Human-Centered Data Science: A New Paradigm for Industrial IoT

MATTHEW YAPCHAIAN
Uptake

Few professions appear more at odds, at least on the surface, than ethnography and data science. The first deals in qualitative "truths," gleaned by human researchers, based on careful, deep observation of only a small number of human subjects, typically. The latter deals in quantitative "truths," mined through computer-executed algorithms, based on vast swaths of anonymous data points. To the ethnographer, "truth" involves an understanding of how and why things are truly the way they are. To the data scientist, "truth" is more about designing algorithms that make guesses that are empirically correct a good portion of the time. Data science driven products, like those that Uptake builds, are most powerful and functional when they leverage the core strengths of both data science and ethnographic insights: what we call Human-Centered Data Science. I will argue that data science, including the collection and manipulation of data, is a practice that is in many ways as human-centered and subjective in nature as ethnographic-based practices. I will explore the role of data, along with its generation, collection, and manipulation by data science and ethnographic practices embedded within organizations developing Industrial IOT software products (i.e. Department of Defense, rail, wind, manufacturing, mining, etc.).

FIGURE DRAWING: OBSERVE AND REFINE

Relating ethnography and data science practices begins in the studio of an artist with a metaphor; figure drawing. An exercise that requires a clothed or nude person (most common), figure drawing is a traditional fine arts practice used by an artist throughout his/her career to continually develop foundational drawing skills. More importantly, figure drawing is an exercise about *refinement of observation*.

An artist drawing a model only glances at their sheet to mark an observation during their session; most of the artist's time is spent observing the model. Drawing is a process of refinement: observe, refine/mark; repeat. The goal is not to render the image in a single pass; rather it is to develop the image over a defined period of time. An artist's gaze fixed too long on the drawing as he/she draws, reveals a practice that results in an image describing what the artist *imagines* the subject to be rather than how he/she exists.

Evidence of a drawing's development (i.e. observations) can be found within the drawn artifact. Each of the figures in Figure 1. and Figure 2. describes a unique pose by a single model that was held over a specific duration of time (i.e. "5 minute pose"). Figure 1's knee is formed over the course of at least 3-4 observations; progression is most visible with this figure in by the number of lines required by the artist to define the knee. Early marks are visible that were refined over time. Development of a drawing is more visible in Figure 2. With the number of marks defining each part of the human form layered over each other.

Observe, refine/mark; repeat. In the context of building and iterating through repeated observations, an artist's practice is similar to that of an ethnographer and data scientist. Not all marks are correct, but they create a set of knowledge that is directional, leading to an accurate solution. All three practices begin with observations and broad marks that are iterated on over time, resulting in a meaningful artifact that is revealed slowly. [1] Images by Tony Cheng

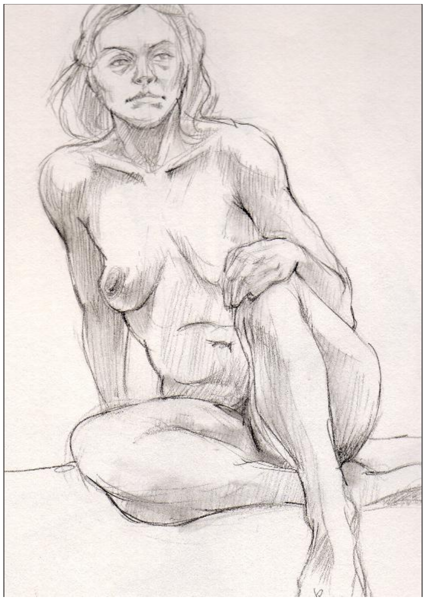

Figure 1. Figure Drawing

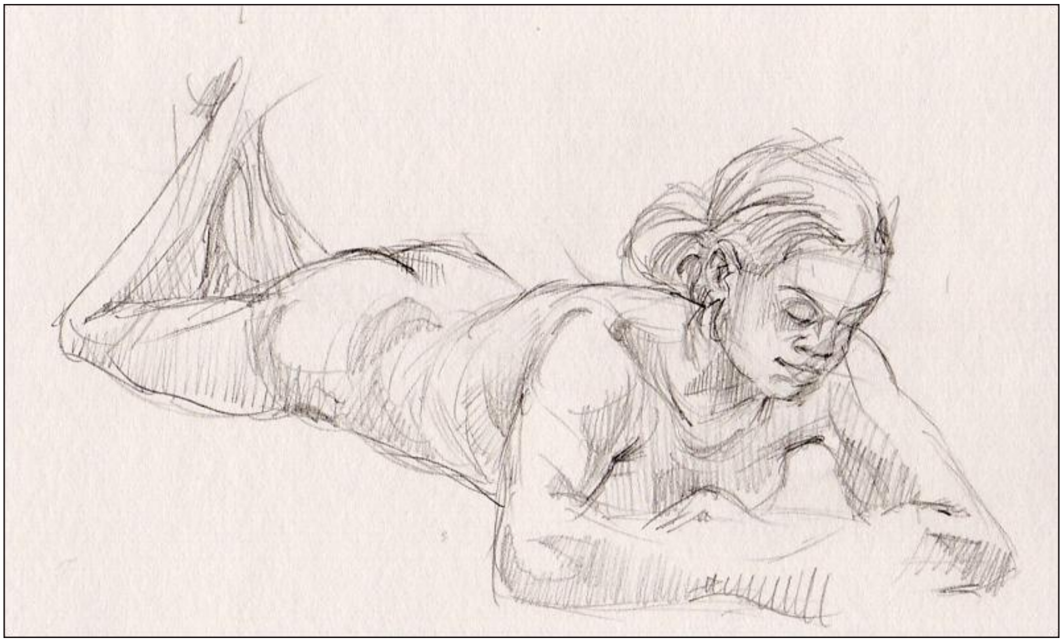

Figure 2. Figure Drawing (ran out of time)

INTRODUCTION

The author of this paper is user experience researcher at Uptake, a company specializing in artificial intelligence in the space of Industrial IoT. Within Uptake, software development leverages core resources throughout an engagement, including data science and user experience research. Due to this ongoing working relationship with the data science team, Uptake's user experience researchers have gained a unique understanding of data scientists and their everyday practices.

This view provides Uptake's UX research team an empathetic understanding of and appreciation for the manual, tedious and often-invisible process occupying most of a data scientist's time. Their effort begins with a problem defined by a client, one self-identified or surfaced through field research. Then an ongoing cycle of observing and refining their subject by identifying a source(s) for the required machine data; generating and logging that data, engaging with it, confirming that it is the correct data, and ultimately crafting purposeful models to yield actionable insights.

Data—quantitative and qualitative, supporting data science or UX—in its raw forms is an observation, not evidence. A machine's gears are grinding, the temperature reading is 400° F, a bus will not start, an asset's fluids levels are x, a component is on/off, etc. A single observation is just as a mark on a sheet a paper that contributes to the rendering of a figure.

Actionable insights emerge from this data when rich contextual and historical data is introduced. For example, an alert for a rail engine's loss of horsepower will be triggered when a certain threshold is met. Observation.

This alert is observed in the data during each run at the same location along a specific route. Does the engine need to be serviced? No. Evidence is forged from quantitative and qualitative observations. Upon further inspection, the train passes through a tunnel that has limited circulation, forcing the engine's intake center to be temporarily overwhelmed by the exhaust system, triggering an alert for loss of horsepower. Horsepower returns to normal levels immediately after exiting the tunnel.

Raw sensor readings coupled with qualitative context becomes evidence supporting—in the example above—a decision not to service the engine and ignore all alerts of this type generated at this location. Action may be taken if the same fault occurs at a different location along the same route.

Observe, refine/mark; repeat.

Industrial IoT's soul is large data sets and real world context that require technical excellence and human curiosity to generate meaning from it through an ongoing cycle of observing, refining and marking.

UPTAKE DATA SCIENTISTS AND USER EXPERIENCE RESEARCHERS

A typical figure drawing session is organized as a ring of people, each with a pad of paper and a drawing instrument around a small rise; at the center of this group is a single model posing on the rise. At the conclusion of the pose, each person has rendered an image of the same model from a unique *view*.

Individuals drawing during these sessions are usually fixed to a bench or easel for the duration of the session as a matter of convenience. Models will change poses and their orientation to the group throughout a session, allowing each person to "see" different views of the same model. An alternative approach is to have a model hold a long pose as a group of artists (or single individual) rotate around the model in timed increments to see the same subject from multiple points of view.

A drawing session is an effort to understand a figure through individual poses and views of the model. An artist restricting his or her a view of the model to only front, side or back poses limits their understanding of how the complete form exists in space.

Poses can last for durations of 30 seconds, 1 minute, 5 minutes, etc. Or a pose may be continuous for hours across multiple sessions. No matter the length of time, the initial marks made a on sheet of paper situate the figure within the page's space. These marks do not commit the artist to a final image, that image develops from continuous observation of the model and piecemeal refinement of the drawing.

A project begins at Uptake with each role positioned around the project's challenge. Uptake's data scientists need access to large sets of data—generated by man and machine. Data is not always received in an organized and tidy package ready to be acted on. Data scientists, working with subject matter experts (SME), begin their effort by understanding the data available to them. (Including the asset(s) generating data.) From this initial view, they can determine what additional data and quantity of it is needed. If the required data does not exist, they develop a plan to generate and log it.

Observe, refine/mark; repeat.

As data scientists engage machine data, user experience researchers are understanding the context surrounding the data through preliminary research and SME engagement. This is critical learning ahead of any fieldwork due to the nature of Industrial IoT sites. They are often dangerous spaces with unique obstacles, including: intense security/regulatory requirements (Department of Defense, rail), remote locations that are costly to access (mining), limited number of candidates to speak with (manufacturing), and a required escort for the duration of a visit (all).

Observe, refine/mark; repeat.

Similar to the initial broad strokes of an early figure drawing, these first steps of discovery are to situate the roles before committing to courses of action. Data sources and end users may change over time just as the figure's final image emerges through a series of observations and marks.

INDUSTRIAL IoT 2018

General Overview

Consumer, enterprise and industrial software products have been produced and sold to meet an array of intended and unintended needs for decades. Industrial operations of all types leverage a suite of these solutions to lead their daily operations. Data science driven products leveraging AI and machine learning within Industrial IoT are new tools being developed today to be included in an organization's current suite of tools. These products provide previously unavailable line of sight into operations down to asset components, enabling organizations to develop predictive maintenance practices that lead to increased efficiency, uptime and cost savings.

Ganesh Bell, the president of Uptake, describes an emerging Industrial IoT market by contrasting enterprise and industrial IoT software: "In the last several decades, enterprise software was all about humans entering data and automating business processes. Now we are in a world where machines are generating data; robots and drones are increasingly at work with wearables and humans augmented by machines are generating more data.–We believe we will be in a paradigm where it will be about automating decisions versus automating business processes."

Industrial IoT is a critical space because of the volume of data generated (and potential to be generated) that can be leveraged to drive value across industries from transportation to mining, rail to the Department of Defense. Data science and user experience research are core practices in the development of the tools driving value.

I will refer to large industrial organizations that operate fleets of heavy machine assets (rail engines, construction vehicles, components of an oil & gas plant, wind turbines, transit buses, etc.) as "Industrial IoT." Within each fleet, Industrial IoT's current state of technology is an intentional or unintentional mix of "connectivity states" that describe an asset's ability to generate and collect data.

- **Connected** assets generate, collect and act against copious sums of machine data. Many rail engines are embedded with sensors by an original equipment manufacturer (OEM). The generation, logging and use of data are deliberate choices made by an asset's owner.
- **Enabled** assets are generating data but are not configured to collect that information. Example: A transit bus OEM will embed assemble a bust with equipment that generates data by default but the customer may not be logging it.
- **Unconnected** assets are without a sensor-based infrastructure to generate data. An older machine in a manufacturing environment may require the addition of sensors or an edge device to generate and log data.

Many of Uptake's clients—including wind farms, mining and rail operations—have sites located in remote regions across the globe. A parts technician or a maintenance team needing to service a site can expect a commute of several hours to a day to access many of these places. Further complicating these visits is lack of sight into component status on the sites they service. Without visibility of an asset's health, routine maintenance is not prioritized by immediate need.

Uptake's software enables line of sight into a site's operation and, at a granular level, down to an asset. For organizations (mining, manufacturing, etc.) that are in continuous operation with limited scheduled downtime, knowing what component might fail, when that failure will occur and the probability of those insights is a critical advantage to an organization. Annual preventative maintenance (PM) is a current practice used to service assets, but asset failures do not wait for PM appointments.

Industrial IoT's problems are ideal opportunities for data science and user experience research to collaborate in an effort to produce meaningful products that will help shape how industrial problems are solved.

Industrial IoT products are designed for an audience of end users smaller than those of consumer or enterprise products. Within an organization, Industrial IoT software can enable a glimpse into operations for Executives, empower engineers and managers to make decisions with confidence and speed that affect operations, help supply managers curating stock rooms know exactly what they need and for when, and empower the maintenance workers on the floor turning wrenches, performing everyday firefighting and maintenance to be more efficient by taking on problems when they are smaller in scale.

While these industries are slow to change, automation and robotics are transforming the shape and size of workforces on the floor across different types of industrial operations. Individuals present on the floors of companies Uptake engages with recognize machine data will be a core tool of their future practice. Many people occupying these roles are embracing data's emerging presence at work because it enables them to take on problems previously too complicated for them to solve with available toolsets, such as Excel-based applications. These are often complicated tools authored by a single person.

End users receiving insights desire to learn more about their systems by understanding the data behind an insight. Some insights are packaged as a "stop light" (red, yellow and green to signal priority) with the ability for the user to drill down into the data driving the insight. End users, not just data scientists and researchers are continually refining the image before them through a practice of observation and marking

Moving Forward

Industrial IoT SaaS products driving significant financial and operational outcomes from machine data will ultimately lead to broader adoption of solutions offered by Uptake and similar companies by organizations throughout industries.

Uptake's data science and user experience teams are two critical practices found at the point where an insight and an end user meet. Data science produces a model that yields an actionable insight based on a set of conditions, user research must identify which role needs that information and at what time and where in that role's workflow to introduce it for it to provide value.

Data science and user experience research teams dedicate a majority of time to understanding differences between observation and evidences through practices of ongoing refining of their observations and mark making. The challenges of Industrial IoT software development is an array of poses that both roles are continually approaching from complimentary views. Some poses are brief in duration, while other extend over time—inherent in SaaS software. Always observing and refining.

THE TRUTH, THE WHOLE TRUTH, NOTHING BUT THE TRUTH

In a world of "big data," the designed, built, engineered nature of machines—specifically industrial machines—in concert with the rigorous mathematical, computer-based nature of data science, can easily lead customers and end users to believe that data and insights derived from data are objective in nature. In reality, insights are the result of a deliberate process of observing, refining and bringing together quantitative and qualitative data so that the probability of an insight's accuracy is high.

Consider an "edge device" (sometimes also called an "event recorder" or "data logger") on freight locomotive. The purpose of such a device is to collect readings (vibrations, pressures, speeds, temperatures, etc.) from an array of sensors located throughout the locomotive, and to relay those back to computer servers where they can be analyzed to identify signs of impending breakdowns. Sensor readings themselves are measured by precision instruments, and they are relayed automatically without human intervention or hand-offs. This would seem like the ideal set-up for a fully reliable transmission of "objective" information.

Not quite.

Since storing and transmitting data can be expensive, sensors on the locomotive only take measurements under specific operating conditions, for example, when the locomotive is running in a particular throttle position. In turn, the edge device transmits small samples of the data at 15-minute or longer intervals, relying on a set of programmed rules to determine what data to send over, at what granularity (anything from just one to several hundred readings per second), and whether in raw or aggregated form. If for any reason the edge device cannot communicate with the servers, for example due to absence of a cell signal, it will accumulate unsent data on its small hard drive until this hard drive is full at which point it will simply delete some data according to programmed rules to write new data.

The parameters governing these behaviors were at some point designed, programmed and informed by a collection of human beings whose job it is to ensure a certain level of reliability of the locomotive as a whole, itself determined whether contractually or incidentally, by yet another collection of human beings and corporate entities driven by a combination of personal and financial incentives. Considered through this lens, the collection and transmission of information from a machine does not seem so objective at all.

This process of deciding which observations to collect and how to collect them is the result of observing and refining a subject, these are the lines of a figure drawing over time forming a hip or elbow, placing it in a space.

An Uptake data scientist looking to build an algorithm to predict locomotive breakdowns based on this data will engage in an iterative and creative process quite similar, in practice, to that of an ethnographer interviewing a human subject. The data scientist needs to understand the conditions and constraints under which the edge device was designed and built, and interpret the data she receives from it accordingly. Alone at her computer, this task can prove inscrutable, which is why she will spend time talking with those people whose work informed the data collection mechanism with all its quirks and idiosyncrasies. Armed with an ethnographer's toolbox, she can be exponentially more effective.

Point of Origin

Uptake's data scientist may or may not be aware that she is using tools from the ethnographer's box. By looking at an early stage of an algorithm's design and development, we can observe how the two practices depend on discovery, iteration, trial and error, and direct engagement with people, including subject matter experts (SMEs), end users, stake holders, etc..

An algorithm's origin is initiated through numerous possible actions, including a client's request to solve a specific problem, emerging organically through a data scientist pursuing her own curiosity, an internally driven effort, previous experiments, or a lead from ethnographic field research. From here, next steps are a series of questions about quantity and quality of available data—if it exists at all—that are critical in the overall development of an algorithm. This is an iterative process that can include a large number of people from the internal and client teams.

At Uptake, the data scientist assigned to a project will ask if data exists, and is the data provided to her the correct data? How much data is available (volume and duration over time)? What is the source of the data? How was it collected? What is the data's quality? Is additional data required? How much "cleaning of the data" is required? How accessible is the desired data? (Accessibility can be contingent on many variables, including: an asset's network connectivity, security, physical location on an asset, etc.) Uptake's data scientist describe the time required to prepare data so that it will be in a state to have math performed on it as "80% time."

Observe, refine/mark; repeat

Sometimes, in this process of focusing in on the correct data, the pursuit of one data source will reveal itself to be less meaningful than initially thought of, challenging the data scientist to look at other sources. This process often involves ongoing conversations with the client's SMEs.

If data is unavailable from an asset, there are two questions to ask. (1) Is data currently being generated by an asset and not logged? A bus may have the capabilities to generate data streams with built-in sensors, but a logging device has not been connected to collect the streaming data. Or (2) are the hardware/software capabilities that facilitate generation and logging of data from an asset absent? An bus' OEM may not have installed assets components with sensors. For both questions, determining methods and cost required to obtain the necessary data will influence the algorithm's design.

For the later asset types, those that do not generate data of any type, what are the potential solutions to enable generation and logging of data? Is an Edge device required? If it is, will a custom device need to be designed or can an off-the-shelf solution work? Who will design and manufacture a custom device? (An in-house IoT team, third party, etc.?) How does the device's design affect data collection? Observe, refine/mark; repeat

Building out that challenge more, for components that need to be monitored but do not have any type of direct output channel to stream data, the component must be observed indirectly. For example, a wheel can be monitored through rotation count; from this value, information related to a wheel's health might be inferred. Identifying the correct peripheral signals to collect and analyze in an effort to gain insight on a component requires (often) engagement with a subject matter expert (SME) and trial and error.

Decisions related to data that will be generated, logged, cleaned and represented as objective evidence are iterated over time, manually. This process produces a complementary data set that remains invisible—information that will not be generated and logged; a deselection set.

How much data will be transmitted to the cloud and in what intervals? Each decision point in the process of determining the appropriate data to use and why, shapes an algorithm's design and output. Similar to ethnographic and design approaches, data science's process begins with a discovery phase that transitions to an iterative process, then toggles between the two as a point of view is refined. Through trial and error— observe, refine/mark, repeat—an algorithm (a human designed object) develops a specific agenda built on a pastiche of human decisions. Decisions later obscured in a veneer of objectivity by users of a product when encapsulated in "data from the machine."

Data-driven AI Industrial IoT software applications are tools that enable engineers, maintenance workers and other roles within industrial operations to observe, refine/mark; repeat.

Data Scientist and Ethnographers: A Future View

Data scientists and social science based researchers are inherently a motley crew of curious individuals. While the roles "data scientist" and "researcher" can be shorthand for a generic job, each discipline is composed of professionals from a variety of backgrounds and interests. A person may be a weather geek with exceptional math skills or an artist passionate about the intersection of people and technology. These backgrounds and curiosities are fundamental attributes driving these practices, rendering them valuable to an organization.

When organizations recognize the complimentary nature of these two practices and foster synergy between them, more robust products will emerge.

Similar to a studio of artists positioned around a figure model, each person practicing a similar practice of learning and building as they render an accurate but different image of a pose; data scientists and user experience researchers have opportunities to be in the field and office together, experiencing the same client, the same site, and problem from very different (though complimentary) perspectives. Through this mode of collaboration, creativity and understanding flourishes, ultimately contributing to more robust and meaningful practices and products.

Observe, refine/mark; repeat.

NOTES

Thank you Chiara Piccinotti and James Lamb for meaningful conversations and ideas.

[1] Figure drawings by Tony Cheng. Images have not been altered other than scale and were found on Flickr. Figure 1.: https://flic.kr/p/G7rUj , Figure 2.: https://flic.kr/p/FhLis . Both images are under Creative Common license with some rights reserved: https://creativecommons.org/licenses/by-nc/2.0/