# Predicting the Past: Digital Art History, Modeling, and Machine Learning

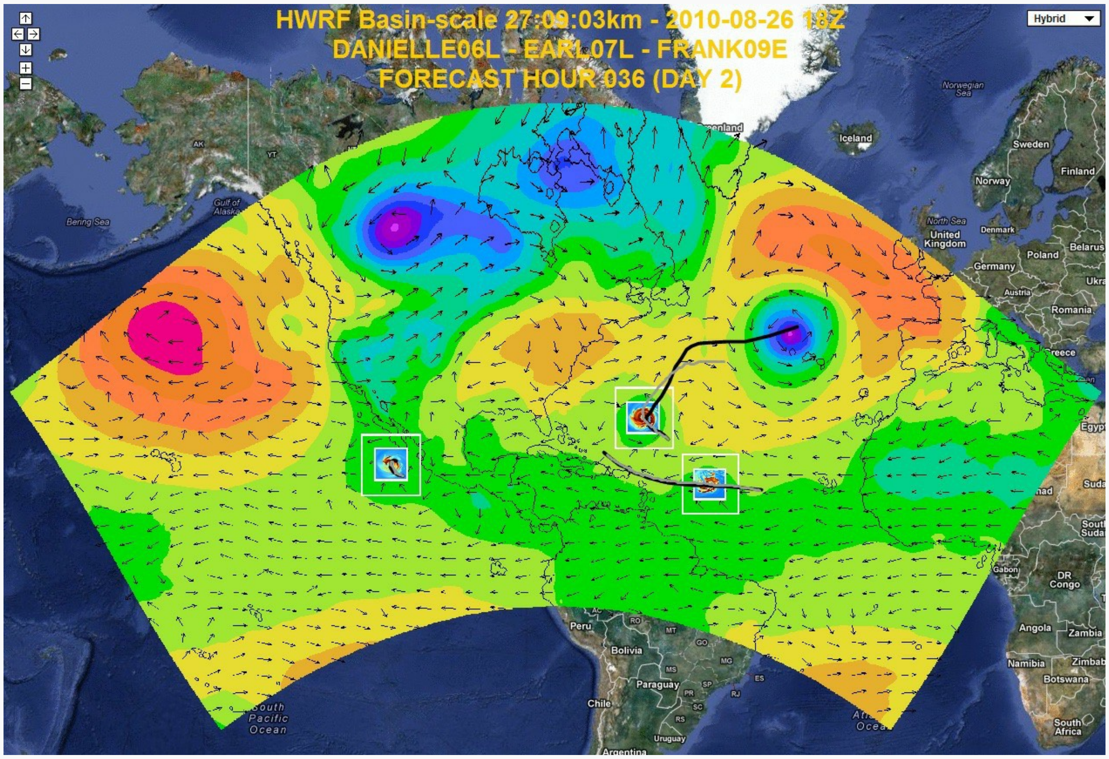
blogs.getty.edu/iris/predicting-the-past-digital-art-history-modeling-and-machine-learning/

Matthew Lincoln

7/27/2017

We are surrounded by models.

When you check the weather forecast in the morning before going to work, you are seeing the result of a model of your local atmosphere. This model is a set of rules and data manipulations built according to past observations of the weather, and refined over many years of testing and evaluation of predictions against actual results. By running current data—like temperature, pressure, humidity, and wind from a variety of locations collected over the past few days or weeks—through this model, you get a resulting prediction about whether it's going to rain that afternoon.


HWRF Basin-scale 27:09:03km - 2010-08-26 18Z
DANIELLE06L - EARL07L - FRANK09E
FORECAST HOUR 036 (DAY 2)
Hybrid

NOAA 5-day hurricane forecast model. Forecast tracks are shown in gray, observed tracks in black. (For more information, see storm.aoml.noaa.gov.)

Similarly, when Facebook recommends "people you might know," it's showing you predictions produced by a model of how people have previously used that site (e.g. in the past, how likely was someone to friend a person who has X mutual friends with them?, etc.) and using that past data to predict what *might* happen next…(1)

But models aren't only good for trying to predict the future. By using them to predict the *past*, they can become powerful mechanisms for reasoning about history. In this post, I'll be introducing some modeling approaches (some aided by *machine learning*) that historians at the Getty and elsewhere are using to grapple with the past.

## Why model?

The sociologist Joshua Epstein once wrote that we are *all* modelers already(2). Anyone who suggests what might happen in the future based on what happened in the past holds a model—however informal—of the world in their head. These mental models are replete with rules and assumptions that chart the course between some set of evidence or past observations, and some prediction for the future.

Thus, Epstein argues, the choice for scholars isn't whether or not to model, but whether or not to do so *explicitly* by specifying our starting premises along with the rules—whether defined mathematically, or in code—that lead from that starting evidence to our final conclusions. Among other benefits, Epstein shows that explicit model buildings help us form better explanations, discover new questions, and highlight uncertainties and unknowns.

While Epstein was talking about modeling in the context of the social sciences, many historians are beginning to think (once again) about the intersections between modeling and historical thinking and argumentation. Historians create theories to explain the processes that may have produced evidence—whether archeological remains, archival documents, or even works of art—that survives today.

We normally evaluate these kinds of theories through logical reasoning and careful validation against primary sources. A historical theory is good if it makes clear, logical sense and if it fits the available evidence. But the greater the volume of those primary sources, and the more subtle and complicated the phenomenon we are theorizing, the harder it can be to thoroughly evaluate whether a good-sounding theory really is a plausible explanation for past evidence. When evidence is particularly abundant, and the potential process highly complex, it can even be difficult to formulate a theory in the first place.

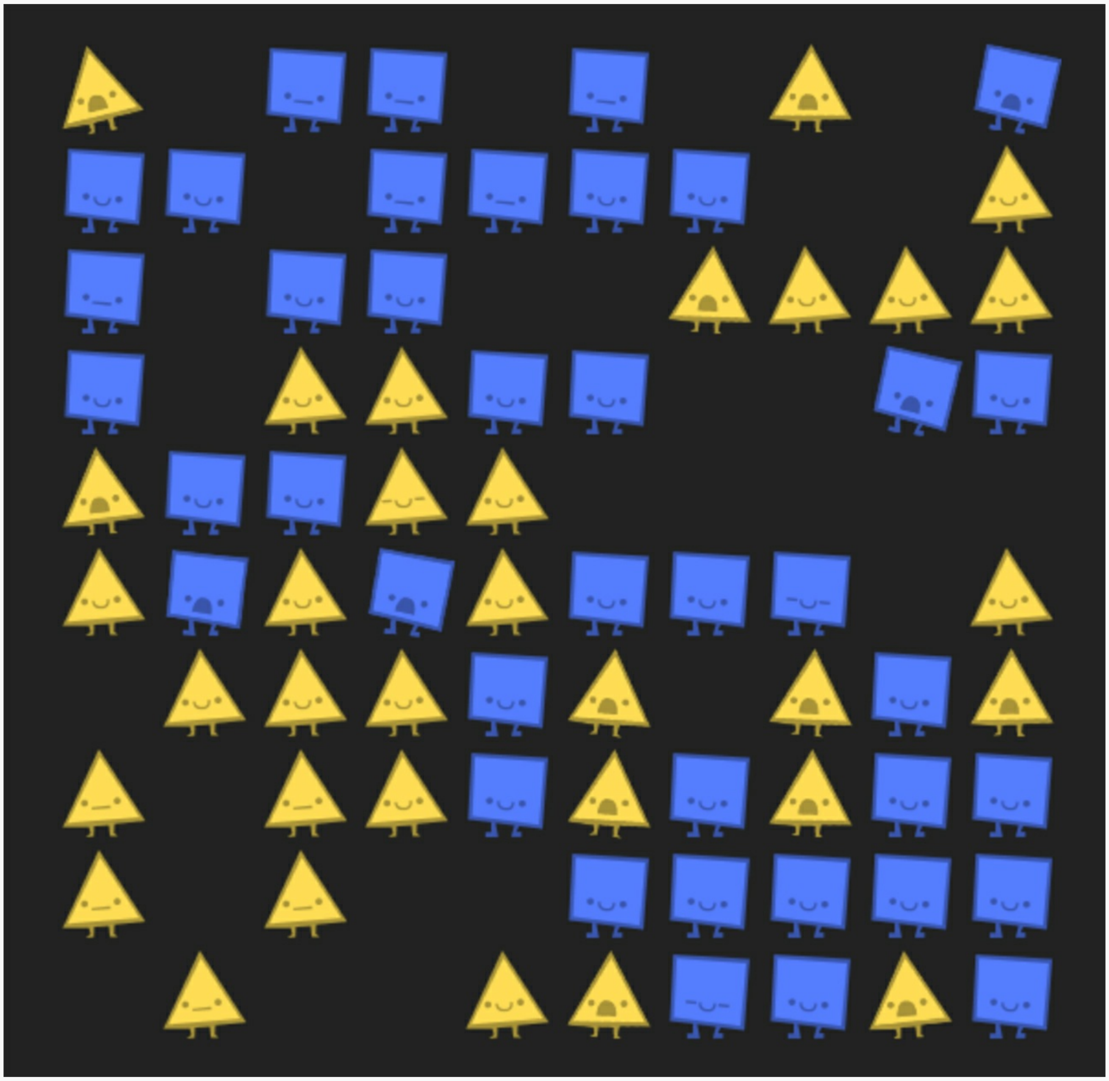
By building data-driven simulations, even highly simplified ones, we have the opportunity to play out these theories and evaluate when they are *plausible*—that is, when they successfully "predict the past"—and when they fail.(3)

## What Makes a Model?

The modeling process comprises three things:

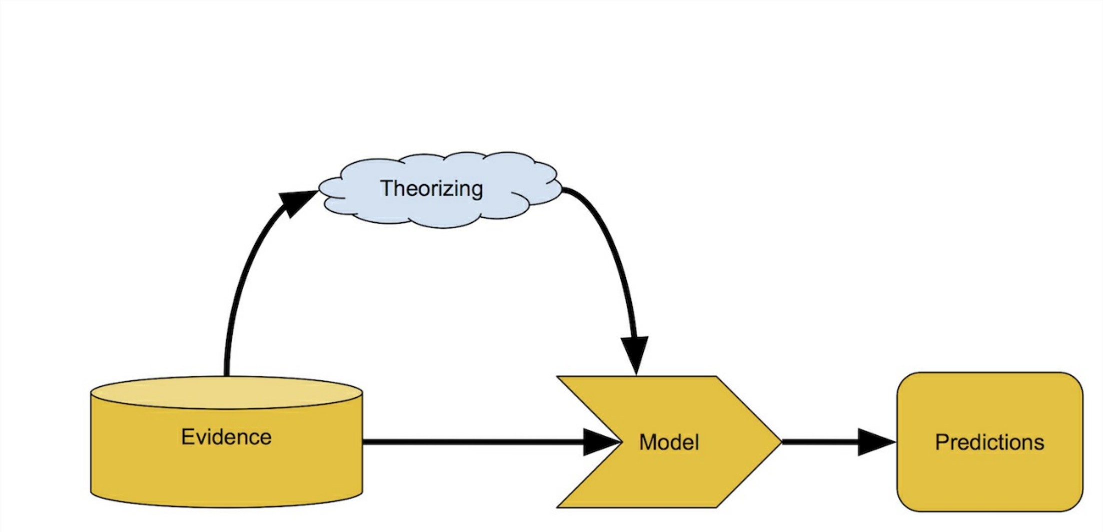
1. A lot of data, or observations, of many different variables, that we will use to build and evaluate our model;
2. The model itself: a set of rules (sometimes simple, sometimes fiendishly complex) that take those observations as input and compute on them; and
3. An output, whether it's a list of suggested friends, the chance of rain during your lunch hour, or a judgment as to how risky your home loan application is.

Let's look at two examples of a model. The first, and most fun one, is Parable of the Polygons, developed by Vi Hart and Nicky Case based on a longstanding sociological model published by Thomas Schelling in 1971.



Screencap from “Parable of the Polygons”

This interactive site walks you through some increasingly complex models, or simulations, of social segregation. In this model the inputs are the two population sizes of squares and triangles. The rules: What percentage of neighbors of a common type am I happy living near? What percentage of a different type am I *unhappy* living near, and which will cause me to move? This model simulates how such a system will evolve over time. Its output, or prediction, is the resulting community-wide level of segregation after some set number of rounds.



A simple model-building process.

The “Polygons” model provides some powerful insights that we might not have expected: Even if each individual only has a very slight bias toward having similar neighbors, the resulting society will rapidly segregate itself. Moreover, an already-segregated society will not suddenly reintegrate if you simply eliminate that bias. Individuals must actively seek diversity in order to produce a more equal society.

Here at the Getty, data-based modeling is also promising to aid the way we think about the history of collecting and the art market. I and my collaborators Sandra van Ginhoven and Christian Huemer are working on models that try to illuminate the kinds of risks that the New York dealer M. Knoedler & Co. faced when deciding to purchase an artwork, and how Knoedler may have managed that risk. Art dealing is a risky business, with few hard and fast rules. Yet, to succeed as a business, Knoedler had to have some feeling for whether some combination of a given artwork, seller history, joint ownership agreement, or buyer lead would give them a better or worse chance of turning a profit. Even in the face of a complex and difficult-to-predict market, what properties had the biggest effect on chance to make a profit?

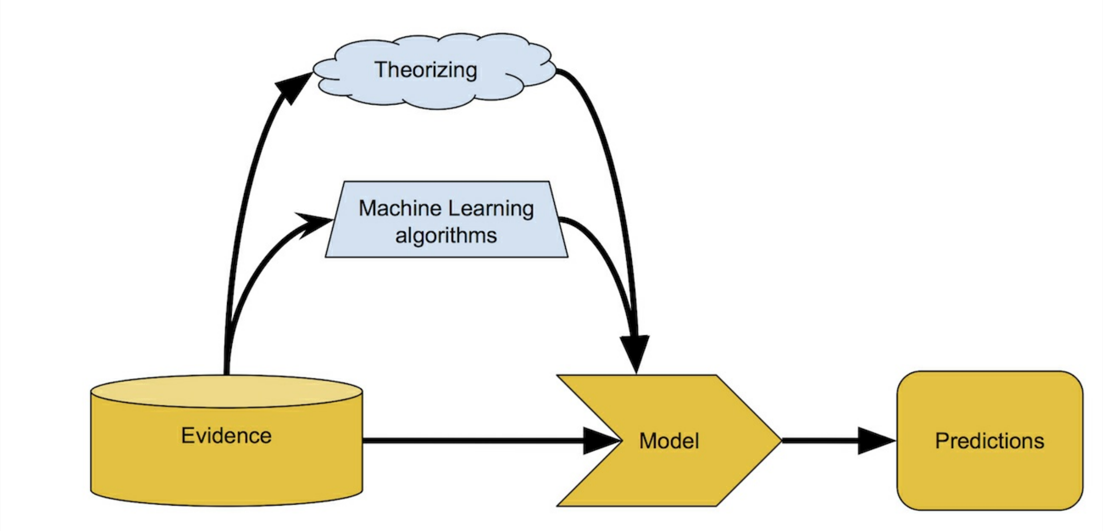
Unlike the “Parable of the Polygons,” we have many, many variables to consider when trying to build a model. That’s why we’ve turned to machine learning to help our research.

## Machine Learning

In very simple models like “Parable of the Polygons,” the model can be explicitly specified by you, the user. You can tweak the rules of the simulation by hand and get different results. You know that you have arrived at a plausible model when the simulated results match up to the observed historical results.

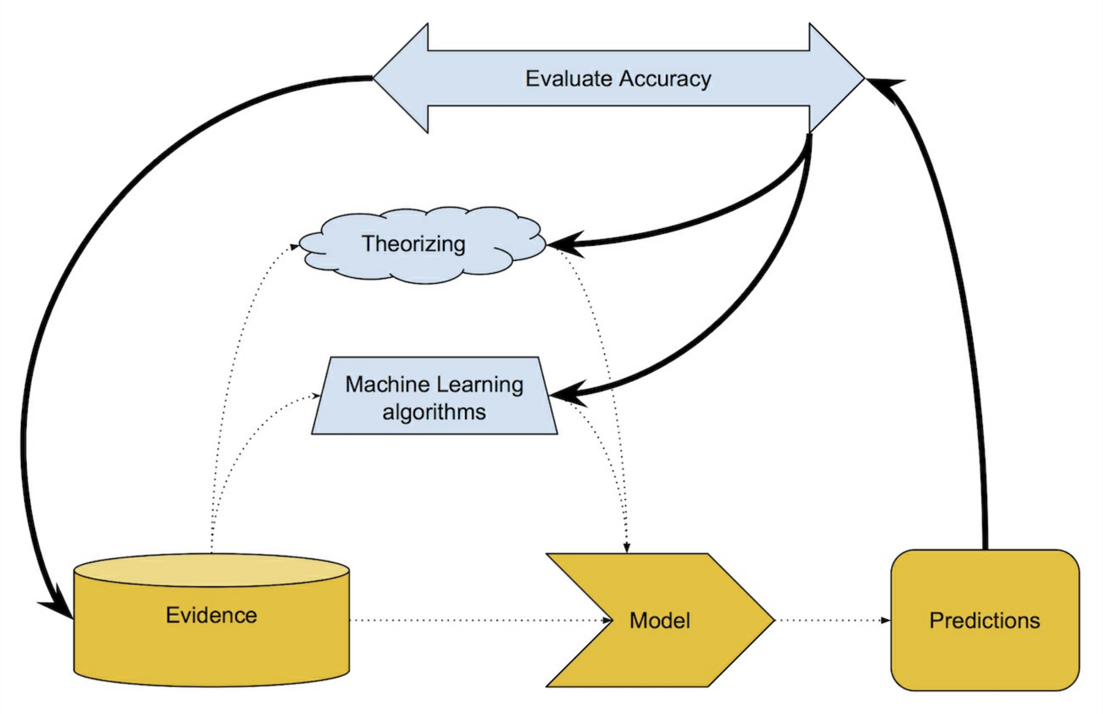
However, in more complex models like the weather forecast, Facebook’s friend suggestions, or our Knoedler profitability model, there are too many variables to tune by hand. In these cases, we can use one or more machine learning algorithms to help build a model that makes accurate predictions. **Machine learning** — the study of algorithms that can “learn,” i.e., detect generalizable patterns in data — is a sprawling field that has been around for several decades, but has recently burst into the limelight as the amount of structured data available to researchers has caught up with today’s impressive computational capabilities.(4)

There are many types of learning algorithms that produce models, from those as simple as linear regression to those as complex as recurrent neural networks. Each has their own strengths and weaknesses for different kinds of problems. While there are far too many types of algorithms to discuss in detail here, understand that these algorithms are used to *produce* the models. If they accurately predict the past, then we can look at the behavior of those models to understand how they work, and build plausible historical interpretations informed by those insights.



Building a model with the help of machine learning.

Both ML-produced models, and those derived through other means, have to be evaluated against reality. This is usually an iterative process, whether you are manually adjusting rules like in the “Parable of the Polygons,” or a machine learning algorithm is adjusting its own parameters as it searches for the model configuration with the most accurate fit to the existing data.

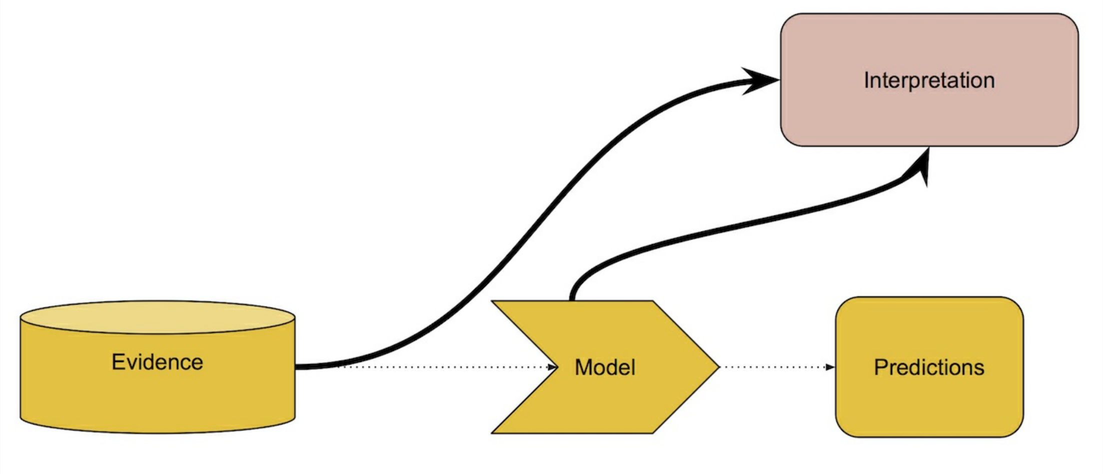


Model-building is an iterative process, where you evaluate model predictions against reality, and then adjust the model to achieve a better result.

For predicting Knoedler’s profitability, we used a random forest algorithm to try and learn how to predict Knoedler’s chance of making a profit on any one work of art that they purchased. As the inputs for our model, we used a host of variables that we derived from Knoedler’s encoded stock books in the Getty Provenance Index databases. For example:

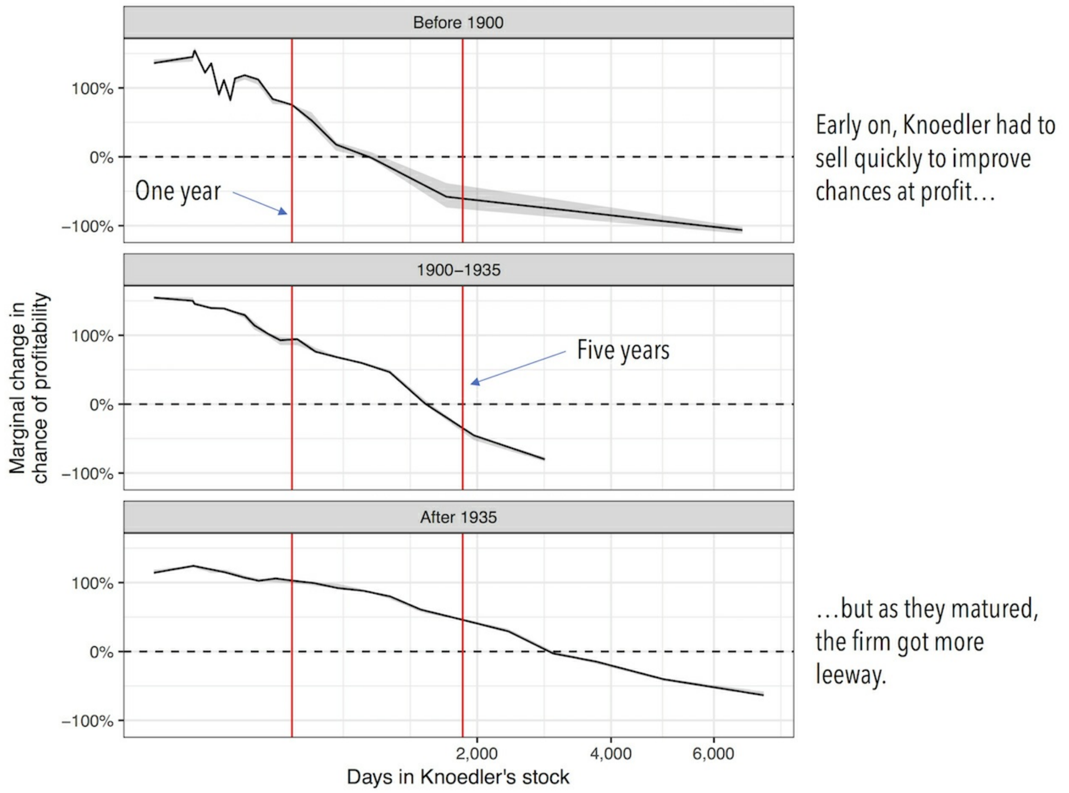
1. How much money did Knoedler spend to purchase the work?
2. Did they buy the work from an artist? A collector? A fellow art dealer?
3. How long did they hold the work in stock before selling it?
4. Was the work a landscape? A portrait? A still life?
5. Was the work by an old master, or a contemporary artist?
6. Did Knoedler sell to a longtime client, or to a customer they'd never sold to before?

Using these and a dozen other variables, we used the random forest algorithm to iterate through hundreds of potential models. The aggregate model shed light on features of the existing evidence that we weren't able to discern before.



Once you've built a model that works reasonably well, you can interpret it in conjunction with your evidence.

For example, while we weren't surprised to learn that works became less and less profitable the longer they sat in Knoedler's inventory, we didn't understand precisely how long that profitability window was, nor did we know whether that window was constant or if changed over the life of the firm. What we found was that, as the firm matured, their window grew longer and longer. This meant that they still had a good chance to make a profit on an artwork even if they'd been holding onto it for upwards of three years—a timeline that would have been unthinkable for Knoedler in their first decade of operation. This raises many new questions for us about how Knoedler developed their own business practices, and how the behavior of the larger art market in New York may have evolved over the same period.



Before 1900

1900–1935

After 1935

One year

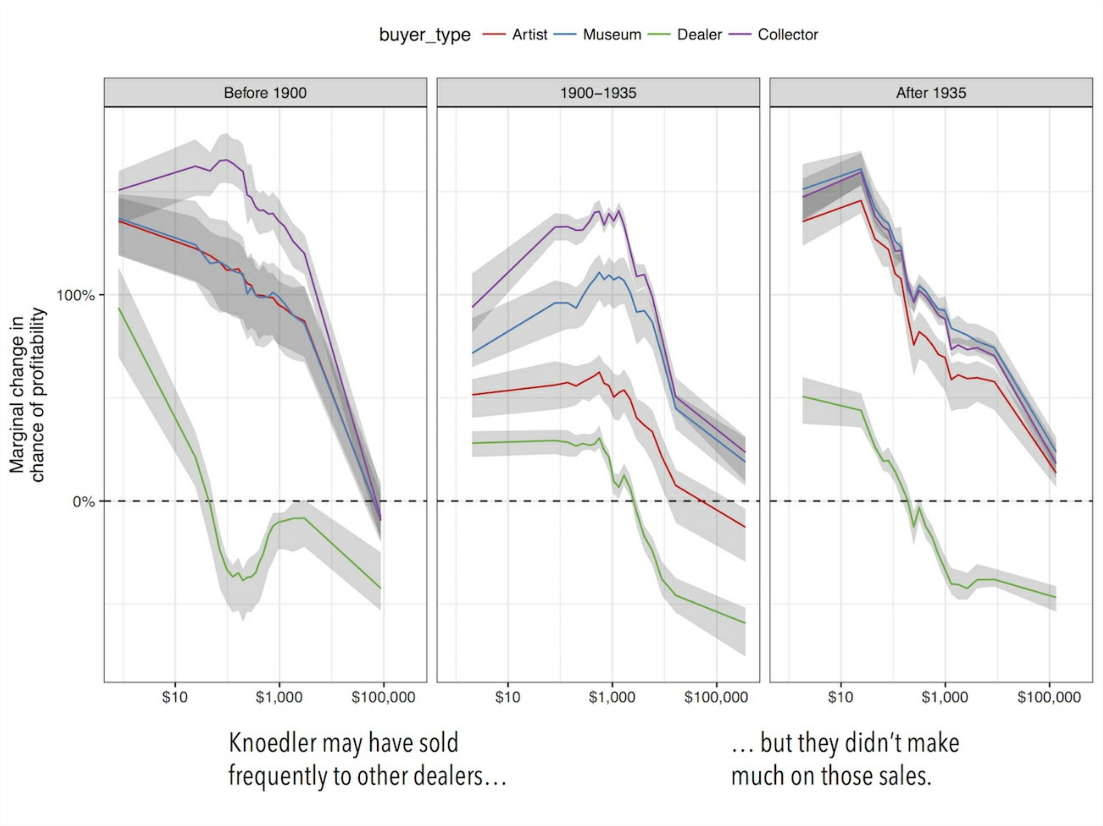
Five years

Early on, Knoedler had to sell quickly to improve chances at profit…

…but as they matured, the firm got more leeway.

The effect of “days in stock” on our model’s profit predictions for Knoedler’s firm during three different periods of the business’s operations.

Another surprising finding is changing the way we think about Knoedler’s relationships with fellow dealers. We knew that many of Knoedler’s sales were going to other dealers, rather than to private collectors or museums. But we didn’t understand how much profit Knoedler was willing to sacrifice in order to maintain business relationships with these partners. Our model suggests that selling to fellow dealers resulted in a big hit to Knoedler’s chances of making a profit.



Knoedler may have sold frequently to other dealers…

… but they didn't make much on those sales.

The effect of buyer type and purchase price on Knoedler's profitability.

This could be for any number of reasons. A fellow art dealer could serve as a buyer of last resort for a work that just wouldn't sell, helping Knoedler cut its losses. Such deals could also cement future collaborative relationships with other firms, such as joint purchasing of particularly expensive works. Thus, taking a financial loss on a few sales might mean a big return in future goodwill.

Like most good historical theories, these explanations may seem intuitive in hindsight. But we didn't even think to ask these kinds of questions until we began explicitly modeling Knoedler's business in a data-driven way. Simulating Knoedler's business, even in this simplified manner, has transformed the way we are researching the firm. Now when we find a letter where Knoedler discusses the opportunity for cultivating a brand-new client, we bring to that letter a broader understanding of how profitable new clients could be for the firm on the whole. Likewise, the firm's surviving correspondence also helps us understand why they might have cultivated so many collaborative relationships with fellow dealers, even when the sales they made to those same dealers weren't immediately profitable.

Wedding these two approaches means that we can refine the assumptions in our data and our model, while at the same better understanding what particular transactions were exceptional vs. run-of-the-mill. Such a perspective is difficult, if not impossible, to adequately capture when sifting page by page through an archive that spans almost 3,000 linear feet.

## What Does Modeling *Not* Do?

Models lend clarity to many complex phenomena. However, they also miss out on a lot of subtlety. Reality will always be messier and more complicated than our explanations for it, whether we form those explanations through data-driven models or through "traditional" narrative argumentation. It is vital that we never mistake models for *complete* representations of reality.

This shortcoming is readily apparent in "Parable of the Polygons." In that digital simulation of society, each shape could move around the board totally unimpeded. In real life, however, we know that not everyone has equal freedom to move where they please. This simulation calls attention to the complex consequences of seemingly-benign social preferences, but it does not try to model the effects of structural segregation like redlining. While it does not tell the full story needed to understand the specific historical phenomenon of, for example, housing segregation, it does supply a crucial part of the puzzle that is relevant to how social segregation manifests in a wide variety of contexts. No account of segregation would be complete if it *only* discussed this model. At the same time, no such story would be complete *without* grappling with this model and its implications for more specific and nuanced arguments.

Likewise, while looking at Knoedler's stock books gives us a crucial and informative perspective on their business decisions, they aren't the *only* source of information about the firm's history. The Knoedler Archive contains a tremendous wealth of additional evidence, from their correspondence with other dealers, to index cards used to catalog their stock, to photographs of artworks in their stock. None of our analyses are complete without merging lessons learned from these predictive models with good old detective work in the physical archives of correspondence and memoranda.

## What Do I Need to Get Started?

The tools you use to build explicit models like this will vary depending on the complexity of the work you're doing. You will always need to do data cleaning (OpenRefine is great for this). Note that the hours or days spent programming and executing a computational model are *easily* dwarfed by the weeks or months spent data cleaning—a process that can be as informative as the modeling itself! Once your data is well formatted, some simple modeling can be done in a tool as everyday as Excel; it actually does linear regression! If you're doing agent-based modeling like in "Parable of the Polygons," NetLogo is a popular platform. More advanced machine-learning algorithms like support vector machines, random forests, gradient boosted trees, or neural networks, though, require competency in a programming language like R (we use this for our work at the Getty) or Python.

In our group, we are trained art historians who are *also* well-versed in programming and computational model building, so we merge both subject knowledge and computational know-how. It's more common, though, to have these skills spread out among several different team members. In order to succeed, such relationships have to be equally collaborative, with each party speaking at least a little bit of the others' language.

## Key Takeaway

Computational simulations will never capture the full complexity of history—but then again, historians aren't here to tell *everything*. We are here to distill evidence into a cogent argument. Explicit models, whether or not they use large datasets or machine learning, are a great help to historians in pursuit of this goal. Like an X-ray, a simplified model does flatten our view of the subject. But it also reveals a perspective that would otherwise have gone unseen.

## Notes

1. Unlike meteorologists, though, Facebook has the power to *modify* people's behavior based on their predictions. This blurs the lines between a model of a real phenomenon and the real phenomenon itself; Kieran Healy writes perceptively about this effect in "The Performativity of Networks," *European Journal of Sociology / Archives Européennes de Sociologie* 56, no. 2 (August 2015): 175–205.
2. Joshua M. Epstein, "Why Model?," *Journal of Artificial Societies and Social Simulation* 11, no. 4 (2008).
3. You'll note that I discuss models providing "plausible" answers, not actual ones. To read more about this distinction and why it's important for historians, take a look at Scott B. Weingart, "Abduction, Falsifiability, and Parsimony: When Is 'Good Enough' Good Enough for Simulating History?" 2013.

4. A concise and general-reader-friendly introduction to machine learning is Ethem Alpaydin, Machine Learning: The New AI, MIT Press Essential Knowledge series (Cambridge: MIT Press, 2016).