Smart Sensemaking Systems, First and Foremost, Must be Expert Counting Systems

Jeff Jonas

INTRODUCTION

Man continues to chase the notion that systems should be capable of digesting daunting volumes of data and making sufficient sense of this data such that novel, specific, and accurate insight can be derived without direct human involvement. While there are many major breakthroughs in computation and storage, advances in sensemaking systems have not enjoyed the same significant gains.

This article suggests that the single most fundamental capability required to make a sensemaking system is the system's ability to recognise when multiple references to the same entity (often from different source systems) are in fact the same entity. For example, it is essential to understand the difference between three transactions carried out by three people versus one person who carried out all three transactions. Without the ability to determine when entities are the same, it quickly becomes clear that sensemaking is all but impossible.

Essentially, sensemaking systems must first and foremost be expert counting systems.

Of course, smart systems must be able to do far more than just count people, places, things, events, groups, etc. Among other things, smart systems must be able to make assertions, reconsider earlier assertions as new evidence is presented, recognise importance, and determine what or whom to notify when such relevance is detected. Fortunately, systems that focus on "counting" first will come to realise that many of the requirements of sensemaking systems become easier, even the hard problems facing the sensemaking community.

WHY COUNTING MATTERS SO MUCH

When someone throws a Frisbee to you, your sense-making faculties are utilised to predict the course of the disc so you can catch it. Based on the vector (direction) and velocity of the Frisbee and one's previous experience of similar events—most folks have sufficient estimation skills to catch the disc even if the park is filled with people flinging Frisbees around. This example of vector and velocity are very straightforward, in part, because there is a single, integrated system—your eyes and brain—that collects and processes the series of observations as the Frisbee makes its arc.

What if you could not use your eyes to watch the Frisbee in first person? Instead you had to rely on a small number of friends presenting observations in the form of photos, Twitter feeds, short stories, essays, heat maps, etc. No matter how "slow motion" this was attempted, it would be hard to establish which observation related to which Frisbee. This makes it impossible to estimate the vector and velocity of your Frisbee. Consequence: In this case a Frisbee may hit your forehead.

Humans involved in more complex tasks, like 911 emergency call centers, rely equally on vectors and velocities to make sense of events. If emergency operators were to receive three calls reporting gun shots fired, a large number of scenarios are possible including: There was one shot reported three times; there is one person who shot three times (possibly while on the run over some distance); maybe three people each fired a shot in three separate incidents. Making sense of this information requires the analyst be able to count discreet entities (people, places, things, etc.) in spite of duplicate, inconsistent, and at times errant reporting. Emergency services personnel address such sensemaking challenges by asking the observer for very specific details such as where, when, and features of the entities (*e.g.,* estimated height, weight, clothing, make and model of the car and its license plate number). Such facts are essential for analysts to differentiate entities , that is count.

Automated sensemaking platforms don't have it so easy. Unlike the Frisbee player, the data presented to sensemaking systems comes from many perspectives (disparate data sources). And unlike the emergency services operator, there is so much data there aren't enough humans to interrogate witnesses in effort to resolve ambiguity.

Bio-surveillance sensemaking systems might draw on newspaper stories, blogs, Twitter feeds, social networking sites, conference papers from international pandemic conferences, etc, to compute emerging threats. Without an ability to count repeated references to the same people and places, it would be impossible to determine macro level trends. Does the open source reporting refer to one person infected with H1N1 reported many times, or many people with H1N1 all in one dense geographical region, or many people in many places with H1N1?

Whether the sensemaking system is intended to improve insight or prediction in bio-surveillance, health care, stability of financial eco-systems, or national security; if the sensemaking system cannot first and foremost count, it will not produce reliable insight.

WHY COUNTING IS SO HARD

Not every white van is the same white van. To determine if it is the same van, one must consider the evidence at hand. If the Vehicle Identification Number (VIN) is the same, this makes for some compelling evidence; unless of course the make, model and year are now somehow different. If you cannot obtain the VIN, a matching license plate number, make, model and year would lead to a high degree of confidence as well.

The process of determining (same) identification involves an evaluation of agreeing and disagreeing features. Accounting for the fact that some features are highly discriminating like a VIN or passport number, other features are not discriminating at all; however, they are lifetime stable, such as a vehicle's make and model or a person's date of birth or place of birth. Some features can change over time, like vehicle owners and license plate numbers or a person's residential address, while some features can change over time in gradual increments, like colour of the car as it fades or the weight or age of a person.

Another complicating factor is that sensors produce different, and often incompatible, features. For example, in people related data one might find these two records:

William Angstrum
PO Box 99811

Bill Angstrum
123 Main Street

If this is all that is known, there is no way to assert with any confidence that they are the same person. If they are the same person, one would only come to realise this if another observation arrived which shared features from both records. For example:

William Angstrum
Current address: PO Box 99811
Former address: 123 Main Street

Yes, it could be a junior and senior. Add a few more observations like dates of birth and most would come to believe (assert) this is the same person.

What makes counting even more difficult is poor data quality (e.g., misspellings, missing fields), intentional deceit (e.g., fabricated identities), and natural variability (e.g., nicknames, handles, abbreviations, alternate spellings). Practically speaking, it is virtually impossible to determine same identity with absolute and permanent certainty. As such, counting involves making assertions—being so sure different observations reflect the same identity—that a claim can be made that they are the same. One must remain ever vigilant to recognise that an earlier assertion was made in error should a new piece of evidence warrant a different conclusion.

Imagine that you work at a law office and meet a nice young lady, well-dressed, pleasant little laugh, who presents a state-issued identification to confirm her identity before you hand her a cheque. Forty minutes later the same young lady returns in the same clothes, carrying the same identification, and exhibiting the same pleasant little laugh and demands that you give her the cheque. With absolute certainty (so certain you may bet money, your reputation, or maybe even "swear on your life") you are convinced that this is the same lady and you are being tricked or she is crazy. Until her identical twin enters the room carrying the same identification document. Question: How is it that two ladies who share exactly every observable feature can instantly be recognised as two different people? Answer: Space-time disagreement—the same thing cannot be in two different places at the same time. Being able to identify this fact is somehow a built-in feature of a human being's innate ability to count and subsequent sensemaking.

This highlights two particularly interesting issues about what makes counting entities a difficult problem.

1. One of the only ways to have absolute certainty about identity involves considering space and time features.[1] However, at this time most data sources are

[1] An exception being some forms of biometrics like DNA.

not collecting this geo-locational and temporal data at all or with sufficient precision to enable more precise identification.

2. As errors in identity assertions will be made, it is essential that smart systems are able to reverse earlier assertions (detected errors) based on new observations. Much in the same way the worker in the law office was certain the lady was one and the same; until presented with evidence she had an identical twin.

If counting ″like″ entities is that easy, everyone would be doing it and the current generation of sensemaking systems would be substantially more intelligent.

EXPERT COUNTING SYSTEMS: ESSENTIAL INGREDIENTS FOR SENSEMAKING

Systems that detect duplicates within a data set and between data sets have been around for years. Match/merge systems, as these have often been called, have been used to ensure that direct marketers don’t waste postage by mailing the same promotion to one person three times. These first-generation counting systems *do not* have the essential ingredients necessary to support smart, sensemaking systems.

What then, are the most essential ingredients of expert counting systems?

Expert counting systems need to rely on incremental learning techniques rather than being dependent on training data. Systems that require training data have to be periodically retrained as underlying data sets evolve. When managing large-scale sensemaking systems, the idea of having to retrain and reevaluate historical observations is impractical.

Choosing between using probabilistic[2] or deterministic[3] algorithms is unnecessary. Expert counting systems perform best when both probabilistic and deterministic methods are applied. The real question is the order in which these methods are applied. Because dependence on training data is less than ideal, leading with deterministic algorithms is appropriate. Then probabilistic methods are applied to learn statistical distributions over time, applying this additional insight in real time.[4]

[2] Simply put, systems that use statistical distributions found in data to make future assertions.

[3] Simply put, systems that have explicit rules that are applied to make future assertions.

[4] Using the flip/flop processes as described in the following page, learning fixes the past to avoid reloads.

Unlike old school counting systems that are designed to compare File A to itself or compare File A to File B, expert counting systems perform a "resolution" process. This means that each inbound entity is not evaluated against individual data sources or individual records, rather, inbound entities are compared to existing entities which may be composed of one or more historical entities now conjoined. Resolved entities accumulate features over time and enable resolutions that are otherwise impossible to establish.

Current Inbound Record
Mark Lawrence Smith
DOB: 06/1976
PP#: 11334455

Historical Record 1
Mark L. Smith
+1 702 555-1212
123 Main Street

Historical Record 2
Mark Smith
DOB: 06/12/1976
PP#: 0011334455
702.555.1212
123 S. Main St

In the above example, the **Current Inbound Record** would have no chance of being recognised as the same identity as **Historical Record 1**. However, it would be obvious that the inbound record is the same identity as **Historical Record 2.** Expert counting systems that use resolution processing deal with this simply by recognising historical records 1 and 2 as "same identity," which means the **Current Inbound Record** is evaluated against this first conjoined entity (on the left) to become the second entity (on the right):

Identity 1 Before Inbound Record	
Mark L. Smith	R1
Mark Smith	R2
+1 702 555-1212	R1
702.555.1212	R2
123 Main Street	R1
123 S. Main St	R2
DOB: 06/12/1976	R2
PP#: 0011334455	R2

Identity 1 After Inbound Record	
Mark L. Smith	R1
Mark Smith	R2
Mark Lawrence Smith	R3
+1 702 555-1212	R1
702.555.1212	R2
123 Main Street	R1
123 S. Main St	R2
DOB: 06/12/1976	R2
DOB: 06/1976	R3
PP#: 0011334455	R2
PP#: 11334455	R3

High performance counting systems make one of two assertions: same or not same … and persist (store/remember) this, for example, in a database. If the counting system attempts to only associate observed instances of entities with degrees of probability/confidence serious scalability issues ensue.[5]

When assertions are made, expert counting systems must favour the false negative[6] over the false positive.[7] If the counting system gets too opportunistic (favoring false positives) in its assertions of same, there is a tendency for the discreet resolved entities to implode, creating what could be characterised as "fur balls." On a more technical note: False negatives have the opportunity to be remedied over time as new data is presented—in an automated fashion through the "flip/flop" property described below.

Because some identity resolution assertions are incorrect, expert counting systems must be able to flip/flop (change their minds) on these earlier assertions. Upon each new record, an expert counting system considers "now that I know this, had I known this in the beginning of time, does this change any earlier assertion, and if so … remedy all such earlier assertions."

> Excerpt from Jeff Jonas Blog entitled: Smart Systems Flip-Flop
> http://jeffjonas.typepad.com/jeff_jonas/2008/06/smart-systems-f.html
>
> Certainty often shifts with observations over time. And this is good.
>
> …
>
> But 'smarts' requires much more than just available data and good correlation. Two additional critical elements of smart systems are:
>
> 1. An ability to make assertions based on new data points
> 2. An ability to use new data points to reverse earlier assertions
>
> …
>
> Smart systems also have to be able to undo earlier assertions made in error. If a new observation is in fact evidence that invalidates earlier assertions, these earlier incorrect assertions must be corrected (there are some caveats, more on this at another time).
>
> Once presented with compelling new data, systems that cannot flip-flop on previous certainties … are dumb. The same goes for humans.

[5] The reason why is beyond the scope of this article. Feel free to write the author for more iunformation on this point.

[6] The term "false negative" is used to describe the condition of not detecting something that is the same. For example, thinking the records belong to two different people when they are in fact the same person.

[7] The term "false positive" is used to describe the condition that occurs when something is detected as the same when it is not. For example, thinking the records belong to the same person, when they are in fact different people.

When an expert counting system reverses an earlier assertion, it must be able to disassemble and reassemble previously established identities. To do this, the counting system must therefore meticulously maintain full attribution[8] of every record and data point. First-generation counting systems that merge records, introduce data survivorship rules and/or other lossee processes,[9] are unable to flip/flop to reverse earlier assertions. Retaining all encountered records and features also means retaining data that is inconsistent, incorrect, or outright designed to be deceiving. Contrary to most current thinking, this is in fact an important property of expert counting systems.

Point being: Bad data is good. By retaining the natural variability of data, sensemaking systems have a significantly better chance of detecting a weak signal.

In addition to collecting both good and bad data, expert counting systems must be screaming fast. Fast enough to keep up with current ongoing transactional data isn't good enough. Rather, these systems must be much faster that that because they must be able to ingest the even larger pile of historical data (*i.e.,* learning one's past). Unlike data warehouses, multi-source data cannot be simply commingled in a big pile; it must be properly counted.

Whether the sensemaking environment is serving real-time missions or periodic analysis, expert counting systems run optimally when designed for real-time streams—regardless of whether they are serving real-time missions or periodic analysis. While beyond the scope of this article, there are deep architecture reasons why batch systems never seem to be able to grow up and become fast real-time engines. On the contrary, streaming engines can ingest and resolve data from real-time streams or batches with indifference. And on a related note, a funny thing about batch analytic systems: The more often they produce valuable insight, the more often the user asks: Can I get these kinds of answers sooner?

And finally, a sensemaking platform is smartest and scales best if relevance and insight are evaluated simultaneously as data is ingested on data streams—as it is computationally most efficient for sensemaking to be made in real-time as observations become available. For this reason, expert counting systems deployed into sensemaking environments must have ultra-low latency and provide deep native

[8] More about full attribution here: http://jeffjonas.typepad.com/jeff_jonas/2006/10/source_attribut.html

[9] Lossee processes are processes that result in the destruction (or loss) of data. An example would be if a record has the name Bill and William associated with it, some systems would drop the word Bill, keeping only William.

integration with downstream algorithms which are evaluating newly contextualised observations for relevance.

EXPERT COUNTING SYSTEM HELP SOLVE OTHER HARD SENSEMAKING PROBLEMS

While expert counting systems are of critical importance to smart sensemaking systems, there are other necessary analytic sensemaking activities that are in themselves their own hard problems. For example, before counting, analytics are required to extract and classify useful features from observations. After incoming entities are counted, different algorithms are used to determine association between resolved entities (e.g., link analysis)—this being the next critical step in contextualisation. Beyond that, other analytic methods are used to perform such activities as relevance detection and insight dissemination.

A number of these additional sensemaking system components are in themselves very hard problems—in fact, sufficiently challenging to blunt major advances in this field, such as:

- Entity extraction and classification,[10] for example, are proving to be rather imprecise. Passing extracted and classified data with low accuracy rates (*e.g.,* less than 90% accuracy) begins to materially degrade expert counting systems.
- Scalability issues are being faced as the volumes of data are staggering.
- Recognising what constitutes relevance and insight has equally challenged sensemaking systems—the production of accurate and novel intelligence has not been easy to come by.

Expert counting systems will bring a great deal of relief to these impediments and more.

Entity extraction and classification algorithms are going to see material improvement in their accuracy as they interact with expert counting engines. While

[10] Entity extraction refers to selecting key features out of unstructured data. Classification in this usage refers to properly characterizing what a feature means. For example, entity extractors and classifiers are can be used to extract names and phone numbers from text and recognize the names are people versus companies and determine what kind of phone number it is (e.g., mobile phone, fax line).

current techniques in this area rely on elaborate, domain specific rules and static training data sets, next-generation extractors will peek ahead into the reconciled view of what has been learned, incrementally, up to the moment. Drawing on this rich context, in what could be characterised as a two-way conversation between the extractors and the world of counted observations, will prove to substantially improve accuracy.

Sensemaking systems with embedded expert counting engines will see not only greater accuracy (lower false positives and lower false negatives), but also may simultaneously enjoy greater performance over more data. While this sounds counterintuitive, there are real world principles that have been seen in production systems whereby more data equates to faster sensemaking. More about this concept explained here:

> Excerpt from Jeff Jonas Blog entitled: The Fast Last Puzzle Piece
> http://jeffjonas.typepad.com/jeff_jonas/2008/09/the-fast-last-puzzle-piece.html
>
> The notion that the more data, the slower the system—ain't always true. My favorite way to explain this very important phenomenon involves the familiar process of assembling a jigsaw puzzle.
>
> The first piece you take out of the box and place on the work surface requires very little computational effort. The second and third pieces require almost equally insignificant mental effort. Then as the number of pieces on the table grows the effort to determine where the next piece goes increases as well. But there is a tipping point where the effort to determine where to place the next piece gets easier and easier … despite the fact the number of puzzle pieces on the table continues to grow.
>
> …
>
> This does not apply to all domains. This behaviour requires: (a) observations from the same universe; (b) observations with enough features to enable contextualisation; (c) observations in which these features can be extracted, enhanced and classified; (d) sufficient saturation of the observational space; and (e) enough smarts to stitch these puzzle pieces together.

When sensemaking platforms are evaluated, errant output can generally be caused by 1) not enough observations, or 2) an inability to make sense of what one knows. If there is not enough data, no analytics will fix the problem. The only remedy is more observations. If the data exists and the problem is analytics, expert counting is of course required. And when such counting is in place, systems accumulate context over time. Counting systems will be shown to substantially improve sensemaking systems as incrementally improving context enables more fine-grained relevance and insight processing.

Other hard problems are also likely to give way, including sentiment analysis[11] and concept classification.

CONCLUSION

Sensemaking platforms that are not equipped to count like entities will have a difficult time producing meaningful intelligence. Counting is hard, which is why it is so often overlooked or put off as a future "to do." To the contrary, it must be done first and it must be done exceedingly well. Once counting is mastered a number of very hard problems facing the sensemaking community are going to become more tractable.

Sensemaking systems that cannot count will miss the obvious, and corrupt all downstream processes (e.g. secondary systems or human analysts who are taking these predictions as inputs). Such systems will also fail to scale. Finally, to the extent an organisation is in the "we want to detect weak signals" business, counting becomes even that much more important.

Smart systems, prediction systems, sensemaking systems, situational awareness systems, incremental learning systems—whatever one calls these thing—sensemaking systems must first be able to count if they are to be relevant.

[11] Algorithms that determine how some feels about something e.g., hate, dislike, indifference, passion, etc.