

Active inference

UCSF NS Orientation 2024 — Stats III

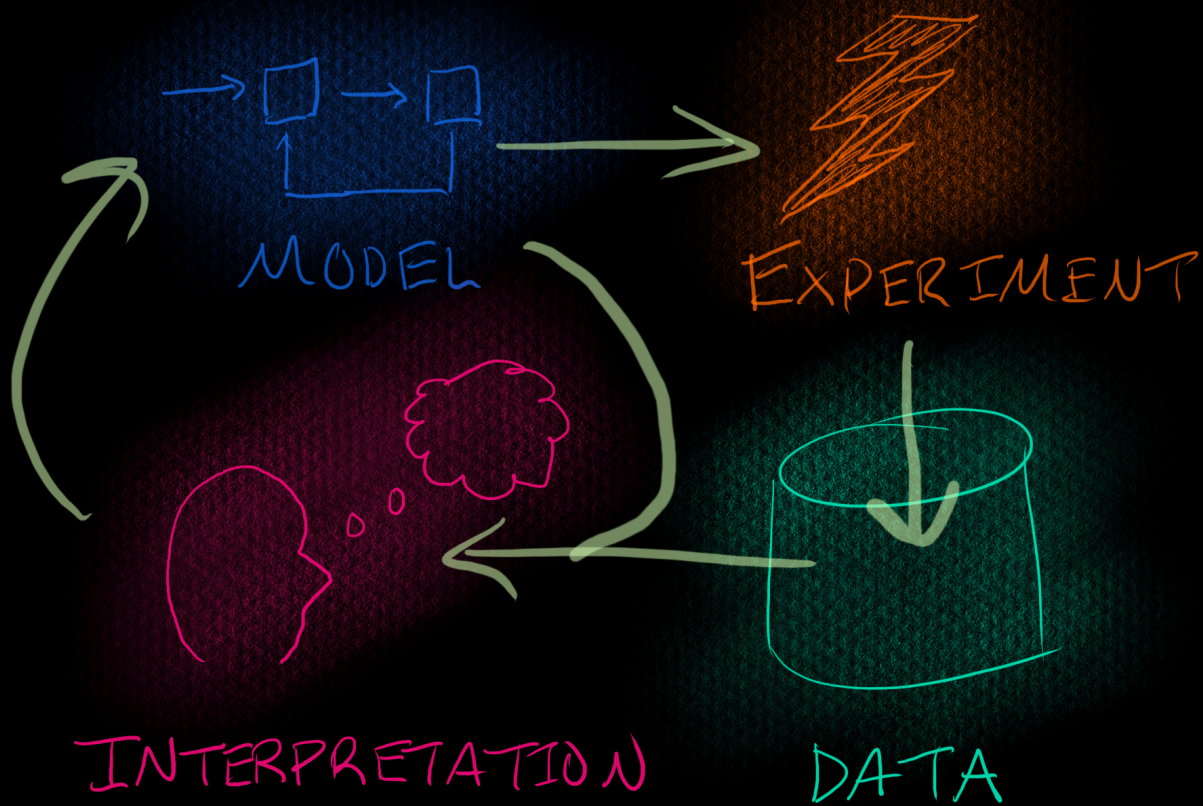
Maxine Collard

Learning about the world

Learning about the world

The “ideal” science centers on the **experimental cycle**:

- Start with **what we currently know** about the world.
- **Observe new data.**
- Use this data to **make inferences** that **update** how we view the world.



Recall

The amount of information provided by a test result is called the result's **Bayes factor**, K . While the PPV—which is the **posterior probability** of disease given a positive test—changes depending on the prior, the following **ratios** are always proportional:

$$\frac{\text{Pr}(\text{actually } + \mid \text{test } +)}{\text{Pr}(\text{actually } - \mid \text{test } +)} = K \frac{\text{Pr}(\text{actually } +)}{\text{Pr}(\text{actually } -)}$$

So, the Bayes factor tells us how much receiving a positive test result **changes** our prior belief. Test results with larger Bayes factors **change our beliefs more**.

In experimental design

The **posterior probability** of a model given data changes depending on the prior.

However, the following **ratios** are always proportional:

$$\frac{\text{Pr}(\text{model 1} \mid \text{data we see})}{\text{Pr}(\text{model 2} \mid \text{data we see})} = K \frac{\text{Pr}(\text{model 1})}{\text{Pr}(\text{model 2})}$$

So, the Bayes factor K tells us how much receiving a positive test result **changes** our prior belief. Test results with larger Bayes factors **change our beliefs more**.

Principles of experimental design

All we can reasonably do is **compare models**.

“All models are wrong; some are useful.”—James Box

Choose comparisons that we think **will tease apart the models** (have a large expected Bayes factor)

Somebody has to tell a Congressional panel why taxpayer money pays for what we do.

What's in an experimental design

When we conduct experiments, we collect **data** that **updates** our beliefs about the world:

$$\begin{array}{c} \text{Pr}(\text{model}) \\ \downarrow \\ \text{Pr}(\text{model} \mid \text{data}) \end{array}$$

When you design an experiment, you specify:

- what **data** you will **collect**, and
- how that will **update** your beliefs about your **model** with any outcome.

...Is that all?

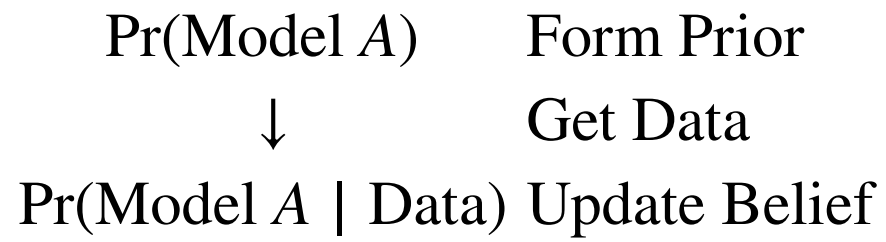
Science:

Pr(Model A) Form Prior

Science:

Pr(Model A) Form Prior
↓ Get Data

Science:



One might wonder...

- What PPV makes a diagnostic test good? What NPV?
- Is it the same for scientific tests?
- Where do tests sit in a larger societal context?
 - i.e., What are we actually **doing** with the knowledge we get from tests?
- Are false positives and false negatives **equally** bad?
 - Since there's a tradeoff between PPV and NPV, which should we prioritize making as large as possible?

One might wonder...

- How do we know whether our belief about the prevalence of Covid is correct?
- How do we know how accurate our prior belief about a statement in science is?

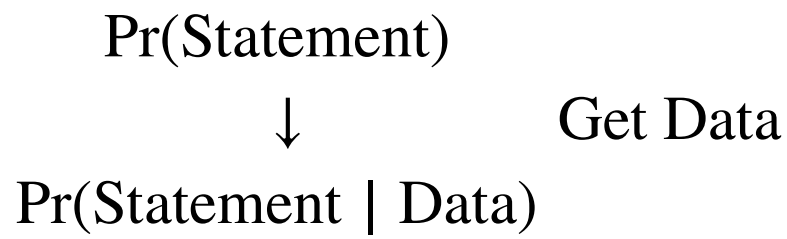
Since, as we established before, these are **crucial** for determining how useful any test is.

The unraveling begins...

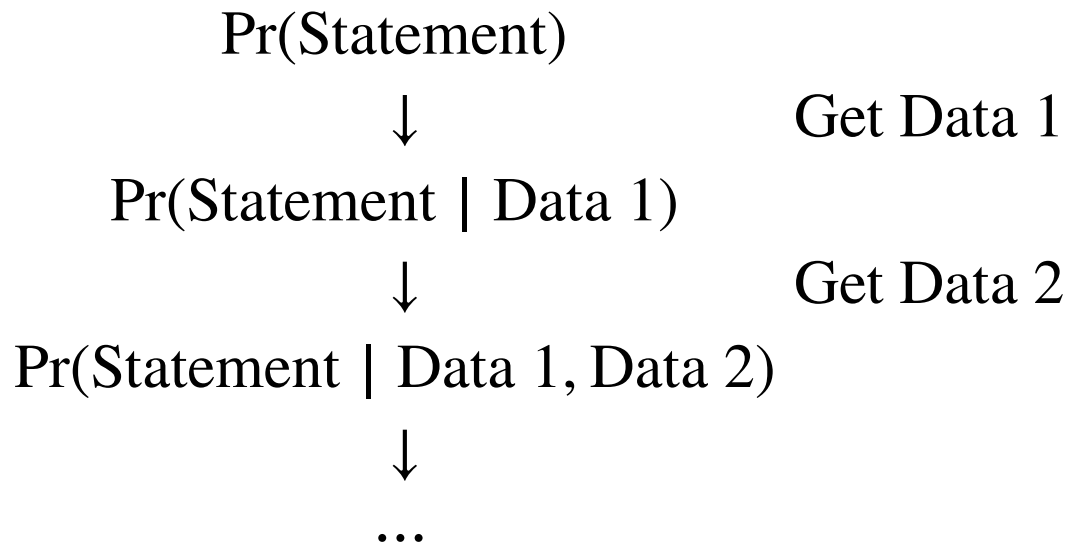
Science as process

Science as process

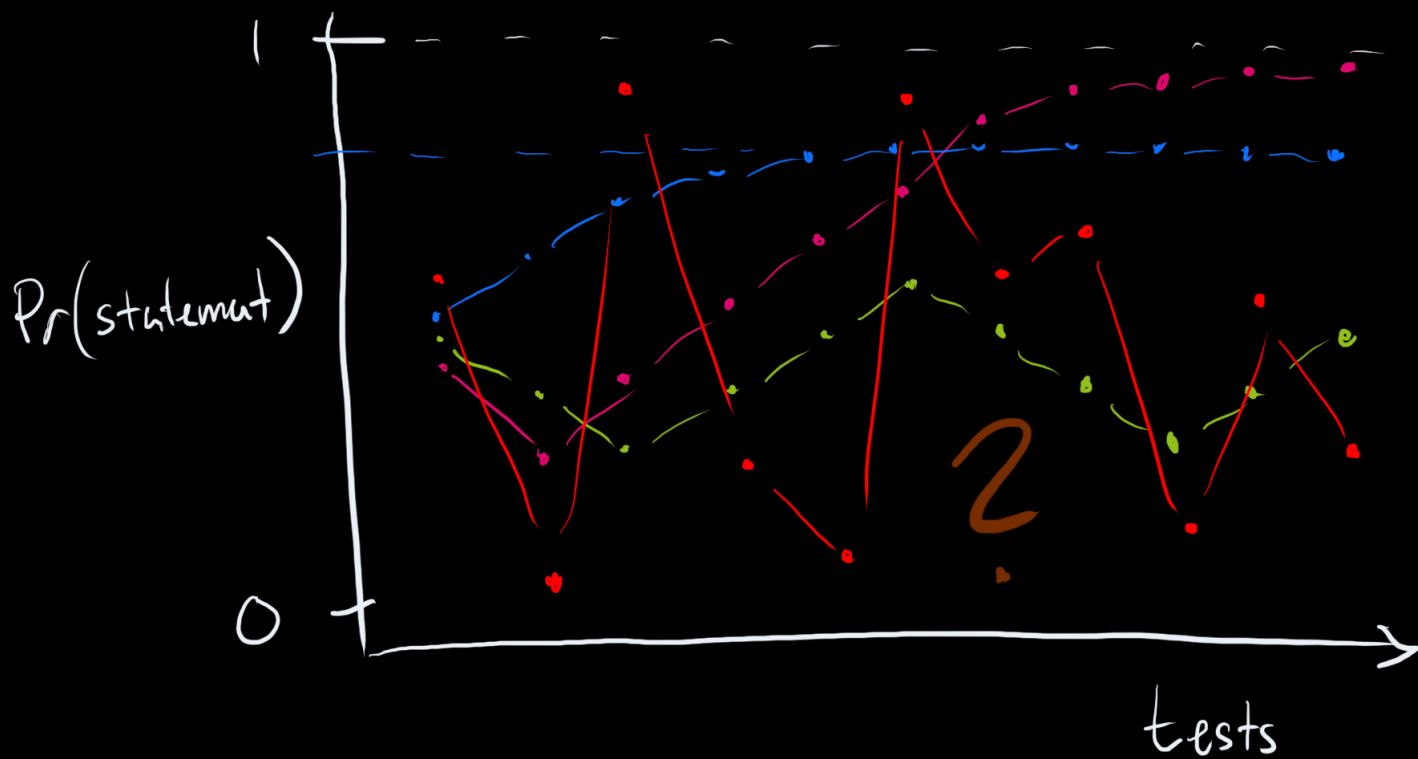
Do we really just do one experiment?



Let's say we run a second experiment with a second test.
Now I can **update** my knowledge **again**:



And so on, for test after test.



One might wonder...

- What does this process,

$$\Pr(\text{Statement} \mid \text{Data } 1, \dots, \text{Data } k)$$

converge to as we do more and more experiments,

$$k \rightarrow \infty$$

- Does it converge at all?
- What does this convergence depend on?

One might wonder...

Wait, how do we choose the next test to run each time?

This is a **big** question, so let's think about it in the context of a few specific examples.

One might wonder...

Let's say we **update** our belief with successive tests, as above, but at each step, we choose the next test such that it will **maximally increase our updated probability**, $\Pr(\text{statement} \mid \text{tests})$.

That is, we choose the experiment that we think will **give us the most evidence supporting our model**.

What does this process converge to?

One might wonder...

Let's say we update our belief with successive tests, as above, but at each step, we choose the next test **completely at random**.

What does this process converge to?

One might wonder...

Let's say we **update** our belief with successive tests, but we are **really wrong** about our initial belief.

- What happens to the limit of $\Pr(\text{Statement} \mid \text{Data } 1, \dots, \text{Data } k)$?
- Will our belief always end up in the same place at the end of the process, regardless of what we initially think?
- What is the best initial belief to have?

Comparing models

Comparing models

Let's say we **update** our belief with successive tests, but this time, we look at **multiple** candidate models; i.e., we obtain

$$\Pr(\text{model 1} \mid \text{test 1}, \dots, \text{test } n)$$

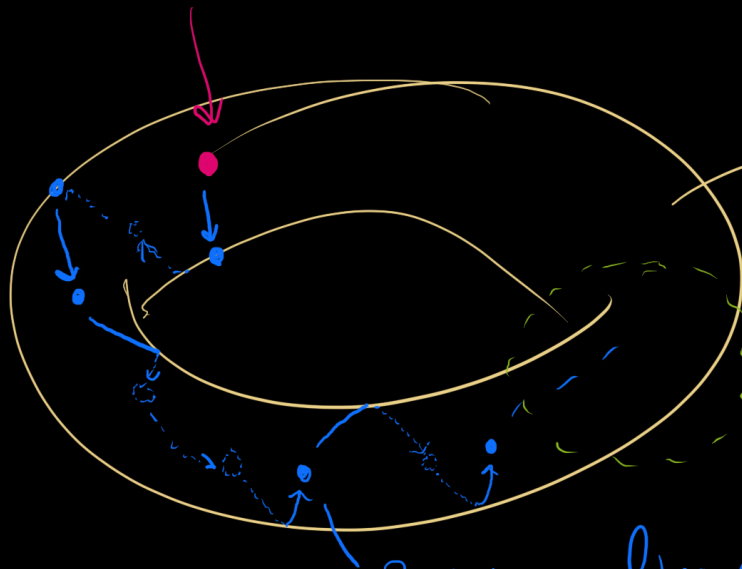
$$\Pr(\text{model 2} \mid \text{test 1}, \dots, \text{test } n)$$

...

$$\Pr(\text{model } k \mid \text{test 1}, \dots, \text{test } n)$$

after a sequence of n tests.

Initial prediction



Space of probability
distributions over
your models

Predictions after each
successive test

Converges?

One might wonder...

**Does the presence of multiple candidate models
change the way we should choose the next test?**

One might wonder...

In the above scheme with multiple candidate models, let's say that the **truth** is **not** among the models that you considered.

(After all, do we have so much hubris as scientists to think that universal truth is guaranteed to be comprehensible by human models?)

- What does **this** process converge to?
- How do you interpret the result you end up at after iterating the experimental cycle a bunch?

One might wonder...

In the above scheme with multiple candidate models, let's say that the **truth** is **not** among the models that you considered.

(After all, do we have so much hubris as scientists to think that universal truth is guaranteed to be comprehensible by human models?)

- How would you **even know** whether you included the truth as a candidate model you were considering in your prior?

History dependence

History dependence

The process of updating our beliefs about our model is kind of like an **optimization problem**.

Imagine that there is some “landscape” painted across the space of all models we’re considering, with the “height” at each point (each model) corresponding to the “badness” of that model.

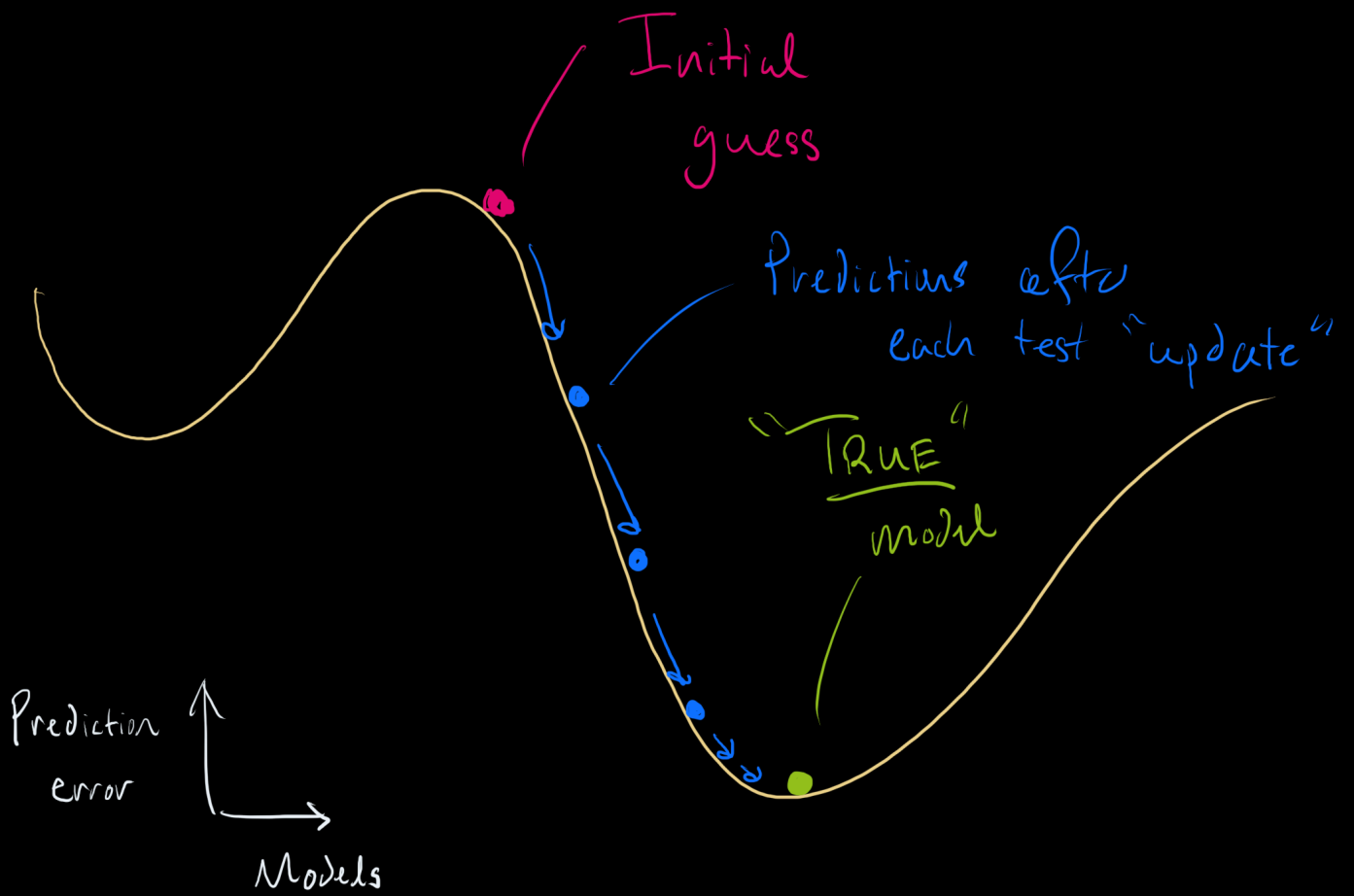
For example, models are worse (higher up) if they have a bigger **prediction error**, lower R^2 , etc.

History dependence

The process of updating our beliefs about our model is kind of like an **optimization problem**.

Imagine that there is some “landscape” painted across the space of all models we’re considering, with the “height” at each point (each model) corresponding to the “badness” of that model.

Our goal is to incrementally update what we *think* the best model is, in order to someday find the lowest point in the entire landscape—the **actual best** model—which should be the **truth**.



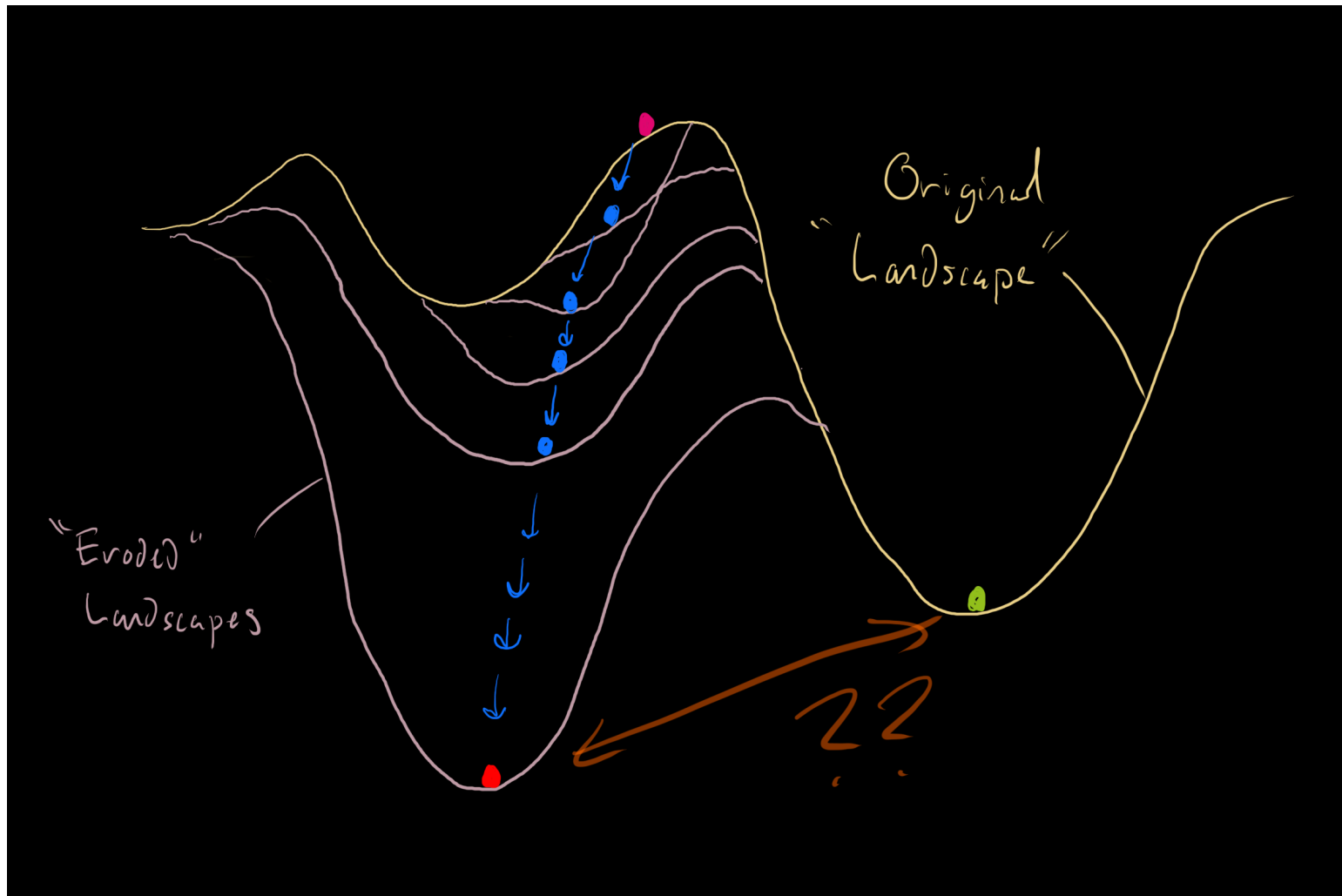
But...

::: {.fragment} After each experiment we run, we collect **new data**. :::

This new data **changes how we measure** the badness or goodness of a model's fit.

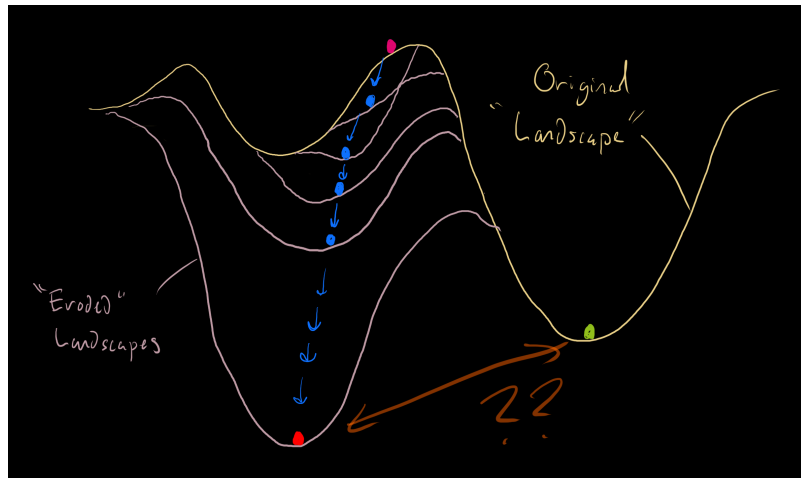
After all, the model should fit the new data **also!**

This **alters the landscape** of our optimization problem.



One might wonder...

- What happens if we start off **very wrong** about our prediction of what model is best?
- Does this optimization process, iteratively deforming the landscape, converge to the truth?
- If not, how can we ever “recover”?



- Is this new minimum *actually* the truth? That is, was our notion of truth—our “landscape”—wrong to begin with?
- Was the evidence that “eroded” the landscape actually *helpful*? Or was this erosion a result of our **bias** from an incorrect initial prediction?

How does the optimum we converge to depend on how we choose the next data to collect?

One might wonder...

- If we had started off the entire process with a **vastly different initial prediction**, would *yet another* minimum have emerged?
 - Is *that* the truth?
 - Which one?
- How does the optimum we converge to depend on our initial beliefs?

One might wonder...

- Is there a criterion for optimality that is **independent** of our starting position?
- Can we design a scheme for conducting tests that makes us more likely to converge to that **universal** model, regardless of our initial beliefs?

One might wonder...

- Is any of this even a problem?
- As we try to design a scientific process:
 - Should we **minimize** the effect of this history dependence, the “erosion” of the landscape?
 - Or is this actually an **important** feature of how science behaves?

Is the body of knowledge we have converged to through the scientific process a statement about the world we live in?

Or is it just a statement about the process itself?

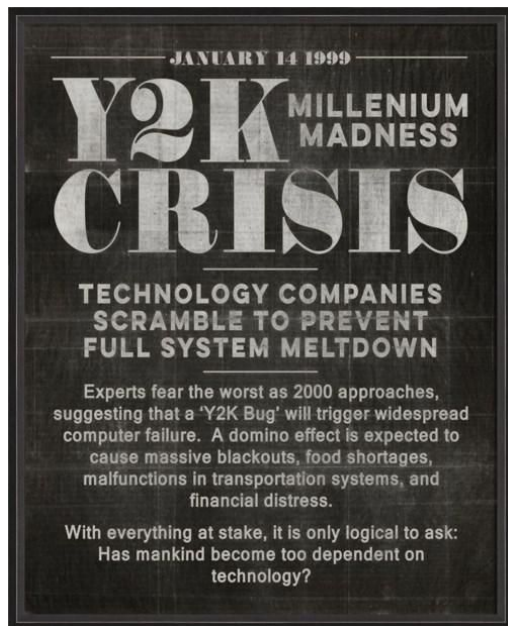
What would scientific “truth” look like today if we started with different structures for how we build new knowledge?

What amazing ideas have we missed out on?

Who benefits from the rules being as they are?

My story

The year 2000

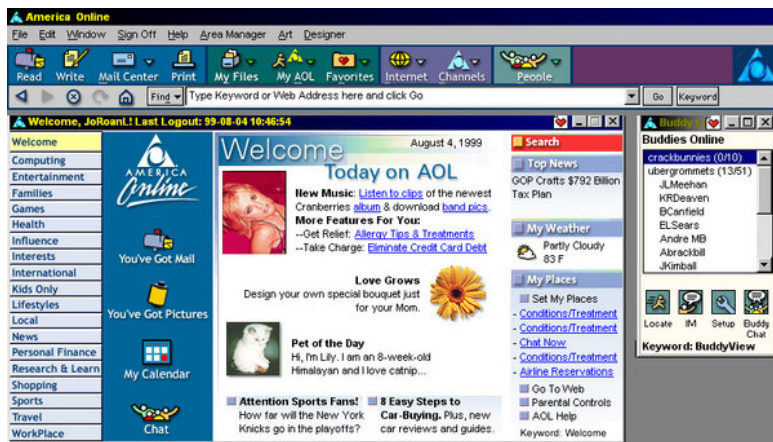


Expectation



Reality

The year 2000



(The Internet)

“Get off the phone, I need to use the Internet”

The year 2000



Maxwell Jordan Collard

(It only got worse by
2011.)



Yikers

The year 2000

When I first went to see a therapist for depression.



Labels I have been given

- Major depressive disorder
- Obsessive-compulsive disorder
- Oppositional defiant disorder
- Post-traumatic stress disorder
- Borderline personality disorder
- Bipolar II disorder
- *Complex* post-traumatic stress disorder
- Autism spectrum disorder
- Attention-deficit/hyperactivity disorder

Labels I would endorse

- Major depressive disorder
- Oppositional defiant disorder
- Post-traumatic stress disorder
- Bipolar II disorder
- Borderline personality disorder

Two buckets:

"Neuro-spicy"

- **Autism spectrum disorder**
- **Attention-deficit/hyperactivity disorder**
- **Obsessive-compulsive disorder**

"Stuff happened"

- ***Complex* post-traumatic stress disorder**

Some medications I tried ...

- fluoxetine
- sertraline
- mirtazapine
- alprazolam
- clonazepam
- lurasidone
- aripiprazole, brexpiprazole
- olanzapine
- quetiapine
- risperidone
- trazadone
- lithium carbonate
- lamotrigine
- valproic acid
- oxcarbazepine
- escitalopram, bupropion

Some medications I tried ... And their side effects

- fluoxetine
 - **diarrhea; sexual dysfunction**
- sertraline
 - **precipitated hypomanic episode** after one half-pill
- mirtazapine
 - somnolence; **severe weight gain**
- alprazolam
 - **vivid, horrifying nightmares**
- clonazepam
 - **depression; worsening anxiety**
- lurasidone
 - **akathisia; drug-induced Parkinsonism**
- aripiprazole, brexpiprazole
 - **mood swings; severe drug-induced Parkinsonism**
- olanzapine
 - somnolence; **impaired cognition; weight gain; akathisia**
- quetiapine
 - somnolence; impaired cognition; akathisia
- risperidone
 - **weight gain**
- trazadone
 - **somnolence**
- lithium carbonate
 - **diarrhea; incontinence; urinating 6+/day; myoclonus; drug-induced Parkinsonism**
- lamotrigine
 - **full-body skin rash** after one half-pill
- valproic acid
 - **episodes of frank derealization-depersonalization**
- oxcarbazepine
 - **impaired cognition**
- escitalopram, bupropion
 - **severe mood swings**

And it still didn't help

Constant battle against **dissociative spirals**

- Small trigger, sometimes subliminal (not consciously aware of it at the start)

And it still didn't help

Constant battle against **dissociative spirals**

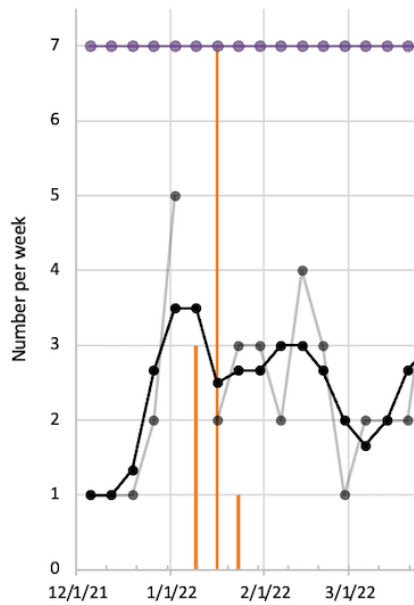
- Progressive worsening of symptoms over ensuing minutes to hours
 - Anxiety, sympathetic activation
 - Reach out to social network for help
 - Lack of response → Worsening agitation
 - **Depersonalization, derealization**
 - ...

And it still didn't help

Constant battle against **dissociative spirals**

- Eventually spontaneously ends within seconds to a minute
- Refractory period of fatigue afterward for several hours

And it still didn't help



In December 2021, after months of struggles, I started having dissociative spirals **every day**. I couldn't function.

Crisis text-message days per week (grey), 4-week rolling average (black), and hospital days per week (orange)

And it still didn't help

In December 2021, after months of struggles, I started having dissociative spirals **every day**. I couldn't function.

I walked myself over what was then the Langly Porter Psychiatric Institute at Parnassus (*since demolished*).



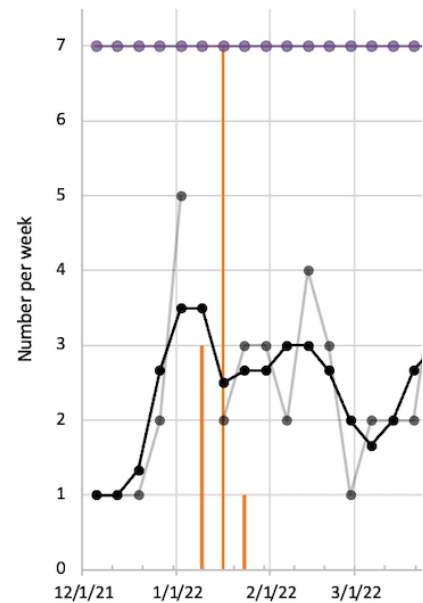
The old Grey line shuttle stop

And it still didn't help

In December 2021, after months of struggles, I started having dissociative spirals **every day**. I couldn't function.

I walked myself over what was then the Langly Porter Psychiatric Institute at Parnassus (*since demolished*).

And even that didn't help!



Crisis text-message days per week (grey), 4-week rolling average (black), and hospital days per week (orange)

But I have a different theory:

What if I'm not mentally ill at all?

What if I grew up with a different brain—an **autistic** brain—that nobody around me understood?

My 1st Grade classmates made fun of me after I brought a stack of CDs to class with a Visual Basic 6 program I had written for practicing their multiplication tables.

...Never lived that one down.

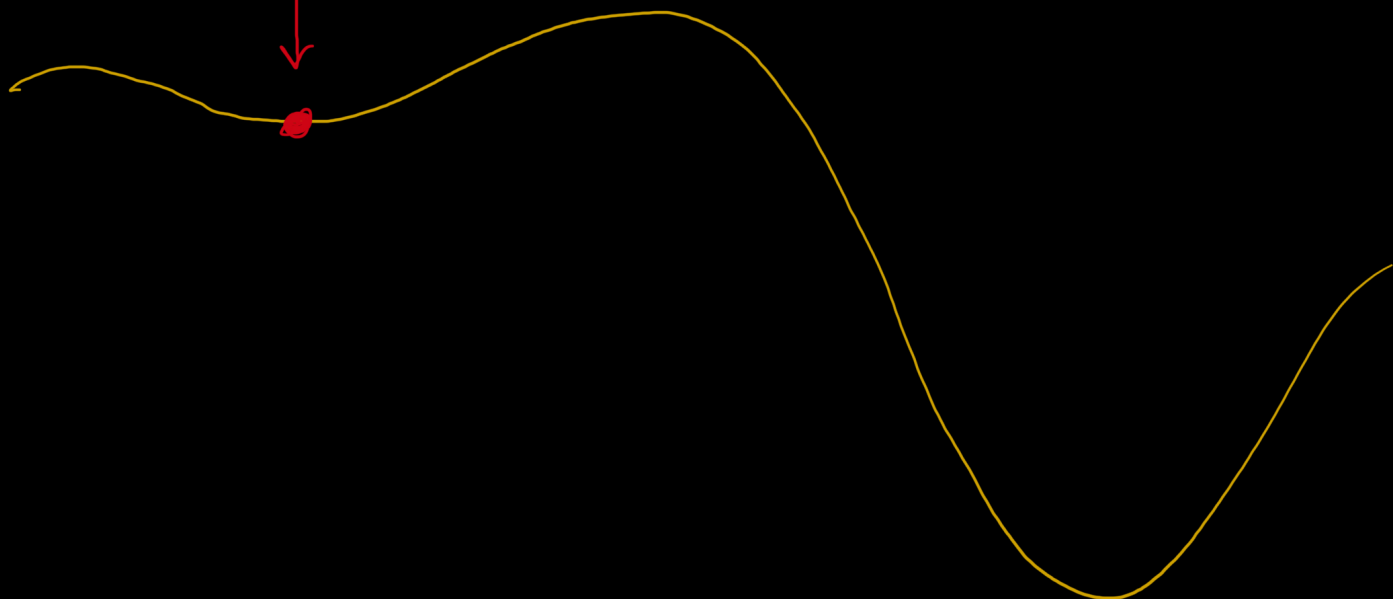
What if—having been socially ostracized and hurt everywhere I went—I never felt **safe** enough as a child to explore my real identity?

What if instead I built my life from an initial belief, learned early, ...

a belief that was **wrong**, ...

but one that at least allowed me to **survive** in that unsafe place?

"I am a man"

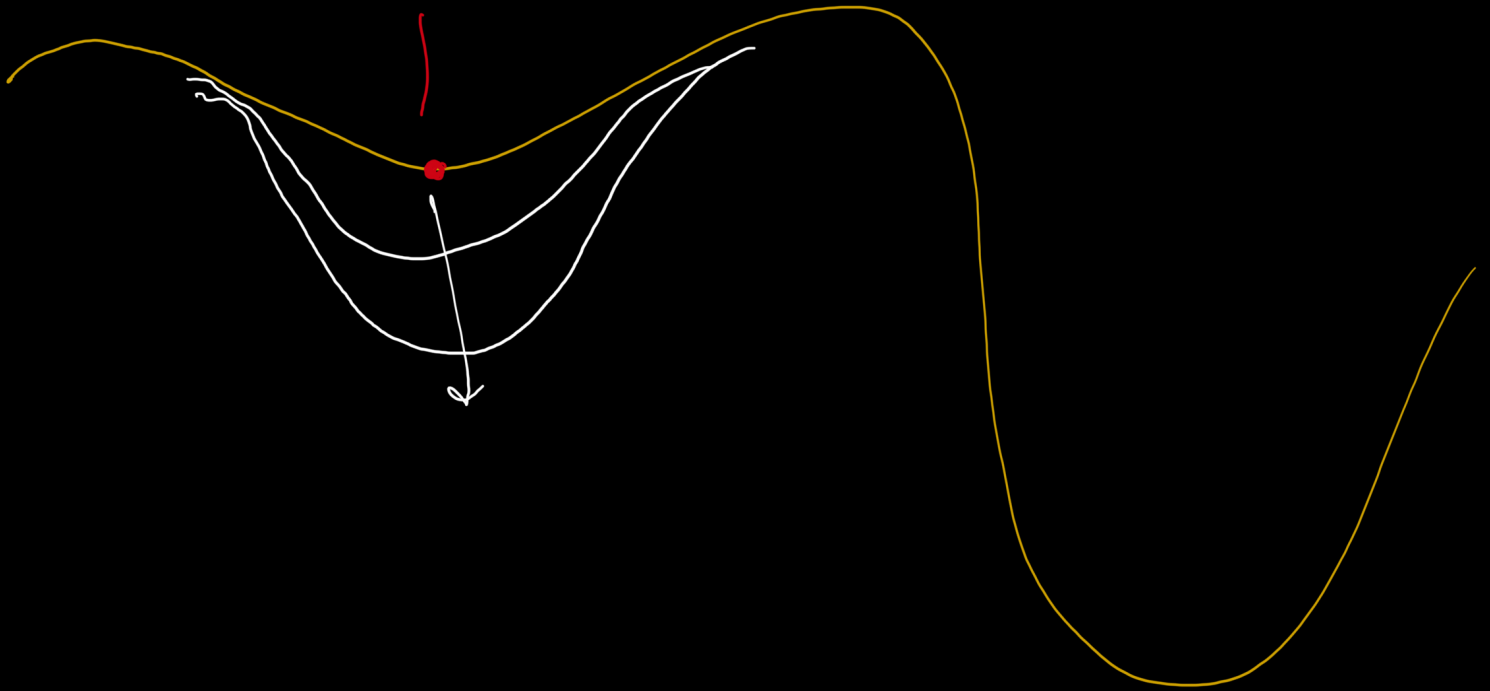


What if I created an **entire universe** of ideas built on top of that one concept—a concept so core to who we are, how we relate to others, and how we relate to the world?

What if I spent decades of my life **digging deeper and deeper** into that wrong belief...

...years going out into the world, **collecting data** that I knew would reinforce it, because it was what had kept me **safe** in that formative time?

"I am a man"



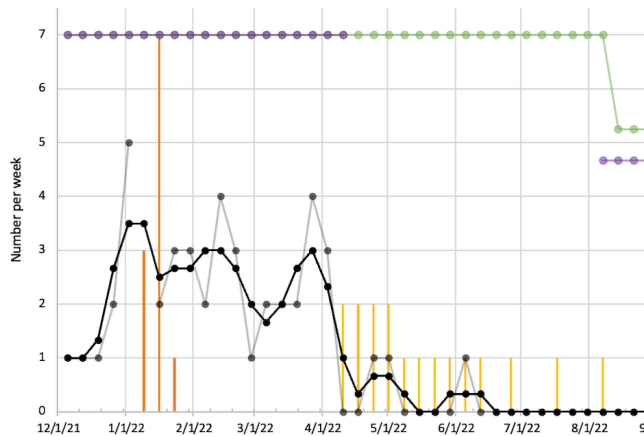
What if there isn't any disease in me that has to be cured at all?

What if my suffering was caused because of one prior belief, one concept I had gotten wrong in the distant past, and the years of erosion that built on top of it?

What if one wrong idea from the past changes everything?

Healing

Healing



Crisis text-message days per week (grey),
4-week rolling average (black), and
hospital days per week (orange)

Well, ...

something worked.

And that thing was
**psychedelic ketamine
therapy** (sessions in **yellow
bars**).

Session structure



One of the treatment rooms at my clinic.

Preparation (30 minutes): Intention journaling, what I would like to receive from the session

Administration (15 minutes): 4–6 intranasal sprays, administered under physician supervision

Acute trip (30–90 minutes):

- Eyes closed, listening to prepared playlist (created by my own simulated annealing music semantic arc playlist, to evoke particular content)
- Florid, ever-evolving, internally generated visual and somatic percepts

Post-acute integration (60–120 minutes):

Spontaneous, free-association journaling on acute trip percepts and evoked relational content on reflection.

Example. Perceptual symbolic content

(Session 43, post-acute journal)

I say:

These things happened.

Let the Ocean
of lifetimes that could have been
wash over me.

Allow—
the unbearable torture of absence,
raging torrent of pink and blue.

My body—
dissolved.

They whisper:

Yes it was, yes it was, yes it was.

Yes it was, yes it was, yes it was.

Yes it was, yes it was, yes it was.

Yes it was, yes it was, yes it was.

Example. Perceptual symbolic content

(Session 43, post-acute journal)

I say:

Take me to see the Child—
my first Vision, so long ago.
The little boy,
curled in a ball,
his face hidden.

It's ok—
I'm here now.
Let me hold you.
Let me share the burden.
Let me lift you up.

I reach out my hand:

Just turn to me.
Show me your face.
Show me who you really are.

*The little girl turns her head
and looks at me.*

Example. Perceptual symbolic content

(Session 43, post-acute journal)

The Ocean

of lives that could have been.

Moments—

a yearbook photo,

a family dinner,

a white coat ceremony,

a trip to Poland,

Her face—

my face,

My body—

the *right* body.

Not dissolved—

here-now.

The wave travels:

I have hair.

I have a face.

I have fingers.

I have a navel.

Farther.

I have thighs.

I have calves.

I have toes.

I am a *human*.

Example. Perceptual symbolic content

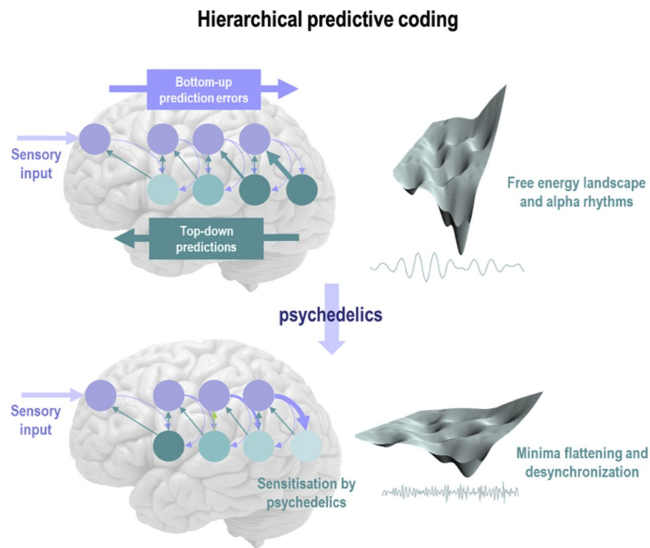
(Session 43, post-acute journal)

Before me,
 the Moon passes in front of the Sun.
In the shadow of Eclipse,
from the Ocean, I rise—
 Anima, resurrected;
 golden daughter;
 Woman of the Water,
 heart of Fire.

The Sun returns.

I dance through sunbeams,
 into the faraway and forever starlight,
 and rejoice that I am *found*.

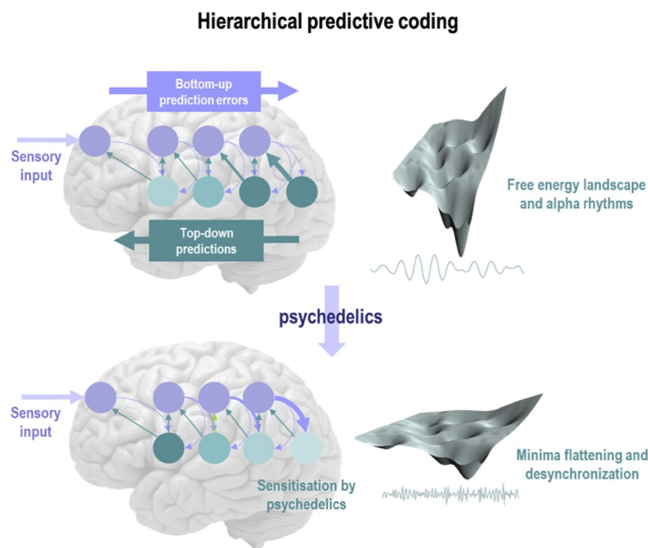
How do psychedelics work?



One contemporary hypothesis: **Relaxed beliefs under psychedelics** (Carhart-Harris and Friston, 2019).

Reproduced from *ibid.*

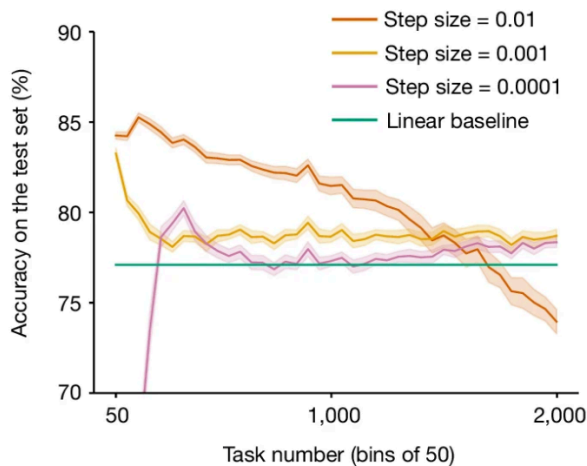
How do psychedelics work?



Reproduced from *ibid.*

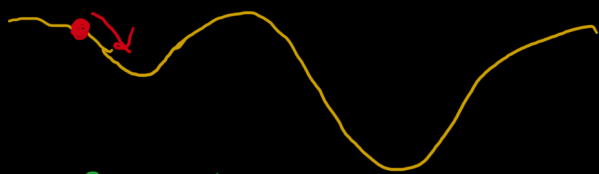
- Psychedelics reduce impact of prior beliefs on shaping experience
- Seen as a transient flattening of the energy landscape of brain dynamics
- Allows exploration of previously-inaccessible parts of world model-space

A larger picture of homeostasis in learning systems

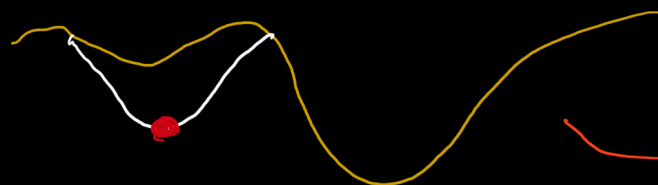
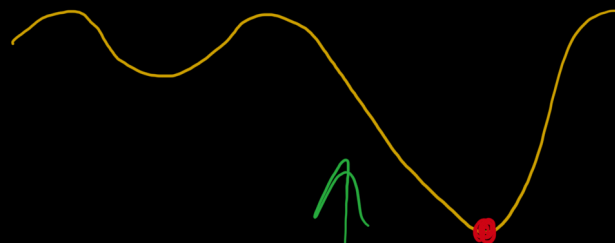


Reproduced from [Dohare et al. 2024](#).

- AI training with gradient descent fails in continuous learning
- Problem is that **early descent into a local minimum** is unrecoverable
- Requires a **continuous regularization** strategy



DA?
NE?



5-HT?

Overheard at UCSF

The 2023 Samuel Barondes Lecture in Biological Psychiatry



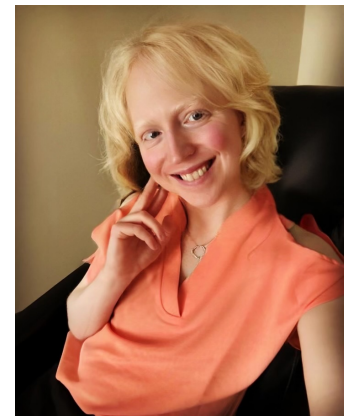


"And so the psychedelic experience—I think it's probably just noise, that's it. And the brain is very good at making meaning out of nothing. ... And so it just makes up a story about what's going on. And for reasons that are not entirely clear to me, this story that's made up when people take psychedelics is very meaningful to them."

(Image: @BryanRoth on X, annotation mine)



(a) Maxwell



(b) Maxine

Figure 1: "Take me to see the Child ... The little boy, curled in a ball, his face hidden. ... Just turn to me. Show me your face. Show me who you really are."

"The little girl turns her head and looks at me."

The origins of bad prior beliefs in psychiatry

Not difficult to locate

When the patient lashes out against "them"—


THORAZINE®
brand of chlorpromazine


quickly puts an end to his violent outburst

'Thorazine' is especially effective when the psychotic episode is triggered by delusions or hallucinations.

At the outset of treatment, Thorazine's combination of antipsychotic and sedative effects provides both emotional and physical calming. Assaultive or destructive behavior is rapidly controlled.

As therapy continues, the initial sedative effect gradually disappears. But the antipsychotic effect continues, helping to dispel or modify delusions, hallucinations and confusion, while keeping the patient calm and approachable.

 SMITH KLINE & FRENCH LABORATORIES
leaders in psychopharmaceutical research

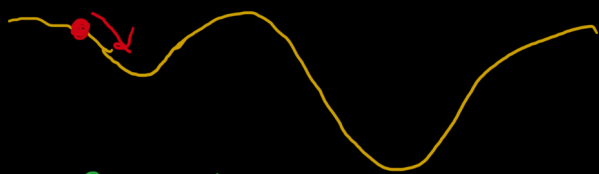


A reminder advertisement — For prescribing information, please see **PDS** or available literature.

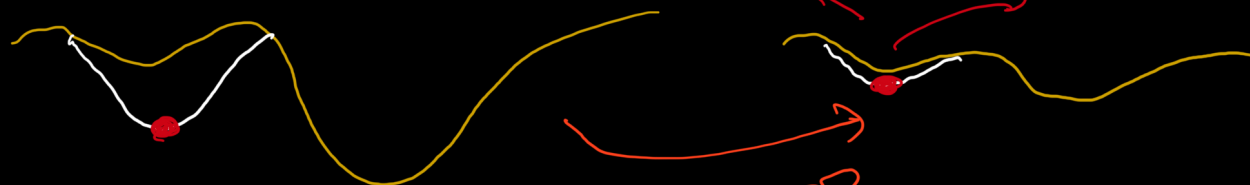
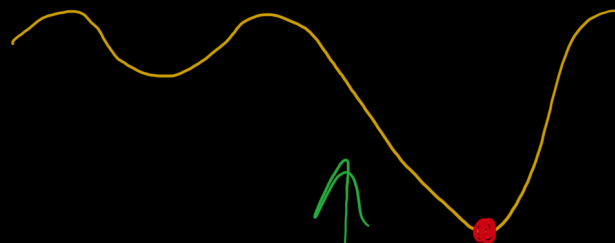
A hopeful proposal

Maybe the reason mental suffering is so difficult to alleviate *isn't* because we haven't found the right neural representation of a **disease**, or the right circuit to stimulate that **cures** it.

Maybe the suffering of the mentally different **arises from that very story** we tell them as scientists and as a society—
—the story that there is a disease to be cured.



DA?
NE?



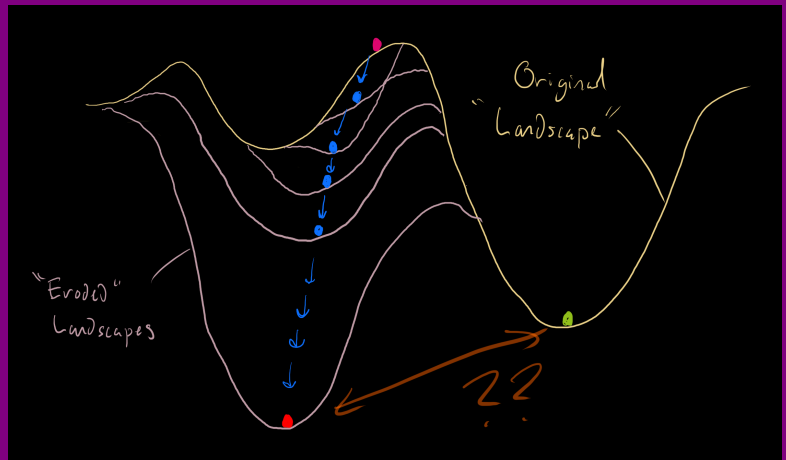
S-HT?

But the good news is, every year we get a special infusion of energy and creative ideas, a flattening of our energy landscape...

**When new students—with fresh new perspectives
—arrive here!**

Summary

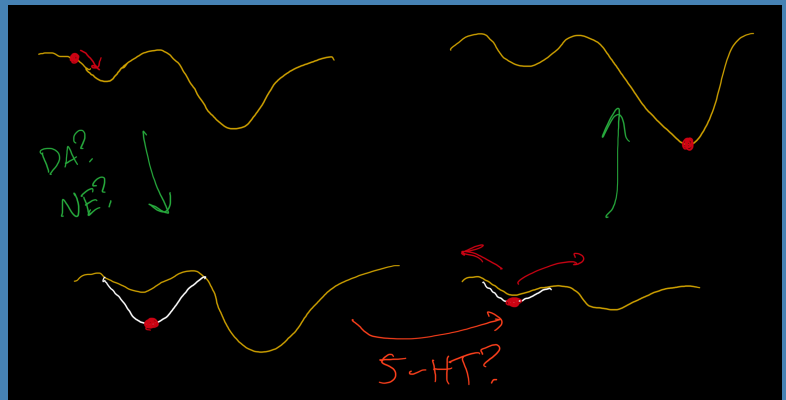
The structures of power were formed through centuries of erosion—



—from initial conditions that bias everything we think to benefit a select few.

They will make you hurt if you do not conform.

But true resistance
lies in allowing:



Allowing yourself the freedom of
stepping outside the constraints of
history—outside who you are told you
must be.

The freedom of becoming you.



Unbreakable you.

And that's Statistics

Welcome to UCSF!



Maxine