Discovering the Language of Data:
Personal Pattern Languages and the
Social Construction of Meaning from
Big Data

Kim Erwin, Maggee Bond and Aashika Jain
IIT Institute of Design, Chicago, USA

This paper attempts to address two issues relevant to the sense-making of Big Data. First, it presents a case study for how a large dataset can be transformed into both a visual language and, in effect, a ‘text’ that can be read and interpreted by human beings. The case study comes from direct observation of graduate students at the IIT Institute of Design who investigated task-switching behaviours, as documented by productivity software on a single user’s laptop and a smart phone. Through a series of experiments with the resulting dataset, the team effects a transformation of that data into a catalogue of visual primitives — a kind of iconic alphabet — that allow others to ‘read’ the data as a corpus and, more provocatively, suggest the formation of a personal pattern language. Second, this paper offers a model for human-technical collaboration in the sense-making of data, as demonstrated by this and other teams in the class. Current sense-making models tend to be data- and technology-centric, and increasingly presume data visualization as a primary point of entry of humans into Big Data systems. This alternative model proposes that meaningful interpretation of data emerges from a more elaborate interplay between algorithms, data and human beings.

KEYWORDS personal pattern language, sense-making processes, sensor-based data, data visualization, big data, qualitative data, analytic methods

Introduction
Many academic, scientific, and business interests have come together under the tent of Big Data. As computing technology and the rapid digitization of vast amounts of both historical and real-time content come together,
researchers from markedly different disciplines seek to make sense of it. Big Data began as an objective description of a technical problem: how to analyse a volume of data that can’t be held in memory or on one computer (Wikipedia). Today the term is used as shorthand for the potential power of large datasets: ‘How Big Data can transform society for the better’ (Pentland 2013); ‘The next frontier for innovation, competition, and productivity’ (Manyika et al. 2011), ‘Changing the way humans solve problems’ (VentureBeat, 2014) ‘Creating Revolutionary Breakthroughs in Commerce, Science and Society’ (Bryant et al. 2008).

To date, attempts to harness Big Data have largely been technology-led: data scientists and engineers labor to develop statistical models, algorithms, and visualizations that can cluster and correlate complex data so as to reveal patterns in diverse datasets. Developers have created important enabling technologies, such as cloud computing infrastructures (enabling large datasets to be hosted and accessed remotely) and software frameworks like Hadoop (designed to support distributed applications and analysis of complex data as if it were a unified repository).

However, there is little evidence in the literature that the human side of this revolution is being as rapidly or systemically addressed. While researchers advocate for the inclusion of human beings as part of the sense-making model in developing Big Data technologies, there remains no clear model in the literature for how humans and technology will meaningfully partner to derive meaning from Big Data.

**Conceptualizing human inclusion in the sense-making system**

We are in the early days of the Big Data revolution, and so the processes and conceptions of how Big Data might to be productively employed are still emerging. A quick tour of industry discussions and academic literature reveals a number of conceptions at work.

One provisional approach at work in industry is the use of a designated individual, armed with a laptop and gifted with numbers sense, as a big data solution. Observed by Erwin at GigaOM’s StructureData 2012 conference in New York City, for example, moderator Malik Om asked a panel of senior executives how they were engaging big data in their organizations. John Sotham, the VP of Finance for BuildDirect, stated that he had a designated ‘guy’ with a laptop who came to every meeting with him and to whom he addressed all Big Data questions (the other two panellists admitted they didn’t yet have what they would call a Big Data strategy). A similar model is also in evidence in the movie Moneyball, based on the true story of Paul Podesta, who transformed the Oakland A’s baseball team from laggard to winner by developing a formula that more successfully predicted player performance than traditional player statistics. Underlying both of these genius-with-a-laptop conceptions at work in industry is a linear service model — one in which executives pose questions to designated others, who return answers after applying carefully chosen calculations and custom algorithms (Figure 1).
A more systematic model is emerging in research and technology fields, where many are pursuing sense-making systems that actively call for human beings to be considered part of a more distributed technical model. Various leaders from relevant fields, such as the data sciences, visual analytics and visual mining, have underscored the critical role of human judgment in sense-making. More specifically, they are coalescing around a belief that what’s needed are tools that enable human beings to bring their considerable perceptual powers to complex data. For example, Thomas and Cook (2006), writing as part of a US government-convened panel tasked with establishing an R&D agenda for using visual analytics in counterterrorism efforts, framed the technical agenda for the research community thusly: ‘Visual analytics strives to facilitate analytical reasoning by creating software that maximizes human capacity to perceive, understand, and reason about complex and dynamic data and situations. It must build upon an understanding of the reasoning process, as well as an understanding of underlying cognitive and perceptual principles, to provide task-appropriate interactions that let users have a true discourse with their information’. In their call to action, Thomas and Cook also note that ‘the research community hasn’t adequately addressed the integration [required] to advance the analyst’s ability to apply their expert human judgment to complex data’.

Respected figures in the visual analytics and HCI communities have also made explicit statements that technical solutions to sense-making must orient to human beings as pattern detectors and arbiters of meaning: ‘The main idea is that software tools prepare and visualize the data so that the human analyst can detect various types of patterns’ (Andrienko and Andrienko 2007). Heer and Shneiderman (2012), both pioneers in HCI, assert that ‘[digital data] analysis requires contextualized human judgments regarding the domain-specific significance of clusters, trends, and outliers discovered in data’ and further state that ‘to be most effective, visual analytics tools must support the fluent and flexible use of visualizations at rates resonant with the pace of human thought’.

What emerges as common across these conceptual models is that the role of technology is to transform and represent the data (Thomas and Cook 2006) and the role of human beings is to appraise that data for patterns and clues, typically in iterative fashion (Heer and Shneiderman, 2012) (see Figure 2). In other words, the distinctive contribution of human beings to the sense-making system is that they can see. As statistician and writer Nate Silver points out in his well-received book, The Signal and the Noise: The Art and Science of Prediction: ‘What is it, exactly, that humans can do better than
computers that can crunch numbers at seventy-seven teraflops? They can see... Distort a series of letters just slightly — as with the CAPTCHA technology that is often used in spam or password protection — and very “smart” computers get very confused. They are too literal-minded, unable to recognize the pattern once it’s subjected to even the slightest degree of manipulation. Humans by contrast, out of evolutionary necessity, have very powerful visual cortices. They rapidly parse distortions in the data to identify abstract qualities like pattern and organization’ (2012, 124).

And here we arrive at a key point: sense-making systems are increasingly conceptualized as a set of computational procedures (algorithms) that convert and then depict data, performing a closed act of intelligence that is open to human beings primarily at the point of visualization. Not surprisingly, effective data visualization has become a white-hot topic in the development of Big Data technologies, and has generated substantive research and investment into related fields, including the science of visual perception for data design (Ware 2004; Ware 2008), principles of visual reasoning and data design (Bertin 1983; Tufte 1990; Tufte 2001; Few 2009; and many others), and the development of libraries of data visualization with algorithms that can draw them (Bostock et al. 2011).

Visualization is obviously a key component in the quest for effective sense-making of Big Data. However, is the sense-making model implied by a visualization-driven engagement sufficient? Might there be other equally meaningful roles for human beings in this endeavour? If this provisional model seems narrow in its conception of human beings in the sense-making system, what might be other models to better inform technology development? This paper offers an alternative model, developed from direct observation of graduate students engaging in a complex sense-making challenge, that organizes the interactions of human beings, technology and data into a more dynamic and integrative sense-making system.

Case study: the search for meaning in 242,000 rows of ‘training’ data
In the spring of 2014, graduate students at the IIT Institute of Design engaged in an open exploration of how to use micro-sensors to understand human
behaviour, led by adjunct professor John Cain and his colleagues at Sapient’s Iota Labs. Student teams were tasked with using sensor-based technologies — anything from activity trackers to arduinos to Wii controllers — to generate a dataset about some aspect of human behaviour. Students were challenged to make sense of that dataset by whatever means, as long as their process advanced their research question.

For clarification, it must be noted that first author Erwin acted as an outside observer of this and other student teams; co-authors Bond and Jain are IIT Institute of Design graduate students who performed the work presented on the following pages.

This paper presents as a case study the experience of co-authors Bond and Jain and their unscripted journey through 242,000 rows of sensor-generated data. What results is a demonstration of how a large data set is evolved — through a series of improvised human and technical interventions — into both a visual language and, in effect, a ‘text’ that can be read and interpreted by human beings. Then, generalizing from the case study, first author Erwin offers an alternative process model for human-technical collaboration in the sense-making of data — one more nuanced and inclusive than seen in the current literature.

**Methodology and approach**

The team framed their research question thusly: ‘We seek to understand people’s habits of task switching (moving between tasks/activities) in order to design tools that help us transition from one task into the next. We plan to do this by exploring task intentionality, how participants switch between both digital and physical activities, and how these activities affect one another’.

The team then hypothesized two patterns in the behaviours of users as they dip in and out of tasks: ‘clean’ switching — fully concluding one task before beginning a new task — and ‘fuzzy’ switching, in which a second task begins while the first task is concluding, creating an overlap of activity (see Figure 3).

Next, they created a plan of study and instrumentation. The objective was to generate a set of training data, defined as data sufficiently representative of a user or a behaviour that an analytic approach could be developed and refined for use at a larger scale. The team recruited two time-pressured graduate students to participate for twenty-four hours a day for seven days in a row.

**Data collection and research protocol**

Quantitative data was generated by RescueTime running on the laptop and mobile phone of each participant. RescueTime is a productivity-tracking application that samples technology usage every five seconds. The output is a file that logs each application, document or webpage engaged by a participant during that time. This instrumentation of participants produced data sets that covered 336 possible hours of technology usage (24 hours x 7 days x 2 devices) per participant. Sampling every five seconds, RescueTime
captured roughly 17,280 entries for each day per device, for a grand total of approximately 242,000 rows of data per participant.

The team then used ethnographic methods to build a qualitative context for the same timeframe. They started with a one hour interview of each participant to establish general behavioural context and to better understand productivity and time management-related behaviours. To inform questions of task-intentionality — what was the difference between what participants said they needed to do and what they actually did — and distraction, participants engaged in three different forms of self-reporting:

- each morning, send screenshots of any daily calendar or ‘to do’ list that captures intentions for the day
- each afternoon, respond to a probe about what’s been on the mind of the participant in the last 20 minutes
- each evening, respond to brief survey to rate how pressing the day’s activities were, write about what made them feel productive/unproductive, and document any non-technology tasks undertaken.

**Analysis: making sense of the training data**

This section presents the efforts of the graduate student team to translate a single study participant’s data set into insight about task-switching behaviours. The team chose to experiment with one participant’s data first, so as to establish an analytic process that might be repeated with the second — or any subsequent — participant. The analytic progression is best described as an unscripted journey through the data, as the team had no rubric or prior examples for how to proceed. Their efforts are outlined in six stages, but it is important to underscore that, in real-time, this process evolved organically and with little certainty or clarity about the kind of results any particular effort would produce.
<table>
<thead>
<tr>
<th>Category (generic)</th>
<th>Activity (site/app)</th>
<th>Duration (in seconds)</th>
<th>Device (brand + type)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure 4** Sample output of the study participant's raw data as collected by RescueTime.
Stage one: assessing the raw data, identifying necessary optimizations
The first stage of the analytic progression looked at the nature of the RescueTime output (see Figure 4). Each line in a RescueTime file catalogs a ‘task’ and records the following:

- unique identifier numbers (one for each instance, participant and device)
- device specifications (brand, type, name, operating system)
- activity information (time/date, duration in seconds, time zone, activity description — such as google.com or ‘newtab’ — and the object of that activity, such as a document or web page title)
- category information, as assigned by RescueTime (category name, such as ‘reference & learning’; followed by a subcategory, such as ‘search’; followed by a productivity score from -2 to +2)

However, after reviewing how RescueTime assigned categories, it became clear to the team that the generic schema used by the software was insufficient for the research objectives. A significant number of entries were categorized as ‘uncategorized’, for example. Other categories, such as ‘reference & learning’ included a diverse set of activities that, on close examination, did not group into a meaningful collection. Creating a categorization scheme that successfully grouped like activities, and better reflected the participant’s intent, became the next stage of work.

Stage two: evolving participant-relevant categories
To solve the category/categorization problem, the team stepped back into the qualitative data to look for patterns of intent — evidence of what the participant was attempting to do — that might help to more accurately characterize the RescueTime tasks (see Figure 5). Using this and the participant’s morning, afternoon, and evening self-reporting activities, the team generated a list of categories that they believed more accurately characterized the participant’s tasks. The team reviewed the categories and recategorized data with the study participant for refinement. This produced 12 final categories and a robust case for each.

---

**Figure 5** Twelve new task categories created by analyzing the qualitative data from the participant’s morning, afternoon and evening logs and conferring with the study participant.
**Stage three: visualization for discovery**

After recategorizing the data, the graduate team ran the revised data file through Excel’s Pivot tables, resulting in visual representations that failed to reveal meaningful patterns. Switching software, the team ran the file through Tableau. One Tableau-enabled visualization (see Figure 6) represents each discrete RescueTime entry as a vertical line, and each line is coloured by the category of the task. Areas of continuous tone suggest periods of focus. The patch highlighted cluster on Saturday afternoon, for example, is an extended effort on the participant’s part to fix and then replace a broken power cord. This was knowable because the team, observing the cluster, returned to the qualitative data — Saturday’s afternoon prompt and evening log — to discover the participant writing about this frustrating aspect of her day. Another pattern: it’s clear from the overall activity level that the participant didn’t sleep much.

**Stage four: optimizing the data—establishing relationships between tasks**

The Tableau visualizations were interesting but did not advance the research question relative to distinctions between types of task switching. The team posed a new question: when the participant switches tasks, does it matter how alike or dissimilar the tasks are? Are all switches equal?

The team created a symmetric matrix to establish relationships between each category (see Figure 7). With qualitative data as a guide, each category was scored for similarity to every other category, including itself, in terms of the participant’s objective. Any category against itself was scored a five, the
highest level of similarity; all others were scored from five (very alike) to zero (least alike). For example, the relationship of communication: social to itself produced a score of five, while the relationship of communication: social to work: school produced a score of zero, as these two categories of tasks were maximally dissimilar in terms of objective.

For this table to aid in computation of meaning, each score needed to be as accurate as possible. To verify scoring logic, the team again went back to the participant for validation and refinement.

Stage five: the breakthrough visualization — a shape is born
The team added the similarity scores to the data file, then ran the newly-optimized data through various Tableau options again. Through a chance experiment with a line graph visualization, the tasks transformed into a series of shapes. And, remarkably, the shapes began to repeat and form identifiable patterns (see Figure 8).

Stage six: decoding the shapes — who is Batman?
With tasks now depicted as connected shapes, the data no longer reads as discrete tasks but instead as a flow of interconnected events. The student
team was challenged to interpret both the individual shapes and the patterns they formed. What did a ‘Batman’ shape mean? Was there a meaningful difference between the shapes that looked like whales and those that resembled the state of Washington? What did it mean that a Batman shape was sometimes preceded by a crown shape?

To see their data as a whole, the team printed the entire dataset and taped it together to create a long scroll of continuous data that stretched across two classroom walls. This depiction organized the data by time on the horizontal axis, with 22 ft of paper required to show 24 hours. Days were stacked vertically, so that 7am Monday was positioned over 7am Tuesday, etc.

Over the course of several weeks, classmates and Sapient staff participated in reviewing the data, asking questions, posing hypotheses, and speculating on aspects of the shapes. Issues included the importance of empty gaps between shapes, what was signified by changes in slope or by different horizontal line lengths and, of course, the meaning of the shapes themselves.

What the team discovered: length, depth, and frequency of dips in the shapes indicate levels of participant focus. Slope is an indicator of relatedness of tasks: Tableau draws the shapes based on the scale of zero to five, so a long vertical line represents a five, and a mild dip down in a shape signifies a switch to a highly-related task (typically a score of 4). A steep change in slope, by contrast, signifies switching to a highly unrelated task (with relatedness of maybe a zero to a three). It was determined that gaps between shapes actually signalled the continuation of a task, with no change or switching.

Iterative reviews of the original RescueTime files against Tableau’s visual depictions eventually suggested the meaning of various shapes. ‘Batman’, for example, most often represents a period of preparation before entering an extended period of focus (see Figure 9). The shape is created when switching between highly-related tasks, such as moving from a ‘work’ task to another ‘work’ task to a ‘work:school’ task (i.e. checking with a teammate about next steps on a project) and back to a ‘work’ task for a more extended period...
of time. This pattern is seen in data from both the participant’s phone and their PC.

The now linear, interconnected nature of the visualization, coupled with a growing understanding of what each shape signifies, allowed the team to ‘read’ the data, as a sort of text in a custom-built alphabet. They could follow a period of activity, understand its constituent parts, and begin to compare patterns across participant devices and days. Now able to read larger stretches of data, the team could identify periods of extended focus and distraction (see Figure 10).

Implications

Witnessing this process and its output offers two significant points of learning.

1. The emergence of a personal pattern language
Because of the short timeframe of the class, the student team worked only with one participant’s data set. In effect, no one knows whether the batman, whale, or the crown shapes are unique to this person’s usage patterns, or whether every individual studied this way might produce a similar set of shapes. It’s provocative to consider, and perhaps even likely, that these shapes might be unique identifiers of the technology habits of the participant,
functioning more like fingerprints than like universal patterns. The graduate team’s process and its output, then, potentially offer a path to identify the personal pattern language of every individual.

The notion of a pattern language is not new. Pattern language for design purposes is most commonly associated with Christopher Alexander, who proposed a system of architectural components that could enable design at any scale, whether a room or a town. In *A Pattern Language: Towns, Buildings, Construction* (1977), Alexander identifies the primary elements of architecture, akin to an alphabet, and describes grammatical rules or principles for combining those elements.

Similarly, the work presented here suggests that human technical behaviour can be captured and translated into a catalogue of *visual primitives* — small-scale patterns that can be described and used in combination to ‘read’ the behaviour of a user interacting with technology. Bond and Jain also began to describe the properties of those primitives (see figure 9), such as dimensions and variations, that allow the primitives to be differentiated and formalized. The last aspect of a pattern language, described by mathematician Nikos A. Salingaros (2000) as the identification of ‘connective rules’ or combinatorial patterns of primitives by which primitives become a true language, has only begun to be explored. Jain and Bond observed, for example, that a Batman shape was often preceded by a crown shape, and then could be followed by either a whale or Washington shape. But much more work is required to link primitives to larger patterns.

What might be the significance of a formalized personal pattern language? First, it could be a potent new means of making sense of personal data. It could, for example, enable new self-knowledge — what are my patterns of technology usage and how do they affect key behaviours, such as focus and distraction? Second, it could transform interactions between humans and their technologies. As the graduate team pointed out, usage patterns could be codified and taught to digital devices so as to enable those devices to act in greater harmony with the human beings that use them. Imagine a digital
device that can identify that its user is in an intense period of focus and flow, and manages interruptions, such as delivery of email, accordingly. Salingaros (2000) noted, in advocating for the use of architectural pattern languages to drive design, ‘The smaller the scale on which a pattern acts, the more immediately it connects to human beings’. A personal pattern language of the sort suggested by Bond and Jain may form the most intimate language yet, connecting humans to themselves and orchestrating the internet of things to an individual’s benefit.

2. A more people-driven model for making sense of data

Observation of the graduate teams overall, including not just the team of Bond and Jain but also other teams given the assignment, in their collection and analysis of their sensor-generated data, suggests a sense-making model that is more varied, integrated and people-driven than those (both implied and stated) in the current literature.

Sense-making is an iterative process (figure 2): As Heer et al. suggest, sense-making demands an iterative dance between computation and discovery, a process which is highly experimental because outcomes are often unknowable in advance. Bond and Jain, for example, state that the output of each iteration only raised more questions — interesting, more refined and ripe questions — but did not produce definitive answers that would have permitted exit from the cycle.

Sense-making is a hybrid process: For all teams in the class, interpretation of their sensor-based data required qualitative context. Each iteration required raw data to be optimized again, necessitating reference to the qualitative data. Bond and Jain, for example, used qualitative data to create meaningful categories; to establish relationships between categories; and to assign weights to those categories in order to move into the next round of computation. While the quantitative data provided incredibly granular information about participant behaviour, it was the qualitative data that allowed the teams to put a frame of meaning around the granules, imputing ‘motivations’, ‘intentions’, ‘surprises’, and ‘distractions’ to the raw numbers.

Sense-making is also a physical process: For these graduate teams, sense-making required an analogue component, an off-screen experience that was as robust as the on-screen experience to advance the work. Every team printed out their data and posted it on a wall for what one social scientist drily termed ‘interocular appraisal’ — or a little eyeball work. The dance between the on-screen and off-screen work proved to be as critical to the sense-making process as the dance between quantitative and qualitative data.

Finally, sense-making is also a fundamentally social process (Figure 11): The collaborative approach to sense-making is the real story here: to advance their work, all teams engaged in discussion, debate, review, guesswork, and even group intuition. Social sense-making is a uniquely human and uniquely powerful means of processing information for meaning. An accurate sense-making model is not complete unless it depicts this factor.
Conclusion

We are, all of us, data scientists included, just at the beginning of the Big Data revolution. While new algorithms, visualizations, and software geniuses are important to incubate, it’s too early to fall into orthodoxies about how sense-making best occurs or needs to be supported.

The work presented here provides a rich example of the unexpected outcomes that are possible when Big Data challenges are pursued outside of conventional, technology-centric practices. In evidence are three discoveries: First, it is possible to transform sensor-generated data into a custom alphabet that can be read and interpreted by human beings, who are then able to engage that data using a more holistic and collaborative analytic process than before. Second, when combined with qualitative data, it is possible to transform sensor-based data into a personal pattern language, potentially capturing the unique behavioural and usage patterns of each individual with their devices. This is useful, as it may produce insights and enable new interactions for technology that we can only begin to imagine now. Third, demonstrated here is a subtle but significant shift in analytic focus: the shape-based visualization transformed discreet tasks (computers engaged in counting) into a continuous flow (humans engaged in activity). This is an important step in a new direction for personal informatics, as a frontier for researchers continues to be how to move from highly-granular units of activity toward meaningful representations of human effort.
The work presented here also illustrates a robust partnership of human beings and technology in making sense of Big Data. This model adds dimension to technology-centric models that are increasingly the conception driving Big Data research and commercial efforts. Given the experimental nature required to make sense of big/thick/complex/dynamic/organic data, it may be prudent to hold off conceptualizing sense-making as a technology or service delivering answers. Instead, we might conceptualize sense-making as more of a lab, a human-technical system that must partner in ways not yet knowable in order to investigate an unclear and murky future.

Acknowledgements

The authors would like to thank Alisa Weinstein for her contributions to early-stage research definition. We would also like to thank Iota team members and adjunct professors, Thomas McLeish, Peter Binggeser, Laura Mast for their guidance through this project.

References


Notes on contributors

Kim Erwin is an assistant professor at IIT’s Institute of Design, with research interests in making complex information easier to understand and use. Her digital investigations include tools and methods for the visual analysis of large qualitative datasets and the exploration of ‘big personal data’ generated by self-tracking devices. Her analog investigations focus on methods for exploring, building and diffusing critical knowledge inside organizations during the innovation planning process. Her recent book, *Communicating the New* (Wiley & Sons), describes these methods.

Correspondence to: Kim Erwin, IIT Institute of Design, 350 N. LaSalle, 4th Floor, Chicago, IL 60202, USA. E-mail: kerwin@id.iit.edu.

Maggee Bond is a design researcher at Gravitytank, an innovation consultancy in Chicago. Her work leverages a wide variety of user research methods to design products and services for global Fortune 500 companies. While pursuing her Masters degree at the IIT Institute of Design, Maggee spent a significant amount of time immersed in both qualitative and quantitative data analysis and now believes that the most profound insights often come from the integration of the two.

Aashika Jain is an Insights Lead at Doblin Inc., a strategy consultancy owned by Deloitte Consulting LLP based in Chicago. She is trained in design, business and user research and works at the intersection of these domains to convert user research into offerings and strategies that create value for users and businesses alike. While pursuing her dual masters at the IIT Institute of Design, she studied and created new ways of visualizing research by combining qualitative and quantitative data. She believes that approachable data visualization can give designers creative super-powers.