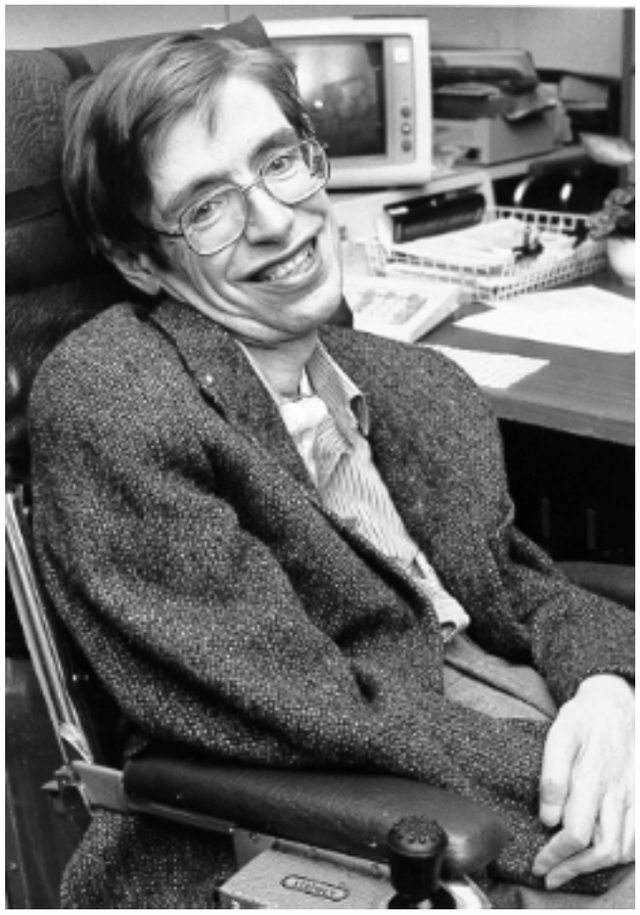
What Bayes (and stats) can do for AI: vignettes from assistive technology, air pollution mitigation, and LLM leaderboards

Tamara Broderick
Associate Professor,
MIT

Bayes and AI for assistive technology

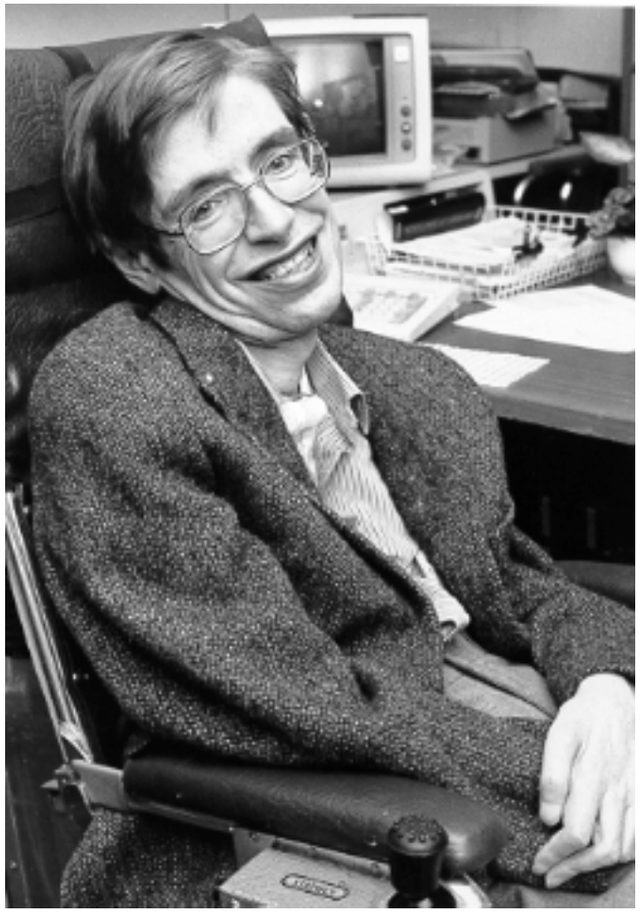
[https://commons.wikimedia.org/wiki/File:Jean-Dominique-Bauby-01.jpg, https://commons.wikimedia.org/wiki/File:Stephen_Hawking.StarChild.jpg, www.itv.com/news/central/story/2014-08-05/locked-in-syndrome-woman-earns-degree-by-blinking]

Bayes and AI for assistive technology



[https://commons.wikimedia.org/wiki/File:Jean-Dominique-Bauby-01.jpg, https://commons.wikimedia.org/wiki/File:Stephen_Hawking.StarChild.jpg, www.itv.com/news/central/story/2014-08-05/locked-in-syndrome-woman-earns-degree-by-blinking]

Bayes and AI for assistive technology



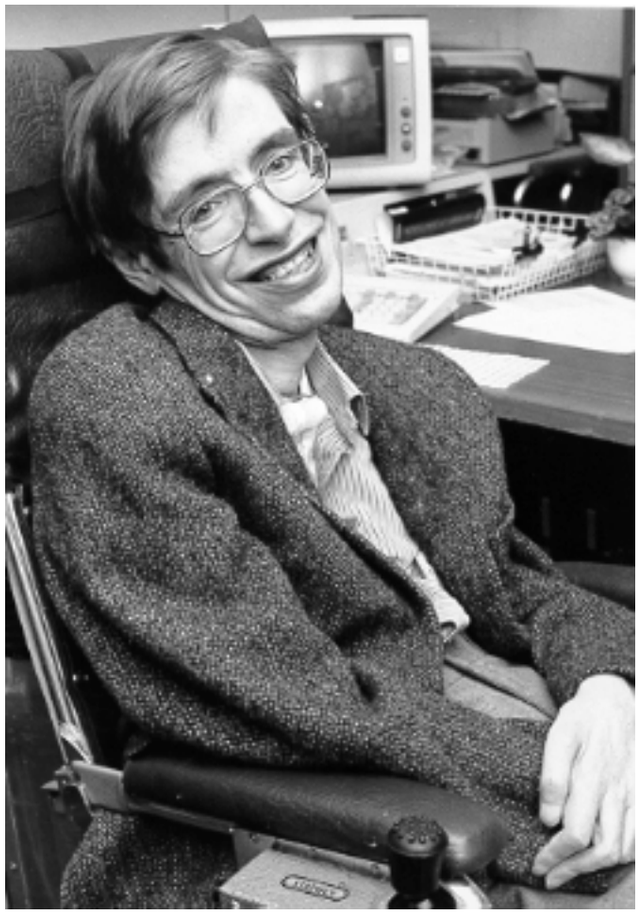


[see profile in May 2026 Sunday Times Magazine article]

[https://commons.wikimedia.org/wiki/File:Jean-Dominique-Bauby-01.jpg, https://commons.wikimedia.org/wiki/File:Stephen_Hawking.StarChild.jpg, www.itv.com/news/central/story/2014-08-05/locked-in-syndrome-woman-earns-degree-by-blinking]

Bayes and AI for assistive technology





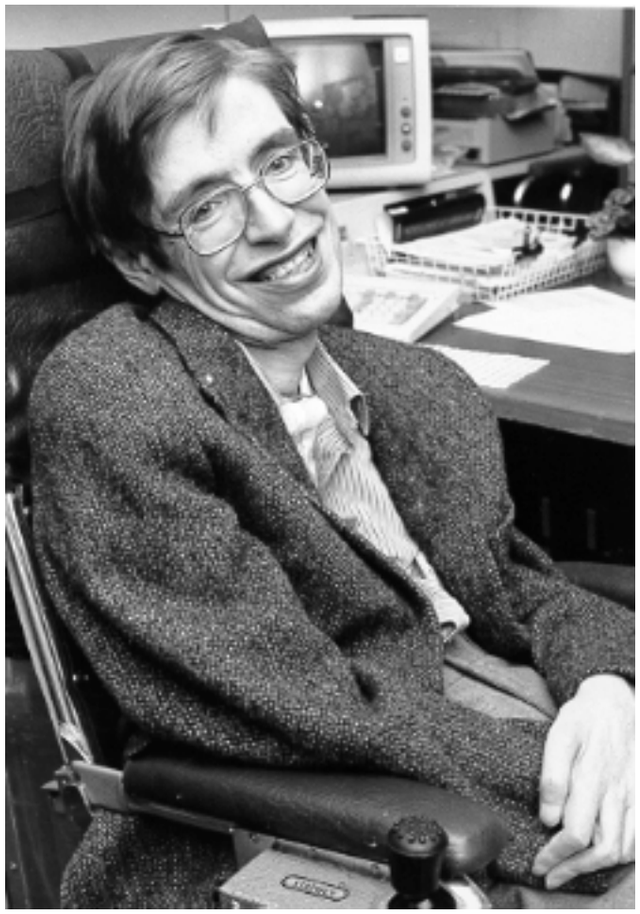


[see profile in May 2026 Sunday Times Magazine article]

[https://commons.wikimedia.org/wiki/File:Jean-Dominique-Bauby-01.jpg, https://commons.wikimedia.org/wiki/File:Stephen_Hawking.StarChild.jpg, www.itv.com/news/central/story/2014-08-05/locked-in-syndrome-woman-earns-degree-by-blinking]

Bayes and AI for assistive technology







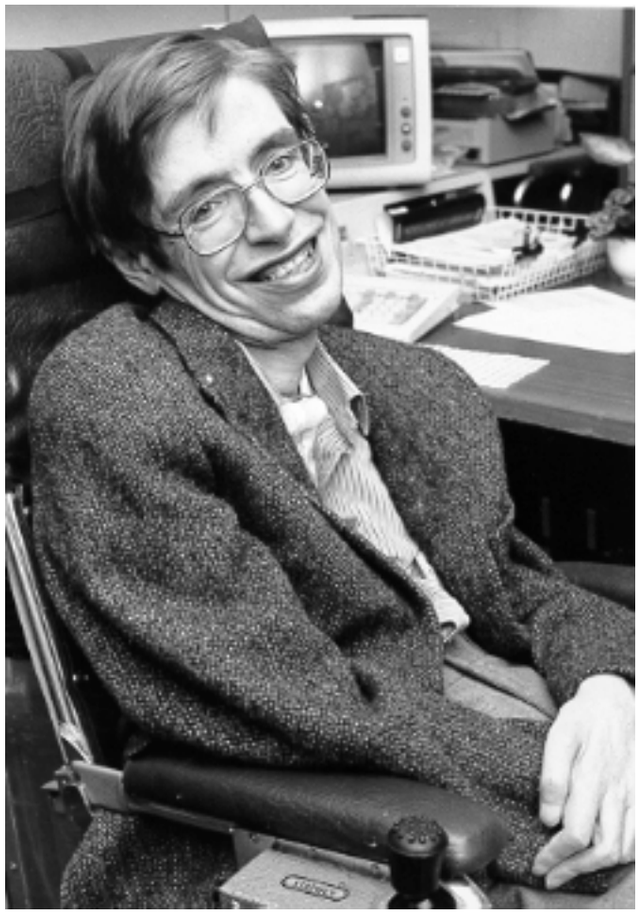
[see profile in May 2026 Sunday Times Magazine article]

- Near-total paralysis: locked-in syndrome, cerebral palsy, advanced ALS, stroke, etc.

[https://commons.wikimedia.org/wiki/File:Jean-Dominique-Bauby-01.jpg, https://commons.wikimedia.org/wiki/File:Stephen_Hawking.StarChild.jpg, www.itv.com/news/central/story/2014-08-05/locked-in-syndrome-woman-earns-degree-by-blinking]

Bayes and AI for assistive technology







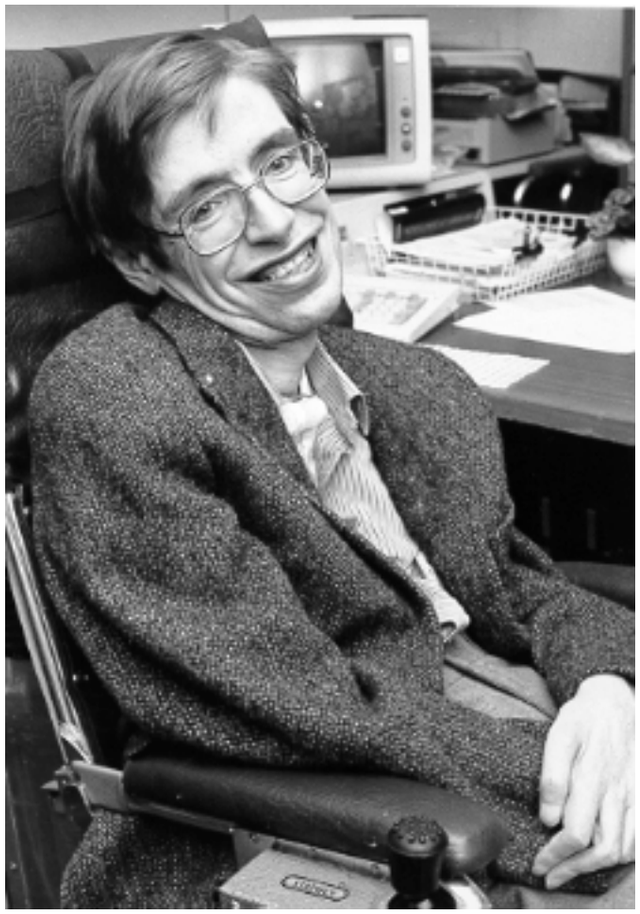
[see profile in May 2026 Sunday Times Magazine article]

- Near-total paralysis: locked-in syndrome, cerebral palsy, advanced ALS, stroke, etc.
- Single switch: blink, button press, puff of air, etc.

[https://commons.wikimedia.org/wiki/File:Jean-Dominique-Bauby-01.jpg, https://commons.wikimedia.org/wiki/File:Stephen_Hawking.StarChild.jpg, www.itv.com/news/central/story/2014-08-05/locked-in-syndrome-woman-earns-degree-by-blinking]

Bayes and AI for assistive technology







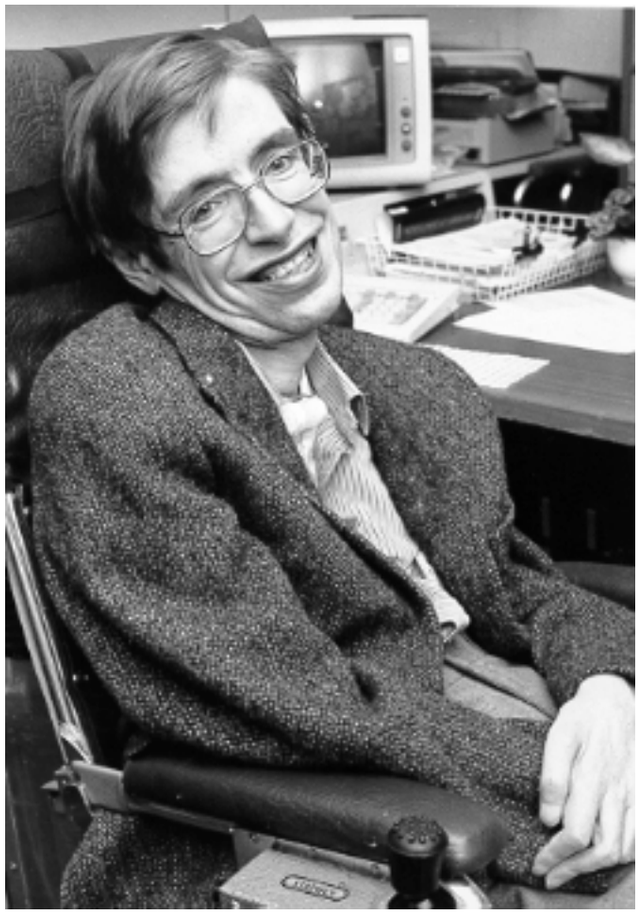
[see profile in May 2026 Sunday Times Magazine article]

- Near-total paralysis: locked-in syndrome, cerebral palsy, advanced ALS, stroke, etc.
- Single switch: blink, button press, puff of air, etc.
- **We provide**: open-source software ("*Nomon*") for writing, gaming, web, drawing, GUIs with a single switch

[https://commons.wikimedia.org/wiki/File:Jean-Dominique-Bauby-01.jpg, https://commons.wikimedia.org/wiki/File:Stephen_Hawking.StarChild.jpg, www.itv.com/news/central/story/2014-08-05/locked-in-syndrome-woman-earns-degree-by-blinking]

Bayes and AI for assistive technology







[see profile in May 2026 Sunday Times Magazine article]

- Near-total paralysis: locked-in syndrome, cerebral palsy, advanced ALS, stroke, etc.
- Single switch: blink, button press, puff of air, etc.
- **We provide**: open-source software ("*Nomon*") for writing, gaming, web, drawing, GUIs with a single switch
  - Language models for language prediction
  - Bayes to infer target from switch input

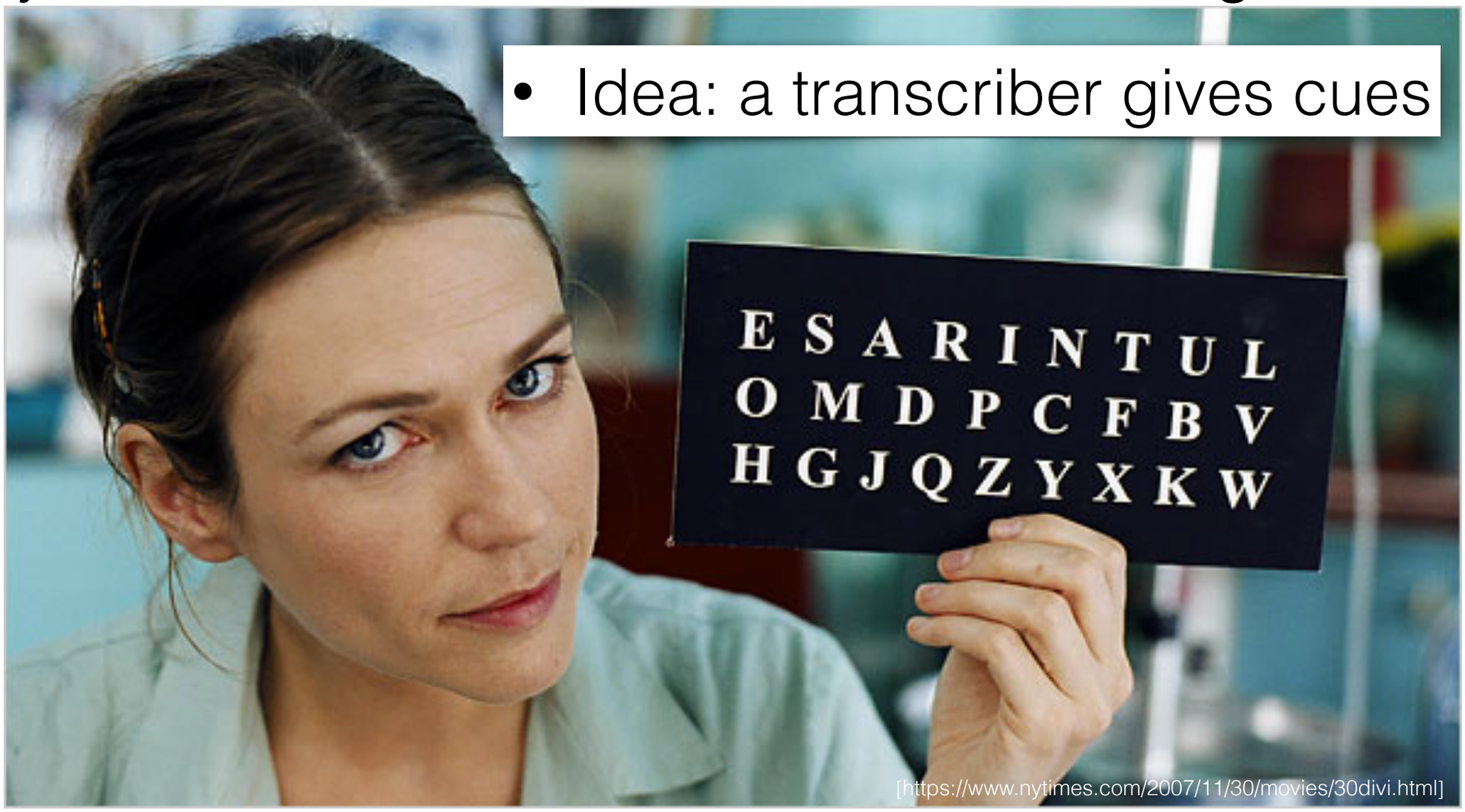
[https://commons.wikimedia.org/wiki/File:Jean-Dominique-Bauby-01.jpg, https://commons.wikimedia.org/wiki/File:Stephen_Hawking.StarChild.jpg, www.itv.com/news/central/story/2014-08-05/locked-in-syndrome-woman-earns-degree-by-blinking]

How to use large language models?

- LLMs are great for predicting characters and words, but they need a start and direction. How can we get that?

How to use large language models?

- LLMs are great for predicting characters and words, but they need a start and direction. How can we get that?

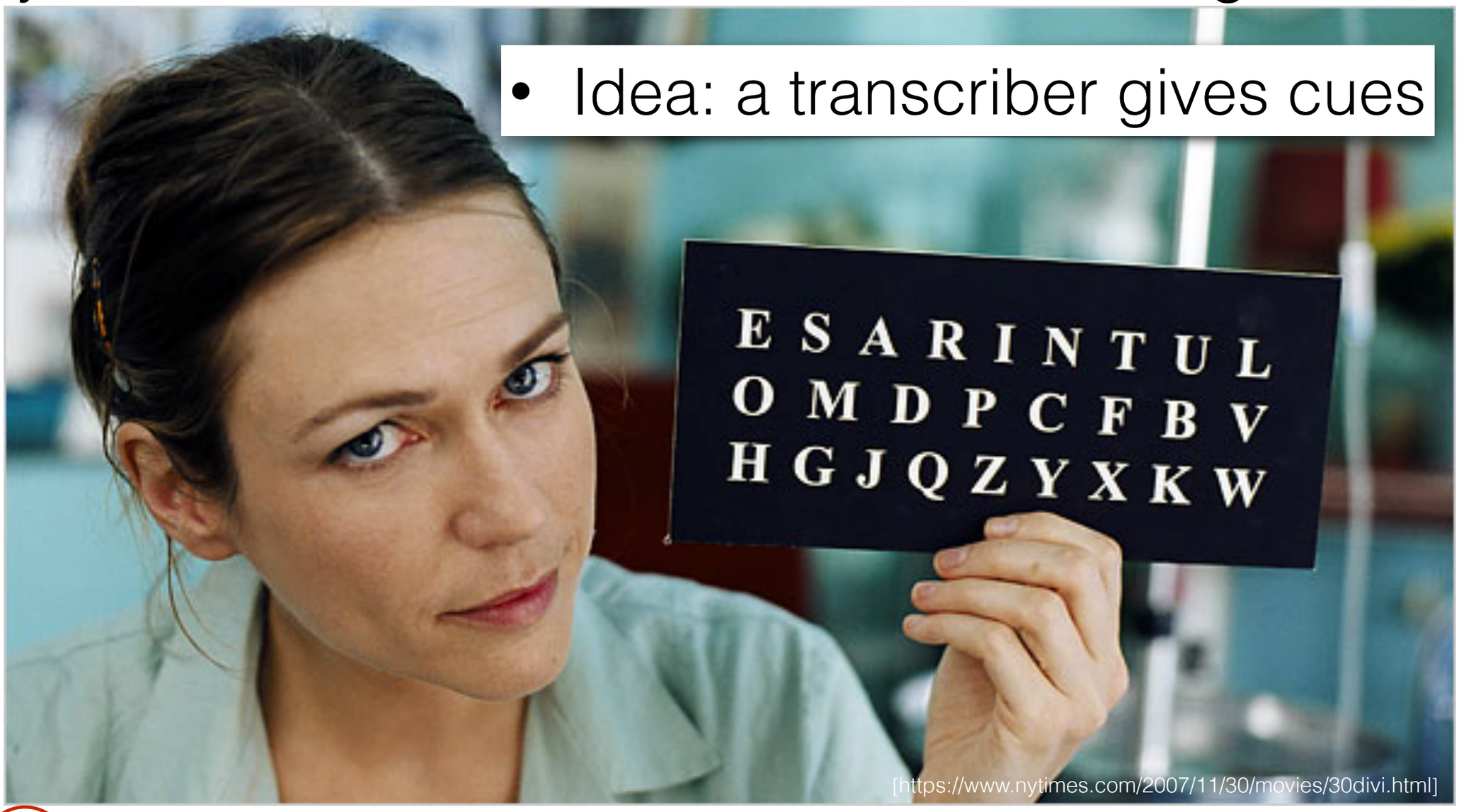


- Idea: a transcriber gives cues

[https://www.nytimes.com/2007/11/30/movies/30divi.html]

How to use large language models?

- LLMs are great for predicting characters and words, but they need a start and direction. How can we get that?

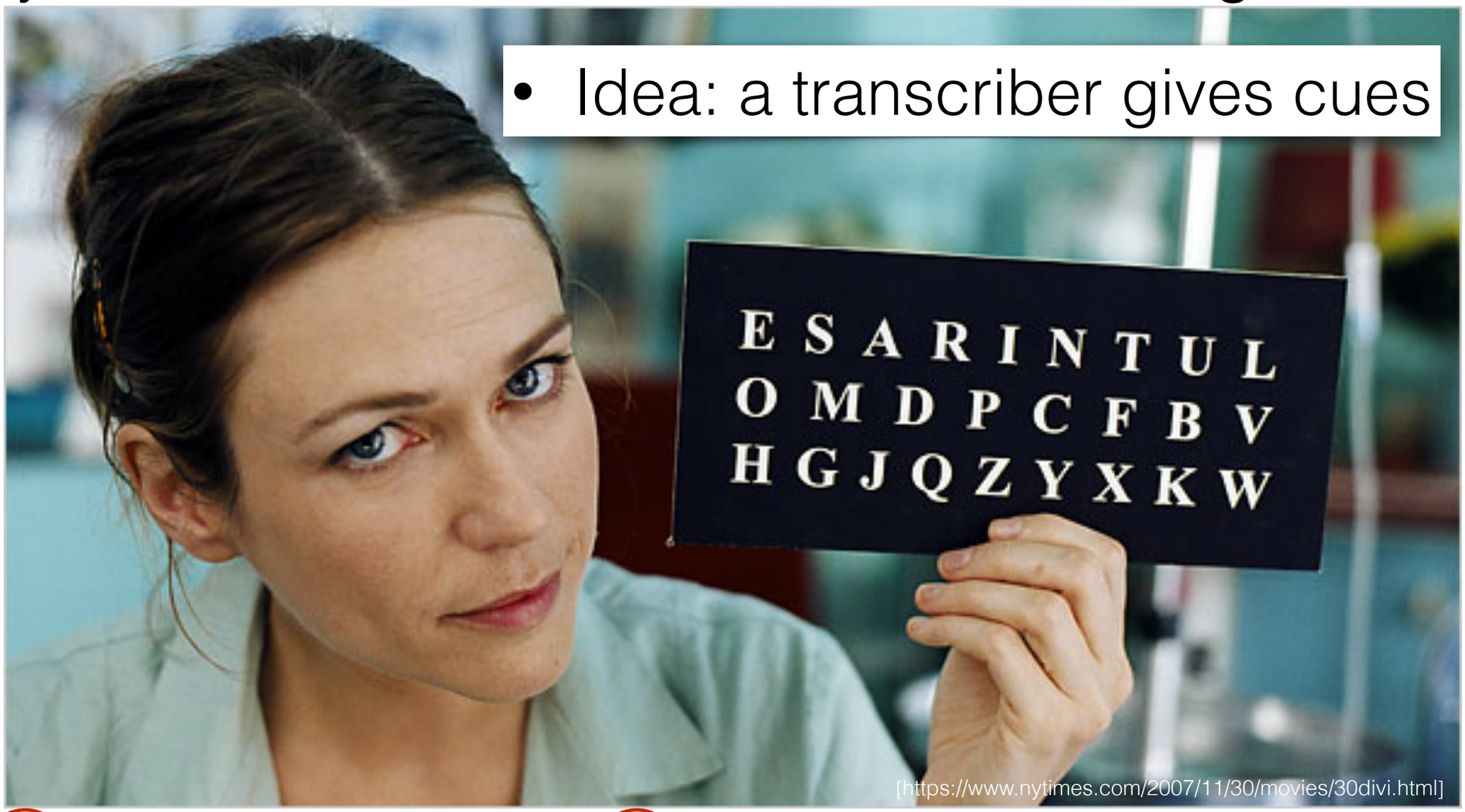


- Idea: a transcriber gives cues

⊗ Need automation

How to use large language models?

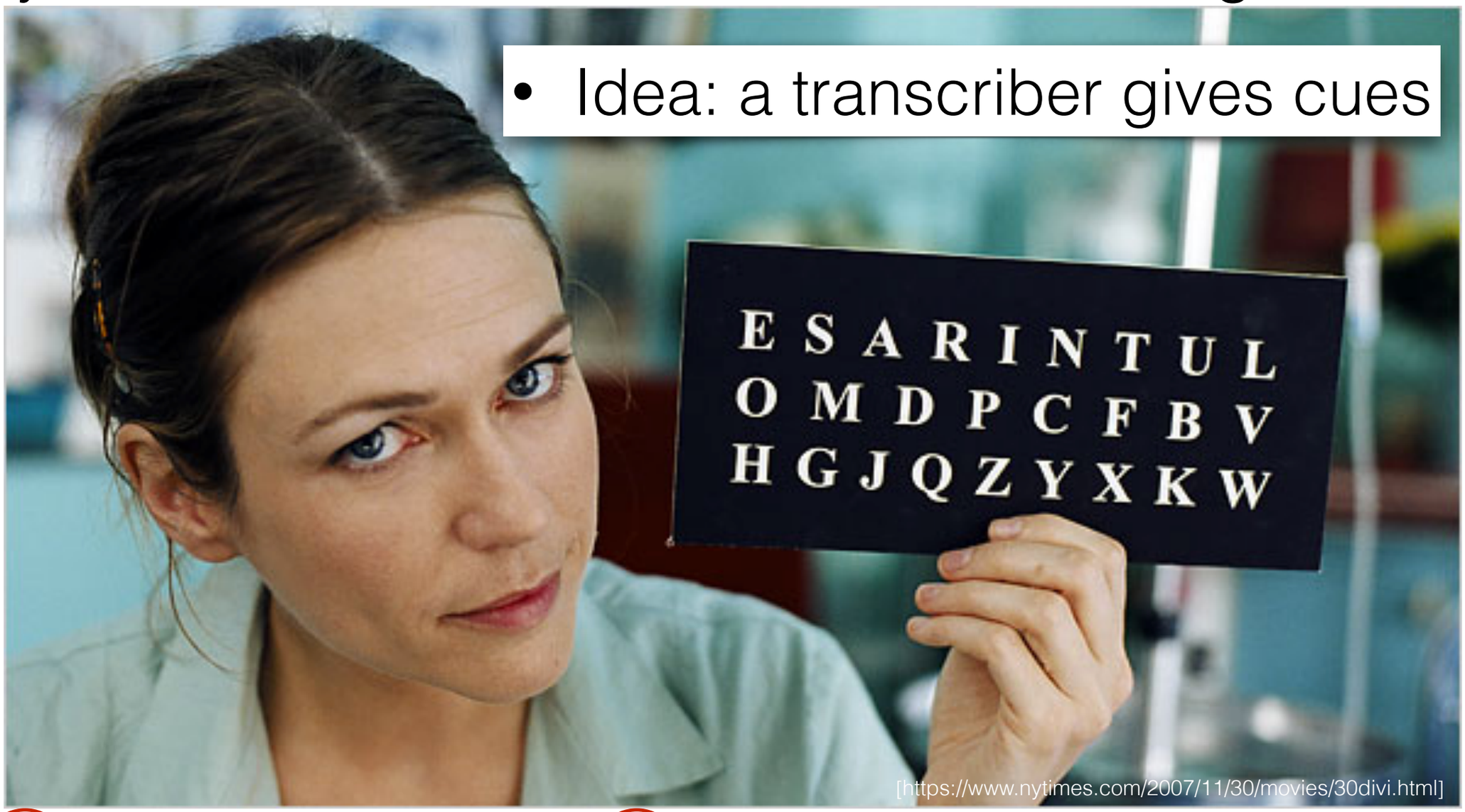
- LLMs are great for predicting characters and words, but they need a start and direction. How can we get that?



⊗ Need automation ⊗ Annoying to use this method

How to use large language models?

- LLMs are great for predicting characters and words, but they need a start and direction. How can we get that?



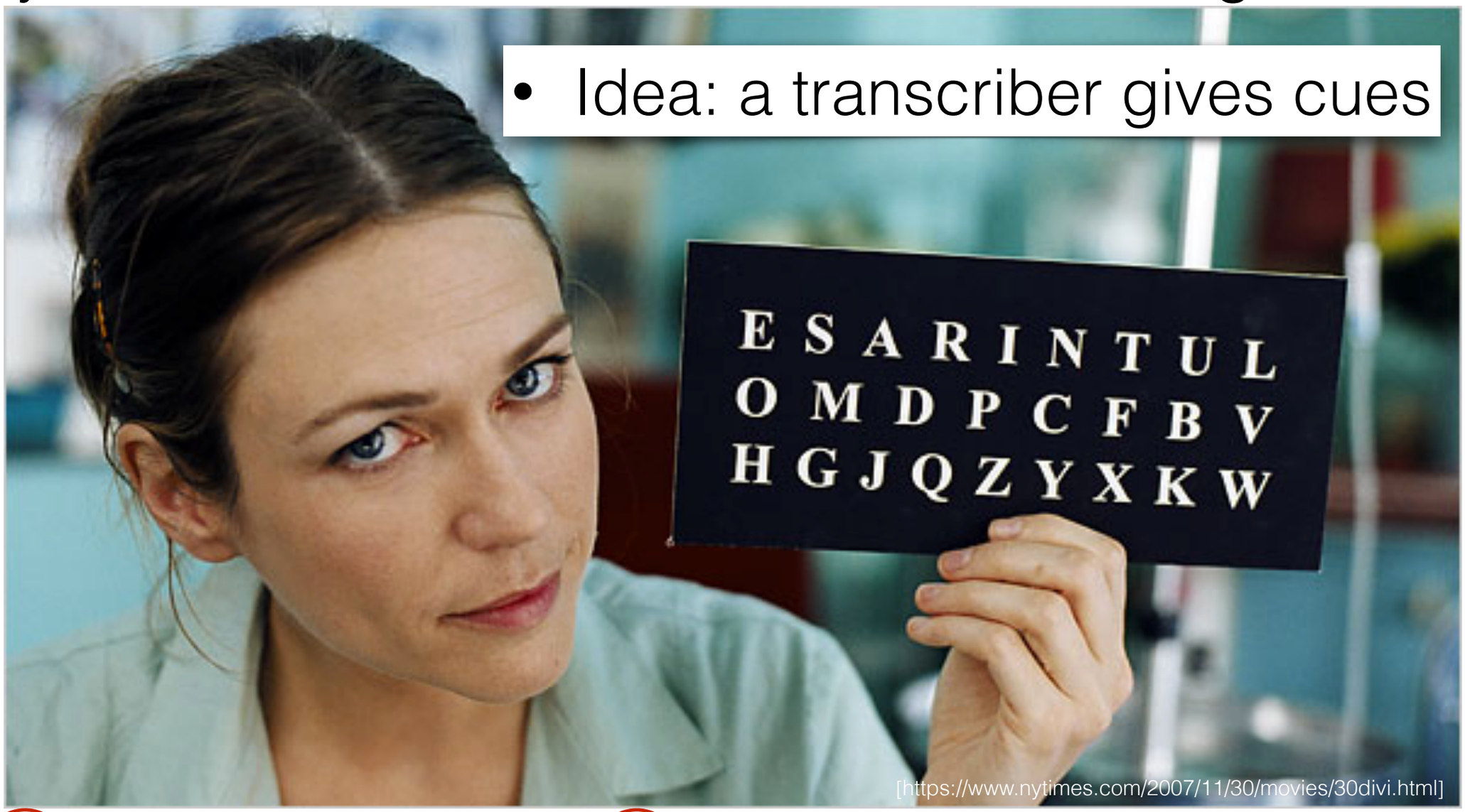
- Idea: a transcriber gives cues

⊗ Need automation ⊗ Annoying to use this method

⊗ Inefficient (words/minute)

How to use large language models?

- LLMs are great for predicting characters and words, but they need a start and direction. How can we get that?

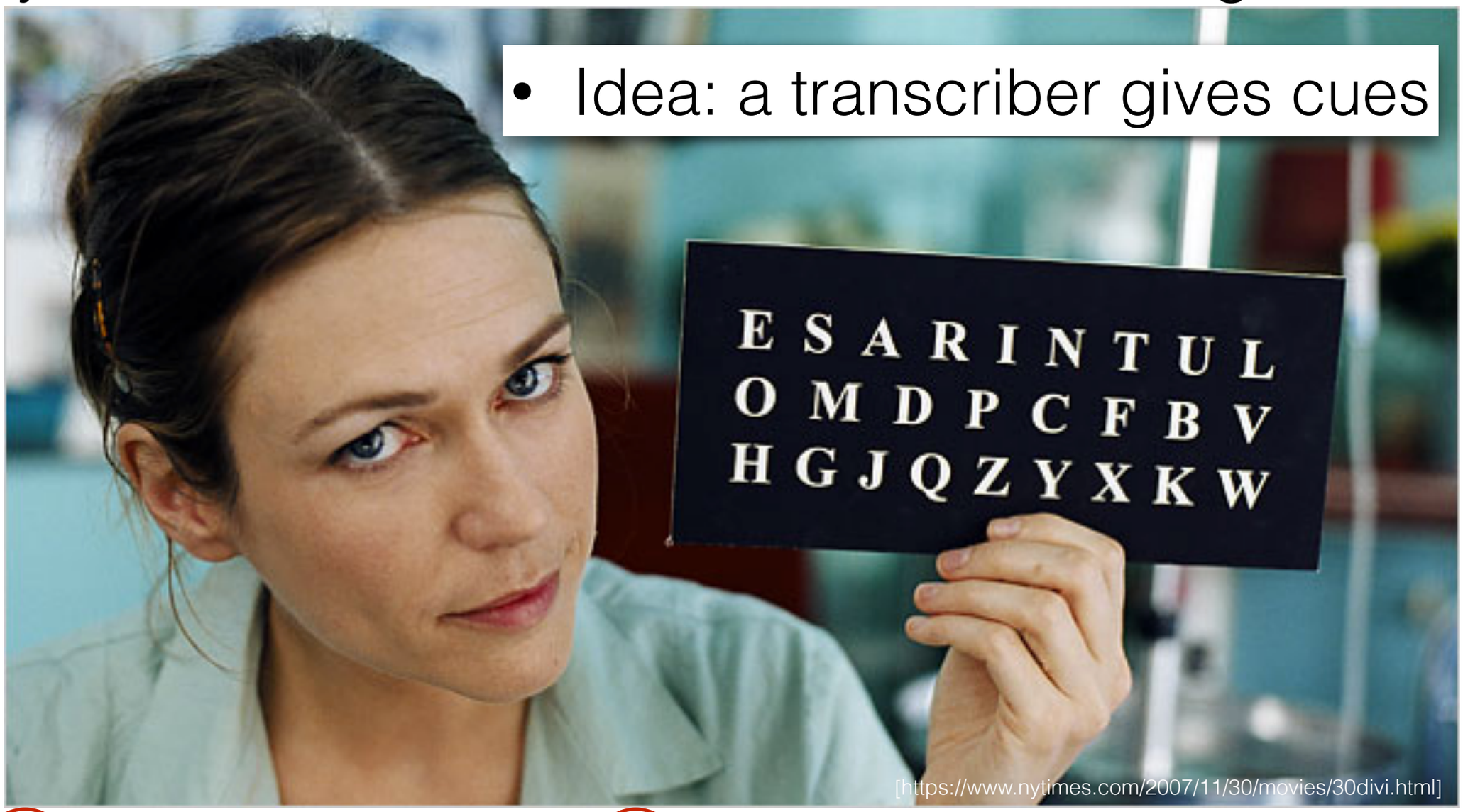


⊗Need automation ⊗Annoying to use this method

⊗Inefficient (words/minute) ⊗How to use LMs

How to use large language models?

- LLMs are great for predicting characters and words, but they need a start and direction. How can we get that?



- Idea: a transcriber gives cues

⊗Need automation ⊗Annoying to use this method

⊗Inefficient (words/minute) ⊗How to use LMs

⊗GUI use beyond text (gaming, web, drawing, etc.)

Row-column scanning



[https://www.theguardian.com/commentisfree/2014/aug/05/in-praise-of-dawn-faizey-webster]

Row-column scanning



and	the	this	do	how	what	my
Undo	e	a	n	d	f	v
t	o	s	l	y	k	.
i	r	u	w	x	‘	
h	c	g	q	_		
m	p	j	Clear			
b	z	Opts				
Back						

[https://www.theguardian.com/commentisfree/2014/aug/05/in-praise-of-dawn-faizey-webster]

Row-column scanning



and	the	this	do	how	what	my
Undo	e	a	n	d	f	v
t	o	s	l	y	k	.
i	r	u	w	x	‘	
h	c	g	q	_		
m	p	j	Clear			
b	z	Opts				
Back						

[https://www.theguardian.com/commentisfree/2014/aug/05/in-praise-of-dawn-faizey-webster]

Row-column scanning



and	the	this	do	how	what	my
Undo	e	a	n	d	f	v
t	o	s	l	y	k	.
i	r	u	w	x	‘	
h	c	g	q	_		
m	p	j	Clear			
b	z	Opts				
Back						

[https://www.theguardian.com/commentisfree/2014/aug/05/in-praise-of-dawn-faizey-webster]

Row-column scanning



and	the	this	do	how	what	my
Undo	e	a	n	d	f	v
t	o	s	l	y	k	.
i	r	u	w	x	‘	
h	c	g	q	_		
m	p	j	Clear			
b	z	Opts				
Back						

[https://www.theguardian.com/commentisfree/2014/aug/05/in-praise-of-dawn-faizey-webster]

Row-column scanning



and	the	this	do	how	what	my
Undo	e	a	n	d	f	v
t	o	s	l	y	k	.
i	r	u	w	x	‘	
h	c	g	q	_		
m	p	j	Clear			
b	z	Opts				
Back						

[https://www.theguardian.com/commentisfree/2014/aug/05/in-praise-of-dawn-faizey-webster]

Row-column scanning



and	the	this	do	how	what	my
Undo	e	a	n	d	f	v
t	o	s	l	y	k	.
i	r	u	w	x	‘	
h	c	g	q	_		
m	p	j	Clear			
b	z	Opts				
Back						

[https://www.theguardian.com/commentisfree/2014/aug/05/in-praise-of-dawn-faizey-webster]

Row-column scanning



and	the	this	do	how	what	my
Undo	e	a	n	d	f	v
t	o	s	l	y	k	.
i	r	u	w	x	‘	
h	c	g	q	_		
m	p	j	Clear			
b	z	Opts				
Back						

[https://www.theguardian.com/commentisfree/2014/aug/05/in-praise-of-dawn-faizey-webster]

Row-column scanning



and	the	this	do	how	what	my
Undo	e	a	n	d	f	v
t	o	s	l	y	k	.
i	r	u	w	x	‘	
h	c	g	q	_		
m	p	j	Clear			
b	z	Opts				
Back						

[https://www.theguardian.com/commentisfree/2014/aug/05/in-praise-of-dawn-faizey-webster]

Row-column scanning



and	the	this	do	how	what	my
Undo	e	a	n	d	f	v
t	o	s	l	y	k	.
i	r	u	w	x	‘	
h	c	g	q	_		
m	p	j	Clear			
b	z	Opts				
Back						

[https://www.theguardian.com/commentisfree/2014/aug/05/in-praise-of-dawn-faizey-webster]

Row-column scanning



and	the	this	do	how	what	my
Undo	e	a	n	d	f	v
t	o	s	l	y	k	.
i	r	u	w	x	‘	
h	c	g	q	_		
m	p	j	Clear			
b	z	Opts				
Back						

Automated

[https://www.theguardian.com/commentisfree/2014/aug/05/in-praise-of-dawn-faizey-webster]

Row-column scanning



and	the	this	do	how	what	my
Undo	e	a	n	d	f	v
t	o	s	l	y	k	.
i	r	u	w	x	‘	
h	c	g	q	_		
m	p	j	Clear			
b	z	Opts				
Back						

- Automated
- Can include text completions, menu options

[https://www.theguardian.com/commentisfree/2014/aug/05/in-praise-of-dawn-faizey-webster]

Row-column scanning



and	the	this	do	how	what	my
Undo	e	a	n	d	f	v
t	o	s	l	y	k	.
i	r	u	w	x	‘	
h	c	g	q	_		
m	p	j	Clear			
b	z	Opts				
Back						

- Automated
- Can include text completions, menu options
  - But must be in a grid (vs. gaming, web, drawing, etc.)

[https://www.theguardian.com/commentisfree/2014/aug/05/in-praise-of-dawn-faizey-webster]

Row-column scanning



and	the	this	do	how	what	my
Undo	e	a	n	d	f	v
t	o	s	l	y	k	.
i	r	u	w	x	‘	
h	c	g	q	_		
m	p	j	Clear			
b	z	Opts				
Back						

- ✓ Automated
- ✓ Can include text completions, menu options
  - ✗ But must be in a grid (vs. gaming, web, drawing, etc.)
- ✓ More time-efficient, especially if add text completions

[https://www.theguardian.com/commentisfree/2014/aug/05/in-praise-of-dawn-faizey-webster]

Row-column scanning



and	the	this	do	how	what	my
Undo	e	a	n	d	f	v
t	o	s	l	y	k	.
i	r	u	w	x	‘	
h	c	g	q	_		
m	p	j	Clear			
b	z	Opts				
Back						

- ✓ Automated
- ✓ Can include text completions, menu options
  - ✗ But must be in a grid (vs. gaming, web, drawing, etc.)
- ✓ More time-efficient, especially if add text completions
  - ✗ But annoying/slow (have to cycle through many options)

[https://www.theguardian.com/commentisfree/2014/aug/05/in-praise-of-dawn-faizey-webster]

Row-column scanning

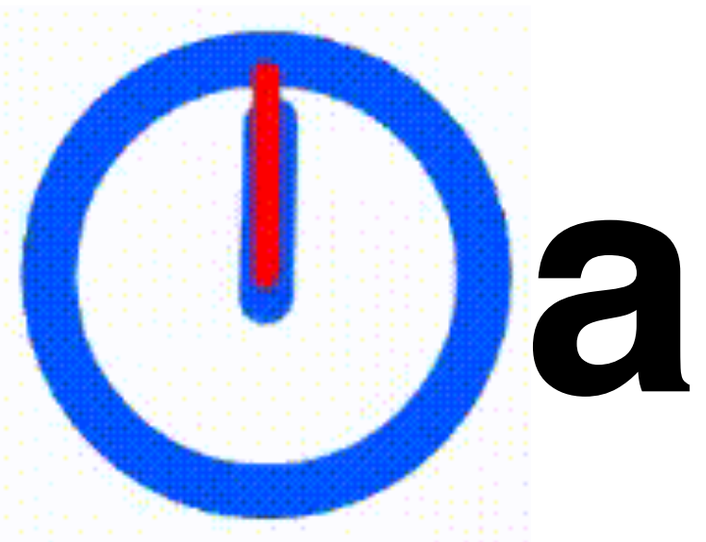


and	the	this	do	how	what	my
Undo	e	a	n	d	f	v
t	o	s	l	y	k	.
i	r	u	w	x	‘	
h	c	g	q	_		
m	p	j	Clear			
b	z	Opts				
Back						

- ✓ Automated
- ✓ Can include text completions, menu options
  - ✗ But must be in a grid (vs. gaming, web, drawing, etc.)
- ✓ More time-efficient, especially if add text completions
  - ✗ But annoying/slow (have to cycle through many options)
- ✗ Have to change focus quickly, continuously (taxing)

[https://www.theguardian.com/commentisfree/2014/aug/05/in-praise-of-dawn-faizey-webster]

Our method: “Nomon”



a

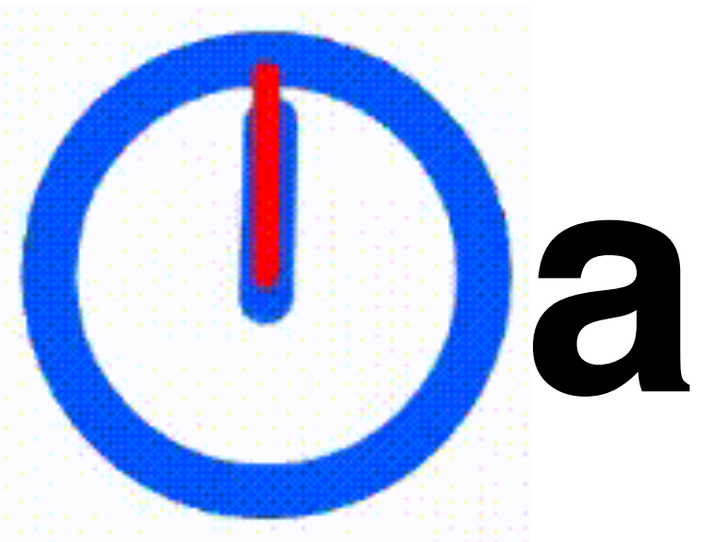


b



c

Our method: “Nomon”

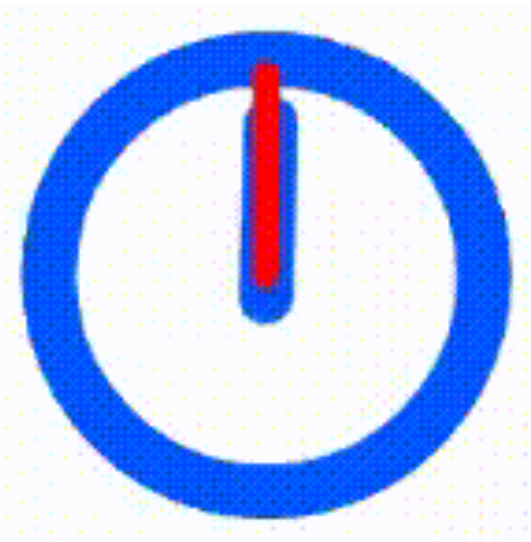

a


b


c

[demo: nomon.app]

Why is Nomon useful?



a

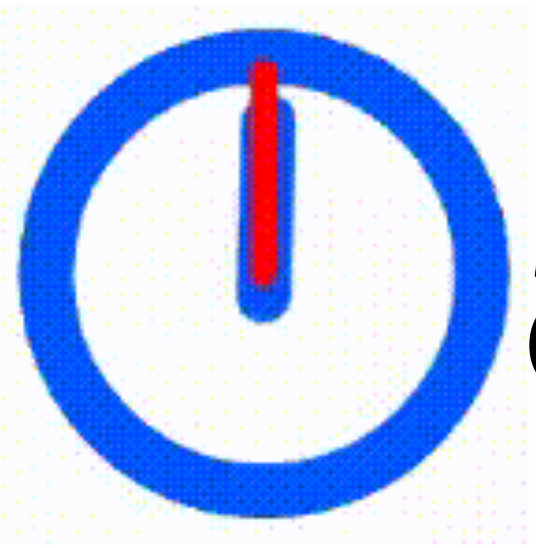


b



c

Why is Nomon useful?



a

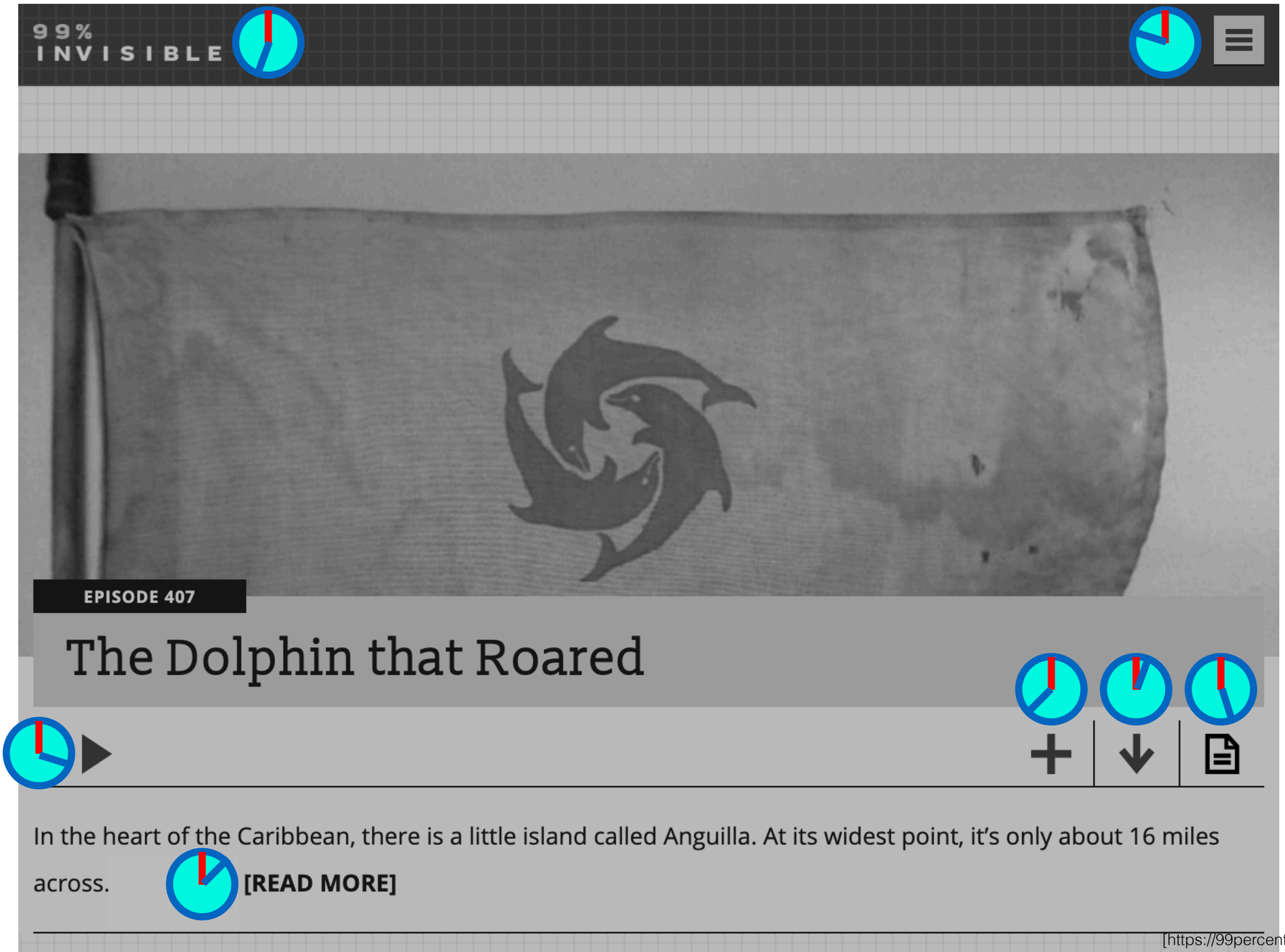


c



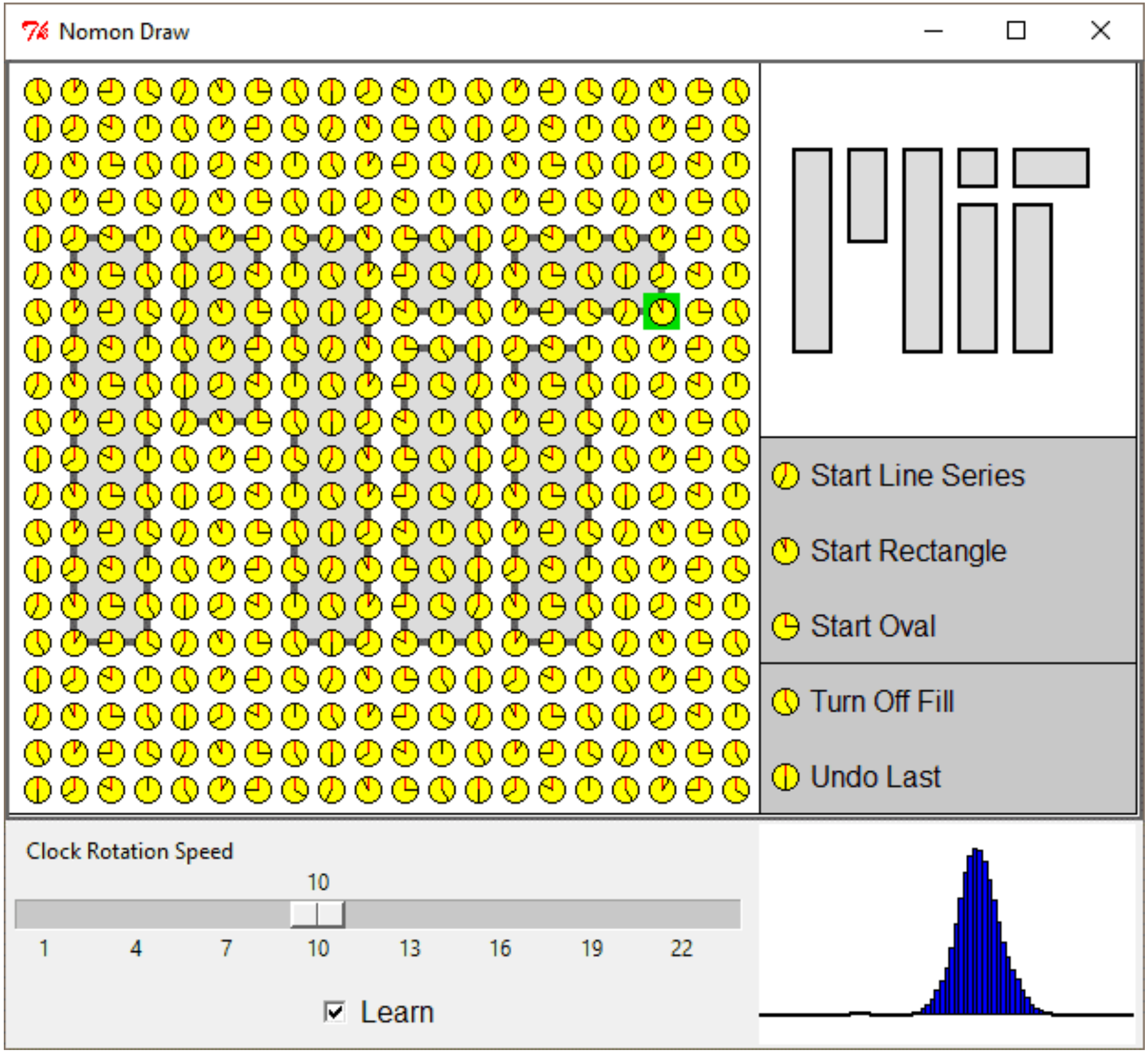
b

Sketch of Nomon for web browsing



[https://99percentinvisible.org/]

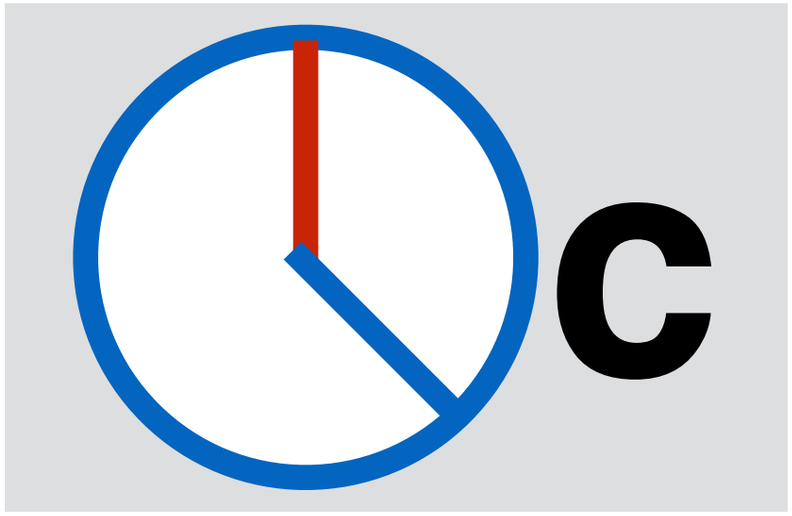
Early version of Nomon for drawing



Why is Nomon useful?



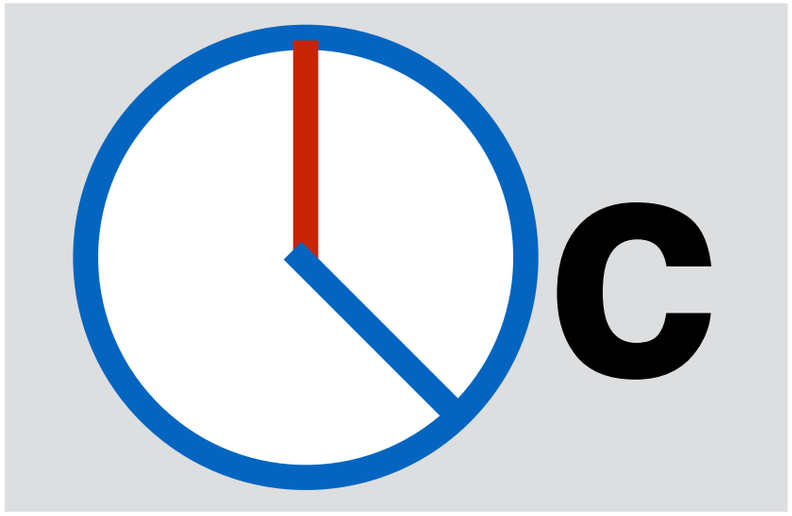




Why is Nomon useful?



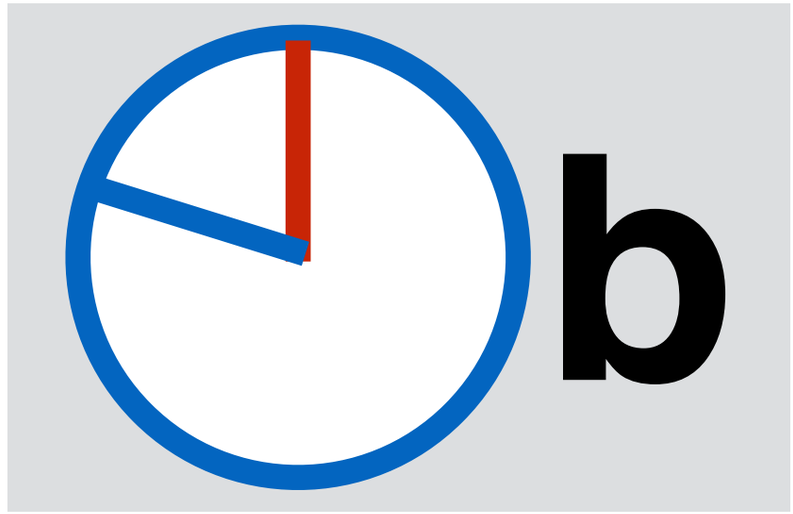


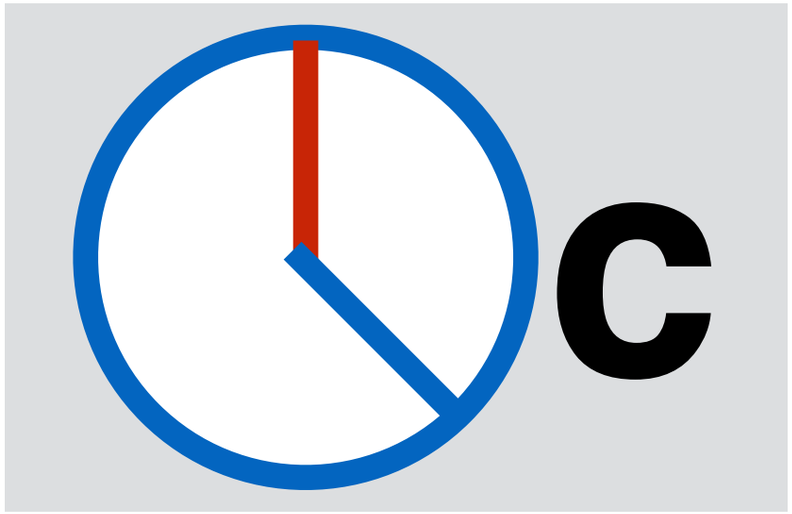


Automated

Why is Nomon useful?





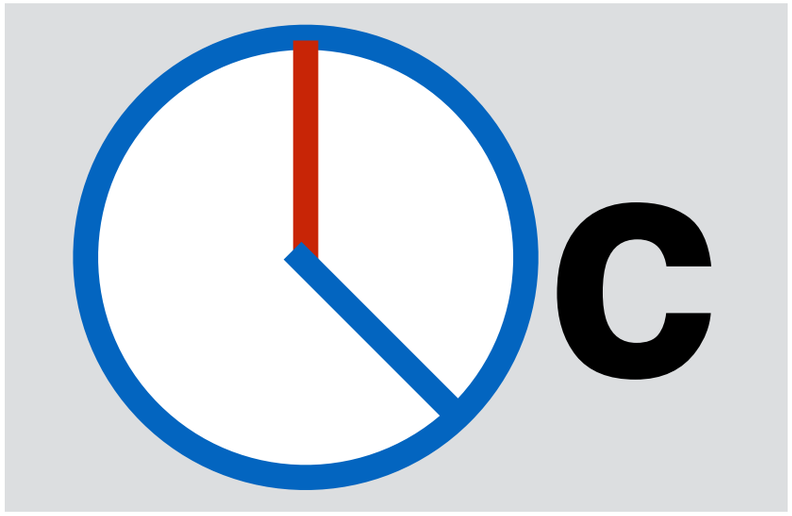


- Automated
- **Flexible**: Can put the clocks anywhere on screen (gaming, web browsing, social media, operating system, drawing, etc.); does not need to be in a grid

Why is Nomon useful?





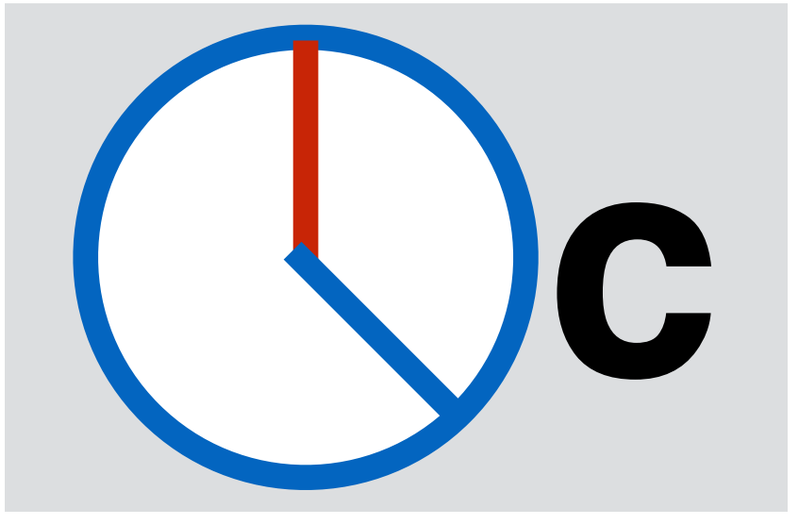


- Automated
- **Flexible**: Can put the clocks anywhere on screen (gaming, web browsing, social media, operating system, drawing, etc.); does not need to be in a grid
- **Relaxed**: can focus on a single clock per selection

Why is Nomon useful?



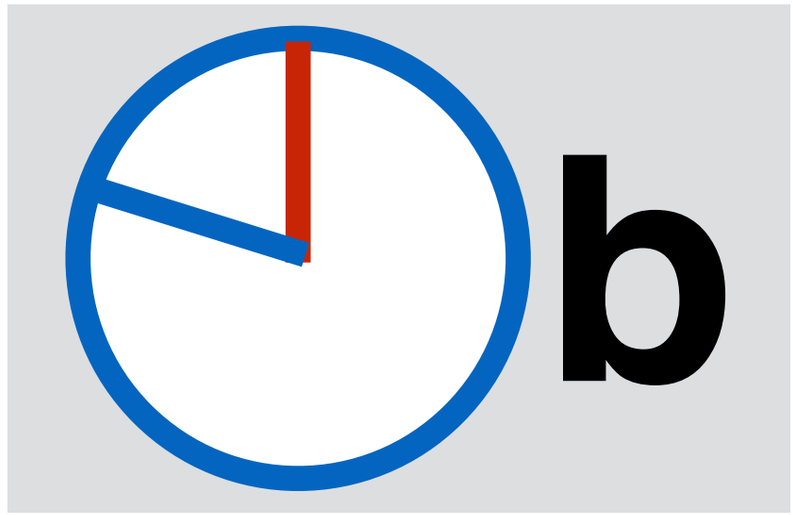


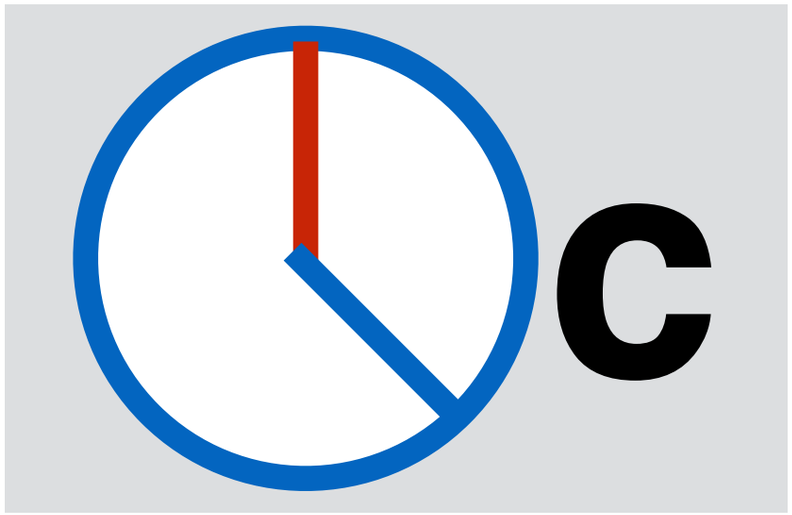


- Automated
- **Flexible**: Can put the clocks anywhere on screen (gaming, web browsing, social media, operating system, drawing, etc.); does not need to be in a grid
- **Relaxed**: can focus on a single clock per selection
- Non-invasive, inexpensive, usable right now (vs. BCI)

Why is Nomon useful?



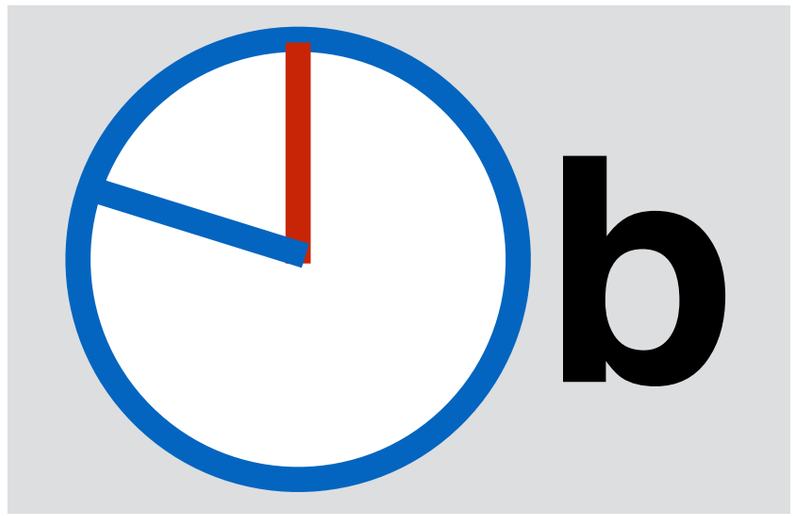


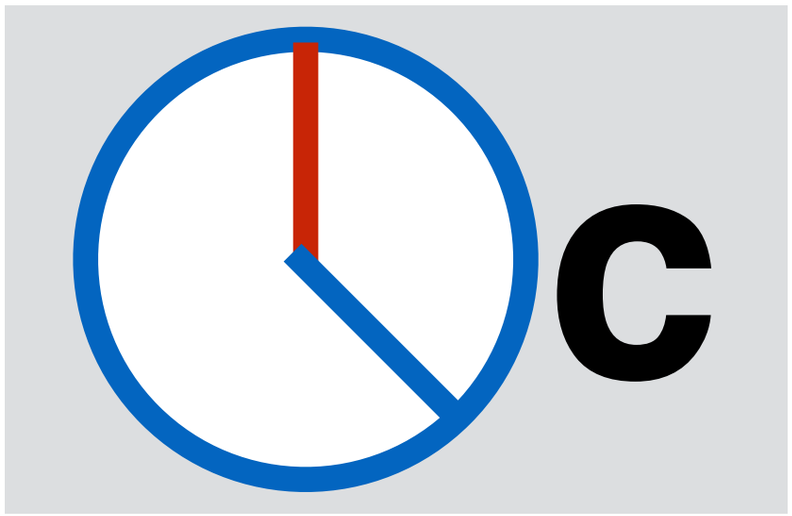


- Automated
- **Flexible**: Can put the clocks anywhere on screen (gaming, web browsing, social media, operating system, drawing, etc.); does not need to be in a grid
- **Relaxed**: can focus on a single clock per selection
- Non-invasive, inexpensive, usable right now (vs. BCI)
- **Efficient** in words per minute: can incorporate language modeling via (1) text completions and (2) clock priors

Why is Nomon useful?







- Automated
- **Flexible**: Can put the clocks anywhere on screen (gaming, web browsing, social media, operating system, drawing, etc.); does not need to be in a grid
- **Relaxed**: can focus on a single clock per selection
- Non-invasive, inexpensive, usable right now (vs. BCI)
- **Efficient** in words per minute: can incorporate language modeling via (1) text completions and (2) clock priors
- **Less "dead time"** than row-column scanning

Nomon: Bayes & language models

Nomon: Bayes & language models

- *c*: desired clock index

Nomon: Bayes & language models

- c: desired clock index; t_r: rth click time

Nomon: Bayes & language models

- c: desired clock index; t_r: rth click time

$$p(c|t_{1:R})$$

Nomon: Bayes & language models

- c: desired clock index; t_r: rth click time

$$p(c|t_{1:R}) \propto p(c)p(t_{1:R}|c)$$

Nomon: Bayes & language models

- c: desired clock index; t_r: rth click time

$$p(c|t_{1:R}) \propto p(c)p(t_{1:R}|c)$$

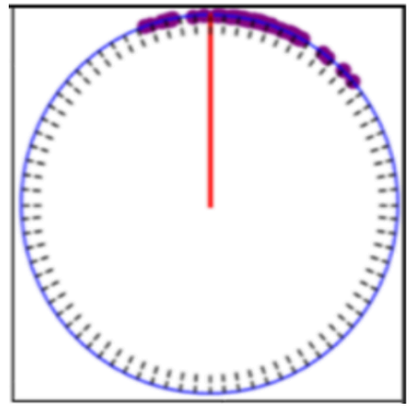
- likelihood

Nomon: Bayes & language models

- c: desired clock index; t_r: rth click time

$$p(c|t_{1:R}) \propto p(c)p(t_{1:R}|c)$$

- likelihood

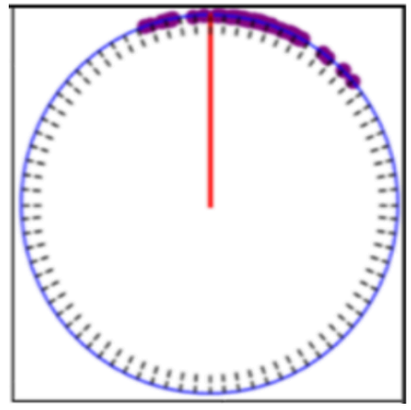


Nomon: Bayes & language models

- c: desired clock index; t_r: rth click time

$$p(c|t_{1:R}) \propto p(c)p(t_{1:R}|c) = p(c)\prod_{r=1}^{R} p(t_r|c)$$

- likelihood

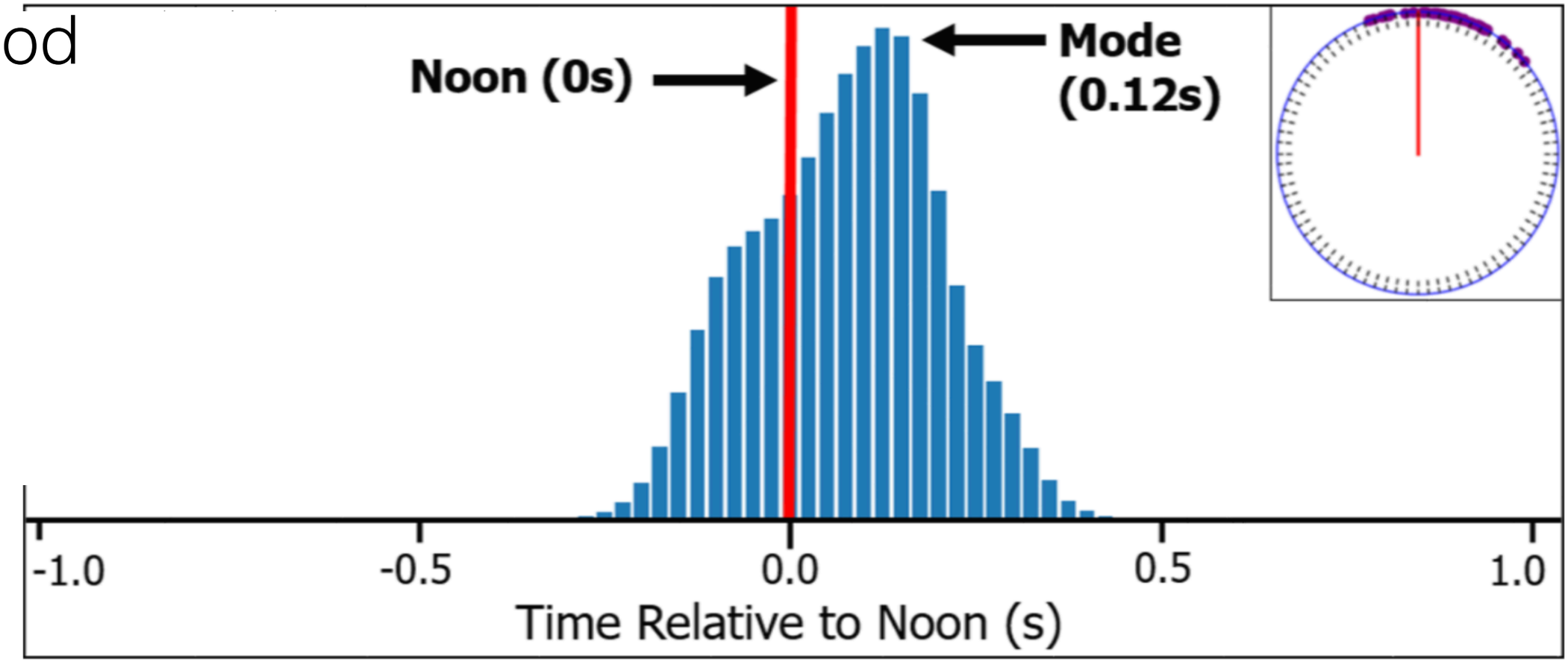


Nomon: Bayes & language models

- c: desired clock index; t_r: rth click time

$$p(c|t_{1:R}) \propto p(c)p(t_{1:R}|c) = p(c)\prod_{r=1}^{R} p(t_r|c)$$

- likelihood

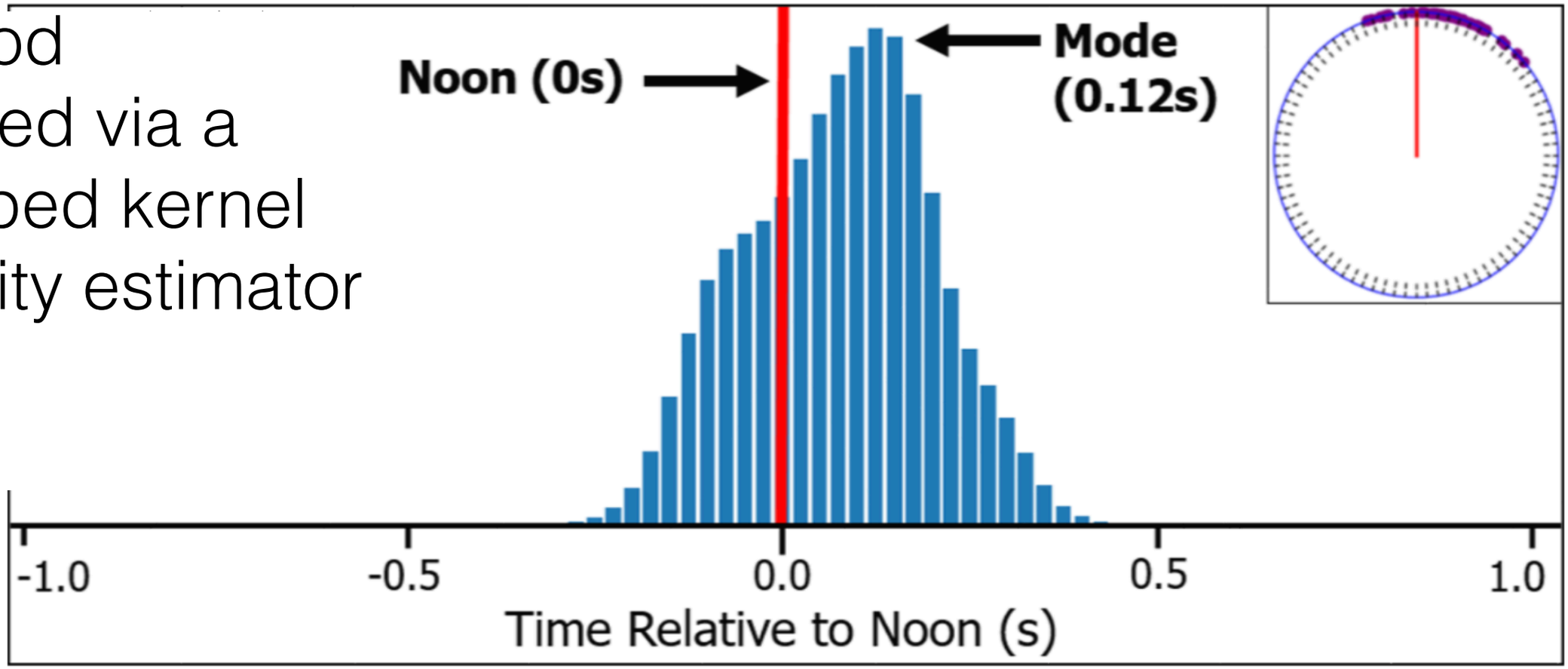


Nomon: Bayes & language models

- c: desired clock index; t_r: rth click time

$$p(c|t_{1:R}) \propto p(c)p(t_{1:R}|c) = p(c)\prod_{r=1}^{R} p(t_r|c)$$

- likelihood
  - learned via a damped kernel density estimator

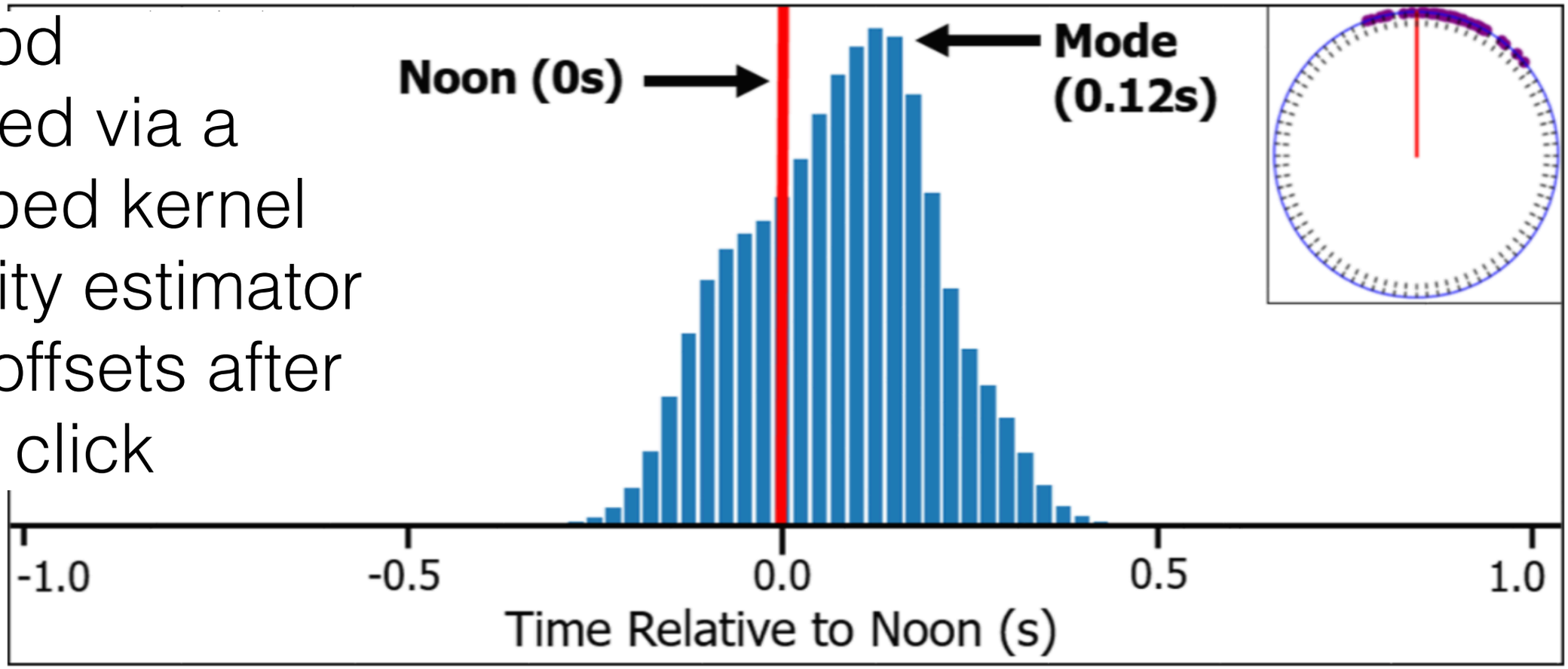


Nomon: Bayes & language models

- c: desired clock index; t_r: rth click time

$$p(c|t_{1:R}) \propto p(c)p(t_{1:R}|c) = p(c)\prod_{r=1}^{R} p(t_r|c)$$

- likelihood
  - learned via a damped kernel density estimator
  - new offsets after each click

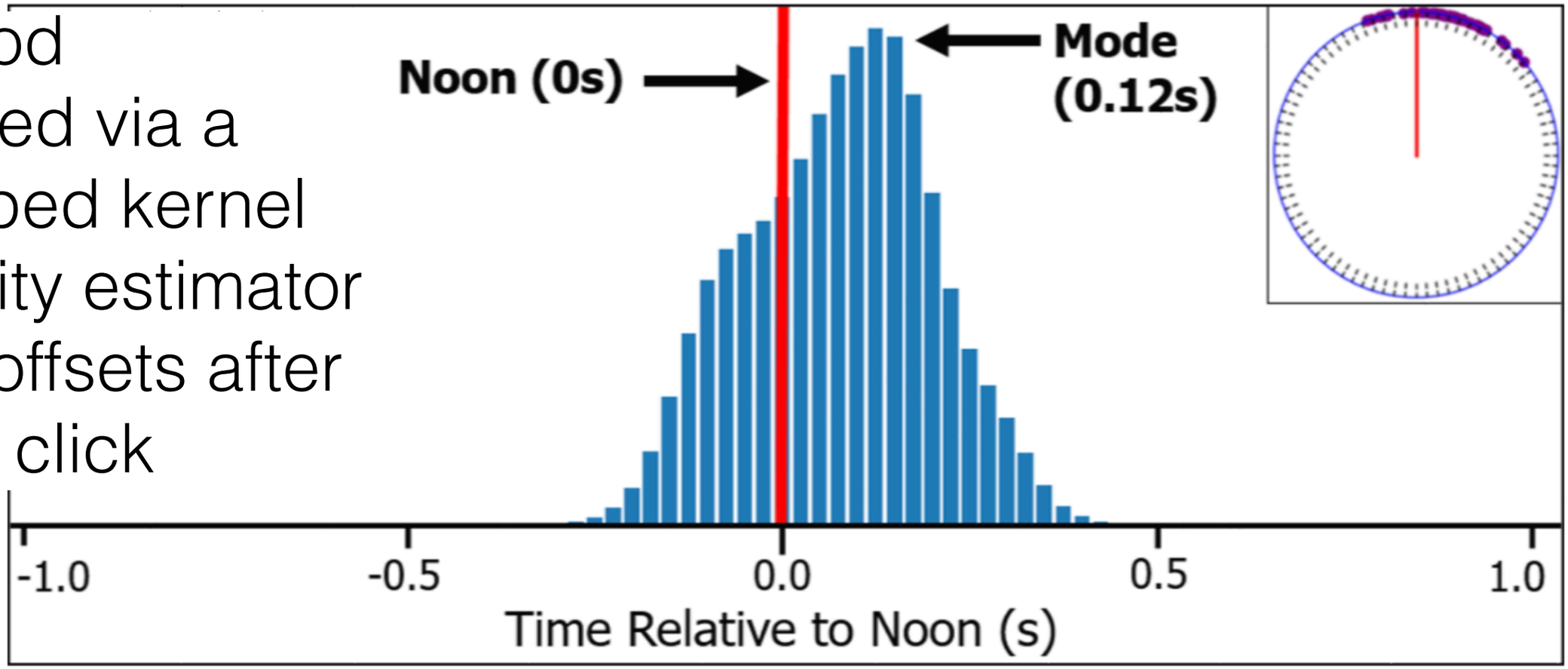


Nomon: Bayes & language models

- c: desired clock index; t_r: rth click time

$$p(c|t_{1:R}) \propto p(c)p(t_{1:R}|c) = p(c)\prod_{r=1}^{R} p(t_r|c)$$

- likelihood
  - learned via a damped kernel density estimator
  - new offsets after each click



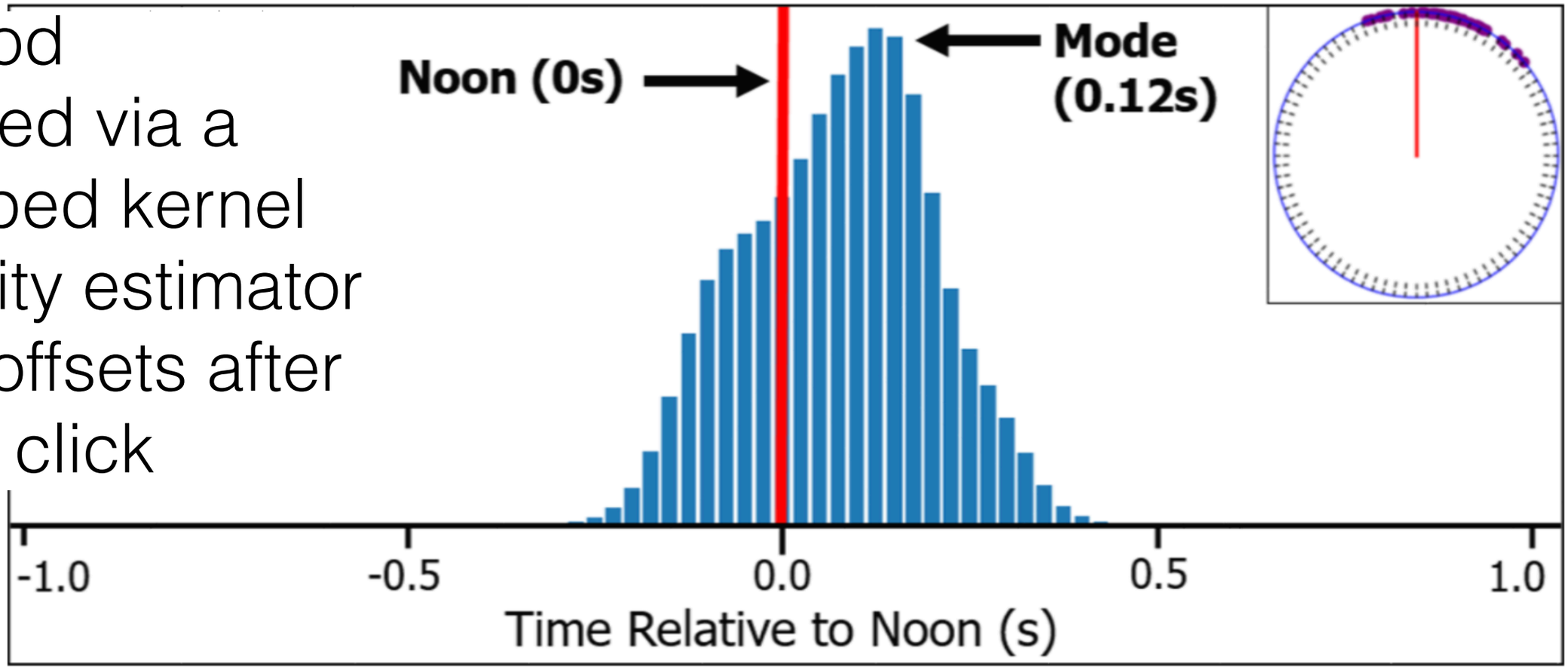
- prior $p(c)$: for writing application, from a language model (both next-character & word probabilities)

Nomon: Bayes & language models

- c: desired clock index; t_r: rth click time

$$p(c|t_{1:R}) \propto p(c)p(t_{1:R}|c) = p(c)\prod_{r=1}^{R} p(t_r|c)$$

- likelihood
  - learned via a damped kernel density estimator
  - new offsets after each click



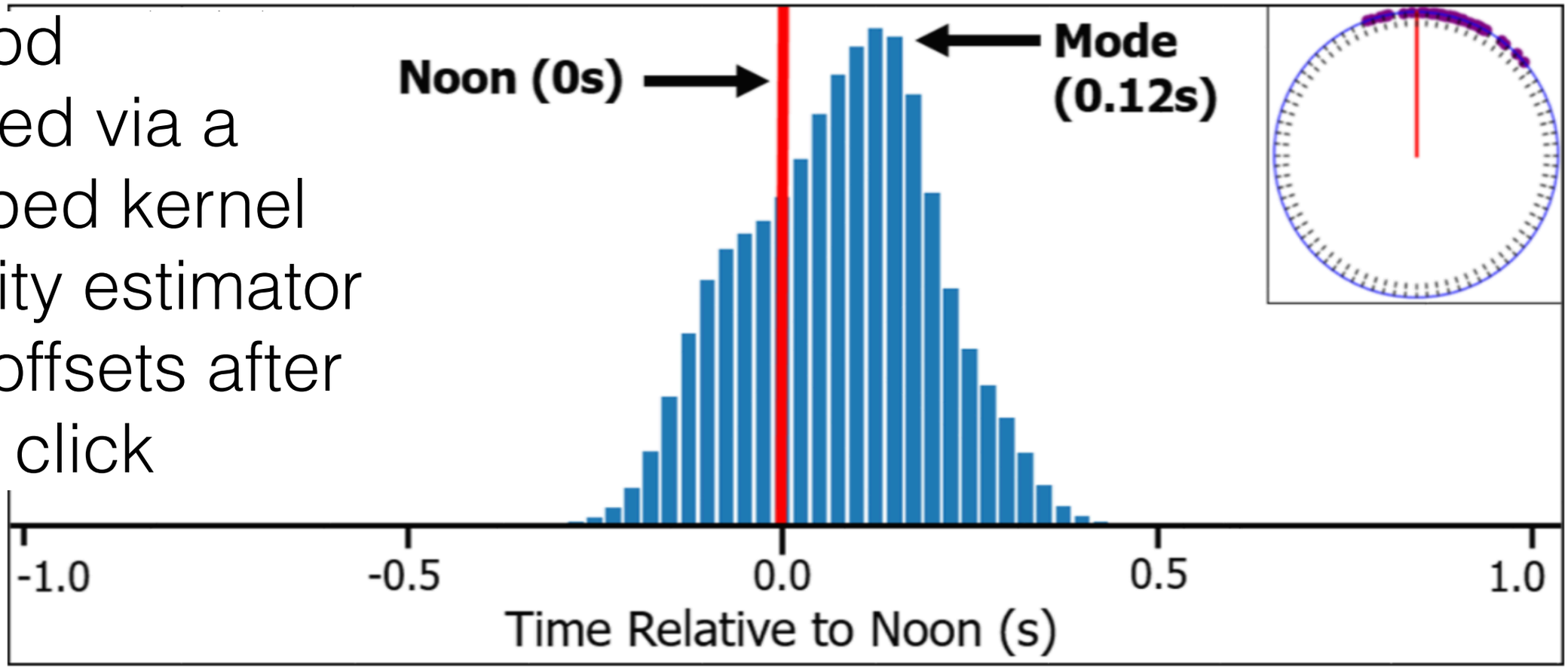
- prior $p(c)$: for writing application, from a language model (both next-character & word probabilities)
- Also choose which word completions to display via LM

Nomon: Bayes & language models

- c: desired clock index; t_r: rth click time

$$p(c|t_{1:R}) \propto p(c)p(t_{1:R}|c) = p(c)\prod_{r=1}^{R} p(t_r|c)$$

- likelihood
  - learned via a damped kernel density estimator
  - new offsets after each click



- prior $p(c)$: for writing application, from a language model (both next-character & word probabilities)
- Also choose which word completions to display via LM
- Challenge isn't (just) compute or modeling given fixed data and parameters; it's *how* to apply Bayes & LM

Why do we need user studies?

Why do we need user studies?

- Evident that Nomon allows more use cases than RCS

Why do we need user studies?

- Evident that Nomon allows more use cases than RCS
  - But need to check for tradeoffs when both apply (ease of use, words per minute, clicks per word)

Why do we need user studies?

- Evident that Nomon allows more use cases than RCS
  - But need to check for tradeoffs when both apply (ease of use, words per minute, clicks per word)
- Goal: avoid wasting time of motor-impaired users

Why do we need user studies?

- Evident that Nomon allows more use cases than RCS
  - But need to check for tradeoffs when both apply (ease of use, words per minute, clicks per word)
- Goal: avoid wasting time of motor-impaired users
  1. Study 13 non–motor-impaired users as check

Why do we need user studies?

- Evident that Nomon allows more use cases than RCS
  - But need to check for tradeoffs when both apply (ease of use, words per minute, clicks per word)
- Goal: avoid wasting time of motor-impaired users
  1. Study 13 non–motor-impaired users as check
  2. Study 7 motor-impaired users (thanks to charities SpecialEffect & The Ace Centre)





Why do we need user studies?

- Evident that Nomon allows more use cases than RCS
  - But need to check for tradeoffs when both apply (ease of use, words per minute, clicks per word)
- Goal: avoid wasting time of motor-impaired users
  1. Study 13 non–motor-impaired users as check
  2. Study 7 motor-impaired users (thanks to charities SpecialEffect & The Ace Centre)



AceCentre

- Participant diagnoses: “complete” spinal cord injury, advanced muscular dystrophy, cerebral palsy/ quadriplegic, Ehlers–Danlos syndrome, spinal muscular dystrophy, stroke/quadriplegic, cerebral palsy

Nomon is useful even before LLMs

Nomon is useful even before LLMs

- Both user studies: (a) writing, (b) choosing pictures

Nomon is useful even before LLMs

- Both user studies: (a) writing, (b) choosing pictures
- Most participants preferred Nomon over RCS

Nomon is useful even before LLMs

- Both user studies: (a) writing, (b) choosing pictures
- Most participants preferred Nomon over RCS
  - Non–motor-impaired: 12 of 13, motor-impaired: 6 of 7

Nomon is useful even before LLMs

- Both user studies: (a) writing, (b) choosing pictures
- Most participants preferred Nomon over RCS
  - Non–motor-impaired: 12 of 13, motor-impaired: 6 of 7
- Nomon: a bit faster entry rate, but a bit more clicking

Nomon is useful even before LLMs

- Both user studies: (a) writing, (b) choosing pictures
- Most participants preferred Nomon over RCS
  - Non–motor-impaired: 12 of 13, motor-impaired: 6 of 7
- Nomon: a bit faster entry rate, but a bit more clicking
- One motor-impaired participant: “it felt faster”

Nomon is useful even before LLMs

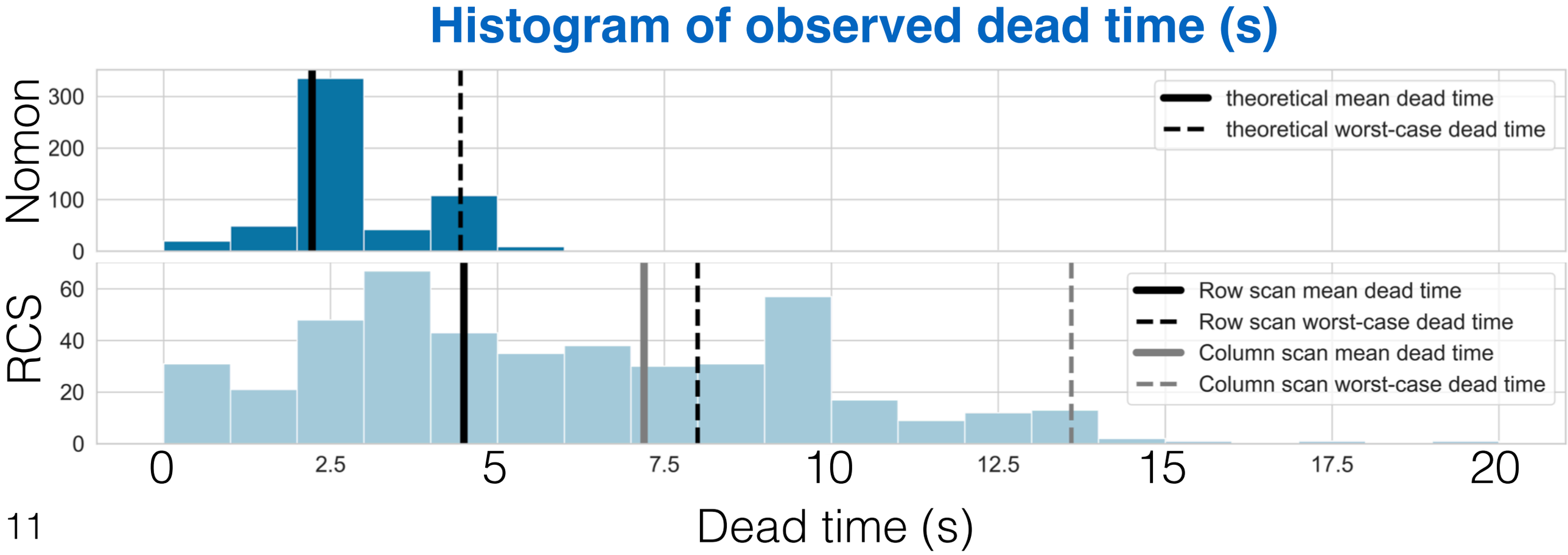
- Both user studies: (a) writing, (b) choosing pictures
- Most participants preferred Nomon over RCS
  - Non–motor-impaired: 12 of 13, motor-impaired: 6 of 7
- Nomon: a bit faster entry rate, but a bit more clicking
- One motor-impaired participant: “it felt faster”
  - All but 1 user: fewer clicks spent correcting with Nomon

Nomon is useful even before LLMs

- Both user studies: (a) writing, (b) choosing pictures
- Most participants preferred Nomon over RCS
  - Non–motor-impaired: 12 of 13, motor-impaired: 6 of 7
- Nomon: a bit faster entry rate, but a bit more clicking
- One motor-impaired participant: “it felt faster”
  - All but 1 user: fewer clicks spent correcting with Nomon
  - Dead time: much better with Nomon

Nomon is useful even before LLMs

- Both user studies: (a) writing, (b) choosing pictures
- Most participants preferred Nomon over RCS
  - Non–motor-impaired: 12 of 13, motor-impaired: 6 of 7
- Nomon: a bit faster entry rate, but a bit more clicking
- One motor-impaired participant: “it felt faster”
  - All but 1 user: fewer clicks spent correcting with Nomon
  - Dead time: much better with Nomon



More directions for Bayes + AI

More directions for Bayes + AI

- In “SureClick” (original) Nomon, a user keeps clicking until the interface is “sure” they wanted a particular clock

More directions for Bayes + AI

- In “SureClick” (original) Nomon, a user keeps clicking until the interface is “sure” they wanted a particular clock
  - Now we’re working on “QuickClick”: one click per character until the language model has a good guess of the word. We have a working prototype (with LLMs).

More directions for Bayes + AI

- In “SureClick” (original) Nomon, a user keeps clicking until the interface is “sure” they wanted a particular clock
  - Now we’re working on “QuickClick”: one click per character until the language model has a good guess of the word. We have a working prototype (with LLMs).
  - In the pipeline: Simulations, real users, target users

More directions for Bayes + AI

- In “SureClick” (original) Nomon, a user keeps clicking until the interface is “sure” they wanted a particular clock
  - Now we’re working on “QuickClick”: one click per character until the language model has a good guess of the word. We have a working prototype (with LLMs).
  - In the pipeline: Simulations, real users, target users
- Open question: how to better use Bayes + LLMs

More directions for Bayes + AI

- In “SureClick” (original) Nomon, a user keeps clicking until the interface is “sure” they wanted a particular clock
  - Now we’re working on “QuickClick”: one click per character until the language model has a good guess of the word. We have a working prototype (with LLMs).
  - In the pipeline: Simulations, real users, target users
- Open question: how to better use Bayes + LLMs
  - Word completions vs longer completions

More directions for Bayes + AI

- In “SureClick” (original) Nomon, a user keeps clicking until the interface is “sure” they wanted a particular clock
  - Now we’re working on “QuickClick”: one click per character until the language model has a good guess of the word. We have a working prototype (with LLMs).
  - In the pipeline: Simulations, real users, target users
- Open question: how to better use Bayes + LLMs
  - Word completions vs longer completions
  - Can we (efficiently & accurately) allow character insertions and deletions?

More directions for Bayes + AI

- In “SureClick” (original) Nomon, a user keeps clicking until the interface is “sure” they wanted a particular clock
  - Now we’re working on “QuickClick”: one click per character until the language model has a good guess of the word. We have a working prototype (with LLMs).
  - In the pipeline: Simulations, real users, target users
- Open question: how to better use Bayes + LLMs
  - Word completions vs longer completions
  - Can we (efficiently & accurately) allow character insertions and deletions?
  - SureClick is truly single-switch. Right now QuickClick needs two separate switches. Could it be one?

More directions for Bayes + AI

- In “SureClick” (original) Nomon, a user keeps clicking until the interface is “sure” they wanted a particular clock
  - Now we’re working on “QuickClick”: one click per character until the language model has a good guess of the word. We have a working prototype (with LLMs).
  - In the pipeline: Simulations, real users, target users
- Open question: how to better use Bayes + LLMs
  - Word completions vs longer completions
  - Can we (efficiently & accurately) allow character insertions and deletions?
  - SureClick is truly single-switch. Right now QuickClick needs two separate switches. Could it be one?
  - Everything needs to work in real time

More directions for Bayes + AI

- In “SureClick” (original) Nomon, a user keeps clicking until the interface is “sure” they wanted a particular clock
  - Now we’re working on “QuickClick”: one click per character until the language model has a good guess of the word. We have a working prototype (with LLMs).
  - In the pipeline: Simulations, real users, target users
- Open question: how to better use Bayes + LLMs
  - Word completions vs longer completions
  - Can we (efficiently & accurately) allow character insertions and deletions?
  - SureClick is truly single-switch. Right now QuickClick needs two separate switches. Could it be one?
  - Everything needs to work in real time
- Open: Improved user interfaces (w modern coding tools)
  - + More (robust) capabilities: gaming, social media, web browsing, using operating system, drawing, etc.

Air pollution forecasts for health

Air pollution forecasts for health

- We compare 6 tools: 4 from US and Europe governments, including CAMS (used by Google & Apple) + 1 ensemble

Air pollution forecasts for health

- We compare 6 tools: 4 from US and Europe governments, including CAMS (used by Google & Apple) + 1 ensemble
- Aurora: AI foundation model trained for weather forecasting and fine-tuned for air pollution forecasting

Air pollution forecasts for health

- We compare 6 tools: 4 from US and Europe governments, including CAMS (used by Google & Apple) + 1 ensemble
- Aurora: AI foundation model trained for weather forecasting and fine-tuned for air pollution forecasting

See paper for analysis over many places and times

Air pollution forecasts for health

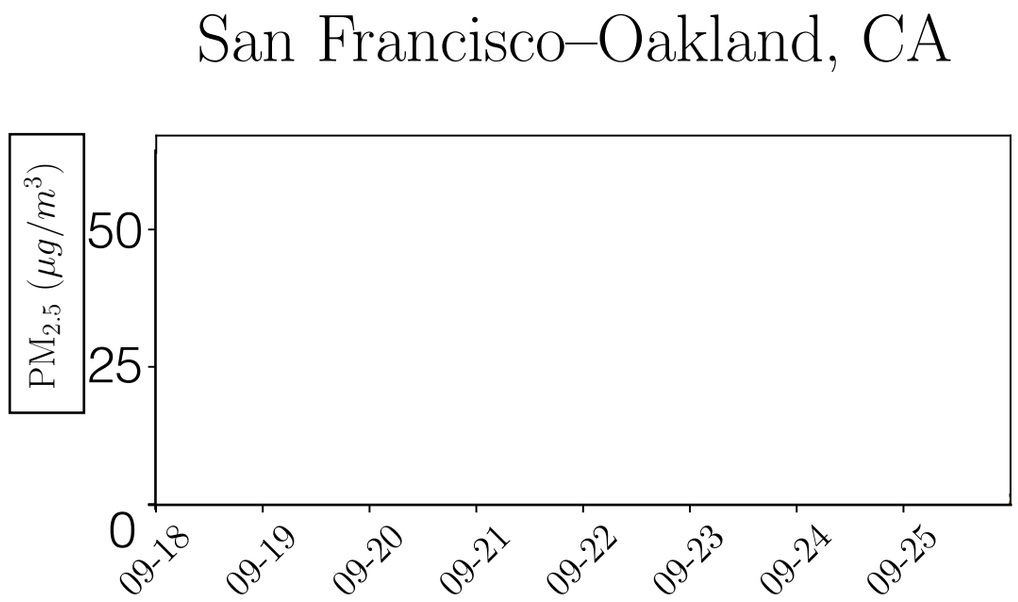
- We compare 6 tools: 4 from US and Europe governments, including CAMS (used by Google & Apple) + 1 ensemble
- Aurora: AI foundation model trained for weather forecasting and fine-tuned for air pollution forecasting

San Francisco–Oakland, CA

See paper for analysis over many places and times

Air pollution forecasts for health

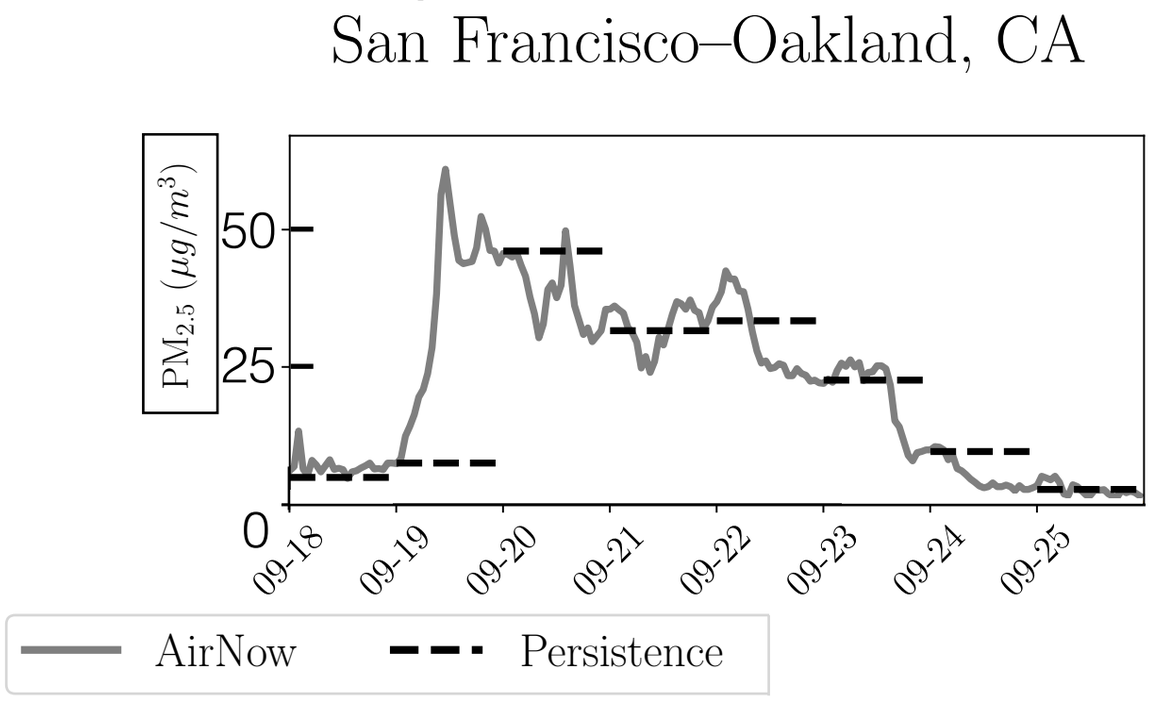
- We compare 6 tools: 4 from US and Europe governments, including CAMS (used by Google & Apple) + 1 ensemble
- Aurora: AI foundation model trained for weather forecasting and fine-tuned for air pollution forecasting



See paper for analysis over many places and times

Air pollution forecasts for health

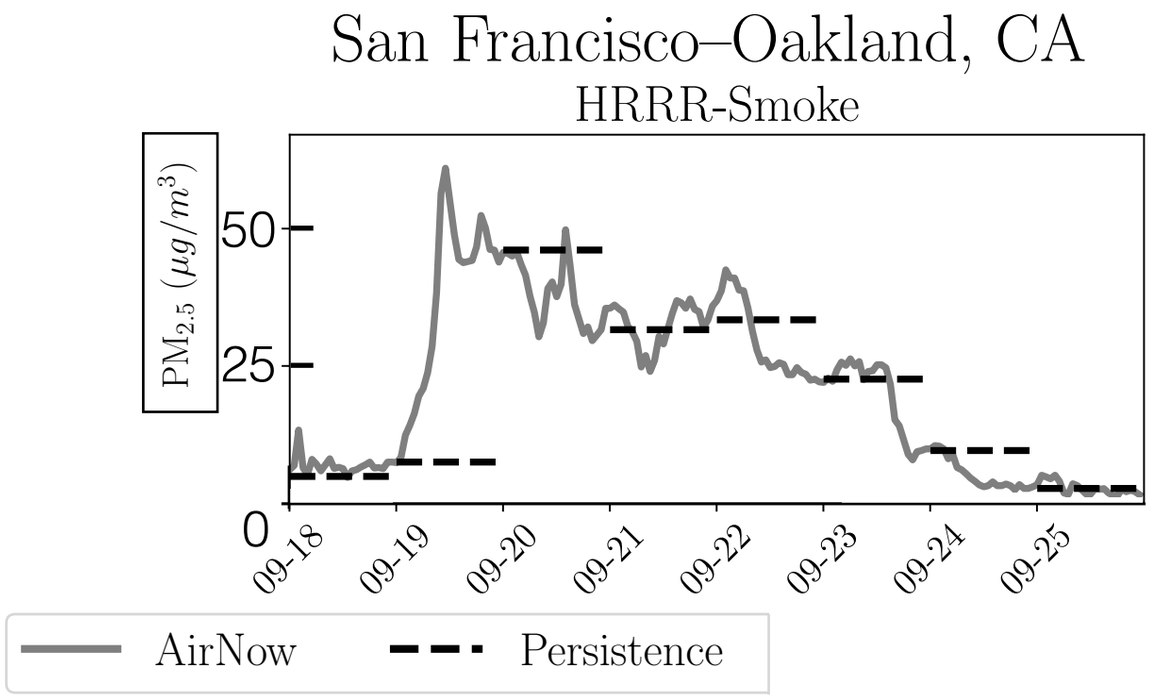
- We compare 6 tools: 4 from US and Europe governments, including CAMS (used by Google & Apple) + 1 ensemble
- Aurora: AI foundation model trained for weather forecasting and fine-tuned for air pollution forecasting



See paper for analysis over many places and times

Air pollution forecasts for health

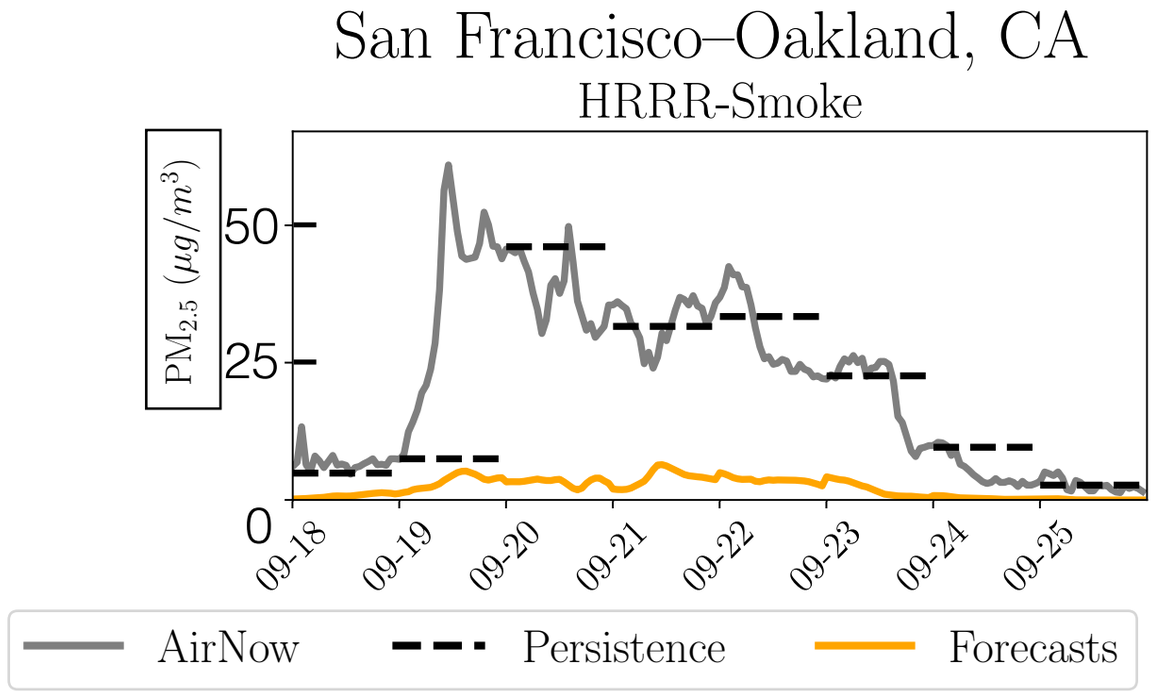
- We compare 6 tools: 4 from US and Europe governments, including CAMS (used by Google & Apple) + 1 ensemble
- Aurora: AI foundation model trained for weather forecasting and fine-tuned for air pollution forecasting



See paper for analysis over many places and times

Air pollution forecasts for health

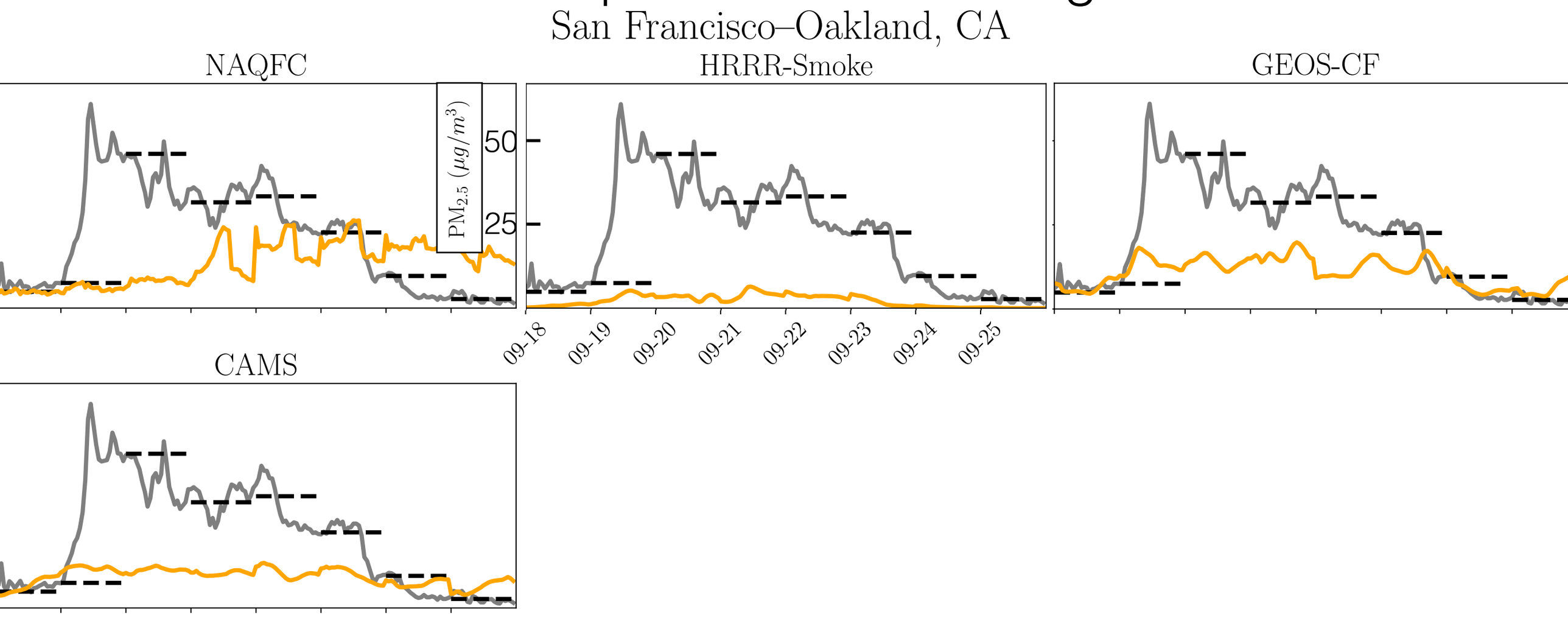
- We compare 6 tools: 4 from US and Europe governments, including CAMS (used by Google & Apple) + 1 ensemble
- Aurora: AI foundation model trained for weather forecasting and fine-tuned for air pollution forecasting



See paper for analysis over many places and times

Air pollution forecasts for health

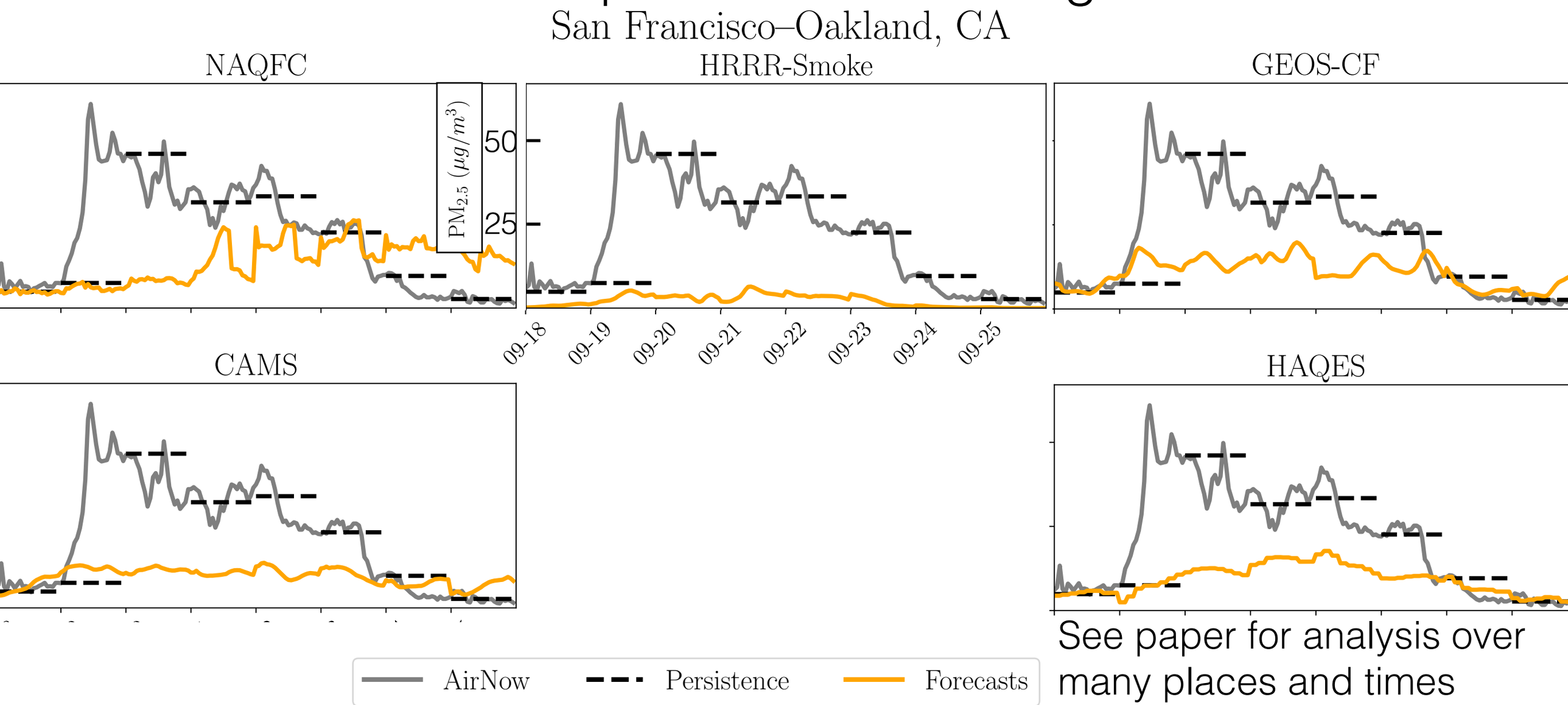
- We compare 6 tools: 4 from US and Europe governments, including CAMS (used by Google & Apple) + 1 ensemble
- Aurora: AI foundation model trained for weather forecasting and fine-tuned for air pollution forecasting



See paper for analysis over many places and times

Air pollution forecasts for health

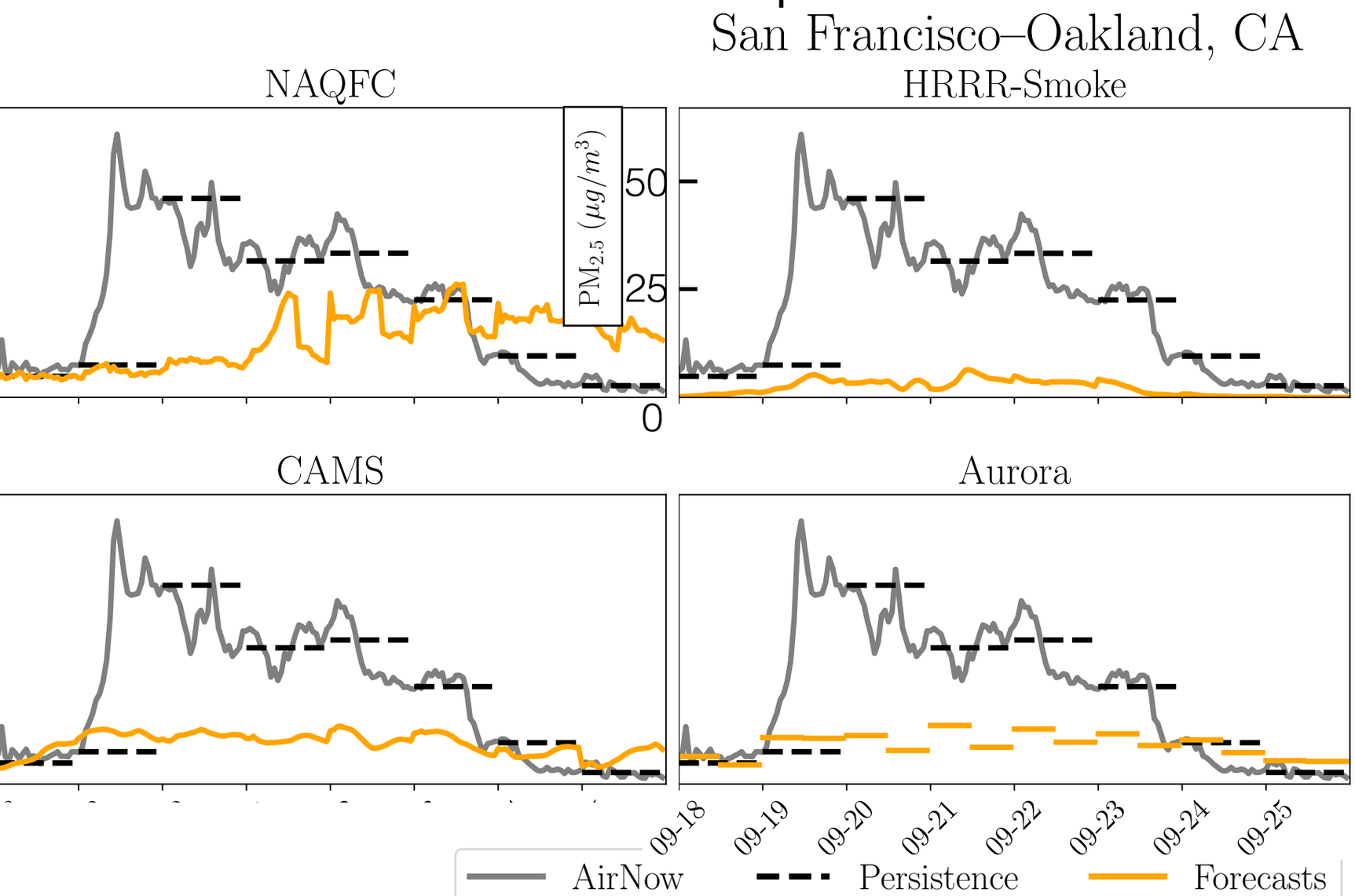
- We compare 6 tools: 4 from US and Europe governments, including CAMS (used by Google & Apple) + 1 ensemble
- Aurora: AI foundation model trained for weather forecasting and fine-tuned for air pollution forecasting



See paper for analysis over many places and times

Air pollution forecasts for health

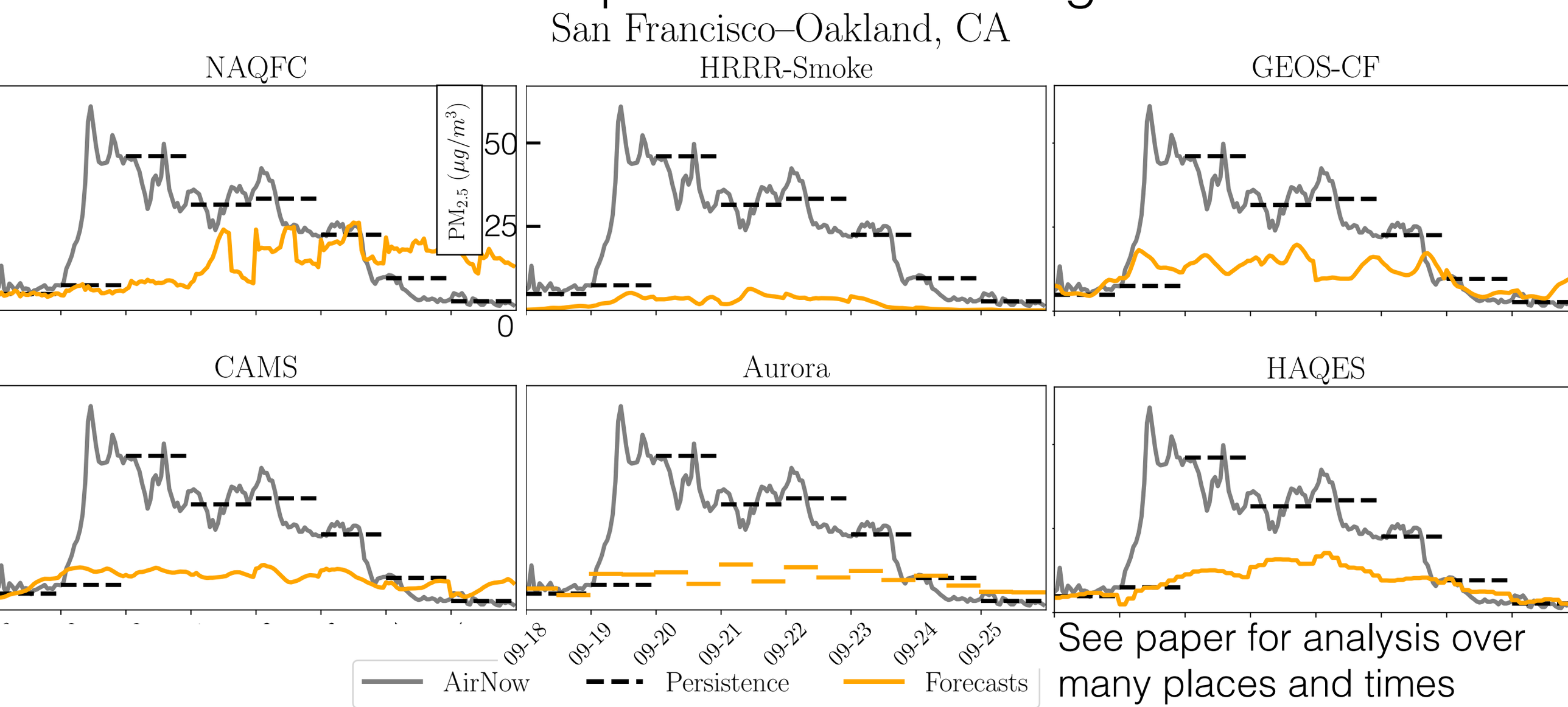
- We compare 6 tools: 4 from US and Europe governments, including CAMS (used by Google & Apple) + 1 ensemble
- Aurora: AI foundation model trained for weather forecasting and fine-tuned for air pollution forecasting



See paper for analysis over many places and times

Air pollution forecasts for health

- We compare 6 tools: 4 from US and Europe governments, including CAMS (used by Google & Apple) + 1 ensemble
- Aurora: AI foundation model trained for weather forecasting and fine-tuned for air pollution forecasting



See paper for analysis over many places and times

- We think AI is a good fit, but still needs work (a role for stats)

Evaluation relies on estimation

- Sometimes underappreciated: methods like cross-validation or averaging loss over held-out data aren't direct measures of quality; they're (good or bad) estimates

Evaluation relies on estimation

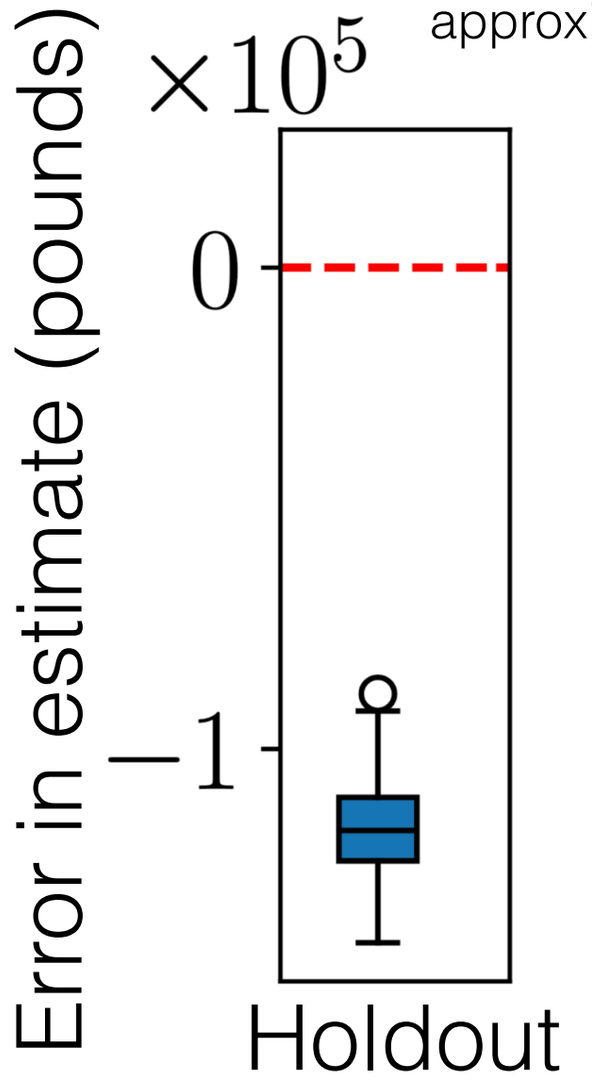
- Sometimes underappreciated: methods like cross-validation or averaging loss over held-out data aren't direct measures of quality; they're (good or bad) estimates
- The standard holdout (and other methods like spatial blocking, covariate shift) can substantially misestimate performance in spatial settings [cf. Roberts+ 2017; Telford & Birks 2005; Wadoux+ 2021; De Bruin+ 2022]

Evaluation relies on estimation

- Sometimes underappreciated: methods like cross-validation or averaging loss over held-out data aren't direct measures of quality; they're (good or bad) estimates
- The standard holdout (and other methods like spatial blocking, covariate shift) can substantially misestimate performance in spatial settings [cf. Roberts+ 2017; Telford & Birks 2005; Wadoux+ 2021; De Bruin+ 2022]
- Our experiments: London flat prices (right), air temperature, wind speed, etc

Evaluation relies on estimation

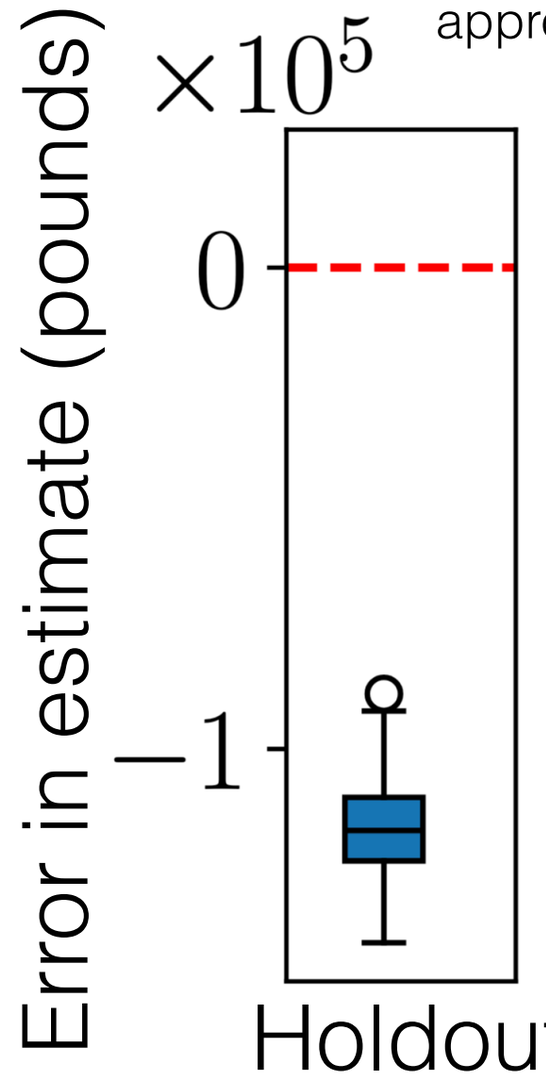
- Sometimes underappreciated: methods like cross-validation or averaging loss over held-out data aren't direct measures of quality; they're (good or bad) estimates
- The standard holdout (and other methods like spatial blocking, covariate shift) can substantially misestimate performance in spatial settings [cf. Roberts+ 2017; Telford & Birks 2005; Wadoux+ 2021; De Bruin+ 2022]
- Our experiments: London flat prices (right), air temperature, wind speed, etc



Evaluation relies on estimation

holdout is estimating performance of: a GP regression using VB approximation

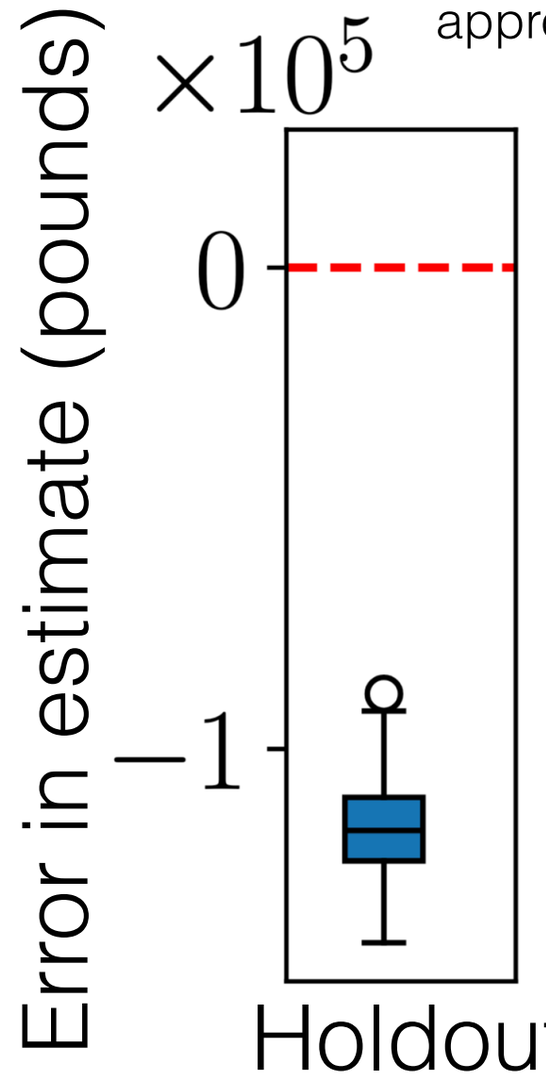
- Sometimes underappreciated: methods like cross-validation or averaging loss over held-out data aren't direct measures of quality; they're (good or bad) estimates
- The standard holdout (and other methods like spatial blocking, covariate shift) can substantially misestimate performance in spatial settings [cf. Roberts+ 2017; Telford & Birks 2005; Wadoux+ 2021; De Bruin+ 2022]
  - Our experiments: London flat prices (right), air temperature, wind speed, etc
  - Our theory: existing methods can be arbitrarily wrong with infinitely dense validation data



Evaluation relies on estimation

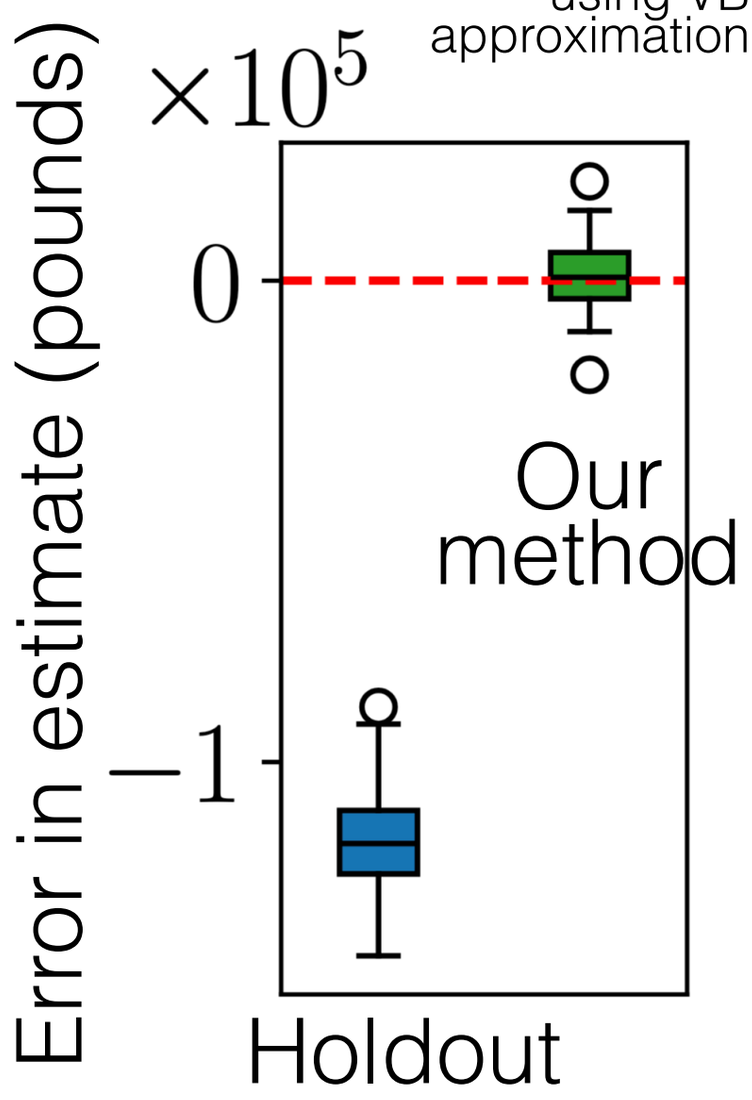
- Sometimes underappreciated: methods like cross-validation or averaging loss over held-out data aren't direct measures of quality; they're (good or bad) estimates
- The standard holdout (and other methods like spatial blocking, covariate shift) can substantially misestimate performance in spatial settings [cf. Roberts+ 2017; Telford & Birks 2005; Wadoux+ 2021; De Bruin+ 2022]
  - Our experiments: London flat prices (right), air temperature, wind speed, etc
  - Our theory: existing methods can be arbitrarily wrong with infinitely dense validation data
- We propose an alternative performance estimator (tailored to spatial tasks)

holdout is estimating performance of: a GP regression using VB approximation



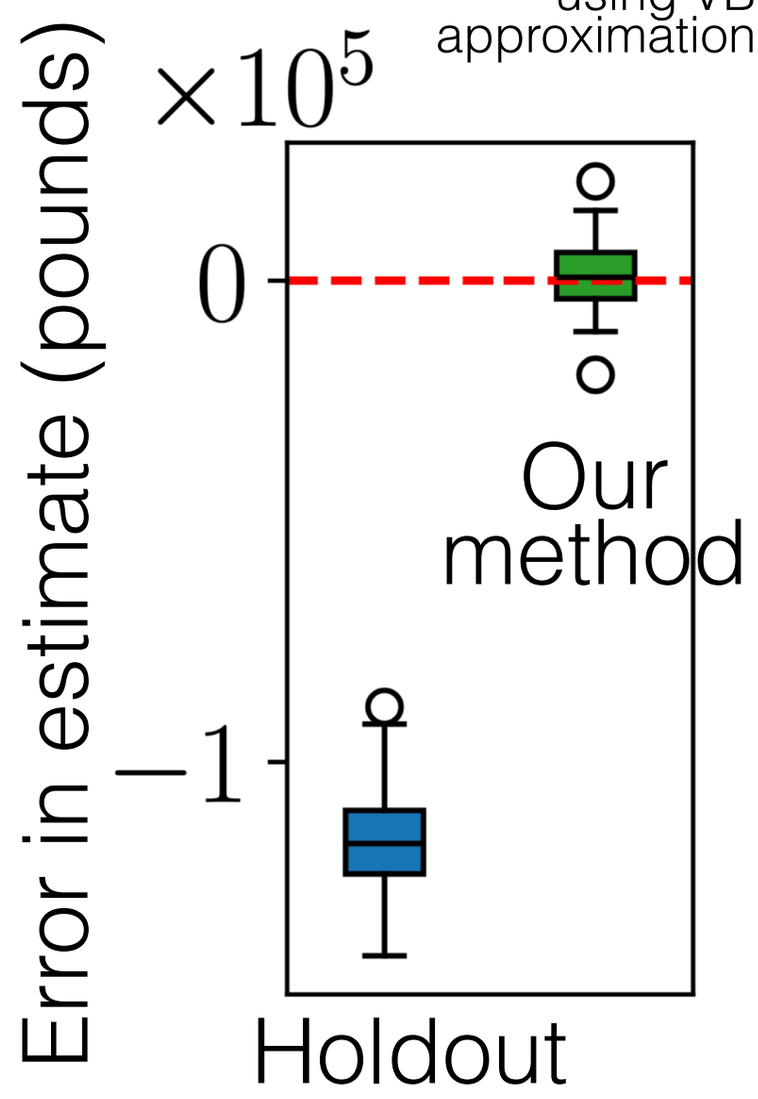
Evaluation relies on estimation

- Sometimes underappreciated: methods like cross-validation or averaging loss over held-out data aren't direct measures of quality; they're (good or bad) estimates
- The standard holdout (and other methods like spatial blocking, covariate shift) can substantially misestimate performance in spatial settings [cf. Roberts+ 2017; Telford & Birks 2005; Wadoux+ 2021; De Bruin+ 2022]
  - Our experiments: London flat prices (right), air temperature, wind speed, etc
  - Our theory: existing methods can be arbitrarily wrong with infinitely dense validation data
- We propose an alternative performance estimator (tailored to spatial tasks)



Evaluation relies on estimation

- Sometimes underappreciated: methods like cross-validation or averaging loss over held-out data aren't direct measures of quality; they're (good or bad) estimates
- The standard holdout (and other methods like spatial blocking, covariate shift) can substantially misestimate performance in spatial settings [cf. Roberts+ 2017; Telford & Birks 2005; Wadoux+ 2021; De Bruin+ 2022]
  - Our experiments: London flat prices (right), air temperature, wind speed, etc
  - Our theory: existing methods can be arbitrarily wrong with infinitely dense validation data
- We propose an alternative performance estimator (tailored to spatial tasks)
  - We show it performs well in experiments + theory (gets the right answer given infinitely dense data)



A lot of AI trust is pushed to evaluation

A lot of AI trust is pushed to evaluation

- And it's big business. Just one example: "Arena Intelligence Inc., the startup behind a popular artificial intelligence model leaderboard, has raised $200 million in funding at a $3.1 billion valuation" Bloomberg 2026 Oct 8
  - "Arena began in 2023 as Chatbot Arena"

A lot of AI trust is pushed to evaluation

- And it's big business. Just one example: "Arena Intelligence Inc., the startup behind a popular artificial intelligence model leaderboard, has raised $200 million in funding at a $3.1 billion valuation" Bloomberg 2026 Oct 8
  - "Arena began in 2023 as Chatbot Arena"

▷	Model ▷	Arena Elo ▷	Coding ▷	Vision ▷	MII ▷	ARC-AGI ▷	Organization ▷	License ▷
🏆	Claude Opus 5.5	1526	1575	1320	56	93.3	Anthropic	Proprietary
🥇	Gemini-4-Argon	1525	1570	1318	53	91.3	Google	Proprietary
🥇	Claude Sonnet 5.5	1521	1568	1315	54		Anthropic	Proprietary
🥇	Claude Fable 5.1	1520	1570	1315	52	90	Anthropic	Proprietary
🥇	GPT-6 Astra	1520	1568	1316	52	95	OpenAI	Proprietary
🥇	Muse Wonder Full	1518	1567	1316	53		Meta	Proprietary
🥇	GPT-6.1 Sol	1516	1568	1315	51	94.2	OpenAI	Proprietary

A lot of AI trust is pushed to evaluation

- And it's big business. Just one example: "Arena Intelligence Inc., the startup behind a popular artificial intelligence model leaderboard, has raised $200 million in funding at a $3.1 billion valuation" Bloomberg 2026 Oct 8
  - "Arena began in 2023 as Chatbot Arena"

▷	Model ▷	Arena Elo ▷	Coding ▷	Vision ▷	MII ▷	ARC-AGI ▷	Organization ▷	License ▷
🏆	Claude Opus 5.5	1526	1575	1320	56	93.3	Anthropic	Proprietary
🥇	Gemini-4-Argon	1525	1570	1318	53	91.3	Google	Proprietary
🥇	Claude Sonnet 5.5	1521	1568	1315	54		Anthropic	Proprietary
🥇	Claude Fable 5.1	1520	1570	1315	52	90	Anthropic	Proprietary
🥇	GPT-6 Astra	1520	1568	1316	52	95	OpenAI	Proprietary
🥇	Muse Wonder Full	1518	1567	1316	53		Meta	Proprietary
🥇	GPT-6.1 Sol	1516	1568	1315	51	94.2	OpenAI	Proprietary

- We show: Out of 57,477 preferences (matchups) in Chatbot Arena (at a particular time), we can drop 2 to change the top-ranked LLM (i.e. 0.003% of matchups)

A lot of AI trust is pushed to evaluation

- And it's big business. Just one example: "Arena Intelligence Inc., the startup behind a popular artificial intelligence model leaderboard, has raised $200 million in funding at a $3.1 billion valuation" Bloomberg 2026 Oct 8
  - "Arena began in 2023 as Chatbot Arena"

▷	Model ▷	Arena Elo ▷	Coding ▷	Vision ▷	MII ▷	ARC-AGI ▷	Organization ▷	License ▷
🏆	Claude Opus 5.5	1526	1575	1320	56	93.3	Anthropic	Proprietary
🥇	Gemini-4-Argon	1525	1570	1318	53	91.3	Google	Proprietary
🥇	Claude Sonnet 5.5	1521	1568	1315	54		Anthropic	Proprietary
🥇	Claude Fable 5.1	1520	1570	1315	52	90	Anthropic	Proprietary
🥇	GPT-6 Astra	1520	1568	1316	52	95	OpenAI	Proprietary
🥇	Muse Wonder Full	1518	1567	1316	53		Meta	Proprietary
🥇	GPT-6.1 Sol	1516	1568	1315	51	94.2	OpenAI	Proprietary

- We show: Out of 57,477 preferences (matchups) in Chatbot Arena (at a particular time), we can drop 2 to change the top-ranked LLM (i.e. 0.003% of matchups)
  - Using bootstrap conf intervals doesn't remove sensitivity

Some roles for Bayes & stats in AI

- **Bayes (and stats) can take the strengths of modern AI and turn them into actionable, trustworthy decisions**

Some roles for Bayes & stats in AI

- **Bayes (and stats) can take the strengths of modern AI and turn them into actionable, trustworthy decisions**
- **Nomon: computer interface for people with near-total paralysis**
  - QuickClick paper in prep. (try it out at **nomon.app** → demos!)
  - Frisch, Wade, Gaines, Kinsella, Peters, Broderick, Vertanen. It's Complicated: On the Design and Evaluation of AI-Powered AAC Interfaces. *CHI Workshop on Speech AI For All: The What, How, and Who of Measurement*, 2026
  - Bonaker, Nel, Vertanen, Broderick. A Usability Study of Nomon: A Flexible Interface for Single-Switch Users. *ASSETS*, 2023
  - Bonaker, Nel, Vertanen, Broderick. A Performance Evaluation of Nomon: A Flexible Interface for Noisy Single-Switch Users. *CHI*, 2022

Some roles for Bayes & stats in AI

- **Bayes (and stats) can take the strengths of modern AI and turn them into actionable, trustworthy decisions**
- **Nomon: computer interface for people with near-total paralysis**
  - QuickClick paper in prep. (try it out at **nomon.app** → demos!)
  - Frisch, Wade, Gaines, Kinsella, Peters, Broderick, Vertanen. It's Complicated: On the Design and Evaluation of AI-Powered AAC Interfaces. *CHI Workshop on Speech AI For All: The What, How, and Who of Measurement*, 2026
  - Bonaker, Nel, Vertanen, Broderick. A Usability Study of Nomon: A Flexible Interface for Single-Switch Users. *ASSETS*, 2023
  - Bonaker, Nel, Vertanen, Broderick. A Performance Evaluation of Nomon: A Flexible Interface for Noisy Single-Switch Users. *CHI*, 2022
- **Burt, Shen, Broderick. Consistent validation for prediction methods in spatial settings, *AISTATS* 2025**
  - **Berlinghieri*, Burt*, Giani, Fiore, Broderick. Are Hourly $PM_{2.5}$ Forecasts Sufficiently Accurate to Plan Your Day? *BAMS* 2026**
  - Burt*, Berlinghieri*, Bates, Broderick. Smooth Sailing: Lipschitz-Driven Uncertainty Quantification for Spatial Association, *NeurIPS* 2026 • Burt, Berlinghieri, Broderick. Wrong Model, Right Uncertainty: Spatial Associations for Discrete Data with Misspecification. *NeurIPS 2025 Workshop on Reliable ML from Unreliable Data*

Some roles for Bayes & stats in AI

- **Bayes (and stats) can take the strengths of modern AI and turn them into actionable, trustworthy decisions**
- **Nomon: computer interface for people with near-total paralysis**
  - QuickClick paper in prep. (try it out at **nomon.app** → demos!)
  - Frisch, Wade, Gaines, Kinsella, Peters, Broderick, Vertanen. It's Complicated: On the Design and Evaluation of AI-Powered AAC Interfaces. *CHI Workshop on Speech AI For All: The What, How, and Who of Measurement*, 2026
  - Bonaker, Nel, Vertanen, Broderick. A Usability Study of Nomon: A Flexible Interface for Single-Switch Users. *ASSETS*, 2023
  - Bonaker, Nel, Vertanen, Broderick. A Performance Evaluation of Nomon: A Flexible Interface for Noisy Single-Switch Users. *CHI*, 2022
- **Burt, Shen, Broderick. Consistent validation for prediction methods in spatial settings, *AISTATS* 2025**
  - **Berlinghieri*, Burt*, Giani, Fiore, Broderick. Are Hourly $PM_{2.5}$ Forecasts Sufficiently Accurate to Plan Your Day? *BAMS* 2026**
  - Burt*, Berlinghieri*, Bates, Broderick. Smooth Sailing: Lipschitz-Driven Uncertainty Quantification for Spatial Association, *NeurIPS* 2026 • Burt, Berlinghieri, Broderick. Wrong Model, Right Uncertainty: Spatial Associations for Discrete Data with Misspecification. *NeurIPS 2025 Workshop on Reliable ML from Unreliable Data*
- **Huang*, Shen*, Wei, Broderick. Dropping Just a Handful of Preferences Can Change Top LLM Rankings, *ICLR* 2026**