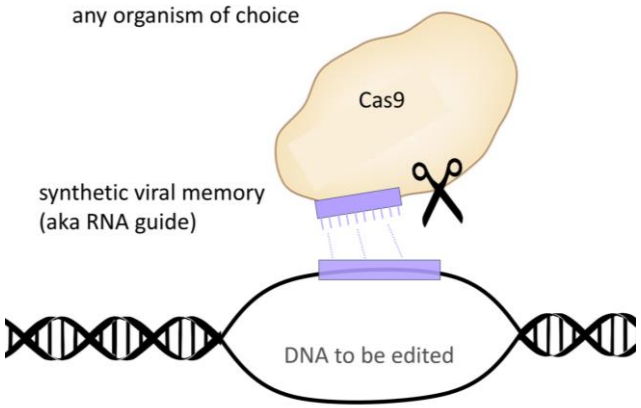
Microsoft

# Machine Learning for Biomedicine at Scale

Jennifer Chayes
Technical Fellow and Managing Director
Microsoft Research New England, New York, and Montreal

# Machine Learning Problems in Biomedicine

## CRISPR gene editing




## Cancer immunotherapy

Memorial Sloan Kettering Cancer Center

Icahn School of Medicine at Mount Sinai

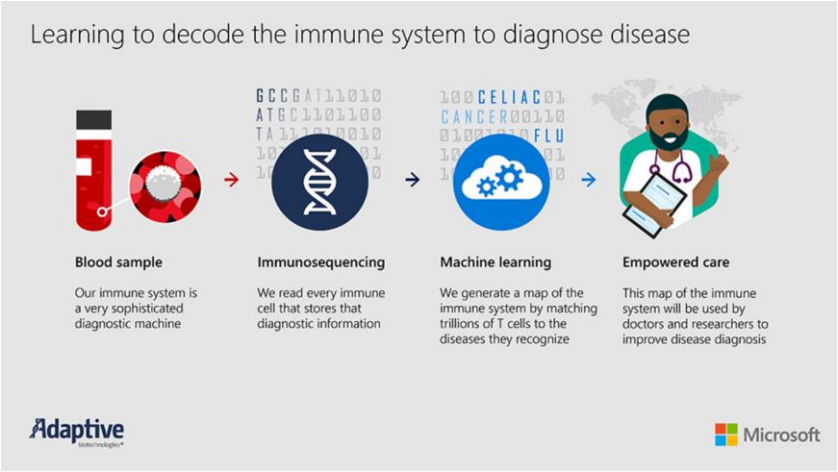
Stanford SCHOOL OF MEDICINE

DANA-FARBER CANCER INSTITUTE

BROAD INSTITUTE

STAND UP TO CANCER

## Decoding the immune system



# CRISPR-ML Team

## Microsoft Research

**Nicolo Fusi**

**Jennifer Listgarten**

**Melih Elibol**
**Luong Hoang**
**Jake Crawford**
**Kevin Gao**

## Broad Institute of MIT and Harvard

**John Doench**

Meagan Sullender
**Mudra Hegde**
Emma W. Vaimberg
Katherine Donovan
Ian Smith
David Root

## Dana-Farber Cancer Institute / Broad

Craig Wilen
Robert Orchard
Herbert W. Virgin

## Washington University School of Medicine

Zuzana Tothova

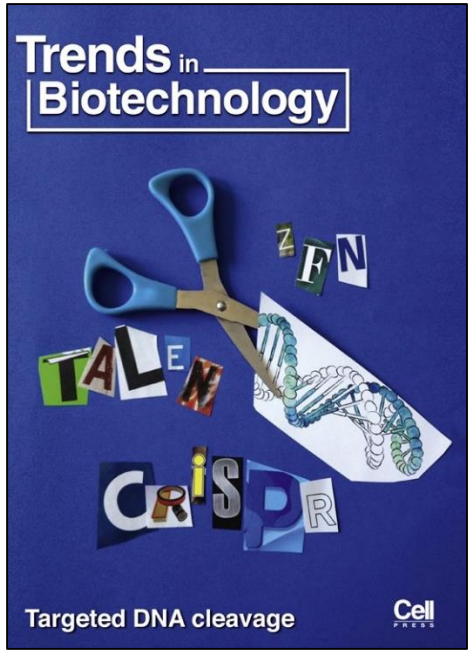
## Harvard / Massachusetts General Hospital

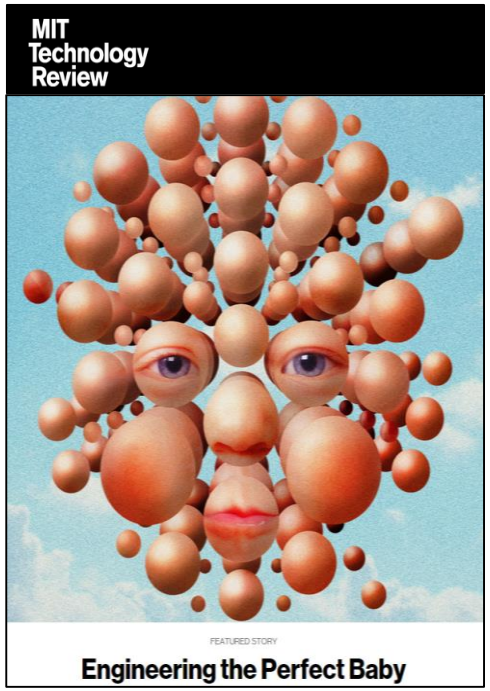
Benjamin Kleinstiver
Keith Young
Alex Sousa

## University of California Los Angeles

**Michael Weinstein**

And thanks to: Carl Kadie (Microsoft Research), Maximilian Haeussler (UCSC)



SCIENCE

The New York Times

## *In Breakthrough, Scientists Edit a Dangerous Mutation From Genes in Human Embryos*

By PAM BELLUCK AUG. 2, 2017

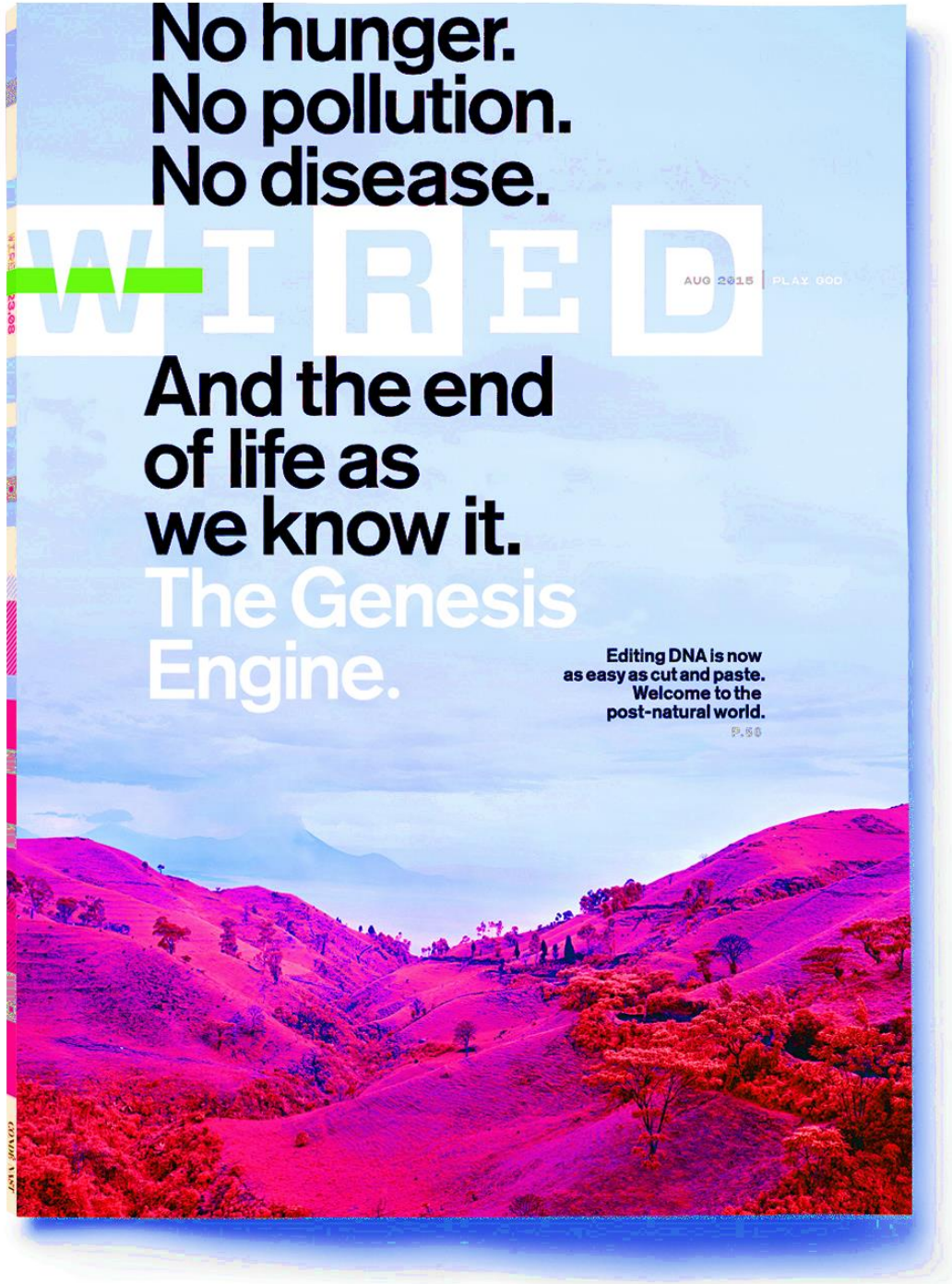
## Could the DNA-editing CRISPR revolutionize medicine?

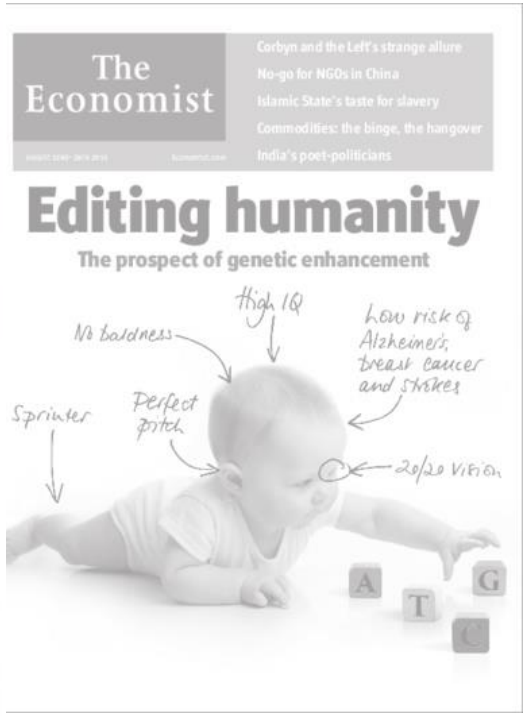
CNN

By Carina Storrs, Special to CNN

Updated 12:22 PM ET, Wed August 12, 2015






HEALTH

*A Powerful New Way*

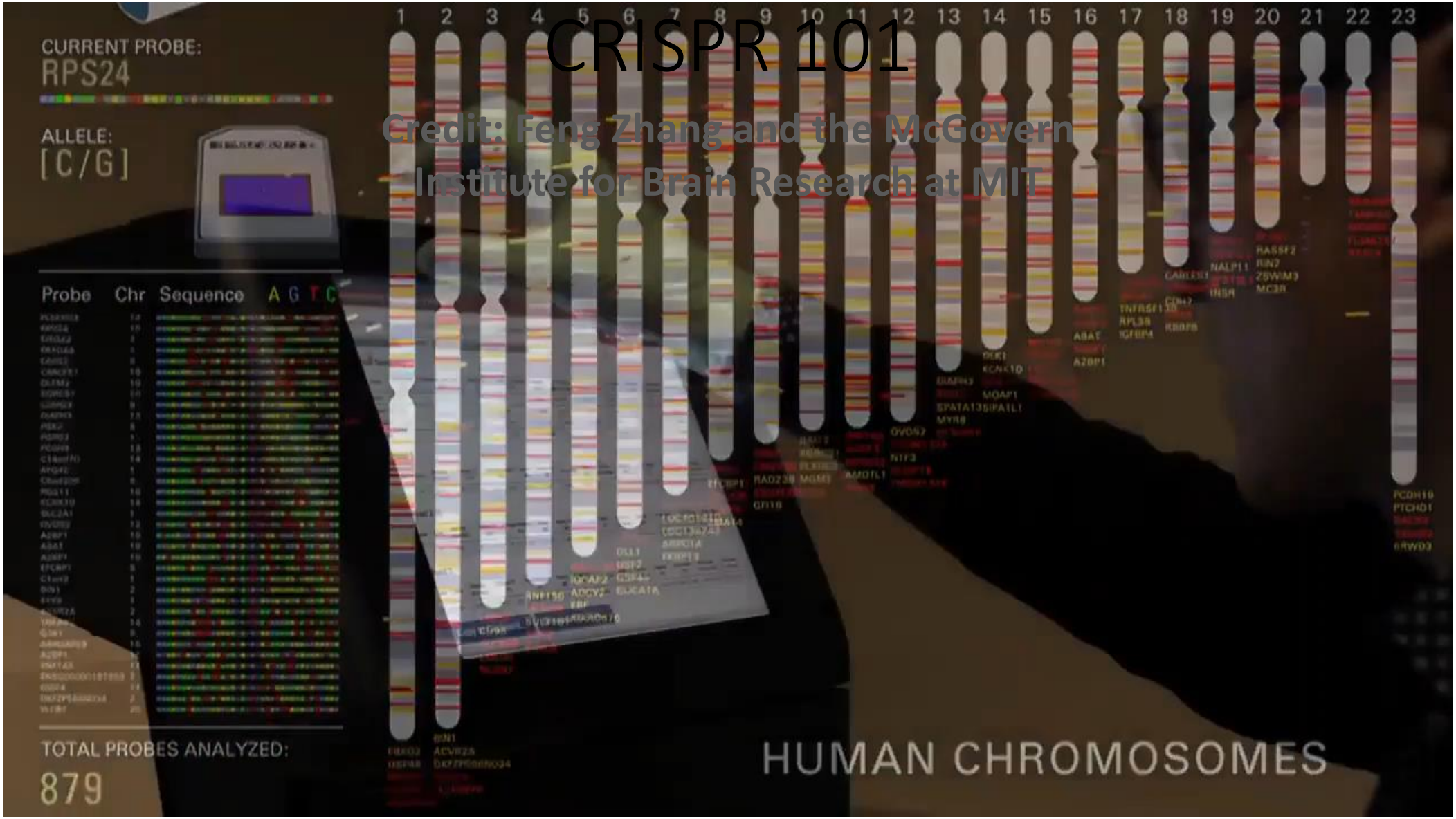
By ANDREW POLLACK MARCH 3, 2014

ting CRISPR
cine?

CNN

# CRISPR 101

Credit: Feng Zhang and the McGovern Institute for Brain Research at MIT

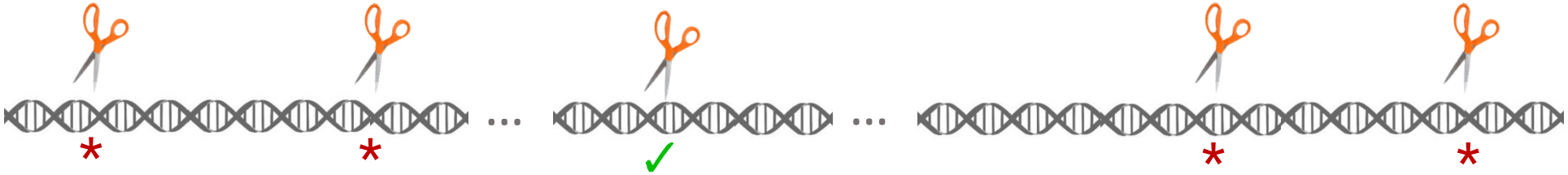


# Matching the Target

<u>Two problems:</u>

1. Better "**on-target**" (knocking out the gene of interest):


2. Elimination/reduction of "**off-target**" effects:

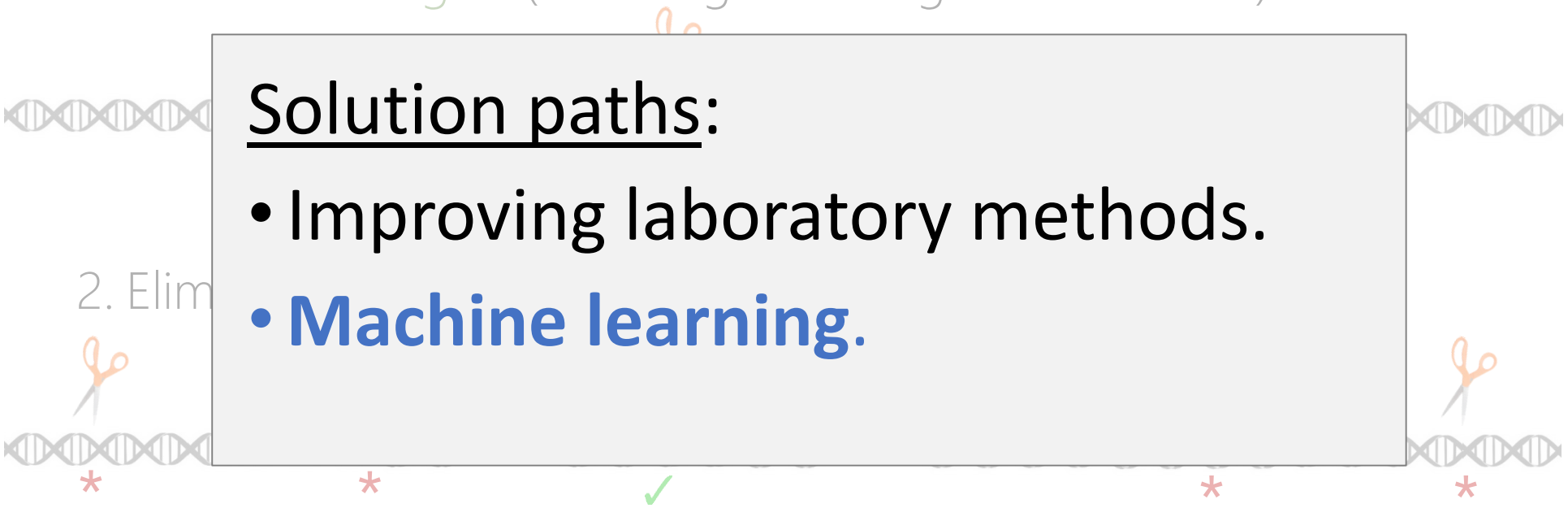


# Matching the Target

Two **problems:**

1. Better "on-target" (knocking out the gene of interest):

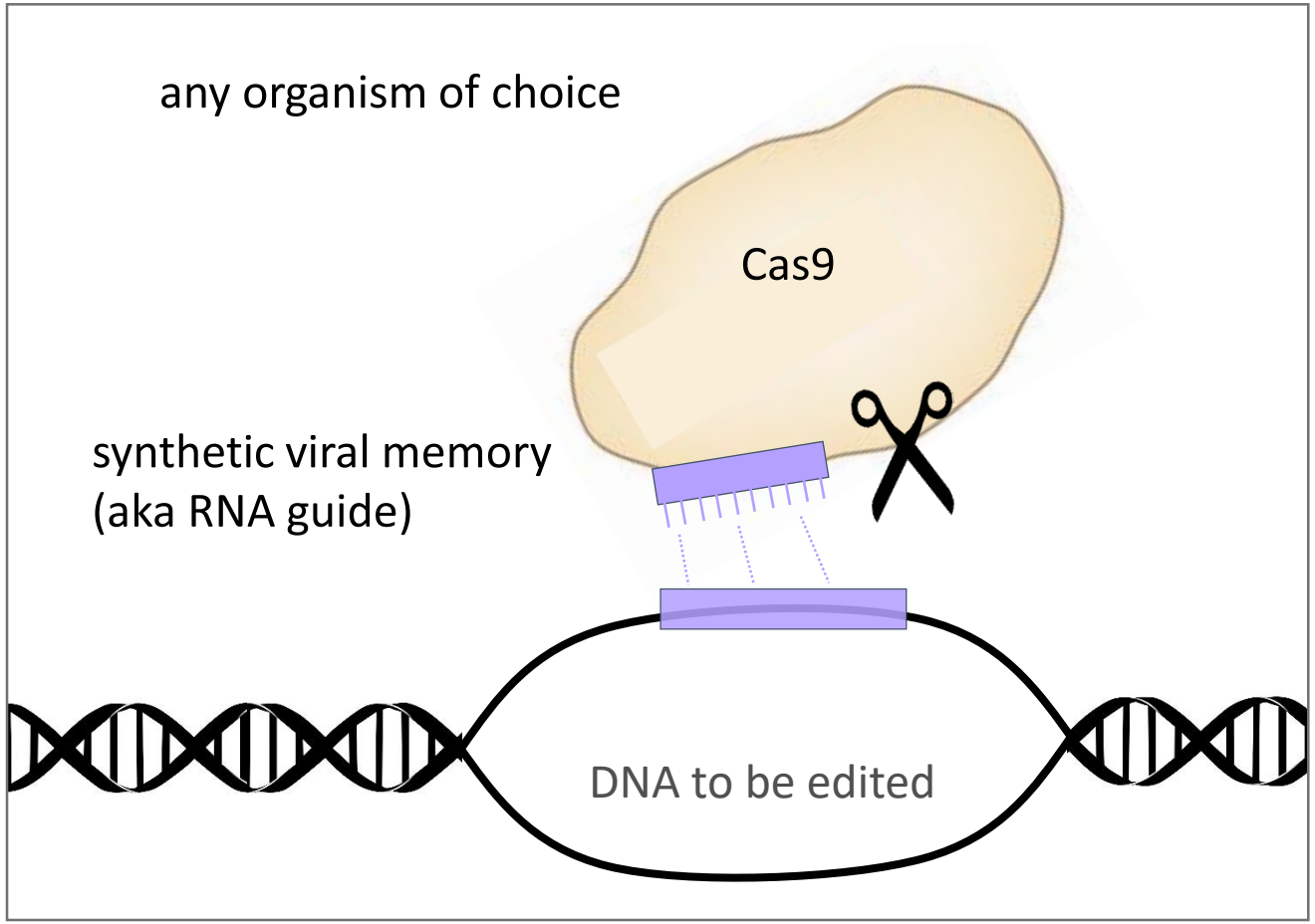
2. Elim



**Solution paths:**

- Improving laboratory methods.
- **Machine learning.**

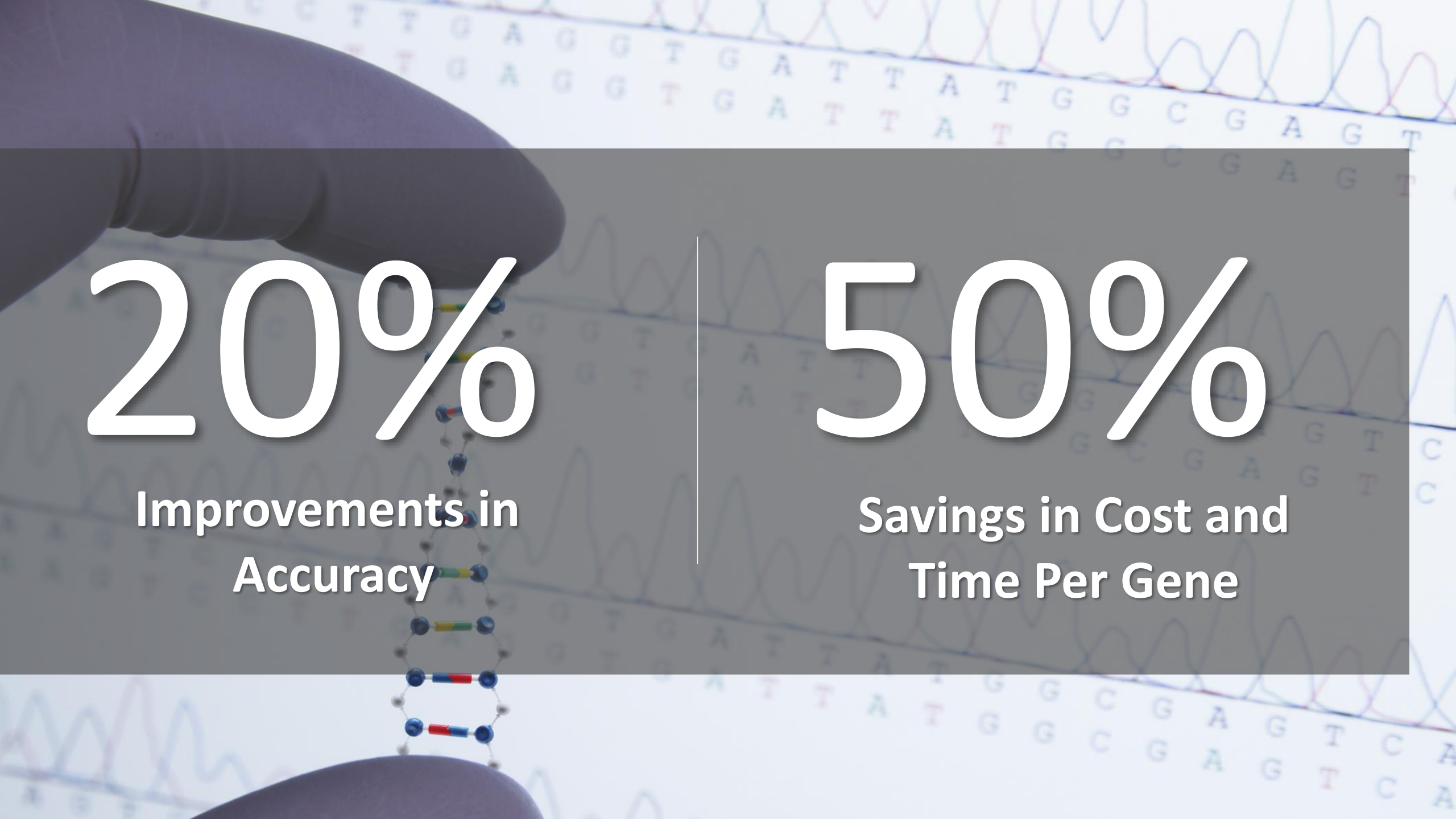
# ML on-target predictive modeling for CRISPR




| RNA guide | | gene edit success |
|---|---|---|
| $[x_1, x_2, \ldots, x_M]$ | $\xrightarrow{f(\vec{x})}$ | $y =$ **not effective** |
| $[x_1, x_2, \ldots, x_M]$ | $\xrightarrow{f(\vec{x})}$ | $y =$ **effective** |
| $[x_1, x_2, \ldots, x_M]$ | $\xrightarrow{f(\vec{x})}$ | $y =$ **very effective** |
| $[x_1, x_2, \ldots, x_M]$ | $\xrightarrow{f(\vec{x})}$ | $y =$ **not effective** |

**Problem scale: ~40M possible guides**

20,000 human genes x 10 potential transcripts (RNA) per gene x 200 potential edit locations per transcript



# 20%

**Improvements in Accuracy**

# 50%

**Savings in Cost and Time Per Gene**

# Strong uptake in biology community

- Recommended by independent studies (Haeussler et al. 2016).
- Adopted by startups and academics/researchers worldwide.
- Azure ML service **~1500 requests/day**, doubling every 3 months
- Web service **~300 requests/day**.
- Over **3000 open-source software downloads**.

*Nature Biotechnology* 2016 (320 citations to date)

https://www.microsoft.com/en-us/research/project/crispr/

(Azimuth)

# Off-target prediction: Much harder problem

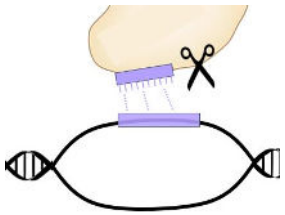
- Need to scan across all 3 billion nucleotides of the genome
- Sparse training data: only measure genes with observable effect
- Combinatorial explosion arising from tolerance of mismatches
- Need to combine into one off-target score for the user

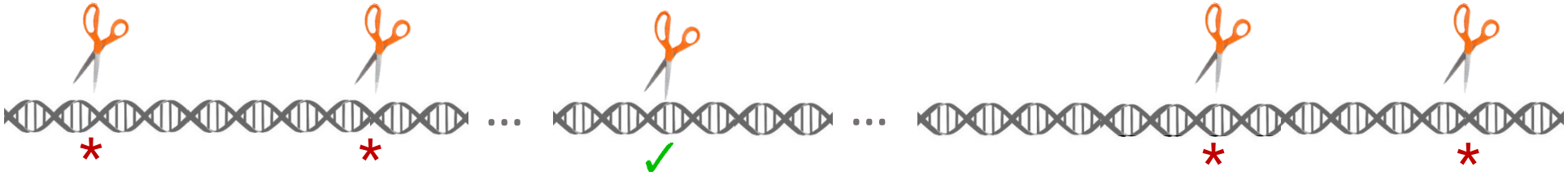
*guide/intended target*

GGCTGCTTTACCCCCTGTGGG

...CTATAACTGGCAGCTCTACCCGGTGTGGGACAAG...

*potential off-target with 2 mismatches*





# Combinatorial explosion (for 1 guide in 1 gene)

**1** mismatch: **69** sites

**2** mismatches: **2277** sites

**3** mismatches: **47,817** sites

**4** mismatches: **717,255** sites

**5** mismatches: **8,176,707** sites

1 full example

**very** sparsely sampled across different genes

Combinatorial explosion + Sparse training data => Cleverness needed!

# Three step-approach to off-target modeling

**Given a guide:**

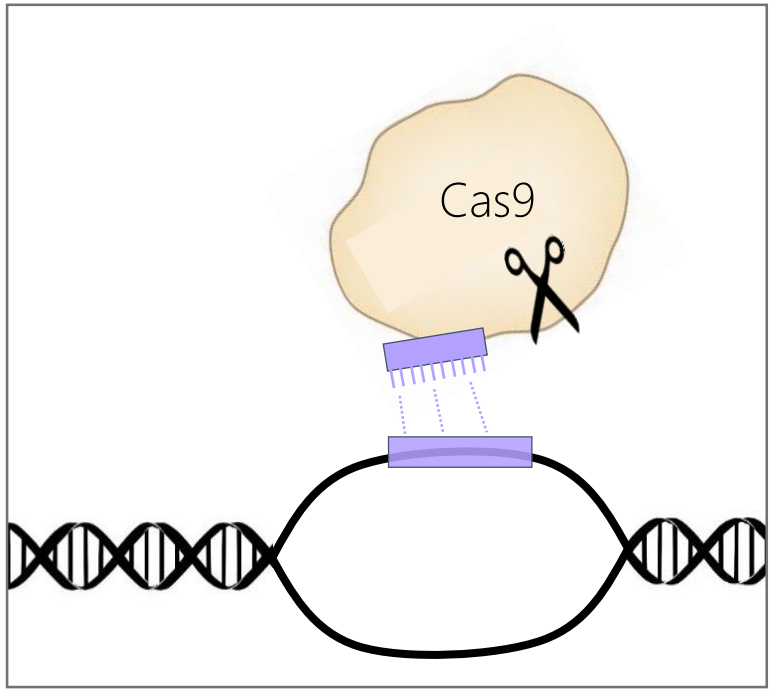
1. Filter genome-wide potential off-targets to assess (all is too many; reduce to ~2,000 most serious)
2. **Score each off-target from (1) for activity using a machine learning predictive model.**
3. Aggregate the scores from (2) into a single overall off-target score.

*Nature Biomedical Engineering* 2018

https://www.microsoft.com/en-us/research/project/crispr/ (Elevation)

(or go to Microsoft AI blog and search for "CRISPR")

# 2. **Score** each off-target for activity

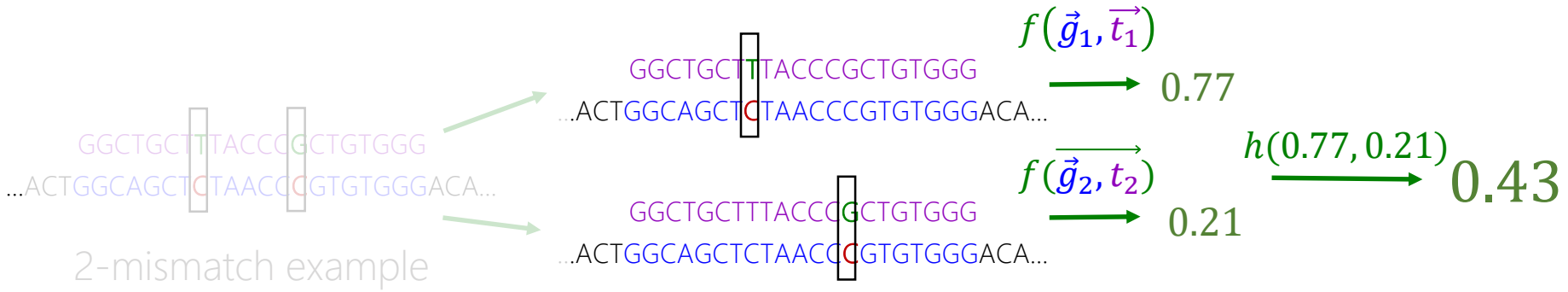



| RNA guide | Target (Mismatch) | machine learning model | Activity |
|---|---|---|---|
| $[g_1, g_2, \ldots, g_M]$ | $[t_1, t_2, \ldots, t_M]$ | $f(\vec{g}, \vec{t})$ | $y = \text{high}$ |
| $[g_1, g_2, \ldots, g_M]$ | $[t_1, t_2, \ldots, t_M]$ | $f(\vec{g}, \vec{t})$ | $y = \text{low}$ |
| $[g_1, g_2, \ldots, g]$ | $[t_1, t_2, \ldots, t_M]$ | $f(\vec{g}, \vec{t})$ | $y = \text{medium}$ |
| | | $\vdots$ | |
| $[g_1, g_2, \ldots, g_M]$ | $[t_1, t_2, \ldots, t_M]$ | $f(\vec{g}, \vec{t})$ | $y = \text{high}$ |

# Off-target: two-step guide-target modelling

i. Build single-mismatch model. $f(\vec{g}, \vec{t})$

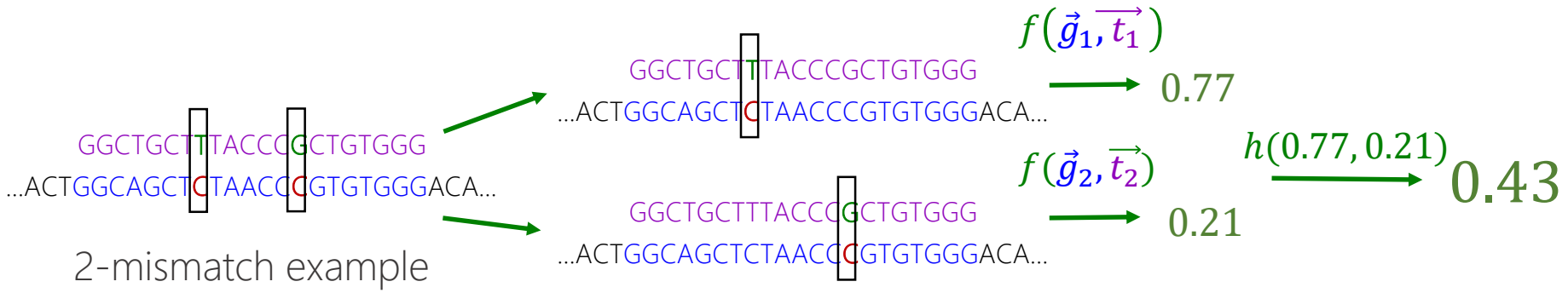
ii. Build multi-mismatch model that combines the output from step (i). $g(f(\overrightarrow{t_1}, \vec{g}_1), f(\overrightarrow{t_2}, \vec{g}_2))$



# Off-target: two-step guide-target modelling

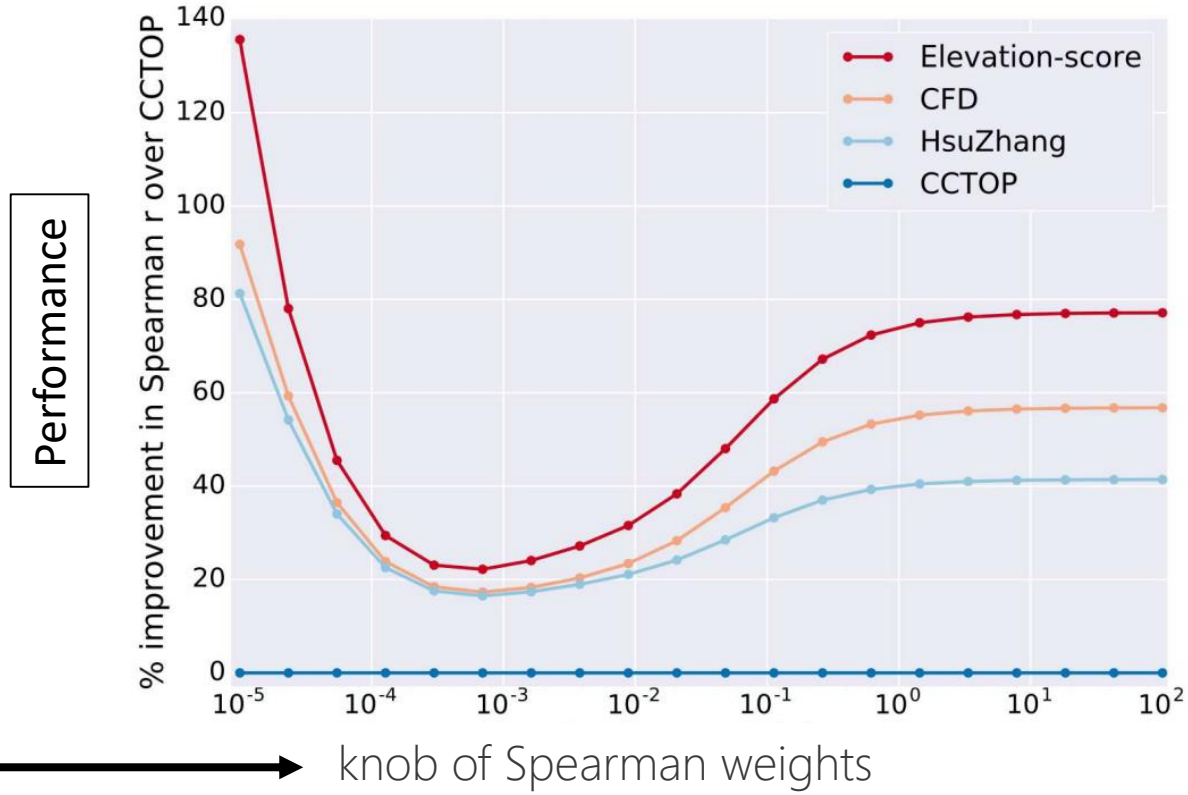
i. Build single-mismatch model. $f(\vec{g}, \vec{t})$

ii. Build multi-mismatch model that combines the output from step (i). $h(f(\vec{g}_1, t_1), f(\vec{g}_2, \vec{t_2},))$



# Evaluation on held out data

- Evaluation depends on usage scenario
- May want to weight off-targets differently depending on their impact
- Our approach (Elevation) outperforms others, no matter what the weighting

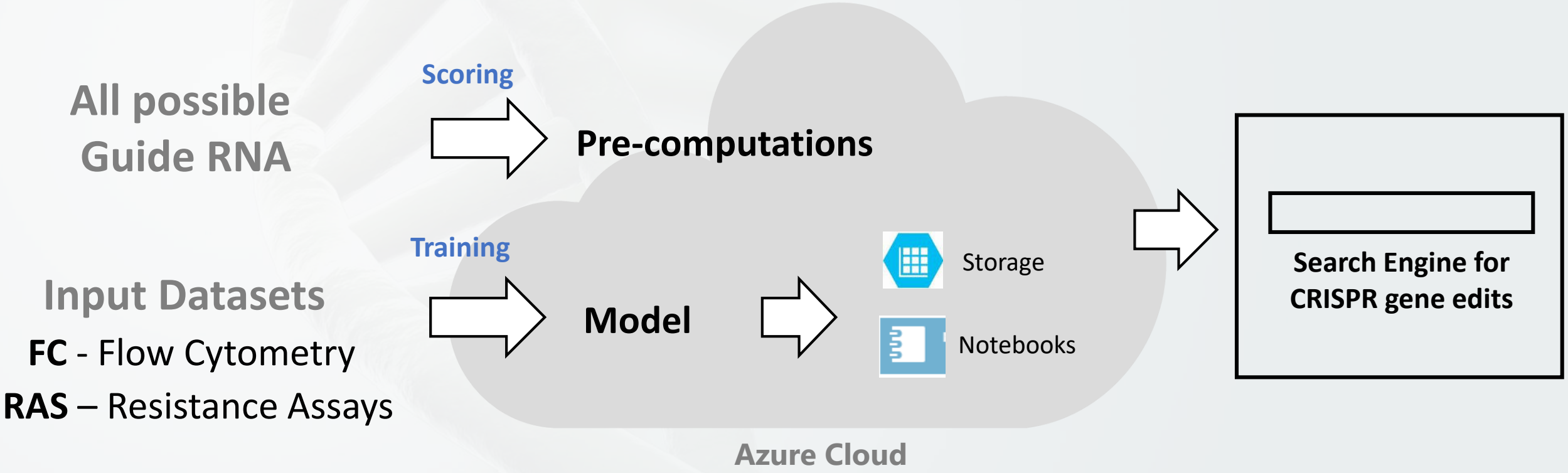



*Has now also been validated as state-of-the-art on two data sets.*

# Problem scale: ~80B possible guides

~2000 mismatches x

(on-target) 20,000 genes x 10 potential transcripts (RNA) per gene x 200 potential edit locations per transcript




**Pre-computation** ran in **18 days** on **16k cores** (~7M CPU-hours)

Results stored in **Azure Tables** with a front-end web interface

**Search** uses hashing to **return** ranked **results instantly**

# CRISPR.ML – Putting it all together

Machine learning-based end-to-end CRISPR/Cas9 guide design

Please cite papers according to these instructions

( On-Target + Off-Target ) ( On-Target Only )

Input Gene / Transcript ID | Input Sequence

Enter value(s) to search, e.g. ENSG00000018510 or ENST00000420982 separated by new line

Model
In Vitro
In Vivo

Log In To Search | Demo

Contact Us | Privacy & Cookies | Terms of Use | Trademarks | © 2018 Microsoft

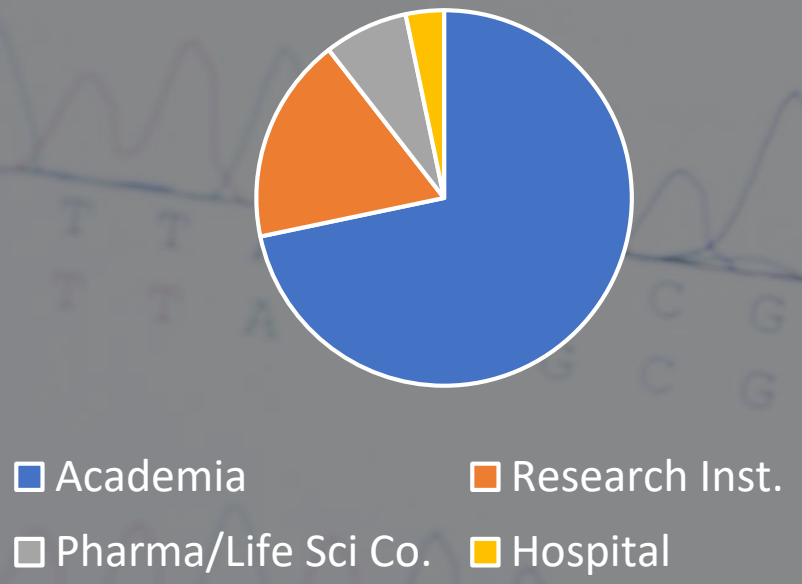
| off-target [?] | on-target [?] |
|---|---|
| 0.38095 | 0.70755 |
| 0.38903 | 0.60016 |
| 0.43129 | 0.34812 |
| 0.44431 | 0.55046 |
| 0.45890 | 0.53405 |
| 0.47083 | 0.38042 |
| 0.47116 | 0.63856 |
| 0.47273 | 0.53924 |

Source code for everything on GitHub

## Usage per Day

15.6K guides queried
9.6M off-targets returned

## Users



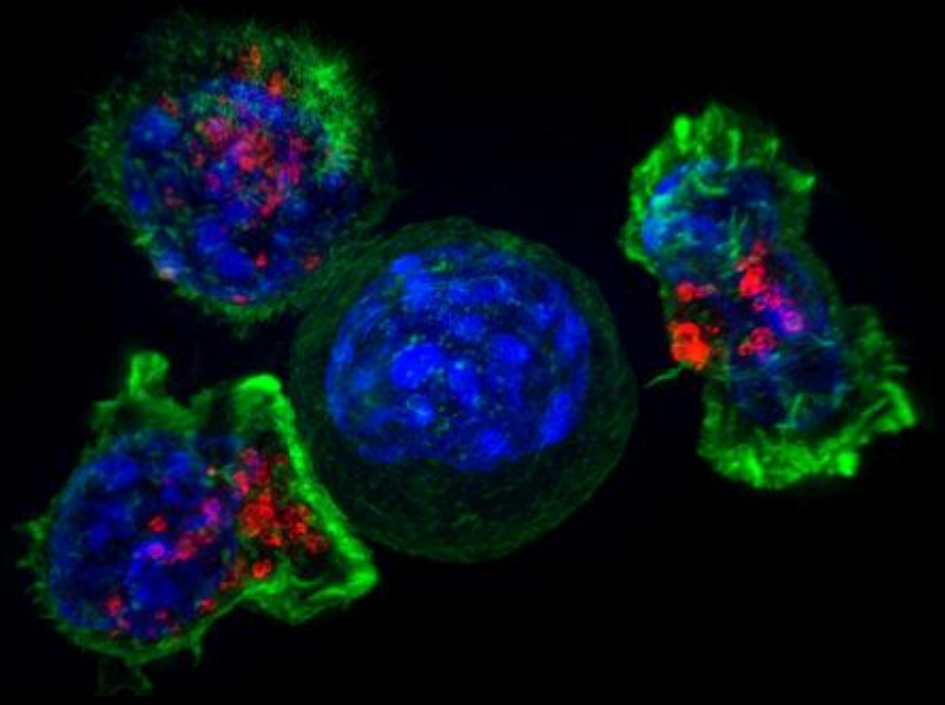
# Cloud-Scale Immunomics

*Reading the immune system to diagnose disease*

Microsoft

Adaptive biotechnologies®

Slides courtesy of Jonathan Carlson, Ph.D.
Microsoft Research Healthcare NExT



T-cells killing a cancer cell
Credit: Alex Ritter, Jennifer Lippincott Schwartz and Gillian Griffiths, National Institutes of Health

# Learning to decode the immune system to diagnose disease

Blood Sample

Immunosequencing

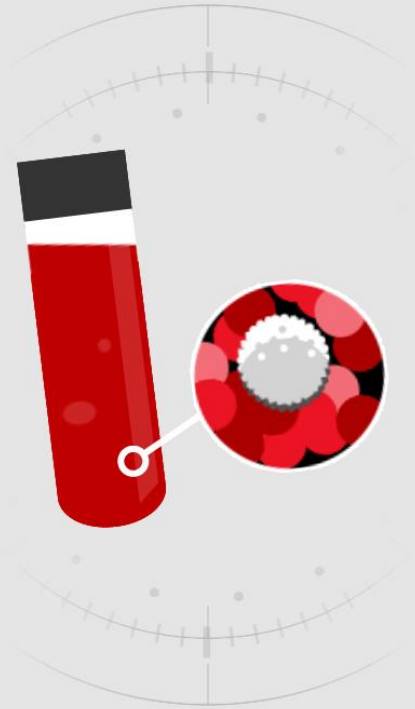

Machine Learning

Empowered care

Adaptive biotechnologies

Microsoft

# Learning to decode the immune system to diagnose disease



**Blood Sample**

Immunosequencing

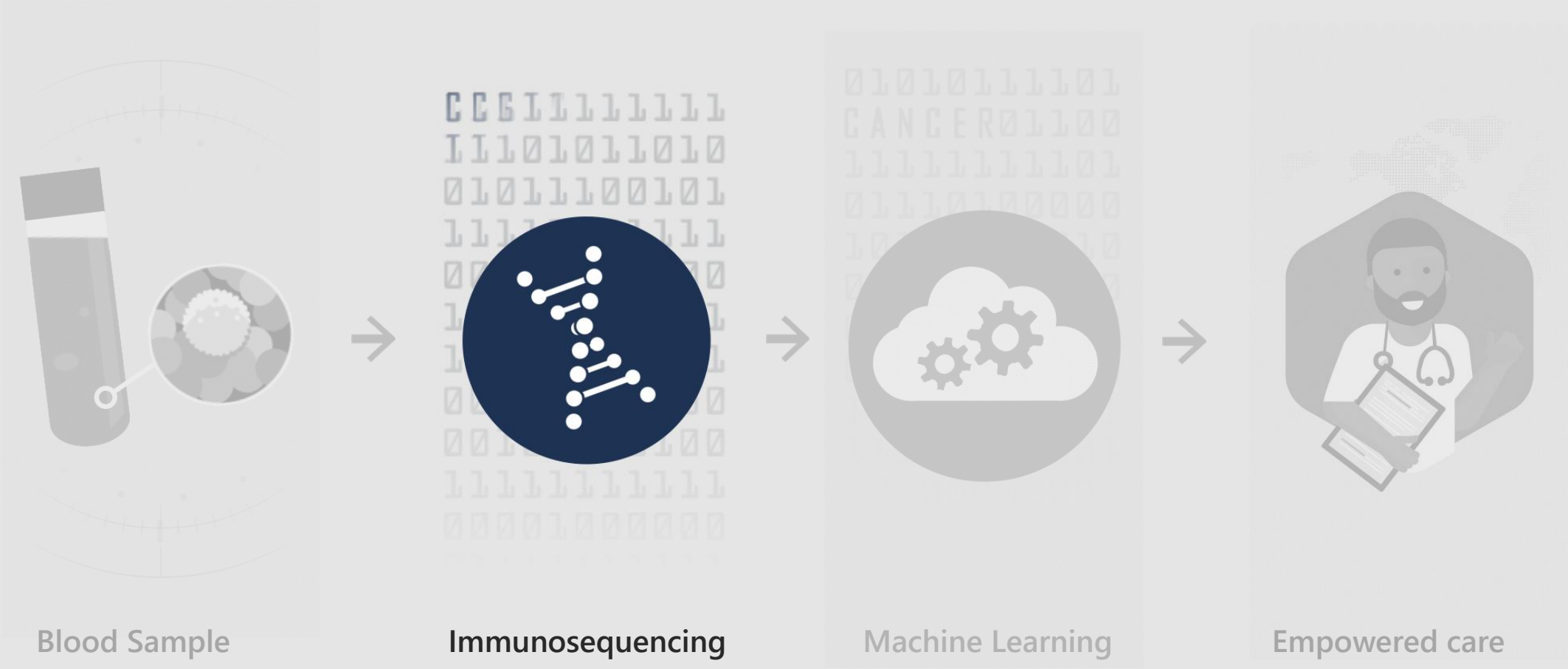
Machine Learning

Empowered care

Adaptive biotechnologies®

Microsoft

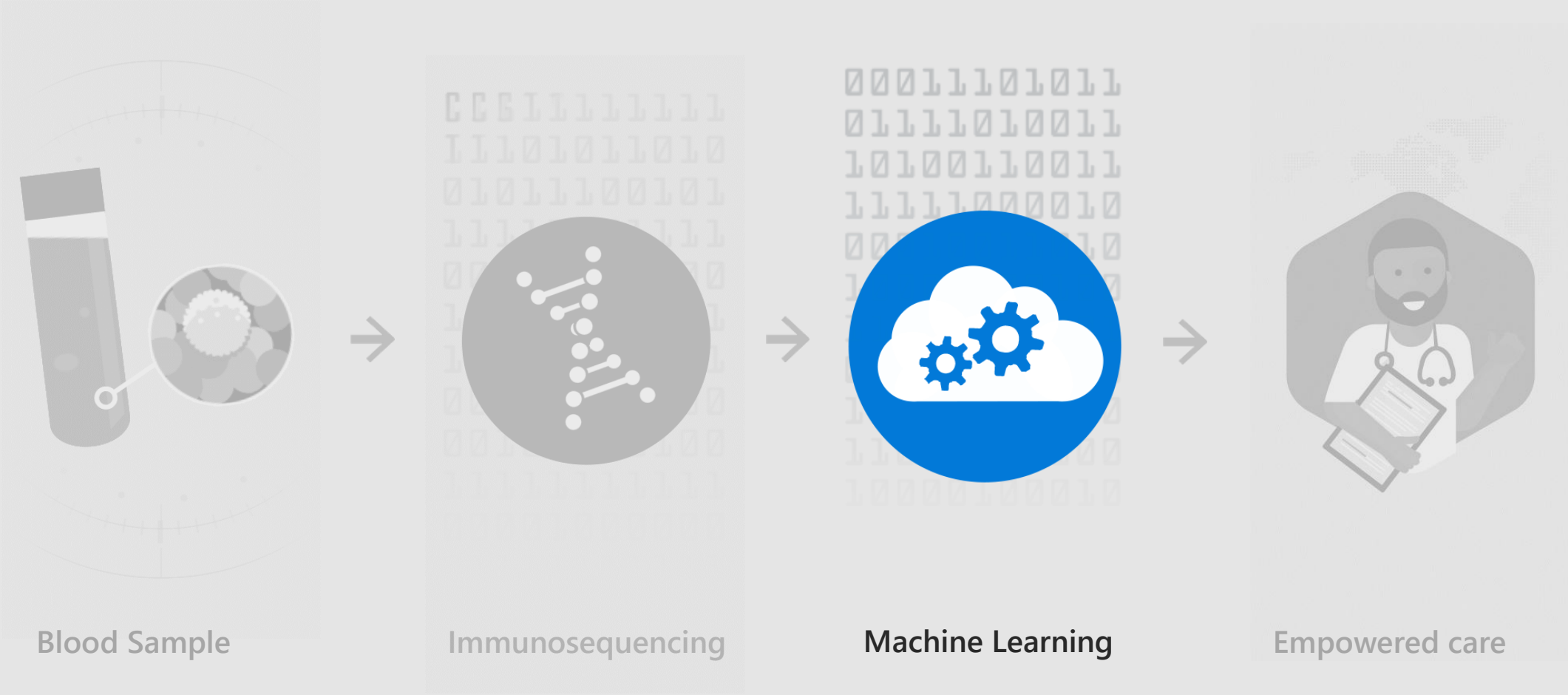
# Learning to decode the immune system to diagnose disease




Adaptive biotechnologies®

Microsoft

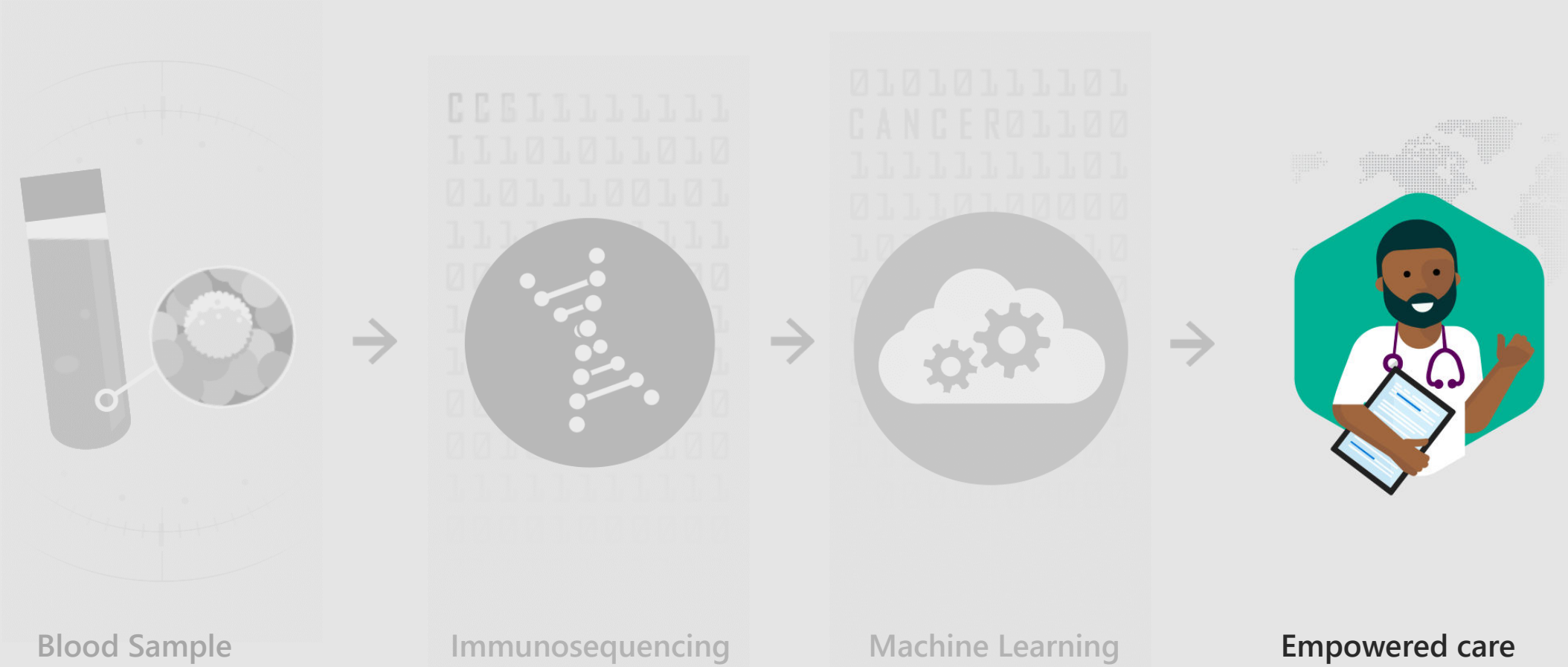
# Learning to decode the immune system to diagnose disease




Adaptive biotechnologies

Microsoft

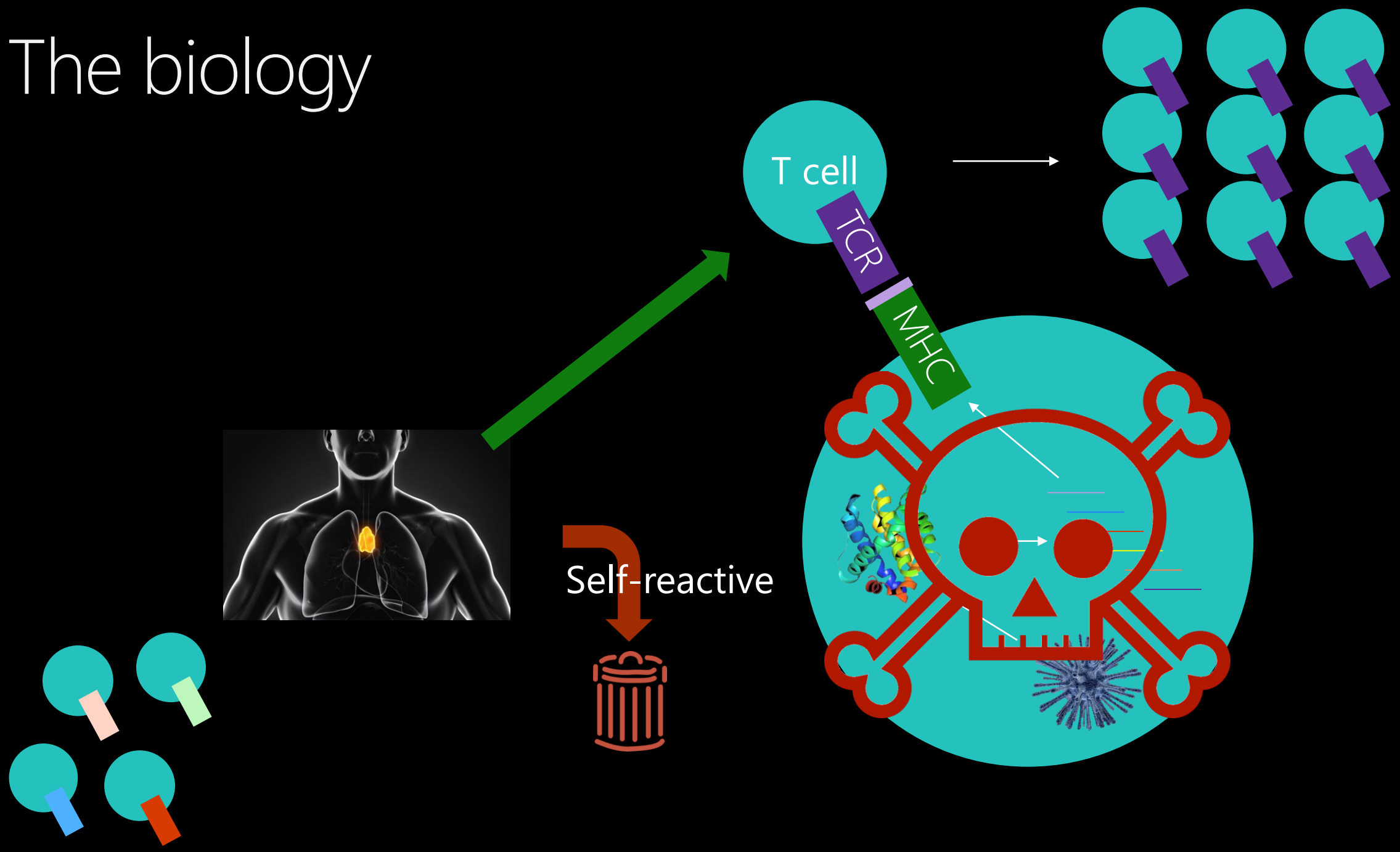
# Learning to decode the immune system to diagnose disease




Adaptive biotechnologies

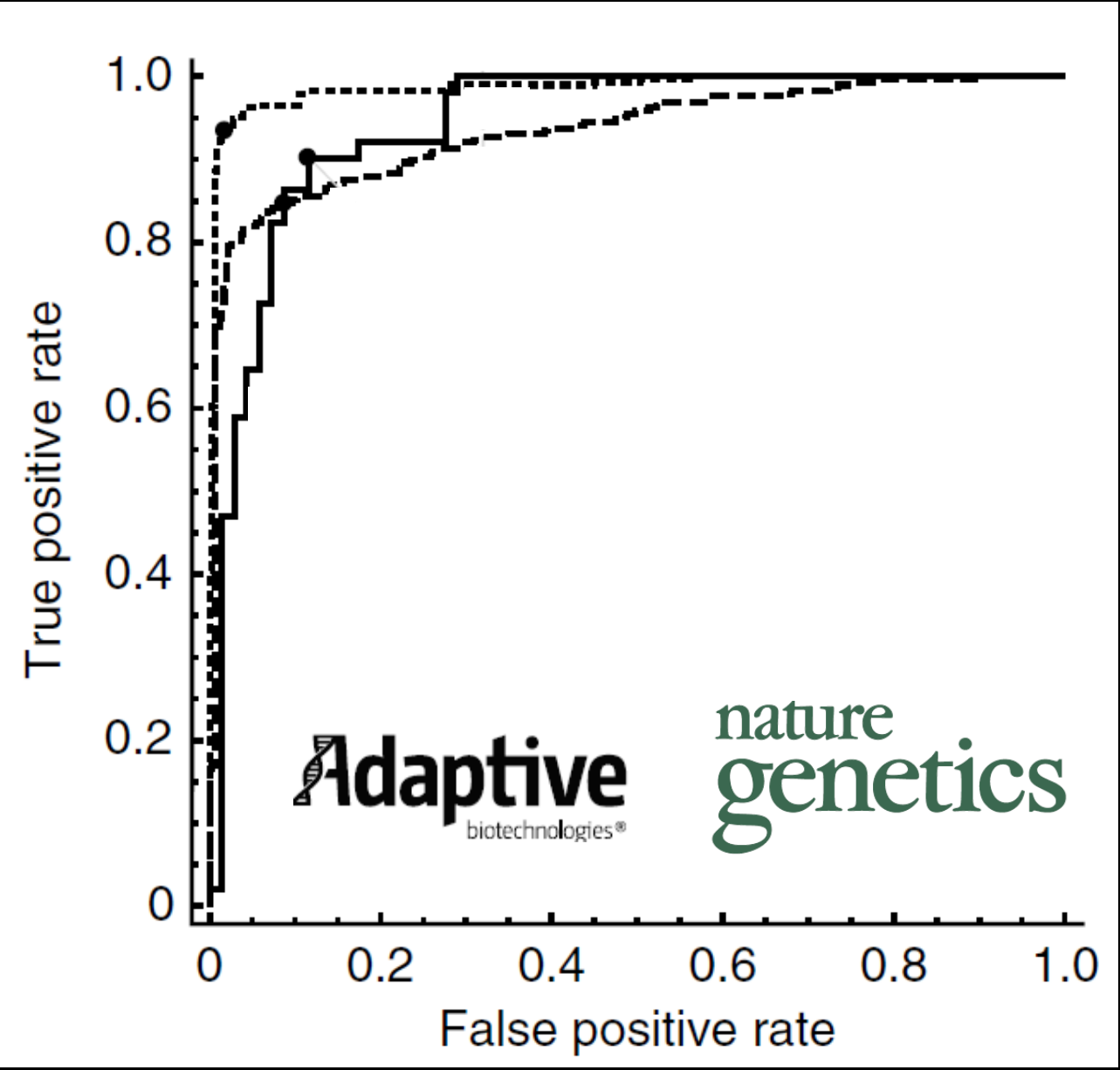
Microsoft

# The biology



# The machine learning

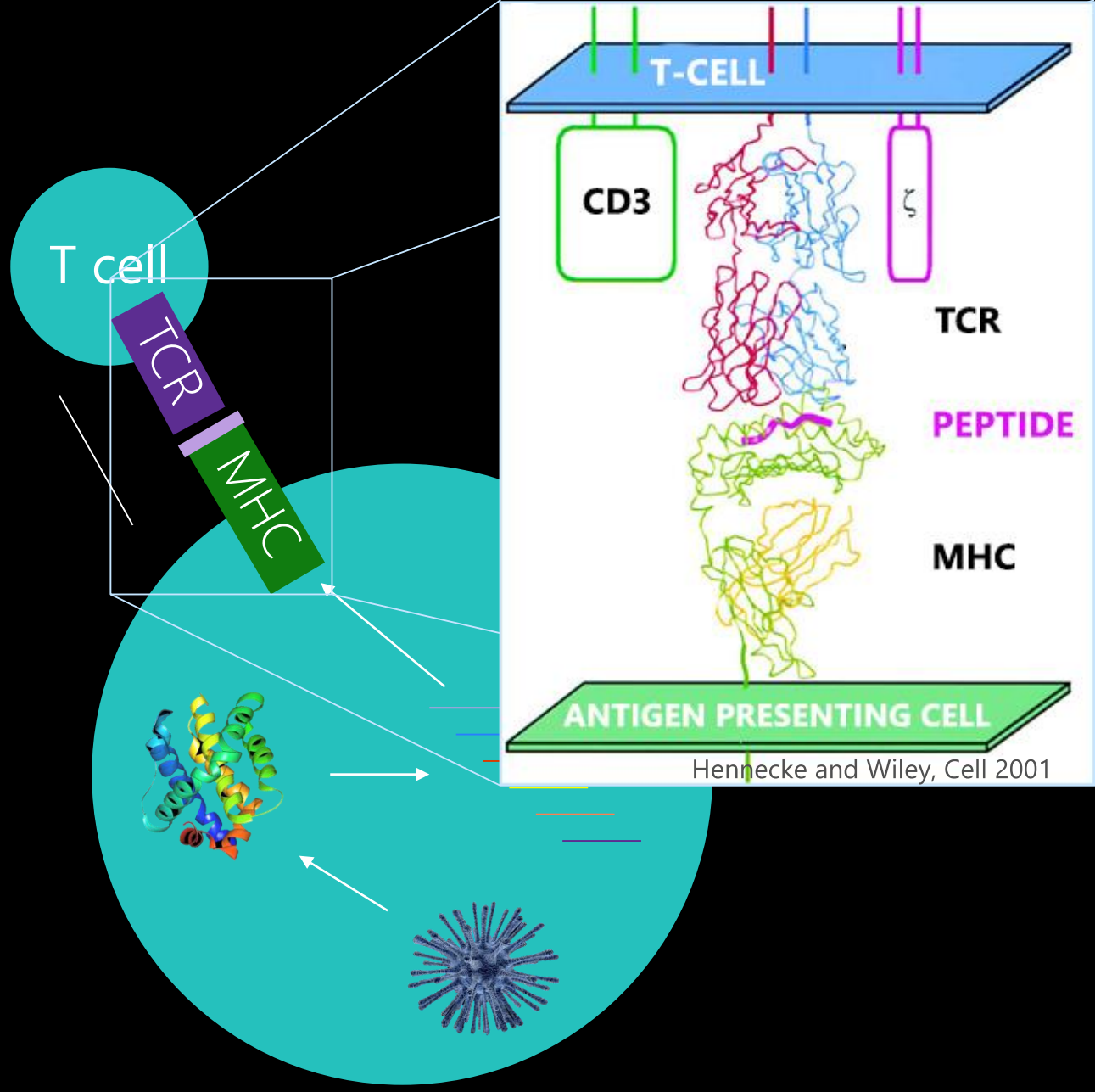
$$f_2(\{TCR\}) \rightarrow \{disease\}$$



# The machine learning

$f_1(TCR, MHC, peptide) \rightarrow \mathbb{R}$ (binding energy)

$f_2(\{TCR\}) \rightarrow \{disease\}$




Hennecke and Wiley, Cell 2001

# The data

$f_1(TCR, MHC, peptide) \rightarrow \mathbb{R}$ (binding energy)

MIRA

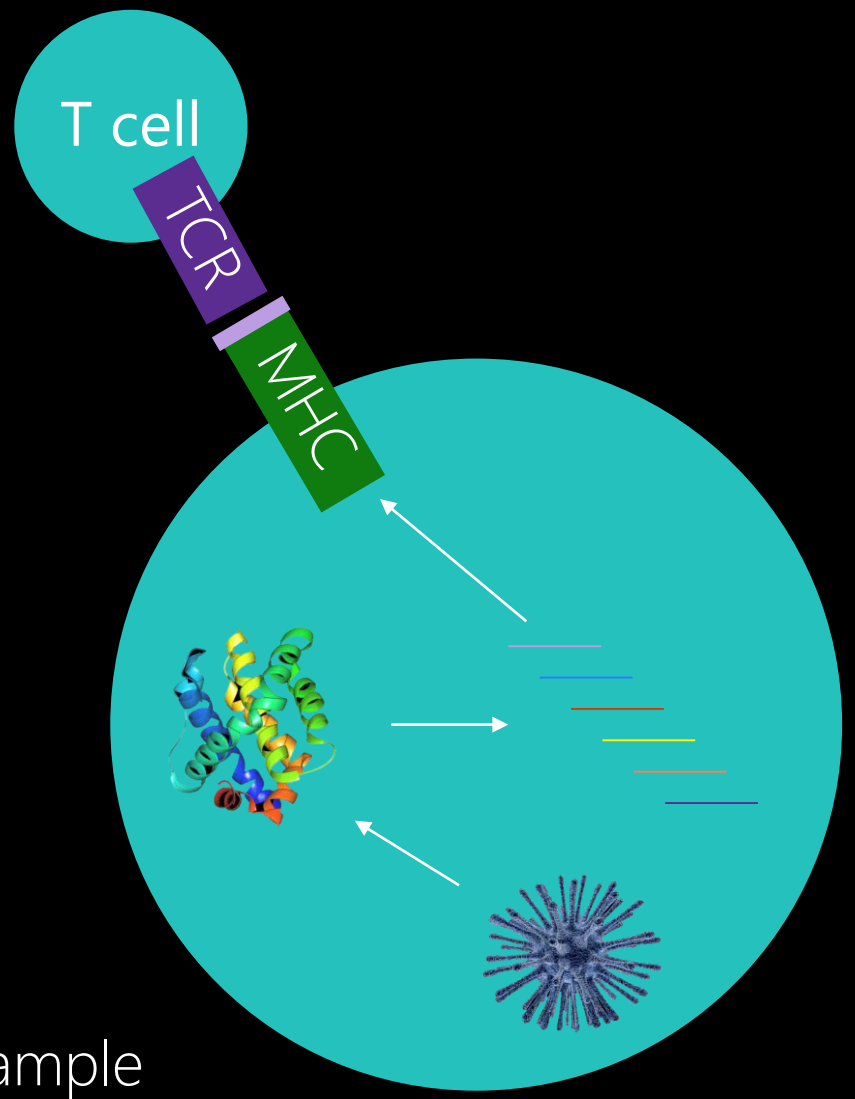
Pairwise binding data for
1k antigens against
1M TCR

$f_2(\{TCR\}) \rightarrow \{disease\}$

immunoSEQ®

pairSEQ

1M TCR per clinical sample



# Toward a universal diagnostic

T-cells are nature's universal diagnostic machine


Will generate billions of training samples per day


Cloud-scale machine learning to translate T-cells to antigens

Empowered care