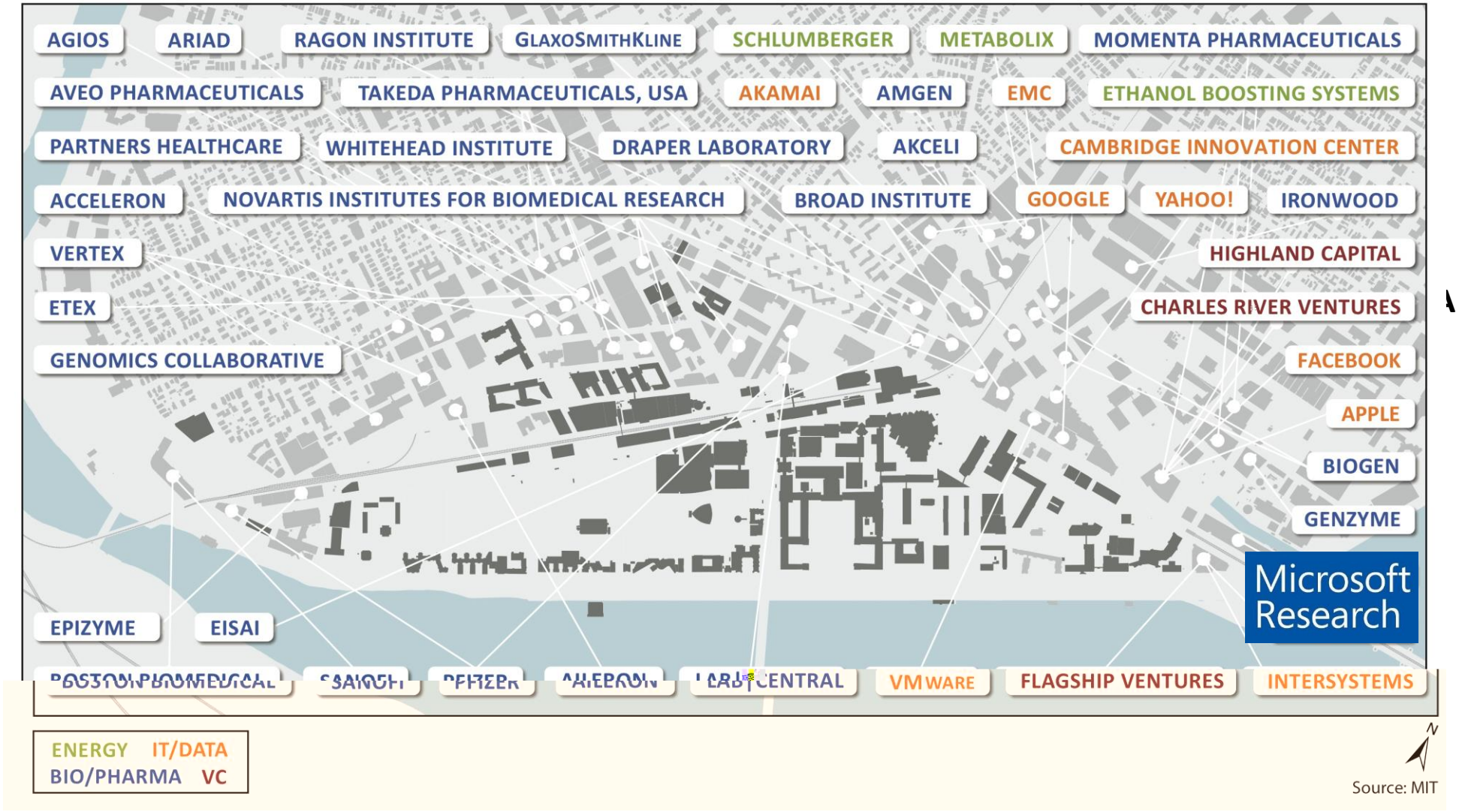

Microsoft
Microsoft Research
Faculty
Summit
2016

Machine learning for CRISPR gene editing

Nicolo Fusi

Microsoft Research

Microsoft Research, New England



Source: MIT

Acknowledgements

Broad Institute of MIT and Harvard

John Doench

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

Microsoft Research

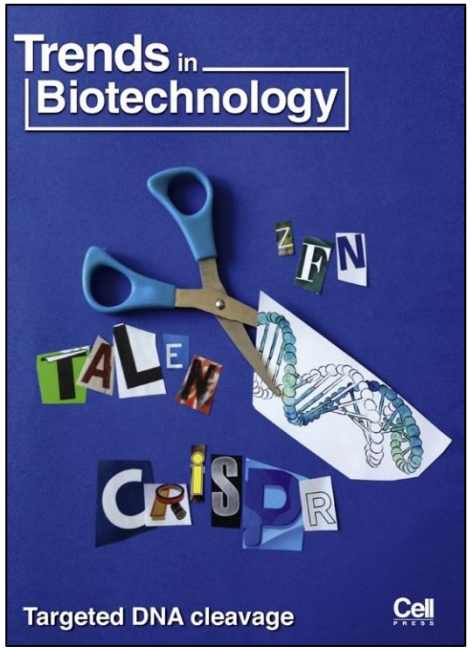
Jennifer Listgarten

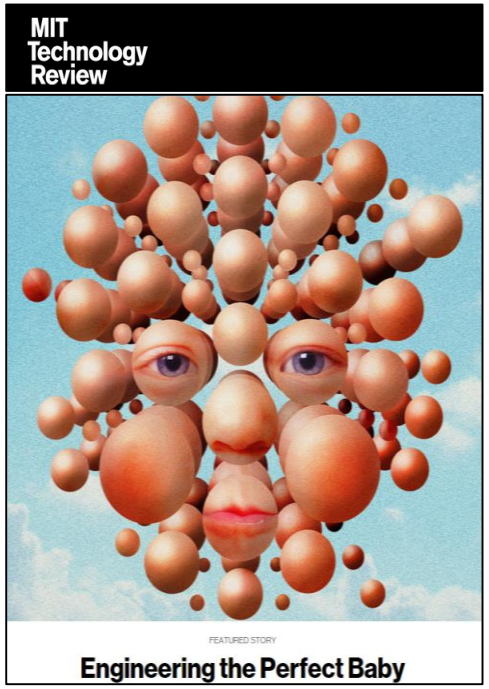
Washington University School of Medicine

Zuzana Tothova

Dana Farber Cancer Institute

Craig Wilen
Robert Orchard
Herbert W. Virgin





HEALTH

The New York Times

A Powerful New Way to Edit DNA

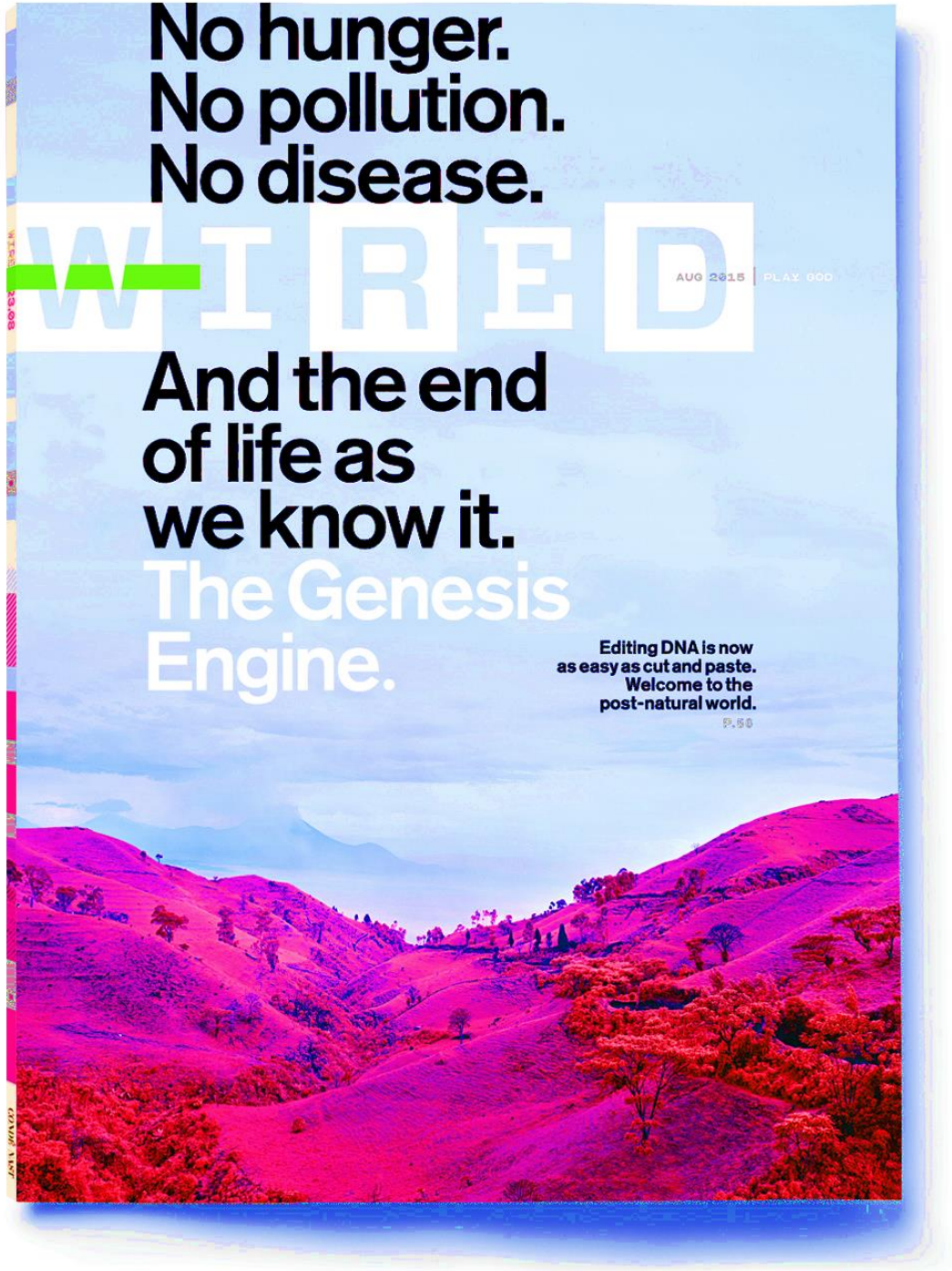
By ANDREW POLLACK MARCH 3, 2014

Could the DNA-editing CRISPR revolutionize medicine?

CNN

By Carina Storrs, Special to CNN

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



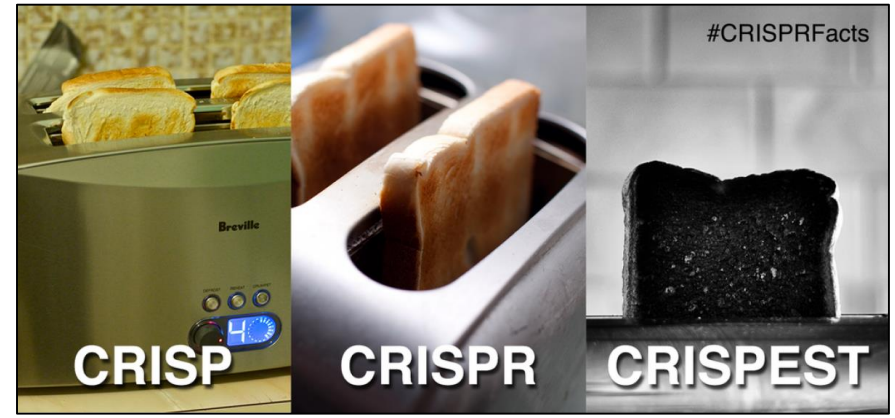
HEALTH

A Powerful New Way

By ANDREW POLLACK MARCH 3, 2014

ting CRISPR
ine?

CNN



Stephen B Montgomery @sbmontgom · 2h

With CRISPRs my lab is picking and choosing what X-men they want to be @wired #**crisprfacts** @dgmacarthur

smarf dos @smarfdoc · Aug 20

CRISPR can turn you into a baby ALL OVER AGAIN #**crisprfacts** #late

Matthew Cobb
@matthewcobb

CRISPR is both gold AND blue #crisprfacts

Chris Dwan
@fdmts

Follow

@dgmacarthur @EricTopol CRISPR cannot be overhyped. CRISPR proves P = NP. #crisprfacts

Henry Scowcroft
@oh_henry

If you genetically edit the lettuce genome, you can make it CRISPR #crisprfacts

Terry D. Johnson
@terrydjohnson

Peter Jackson worked with CRISPR to edit The Lord of the Rings. CRISPR was unavailable for The Hobbit. #crisprfacts

Promising results for translational medicine

Proof of principle in stem cells/model organisms:

- Remove CCR5 receptor used by **HIV**.[1]
- Correct a CFTR defect associated with **cystic fibrosis**.[2]
- Corrected **muscular dystrophy** gene to produce cured mice.[3]

1. Mandal *et al*, Cell Stem Cell 2014
2. Schwank *et al*, Cell Stem Cell 2013
3. Long *et al*, Science 2014

Not quite ready for prime time

Want

Have

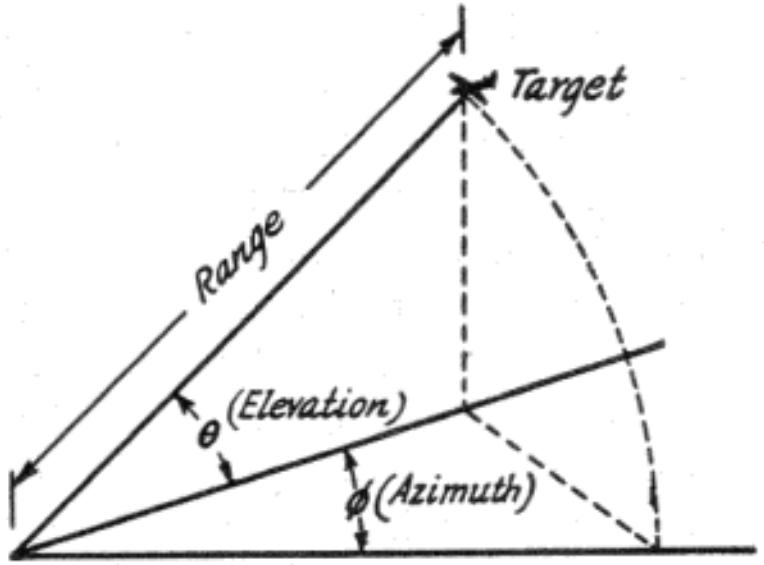
Not quite ready for prime time

Two problems and two solutions:

1. Better "on-target" efficiency needed: *Azimuth*.
2. Elimination/reduction of "off-target" effects: *Elevation*.

Solution paths:

- Smarter/improved lab protocols.
- Machine learning.



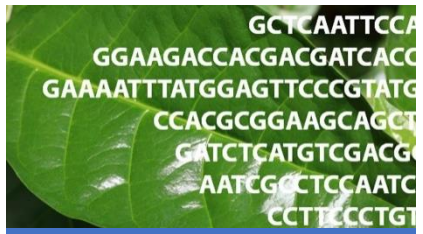
A short intro to CRISPR for gene editing

naturally-occurring bacterial immune defense mechanism

co-opted for generic editing in any organism

human disease

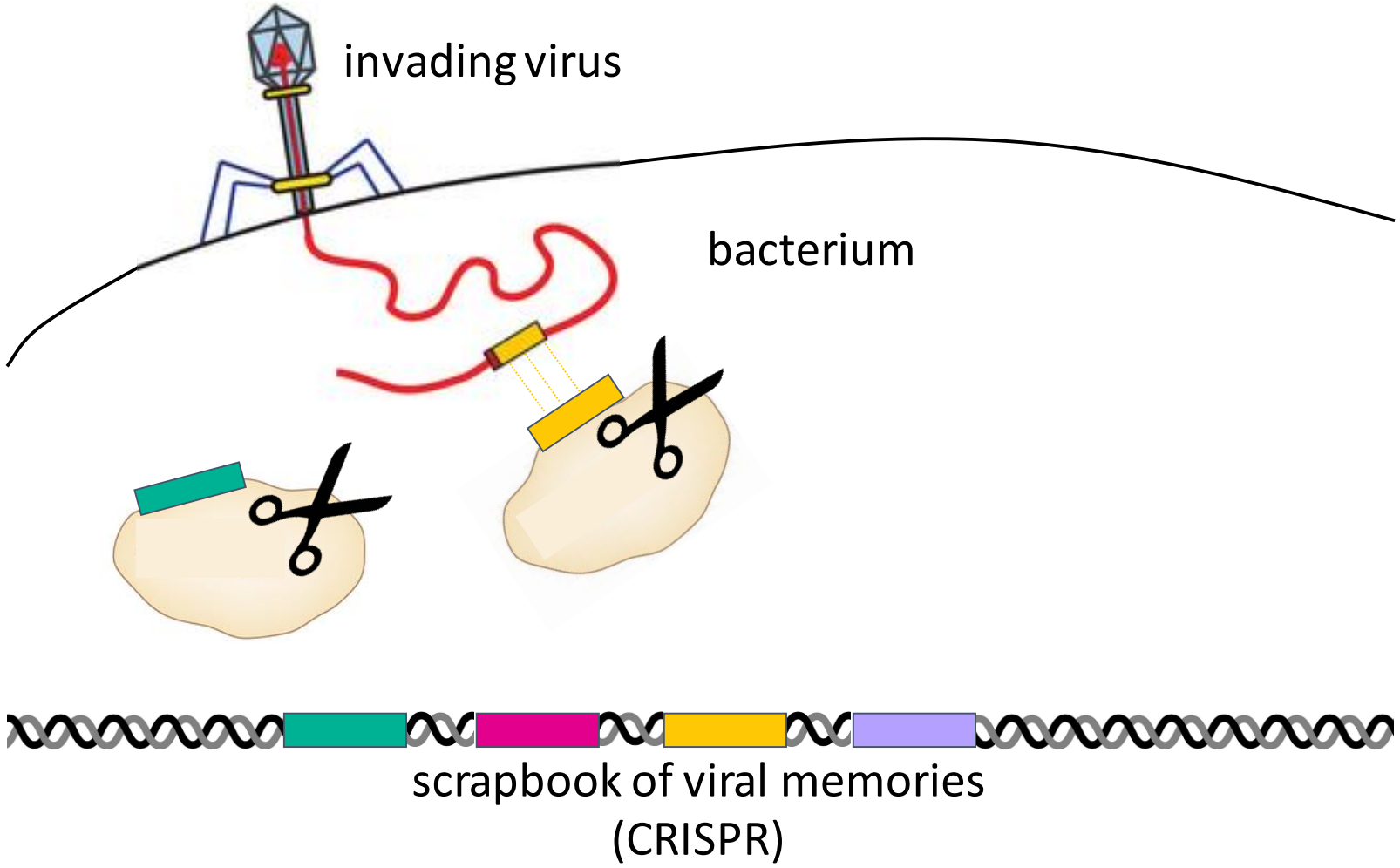
drug development



agriculture

CRISPR = Clustered **R**egularly **I**nterspaced **S**hort **P**alindromic **R**epeats

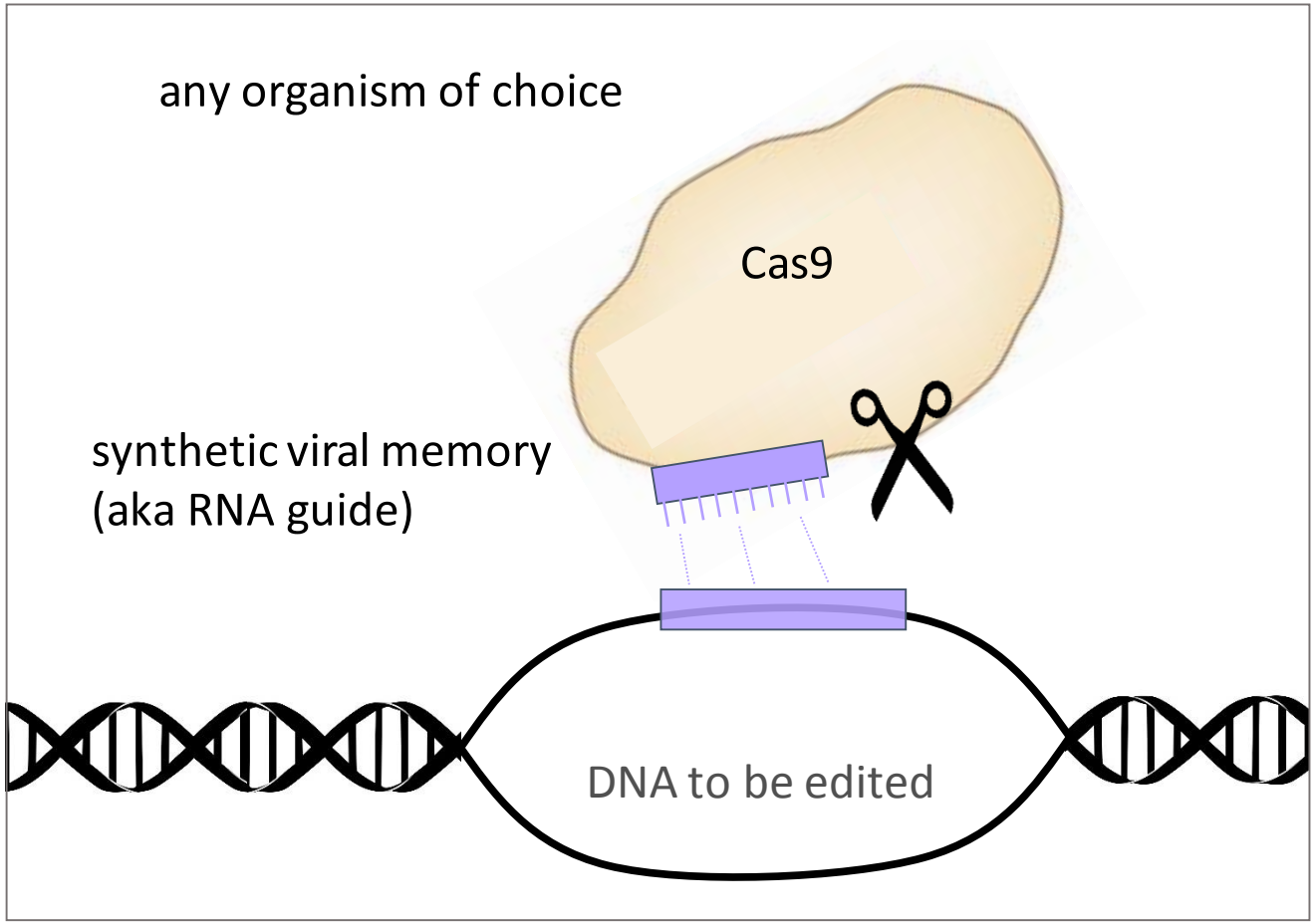
Originates from two-part bacterial defense mechanism



2 Cut & paste mechanism

1 Viral scrapbook

Gene editing using CRISPR



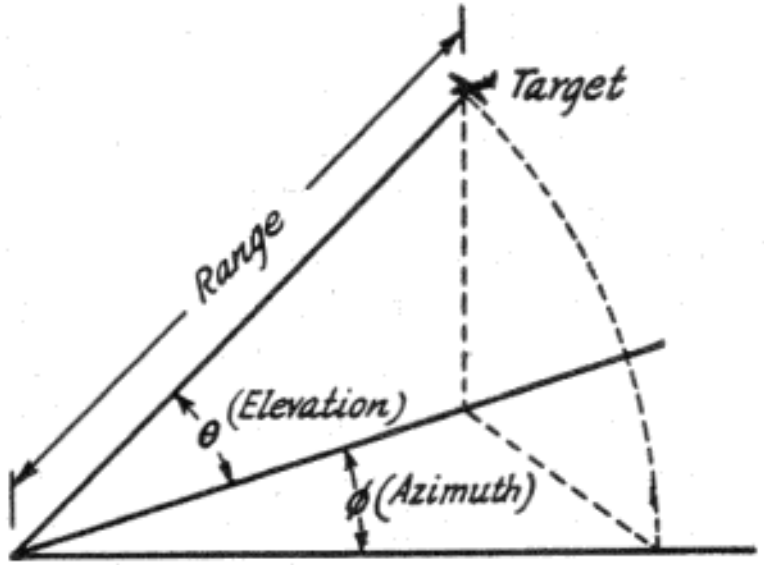
Not quite ready for prime time

Two problems and two solutions:

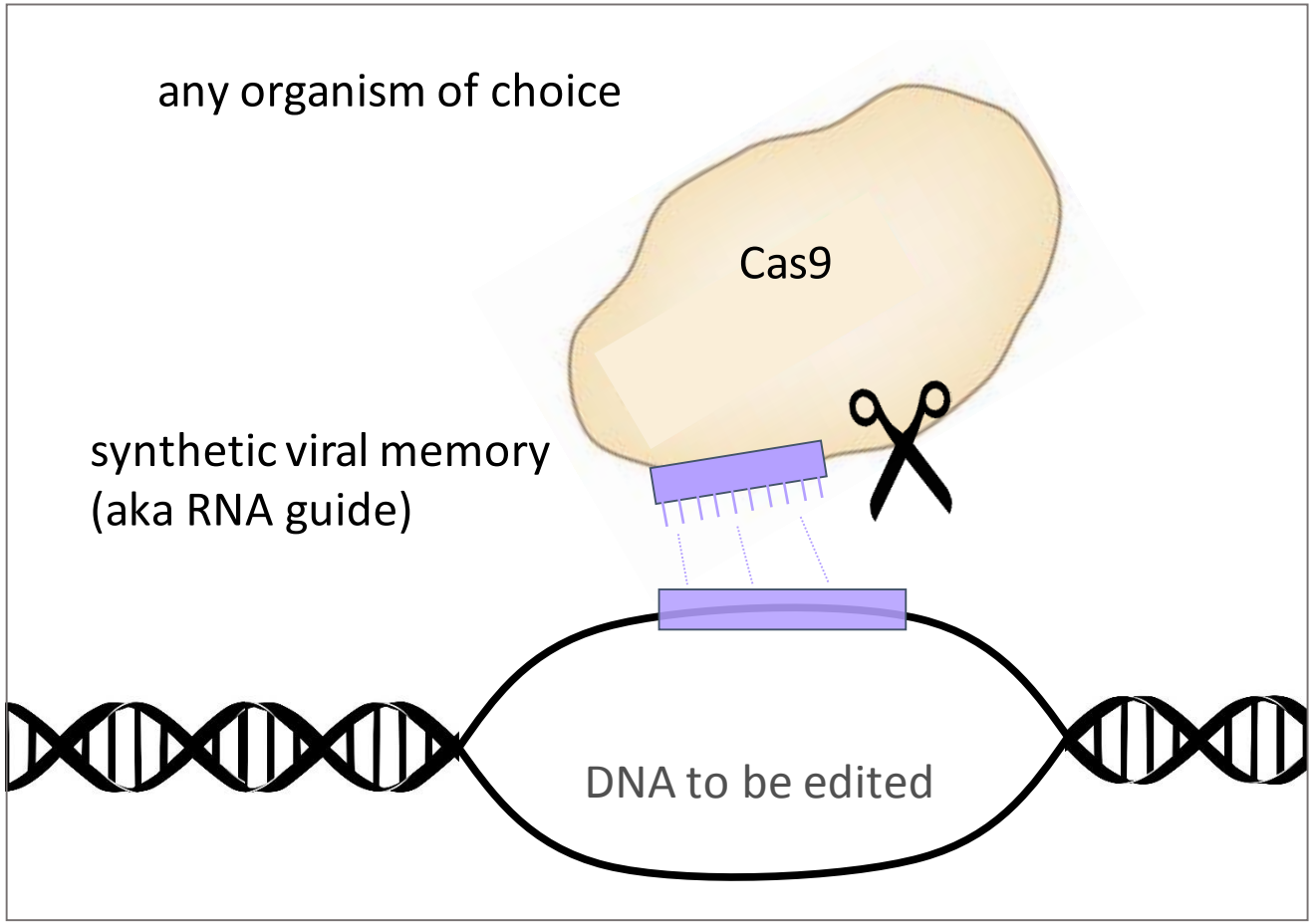
1. Better "on-target" efficiency needed: *Azimuth*.
2. Elimination/reduction of "off-target" effects: *Elevation*.

Solution paths:

- Smarter/improved lab protocols.
- Machine learning.



Machine learning predictive modelling for CRISPR



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

In silico prediction of guide efficiency

$[x_1, x_2, \ldots, x_M]$

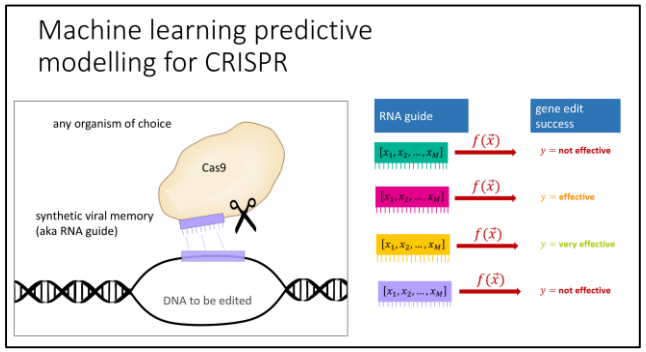
Input features
(e.g. guide sequence,
GC content of target gene)

$y =$ **effective**

Measured guide efficacy
(e.g. "working" vs "not working")

$f(\vec{x})$

Model
(e.g. Logistic
Regression)



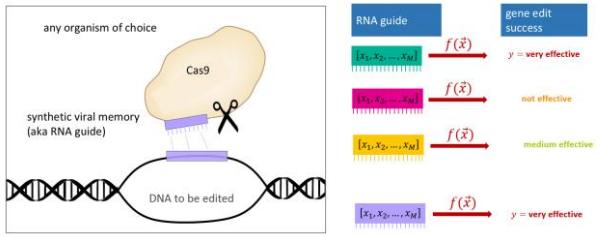
Azimuth: our state-of-the art approach

- Investigate and use richer features of the RNA guide.
- Removed information bottlenecks to the supervised signal.
- Investigate richer model classes.

Azimuth: our state-of-the art approach

- Investigate and use richer features of the RNA guide.
- Removed information bottlenecks to the supervised signal.
- Investigate richer model classes.

Featurization of a guide



20mer guide NGG PAM

TGGAGGCTGCTTTACCCGCTGTGGG**GG**CGC

4mer extra context 3mer extra context

$$\vec{x} = [x_1, x_2, \ldots, x_M] = [0,1,1,0, \ldots 3.4, 0,1,0,0,0,9.8, 0,0,0.1]$$

J.A.J.

Just Ask John

Melting temperatures

temperature at which half of the DNA strands are in the random coil or single-stranded (ssDNA) state.

TGGAGGCTGCTTTACCCGCTGTGGGGGCGC

30mer

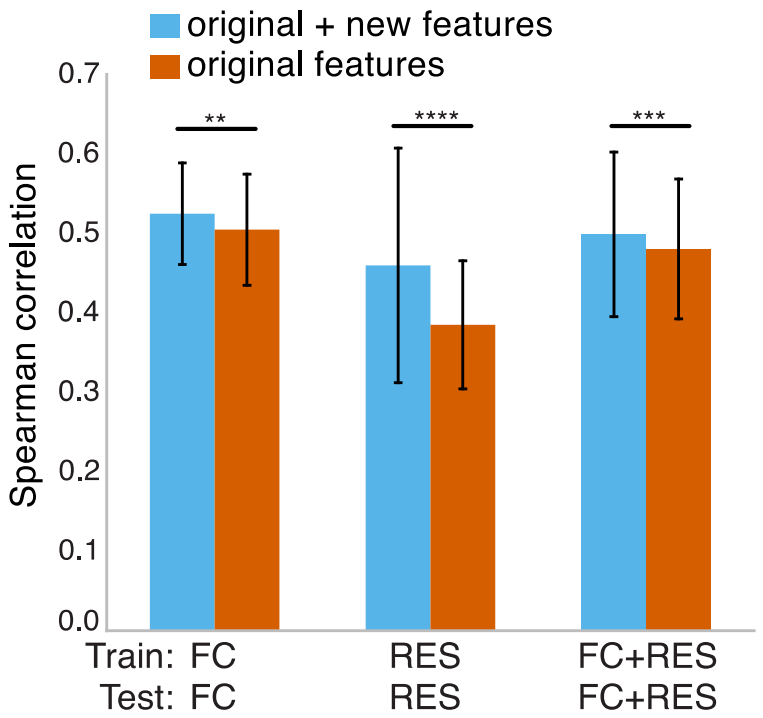
5mer proximal to PAM

8mer in position 8-15 of 20mer guide

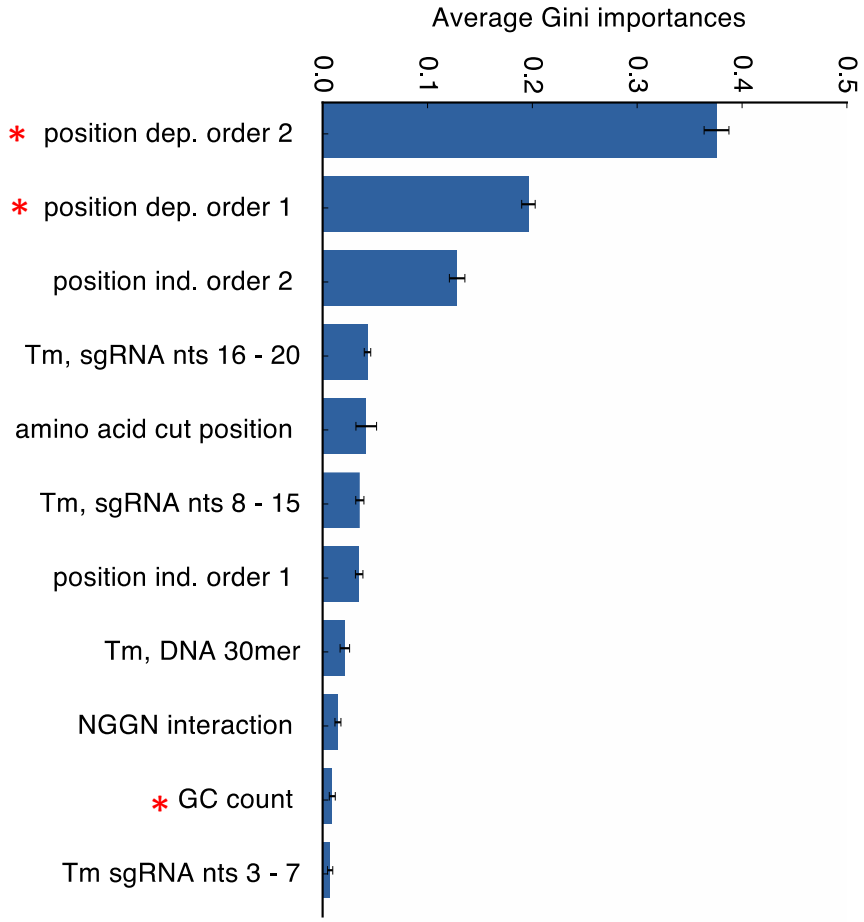
5mer in position 3-7 of 20mer guide

[credit: McGovern Institute for Brain Research at MIT]

Additional features improve performance



evaluated using L1 regression



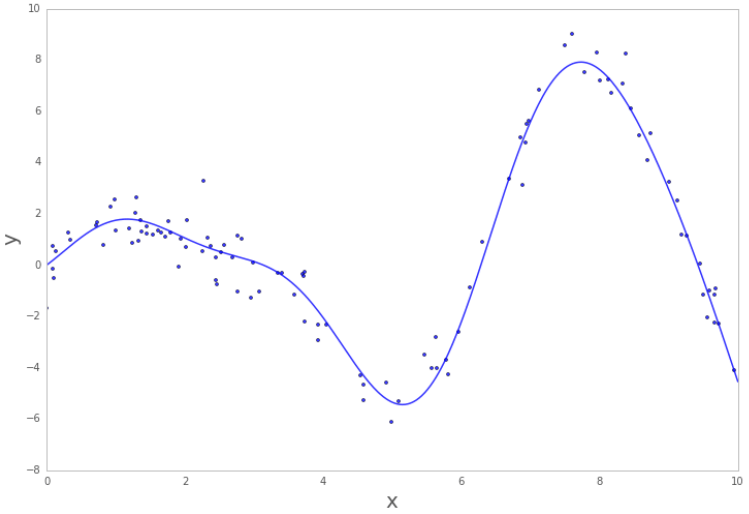
*old feature

Azimuth: our state-of-the art approach

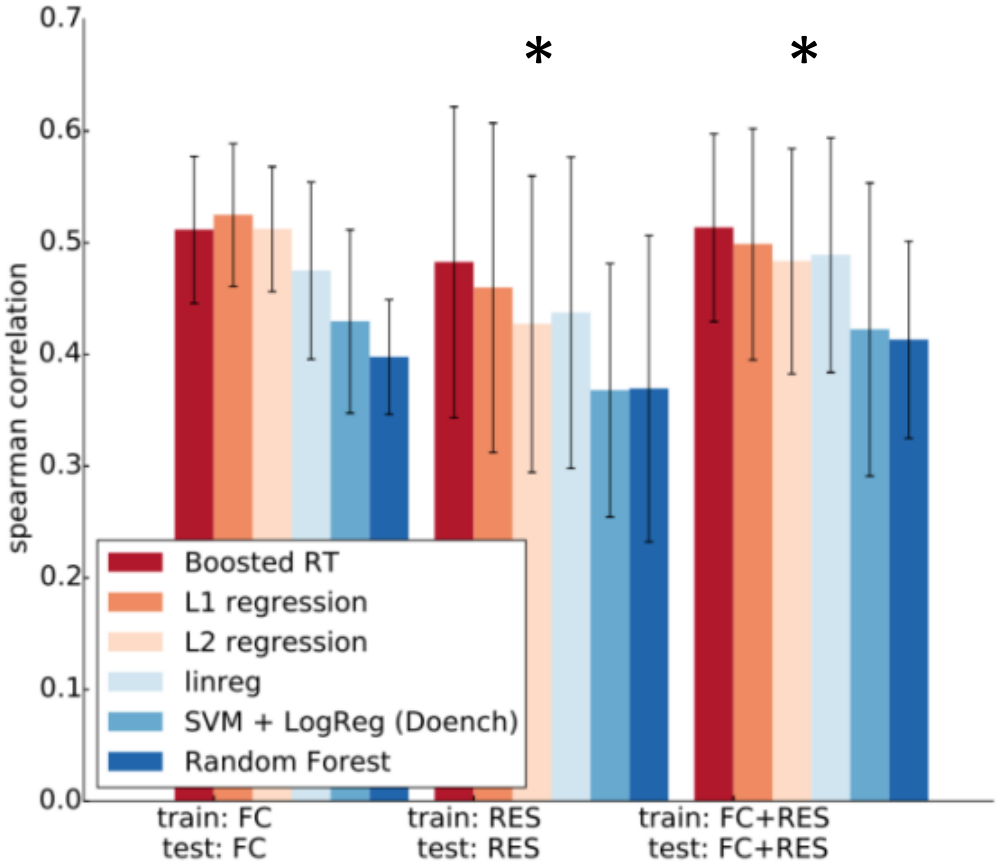
- Investigate and use richer features of the RNA guide.
- Removed information bottlenecks to the supervised signal.
- Investigate richer model classes.

Non-linear modelling

- Simple linear models are incapable of representing or capturing complex interactions between the variables.
- For the final model we use **Gradient-Boosted Regression Trees** (GBRTs)
- An ensemble of weak predictors (regression trees).
- Each RT is trained on the residuals of the previous one.
- GBRTs can easily handle non-homogeneous data (mix of categorical and continuous).



Systematic comparison of models



Impact of our Azimuth model

- *Nature Biotechnology* 2016.
- Recommended by independent studies (Haeussler et al. 2016).
- **Adoption** by two startups and academics/researchers worldwide.
- Azure ML service **~1000 requests/day**, doubling every 3 months
- Web service **~300 requests/day**.
- Over **1000 open-source software downloads**.

http://research.microsoft.com/en-us/projects/azimuth

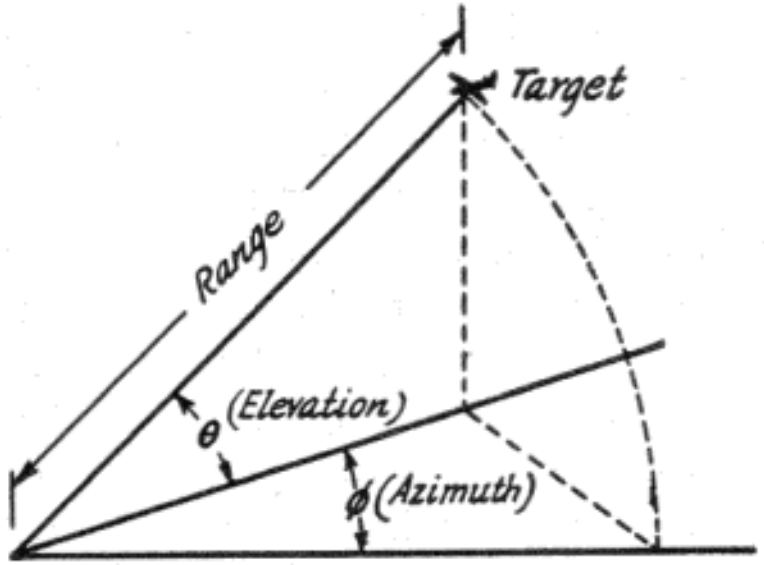
Not quite ready for prime time

Two problems and two solutions:

1. Better "on-target" efficiency needed: *Azimuth*.
2. Elimination/reduction of "off-target" effects: *Elevation*.

Solution paths:

- Smarter/improved lab protocols.
- Machine learning.



Elevation: prediction of off-target effects

Much more challenging than on-target:

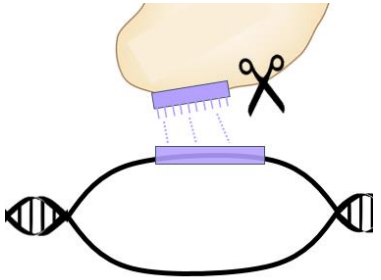
- For just one single guide need to check for imperfect matches genome-wide.
- Combinatorial explosion of mismatches, hard to get enough training data.

intended target

GGCTGCTTTACCCGCTGTGGG

...CTATAACTGGCAGCTCTACCCGGTGTGGGACAAG...

whole genome—potential off-targets



Combinatorial explosion (for 1 guide in 1 gene)

1 mismatch: **69** sites

2 mismatches: **2277** sites

3 mismatches: **47,817** sites

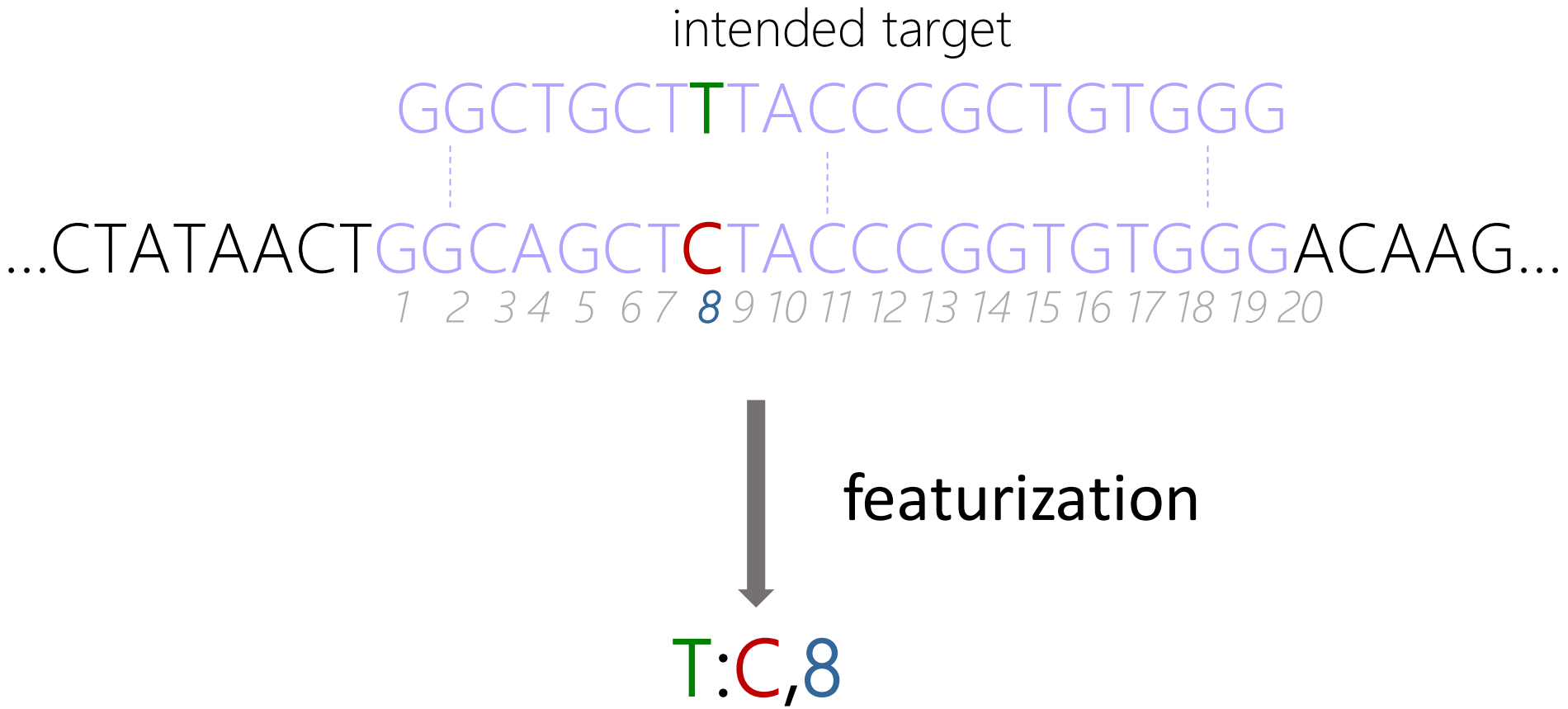
4 mismatches: **717,255** sites

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

1 full example

very sparsely sampled across different genes

Previous state-of-the-art approach: *CFD* (Doench et al 2016)



Previous state-of-the-art approach: *CFD (Doench et al 2016)*

$$CFD \approx \prod_i P(Y = 1 | X_i = 1)$$

- Measured off-target activities (on a continuous scale) are **discretized** in present (1) vs not present (0).
- CFD computes probability of off-target given mismatch.
- Probabilities are aggregated assuming conditional independencies.

Elevation: generalizations of *CFD*

$$CFD \approx \prod_i P(Y = 1|X_i = 1)$$

1. Change from classification to regression for $P(Y = 1|X_i = 1)$.
2. Augment the feature space from T:C,8.
3. Use non-linear regression model for $P(Y = 1|X_i = 1)$, in particular Boosted Regression trees.
4. Refine predictions with a second model layer using the multi-mismatch data.

Elevation: generalizations of *CFD*

$$CFD \approx \prod_i P(Y = 1|X_i = 1)$$

Goal 1: make better use of the better-sampled 1 mismatch data

1. Change from classification to regression for $P(Y = 1|X_i = 1)$.
2. Augment the feature space from T:C,8.
3. Use non-linear regression model for $P(Y = 1|X_i = 1)$, in particular Boosted Regression trees.
4. Refine predictions with a second model layer using the multi-mismatch data.

Elevation: generalizations of *CFD*

Goal 2: relax independence and other assumptions using sparsely-sampled data

$$CFD \approx \prod_i P(Y = 1 | X_i = 1)$$

1. Change from classification to regression for $P(Y = 1 | X_i = 1)$.
2. Augment the feature space from T:C,8.
3. Use non-linear regression model for $P(Y = 1 | X_i = 1)$, in particular Boosted Regression trees.
4. Refine predictions with a second model layer using the multi-mismatch data.

Cascading from single mismatch to multi-mismatch

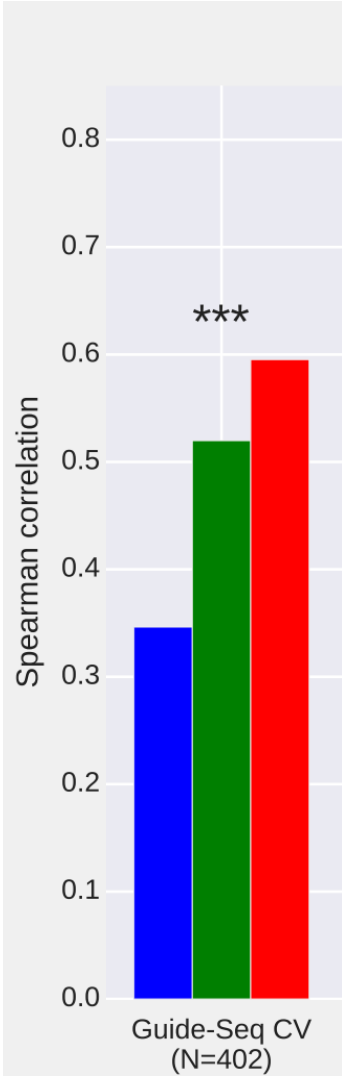
1. **Non-linear** regression model trained on **1-mismatch data**.
   - Complex model capturing interactions
   - Can only compute predictions for 1 mismatch at a time

Elevation-naive

2. **Linear** model trained on scarce **multi-mismatch data**
   - Relatively simple model
   - Trained on individual and aggregated predictions (e.g. product, sum) from layer 1

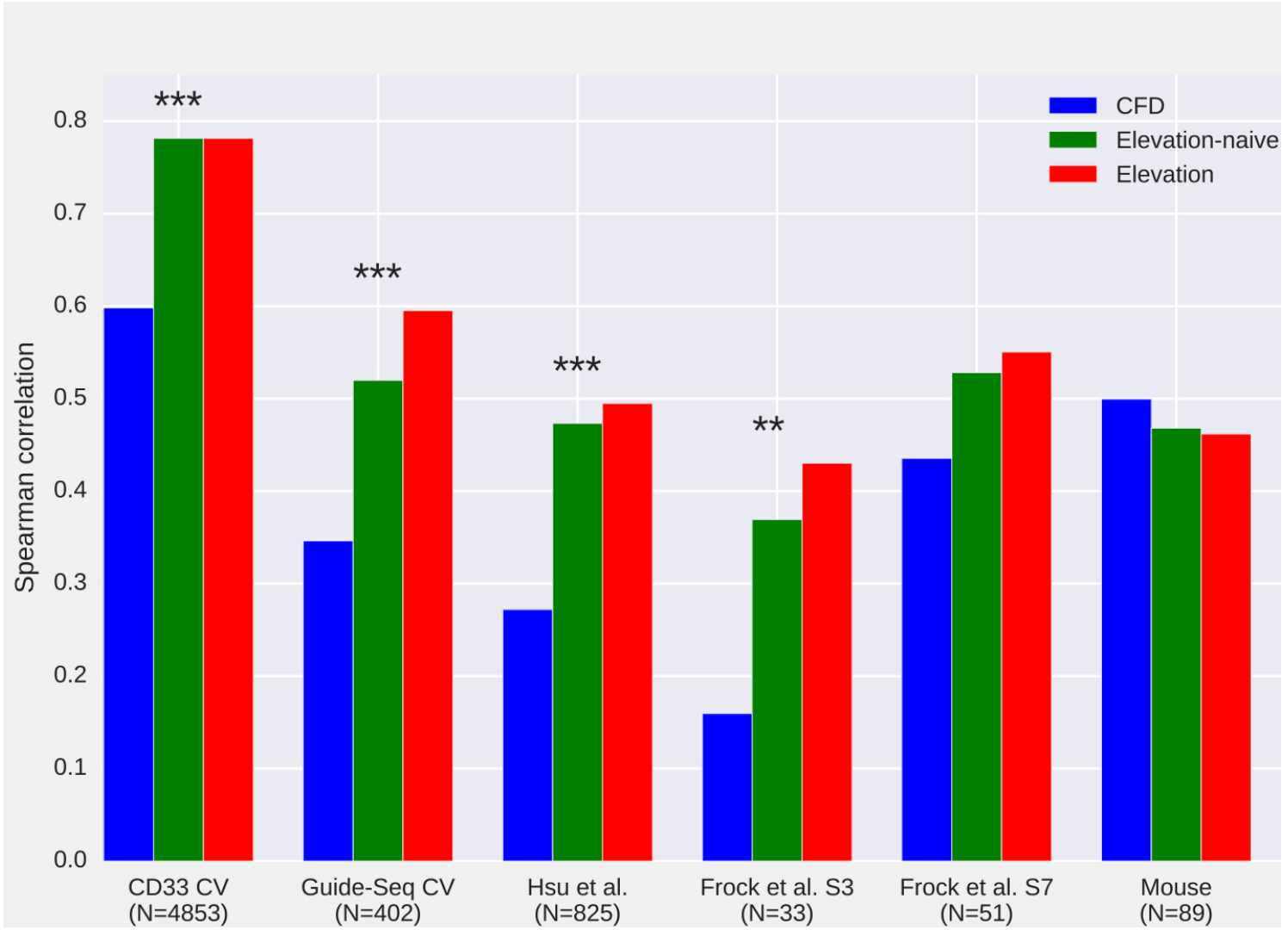
Elevation

Elevation outperforms *CFD* by 64%



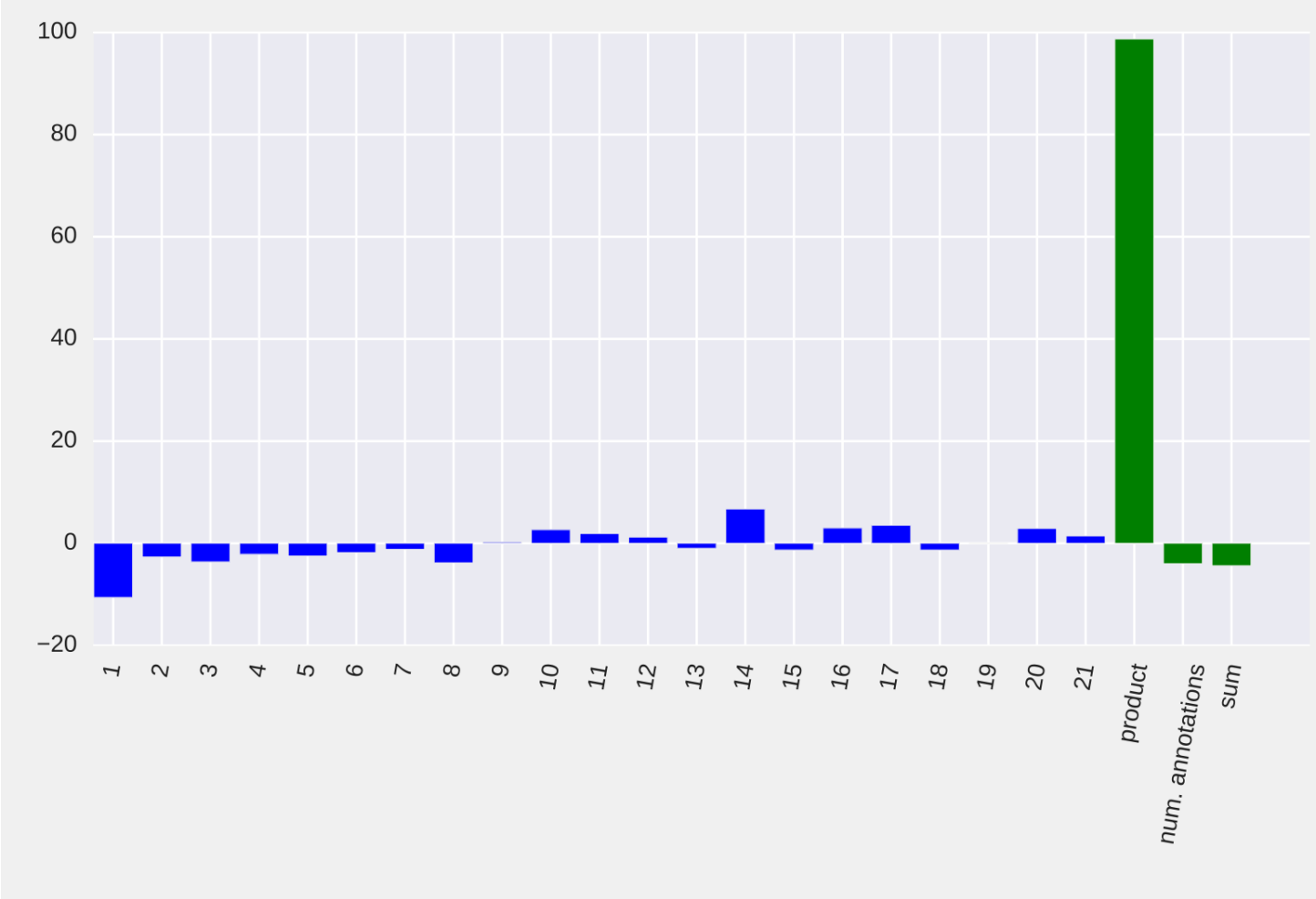
- Elevation spearman $\rho = 0.59$
- CFD spearman $\rho = 0.36$
- 64% improvement $(p = 5.5 \times 10^{-5})$

Elevation performs best on 4/5 other data sets



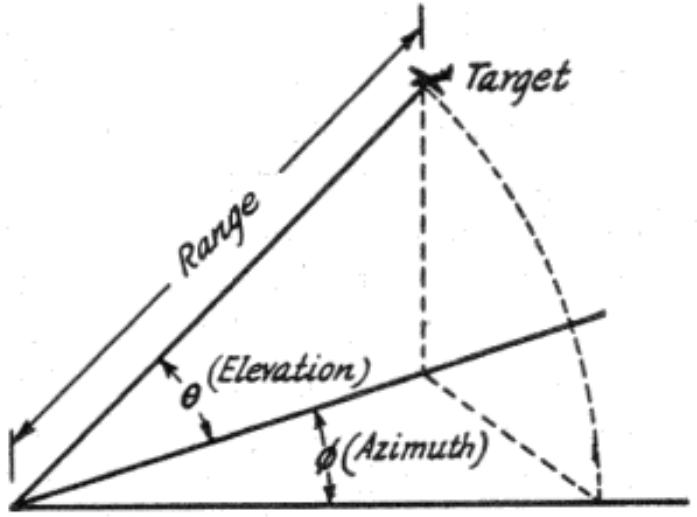
Mitigation of assumptions

Quantitative correction from the full-assumptions model



Putting it all together

- Elevation cloud prediction server.
- Open source code.
- Framework to efficiently search genome-wide for mismatches and call **Azimuth** & **Elevation**.



Acknowledgements

Broad Institute of MIT and Harvard

John Doench

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

Microsoft Research

Jennifer Listgarten

Washington University School of Medicine

Zuzana Tothova

Dana Farber Cancer Institute

Craig Wilen
Robert Orchard
Herbert W. Virgin