Now What?

Model Building, Science, and Analytics

Phil Henrickson, PhD
AE Business Solutions

Presented 05/20/21

First, some context.

The genesis of this talk is from a conversation with Olga, who remembered a previous presentation I gave at a MKE TUG.

I said at the time:

The genesis of this talk is from a conversation with Olga, who remembered a previous presentation I gave at a MKE TUG.

I said at the time:

Tableau enables companies to begin their journey of learning from data.

Visualization and reporting are not the end.

I may have also said something along the lines of:

I may have also said something along the lines of:

Every time I ask a client what they plan to do with a dataset and they say, 'oh we're going to put it into a dashboard and that'll be it', a part of me dies inside.

Olga asked us the following:

Our Q2 TUG is going to be all about Analytics. Like the real stuff, the stuff that everyone wants, but no one seems to get to…

We need more analytics than just visualization and reporting. Getting a dashboard up is great, but what you do with it is better.

Do you think that Dr Phil would be interested?

Olga asked us the following:

Our Q2 TUG is going to be all about Analytics. Like the real stuff, the stuff that everyone wants, but no one seems to get to…

We need more analytics than just visualization and reporting. Getting a dashboard up is great, but what you do with it is better.

Do you think that Dr Phil would be interested?

To which I, essentially, said:

Absolutely.

Can I talk about philosophy of science, astronomy, and/or hockey?

She said, ‘Sure, why not?’.

Can I talk about philosophy of science, astronomy, and/or hockey?

She said, ‘Sure, why not?’.

If this talk bores you to tears and/or confuses the heck out of you, I’m only partly to blame.

Can I talk about philosophy of science, astronomy, and/or hockey?

Now What?

Model Building, Science, and Analytics

Phil Henrickson, PhD
AE Business Solutions

Presented 05/20/21

A little bit about my background.

Phil Henrickson

Data Scientist
3 years consulting in analytics
PhD, Political Science
Predictive Modeling, Machine Learning, and Causal Inference
Research interests include international conflict, political violence, and minor league hockey
Obsessed with board games and miniature painting

AE BUSINESS SOLUTIONS

Data is hard.

We can help.

BUSINESS INTELLIGENCE

& ANALYTICS

Helping you ask and answer new questions.

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**?"

Some clients kind of think they’re just done:

Some clients kind of think they're just done:

"Everything that can be invented has been invented"

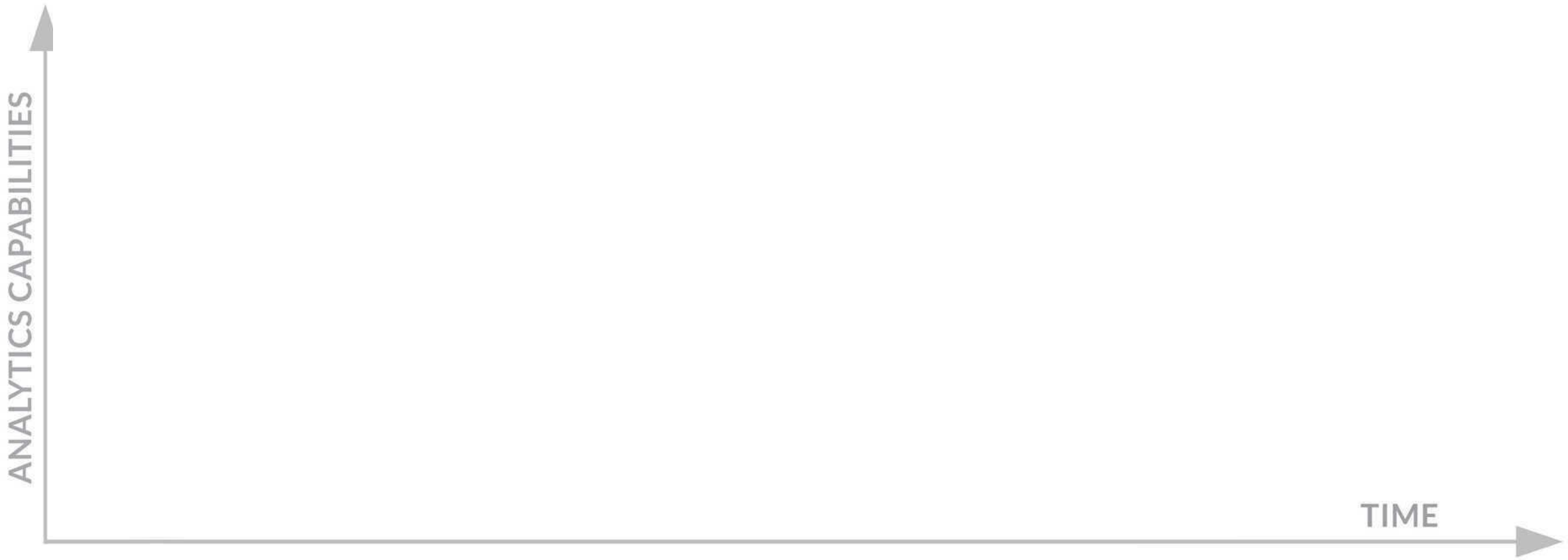
Charles H. Duell, 1899
Commissioner, U.S. Patent Office

Others are kind of just wondering if they've reached the end of the journey:

Others are kind of just wondering if they've reached the end of the journey:

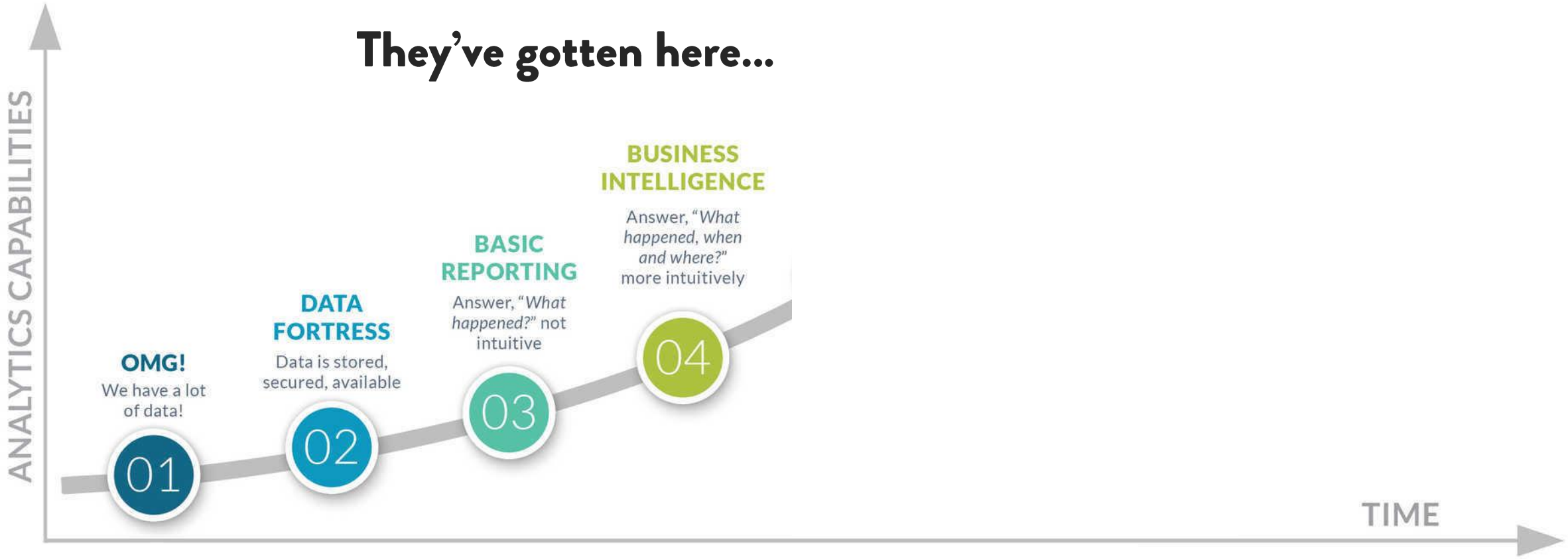
But most clients have discovered problems that dashboards and reports cannot solve.

But most clients have discovered problems
that dashboards and reports cannot solve.



But most clients have discovered problems that dashboards and reports cannot solve.

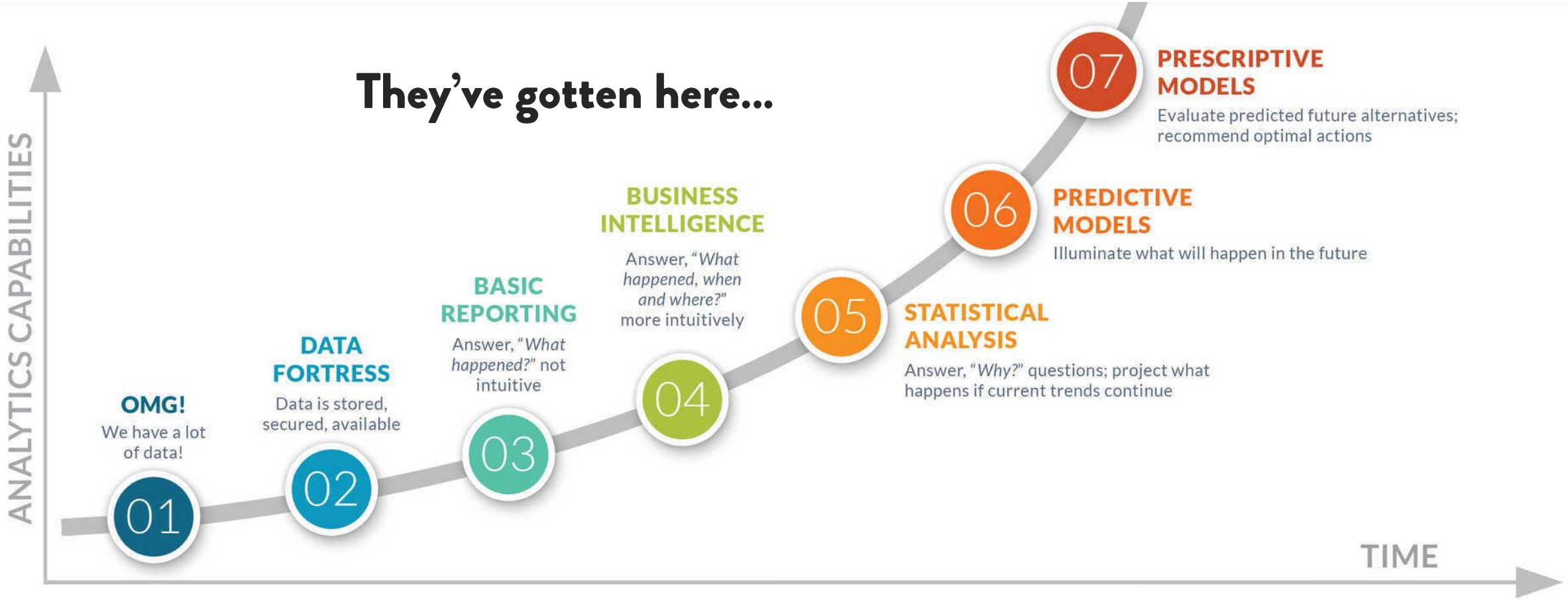
They've gotten here...



Raw Data -> Clean Data -> Dashboard -> ??? -> Data Driven Insight™

something was supposed
to happen here

... now they want to get here.



Raw Data -> Clean Data -> Model -> Dashboard -> ??? -> #AI Driven Insight™

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**.

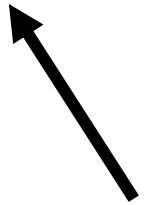
We need to ignore this for a second.

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)

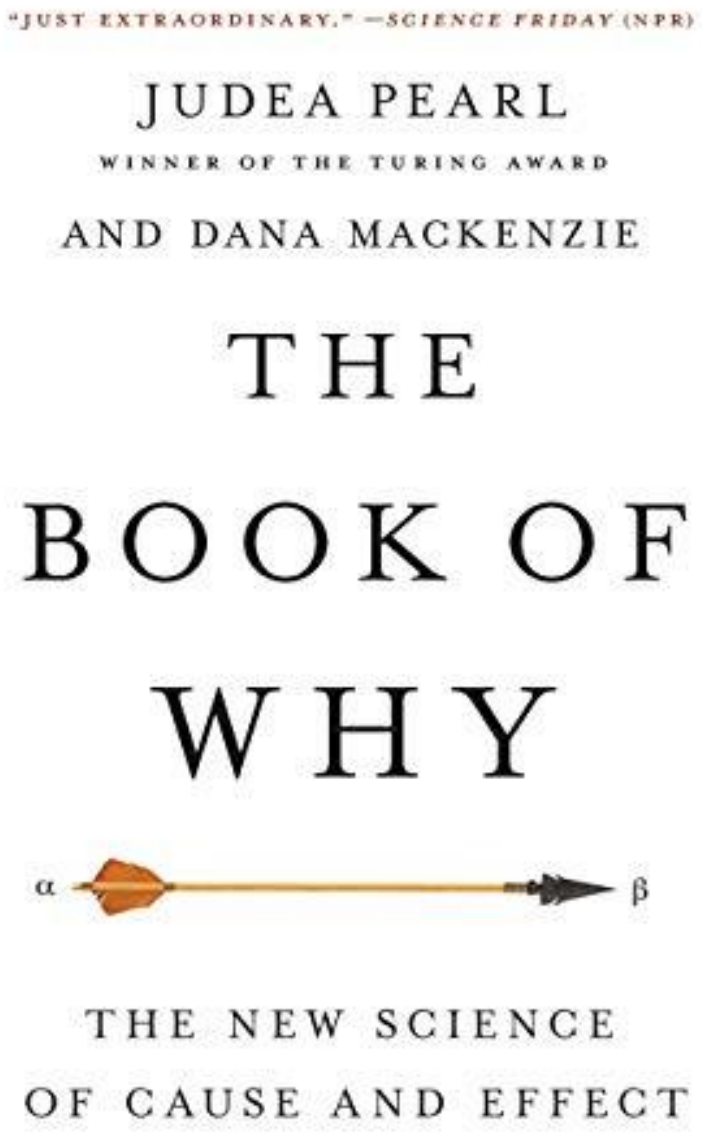
JUDEA PEARL

WINNER OF THE TURING AWARD

AND DANA MACKENZIE

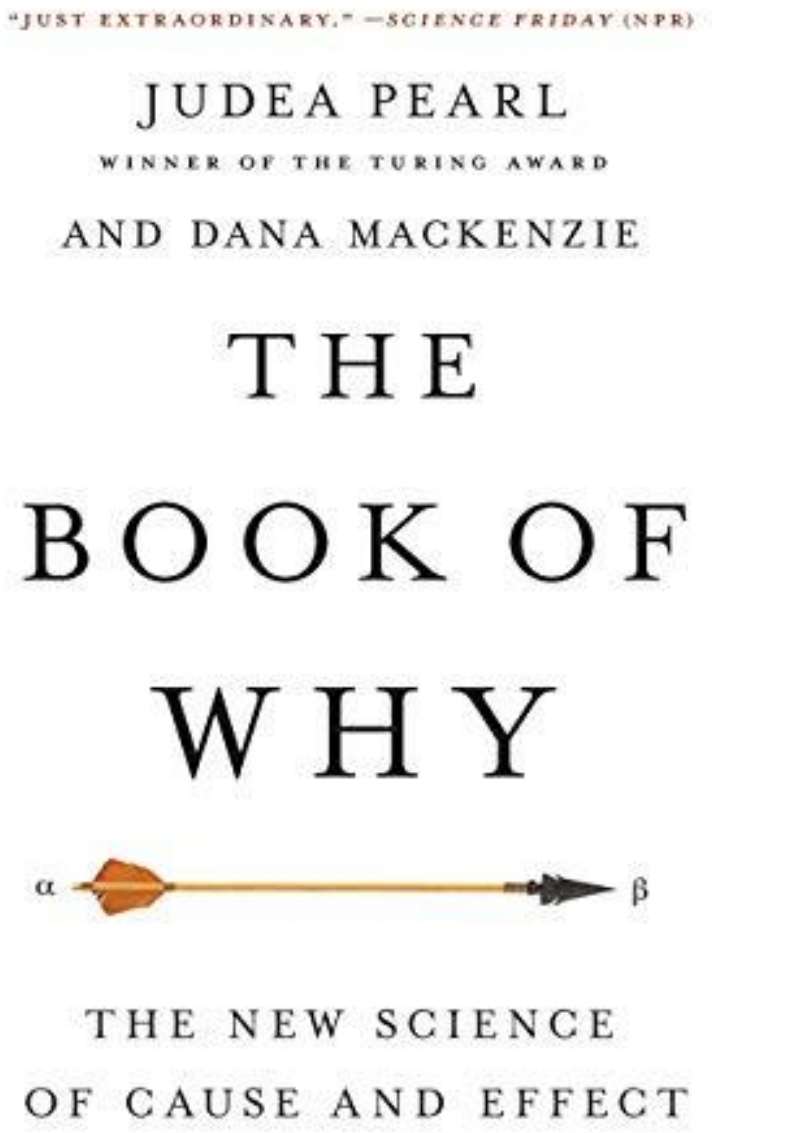
THE BOOK OF WHY

THE NEW SCIENCE
OF CAUSE AND EFFECT



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

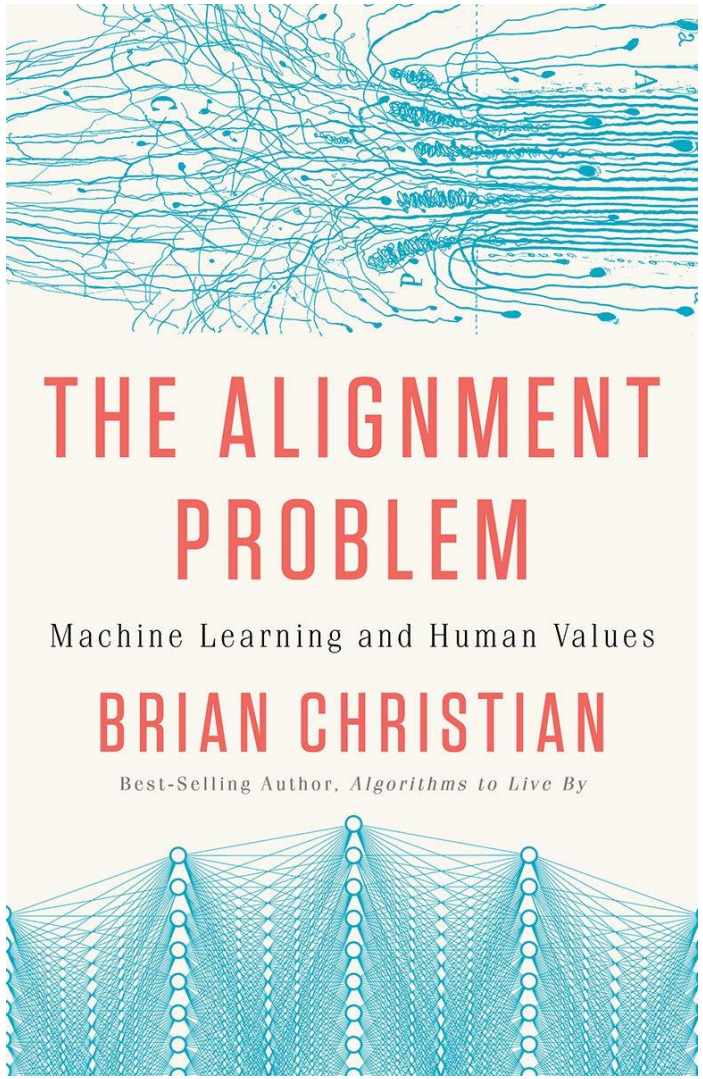
No machine can derive explanations from raw data… Data can tell you that the people who took a medicine recovered faster than those who did not take it, but **they can't tell you why**.

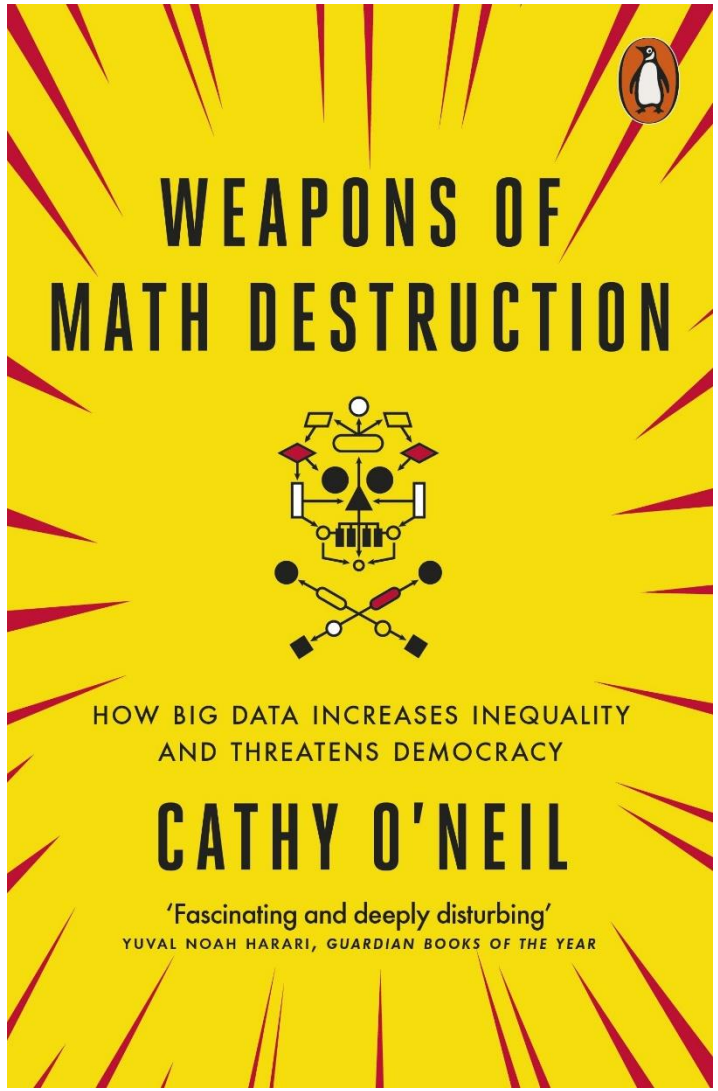


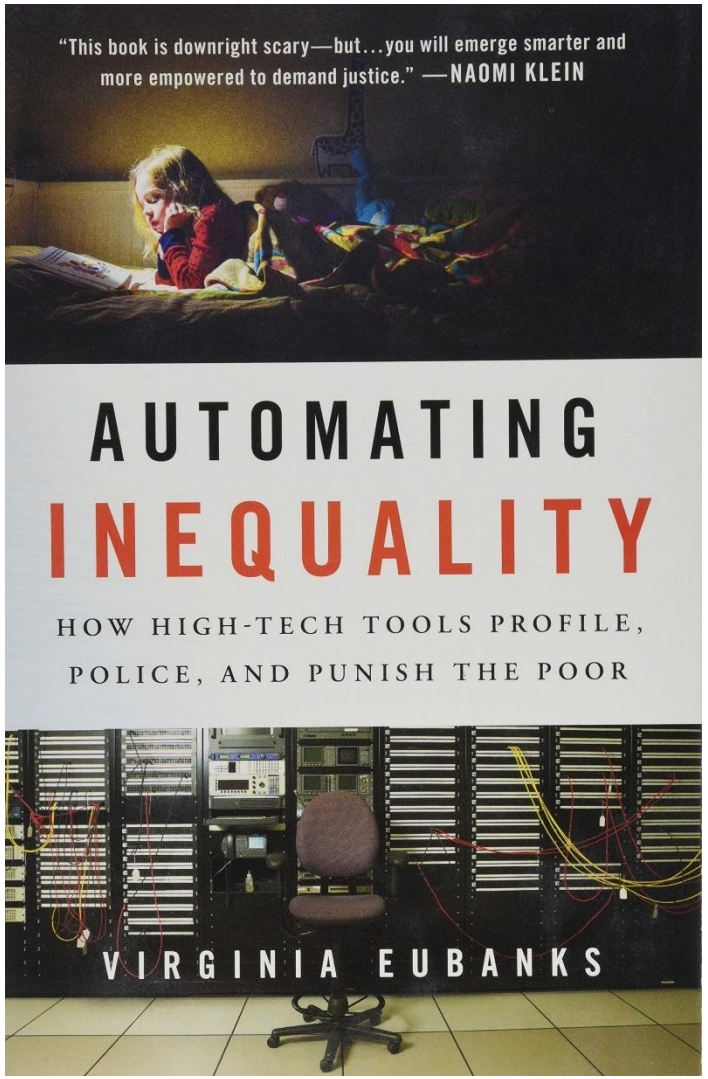
Over and over again, in science and in business, we see situations where mere **data aren't enough**.

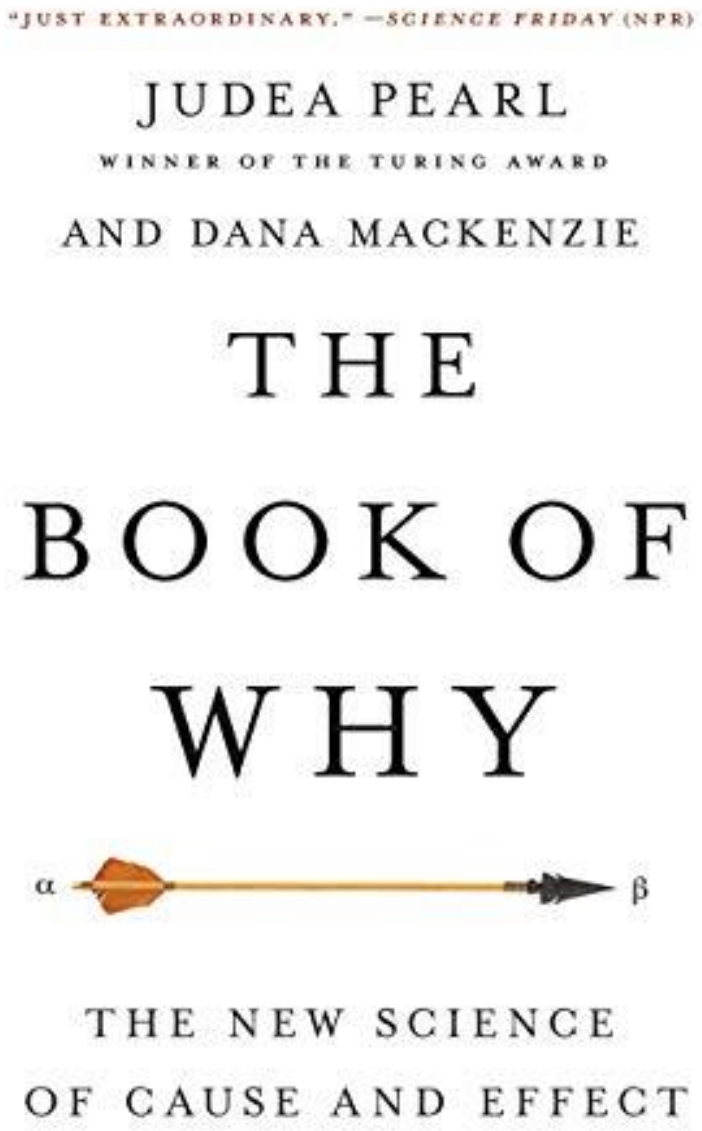
The hope… is that **the data themselves will guide us to the right answers** whenever causal questions come up.

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 method.

Raw Data -> Clean Data -> Dashboard -> ??? -> Data Driven Insight™

we were supposed to be learning from data

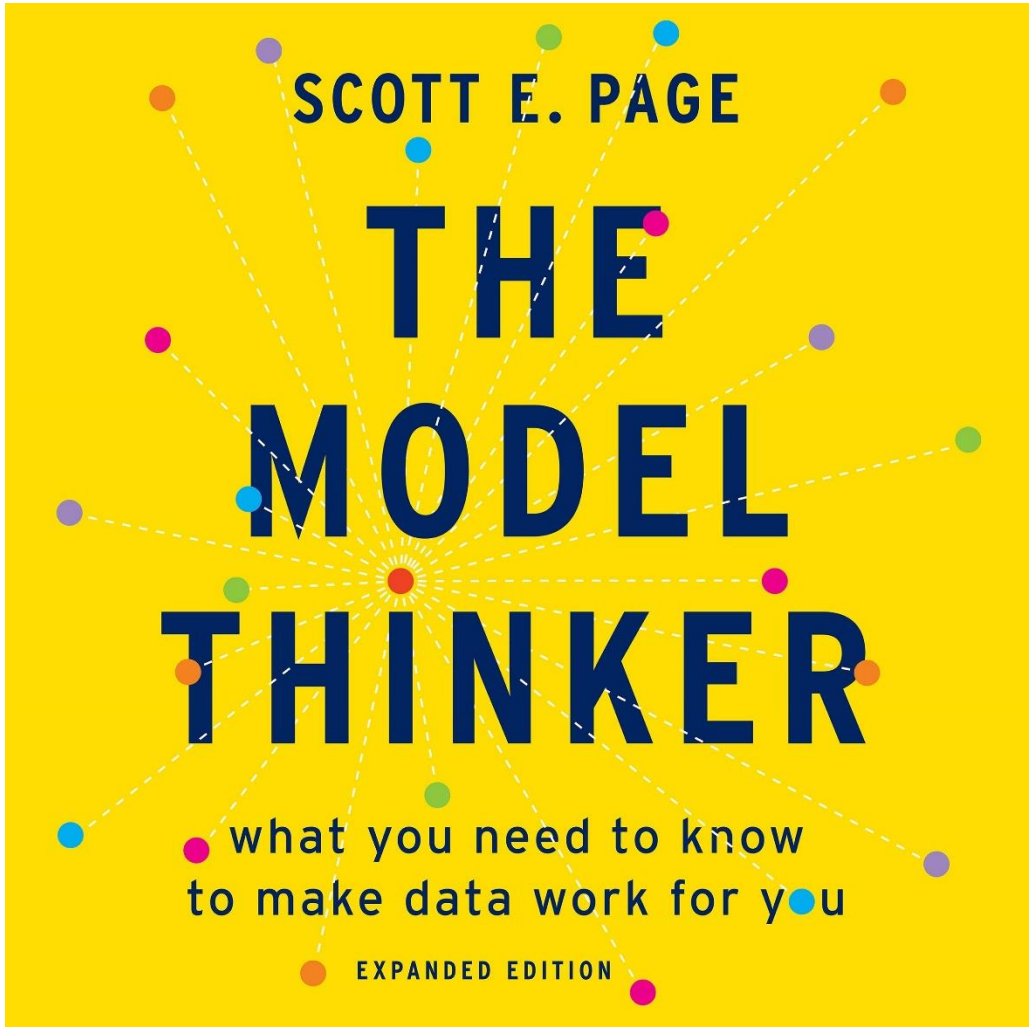
Raw Data -> Clean Data -> Model -> Dashboard -> ??? -> #AI Driven Insight™

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.

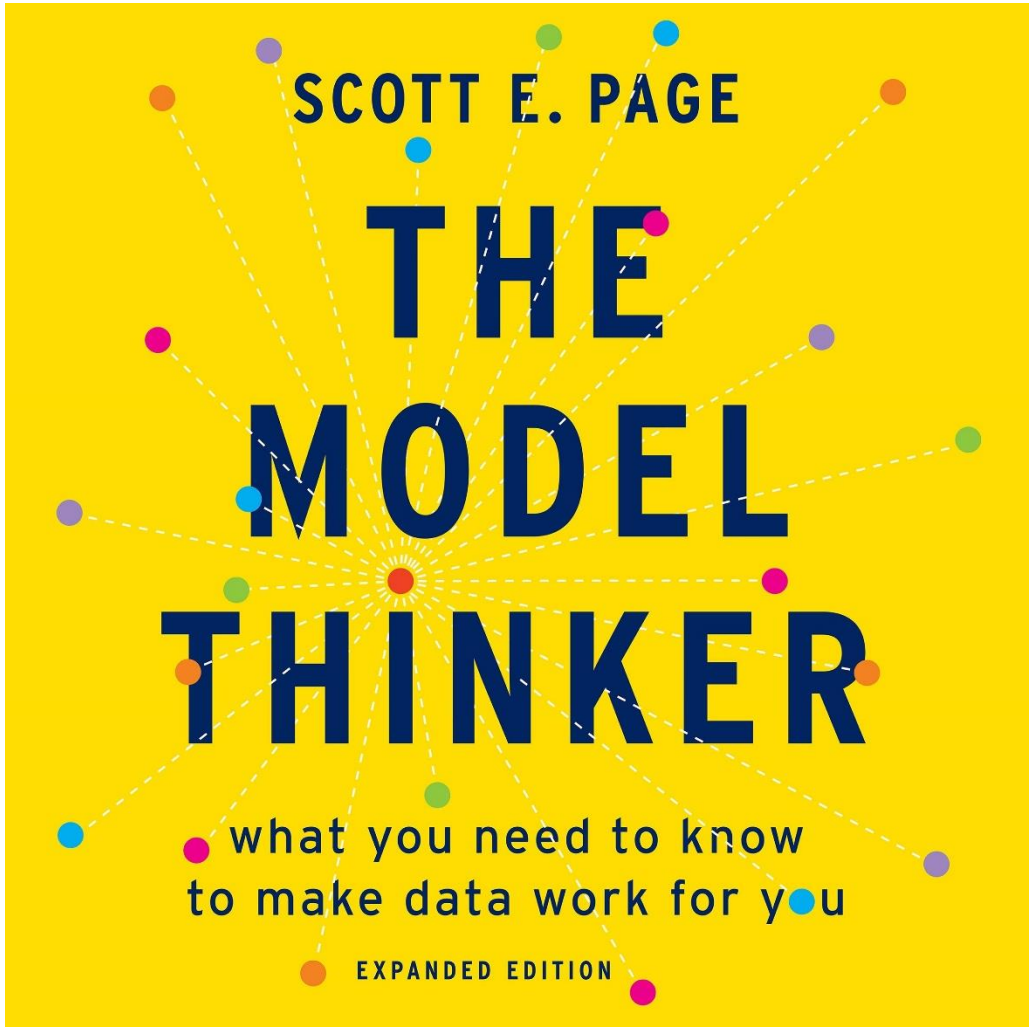
That brings us to the topic of the day:
building models.


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, **[humans] 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.

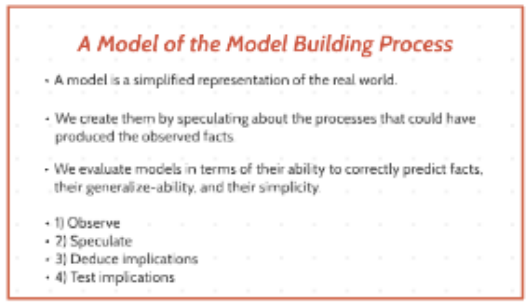
1 Building Models of the Stars

And Everything Else

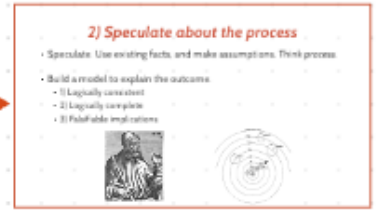
In a former life, I taught classes on research methods, international conflict, political violence, and civil war.

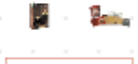
Regardless of the subject matter, I started every semester with the same lecture.

How to be a (Social) Scientist



Let's play a game!





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.

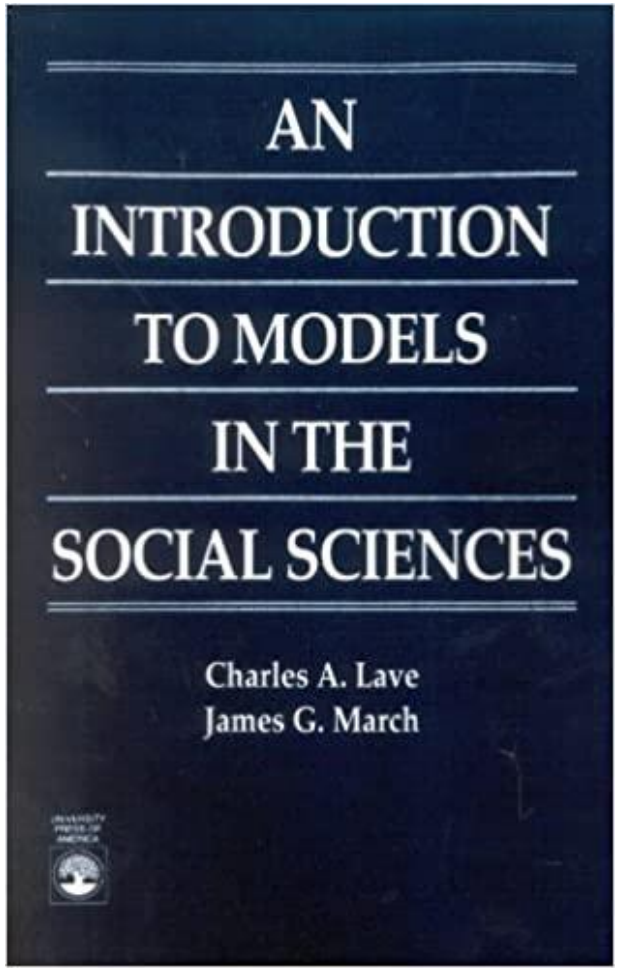
So, naturally, I immediately tried to win them over using the coolest possible method:

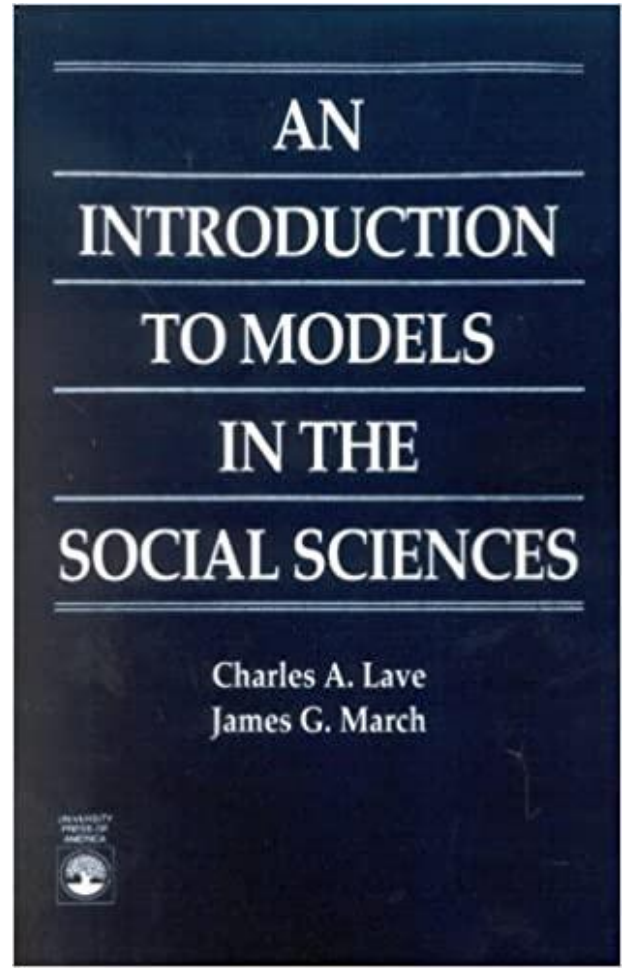
How to be a (Social) Scientist

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:

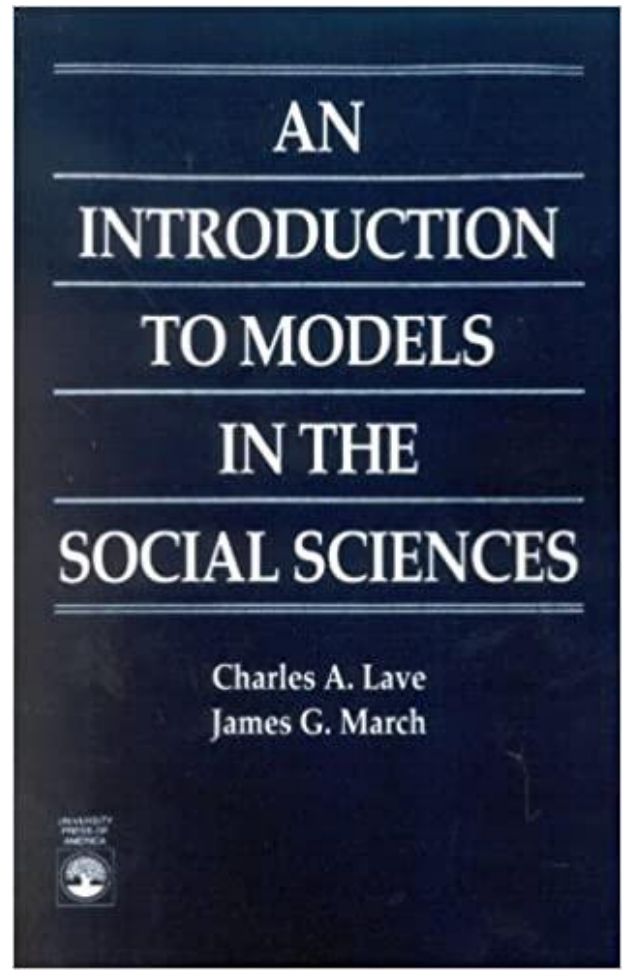
a social sciences textbook from 1975




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

chapter one

what we are up to

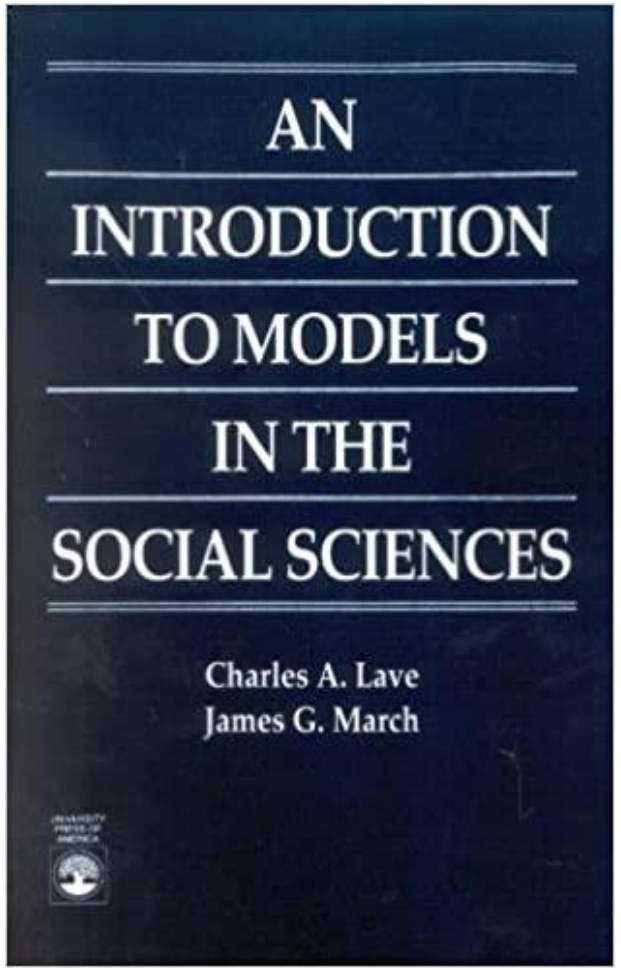


chapter one

what we are up to

God has chosen to give the easy problems to the physicists.

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



More than anything else, Lave and March are known for their emphasis on **speculation** in looking at data.

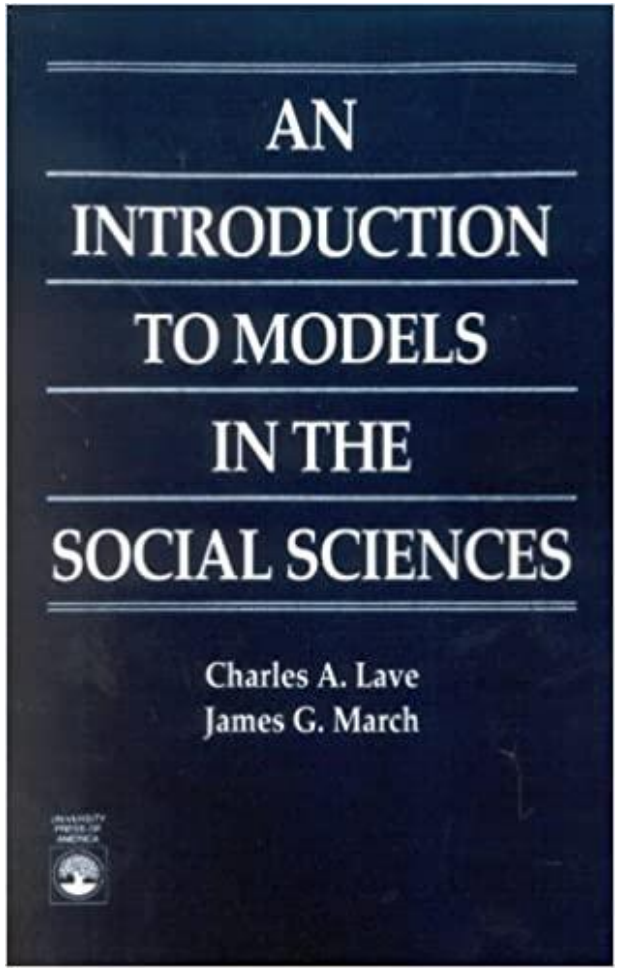
They point out that we are all building models all the time, but we need to adopt a consistent approach in doing so. And we need to '**stop and think**' a bit more.

chapter one

what we are up to

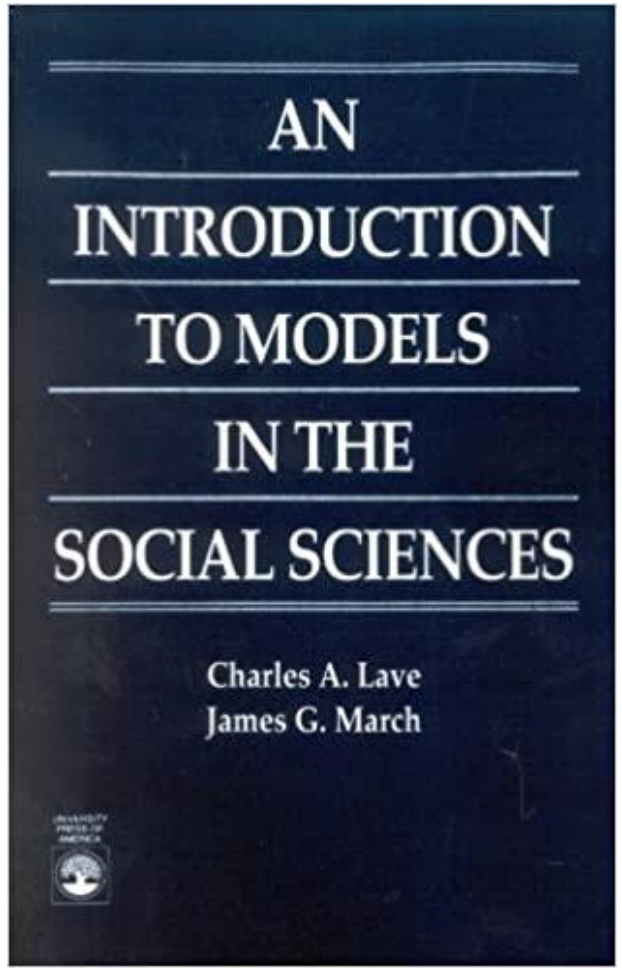
Speculative models are central to science, history, and literature. They are also part of normal existence. We are **constantly forming partial interpretations of the world** in order to live in it.

Because we do not always label our daily guesses about the world as "models", we sometimes overlook the extent to which **we are all theorists of human behavior**.



chapter two

an introduction to speculation

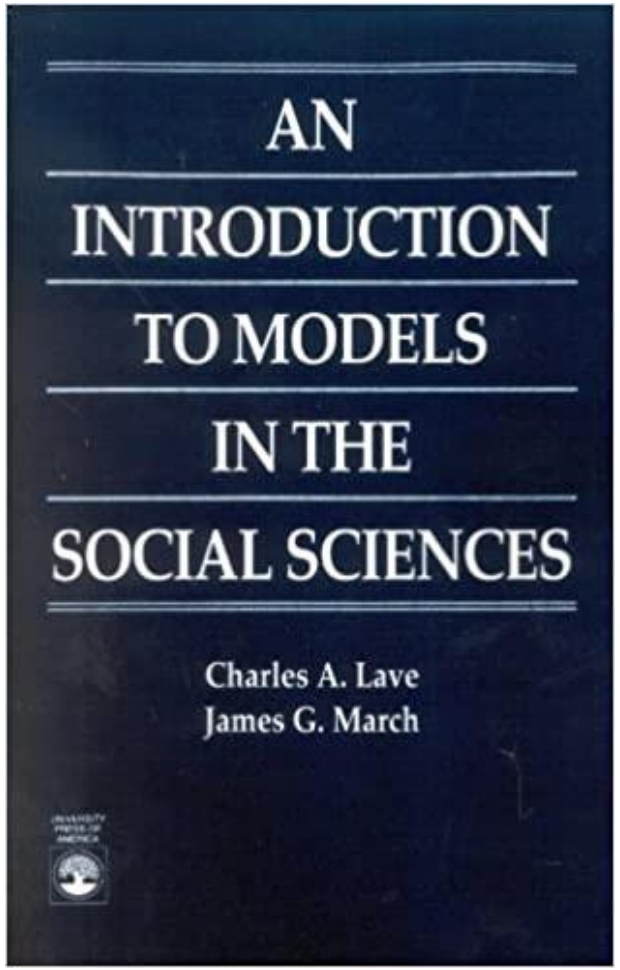


chapter two

an introduction to speculation

One feature of this book is that we will often ask you **to stop and do some thinking**. **We are serious.**

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 outcome 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.

The students **would have to have known each prior to the start of the semester**. But, we observe a freshman dormitory and notice the same pattern.

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?

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 what 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 what 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.

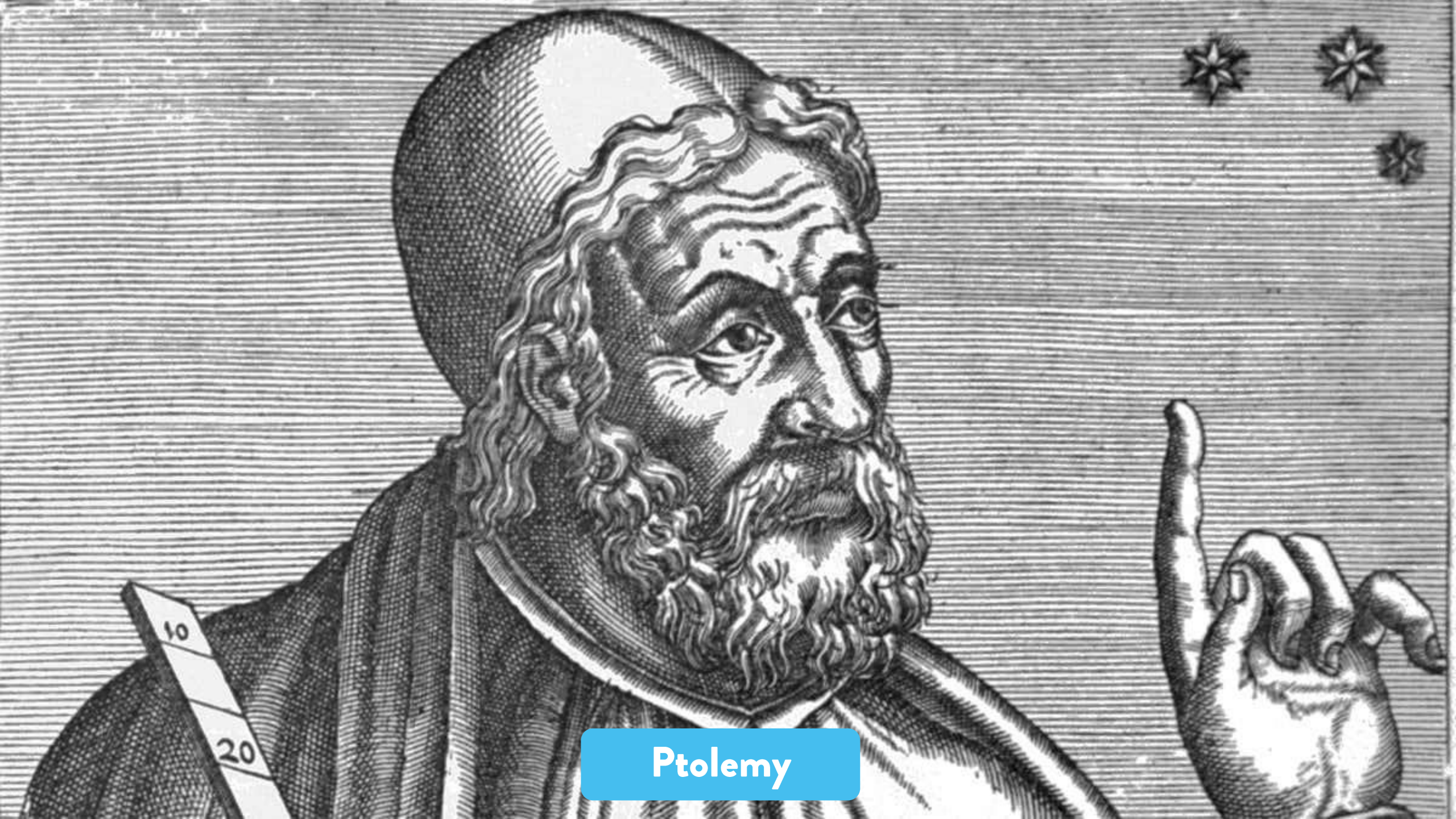
Why does Mars move backwards in the nighttime sky?

Model
(Speculation)

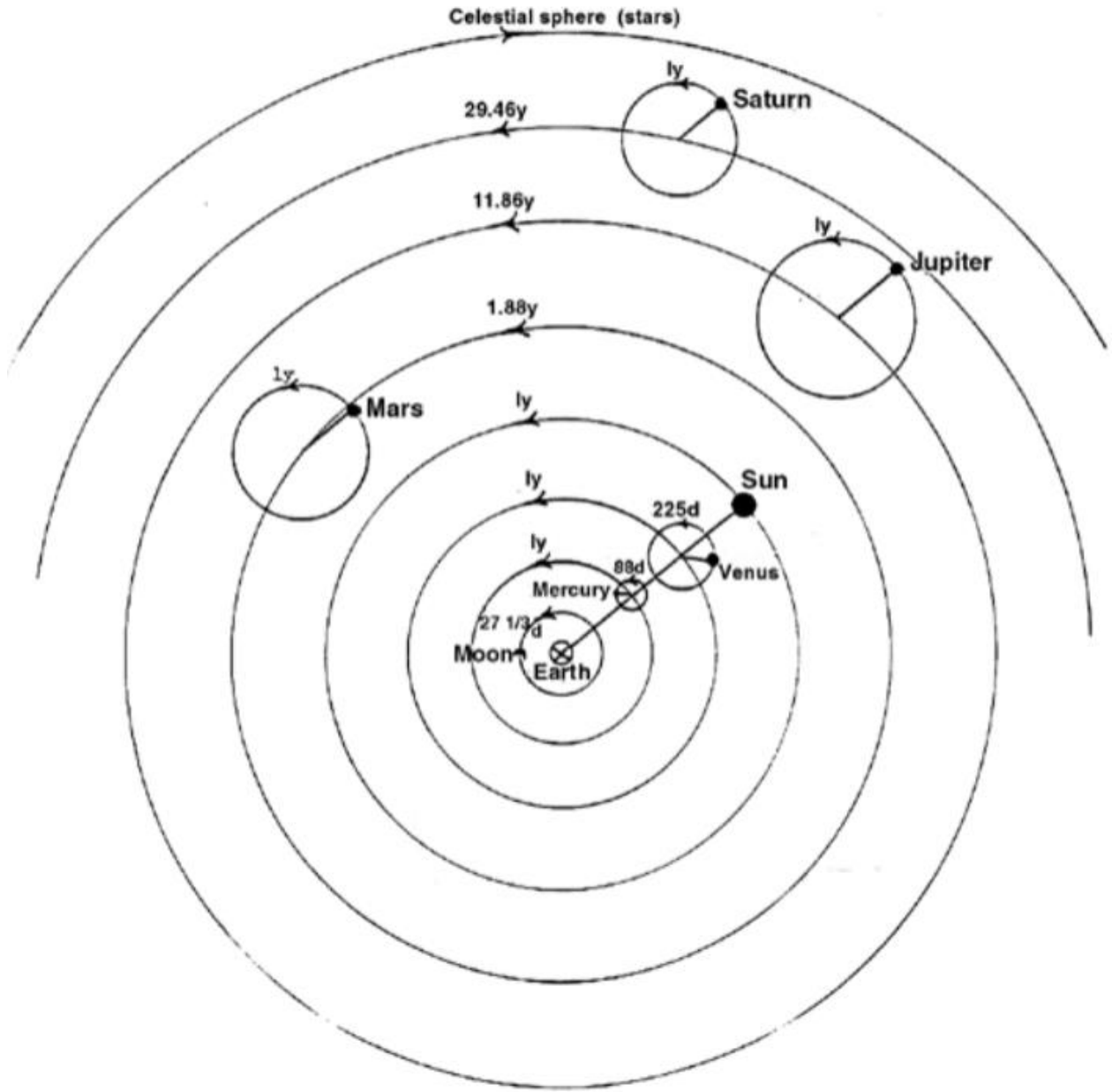
The Earth is at the center of the solar system.
The heavens are in perfect harmony and
objects must orbit the Earth in circles.

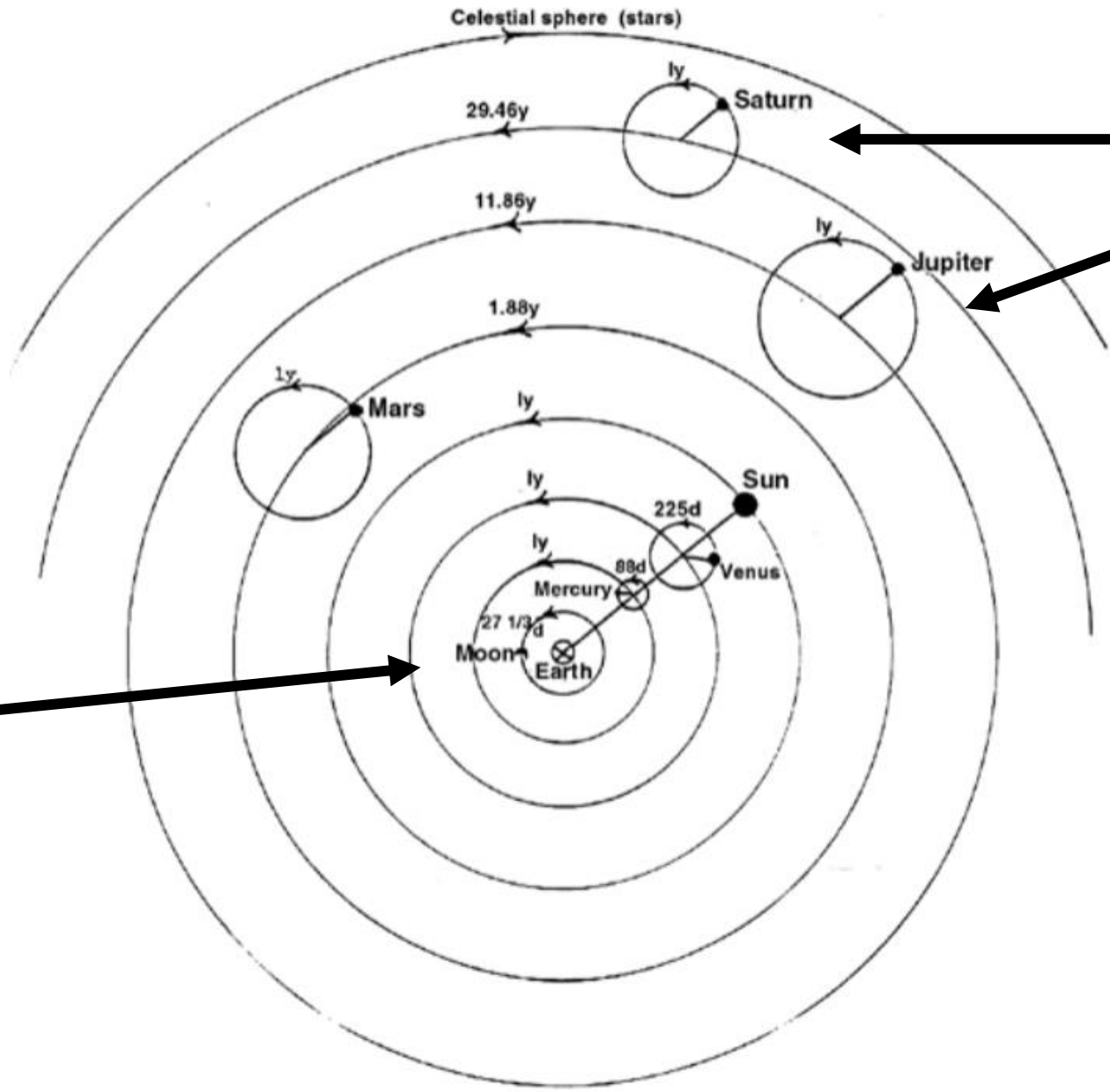
Model (Speculation)	The Earth is at the center of the solar system. The heavens are in perfect harmony and objects must orbit the Earth in circles.
Implication	If that's the case, Mars shouldn't move backwards.

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



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



Planets moving in 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.

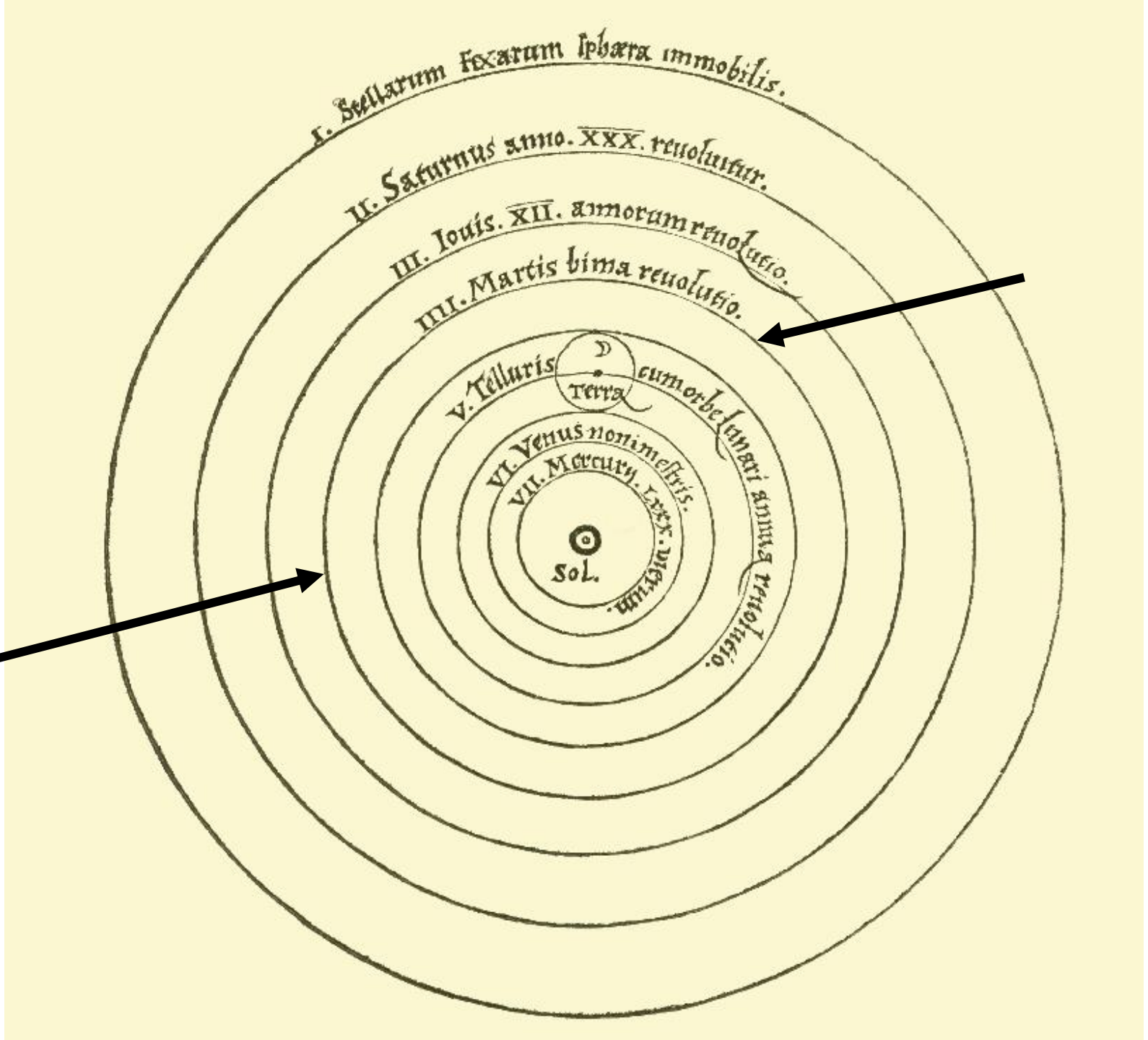
And it does. But not perfectly.

More observation led to more data collection which led to more inaccuracies.

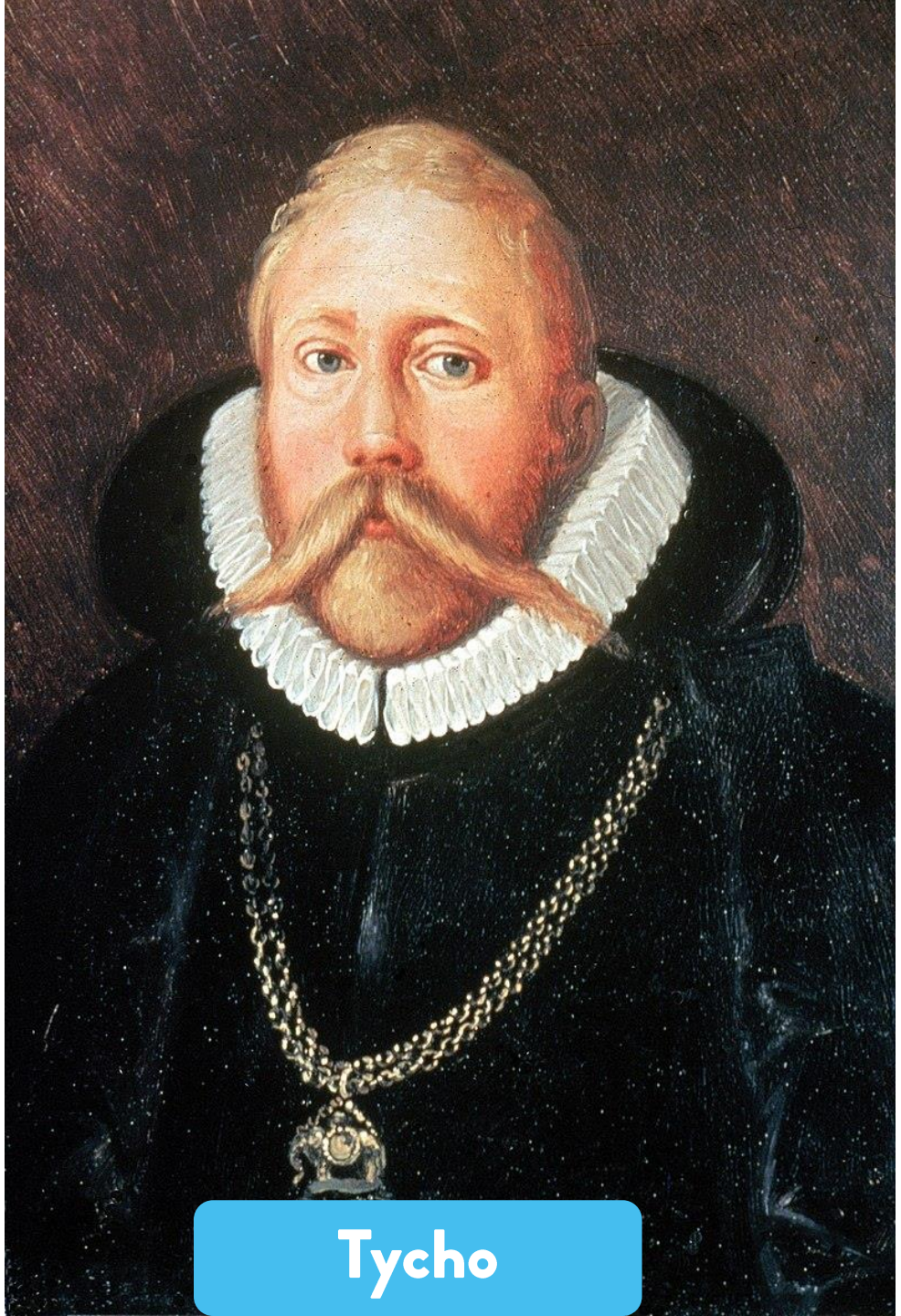
New models for the solar system were proposed.


Copernicus

Sun at the center

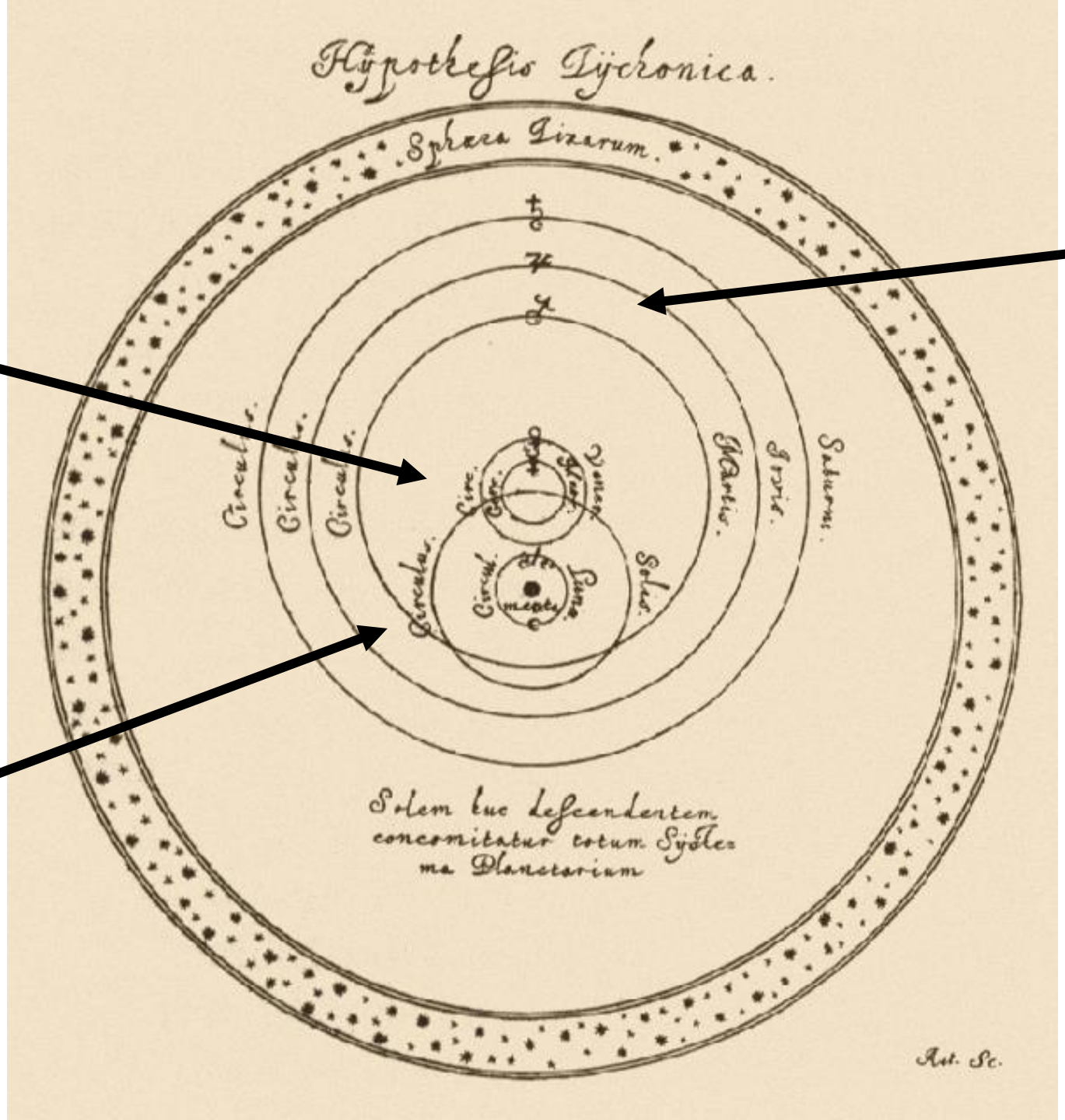


Objects, including Earth, move in circles


Tycho

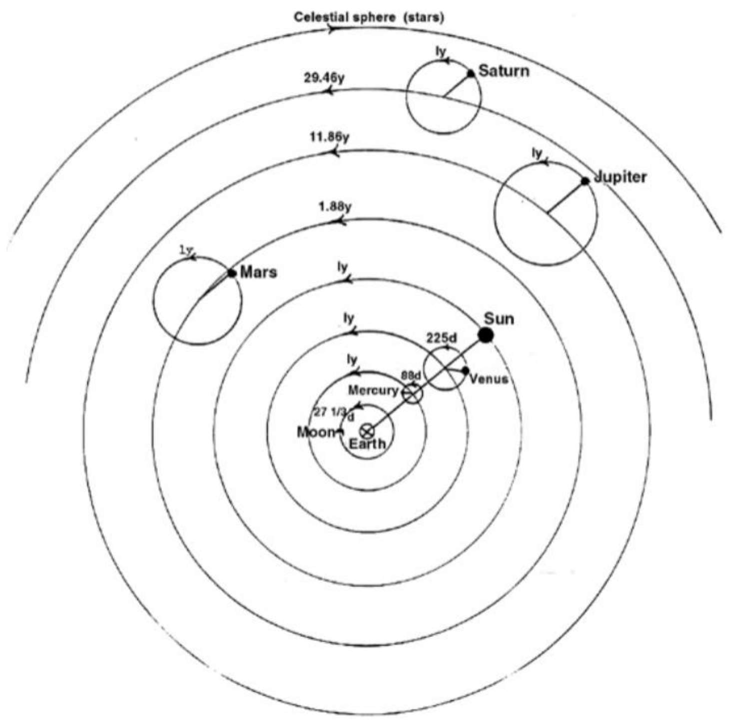
Sun orbits the Earth

Earth at the center

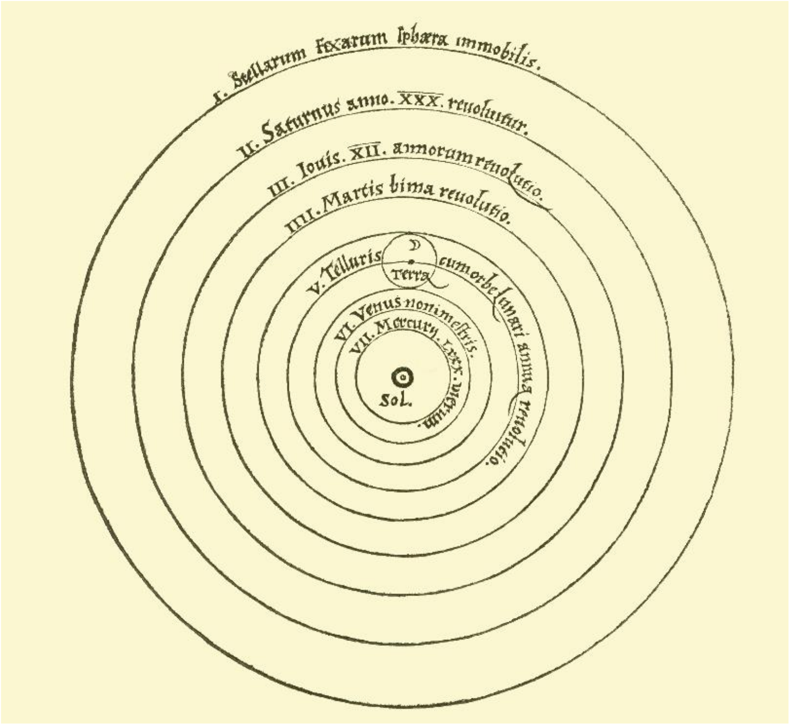


Planets orbit sun

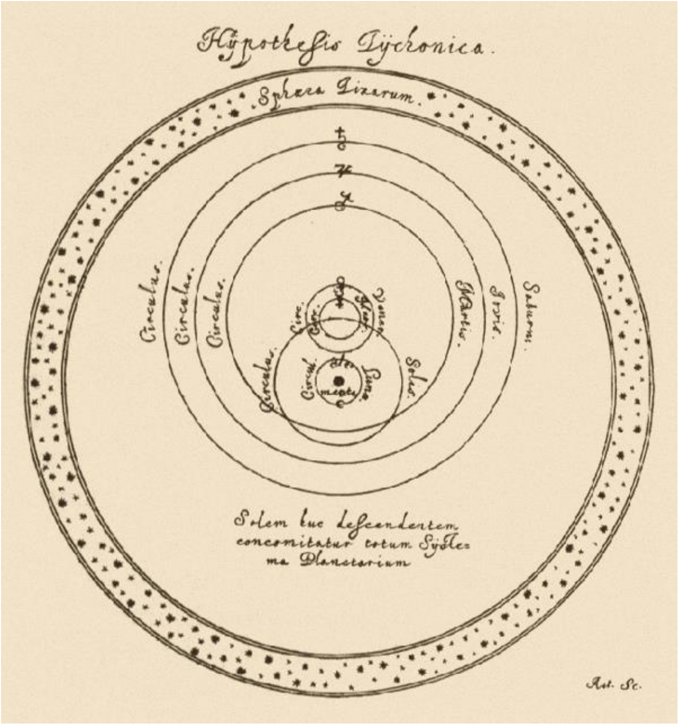
Ptolemy



Copernicus



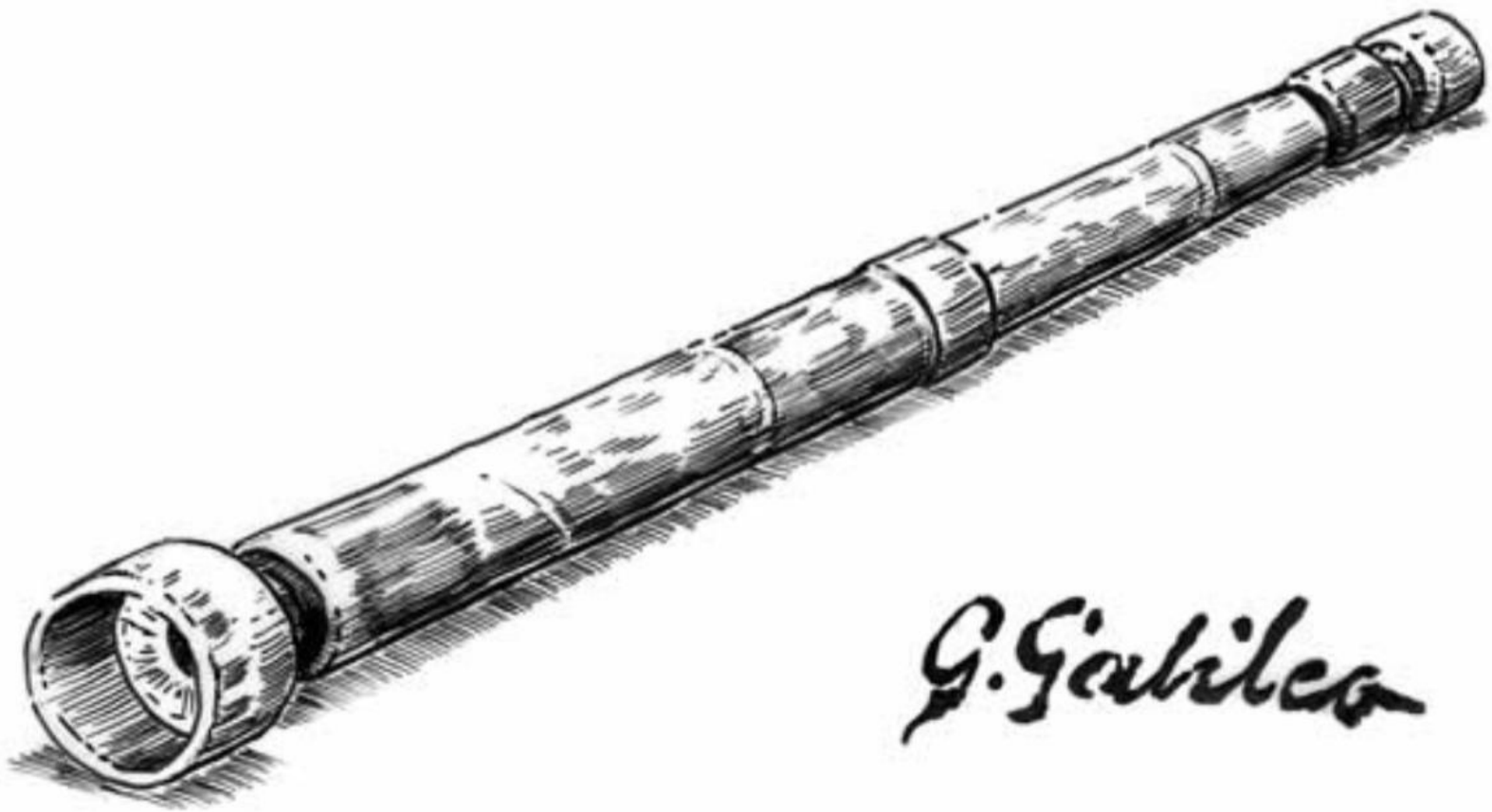
Tycho

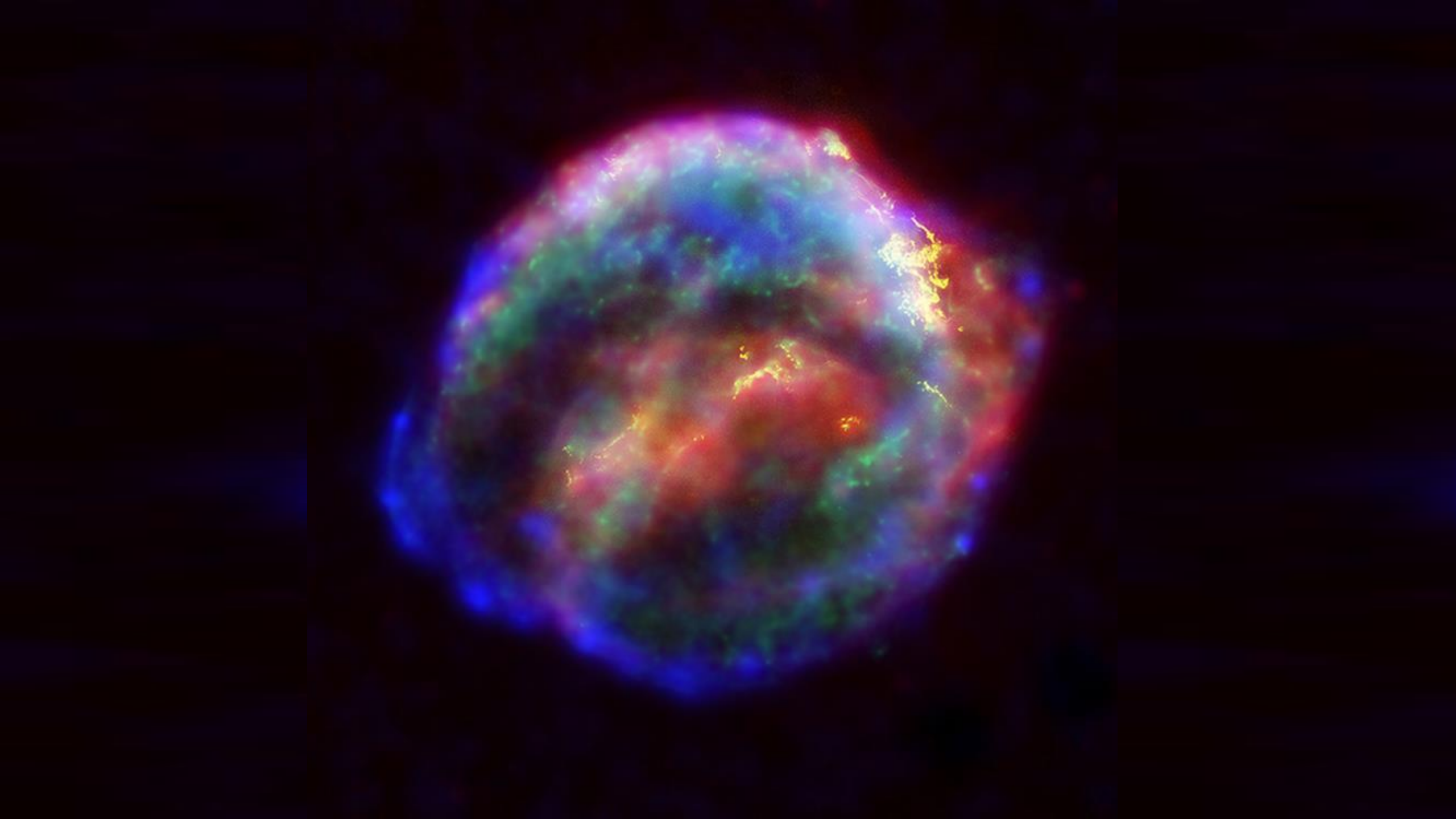


How do we determine which model is ‘the most correct’?

We think about implications. We gather data.

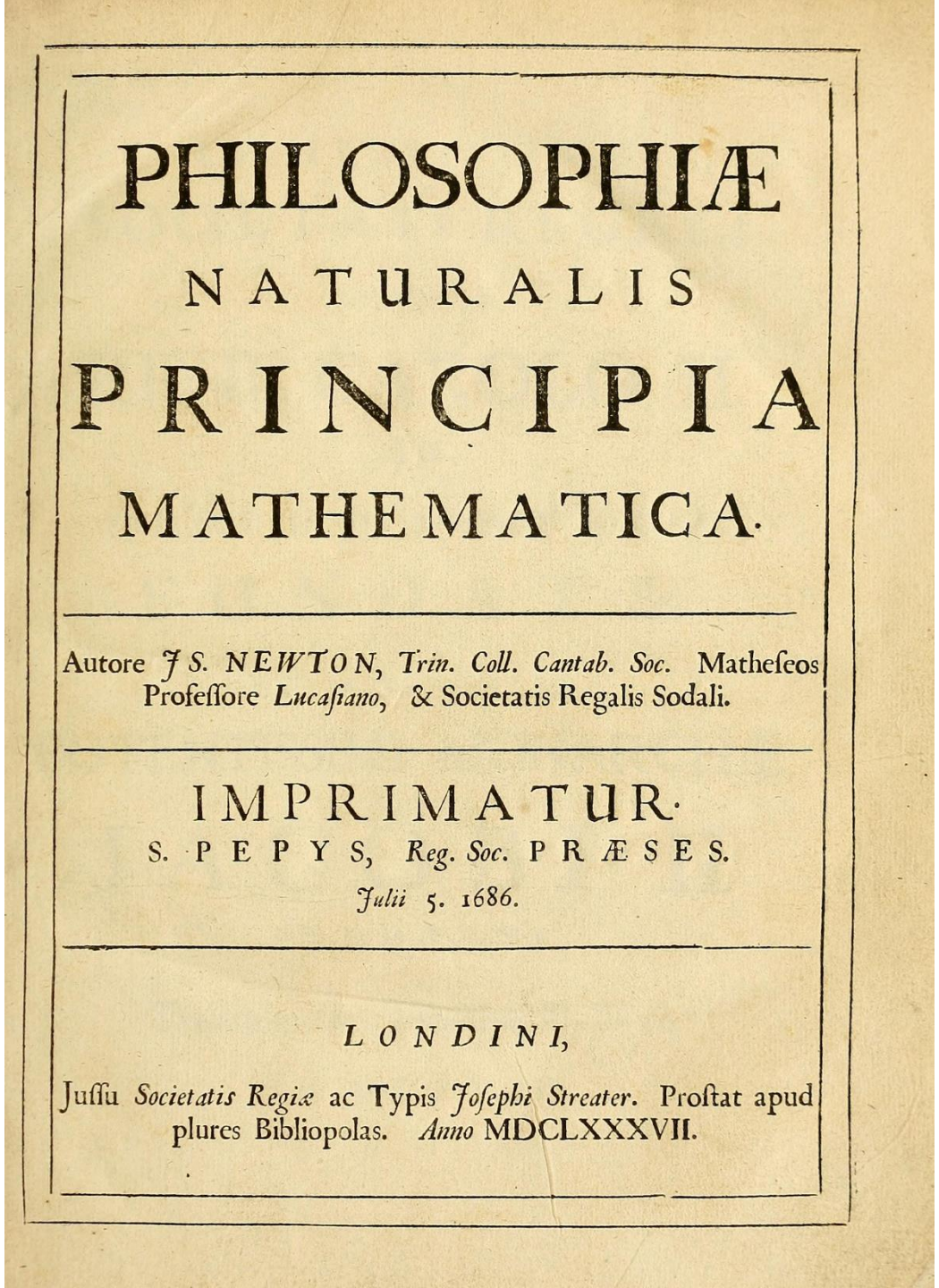
New methods of data collection challenged implications of Ptolemy's model.


G. Galilea



New developments in physics and mathematics offered a new process to explain the motion of the planets.

Newton's Principia



PHILOSOPHIÆ
NATURALIS
PRINCIPIA
MATHEMATICA.

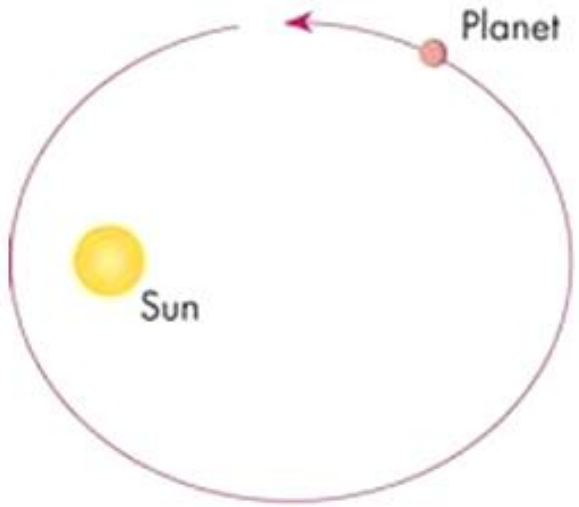
Autore *J S. NEWTON, Trin. Coll. Cantab. Soc.* Matheſeos Profeſſore *Lucaſiano*, & Societatis Regalis Sodali.

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

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

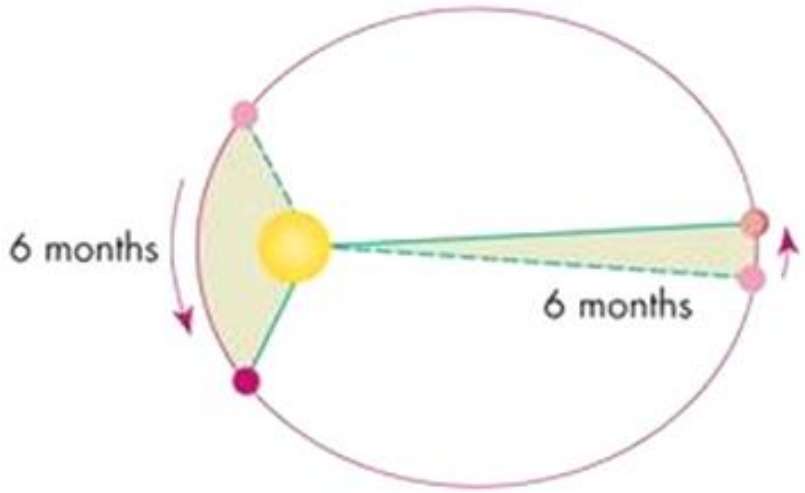

Kepler

Kepler's Laws of Planetary Motion



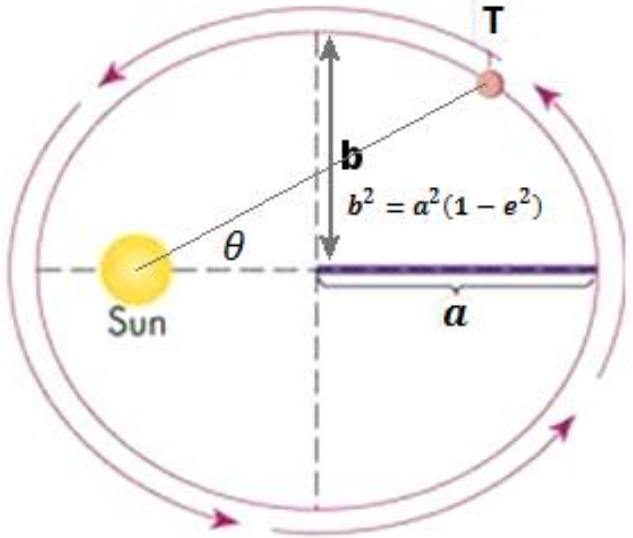
(1)

The orbits are ellipses



(2)

Equal areas in equal time

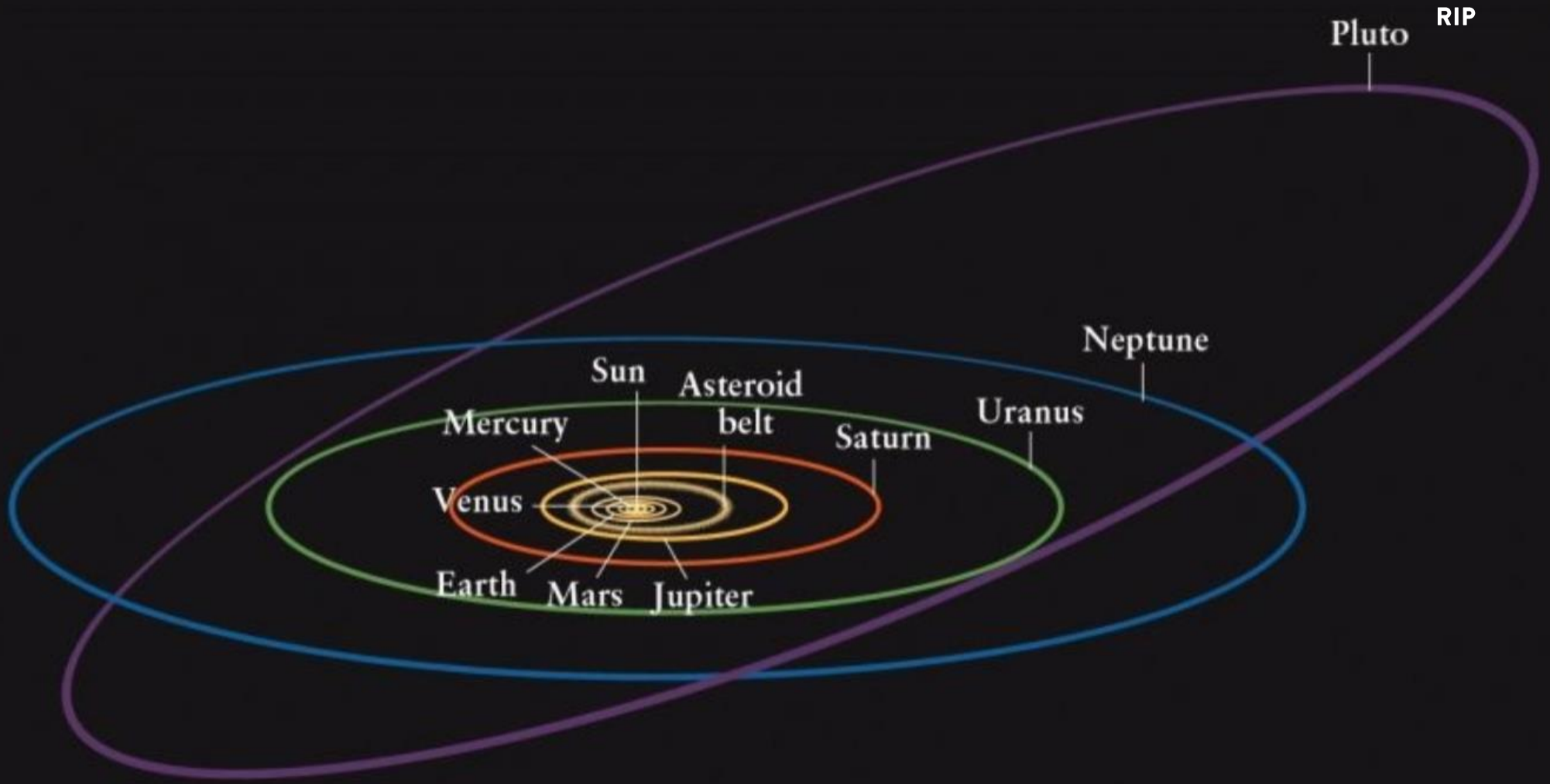


(3)

$T^2 \propto a^3$

T = time to complete orbit

a = semi-major axis


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

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?

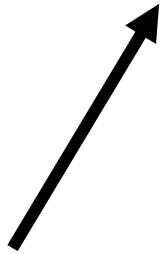
4) **We test**. We look to see whether the other implications of our model are supported in the data.

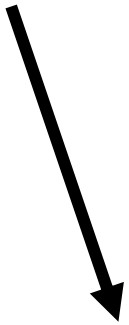
A model of the model building process:

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.

Data is used here.





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** enable learning.

Data does not **by itself** enable learning.

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

Example:

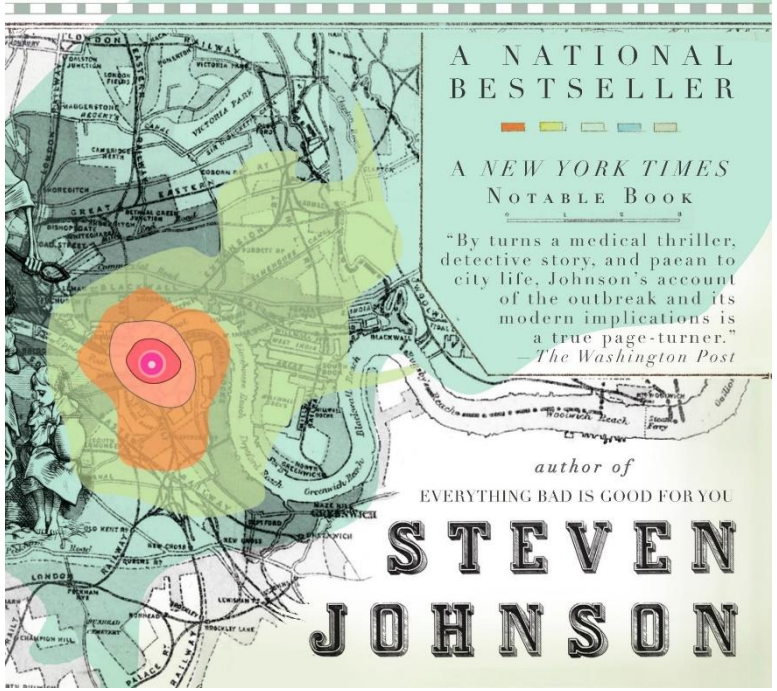
John Snow and the London Cholera Outbreak of 1854


GREAT MARLBOROUGH STREET
PORTLAND STREET
WORK HOUSE
BROAD STREET
GOLDEN SQUARE
WARWICK STREET
REGENT STREET
OLD COMPTON
QUEEN STREET
PUMP

John Snow's map of the London cholera outbreak of 1854

THE GHOST MAP

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



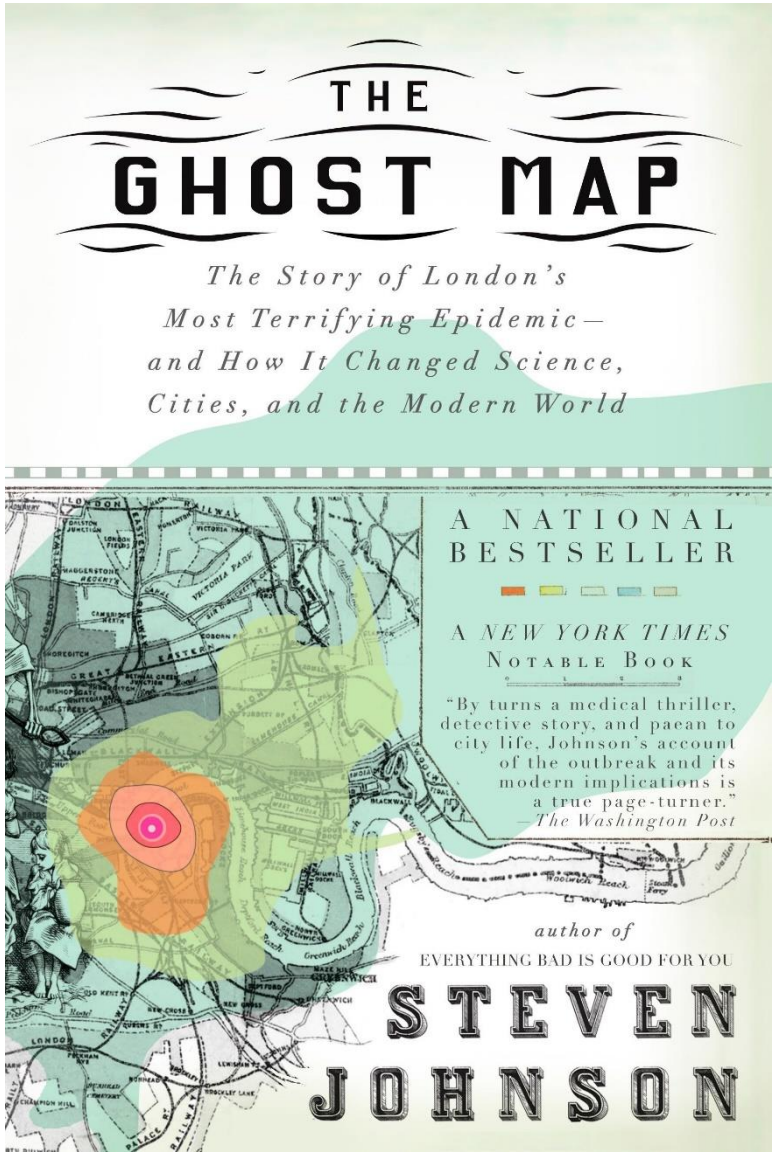
A NATIONAL BESTSELLER

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



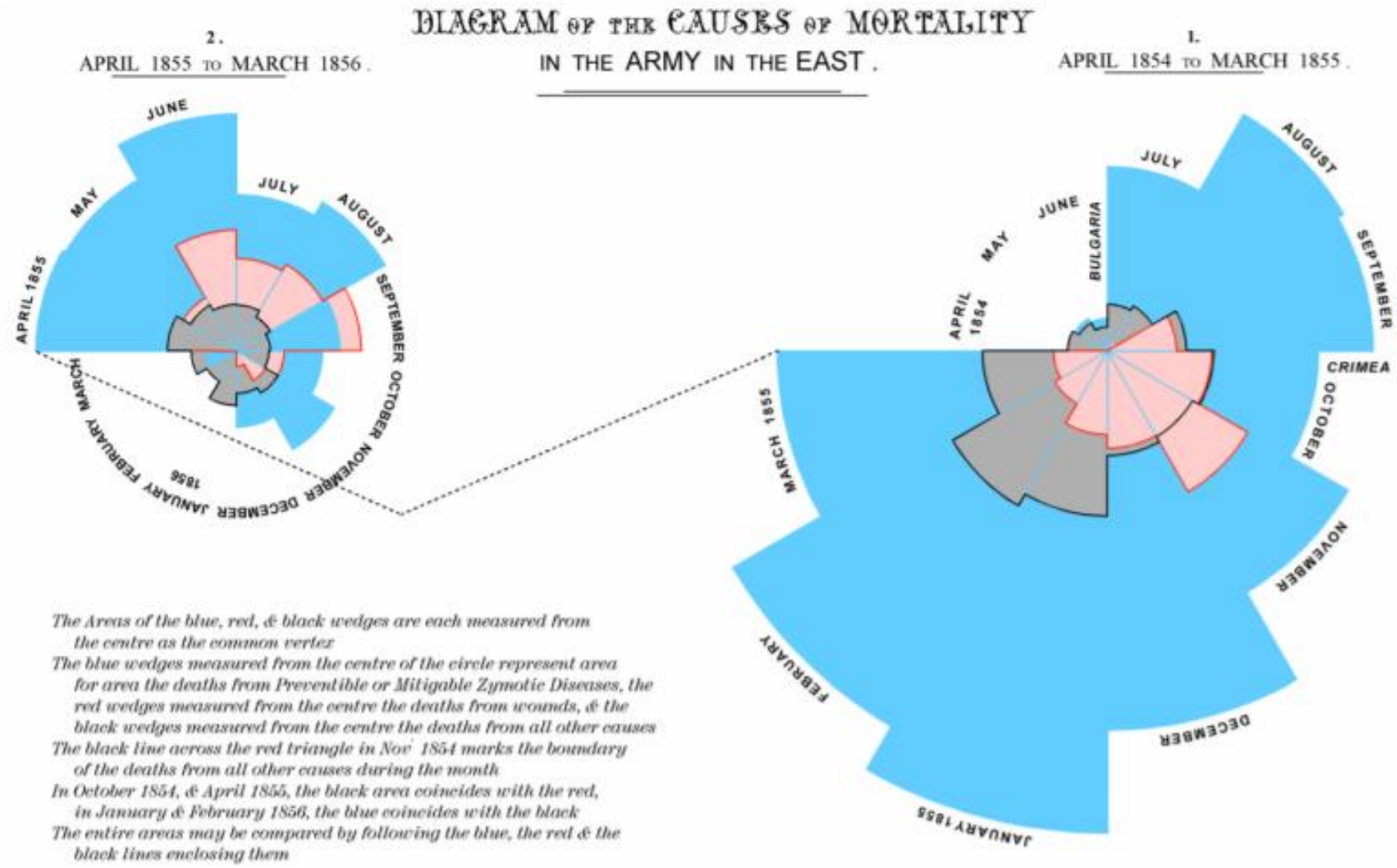
The dominant theory of disease at the time was **miasma** – disease was spread through bad air.

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**.

Example:

Florence Nightingale and the Fatalities in the Crimean War

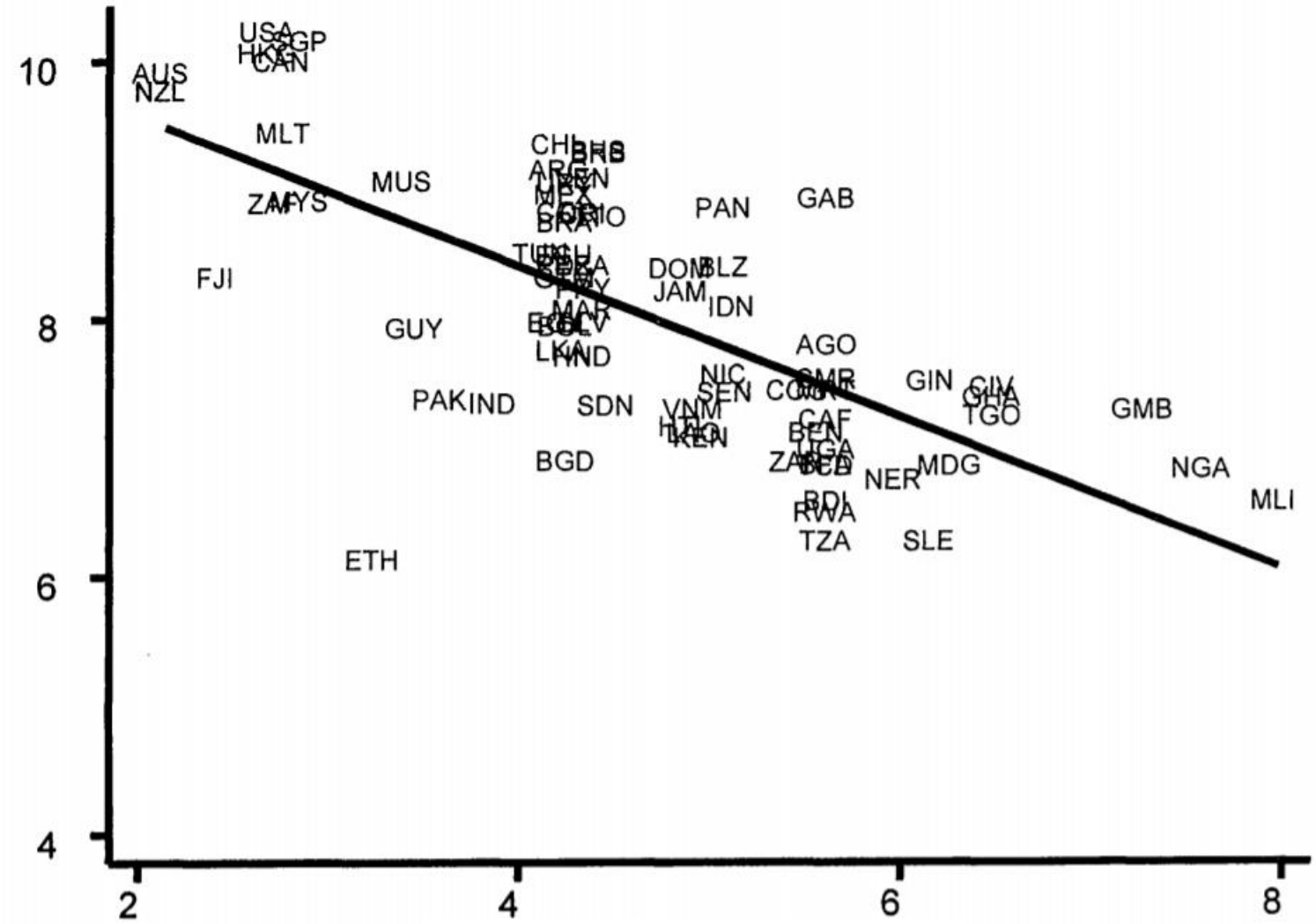


Source: Data Visualization as Narrative, Gelman and Dewitt

Example:

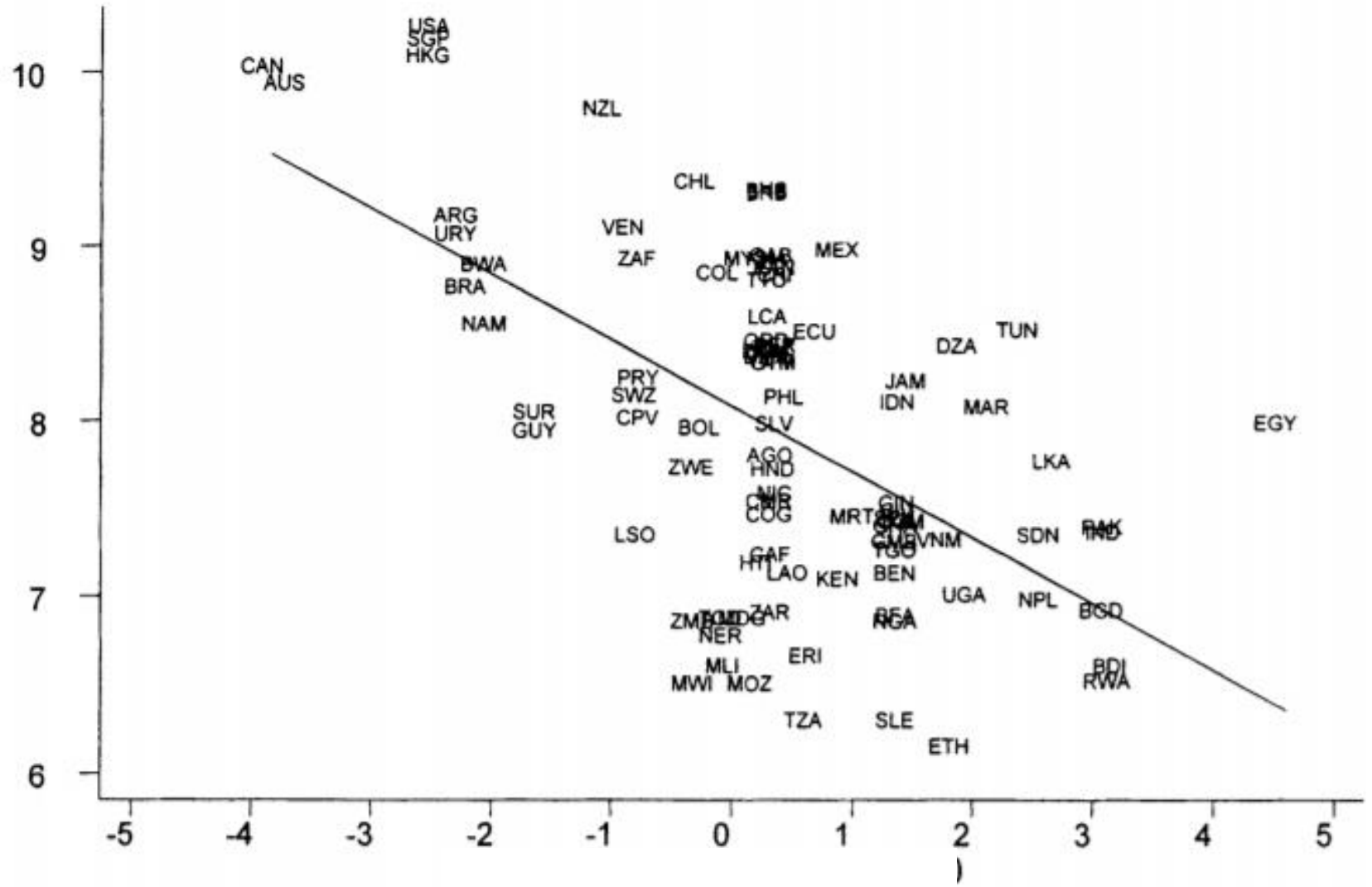
Acemoglu, Johnson, and Robinson, and Why Some Nations Became Richer Than Others

GDP per capita (logged) in 1995



Settler Mortality in 1500

GDP per capita (logged) in 1995

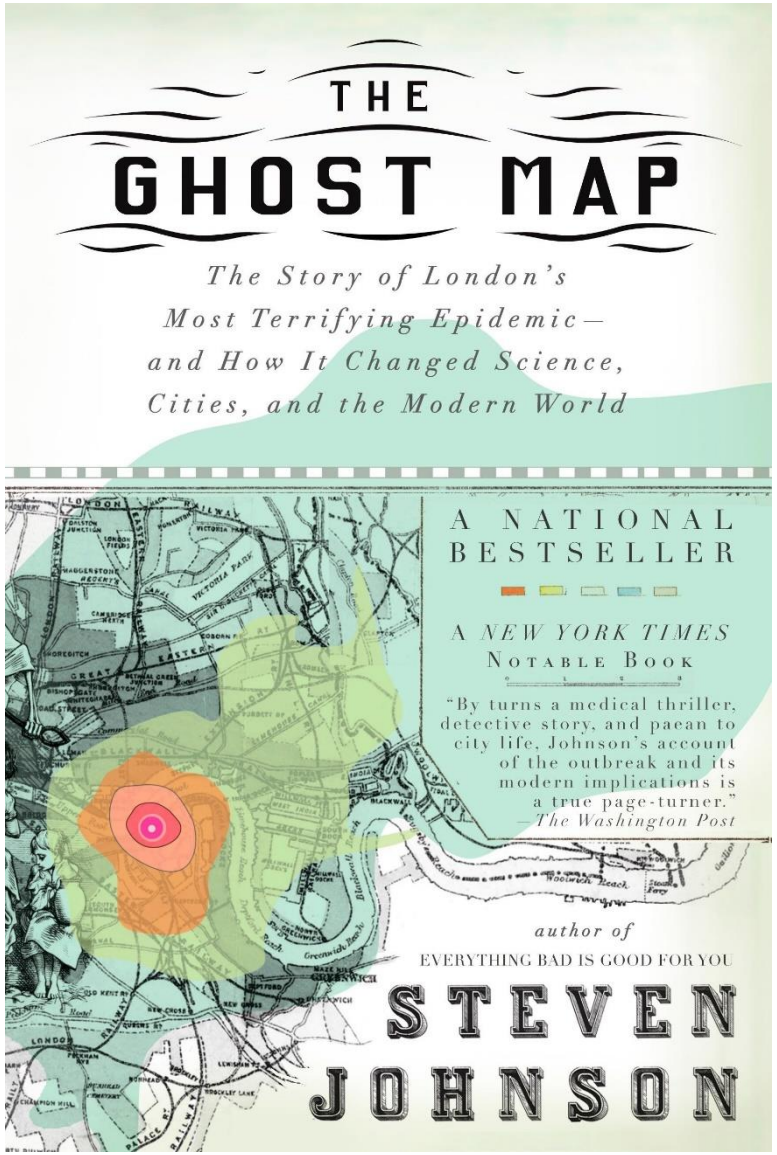


Population Density (logged) in 1500

Data does not by itself enable learning.

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

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



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

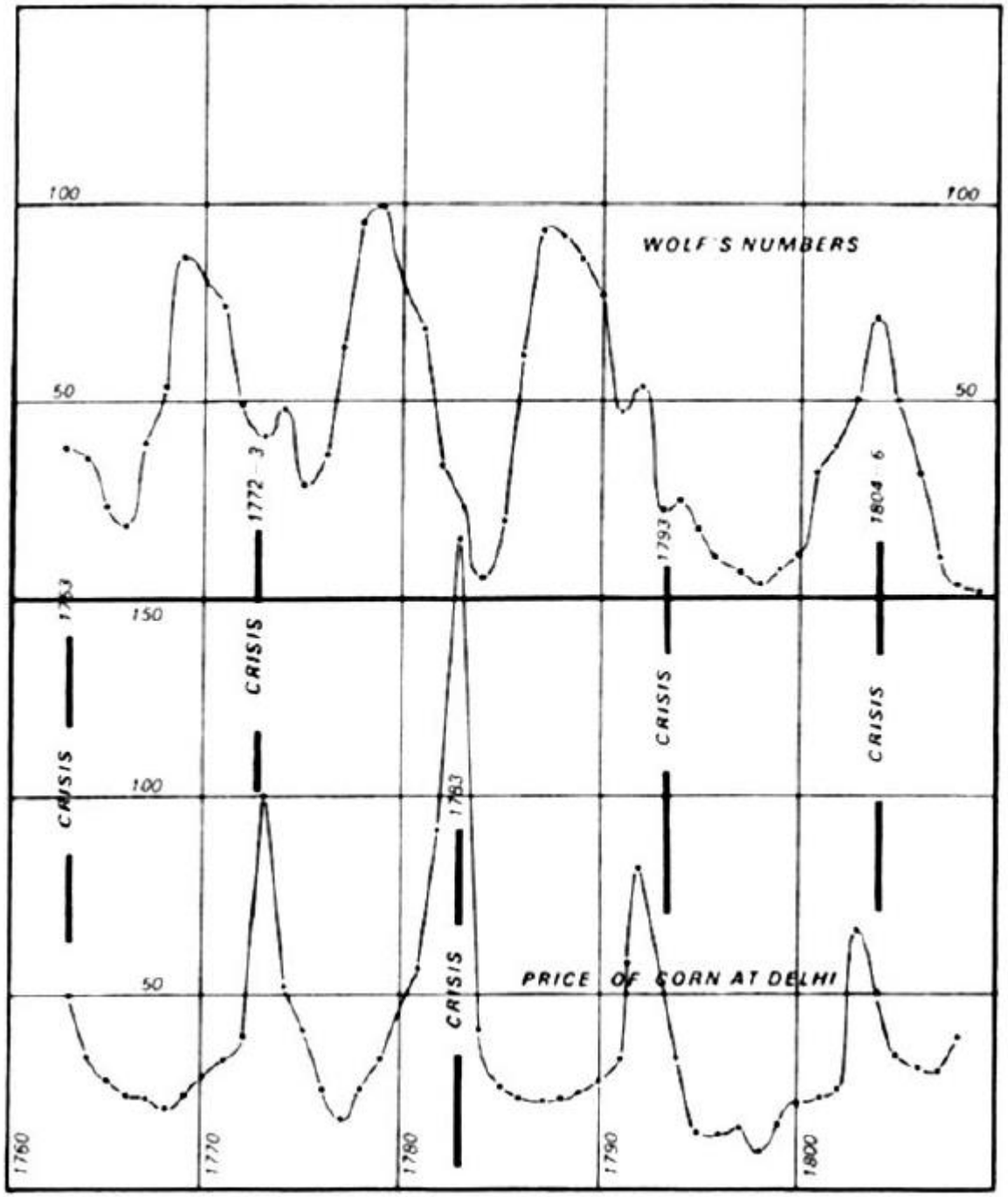
How could they ignore so much overwhelming evidence that contradicted their most basic theories?"

In order to learn from data we need to both speculate and **be willing to be wrong**.

Discovering that our models are wrong is **good**!

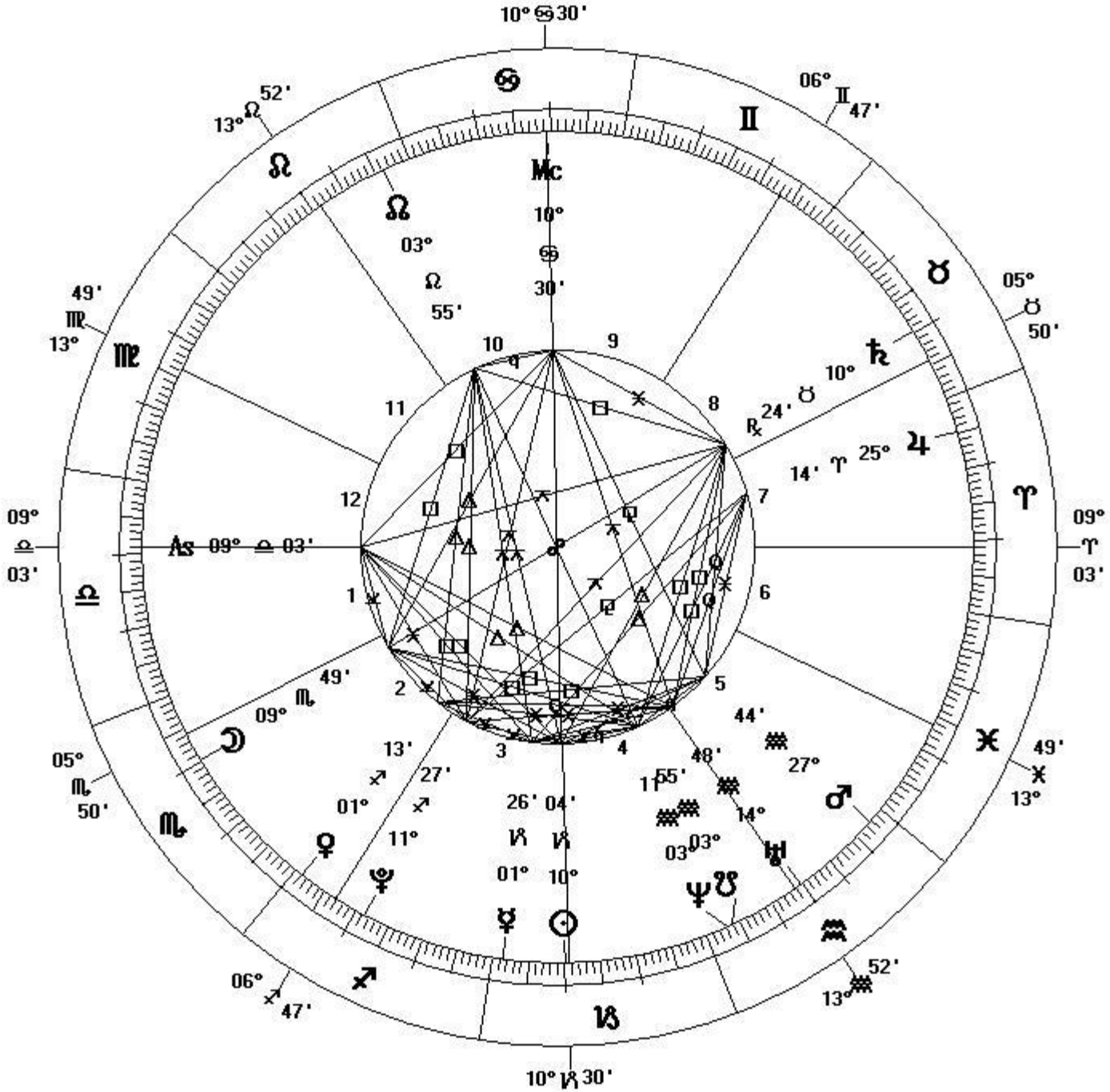
Refusing to acknowledge that they are wrong is **bad**!

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 evidence, **they 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"

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

Speculating about the world, building models, developing their implications, and putting them to the test is **how we learn**.

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

Let’s look at an example in some depth.

Let’s look at an example in some depth.

hold onto your butts.

2 Building Models of the Texas Stars

2 Building Models of the

And Every Other Minor League Hockey Player

As a side project, a few members of our team at AE started working with **AHL (minor league hockey) data**.

We discovered we had a contact within the **Milwaukee Admirals** organization.

As a side project, a few members of our team at AE started working with **AHL (minor league hockey) data**.

We discovered we had a contact within the **Milwaukee Admirals** organization.

We eventually ended up meeting with their coaches and front office to show them some advanced analytics. **I'll show that to you, but mainly I want to show the process of how we got there.**


BRADLEY CENTER
BRADLEY CENTER
WORLD CHAMPIONS
1971

If you observe enough hockey games, you start to ask questions:

Why are some **hockey teams and players** better than others?

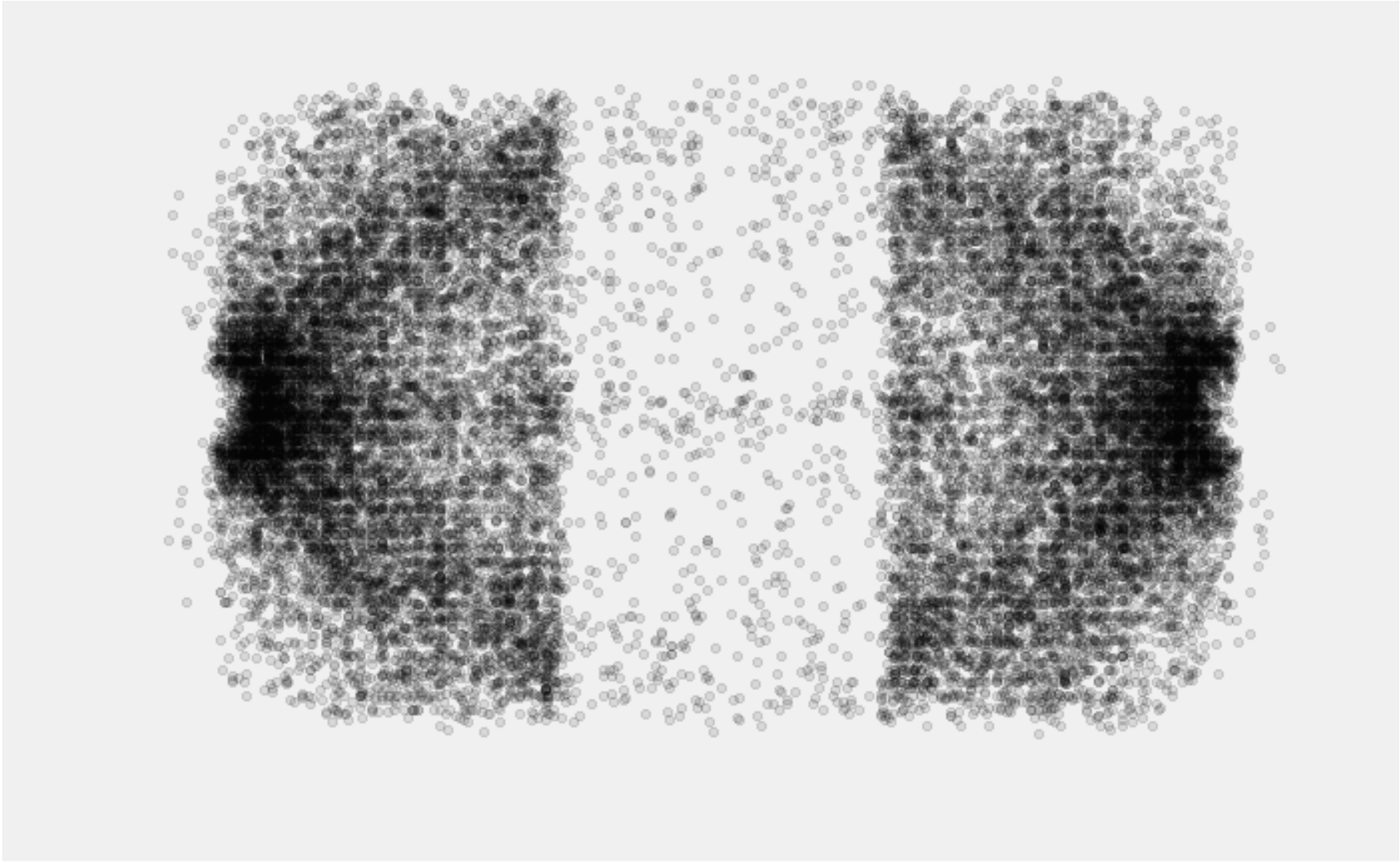
You win hockey games by scoring more goals than your opponent. Better players are able to generate more shots and goals than others.

You win hockey games by scoring more goals than your opponent. Better players are able to generate more shots and goals than others.

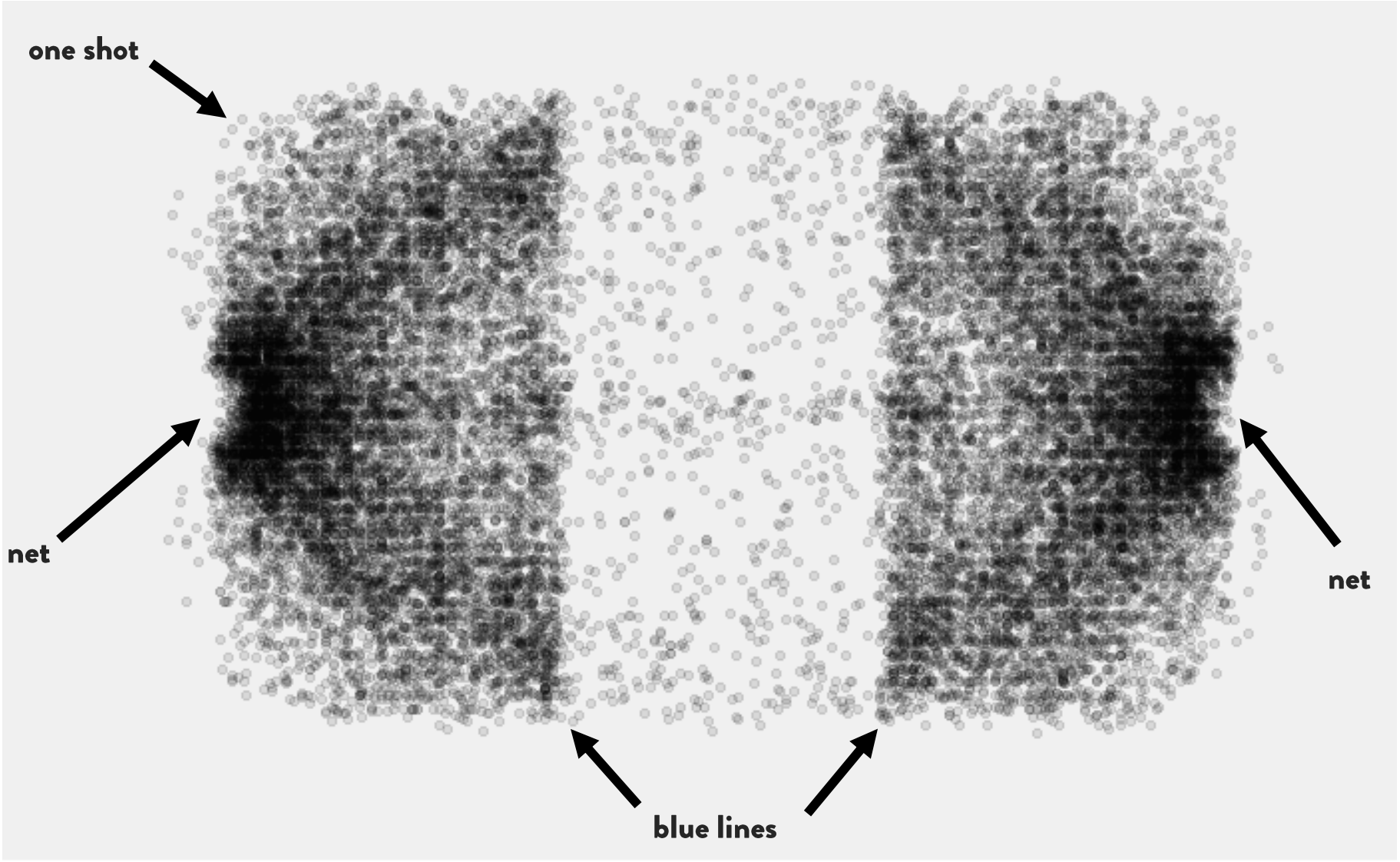
Why do some **hockey players** score more goals than others?
Why do some **hockey goalies** stop more goals than others?

We can get data on every shot taken by every player for every AHL team in every game from 2018-2020.

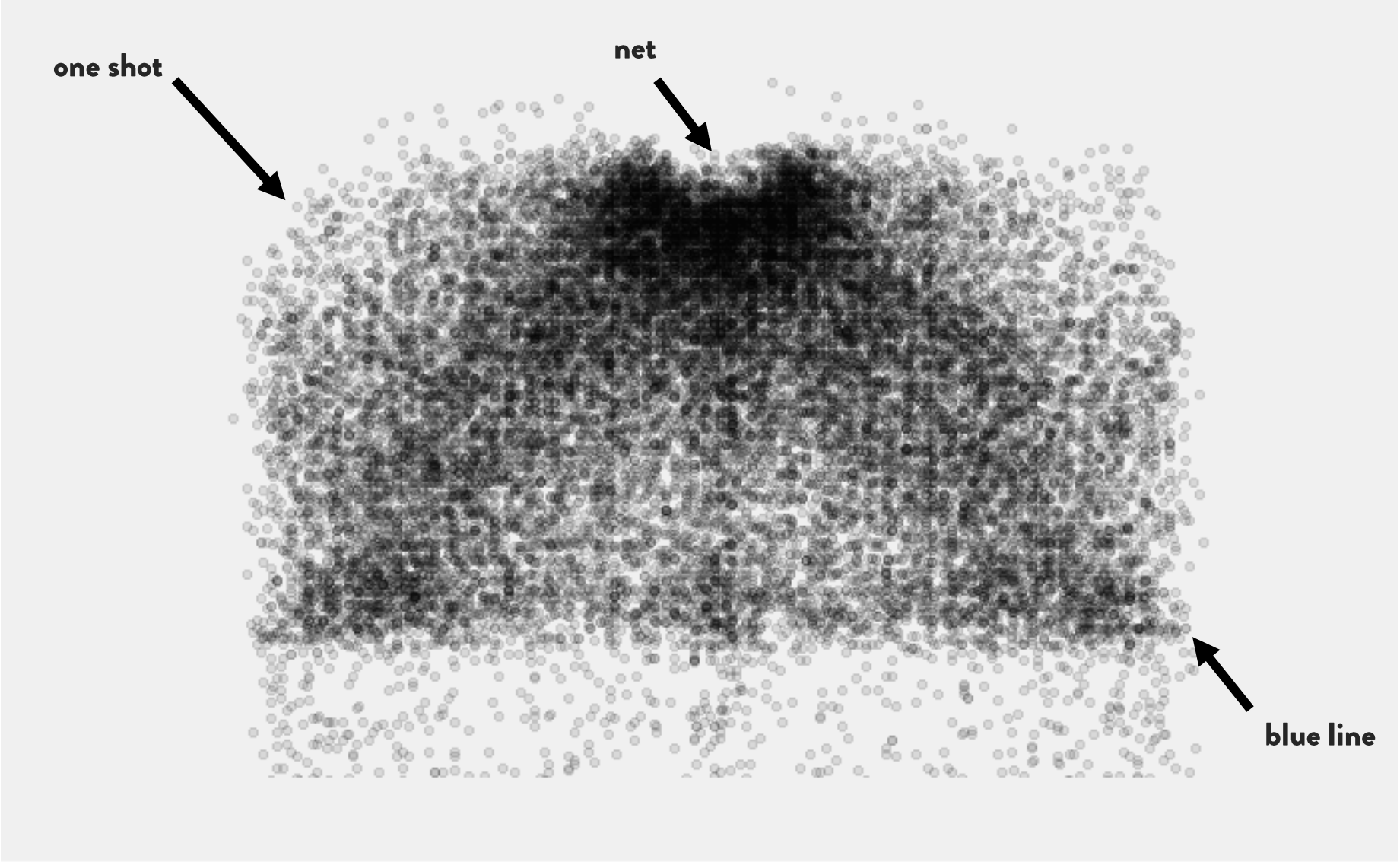
We can get data on every shot taken by every player for every AHL team in every game from 2018-2020.



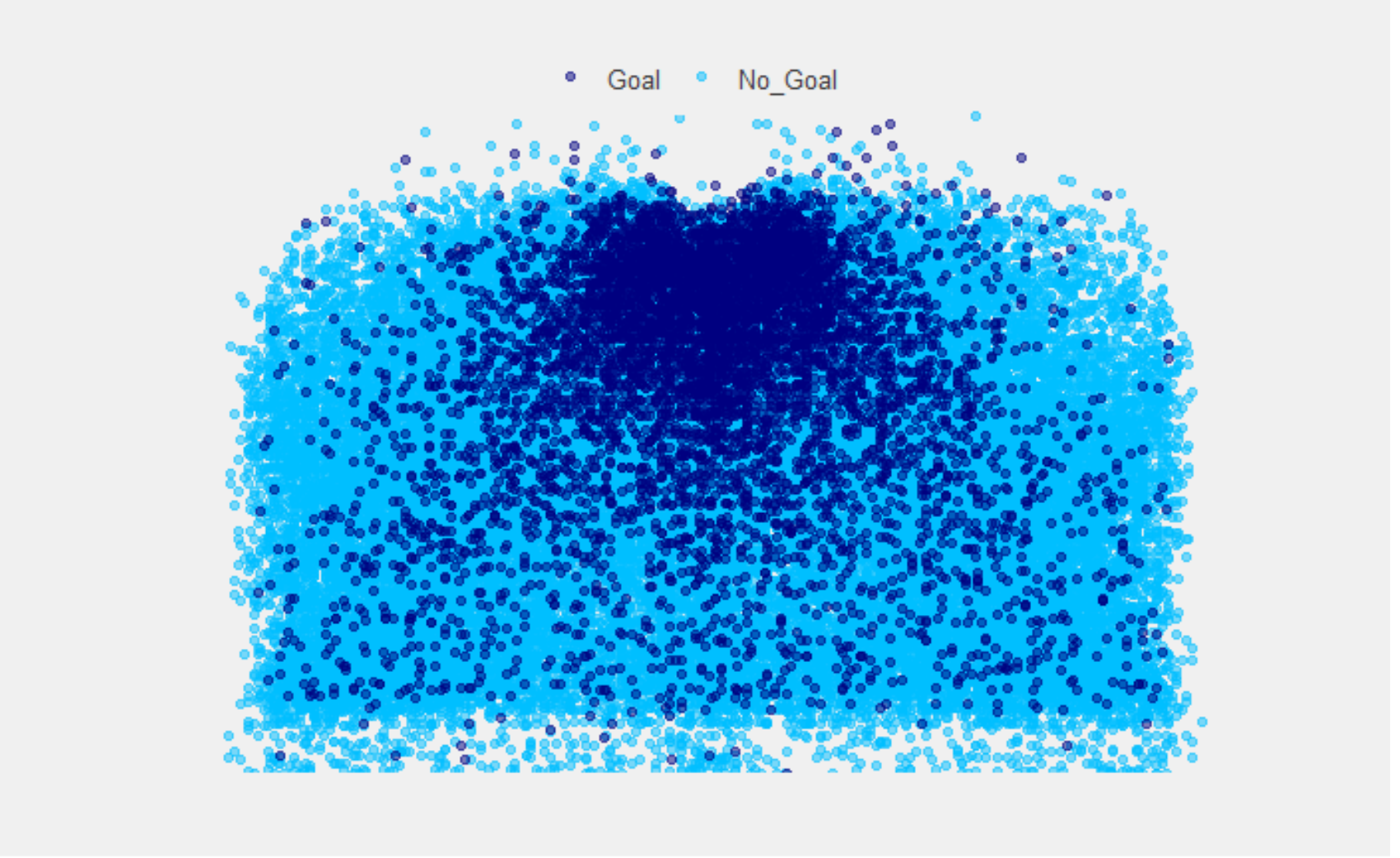
We can get data on every shot taken by every player
for every AHL team in every game from 2018-2020.



We can get data on every shot taken by every player for every AHL team in every game from 2018-2020.



We know the location, the player, the goalie, and the situation in the game. But most importantly, we know whether a goal was scored.



Does this data tell us which players are better than others?

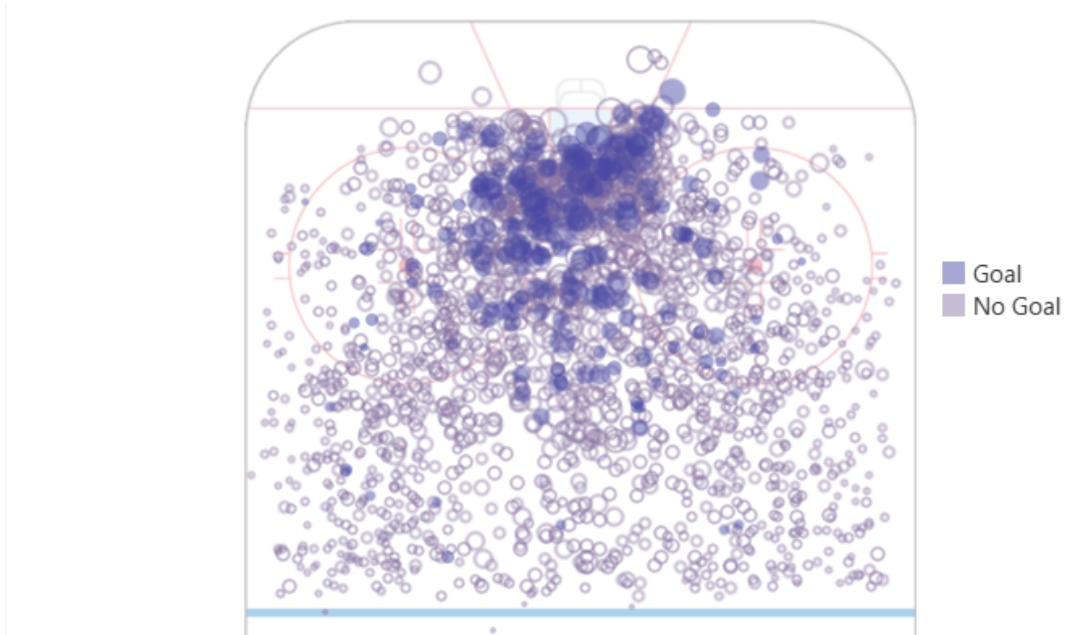
Does this data tell us which players are better than others?

Kind of. We can look at who scores the most goals, which goalies have the most saves, that sort of thing.

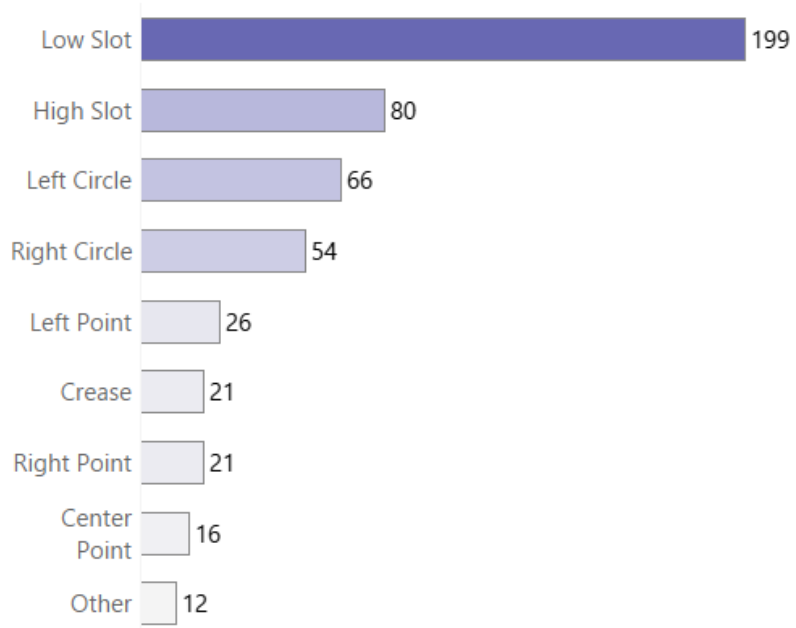
MIL Offensive Breakdown 2018

Games	Shots	Goals	Shooting %
71	1,958	176	8.99%

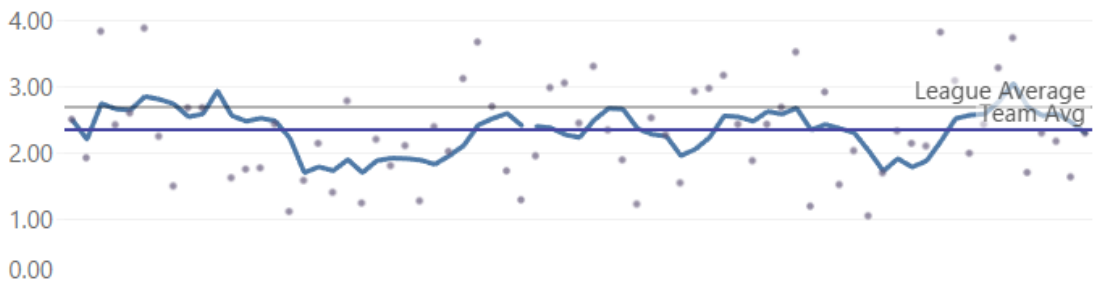
Goals



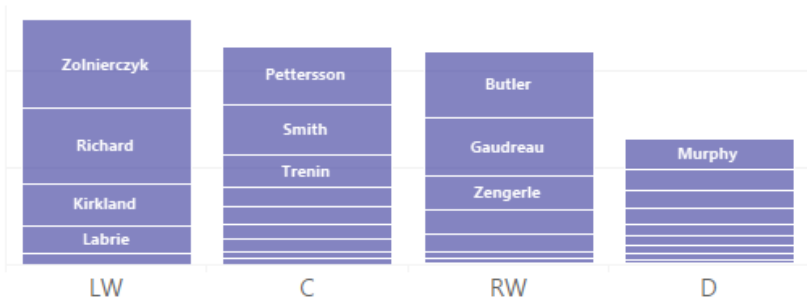
Goals by Scoring Area



Goals by Game



Goals by Player



Filters

Season: 2018

Team Abbr: MIL

Opp: (All)

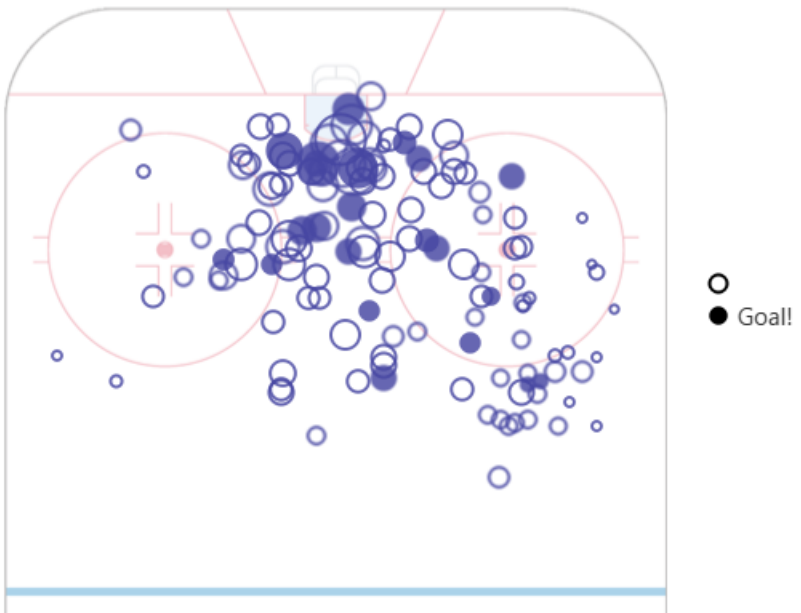
xGoals Model: Baseline

Algorithm Selector: Gradient Boosted...

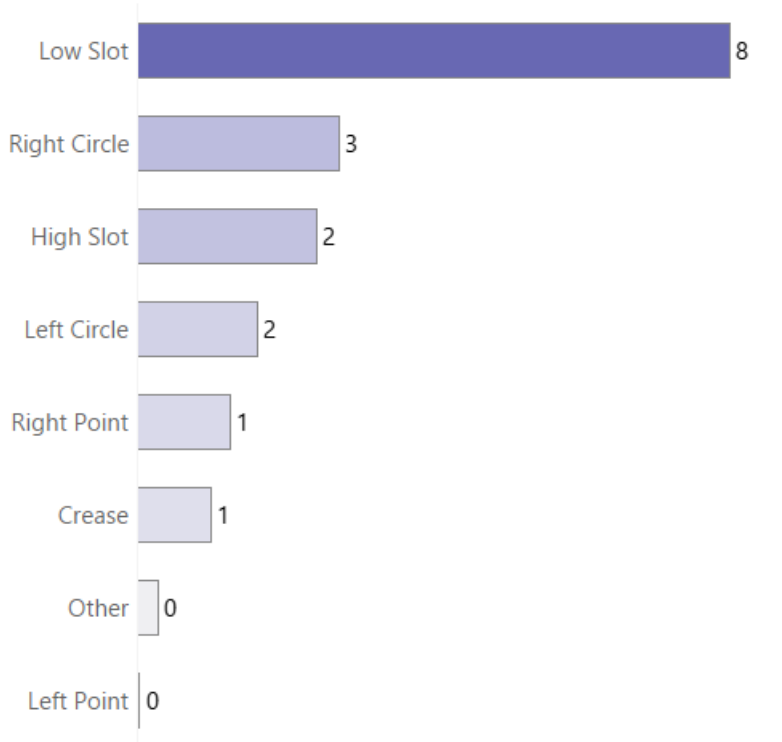
Daniel Carr Offensive Breakdown

Position	Games	Shots	Goals	Shooting %	Goals per Game
LW	45	131	22	16.79%	0.361

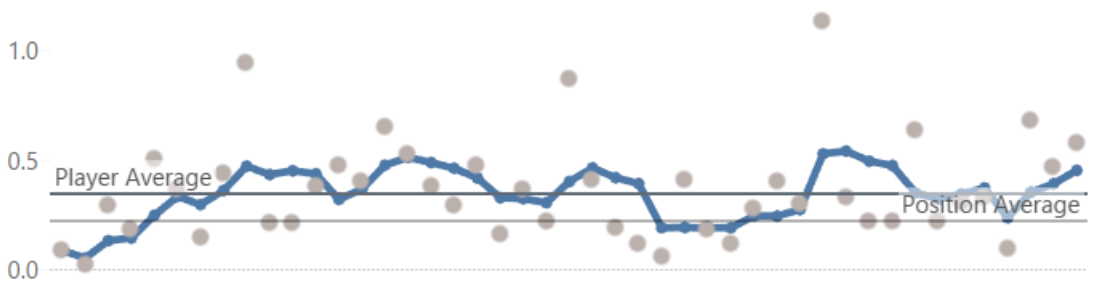
Shots and Goals



Goals by Scoring Area



Goals per Game



Filters

Select Season

(All)

Select a Team

MIL

Select a Player

Daniel Carr

Select Opponent

(All)

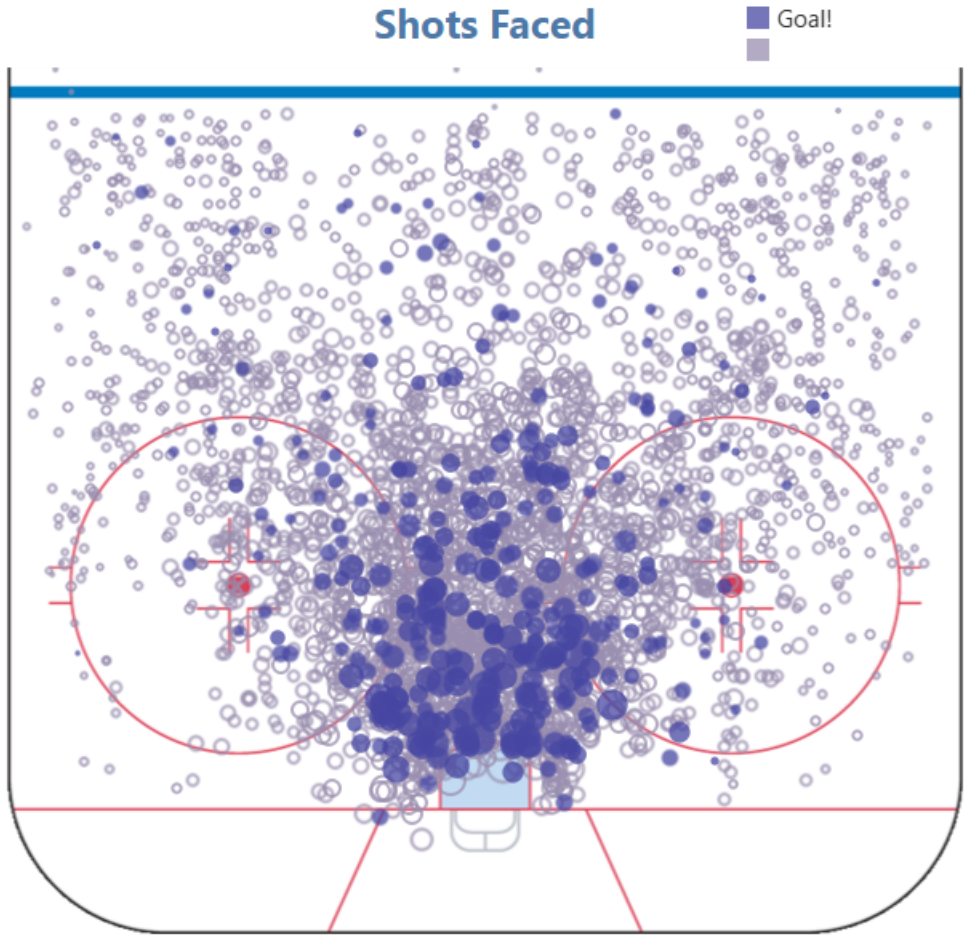
Model Selector

Baseline

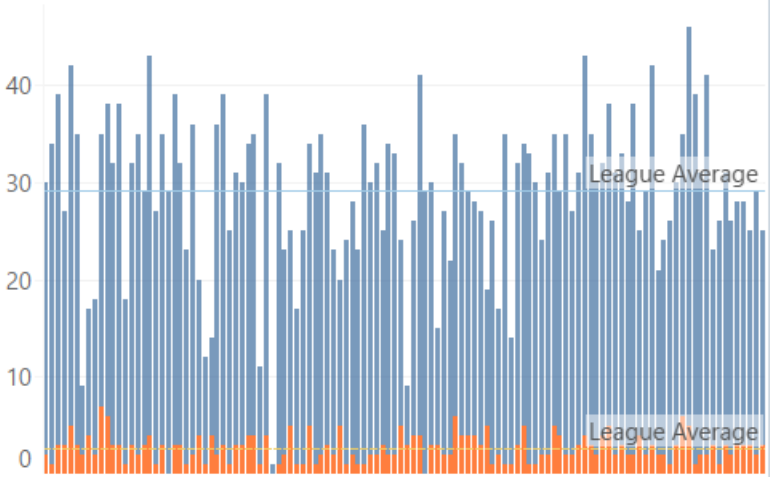
Goalie: Landon Bow

Games	Shots	Goals Allowed	Save %
118	1,091	318	93.03%

Shots Faced



Shots and Goals by Game



Goalie Save % Over Time

Filters

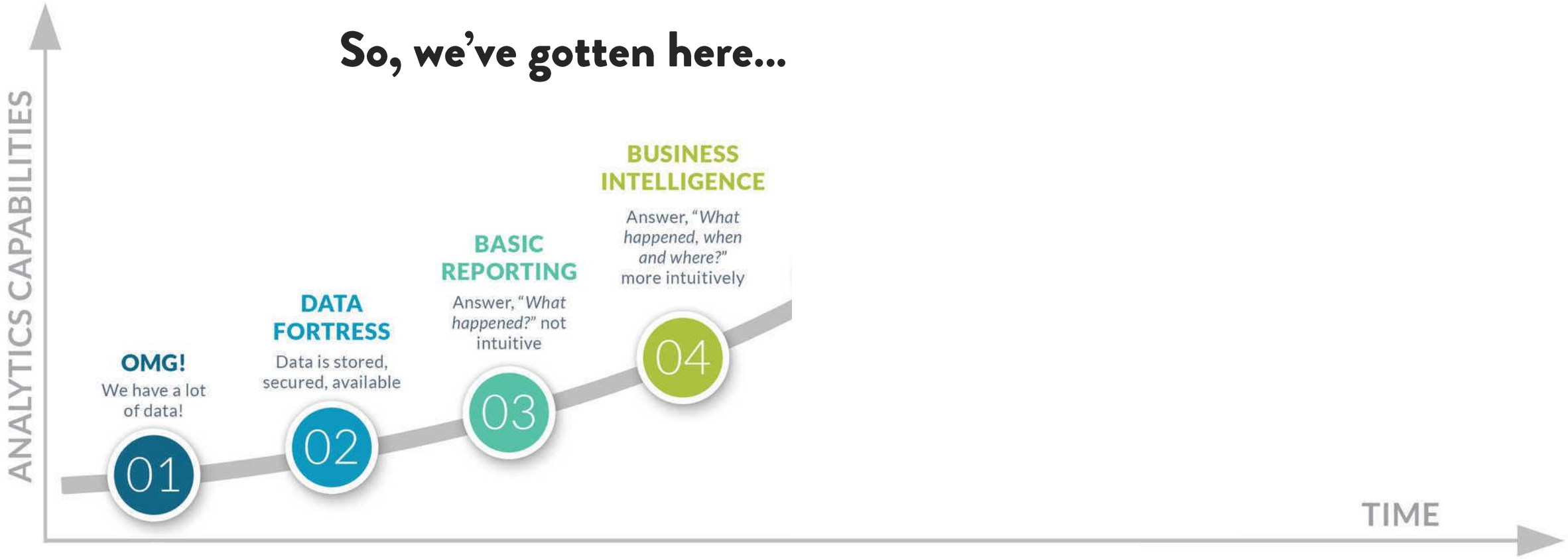
Season

(All)

Select Goalie

Landon Bow

So, we've gotten here...



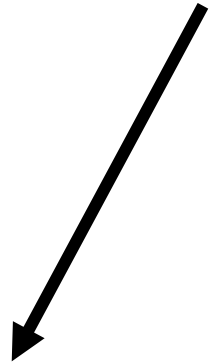
But the data, by itself, doesn't tell us things we want to know.

But the data, by itself, doesn't tell us things we want to know.

Hockey is a game of **luck** and **skill**.
Sometimes teams and players just **get lucky**.

Hockey is a game of **luck** and **skill**.
Sometimes teams and players just **get lucky**.

We want to try to measure this.



Hockey is a game of **luck** and **skill**.
Sometimes teams and players just **get lucky**.

This is where advanced analytics come in –
we can build models.

For **every shot taken**, we want to know:
how likely was it become a goal?

In other words, if you took that exact shot,
say, 1000 times, **how many times would you expect it become a goal**?

December 6, 2017
San Diego Gulls vs Chicago Wolves
Giovanni Fiore 1:33 into the 1st Period

Goals: 0
xGoals: ?

December 6, 2017
San Diego Gulls vs Chicago Wolves
Giovanni Fiore 1:33 into the 1st Period

Goals: 0
xGoals: ?

March 13, 2019
Texas Stars vs Chicago Wolves
Joel Hanley 11:56 into the 2nd Period

Goals: 1
xGoals: ?

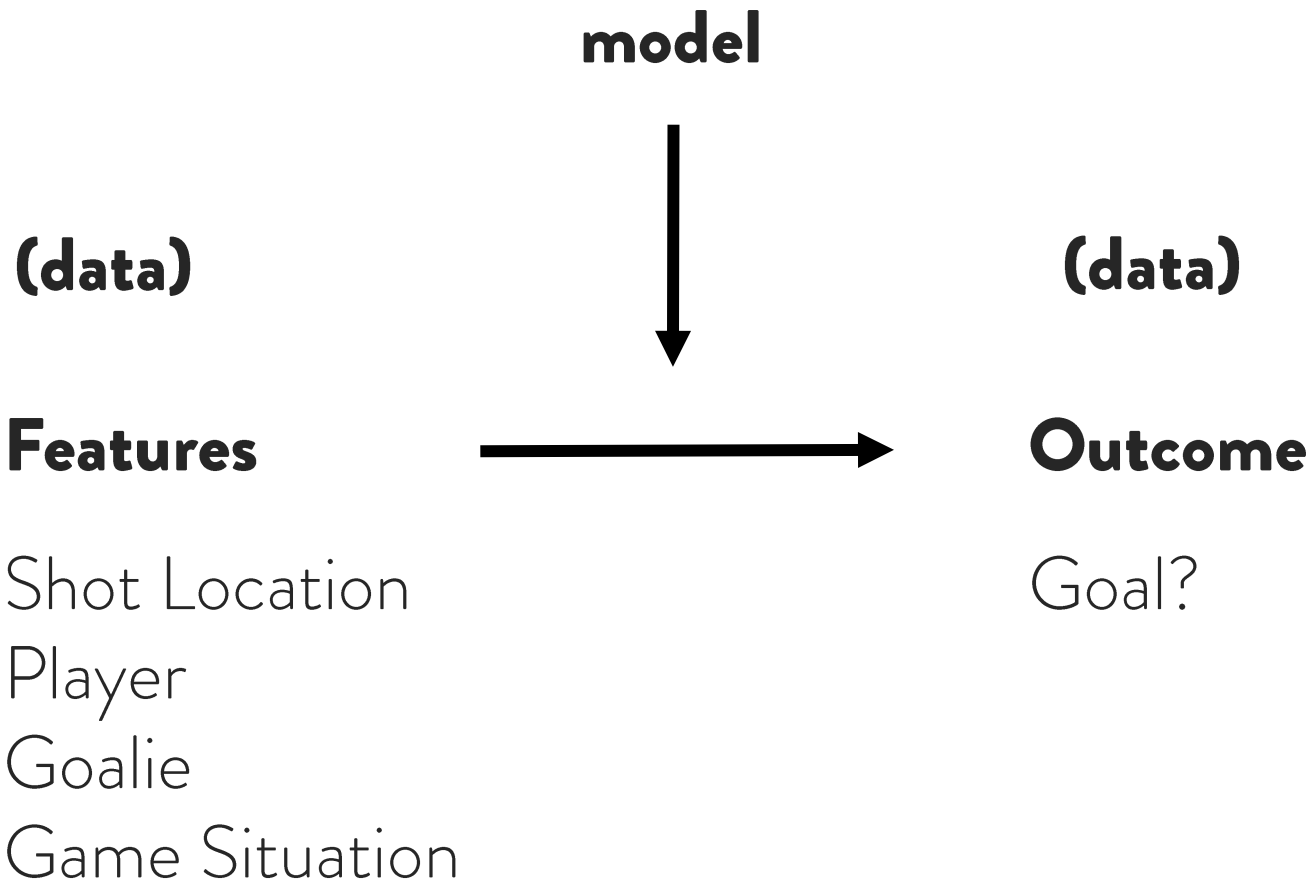
March 13, 2019
Texas Stars vs Chicago Wolves
Joel Hanley 11:56 into the 2nd Period

Goals: 1
xGoals: ?

We need a model that predicts the probability that a shot will become a goal.

We need a model that predicts the probability that a shot will become a goal.

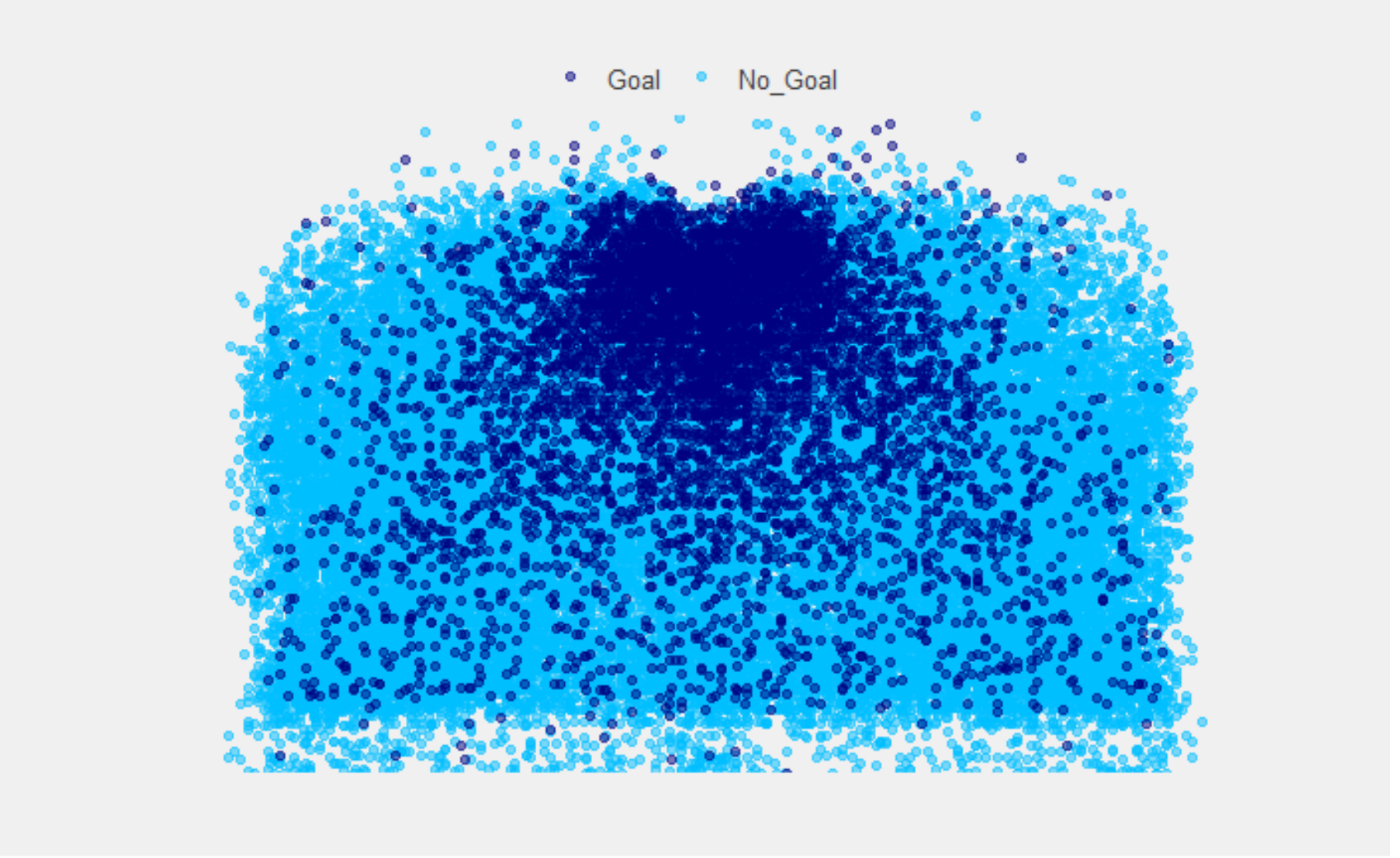
Features		**Outcome**
Shot Location Player Goalie Game Situation	→	Goal?


model
(data)
(data)
Features
Outcome
Shot Location
Player
Goalie
Game Situation
Goal?

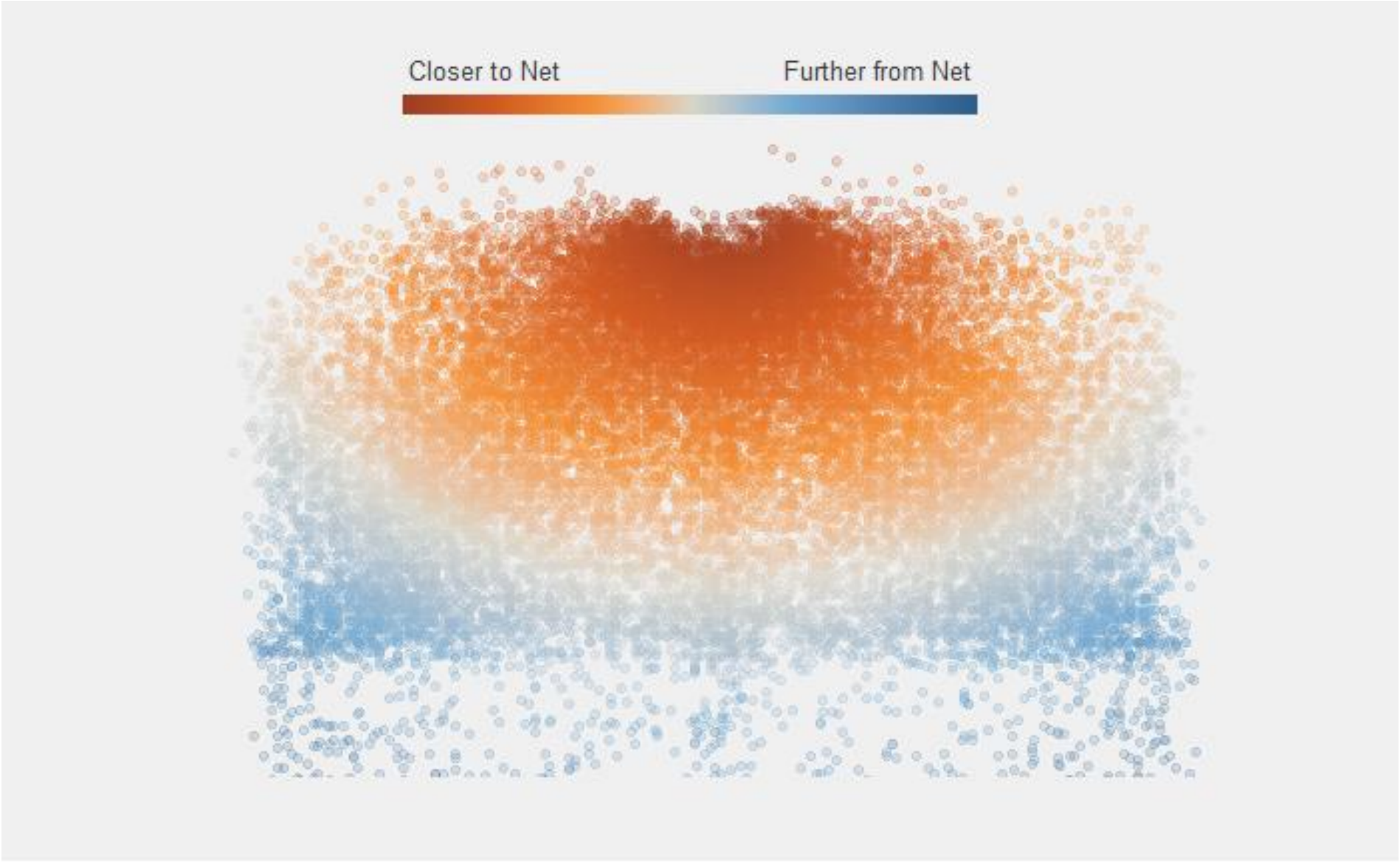
In order to get a model to learn this relationship, we need to **speculate about features** that could be useful in predicting goals.

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

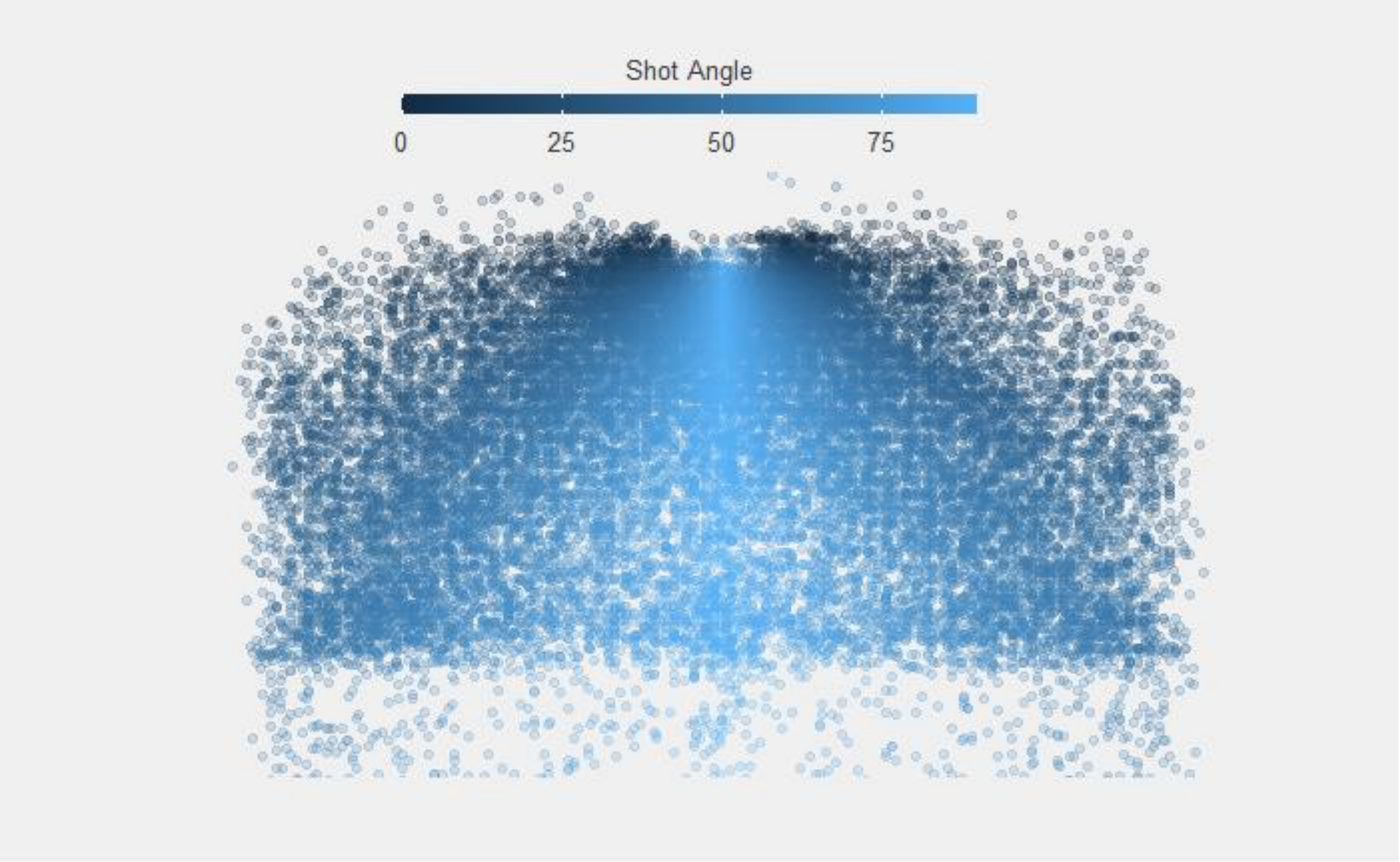
We observe some of the data and start to speculate about the process that leads to goals.



First, we can compute the distance of each shot to the net.



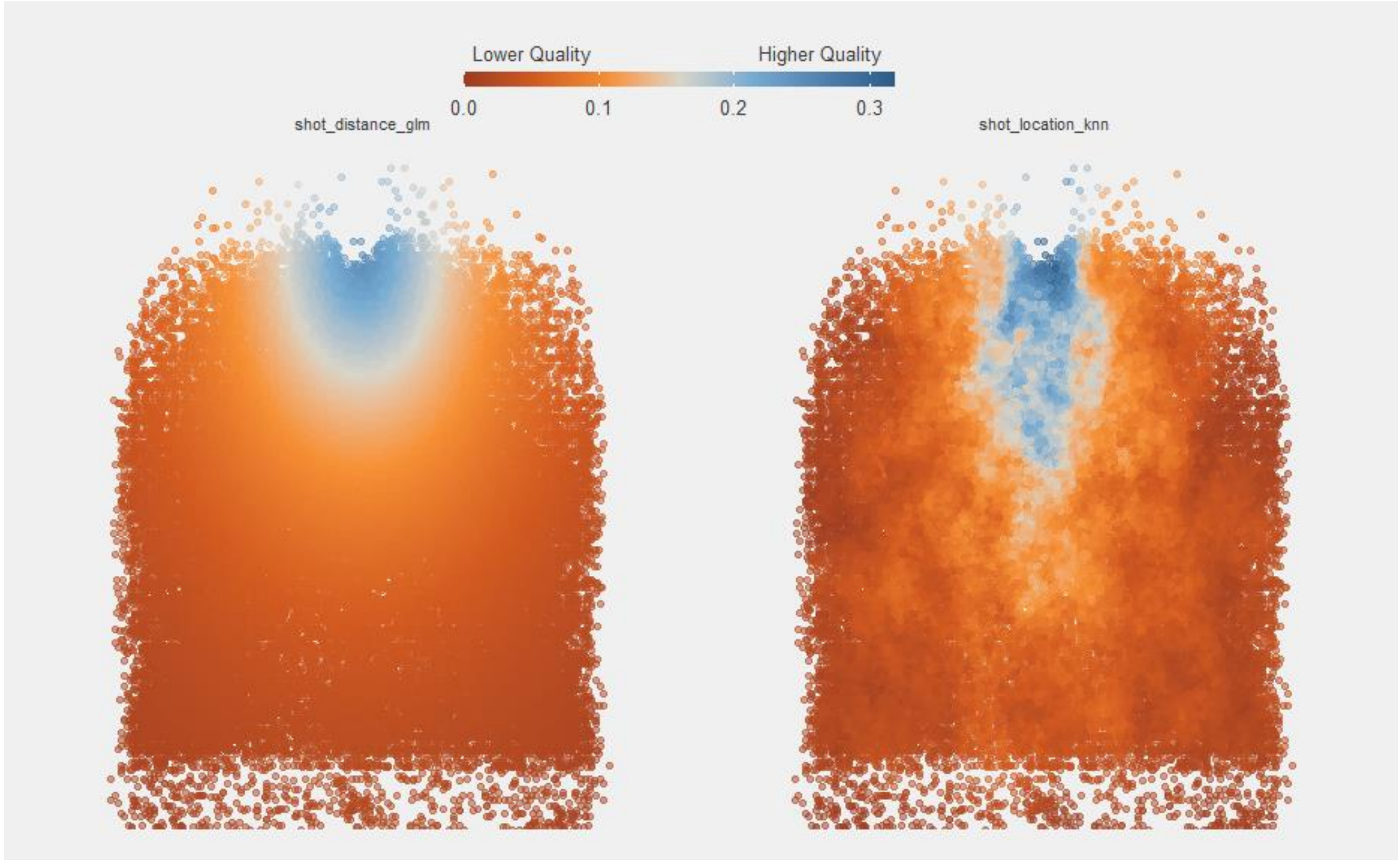
We can then struggle to recall high school geometry and eventually calculate the shot angle.



Using just shot distance and angle, can we predict which shots become goals?

To answer this, we build a couple models.

This is the pattern they learn:


Lower Quality
Higher Quality
0.0
0.1
0.2
0.3
shot_distance_glm
shot_location_knn

December 6, 2017
San Diego Gulls vs Chicago Wolves
Giovanni Fiore 1:33 into the 1st Period

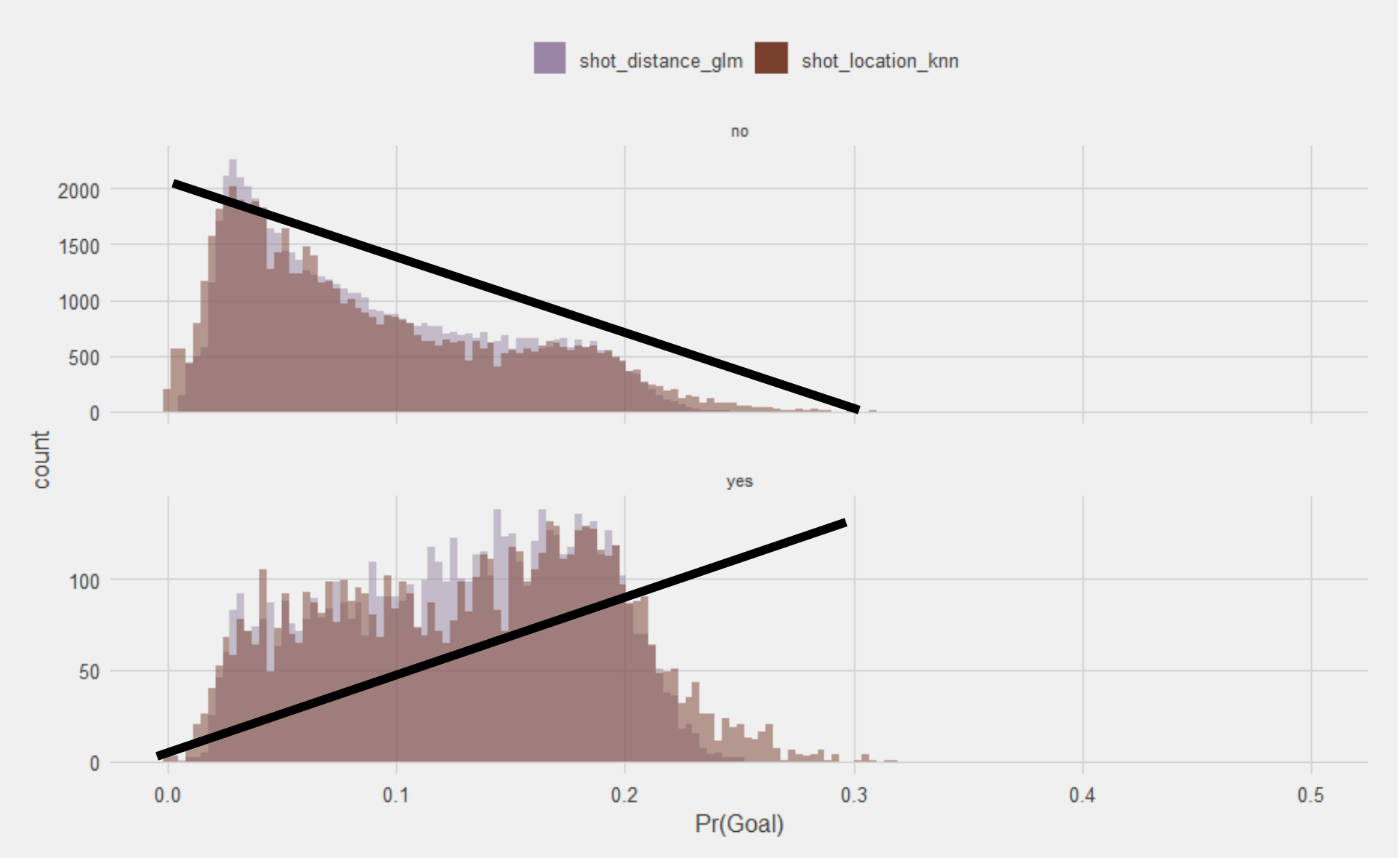
xGoals Distance Model: 0.251
xGoals Location Model: 0.232

March 13, 2019
Texas Stars vs Chicago Wolves
Joel Hanley 11:56 into the 2nd Period

xGoals Shot Situation Model: .0041

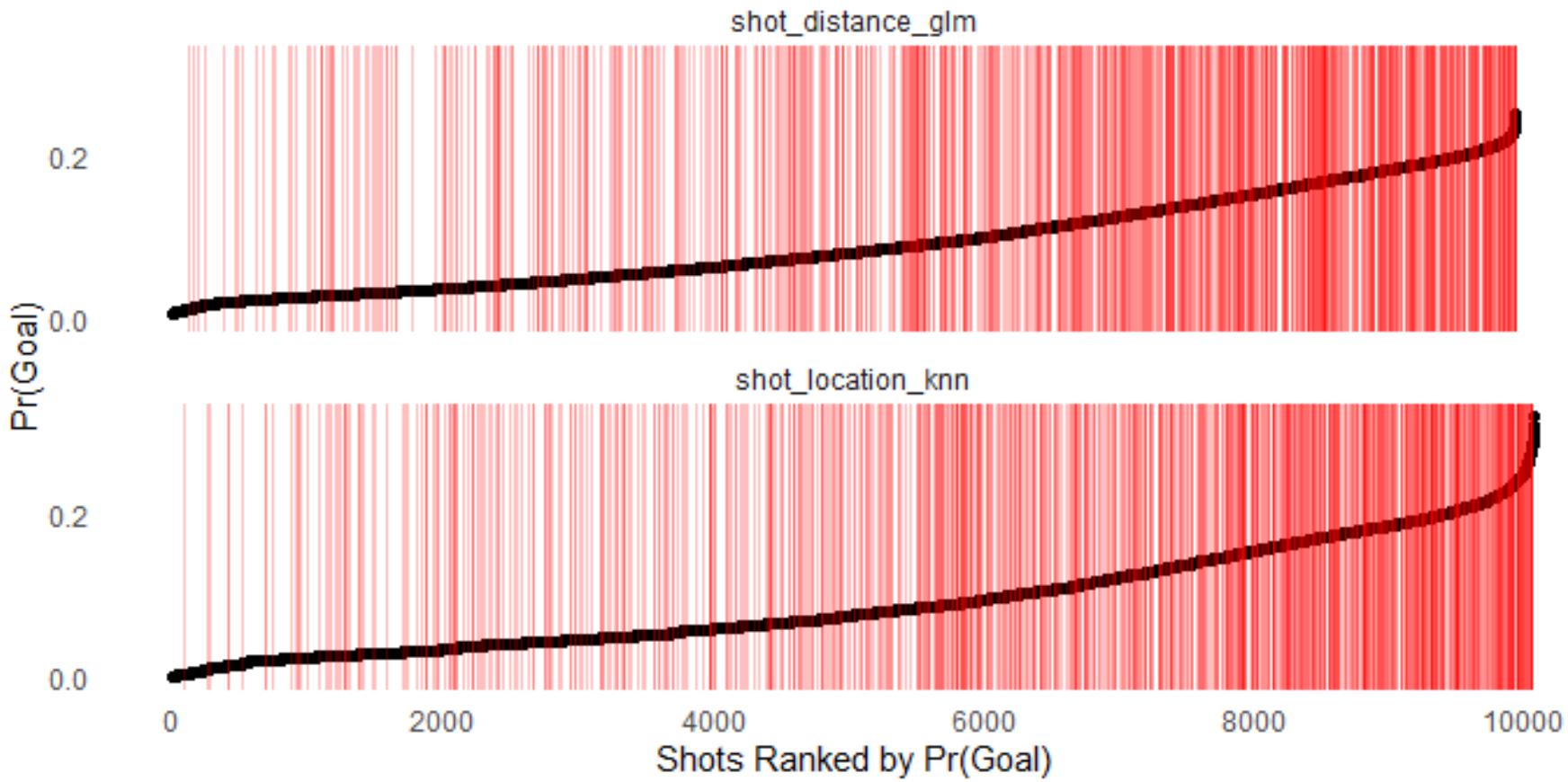
Are the models good?

Well, they're better than *not having one*.


shot_distance_glm
shot_location_knn
no
yes
count
2000
1500
1000
500
0
100
50
0
0.0
0.1
0.2
0.3
0.4
0.5
Pr(Goal)

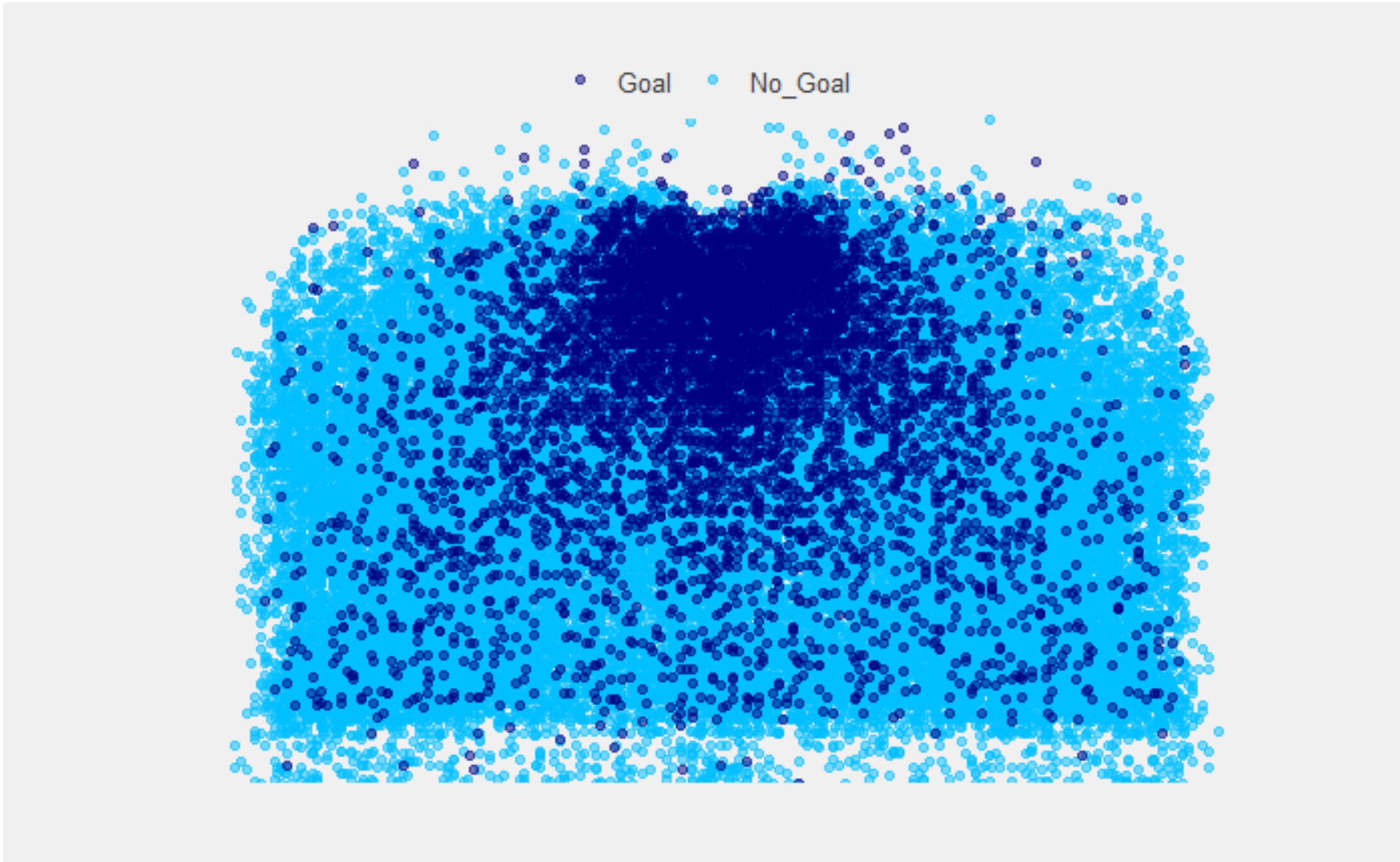
Separation Plots - Simple Models

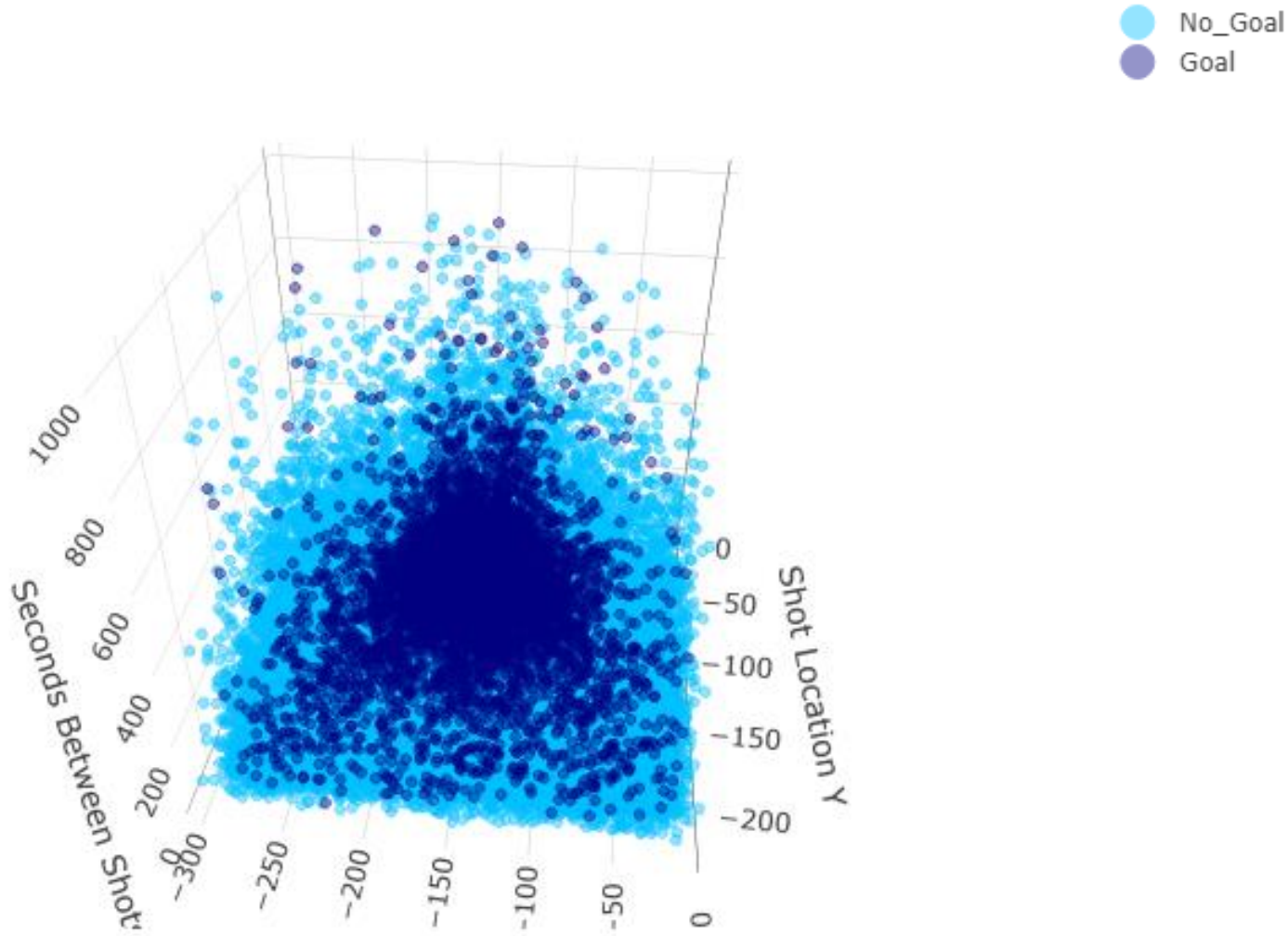
Shots ranked by probability of goal from model. Red line indicates a goal was scored. Random sample of 20,000 shots from out of sample (CV) predictions in training set.

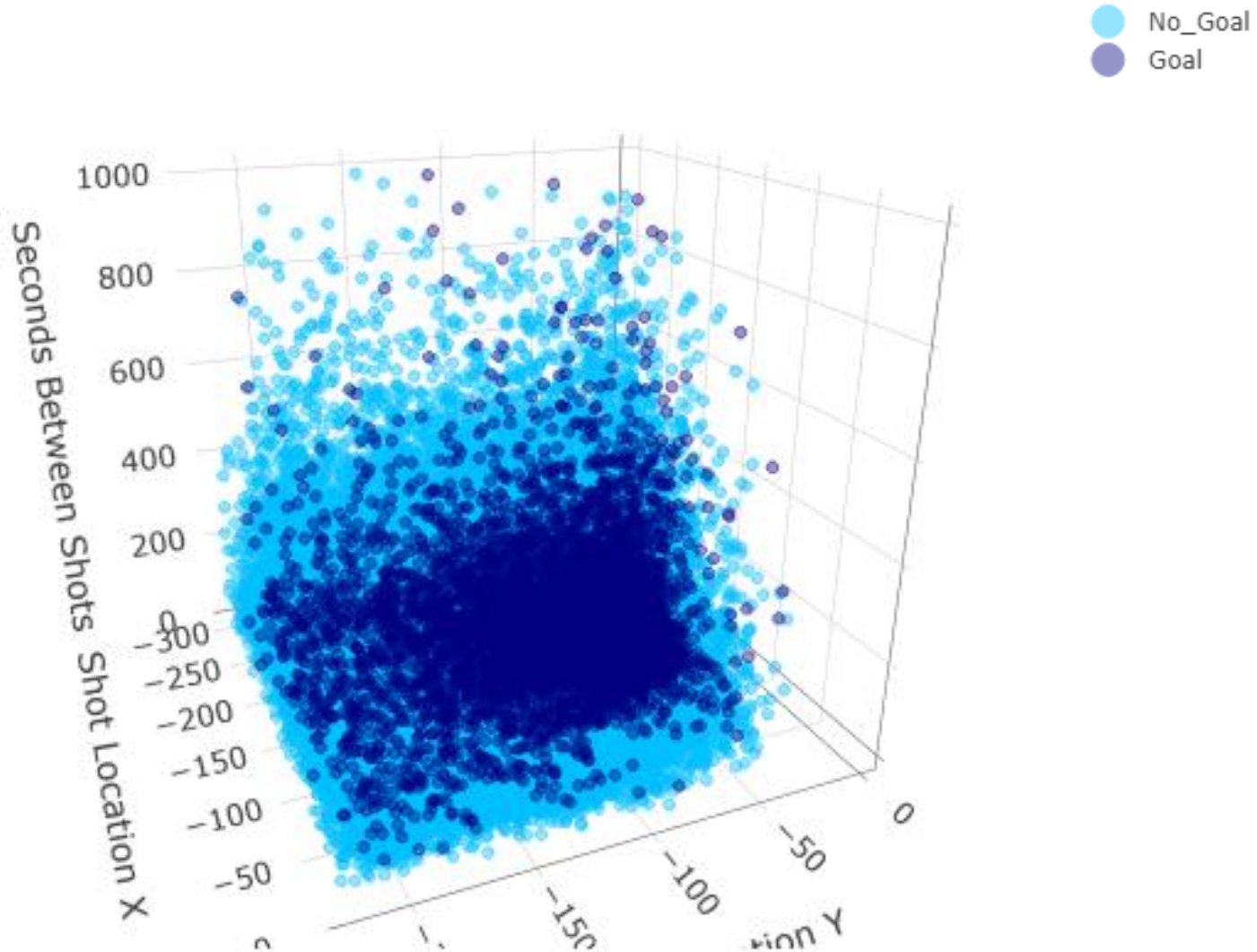


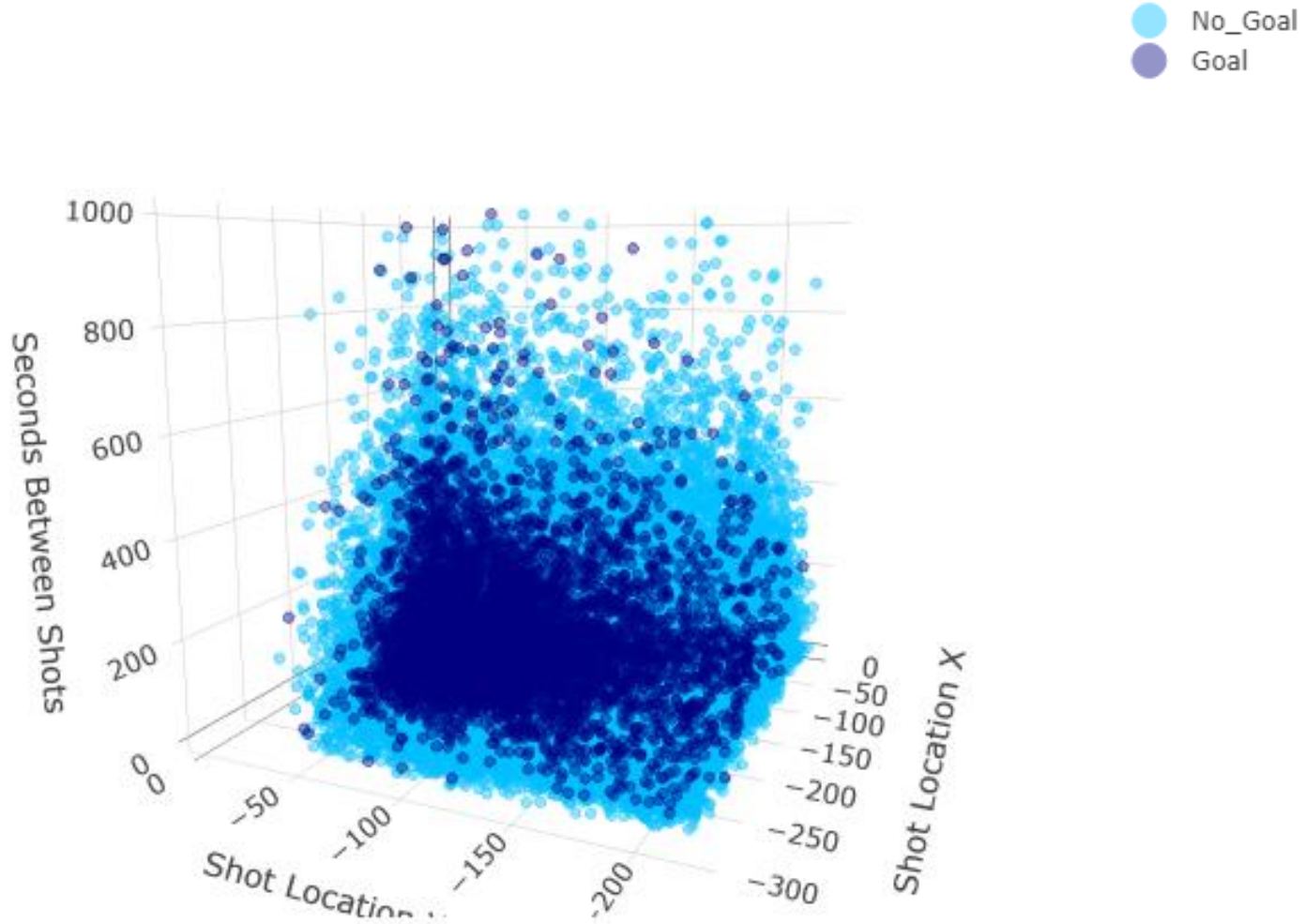
They do a decent job, but they're missing a lot of things that we would expect to matter for predicting goals.

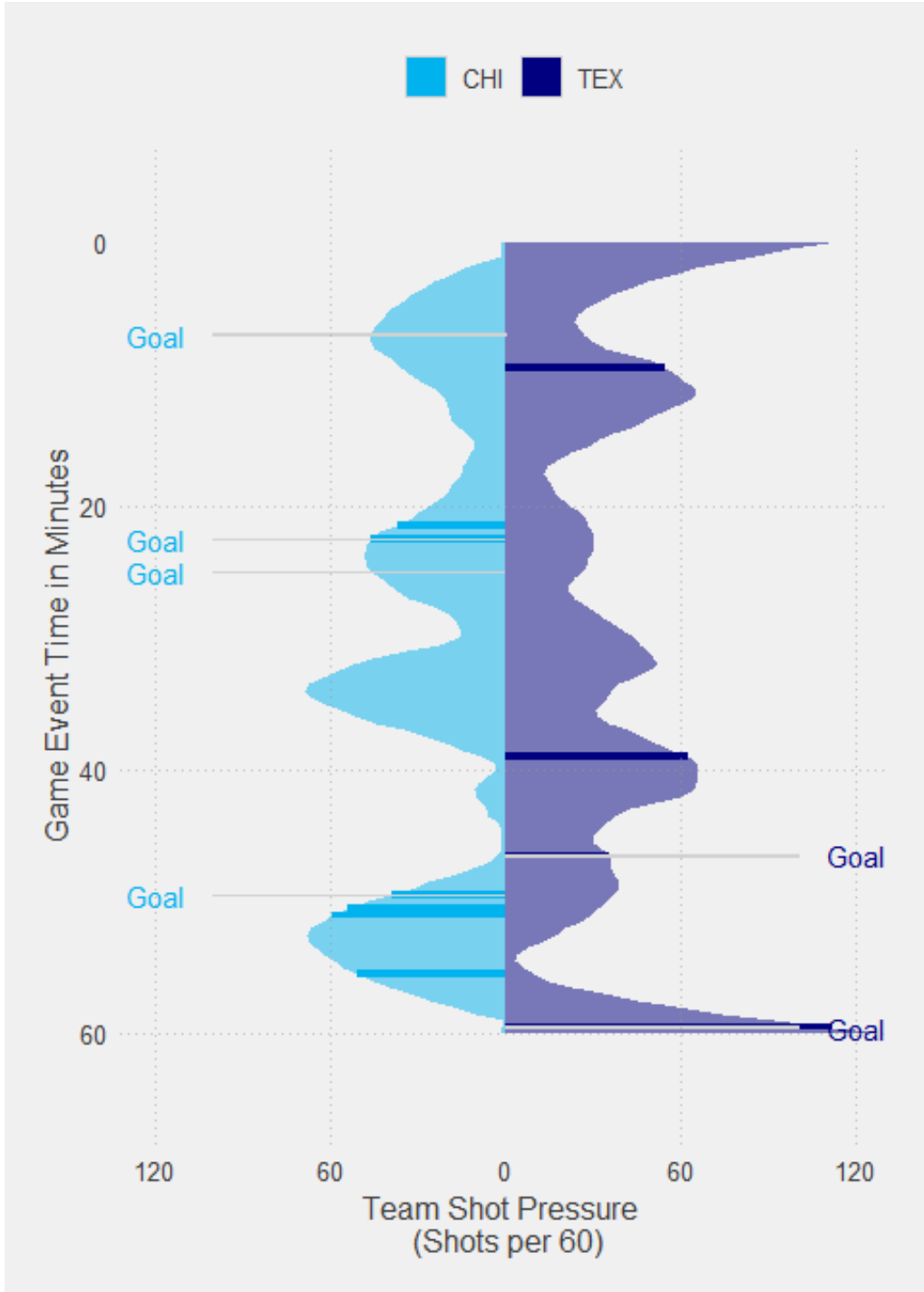
We can speculate about more features to include in a model.

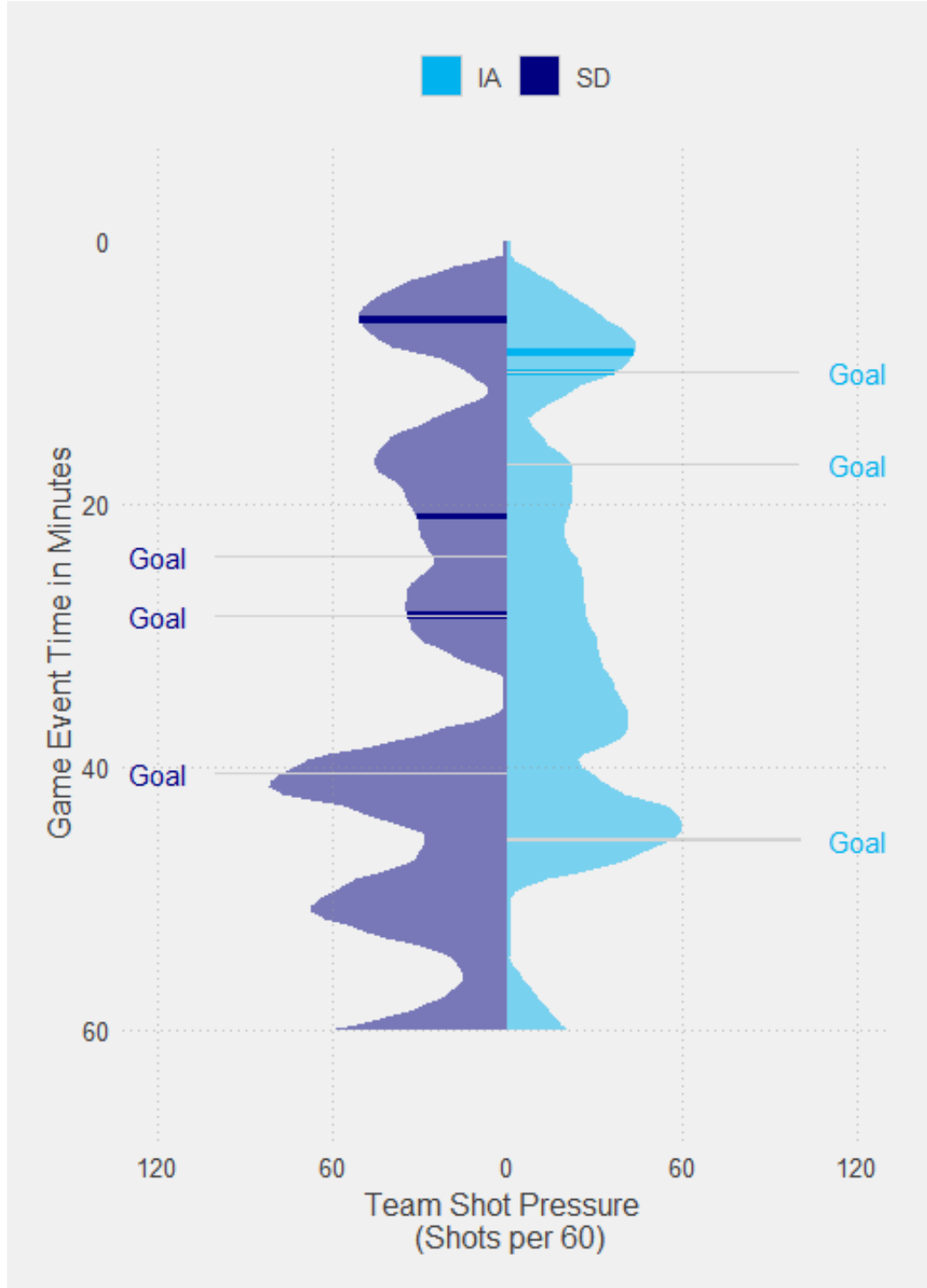

Goal
No_Goal


No_Goal
Goal
Seconds Between Shot
1000
800
600
400
200
0
-300
-250
-200
-150
-100
-50
0
Shot Location Y
0
−50
−100
−150
−200

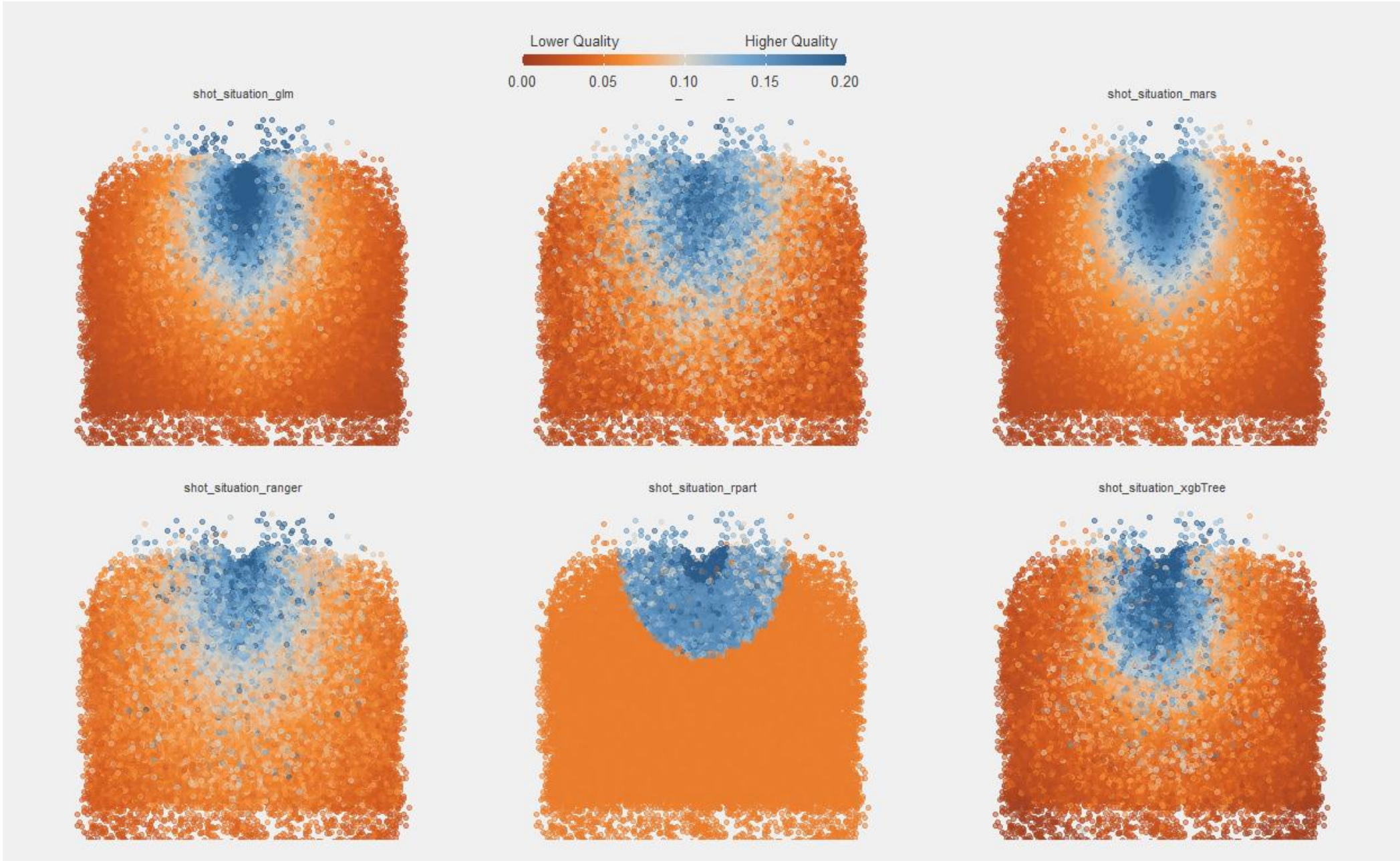

No_Goal
Goal
Seconds Between Shots
1000
800
600
400
200
0
Shot Location X
−300
−250
−200
−150
−100
−50
0
−50
−100
−150


No_Goal
Goal
Seconds Between Shots
1000
800
600
400
200
0
0
−50
−100
−150
−200
Shot Location X
0
−50
−100
−150
−200
−250
−300


CHI
TEX
0
20
40
60
Game Event Time in Minutes
Goal
Goal
Goal
Goal
Goal
Goal
120
60
0
60
120
Team Shot Pressure
(Shots per 60)


IA
SD
0
20
40
60
Game Event Time in Minutes
Goal
Goal
Goal
Goal
Goal
120
60
0
60
120
Team Shot Pressure
(Shots per 60)

We can come up with more features and fit more flexible models, then see whether we improve over the initial models.


Lower Quality
Higher Quality
0.00
0.05
0.10
0.15
0.20
shot_situation_glm
shot_situation_mars
shot_situation_ranger
shot_situation_rpart
shot_situation_xgbTree

December 6, 2017

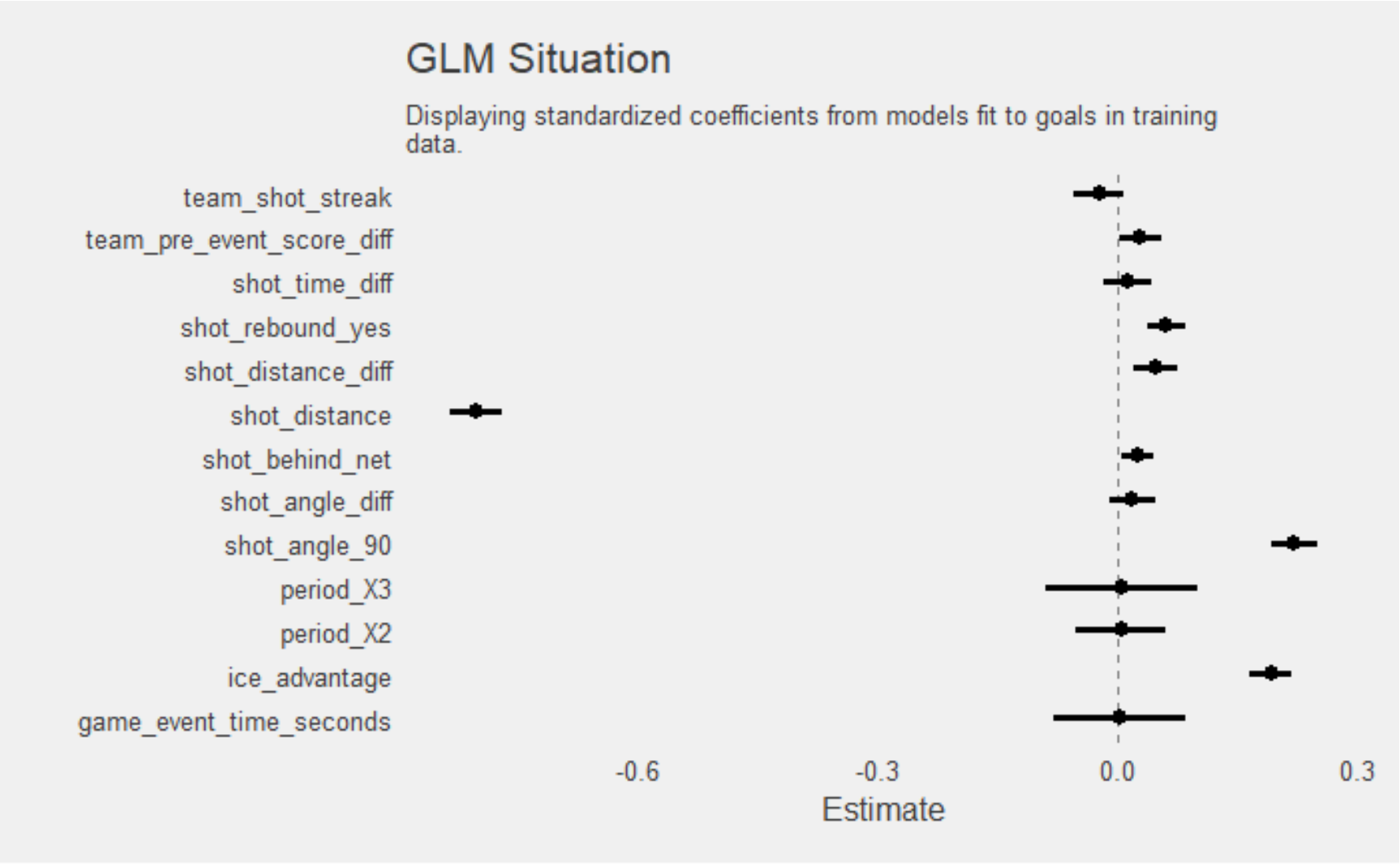
Giovanni Fiore vs 1:33 into the 1st Period

xGoals Distance Model: 0.251

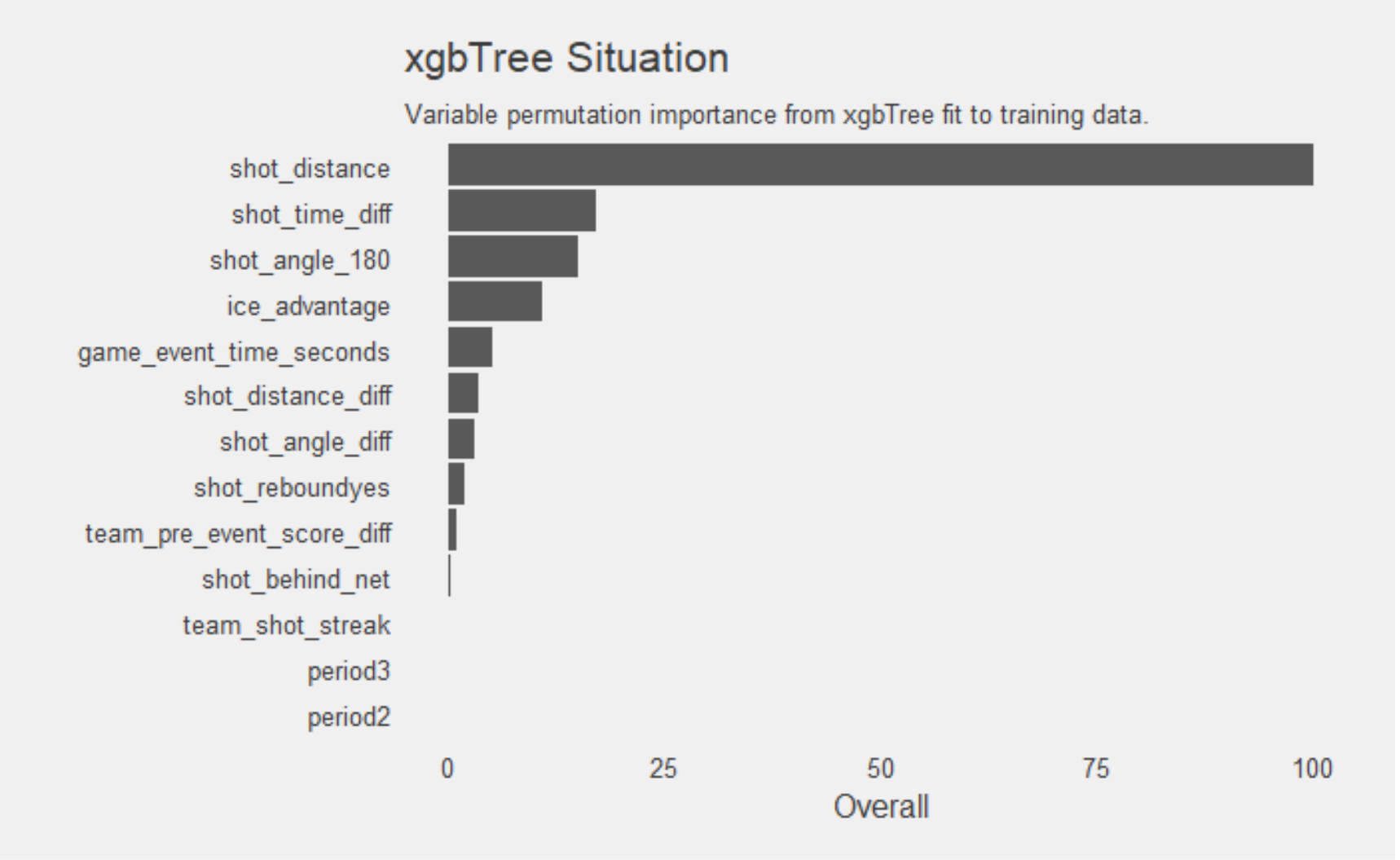
xGoals Location Model: 0.232

xGoals Situation Model: 0.649

We can examine the relationship between the features we put into the model and the outcome.



We can examine the importance of the features we put into the model.



Ultimately, we can use all of these models to predict shots in games in order to compute expected goals (xGoals).

We **learned something from the data** so we can start answering some of our questions.

Ultimately, we can use all of these models to predict shots in games in order to compute expected goals (xGoals).

We **learned something from the data** so we can start answering some of our questions.

This is where the fun begins (and where Tableau enters the picture).

We can look at every game and ask,
did we play better than our opponent?

Or did we get lucky/unlucky?

AHL Post Game Breakdown

Game Info

CHI vs **MIL**
November 11, 2018
Allstate Arena
Attendance: 11,794

Shots on Goal

MIL	CHI
25	29

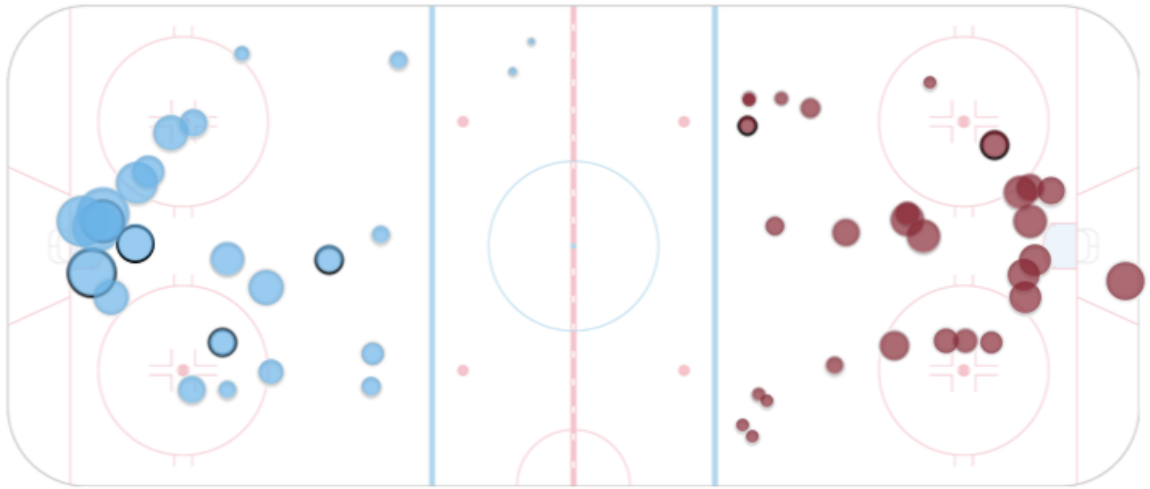
Regulation Goals

MIL	CHI
5	2

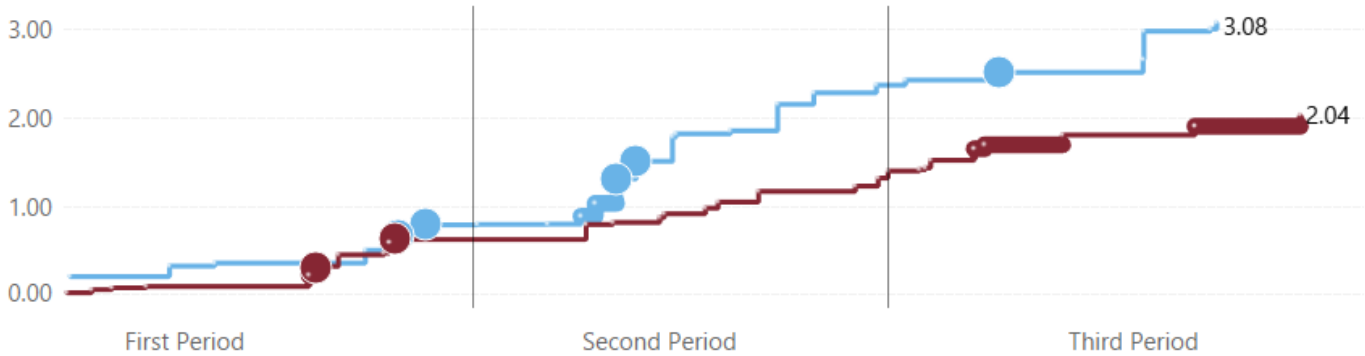
Expected Goals

MIL	CHI
3.08	2.04

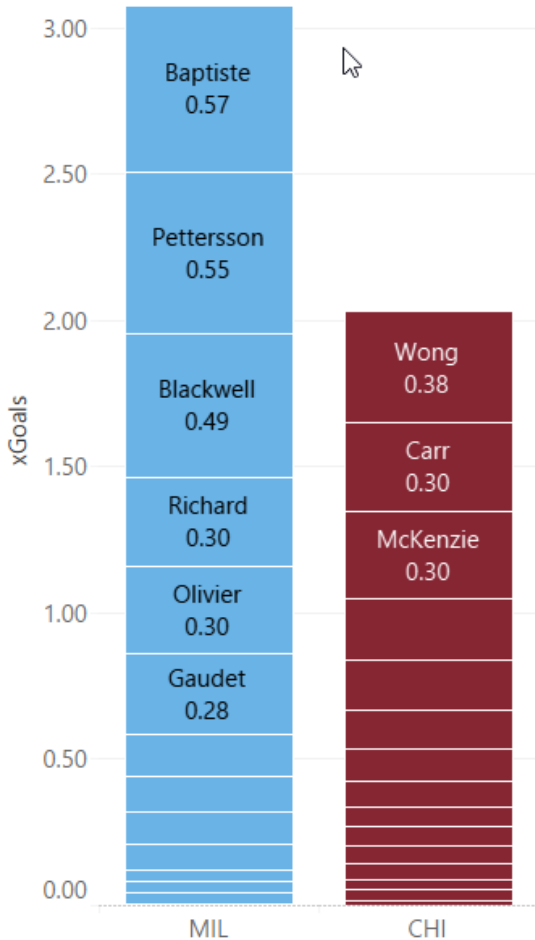
Expected Goals by Location



Expected Goals in Game *(Shaded by Power Play)*



(click bar to highlight)



Filters

Select Season
(All)

Select a Date
November 11, 2018

Select a Game ID
- 1400
- 1401
- 1402
- 1403
- 1404
- 1405

Algorithm Selector
Gradient Boosted ...

Model Selector
Baseline

played pretty well

AHL Post Game Breakdown

Game Info

MIL vs **RFD**
November 9, 2018
UW-Milwaukee Panther Arena
Attendance: 6,147

Shots on Goal

RFD	MIL
19	20

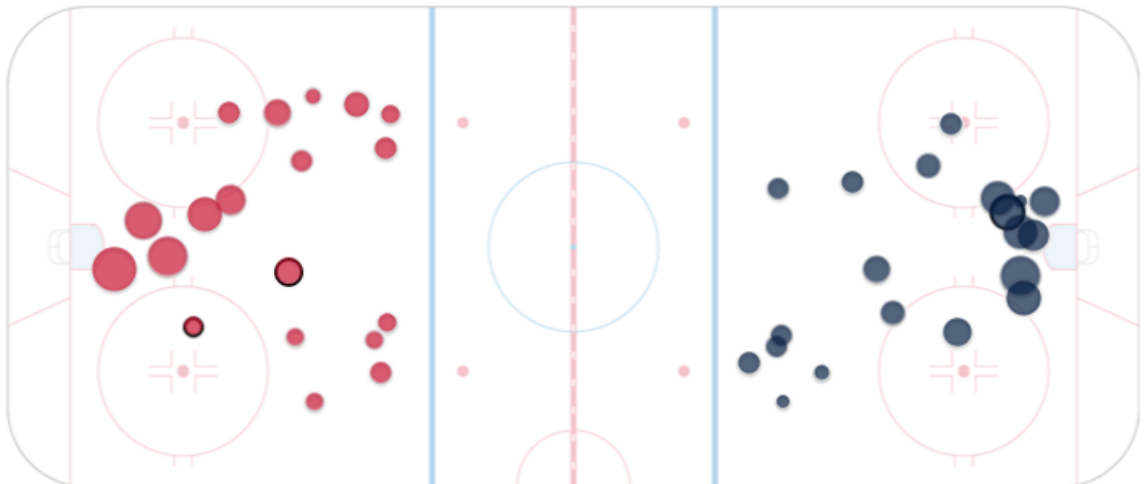
Regulation Goals

RFD	MIL
2	1

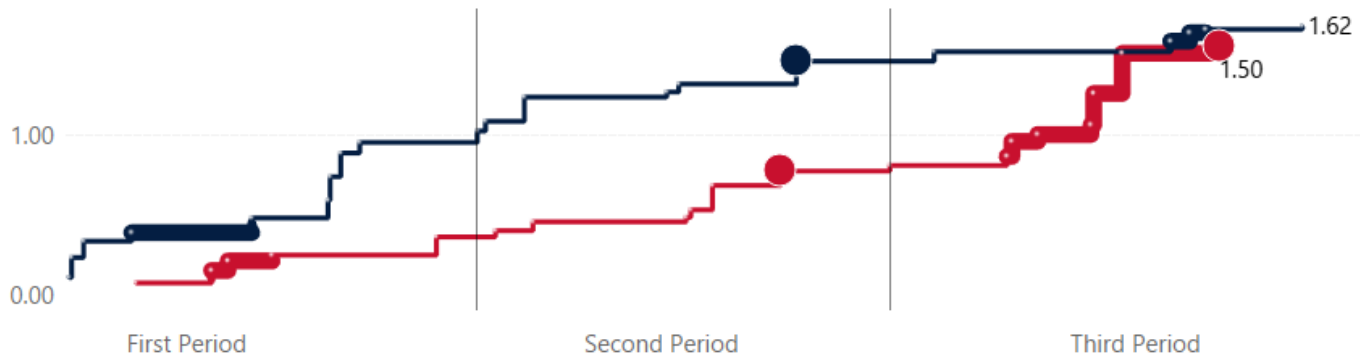
Expected Goals

RFD	MIL
1.50	1.62

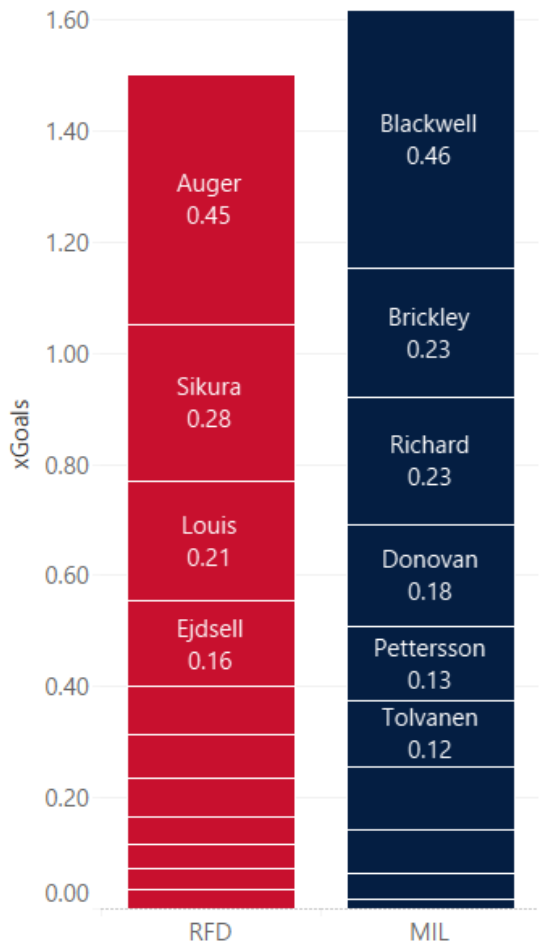
Expected Goals by Location



Expected Goals in Game *(Shaded by Power Play)*



(click bar to highlight)



Filters

Select Season
(All)

Select a Date
November 9, 2018

Select a Game ID
- 1376
- 1377
- 1378
- 1379
- 1380
- 1381
- 1382
- 1383
- 1384
- 1385
- 1386

Algorithm Selector
Gradient Boosted ...

Model Selector
Baseline

eh, maybe a bit unlucky

AHL Post Game Breakdown

Game Info

ROC vs **SPR**
November 16, 2018
MassMutual Center
Attendance: 5,044

Shots on Goal

ROC	SPR
37	34

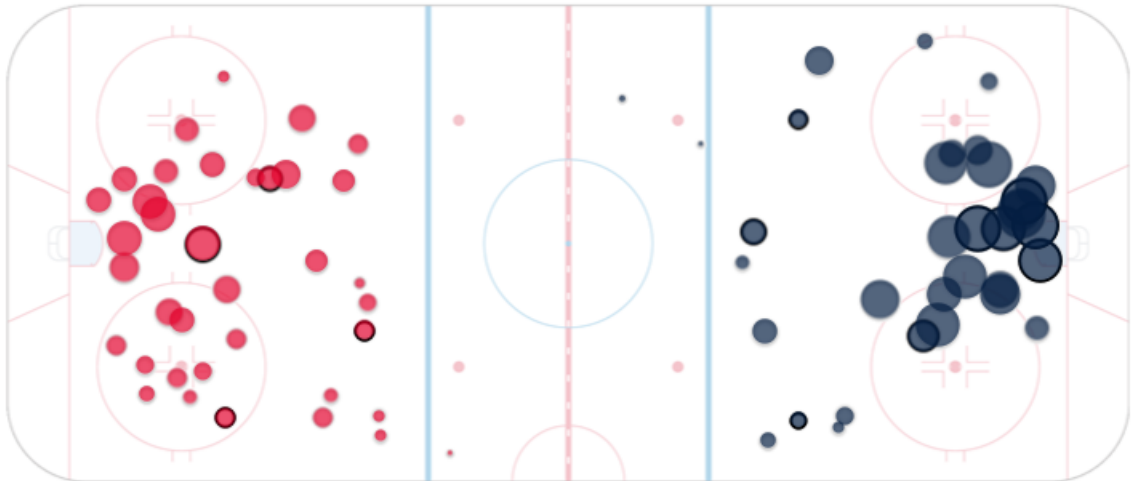
Regulation Goals

ROC	SPR
4	9

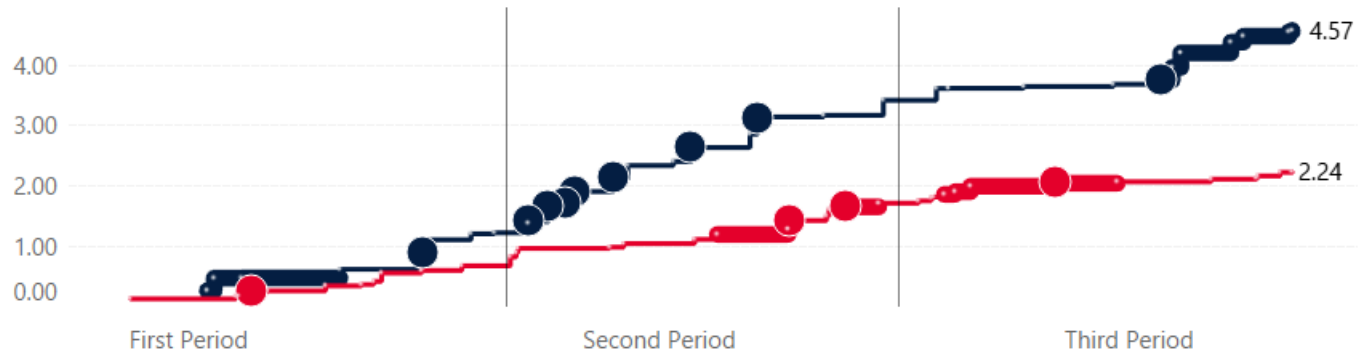
Expected Goals

ROC	SPR
2.24	4.57

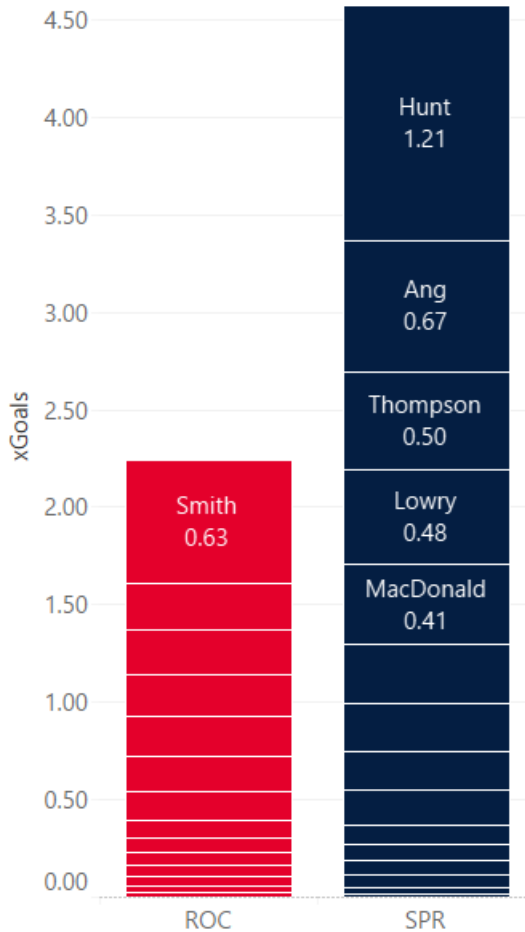
Expected Goals by Location



Expected Goals in Game *(Shaded by Power Play)*



(click bar to highlight)



Filters

Select Season
(All)

Select a Date
November 16, 2018

Select a Game ID
- 1418
- 1419
- 1420
- 1421
- 1422
- 1423
- 1424
- 1425
- 1426
- 1427
- 1428
- 1429

Algorithm Selector
Gradient Boosted ...

Model Selector
Baseline

the hell happened in that second period?

We can ask questions, like how does a team's offense compare to the rest of the league?

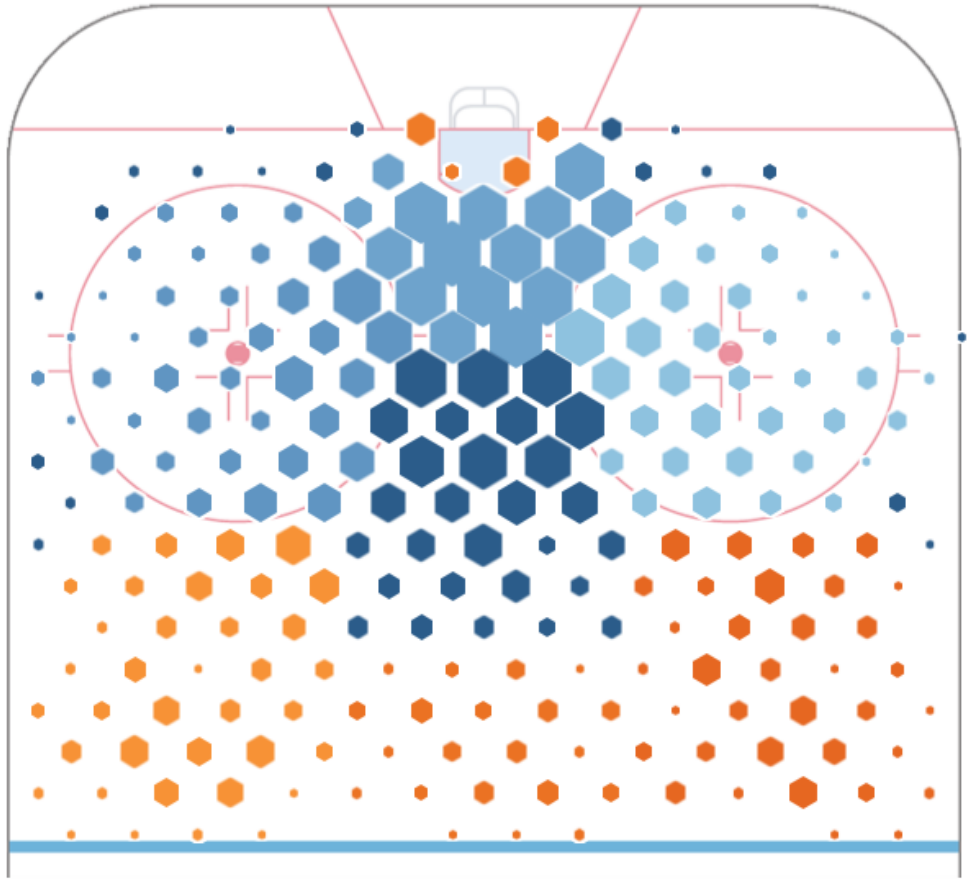
Have they improved year over year?

TEX Offense by Scoring Areas 2019

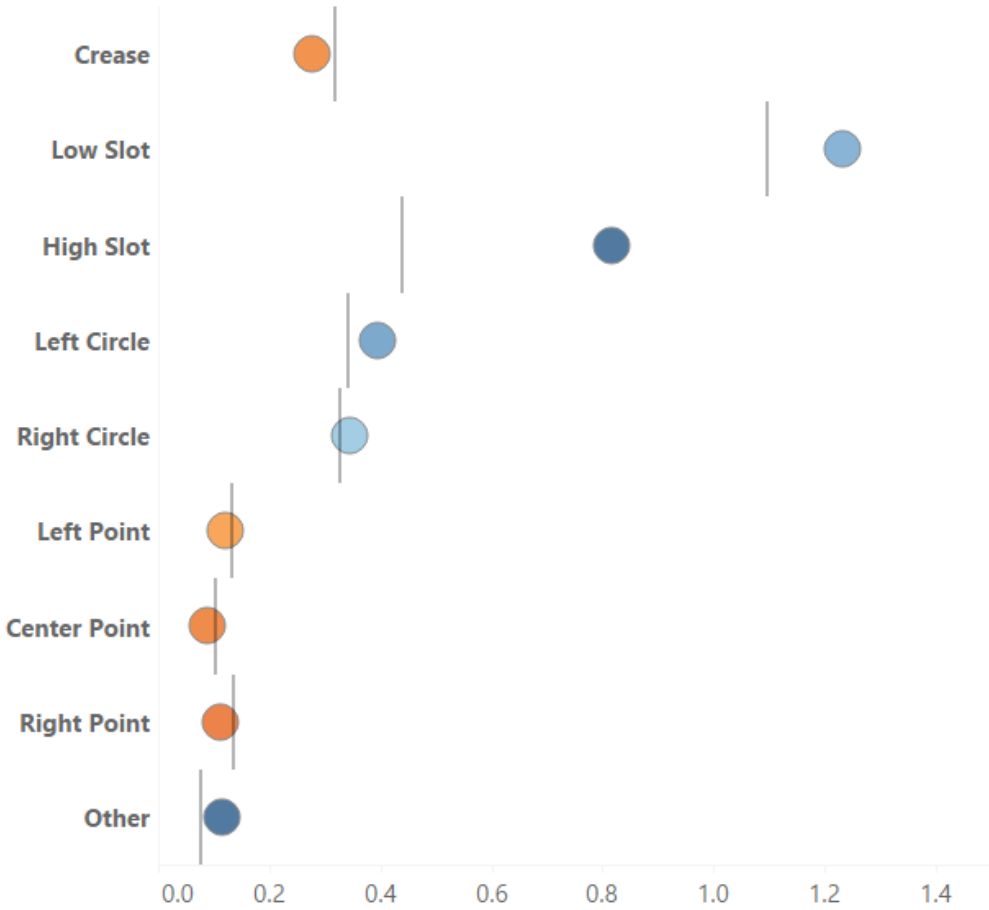
Games	TEX xGoals	per Game	AHL xGoals	per Game	xGoal Diff
75	235.46	3.139	196.0	2.661	0.478

Below Average -30.00% — 30.00% *Above Average*

Team xGoals per Game vs League xGoals per Game



Team xGoals by Scoring Area vs League Avg



Filters

Season: 2019

Select Team: TEX

Algorithm Selector: Gradient Boosted...

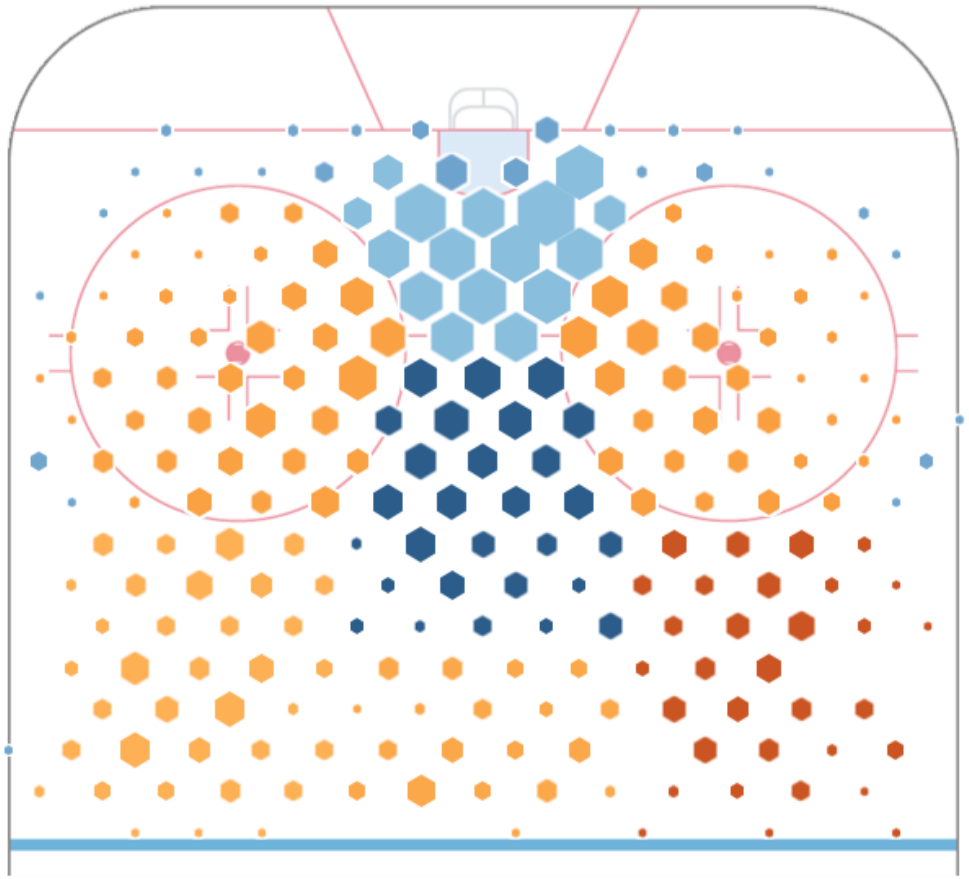
Model Selector: Baseline

TEX Offense by Scoring Areas 2020

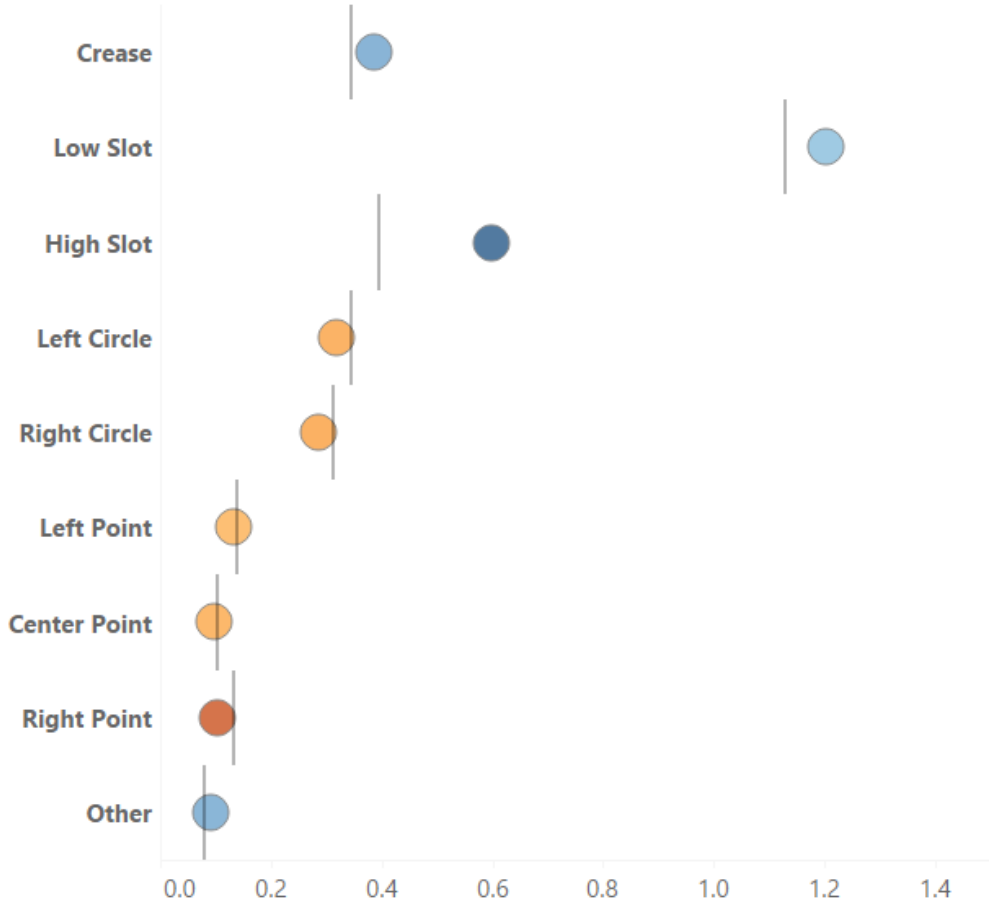
Games	TEX xGoals	per Game	AHL xGoals	per Game	xGoal Diff
62	174.42	2.813	162.7	2.684	0.129

Below Average -30.00% 30.00% *Above Average*

Team xGoals per Game vs League xGoals per Game



Team xGoals by Scoring Area vs League Avg



Filters

Season: 2020

Select Team: TEX

Algorithm Selector: Gradient Boosted...

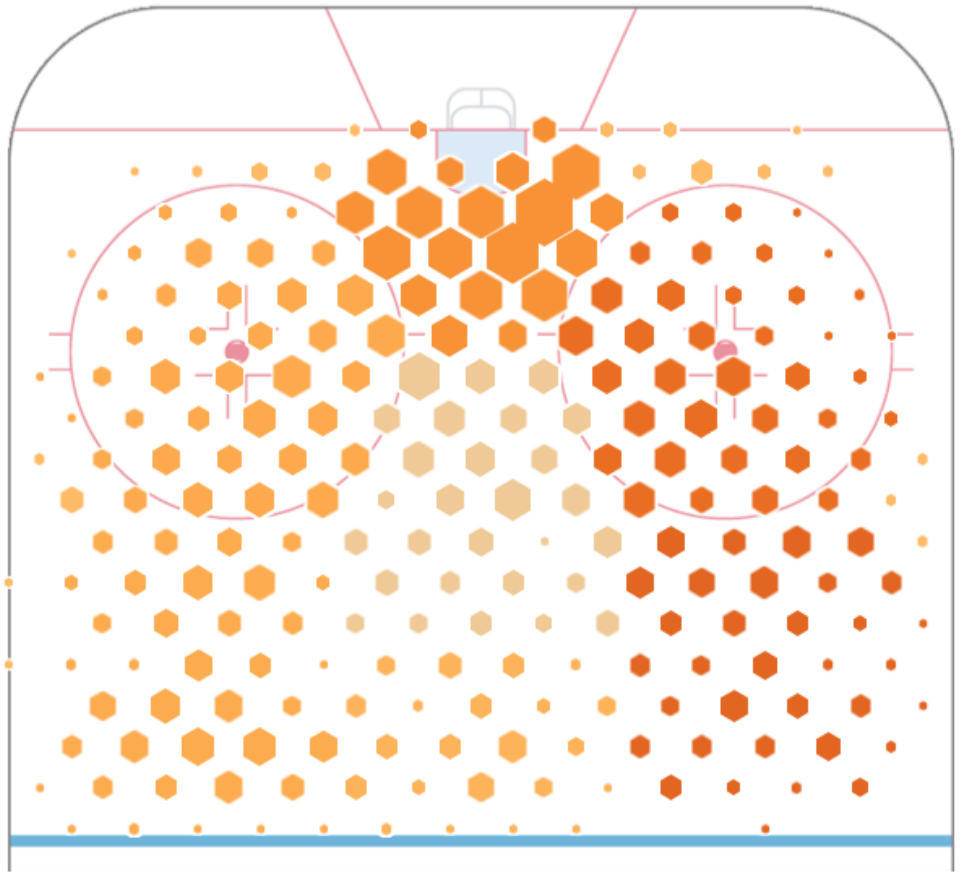
Model Selector: Baseline

MIL Offense by Scoring Areas 2019

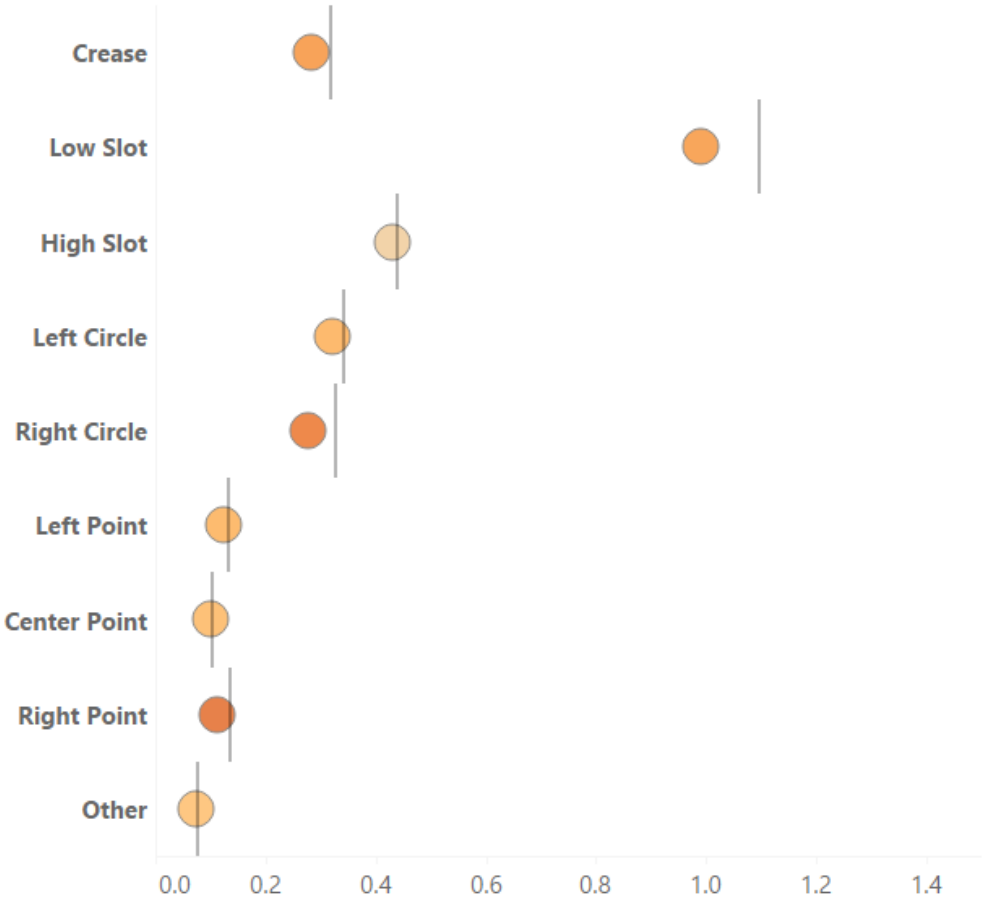
Games	MIL xGoals	per Game	AHL xGoals	per Game	xGoal Diff
76	182.31	2.399	196.0	2.661	-0.262

Below Average -30.00% — 30.00% *Above Average*

Team xGoals per Game vs League xGoals per Game



Team xGoals by Scoring Area vs League Avg



Filters

Season: 2019

Select Team: MIL

Algorithm Selector: Gradient Boosted...

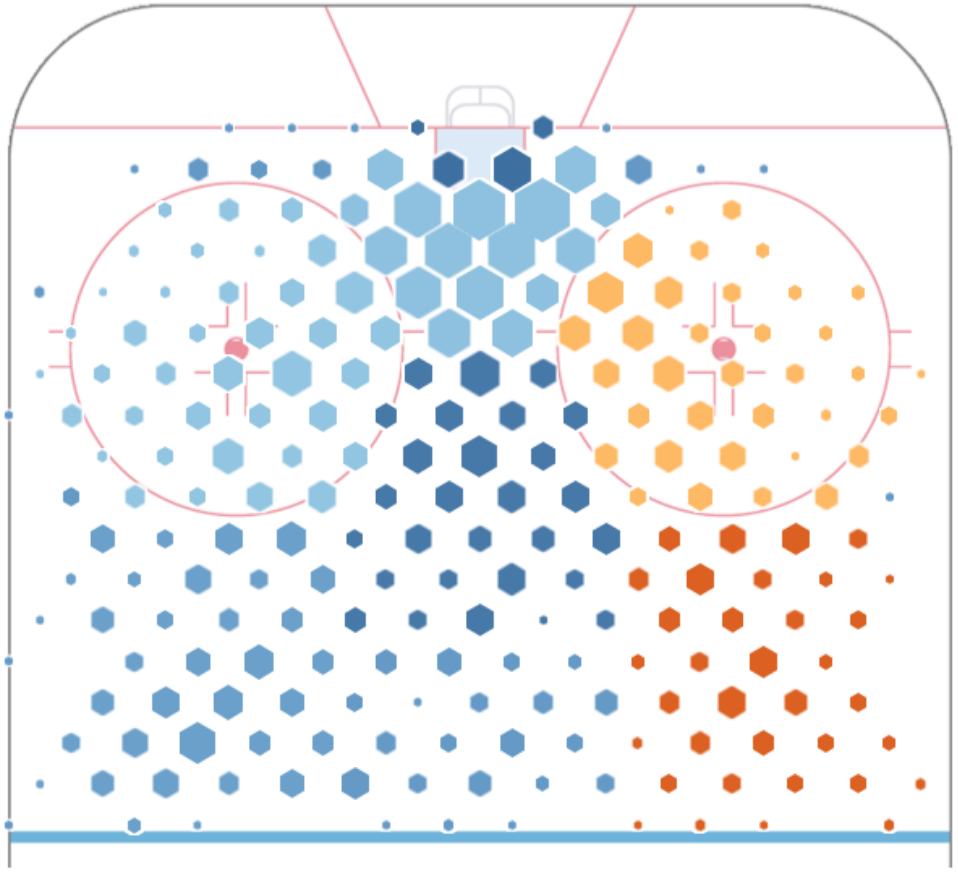
Model Selector: Baseline

MIL Offense by Scoring Areas 2020

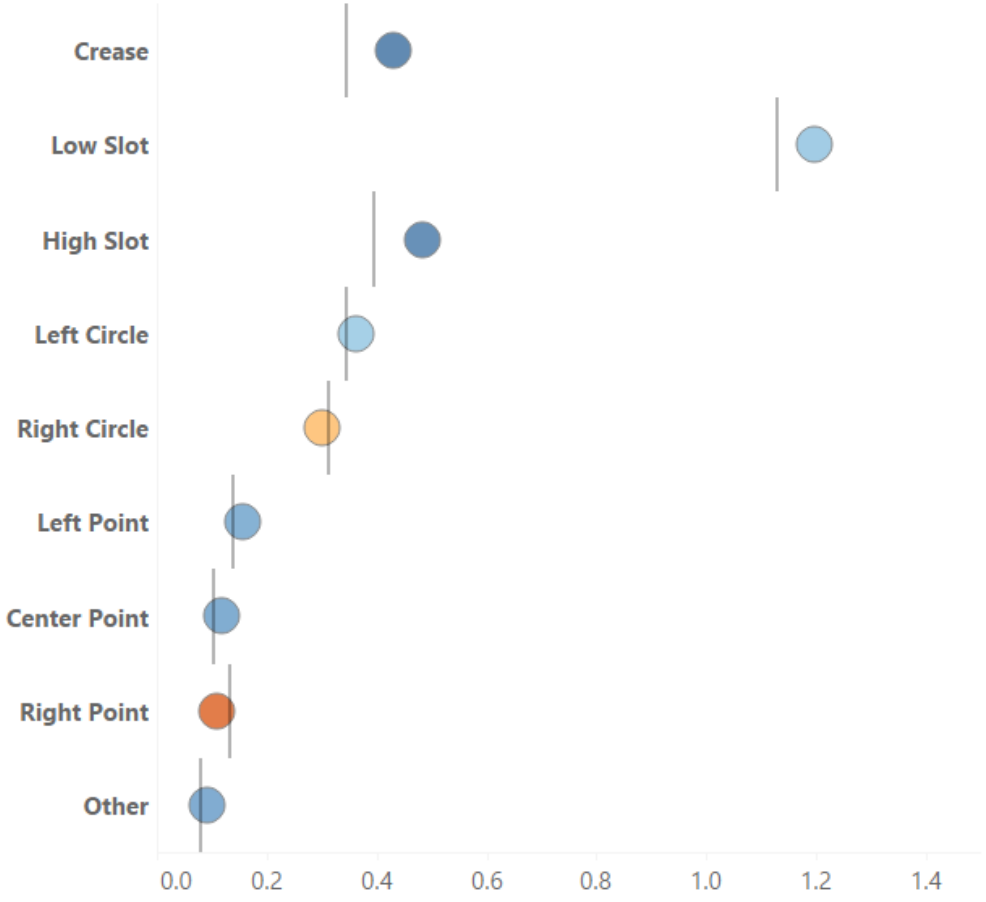
Games	MIL xGoals	per Game	AHL xGoals	per Game	xGoal Diff
63	182.93	2.904	162.7	2.684	0.219

Below Average -30.00% 30.00% *Above Average*

Team xGoals per Game vs League xGoals per Game



Team xGoals by Scoring Area vs League Avg



Filters

Season: 2020

Select Team: MIL

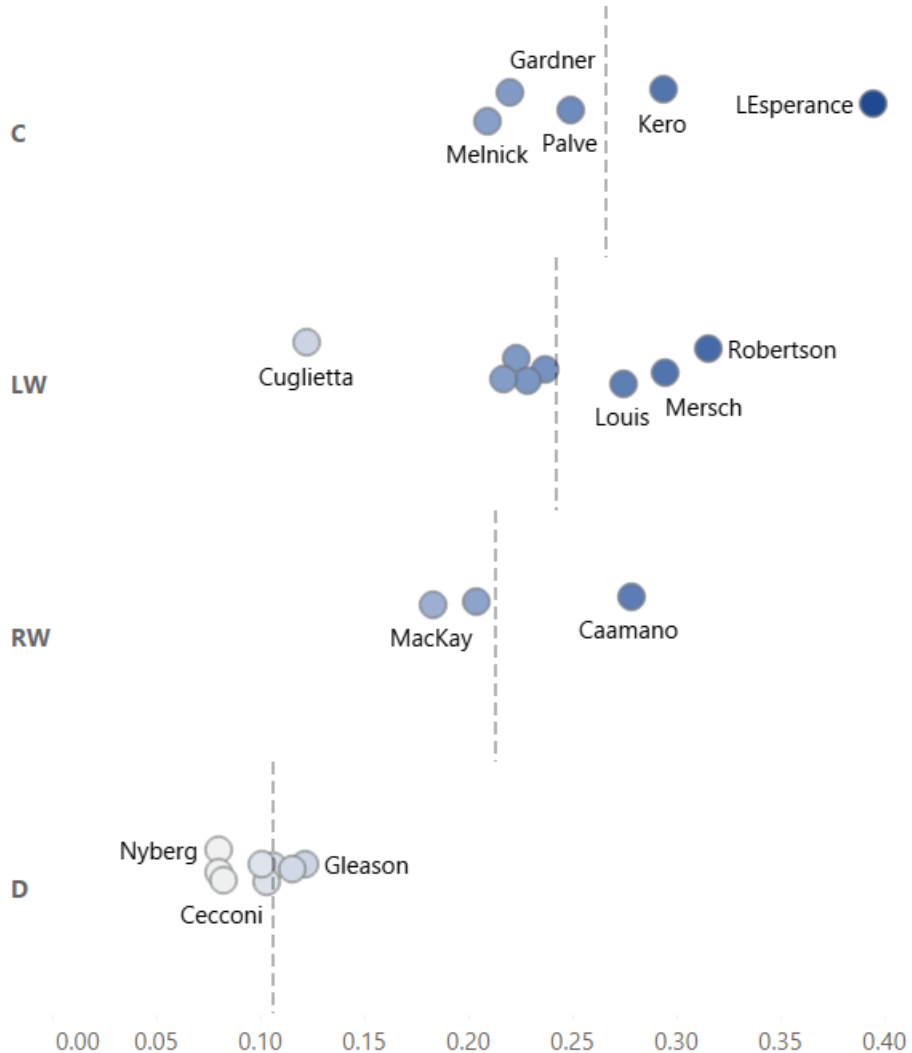
Algorithm Selector: Gradient Boosted...

Model Selector: Baseline

Which of our players are generating the most offense?

Player xGoals Leaders

xGoals per Game 2020 Season(s)



click to highlight

		Games	Shots	Goals	xGoals	Diff	xGoals per G..
Joel LEsperance	C	55	185	23	21.69	1.3	0.39
Jason Robertson	LW	51	136	25	16.07	8.9	0.32
Michael Mersch	LW	53	119	13	15.62	-2.6	0.29
Tanner Kero	C	36	91	7	10.58	-3.6	0.29
Nicholas Caam..	RW	29	75	8	8.06	-0.1	0.28
Anthony Louis	LW	32	79	9	8.78	0.2	0.27
Oula Palve	C	19	44	2	4.73	-2.7	0.25
Riley Tufte	LW	36	72	3	8.53	-5.5	0.24
Joel Kiviranta	LW	43	80	12	9.83	2.2	0.23
Adam Mascheri..	LW	22	47	4	4.90	-0.9	0.22
Rhett Gardner	C	38	72	9	8.35	0.6	0.22
Tye Felhaber	LW	33	62	2	7.14	-5.1	0.22
Josh Melnick	C	32	57	4	6.70	-2.7	0.21
Brad McClure	RW	39	79	5	7.96	-3.0	0.20
Parker MacKay	RW	17	29	2	3.11	-1.1	0.18
Diego Cuglietta	LW	12	17	1	1.47	-0.5	0.12
Ben Gleason	D	37	77	2	4.49	-2.5	0.12
Gavin Bayreuth..	D	55	103	5	6.36	-1.4	0.12
Reece Scarlett	D	28	49	5	2.97	2.0	0.11
Joel Hanley	D	28	49	0	2.88	-2.9	0.10
Emil Djuse	D	33	52	4	3.33	0.7	0.10
Joseph Cecconi	D	29	49	0	2.38	-2.4	0.08
Dillon Heatheri..	D	40	70	1	3.18	-2.2	0.08
John Nyberg	D	10	15	2	0.80	1.2	0.08

Filters

Season
2020

Select xGoals
per Game

Select Position
(All)

Filter to Team
TEX

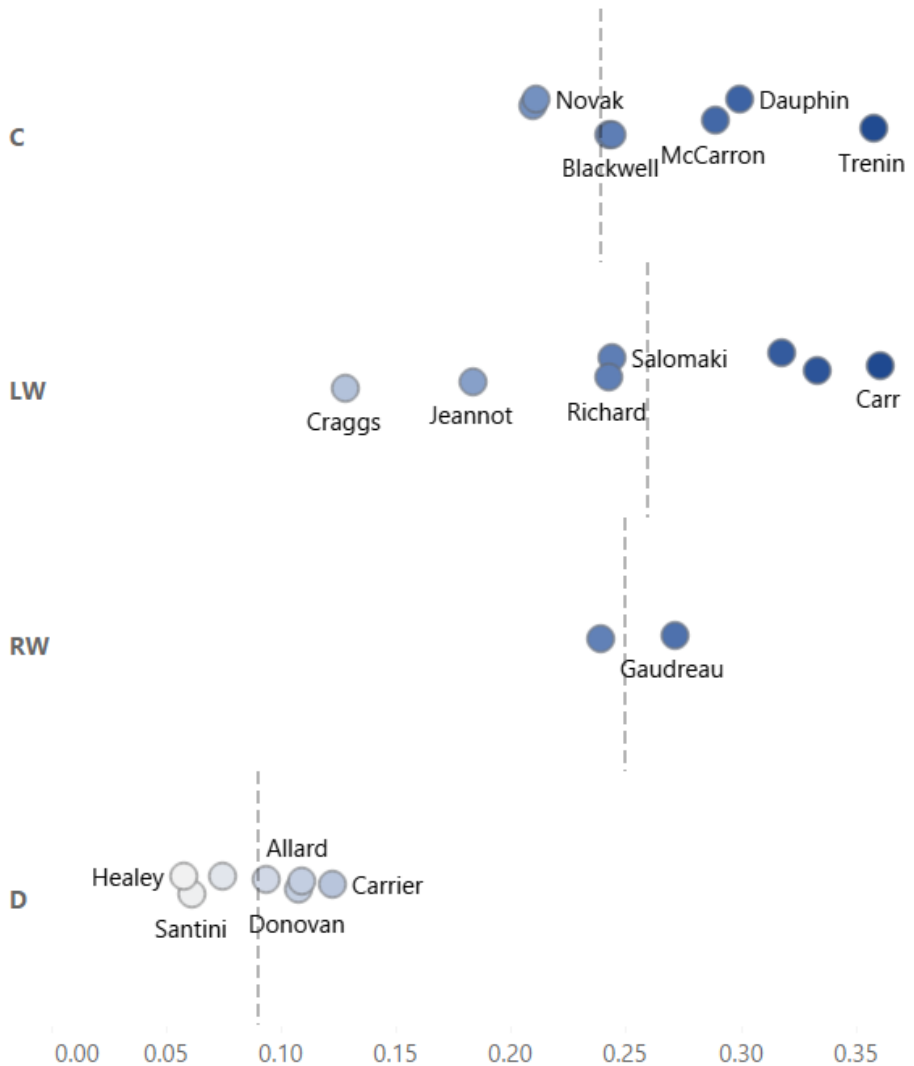
Highlight Player
Highlight Player N...

Model Selector
Baseline

Minimum Games
10 55

Player xGoals Leaders

xGoals per Game 2020 Season(s)



click to highlight

		Games	Shots	Goals	xGoals	Diff	xGoals per G..
Daniel Carr	LW	45	131	22	16.24	5.8	0.36
Yakov Trenin	C	26	67	18	9.30	8.7	0.36
Cole Schneider	LW	49	124	17	16.32	0.7	0.33
Eeli Tolvanen	LW	58	187	20	18.42	1.6	0.32
Laurent Dauphin	C	25	56	6	7.49	-1.5	0.30
Michael McCarr..	C	25	60	9	7.22	1.8	0.29
Frederick Gaud..	RW	37	86	10	10.04	0.0	0.27
Colin Blackwell	C	23	53	4	5.62	-1.6	0.24
Miikka Salomaki	LW	34	82	5	8.30	-3.3	0.24
Anthony Richard	LW	50	106	13	12.13	0.9	0.24
Rem Pitlick	C	42	83	18	10.18	7.8	0.24
Mathieu Olivier	RW	41	76	9	9.82	-0.8	0.24
Tommy Novak	C	45	83	11	9.49	1.5	0.21
Josh Wilkins	C	28	45	3	5.86	-2.9	0.21
Tanner Jeannot	LW	40	67	4	7.35	-3.3	0.18
Lukas Craggs	LW	13	22	2	1.66	0.3	0.13
Alexandre Carri..	D	46	84	4	5.62	-1.6	0.12
Jeremy Davies	D	48	114	4	5.24	-1.2	0.11
Matt Donovan	D	53	116	5	5.71	-0.7	0.11
Frederic Allard	D	47	89	2	4.40	-2.4	0.09
Jarred Tinordi	D	22	39	0	1.63	-1.6	0.07
Steven Santini	D	38	63	2	2.33	-0.3	0.06
Josh Healey	D	11	19	0	0.64	-0.6	0.06

Filters

Season
2020

Select xGoals
per Game

Select Position
(All)

Filter to Team
MIL

Highlight Player
Highlight Player N...

Model Selector
Baseline

Minimum Games
10 58

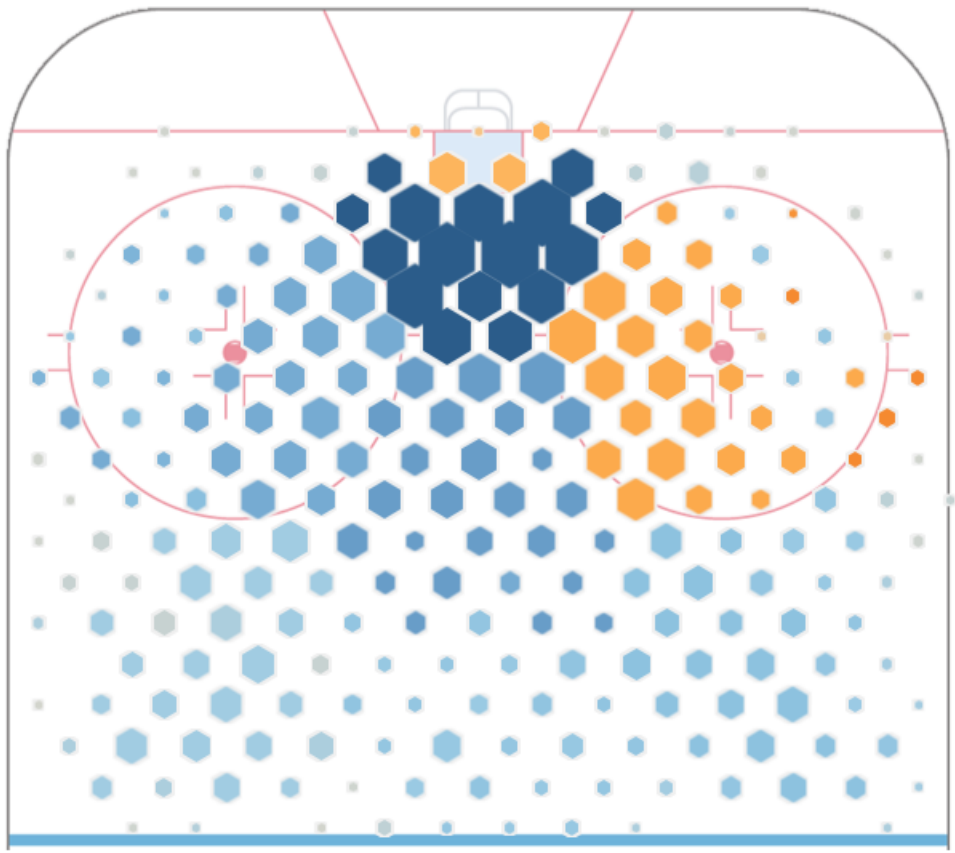
How good is our goalie? Where are his weaknesses?

Goalie: Connor Ingram

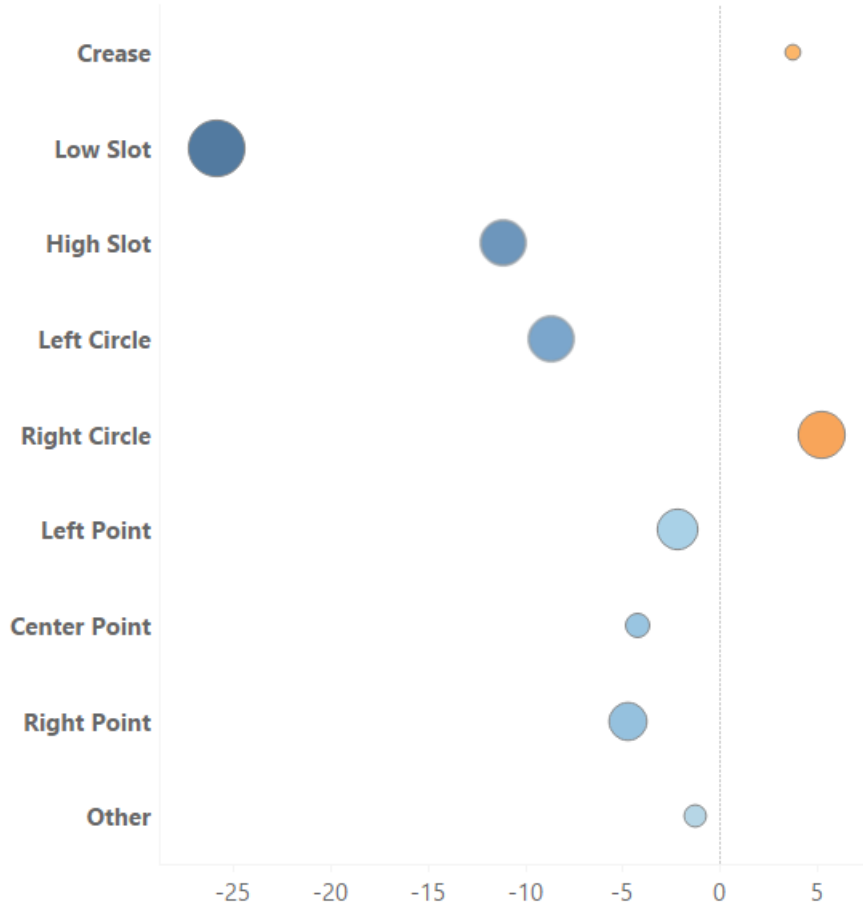
Games	Save %	Shots	xGoals	Goals	xGoal Diff
86	92.58%	2,290	218.87	170	-48.87

Saved More Goals — xGoal Diff: -25.00 to 25.00 — *Allowed More Goals*

xGoals Diff by Scoring Area



xGoals Diff by Scoring Area



Filters

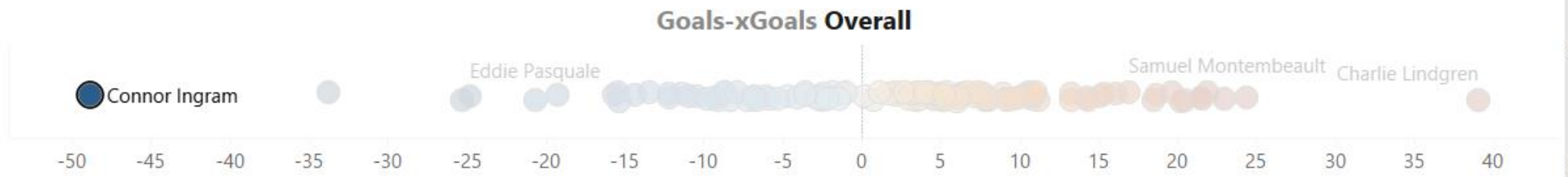
Season: (All)

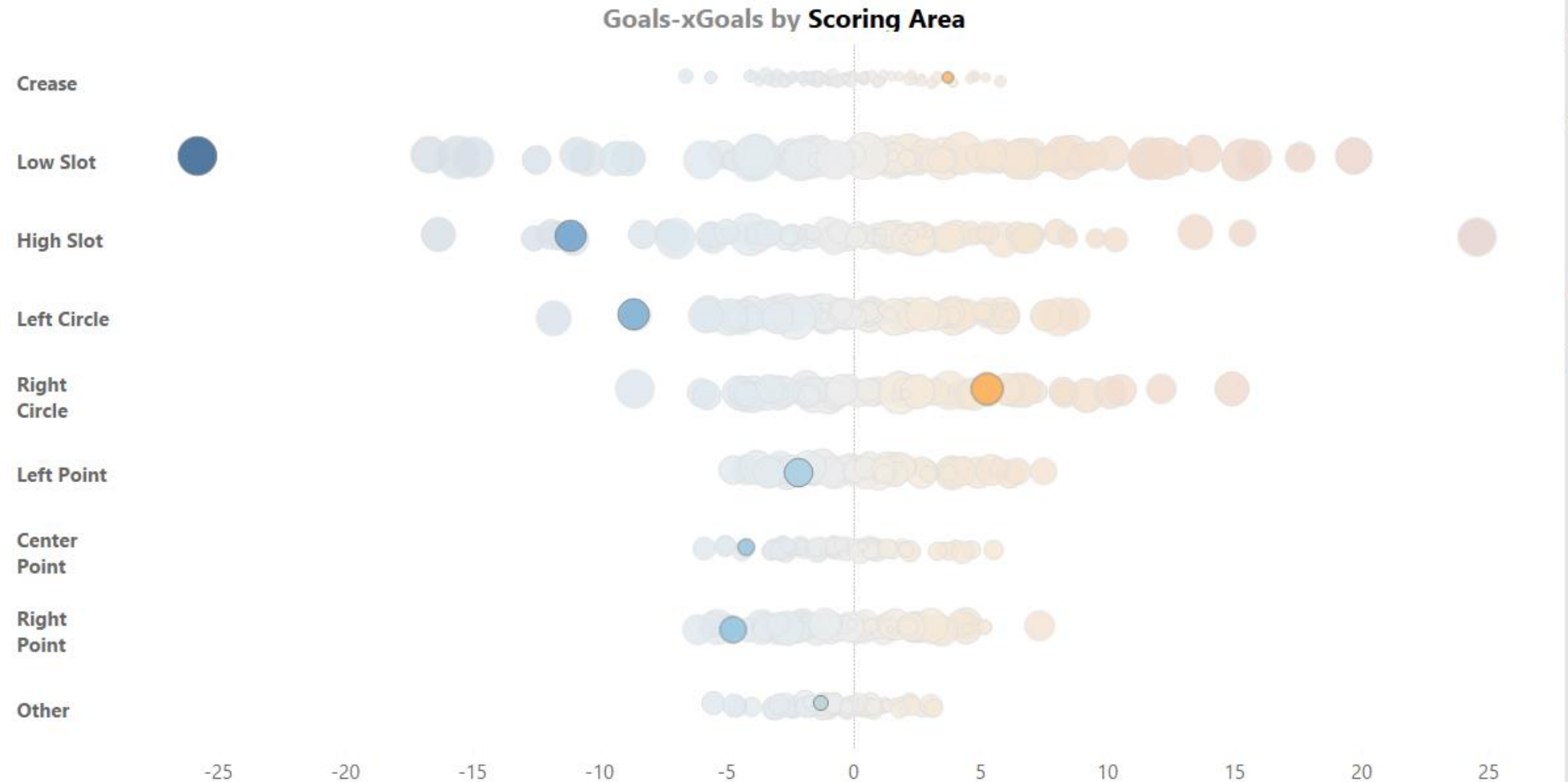
Select Goalie: Connor Ingram

Shot Minimum: 150

# of Shots
- 40
- 100
- 200
- 300
- 400
- 527

Goalie Breakdown by Scoring Area





Filters

Select Season

(All)

Select Metric

Goals-xGoals

Select Goalies

(Multiple values)

# of Shots

15
200
400
600
741

Highlight Goalie Name

Highlight Goalie N...

Highlight Goalie Name

Highlight Goalie N...

We can see that our favorite team's opponent will be playing their 3rd string goalie in the playoffs and realize we have a lot of AHL data on him and decide to scout for his weaknesses.

Goalie: Michael Hutchinson

Games	Save %	Shots	xGoals	Goals	xGoal Diff
60	92.44%	1,798	151.70	136	-15.70

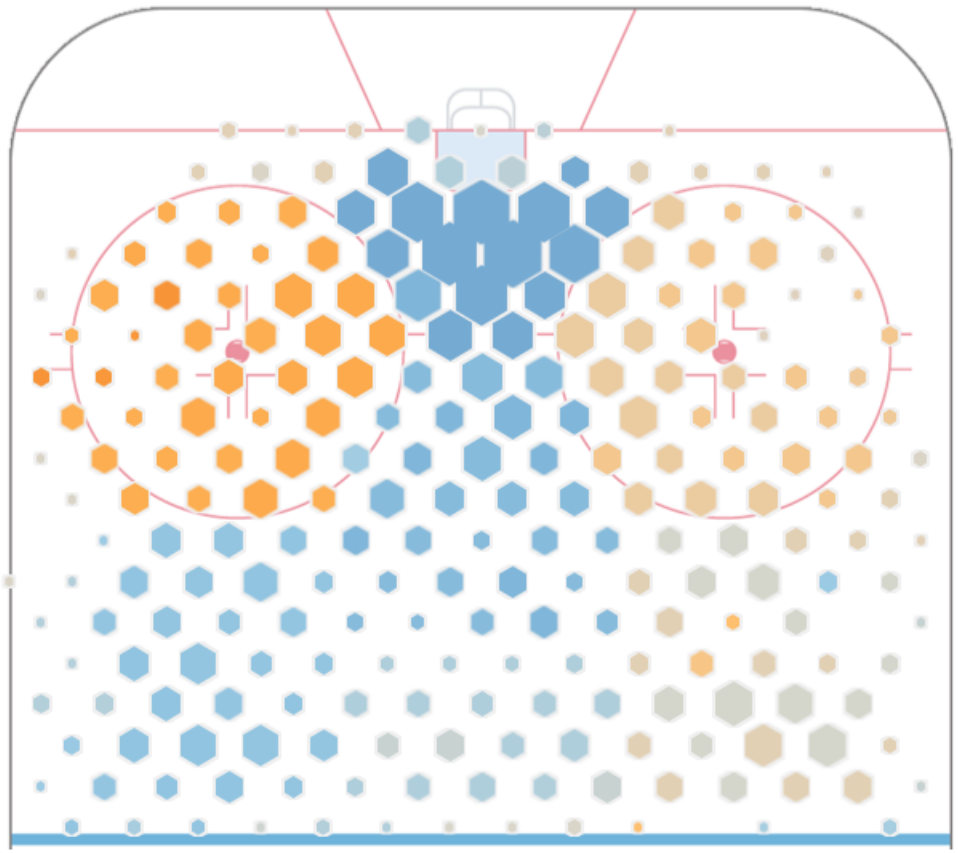
Saved More Goals

xGoal Diff

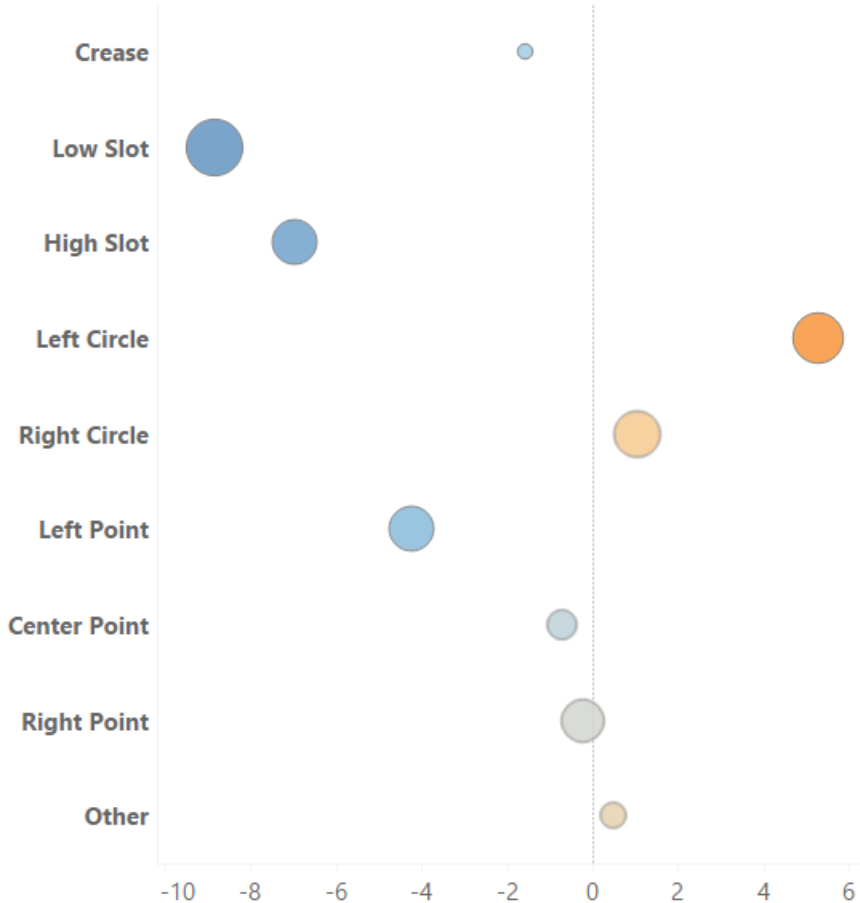
-25.00 — 25.00

Allowed More Goals

xGoals Diff by Scoring Area



xGoals Diff by Scoring Area



Filters

Season

(All)

Select Goalie

Michael Hutchins...

Shot Minimum

150

# of Shots

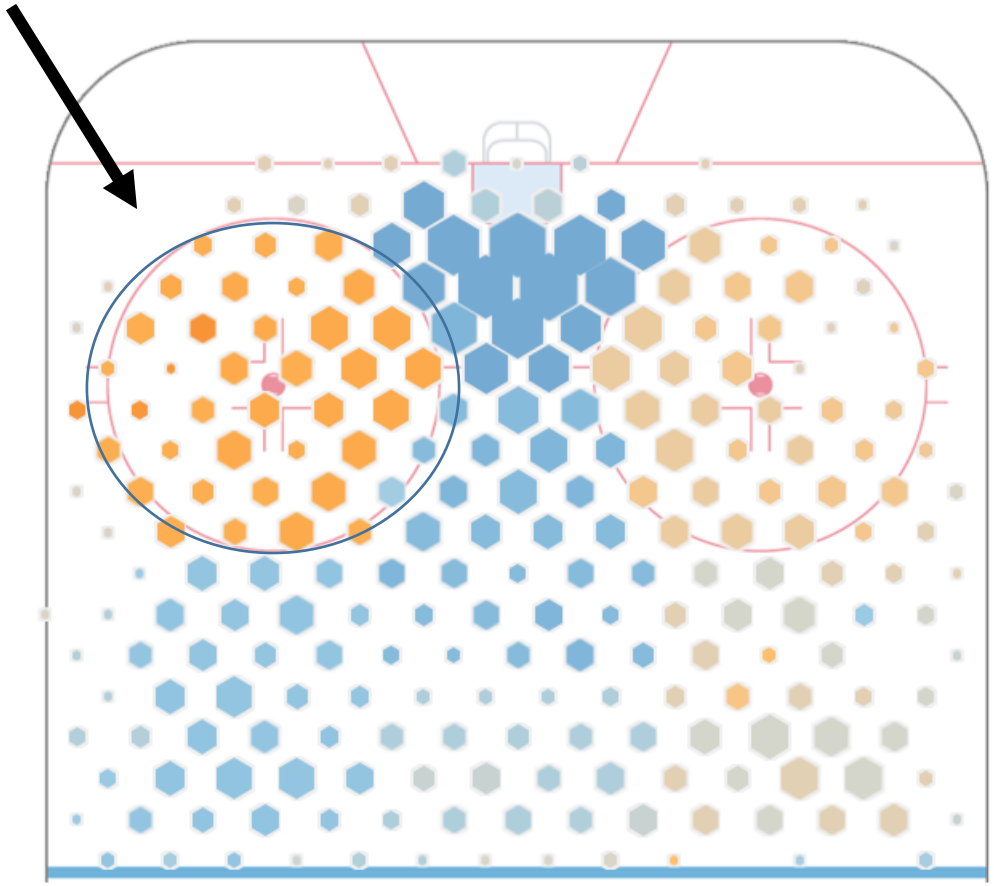
27
100
200
300
371

Goalie: Michael Hutchinson

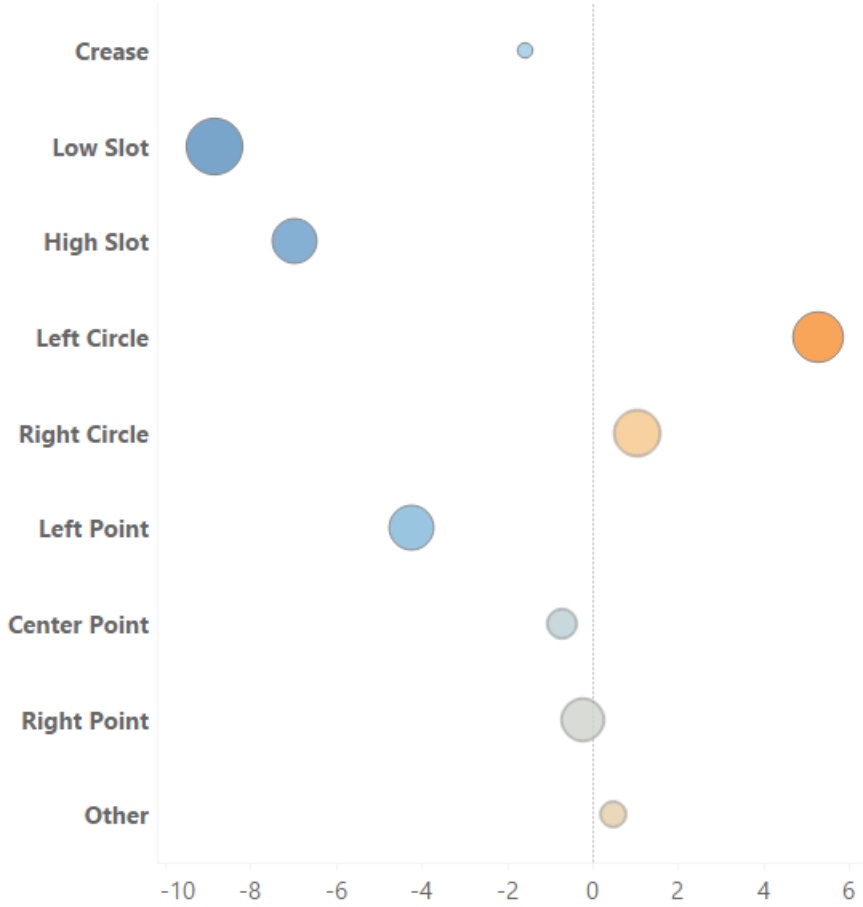
Games	Save %	Shots	xGoals	Goals	xGoal Diff
60	92.44%	1,798	151.70	136	-15.70

Saved More Goals — xGoal Diff: -25.00 to 25.00 — *Allowed More Goals*

xGoals Diff by Scoring Area



xGoals Diff by Scoring Area



Filters

Season: (All)

Select Goalie: Michael Hutchins...

Shot Minimum: 150

# of Shots
- 27
- 100
- 200
- 300
- 371

We can then fruitlessly try to contact the Dallas Stars coaching staff to give them our scouting report.

But they seemed to have figured it out anyway.

September 2, 2020
Miro Heiskanen vs Michael Hutchinson
Dallas vs Colorado, Game 6
17:34 into the 1st Period

September 2, 2020
Joel Kiviranta vs Michael Hutchinson
Dallas vs Colorado, Game 7
16:30 into the 3rd Period

September 2, 2020
Joel Kiviranta vs Michael Hutchinson
Dallas vs Colorado, Game 7
7:24 into the 1st OT

We build models in order to learn about ~~the weaknesses of our opponents goaltenders~~ the world around us.

The process of building models requires us to speculate about the processes we are observing.

Subject matter expertise is incredibly valuable, so long as **we put it to the test**.

Once we have a model in place, we aren't done. We need to look at what it produces. See where it does well. See where it makes mistakes.

We can then speculate again and **try to improve in the next model**.

Science, it turns out, is just as useful for studying the stars as it is for studying the Texas Stars.

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

The basic question of the show was, '**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 data scientist.

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

Mike Schur, Writer

Phil Henrickson, Stealer of Quotes

The basic question of the talk was, '**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 data 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.

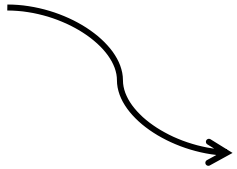
Phil Henrickson, Stealer of Quotes

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

Phil Henrickson, Stealer of Quotes

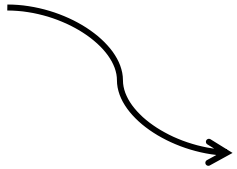
Try to be a little bit better today than you were yesterday.

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



This is the secret to building models.

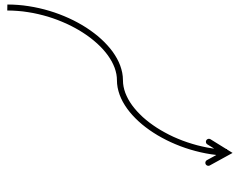
Try to be a little bit better today than
you were yesterday.



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.

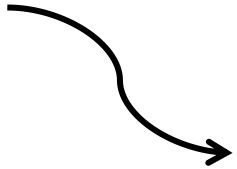


This is the secret to building models.

This is the secret to good science.

This is the answer to the now what of analytics.

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.

For data science advising,
board game recommendations,
and/or AHL scouting reports:

phil.henrickson@aebs.com

Questions?

Appendix

To design interfaces for exploratory data analysis, we need theories of graphical inference*

Jessica Hullman† Andrew Gelman‡

21 Mar 2021

Abstract

Research and development in computer science and statistics have produced increasingly sophisticated software interfaces for interactive and exploratory analysis, optimized for easy pattern finding and data exposure. But design philosophies that emphasize exploration over other phases of analysis risk confusing a need for flexibility with a conclusion that exploratory visual analysis is inherently "model free" and cannot be formalized. We describe how without a grounding in theories of human statistical inference, research in exploratory visual analysis can lead to contradictory interface objectives and representations of uncertainty that can discourage users from drawing valid inferences. We discuss how the concept of a model check in a Bayesian statistical framework unites exploratory and confirmatory analysis, and how this understanding relates to other proposed theories of graphical inference. Viewing interactive analysis as driven by model checks suggests new directions for software and empirical research around exploratory and visual analysis. For example, systems should enable specifying and explicitly comparing data to null and other reference distributions and better representations of uncertainty. Implications of Bayesian and other theories of graphical inference should be tested against outcomes of interactive analysis by people to drive theory development.

Media Summary: Novel interactive graphical user interface tools for exploratory visual data analysis provide analysts with impressive flexibility in how to look at and interact with data. Often these systems are designed to make patterns in data as easy to see as possible. However, there are risks to designing systems for easy pattern finding alone. One risk is that the techniques used to emphasize patterns, like aggregating data by default, lead analysts to overlook variation and uncertainty in their data, leading them to draw conclusions that aren't well supported by the data. Another is that some analysts may fail to recognize the importance of confirming any insights they arrive at through visual search using confirmatory statistical modeling to make sure they are valid. One reason that graphical user interface systems for interactive analysis may not be designed to enforce strong connections between exploratory and confirmatory analysis is because there aren't well-established theories of how these two types of activities are related. We propose a perspective that unites exploratory and confirmatory analysis through the idea of graphs as model checks in a Bayesian statistical framework, and describe how in light of this view, it becomes clear that systems for exploratory visual analysis should better support model-driven inference and representation of uncertainty.

Keywords: Exploratory data analysis, interactive analysis, graphical inference, Bayesian model check.

Source: Data Visualization as Narrative, Gelman and Dewitt

Why is machine learning proving to be so useful?

In a nutshell, it starts simple and then learns from its mistakes really quickly.

Boosting

1. Start with a simple model
2. Figure out where you missed
3. Adjust for your misses in the next model