Evaluating Buildings with Computation and Machine Learning

Daniel Davis
WeWork

1

ABSTRACT

Although computers have significantly impacted the way we design buildings, they have yet to meaningfully impact the way we evaluate buildings. In this paper we detail two case studies where computation and machine learning were used to analyze data produced by building inhabitants. We find that a building's 'data exhaust' provides a rich source of information for longitudinally analyzing people's architectural preferences. We argue that computation-driven evaluation could supplement traditional post occupancy evaluations.

1 WeWork's Wonderbread office in Washington, DC is one of seventy offices studied as part of this paper.

INTRODUCTION

Although academia and practice have found many novel ways to improve the design process with computation, almost none of this work has filtered through to impact how designers evaluate buildings. Searching for 'post occupancy' in CuminCAD, an index of over 12,000 papers from ACADIA, CAADRIA, eCAADe, SIGraDi, ASCAAD, and CAAD Futures, returns 22 results, only 9 of which were published in the past 10 years. While computers are used in some aspects of modern building evaluation, such as emailing surveys and taking temperature readings, the use of computation in building evaluations hasn't had nearly the same impact, or warranted the same degree of attention, as the use of computation in the design process.

In some ways, the focus on 'computation for design' reflects a larger preference in the industry for doing design instead of evaluating design. In a survey of 29 mid-sized American architecture firms in 2015, Julie Hiromoto found that "post occupancy evaluation is currently rare" because of the "design team time and cost required to produce meaningful results" (Hiromoto 2015). And when firms find the budget, it generally only allows for a short-term study, often at the start of the building's life, making long-term longitudinal studies of building performance exceptionally rare. Bordass, Leaman, and Ruyssevelt (1999), who have spent their careers evaluating buildings, conclude that "the sad fact is that hardly any architectural or engineering design practices consistently collect information on whether or not their buildings work."

Although designers seldom evaluate the performance of their designs, they are under increasing pressure to create designs that perform. In particular, there seems to be increasing pressure from clients, governments, and society to attain environmental, sociological, and economic performance targets. But since architects rarely study their completed projects, they don't necessarily understand the impact of design decisions on previous projects, putting them in a difficult position to forecast the impact of future decisions. Frank Duffy says that "because our heuristic seems to be 'Never look back,' we are unable to predict the longer term consequences in use of what we design" (Duffy 2008).

In this paper, we explore new potentials for applying computation to the evaluation of buildings. Using two longitudinal case studies based on real data collected from seventy buildings, we consider whether metadata from a room booking application can be used to evaluate meeting rooms and whether machine learning can be used to identify building maintenance patterns. Throughout this paper, we will make the case that people's behavior online often carries with it latent information about how they perceive their physical environment, and that this data can be a valuable tool in the long-term analysis of physical environments.

EXISTING METHODS OF BUILDING EVALUATION

The process of building evaluation can vary dramatically. In some instances it involves a quick survey sent to the client, while in other instances it comprises = an in-depth study involving extensive data collection. Although there is variety in the scope of an evaluation, for the most part all evaluations rely on a few established research methods. In the book, *Learning from Our Buildings: The State of Practice Summary of Post Occupancy Evaluation*, the Federal Facilities Council writes that "traditionally post occupancy evaluations are conducted using questionnaires, interviews, site-visits, and observations of building users" (2001).

These methods of building evaluation have been in development since the 1960s (Preiser 2002), and as such, they tend to focus on aspects of the building that were readily available five decades ago. This is evidenced in Hiromoto's (2015) survey of 29 architecture firms, which found that building evaluations tended to rely on a few well established techniques:

- Quantitative measurements primarily of daylight, acoustics, and the thermal environment.
- Observational studies of density, utilization, efficiency, and differences between plans and occupation.
- Data collected from facility managers on energy and water usage.
- User surveys focused primarily on happiness, energy levels, perceived health benefits, and personal perceptions of the space.

In recent years there has been significant commercial interest in the possibility of adding sensors to the built environment in order to automate the process of taking quantitative measurements. Currently there are many vendors selling varieties of sensors that enable building owners to monitor everything from electricity usage to air quality, occupation rates, and light levels. While these systems are often an effective means of collecting high-fidelity, longitudinal data, the cost of installing and maintaining a sensor network means that this type of analysis is still the exception rather than the norm. Furthermore, the sensors are designed to capture only the most readily measurable aspects of the built environment, which means that more subjective interpretations of the space are not captured. For example, sensors can tell a researcher about a room's mean temperature and CO2 levels, but this data only gives the researcher a rudimentary understanding of whether the room is successful and gives almost no insight into whether the user likes it.

2 Two meeting rooms at WeWork's City Hall location in New York.

In this paper we explore ways of longitudinally collecting data about people's spatial preferences. In other fields, such as advertising, traditional methods of capturing a person's opinion, such as surveys and interviews, have been supplemented with modern methods of inferring people's preferences through their online actions (Neef 2014). Google, for instance, has developed sophisticated machine-learning techniques to predict which advertisements people will prefer, not by surveying them about their preferences, but by observing which websites they visit (McMahan et al. 2013).

There is reason to suspect that a similar analysis could help evaluate buildings by using people's online behavior as a means to understand their preferences in the physical environment. To date, there seem to be no previous studies in this specific area, with prior research tending to focus on traditional post-occupancy evaluation methods coupled with electronic sensors (such as the research from Berkley's Centre for the Built Environment, Stanford University's Space-Mate program, and from Humanyze). One reason for the lack of previous investigation may be that only a few companies currently have access to the quantity and quality of data used in this paper (which captures how tens of thousands of people interacted with over seventy buildings during a span of three years), although the growing connections between the digital and physical environment means that many more organisations should have access to this type of data in the near future.

IDENTIFYING THE BEST MEETING ROOM

WeWork is a company that provides shared workspace, community, and services for entrepreneurs, freelancers, startups, and small businesses. As of April 2016, WeWork has over seventy locations in six countries. Each building has a similar mix of room types—offices, meeting rooms, lounges, phone booths, and other common WeWork amenities—which are customized for the local market.

One critical component of every WeWork location is the meeting rooms (Figure 2). In total, WeWork manages over a thousand meeting rooms spread across seventy locations. Functionally, each meeting room is similar, providing a space for teams to gather and discuss work with the aid of a whiteboard and projector. Although the rooms are functionally similar, they are each customized for their location, differing from other rooms in terms of capacity, furniture, visual appearance, and location within the building.

In 2015, the WeWork Research team was asked to identify whether any of the room variations were more successful than others. In other words, do people prefer meeting rooms with windows? Do they favor large rooms? Do they like the room in Boston that has swings instead of chairs?

Method

In order to understand the components of a successful meeting room, we first set out to identify which meeting rooms the

WeWork members most preferred. Given the scale at which WeWork operates, it was not practical to physically visit the hundreds of meeting rooms and conduct interviews with people spread across seventy buildings. To identify the best meeting rooms we instead examined the data that members of WeWork produce when they book a meeting room.

Members of WeWork are able to reserve meeting rooms for periods of time using either the WeWork website or the WeWork app for iOS and Android. Each reservation is stored in a central database, which, as of April 2016, has over a million reservations going back three years. For this study we were not interested in historic trends and therefore only examined reservations from a three month period (13 weeks) between June 6, 2015 and October 4, 2015—the most recent data at the time of the study. In total we analysed 158,000 meetings from 728 rooms in 44 buildings (WeWork has opened a number of locations since this study was conducted). Based on this data we were able to extract two pieces of information that we assumed may be indicators of which rooms people preferred:

1. Room utilization, which is the number of hours a particular meeting room was booked. We assumed the best rooms would be more popular and therefore utilized more often. To measure the utilization we calculated the percentage of time the room was reserved during peak hours (11 am to 4 pm on weekdays). We excluded meetings that occurred off-peak (before 11 am, after 4 pm, on the weekends, and on public holidays) since these tend to be more sporadic and inconsistent.

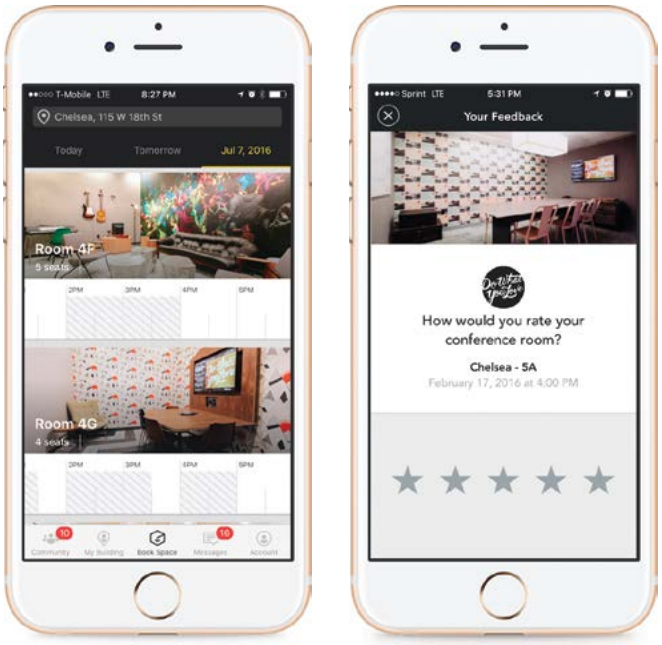
2. Lead times, which is the time between when a booking was made and when the meeting occurred. If a person books a room well in advance (a long lead time) they can generally pick their favourite room since most of the rooms will be available. But if a person books a room at the last minute, many of the best rooms will be taken and they'll have to make do with what's left. We therefore assumed that the best rooms would have long lead times. We calculated a room's lead time as the median of all lead times for meetings that occurred during peak hours in that room.

In addition to the lead time and room utilization rates, we asked members to rate meeting rooms. Members were recruited through the WeWork website and app. When they opened the website or app after a scheduled meeting, they were presented with popup that asked them "how would you rate your conference room?" with the option to rate the room on a scale of 1–5 (Figure 3). If a member rated the room 3 or less, they were given the option to leave additional comments. In total, we collected 23,140 ratings between June 6, 2015 and October 4, 2015.

Results

We began our analysis by looking at the individual buildings and seeing if there were any general patterns between buildings in terms of utilization, lead times, and ratings.

There is a strong correlation (r=0.83) between the median lead time of meeting rooms in a building and the median utilization of rooms in the building. This means that rooms in heavily utilized buildings generally need to be booked further in advance, which is to be expected since heavily utilized rooms are almost constantly booked and require advance booking to find available time.



3 The interface for booking a room (left) and for rating a room (right).

There is, however, a negative correlation between a building's median utilization and a building's median ratings (r=-0.4). This is an interesting finding because it demonstrates that ratings, lead times, and utilizations rate are not analogous and not measuring the same thing. One reason for this negative correlation may be that members tend to always give the room a rating of four or five unless there is an immediate problem with the room (such as the whiteboard markers running out). In this case, buildings with heavy utilization may generally have a lower median rating simply because there are more people using the rooms and therefore more opportunities for short-term maintenance issues to arise.

We also examined each of the 728 meeting rooms in relation to one another. Given the differences between buildings, we normalized the lead times, utilization, and ratings for each building, which let us identify whether a room was utilized more or less than the median room in a particular building.

There was a correlation (r=0.47) between a meeting room's lead times and its utilization, but there was no statistically significant relationship between the room's rating and either its utilization or its lead time. In other words, members did not appear to use highly rated rooms more than any other room. Based on these results, we believe that, in this instance, utilization and lead times seem to be better measures of people's preferred meeting rooms.

Had we done this study without access to the booking data, and instead relied solely on survey data, we would have come to a different conclusion about which rooms were most successful. But because we had both sets of data, the booking data and the survey data, we were able to triangulate a better understanding of the spaces. There are still imperfections to this measurement technique since there are clearly other factors driving a person's decision to book a room that are not accounted for in this iteration of the research (such as the way rooms are presented in the booking app). Nevertheless, this data, latent in WeWork's reservation system, proves to be a quick way to assess hundreds of meeting rooms without installing sensors, physically observing the rooms, or conducting interviews. Based on this research, WeWork was able to identify low performing rooms to upgrade, and update the WeWork design standards based on analysis of which room size was most preferred.

USING MACHINE LEARNING TO UNCOVER COMMON BUILDING ISSUES

When a member at WeWork has a problem, they can inform WeWork of the issue by speaking to a WeWork employee either in-person, through email, or via the WeWork app and website. As soon as the issue is raised, WeWork tracks the issue through to its resolution in a database. In the database, each issue is referred to as a 'ticket,' with each ticket containing data about the time the ticket was created, the subject of the ticket (taken directly from the text in the email or app), and other metadata related to the issue. As of April 2016, this database contained just over 180,000 resolved and open tickets.

The tickets in the database vary widely—there are requests for new keys, members notifying WeWork that their company has outgrown its office, people who can't connect to the wifi, and others who would like a different type of milk for their coffee. A subset of the tickets relate to spatial design—people informing us that they are too cold, that their door is broken, or that it is too noisy in their office.

In early 2016, the WeWork Research team was asked to look back through the tickets to identify whether there were any overall trends in WeWork's spatial design. Were certain HVAC systems leading to less complaints? Were particular buildings more successful than others? Were there trends in maintenance issues?

Method

To find trends in the tickets, we first had to identify the subject of each ticket. We first categorized a thousand tickets by hand to identify the most common topics. We found five recurring themes: HVAC, noise, lighting, maintenance, and tickets not related to spatial design. Given these topics, we set about categorizing the remaining 180,000 tickets into one of these groups.

We initially tried to categorize the tickets using keywords. The algorithm we developed would find all the tickets that mentioned a particular keyword in the text related to the ticket. For instance, a ticket with the subject "I feel too hot," would be assisted the 'HVAC' category because it contained the keyword 'hot,' whereas a ticket containing the word 'loud' would be assigned the category 'noise.'

We tested the keyword method using 2000 randomly selected tickets, comparing the machine-generated classification to the classification assigned manually by the researchers. Although the keyword method was fairly crude, it was surprisingly accurate, particularly for categories that have keywords distinct from the other categories (HVAC, for instance, has words like freezing, warmer, and air conditioning, which tended to only appear in that category).

For HVAC, the algorithm had a precision of 82% and a recall of 92%, meaning that most HVAC tickets were correctly identified by the algorithm with few false positives. For categories that do not have distinct keywords, such as maintenance, which encompasses a range of issues that share few common words, the keyword method fared much worse. For maintenance it has a precision of 46% and a recall of 27%, meaning that only a quarter of the maintenance tickets were correctly identified and only half of the tickets flagged as maintenance actually were about maintenance.

In an attempt to more accurately classify the tickets, we tested the effectiveness of applying naive bayes classification to the tickets. Naive bayes is a classifier based on Bayes' Theorem that calculates the probability that an object belongs to a class given the features of the object and the frequency with which those features occur on other objects of a particular class. In this case, we wanted to classify the ticket based on the words in the ticket. For textual classification, naive bayes is a popular method often used "to define a baseline accuracy" (Kelleher, Namee, and D'Arcy 2015). We used the 'bag of words' method, whereby each word in the subject and description of a ticket became a feature

of the ticket (Zhang and Zhou 2010). We trained the classifier on 3000 hand-categorized tickets, which allowed the classifier to learn the probability that tickets classified as being noise-related also contained the word 'sound' somewhere in the bag of words from the subject and description. Another 2,000 tickets were used as a test set, with each ticket classified by both the classifier and a human so we could understand the accuracy of the classifier.

The naive bayes classifier was generally better than the keyword classifier for every ticket category except those tickets related to noise (where it had a recall of 43% and a precision of 31%). The noise category seemed to be difficult for the naive bayes classifier because many of the noise-related tickets were related to other categories. For example, a ticket that states "my air conditioner is making a loud noise," could be categorized as an HVAC issue, but we categorized it as relating to noise since it is primarily a noise complaint. Noise was the only category the naive bayes classifier struggled with, and in other categories, like maintenance, it performed significantly better than the keyword classified, achieving a recall of 96% and a precision of 70%.

Outcomes

The naive bayes classifier was applied to the 180,000 tickets in the database. Although the classifier isn't perfectly accurate, at a macro level the classification is accurate enough to allow WeWork to identify which aspects of the built environment cause the most problems for WeWork members. This has enabled us to prioritise our research so that it will have the most impact on member experience. We have also setup a dashboard that allows WeWork employees to compare how each building is performing, letting them identify macro issues as they emerge in real-time and get fixes out to them quickly.

There are some limitations to this method. Most notably, members tend to only create tickets for things they know can be fixed, such as the temperature, and they might find it hard write a ticket that articulates more subtle and less actionable feelings about their environment. The data is also biased towards negative experiences of the built environment; it tells us what frustrates members, but it doesn't tell us what delights them since people don't create tickets when everything is going well. The data therefore isn't a complete record of how people feel about their environment. That said, the data does provide a foundation for continuously monitoring how people feel about the built environment without needing to constantly send surveys gauging people's preferences.

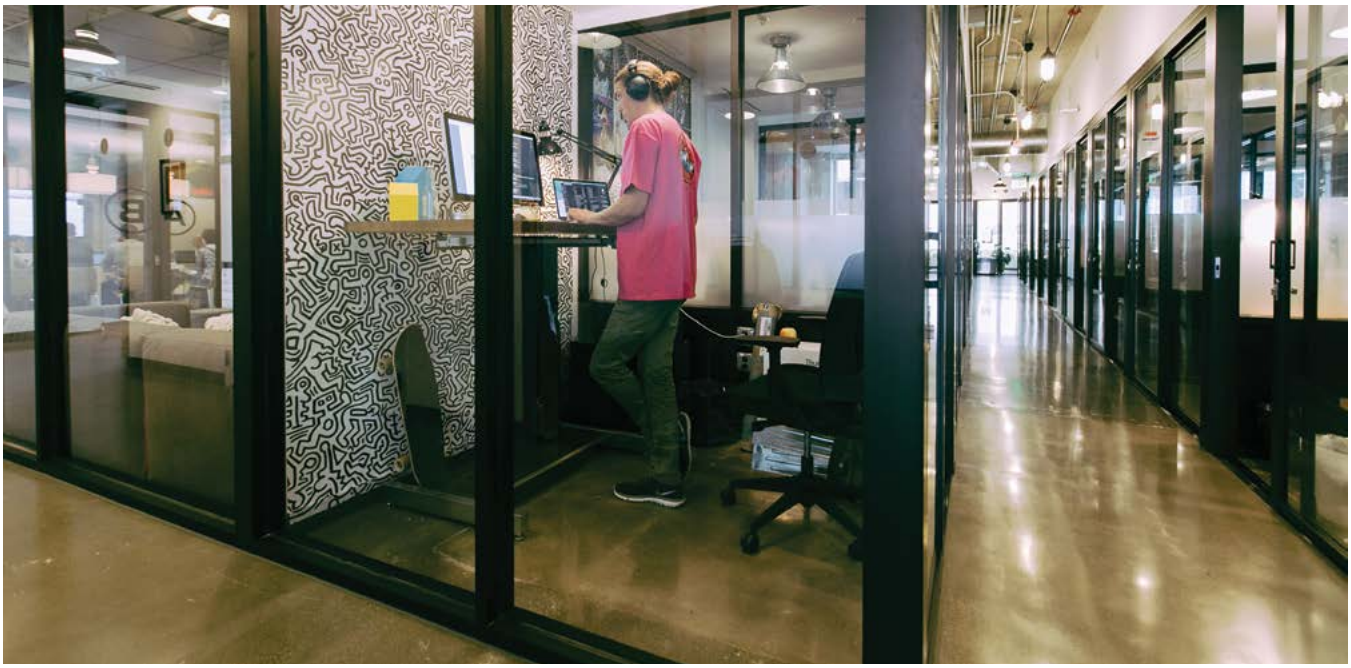
DISCUSSION: WHAT THE 'DATA EXHAUST' TELLS US ABOUT BUILDINGS

The term 'data exhaust' has appeared recently as a way of describing the data that a person leaves behind in the digital world as they go about their lives. This includes everything from data stored in server logs when a person visits a website to the data recorded for financial purposes about a credit card transaction. These pieces of information often tell detailed stories about the actions of individuals, and many companies and researchers have been using this data to understand people's preferences for everything from search results to advertisements (Neef 2014).

The two case studies in this paper, one analyzing meeting rooms through booking data, and the other evaluating buildings using a ticketing system, begin to articulate how this data exhaust may be used in the evaluation of architecture. The original contribution is not the methods of analysis used, but rather their application to the built environment. They show how data already produced by the people inhabiting buildings, data that is a latent product of the built environment, can be used to analyse people's preferences for particular aspects of the built environment. It shows how traditional means of post-occupancy evaluation, such as surveys, interviews, and site visits, as well as modern methods of electronic sensing, may soon be complemented by this new way of understanding architecture and the people that inhabit it.

In many ways, this type of analysis seems long overdue. Architects have long experimented with computation in the early stages of design, findings ways to create and manufacture buildings using algorithms—some going as far as to claim that computation has given rise to a new style of architecture (Schumacher 2009). But while the process of designing buildings has evolved to embrace computation, the process of evaluating buildings has remained largely unaffected by the advent of computation and the rise of the data exhaust.

In part, research in this area may be limited because of the challenges associated with creating a data set. Most machine learning techniques require fairly large data sets. In this paper, the meeting room analysis looked at data from 158,000 meetings and the ticket analysis looked at 180,000 tickets. Collecting this data required that data collection systems were implemented across multiple buildings for many years prior to the analysis. A lot of architects and researchers may not have access to this scale of data. Even if they do have access to the prerequisite buildings and data, a major limitation of this type of analysis is that it requires a building to be settled, operational, and producing data before it can be evaluated.



4 Seattle - WeWork South Lake Union in Seattle.

If the right data can be gathered, the potential payoff is that aspects of the building can be analyzed without requiring that the building's inhabitants complete a survey or interview. In certain instances it may even provide a more complete picture. For instance, the tickets related to HVAC issues in the second case study give insight into how people felt about their thermal environment over a number of years. While traditional survey methods give insight into how people feel at a particular moment in time, these new methods lend themselves to longitudinal studies that allow researchers to passively collect data about people's perceptions over a period of years. Analysis of this longitudinal data exhaust potentially opens up new avenues of research, and new ways of understanding buildings that take into account the evolution of the project over time.

CONCLUSION

Most buildings are constructed without any form of post occupancy evaluation. In part, this may be due to the typical methods of post occupancy evaluation, which have not benefited from advances in computation to the same extent as other parts of the design process.

In this paper, we have presented two ways that computation and machine learning can be applied to datasets generated by building inhabitants in order to extract information about how people perceive their built environment. Both of these methods have proved successful in practice and have demonstrated tangible impacts on the way similar buildings are designed and operated. This research shows, in two instances, that data generated by people through the course of their day often carries with it latent information about how they perceive their physical environment, which can be a valuable tool in the long-term analysis of physical environments.

REFERENCES

Bordass, Bill, Adrian Leaman, and Paul Ruyssevelt. 1999. *PROBE STRATEGIC REVIEW 1999 FINAL REPORT 4: Strategic Conclusions*. London: Department of the Environment, Transport and the Regions. http://www.usablebuildings.co.uk/Probe/ProbePDFs/SR4.pdf

Duffy, Frank. 2008. *Work and the City*. London: Black Dog Publishing.

Federal Facilities Council. 2001. *Learning From Our Buildings: A State-of-the-Practice Summary of Post-Occupancy Evaluation*. Washington, D.C.: National Academy Press.

Hiromoto, Julie. 2015. *Architect & Design Sustainable Design Leaders Post Occupancy Evaluation Survey Report*. SOM: New York. http://www.som.com/FILE/22966/post-occupancy-evaluation_survey-report_update_2.pdf

Kelleher, John D., Brian Mac Namee, and Aoife D'Arcy. 2015. *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies*. Cambridge, MA: MIT Press.

McMahan, H. Brendan, Gary Holt, D. Sculley, Michael Young, Dietmar Ebner, Julian Grady, Lan Nie, Todd Phillips, Eugene Davydov, Daniel Golovin, Sharat Chikkerur, Dan Liu, Martin Wattenberg, Arnar Mar

Hrafnkelsson, Tom Boulos, and Jeremy Kubica. 2013. "Ad Click Prediction: A View From the Trenches." In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, edited by Inderjit S. Dhillon, Yehuda Koren, Rayid Ghani, Ted E. Senator, Paul Bradley, Rajesh Parekh, Jingrui He, Robert L. Grossman, and Ramasamy Uthurusamy. New York: ACM. 1222–1230. http://doi.org/10.1145/2487575.2488200

Neef, Dale. 2014. *Digital Exhaust: What Everyone Should Know About Big Data, Digitization and Digitally Driven Innovation*. Upper Saddle River, NJ.: Pearson FT Press.

Preiser, Wolfgang F. E., and Ulrich Schramm. 2002. "Intelligent Office Building Performance Evaluation." *Facilities* 20 (7): 279–287. doi:10.1108/02632770210435198.

Schumacher, Patrik. 2009. "Parametricism: A New Global Style for Architecture and Urban Design." *Architectural Design* 79 (4): 14–23.

Zhang, Yin, Rong Jin, and Zhi-Hua Zhou. 2010. "Understanding Bag-of-Words Model: A Statistical Framework." *International Journal of Machine Learning and Cybernetics* 1 (1): 43–52. http://doi.org/10.1007/s13042-010-0001-0

IMAGE CREDITS

Figures 1–4: © WeWork

Daniel Davis – Director of Spaces and Cities Research at WeWork. Based out of WeWork's headquarters in New York, Daniel leads a team of researchers investigating how to design workplaces so that people feel happier, more productive, and more connected to their community. He originally trained as an architect in New Zealand and later did a PhD in computational design at RMIT University's Spatial Information Architecture Laboratory in Australia.