

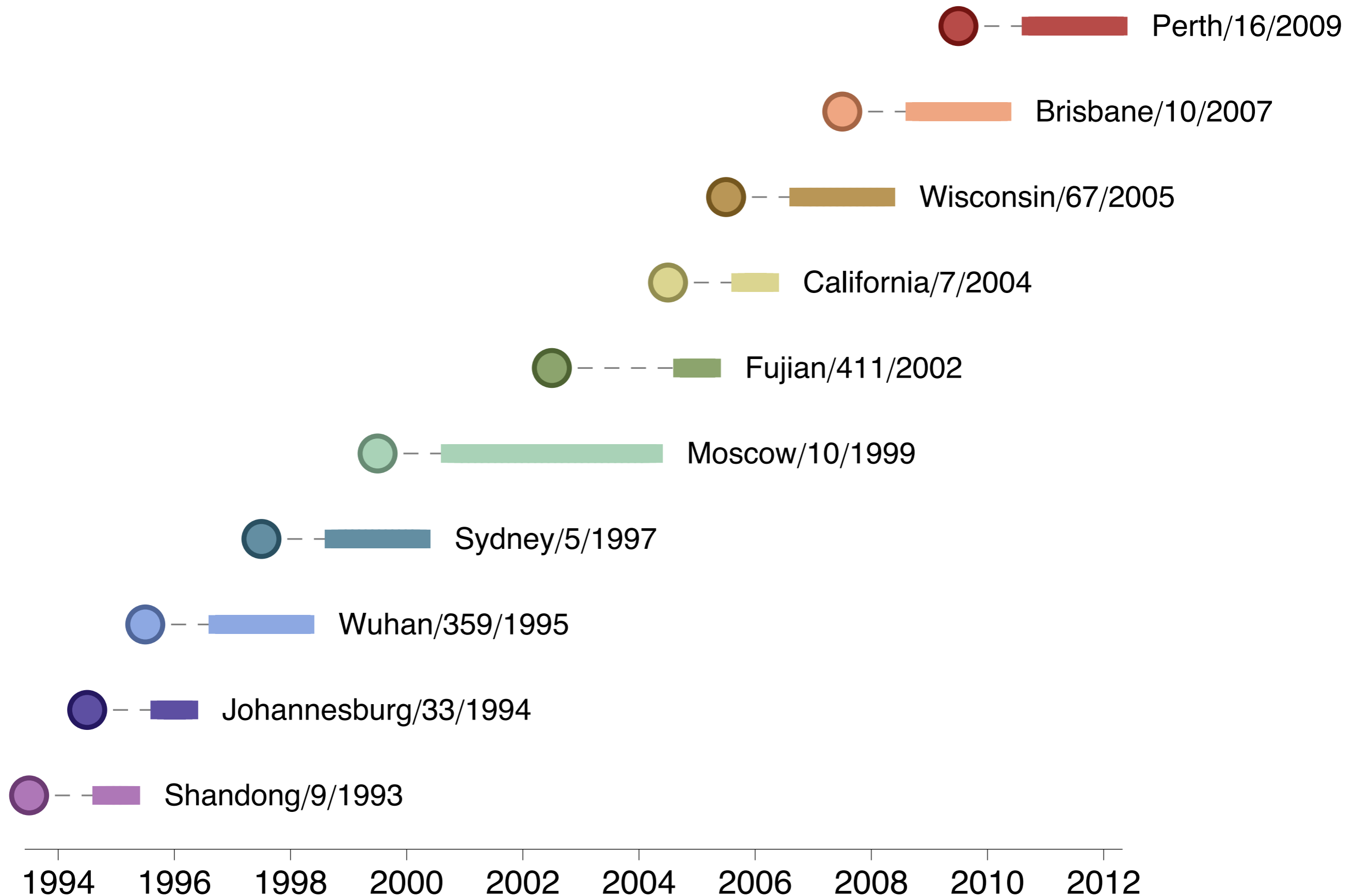
# Influenza antigenic evolution and vaccine strain selection

Trevor Bedford (@trvrb)

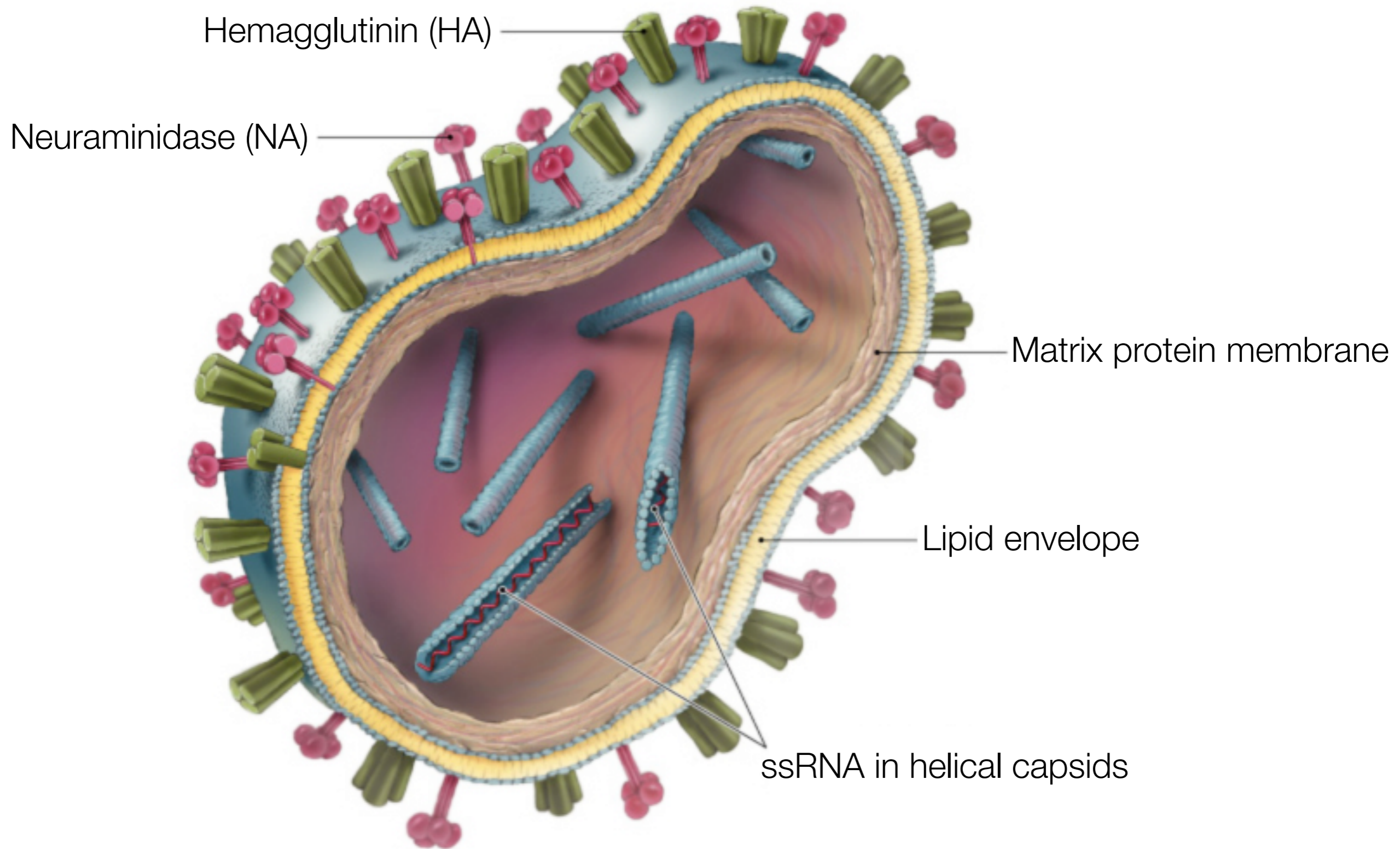
<http://bedford.io>



# Influenza H3N2 vaccine strains from 1994 to 2011



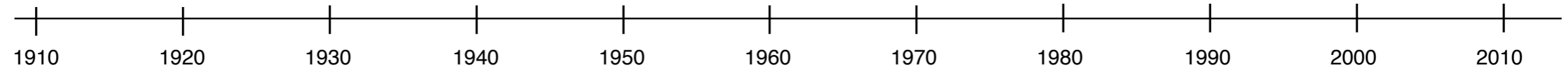
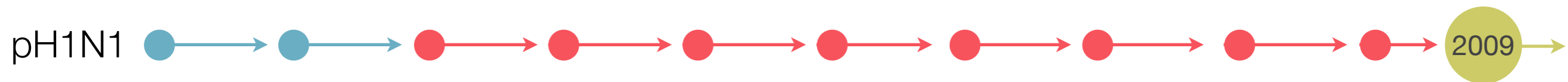
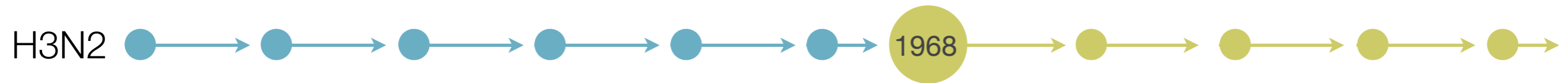
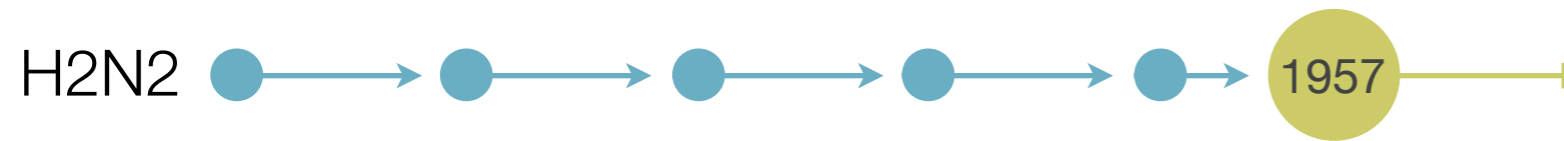
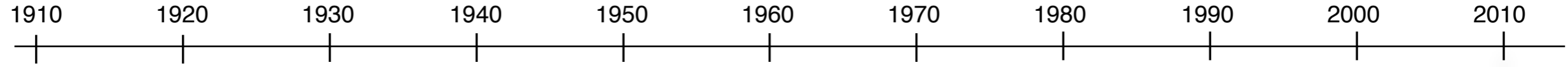
# The influenza virus



# Influenza A pandemics caused by host switch events

Avian  
Swine  
Human

Hemagglutinin (HA) lineage



# Gene flow in the influenza population

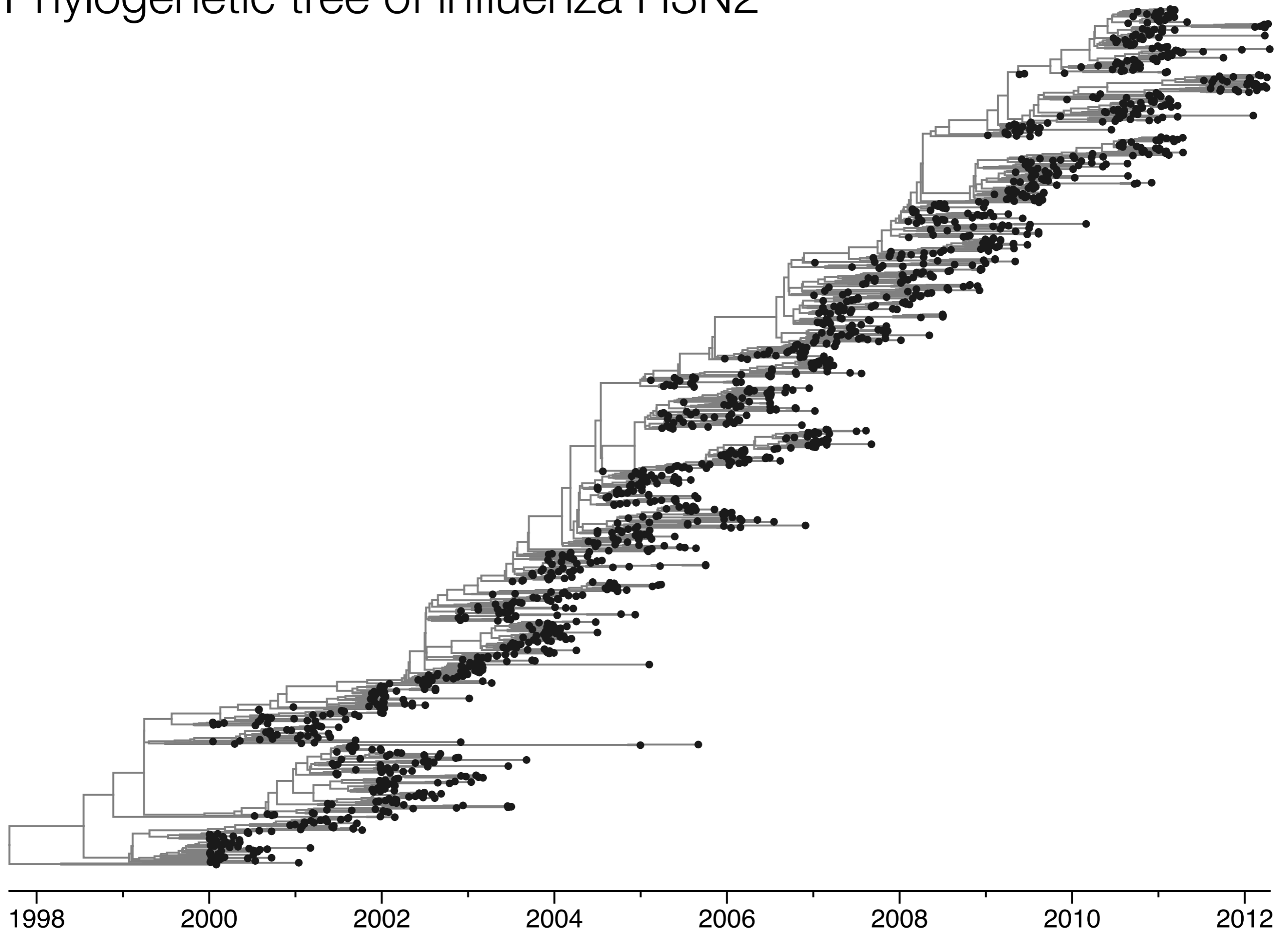
HA amino acid sequences taken from H3N2 influenza at 2 month intervals



Approximately 1 in 20 sites change over the course of 10 years

Provides a chronological record of evolution

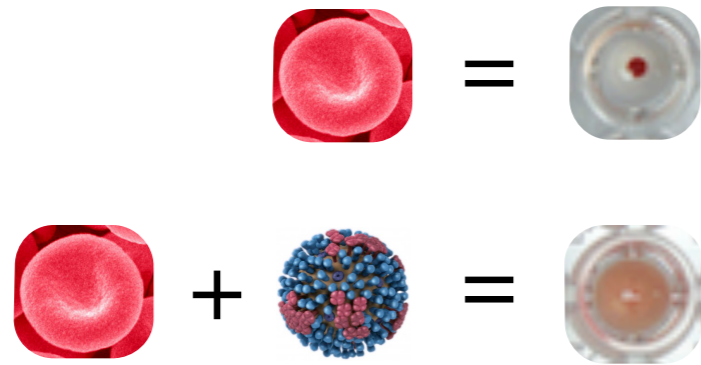
# Phylogenetic tree of influenza H3N2



# Antigenic cartography

# Influenza hemagglutination inhibition (HI) assay

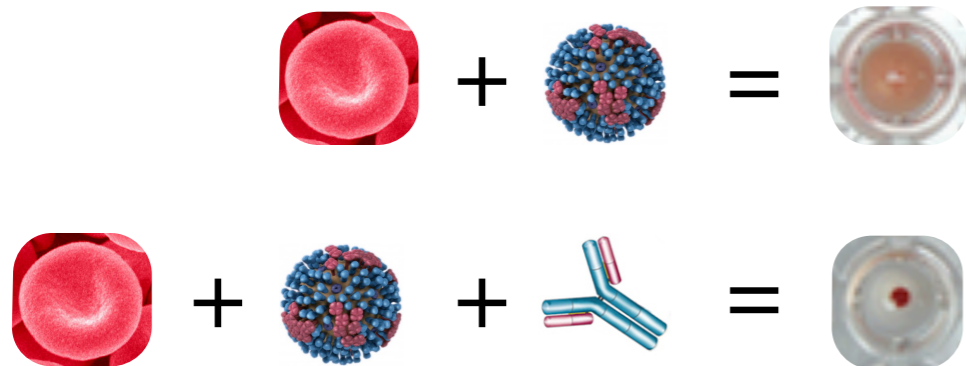
Hemagglutination assay:



Without virus, red blood cell sink to bottom of well

With virus, cells form diffuse lattice

Hemagglutination inhibition assay:

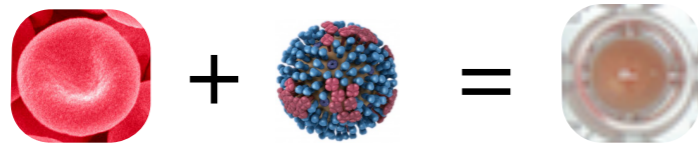


Without antibodies, agglutination of virus to RBC

Antibodies bind viruses, preventing agglutination



# Influenza hemagglutination inhibition (HI) assay

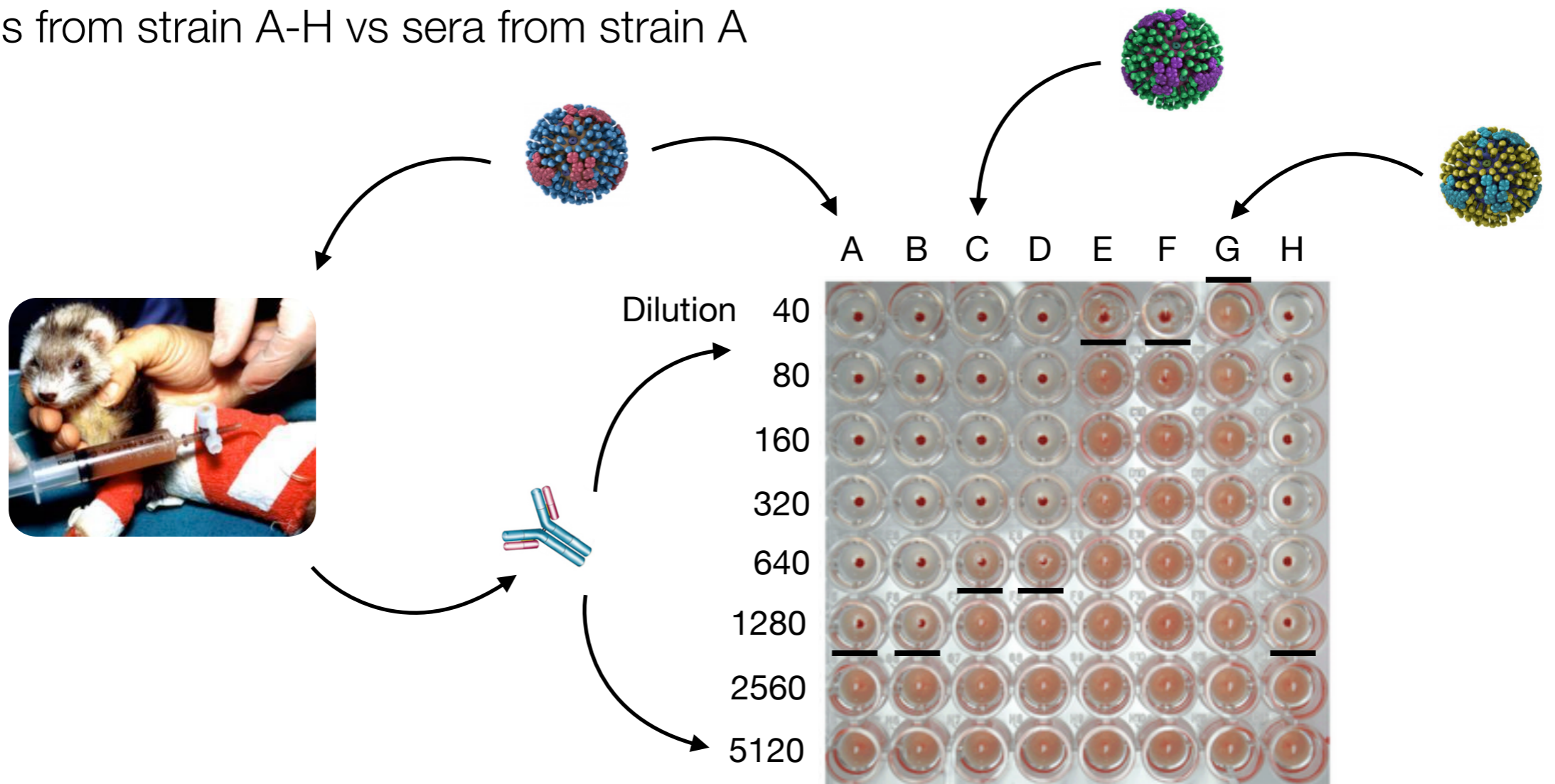


Without antibodies, agglutination of virus to RBC



Antibodies bind viruses, preventing agglutination

Reacting virus from strain A-H vs sera from strain A



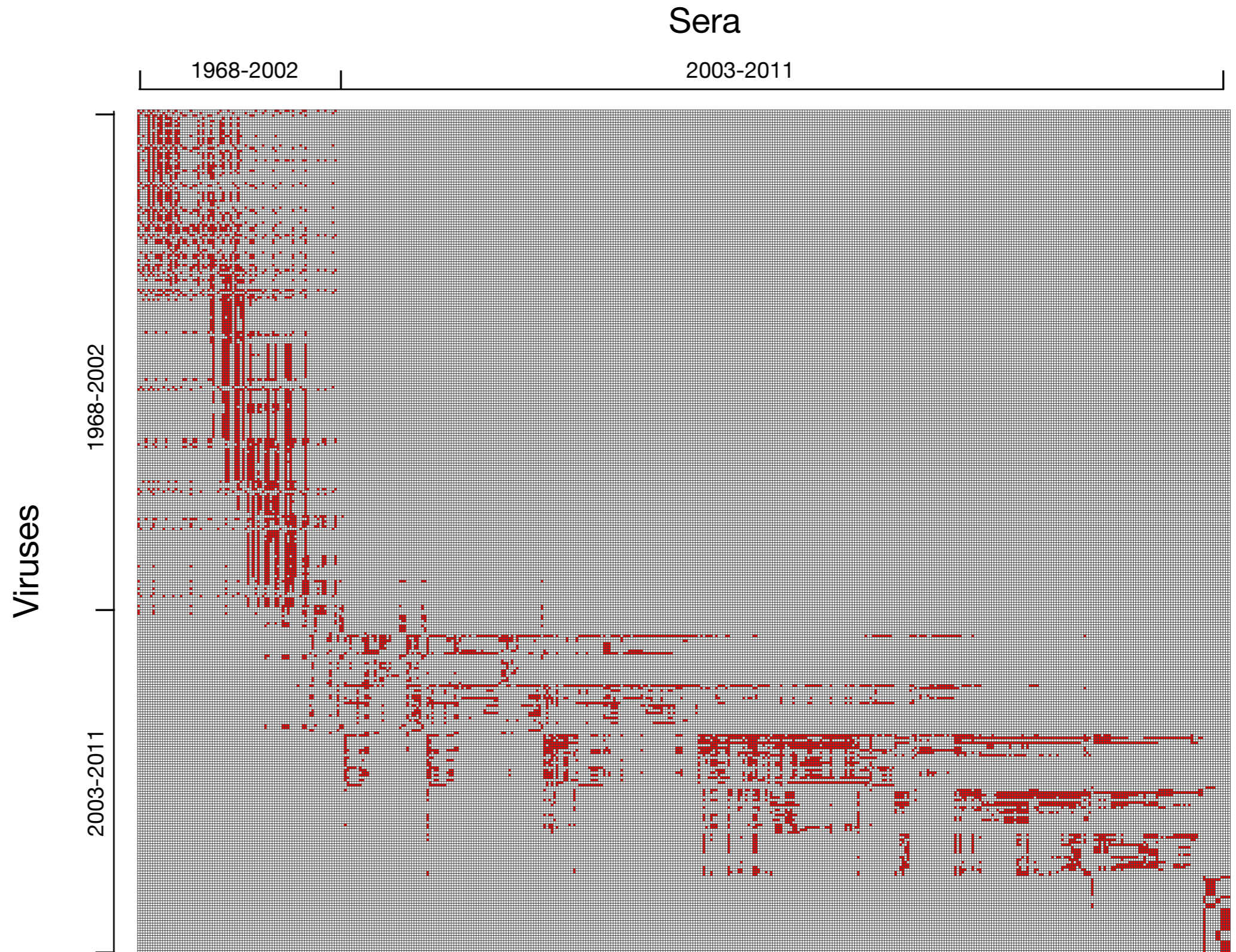
# Data in the form of table of maximum inhibitory titers

Viruses	Collection Date	Passage History	Haemagglutination inhibition titre <sup>1</sup>						
			Post infection ferret sera						
			A/Wis 67/05 F1/06	A/Bris 10/07 F29/08	A/Uru 716/07 F26/08	A/HK 1985/09 F21/09	A/Perth 16/09 F25/09	A/Wis 15/09 F24/09	A/HK 34430/09 F4/10
<b>REFERENCE VIRUSES</b>									
A/Wisconsin/67/2005	2005-08-31	SpfCk3E3/E7	1280	1280	1280	40	<	160	40
A/Brisbane/10/2007	2007-02-06	E2/E3	2560	2560	2560	80	<	160	160
A/Uruguay/716/2007	2007-06-21	SpfCk1, E3/E3	640	1280	2560	<	<	80	40
A/Hong Kong/1985/2009	2009-04-01	MDCK2/SIAT1	40	80	160	1280	640	2560	1280
A/Perth/16/2009	2009-07-04	E3/E2	<	<	40	640	640	640	640
A/Wisconsin/15/2009	2009-07-06	E2/E3	<	<	40	640	640	1280	1280
A/Hong Kong/34430/2009	2009-11-22	MDCK2/SIAT2	<	80	160	5120	640	1280	1280
<b>TEST VIRUSES</b>									
A/Hong Kong/1737/2010	2010-03-24	MDCK2/SIAT1	40	80	320	5120	1280	1280	1280
A/Hong Kong/1775/2010	2010-03-28	MDCK2/SIAT1	<	80	160	5120	640	2560	1280
A/Hong Kong/1837/2010	2010-03-30	MDCK2/SIAT1	40	80	160	5120	640	2560	1280
A/Hong Kong/1888/2010	2010-04-19	MDCK2/SIAT1	160	320	320	5120	1280	2560	2560

# Collect data from many HI assays

Using 26,923 HI measurements

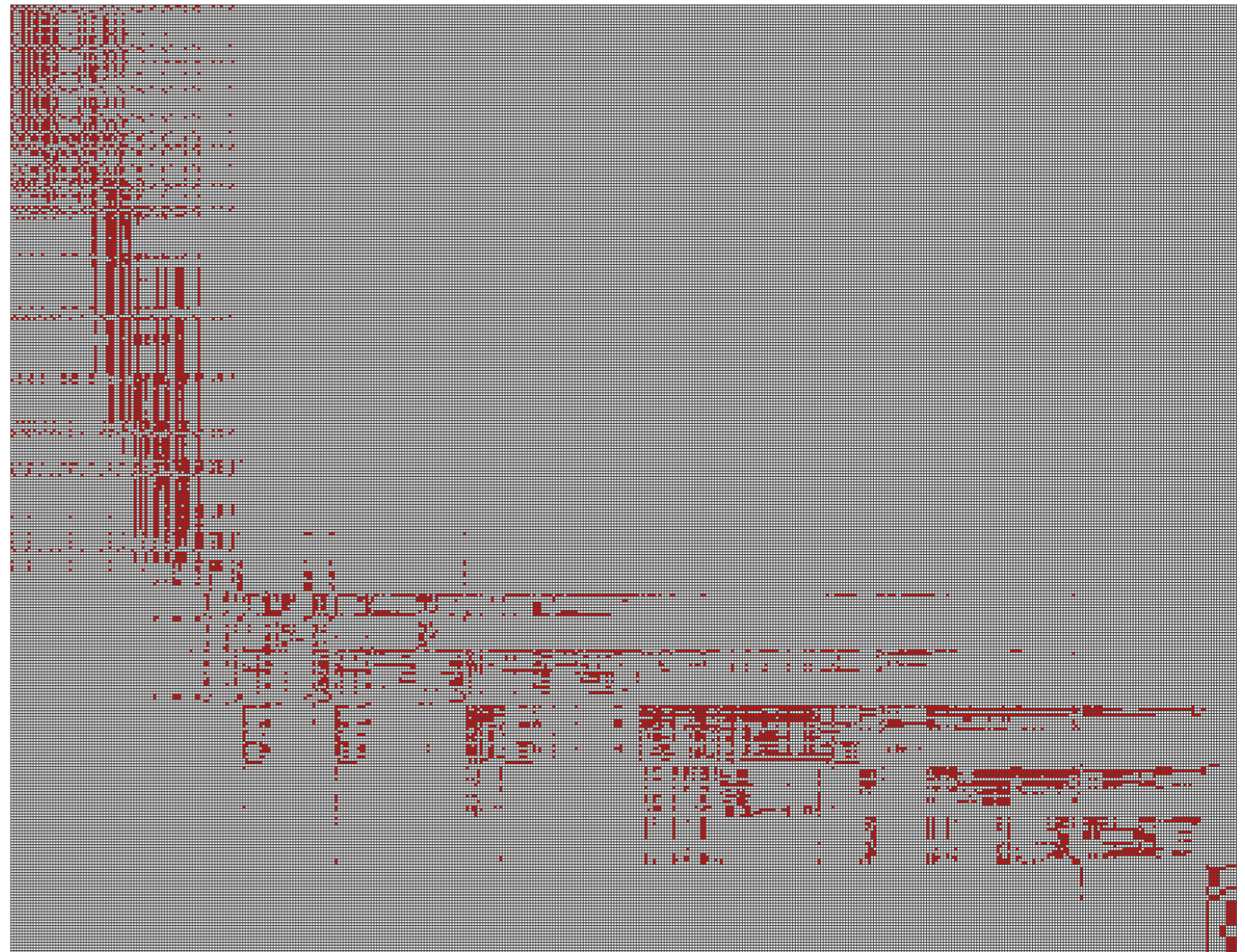
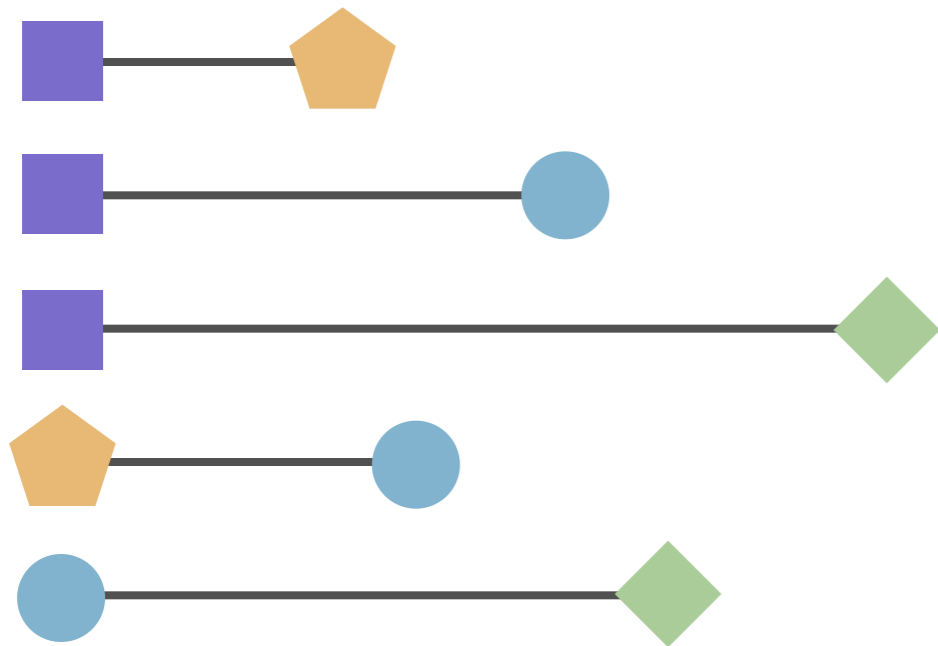
Difficulties: sparse, censored, noisy, high-dimensional



# Antigenic cartography

Developed by Derek Smith and colleagues

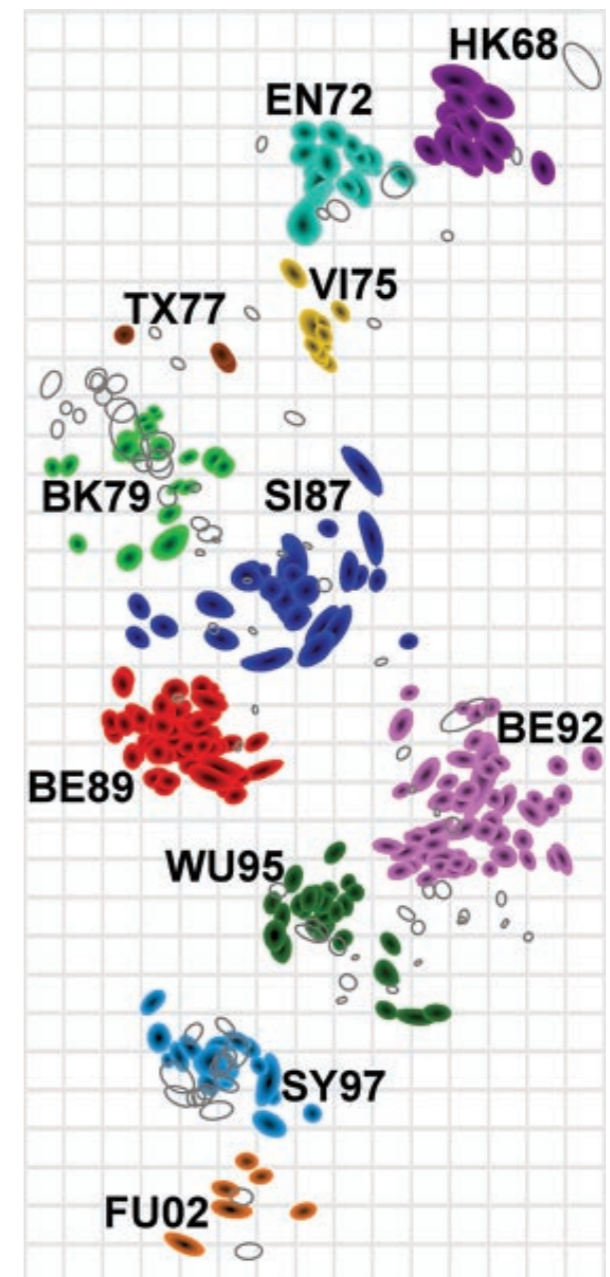
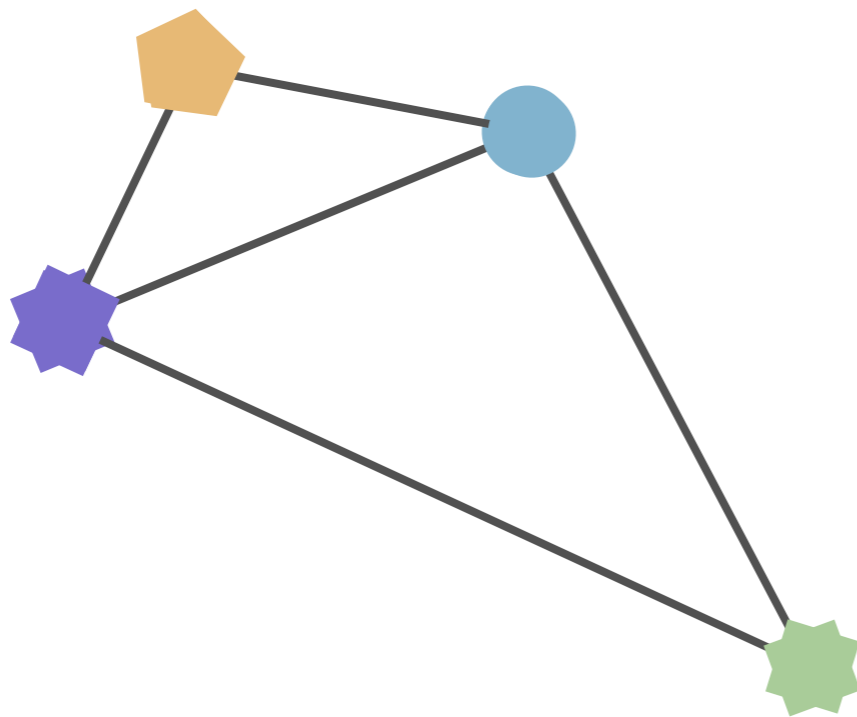
Uses multidimensional scaling (MDS) to position viruses in 2D space such that the distances in this space best fit the HI assay titers.



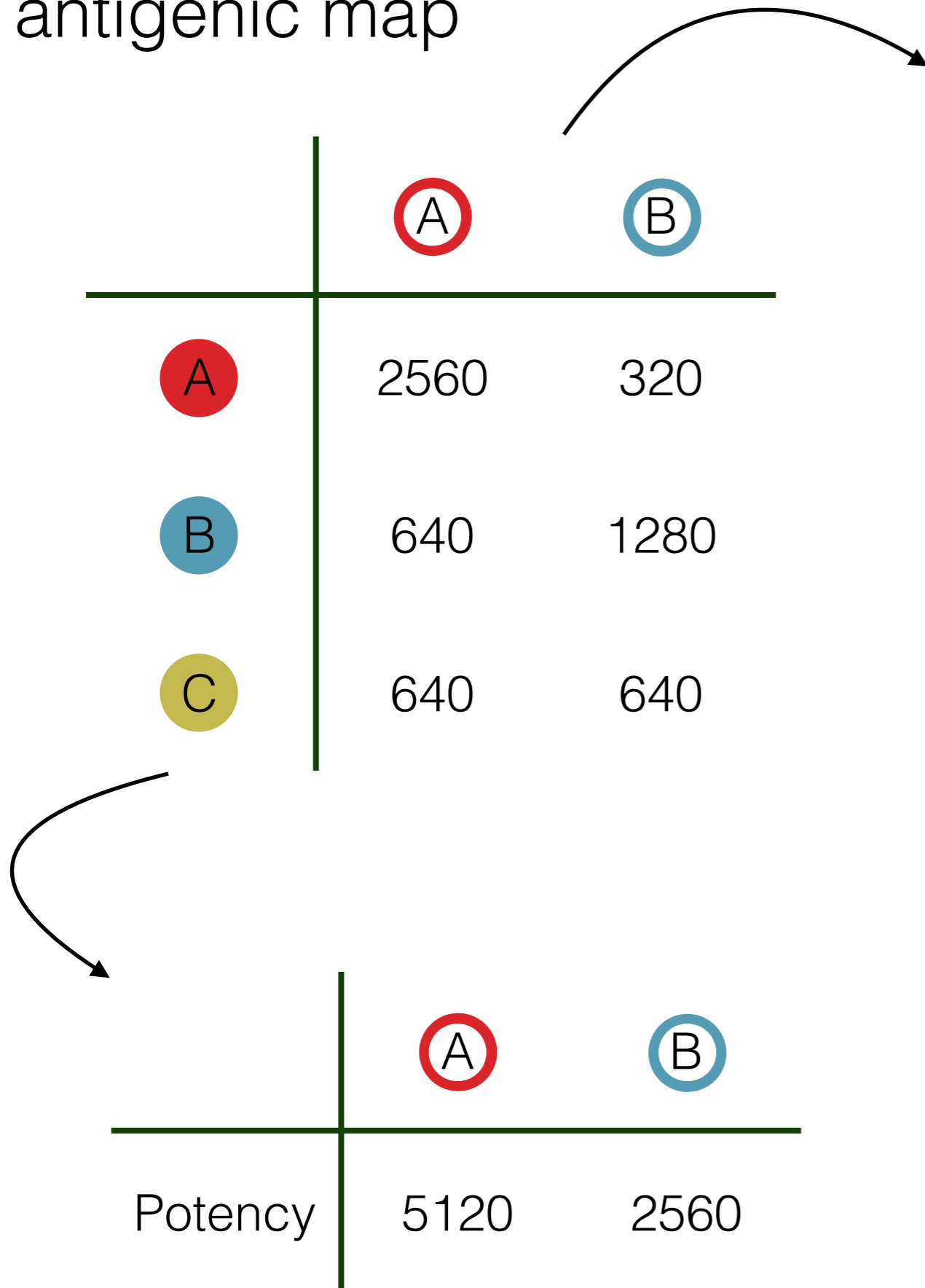
# Antigenic cartography

Developed by Derek Smith and colleagues

Uses multidimensional scaling (MDS) to position viruses in 2D space such that the distances in this space best fit the HI assay titers.



# Schematic HI table and antigenic map

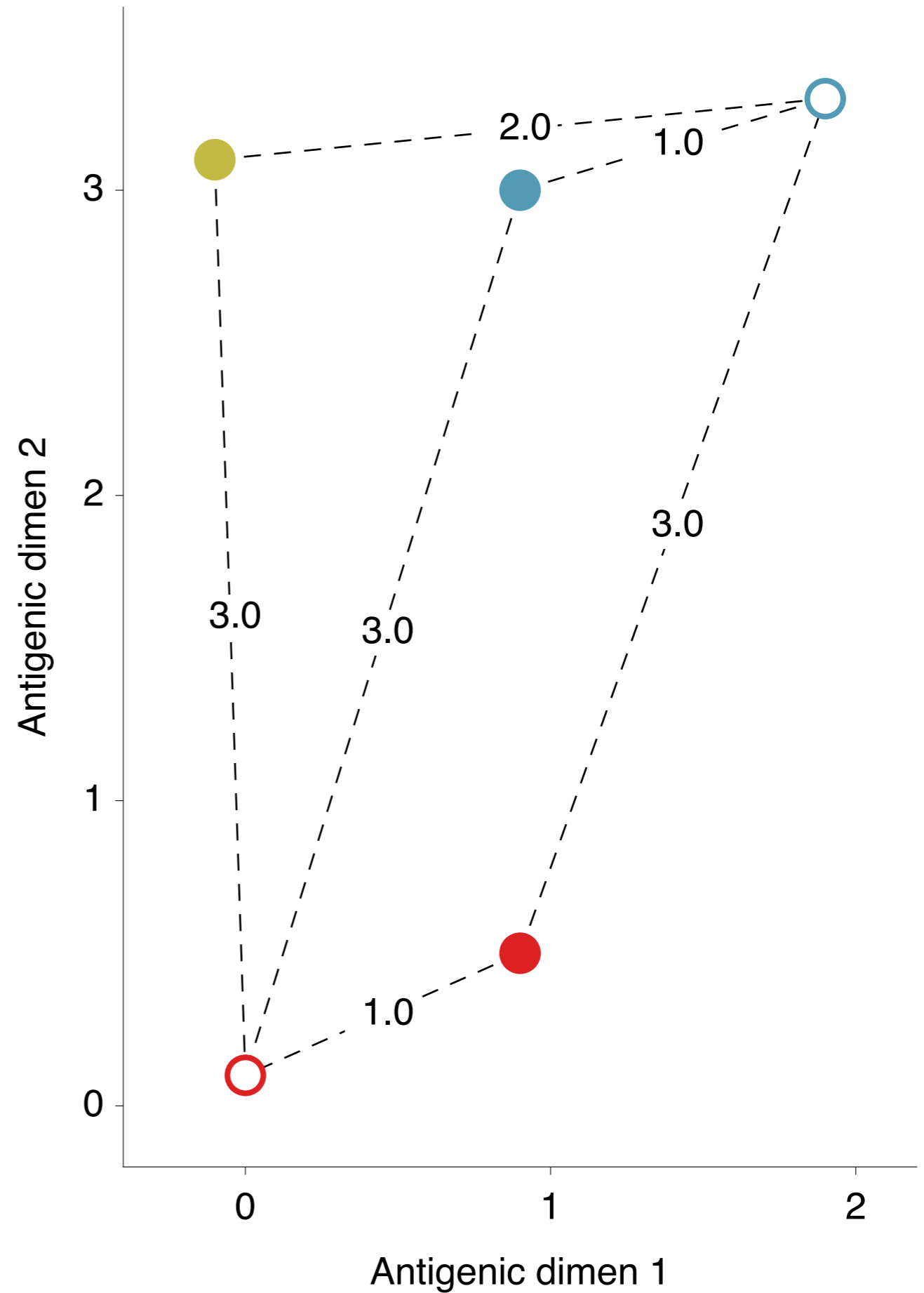


The diagram shows a schematic HI table and an antigenic map. The HI table is a 3x2 grid with rows labeled A, B, and C, and columns labeled A and B. The antigenic map is a 2D plot with axes labeled Antigenic dimen 1 and Antigenic dimen 2. The HI table is connected to the antigenic map by two curved arrows: one from the top-right cell (A, B) to the antigenic map, and another from the bottom-left cell (C, A) to the HI table.

	A	B
A	2560	320
B	640	1280
C	640	640

	A	B
Potency	5120	2560



# Bayesian multidimensional scaling (BMDS)

Titer between virus  $i$  and serum  $j$

$$H_{ij}$$

Maximum titer for serum  $j$ , i.e. serum 'potency'

$$S_j$$

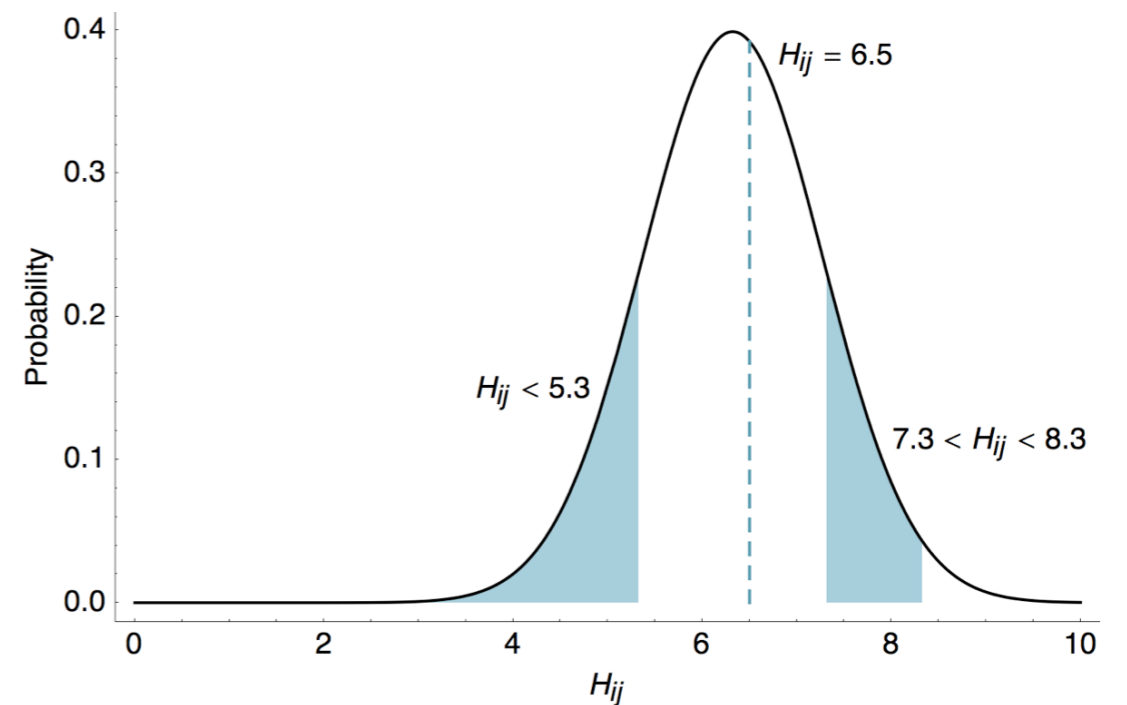
Map distance between virus  $i$  at  $X_i$  and serum  $j$  at  $Y_j$

$$\delta_{ij} = \|X_i - Y_j\|_2$$

Probability of observing data

$$H_{ij} \sim \text{Normal}(S_j - \delta_{ij}, \sigma^2)$$

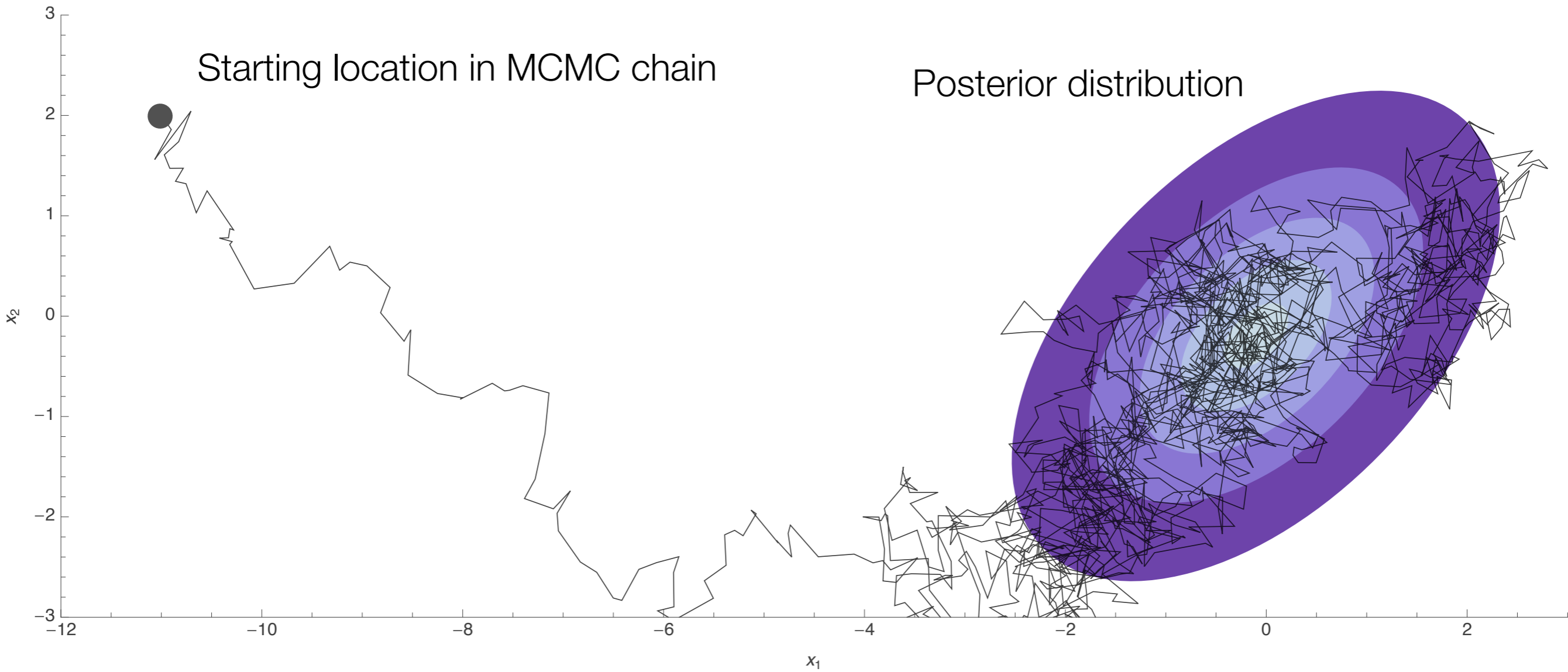
For example, given  $\delta_{ij} = 4$   
and  $S_j = 10.3$



Likelihood of observing HI data

$$L(\mathbf{X}, \mathbf{Y}) = \prod_{(i,j) \in \mathcal{I}} f(H_{ij})$$

# Integration through Markov chain Monte Carlo (MCMC)



BEAST: Bayesian Evolutionary Analysis by Sampling Trees



# Bayesian MDS results

# Predicting HI measurements

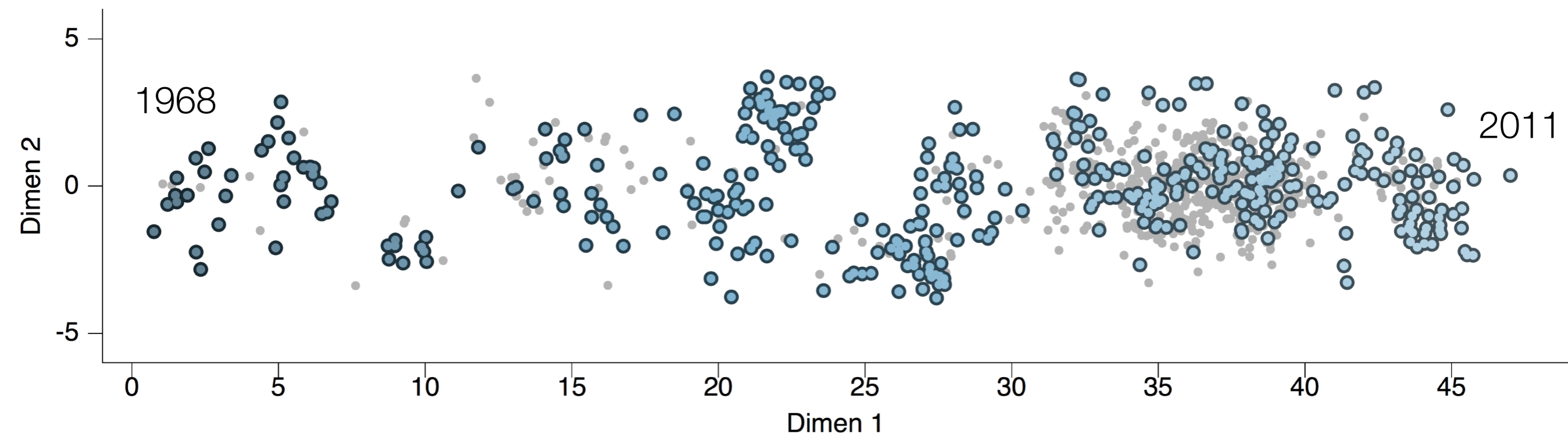
Training dataset: 6545 measurements

Test dataset: 723 measurements

Errors are average absolute prediction errors for  $\log_2$  HI titers

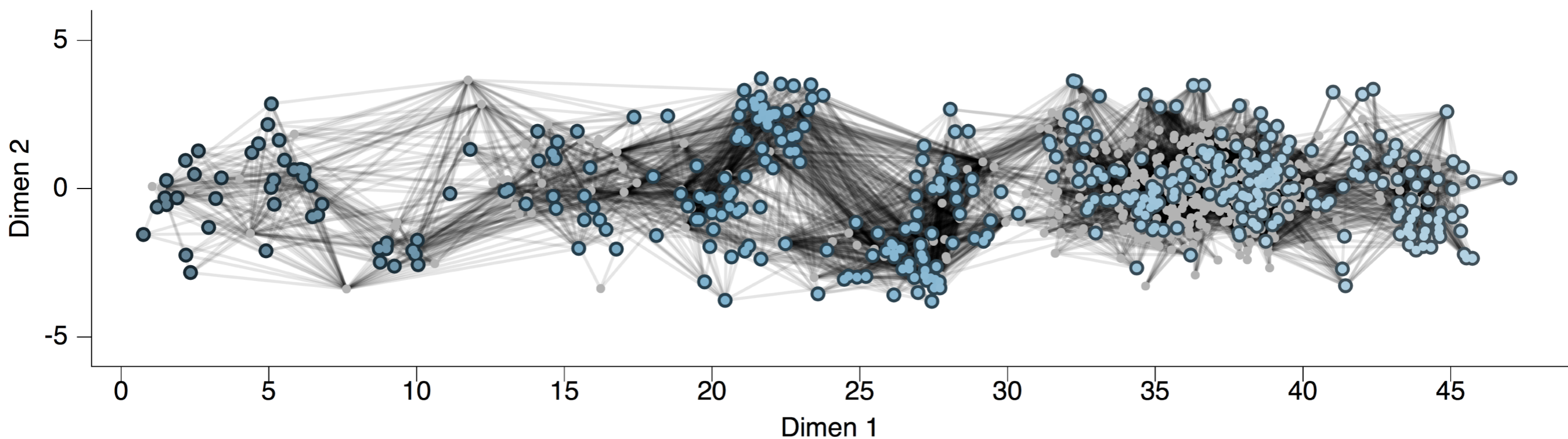
<b>Dimen</b>	<b>Serum effects</b>	<b>Virus effects</b>	<b>Test error</b>
1D	Fixed	None	1.03
2D	Fixed	None	0.86
3D	Fixed	None	0.88
4D	Fixed	None	0.96
5D	Fixed	None	1.06
2D	Estimated	None	0.77
2D	Estimated	Estimated	0.75

# Antigenic map of H3N2 influenza from 1968 to 2011

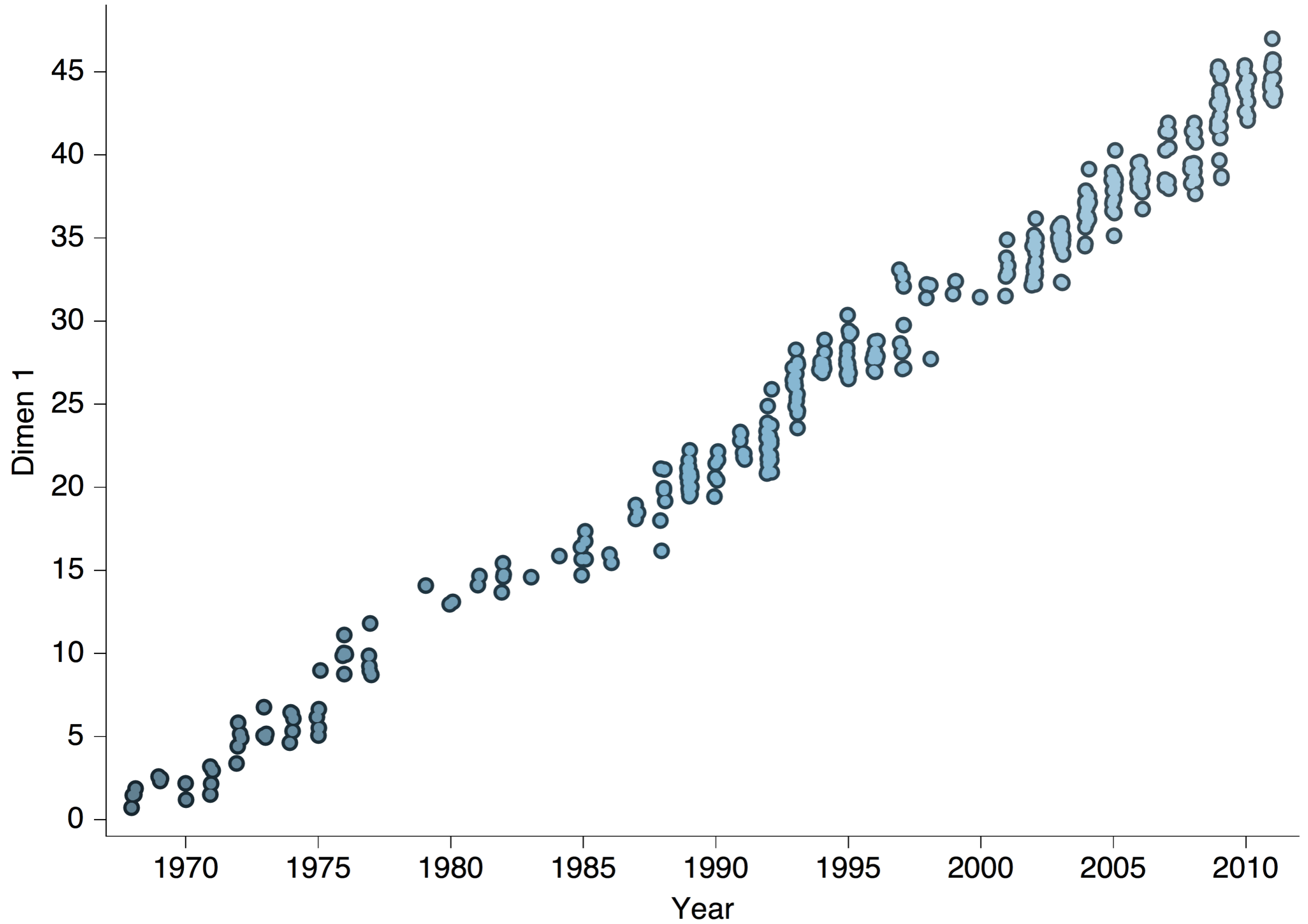


# Antigenic map of H3N2 influenza from 1968 to 2011

Local HI measurements are used to construct the global map



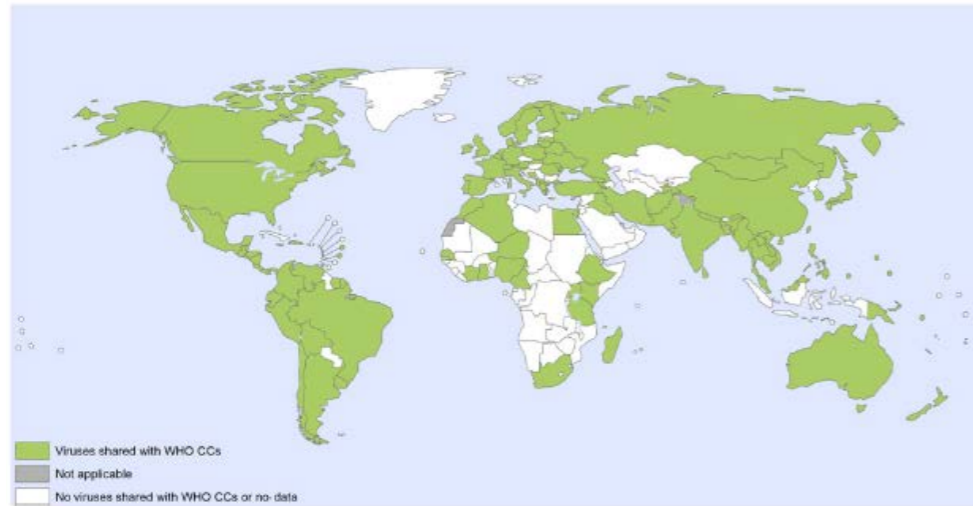
# Antigenic drift of H3N2 influenza



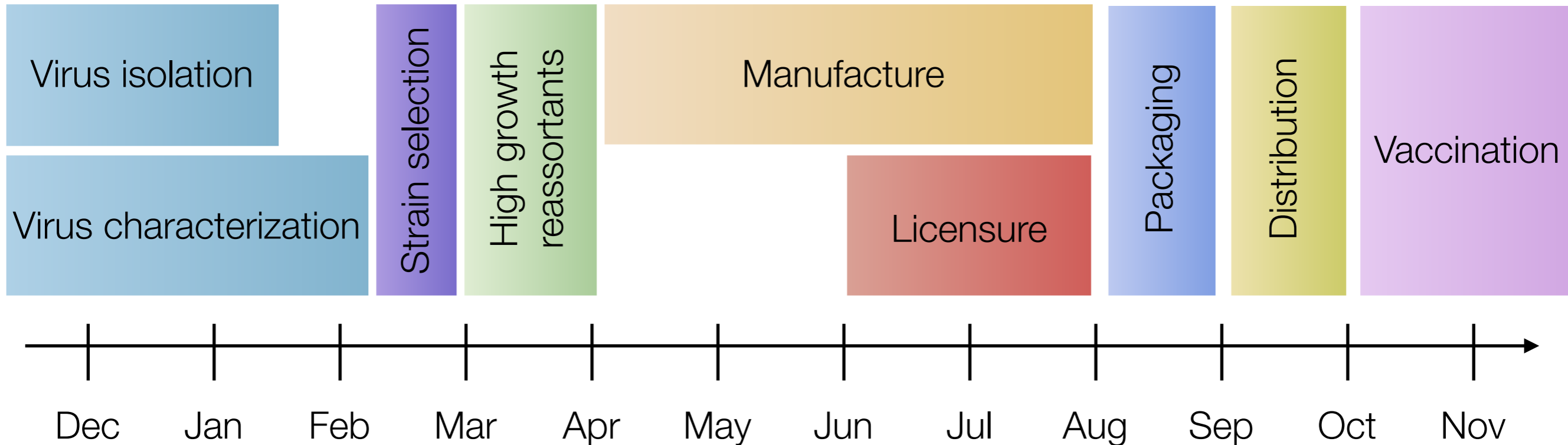


# Vaccine strain selection timeline

Collection by WHO National Influenza Centres



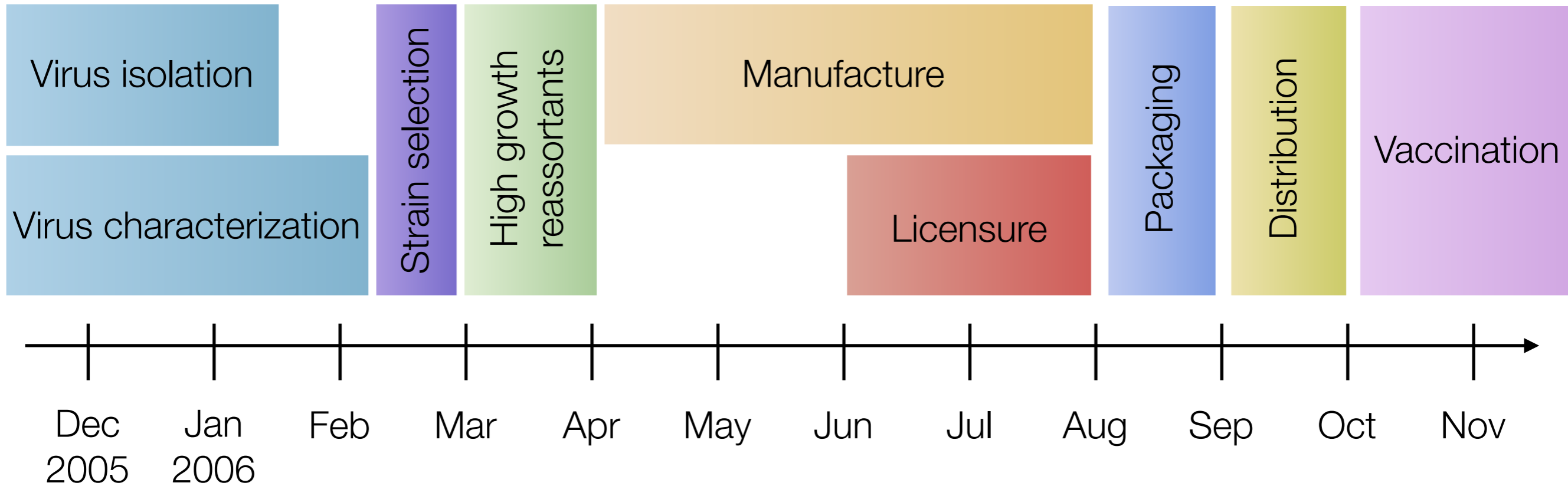
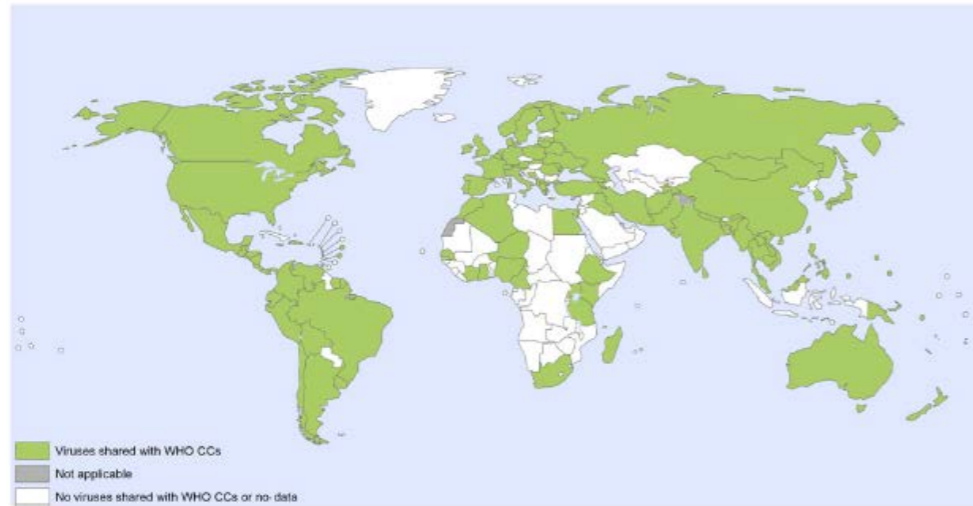
Characterization by WHO Collaborating Centres



# Vaccine strain selection timeline for 06-07 season

Collection by WHO National Influenza Centres

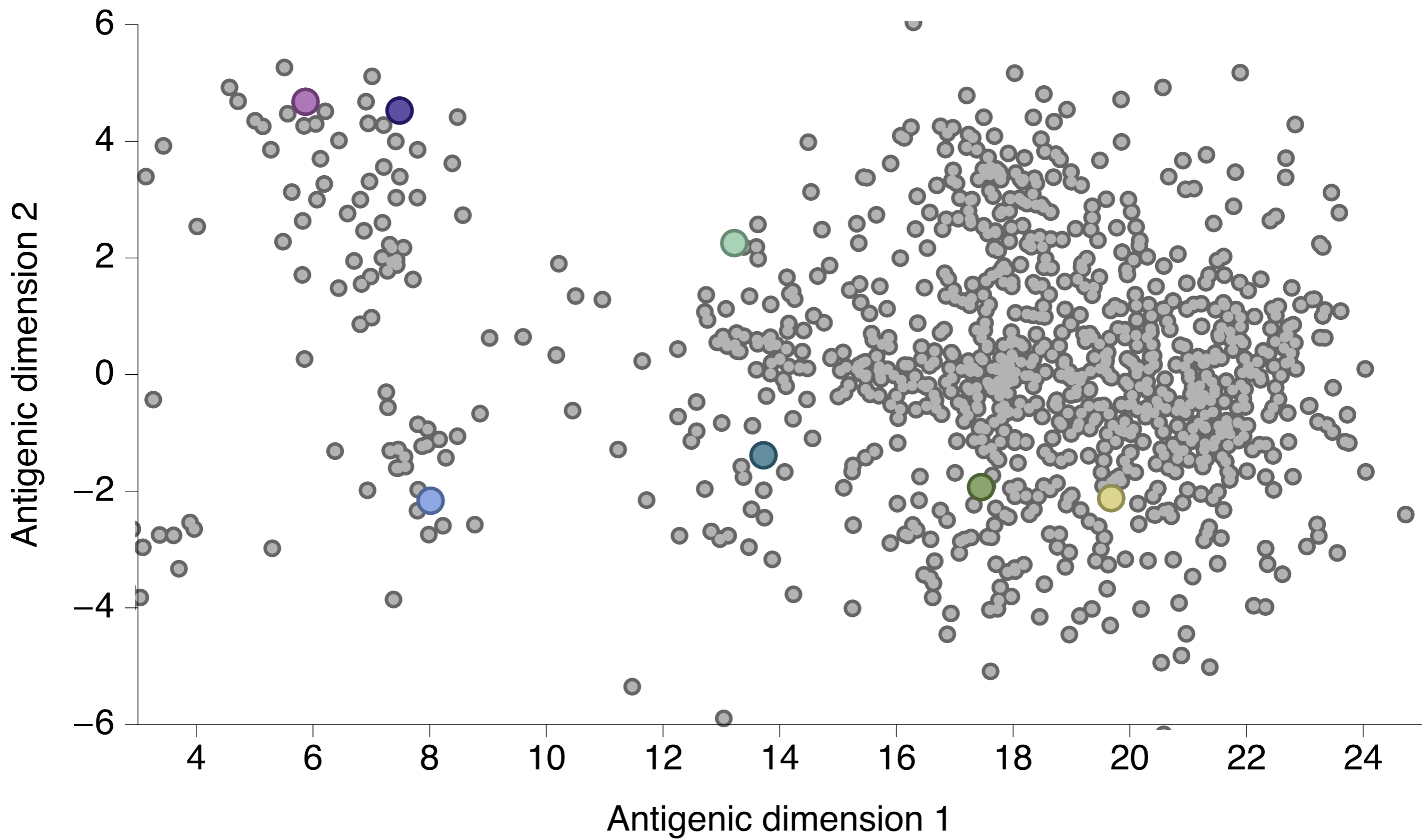
Characterization by WHO Collaborating Centres





# Antigenic map for viruses up to Dec 2005 for Feb 2006 meeting

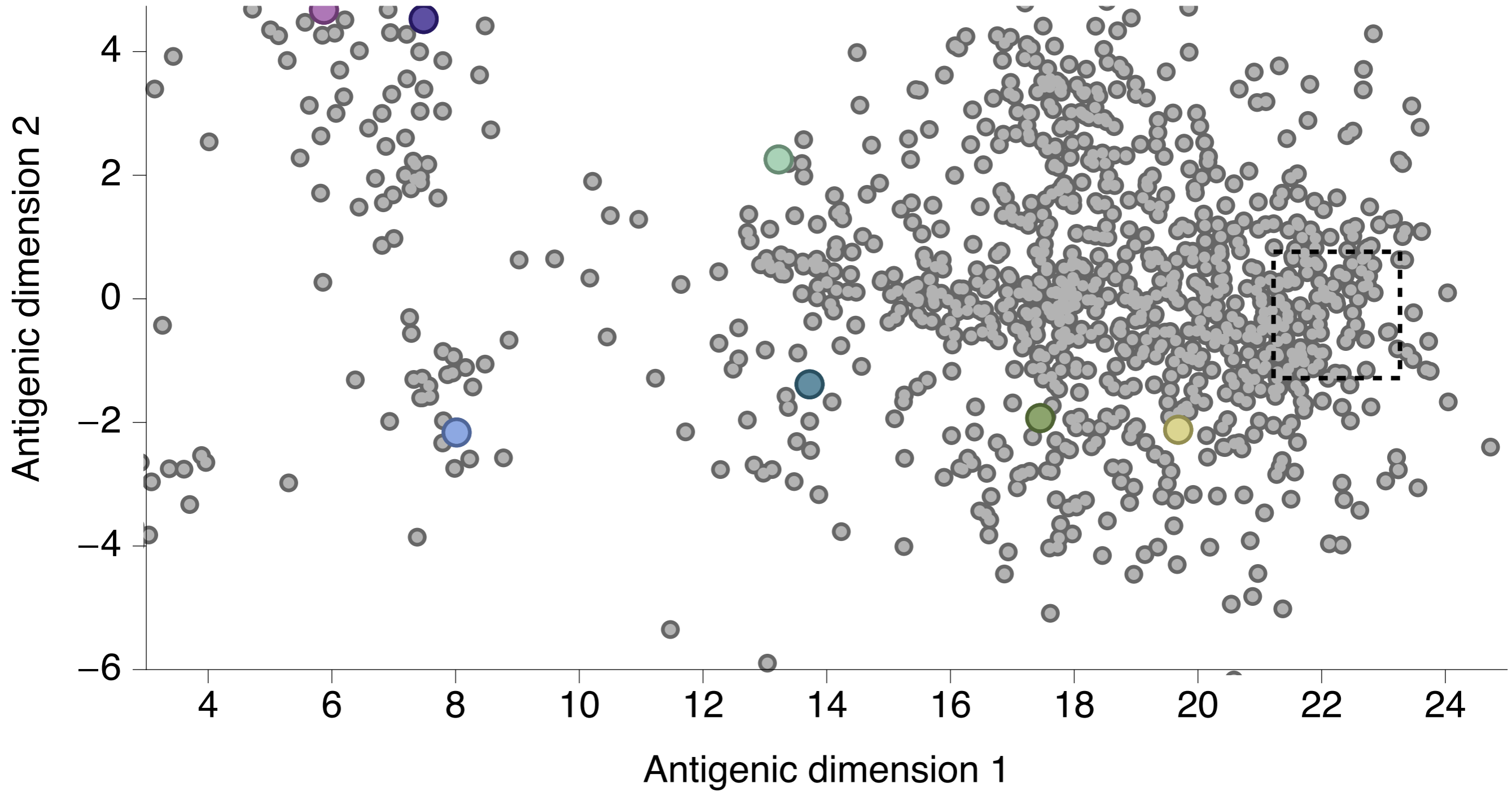
- Shandong/9/1993
- Johannesburg/33/1994
- Wuhan/359/1995
- Sydney/5/1997
- Moscow/10/1999
- Fujian/411/2002
- California/7/2004



# Antigenic map for viruses up to Dec 2005 for Feb 2006 meeting

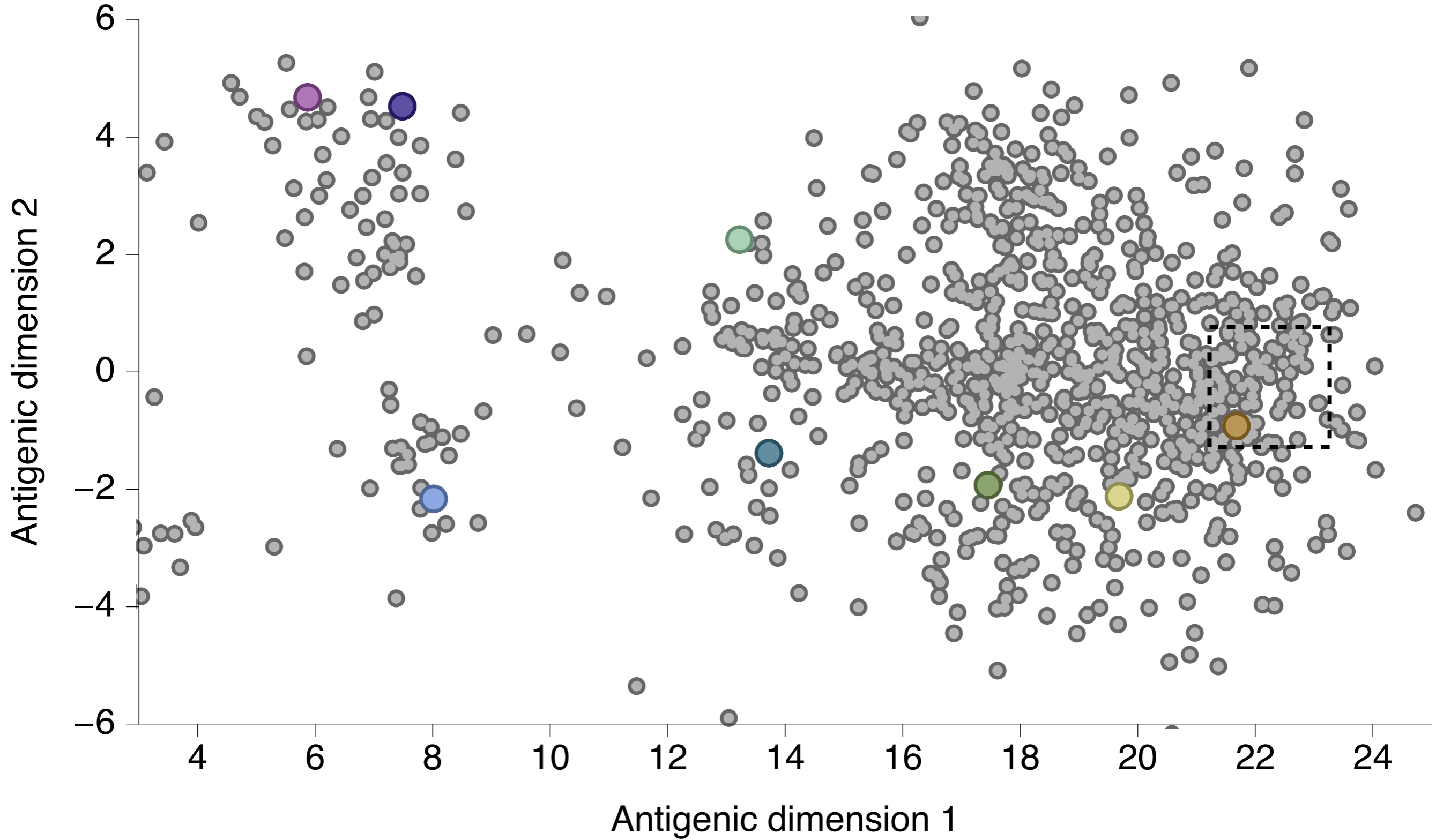
- Shandong/9/1993
- Johannesburg/33/1994
- Wuhan/359/1995
- Sydney/5/1997
- Moscow/10/1999
- Fujian/411/2002
- California/7/2004

A four-fold difference in HI (2 units on the antigenic map) is generally considered sufficient to warrant a vaccine strain update



# Antigenic map for viruses up to Dec 2005 for Feb 2006 meeting

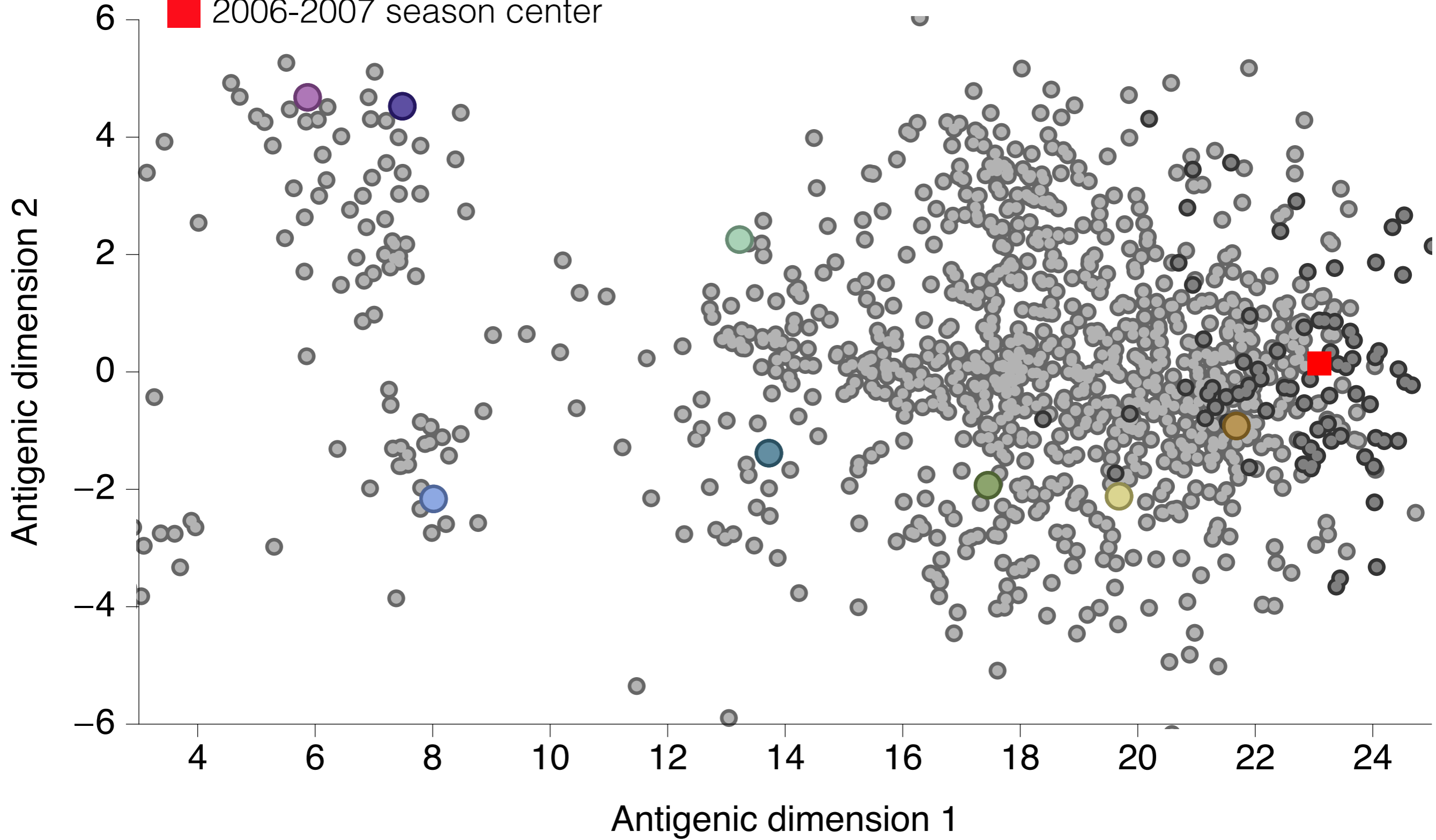
- Shandong/9/1993
- Johannesburg/33/1994
- Wuhan/359/1995
- Sydney/5/1997
- Moscow/10/1999
- Fujian/411/2002
- California/7/2004
- Wisconsin/67/2005



# Antigenic map for viruses up to March 2007

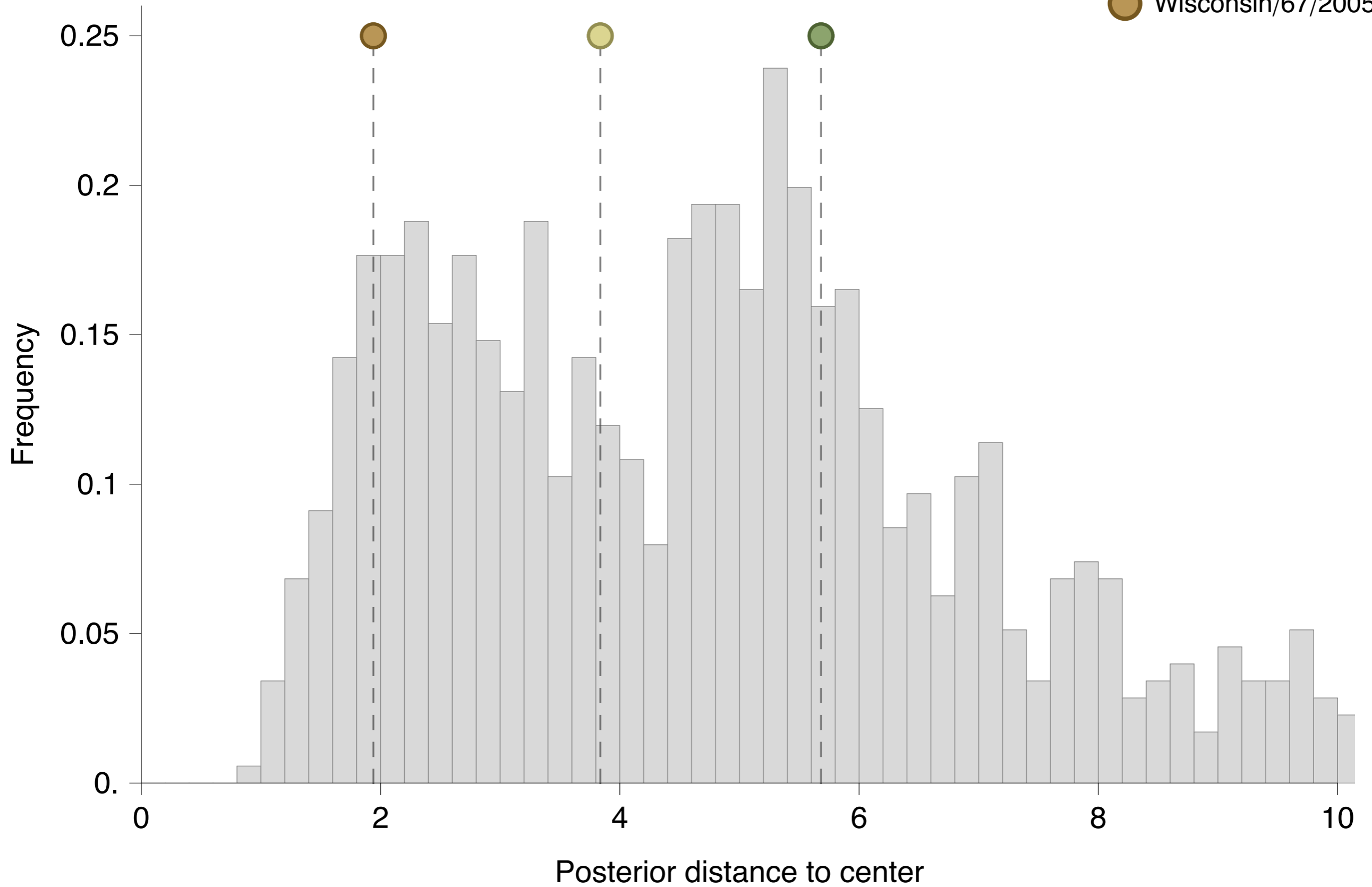
- Shandong/9/1993
- Johannesburg/33/1994
- Wuhan/359/1995
- Sydney/5/1997
- Moscow/10/1999
- Fujian/411/2002
- California/7/2004
- Wisconsin/67/2005

- 2006-2007 season viruses
- 2006-2007 season center



# Ranking of virus posterior distances

- Moscow/10/1999
- Fujian/411/2002
- California/7/2004
- Wisconsin/67/2005

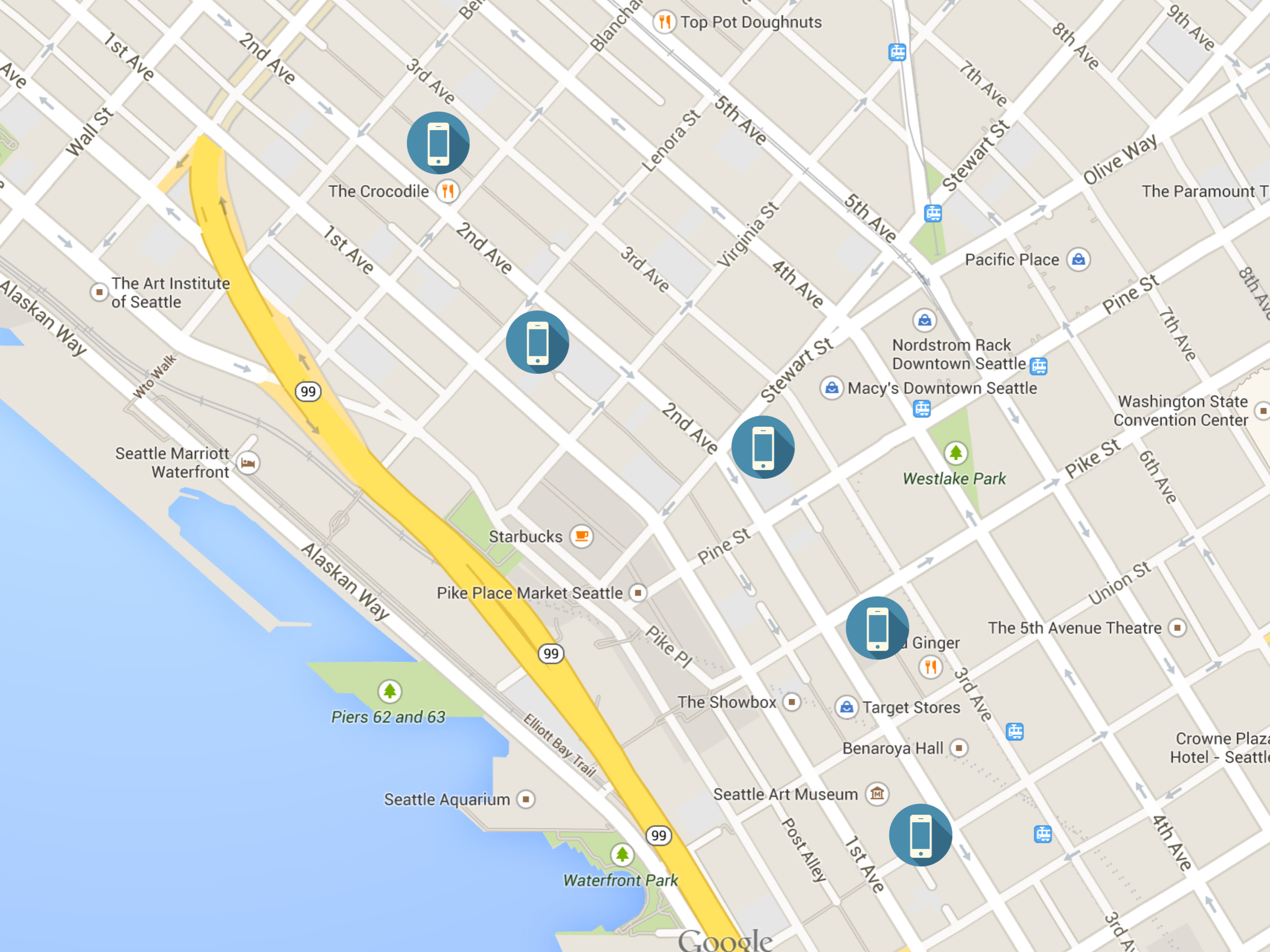


# Kalman filters









Top Pot Doughnuts



The Crocodile



The Art Institute of Seattle

Seattle Marriott Waterfront

99

99

99

Piers 62 and 63

Pike Place Market Seattle

Starbucks

The Showbox

Target Stores

Benaroya Hall

Seattle Art Museum

Seattle Aquarium

Waterfront Park

Westlake Park

The 5th Avenue Theatre

1st Ginger

Crowne Plaza Hotel - Seattle

Washington State Convention Center

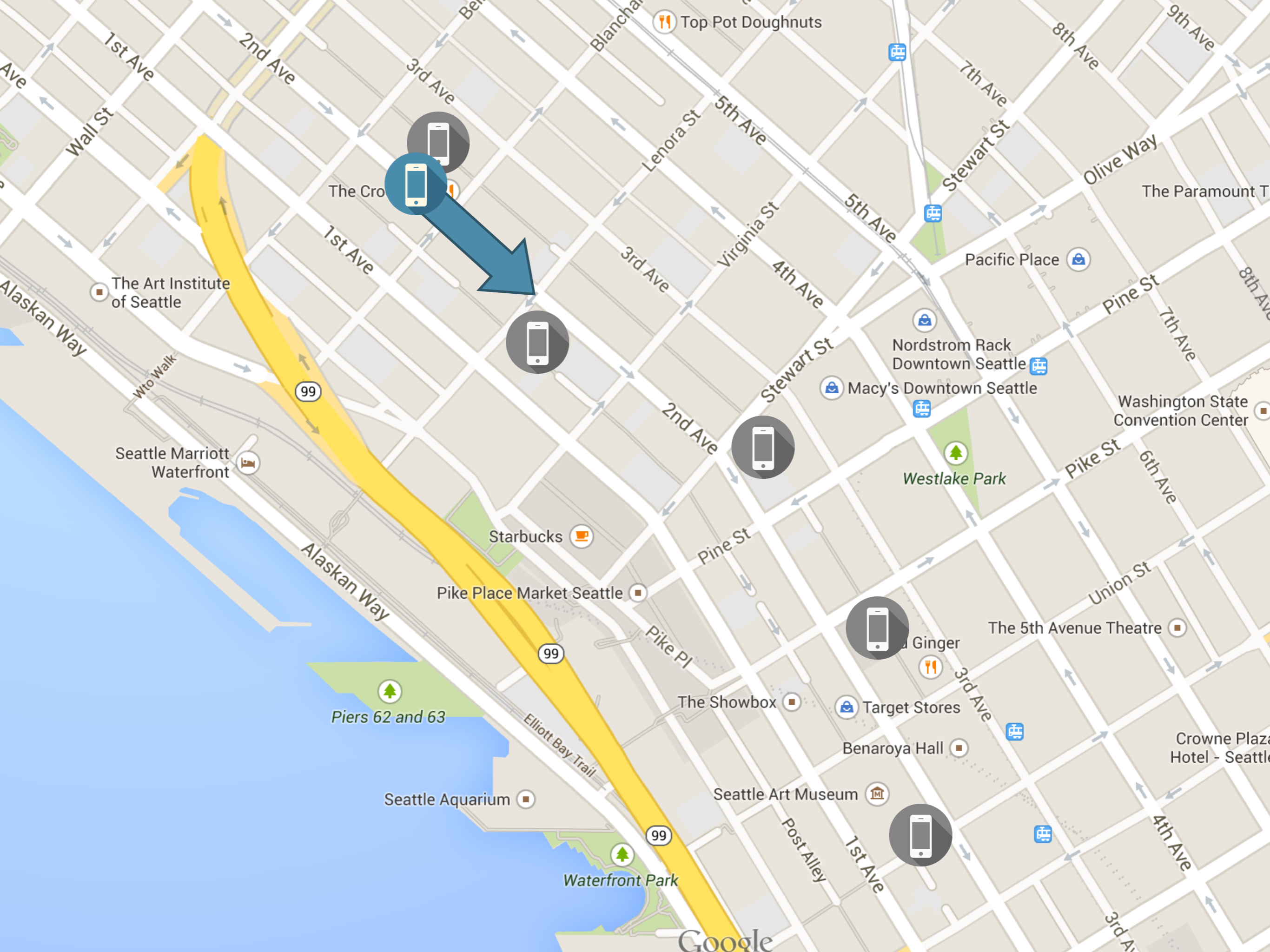
Macy's Downtown Seattle

Nordstrom Rack Downtown Seattle

Pacific Place

The Paramount Theatre

Google



Top Pot Doughnuts

The Art Institute of Seattle

The Cro

Seattle Marriott Waterfront

Starbucks

Pike Place Market Seattle

Piers 62 and 63

Seattle Aquarium

Waterfront Park

The Showbox

Seattle Art Museum

Google

Target Stores

Benaroya Hall

Ginger

The 5th Avenue Theatre

Crowne Plaza Hotel - Seattle

Westlake Park

Nordstrom Rack Downtown Seattle

Macy's Downtown Seattle

Pacific Place

Washington State Convention Center

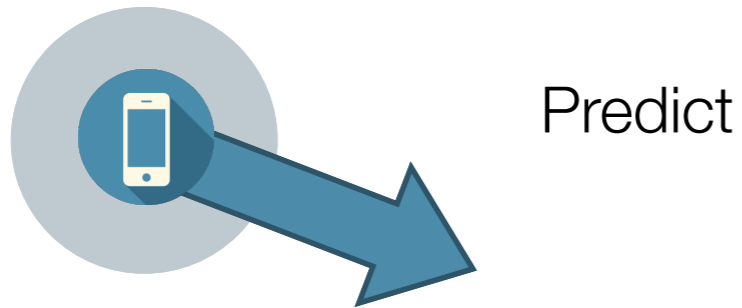
The Paramount T

# Kalman filter algorithm

Keep estimates of system state (location and velocity vectors) and associated precisions.

Predict step: Use current state of the system at time  $t$  to predict state at time  $t+1$

Correct step: Use measurement of state to correction location

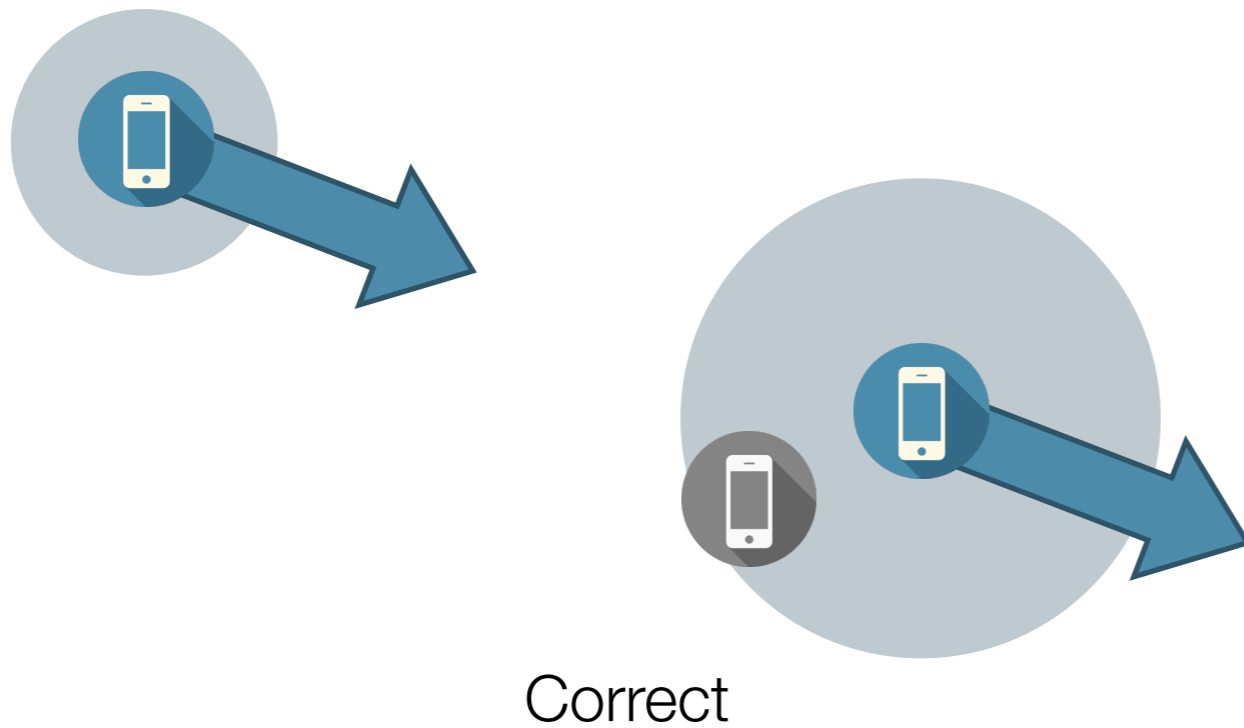


# Kalman filter algorithm

Keep estimates of system state (location and velocity vectors) and associated precisions.

Predict step: Use current state of the system at time  $t$  to predict state at time  $t+1$

Correct step: Use measurement of state to correction location

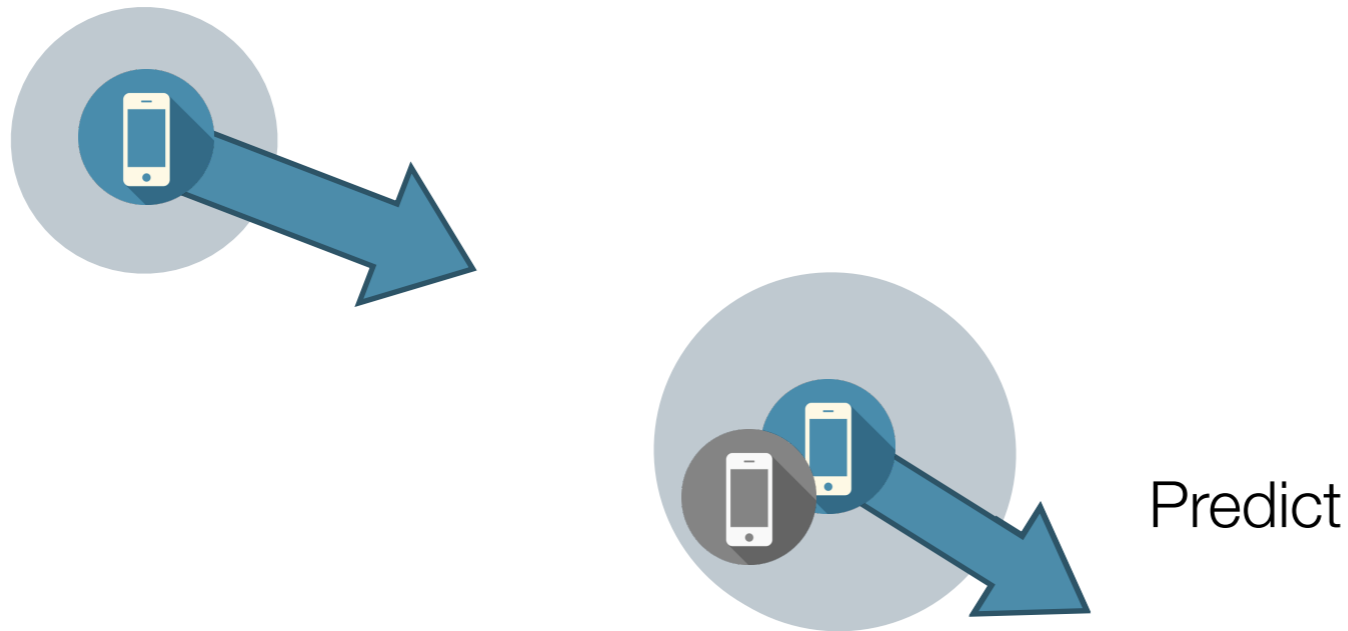


# Kalman filter algorithm

Keep estimates of system state (location and velocity vectors) and associated precisions.

Predict step: Use current state of the system at time  $t$  to predict state at time  $t+1$

Correct step: Use measurement of state to correction location

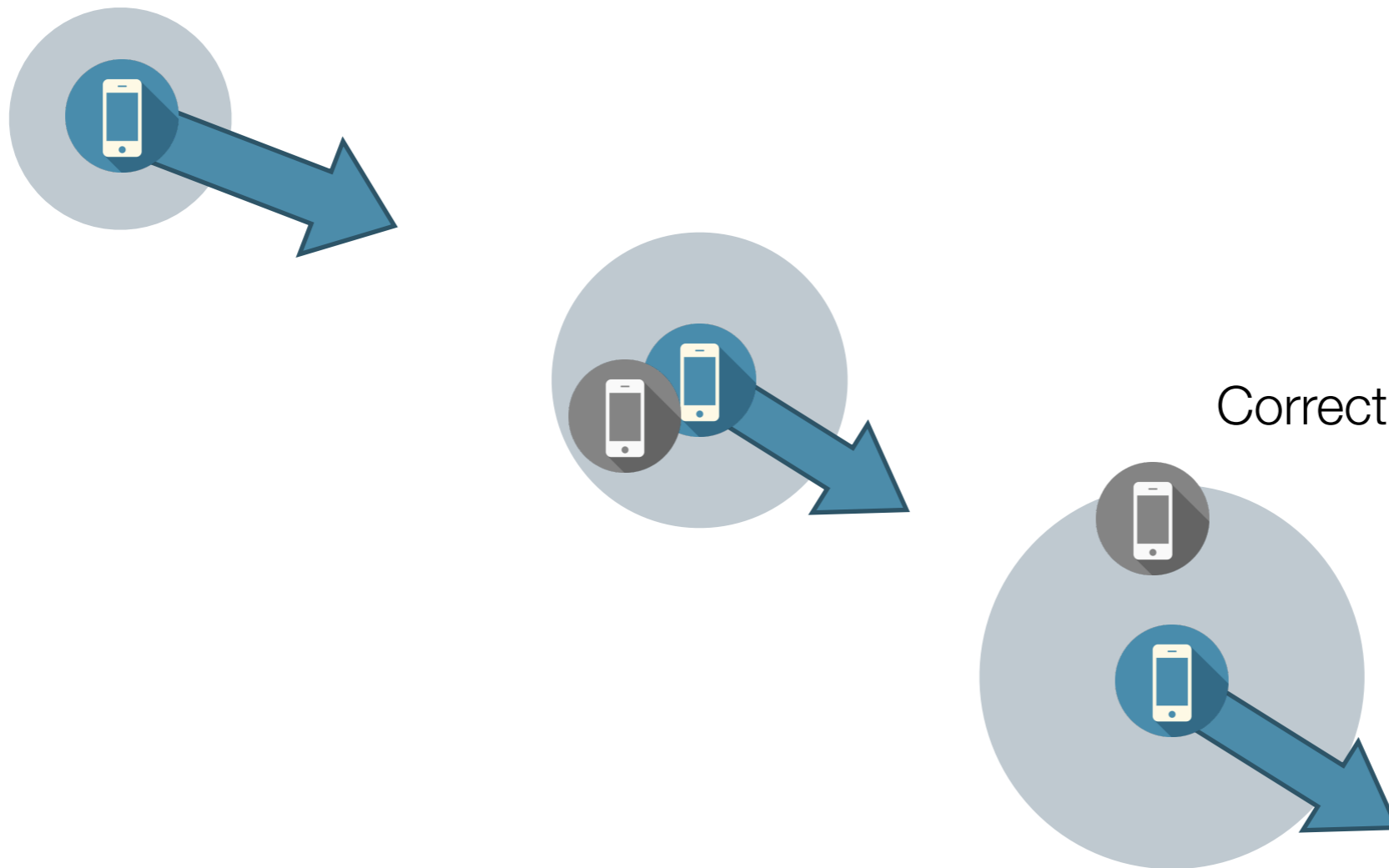


# Kalman filter algorithm

Keep estimates of system state (location and velocity vectors) and associated precisions.

Predict step: Use current state of the system at time  $t$  to predict state at time  $t+1$

Correct step: Use measurement of state to correction location

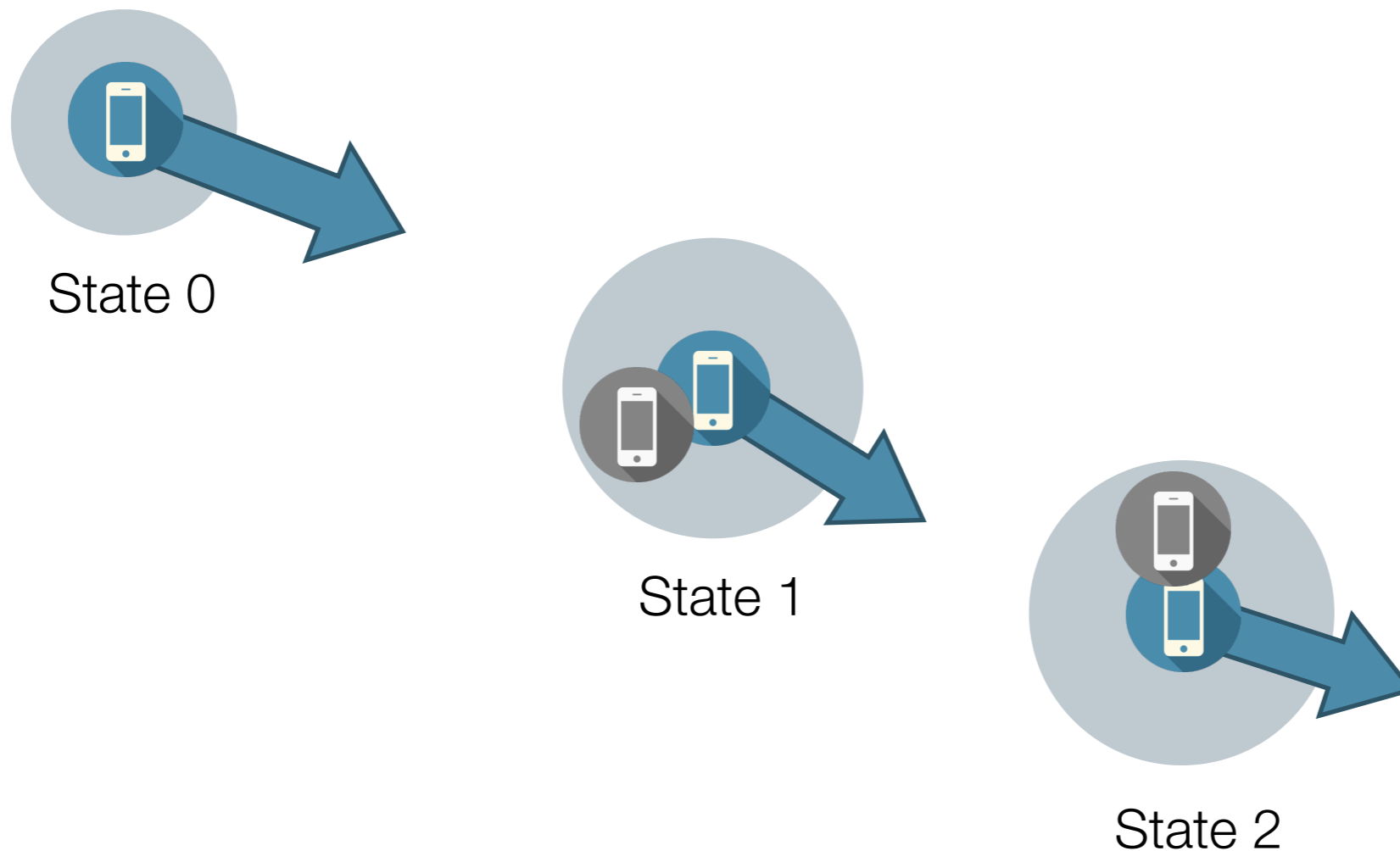


# Kalman filter algorithm

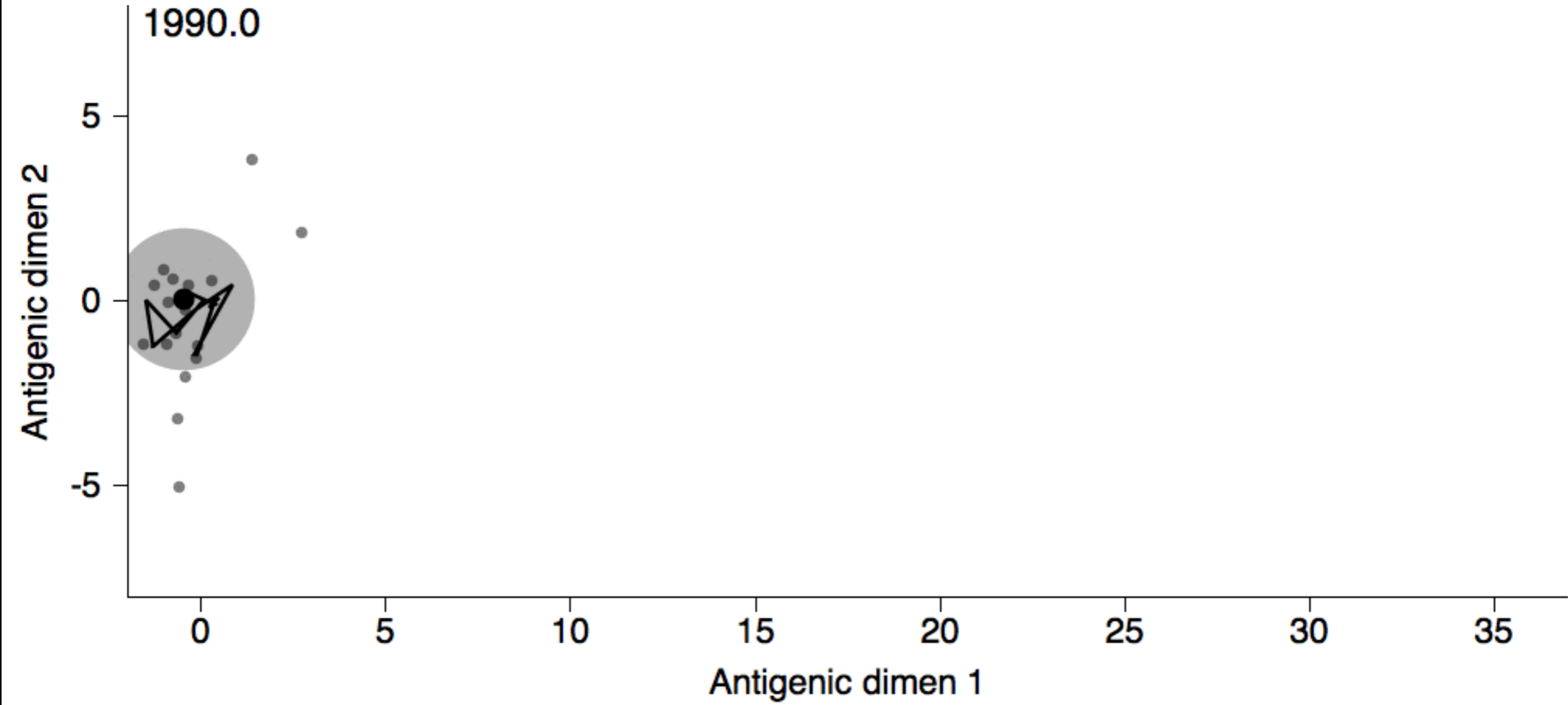
Keep estimates of system state (location and velocity vectors) and associated precisions.

Predict step: Use current state of the system at time  $t$  to predict state at time  $t+1$

Correct step: Use measurement of state to correction location

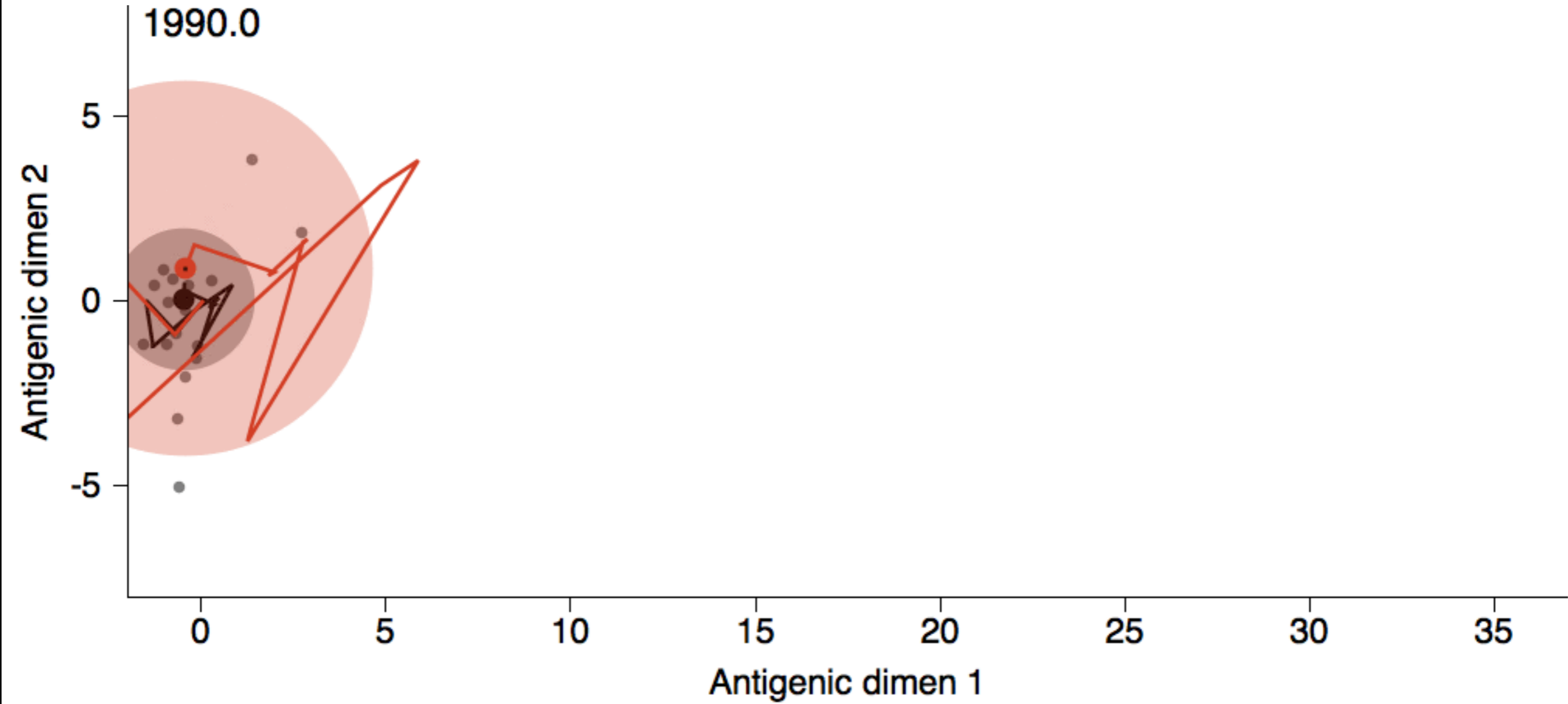


# Estimated antigenic trajectory

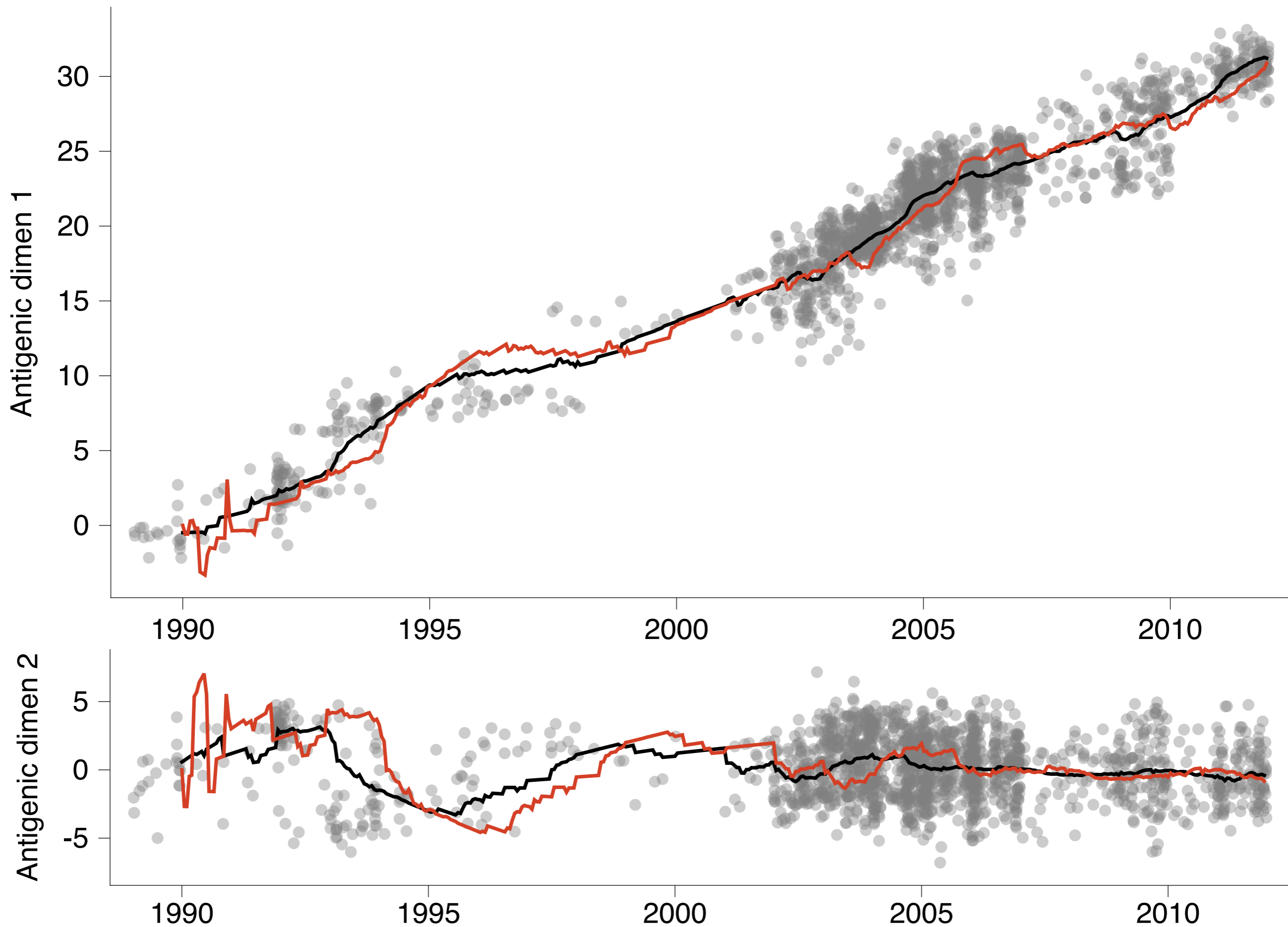




# Estimated antigenic trajectory with 1 year look ahead



# Estimated antigenic trajectory with 1 year look ahead



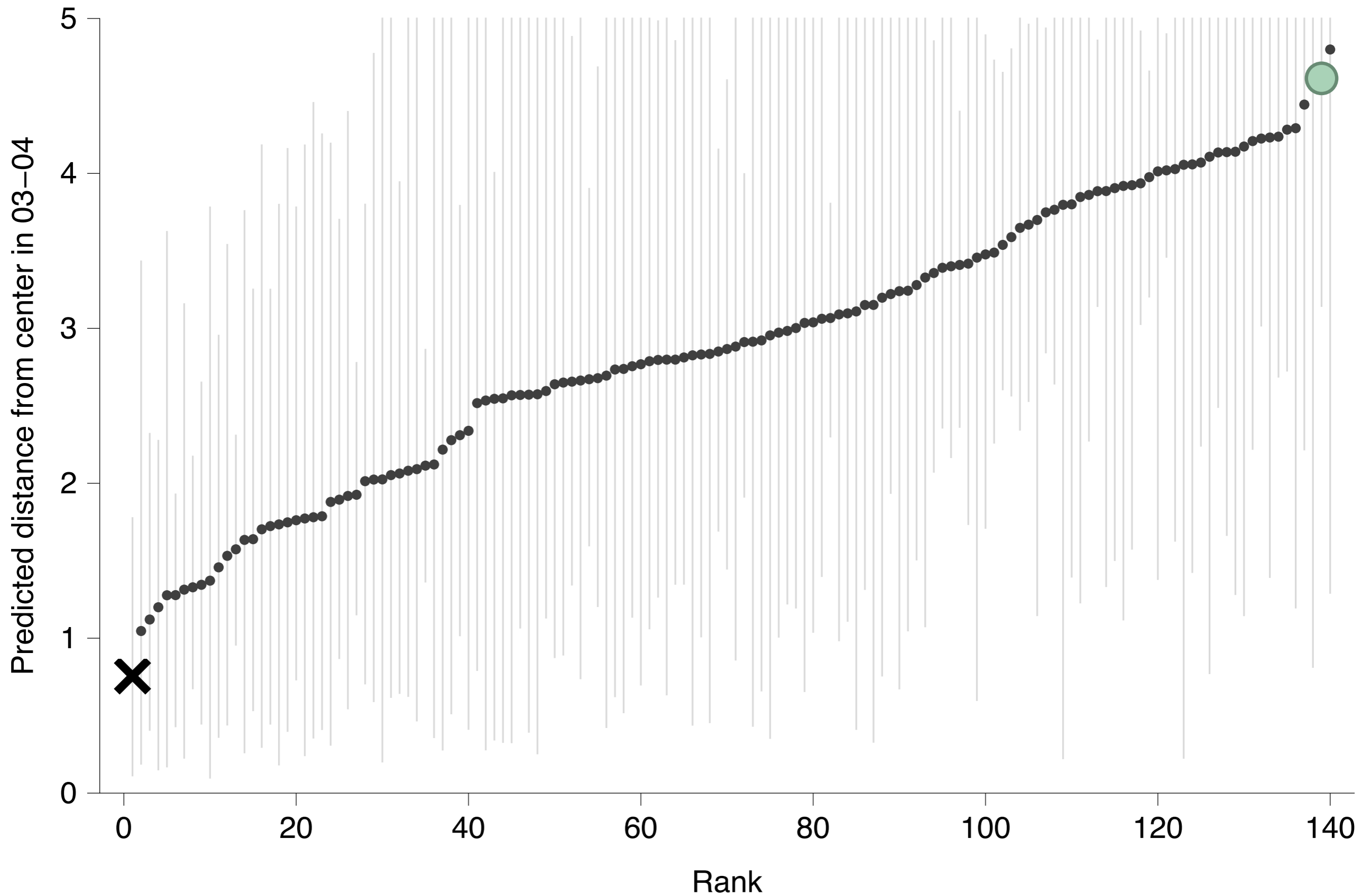
# Vaccine strain prediction

# Choice of 2002 viruses for 03-04 season

● Moscow/10/1999

✕ Guam/228/2002

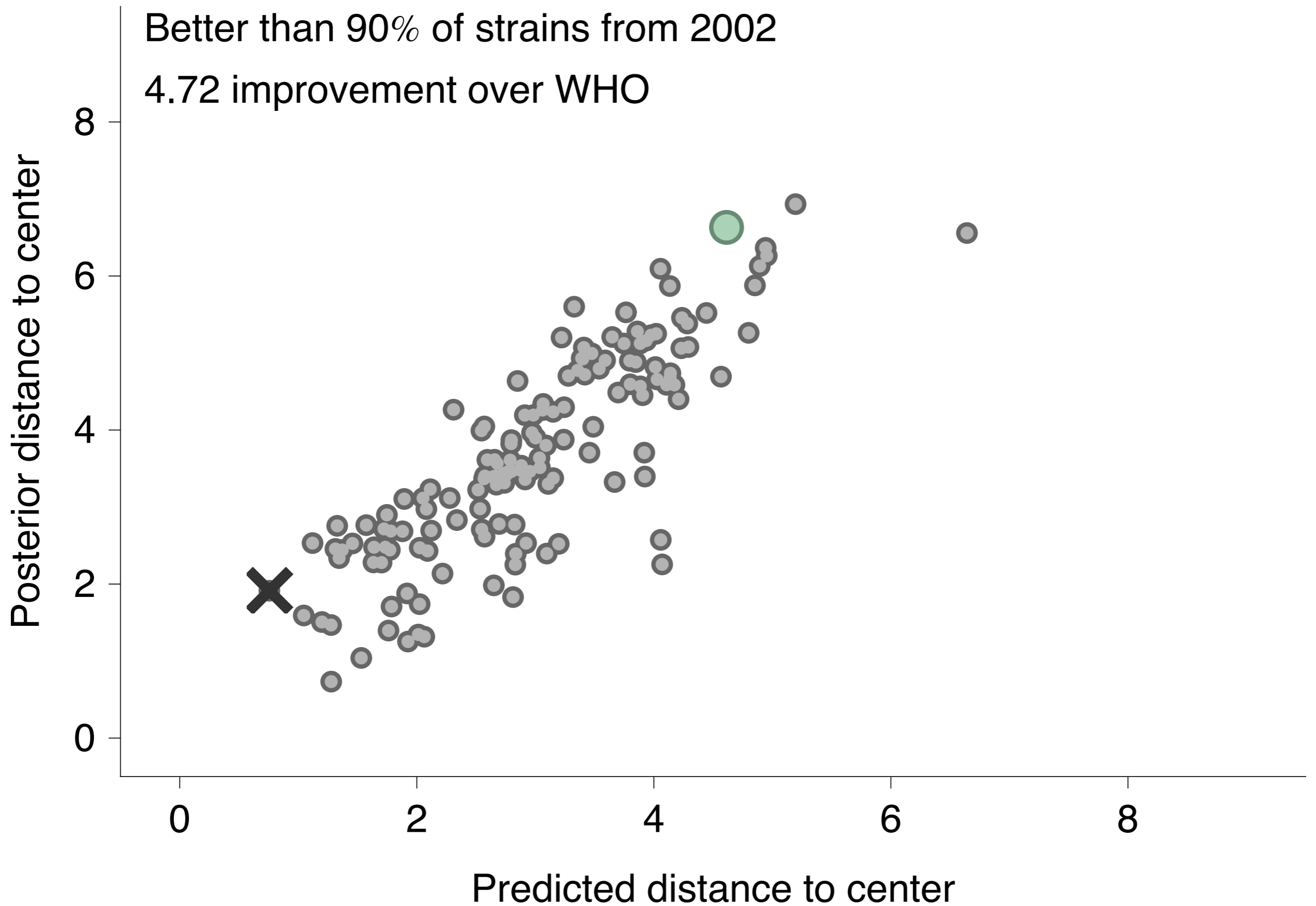
Pick virus with with lowest 95% upper HPD



# Choice of 2002 viruses for 03-04 season

- Moscow/10/1999
- ✕ Guam/228/2002

Better than 90% of strains from 2002  
4.72 improvement over WHO

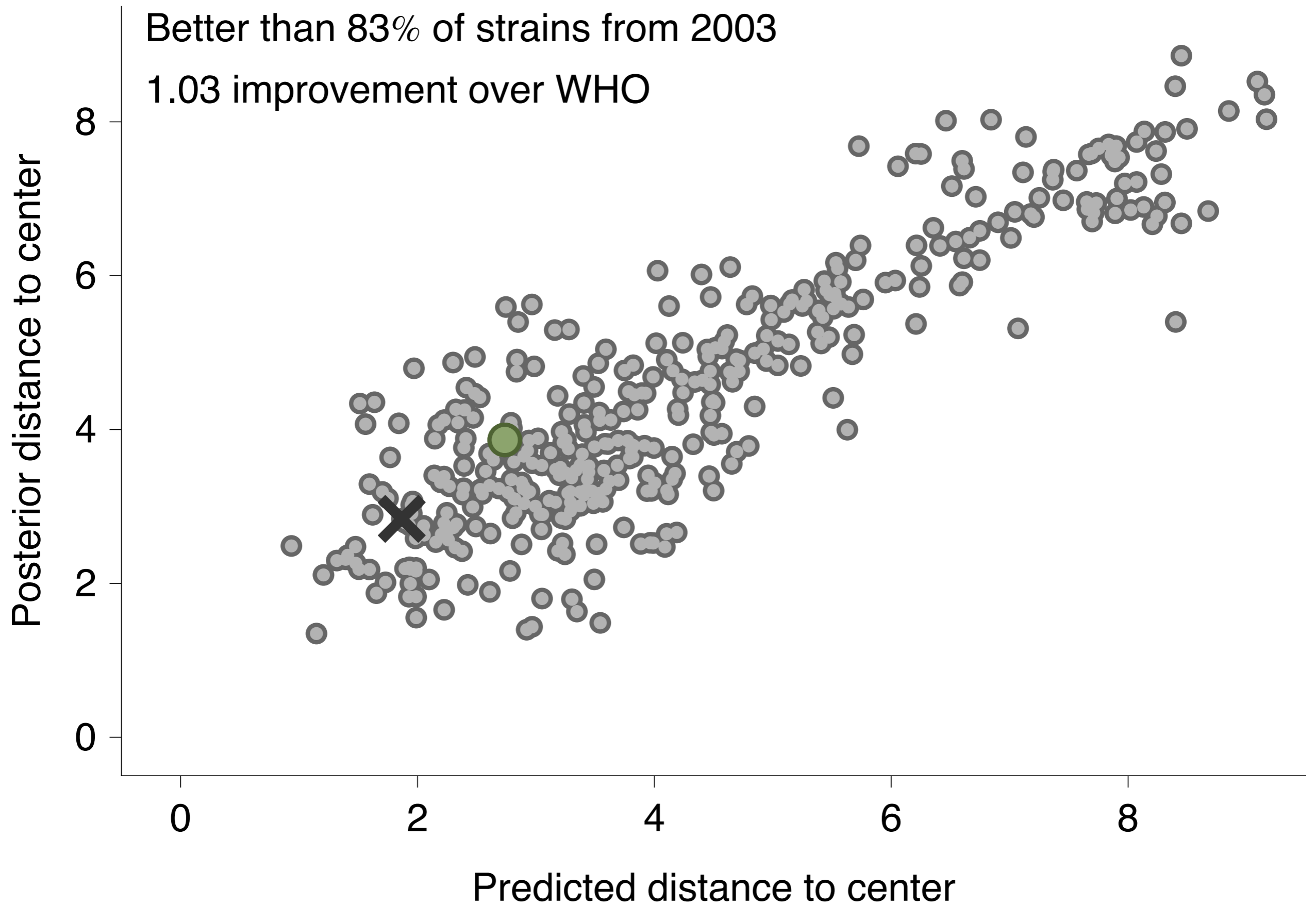


# Choice of 2003 viruses for 04-05 season

- Fujian/411/2002
- ✕ Brisbane/7/2003

Better than 83% of strains from 2003

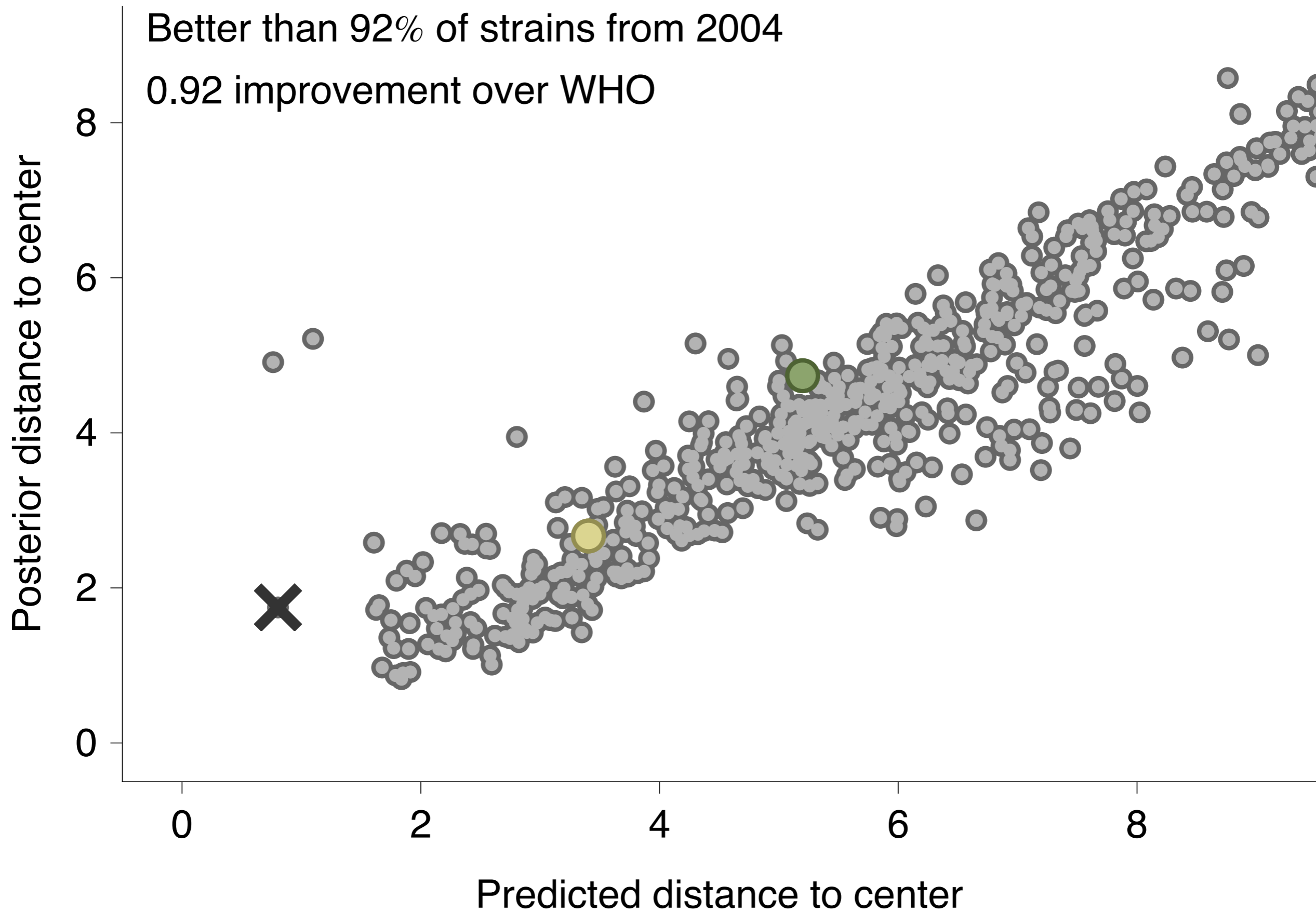
1.03 improvement over WHO



# Choice of 2004 viruses for 05-06 season

- California/7/2004
- ✕ Christ Church/280/2004

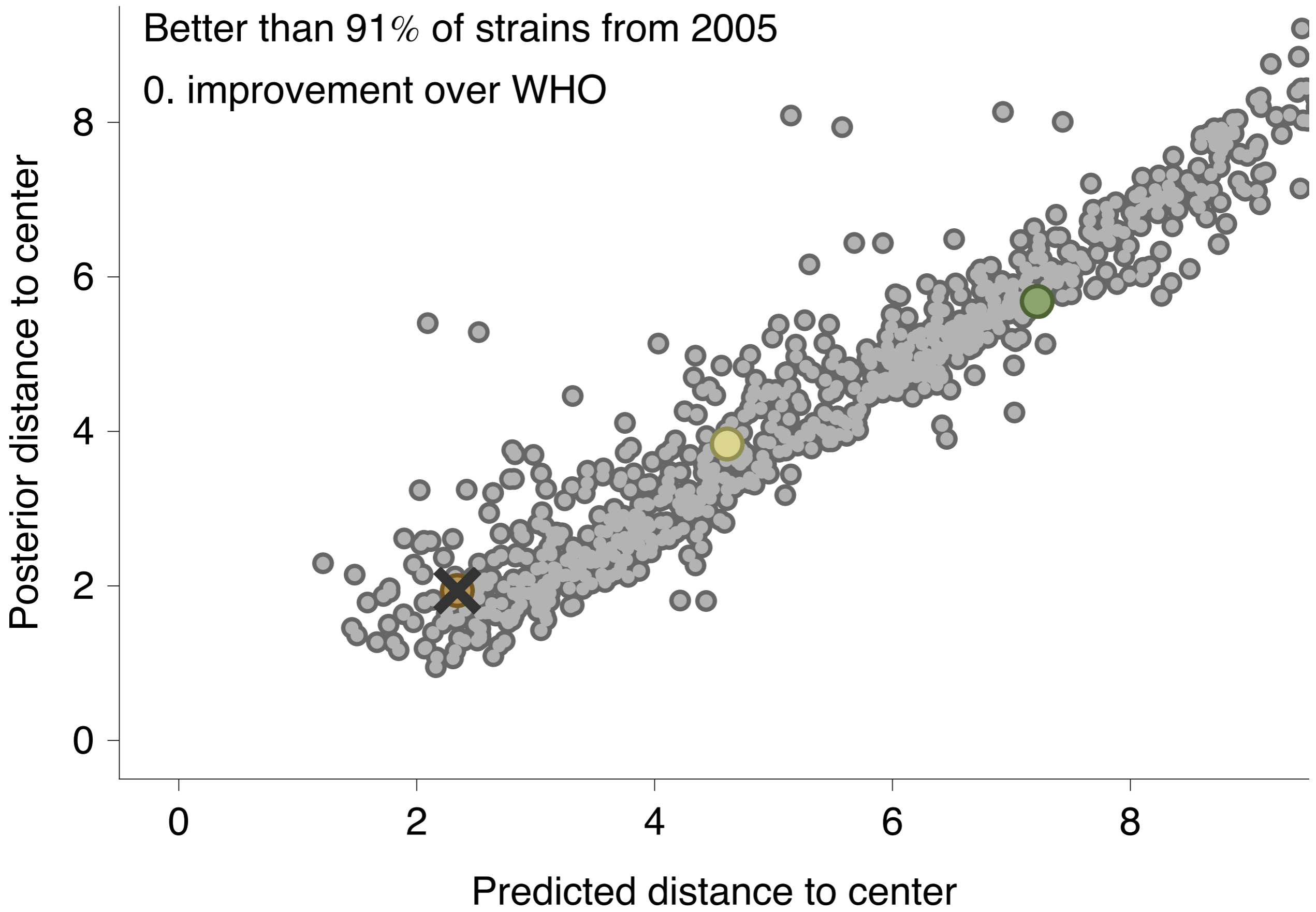
Better than 92% of strains from 2004  
0.92 improvement over WHO



# Choice of 2005 viruses for 06-07 season

- Wisconsin/67/2005
- ✕ Wisconsin/67/2005

Better than 91% of strains from 2005  
0. improvement over WHO





# Choice of 2006 viruses for 07-08 season

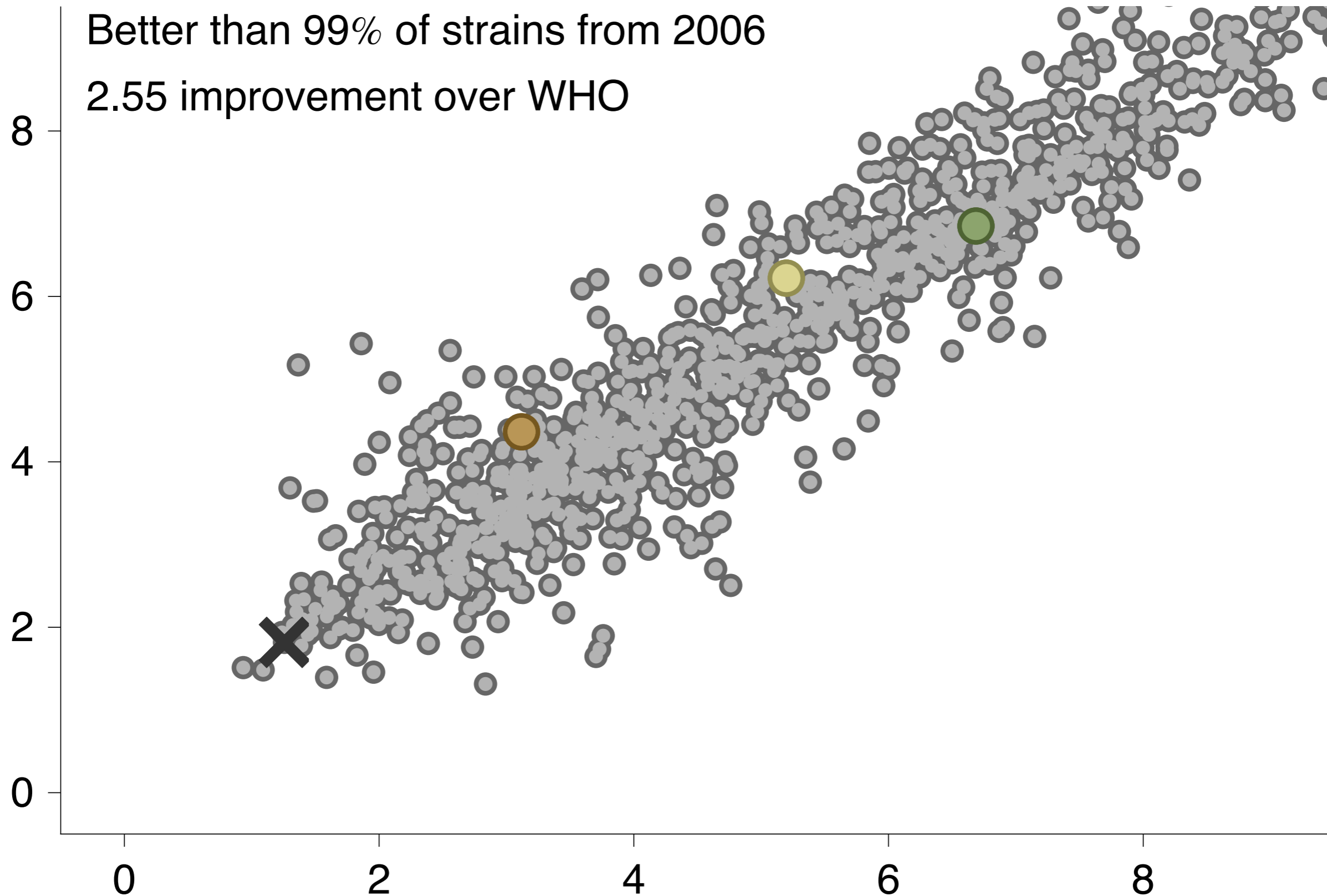
● Wisconsin/67/2005

✕ Gyeongnam/740/2006

Better than 99% of strains from 2006

2.55 improvement over WHO

Posterior distance to center



Predicted distance to center

# Choice of 2007 viruses for 08-09 season

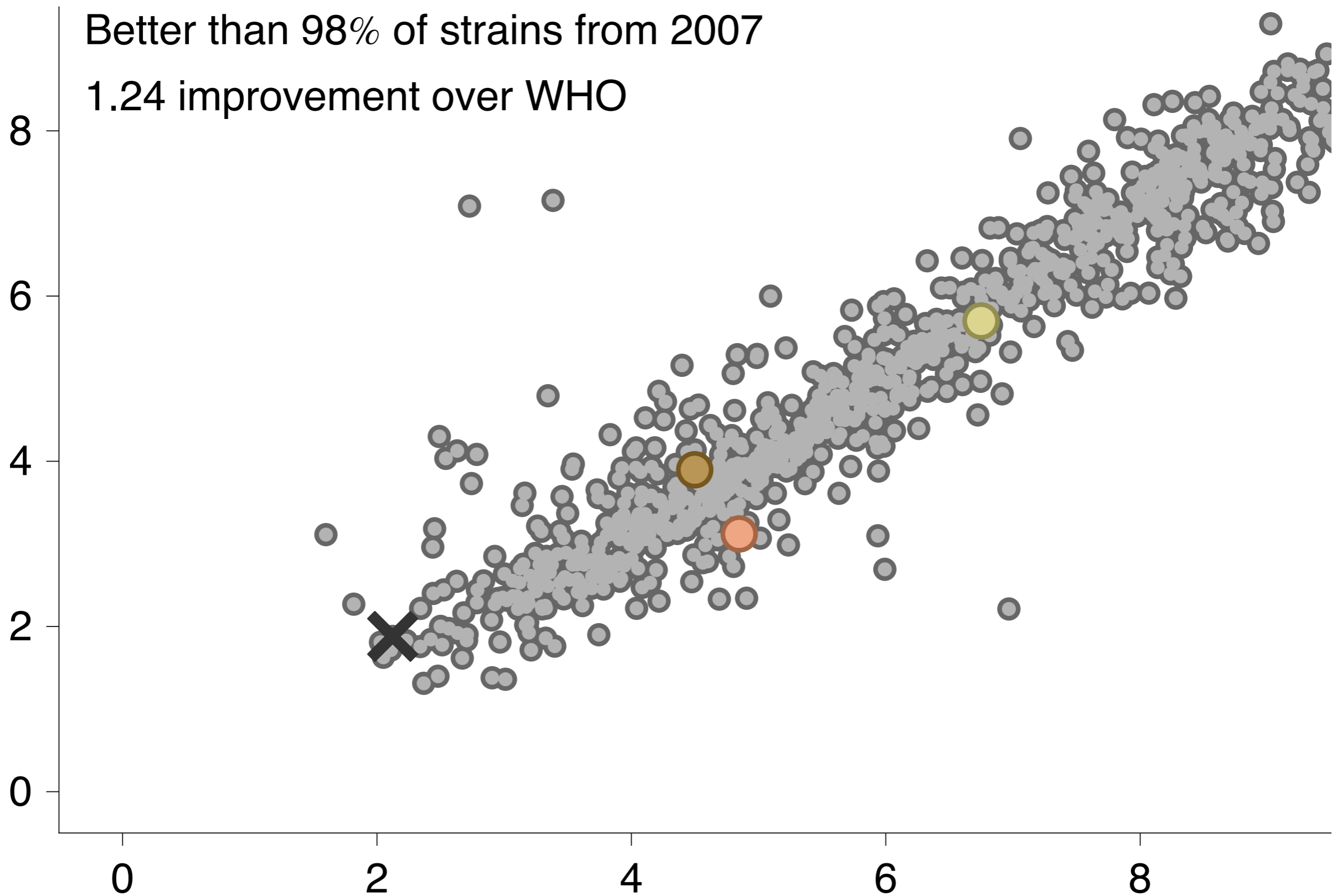
○ Brisbane/10/2007

✕ Lyon/1359/2006

Better than 98% of strains from 2007

1.24 improvement over WHO

Posterior distance to center



Predicted distance to center

# Choice of 2008 viruses for 09-10 season

○ Brisbane/10/2007

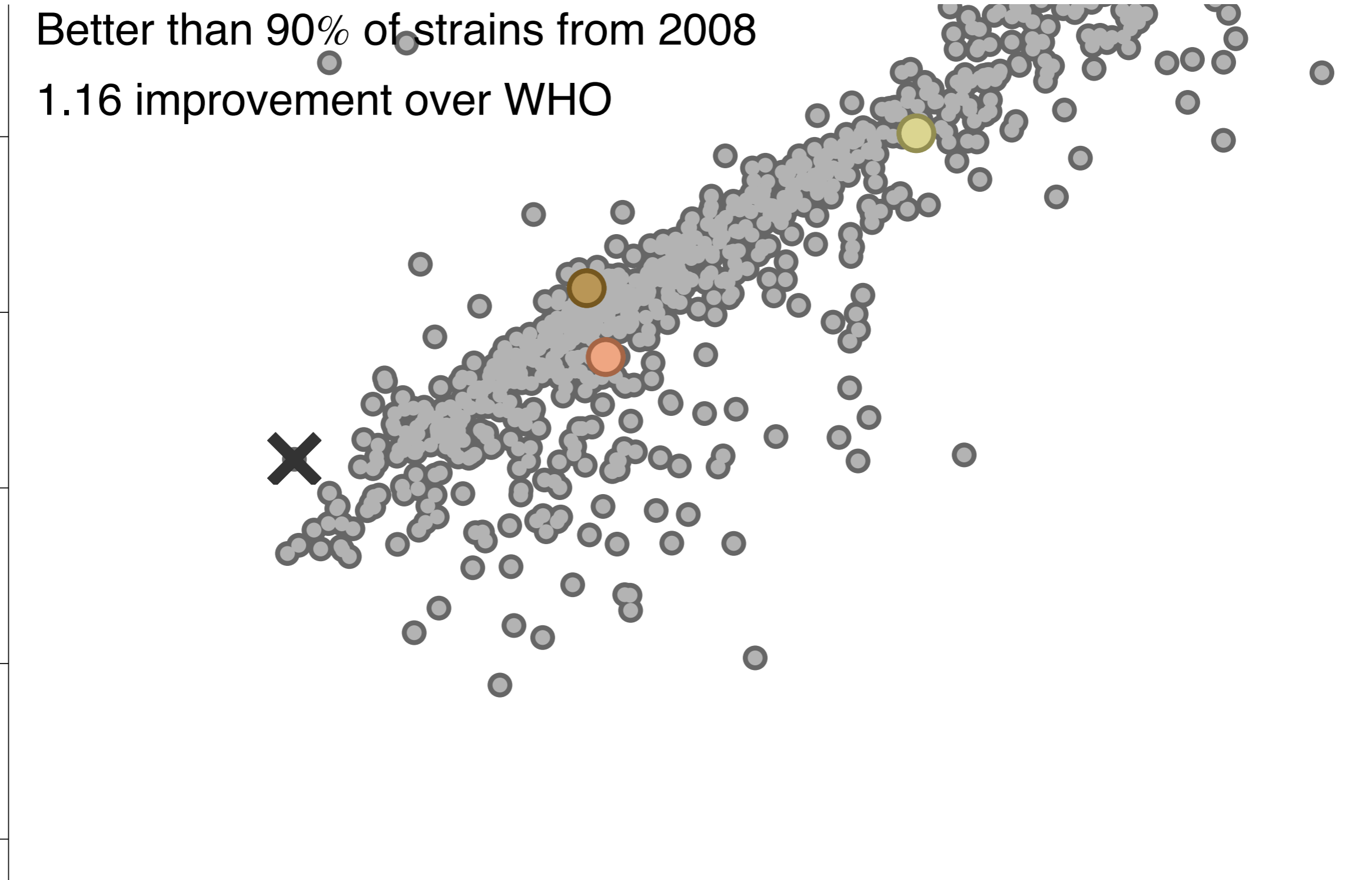
✕ Toulon/1244/2006

Better than 90% of strains from 2008

1.16 improvement over WHO

Posterior distance to center

8  
6  
4  
2  
0



0

2

4

6

8

Predicted distance to center

# Choice of 2009 viruses for 10-11 season

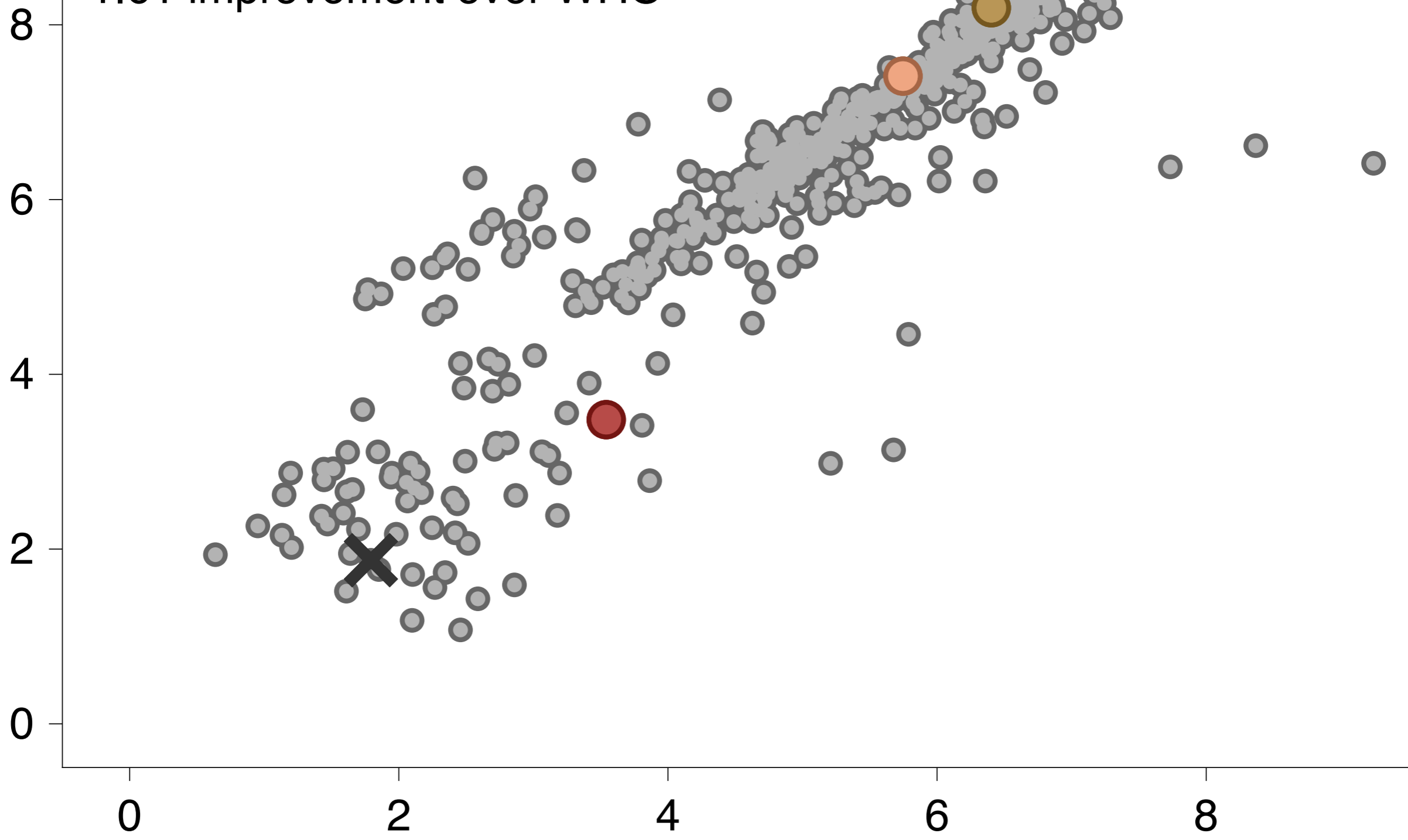
● Perth/16/2009

✕ Cameroon/675/2009

Better than 99% of strains from 2009

1.61 improvement over WHO

Posterior distance to center

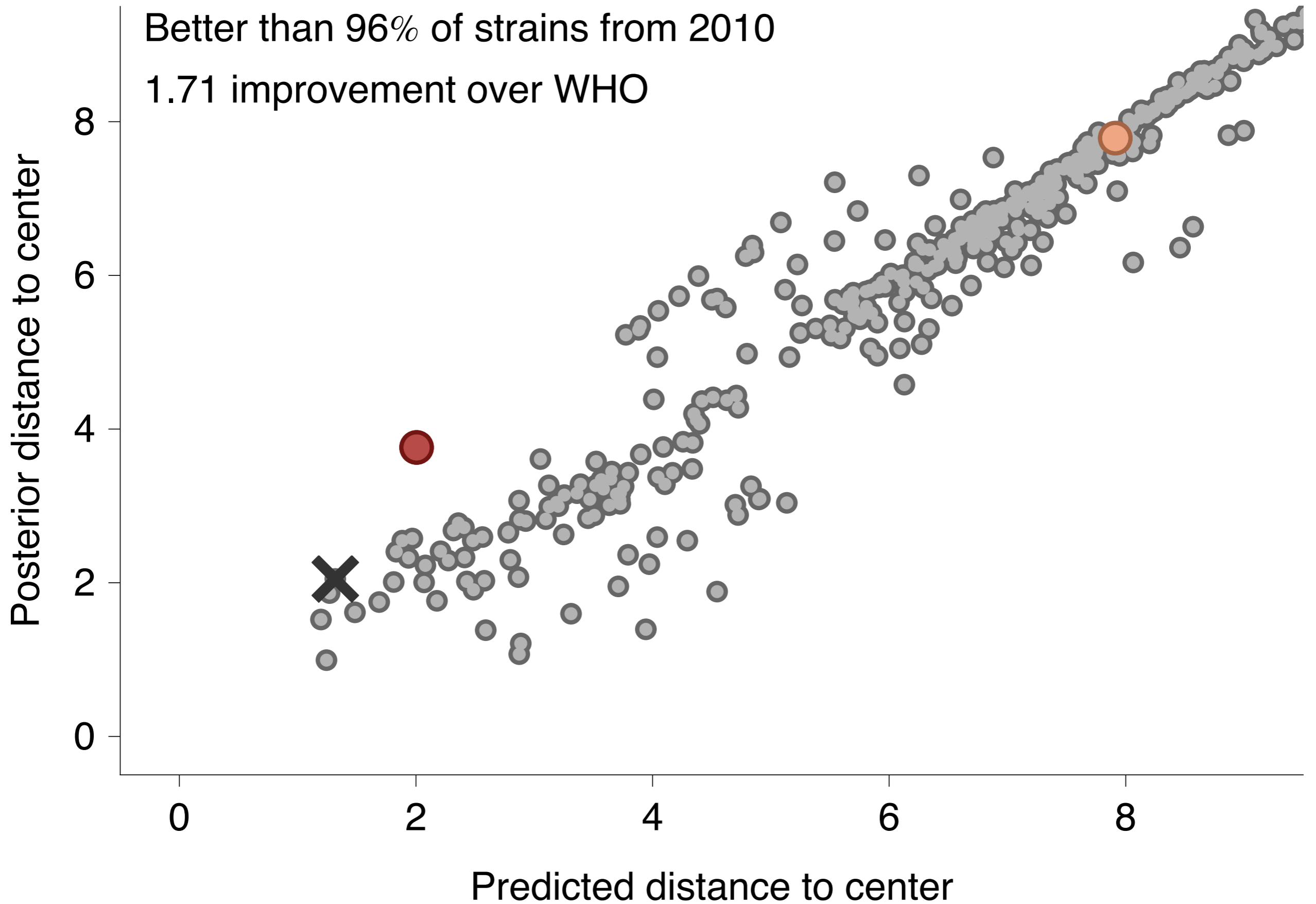


Predicted distance to center

# Choice of 2010 viruses for 11-12 season

- Perth/16/2009
- ✕ Turkey/26/2009

Better than 96% of strains from 2010  
1.71 improvement over WHO



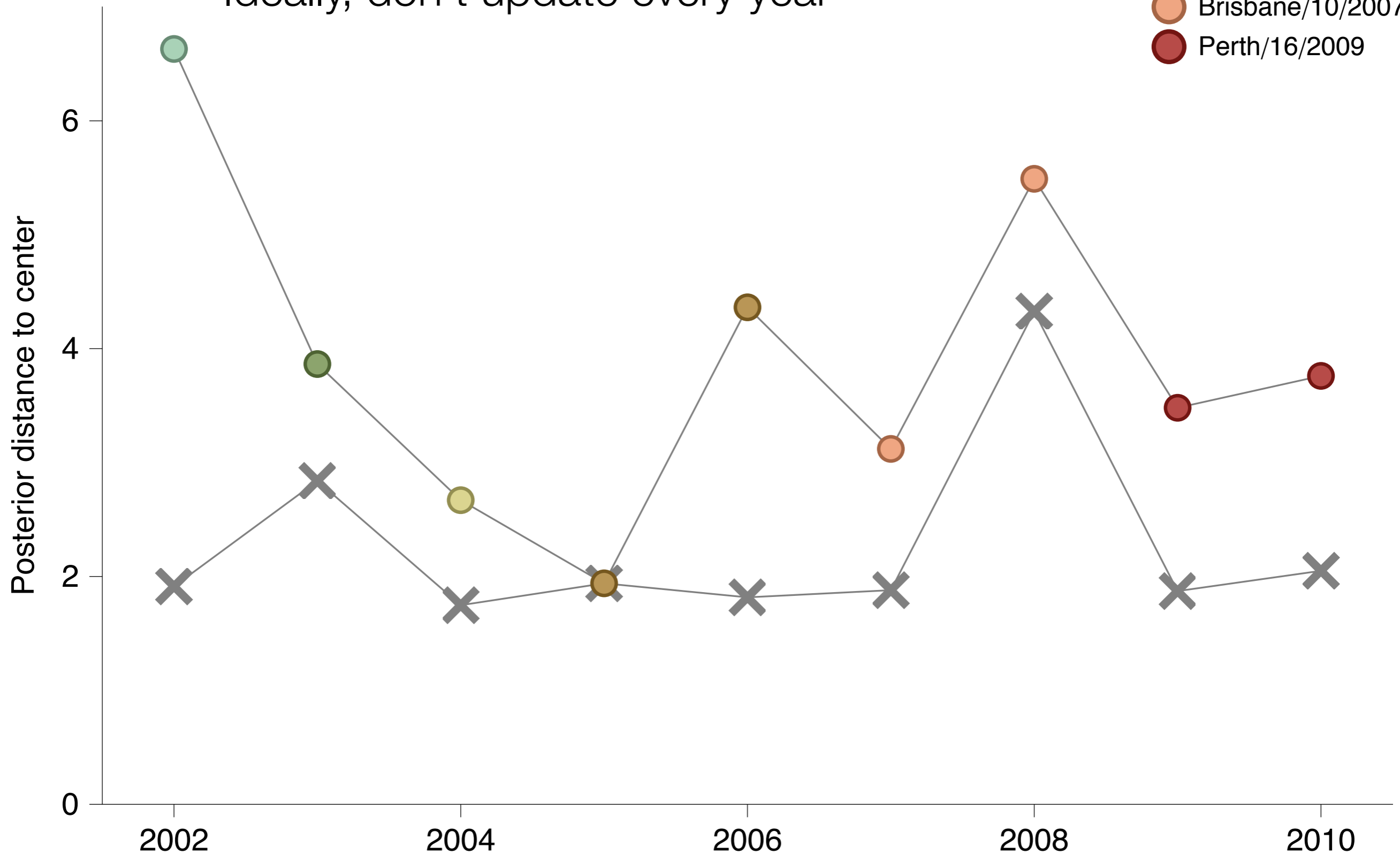
# Comparison with WHO predictions

- Improvement in 8/9 cases (equal in 1) and two mis-matches (2002 & 2006) avoided
- Average improvement of 1.7 HI units
- Roughly, 1 HI unit of mismatch is expected to translate to loss of 5-10% VE
- Expect 8-16% improvement in VE
- Assuming an 8% H3N2 attack rate and 43% coverage, an improvement of 8% VE is expected to translate to ~850 thousand cases prevented each year in the USA

# Comparison with WHO predictions

- Moscow/10/1999
- Fujian/411/2002
- California/7/2004
- Wisconsin/67/2005
- Brisbane/10/2007
- Perth/16/2009

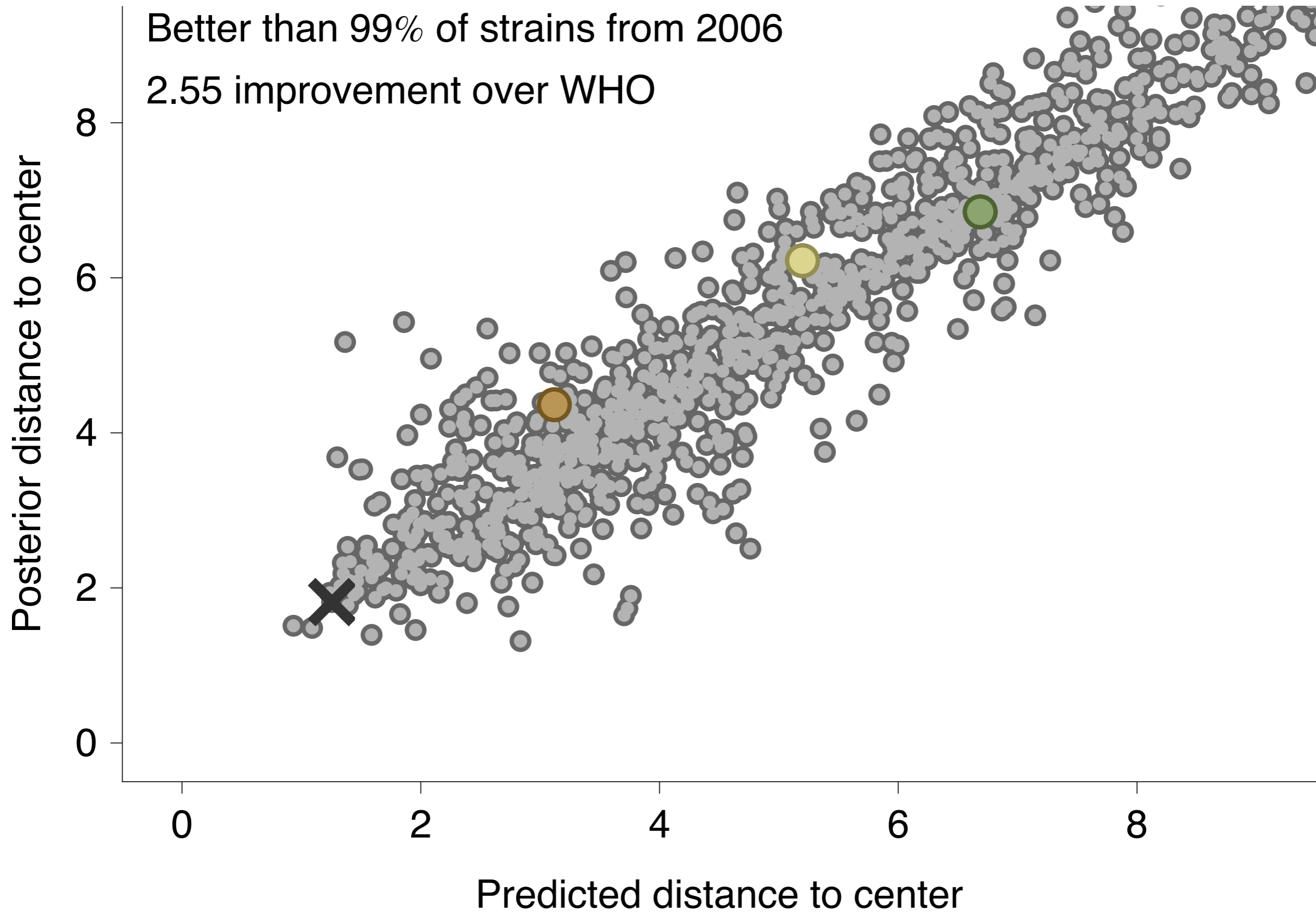
Ideally, don't update every year



# Choice of 2006 viruses for 07-08 season

- Wisconsin/67/2005
- ✕ Gyeongnam/740/2006

Better than 99% of strains from 2006  
2.55 improvement over WHO



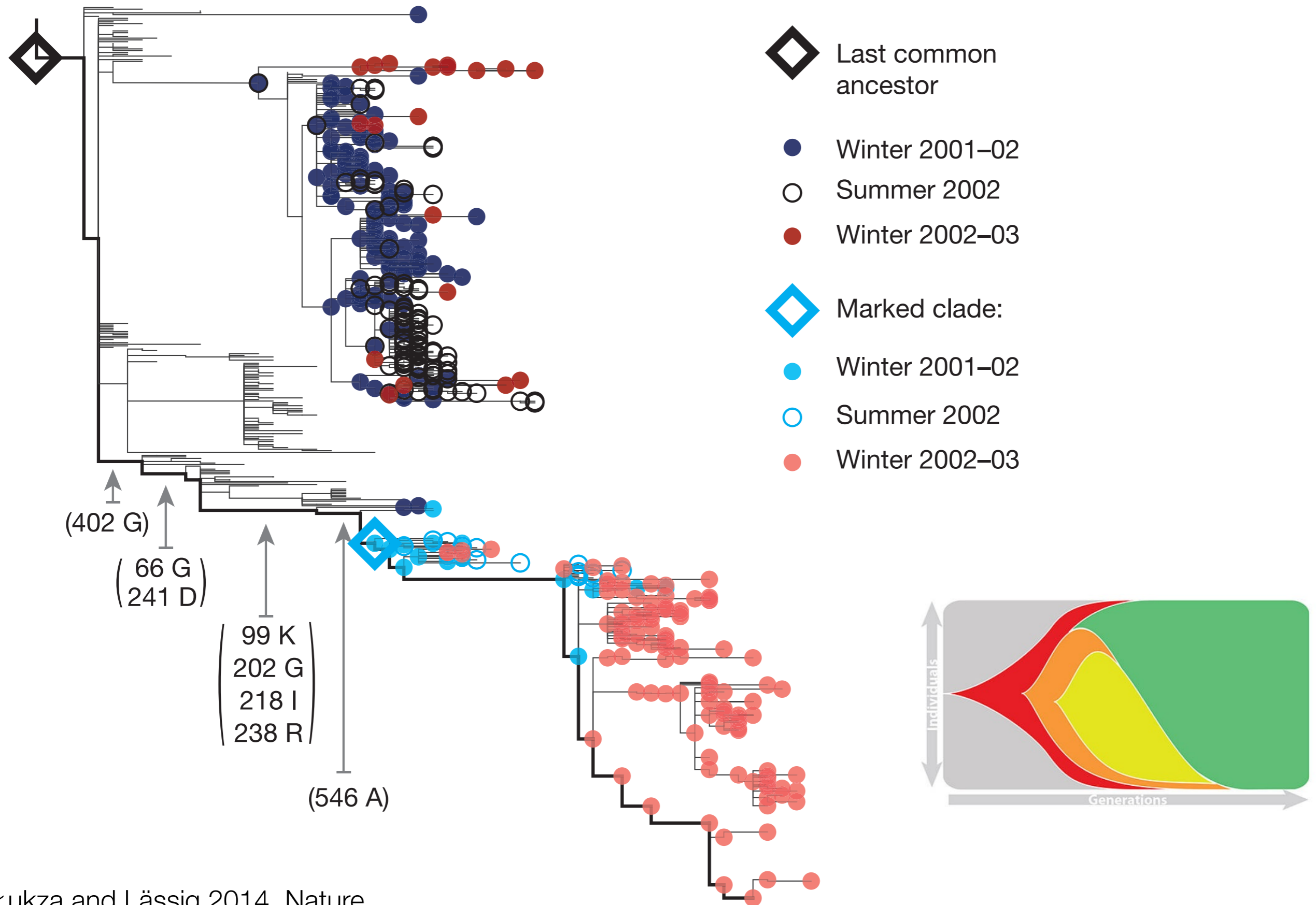


Like to do this for 2014/2015

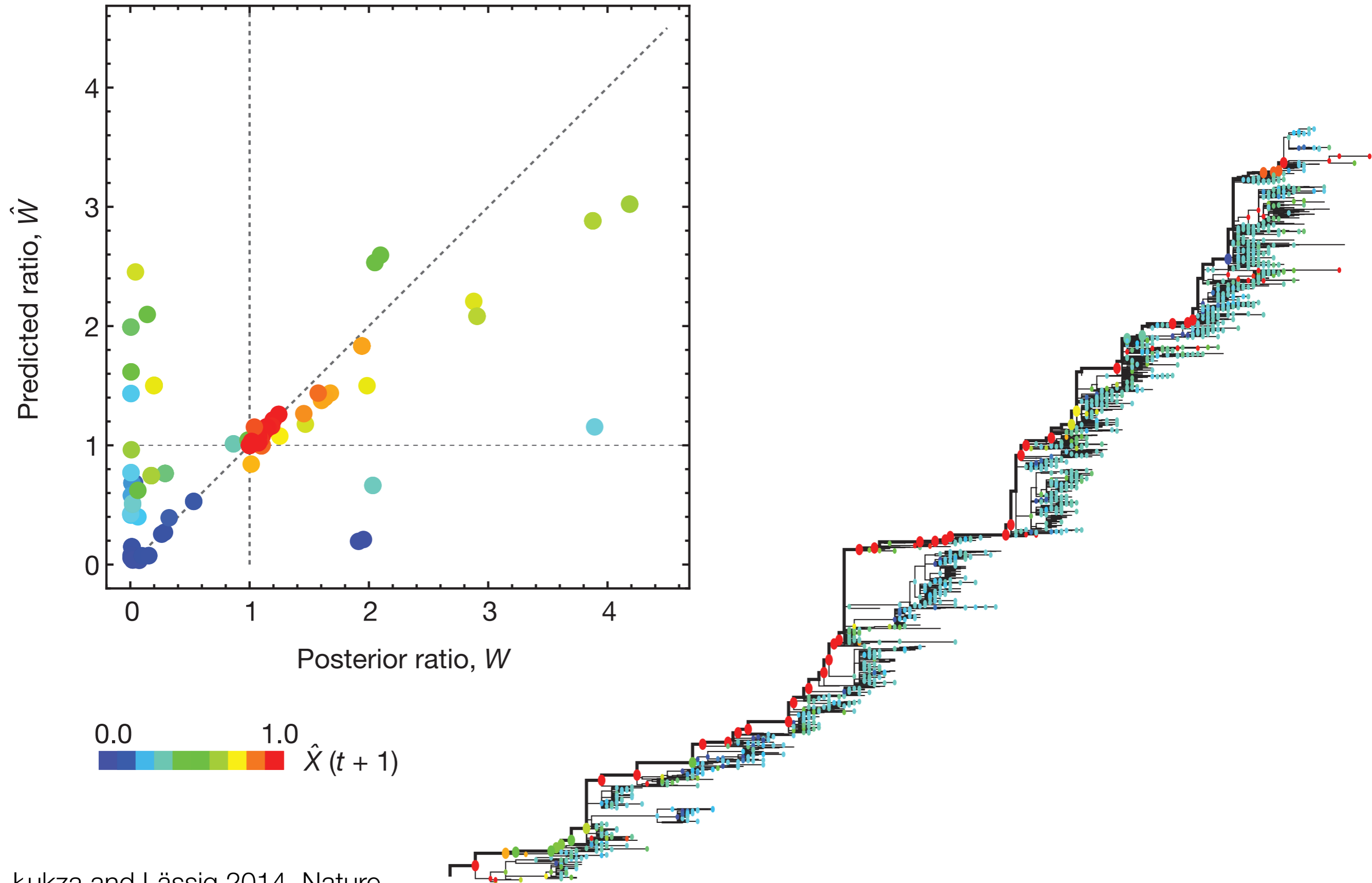
Data problem  
HI is not publicly available

1437 HA sequences for the  
2013/2014 season

# Clonal interference and prediction of clade frequencies



# Clonal interference and prediction of clade frequencies



## Prediction

$$\hat{X}_v(t+1) = \sum_{i:v,t} x_i \exp(f_i)$$

## Strain fitness

$$f_i = f_0 - \mathcal{L}(\mathbf{a}_i) - \sum_{j: t_j < t_i} x_j \mathcal{C}(\mathbf{a}_i, \mathbf{a}_j)$$

$x_i$	Frequency of strain $i$	$\mathcal{L}(\mathbf{a}_i)$	Fitness load of del. mutants
$X_v$	Frequency of clade $v$	$\mathcal{C}(\mathbf{a}_i, \mathbf{a}_j)$	Cross-immunity between $i$ and $j$
$\mathbf{a}_i$	Sequence of strain $i$		

# Fitting strain fitness

$$f_i = f_i^{\text{ep}} + f_i^{\text{ne}} + f_i^{\text{nl}}$$

$$\hat{f}_i = \beta^{\text{ep}} f_i^{\text{ep}} + \beta^{\text{ne}} f_i^{\text{ne}} + \beta^{\text{nl}} f_i^{\text{nl}}$$

$f_i^{\text{ep}}$  fitness from cross-immunity (shared epitope sites)

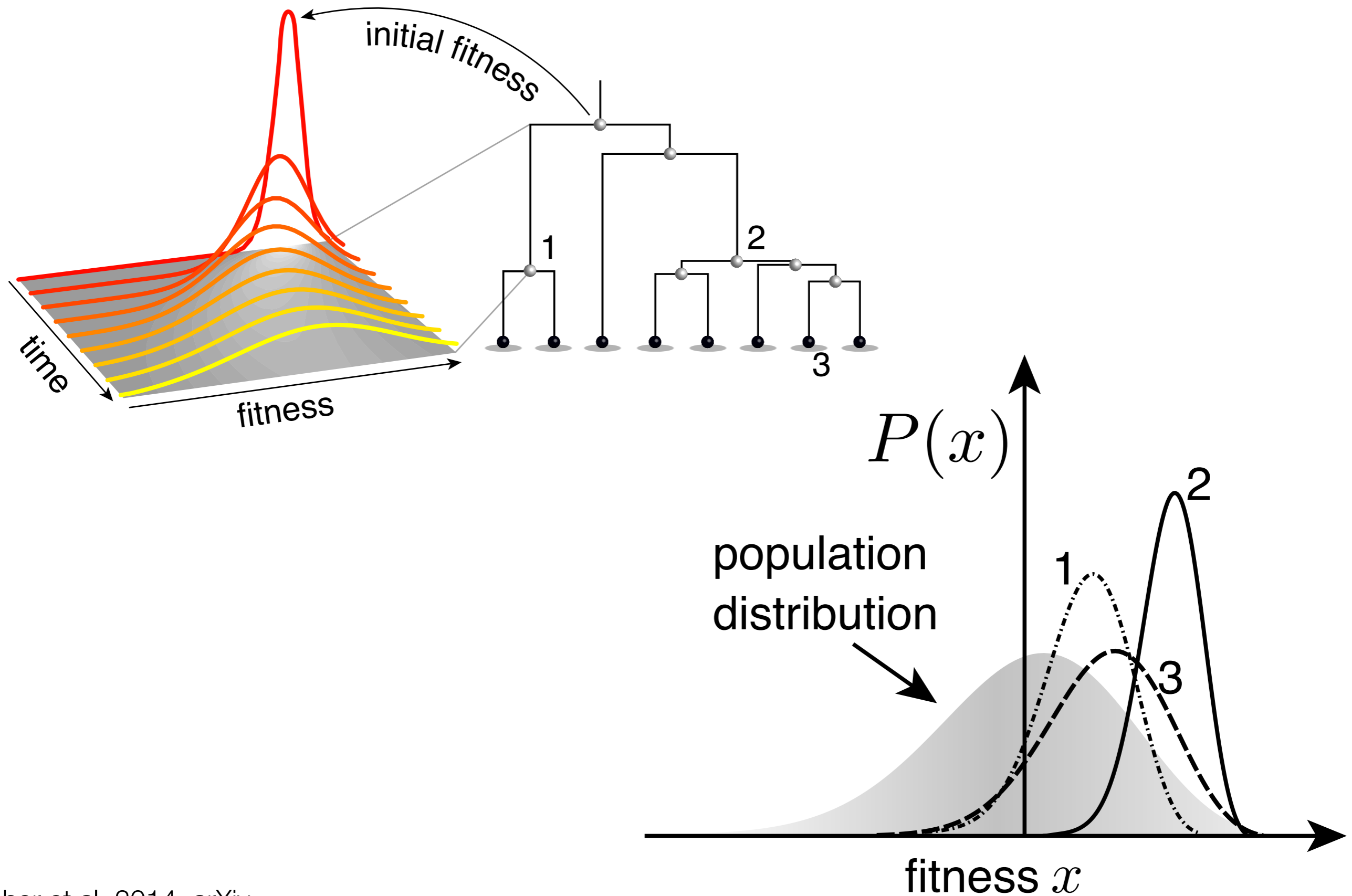
$f_i^{\text{ne}}$  fitness from mutational load (differences from consensus at non-epitope sites)

$f_i^{\text{nl}}$  fitness from clade age (older clades as estimated by synonymous mutations are more fit)

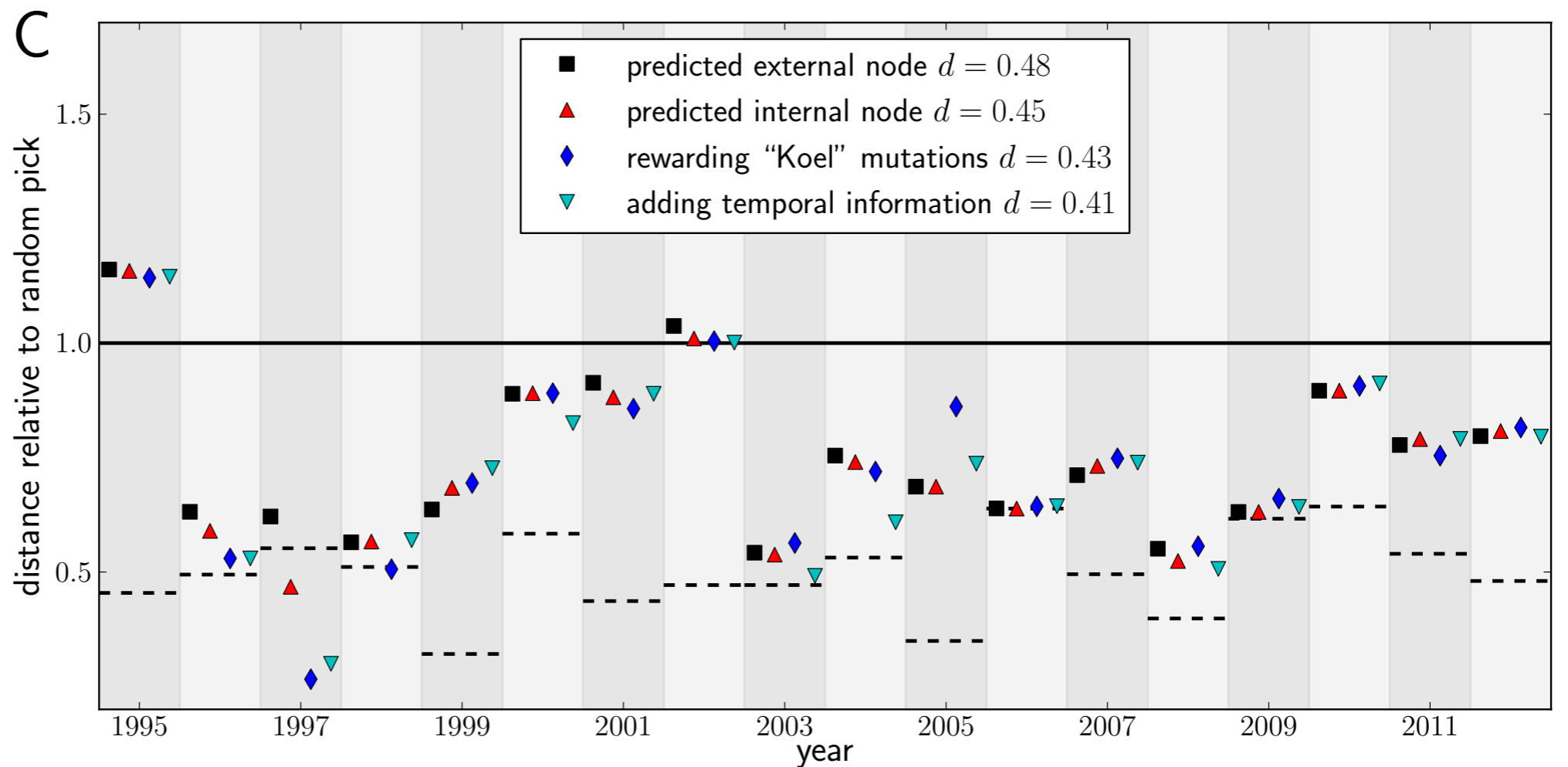
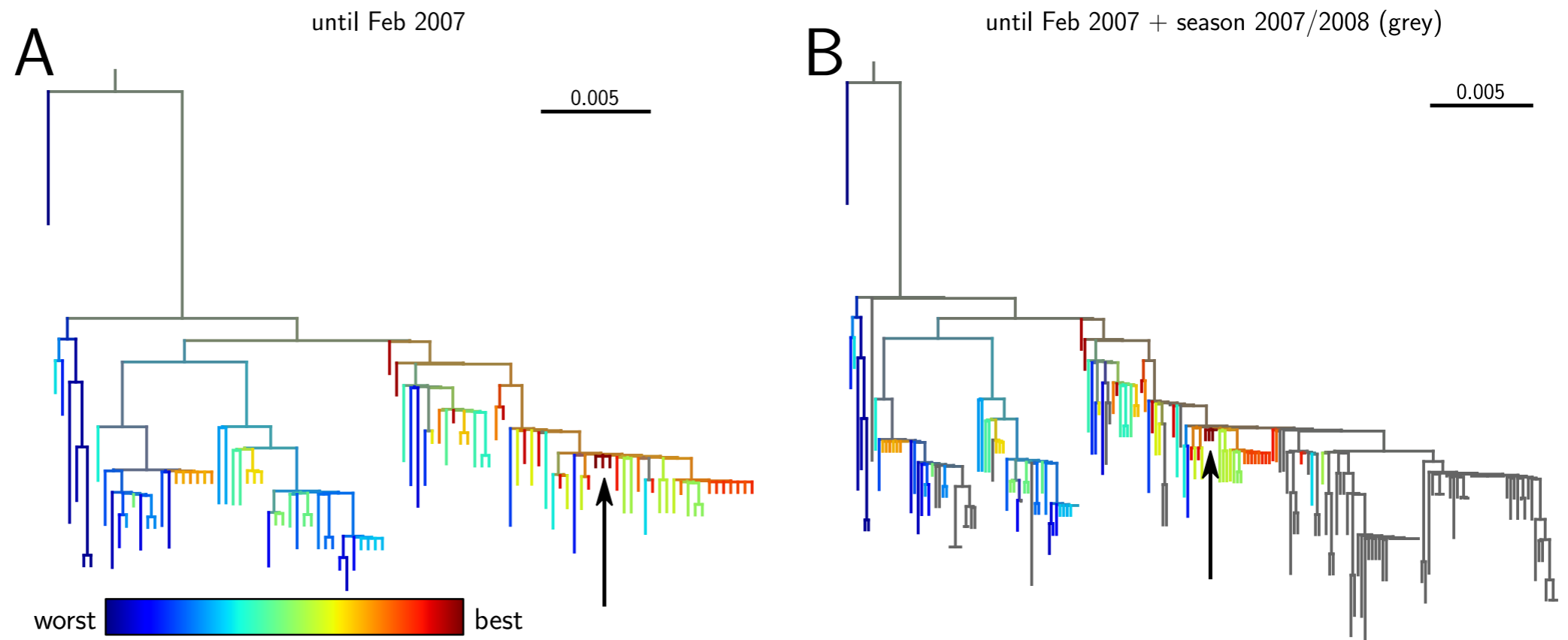
The winning entries for the Netflix prize had 100+ predictors. Winning teams were ensembles.



# Predicting clade fitness from tree shape



# Predicting clade fitness from tree shape



## Fitting strain fitness

$$f_i = f_i^{\text{ep}} + f_i^{\text{ne}} + f_i^{\text{nl}} + f_i^{\text{tr}}$$

$$\hat{f}_i = \beta^{\text{ep}} f_i^{\text{ep}} + \beta^{\text{ne}} f_i^{\text{ne}} + \beta^{\text{nl}} f_i^{\text{nl}} + \beta^{\text{tr}} f_i^{\text{tr}}$$

$f_i^{\text{ep}}$  fitness from cross-immunity

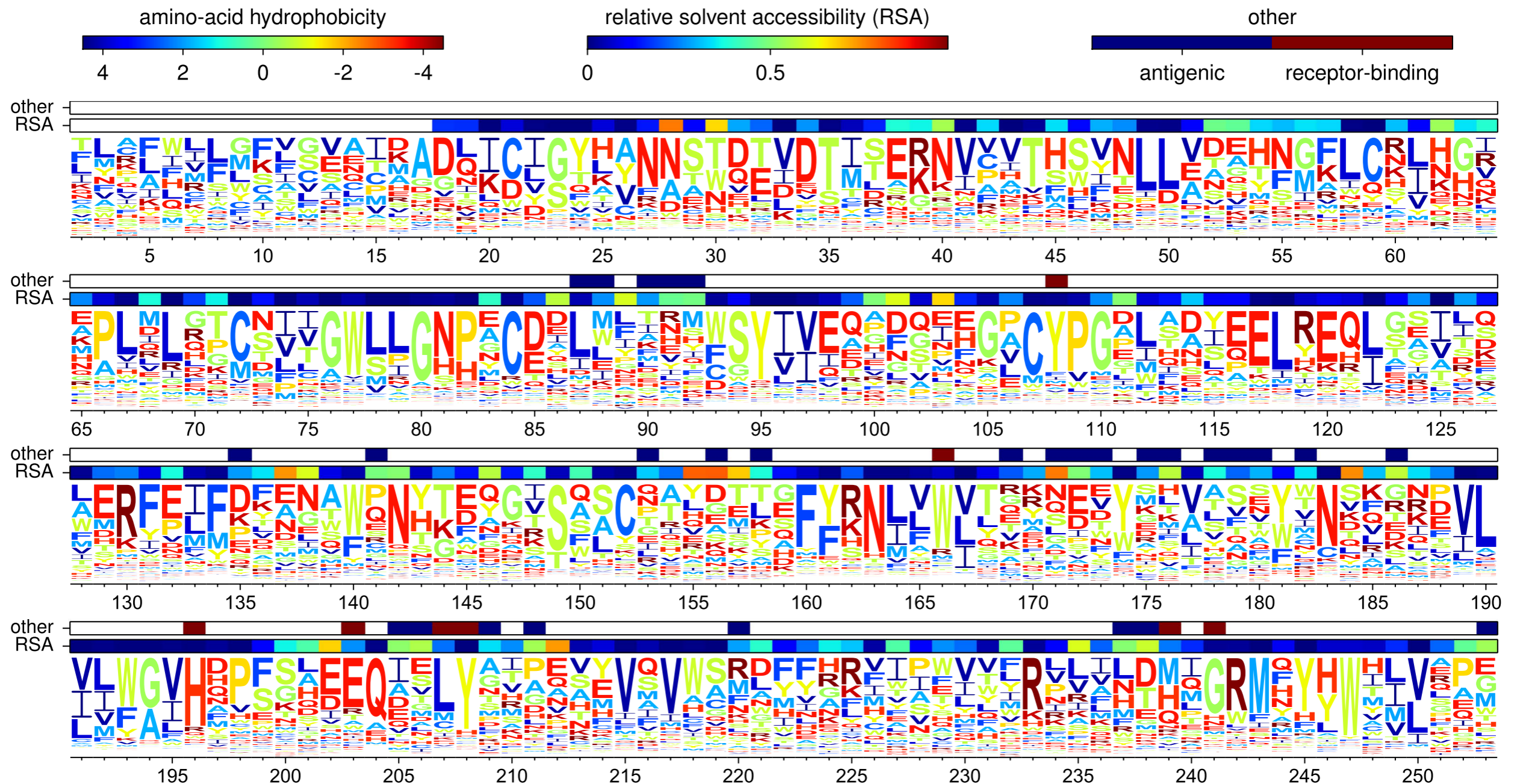
$f_i^{\text{ne}}$  fitness from mutational load

$f_i^{\text{nl}}$  fitness from clade age

$f_i^{\text{tr}}$  fitness from tree shape

# Experimental characterization of HA fitness

$$f_i = f_0 - \mathcal{L}(\mathbf{a}_i) - \sum_{j: t_j < t_i} x_j \mathcal{C}(\mathbf{a}_i, \mathbf{a}_j)$$



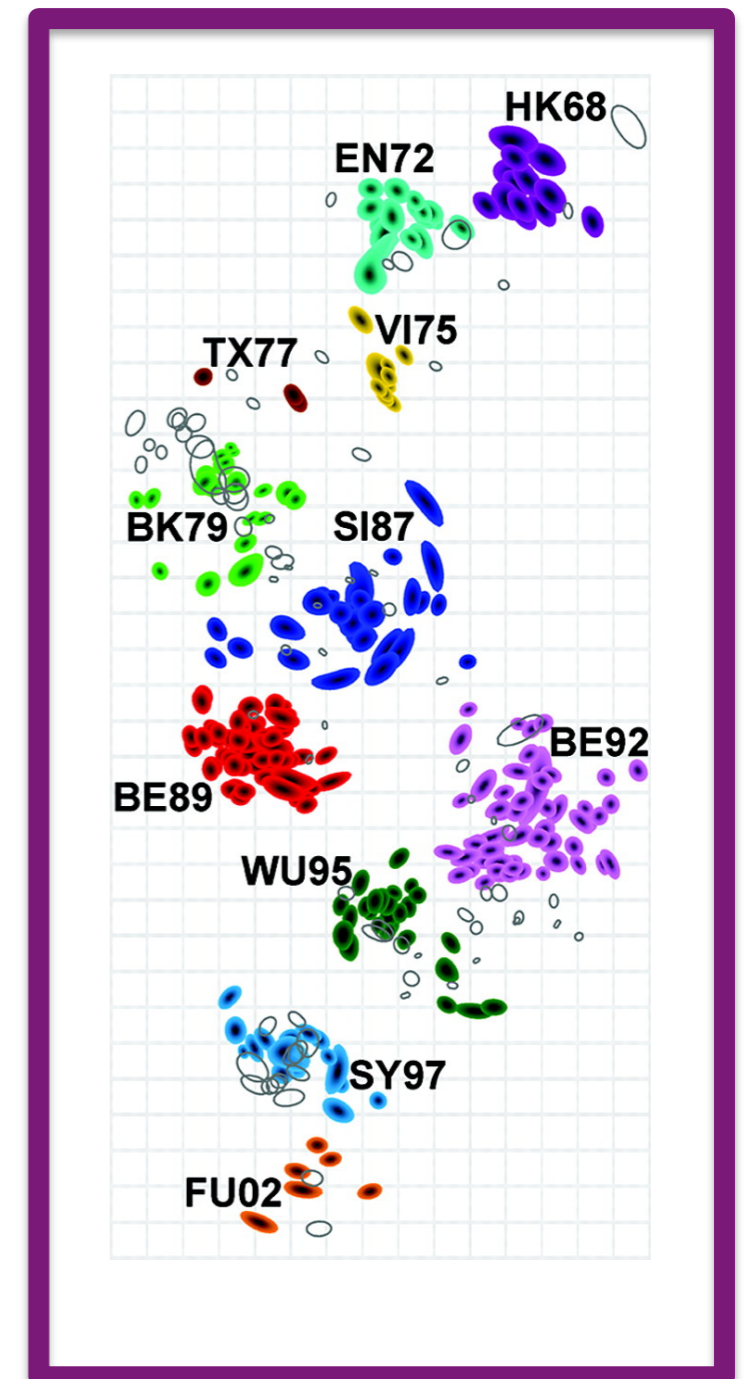
# Uncovering antigenic mutations

$$f_i = f_0 - \mathcal{L}(\mathbf{a}_i) - \sum_{j: t_j < t_i} x_j \mathcal{C}(\mathbf{a}_i, \mathbf{a}_j)$$

X

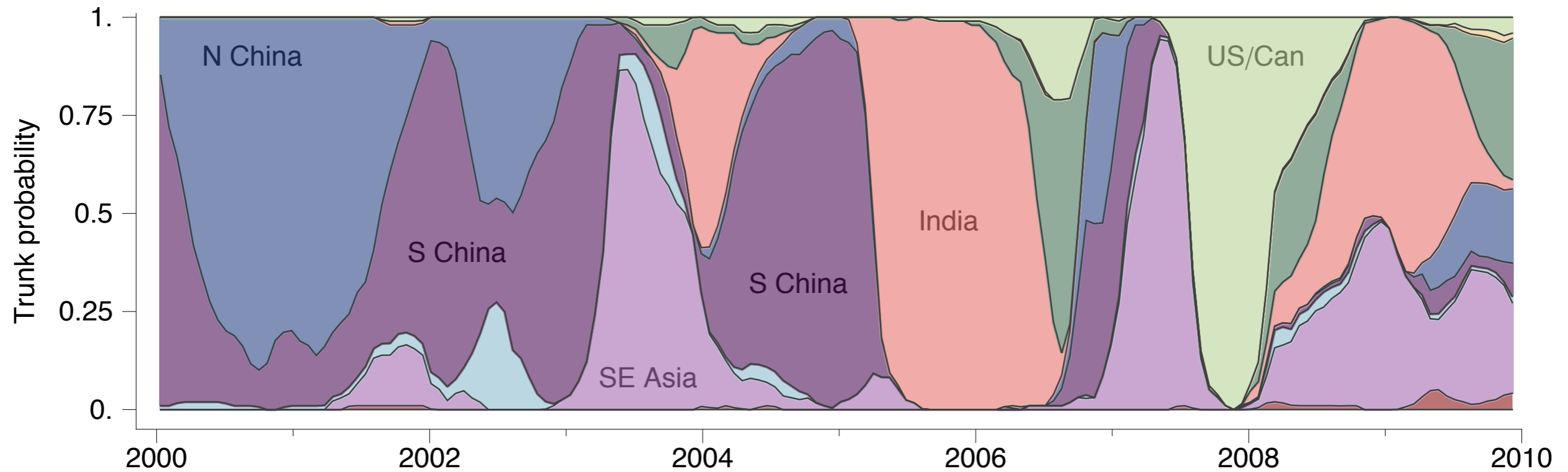
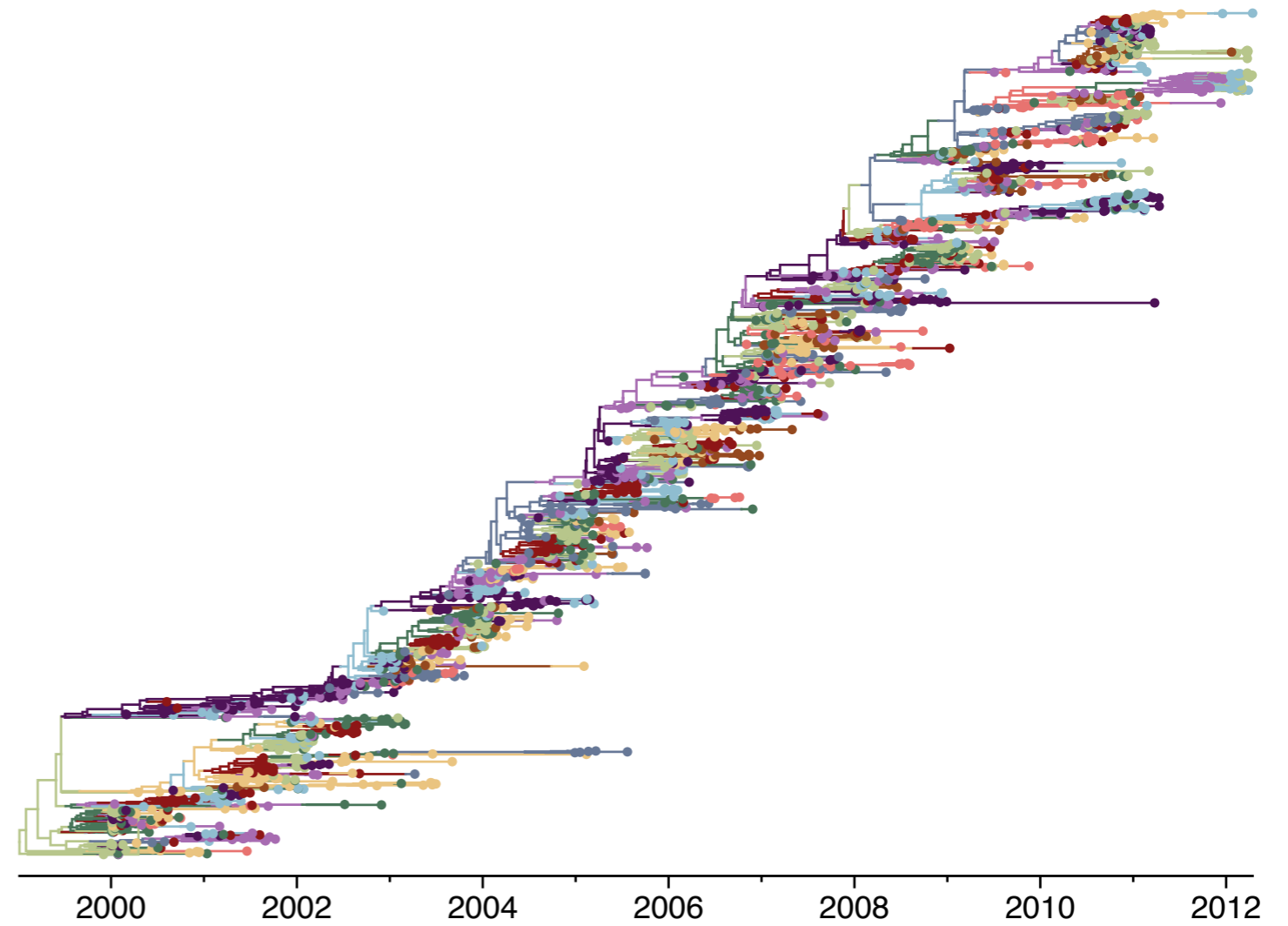
Y

Virus Strain	Sequence
A/Akita/4/1993	S T G R I C D ..
A/Alabama/5/2010	S T G <b>E</b> I C <b>N</b> ..
A/Alaska/5/2010	S T G <b>E</b> I C <b>N</b> ..
A/Algeria/G202/2009	S T G <b>E</b> I C D ..
A/Amsterdam/1609/1977	S T G <b>K</b> I C D ..
A/Anhui/1238/2005	S T G <b>G</b> I C D ..
A/Anhui/1239/2005	S T G <b>G</b> I C D ..
A/Atlanta/211/1989	S T G R I C <b>N</b> ..
...	



# Migration dynamics

- USA & Canada
- Central America
- South America
- Europe
- Africa
- India
- China
- Japan & Korea
- Southeast Asia
- Oceania



Goal is always-on server that does daily builds, pulling in sequence data and predicting clade trajectories

# Acknowledgements

## Data sources:

- WHO Global Influenza Surveillance Network
- John McCauley at the WHO Collaborating Centre in London

## Collaborators:

- Andrew Rambaut (University of Edinburgh)
- Charles Cheung (FHCRC)
- Colin Russell (Cambridge University)
- Philippe Lemey (Katholieke Universiteit Leuven)
- Marta Łukza (Columbia University)
- Marc Suchard (UCLA)
- Gytis Dudas (University of Edinburgh)
- Derek Smith (Cambridge University)
- Michael Lässig (University of Cologne)
- Richard Neher (Max Planck Institute)

