

# Data Science Strategies for Real Estate Development





# WHY IS DATA IMPORTANT FOR REAL ESTATE?

For over thirty years now, there has been a push within the real estate industry to gather a sufficient amount of real estate financial data in order to enable the sector to compete with other financial assets, like that of equities, bonds, commodities and even FX trading. Starting at the turn of the 21st century, a growing number of core data companies began collecting information about real estate transactions and rents, but this notion of a systemic approach to understanding patterns of risk and return in real estate was quite new. After the Great Financial Crisis (GFC), however, there was a tremendous push towards developing information that could help explain the patterns of events that systemically impact society, and in what ways we can learn how to avoid poor governance in the commercial and residential real estate markets.

Fast forward to more than a decade past the GFC: what we see is a data science movement in real estate. There is a shift from just a few data science providers to well over 300 data providers in the marketplace today. This rapid expansion and growth in the real estate sector has created a transformation in how business transpires and the role of data scientists, statisticians, econometricians and machine learning specialists in helping decision makers to answer questions.

One core area where data science for real estate can improve efficiency is in the very disaggregated and fractured process of real estate development, where so many different facets of the built environment domain must come together with one common purpose: to

construct a building. Geographers, planners, architects, contractors, banks and developers must coordinate over a melange of documents and spreadsheets to reinvent the making of a building each time, relying on the tacit knowledge of a few individuals. However, data science is working to expand the margins of knowledge to help coordinate and organize the industry to unlock better buildings for sustainability, equitability and health. By the same measure, it also helps to deliver more financially efficient and profitable outcomes.

For data science to be truly helpful, data scientists must really listen to what the data is saying about the events and experiences being captured. Although data science is a field more commonly known for complex algorithms, its fundamental tenants are embedded in basic, statistical practices like understanding the distribution of experiences and minimizing bias.

In this brief, we outline what data can be for real estate development in a broader sense. We ask the following: what real estate data is out there? Where is data science headed in helping us to answer questions? And, ultimately, how can data science help us build better buildings and districts through the real estate development process?

“CORE SKILLS FOR DEVELOPMENT, DESIGN AND PLANNING ARE SHIFTING TO ENCOMPASS ANALYTICS IN DATA SCIENCE AND MACHINE LEARNING.”

DR., ANDREA CHEGUT  
MIT REAL ESTATE INNOVATION LAB



# FIRST, WHAT IS DATA?

## Data is not about 1/0s

Data has the unfortunate reputation for being something very intangible and abstract. Yes, it is a representative set of information that we use to signal that an event has occurred. But true statisticians, economists, and data scientists are always working to get at something deeper. We are all working to uncover something called the true data generating process.



## Data is about us

Data is, in fact, really about us. It is about our human, emotional, mental, physical, connected and disconnected experiences. It is our collective story - what is common and what is quirky and unique about all of us. If data is truly going to be representative of our data generating process then it has to be mindful of us.

## Data must listen

As data scientists, we are not all that different from an empathetic friend or a psychologist. We spend a lot of time listening and observing what the data has to tell us. Great data scientists and modelers listen, and when they get the story wrong they go back to the data where people told their stories and try and get it right again.



## Data can minimize and expose bias

Data shows us our “shared” collective experience. When we use models and statistics, we uncover facets of the data that can display our bias or hide it from relevant stakeholders. As data scientists, we have an ethical and technical responsibility to remove bias from our models to improve the understanding of our results and their impact on stakeholders.

# SEIZING THE DATA SCIENCE OPPORTUNITY

In the real estate industry, the adaptation of data science has been mainly focused on commercial real estate pricing and valuation. Researchers and practitioners use data science to calculate price indices that determine property market performance, to estimate property value based on numerous property characteristics, for forecasting real estate performance based on economic trends, and more. However, the real estate industry is not made up of just financing and valuation. There are vast areas of opportunities in applying data

science to the rest of the real estate functions and product types, both commercial and publicly funded properties.

Some opportunities that we explore in this paper, pertain to real estate development specific functions like land selection, design and construction. In addition, we explore the intersection of design and financial forecasting to enable the physical production of buildings in the development process.

## THE VALUE PROPOSITION

NEARLY 60% OF PREDICTIVE POWER CAN COME FROM NONTRADITIONAL VARIABLES IN THE REAL ESTATE INDUSTRY.

SOURCE: MCKINSEY & COMPANY, 2018

## WIDE DATA IS THE NORM

VARIOUS PRODUCT TYPES, FUNCTIONS, STAKEHOLDERS AND REGIONALITY IN REAL ESTATE ADD CHALLENGES IN THE APPLICATION OF DATA SCIENCE.

## SCATTERED DATA SOURCES

NUMEROUS DATA AND SOLUTION PROVIDERS, EXISTING AND NEW, ARE AVAILABLE BUT THERE'S NO SINGLE DEPOSITORY OF DATA FOR REAL ESTATE PROFESSIONALS TO EASILY ACCESS.

## FORECASTING POWER

MORE THAN 80% OF INVESTORS THINK THAT PREDICTIVE ANALYTICS AND BUSINESS INTELLIGENCE SHOULD BE PRIORITIZED FOR CRE BUSINESSES.

SOURCE: DELOITTE, 2019

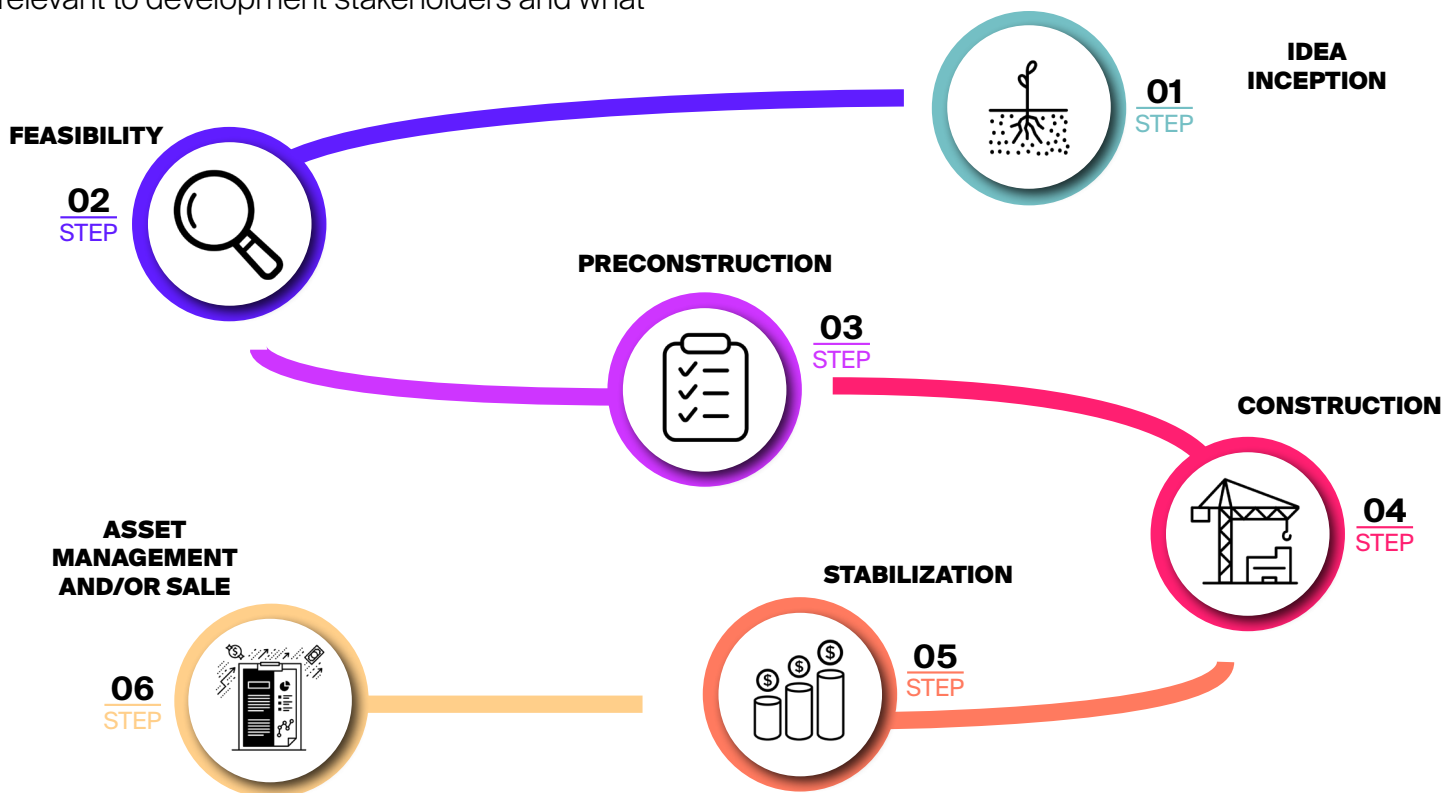
# DATA SCIENCE FOR REAL ESTATE DEVELOPMENT

Data science is at work for real estate more generally, but not for the real estate development process overall. In identifying relevant data for the real estate development process, comprehension of the overall process and individual tasks is essential to understand where data science can help stakeholders. Additionally, the differences arising from regionality, real estate function, and product type need to be considered together to determine a wider framework for a data science strategy.

To do our analysis, we endeavored to understand real estate development through six stages: idea inception, feasibility, preconstruction, construction, stabilization, asset management and/or sale. Further within each stage, we worked through Bulloch and Sullivan (2009)'s 98 development tasks that suggest each developer works through from start to build. By understanding phases and tasks, we deconstruct the outcomes that are relevant to development stakeholders and what

are necessary to help make a viable decision to drive that outcome. These features of decision making help us to dig into what data providers and variables are needed for developers to curate data science strategies.

Much of the real estate development process is spread across various stakeholders and decision makers, so data science becomes an understanding of how we can learn about our own development experiences from various perspectives. Ultimately, real estate development will need its own data science platform that can integrate data across geography, geometry and time. These types of strategies are strong in different thought leader domains like architecture, planning or finance, but not across all domains. Additionally, they are aimed at isolating the data that is needed and how integration across these different groups can begin to make progress over the next decade.



SOURCES: MIT REAL ESTATE INNOVATION LAB, BULLOCH, B.B.E. AND SULLIVAN, J., 2009. APPLICATION OF THE DESIGN STRUCTURE MATRIX (DSM) TO THE REAL ESTATE DEVELOPMENT PROCESS, DISSERTATION, MASSACHUSETTS INSTITUTE OF TECHNOLOGY).

# BUILDING A DATA SCIENCE ACUMEN: LEARNING ABOUT OUR OWN DEVELOPMENT EXPERIENCES

One critical aspect for any data science strategy for the real estate development process is the level of usage of internal and external data sources in each phase of real estate development. Namely, are we already gathering data about our own development experiences to be able to listen and learn from our own development portfolio? Furthermore, are there external experiences of development events outside of my own firm? It's important to keep in mind that many firms are still in the data collection phase for their own internal development events, but several have moved forward with integrating their data from development events with those external development projects.

The understanding of internal and external data can significantly help developers decide where and when to deploy resources for data science, i.e., purchase external data for market analysis or focus on internal data using data management solutions. For example, a developer needs more external data than internal data during the initial planning stage because he or she needs to analyze the market and analyze potential sites before deciding on one. However, when in the feasibility stage there is a great need to work at the intersection of internal and external data with prior experiences and return expectations.

The first step in any real estate development strategy for data science is using the data a firm already has based on their prior development experiences. Instead of letting spreadsheets, development proposals (both failed and successful) and operational details live on a hard drive, work towards creating an architecture that can be shared in the cloud. This can help to build a data science acumen.

## Recognition

### Step 1

#### From

#### Basic Internal Data

- Core financial data
- Business operations
- Back office organization
- In the cloud!



#### Towards

#### Data Integration

- Across systems
- Between knowledge domains
- Integrated across stakeholder tasks
- Towards customer engagement





# UNDERSTANDING THE CONTEXT: DATA PROVIDERS FOR THE DEVELOPMENT PROCESS

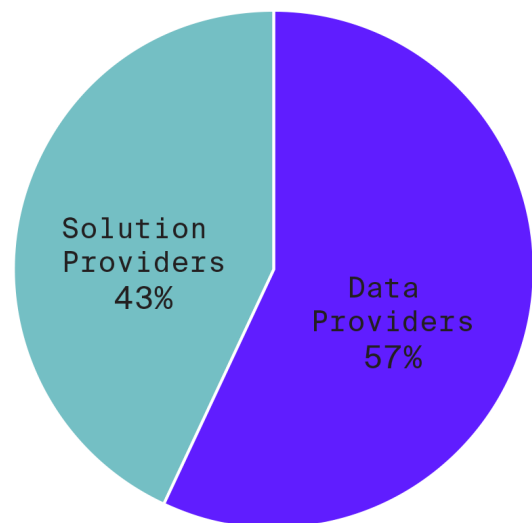
Understanding the current landscape of available data providers can help prepare real estate stakeholders in making better decisions on a data science strategy. We broke down data providers, who collect, aggregate and disseminate data, and solution providers, who provide systems to organize a firm's data. We surveyed both relevant data and solution providers and analyzed their pertinent characteristics for decision makers. Some of the variables captured included interactive platform availability, application programming interface (API) availability, real estate product type coverage, regional coverage, data volume and frequency, and underlying data collection and verification methods. Among these factors, API is a relatively new characteristic for real estate data companies. It acts as an intermediary function that allows users to connect source data or softwares directly onto their own platform.

The analysis of a database of data and solution providers in this research revealed:

- The distribution of data and solution providers showed similar importance in utilizing both internal and external data in making business decisions.
- Among the organizations that provide external data, data platforms were the most popular style of data delivery.
- Almost a quarter of companies relied on crowdsourcing for data collection. This style of data gathering is growing and could be considered more reliable as databases develop.

- One-tenth of the companies provided hardware to enable internal data collection, and this is critical for companies who have not started their data science strategy yet.
- More than half of the data companies provided API, indicating the industry's direction towards a more fluid and connected use of data from various sources.

## Distribution of Data and Solution Providers



“THERE IS CURRENTLY A LACK OF RESEARCH INTO BIG DATA’S ROLE IN BETTER UNDERSTANDING COMMERCIAL AND INDUSTRIAL MARKETS, , AS WELL AS INTO REAL ESTATE DEVELOPMENT OR INVESTMENT TRENDS.”

DR. KIMBERLY WINSON-GEIDEMAN  
UNIVERSITY OF MELBOURNE



# DATA PROVIDERS AND DEVELOPMENT STAGES

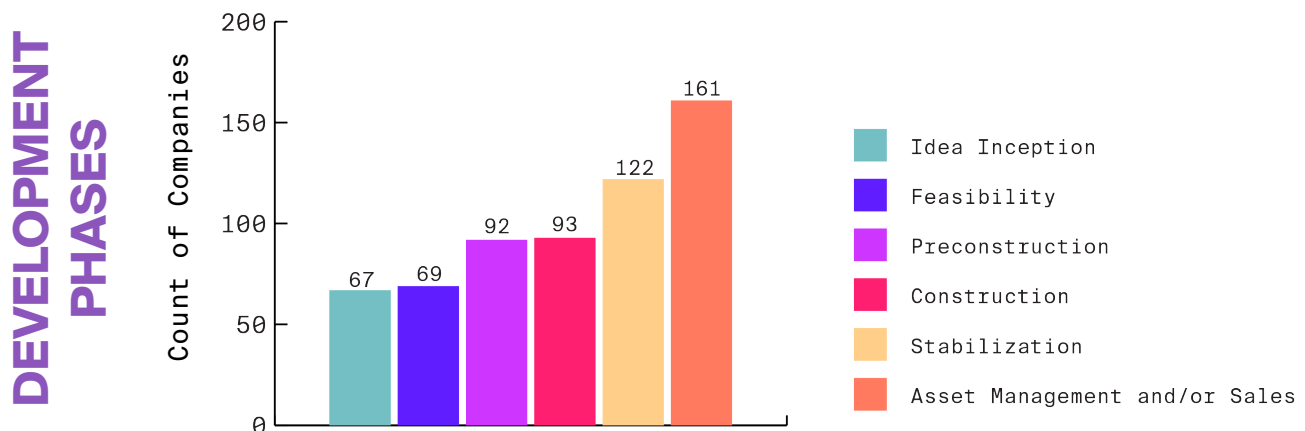
We also found that the number of applicable data and solution providers increases as development process progresses from Idea Inception to Asset Management and/or Sales. This observation can be attributed to the fact that while tasks related to conducting market analysis or feasibility analysis, as well as updating project underwriting, continue throughout the real estate development process, tasks related to construction and operations form newly created internal data that need to be additionally managed. What this means for real estate development is that there is a need for internal data solution providers to help design internal data architectures for efficient real estate development processes.

The most frequently used solution type that helps developers manage internal data during construction is image capturing. This solution helps developers to record the status of construction progress. Doxel.AI deploys a robot surveyor with cameras, and Openspace uses smartphone cameras - or cameras attached to construction hard hats - to capture construction

site images and inspect construction progress. Through computer vision, these solutions can analyze and process images into actionable data that site managers can easily implement into their daily operations. Similarly, Aspec Scire utilizes drone-attached cameras to capture site images for various construction and development related tasks, such as land survey, construction progress tracking, and project completion assessment.

Moving from the construction stage to stabilization, developers need to listen to internal data more than ever. The most frequently used solution type is sensor and tracking technology that detects and captures built environment data such as temperature, water leakage, energy consumption, machine productivity, and space usage. These solutions are often provided as part of an integrated building operations and property management platform that allows property managers to act upon insights as a preventive or reactive maintenance measure.

**Distribution of Data Companies by Real Estate Development Phase**



Note: Double counting allowed as certain data could be applied to multiple phases.

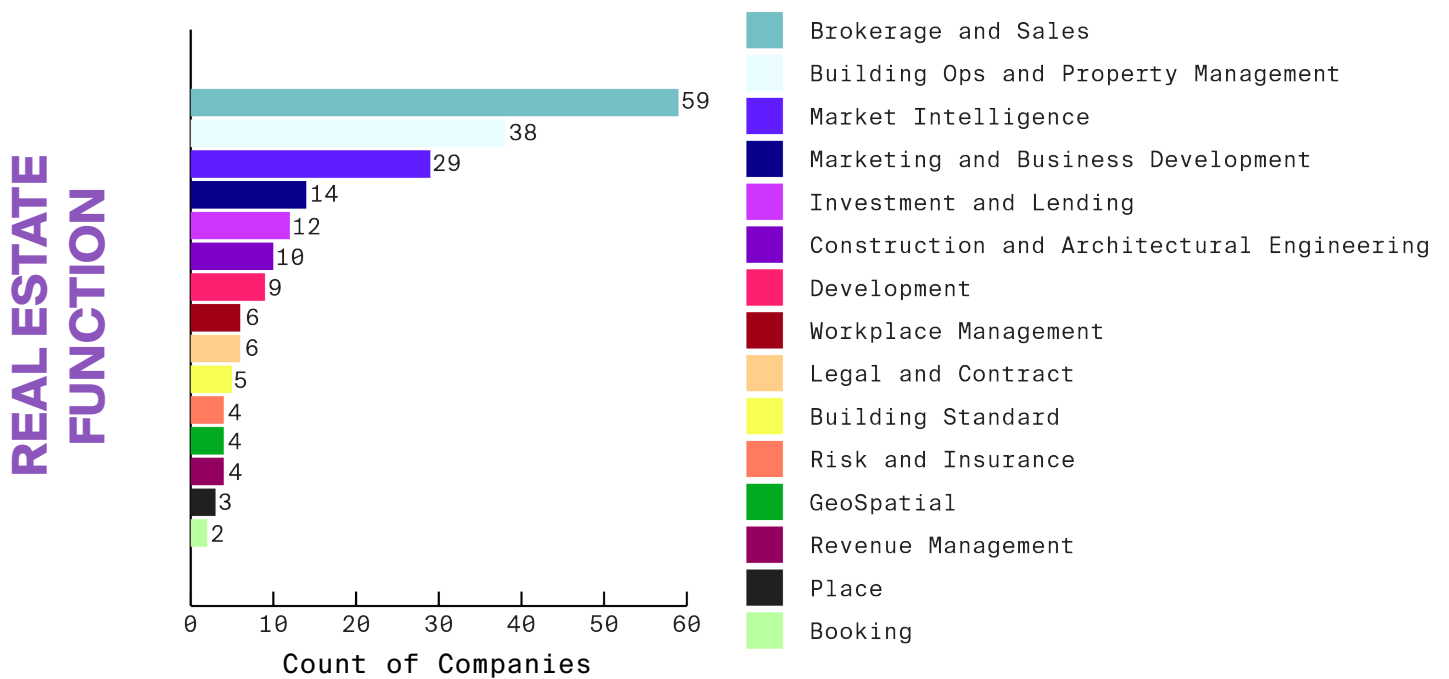
# FIRST DATA TO SCALE THE NUMBER OF USERS

The largest number of data and solution provider companies we have documented and analyzed related to the Brokerage and Sales function of real estate. This observation is aligned with the current market sentiment that an app exists for every step of the home-buying and home-ownership experience.

Evidently, Brokerage and Sales data is now prevalent and easily accessible by the public, and often at no cost. This data group includes data points like a property's physical characteristics, the neighborhood it's located in, boundaries, title and ownership history, valuation, and loan options. While several big players are

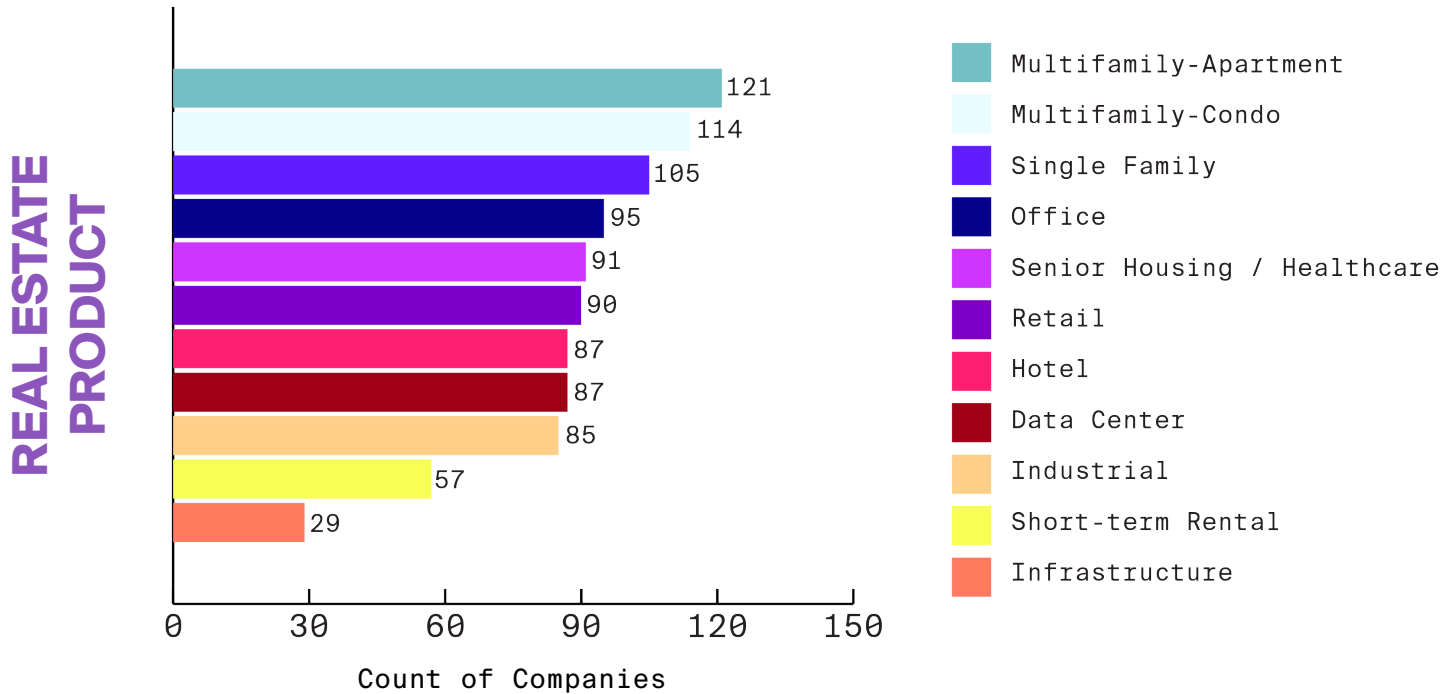
dominating the buyer-seller marketplace portals such as Realtor.com, Redfin, and Zillow, some companies are changing the way the public uses data. For example, Realscout and Remine both encourage the home search process to be conducted by agents and clients together from the beginning, so that they collaborate through the home search process and seamlessly transition to transactions based on trust built through their experience together.

**Distribution of Data Companies by Real Estate Function**



# DATA AND SOLUTION PROVIDERS FILLING THE GAP

Distribution of Data Companies by Real Estate Product Type



“WITH PROXIMITY TO CONSUMERISM, RESIDENTIAL REAL ESTATE HAS MORE DATA AND SOLUTION PROVIDERS, AND MANY OF THEM STRIVE TO SOLVE OPERATIONAL ISSUES. HOWEVER, THERE IS SO MUCH MORE TO BE DONE FOR DATA SCIENCE AND REAL ESTATE DEVELOPMENT. “

SUNNIE (SUN JUNG) PARK  
MIT REAL ESTATE INNOVATION LAB



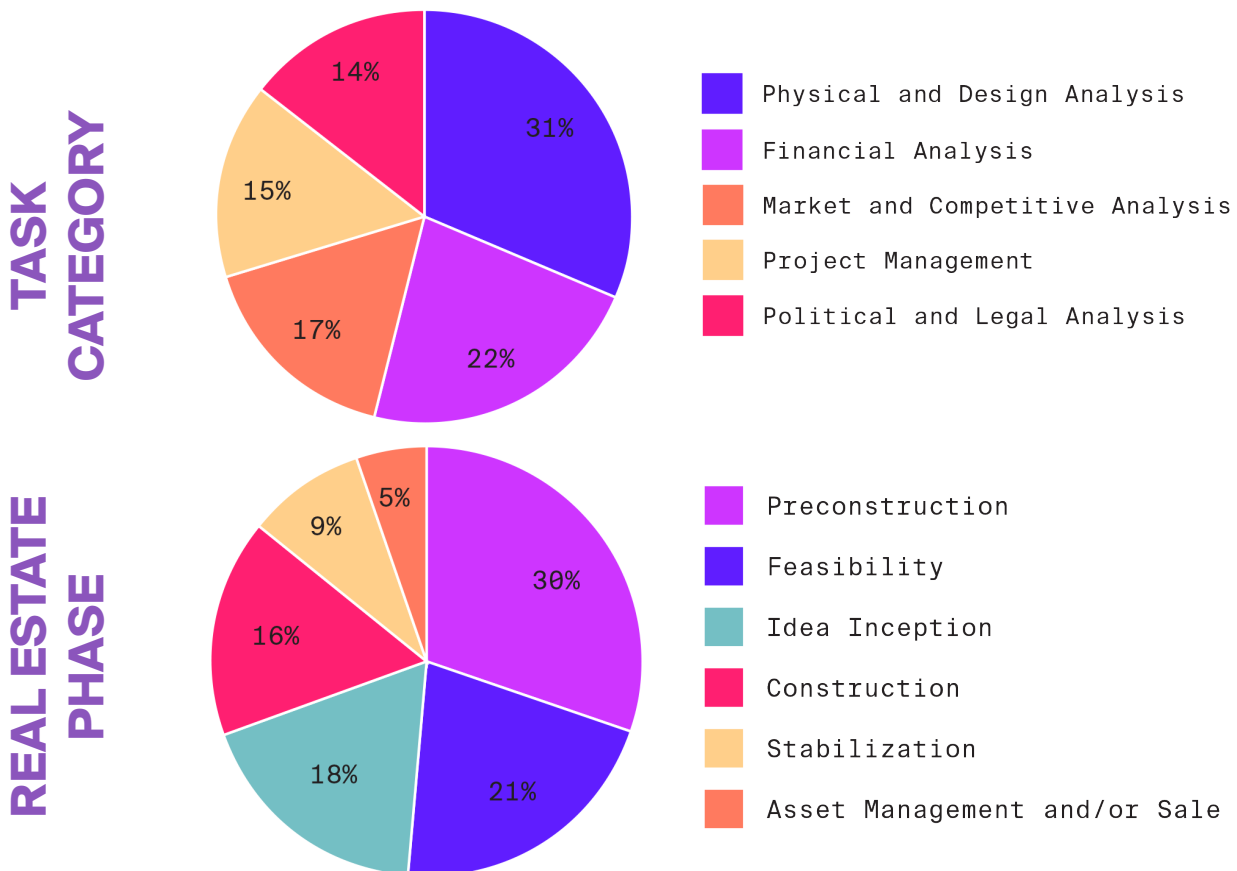
# DEVELOPMENT TASK ANALYSIS

We use the study by Benjamin Bulloch and John Sullivan to break down the number of real estate development phases and tasks, which indicates that Physical and Design analysis have the highest number of tasks. In short, where you are going to build and what you are going to build takes up the largest amount of task time for a developer and the team. This process is currently undergoing a transformation, as geographers, computational architects, planners and data scientists increasingly turn towards data science to expand their decision capacity. Furthermore, when we look at real estate development phases and the number of tasks within them, we find that more than

half of the tasks fall under Preconstruction and Feasibility phases.

These findings suggest priority areas from a developer's perspective and can help data scientists prioritize data collection and analysis. Currently, data scientists search for pain points in the data collection and analysis stages. Developers can help data scientists by getting educated about difficult decisions in the development process where data can be used.

**Distribution of Tasks by Task Category and Real Estate Development Phase**



# A DATA COLLECTION FRAMEWORK FOR EVERY DEVELOPMENT

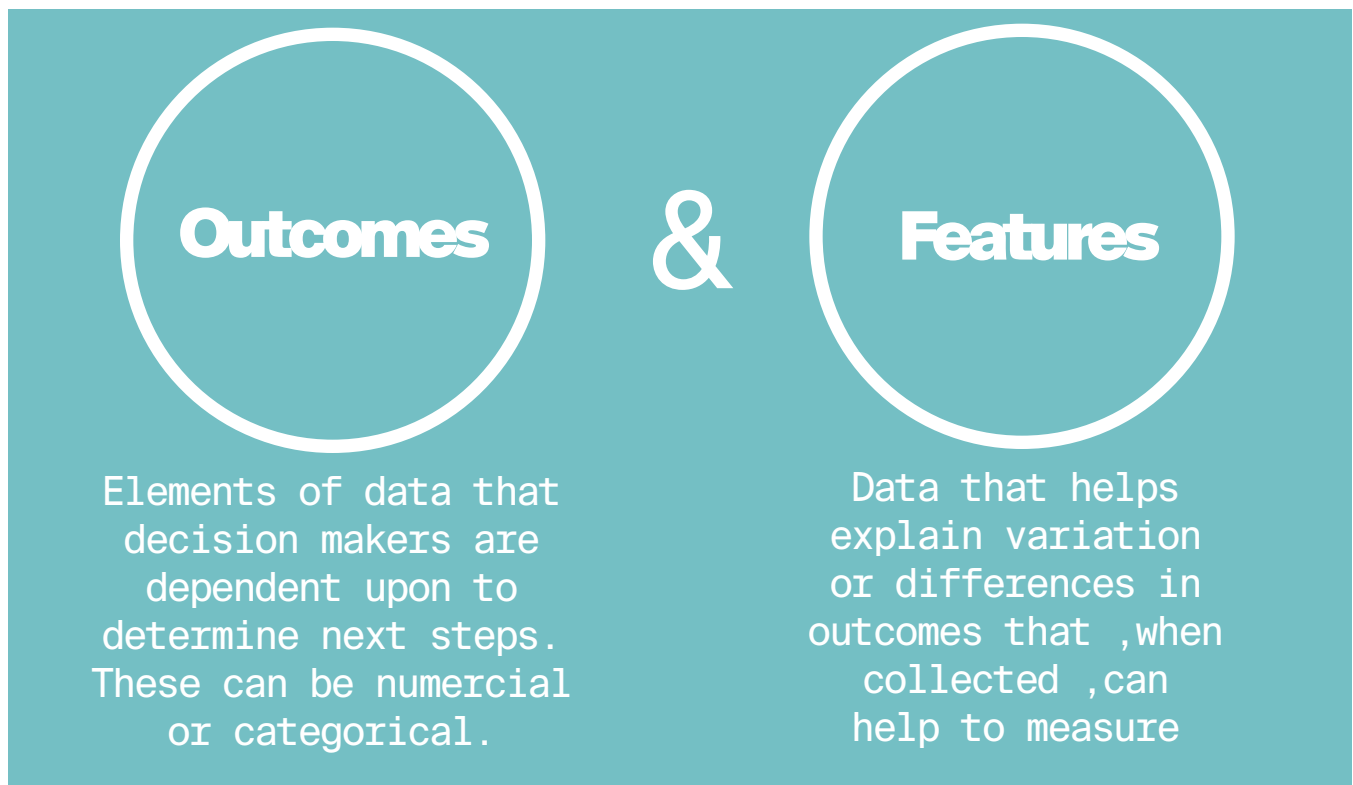
As data scientists, we want to explore how we can bring a modeling lens to the real estate development process. For each real estate development, we looked at the phases and the tasks within them. Furthermore, for each development we have identified relevant outcomes for decision makers that are needed to move the needle towards a relevant development decision. Lastly, we have explored what features or factors are necessary to explain that outcome.

We have looked deeply at each of the outcomes and feature types to discern where data and solution providers are delivering, where real estate developers can be capturing their own data, and where data science needs to expand to understand the data needed for the real estate development process.

For the real estate development process, it is important to identify critical outcomes that push decision making forward. What specific numeric outcome needs to be collected or, in some instances, what outcome category needs to be collected to enable a decision?

Outcomes are driven by specific features or factors that help drive or model their outcome. We collect feature data to help us make decisions and, in traditional frameworks, real estate developers do not systematically collect them. These can range from the location of developments, the urban context of the development, the allowable FAR and even design decision.

Combining the outcomes and features collected within an organization and across development projects can empower better decision making and more profitable development.



# OUTCOMES FOR DECISION MAKING

We looked at 98 individual tasks in the real estate development process and from these tasks there are 71 unique outcomes that developers need to calculate, measure or sort to make a decision.

- Project IRR (Internal Rate of Return) is the most common outcome necessary for making a decision in real estate development.
- Half of the tasks listed required categorical outcomes compared to numerical outcomes, or a combination of both.

The resulting dataset of outcome and feature variables for each step of the real estate development process can be used for further studies, using regression to understand which feature affects an outcome the most, highlighting critical variables for businesses.

**Frequency of Outcome used in Real Estate Development Process**



Note: Other non-repeated outcomes were omitted for the purpose of better visualisation.

DATA SCIENTISTS UTILIZE **ECONOMETRICS** TO EXPLAIN THE DATA SET'S **CAUSALITY** BY TESTING HYPOTHESES. AND, THEY USE **MACHINE LEARNING** TO PREDICT THE FUTURE BY LEARNING THE PATTERNS OBSERVED IN THE PAST.

SOURCE: FORBES

**ECONOMETRICS** AND **MACHINE LEARNING** UTILIZE **REGRESSION** AS A STANDARD TOOL TO UNDERSTAND THE RELATIONSHIP BETWEEN VARIABLES.

SOURCE: FORBES

**BIG DATA** IS NOT ONLY MASSIVE IN AMOUNT BUT ALSO CHARACTERIZED BY ITS VARIETY, COMPLEXITY, AND THE SPEED AT WHICH IT NEEDS TO BE ANALYZED OR DELIVERED.

SOURCE: HARRY PENCE, JOURNAL OF EDUCATIONAL TECHNOLOGY SYSTEMS

THE PROCESS OF CATEGORIZING DATA TYPE IS ESSENTIAL IN DATA SCIENCE AS SOME STATISTICAL COMPUTATION IS ONLY APPLICABLE TO SPECIFIC DATA TYPE.

SOURCE: UCLA INSTITUTE FOR DIGITAL RESEARCH & EDUCATION STATISTICAL CONSULTING

REAL ESTATE DATA IS WIDE; ANOTHER STUDY FOCUSING ON COMMERCIAL REAL ESTATE FOUND 903 UNIQUE FEATURES THAT DESCRIBE PROPERTY CHARACTERISTICS.

SOURCE: RYAN STROUD, MIT MSRED 2017

AS AN INTERDISCIPLINARY FIELD, DATA SCIENCE AMALGAMATES SCIENTIFIC LEARNINGS FROM STATISTICS, ADVANCED MATH, ALGORITHMS, AND MODELING. COMBINED WITH BUSINESS KNOWLEDGE, IT CAN FIND PATTERNS AND CONSEQUENTLY MEANINGFUL INFORMATION FROM LARGE SETS OF DATA.

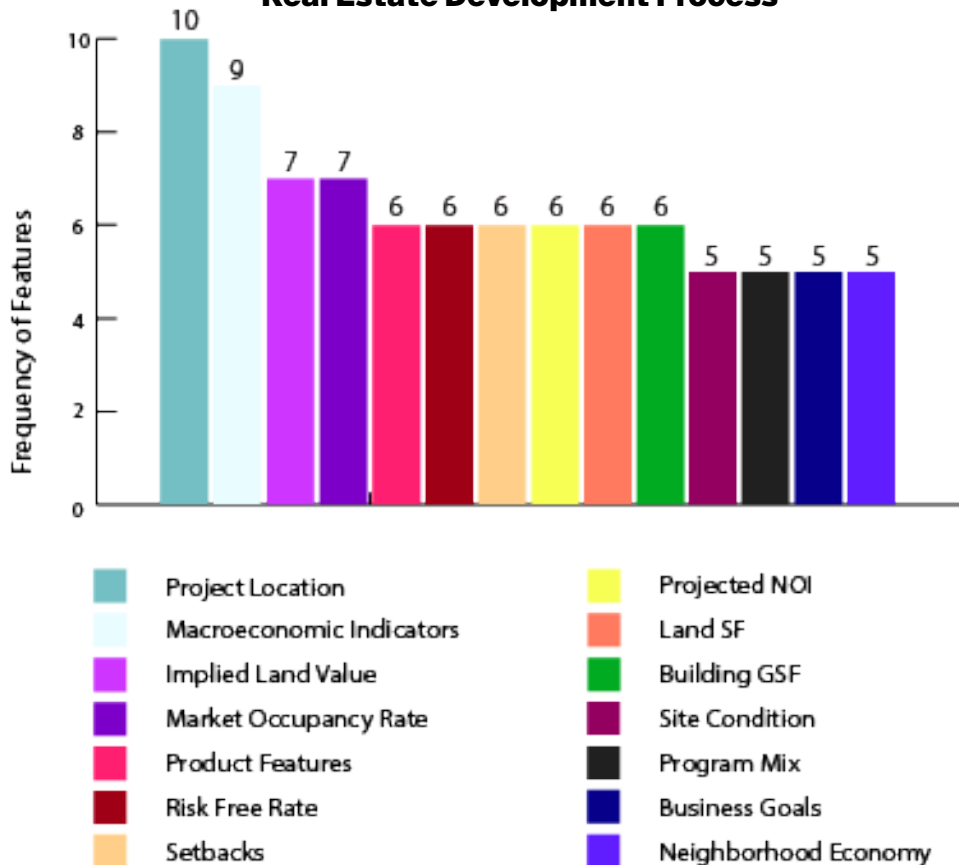
SOURCE: DATA SCIENCE ASSOCIATION

## FEATURES FOR DEVELOPMENT DECISION MAKING

We identified, at least 588 unique features or data points that are required across the real estate development process. Some are consistently needed and others are not. For example, relative project location is the most relevant feature for every stage of real estate development. However, there are other features consistently needed, but with only limited data providers who can get access it. The result of feature analysis showed:

- Each task outcome had, on average, 11 features, a median of 9 features, and ranged from two to 44 features.
- Out of 588 unique features, Project Location was the most frequently appearing feature.
- The majority of unique features (96%) appeared four times or less, and 65% of unique features appeared only once.

**Frequency of Features used in the Real Estate Development Process**



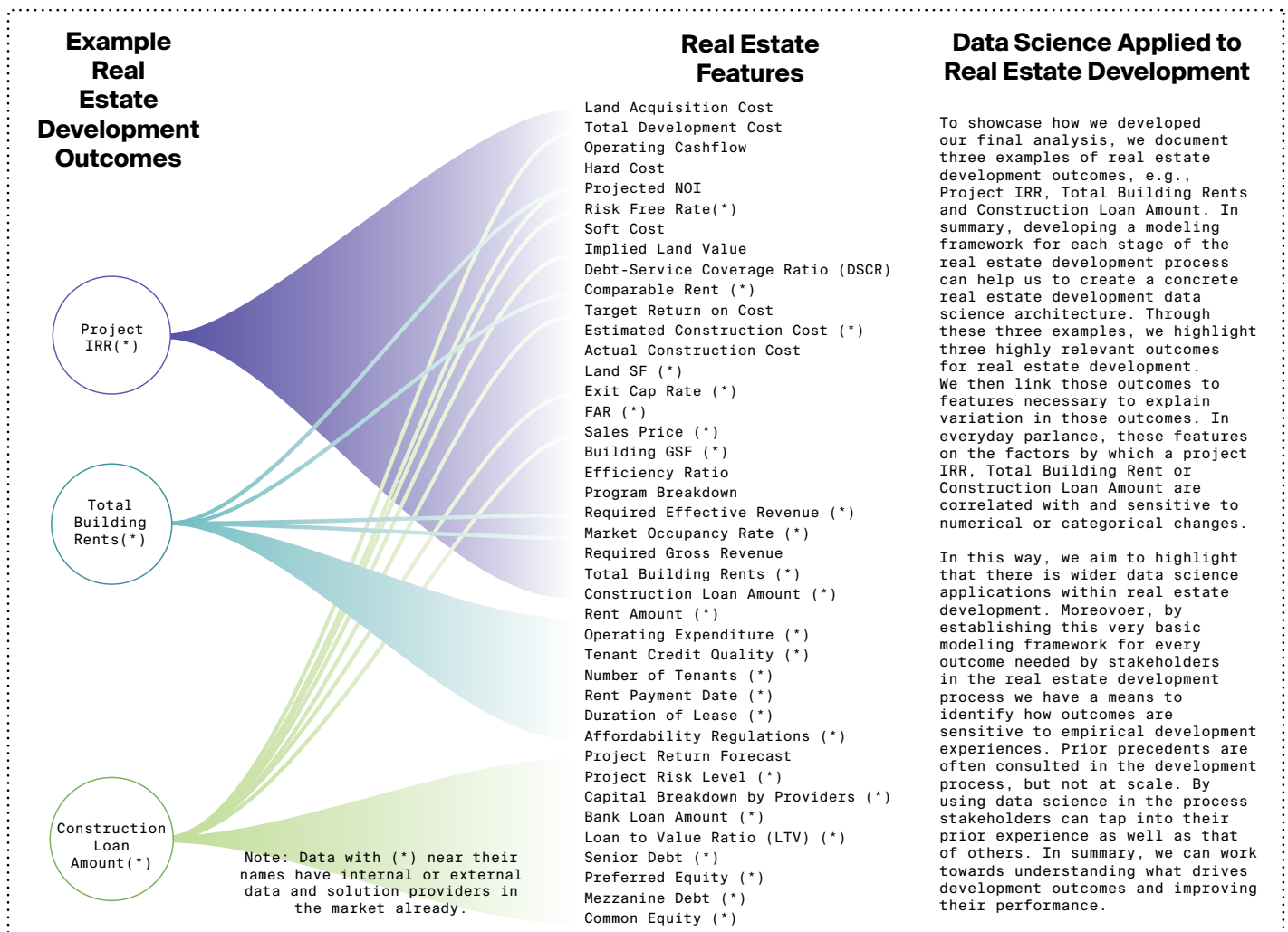


# DATA SCIENCE MODELING FOUNDATIONS

To really link data science with econometric and machine learning modeling. We have looked at the relationship between outcomes and features across every stage of the real estate development process. To do so, we developed a model for every real estate development outcome to understand what features were needed for stakeholders to understand the variation in outcomes. The result of the analysis demonstrates some basic modeling expectations for real estate development. Some key findings of our analysis are outlined below:

- Each real estate development outcome had, on average, 11 features that were required to deliver a model of the real estate development outcome.
- The number of features needed to model a given real estate development outcome, ranged from two to 44 features.
- We identified 588 unique features. Project Location was the most frequently appearing feature, which further confirms the extent to which data science for real estate is firmly embedded in geographic science too.

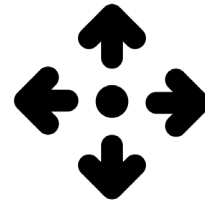
## Modeling Real Estate Development



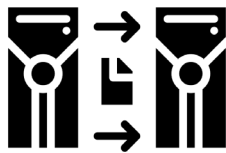
# RECOMMENDATIONS



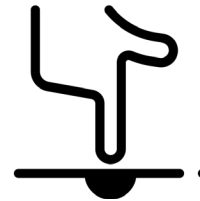
Similar numbers of external data providers and internal data management solution providers suggest that developers need to **balance resource and investment allocation** between them.



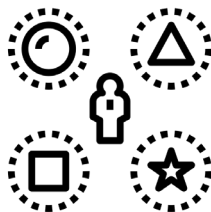
The breakdown of outcome and features of every task in the real estate development process is a **foundation from which data scientists can choose to help decision makers answer questions.**



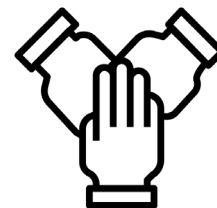
The overall methodology used in this research in analyzing individual tasks and identifying outcomes and features demonstrates an approach that can be **replicated in mapping out data requirements** throughout the development process.



Areas like physical and design analysis with more related tasks but not enough data need **more attention from data scientists to help integrate across the development process.**



The distribution patterns of data and solution providers indicate **various factors that developers should consider**, such as interactive platform and API availability, real estate product type and regional coverage, data volume and frequency, and underlying data collection and verification methods.



Half of development tasks have categorical outcomes and thus calls for **a stronger collaboration between data scientists and real estate developers** in defining appropriate measurements together.

“THE NEXT TECHNOLOGY DEVELOPMENT I’M WATCHING IS REAL ESTATE INTELLIGENCE... THERE’S AI, BIG DATA, WIDE DATA, MACHINE LEARNING, NEURALEARNING AND SO MUCH MORE IMPACTING OTHER INDUSTRIES, AND WE’RE GOING TO SEE IT ACTUALLY IMPACT THE REAL ESTATE BUSINESS, TOO.”

STEVE WEIKAL  
MIT REAL ESTATE INNOVATION LAB  
MIT CENTER FOR REAL ESTATE



# NEXT STEPS IN DATA SCIENCE FOR REAL ESTATE

When applying data-driven real estate development, developers should examine their data requirements across the development process, balance their investment between external data and internal data management, and understand product variety and limitations.

We suggest that developers start with creating a data science architecture that moves their development projects from Excel to a data architecture and platform in the Cloud. This has many benefits: shared access amongst the team, organizational frameworks for reporting and performance, but also systematic understanding of outcomes and features over time.

The overall methodology used in this research in analyzing individual tasks and identifying outcomes and features demonstrates an approach that can be replicated in mapping out data requirements throughout the development process. Similar numbers of external data providers and internal data management solution providers suggest that developers need to balance resource allocation between them. The distribution patterns of data and solution providers based on product characteristics indicate various factors that developers should consider in engaging data and solution providers, such as interactive platform availability, API availability, real estate product type coverage, regional coverage, data volume and frequency, and underlying data collection and verification methods.

Importantly, building data science knowledge about one's own organization should not be abstract, but help developers better plan and engage relevant data sources in a more formal

way for future projects. Namely, this exercise in data science is about not reinventing the sourcing and collection of data each time a new development is proposed, only to store and discard this data and information on an Excel spreadsheet.

Real estate development data scientists can use results from this research as a starting point for their own data science projects, target future areas of study for under-developed outcomes and features, as well as collaborate more with real estate developers in setting up specific areas of data science modeling.

The breakdown of outcome and features of every task in the real estate development process is a foundation from which data scientists can choose to perform analytic and machine learning modeling. Areas like physical and design analysis with more related tasks, but not enough data need more attention from data scientists. Half of the development tasks have categorical outcomes and thus calls for a stronger collaboration between data scientists and real estate developers in defining appropriate measurements together. To accelerate the application of data science in real estate development, data scientists should apply these understandings in advancing further studies.

These recommendations highlight areas of opportunity and suggestions for real estate developers towards data science. The understanding of data and solution providers in relation to real estate development can help various stakeholders in approaching data and extending what they need to create a more efficient real estate development process.



## AUTHORS



Sunnie (Sun Jung) Park is a Technology Analyst for the MIT Real Estate Innovation Lab. She is a MS real estate development alumni at MIT.



Dr. Andrea Chegut is the Director of the MIT Real Estate Innovation Lab. She holds a PhD in financial economics and studies how technology, design, and innovation impact the economic outcomes of the built environment.



Erin Glennon is the Lab Manager of the MIT Real Estate Innovation Lab. She holds an MFA and works with lab members to develop and publish research.



Financial support for this research was provided by MIT School of Architecture and Planning COVID-19 Relief Fund, the MIT Center for Real Estate and supporters of the MIT Real Estate Innovation Lab.

© MIT Real Estate Innovation Lab 2020

Any use of this material without permission is strictly forbidden. For more information contact us at [reilabcontact@mit.edu](mailto:reilabcontact@mit.edu).