Now What?

Model Building, Science, and Analytics

Phil Henrickson, PhD
AE Business Solutions
Delivered 05/04/2022

First, a brief survey.

Raise your hand if...

Raise your hand if...

You have taken an astronomy class.

Raise your hand if...

You have taken an astronomy class.

You have played a board game in the last year.

Raise your hand if...

You have taken an astronomy class.

You have played a board game in the last year.

You have watched **The Good Place**.

More to come on this later.

Now What?

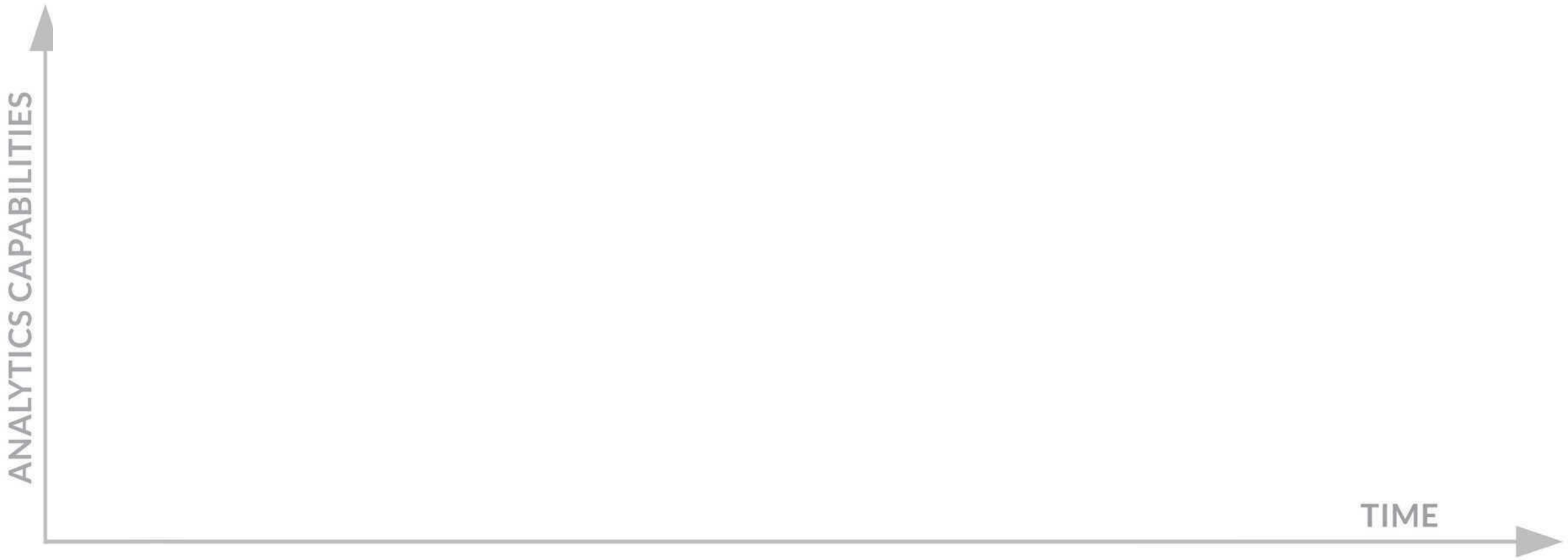
Model Building, Science, and Analytics

Phil Henrickson, PhD
AE Business Solutions

We've been seeing the same sentiment expressed amongst our clients more and more:

"We've modernized our data warehouse, we have all this data, we built all these dashboards... **Now what do we do**?"

Most of our clients have discovered problems that dashboards and reports cannot solve.



Most of our clients have discovered problems that dashboards and reports cannot solve.

They've gotten here...

Raw Data -> Clean Data -> Dashboard -> ??? -> #DataDrivenInsight™

something was supposed
to happen here


... now they want to get here..
They've gotten here...
ANALYTICS CAPABILITIES
TIME
01
OMG!
We have a lot of data!
02
DATA FORTRESS
Data is stored, secured, available
03
BASIC REPORTING
Answer, "What happened?" not intuitive
04
BUSINESS INTELLIGENCE
Answer, "What happened, when and where?" more intuitively
05
STATISTICAL ANALYSIS
Answer, "Why?" questions; project what happens if current trends continue
06
PREDICTIVE MODELS
Illuminate what will happen in the future
07
PRESCRIPTIVE MODELS
Evaluate predicted future alternatives; recommend optimal actions
08
PROCESS OPTIMIZATION
Enabled by analytics-derived insights: transformation underway

Raw Data -> Clean Data -> Model -> Dashboard -> ??? -> #AIDrivenInsight™

surely, something will happen here

The analytics community has been in a bit of a frenzy over **data science**.

The analytics community has been in a bit of a frenzy over **data science**.

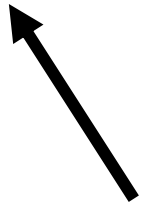
We need to ignore this for a second.

The analytics community has been in a bit of a frenzy over **data science**.

We need to ignore this for a second.

We need to spend more time talking about this.

science.



We need to spend more time talking about this.

science.

We are all pretty comfortable calling ourselves data people.

We hear the term ‘data-driven’ all the time.

science.

Should we strive to be data driven?

science.

Should we strive to be data driven?

I would argue no.

"JUST EXTRAORDINARY." —*SCIENCE FRIDAY* (NPR)

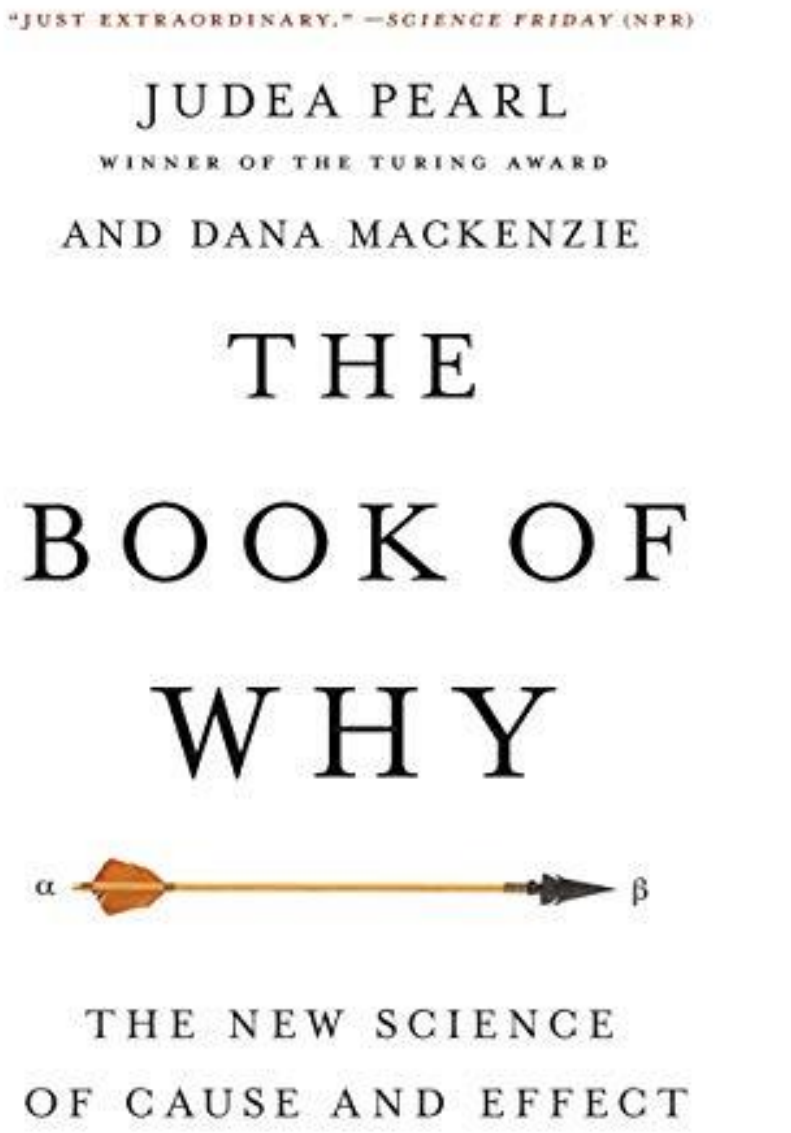
JUDEA PEARL

WINNER OF THE TURING AWARD

AND DANA MACKENZIE

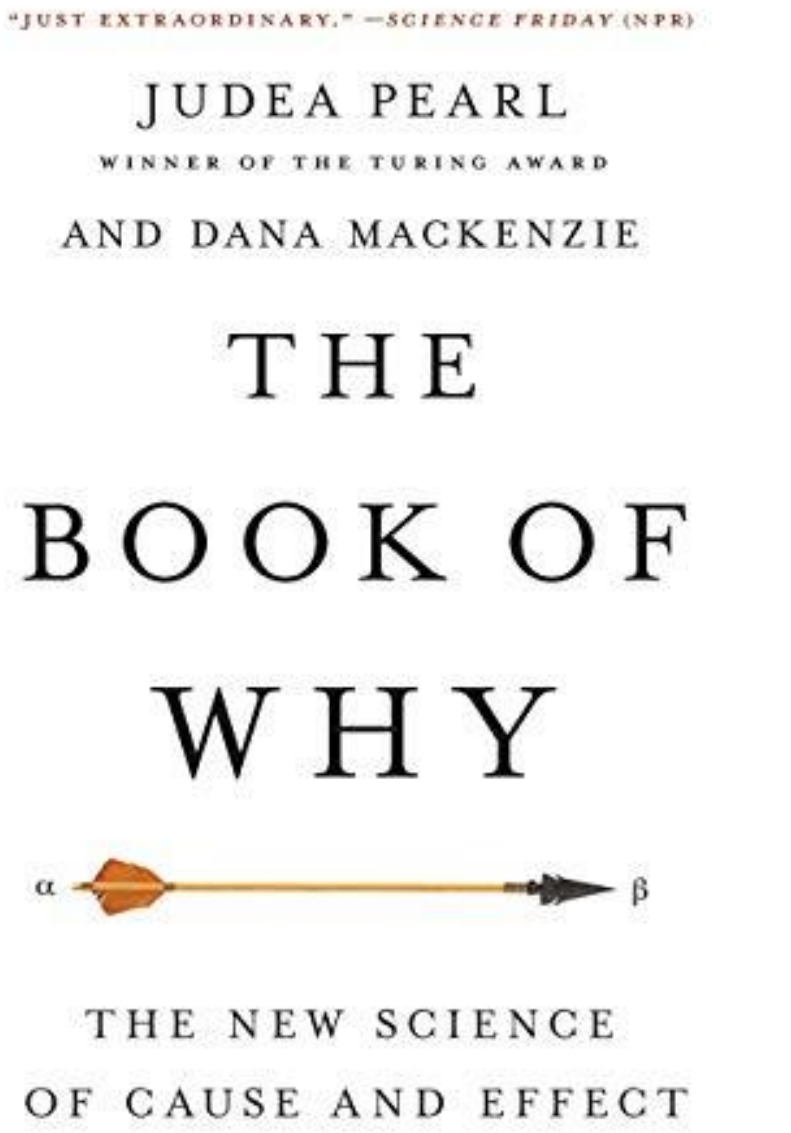
THE BOOK OF WHY

THE NEW SCIENCE
OF CAUSE AND EFFECT



I hope to convince you that **data are profoundly dumb**.

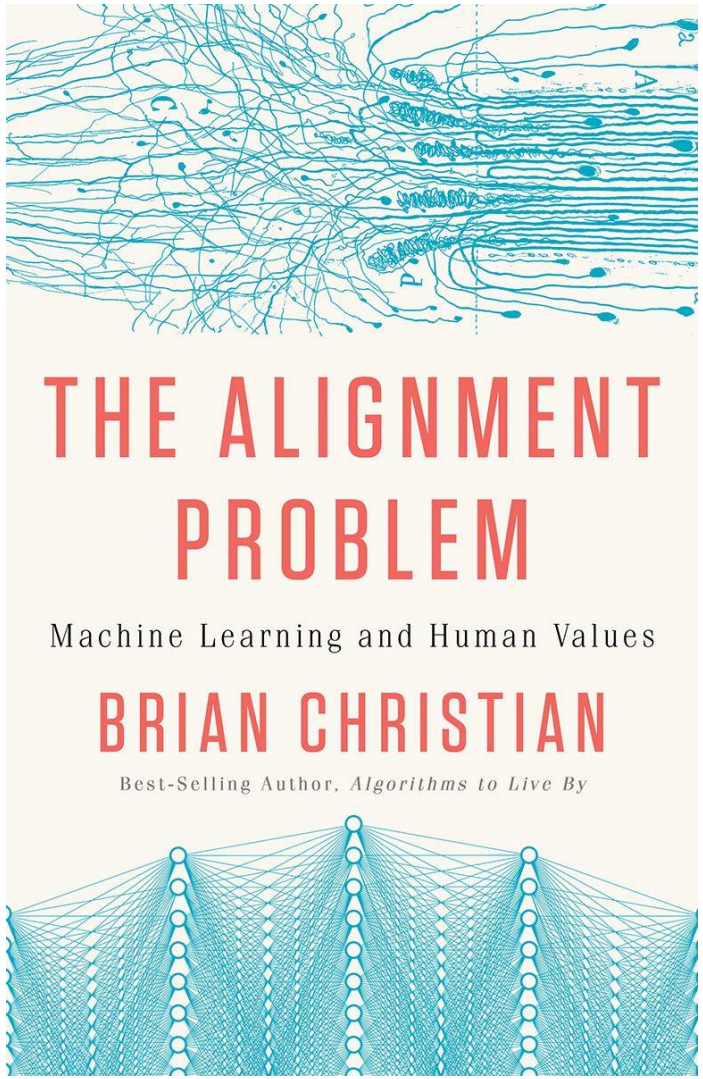
No machine can derive explanations from raw data… Data can tell you that the people who took a medicine recovered faster than those who did not take it, but **they can't tell you why**.

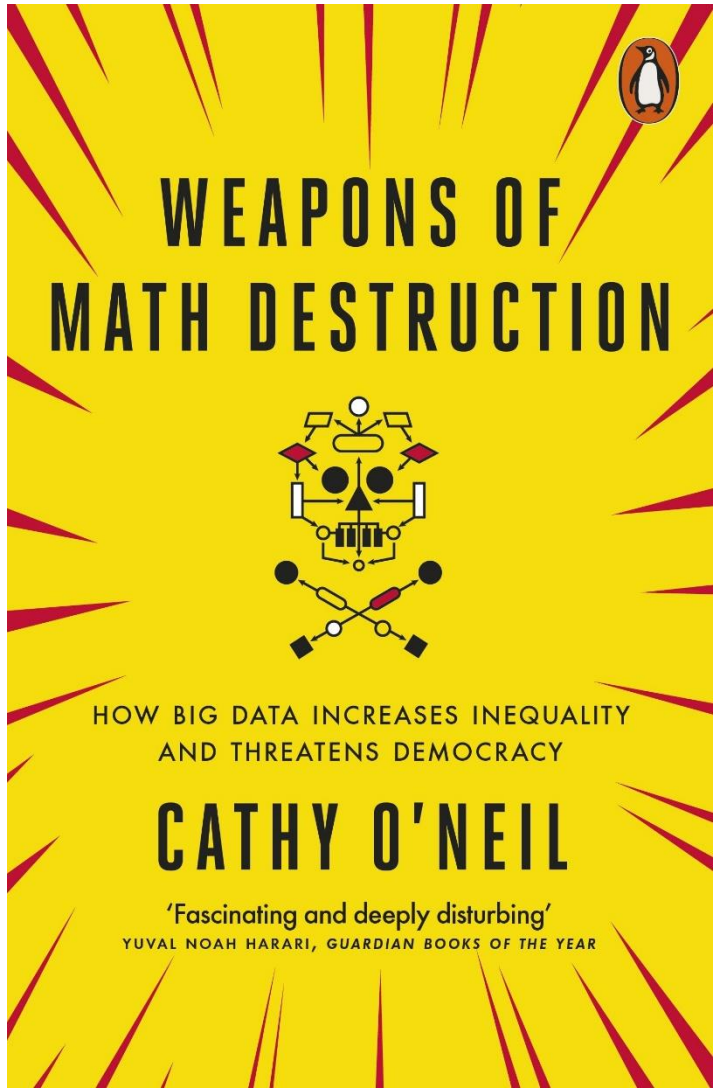


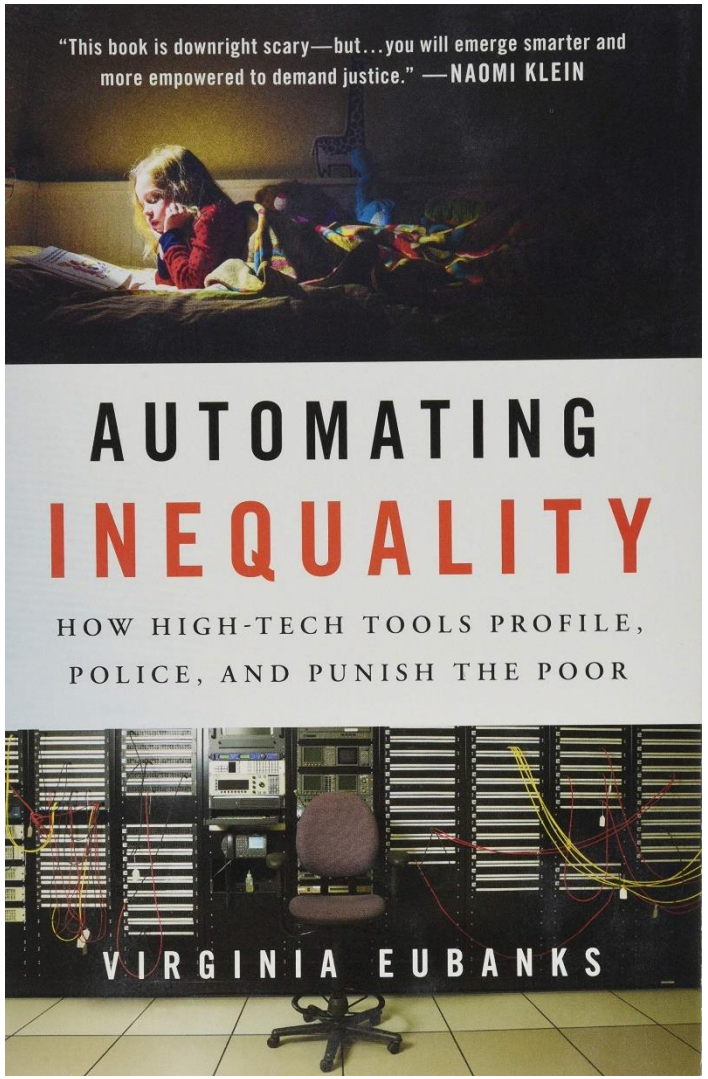
Over and over again, in science and in business, we see situations where mere **data aren't enough**.

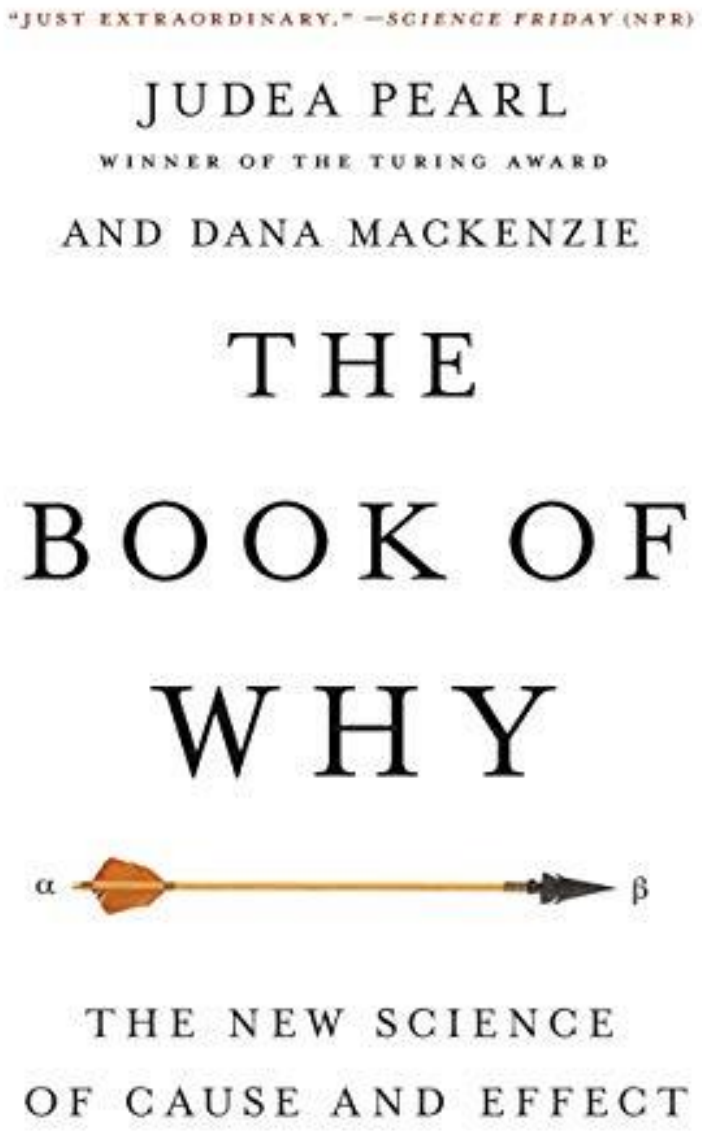
The hope… is that **the data themselves will guide us to the right answers** whenever causal questions come up.

If you'd like to read examples of data guiding to wrong answers:









If I could sum up the message of this book in one pithy phrase, it would be that **you are smarter than your data**.

Data do not understand causes and effects; **humans do**.

science.

science.

Data is an ingredient. It isn't a recipe.

By itself, data offers no guarantee of learning.

science.

Data is necessary for learning.
But it is not sufficient.

science.

Data is necessary for learning.
But it is not sufficient.

In order to learn from data, we have to use a methodology.

Raw Data -> Clean Data -> Dashboard -> ??? -> #DataDrivenInsight™

we were hoping to learn
something from data

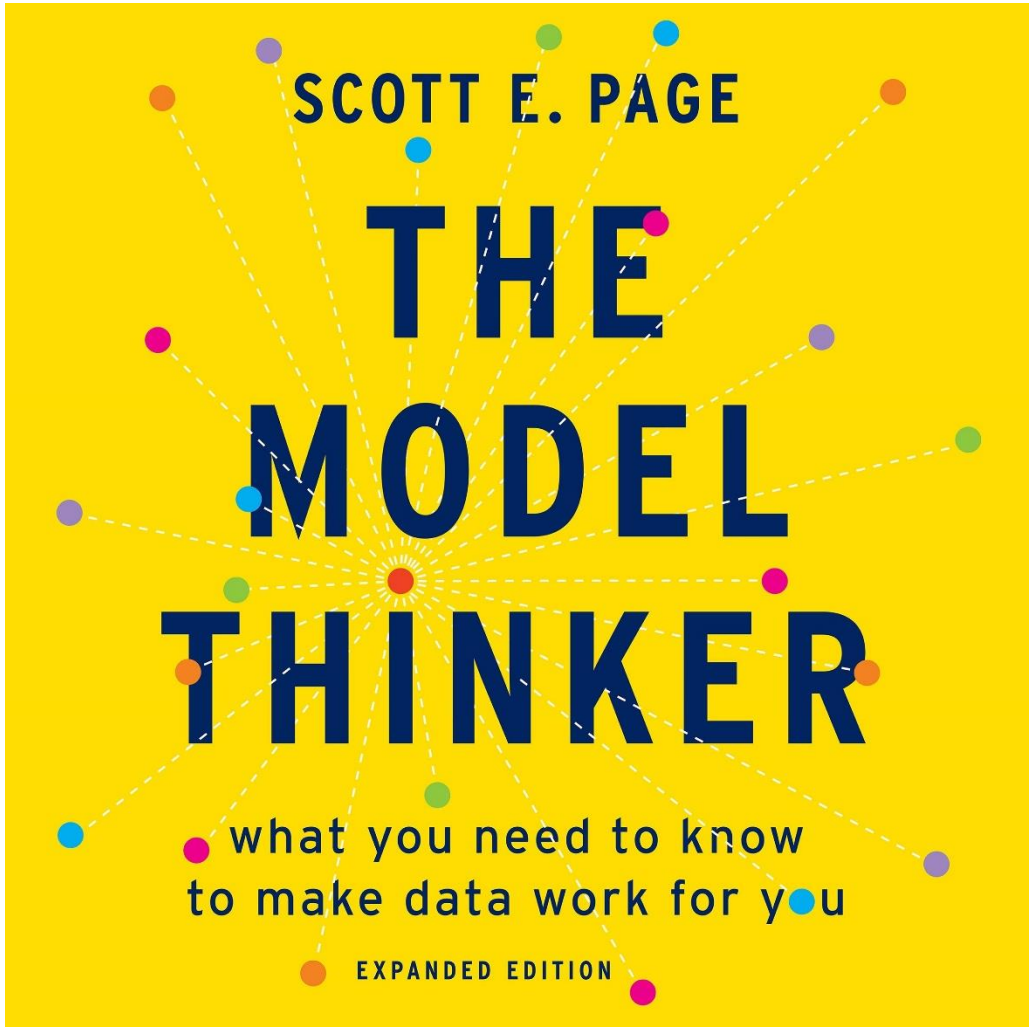
Raw Data -> Clean Data -> Model -> Dashboard -> ??? -> #AIDrivenInsight™

science.

For all of the time we spend working with data, I don't think we spend enough time talking about how we use it to learn.

science.

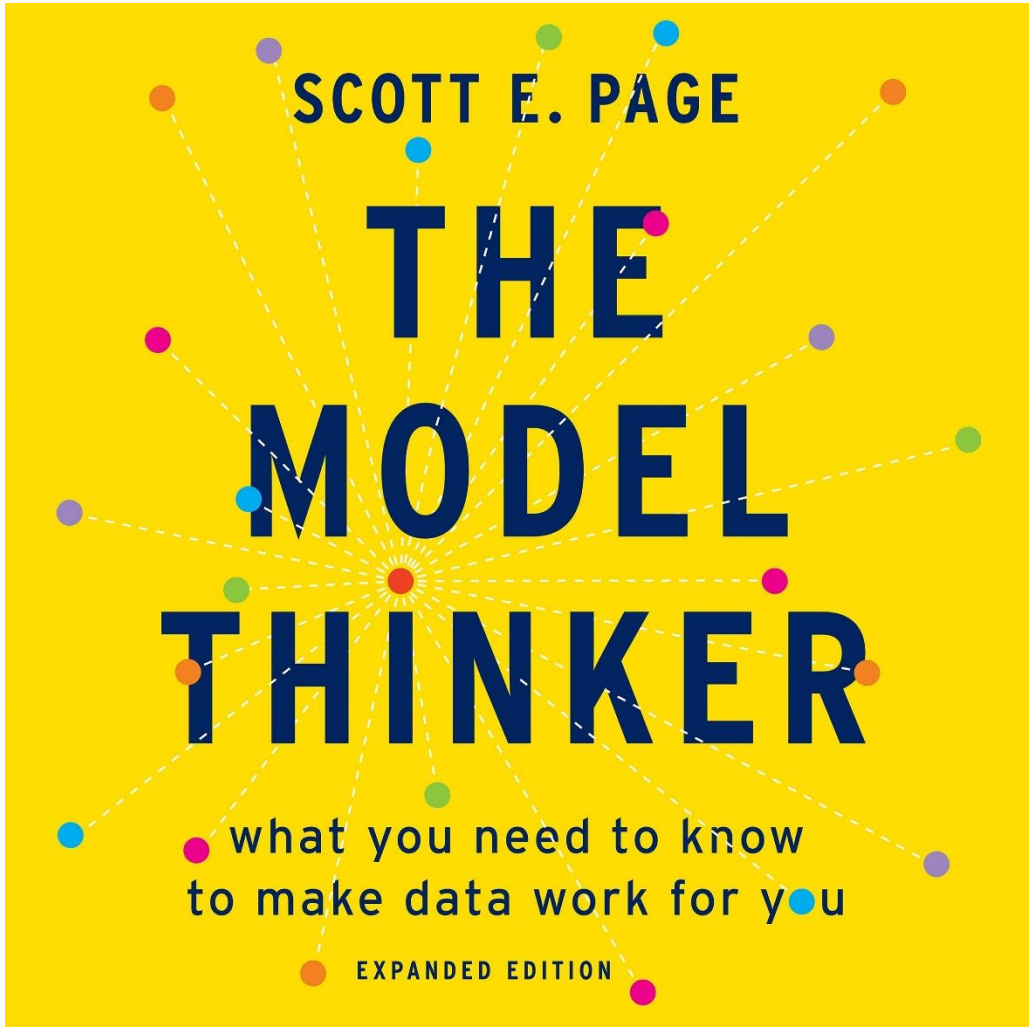
That brings us to the topic of the day:
building models.


SCOTT E. PAGE
THE MODEL THINKER
what you need to know to make data work for you
EXPANDED EDITION

The rise of model thinking has [a simple explanation]: **models make us smarter**.

Without models, we have limited capacity to include data. With models, we clarify assumptions and think logically.

With models, we think better.



today

We need to talk about models.

today

We need to talk about why we build models.

today

We need to talk about why we build models.
We need to talk about how we build models.

today

We need to talk about why we build models.
We need to talk about how we build models.
We need to talk about using models.

today

Part 1	We need to talk about why we build models.
Part 1 & 2	We need to talk about how we build models.
Part 2	We need to talk about using models.

let’s begin.

let’s begin.

hold onto your butts.

1 Building Models of the Solar System

And Everything Else Along the Way

In a former life, I taught classes on research methods, international conflict, political violence, and civil war.

Regardless of the subject matter, I started every semester with the same lecture.

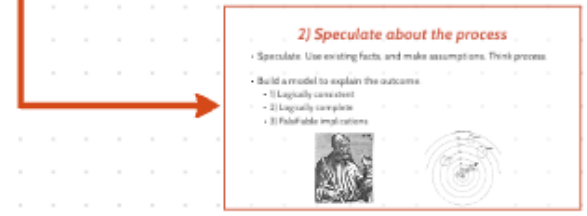
How to be a (Social) Scientist

A Model of the Model Building Process

- A model is a simplified representation of the real world.
- We create them by speculating about the processes that could have produced the observed facts
- We evaluate models in terms of their ability to correctly predict facts, their generalize-ability, and their simplicity
- 1) Observe
- 2) Speculate
- 3) Deduce implications
- 4) Test implications

1) Observe

- Observe some facts about the world
  - A dictatorship engages in mass atrocities against his own citizens
  - An individual bombs a building in his own country
- Puzzles



2) Speculate about the process

- Speculate. Use existing facts, and make assumptions. Think process
- Build a model to explain the outcome
  - 1) Logically consistent
  - 2) Logically complete
  - 3) Falsifiable implications

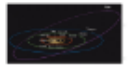
Let's play a game!

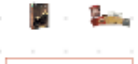
3) Think of other implications

- If our model is correct, what else would we expect to observe?
- This is where we develop hypotheses to test falsifiable implications which should come from our model.
  - If X, then Y
  - If X, then Z

4) Test these implications

- Implications must be Falsifiable
- Gather data. Put them to the test.





How to be a (Social) Scientist

How to be a (Social) Scientist

People tend to diss on the social sciences, mostly because they have a misunderstanding of what social scientists are up to.

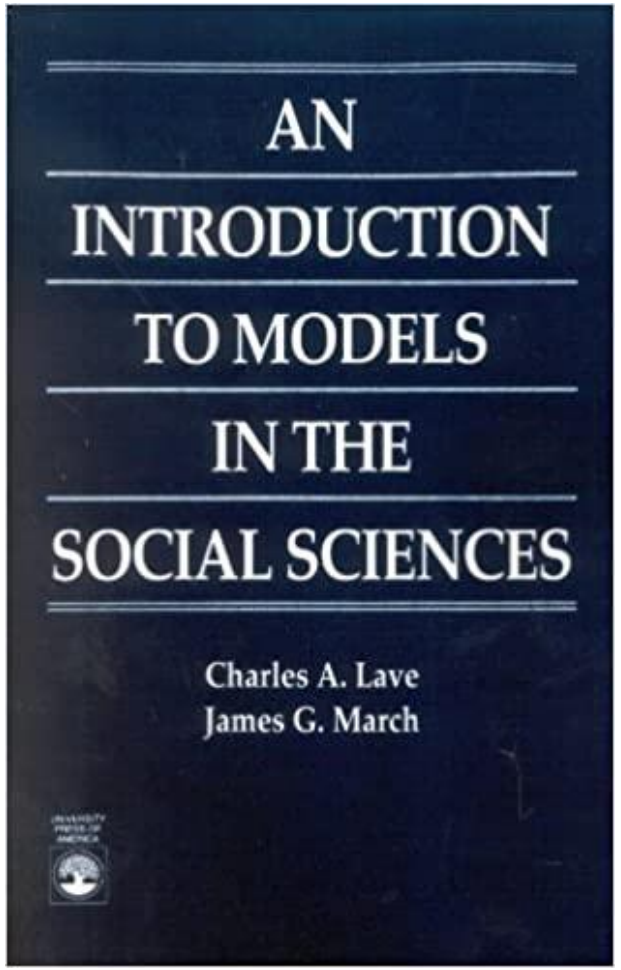
So, naturally, I immediately tried to win them over using the coolest possible method:

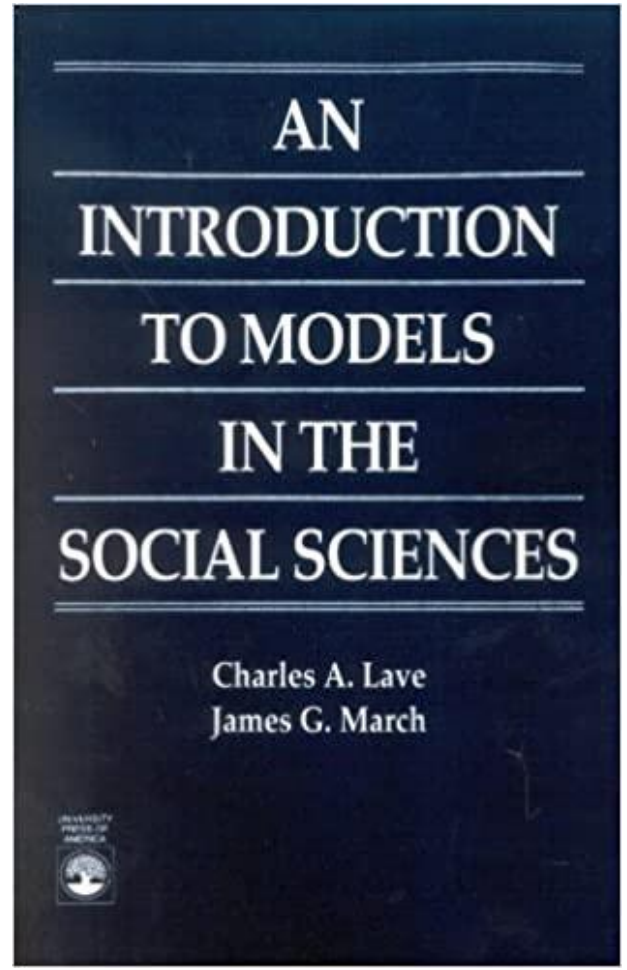
How to be a (Social) Scientist

People tend to diss on the social sciences, mostly because they have a misunderstanding of what social scientists are up to.

So, naturally, I immediately tried to win them over using the coolest possible method:

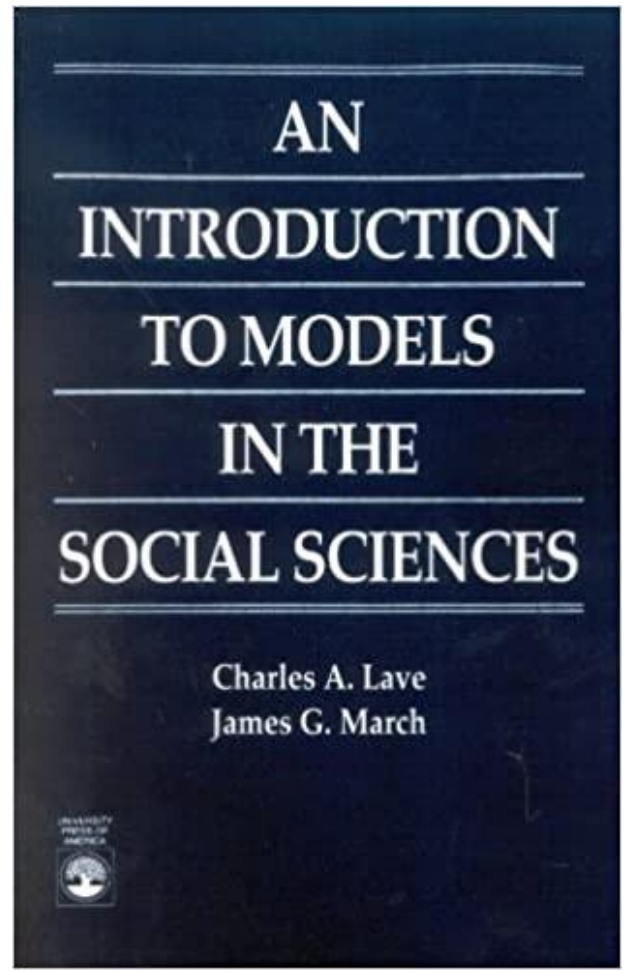
a social sciences textbook from 1975




AN
INTRODUCTION
TO MODELS
IN THE
SOCIAL SCIENCES
Charles A. Lave
James G. March

chapter one

what we are up to



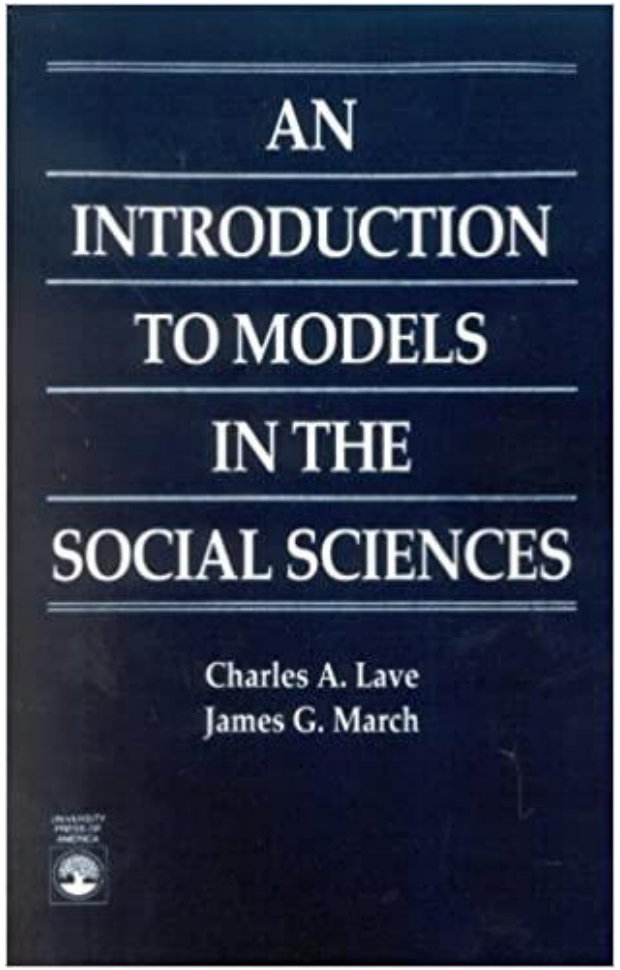
chapter one

what we are up to

What is a model? How do you build one? What makes a good model?

Speculative models are central to science, history, and literature. We are **constantly forming partial interpretations of the world** in order to live in it.

We think **that an increase in the quality of speculation** both in the social sciences and in everyday life would be good.



Lave and March are known for their emphasis on **speculation**.

They tell us to look at data as the end result of some process. Our goal is to **understand the process that produced the data**.

To do this, we need to speculate. We need to **stop and think**.

The best way to learn about building models is to do it.

This is the first example from their book.

Suppose we are interested in understanding why **some people are friends and not others**.

Suppose we are interested in understanding why **some people are friends and not others**.

We gather data on a college campus by asking residents of dormitories to give us a list of their friends.

Suppose we are interested in understanding why **some people are friends and not others**.

We gather data on a college campus by asking residents of dormitories to give us a list of their friends.

We notice a pattern: **friends tend to live close to each other**; they tend to have adjacent dormitory rooms.

Why might this be? **Stop and think**.

One possible explanation:

Campus housing lets students choose where to live in their dormitories. Students prefer to live by their friends. So, students ask campus housing to have friends as roommates or be put in adjacent rooms.

One possible explanation:

Campus housing lets students choose where to live in their dormitories. Students prefer to live by their friends. So, students ask campus housing to have friends as roommates or be put in adjacent rooms.

This is **our speculation about the process** that produced the data we observed.

That is, we have a basic model of the **prior** state of the world which may be able to account for what we observe in the **current** state of the world.

One possible explanation:

Campus housing lets students choose where to live in their dormitories. Students prefer to live by their friends. So, students ask campus housing to have friends as roommates or be put in adjacent rooms.

Is this a good model?

We have to ask: **if this model is correct, what else should we expect to observe**?

One possible explanation:

Campus housing lets students choose where to live in their dormitories. Students prefer to live by their friends. So, students ask campus housing to have friends as roommates or be put in adjacent rooms.

One possible explanation:

Campus housing lets students choose where to live in their dormitories. Students prefer to live by their friends. So, students ask campus housing to have friends as roommates or be put in adjacent rooms.

This model assumes that students **had already known each prior to the start of the semester**. It would predict different patterns of friendships between freshman and upperclassman dorms.

But, we notice this same pattern in a freshman dormitory.

One possible explanation:

Campus housing lets students choose where to live in their dormitories. Students prefer to live by their friends. So, students ask campus housing to have friends as roommates or be put in adjacent rooms.

Can our model still explain the pattern? Probably not.

So what do we? Try again. **We stop and think**.

Another possible explanation:

College students come from similar backgrounds and have a lot in common. Students who live near each other will frequently interact and discover what they have in common, leading to friendship..

Another possible explanation:

College students come from similar backgrounds and have a lot in common. Students who live near each other will frequently interact and discover what they have in common, leading to friendship..

This would explain why we observe clusters of friends in all dormitories, including freshman.

Does that mean our model is correct?

Another possible explanation:

College students come from similar backgrounds and have a lot in common. Students who live near each other will frequently interact and discover what they have in common, leading to friendship..

This would explain why we observe clusters of friends in all dormitories, including freshman.

Does that mean our model is correct?

No! We need to **develop more implications**, then **gather data to put them to test**.

How to be a (Social) Scientist

A model is a simplified representation of the world.

We create models by **speculating about the processes** that could have **produced the data that we observe**.

How to be a (Social) Scientist

A model is a simplified representation of the world.

We create models by **speculating about the processes** that could have **produced the data that we observe**.

We evaluate models in terms of their ability to predict what we observe, their ability to generalize, and their simplicity.

How to be a (Social) Scientist

A model of the model building process looks like this:

How to be a (Social) Scientist

A model of the model building process looks like this:

1) **We observe**. We notice a pattern or result that has occurred in the world.

How to be a (Social) Scientist

A model of the model building process looks like this:

1) **We observe**. We notice a pattern or result that has occurred in the world.

2) **We speculate**. We develop an explanation for the process that could have produced our observation.

How to be a (Social) Scientist

A model of the model building process looks like this:

1) **We observe**. We notice a pattern or result that has occurred in the world.

2) **We speculate**. We develop an explanation for the process that could have produced our observation.

3) **We develop implications**. We ask, if our speculation is correct, what else should we expect to observe?

How to be a (Social) Scientist

A model of the model building process looks like this:

1) **We observe**. We notice a pattern or result that has occurred in the world.

2) **We speculate**. We develop an explanation for the process that could have produced our observation.

3) **We develop implications**. We ask, if our speculation is correct, what else should we expect to observe?

4) **We test**. We look to see whether the other implications of our model are supported in the data.

How to be a (Social) Scientist

A model of the model building process looks like this:

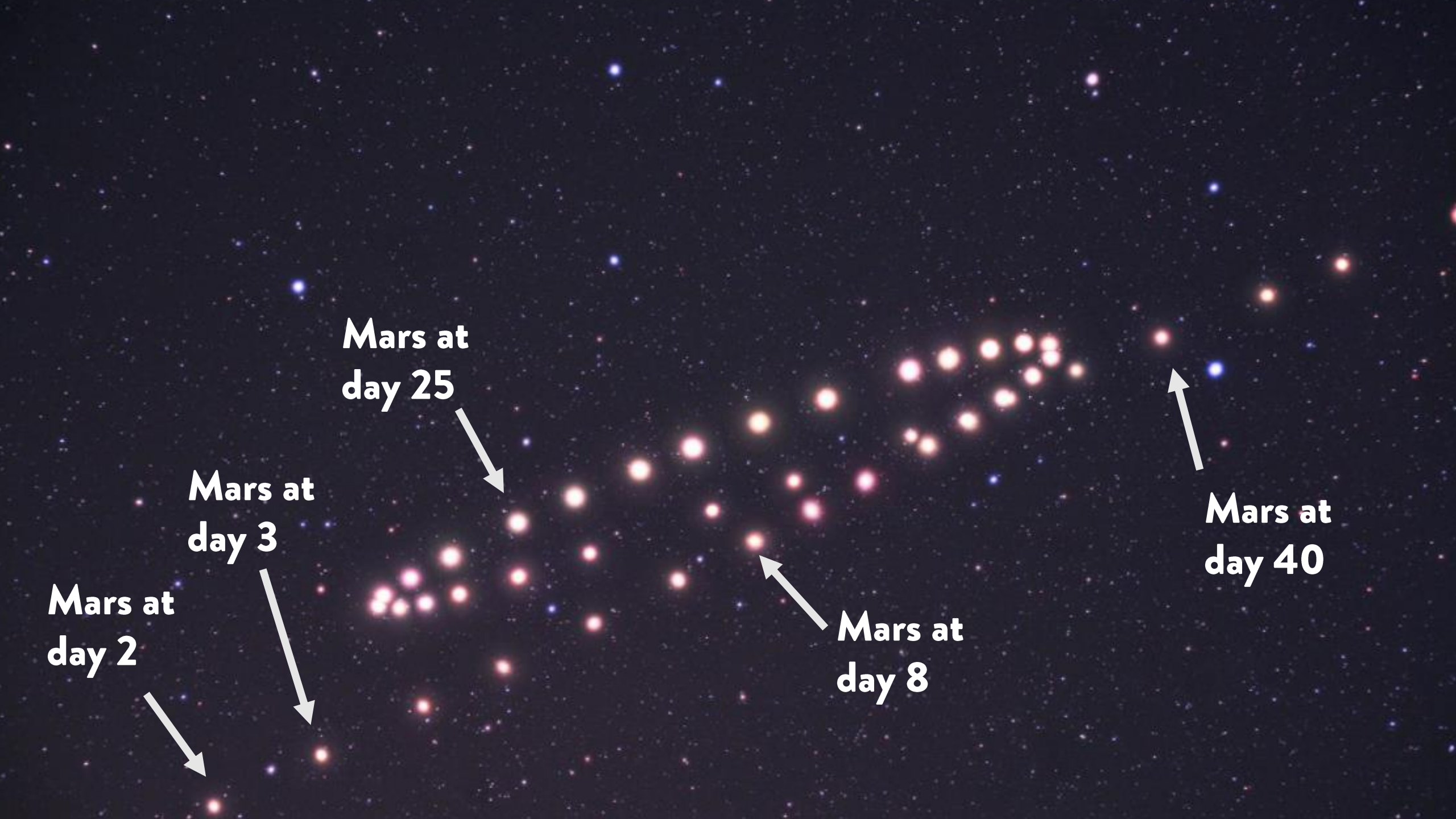
1) **We observe**.

2) **We speculate**.

3) **We develop implications**.

4) **We test**.

This model building process is applicable far beyond the social sciences.


Mars at day 25
Mars at day 3
Mars at day 2
Mars at day 8
Mars at day 40

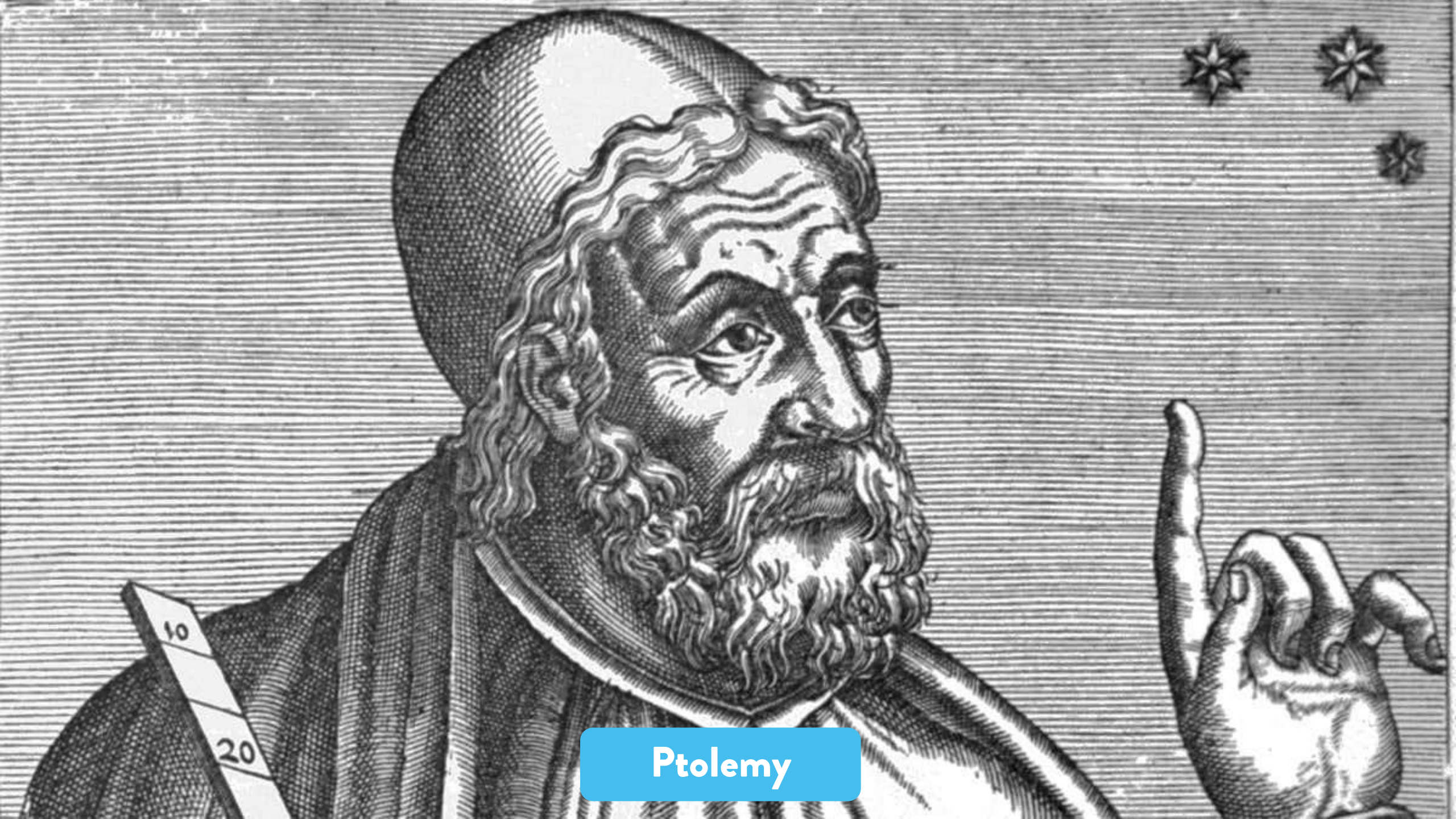
Why does Mars move backwards in the nighttime sky?

Model
(Speculation)

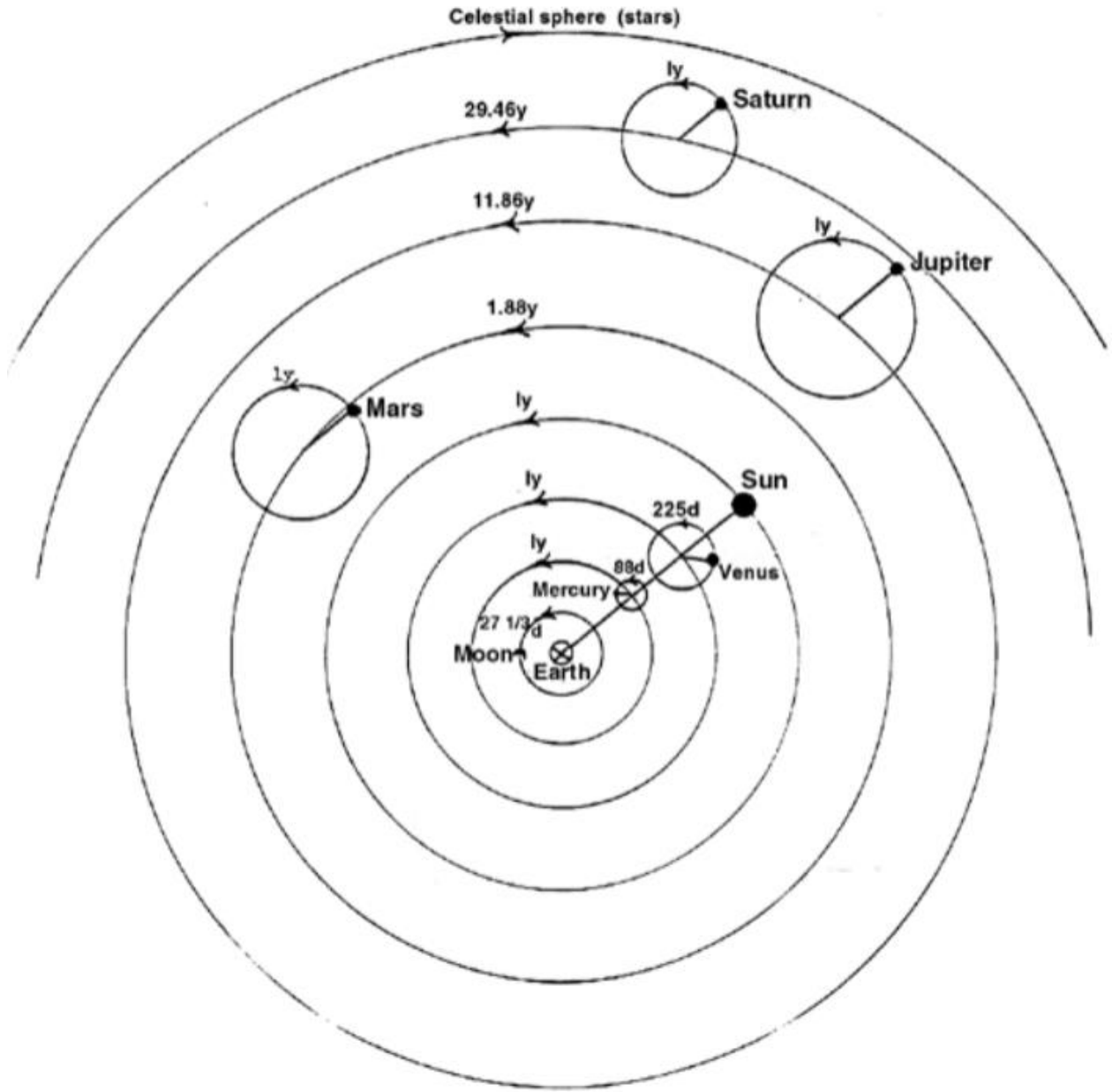
The **Earth is at the center** of the solar system.
The heavens are in perfect harmony and **objects orbit the Earth in circles**.

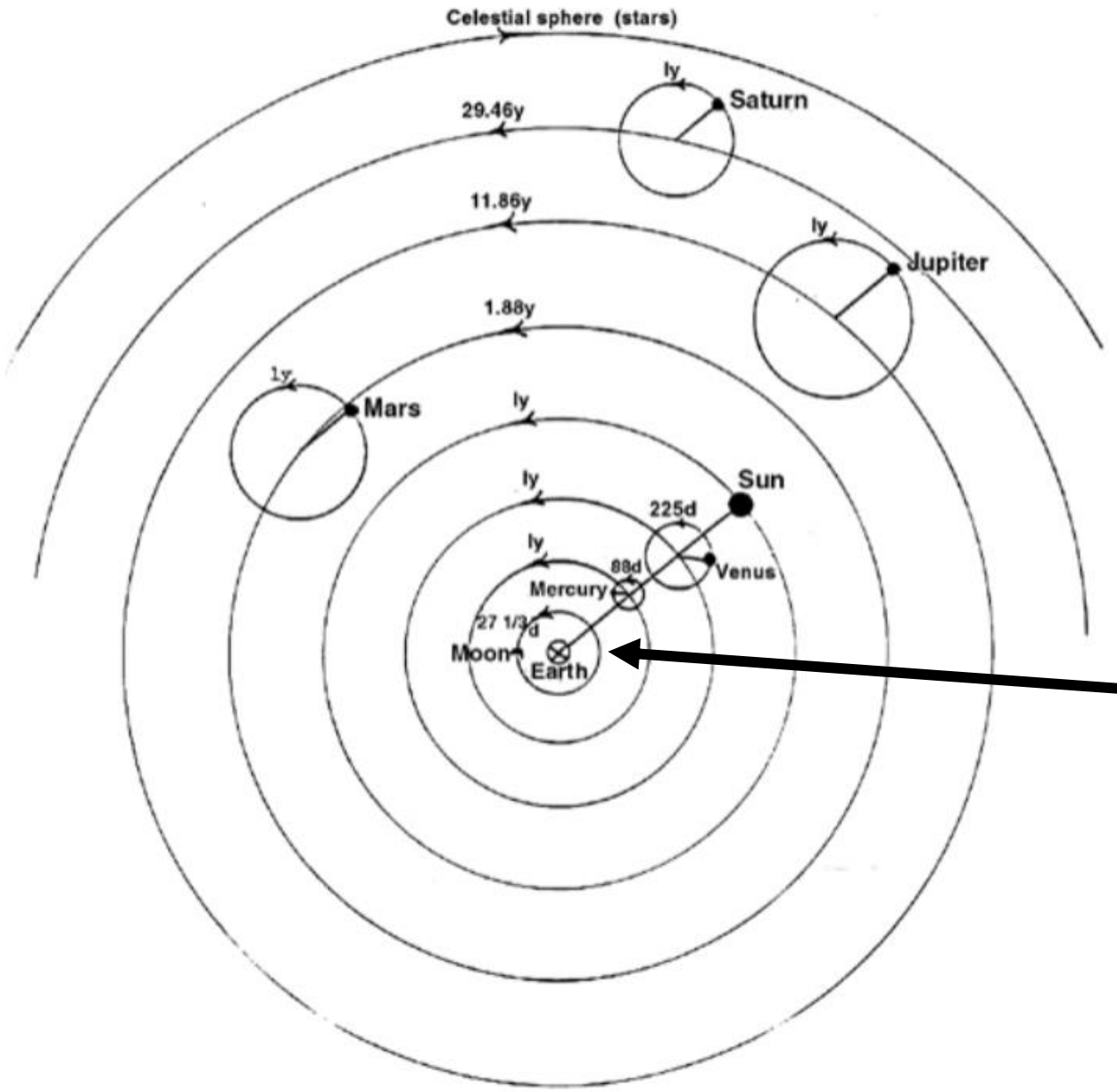
Model (Speculation)	The **Earth is at the center** of the solar system. The heavens are in perfect harmony and **objects orbit the Earth in circles**.
Implication	If that's the case, planets shouldn't move backwards.

Model (Speculation)	The **Earth is at the center** of the solar system. The heavens are in perfect harmony and **objects orbit the Earth in circles**.
Implication	If that's the case, planets shouldn't move backwards. But they do.



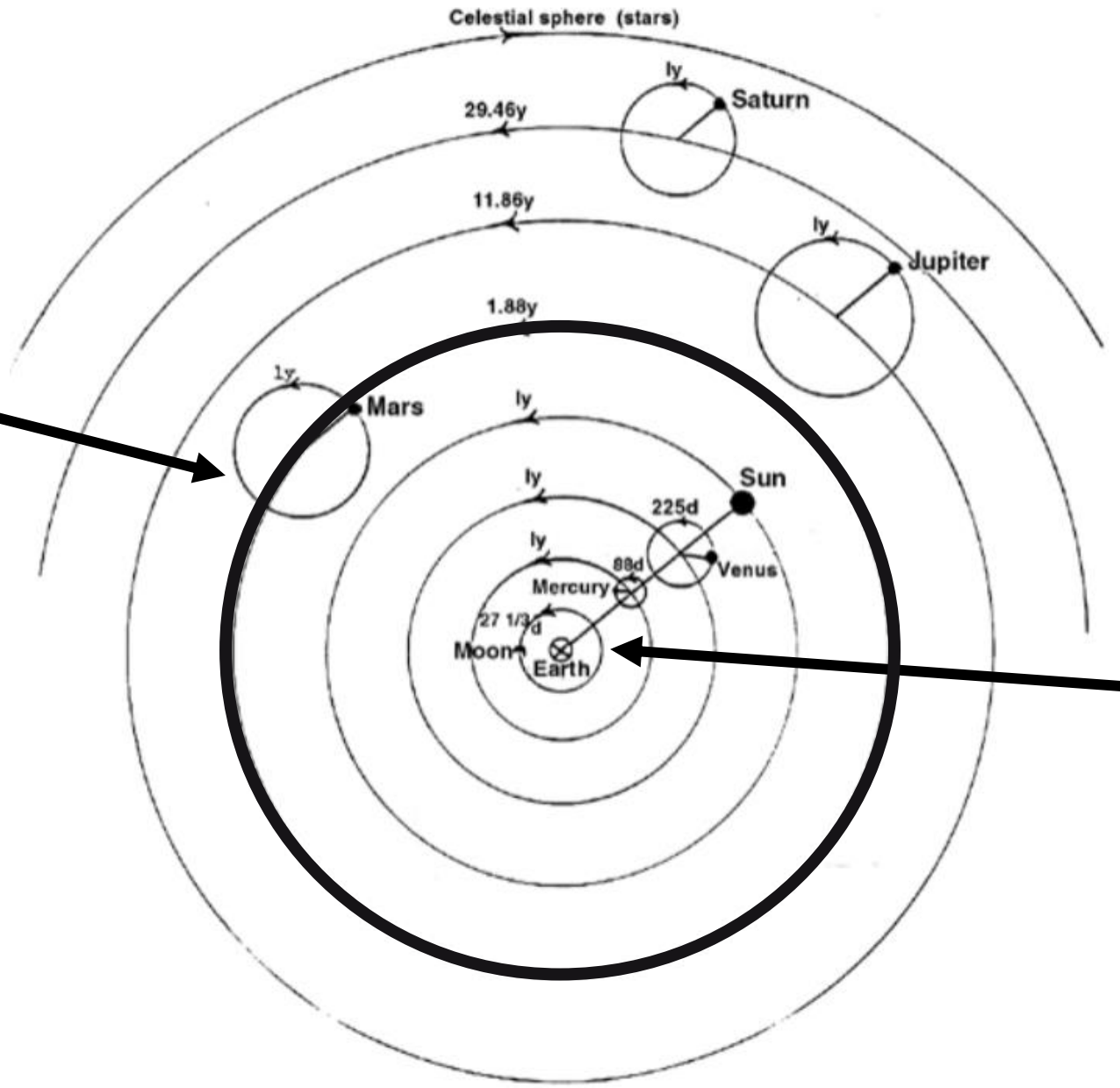
Ptolemy


Celestial sphere (stars)
29.46y
1y
Saturn
11.86y
1y
Jupiter
1.88y
1y
Mars
1y
1y
Sun
225d
1y
88d
Venus
Mercury
27 1/3d
Moon
Earth



Earth at the center

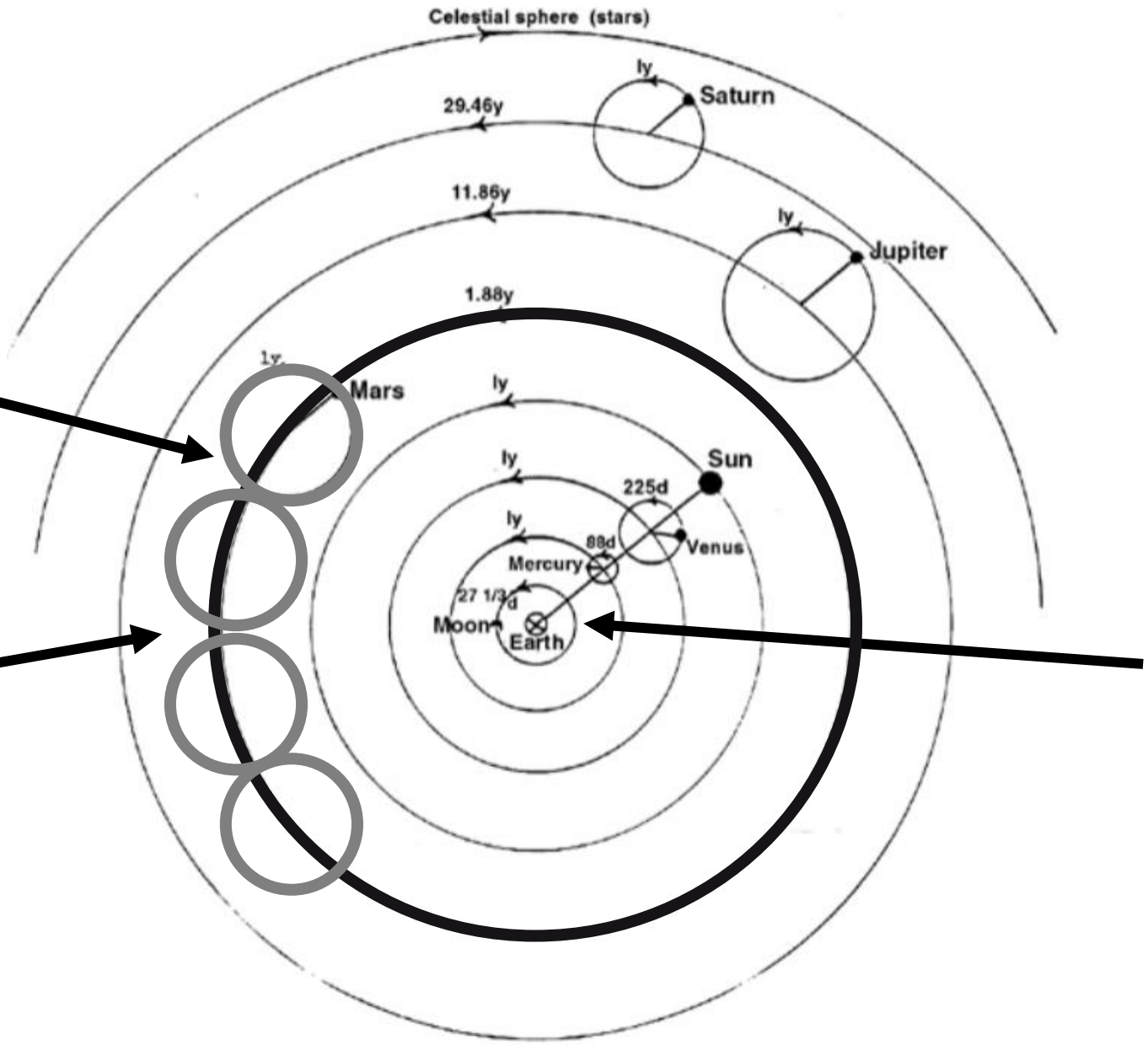
Planets orbit the earth in circles..



Earth at the center

Planets orbit the earth in circles..

...but they're also making other circles.
It's circles on circles!



Earth at the center

If Ptolemy's model is correct, what should we observe?

If Ptolemy's model is correct, what should we observe?

It should be able to predict the movement of the planets.

If Ptolemy's model is correct, what should we observe?

It should be able to predict the movement of the planets.

And it does!

If Ptolemy's model is correct, what should we observe?

It should be able to predict the movement of the planets.

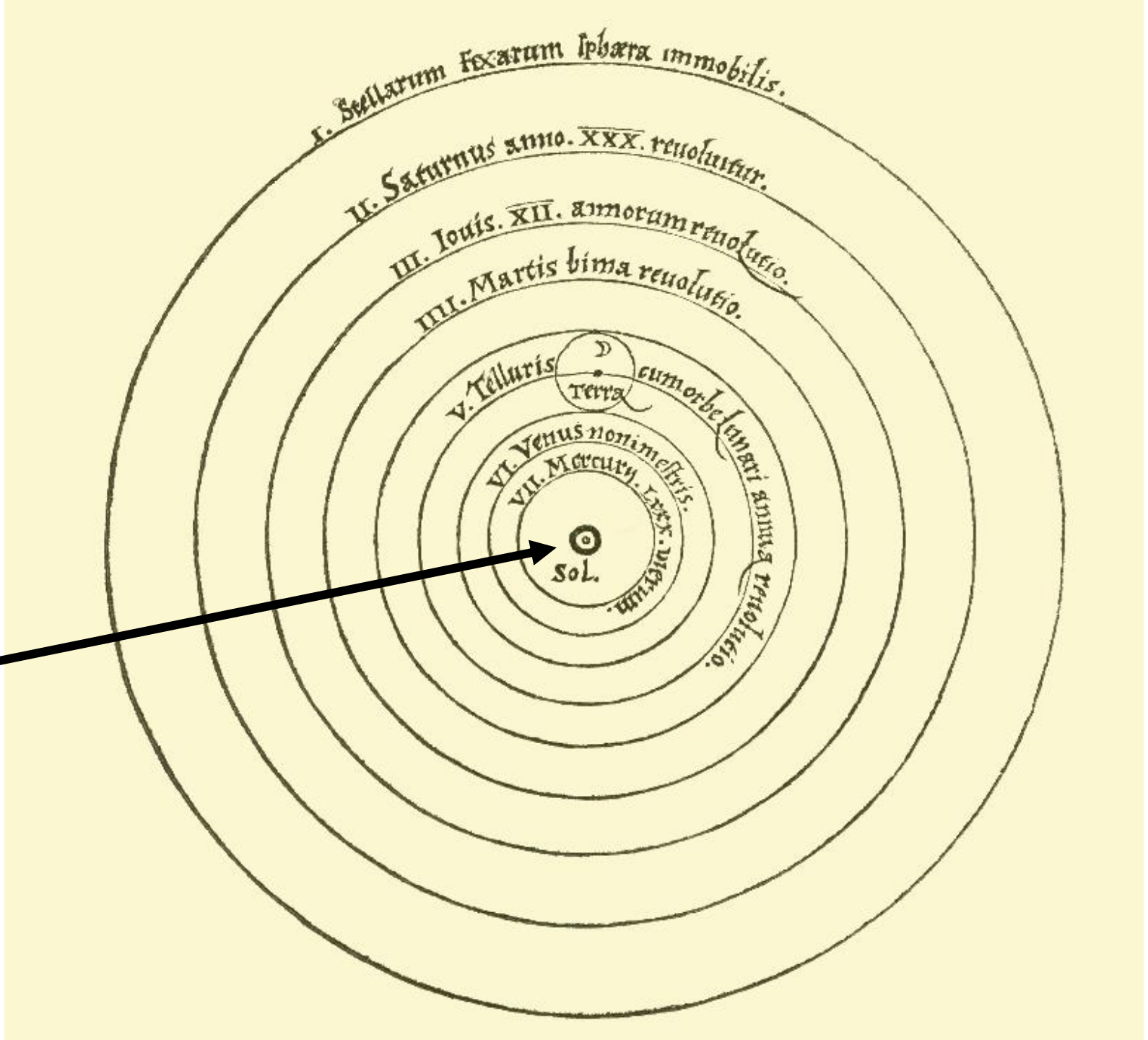
And it does! But not perfectly.

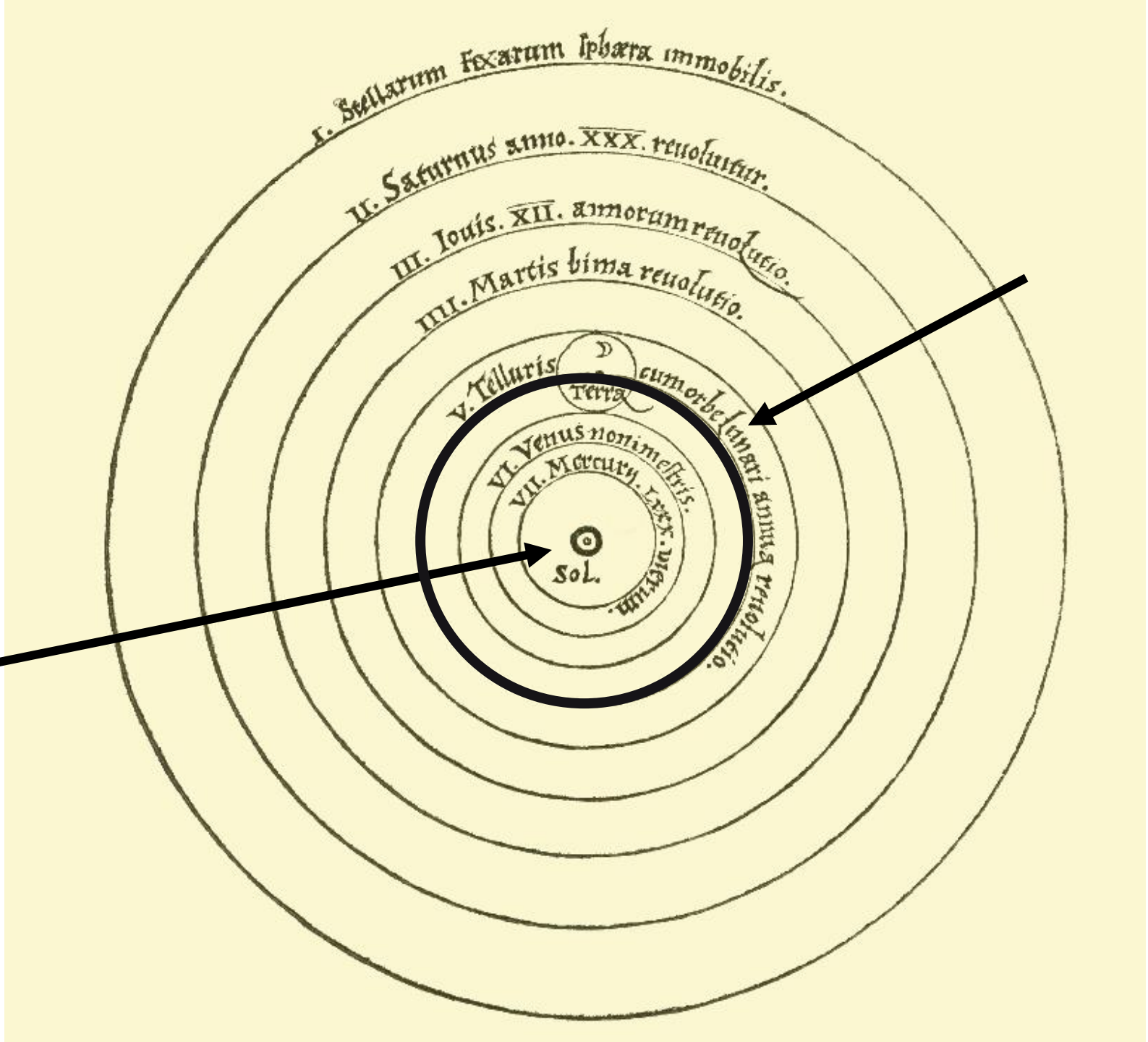
In the ensuing centuries, more observation led to more data collection which led to more inaccuracies with the Ptolemaic model.

New models for the solar system were proposed.


Copernicus

Sun at the center

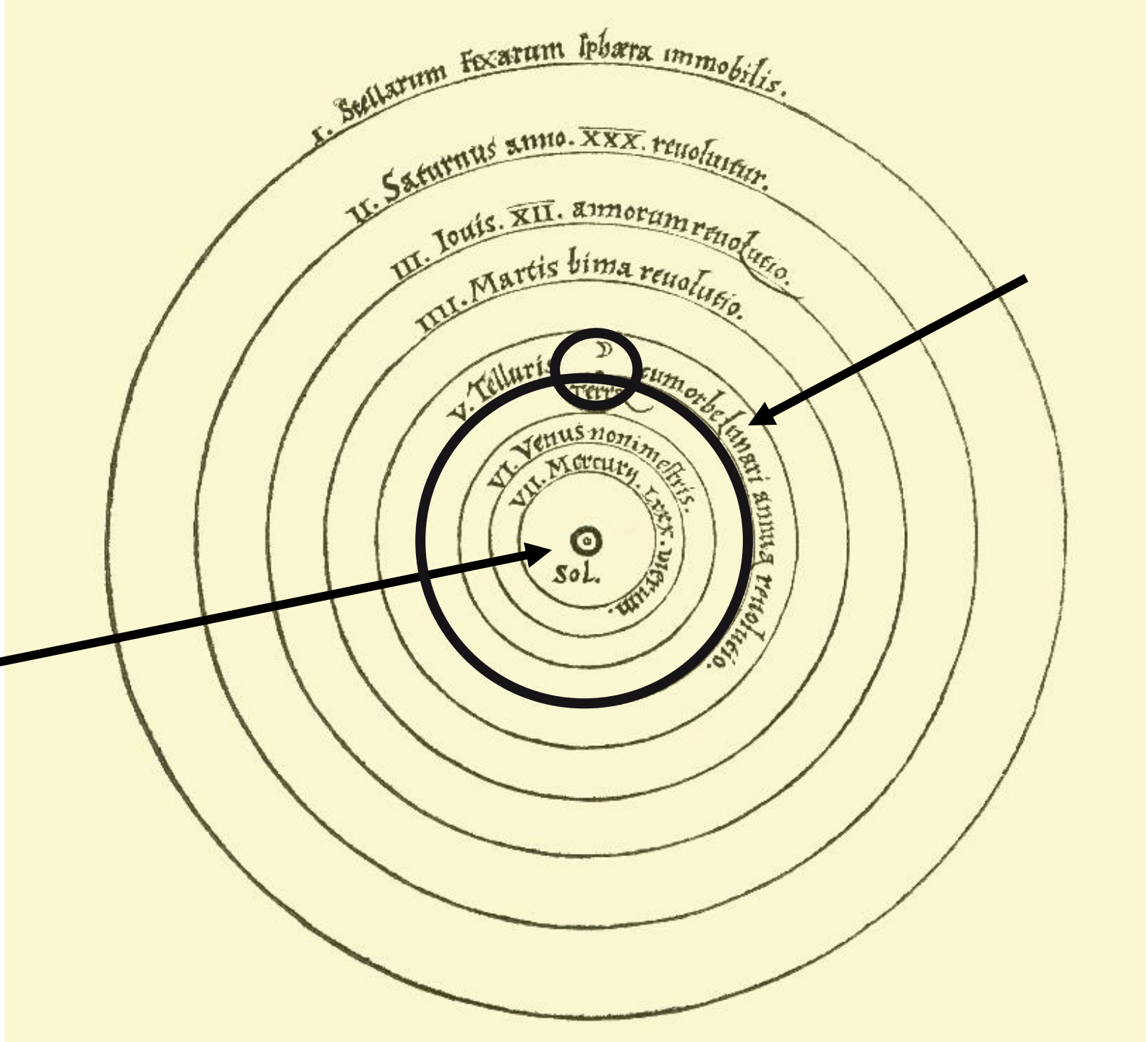




Sun at the center

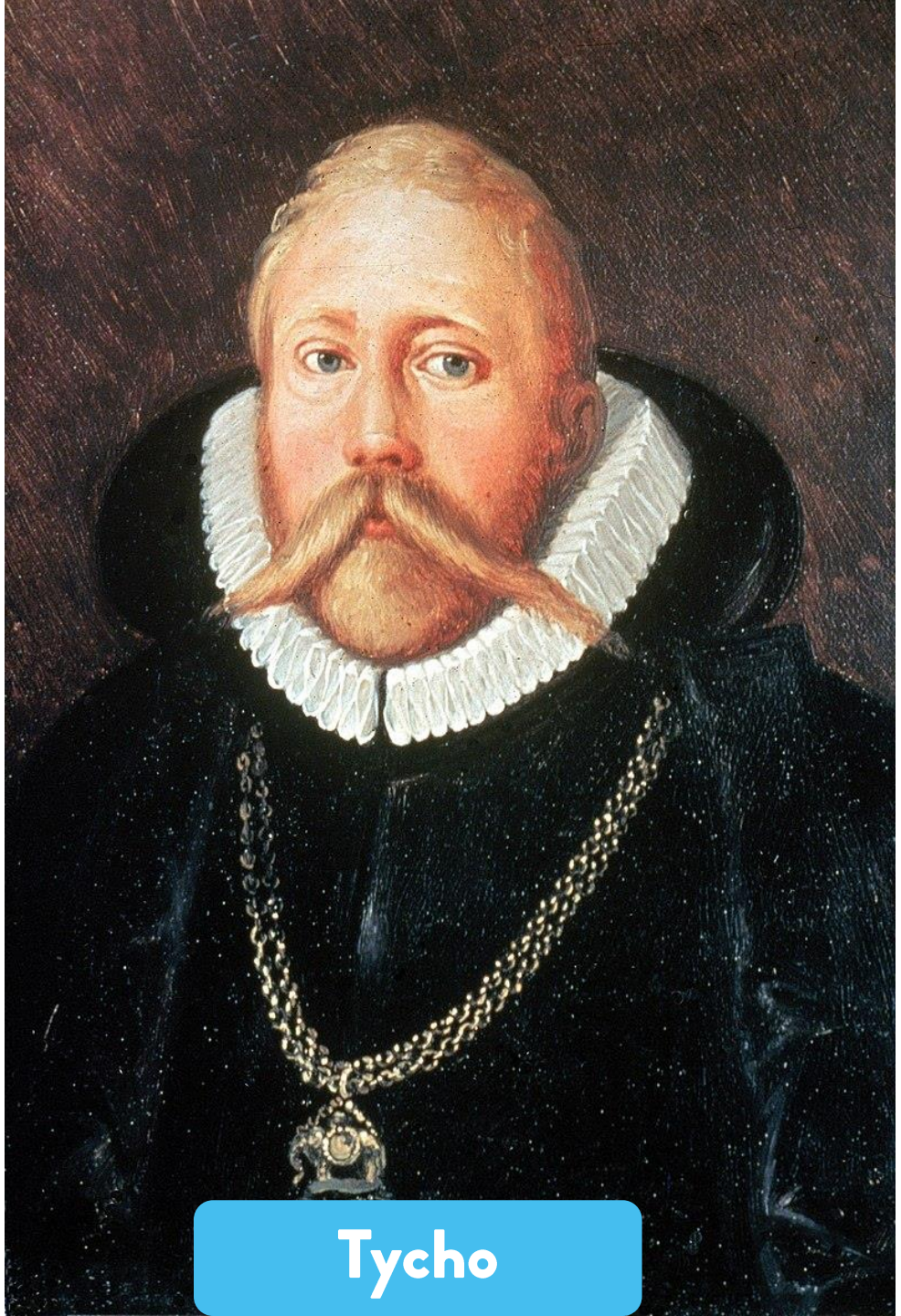
The planets, including Earth, orbit the sun in circles

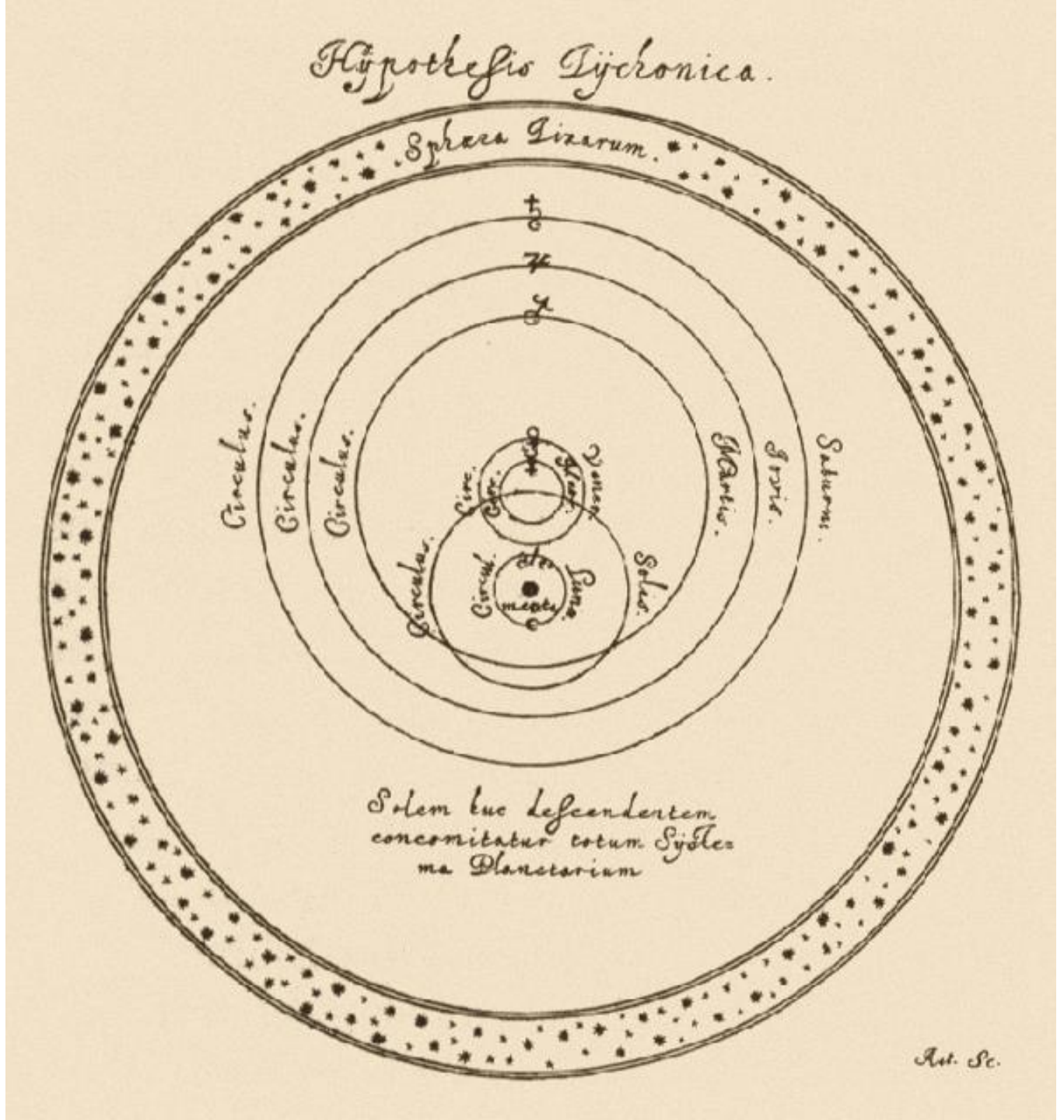
Sun at the center



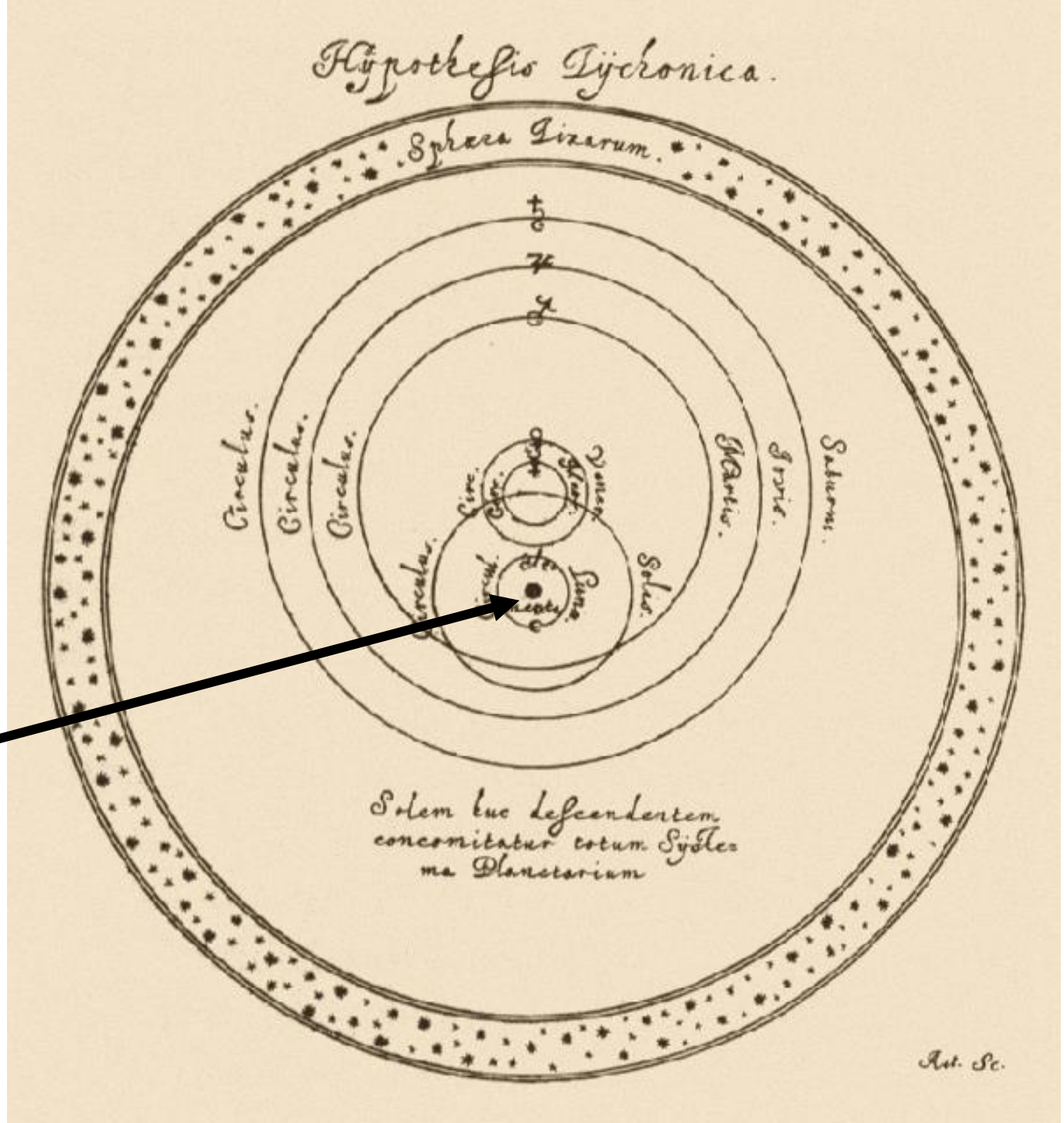
The planets, including Earth, orbit the sun in circles

Except for the moon, which orbits Earth.


Tycho

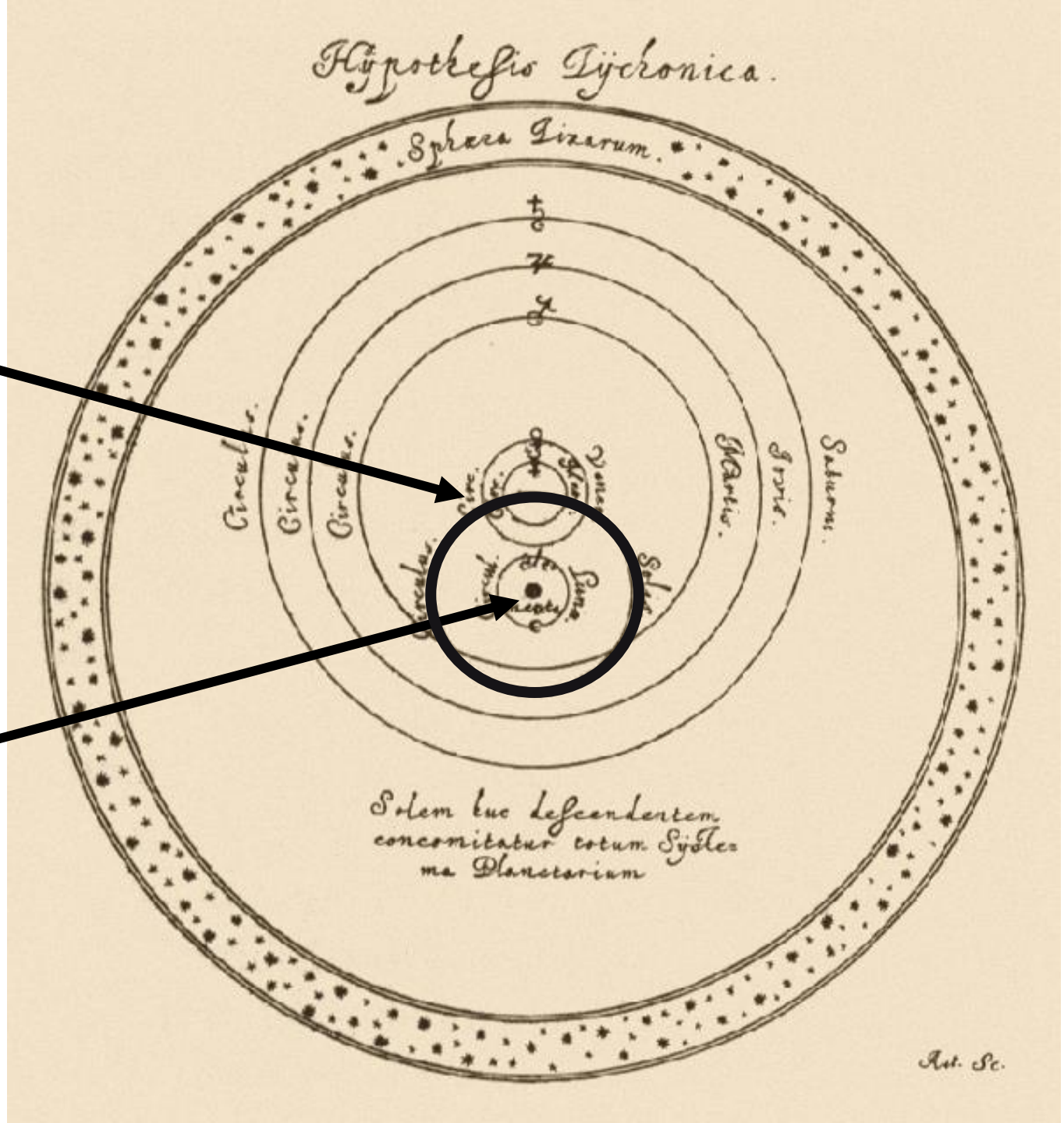

Hypothesis Tychonica.
Sphæra Fixarum.
Circular.
Circular.
Circular.
Circular.
Circul.
Circ.
Circ.
Circ.
Venus.
Mercur.
Sol.
Circul.
Luna.
Mercur.
Solar.
Martis.
Jovis.
Saturni.
Solem hic descendentem
concomitatur totum Systē-
ma Planetarium
Aut. Sc.

Earth at the center



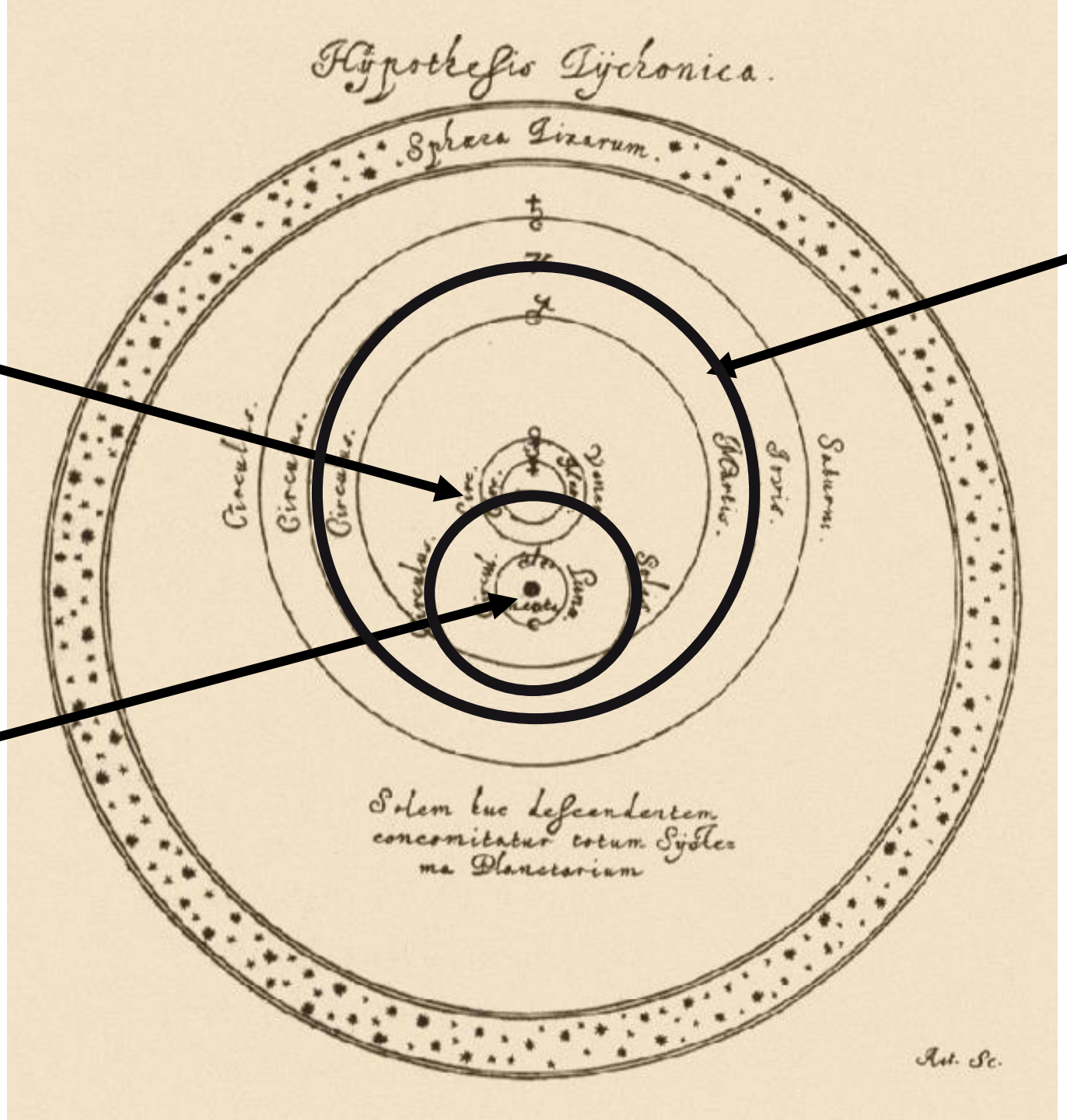
Sun and moon orbit the Earth

Earth at the center



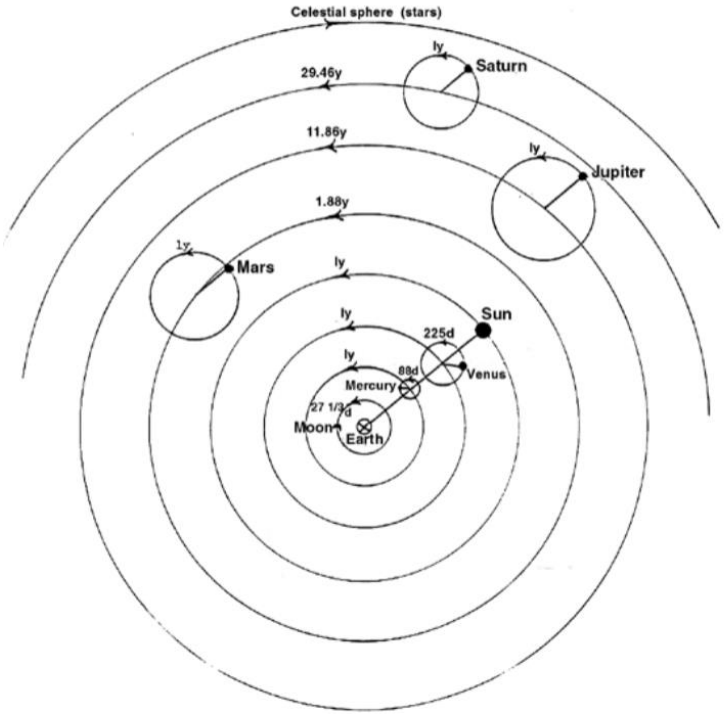
Sun and moon orbit the Earth

Earth at the center

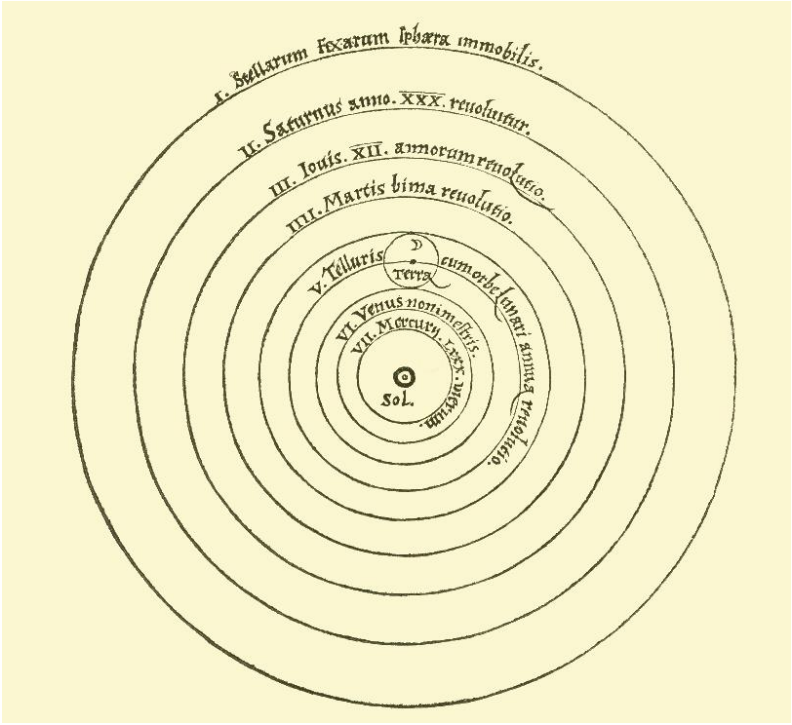


The other planets orbit the Sun

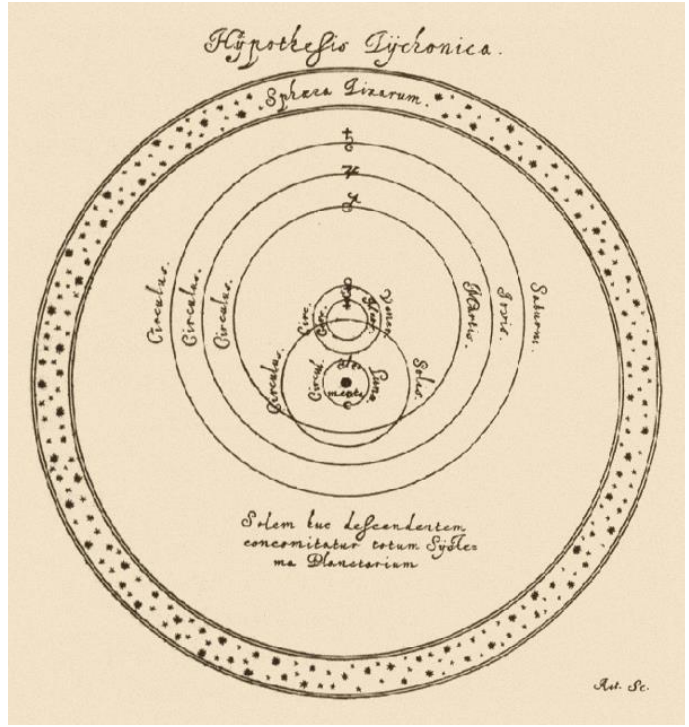
Why does Mars move backwards in the nighttime sky?



Ptolemy



Copernicus

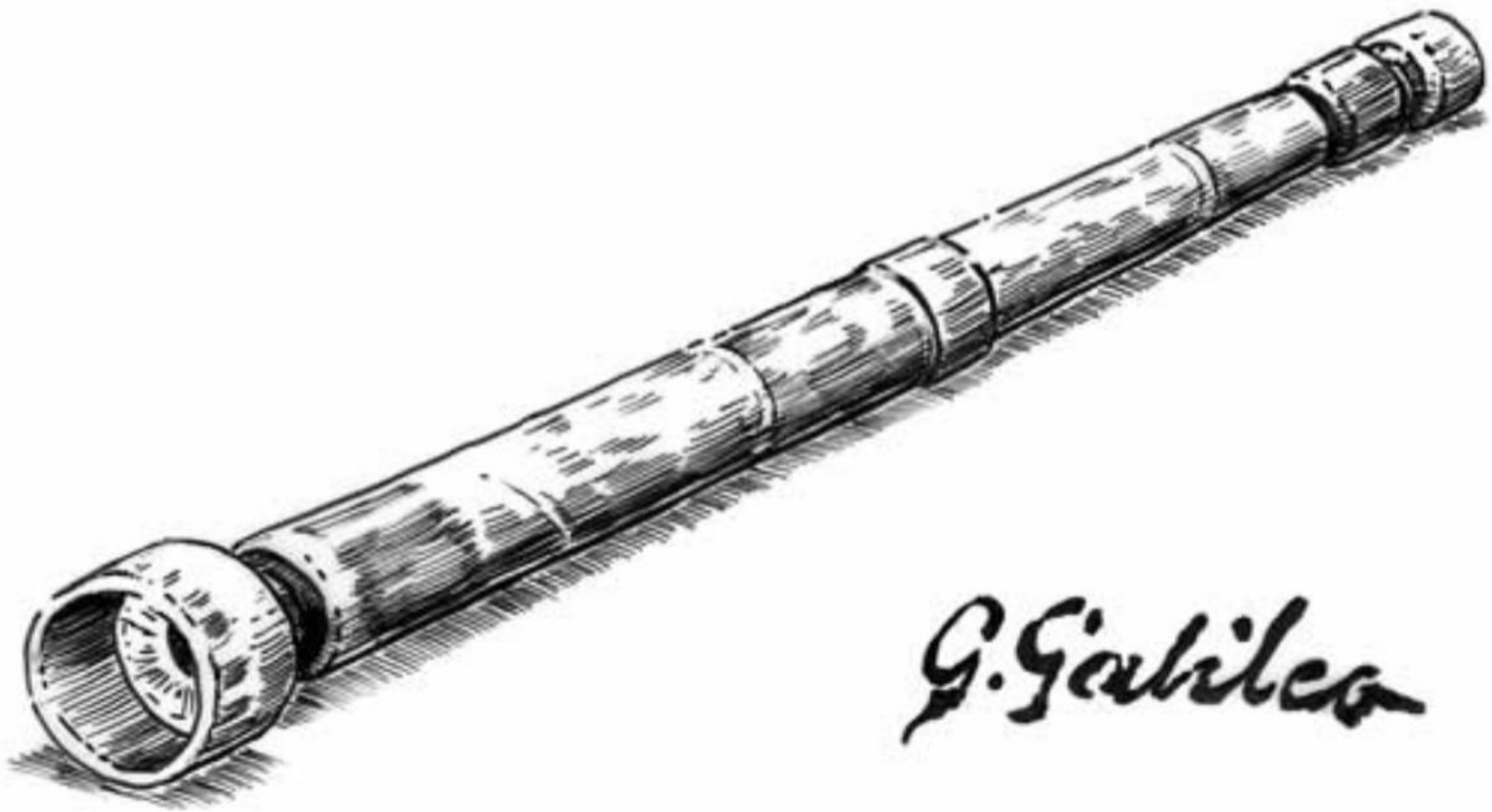


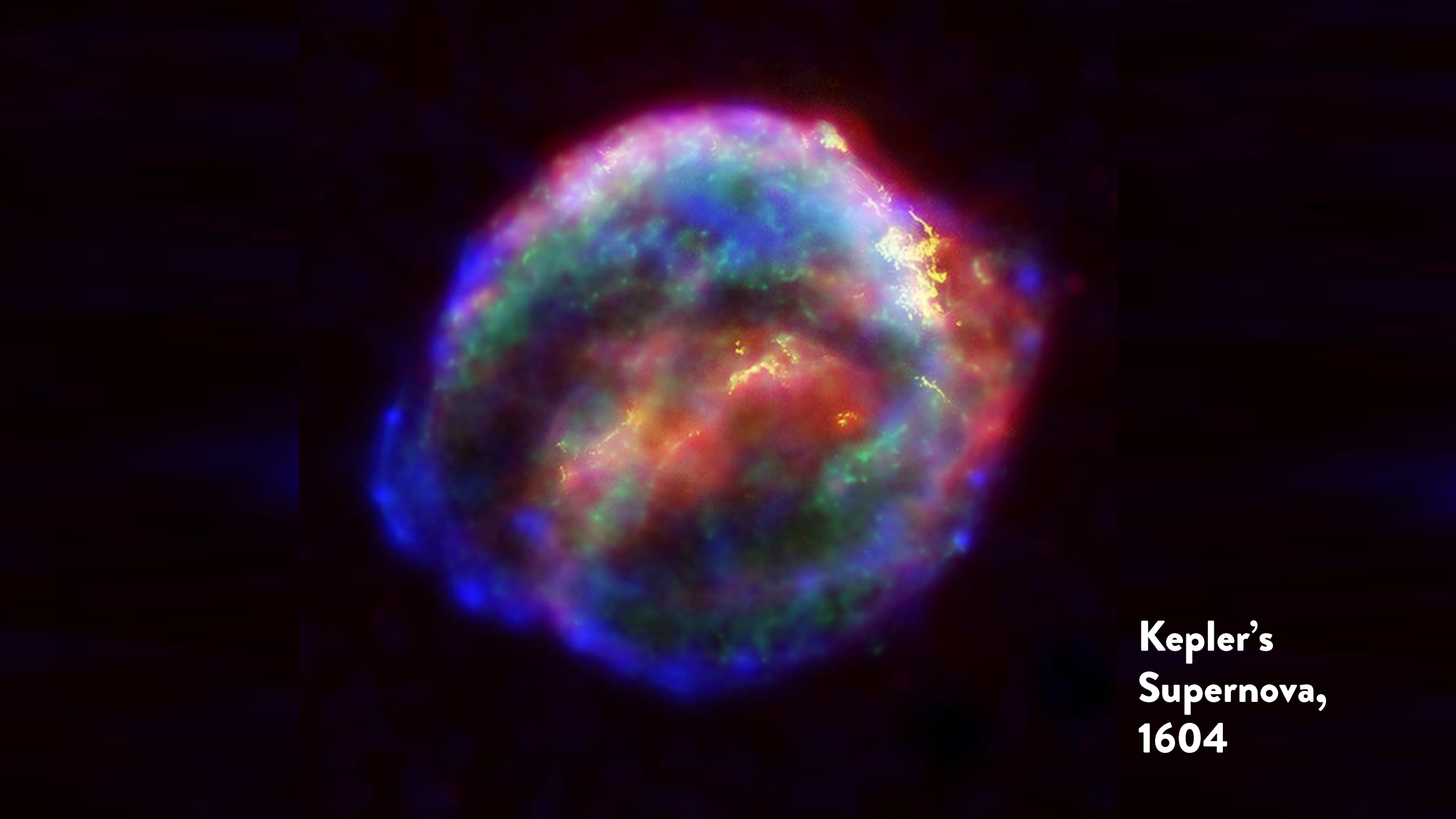
Tycho

How do we determine which model is ‘correct’?

We think about implications. We gather data.

New methods of data collection challenged implications of Ptolemy's model.

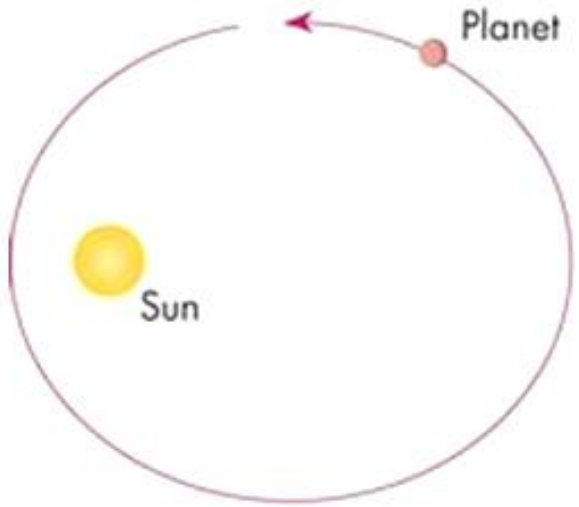

G. Galilea



Kepler's Supernova, 1604

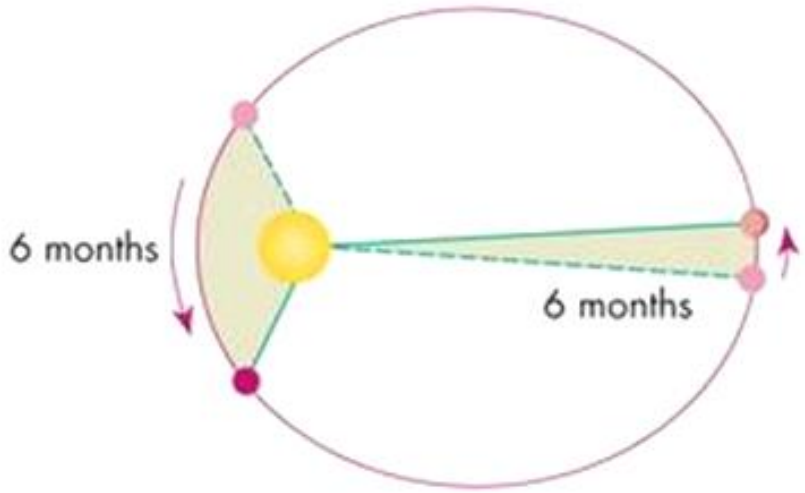

Kepler

Kepler's Laws of Planetary Motion



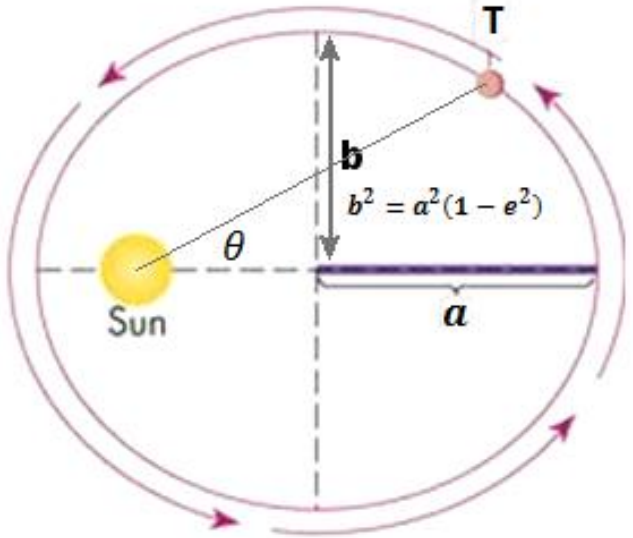
(1)

The orbits are ellipses



(2)

Equal areas in equal time



(3)

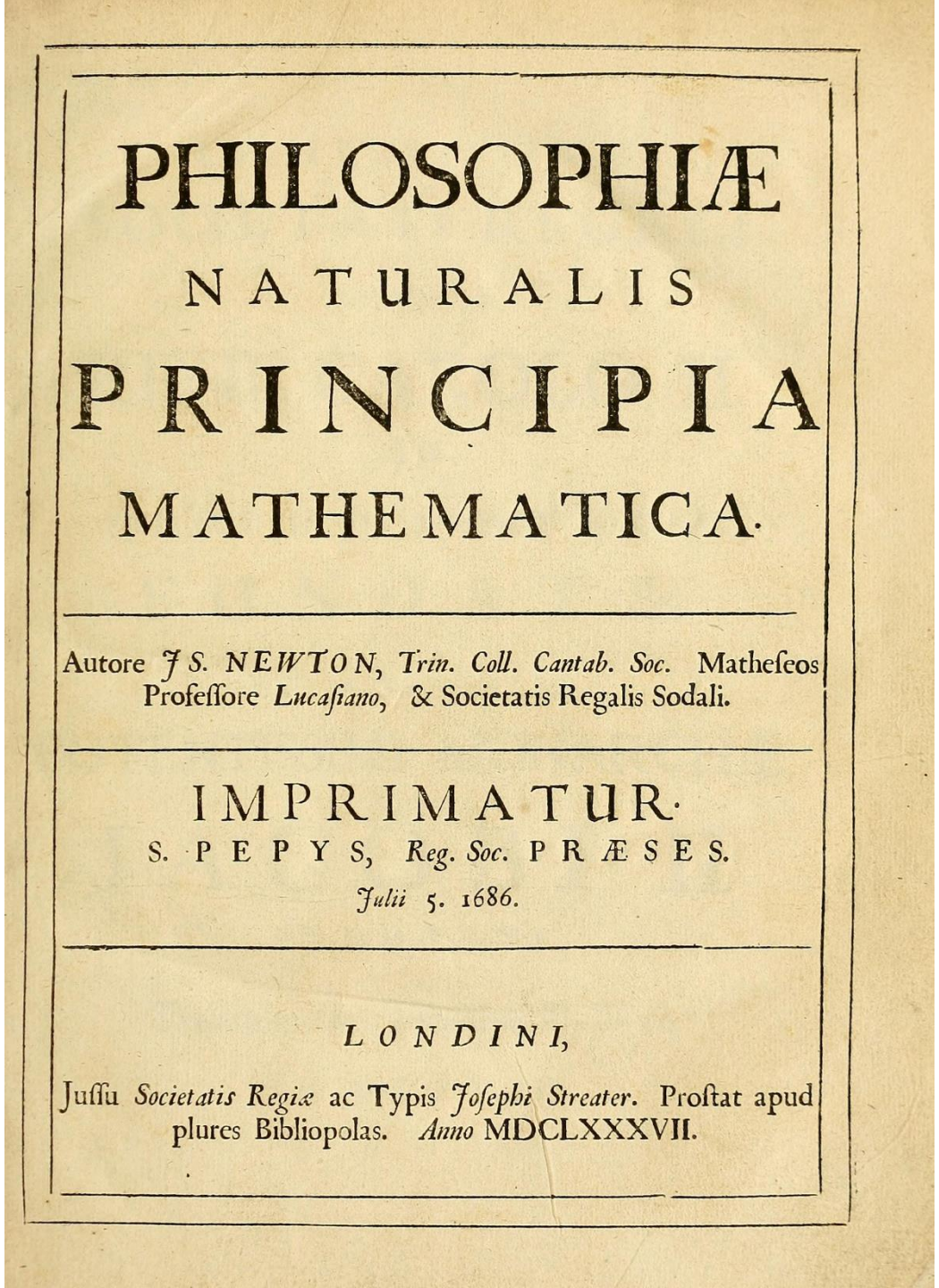
$T^2 \propto a^3$

T = time to complete orbit

a = semi-major axis

Then some guy named **Isaac Newton** came along to explain why Kepler's model was evidence of a more general theory...

Newton's Principia in 1687



PHILOSOPHIÆ
NATURALIS
PRINCIPIA
MATHEMATICA.

Autore *J S. NEWTON, Trin. Coll. Cantab. Soc.* Matheſeos Profeſſore *Lucaſiano*, & Societatis Regalis Sodali.

IMPRIMATUR.
S. PEPYS, *Reg. Soc.* PRÆSES.
Julii 5. 1686.

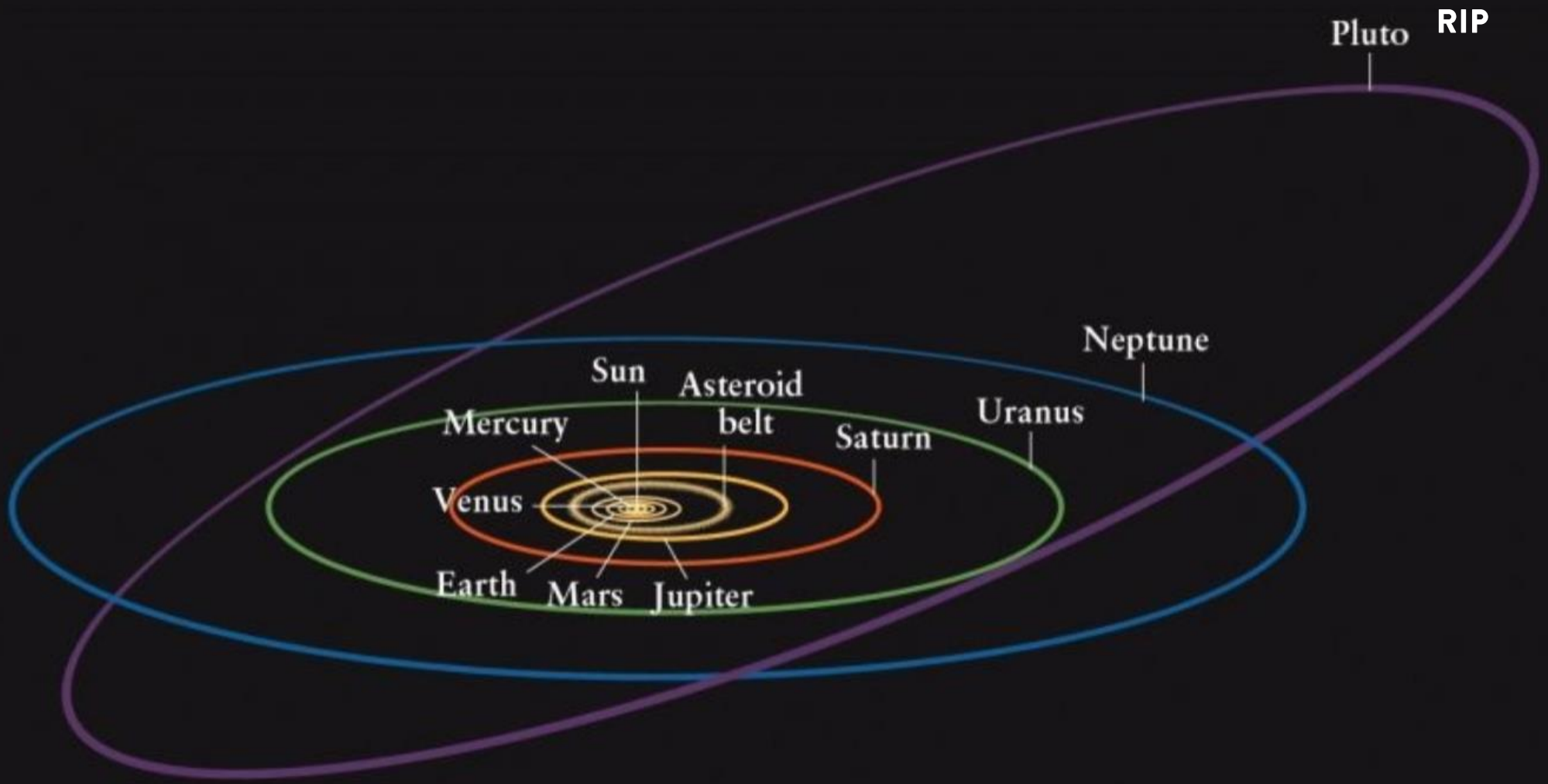
LONDINI,
Juſſu *Societatis Regiæ* ac Typis *Joſephi Streater.* Proſtat apud plures Bibliopolas. *Anno* MDCLXXXVII.

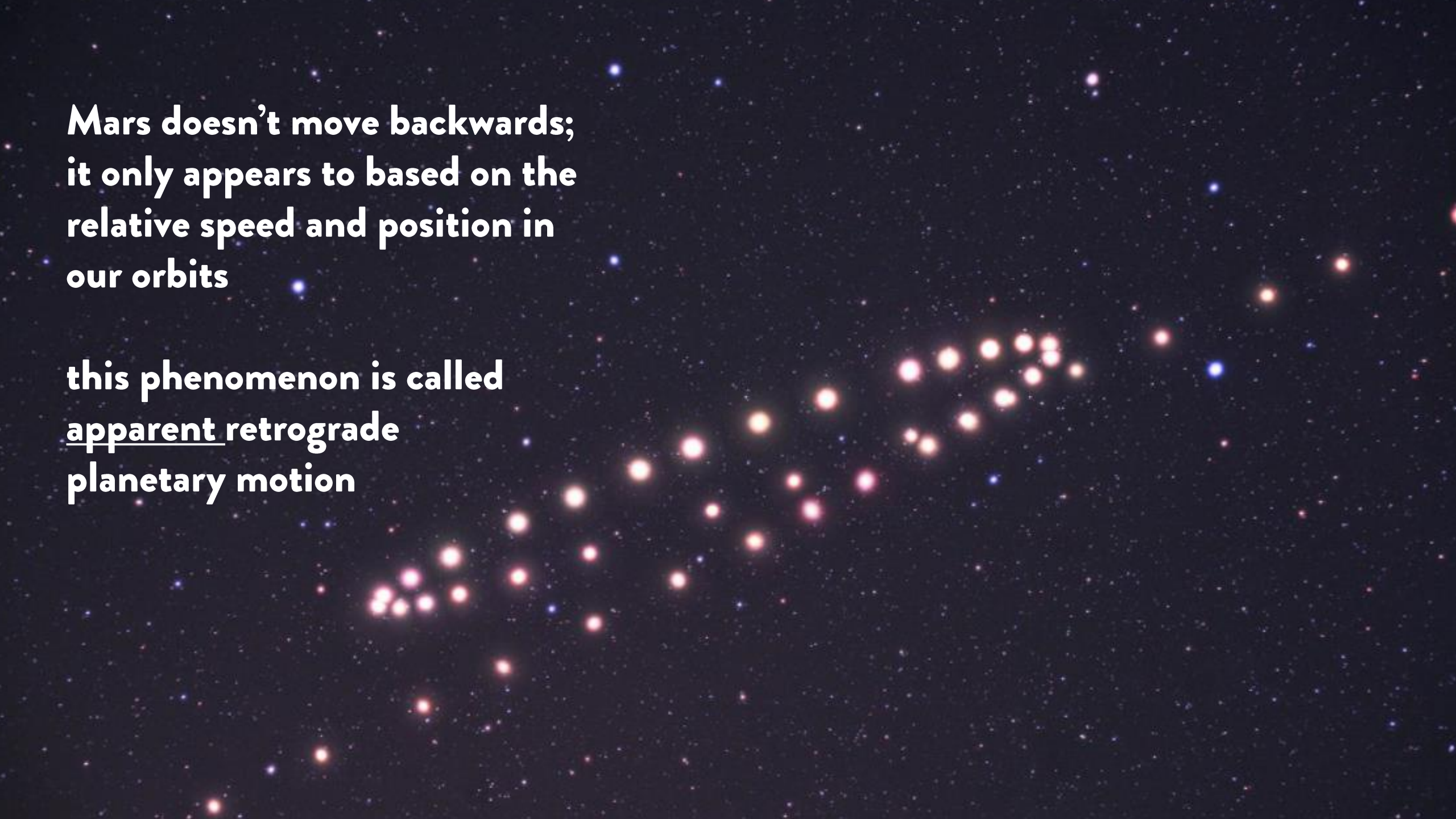
Newton's Laws of Motion

$$\sum \mathbf{F} = 0 \Rightarrow \frac{dv}{dt} = 0$$

$$F = ma$$

$$F_A = -F_B$$


Pluto
RIP
Neptune
Sun
Asteroid
belt
Mercury
Uranus
Saturn
Venus
Earth
Mars
Jupiter



Mars doesn't move backwards; it only appears to based on the relative speed and position in our orbits

this phenomenon is called apparent retrograde planetary motion

A model of the model building process:

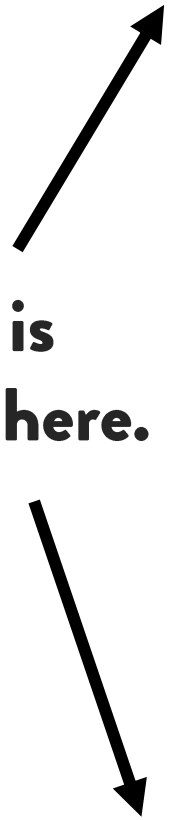
1) **We observe**. We notice a pattern or result that has occurred in the world.

2) **We speculate**. We develop explanation for the process that could have produced our observation.

3) **We develop implications**. We ask, if our speculation is correct, what else should we expect to observe?

4) **We test**. We look to see whether the other implications of our model are supported in the data.

A model of the model building process:



Data is used here.

1) **We observe**. We notice a pattern or result that has occurred in the world.

2) **We speculate**. We develop explanation for the process that could have produced our observation.

3) **We develop implications**. We ask, if our speculation is correct, what else should we expect to observe?

4) **We test**. We look to see whether the other implications of our model are supported in the data.

A model of the model building process:

1) **We observe**. We notice a pattern or result that has occurred in the world.

2) **We speculate**. We develop explanation for the process that could have produced our observation.

3) **We develop implications**. We ask, if our speculation is correct, what else should we expect to observe?

4) **We test**. We look to see whether the other implications of our model are supported in the data.

Data is used here.

The science, how we learn from data, is here.

Data does not, **by itself**, lead to learning.

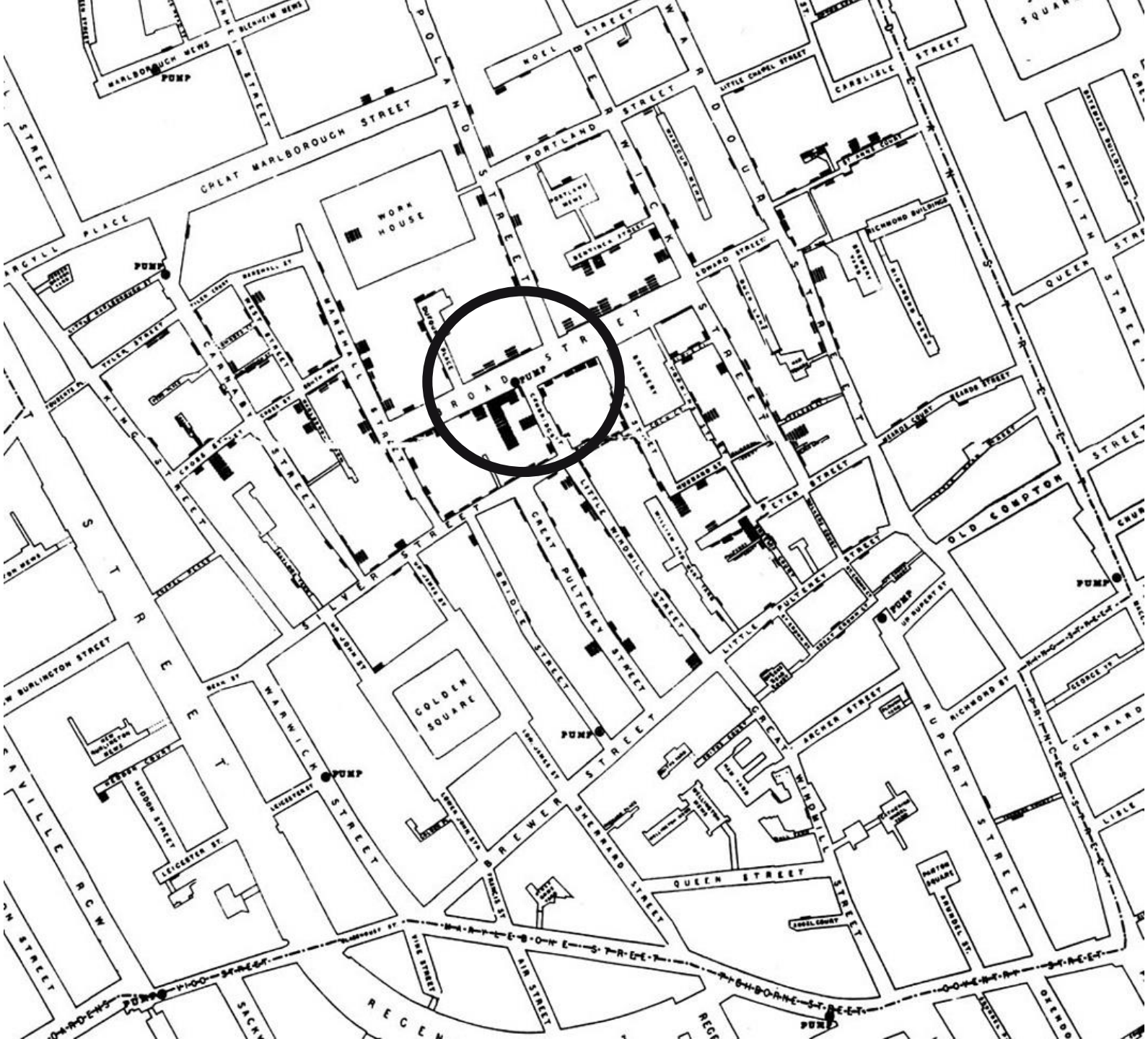
Data is the means of **testing the implications of our models**.

Example:

John Snow and the London Cholera Outbreak of 1854

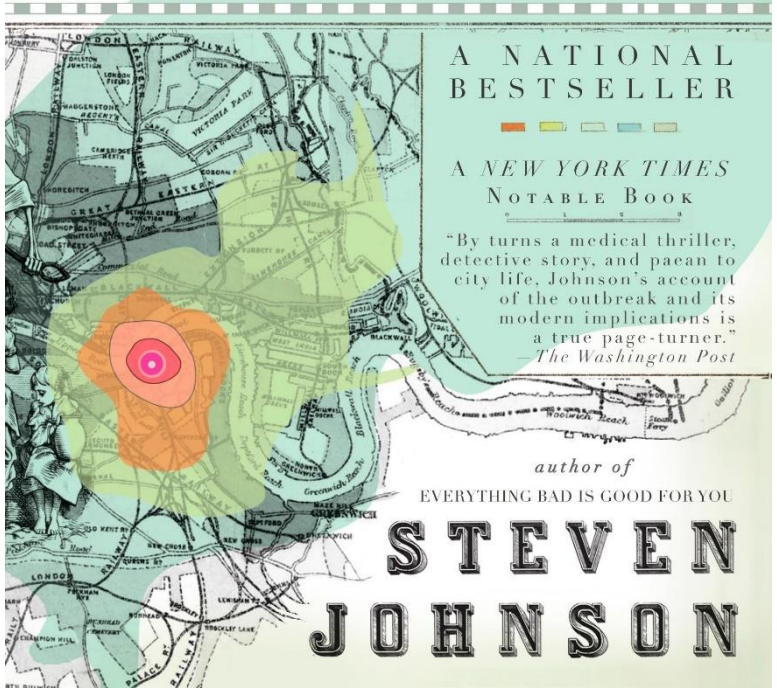
John Snow's map of cholera cases linked to the outbreak

John Snow's map of cholera cases linked to the outbreak



THE GHOST MAP

The Story of London's Most Terrifying Epidemic— and How It Changed Science, Cities, and the Modern World



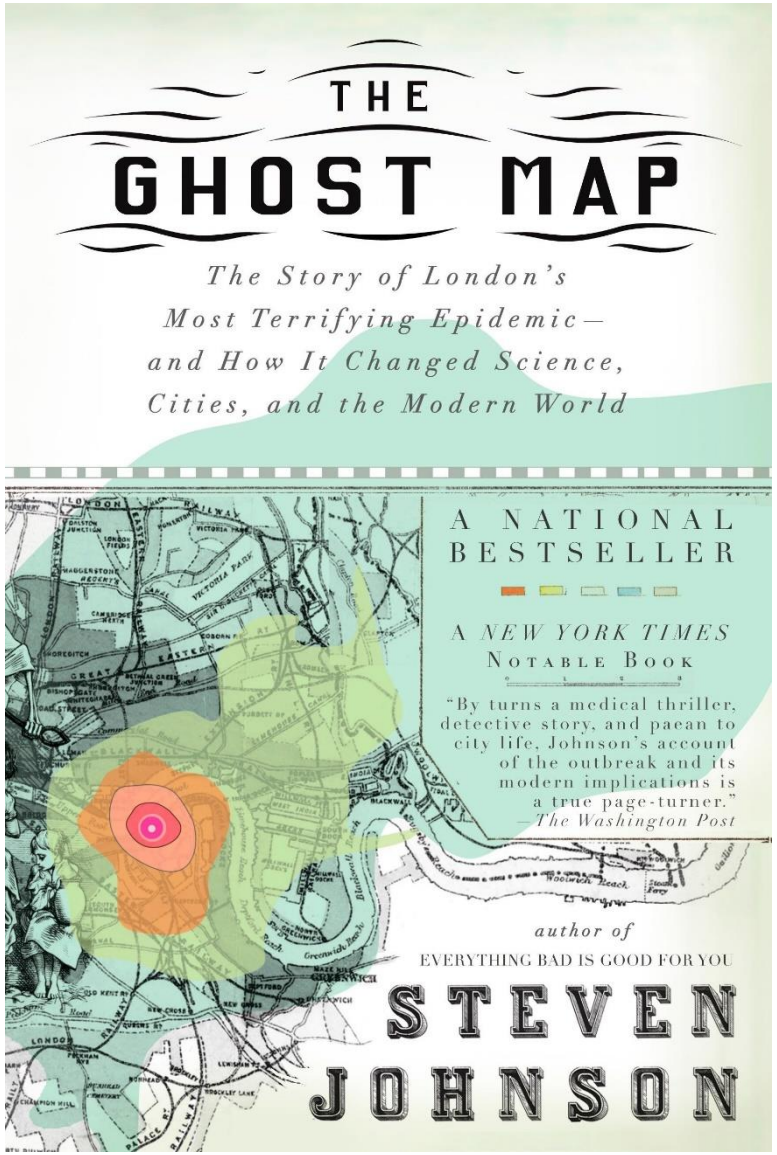
A NATIONAL BESTSELLER

A *NEW YORK TIMES* NOTABLE BOOK

"By turns a medical thriller, detective story, and paean to city life, Johnson's account of the outbreak and its modern implications is a true page-turner."
—*The Washington Post*

author of
EVERYTHING BAD IS GOOD FOR YOU

STEVEN JOHNSON



The dominant theory of disease at the time was **miasma** – disease was spread through bad air.

John Snow spent years accumulating evidence which contradicted this theory.

The Broad Street outbreak of 1854 offered the data to support his **speculative model of how cholera spread: water**.

By itself, data does not tell us anything about the **process by which it was produced**.

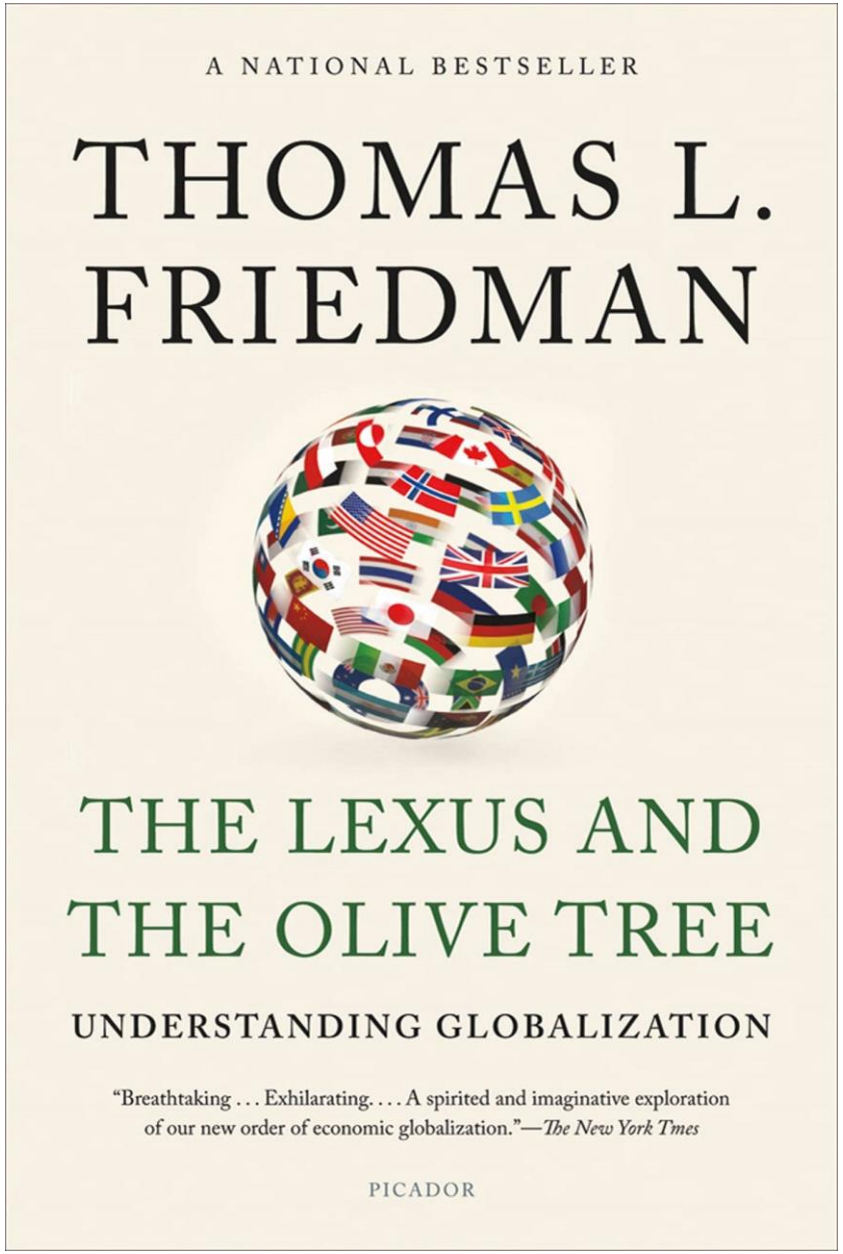
It is incredibly easy to observe data and reach a conclusion about the process that is **completely wrong**.

Example:

The Golden Arches Theory of War

Thomas Friedman, writing in 1996, observed:

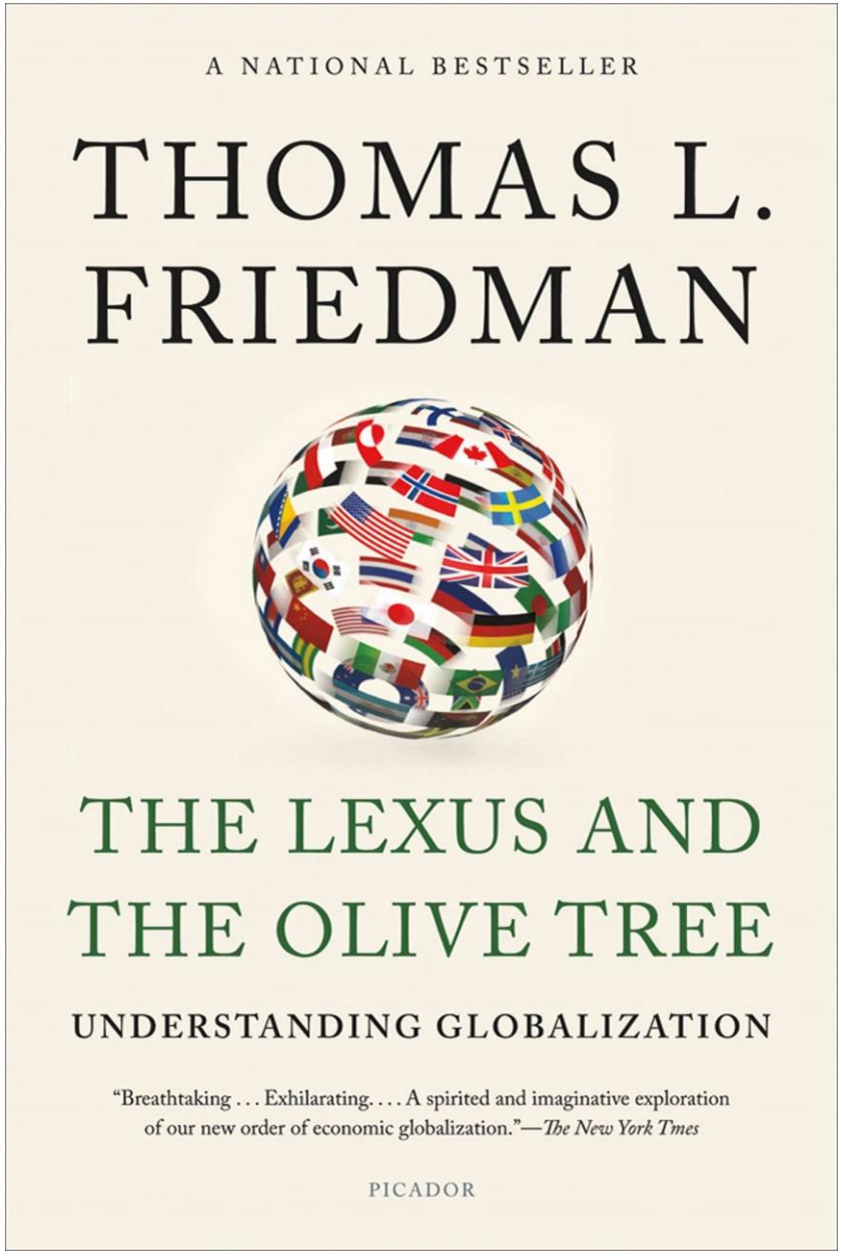
No two countries that both had McDonald's had fought a war against each other since each got its McDonald's.



Thomas Friedman, writing in 1996, observed:

No two countries that both had McDonald's had fought a war against each other since each got its McDonald's.

Our **data-driven solution** to preventing war:
Put a McDonalds in every country!
War will never happen again!



— Opinion

Ukraine war spells the end of the Golden Arches peace theory

With hundreds of McDonald's outlets in both Russia and Ukraine, the war in Europe may mark the final break in theories about the process of perpetual globalisation.

John Roskam *Columnist*

makebigmacsnotwar Follow

540 posts **276** followers **0** following

A humble McDonald's evangelist

Chasing the M around the globe to prove Friedman's "Golden Arches Theory of Conflict Prevention"

🍟🍔☮🌎

#AlguienDijoMcDonalds? #SureIsGoodToHaveAround

makebigmacsnotwar

Follow

540 posts **276** followers **0** following

A humble McDonald's evangelist

Chasing the M around the globe to prove Friedman's "Golden Arches Theory of Conflict Prevention"

🍟🍔☮🌎

#AlguienDijoMcDonalds? #SureIsGoodToHaveAround

This Account is Private

Already follow makebigmacsnotwar?

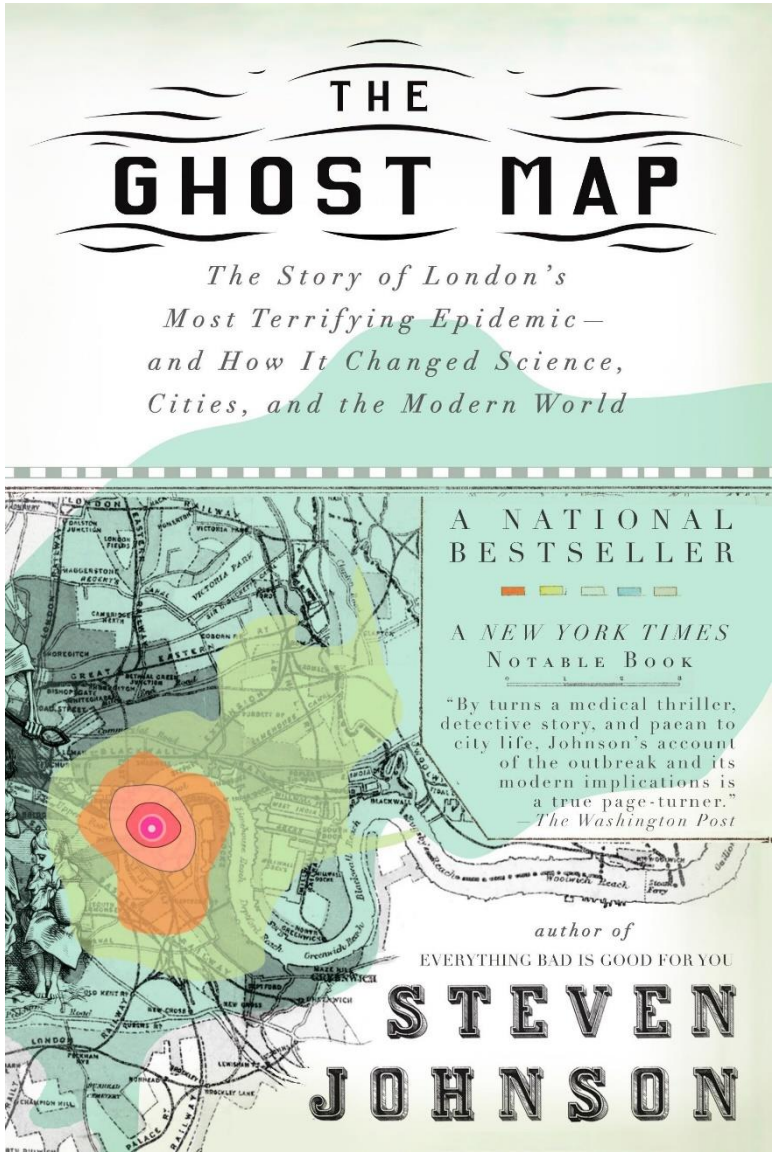
Log in

to see their photos and videos.

Being wrong is okay – it is essential to the model building process!

In order to learn from data, we must **put our speculations to the test** and **be willing to be wrong.**

It is by figuring out **when our models are wrong that we learn.**



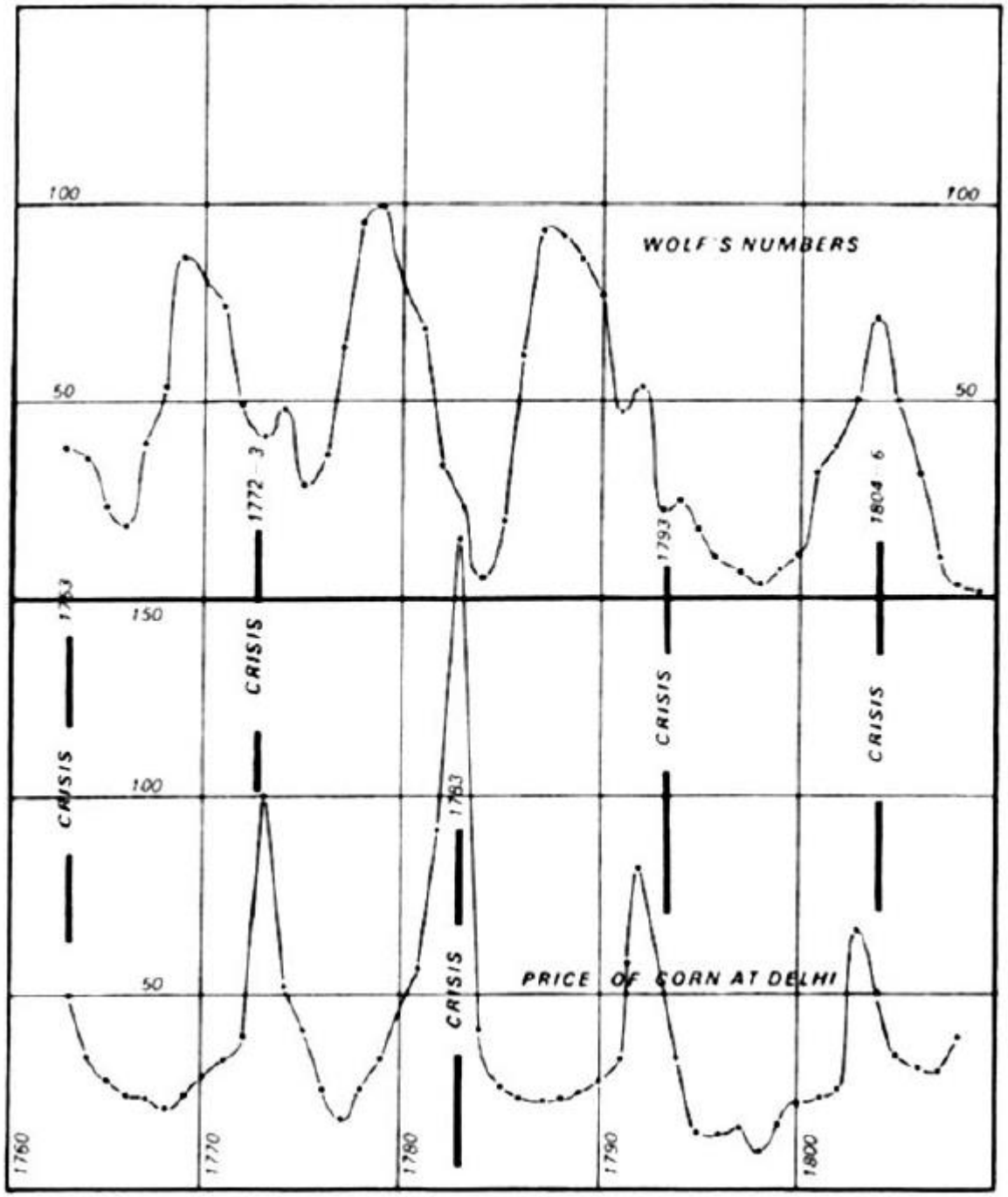
"How could so many intelligent people be so grievously wrong for such an extended period of time?

How could they ignore so much overwhelming evidence that contradicted their most basic theories?"

Discovering that our predictions are wrong can be **good**...

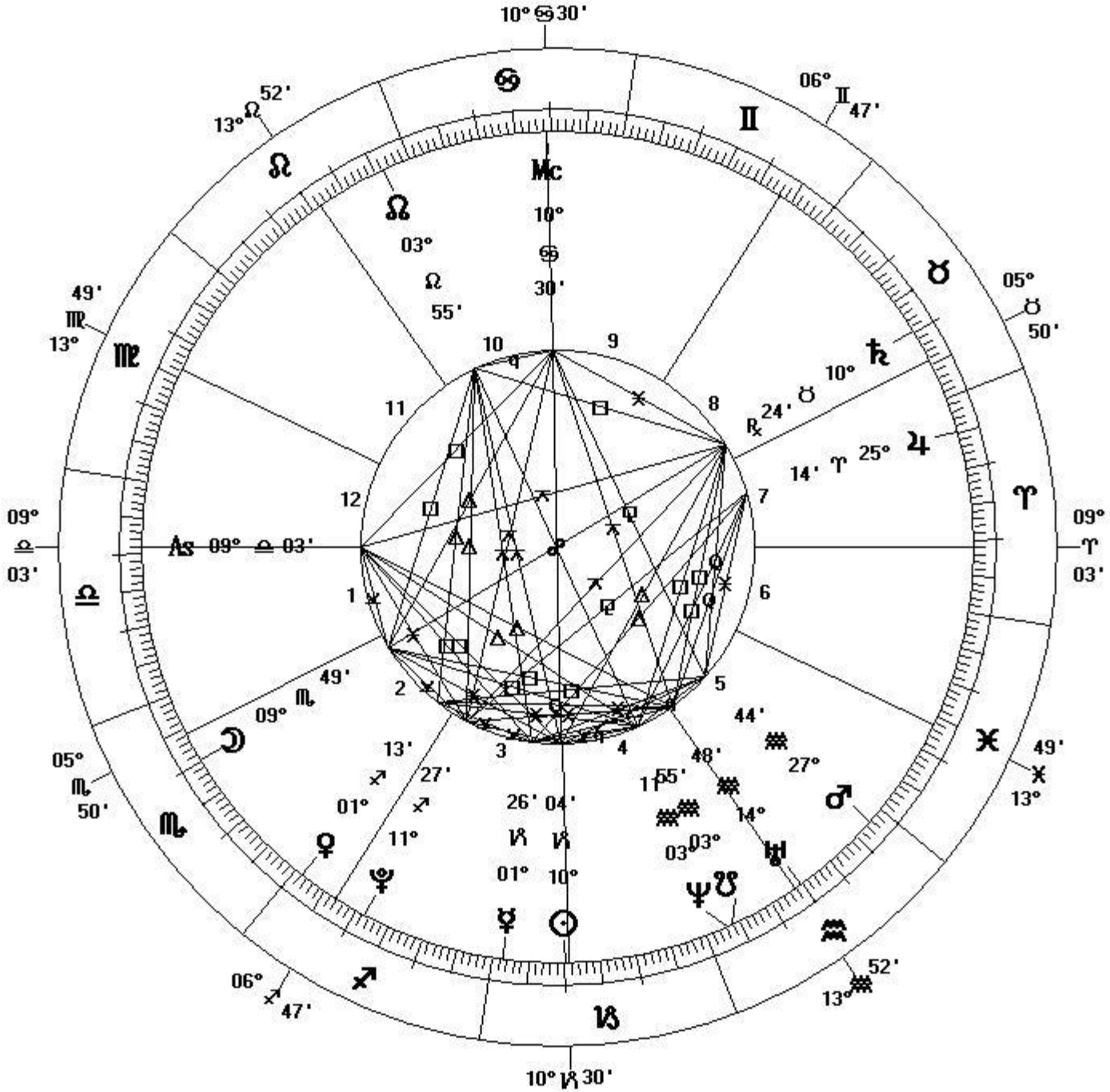
...unless we ignore the data we don't like and **select only the data that shows we're right**.

William Jevons, an economist, was adamant that **sunspots caused economic cycles on earth**.



Astrology was at one point a respected scholarly inquiry: do patterns in the stars determine behavior on earth?

In spite of the lack of evidence, **astrologers refused to say no** (but they did collect a lot of good data).



Ronald Fisher, the father of statistical science and experimental methods, was **adamant that smoking had no effect on rates of cancer**.

LETTERS TO THE EDITORS

The Editors do not hold themselves responsible for opinions expressed by their correspondents. No notice is taken of anonymous communications.

Cancer and Smoking

THE curious associations with lung cancer found in relation to smoking habits do not, in the minds of some of us, lend themselves easily to the simple conclusion that the products of combustion reaching the surface of the bronchus induce, though after a long interval, the development of a cancer. If, for example, it were possible to infer that smoking cigarettes is a cause of this disease, it would equally be possible to infer on exactly similar grounds that inhaling cigarette smoke was a practice of considerable prophylactic value in preventing the disease, for the practice of inhaling is rarer among patients with cancer of the lung than with others.

Such results suggest that an error has been made, of an old kind, in arguing from correlation to causation, and that the possibility should be explored that the different smoking classes, non-smokers, cigarette smokers, cigar smokers, pipe smokers, etc., have adopted their habits partly by reason of their personal temperaments and dispositions, and are not lightly to be assumed to be equivalent in their genotypic composition. Such differences in genetic make-up between these classes would naturally be associated with differences of disease incidence without the disease being causally connected with smoking. It would then seem not so paradoxical that the stronger fumes of pipes or cigars should be so much less associated with cancer than those of cigarettes, or that the practice of drawing cigarette smoke in bulk into the lung should have apparently a protective effect.

Since my letter was written, however, I have received from Dr. Eliot Slater, of the Maudsley Hospital (London, S.E.5), some further data, the greater part of which concern girl twins, and in this way supply a valuable supplement to Verschuer's data, and in which, moreover, a considerable number of pairs were separated at or shortly after birth.

For the resemblance in smoking habit, these female pairs give :

	Alike	Unlike	
Monozygotic	44	9	53
Dizygotic	9	9	18

So far, there is only a clear confirmation of the conclusion from the German data that the monozygotics are much more alike than the dizygotics in their smoking habits. The peculiar value of these data, however, lies in the subdivision of the monozygotic pairs into those separated at birth and those brought up together. These are :

	Alike	Unlike	
Separated	23	4	27
Not separated	21	5	26

Of the nine cases of unlike smoking habit, only four occur among the twenty-seven separated at birth. It would appear that the small proportion unlike among these 53 monozygotic pairs is not to be ascribed to mutual influence.

There is nothing to stop those who greatly desire it from believing that lung cancer is caused by smoking cigarettes. They should also believe that inhaling cigarette smoke is a protection. To believe either is, however, to run the risk of failing to recognize, and therefore failing to prevent, other and more genuine causes.

RONALD A. FISHER

Department of Genetics,
Cambridge.

[1] Fisher, R. A., *Nature*, **182**, 108 (1958).
[2] "Geminus", *New Scientist*, **4**, 440 (1958).

"If, for example, it were possible to infer that smoking cigarettes is a cause of this disease, it would equally be possible to infer on exactly similar grounds that inhaling cigarette smoke was a practice of considerable prophylactic value in preventing the disease"

Why should we do more than visualization and reporting? Why build models?

Our goal is to **learn from data**.

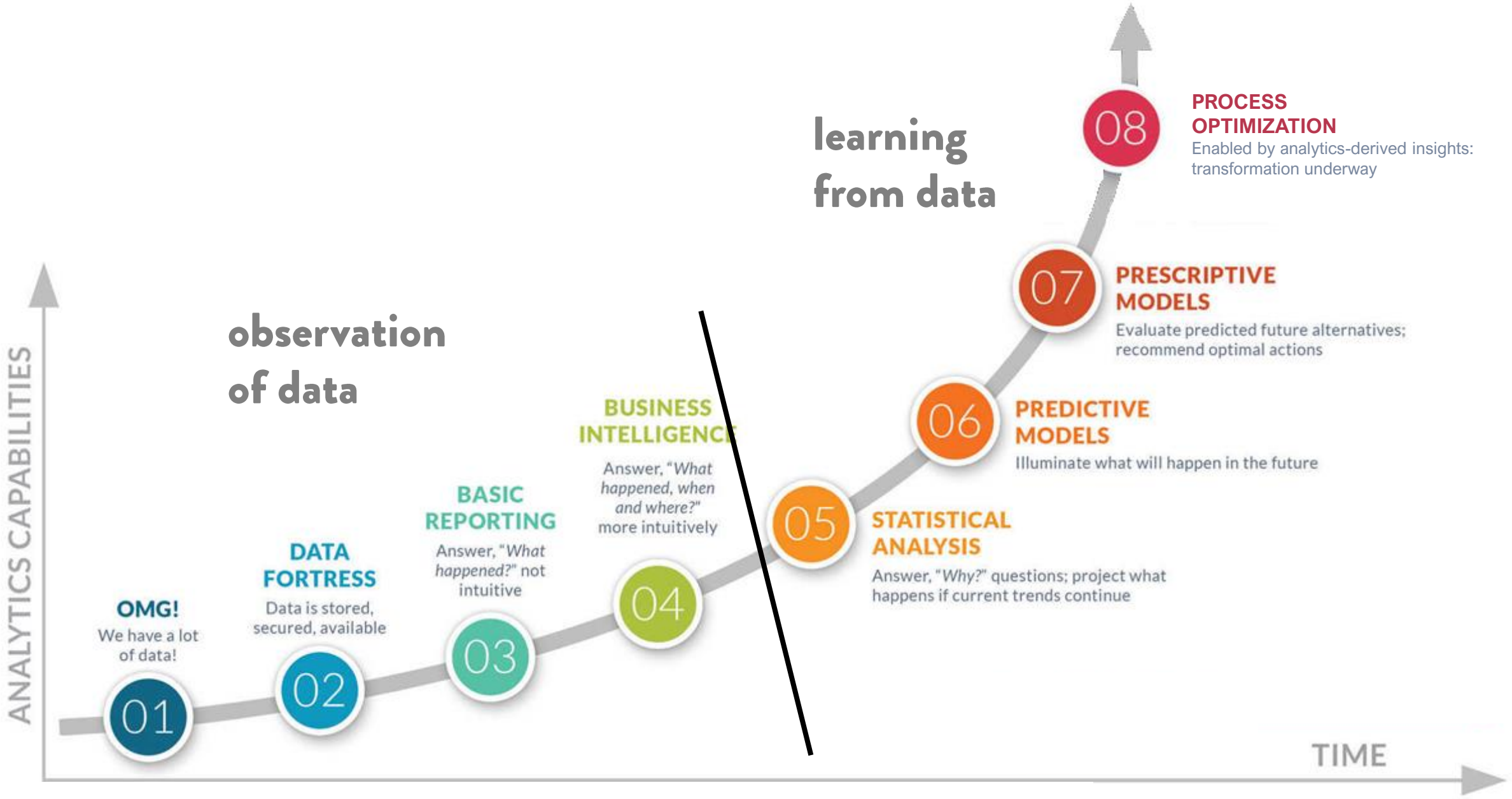
Why should we do more than visualization and reporting? Why build models?

Our goal is to **learn from data**.
Observation is the **start**, it is not the **end**.

Why should we do more than visualization and reporting? Why build models?

Our goal is to **learn from data**.
Observation is the **start**, it is not the **end**.

Speculating, developing implications, and putting them to the test is **how we learn**.


learning
from data
observation
of data
ANALYTICS CAPABILITIES
TIME
01
OMG!
We have a lot of data!
02
DATA FORTRESS
Data is stored, secured, available
03
BASIC REPORTING
Answer, "What happened?" not intuitive
04
BUSINESS INTELLIGENCE
Answer, "What happened, when and where?" more intuitively
05
STATISTICAL ANALYSIS
Answer, "Why?" questions; project what happens if current trends continue
06
PREDICTIVE MODELS
Illuminate what will happen in the future
07
PRESCRIPTIVE MODELS
Evaluate predicted future alternatives; recommend optimal actions
08
PROCESS OPTIMIZATION
Enabled by analytics-derived insights: transformation underway

How do we move from visualization and reporting to building models?

The best way to learn about building predictive models is to do it.

Let's look at an example in some depth.

The best way to learn about building predictive models is to do it.

Let's look at an example in some depth.

~~**I've got a bad feeling about this**~~
This is where the fun begins

2 Building Models of the Universe

Board Game

Those of you have worked with me on a project over the last couple years will have noticed two things about my workstation:

Tons of board games

Biscuit

Star Wars helmets

Massive shelf of board games

Star Wars boardgames

This massive shelf of board games exists for two reasons:

1) I love board games.

This massive shelf of board games exists for two reasons:

1) I love board games.

2) It is part of an ongoing scientific research agenda: why are some games better than others?

This massive shelf of board games exists for two reasons:

1) I love board games.

2) It is part of an ongoing scientific research agenda: why are some games better than others?

FLORIDA STATE UNIVERSITY

AE BUSINESS SOLUTIONS
EST. 1949

Funding Provided By

When a global pandemic hits and you're physically separated from people and can't play games for over a year, you do what anyone would do:

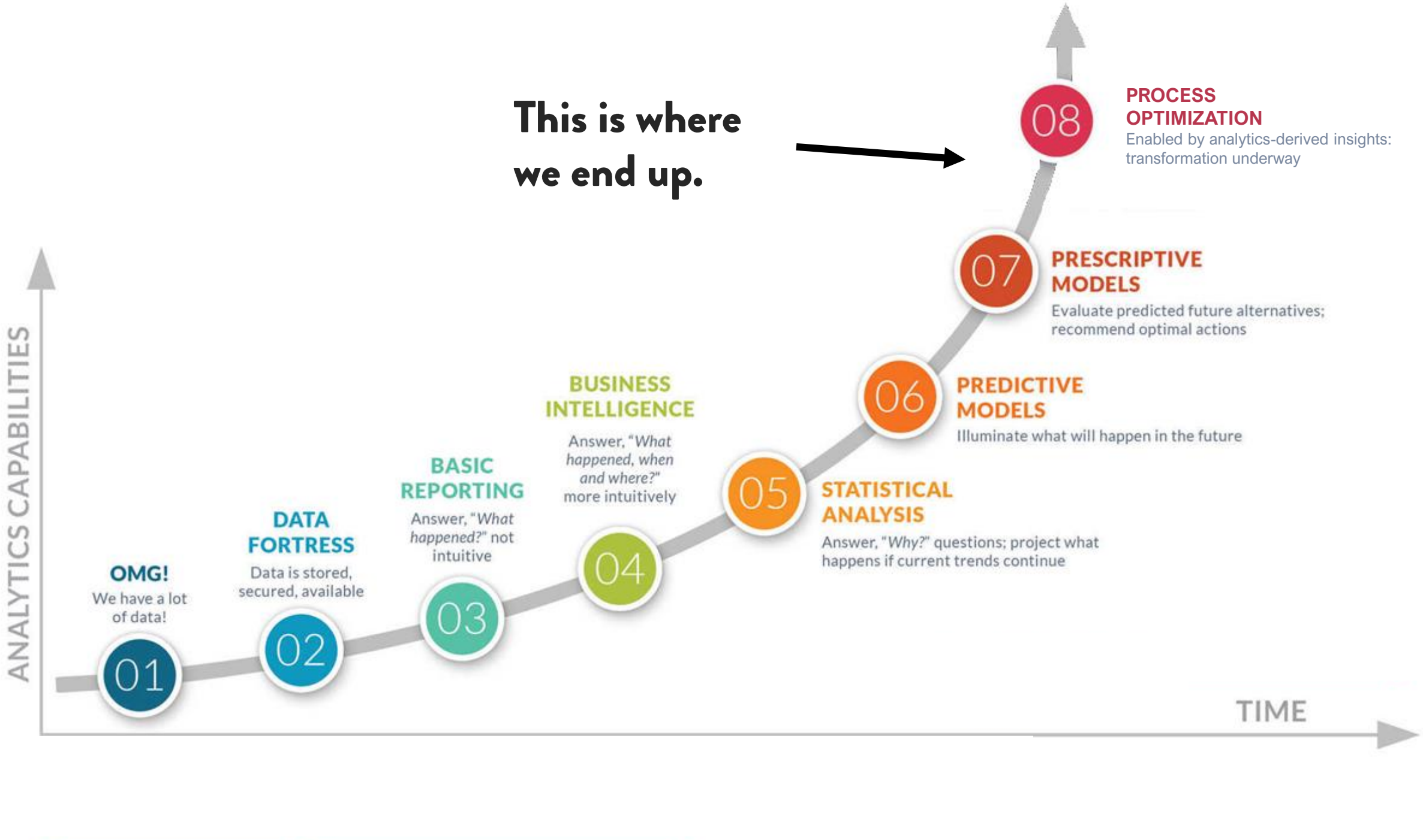
When a global pandemic hits and you're physically separated from people and can't play games for over a year, you do what anyone would do:

Collect data on every boardgame in existence and use the entirety of your scientific training to learn about boardgames so that you can **predict upcoming games** and **recommend them to people** so you have something to look forward to when the pandemic is over.

When a global pandemic hits and you're physically separated from people and can't play games for over a year, you do what anyone would do:

while ridiculous, this is a start to finish data science project

Collect data on every boardgame in existence and use the entirety of your scientific training to learn about boardgames so that you can **predict upcoming games** and **recommend them to people** so you have something to look forward to when the pandemic is over.


This is where we end up.
ANALYTICS CAPABILITIES
TIME
01
OMG!
We have a lot of data!
02
DATA FORTRESS
Data is stored, secured, available
03
BASIC REPORTING
Answer, "What happened?" not intuitive
04
BUSINESS INTELLIGENCE
Answer, "What happened, when and where?" more intuitively
05
STATISTICAL ANALYSIS
Answer, "Why?" questions; project what happens if current trends continue
06
PREDICTIVE MODELS
Illuminate what will happen in the future
07
PRESCRIPTIVE MODELS
Evaluate predicted future alternatives; recommend optimal actions
08
PROCESS OPTIMIZATION
Enabled by analytics-derived insights: transformation underway

Predicting which new games are going to be highly rated/popular

Predicting Upcoming Boardgames

1 Background

What upcoming games on boardgamegeek are expected to be popular?

I trained a variety of models on historical data from BGG in order to make predictions for upcoming games. My models look at information about games that are known at the time of release - mechanics, player count, playingtime, designer, artist, and selected publishers - and estimates four outcomes on BGG: average weight (complexity), average rating, number of user ratings, and geek rating. I trained the models on games published through 2020 in order to predict upcoming games, which for now refers to games published from 2021 onwards.

For more details on my methodology, see my write up at:
https://phenrickson.github.io/bgg_reports/predicting_bgg_outcomes.html

2 Top 100s

In this section, I display the model's top 100 games for the geek rating, average rating, and number of users rated. In each of the tables below, click on the name of the game in order to go directly to the game's boardgamegeek profile, or click on any of the estimated values to go to a report showing how the model arrived at its predictions, along with some other analyses.

2.1 Highest Expected Geek Rating

				Estimated			
Rank	Published	ID	Name	UserRatings	Weight	Average	GeekRating
1	2022	331106	The Witcher: Old World	26,800	3.6	8.5	8.3
2	2021	343905	Boonlake	8,700	4.0	8.2	7.7
3	2021	285967	Ankh: Gods of Egypt	8,600	2.8	8.1	7.7
4	2022	310873	Carnegie	4,900	3.6	8.4	7.6
5	2021	329841	Ticket to Ride: Europe – 15th Anniversary	7,700	2.0	8.0	7.6
6	2022	319807	Shogun no Katana	4,300	3.9	8.4	7.5
7	2022	331224	Zombicide: Undead or Alive	7,000	2.4	8.1	7.5
8	2021	351735	Newton & Great Discoveries	6,900	3.0	8.1	7.5
9	2021	344277	Corrosion	7,500	3.6	8.0	7.5
10	2022	266064	Trudvang Legends	5,200	2.5	8.1	7.5
11	2022	314582	Amsterdam	9,200	3.4	7.8	7.5
12	2021	339906	The Hunger	17,800	2.0	7.6	7.4

1 Game Profile

For any new board game, predicting how it will be rated (and why)

Frozen Frontier
ID: 302892
Published: 2022
Player Count: 1-4
Playing Time: 180 min

1.1 Estimated Outcomes on BGG

This table displays the selected game's **current** values on four BGG outcomes (UsersRated, Average, GeekRating, Weight) along with my predictive model(s) **estimated** values for where these games are likely to end up.

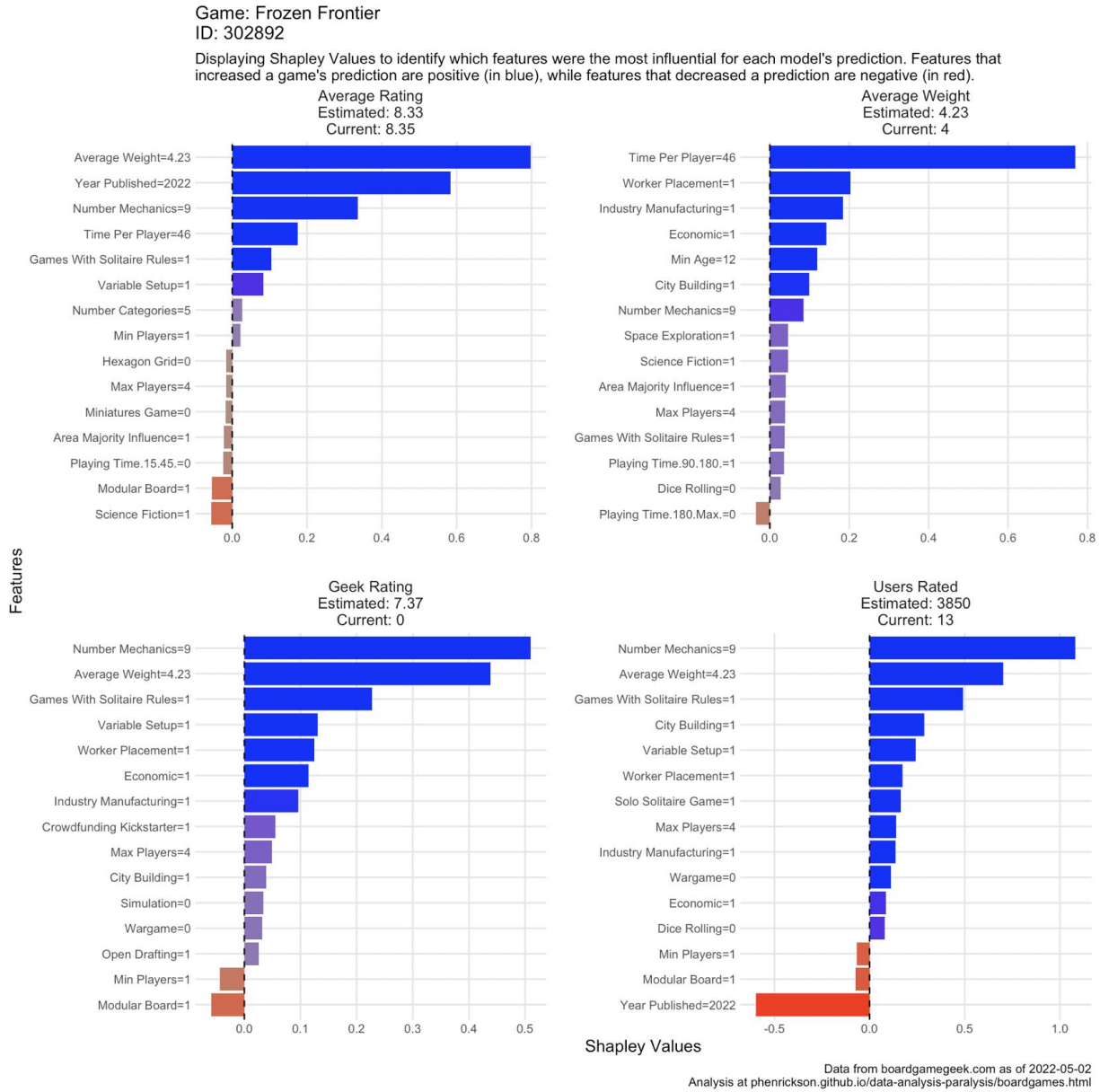
Published	ID	Name	Type	UserRatings	Average	GeekRating	Weight
2022	302892	Frozen Frontier	Current	13	8.35	0.00	4.00
			Estimated	3,900	8.33	7.43	4.23

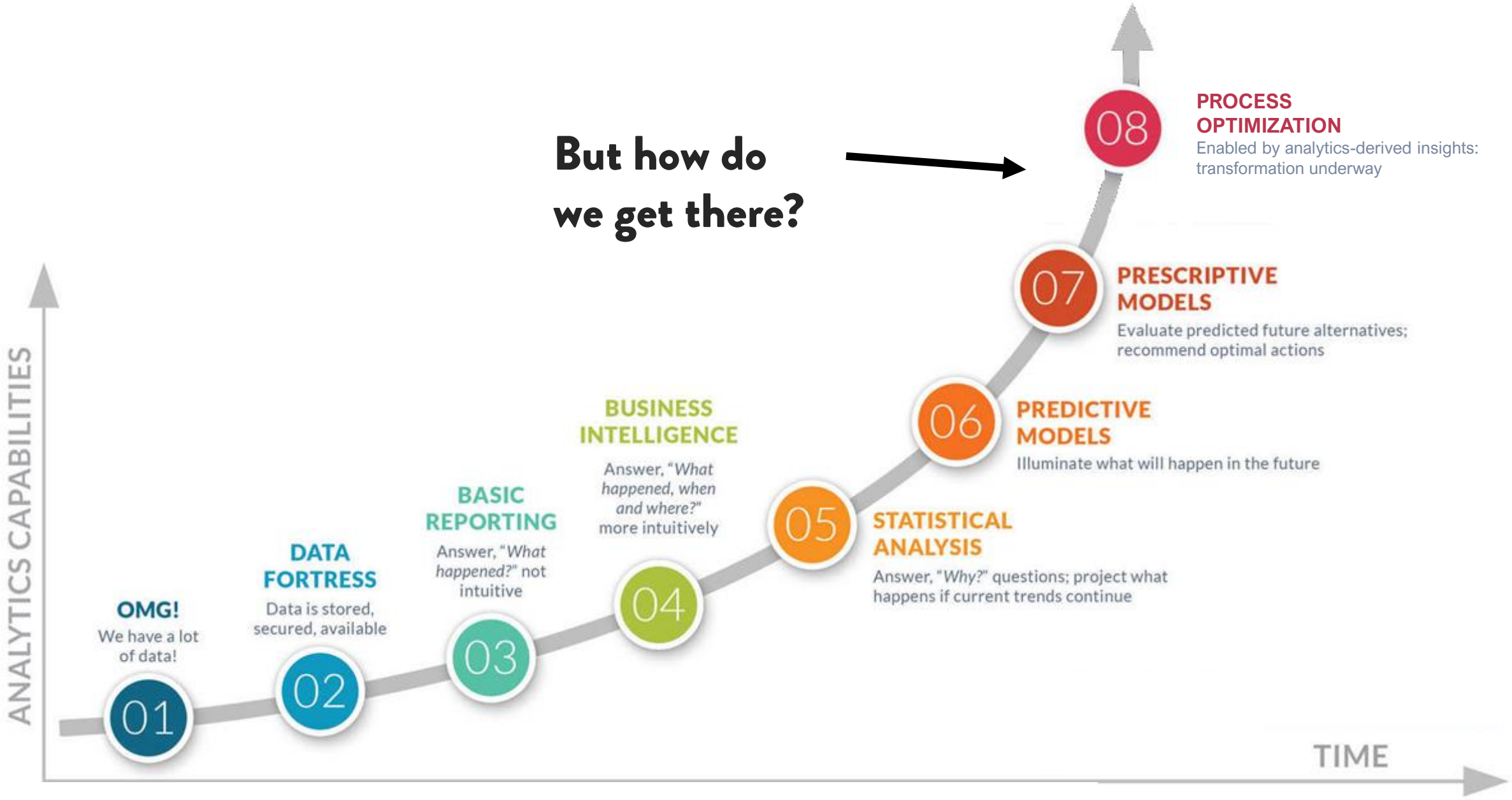
To see more information about the game on boardgamegeek, click on the game's ID or Name to go straight to the game's profile page.

For any new board game, predicting how it will be rated (and why)

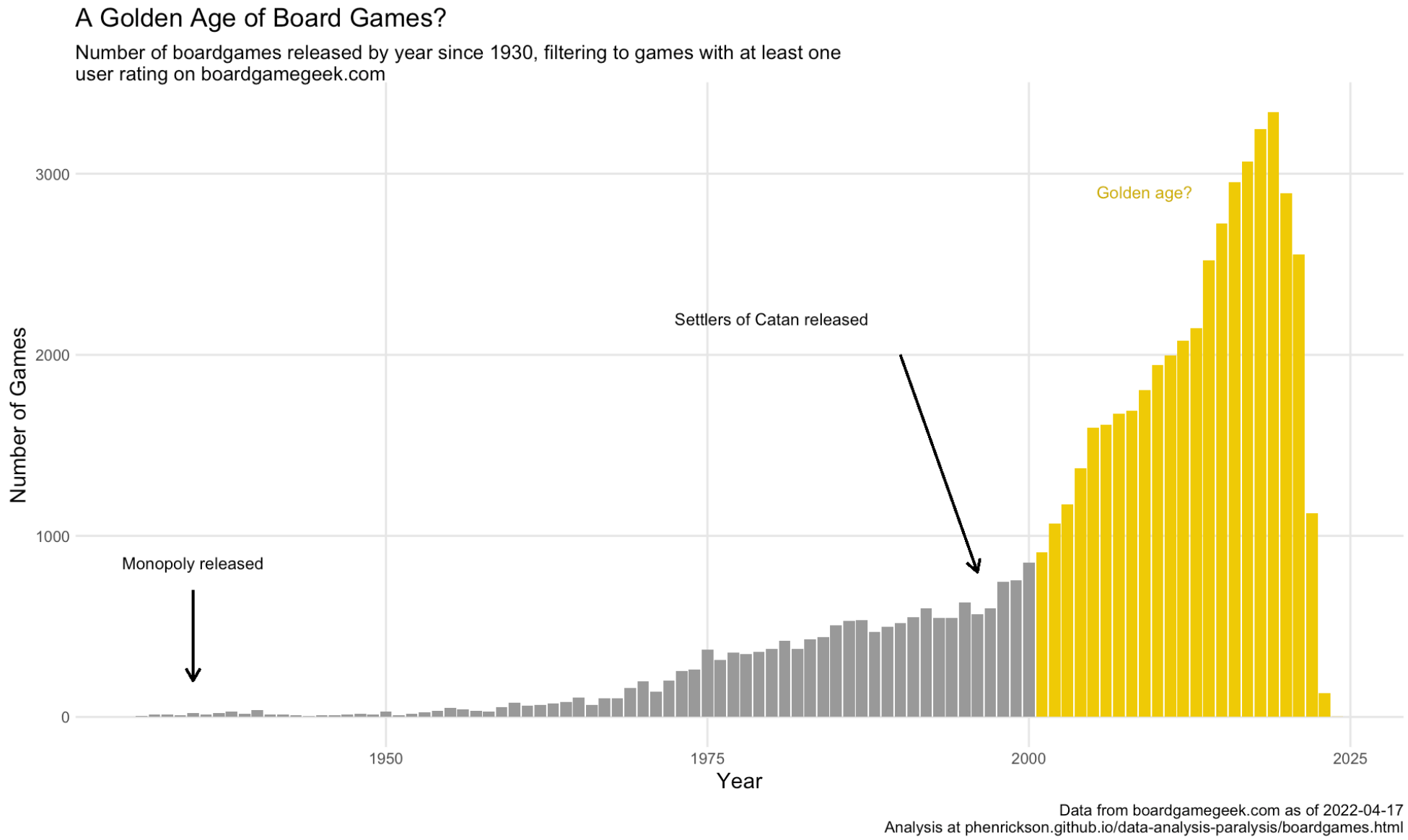
1.4 Explaining Predictions

For each of the predictions above, I used models that were trained on historical boardgamegeek data in order to predict the selected game. How did the model(s) arrive at their predictions? The following plot displays Shapley values to indicate what features were most influential for the model's predictions. Anything in blue increased the model's predictions, anything in red decreased the model's predictions.




But how do we get there?
ANALYTICS CAPABILITIES
TIME
01
OMG!
We have a lot of data!
02
DATA FORTRESS
Data is stored, secured, available
03
BASIC REPORTING
Answer, "What happened?" not intuitive
04
BUSINESS INTELLIGENCE
Answer, "What happened, when and where?" more intuitively
05
STATISTICAL ANALYSIS
Answer, "Why?" questions; project what happens if current trends continue
06
PREDICTIVE MODELS
Illuminate what will happen in the future
07
PRESCRIPTIVE MODELS
Evaluate predicted future alternatives; recommend optimal actions
08
PROCESS OPTIMIZATION
Enabled by analytics-derived insights: transformation underway

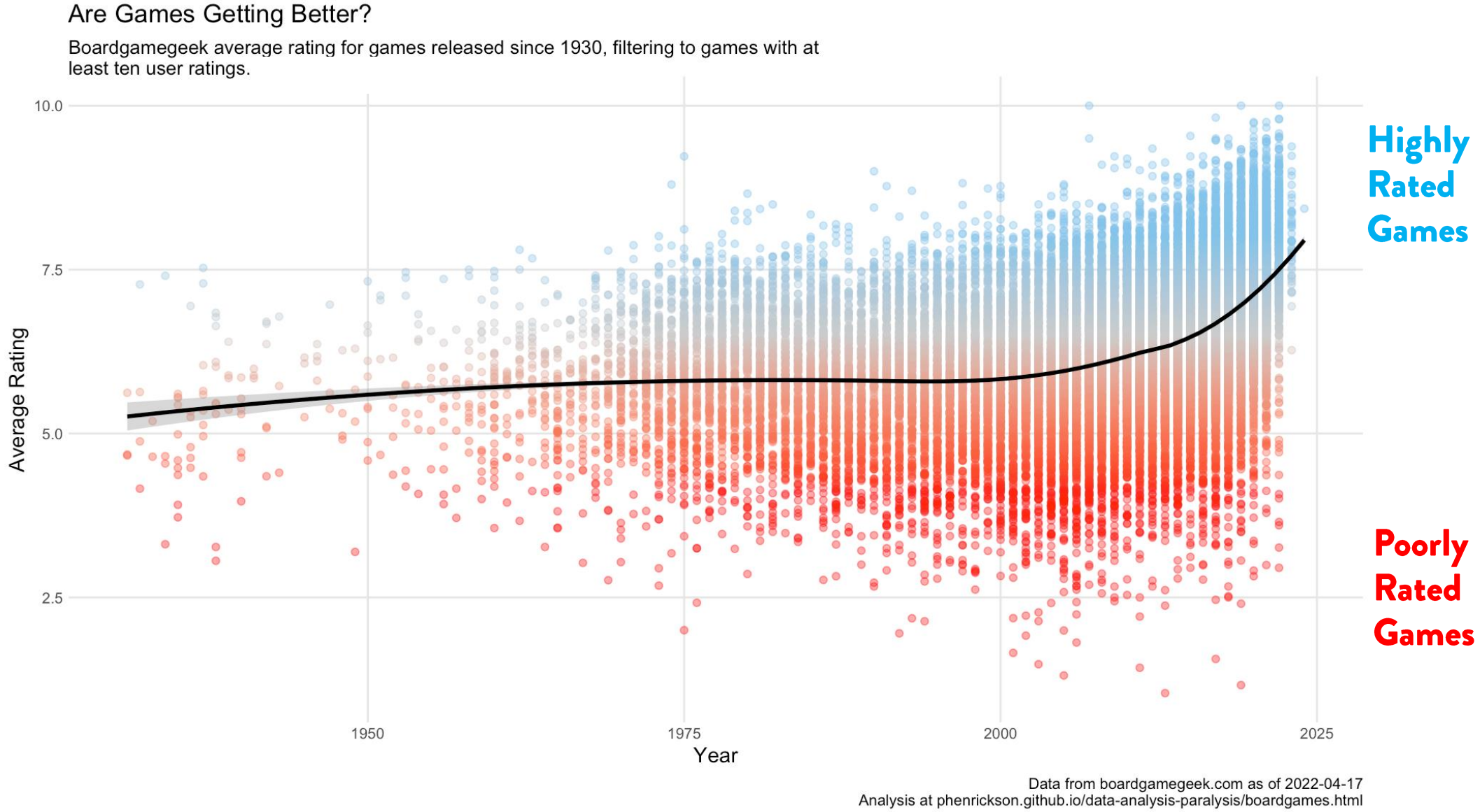
We start where any project begins:
observation.



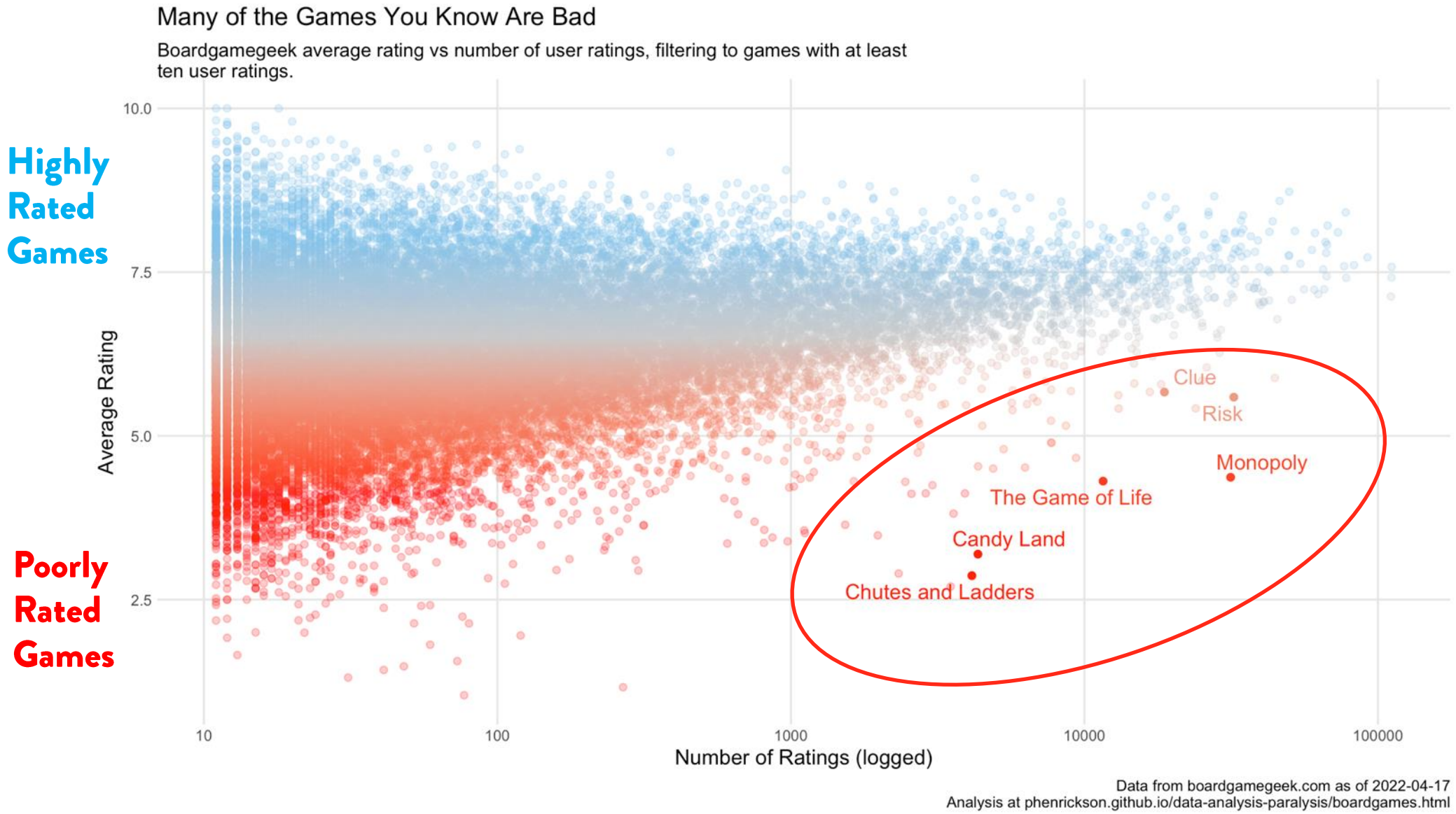
Board games exploded in popularity in the mid 2000s;
thousands of new games come out every year.

Released in 2007
Over 5 million copies sold
Play as the CDC facing a novel virus spreading – oh God

Released in 2019
Over 1 million copies sold
Build an aviary and get the prettiest and best birds



New games tend to be higher rated than old games – games themselves might* be getting better.



Many of the games you know are ~~bad~~ rated quite poorly by the boardgame community.

If you look at enough board games, you start to ask questions: **why are some games better than others**?

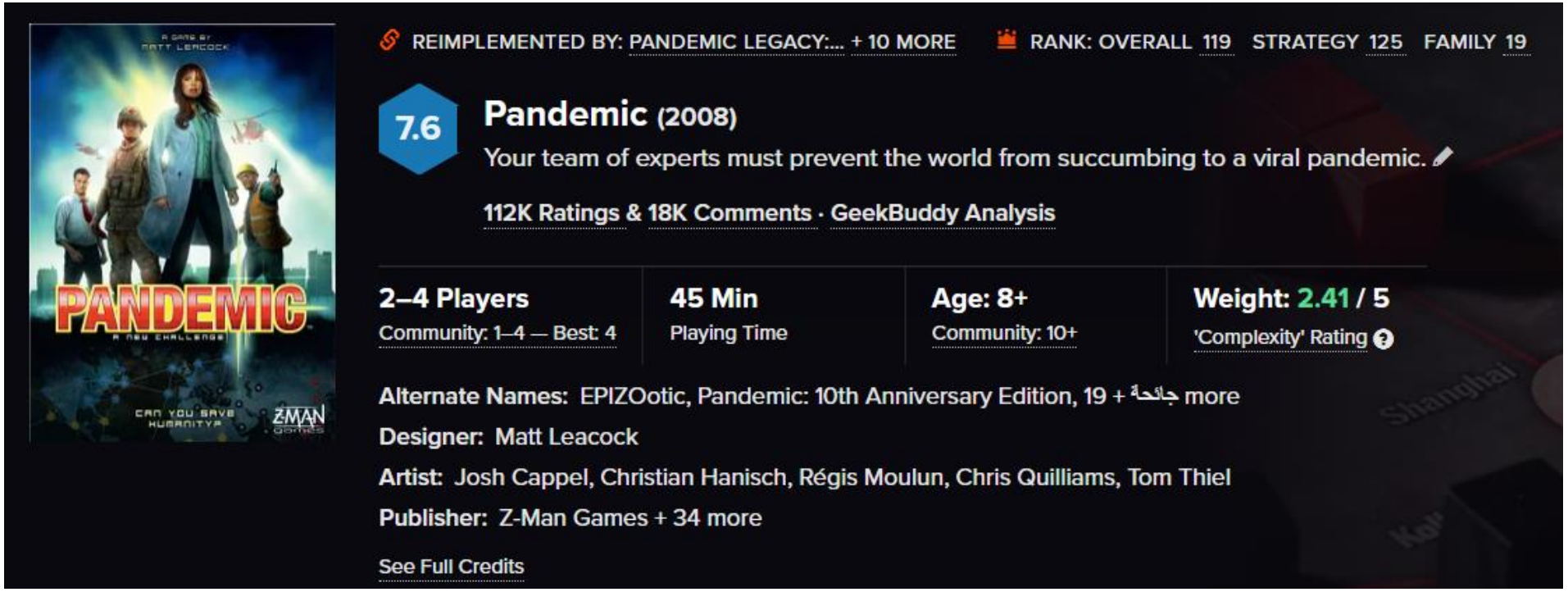
There are so many games out there that it can be overwhelming trying to find games you or your friend might like.


SCYTHE
TAVERNS OF TIEFENTHAL
NEMESIS
BEEZ
LOOP
GREAT WESTERN TRAIL
CAPE MAY
IMPERIAL STEAM
UNDERWATER CITIES
L.A. CRIMES
RESCUING ROBIN HOOD
MY FARM SHOP
HADRIAN'S WALL
PARKS
CALICO

How can I help people find games?

I can **collect data** on every game ever made and how they've been rated, **set up a data warehouse**, and build some **interactive dashboards** to let people explore the data.

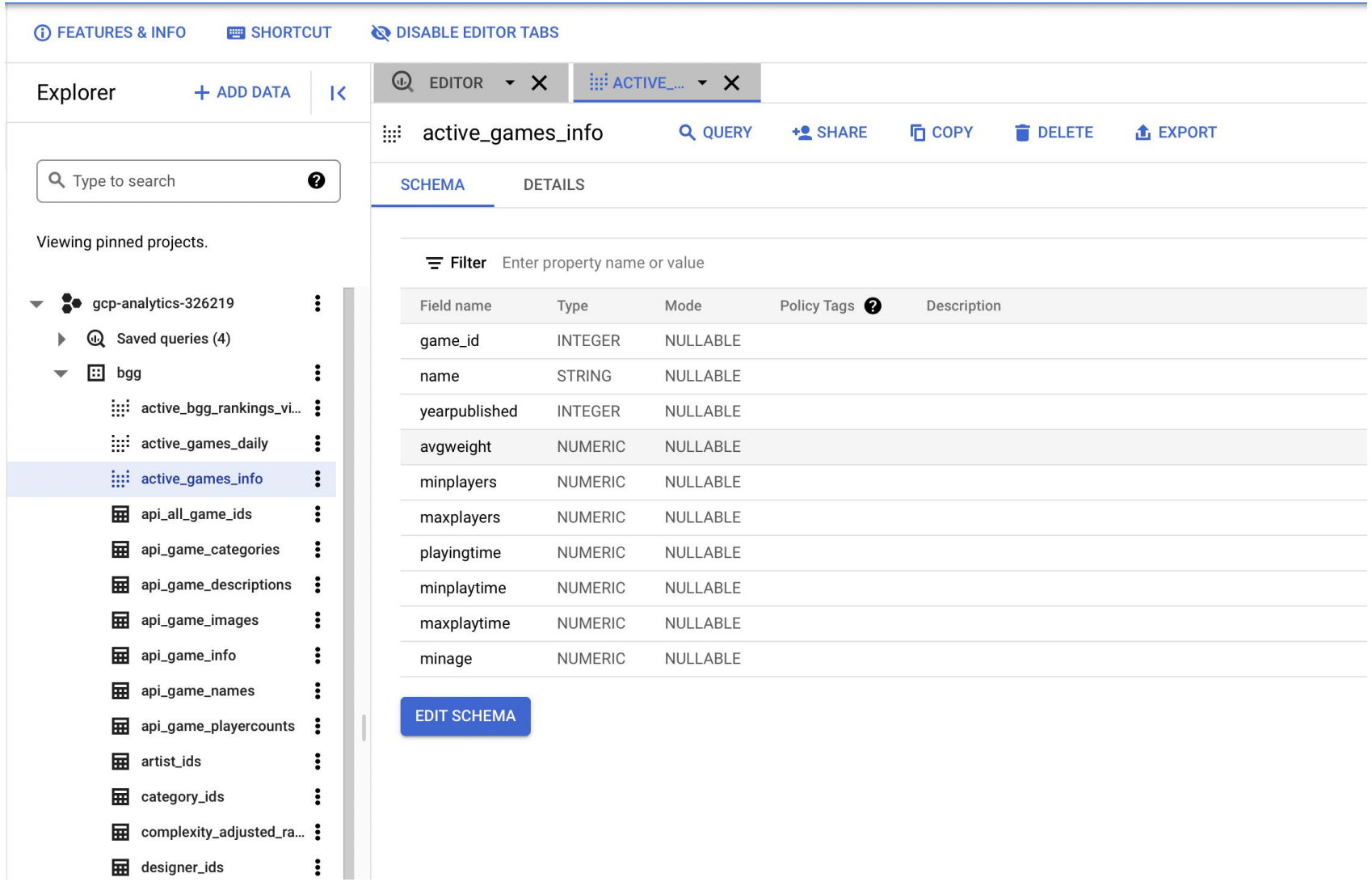
Every single game has a page with data that looks like this; we'll scrape all of them...



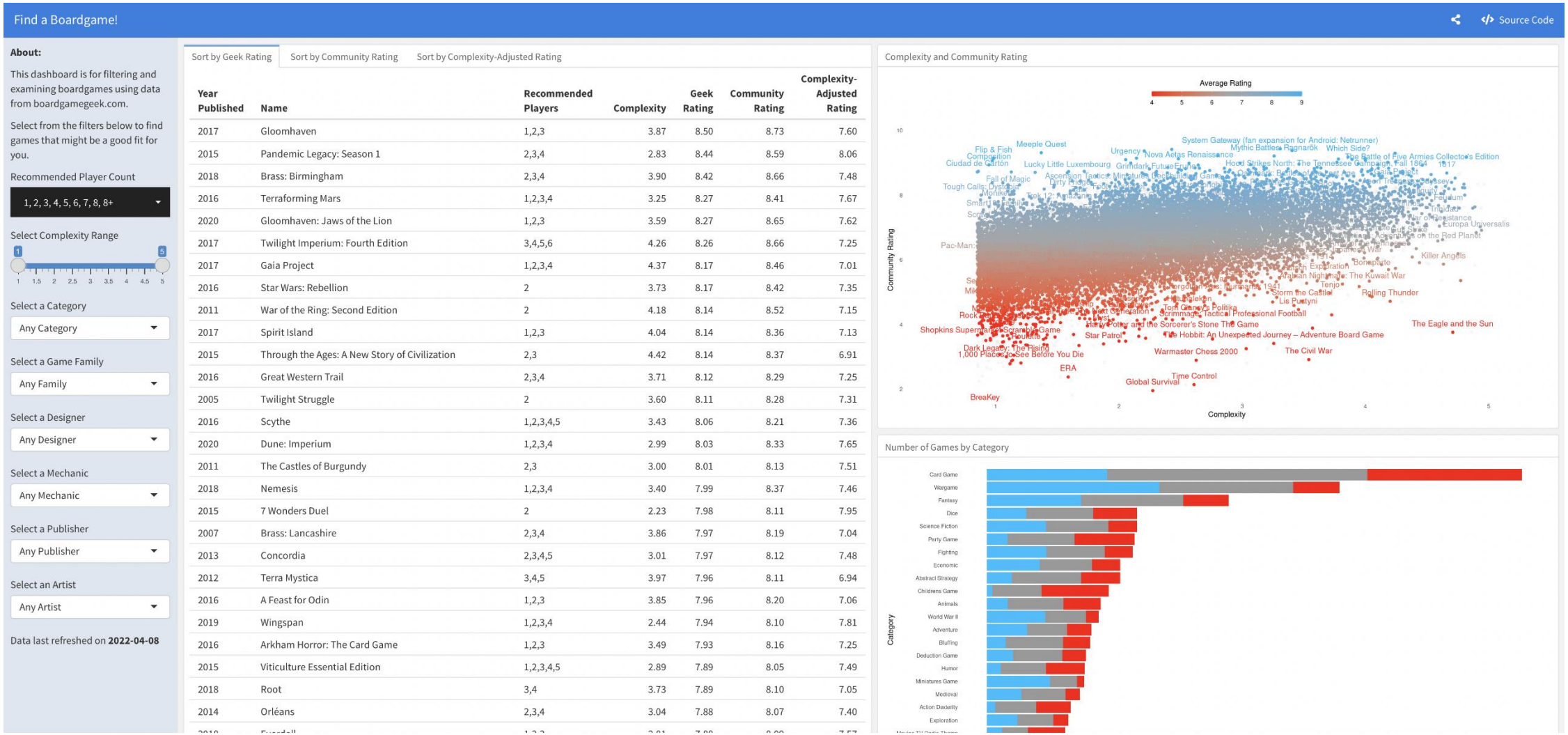
Year Released	2008
Designer	Matt Leacock
Solo Designer	N/A
Artists	Josh Cappel Christian Hanisch Régis Moulun Chris Quilliams Tom Thiel

Mechanisms	Action Points Cooperative Game Hand Management Point to Point Movement Set Collection Trading Variable Player Powers

… which we'll load to a cloud data warehouse…

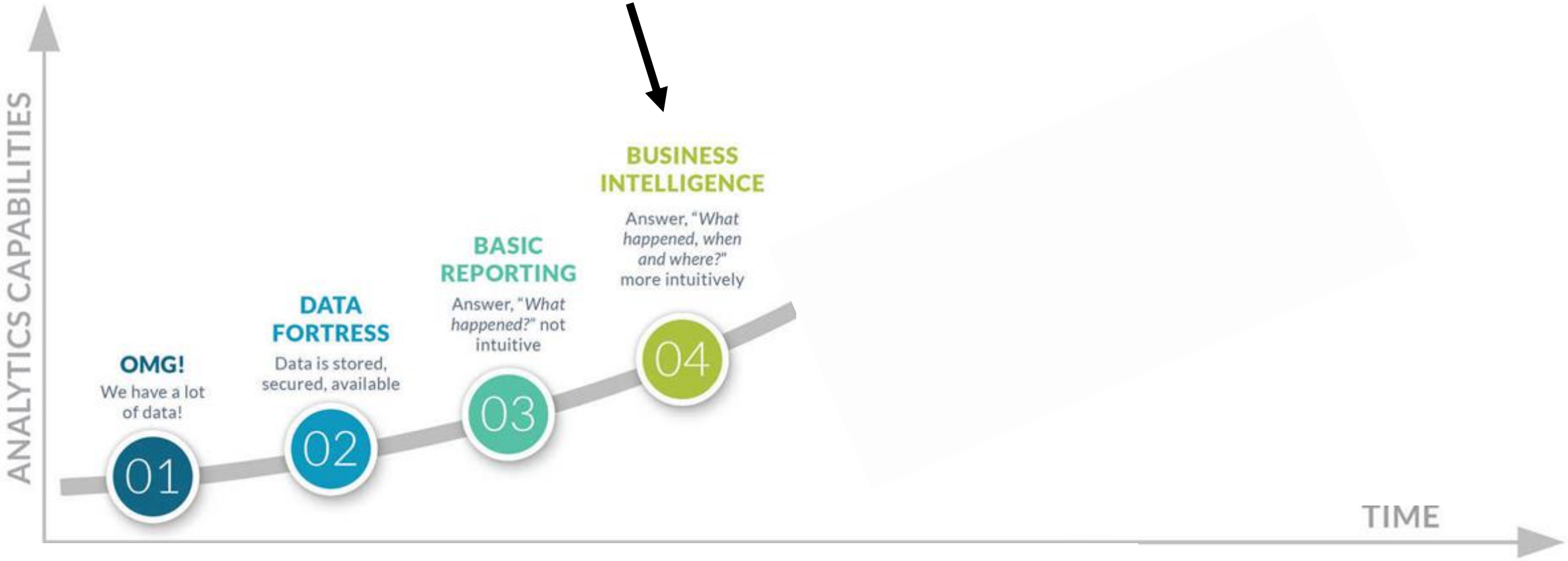


...and then we can build out some interactive dashboards...



https://phil-henrickson.shinyapps.io/explore_bgg/

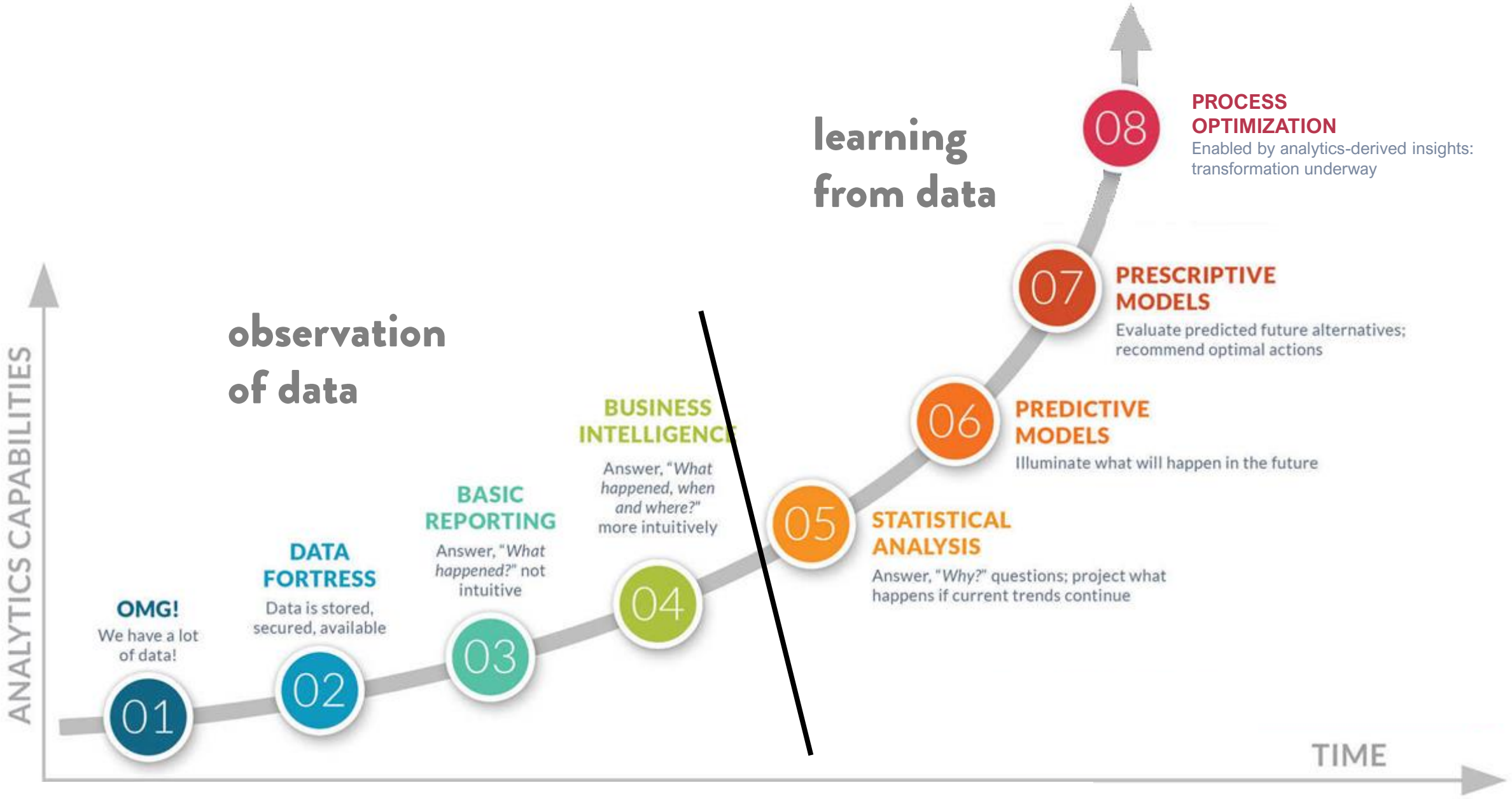
..alright, so we're here.

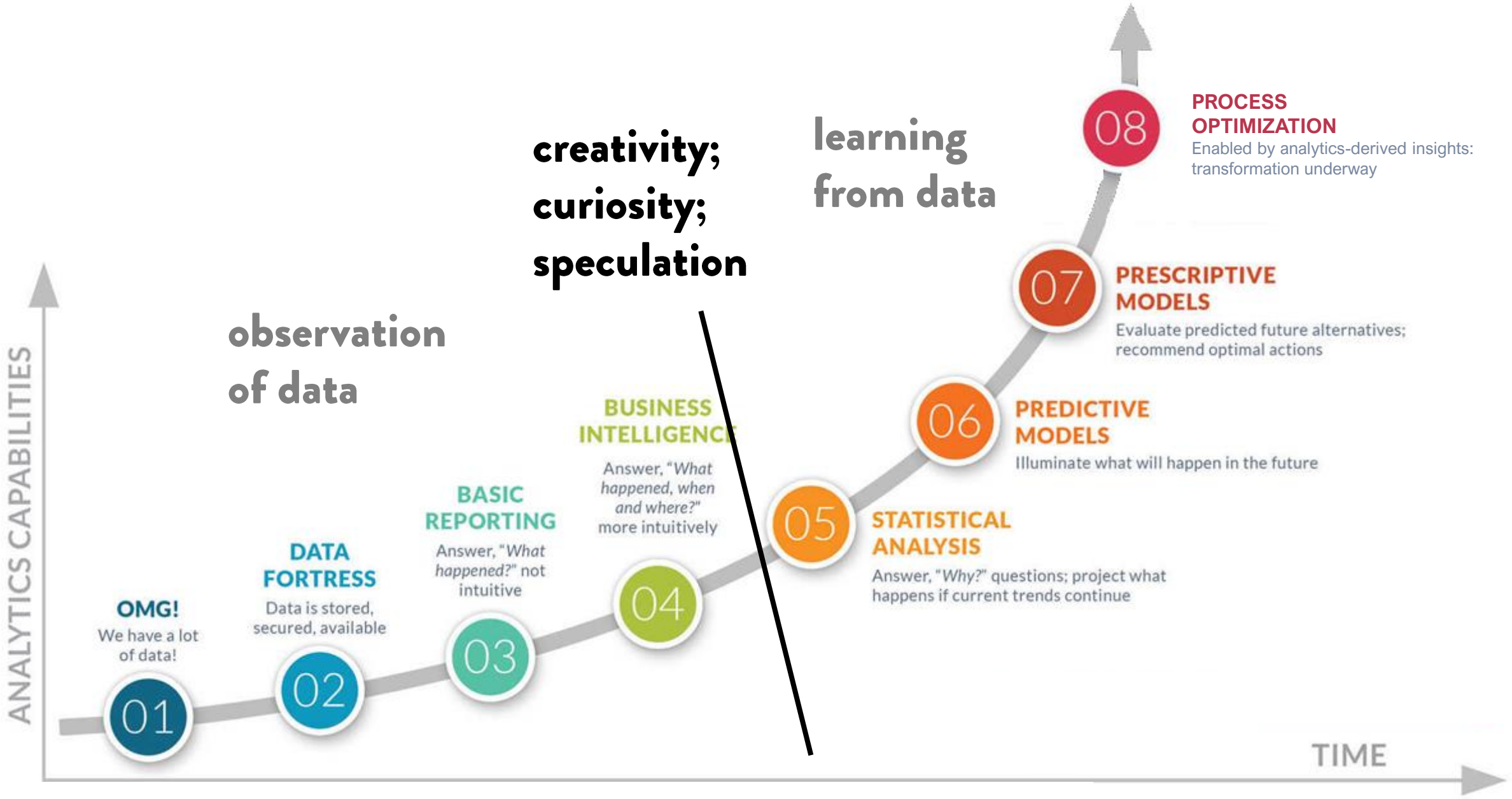


But the data, by itself, can’t answer some of our most important questions:

But the data, by itself, can't answer some of our most important questions:

Why are some games highly rated but not others?
What **new games** are likely to be good?
What games will **I like?**


learning
from data
observation
of data
ANALYTICS CAPABILITIES
TIME
01
OMG!
We have a lot of data!
02
DATA FORTRESS
Data is stored, secured, available
03
BASIC REPORTING
Answer, "What happened?" not intuitive
04
BUSINESS INTELLIGENCE
Answer, "What happened, when and where?" more intuitively
05
STATISTICAL ANALYSIS
Answer, "Why?" questions; project what happens if current trends continue
06
PREDICTIVE MODELS
Illuminate what will happen in the future
07
PRESCRIPTIVE MODELS
Evaluate predicted future alternatives; recommend optimal actions
08
PROCESS OPTIMIZATION
Enabled by analytics-derived insights: transformation underway

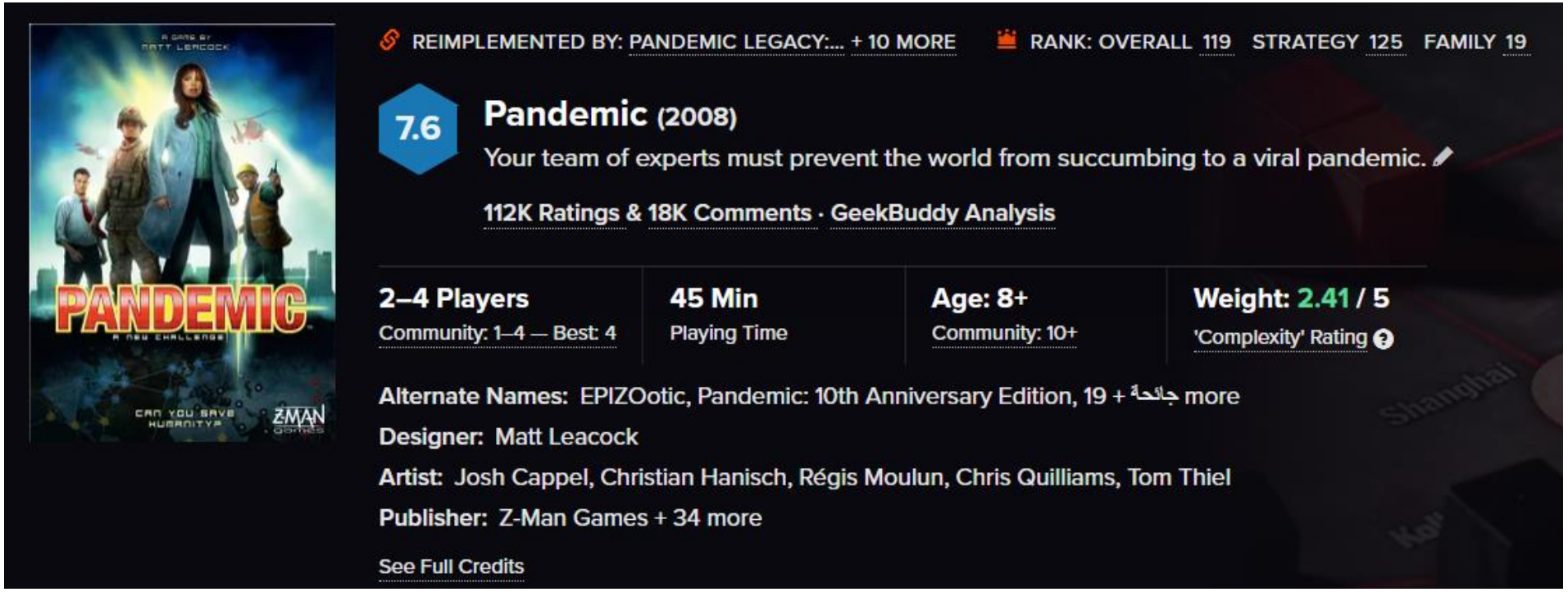

creativity;
curiosity;
speculation
learning
from data
observation
of data
ANALYTICS CAPABILITIES
TIME
01
OMG!
We have a lot of data!
02
DATA FORTRESS
Data is stored, secured, available
03
BASIC REPORTING
Answer, "What happened?" not intuitive
04
BUSINESS INTELLIGENCE
Answer, "What happened, when and where?" more intuitively
05
STATISTICAL ANALYSIS
Answer, "Why?" questions; project what happens if current trends continue
06
PREDICTIVE MODELS
Illuminate what will happen in the future
07
PRESCRIPTIVE MODELS
Evaluate predicted future alternatives; recommend optimal actions
08
PROCESS OPTIMIZATION
Enabled by analytics-derived insights: transformation underway

Can I build models on historical games to predict which new games will be highly rated/popular?

Can I build models on historical games to predict which new games will be highly rated/popular?

I might not be able to!
But I can speculate about what might be useful and then try.

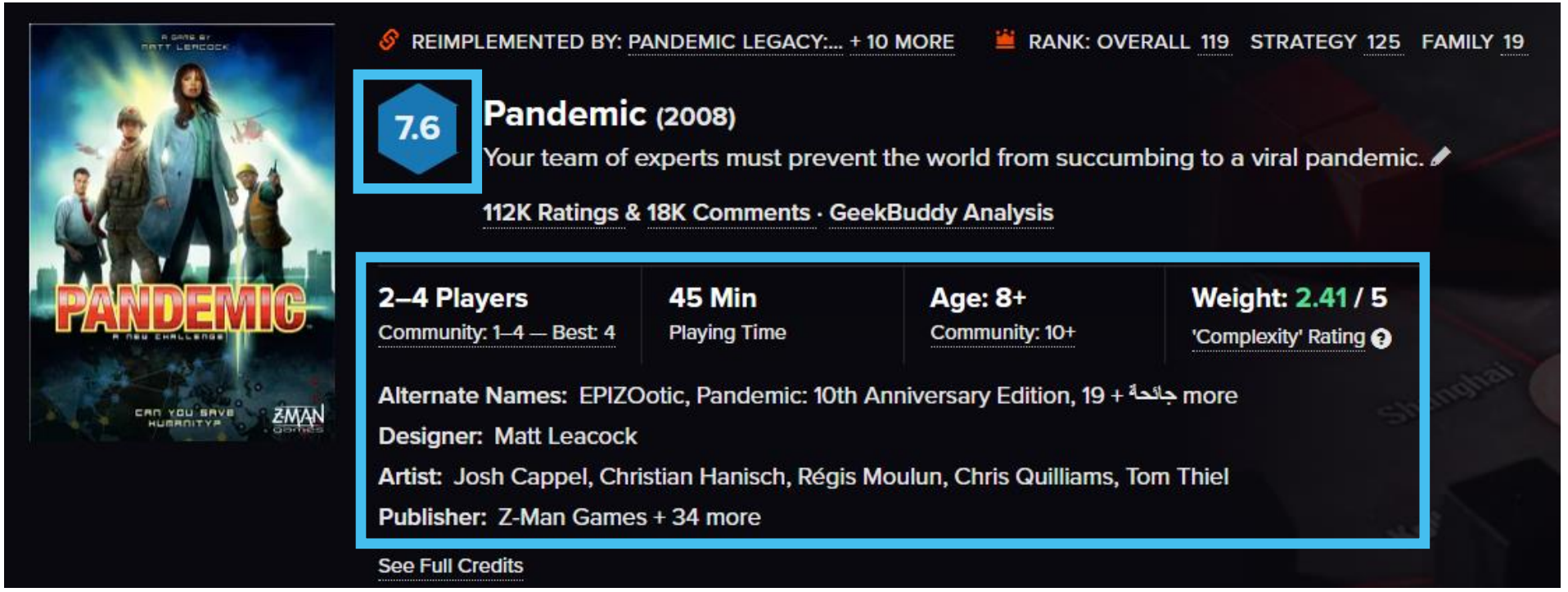
What data do we have?



Year Released	2008
Designer	Matt Leacock
Solo Designer	N/A
Artists	Josh Cappel Christian Hanisch Régis Moulun Chris Quilliams Tom Thiel

Mechanisms	Action Points Cooperative Game Hand Management Point to Point Movement Set Collection Trading Variable Player Powers

What data do we have?



Year Released	2008
Designer	Matt Leacock
Solo Designer	N/A
Artists	Josh Cappel Christian Hanisch Régis Moulun Chris Quilliams Tom Thiel

Mechanisms	Action Points Cooperative Game Hand Management Point to Point Movement Set Collection Trading Variable Player Powers

The Modeling Process

Features

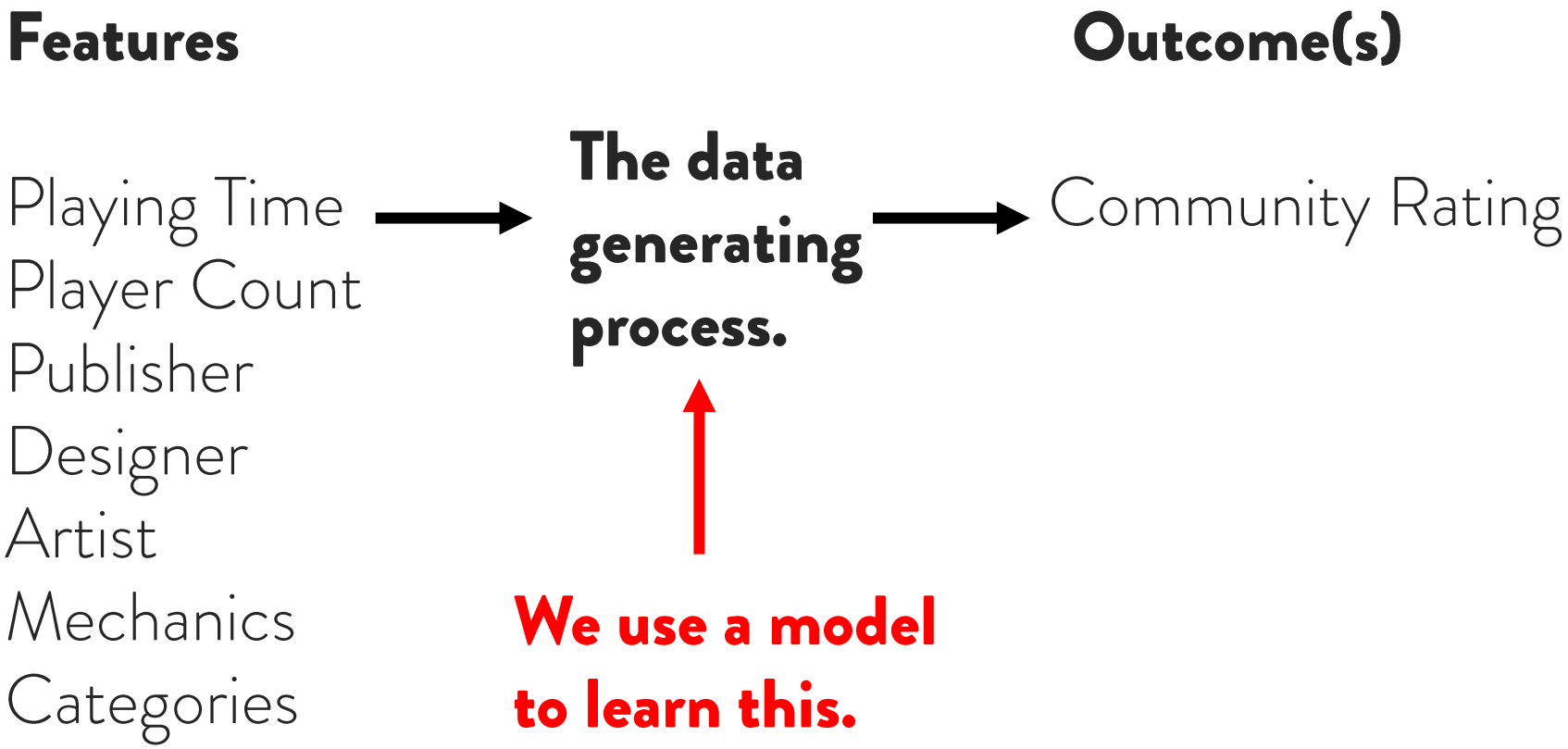
Playing Time
Player Count
Publisher
Designer
Artist
Mechanics
Categories

The data generating process.

Outcome(s)

Community Rating

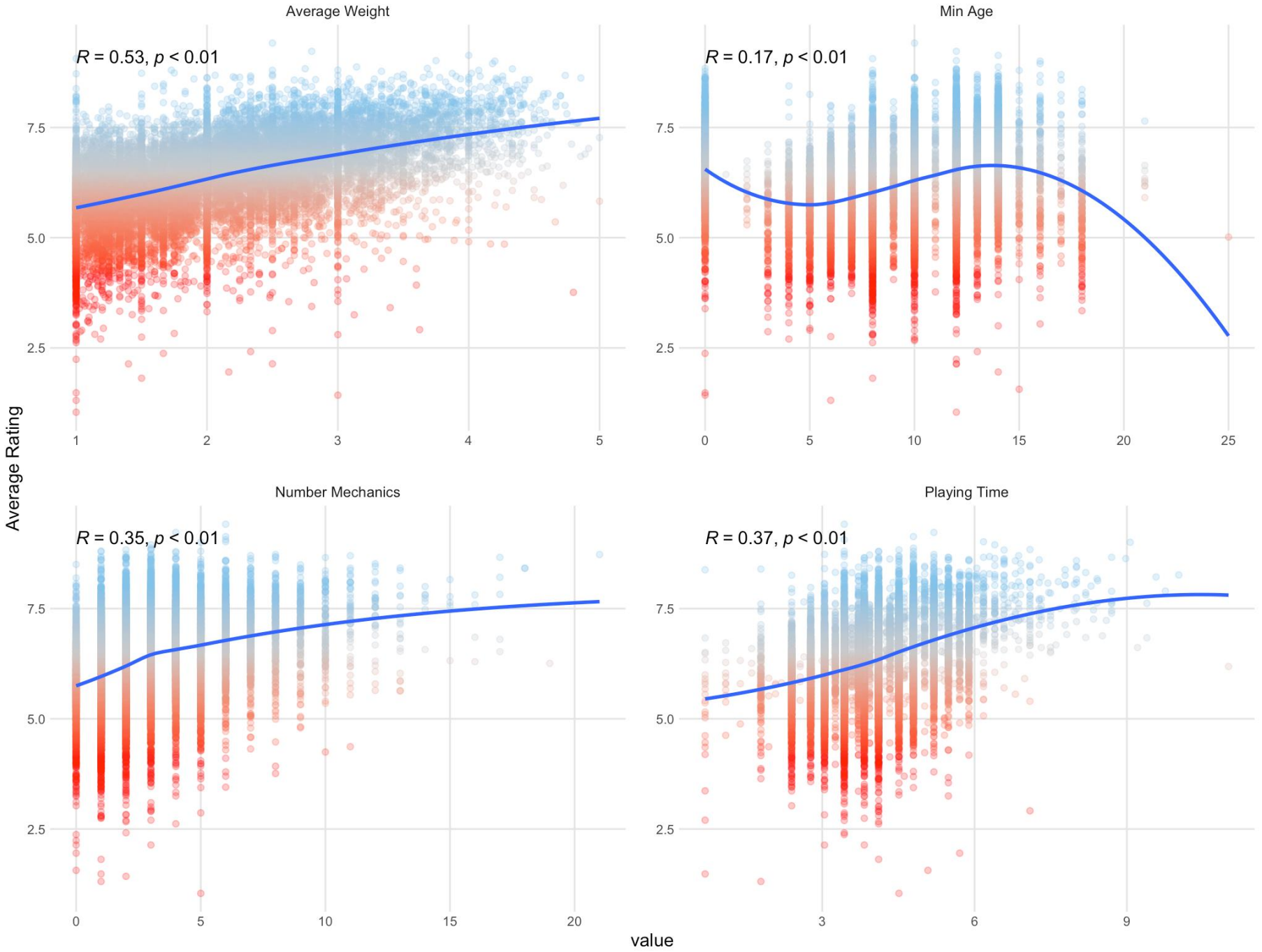
The Modeling Process



In order to get a model to learn this relationship, we need to **speculate about features** involved in the data generating process.

The data will not do this for us, and a model can only do so much.

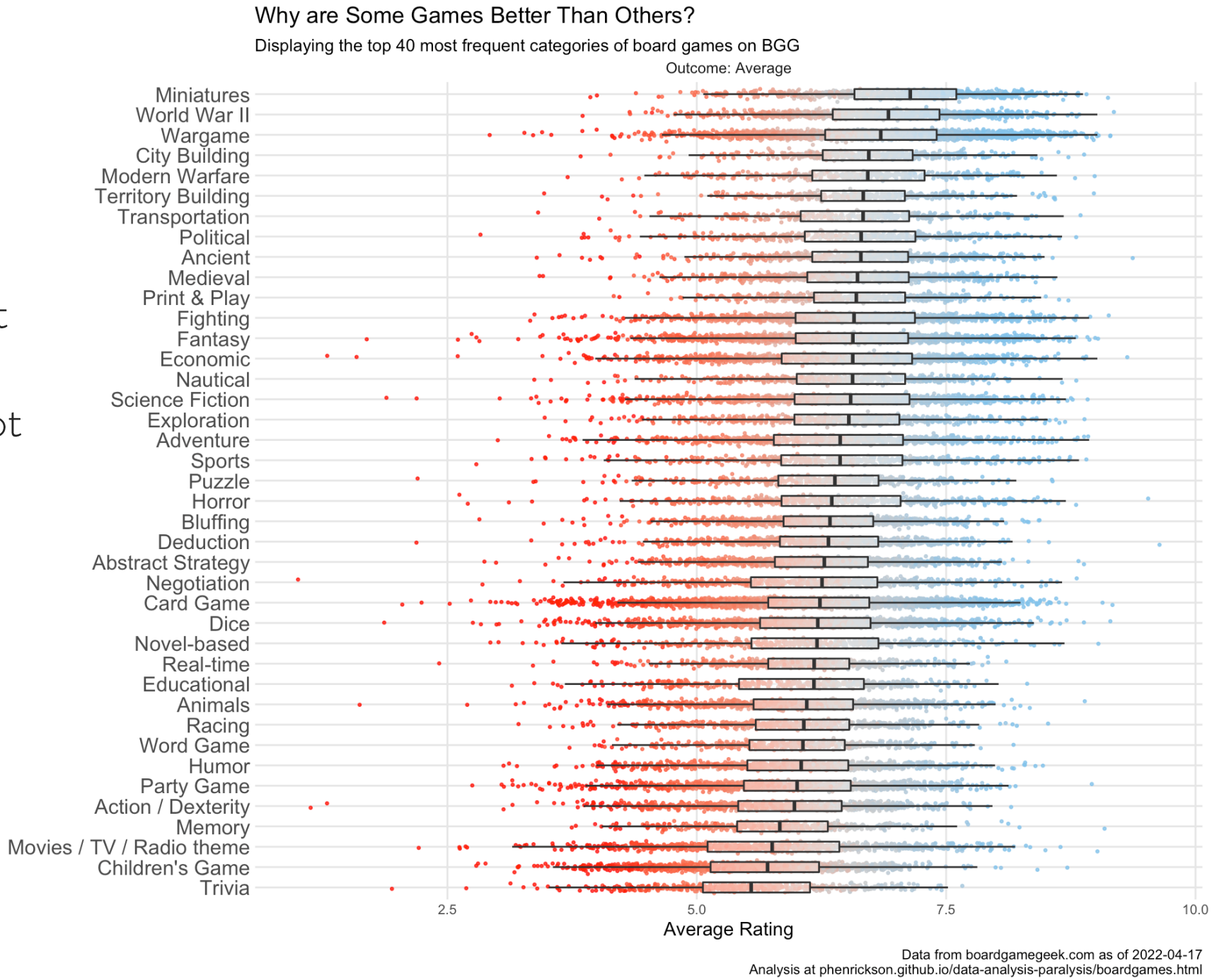
We start looking at the data, and we notice some patterns.



Data from boardgamegeek.com as of 2022-04-27
Analysis at phenrickson.github.io/data-analysis-paralysis/boardgames.html

We make decisions about what data to include (and not include) in the model.

Then we train the model.



How does a model learn?

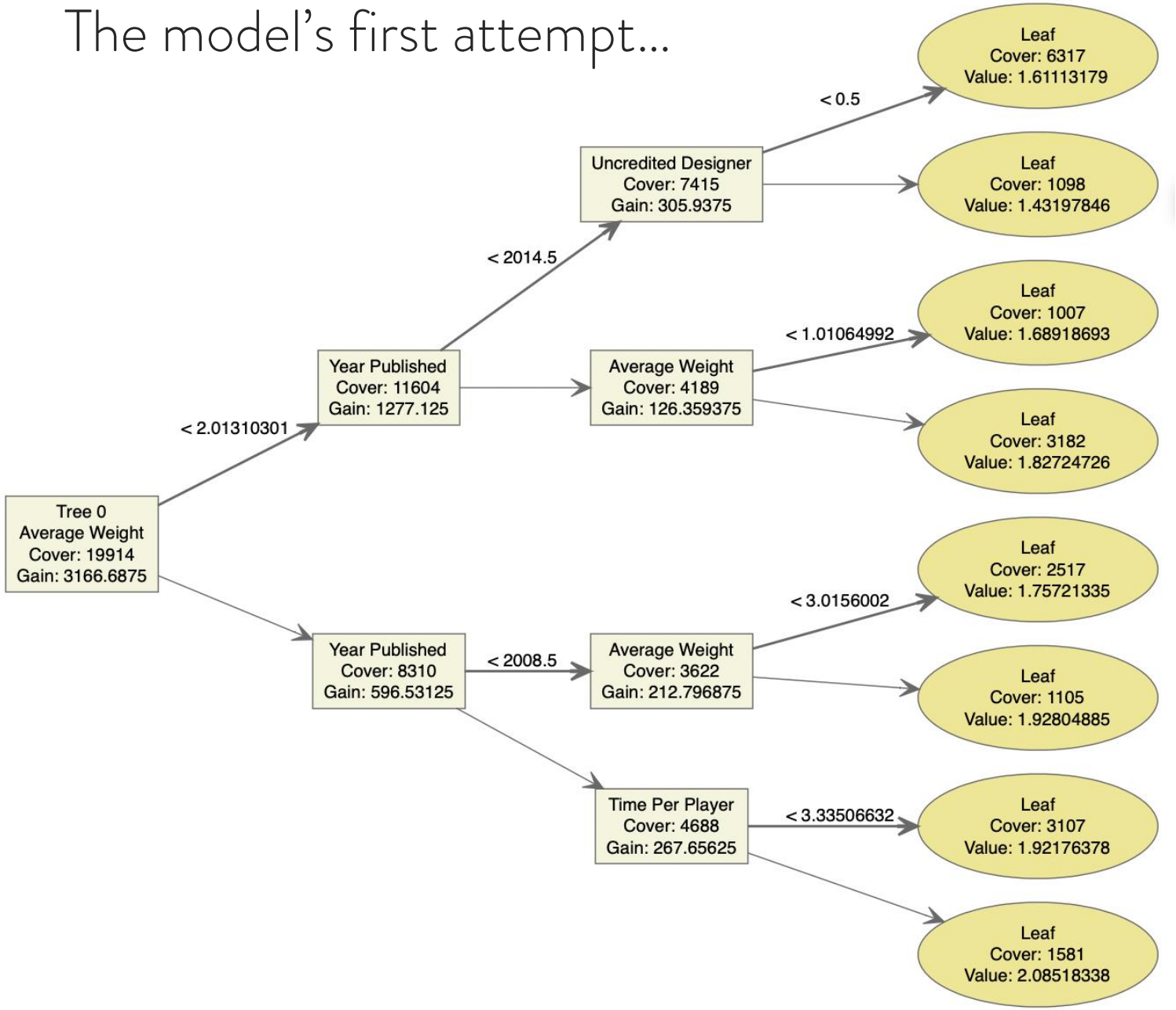
The same way we do. It looks at the data, finds a pattern, then makes a prediction. It tries to figure out where it went wrong, and then it tries again.

How does a model learn?

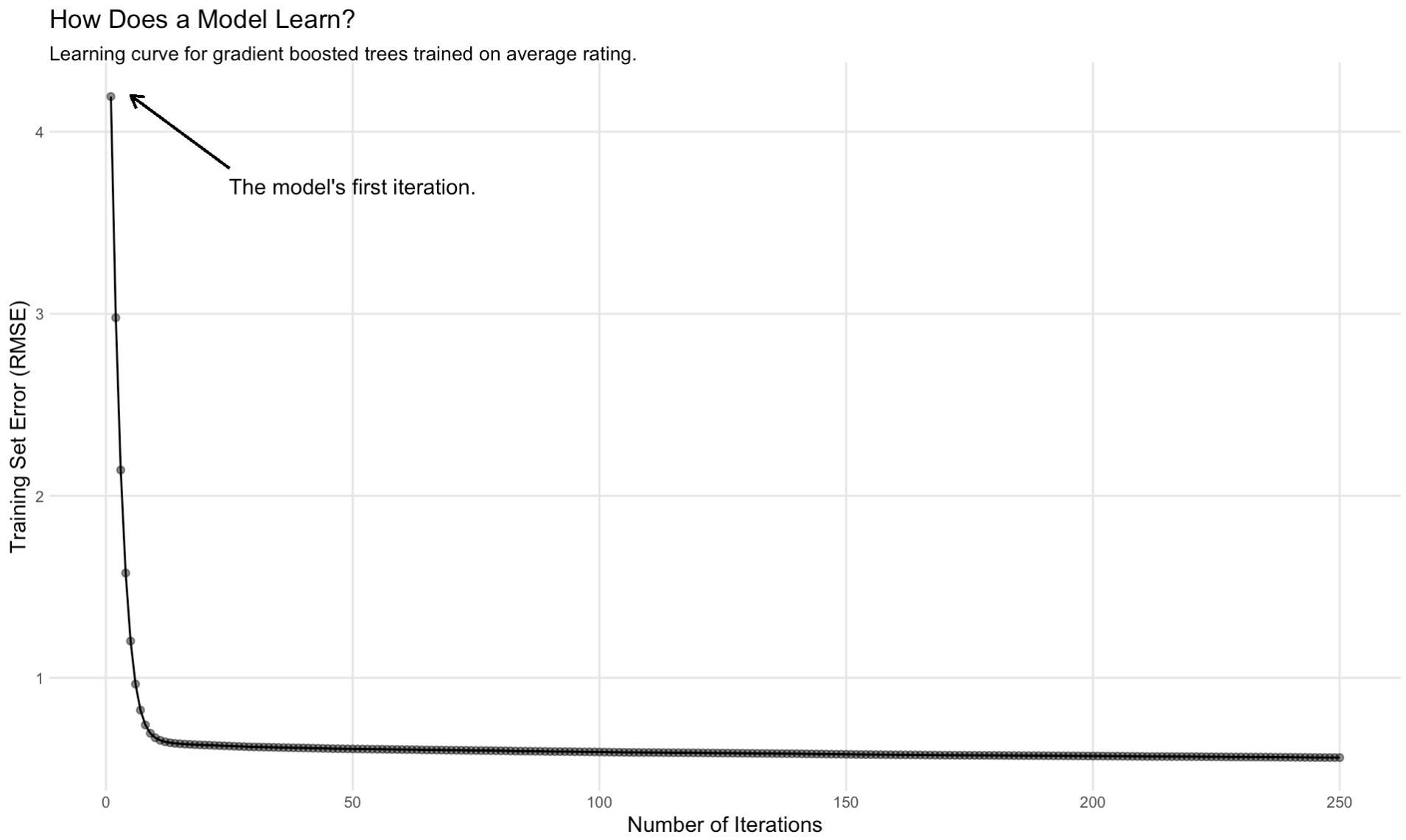
The same way we do. It looks at the data, finds a pattern, then makes a prediction. It tries to figure out where it went wrong, and then it tries again.

I set the model to look at games published before 2019. Here is the model's first attempt at making predictions.

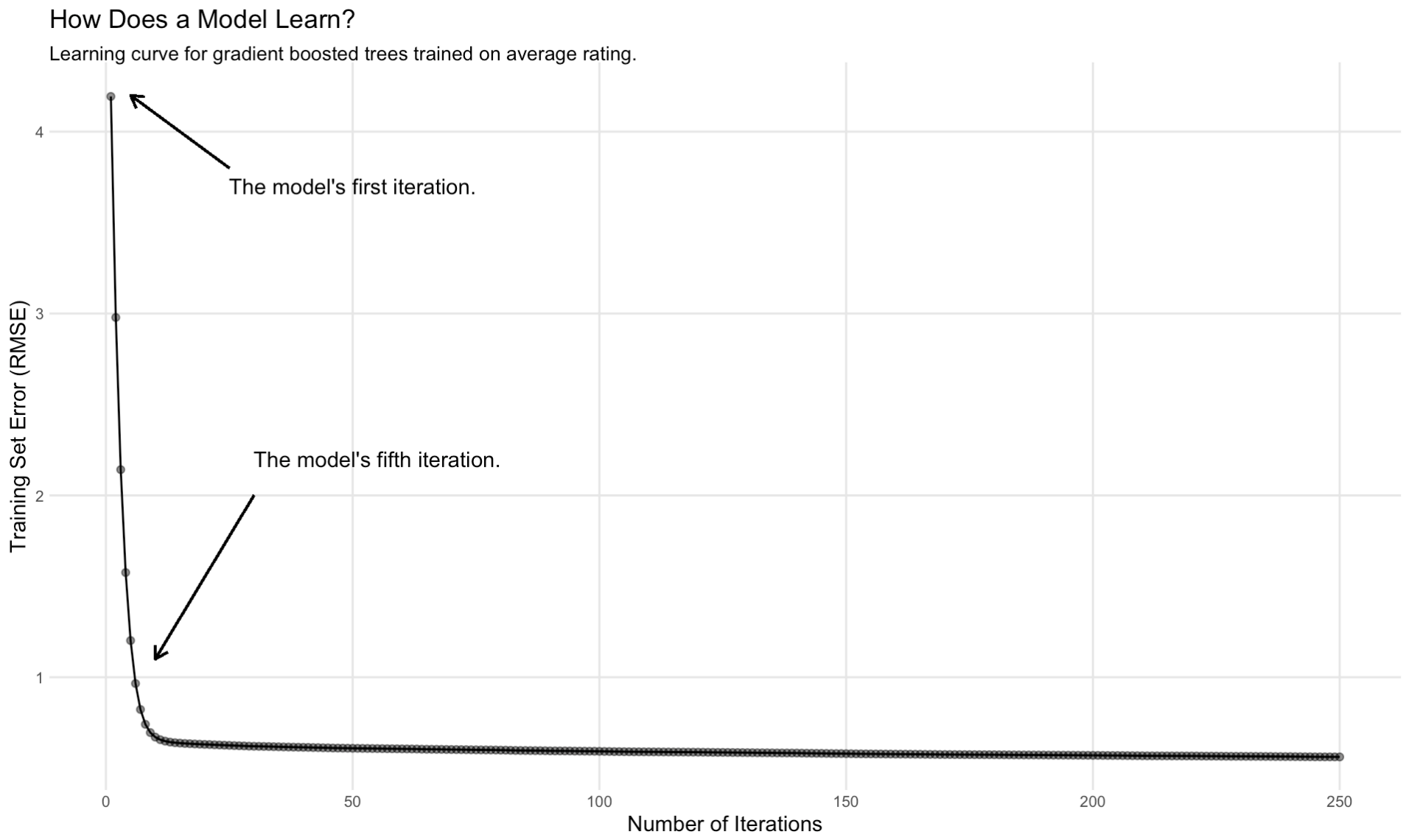
The model's first attempt...



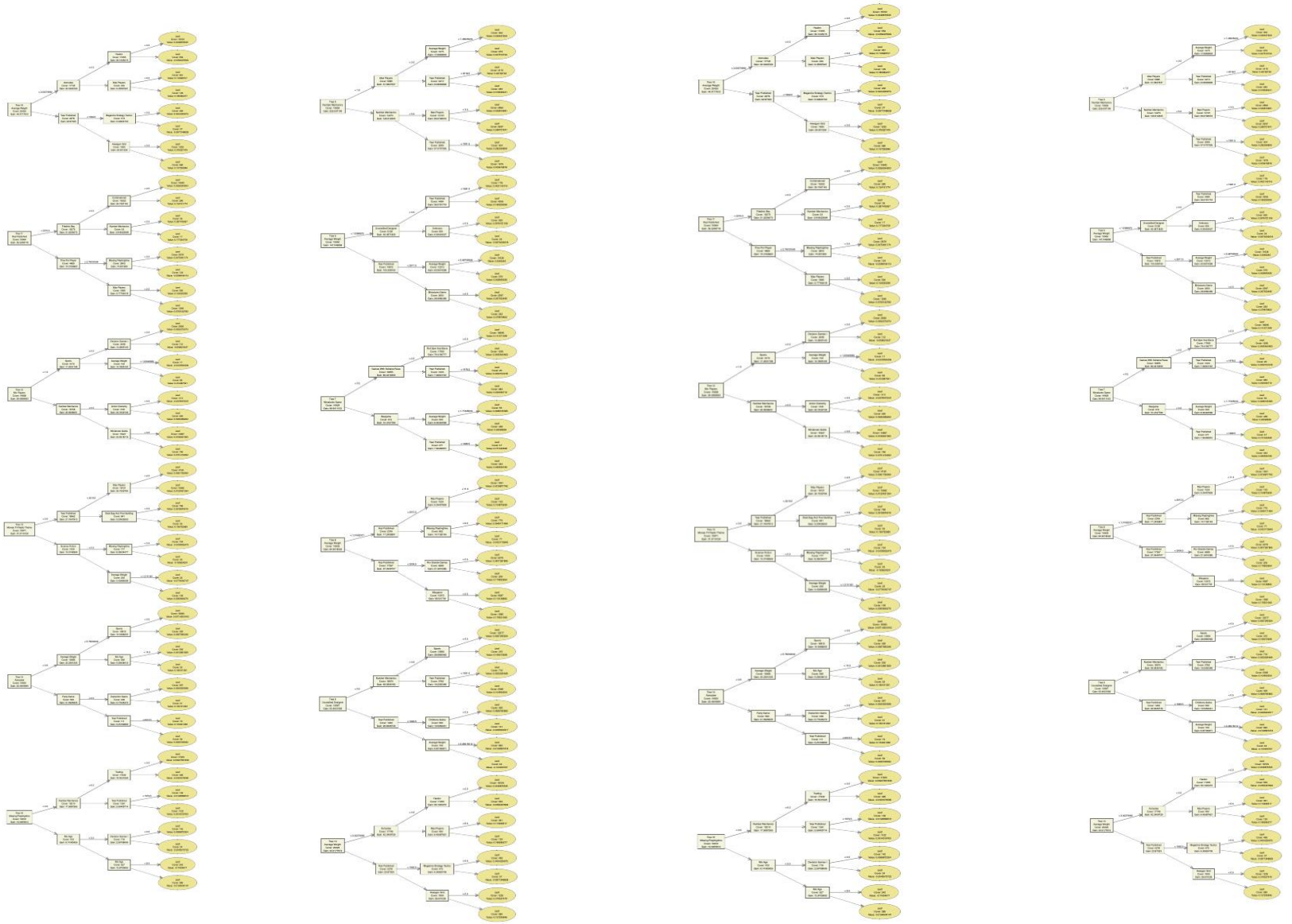
The model's first attempt... **is super inaccurate.**



But that's okay! It learns, tries again, and gets better.



The model keeps trying, until I tell it to stop.



At this point comes **the test**:

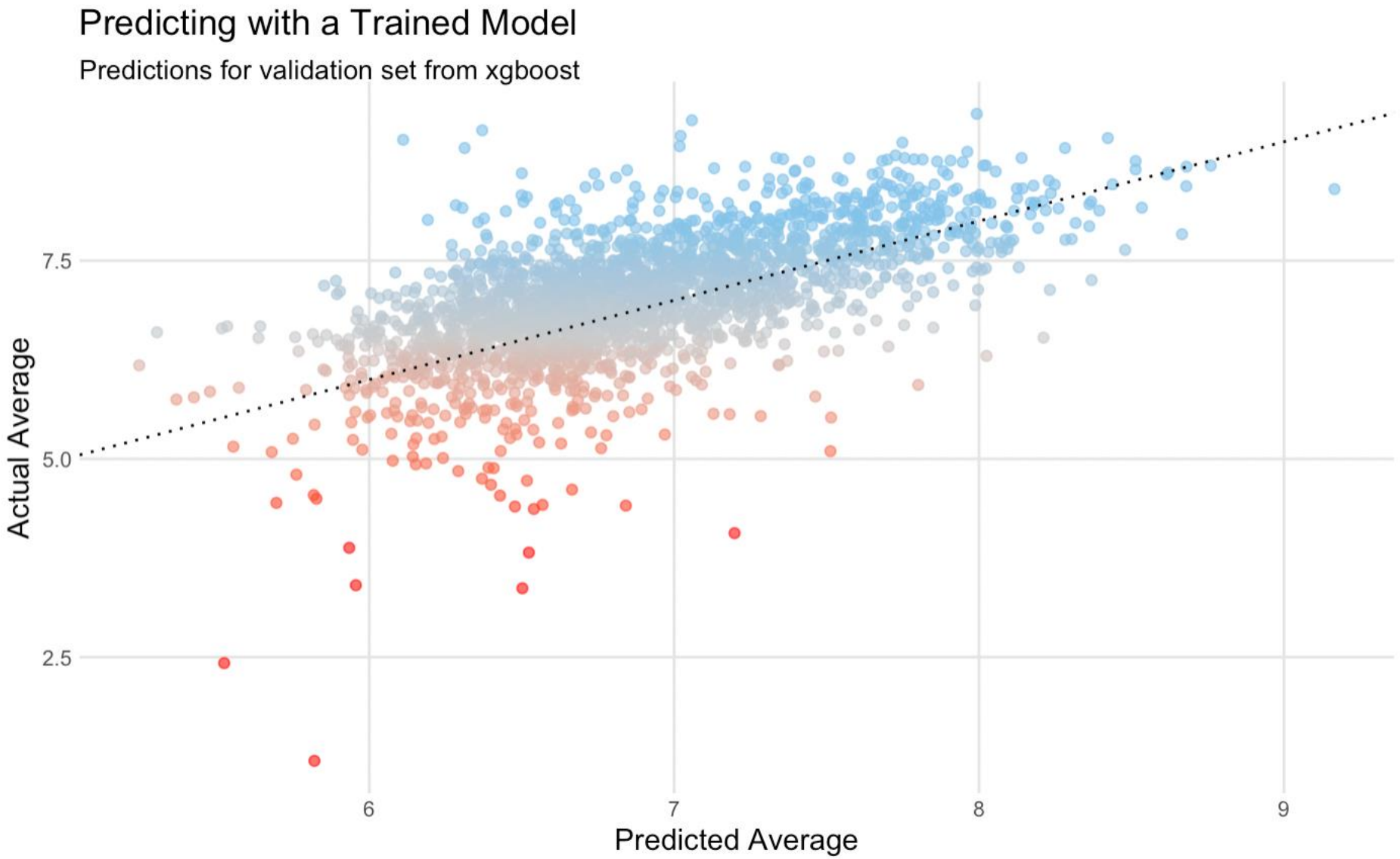
If the model has learned something about
the data generating process,
it should be able to predict new games.

At this point comes **the test**:

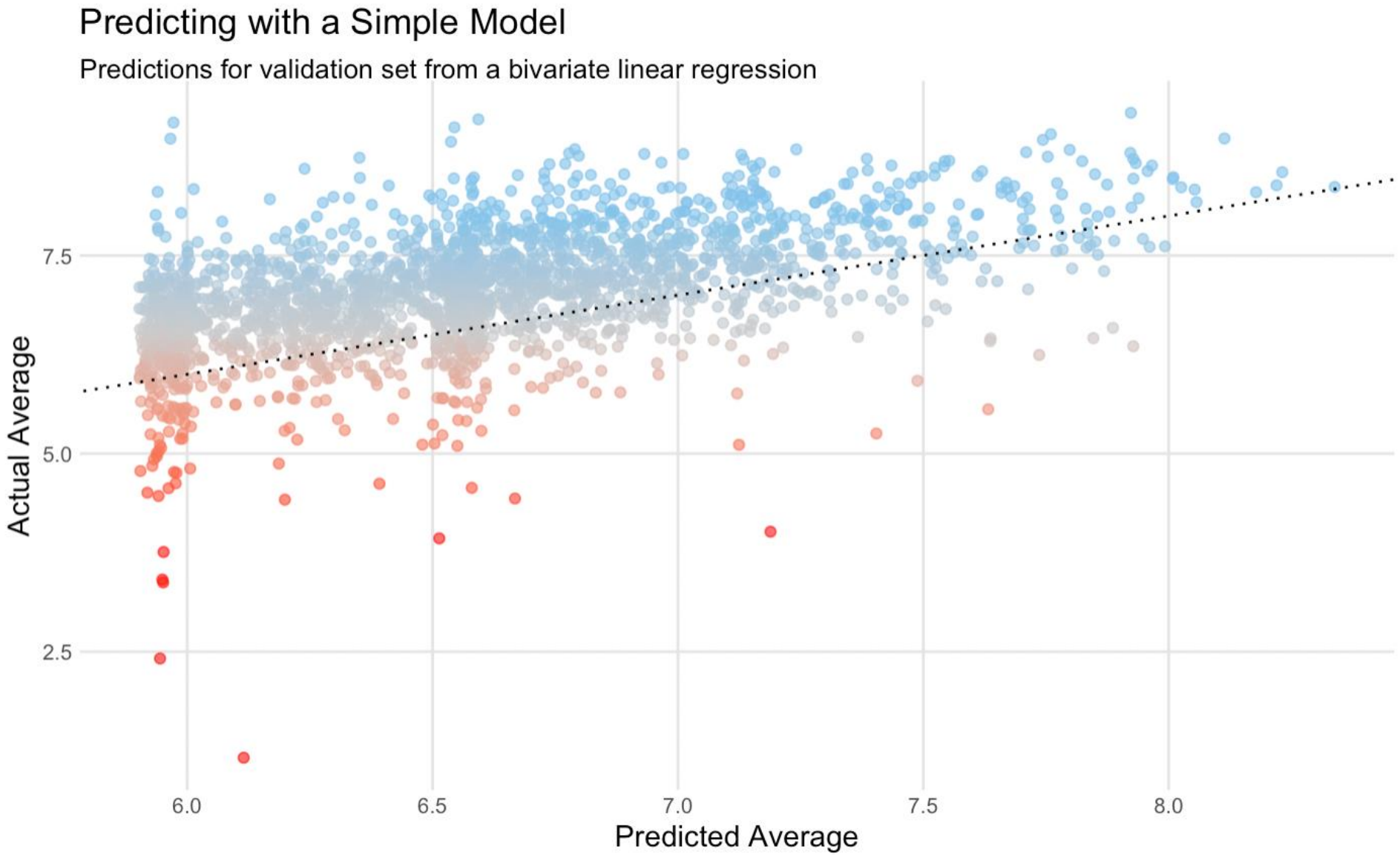
If the model has learned something about the data generating process,
it should be able to predict new games.

So, we validate the model by predicting games it hasn't seen before, those released in 2019-2020.

How did it do?



The Complex Model
Typical Error: .62 ~25% improvement



Simple Model
Typical Error: .81

Hooray! The model we trained performed pretty well, better than simpler models.

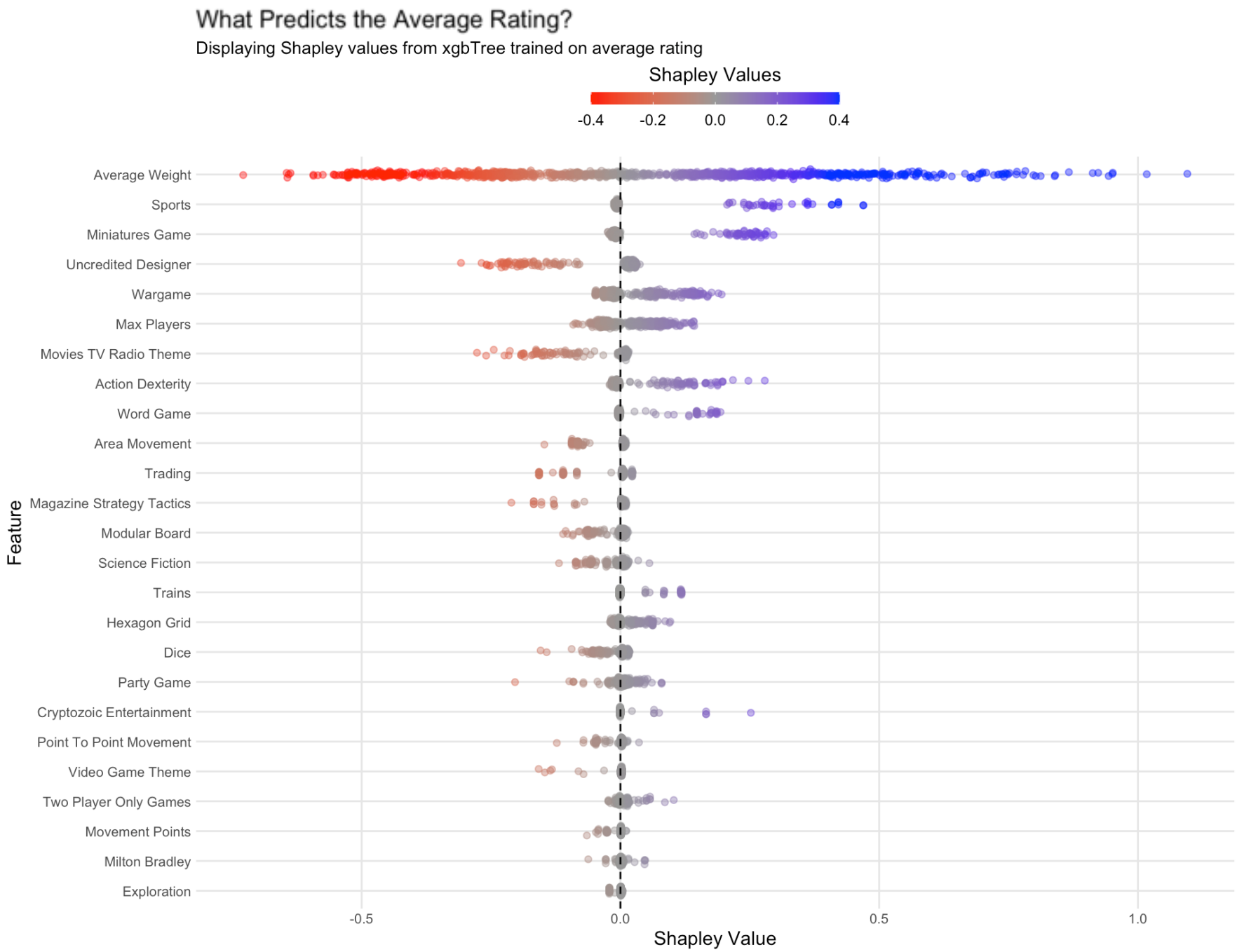
Are we done? Do we just roll with the model?

Hooray! The model we trained performed pretty well, better than simpler models.

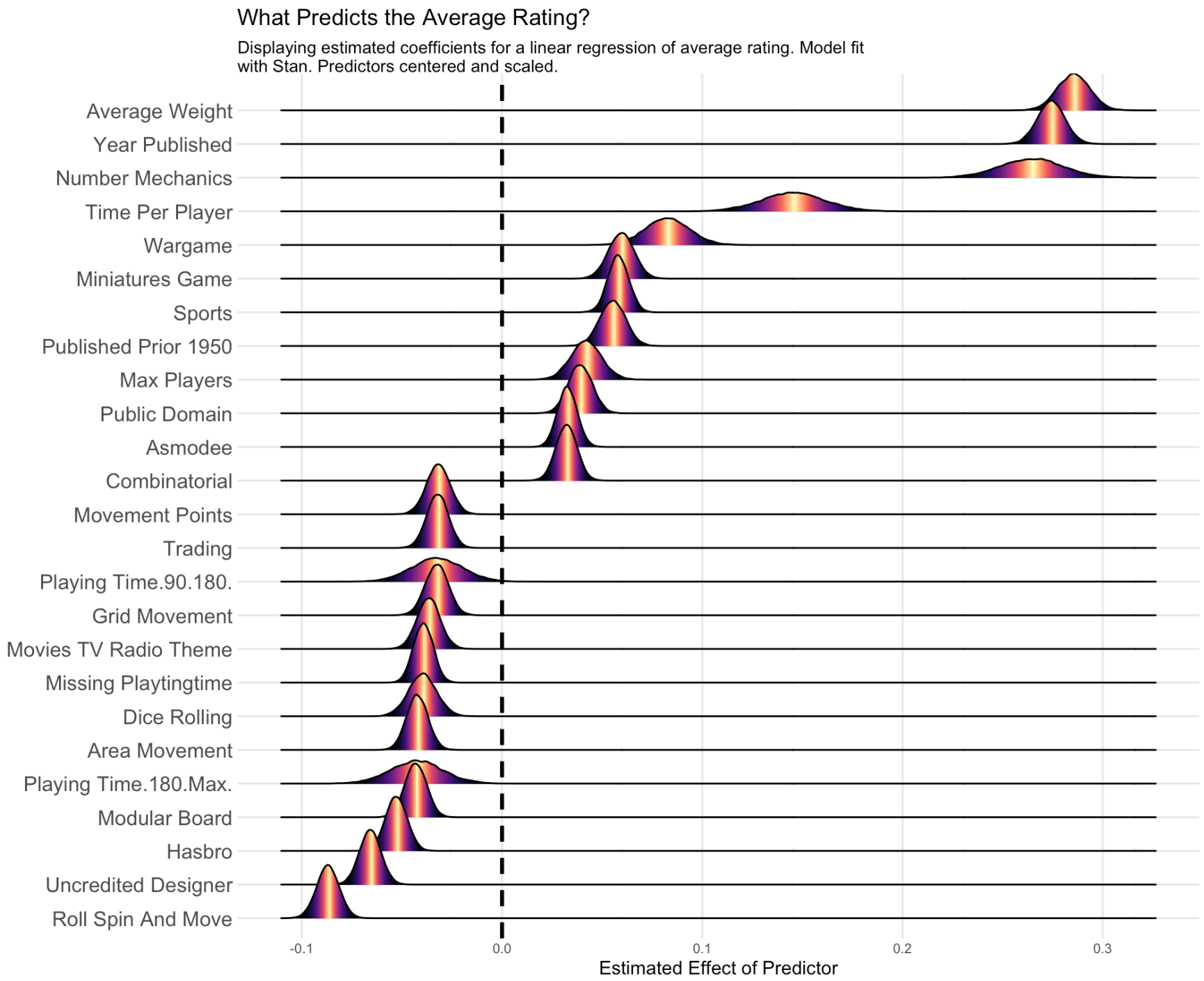
Are we done? Do we just roll with the model?

No! We need to ask, **what did the model learn**? **Why is it better?**

Why are some games rated higher than than others?



A model can only tell us what it learned; **we have to interpret and understand what this means!**



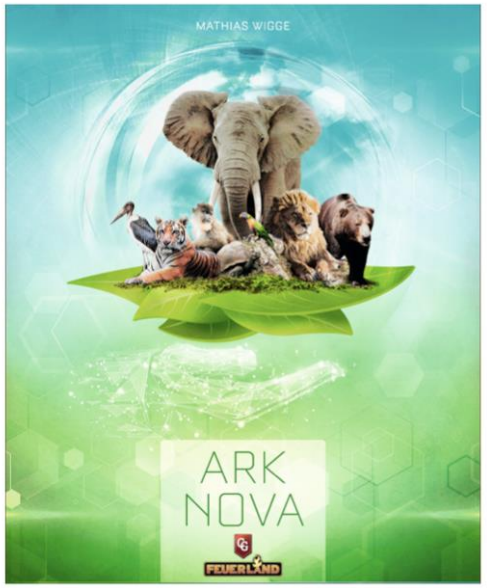
Once we have a model in place, we aren't done. We use it to predict new data and look at what it produces.

See where it does well. See where it makes mistakes.

The best games for 2021-2023, according to the model(s).

				Estimated			
Rank	Published	ID	Name	UserRatings	Weight	Average	GeekRating
1	2022	331106	The Witcher: Old World	26,800	3.60	8.48	8.32
2	2021	343905	Boonlake	8,400	4.03	8.18	7.77
3	2021	285967	Ankh: Gods of Egypt	8,600	2.81	8.13	7.74
4	2022	310873	Carnegie	4,900	3.63	8.36	7.69
5	2022	319807	Shogun no Katana	4,300	3.87	8.41	7.66
6	2021	329841	Ticket to Ride: Europe – 15th Anniversary	7,700	2.05	8.04	7.62
7	2022	331224	Zombicide: Undead or Alive	7,000	2.44	8.06	7.61
8	2021	351735	Newton & Great Discoveries	6,900	2.95	8.07	7.61
9	2022	266064	Trudvang Legends	5,200	2.54	8.12	7.54
10	2022	302892	Frozen Frontier	3,900	4.23	8.33	7.54
11	2021	344277	Corrosion	7,500	3.58	7.95	7.54
12	2023	357212	Fire for Light	2,900	2.84	8.60	7.54
13	2022	314582	Amsterdam	9,200	3.38	7.84	7.51
14	2021	340466	Unfathomable	7,200	3.24	7.91	7.49
15	2021	339906	The Hunger	17,800	2.03	7.64	7.47
16	2022	349067	The Lord of the Rings: The Card Game – Revised Core Set	5,400	2.92	8.00	7.46
17	2022	350198	Terminus	3,000	4.21	8.42	7.45
18	2021	342942	Ark Nova	3,400	3.65	8.22	7.39
19	2021	322708	Descent: Legends of the Dark	3,200	3.27	8.20	7.34
20	2022	359999	Agricola 15	3,400	3.54	8.14	7.33
21	2022	295770	Frosthaven	2,300	3.64	8.40	7.25
22	2021	332075	Warhammer Quest: Cursed City	2,500	3.33	8.29	7.25
23	2022	340520	Ronin Warrior	2,100	3.47	8.41	7.20
24	2021	299255	Vienna Connection	3,800	2.85	7.85	7.19
25	2021	262201	Sword & Sorcery: Ancient Chronicles	2,700	2.97	8.10	7.17

Some games the model has predicted pretty well.



Ark Nova
ID: 342942
Published: 2021
Player Count: 1-4
Playing Time: 150 min

1.1 Estimated Outcomes on BGG

This table displays the selected game's **current** values on four BGG outcomes (UsersRated, Average, GeekRating, Weight) along with my predictive model(s) **estimated** values for where these games are likely to end up.

Published	ID	Name	Type	UserRatings	Average	GeekRating	Weight
2021	342942	Ark Nova	Current	6,383	8.70	7.94	3.74
			Estimated	3,400	8.22	7.28	3.65

To see more information about the game on boardgamegeek, click on the game's ID or Name to go straight to the game's profile page.

Some games the model hasn't predicted so well.

Sleeping Gods
ID: 255984
Published: 2021
Player Count: 1-4
Playing Time: 1200 min

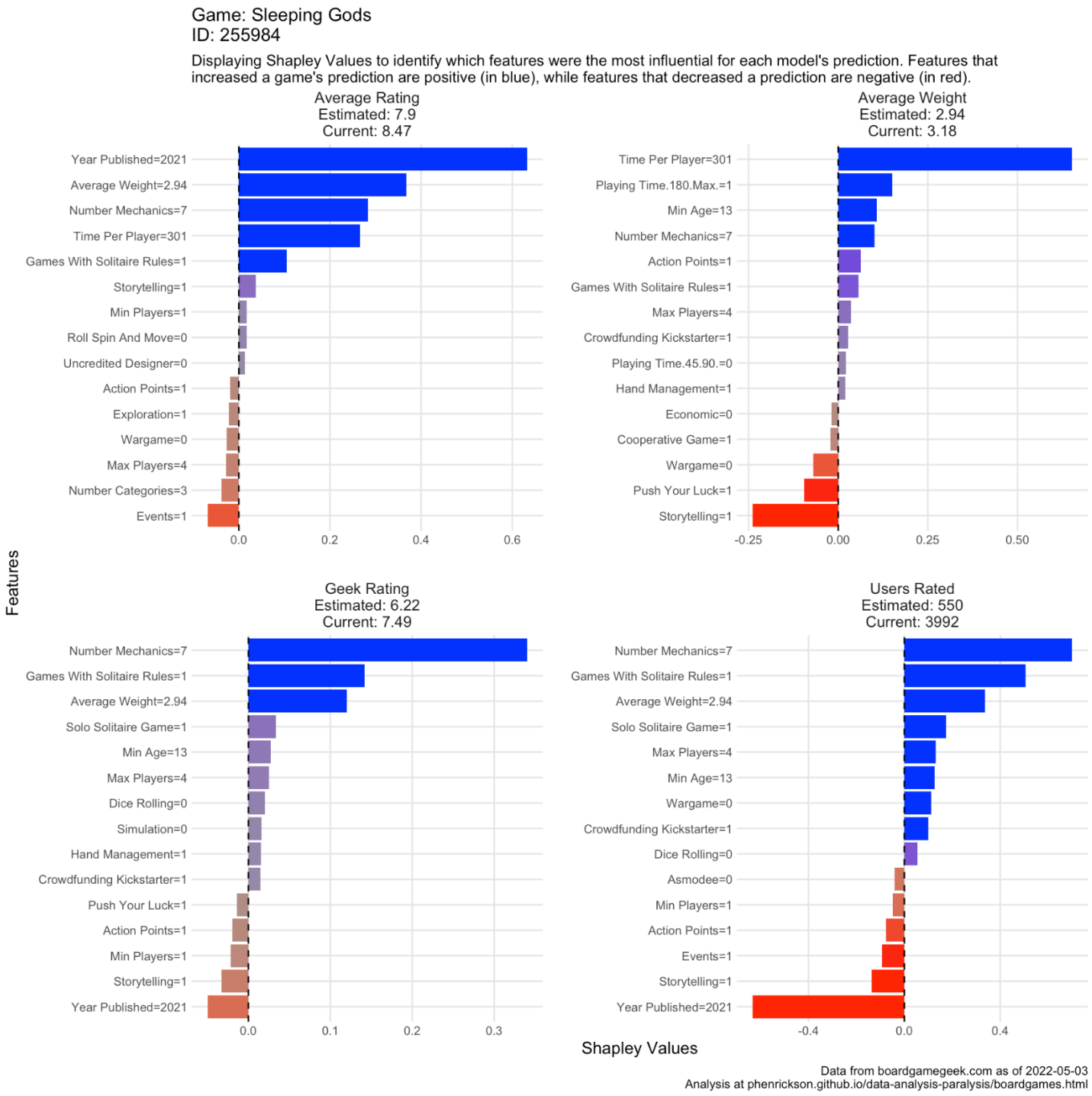
1.1 Estimated Outcomes on BGG

This table displays the selected game's **current** values on four BGG outcomes (UsersRated, Average, GeekRating, Weight) along with my predictive model(s) **estimated** values for where these games are likely to end up.

Published	ID	Name	Type	UserRatings	Average	GeekRating	Weight
2021	255984	Sleeping Gods	Current	3,968	8.47	7.49	3.17
			Estimated	600	7.90	6.10	2.94

To see more information about the game on boardgamegeek, click on the game's ID or Name to go straight to the game's profile page.

Why? We investigate the model's predictions **so that we can figure out how to improve**.

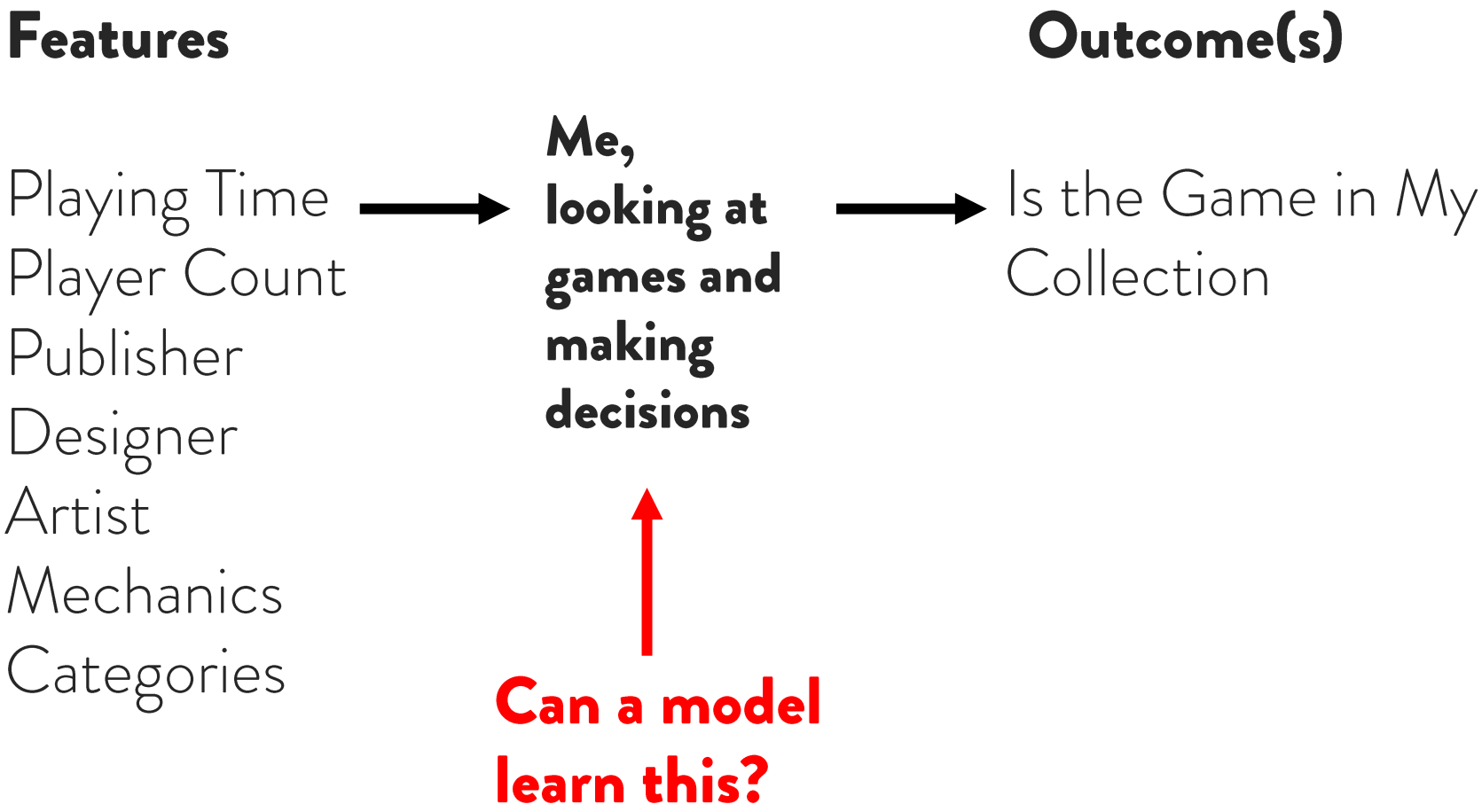


Building one model tends to generate **new ideas** for other models, or new data to collect.

At some point I wondered, could I build a model to predict games that **I like?**

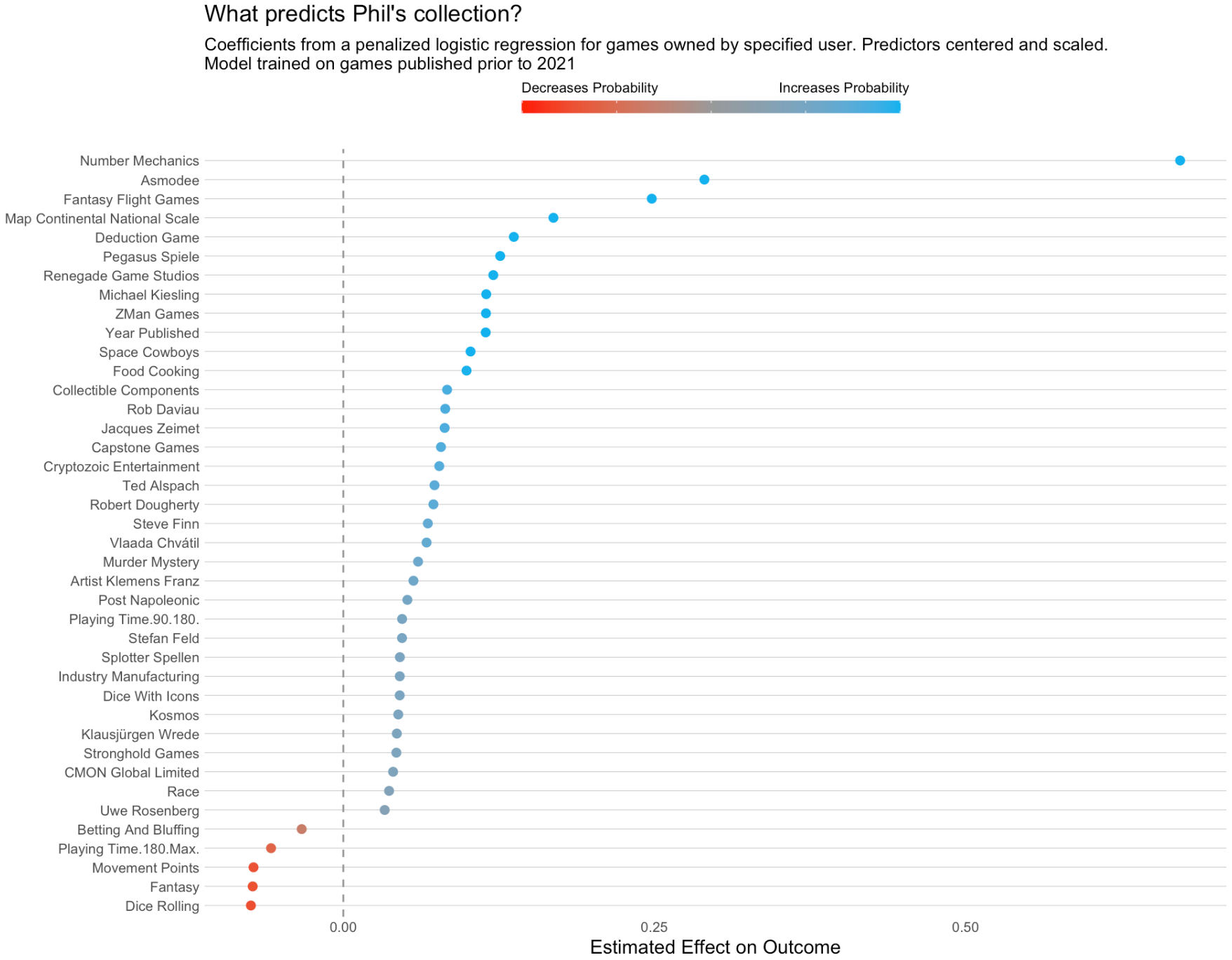
What's the data generating process?

The Modeling Process



Artificial Phil:

The model of myself and what it learned about what makes me likely to own a game

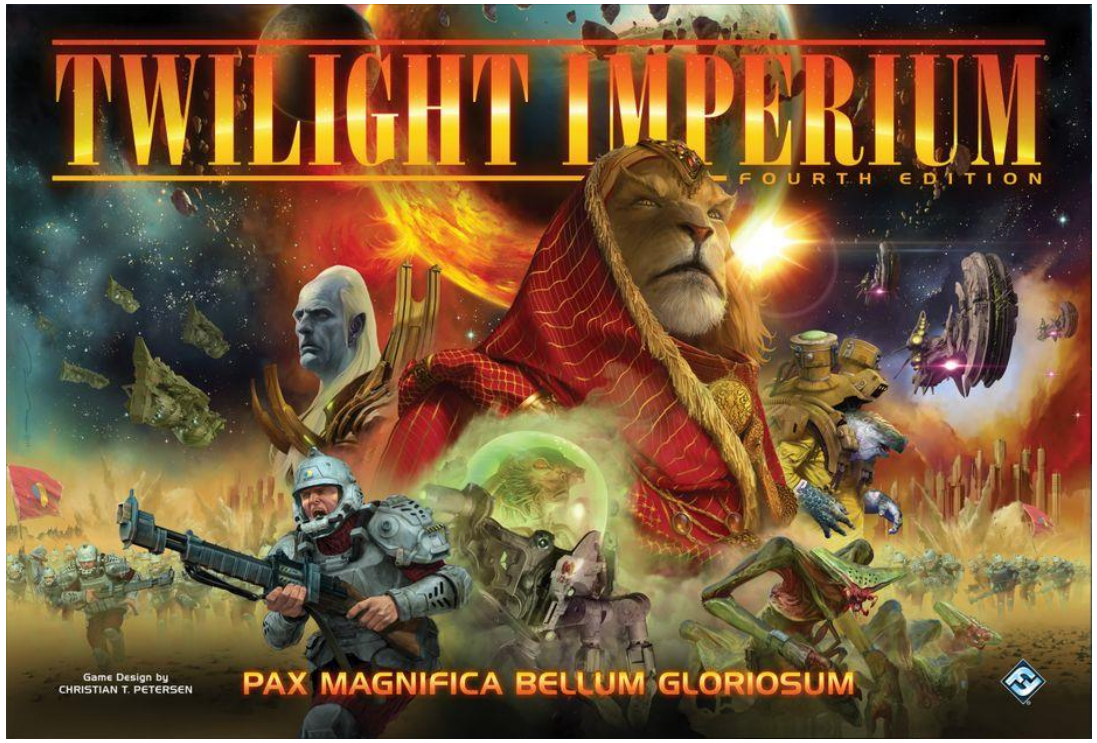


Older games Artificial Phil thought I was most likely to own

3.1 Top Games from Training Set

Displaying the 100 games from the training set with the highest probability of ownership, highlighting in blue games the user has owned.

Rank	Published	ID	Name	Pr(Owned)	Owned
1	2017	233078	Twilight Imperium: Fourth Edition	0.936	yes
2	2013	143693	Glass Road	0.912	no
3	2018	205896	Rising Sun	0.884	no
4	2016	205637	Arkham Horror: The Card Game	0.840	yes
5	2016	187645	Star Wars: Rebellion	0.819	no
6	2019	276025	Maracaibo	0.809	yes
7	2017	220308	Gaia Project	0.764	no
8	1999	552	Bus	0.760	no
9	2012	124742	Android: Netrunner	0.721	no
10	2014	148228	Splendor	0.669	no
11	2011	96848	Mage Knight Board Game	0.640	no
12	2004	9659	War of the Ring	0.615	no
13	2017	221107	Pandemic Legacy: Season 2	0.613	no
14	2016	176083	Hit Z Road	0.593	no
15	2017	174430	Gloomhaven	0.592	yes
16	2010	25292	Merchants & Marauders	0.578	yes



You and your friends bring a warring universe to life on your kitchen table

Takes roughly 8-12 hours to play

Twilight Imperium: Fourth Edition
Released in 2017
Has a Space Lion on the Cover

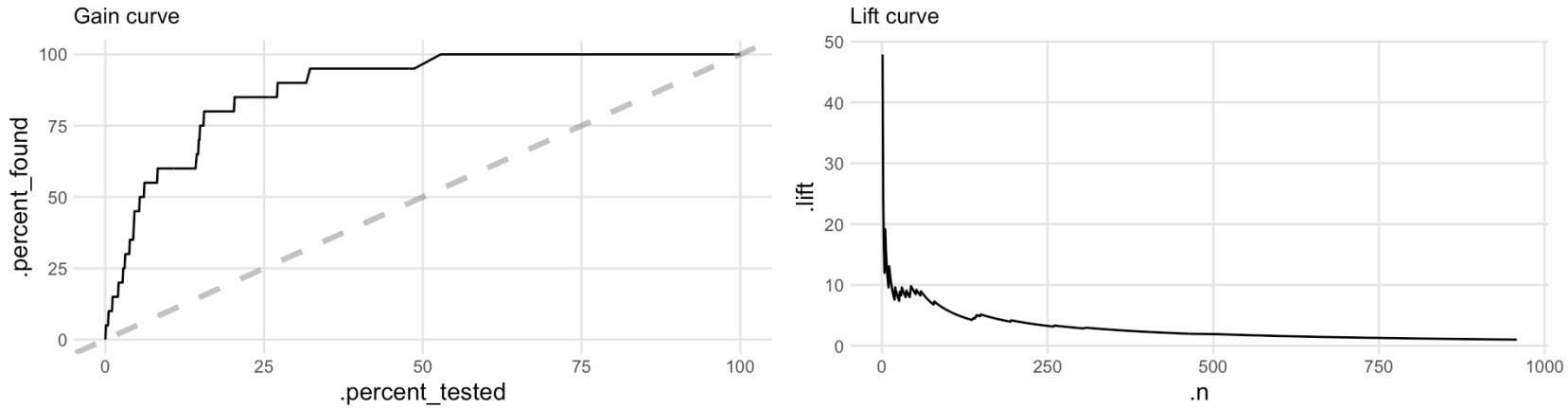
4.1 Model Assessment

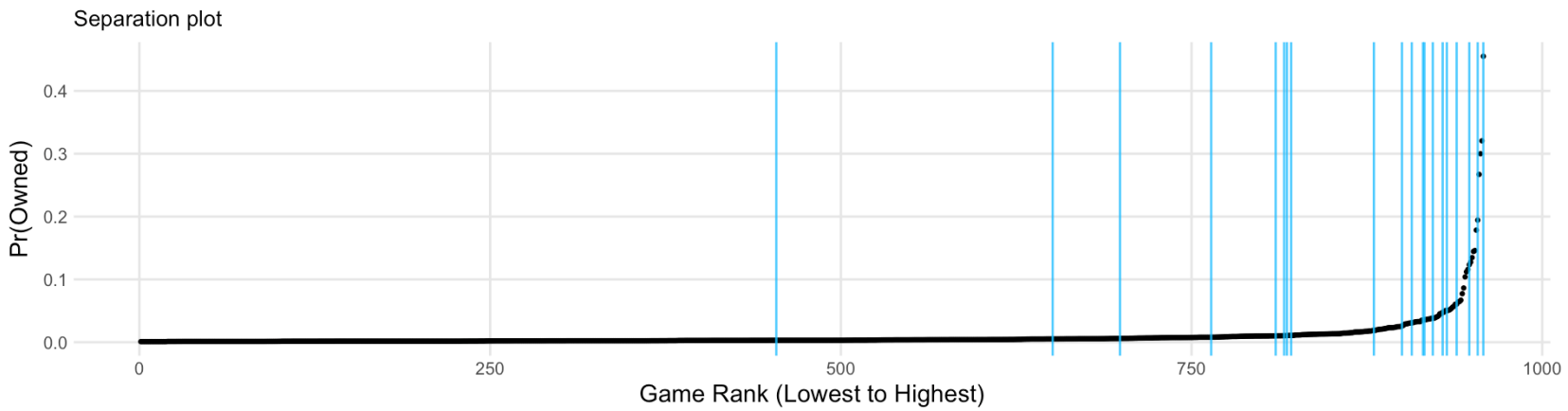
username	outcome	dataset	method	.metric	.estimate
Phil	owned	validation	GLM	roc_auc	0.893
Phil	owned	validation	Decision Tree	roc_auc	0.715

How well Artificial Phil did at predicting myself

How well did model predict the validation set?

Displaying a gain curve (left) and a lift curve (right) and separation plot (bottom) to illustrate the best performing model's performance in predicting games from 2020





New games Artificial Phil thinks I'm likely to buy

5.1 Top Upcoming Games

Examine the top 100 upcoming games, highlighting in blue ones the user already owns.

Published	ID	Name	Pr(Owned)	Owned
2022	331106	The Witcher: Old World	0.593	no
2022	349067	The Lord of the Rings: The Card Game – Revised Core Set	0.592	no
2021	343905	Boonlake	0.574	yes
2021	339484	Savannah Park	0.434	no
2021	340466	Unfathomable	0.402	yes
2022	310873	Carnegie	0.344	no
2021	285967	Ankh: Gods of Egypt	0.325	yes
2023	347909	Rogue Angels: Legacy of the Burning Suns	0.302	no
2021	256680	Return to Dark Tower	0.254	no
2021	344277	Corrosion	0.229	no
2021	342942	Ark Nova	0.224	no
2021	291859	Riftforce	0.217	yes
2022	317511	Tindaya	0.184	no
2022	295770	Frosthaven	0.152	no
2021	338980	Eastern Empires	0.139	no
2022	322524	Bardsung	0.120	no
2021	339906	The Hunger	0.106	no

We use models in order to
~~predict upcoming games so that we can buy them before other people~~
learn about the world around us.

How do we learn from data?
How do we build models?

We speculate.
We test.
We learn from our mistakes.
Science proceeds with trial and error.

Or, as someone very wise once said:

Or, as someone very wise once said:

Pass on what you have learned. Strength.
Mastery. But weakness, folly, failure also.

Or, as someone very wise once said:

Pass on what you have learned. Strength.
Mastery. But weakness, folly, failure also.

Yes, failure most of all.
The greatest teacher, failure is.

Or, as someone very wise once said:

Pass on what you have learned. Strength.
Mastery. But weakness, folly, failure also.

Yes, failure most of all.
The greatest teacher, failure is.

Yoda,
Jedi Master

Wrapping up.

So, now what?

Now What?

Model Building, Science, and Analytics

What should we do?

I'll leave you with one thought.

I'll leave you with one thought.

The answer isn't from computer science, statistics, mathematics, or the social sciences.

I'll leave you with one thought.

The answer isn't from computer science, statistics, mathematics, or the social sciences.

The answer, naturally, comes from television.

The Good Place

Mike Schur, Writer

What does it mean to be a good person?

We can run the full gamut on this, and **explore every possible theory** about how to be a good person, **and it starts to get exhausting**.

It is asking too much of people to become monks and shed all of their earthly possessions.

Mike Schur, Writer

You know what's important? **If you're trying**.

If you're just trying to be a good person, **if that's at the front of your brain all the time**, if you're asking yourself, am I doing okay, am I generally doing things that are good or bad, could I be improving somehow?

If you're just asking the questions, that's kind of the key.

Mike Schur, Writer

Try to be a little bit better today than you were yesterday.

Mike Schur, Writer

It's more important to be a good person than a good scientist.

But, we can easily amend this for the journey into data science.

Mike Schur, Writer

Phil Henrickson, Stealer of Quotes

What does it mean to be a good scientist?

We can run the full gamut on this, and **explore every possible theory** about how to be a good scientist, **and it starts to get exhausting**.

It is asking too much of people to... **become grad students and shed all of their earthly possessions**.

Phil Henrickson, Stealer of Quotes

You know what's important? **If you're trying**.

If you're just trying to learn about the world around you, **if that's at the front of your brain all the time**, if you're asking yourself, am I doing okay, am I generally doing things that are good or bad, could I be improving somehow?

If you're just asking the questions, that's kind of the key.

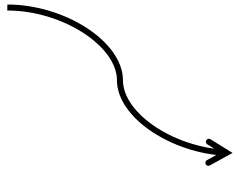
Phil Henrickson, Stealer of Quotes

Try to be a little bit better today than you were yesterday.

Phil Henrickson, Stealer of Quotes

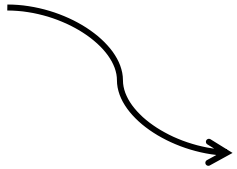
Try to be a little bit better today than you were yesterday.

Try to be a little bit better today than
you were yesterday.



This is the secret to building models.

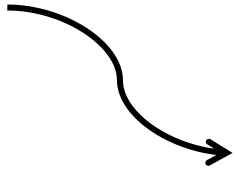
Try to be a little bit better today than
you were yesterday.



This is the secret to building models.

This is the secret to good science.

Try to be a little bit better today than
you were yesterday.



This is the secret to building models.

This is the secret to good science.

This is the answer to the now what of analytics.

Thanks for listening.

Questions?

Appendix

Prerequisites for a data science project

Two Things You Need for a Data Science Project

1) There is some unknown pattern that would be useful to learn.

2) There is data to learn that pattern.

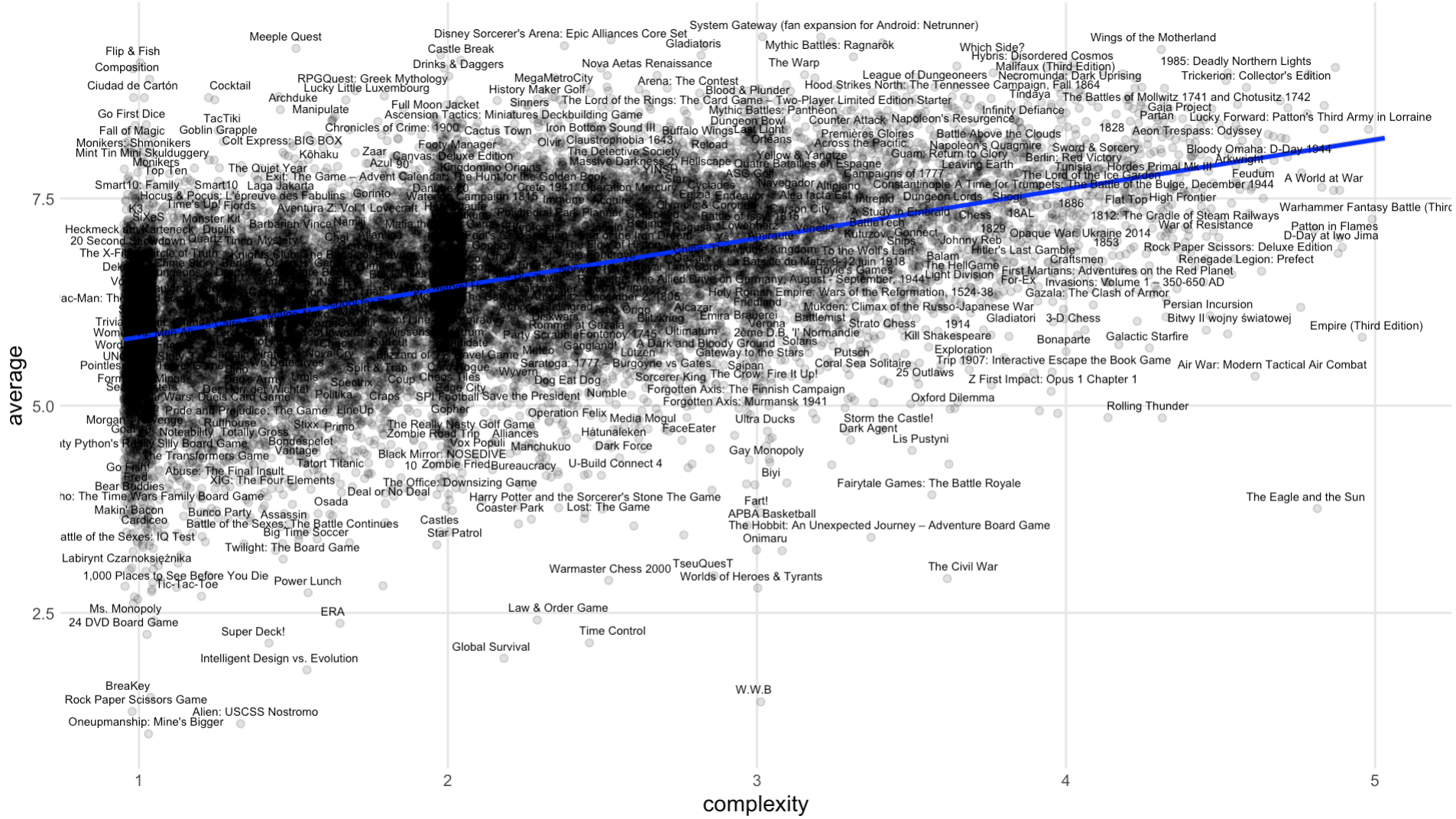
What games should people check out?

2.4 Examining the Top Complexity-Adjusted Games

We can put this all together to now look at games that are rated highly after adjusting for the effect of complexity. We'll look at the top 250 to keep it simple.

published	game_id	name	geekrank	adjustedrank	rankdiff	geekrating	adjustedrating	complexity
1876	521	Crokinole	60	1	59	7.69	8.22	1.25
2015	161936	Pandemic Legacy: Season 1	2	2	0	8.44	8.06	2.83
2018	254640	Just One	145	3	142	7.40	8.04	1.05
2015	178900	Codenames	107	4	103	7.50	7.99	1.28
2015	173346	7 Wonders Duel	18	5	13	7.98	7.95	2.23
2020	318977	MicroMacro: Crime City	212	6	206	7.28	7.92	1.13
2017	230802	Azul	63	7	56	7.68	7.90	1.77
2014	160069	Ticket to Ride: 10th Anniversary	99	8	91	7.53	7.89	1.90
2019	284083	The Crew: The Quest for Planet Nine	47	9	38	7.77	7.88	1.99
2014	165722	KLASK	248	10	238	7.21	7.87	1.08
2019	266507	Clank!: Legacy – Acquisitions Incorporated	29	11	18	7.88	7.85	2.69
2015	156546	Monikers	317	12	305	7.12	7.83	1.06
2018	244521	The Quacks of Quedlinburg	59	13	46	7.69	7.82	1.95
2014	163412	Patchwork	100	14	86	7.52	7.82	1.62
2019	266192	Wingspan	23	15	8	7.94	7.81	2.44
2011	92828	Dixit: Odyssey	242	16	226	7.22	7.80	1.18
2009	46213	Telestrations	281	17	264	7.16	7.79	1.08
2018	225694	Decrypto	104	19	85	7.52	7.78	1.79
2017	224037	Codenames: Duet	189	18	171	7.32	7.78	1.34
2016	192291	Sushi Go Party!	216	20	196	7.28	7.77	1.31
2009	54043	Jaipur	153	21	132	7.39	7.76	1.48
2016	204583	Kingdomino	239	22	217	7.22	7.76	1.22
2012	129622	Love Letter	297	23	274	7.14	7.68	1.19

The boardgamegeek community really tends to like complex games; complexity is highly correlated with the average rating

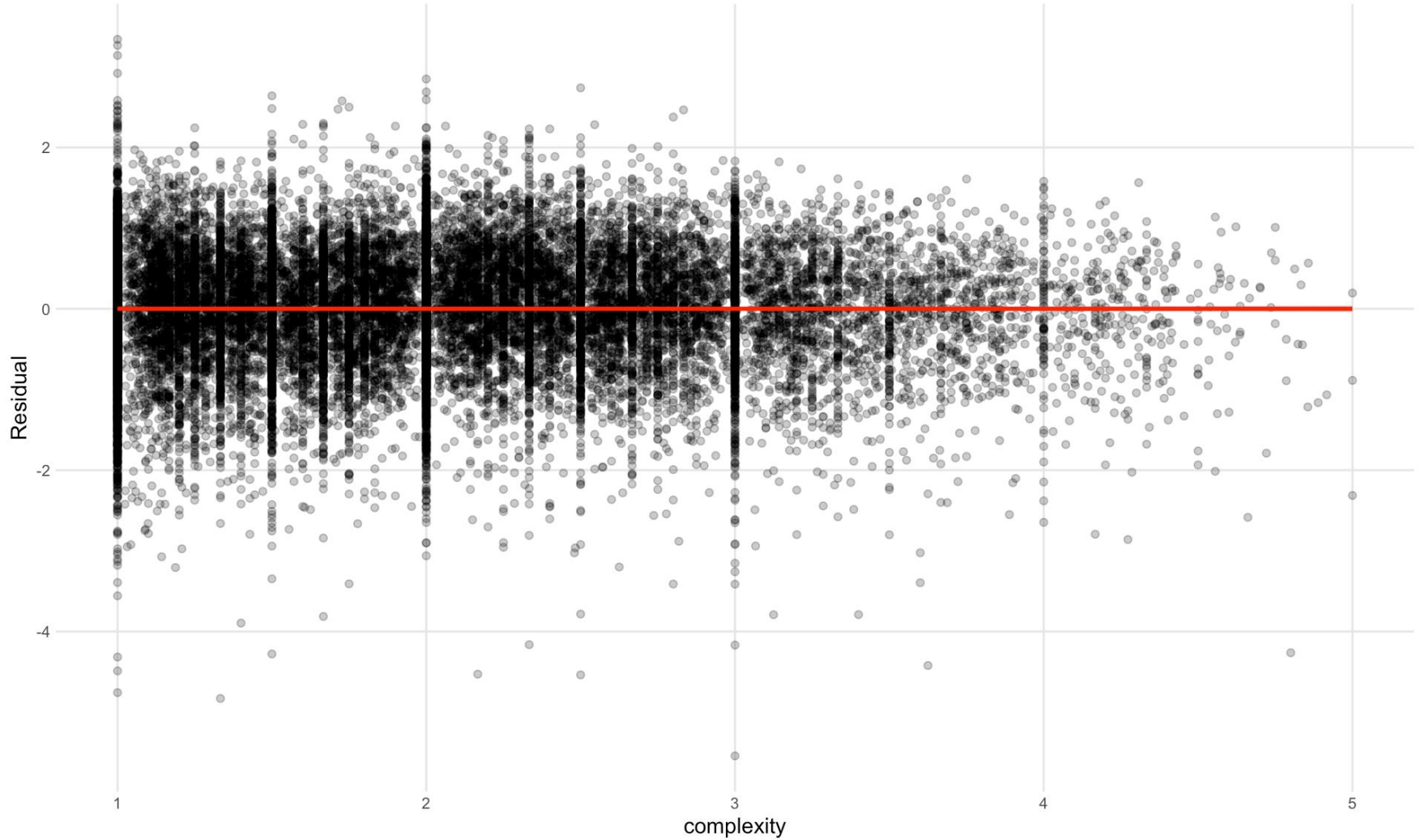


Data from boardgamegeek.com as of 2022-04-17
Data and analysis at phenrickson.github.io/data-analysis-paralysis/boardgames.html

This means if you look online, the “best” games tend also be very complex and difficult to learn

yearpublished	game_id	name	rank	bayesaverage	complexity
2017	174430	Gloomhaven	1	8.49	3.87
2015	161936	Pandemic Legacy: Season 1	2	8.44	2.83
2018	224517	Brass: Birmingham	3	8.42	3.90
2016	167791	Terraforming Mars	4	8.27	3.25
2020	291457	Gloomhaven: Jaws of the Lion	5	8.27	3.59
2017	233078	Twilight Imperium: Fourth Edition	6	8.25	4.26
2017	220308	Gaia Project	7	8.17	4.37
2016	187645	Star Wars: Rebellion	8	8.17	3.73
2017	162886	Spirit Island	9	8.14	4.04
2011	115746	War of the Ring: Second Edition	10	8.14	4.18
2015	182028	Through the Ages: A New Story of Civilization	11	8.14	4.42
2016	193738	Great Western Trail	12	8.12	3.71
2005	12333	Twilight Struggle	13	8.11	3.60
2016	169786	Scythe	14	8.06	3.43
2020	316554	Dune: Imperium	15	8.03	2.99
2011	84876	The Castles of Burgundy	16	8.01	3.00
2018	167355	Nemesis	17	7.99	3.39
2015	173346	7 Wonders Duel	18	7.98	2.23
2013	124361	Concordia	19	7.97	3.01
2007	28720	Brass: Lancashire	20	7.97	3.86
2012	120677	Terra Mystica	21	7.96	3.97
2016	177736	A Feast for Odin	22	7.95	3.85
2019	266192	Wingspan	23	7.94	2.44
2016	205637	Arkham Horror: The Card Game	24	7.93	3.49
2018	237182	Root	25	7.89	3.73

Solution: complexity-adjusted ratings.
Adjust for the effect of complexity by running a simple regression and taking the residuals



2.4 Examining the Top Complexity-Adjusted Games

We can put this all together to now look at games that are rated highly after adjusting for the effect of complexity. We'll look at the top 250 to keep it simple.

published	game_id	name	geekrank	adjustedrank	rankdiff	geekrating	adjustedrating	complexity
1876	521	Crokinole	60	1	59	7.69	8.22	1.25
2015	161936	Pandemic Legacy: Season 1	2	2	0	8.44	8.06	2.83
2018	254640	Just One	145	3	142	7.40	8.04	1.05
2015	178900	Codenames	107	4	103	7.50	7.99	1.28
2015	173346	7 Wonders Duel	18	5	13	7.98	7.95	2.23
2020	318977	MicroMacro: Crime City	212	6	206	7.28	7.92	1.13
2017	230802	Azul	63	7	56	7.68	7.90	1.77
2014	160069	Ticket to Ride: 10th Anniversary	99	8	91	7.53	7.89	1.90
2019	284083	The Crew: The Quest for Planet Nine	47	9	38	7.77	7.88	1.99
2014	165722	KLASK	248	10	238	7.21	7.87	1.08
2019	266507	Clank!: Legacy – Acquisitions Incorporated	29	11	18	7.88	7.85	2.69
2015	156546	Monikers	317	12	305	7.12	7.83	1.06
2018	244521	The Quacks of Quedlinburg	59	13	46	7.69	7.82	1.95
2014	163412	Patchwork	100	14	86	7.52	7.82	1.62
2019	266192	Wingspan	23	15	8	7.94	7.81	2.44
2011	92828	Dixit: Odyssey	242	16	226	7.22	7.80	1.18
2009	46213	Telestrations	281	17	264	7.16	7.79	1.08
2018	225694	Decrypto	104	19	85	7.52	7.78	1.79
2017	224037	Codenames: Duet	189	18	171	7.32	7.78	1.34
2016	192291	Sushi Go Party!	216	20	196	7.28	7.77	1.31
2009	54043	Jaipur	153	21	132	7.39	7.76	1.48
2016	204583	Kingdomino	239	22	217	7.22	7.76	1.22
2012	129622	Love Letter	297	23	274	7.14	7.68	1.19

2.4 Examining the Top Complexity-Adjusted Games

We can put this all together to now look at games that are rated highly after adjusting for the effect of complexity. We'll look at the top 250 to keep it simple.

published	game_id	name	geekrank	adjustedrank	rankdiff	geekrating	adjustedrating	complexity
1876	521	Crokinole	60	1	59	7.69	8.22	1.25
2015	161936	Pandemic Legacy: Season 1	2	2	0	8.44	8.06	2.83
2018	254640	Just One	145	3	142	7.40	8.04	1.05
2015	178900	Codenames	107	4	103	7.50	7.99	1.28
2015	173346	7 Wonders Duel	18	5	13	7.98	7.95	2.23
2020	318977	MicroMacro: Crime City	212	6	206	7.28	7.92	1.13
2017	230802	Azul	63	7	56	7.68	7.90	1.77
2014	160069	Ticket to Ride: 10th Anniversary	99	8	91	7.53	7.89	1.90
2019	284083	The Crew: The Quest for Planet Nine	47	9	38	7.77	7.88	1.99
2014	165722	KLASK	248	10	238	7.21	7.87	1.08
2019	266507	Clank!: Legacy – Acquisitions Incorporated	29	11	18	7.88	7.85	2.69
2015	156546	Monikers	317	12	305	7.12	7.83	1.06
2018	244521	The Quacks of Quedlinburg	59	13	46	7.69	7.82	1.95
2014	163412	Patchwork	100	14	86	7.52	7.82	1.62
2019	266192	Wingspan	23	15	8	7.94	7.81	2.44
2011	92828	Dixit: Odyssey	242	16	226	7.22	7.80	1.18
2009	46213	Telestrations	281	17	264	7.16	7.79	1.08
2018	225694	Decrypto	104	19	85	7.52	7.78	1.79
2017	224037	Codenames: Duet	189	18	171	7.32	7.78	1.34
2016	192291	Sushi Go Party!	216	20	196	7.28	7.77	1.31
2009	54043	Jaipur	153	21	132	7.39	7.76	1.48
2016	204583	Kingdomino	239	22	217	7.22	7.76	1.22
2012	129622	Love Letter	297	23	274	7.14	7.68	1.19

2 Finding Similar Games

What games are similar to **Frozen Frontier**? I compare the selected game to every other game on boardgamegeek based on its mechanics, playing time, complexity, and game type. The tables below display the games which had the most similar data to **Frozen Frontier**. The reported similarity is the cosine similarity between the selected game and all other games published before 2021.

This similarity score ranges from -1 to 1, where similarity of 1 indicates that games are exactly identical (a game will have a score of 1 with itself).

Note: This analysis does not look at a game's ratings on BGG, publisher, artist, or designer - it focuses on finding similar games based on how they play, not how they are rated or by their theme.

2.1 Most Similar Games

For any new game, finding existing games that it resembles

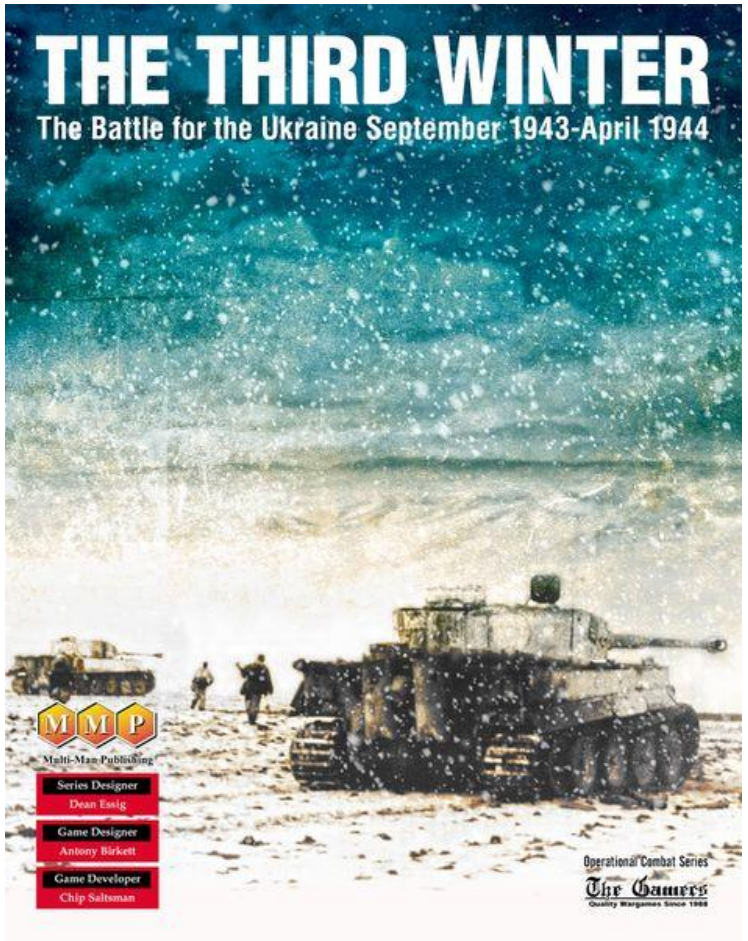
The table below displays games with the most similar data to Frozen Frontier. Click on a game's ID or Name to go straight to its profile on boardgamegeek.

Rank	Published	ID	Name	**Similarity**	GeekRating	Average	Weight	UserRatings
--	2022	302892	Frozen Frontier	**1.00**	0.00	8.35	4.23	13
1	2019	210296	DinoGenics	**0.91**	6.63	7.68	2.90	2,009
2	2020	300322	Hallertau	**0.90**	7.16	8.05	3.31	3,392
3	2012	72225	CO_2	**0.90**	6.60	7.13	3.87	3,791
4	2020	306481	Tawantinsuyu: The Inca Empire	**0.90**	6.52	7.57	4.07	1,738
5	2018	214887	CO_2: Second Chance	**0.89**	6.75	7.61	4.08	2,723
6	2018	229853	Teotihuacan: City of Gods	**0.89**	7.64	7.91	3.77	15,994
7	2020	271055	Dwellings of Eldervale	**0.89**	7.27	8.26	3.23	3,532
8	2020	184267	On Mars	**0.89**	7.72	8.25	4.65	8,527
9	2019	251247	Barrage	**0.88**	7.80	8.23	4.07	10,545
10	2019	282414	Pharaon	**0.87**	6.23	7.43	2.74	1,120
11	2007	31260	Agricola	**0.87**	7.80	7.92	3.64	67,211
12	2013	143693	Glass Road	**0.87**	7.15	7.45	2.97	10,039
13	2017	161533	Lisboa	**0.87**	7.69	8.20	4.57	8,307
14	2019	230244	Black Angel	**0.87**	6.75	7.35	3.85	3,862
15	2018	199792	Everdell	**0.87**	7.88	8.09	2.81	32,930
16	2011	97915	Bios: Megafauna	**0.87**	5.93	6.90	3.69	789

On the importance of looking at your model's predictions

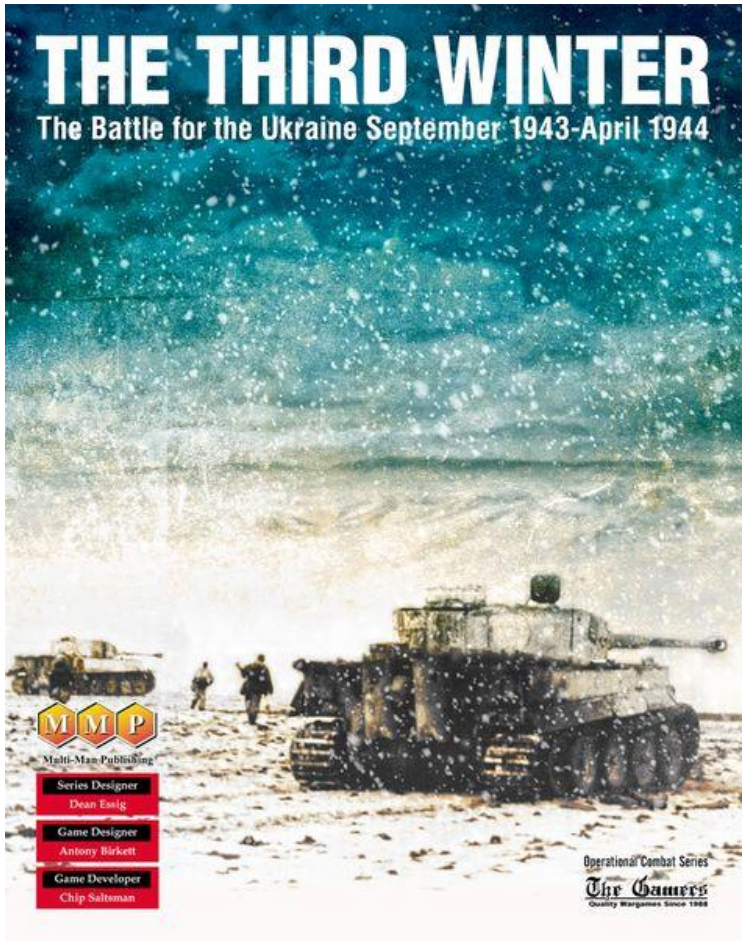
Here are the games the model predicted to be the highest rated:

Here are the games the model predicted to be the highest rated:

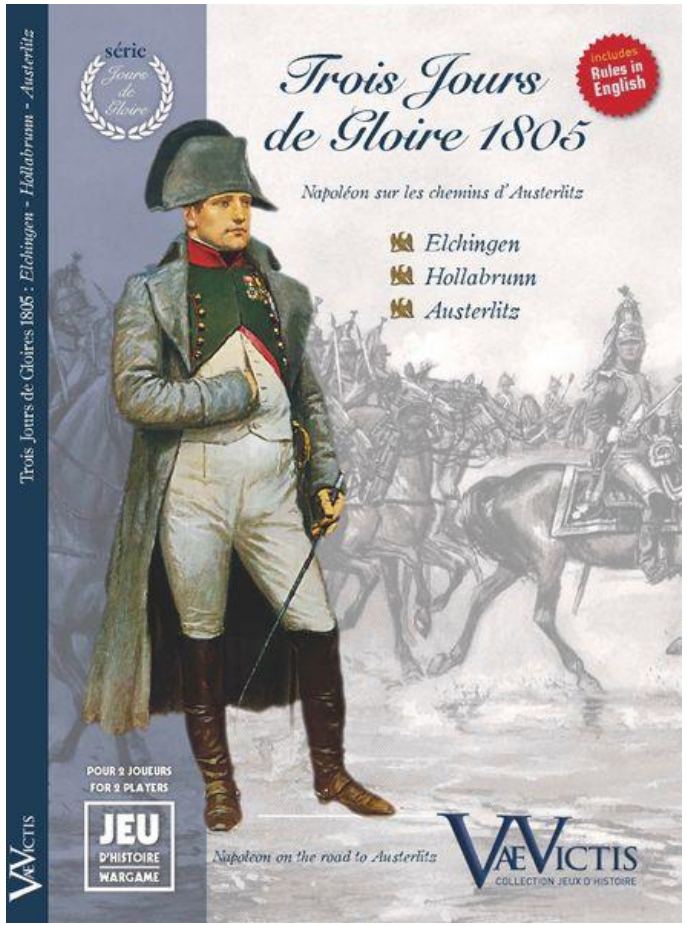


Predicted: 8.7
Actual: 9.0

Here are the games the model predicted to be the highest rated:

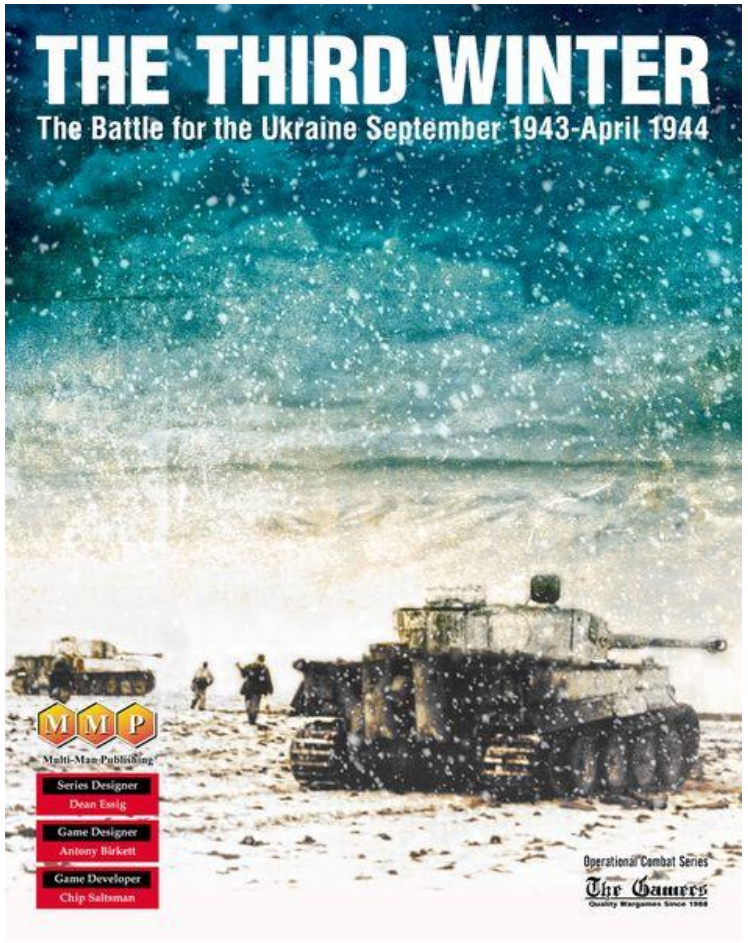


Predicted: 8.7
Actual: 9.0

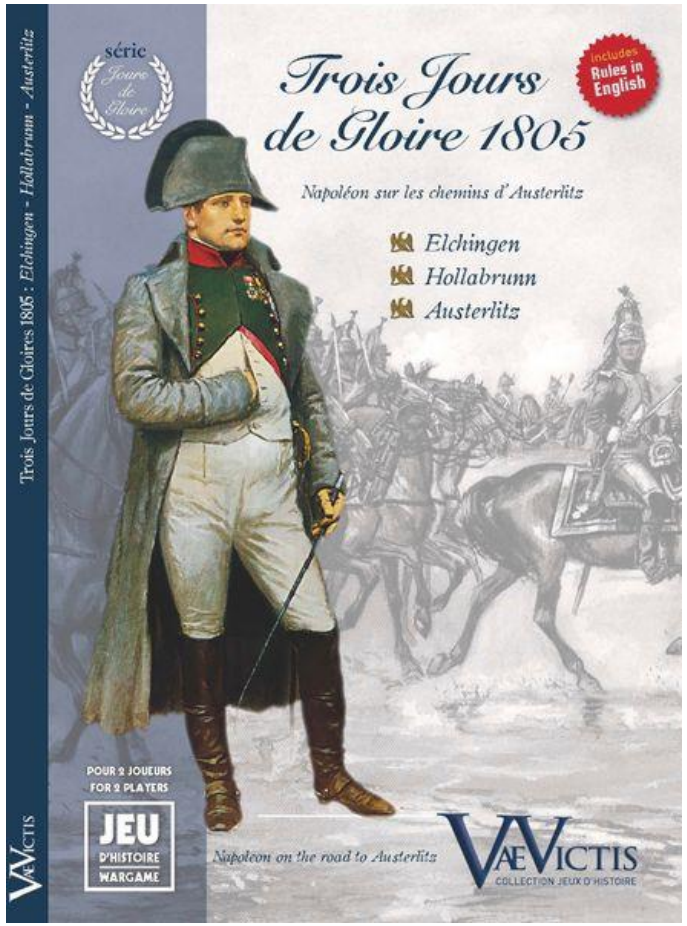


Predicted: 8.6
Actual: 8.3

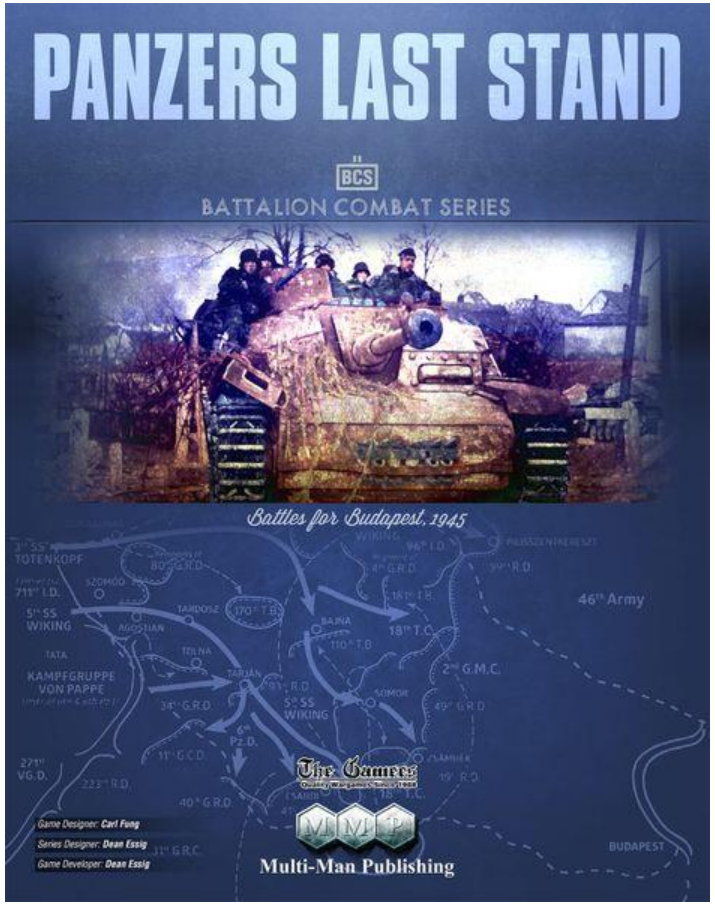
Here are the games the model predicted to be the highest rated:



Predicted: 8.7
Actual: 9.0



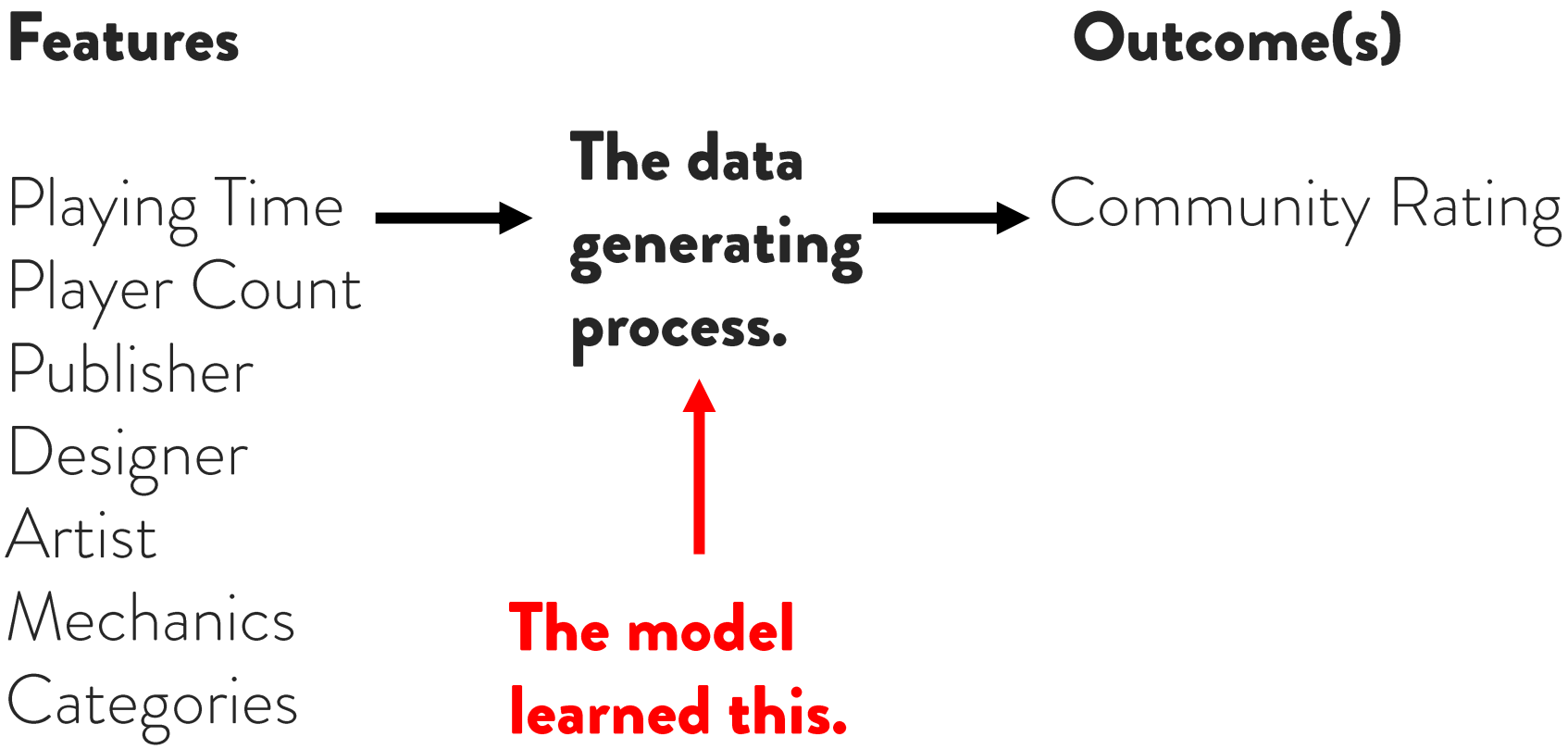
Predicted: 8.6
Actual: 8.3



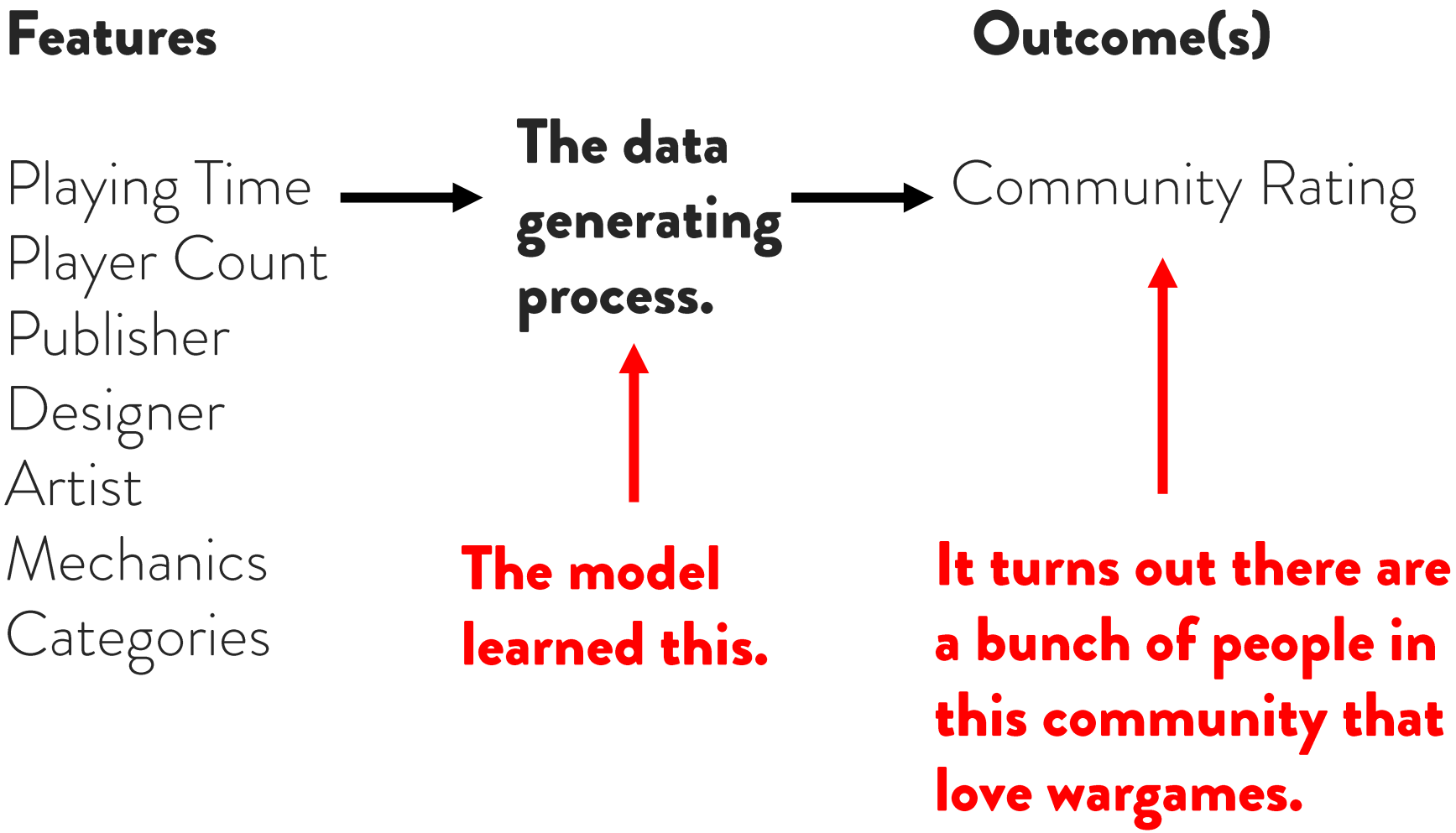
Predicted: 8.5
Actual: 9.0

...why does the model seem to be obsessed with games about war?

The Modeling Process



The Modeling Process



What data should we collect?

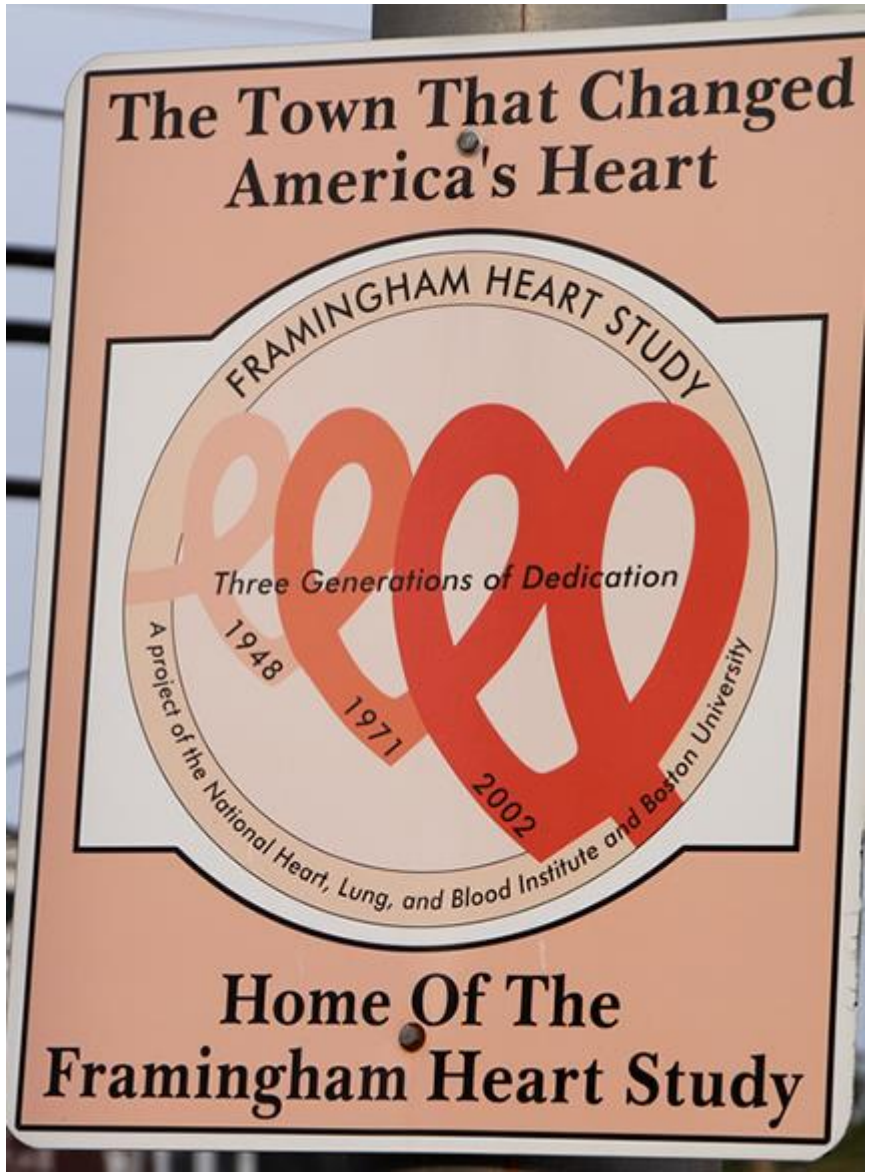
Data does not, by itself, lead to learning.

Data is the means of **testing the implications of our models**.

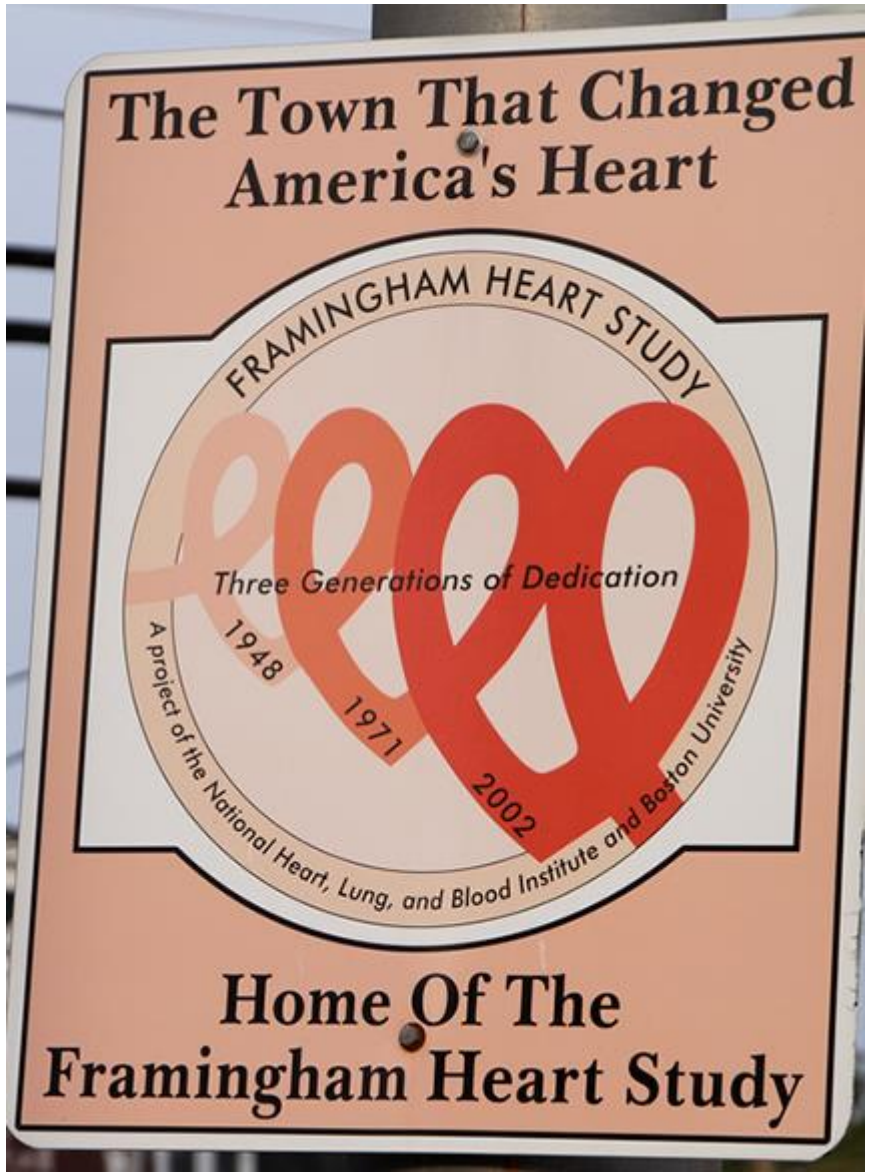
This means, in many cases, **the data we collect depends on our speculations**, what we are trying to test.

Example:

How Do We Estimate Someone's Risk of Heart Failure? The Framingham Heart Study



Source: Mahmood SS, Levy D, Vasan RS, Wang TJ. The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective.

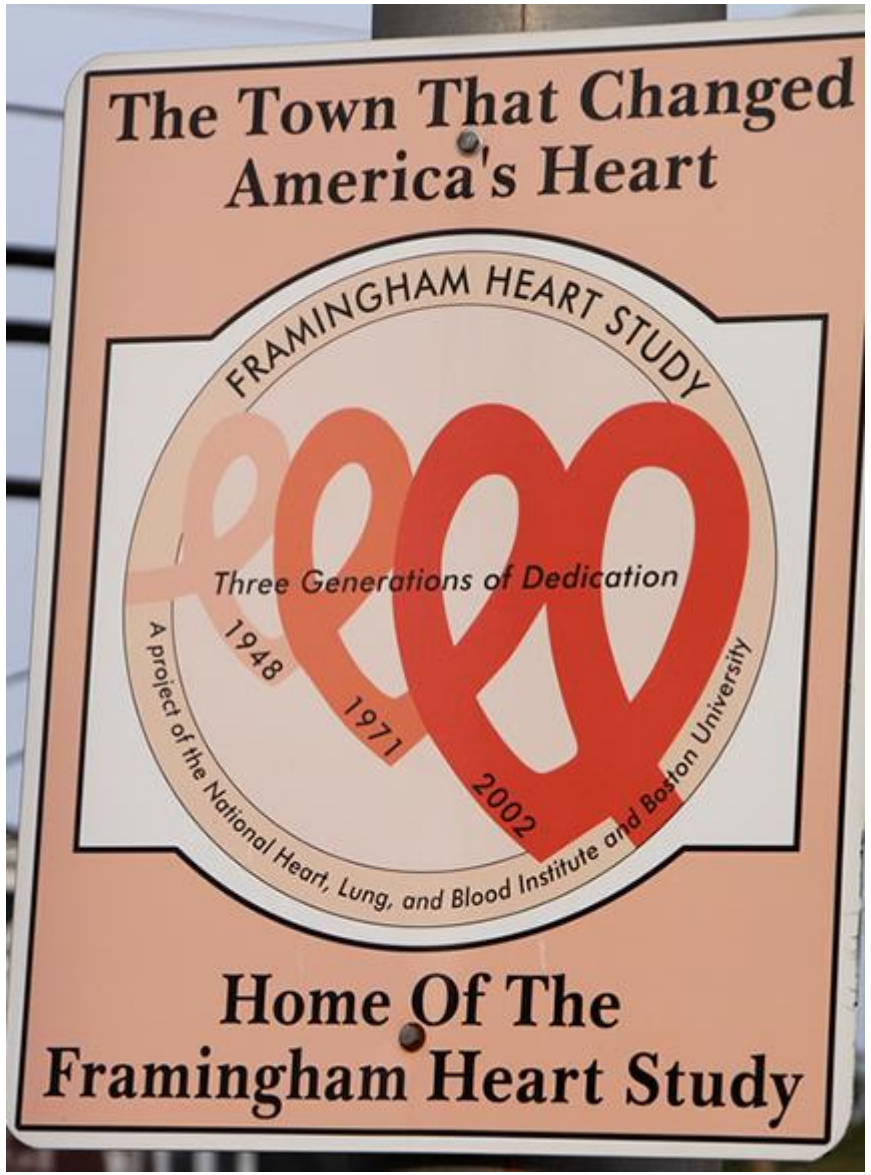


By the 1940s, **cardiovascular disease had become the number one cause of mortality** among Americans, accounting for 1 in 2 deaths.

Prevention and treatment were so poorly understood that most Americans accepted early death from heart disease as unavoidable.

"On June 16, 1948, President Harry Truman signed into law the 'National Heart Act.' The law allocated a $500,000 seed grant for **a twenty-year epidemiological heart study**.

Source: Mahmood SS, Levy D, Vasan RS, Wang TJ. The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective.

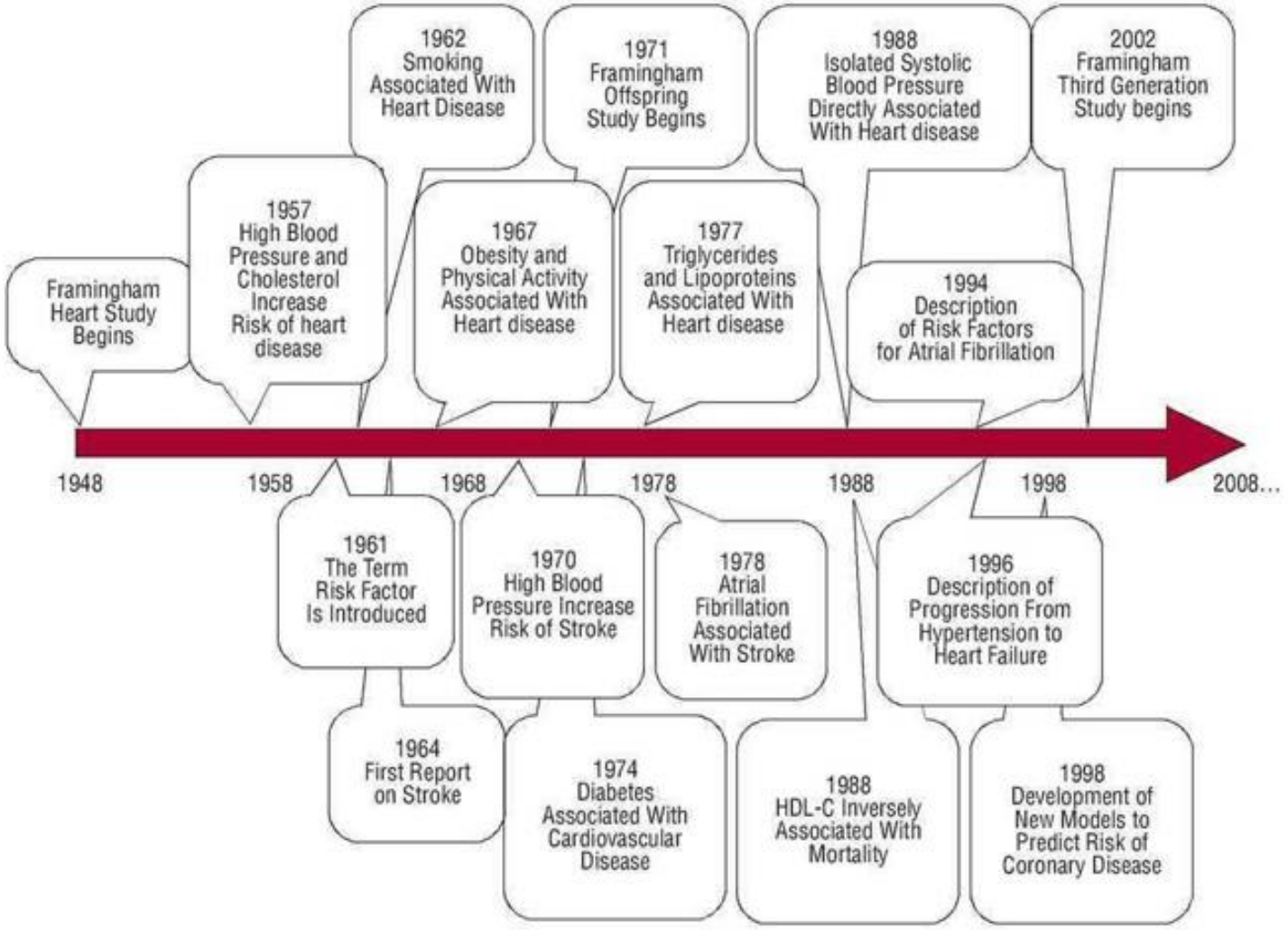


[Framingham] was a factory town of 28,000 middle-class residents of predominantly European origin… and was **therefore considered to be representative of the United States in the 1940s**.

The original cohort was recruited between 1948 and 1952 and consisted of **5209 residents aged 28 to 62 years**."

The first major study findings were published in 1957, almost a decade after the initial participant was examined… **They found a nearly 4 fold increase in coronary heart disease incidence per 1000 persons among hypertensive participants**.

Source: Mahmood SS, Levy D, Vasan RS, Wang TJ. The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective.



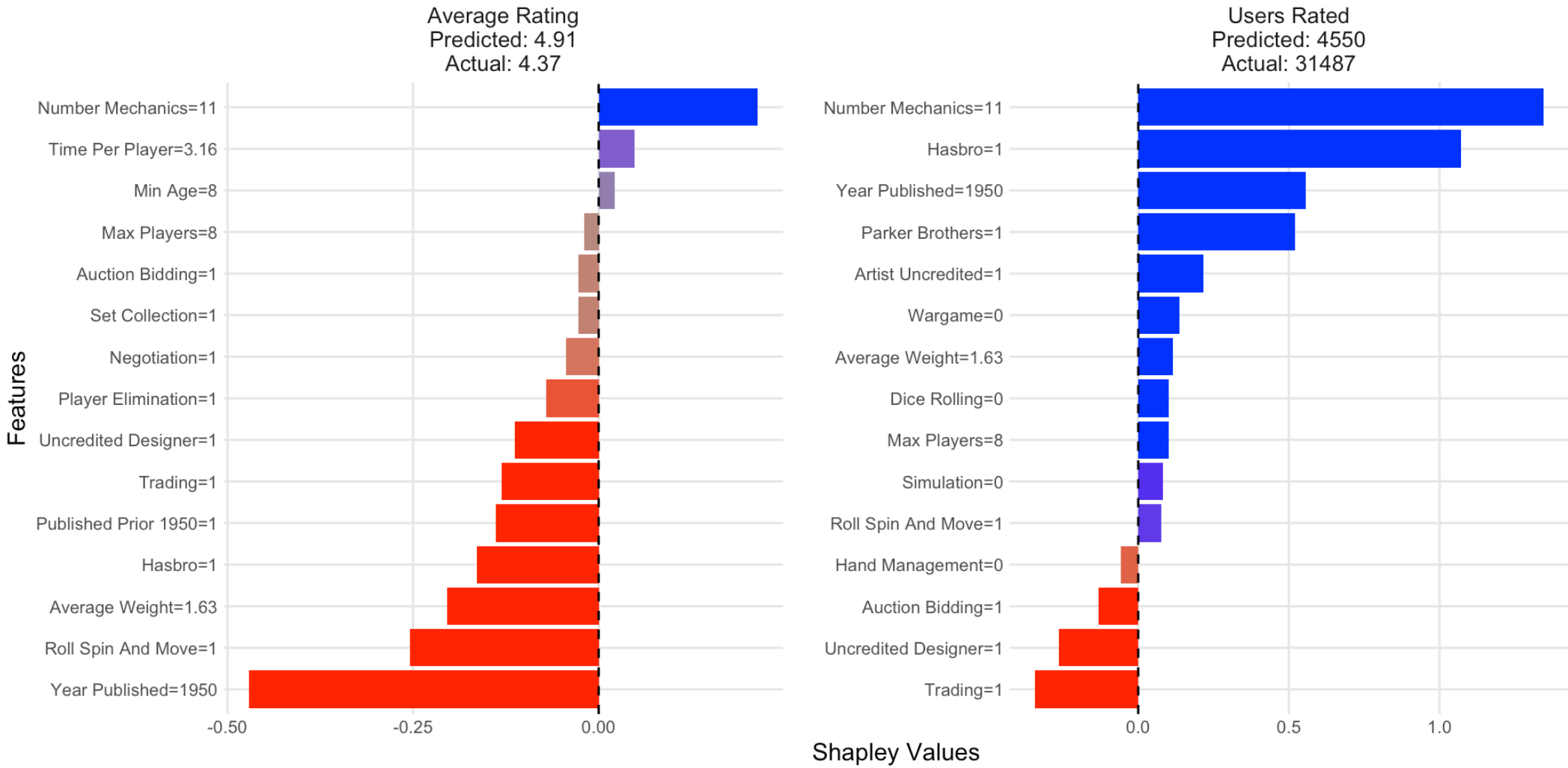
The data has been used in over 3000 scientific papers and is the basis for much of what we know about the long term risk of cardiovascular disease.

The study is now on its third generation of Framingham residents.

Why is Monopoly a Bad Game?

Game: Monopoly
ID: 1406

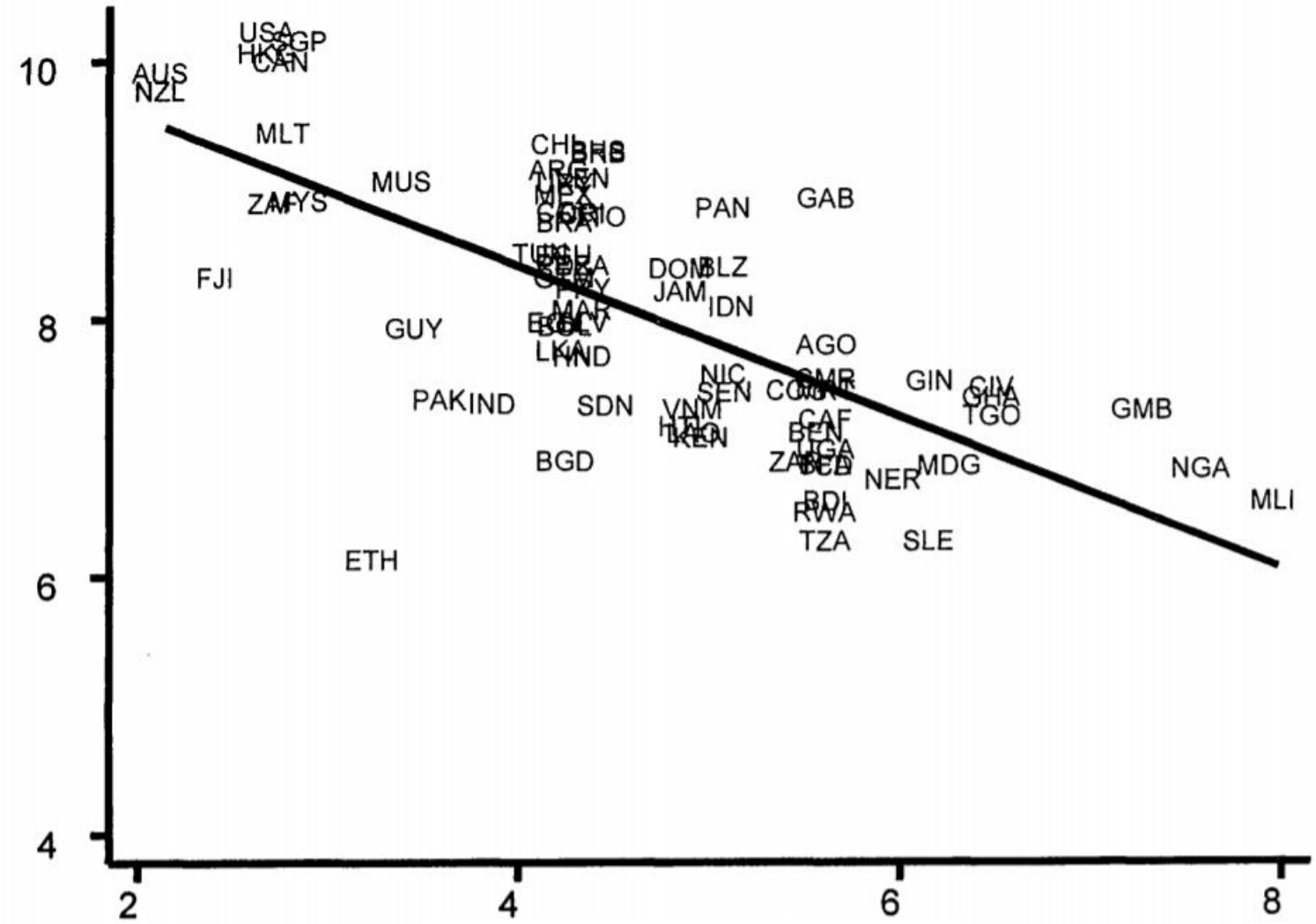
Displaying Shapley Values to identify which features were the most influential for each model's prediction. Features that increased a game's prediction are positive (in blue), while features that decreased a prediction are negative (in red).



Data from boardgamegeek.com as of 2022-04-17
Analysis at phenrickson.github.io/data-analysis-paralysis/boardgames.html

“Won’t the data tell us everything we need to know?”

GDP per capita (logged) in 1995



European Mortality Rates in 1500

Data from:

Acemoglu, Johnson, and Robinson,
and **The Colonial Origins of Comparative Development**

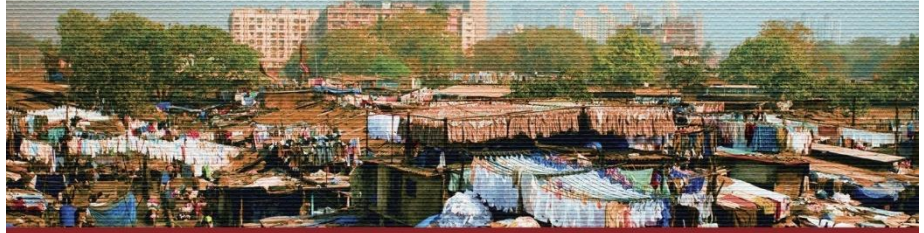
A precursor to...

A *NEW YORK TIMES* AND *WALL STREET JOURNAL* BESTSELLER

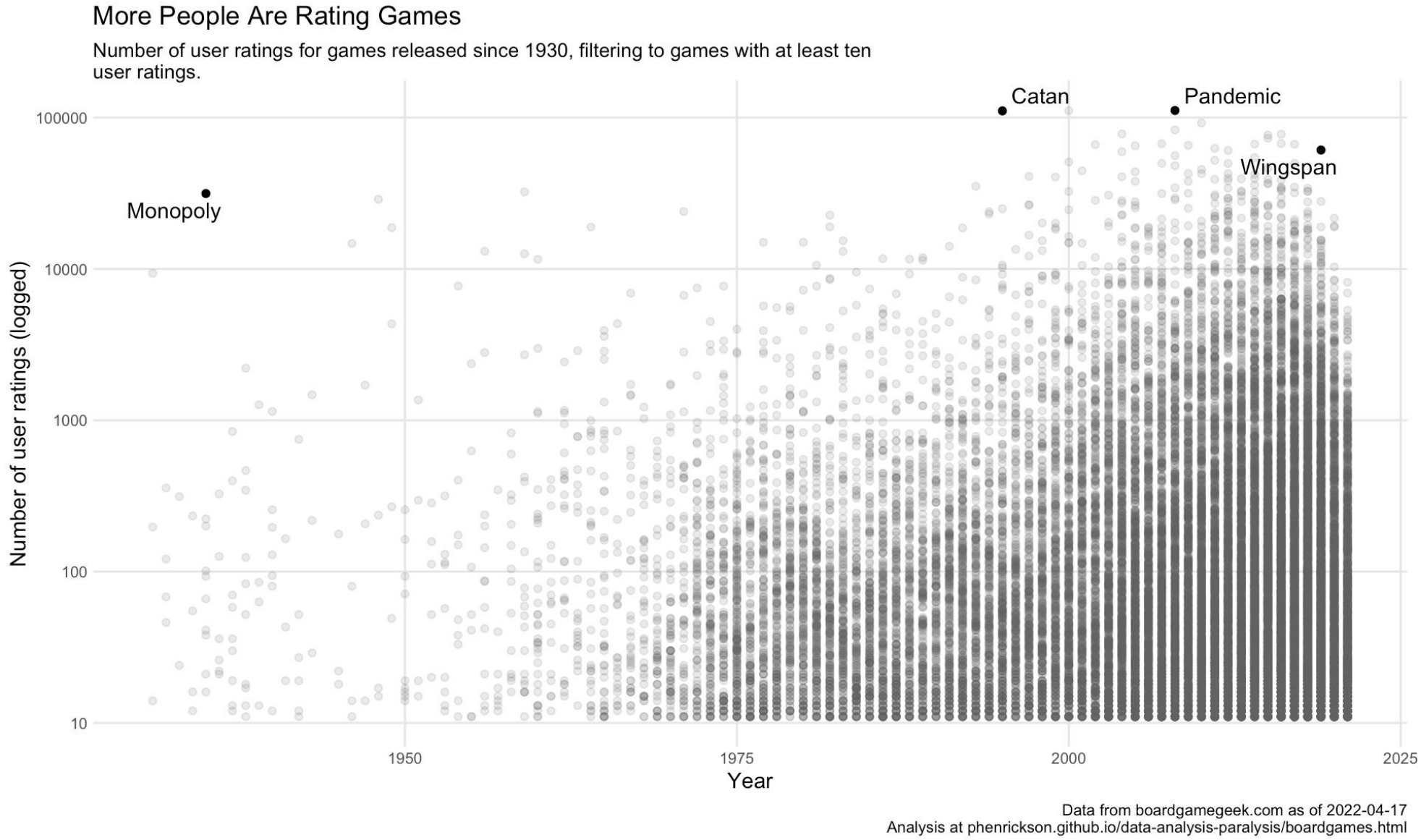
THE ORIGINS OF
POWER, PROSPERITY, AND POVERTY

WHY NATIONS FAIL

DARON ACEMOGLU JAMES A. ROBINSON



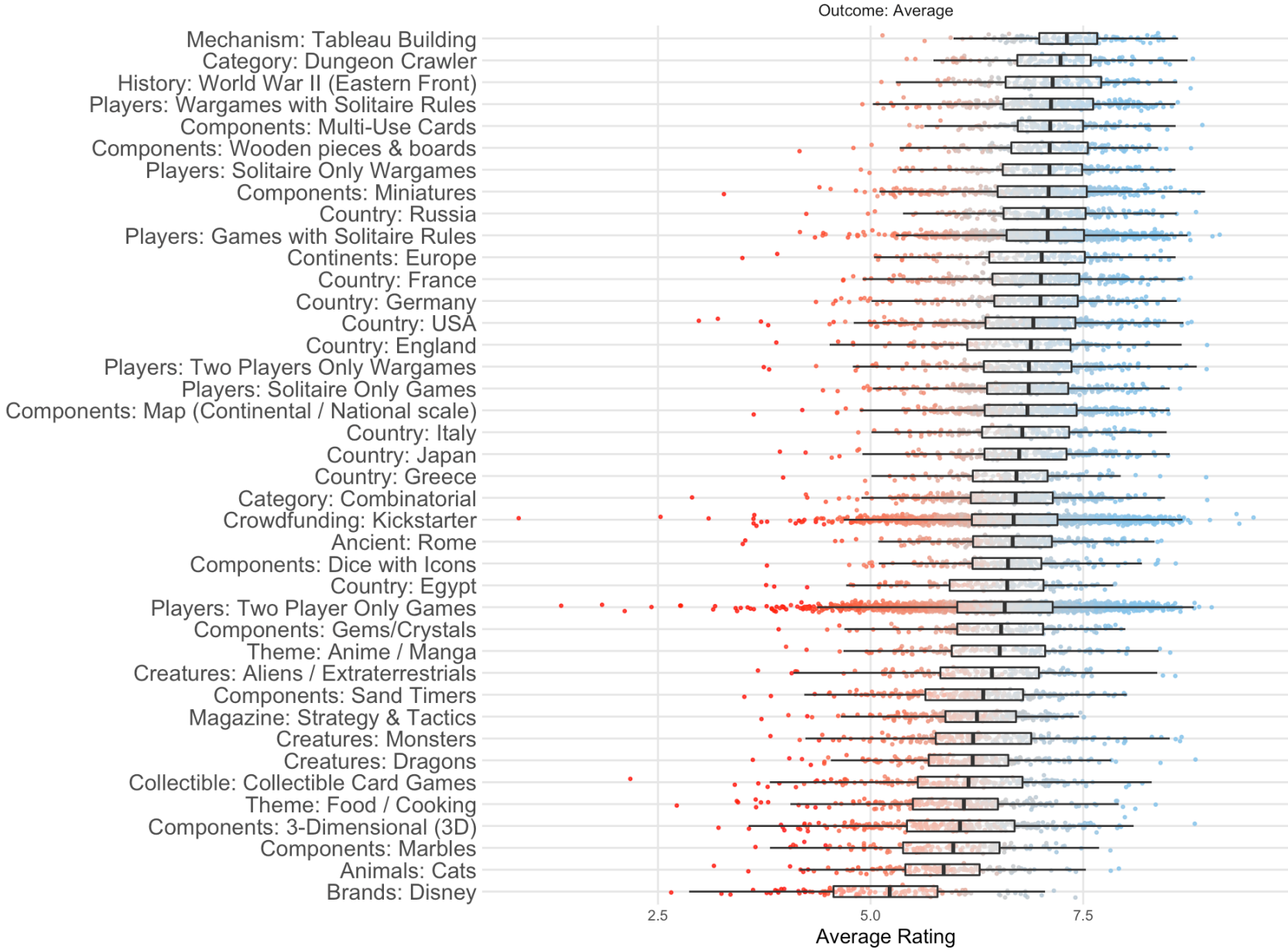
"A wildly ambitious work that hopscotches through history and around the world to answer the very big question of why some countries get rich and others don't."
—*NEW YORK TIMES*



More people are playing and rating games than ever before.

Why are Some Games Better Than Others?

Displaying the top 40 most frequent types of board games on BGG



Data from boardgamegeek.com as of 2022-04-17
Analysis at phenrickson.github.io/data-analysis-paralysis/boardgames.html