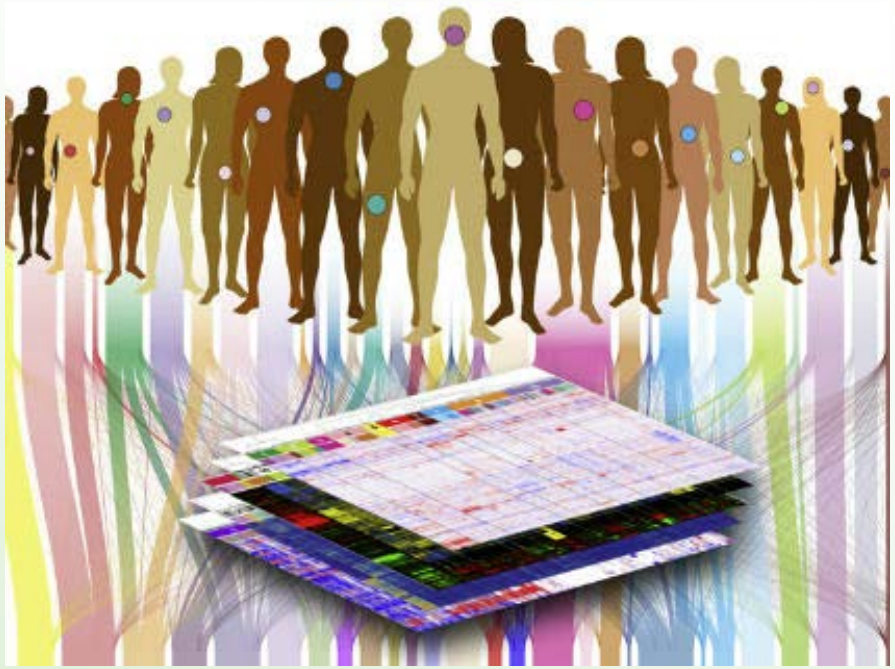
https://ncbiinsights.ncbi.nlm.nih.gov/2021/08/09/espsss-workshop/

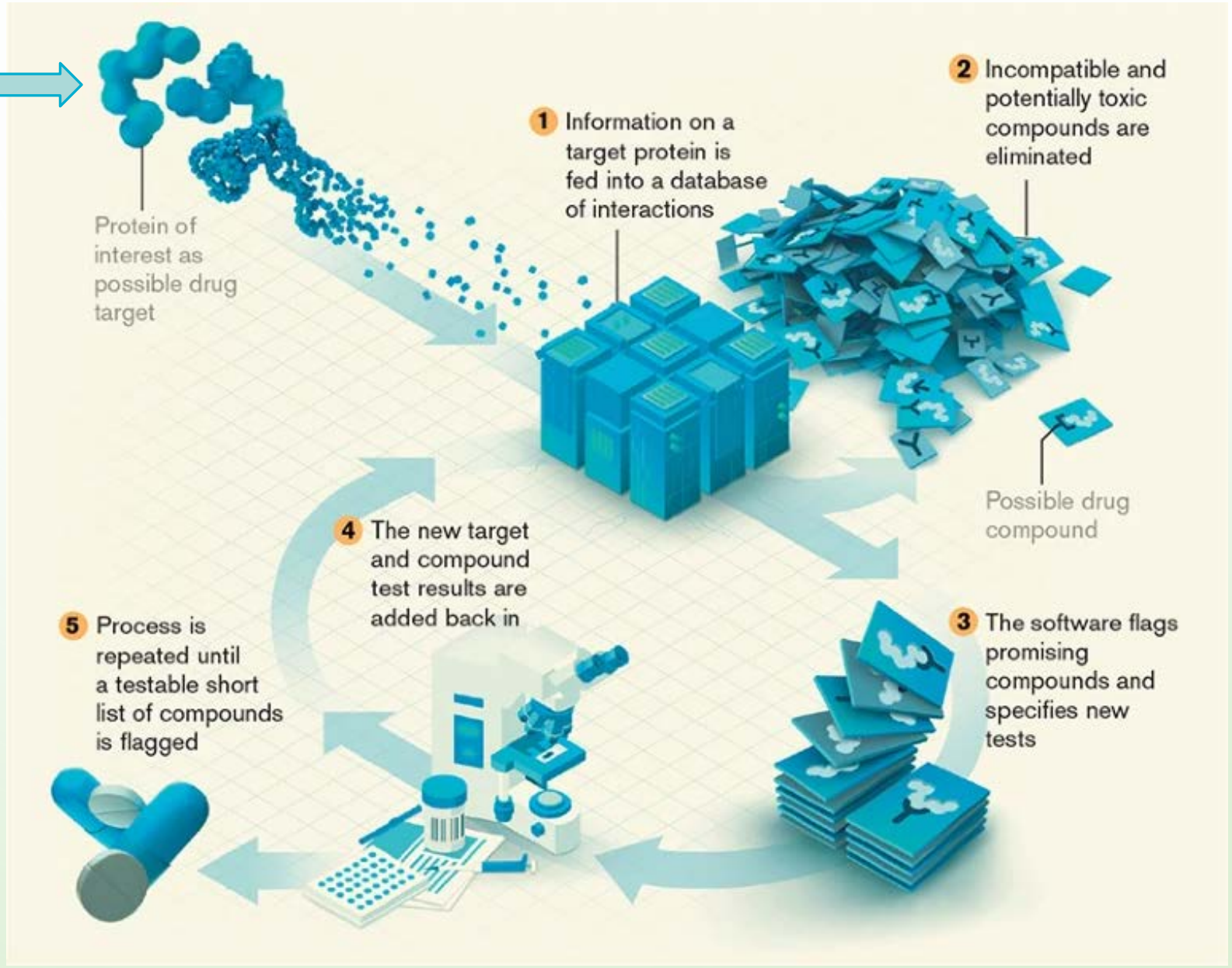
Your genes +

Your environment & lifestyle =

YOU!

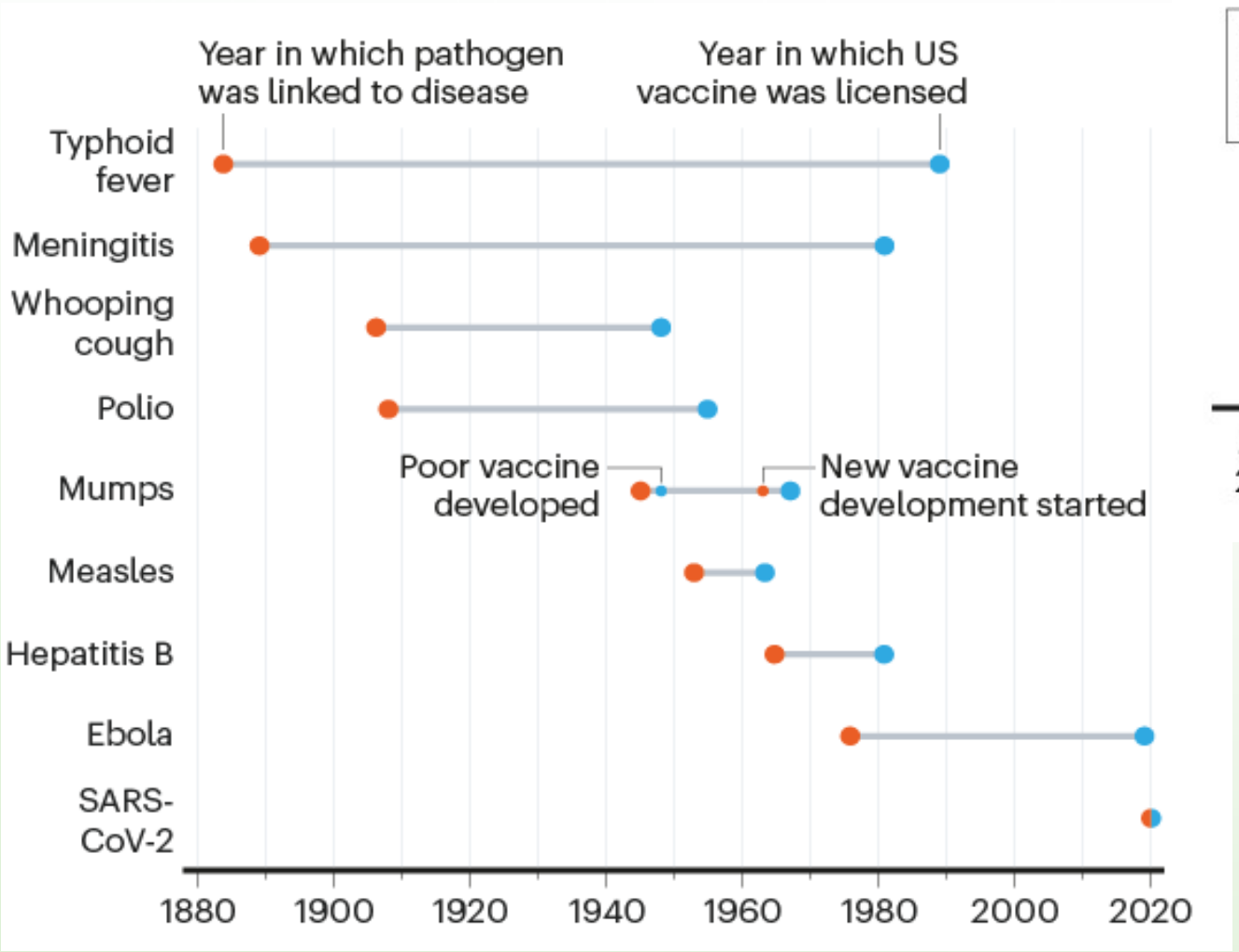
SPEEDING UP THE SEARCH FOR DRUGS WITH AI

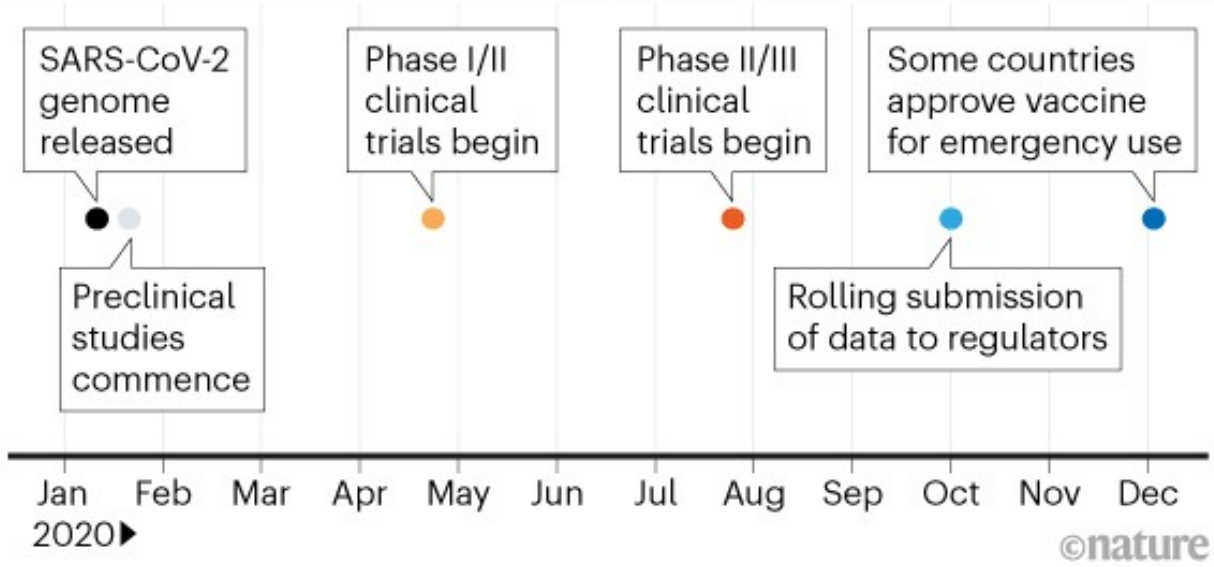




https://www.nature.com/articles/d41586-019-03846-0

AI and vaccine development



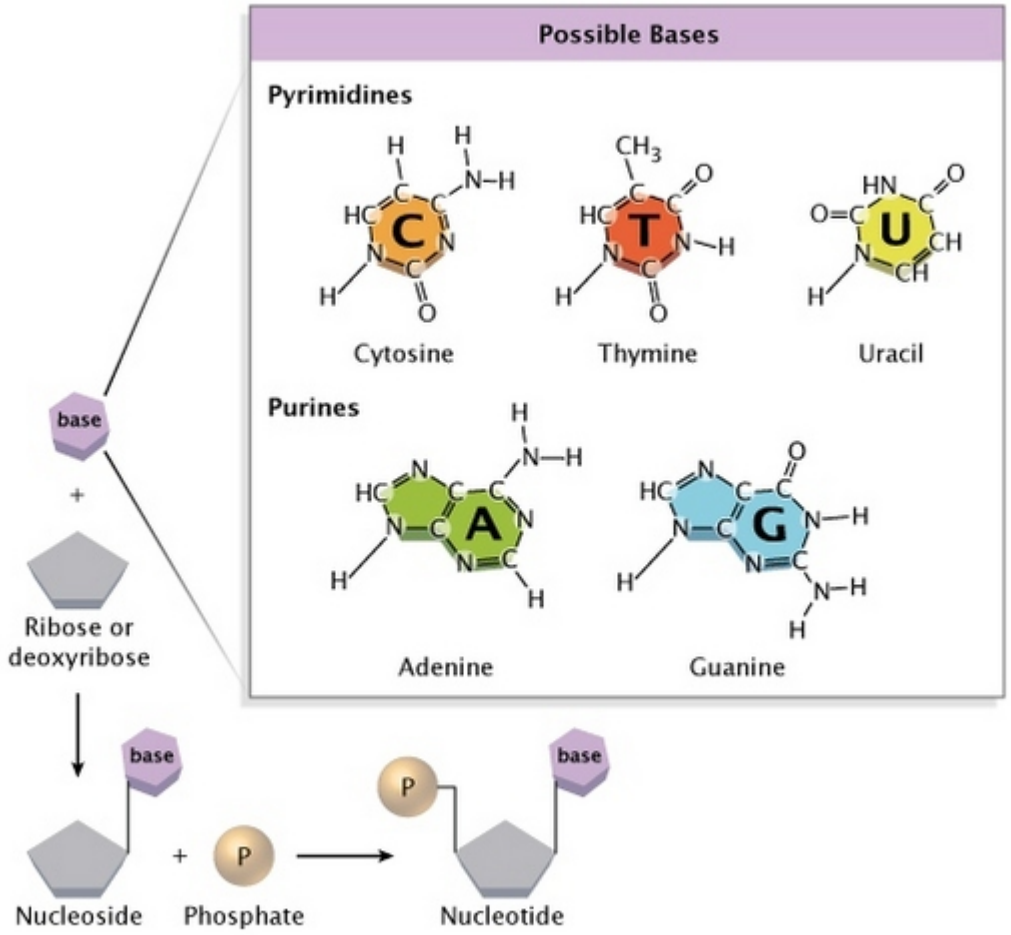


Dave Johnson, PhD in Information Physics, chief data and AI officer at Moderna: "We've seen how this digital infrastructure and how these algorithms can really help push things forward".

https://sloanreview.mit.edu/audio/ai-and-the-covid-19-vaccine-modernas-dave-johnson/

Nature News, 18 December 2020

DNA/RNA STRING REPRESENTATION



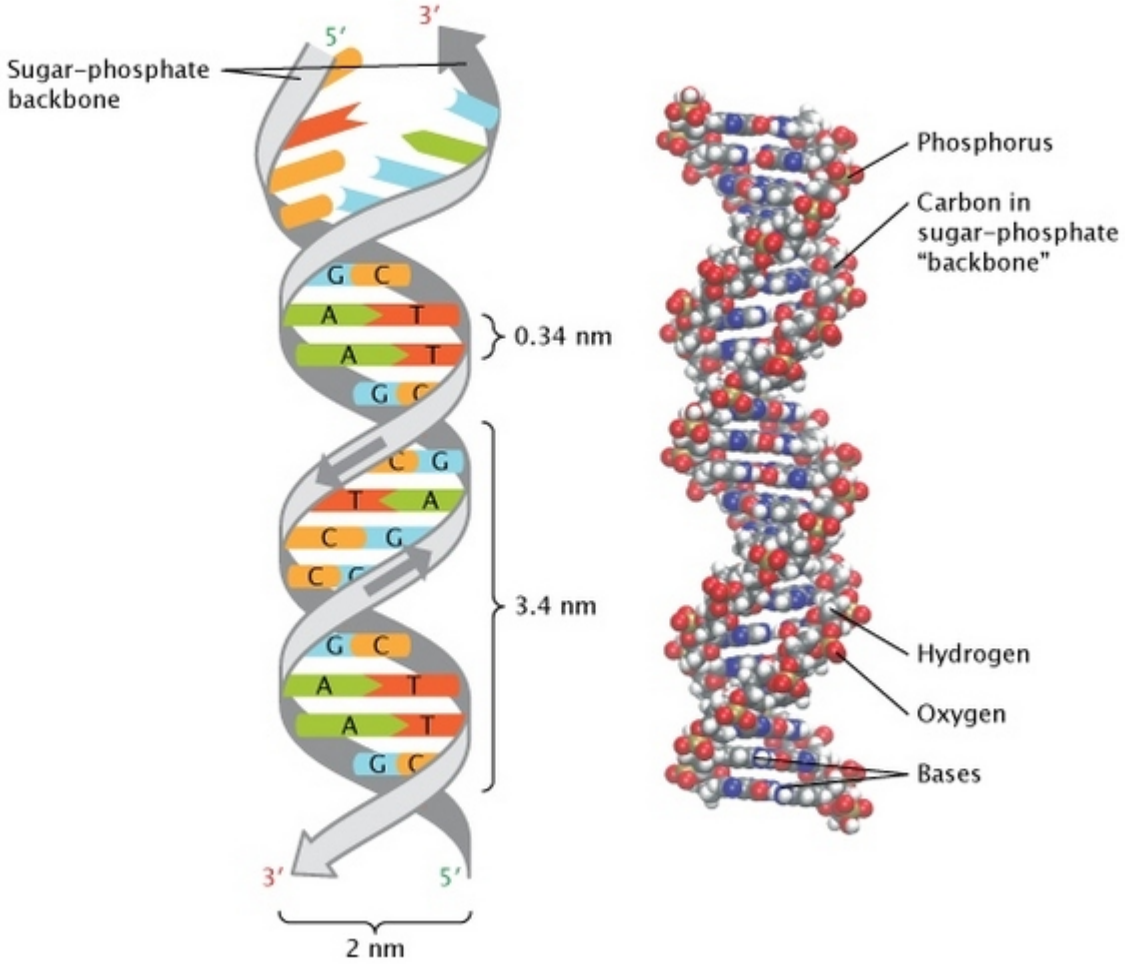
- A (ADENINE)
- T (THYMINE)
- C (CYTOSINE)
- G (GUANINE)

- ORGANIZED IN A DOUBLE HELIX OF PAIRED BASES: A-T, C-G

- IN RNA T -> U (URACIL)

https://www.nature.com/scitable/topicpage/discovery-of-dna-structure-and-function-watson-397/

DNA/RNA STRING REPRESENTATION



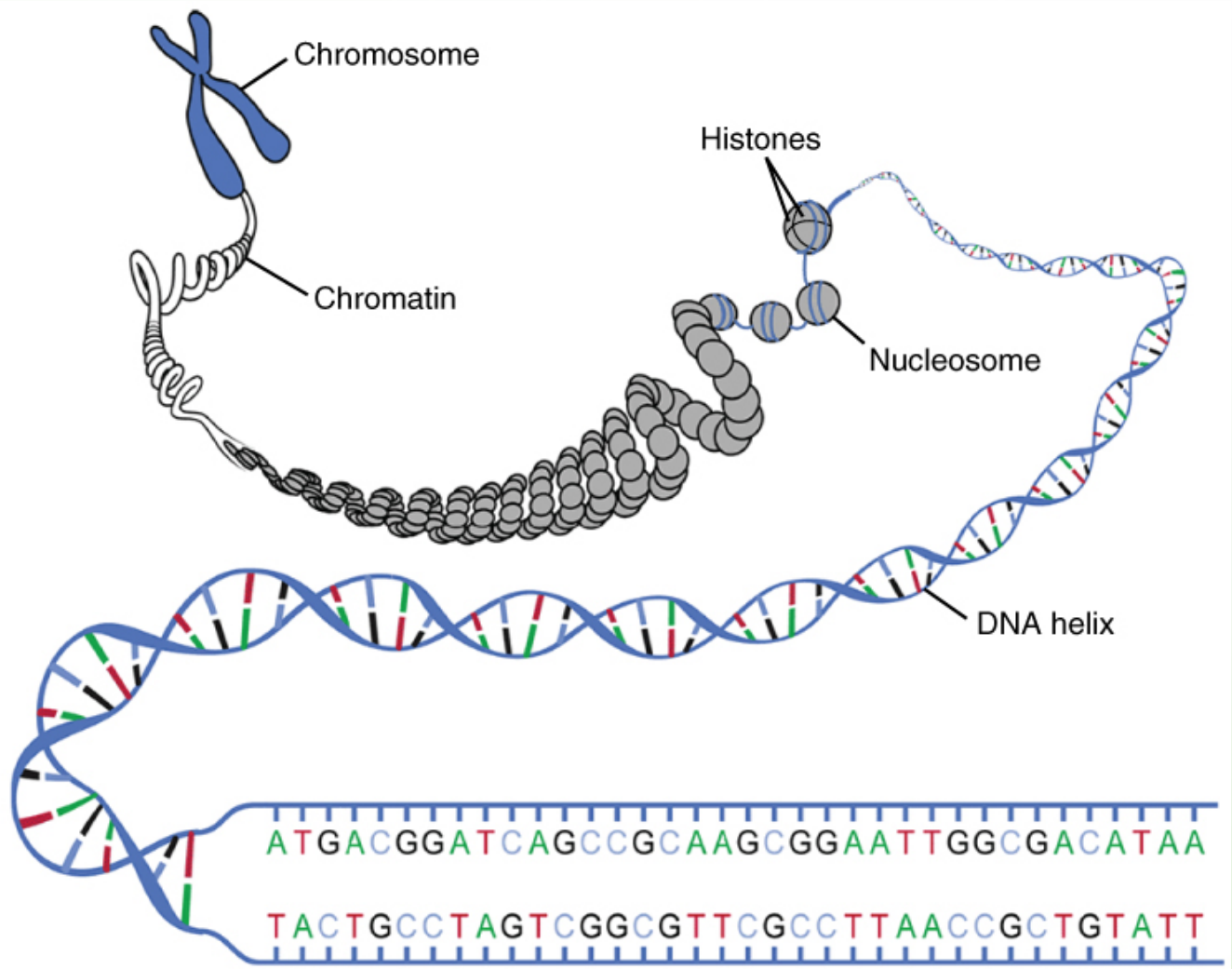
- A (ADENINE)
- T (THYMINE)
- C (CYTOSINE)
- G (GUANINE)

- ORGANIZED IN A DOUBLE HELIX OF PAIRED BASES: A-T, C-G

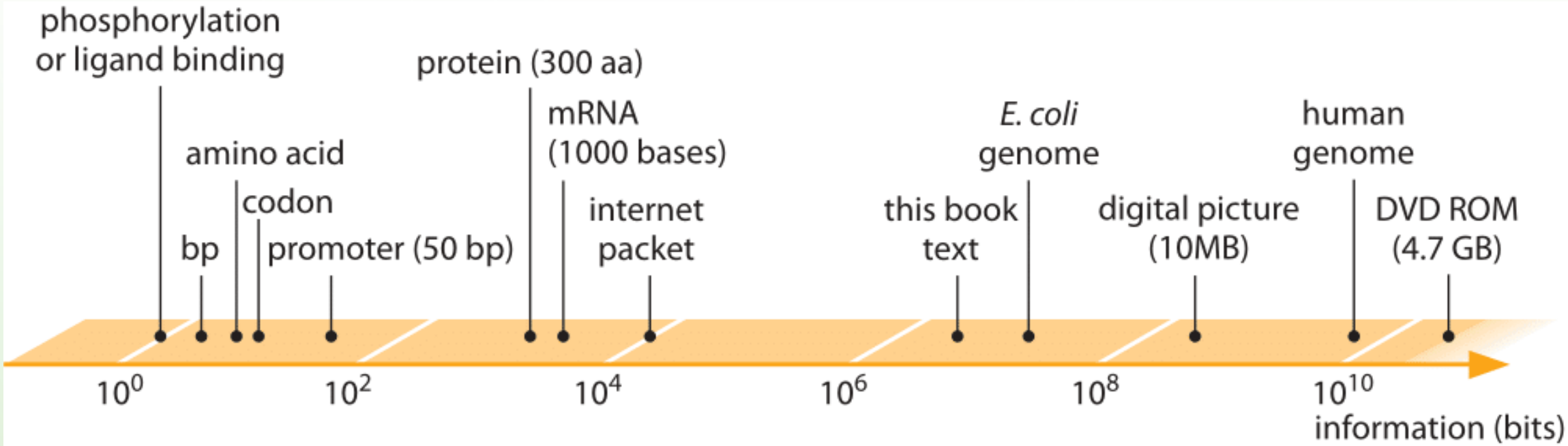
- IN RNA T -> U (URACIL)

https://www.nature.com/scitable/topicpage/discovery-of-dna-structure-and-function-watson-397/

THE DIGITAL CODE OF DNA



INFORMATION CONTENT: BIOLOGICAL ENTITIES VS STORAGE DEVICES

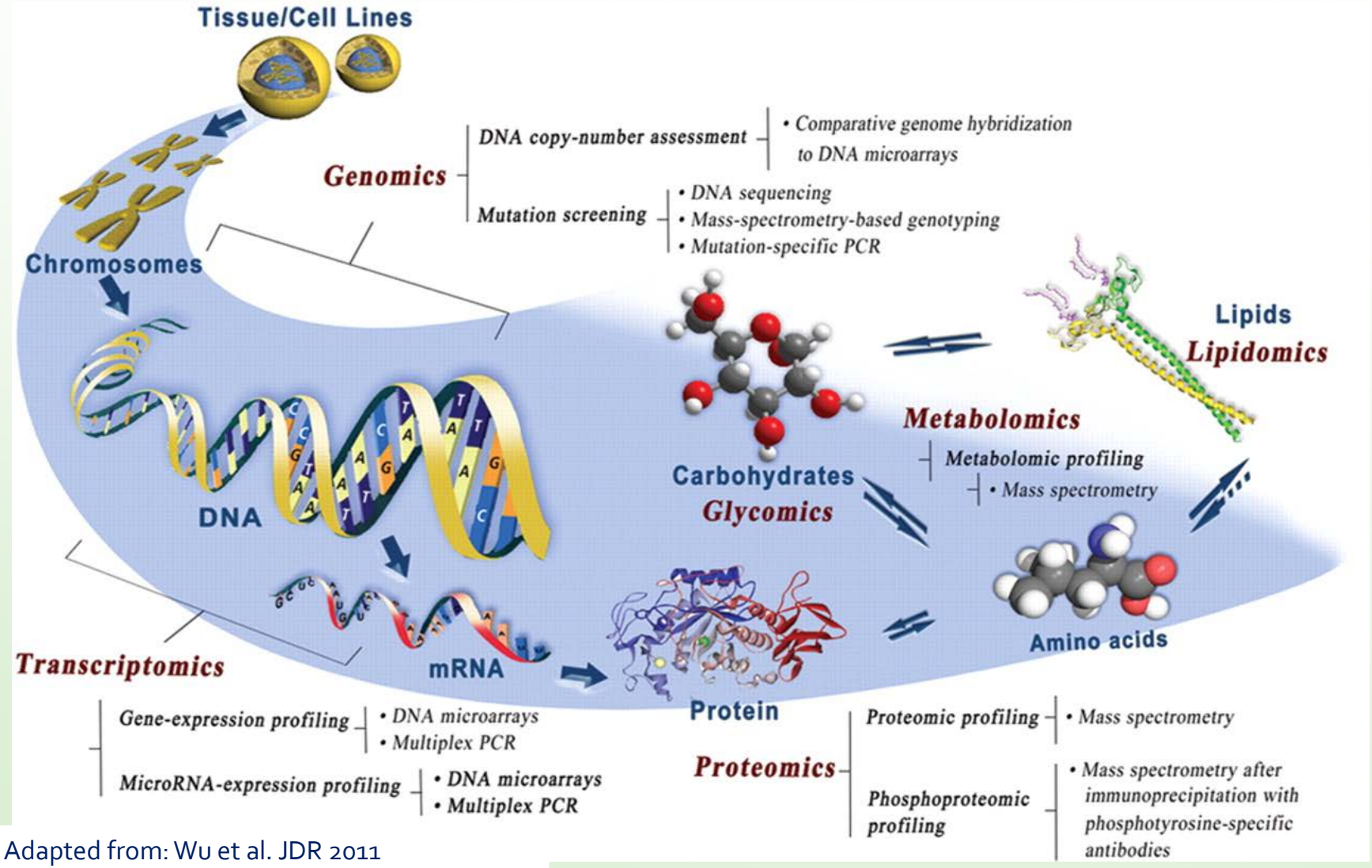


information, $I = \log_2$ (# possible configurations)

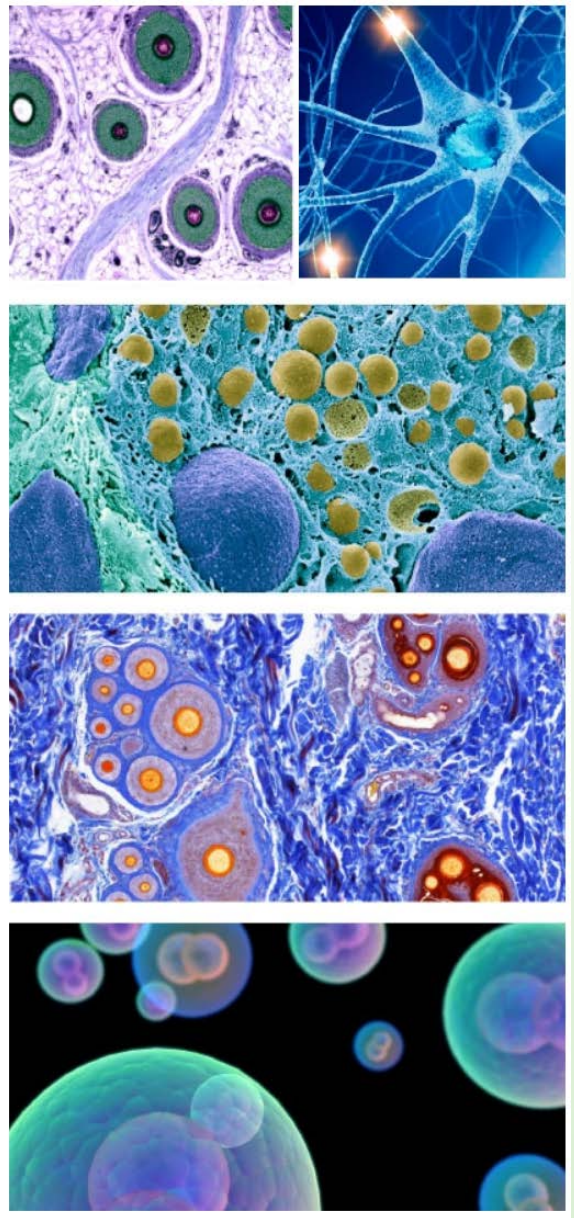
e.g. base pair (bp) has four possibilities $\Rightarrow I = \log_2(4) = 2$

http://book.bionumbers.org/wp-content/uploads/2014/08/500-f1-InformationAxis-1.png

From genotype to phenotype

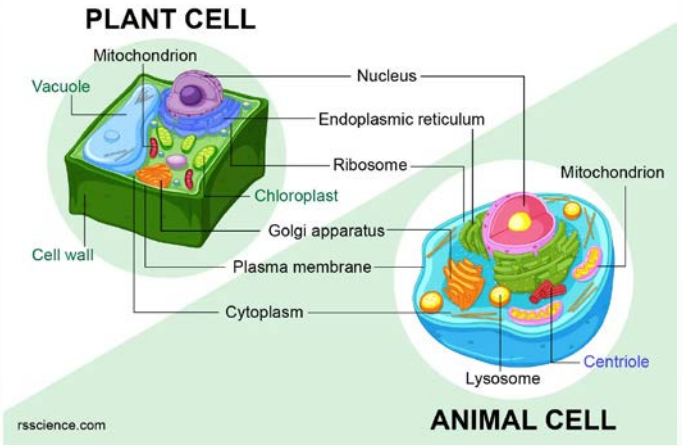


Adapted from: Wu et al. JDR 2011



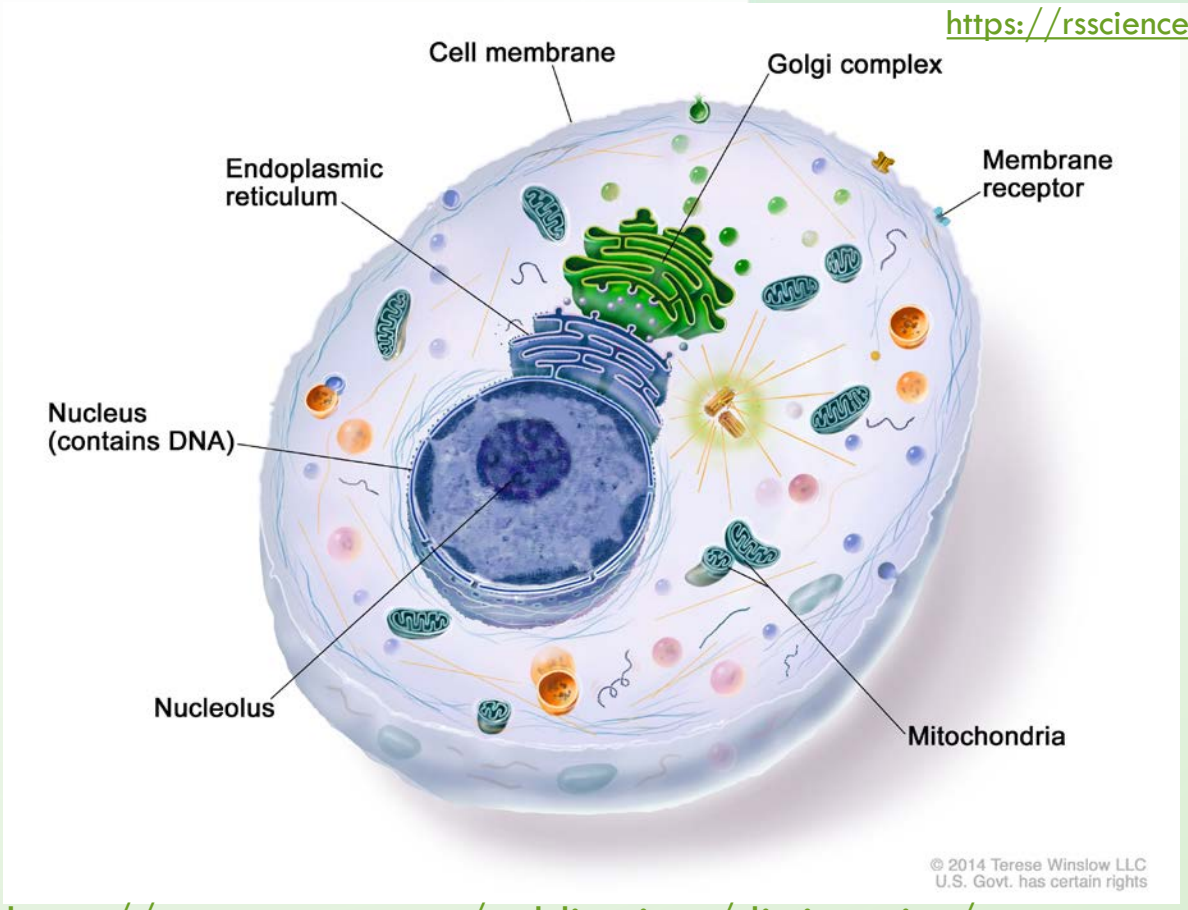
Adapted from https://www.broadinstitute.org

THE CELL



https://rsscience.com/about-me/

https://www.youtube.com/watch?v=2KQbVr9kFO0



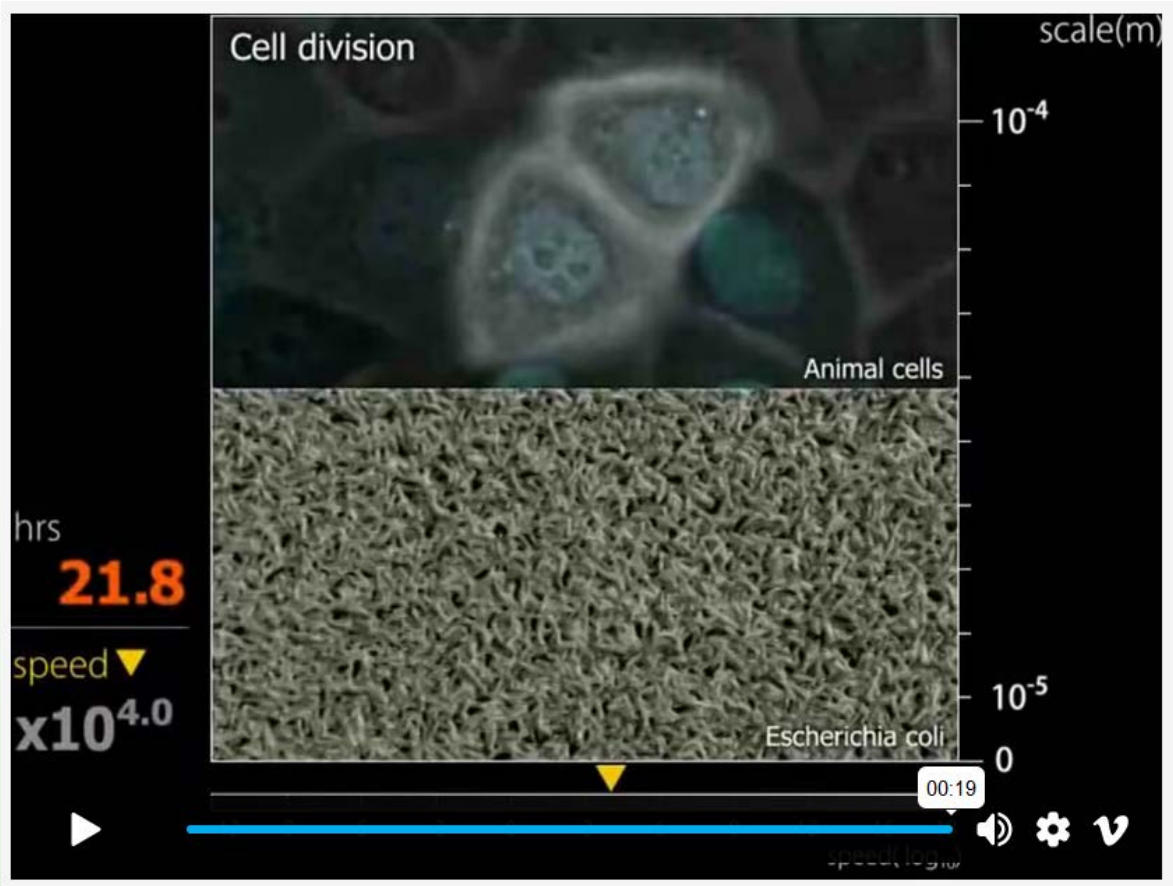
https://www.cancer.gov/publications/dictionaries/cancer-terms/def/cell

GLOSSARY FOR VIDEO

https://www.genome.gov/genetics-glossary

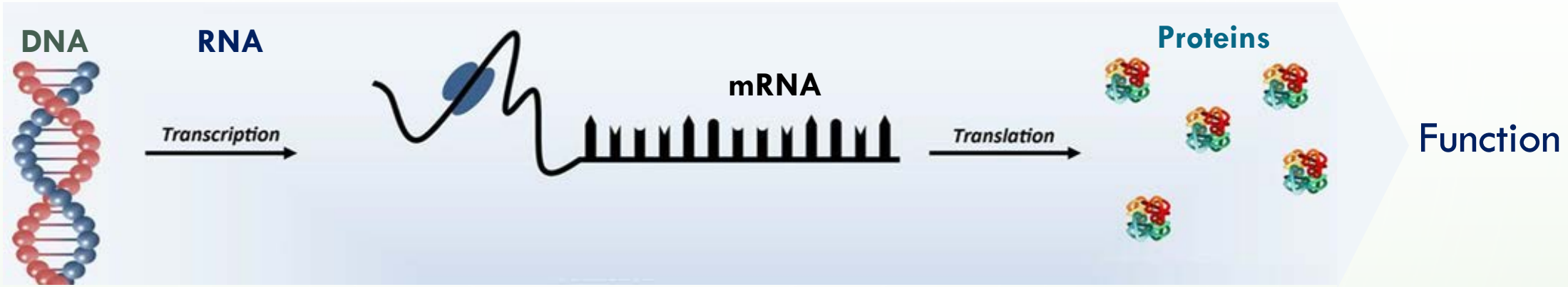
- PLASMA MEMBRANE: MEMBRANE FOUND IN ALL CELLS THAT SEPARATES THE INTERIOR OF THE CELL FROM THE OUTSIDE ENVIRONMENT
- EXTRACELLULAR MATRIX: LARGE NETWORK OF PROTEINS AND OTHER MOLECULES THAT SURROUND, SUPPORT, AND GIVE STRUCTURE TO CELLS AND TISSUES IN THE BODY, COMMUNICATION AND MOVEMENT
- GOLGI: ORGANELLE THAT PROCESSES AND PACKAGES PROTEINS AND MOLECULES, ESPECIALLY THOSE TO BE EXPORTED FROM THE CELL
- LYSOSOME: ORGANELLE CONTAINING ENZYMES CAPABLE OF BREAKING DOWN PROTEINS, NUCLEIC ACIDS, CARBOHYDRATES, LIPIDS
- CENTRIOLES: ORGANELLES LOCATED IN THE CYTOPLASM NEAR THE NUCLEAR ENVELOPE. CENTRIOLES PLAY A ROLE IN ORGANIZING MICROTUBULES THAT SERVE AS THE CELL'S SKELETAL SYSTEM.
- MICROTUBULES: MAJOR COMPONENTS OF THE CYTOSKELETON. INVOLVED IN MITOSIS, CELL MOTILITY, INTRACELLULAR TRANSPORT, AND MAINTENANCE OF CELL SHAPE.

CELL DIVISION

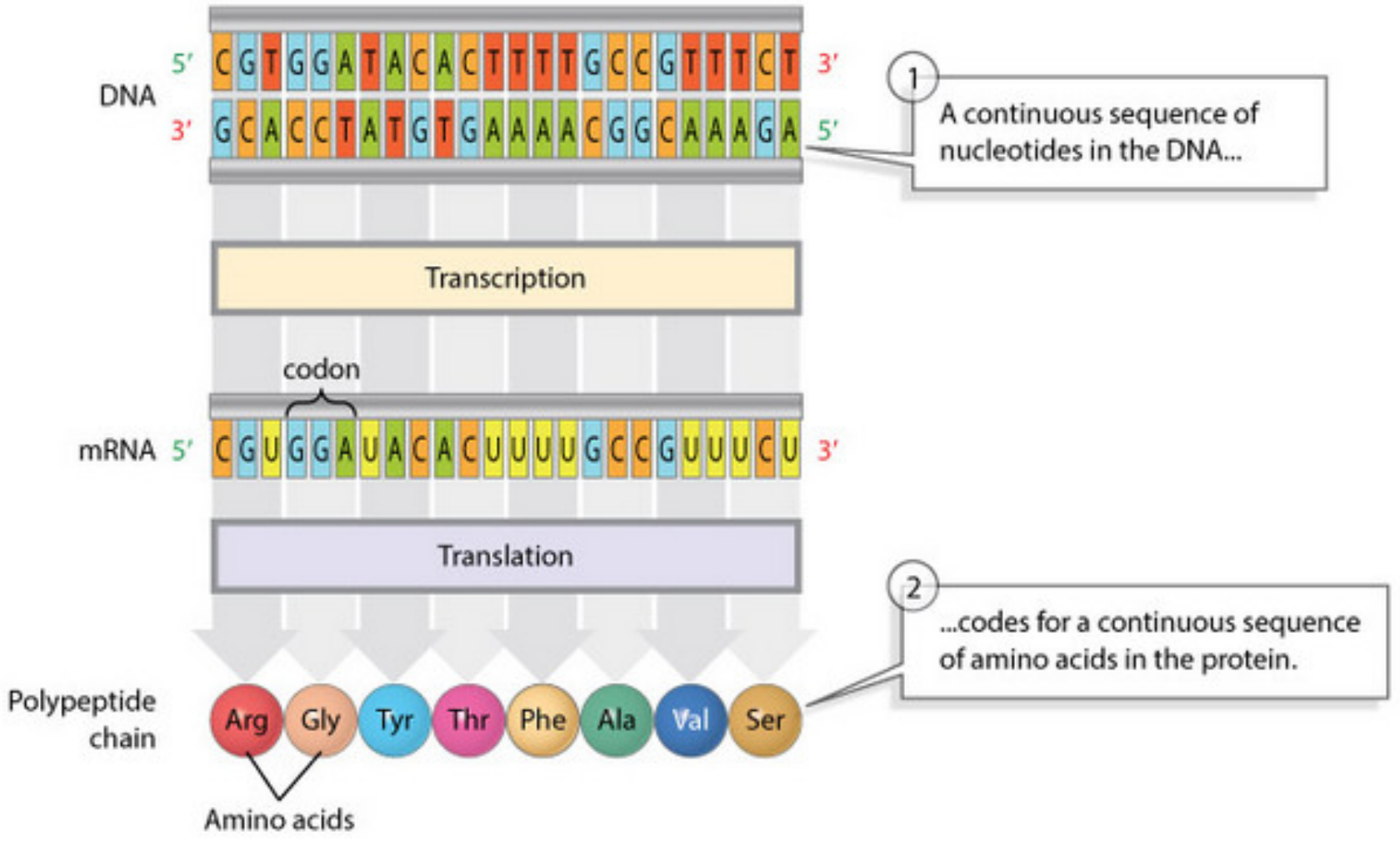


https://vimeo.com/89112986?embedded=true&source=vimeo_logo&owner=9445085

DNA, RNA AND PROTEINS



FROM DNA TO FUNCTION



With colinearity, the number of nucleotides in the gene is proportional to the number of amino acids in the protein.

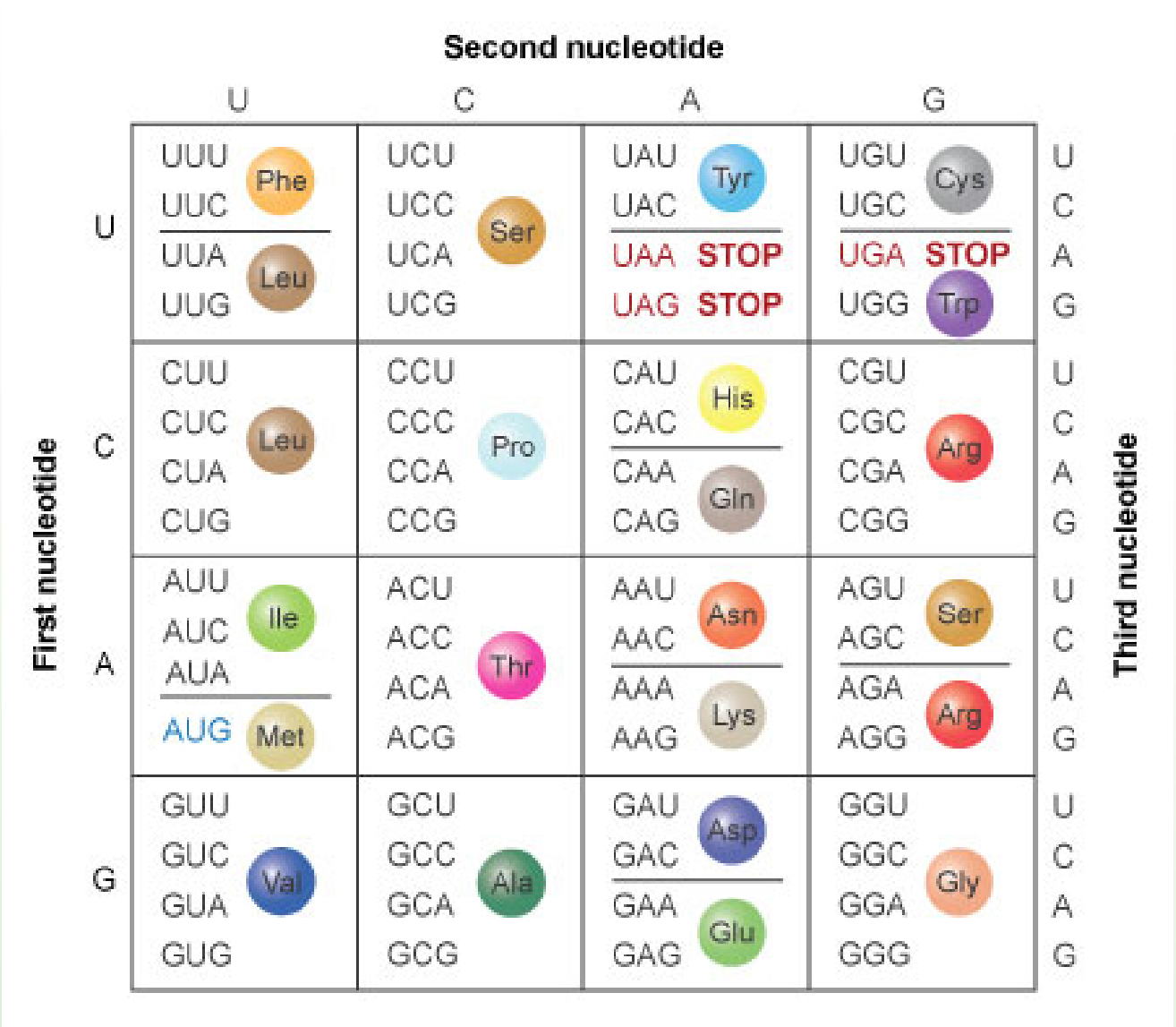
4 bases

3 letters (triplet – a codon) code for one amino acid in a protein

Codons (3 leter strings) specify 20 'words' (amino acid)

https://www.nature.com/scitable/topicpage/what-is-a-gene-colinearity-and-transcription-430/

$4^3 = 64$ possible ways (order important, repetition allowed) to code for 20 amino acids and stop codons



- Alanine
- Arginine
- Asparagine
- Aspartic Acid
- Cysteine
- Glutamic acid
- Glutamine
- Glycine
- Histidine
- Isoleucine
- Leucine
- Lysine
- Methionine
- Phenylalanine
- Proline
- Serine
- Threonine
- Tryptophan
- Tyrosine
- Valine

https://www.nature.com/scitable/topicpage/translation-dna-to-mrna-to-protein-393/

A SINGLE BASE CHANGE CAN CREATE DEVASTATING GENETIC DISORDERS

e.g. Sickle-Cell Anemia

In sickle-cell anemia, the gene for the beta chain of the hemoglobin protein (the oxygen-carrying protein that makes blood red) is mutated.

Beta hemoglobin (beta globin) is a single chain of 147 amino acids.

One single-base mutation causes the sixth amino acid in the chain to be valine, rather than glutamic acid. This changes the red blood cell shape and ability to carry oxygen, with huge consequences for the individual.

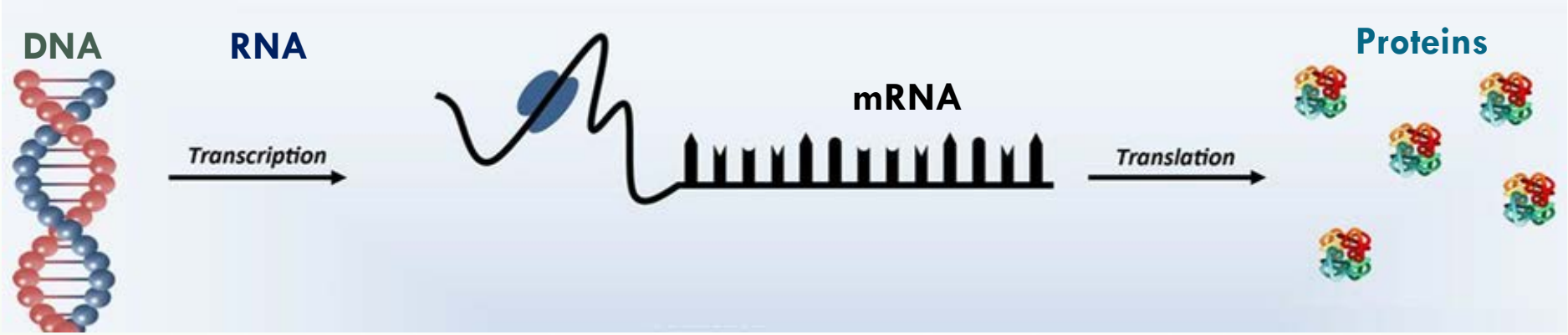
Wild-Type

ATG	GTG	CAC	CTG	ACT	CCT	GAG	GAG	AAG	TCT	GCC	GTT	ACT
Start	Val	His	Leu	Thr	Pro	Glu	Glu	Lys	Ser	Ala	Val	Thr

Mutant

ATG	GTG	CAC	CTG	ACT	CCT	GTG	GAG	AAG	TCT	GCC	GTT	ACT
Start	Val	His	Leu	Thr	Pro	Val	Glu	Lys	Ser	Ala	Val	Thr

DNA, RNA AND PROTEINS

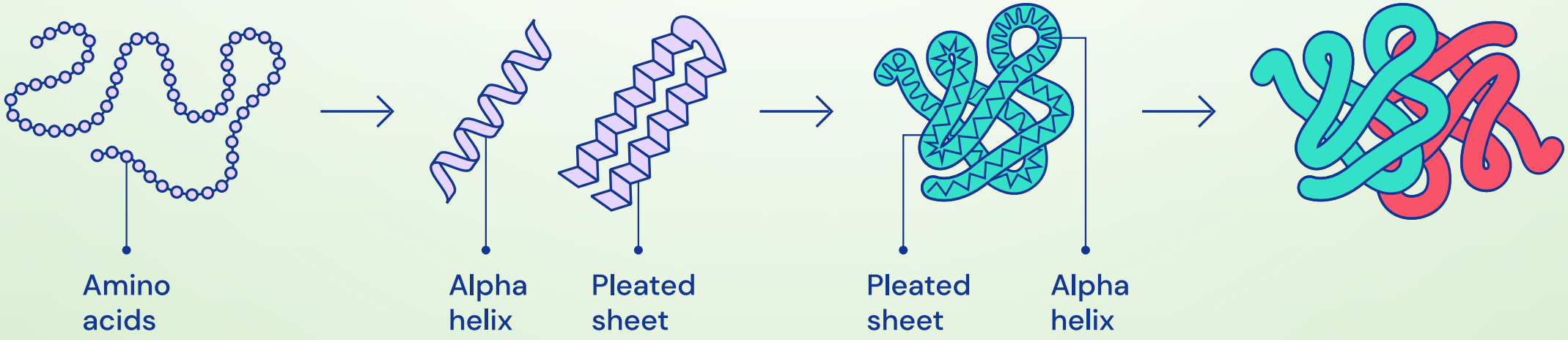


Every protein is made up of a sequence of amino acids bonded together

These amino acids interact locally to form shapes like helices and sheets

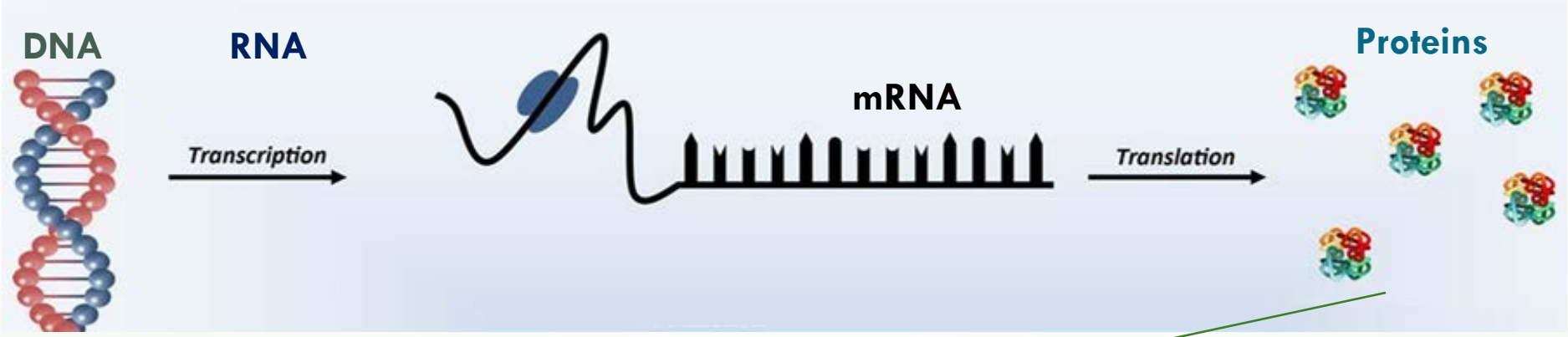
These shapes fold up on larger scales to form the full three-dimensional protein structure

Proteins can interact with other proteins, performing functions such as signalling and transcribing DNA



https://deepmind.com/blog/article/AlphaFold-Using-AI-for-scientific-discovery

DNA, RNA AND PROTEINS



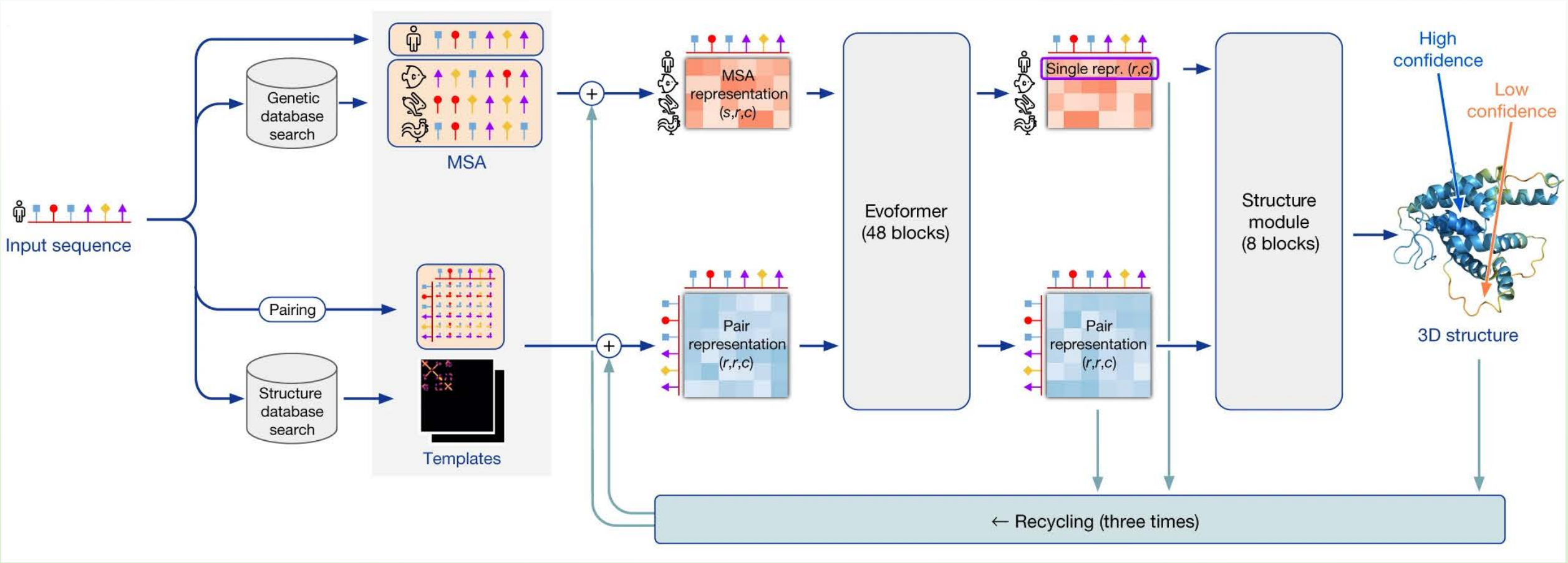
Protein 3D structure

Protein function

https://deepmind.com/blog/article/AlphaFold-Using-AI-for-scientific-discovery

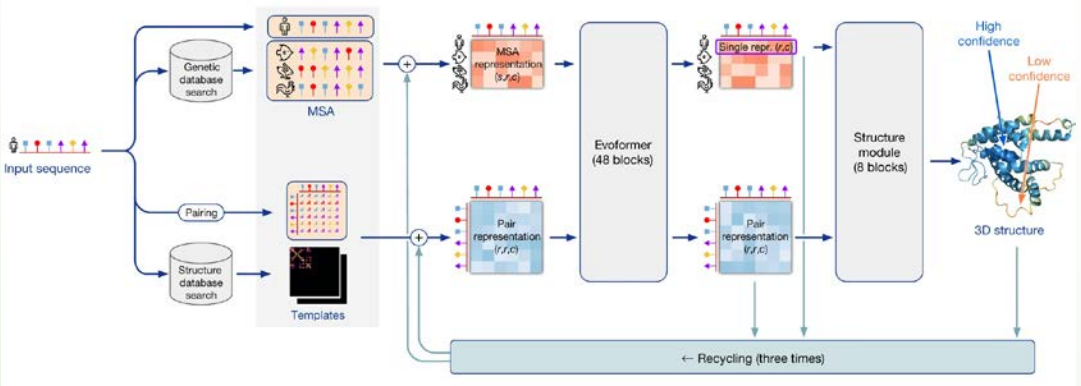
USING AI TO PREDICT PROTEIN STRUCTURE

Can we determine a protein's 3D shape from its amino-acid sequence?



https://www.nature.com/articles/s41586-021-03819-2

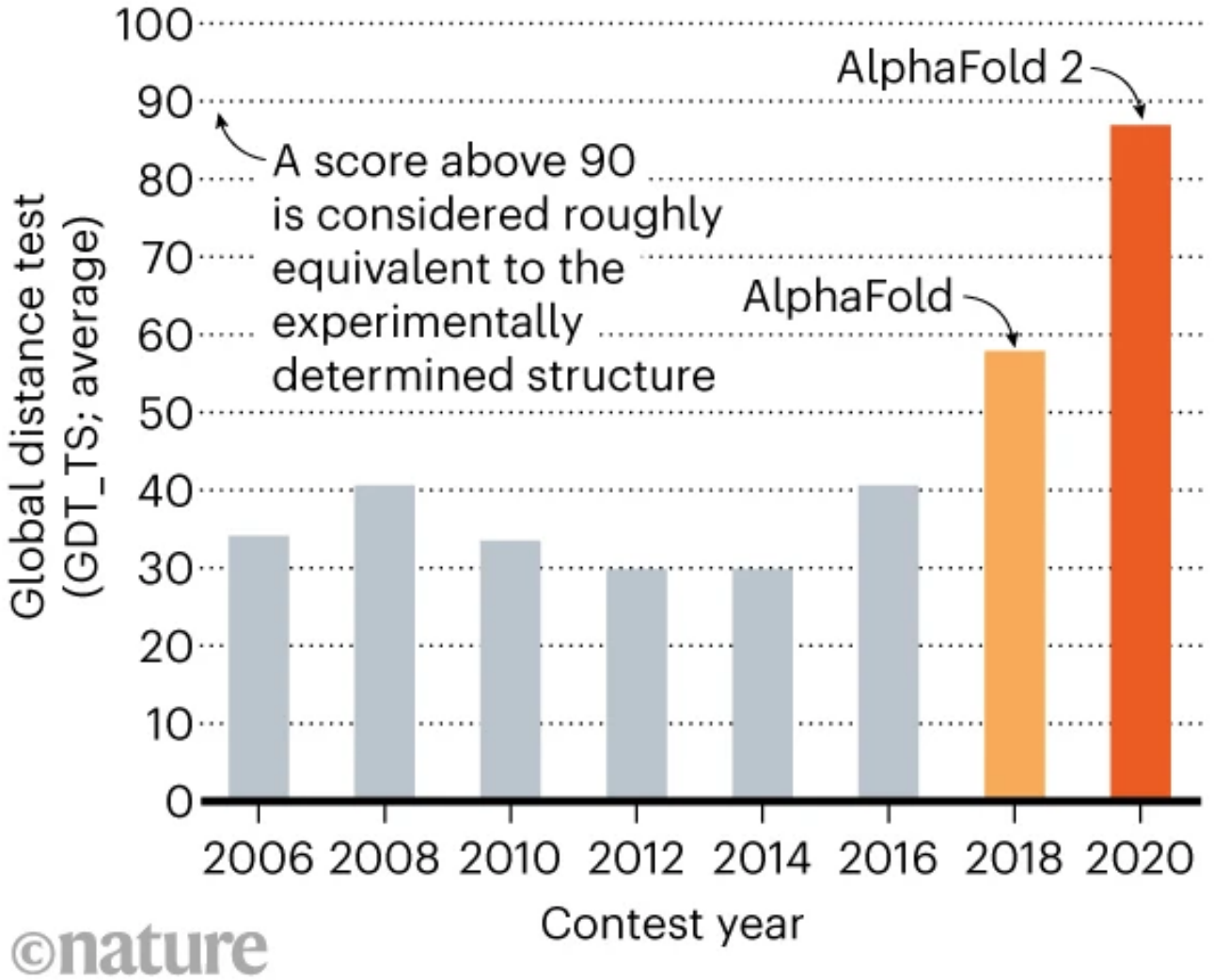
ALPHAFOLD: USING AI TO PREDICT PROTEIN STRUCTURE



AlphaFold
Protein Structure Database

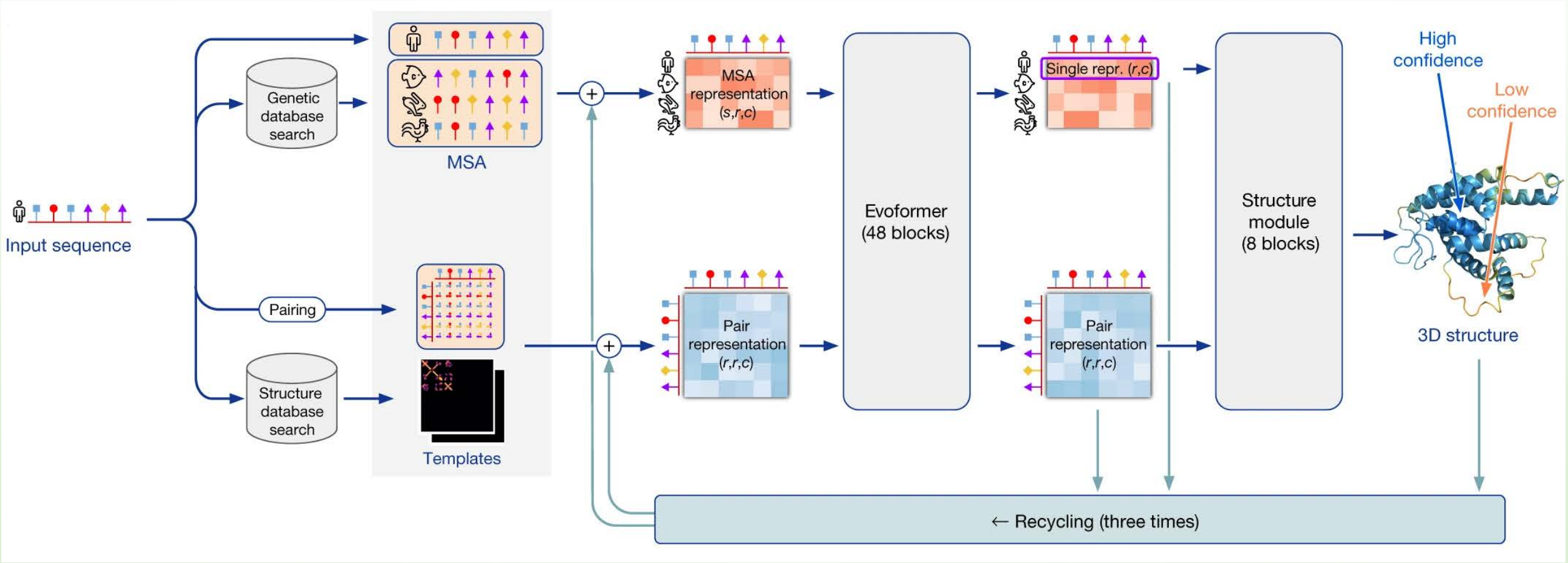
Developed by DeepMind and EMBL-EBI

https://alphafold.ebi.ac.uk/



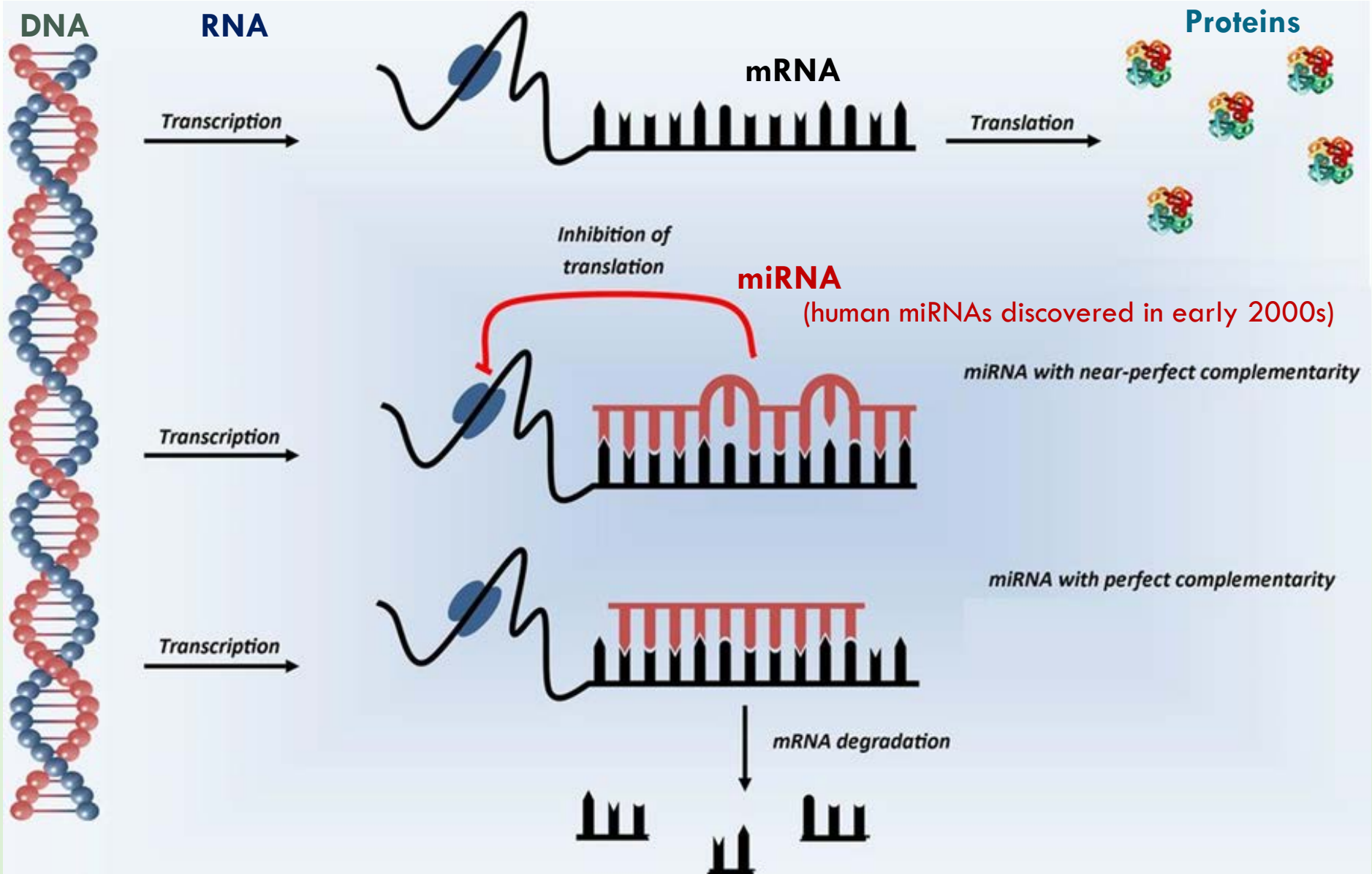
USING AI TO PREDICT PROTEIN STRUCTURE

Can we determine a protein's 3D shape from its amino-acid sequence?



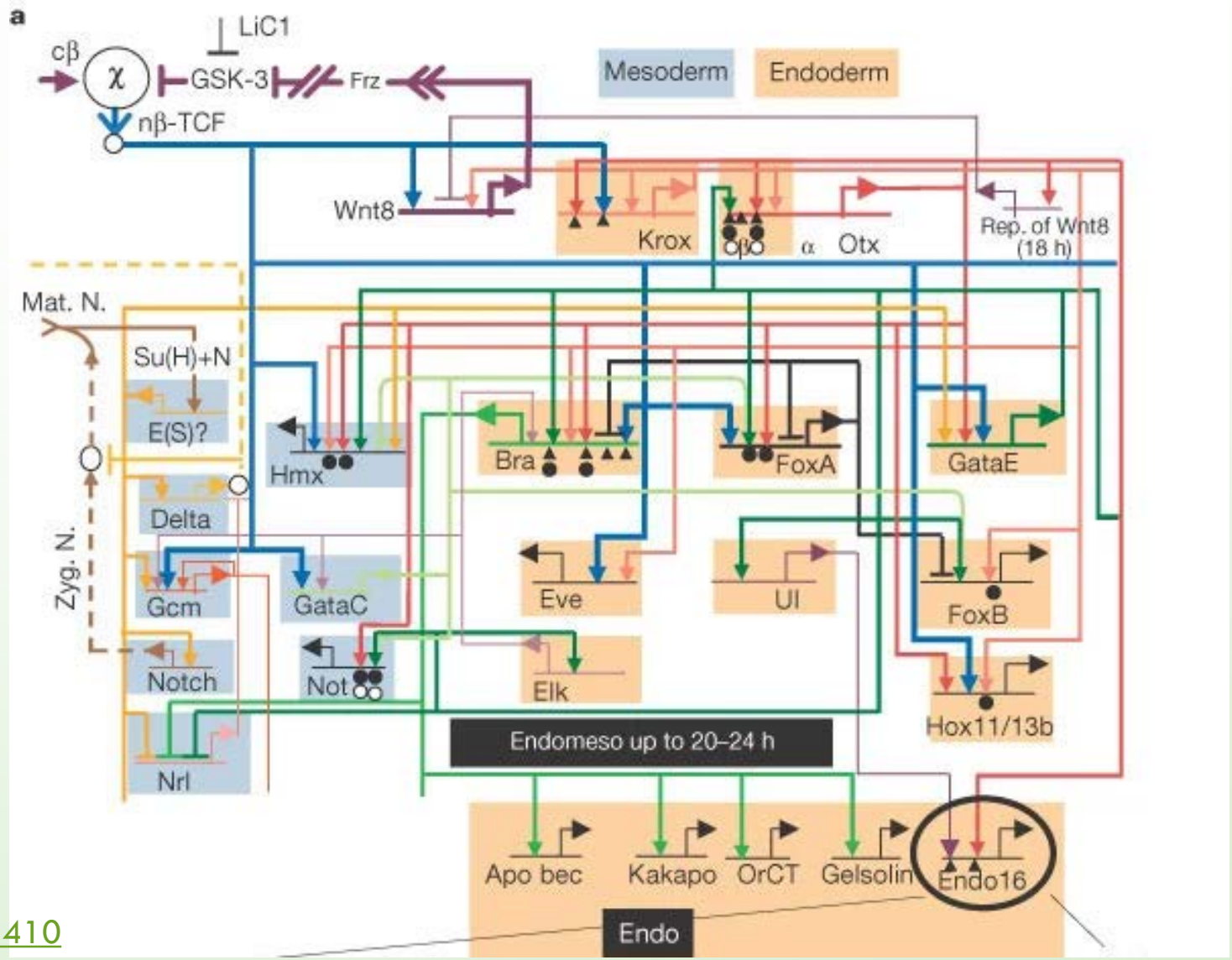
https://www.nature.com/articles/s41586-021-03819-2

DNA, RNA AND PROTEINS



CODING NETWORK OF GENES

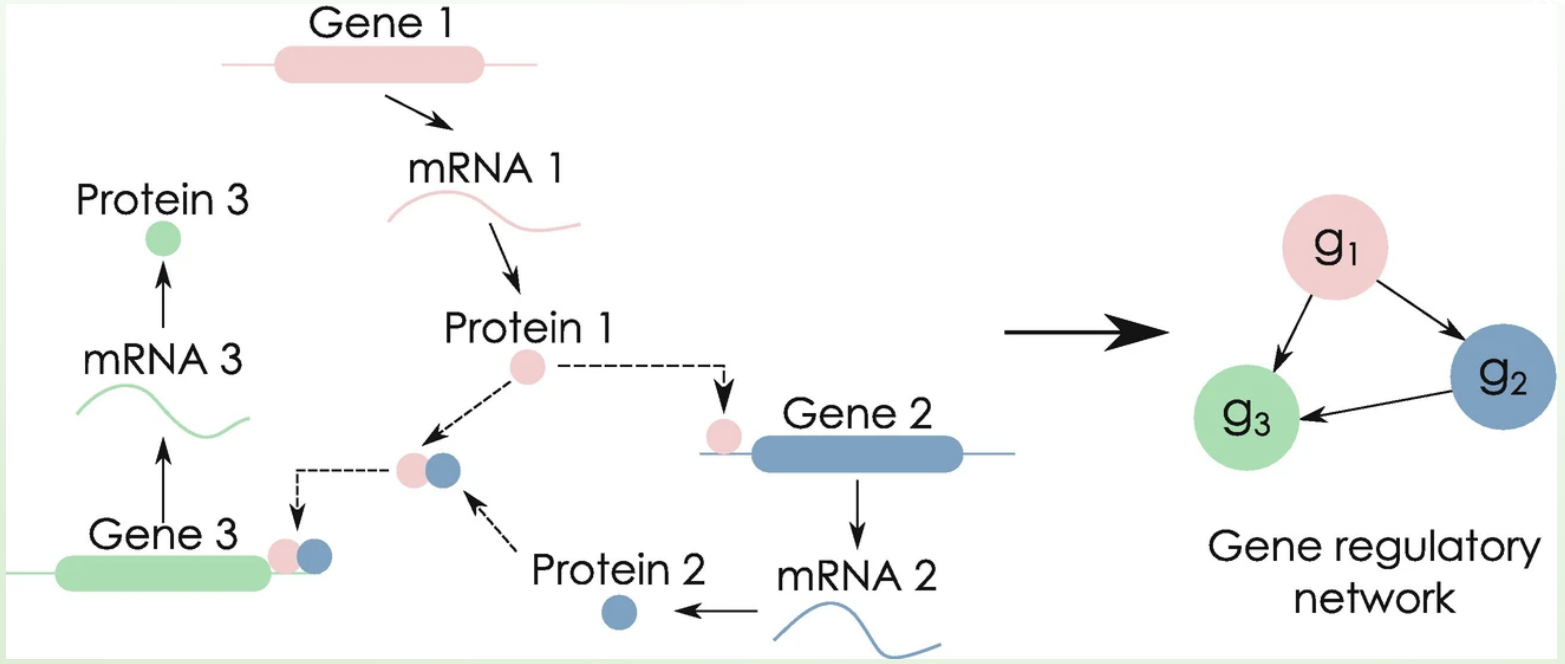
Gene regulatory networks that specify the behaviour of the genes can be represented computationally using methods from STEM

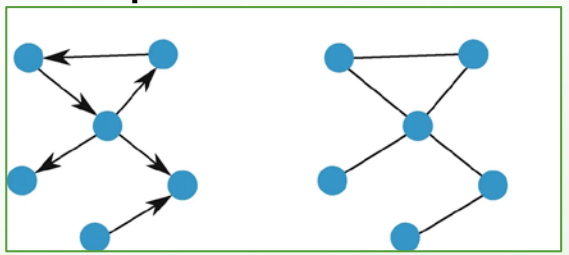


Hood and Galas, https://www.nature.com/articles/nature01410

REGULATION OF GENE EXPRESSION IN EUKARYOTES

At any given time, a complex set of interactions between genes, RNA molecules and proteins determine which genes are activated, and the amount of protein or RNA product.



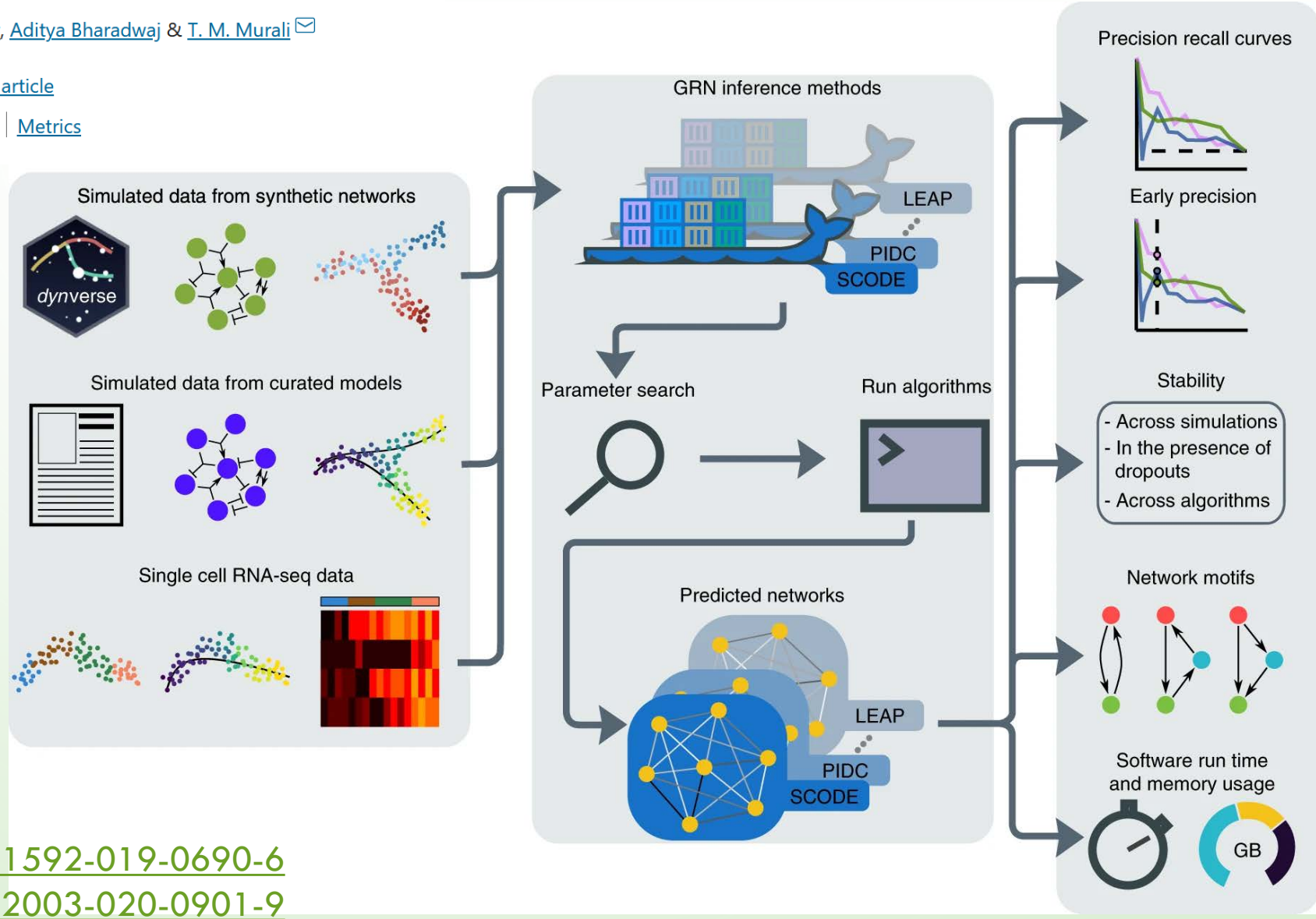


https://link.springer.com/protocol/10.1007/978-1-4939-8882-2_1

Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data

Aditya Pratapa, Amogh P. Jalihal, Jeffrey N. Law, Aditya Bharadwaj & T. M. Murali

Nature Methods **17**, 147–154 (2020) | Cite this article



https://www.nature.com/articles/s41592-019-0690-6
https://www.nature.com/articles/s42003-020-0901-9