BREADCRUMB RECOMMENDATION SYSTEM: THE NEXUS OF USER
INTUITION ON CONCEPTUAL AND STATISTICAL ASSOCIATION

A Thesis
Presented to the
Faculty of
San Diego State University

In Partial Fulfillment
of the Requirements for the Degree
Master of Science
in
Computer Science

by
Atul Belur Vijayakumar
Summer 2013
SAN DIEGO STATE UNIVERSITY

The Undersigned Faculty Committee Approves the
Thesis of Atul Belur Vijayakumar:

Breadcrumb Recommendation System: The Nexus of User Information on
Conceptual and Statistical Association

Joseph Lewis, Chair
Department of Computer Science

Kris Stewart
Department of Computer Science

Amy Schmitz Weiss
School of Journalism and Media Studies

5/16/2013
Approval Date
Copyright © 2013
by
Atul Belur Vijayakumar
All Rights Reserved
DEDICATION

This work is dedicated to my parents, Jayanthy and Vijayakumar, for having the patience, trust and abundance of love without which my journey in life and academics on this pale blue dot would have been staggered and lost.
The vast majority of intelligent behavior we see in nature comes not from deductive reasoning or any such logical processes, but from the ability to make good decisions in the face of information that is simultaneously vast, incomplete, and contradictory.

--Joseph A. Lewis

Making decisions, associations, conceptual linkages and such arise from the myriad contributions of agent processes creating opportunities to continue to make yet other decisions and associations. By asking why, by being curious and being impatient with ignorance, we hope to succeed in bringing to the aid of humans, a helping hand via machines which have been helped by the occasional clarity and prowess of the human mind.

--Lewis and Belur
ABSTRACT OF THE THESIS

Breadcrumb Recommendation System: The Nexus of User Intuition of Conceptual and Statistical Association
by
Atul Belur Vijayakumar
Master of Science in Computer Science
San Diego State University, 2013

The research objective here is to demonstrate as a proof of concept that an emergent distributed agent recommendation system could deliver concept based justifications rather than statistical based justifications from available document sources. We discuss the role of emergence in computational systems and how conceptual networks are realized differently in such systems. We also discuss how journalism is a particularly relevant case study where big data presents itself as a place where recommendation is necessary but also because of big data, a place where people have difficulty and mistrust of systems that try to give them answers. We present test scenarios for rapid prototyped implementation in Python or similar environment. Although not explicitly built on the Starcat framework, the test scenarios rely on features that follow its style of architecture. Finally we discuss results of this proof of concept level of implementation and how it can be advanced with projects that help realize more fully this novel type of recommendation system.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT .......................................................................................................................... vi</td>
</tr>
<tr>
<td>LIST OF TABLES ............................................................................................................... ix</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENTS ......................................................................................................... x</td>
</tr>
<tr>
<td>CHAPTER</td>
</tr>
<tr>
<td>1 THE BREADCRUMBS RECOMMENDATION SYSTEM ................................................................. 1</td>
</tr>
<tr>
<td>A Tool for Data Source Discovery ..................................................................................... 1</td>
</tr>
<tr>
<td>An Overview ....................................................................................................................... 3</td>
</tr>
<tr>
<td>2 BACKGROUND RESEARCH ON PROBLEM SPACES ............................................................ 8</td>
</tr>
<tr>
<td>Of Data and Breath ............................................................................................................ 8</td>
</tr>
<tr>
<td>Signal and Noise ................................................................................................................. 10</td>
</tr>
<tr>
<td>Big Data Domains ............................................................................................................. 12</td>
</tr>
<tr>
<td>Journalism and Big Data .................................................................................................. 12</td>
</tr>
<tr>
<td>Recommenders .................................................................................................................. 14</td>
</tr>
<tr>
<td>Search Engines and Data .................................................................................................. 15</td>
</tr>
<tr>
<td>Google and PageRank ....................................................................................................... 16</td>
</tr>
<tr>
<td>Travails of the Result List ............................................................................................... 17</td>
</tr>
<tr>
<td>Databases and Querying .................................................................................................... 18</td>
</tr>
<tr>
<td>Statistics and Contexts ...................................................................................................... 19</td>
</tr>
<tr>
<td>Recipe for a Story .............................................................................................................. 21</td>
</tr>
<tr>
<td>The Newsmaking Process ................................................................................................. 22</td>
</tr>
<tr>
<td>Computation in Living Systems ....................................................................................... 23</td>
</tr>
<tr>
<td>The Immune System ......................................................................................................... 23</td>
</tr>
<tr>
<td>Ant Colonies ....................................................................................................................... 24</td>
</tr>
<tr>
<td>Aspects of Information Processing .................................................................................... 25</td>
</tr>
<tr>
<td>Complex Systems .............................................................................................................. 27</td>
</tr>
<tr>
<td>The Myth of Precision ...................................................................................................... 28</td>
</tr>
<tr>
<td>Emergence ......................................................................................................................... 30</td>
</tr>
</tbody>
</table>
Our Proposal for Incorporating Conceptual Linkages ........................................30
Embodiment and User Feedback ......................................................................32
Metadata .........................................................................................................32
3 DESIGN AND TESTING METHODOLOGY ..............................................34
  Breadcrumbs Computational Overview ..........................................................34
  Anchoring the Breadcrumbs Implementation with Starcat .............................35
  Future Phases ..............................................................................................41
    Phase 1 .......................................................................................................41
    Phase 2 .......................................................................................................42
    Phase 3 .......................................................................................................42
    Phase 4 .......................................................................................................42
REFERENCES ..................................................................................................43
APPENDIX
  GLOSSARY ..................................................................................................48
LIST OF FIGURES

PAGE

Figure 1. Cognitive processes in journalism are characterized.................................................6
ACKNOWLEDGEMENTS

I would like to thank Dr. Joseph Lewis for providing an engaging research idea and being my mentor and inspiration in times of need. A grateful thanks go to Dr. Amy Weiss and Lorie Hearn for showering me with trust and confidence and giving me experiences in the field of journalism that I could previously only imagine. I would also like to thank Dr. Kris Stewart for having the patience to pick out my ideas hidden among my ramblings and push it forth and for accepting to serve on my defense committee.
CHAPTER 1

THE BREADCRUMBS RECOMMENDATION SYSTEM

A TOOL FOR DATA SOURCE DISCOVERY

Breadcrumbs is a Recommendation System inspired by the emergent computation demonstrated by the Starcat framework developed by Dr. Joseph Lewis [1]. The objective of the Breadcrumbs system is to demonstrate at a proof of concept level that an emergent, distributed agent implementation of a recommendation system can deliver suggestions of artifacts or inventory items based on a network of concepts rather than via statistical methods to make decisions. This system comes under a nascent but promising computational paradigm usually denoted Complex Adaptive Systems, where simple agents with limited information and protocols interact with other agents and the environment they are working in [2]. Here, individual behavior among the components in the system aggregate to collective complex behavior. A few terms referenced in this report have different colloquial meanings. To minimise ambiguity, a glossary/dictionary has been provided in the Appendix that explains what definition is being referred to. It would be helpful to refer to that dictionary whenever in doubt.

One of the most significant contributions of this preliminary groundwork for a novel product of this type has to do with three important aspects of journalism, both generally and in particular in the face of Big Data—often a daunting challenge for journalists especially nascent students of the art. First and foremost journalists need an aid to the process of discovering leads, developing intuitions and identifying means of independent validation of information. Sometimes addressed at later points in the process is also the ever-present undeniable need to keep the discovered leads, validated facts and galvanizing intuitions close to the pulse of the community and people that the journalistic product serves. Mentioned in some more detail later, computer assisted reporting has a long history taking form in machine support for journalists even back in the 1950’s with mainframes providing the first reliable statistics about election results. Today we take for granted to some degree that the same term
applies to what a journalist does when they sit at a screen with keyboard, mouse and access to search engines (e.g. Google these days) and databases and such. But the point of all this computer assistance is to accelerate, amplify and invigorate the otherwise opaque novel connections the process of discovery of related ideas that embellish a journalist’s “lead”, that essence which defines the story they are working on. Another point then, once such discovery has been embellished as an aid to, in no way a replacement of, the journalists skills, is that the investigator will, presumably, have at their disposal a broader set of possible ways to reach out for independent validation; this point cannot be overemphasized as journalists must insist on the highest standard of certainty and precision when it comes to the data they present, the validation of its sources and the utility, even if expressed in personal opinion of what they offer the reader. So Breadcrumbs, for all the other things we will say that it is or does is, at core, a new possibly usefully different tool for lead and support discovery, validation seeking and connection with the community. It is a new tool to aid journalists in their craft in these two ways, in initial discovery and in subsequent work to substantiate the story and make it accessible, meaningful.

There is a third subtle but critical piece of what Breadcrumbs may be able to bring to the journalist user, and although it is generally useful it comes into play ever more clearly when the world of Big Data presents itself to the journalist, as is increasingly the case. Journalists, like almost all other users of traditional search engines such as Google, often have only the most basic understanding of the tool’s capabilities and the model by which it operates. Strictly statistical methods for identifying related documents pose limitations on the recommended connections that only certain “hybrid” informatics-journalists may appreciate—not to mention certain anomalous phenomena such as “sponsored links”. Users find themselves happy with a few hits but dissatisfied with so many more, often wasting time to discover that the connection is incidental or superfluous. They may attempt to modulate the search results but because the underlying algorithms often work very differently from our own conceptualization of the “how” and “why” topics should be associated, this quickly comes to feel fruitless. Many give up and return to other less modern or sophisticated techniques. The most important point to take from this is that users, journalists, actually lose confidence in their tools because the tools do not make their model of operation transparent nor do they have an intuitive interface that can draw out of the journalist exactly what they
need to be doing and allow them to discover on their own—to learn—even better how to use the tool. A very special subclass of hybrid journalists with informatics knowledge can overcome this; but we feel Breadcrumbs and the groundwork laid here for moving toward such a product for this needful community facing these tough issues can not only solve those technical problems with our novel computational approaches, but also help develop in the user a greater sense of confidence in the tool. We want them to think “once I felt helpless because the tool did not reveal enough to me about how to modulate it and its functioning does not match my own process, whereas a new tool such as this which offers me results based on the same kinds of connections I make mentally in my craft allows me to own the skill of modulating the tool and thereby trust it more.” Note this is NOT because the technology is better nor necessarily closer to a human and certainly in no way likely to replace their process, but it works with them rather than orthogonally to them.

AN OVERVIEW

Data has inundated many fields of work as a result of technological advances in storage capacity, computational efficiency, hardware technology and processing software. As a result, tremendous amount of information is now available to any interested party on any topic under the sun. However, this deluge of information has brought along with it the problem of a great deal of useless data which hides the required information that users forage for their specific purposes. The solution for this lies in establishing filtering mechanisms that help the user to pick the wanted from the unwanted information according to the desired user criteria.

When it comes to reading data though, the typical newsreader does not want to sift through data, he wants some analysis on the data and the numbers associated with it. For instance, without journalists analyzing the WikiLeaks logs and filtering out the provocative and poignant material, the logs may have not been read at all except for the rare information hound [3]. The regular reader sees the data but is also interested in why he wants to look at the data. This answering of questions related to relevance and interpretation is one of the main pillars of journalism. The idea of concepts and metadata about the raw information feed is very relevant in the field of Journalism along with the very important question of context.
Journalism is a domain where the ability to draw connections in data are difficult. This is because the journalists ask very open ended questions when confronted with a set of facts. The context, spatial and temporal characteristics of the data and even the source of data can have different interpretations depending on the content being scrutinized. The usual routes of statistics and stochastic analysis used to manipulate data to get patterns, anomalies or a sub set of relevant data is not always the perfect answer to a journalist for analysis. This is because the result set is bereft of an underlying reasoning that tells why that particular result set is relevant to the journalist.

The result set might be the answer derived out of a set of functions, methodologies or calculations based on a set of queries but they are usually precise answers to a standard question. The relevance or the chains connecting these sets is something that the journalist has to conjure up based on his experience and intuition in the field. The tools do not help him in this journalistic task, they might only give him a smaller or more relevant subset to analyze but the proper questions have to be asked by the journalist, they are not prompted by the computing mechanisms or systems.

These challenges in journalism interpreting data, coupled with the advent of Big Data disrupting journalistic processes so much so that data analytics is a need and not just a different way of doing journalism provide us with using Journalism as a very good case study. Enabling a recommendation system which provides context and meaning to the result sets obtained via a concept network process will be a new approach to heed the calls of the journalism community for tools to make sense and help in the workflow. Though there are numerous tools, techniques and software available to the journalist to procure, manage and present data to the user in an interesting way, there aren’t many that actually help the journalist build the story he wants which engages in the style of cognition that a journalist is used to when writing a story; namely answering contextual and conceptual questions that give definite leads and can be tested to withstand or break and justify the hypothesis that will lead to a story, or a different angle of approach than was conceived earlier.

Statistics is seen as the primary tool in dealing with such huge amounts of data or Big data. When the information being worked on is incomplete or has a lot of elements at play, statistical reasoning comes to play. Statistical reasoning uses ideas about data and chance and
infers and interprets statistical results. To do this correctly, the interpreter should be well versed with the ideas of distribution, uncertainty, spread, randomness and sampling [4].

Statistical arbitration or decision making encompasses the following characteristics:

- Critically assess data quality
- Correctly apply data for problem solving
- understand limits to generalizing and interpretation
- understand concepts of risk and probability
- When faced with incomplete data, to recognize that more material is needed to make decisions and to understand what information is missing [5]

Even with proper training imparted to deal with statistical methodologies and experience in dodging its pitfalls of ignoring the baseline, arbitrary comparisons, selection bias or other hurdles of judgment [6], data scientists stumble often. And this is with them recognizing the ability or inability of a statistical method to solve for a given problem at hand, its constraints, areas where the success rate is markedly better or worse and other related knowledge of the elements of statistical science.

Journalists perform a variety of cognitive tasks as they go about with living a life of a journalist that capacitates them to perform their work as shown in Figure 1 [7]. They include categorizing events and people as newsworthy, hypothesis or theory generation, selecting information to test the formulated hypothesis or theory and finally integrating the gathered information to bring out a story [8]. Though they make use of numerous tools to help them in their line of work, the tools or software used can answer only the questions asked and do not help in building a thread of thought that is vital to pursue and form news stories. That has to come from the journalist’s experience and training.

Statistical thinking is a non-human way of cognition. Human thinking or cognitive experience should be perceived as “a phenomena where emergent statistical effects of a large number of small, local, and distributed subcognitive events with no global executive” [9].

But what is emergence? Emergent phenomena are characterized by globally coordinated behavior arising out of the activities of many local agents acting on local information [10]. Many living systems like the immune system or ant colonies, to name a few, use this system of information processing to adapt and survive in their respective environments. The main

Point being that they do not deal with rigid definitions or input and output or even computation, in fact, what the systems perceive as information depicts the method of processing that has to be developed to solve the issues at hand and the communication required to solve for such problems. The whole aspect of processing becomes a constant activity of improvisation among the processing components and this interaction solves the problem and not the independent components themselves.

In a conventional single arbiter oriented Information processing system, like a single algorithm based unit, the decision making component or the arbiter has to have all the facts at hand to take any decision or keep on monitoring the facts at hand from time to time to make incremental decisions. If there is a central arbiter, looking up the variables for formulating a decision will be a constant activity along with monitoring the oncoming of new or changed variable states, the priorities that need to be given to different states of the process need to be constantly monitored, if there is external feedback, then the context of the problem that changes with the user feedback has to be taken into account, clearly a lot of pressures on that central arbiter.

With distributed agent based architecture, various sets of agents with definite simple protocols react and interact on information that they perceive or are local to them. Along with constraints that determine what they do with the information they see and what they leave behind and communication protocols that govern the type of interaction they have within
themselves, which has to be goal oriented, and with the concepts of emergence interlaced in such a complex agent environment, the only arbiter is the networked interaction and the pressures introduced into the system. This model also gives a chance for the development of “behavior” to arise from such a system since the agents are allowed to improvise and navigate among the problem space.

We intend to investigate the utility of this idea of using emergence with a distributed agent based model that will explore a network of concepts that are relevant to the journalistically trained user. The idea that this will build on the intuitive concept recognition aspect of the user and will evoke an emotional response of finding a thread of thought to explore data via this system is the eventual goal.

This thesis is an attempt to bring together three actors to play over a scenario which is of practical use and necessity in the form of a community asking for help over an issue and an approach that we hope will help in the roles of Journalism, Big Data and Emergence based Recommendation Systems.
CHAPTER 2

BACKGROUND RESEARCH ON PROBLEM SPACES

We live in an age of interpretation, where facts are not black and white but have shades of grey depending on the context they are formed in and the person who understands the meaning of the utterance or occurrence of the situation.

Compounded with the fact that we now have a barrage of information from various media sources and formats [11], the amount of filtering that has to be done to get at what one needs to read or report at is such an enormous task that many a time skewed results are generated where the reasoning behind the fact generated or report is neglected.

Data can also be skewed in the process of intake. Surveys, for example, can mislead people by asking questions which lead to false comparisons or vague questions. Or the surveyor may not be clear about what is being surveyed. Data can be misrepresented to show only one side of the results to the detriment of the truth seeker [12].

OF DATA AND BREATH

Though the rise of data may seem as a recent phenomenon, its presence has always been there. Over the years, the production, accumulation and usage of data seemingly mirror the same phases as that of oxygen in the earth’s atmosphere; this gives us a good analogy to explain the history of data.

Oxygen was initially produced in small quantities by cyanobacteria as a waste product but with so much of the bacteria around that it accumulated to quantities to be reckoned with. But it was captured by minerals like limestone, iron, etc, which was later oxidized and let the oxygen get into the atmosphere.

This led to the start of the creation of the ozone layer which helped organisms survive the ultraviolet radiation from the sun by shielding it. This was very helpful to later organisms like plants to develop and release much more oxygen via photosynthesis into the air enough for organisms to grow in size and capacity which depend on oxygen for survival [13]. So,
eventually, we have a scenario where oxygen is needed for survival of life, and it started as a waste product of life!

Come to think of it, data has had a very similar evolution since at least the early decades of post industrial age times. You could argue that data accumulation started from the Library of Alexandria but the turbulent period of dark ages and tumult that occurred during and after its burning and general chaos that followed put an end to any linearity in that trend of data gathering even if pockets of knowledge existed, like the church. So we feel confident of speaking about the timeframe of a few decades after the invention of the printing press. Initially came the cyanobacteria equivalents, the academic, governmental or military researches on various subjects – science, sociology, demographics, etc. The measurement figures, data and conclusions reached were distributed only among peers or locked away in lockers named “CONFIDENTIAL”. Like the bacteria which let off oxygen but were captured by the minerals in the vicinity, lots of data were being collected but it was retained by governments or corporations or academia that did not have very many ideas what to do with it other than shelve it and the general public could not even have a sniff at it.

Then came a series of events similar to the oxidizing events that released the oxygen into the atmosphere; military technology permeated into civilian areas, the thought processes of governments evolved to one of embracing the public in decision making, and that needed them to know facts, hence data had to be released. Technology improved to give alternatives to storing data in formats other than ledgers and cabinets and in places other than huge warehouses, so access to information was physically eased and departments could be set up by the data hoarders which dealt only with information access.

Technological access and changing mindsets, along with the invention and gradual spread of internet, and information feeding agencies provided a data network very much like the ozone layer built on oxygen. And like the ozone layer, this data layer which was becoming more easily accessible also became more and more relevant and a necessity to the data consuming organizations and people. As the concept of active citizenship and open governance was developed, along with science developments consuming and releasing more data than ever before, this network of data became utilized more and also more data was fed into it. This data layer was not something that could be ignored as an interesting object out there in the world but a vital part of life for its betterment, protection and decision making.
Like how the ozone layer protects the organisms from obliteration by UV radiation and where holes in it can cause devastation, so too is data now a protecting shield and holes in data, a cause for more than slight discomfort.

So we have a scenario where data was something generated as waste as a consequence of other activities but now activities are generated as a consequence of data!

**Signal and Noise**

If the climb in the availability and usage of data as seen in the timeline above is a steady growth, the amount of its dispersion and absorption in recent years can only be called as an explosion. Data, can be broadly bifurcated as traditional data containing documents, finances, stock records, personnel files and Big Data, the term associated with photographs, audio and video, 3d modeling, simulation, location data [11]. The problem with the latter is that we are now dealing with unstructured data sources that cannot be categorized or recognized as easily by traditional techniques. And now because you cannot handle the amount of data flowing through your systems, you are letting slip away valuable data that might give better insights into handling and solving problems that you are interested in. In other words, there is too much noise for you to identify the signal.

Consider this nugget, when the Sloan Digital Sky Survey (SDSS) began operating in 2000, it gathered more data in its first few weeks than all data collected in the history of astronomy [14]. So, big data is not just a fancy name thrown around but a challenge. Big Data is defined as high-volume, -velocity and -variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making [15]. Much has been written about the 3V- volume, velocity, variety characteristics of big data and that trinity is now accepted as the identity of Big Data. But considering the fact that an organization or a group of people have the capacity to store high volumes of data, can streamline the flow rate and have a handle over controlling and managing the rate of flow of data into a system and recognize the categories that the data fall into and make way for data to sort itself into buckets to rest in once it is inside the system, what do you do with it?

Data of this magnitude and this diversity now starts breaking the boundaries of well defined traditional statistical approaches to analyze the data. When the size of the data itself becomes a problem as in Big Data [16], using a statistical approach to find all your answers
leaves gaps and glaring holes in this unstructured swamp where well defined analytical approaches can break down quite easily. Either the rules cannot accommodate the massive set, or many elements of the data are thrown into overlapping sets of identity where the traditional statistical techniques might not work for all elements of the set, or there may be new radical data elements that they don’t accommodate for. It may be quite logical and rational to confidently say that computationally aided pattern recognition, goodness of fit functions or clustering model methodologies are now good enough to aid in our analytical issues to replace the traditional methods of thinking as well as to solve problems in the fields of sciences like physics, medicine or even economics; but data has no voice of its own to help you select the elements within it to run these techniques on in the first place [17].

Also, the biggest enemy of statistics based analysis and subsequent judgment or misjudgment arises out of the fact that correlation does not mean causation. This may throw results and interpretations that are just not true but seem very obvious due to the overwhelming numbers in the result. The context of the signal is lost in the noise of the content.

There is a reason why the signal to noise metaphor has been repeated here, once in the context of the amount of information available and again in the context of the results obtained after analysis of the procured information. And the reason is to acknowledge the fact that to get some sense out of the menagerie of data out there and again to make sense of what the results of the analysis are trying to say, a vital component is necessary. And that component is the human operator. In a very insightful essay [17], under the heading how to discover theories, Emanuel Derman (Previous quant at Goldman Sachs and current director of the Financial Engineering program at Columbia University) speaks about how intuition holds the key to realization. The interaction between the information spread before the observer and the observer himself becomes a series of discoveries and consequences that will eventually culminate in a solution. He mainly makes a point about making a judgment of what it is about the data you want to work on. In his words, you can only get out of data what you perceive of it. No matter how sophisticated the tools used to operate the data, the operator should know on what data to use the tools on. The tools are now so good that they will get you an answer, whether you use it on the signal or noise in the data. Hence more is the need for the filtration process to be a task that the tools perform along with the user.
**BIG DATA DOMAINS**

Big data being the opportunity and problem that it is, and recommendation being the solution that we look forward to, let us look at some domains where both are major factors in workflow.

Now that we have had a taste of data and the problems and benefits it brings to the table, what domains of industry and knowledge is Big Data a major factor? The main utility of Big Data is in prediction. In retail outlets, recommendation systems are used to predict your taste and interests, in stock markets and demographic arenas Big Data is used to predict trends, e-commerce uses it to predict demand and supply, meteorological departments use it to predict weather, freakish or otherwise and scientific endeavors use it to predict causes and consequences, to name a few. The main idea being that data, and usually a lot of it, is used for accurate predictions. You have now had Big Data that helped predict whom you will cast your vote to in an election even [18]. But what about interpretation? In medicine, big data can be used to interpret your symptoms into disease conditions, genomics use big data to analyze the gene sequences better [19], semantic systems help build domain knowledge for machine learning systems to improve using expert classifier systems and in journalism, Big Data is seemingly the 800 lb gorilla in the newsroom.

**JOURNALISM AND BIG DATA**

In spite of the hue and cry that attaches itself to any industry that has Big data as a part of its label, big data in journalism is just data journalism which again is just journalism. The only difference being that, apart from adapting the “Big” buzzword as a label, people are recognizing the part that data is playing in the act of journalism. When you combine the traditional reporting and investigative abilities of journalists to spell stories along with the scale and range of digitized information available, new avenues develop [20]. And such opportunities can come at any stage of a journalist’s work flow; may it be using programming skills to automate the data gathering process or using software and computation methods to find the connections to make the story between the gathered data [20].

The flood of data into the sphere of journalism and reporting has been made possible with the phenomenon of open governance being observed across the globe. With pledges of
releasing data to the public for scrutiny, countries are adding to the idea of open governance and the open data movement is picking up pace with data being collected on governmental agencies by news agencies like the Guardian[21] and the New York Times [22]. The idea of open data is defined as the concept that certain data are available for use and republishing by anyone without copyright constraints, patents or other mechanisms of control [23]. Accessing these kinds of data and using them to provide empirical evidence to report stories that empower the society is a new way of broadcasting the responsibility of the media [24].

Journalism is also a means of introducing the readers and the society about new sources of data that they can look into. The Wikileaks phenomenon where thousands of confidential diplomatic data cables were leaked to the public surely could not have collected more than a handful of readership if not for the rigorous analysis done by journalists who then wrote stories on them. Just dumping the raw data of those cables would not have shook the public’s mind as much as critical analysis and discussion by the media did. A significant part of data journalism, due to the abundance of data has led to the act of data journalism becoming data curation [25].

Also, unstructured and nontraditional data is being actively sought after to analyze by the journalistic community. There are more stories just as likely (if not more likely) to come from the unstructured information that comes from documents, audio and video, tweets, other social media — from government and non-government sources [26]. Open source, may it be tools or data sources are shaping journalism. They are eliminating the lack of resources and technical barriers that is pre requisite in dealing with data. All you need to know is learning to use and recognize them, and even learning is free nowadays. Data journalism is now a mix of two different factions of journalists. One who know the skills of extracting data by recognizing sources, creating sources, feeding off APIs and generally bothered about pushing the data out to the reader’s domain. The data can be unfiltered or a bit easier than what the government reports but at least it’s out there and accessible. The others are bothered about reporting, constructing stories, presentation, making the passive reader also knowledgeable with visualization techniques and getting their stories online in more interesting ways [26].

Data Journalism, as described by Steve Doig, Arizona State University journalism professor and data expert, is just another way of gathering information. Earlier journalists on the beat had to interview sources and look up documents but with data journalism, they have
to interview the data for getting their scoops [27]. Maybe that 800 lb gorilla has always been there, since it was an 8 lb kid, in the newsroom.

**RECOMMENDERS**

A unique way of sorting and filtering are done by recommendation systems nowadays. The usual context in which the concept of a recommendation system is heard of is in online retail shopping experiences. Based on your search keywords, purchased items, tagged items present in your shopping cart, items you have browsed through, even your location and time of the year, the system can make educated guesses about the range of products you are looking for and suggest other items that people making similar purchases have looked at or bought. Recommendation is a complicated science because it does not deal with precision in terms of input data or the output it comes up with. Due to the diversity in input that the algorithms deal with, very accurate guesses are the results and not precise recommendations in most cases, especially if the end user who benefits from the recommendation is a human, since humans can claim to have varied tastes and satisfaction unlike precision based machines.

Approaches to solve problems that have recommendations as solutions usually come under traditional collaborative filtering, cluster models or search based methods. There are many variations of the above but we can call these as the super class of those methodologies [28]. In traditional collaborative filtering and cluster models, the algorithms work by finding sets of users who overlap the same purchases and rated items as other users [29]. They mainly work to find other customers similar to the user. The collaborative filtering mechanism measures the similarity of customers by measuring the cosine of the angle between two vectors, say customer A and B[30]. Cluster models work by assigning customers similar to segments present in an already segmented customer base. The segments can be created with another cluster algorithm or manually built. This is possible with greedy clustering in very large data sets or sampling as clustering is a problem in big data sets. Users can also be classified into multiple segments and the strength of each relationship can be noted [31].

In search or content based methods, recommendations are converted to search queries for wanted or related items [32]. But here again, the larger the database to search for, the
more items can be missed and the larger the user’s previous purchases to guide the algorithm, the more impractical it becomes to make a query based on all the history of the user. Since the history of their purchases, even for a single domain, may be very diverse. The algorithm must then compromise using subsets of the database and thus compromise in quality of the results. [28]. Content based recommendation systems carry out suggestions based on probabilistic calculations or naïve Bayesian models (Joeseph Lewis, personal communication, November 2012). But the main problem with recommender algorithms remains that, all kinds of input diversity have to be fed in or the algorithm will struggle with outlier inputs. Scalability issues, sampling and reducing dimensionality problems also plague these algorithms [33].

**SEARCH ENGINES AND DATA**

Coming from all the data out there in journalism, the main source of data is still via the internet. Government sites, open data collection hubs online, fellow journalists’ blogging about good sources of information, new articles across the globe being published online are but a few varieties of news available over the web. Journalists use online resources for background research for interviews, to find or identify sources, to check or verify facts, to read their competition, to become informed about current events, and to identify story ideas [34, 35, 36]. A few big news organizations may have their own local servers and information warehouses which keep updated but a majority of the source of Big Data, any data, remains the internet. The conventional way to make sense of this data available is via search engines to narrow down these sources to what you are interested in.

When confronted with a ton of complex information and having to find authoritative sources, web users usually declare that there is nothing about the subject on the internet [37] . To help them out, search engines help in weeding out the authenticated and required source material from the referred or sponsored links and they do so with various tools. Advanced search options include applying Boolean logic to combine keywords, adding constraints to dates, searching by trusted domains, searching by file type and even the location of the content in the web page. All this is to get at authenticated sources identified by site stability, currency, usability, searchability, listing and usage fees and link descriptions as indicators of quality search and content sites [38].
**GOOGLE AND PAGE RANK**

The phrase Google it! Has now replaced the word search, in day to day communication. In fact, these two words tell people where to look for the answer along with asking them to look for the answer in one phrase itself! Let us have a look at the mechanism with which Google pops out answers for all the questions we type into the single text bar in the site Google.com. Web crawlers, little pieces of code that look up web pages and follow the links in them, gather data to Google's servers. Not all pages are crawled and hence information from all pages do not get stored in Google's servers [39]. There exist rules and constraints to collect selective information, and some pages can’t be crawled upon if the owners do not permit them (or pay walls would not work). But since the crawling happens all over the internet, the selective information is quite a lot.

After collecting the information, though updating is a continuous process, the information about the web pages in indexed for easy searching and look up. The different categories of information in the indices pertaining to media type, location, weight of content, location and even recently, semantic content help in giving appropriate results to search keys. note that all this information is gathered from the content in the web pages themselves.

The content in the servers are then used by the various algorithms that Google uses, including the PageRank [40] algorithm. Most algorithms used to tweak the results are data driven in this context [41]. Nowadays, the PageRank algorithm may play only a modest role along with a host of other algorithms in evaluating web pages to be on the result list but it is still a good place to look at the backbone of Google's logic.

The PageRank uses the academic citation method of valuing a page. Mainly, the value of a page increases by counting the citations or back links to the page [40]. Of course, the value does not increase linearly or all pages would link to themselves and rise in the ranking. The score in PageRank of other pages that link to the page in question also add to the score of the page.

From the original paper of the creators of Google, PageRank is defined as follows:

We assume page A has pages T1...Tn which point to it (i.e., are citations). The parameter d is a damping factor which can be set between 0 and 1. We usually set d to 0.85. There are more details about d in the next section. Also C(A) is defined as the number of links going out of page A. The PageRank of a page A is given as follows:
PR(A) = (1-d) + d (PR(T1)/C(T1) + ... + PR(Tn)/C(Tn))

Note that the PageRanks form a probability distribution over web pages, so the sum of all web pages' PageRanks will be one. [40]

Where

- PR(Tn) - PageRank of the nth page on the web. Starting from the 1st page T1 to the nth page Tn
- C(Tn) - The count of back links from the page. For page 1 it is C(1)
- d - damping factor of 0.85; this is to prevent one or more pages from having too much influence.
- (1-d) - With this correction, sum of all the page ranks will add up to 1 and the page with no back links will have a score of 0.15 at least [42]

The critical difference when the PageRank algorithm came out compared to other search engines was the use of back links to give scores to the individual pages. Thus each page was dependent on other well ranked or less ranked pages to score and it was a control mechanism so that a page with relevant keywords plastered across it could not shoot up the result list. It also focused on the metadata content in the pages to increase relevance and also identify content which is non textual.

**TRAVAILS OF THE RESULT LIST**

Now that the journalist, in requirement of an assignment has identified elements required to search, has identified a database source or search engine to search through to find information and obtained a result list, there are both comforts and pitfalls in dealing with the contents of that result list. Some of the most appealing reasons to turn to search engines and the internet to collect data is the depth of the information and speed with which it can be collected [43]. But to deal with that much information available that fast, the journalist needs to acquire analytical skills and higher levels of computer literacy knowledge than his counterpart of a decade ago [44]. The other skill set required is judging and trusting the quality of online resources and databases [45]. Even government databases have errors and flaws, as experienced users have observed [46, 47, 48]. And for journalists working on a deadline, swimming through the enormous swamplands of the internet to get information is a barrier [49] and technophobia is also a concern [50, 51]. Among the elements in the result list, linked and chat room content, credibility of the content, signs of privacy invasion and separating advertised content and editorial viewpoints are also of concern to the journalist.
But in spite of all these hurdles, online resources are helpful in background stories, locating hard to find information and records, sources and extending government coverage.

**DATABASES AND QUERYING**

Another source of data for journalists has been databases, those available freely and those which involve fees for usage. As use of database-oriented news research grew, so did the amount of information available [53, 54]. But even this has not been embraced wholeheartedly; even though databases provide yet more new sources of information. Some journalists have felt that the user interface to manage querying databases have been too cumbersome and chose to ignore computer based research. For some, online searches lengthened the time committed to research or decreased the local perspective of a story. Online searches and databases have also been blamed for increased errors and for discouraging original work [55, 56, 57]. Indeed, the skill sets required to utilize these resources when they came on the scene were not usual among the news teams. Some news managers were deciding whom to assign online search to. Would it go to anyone in the staff with the skill, the reporter doing the story, the supervising editor, a designated online news editor or to news researchers [43]. But to assign a person to use such skills has been necessary because online research gives a higher quality of context – economic, scientific, legal or social to stories [58].

When it comes to databases, online or news agency owned, they are under a domain of structured resources at the disposal of the news team. Querying a structured artifact is easier if you have at your disposal a tool that is designed to interact with such a specific need. And so, enter the Structured Query Language or SQL. Even wizards in excel sheets will need a relational database manager to merge spread sheets or query huge data sets. SQL gives the user the ability to be very specific about the information type and tags involved, including timestamps and with the inclusion of Postgres, geo tag structures. With structured commands from the query language, one needs to know only a few basic key words and some grammar to start sorting datasets, joining datasets, creating subsets and performing CRUD (Create, read, update delete) functions on structured datasets. The additional advantage with these commands is that they can be saved as a script and reused as many times as needed by simply
changing the dataset names being operated upon. This also leads to good documentation features available so that the scripts can be recycled among the news room personnel for any data sets. SQL is very good at generating reports and performing aggregate statistical functions that are simple to use over small to intermediate data sets. For performing multiple search conditions, something like select a name from department X and city Y, SQL queries provide a very easy interface to compose precise search conditions over datasets. SQL is also compatible with many third party tools and languages and provides a versatile toolkit for journalists to explore.

**STATISTICS AND CONTEXTS**

Weather it is via algorithms used by search engines to show relevant data, recommendation algorithms trying to categorize the user themselves or data searched by the user or calculations of pattern matching and probability used by the user to manipulate data, it is by using statistical calculations and stochastic approaches that sense is being made out of the data. This urge to do the sense deriviation, so to speak, either by a data scientist or a data journalist, or a plain scientist and journalist confronted by a lot of data has a single drive across professions. And that is to tell a story using the characteristics of data as elements. But a strong desire to tell a story using only numbers and not the context of the situation can be very misleading [59]. A good example of this particular pitfall was seen in the Google Flu Trends experiment carried out by Google [60]. Google Flu Trends defines itself as usage of “aggregated Google search data to estimate current flu activity around the world in near real-time” [60]. The way it works is that it monitors words typed into the google search bar by users along with flu trend information captured by the Centers for Disease Control and Prevention (CDC). It then identifies a sub set of search queries that are popular when actual flu season happens. By tracking the identified search queries passing through its servers in real time, Google can, in theory predict flu trends faster than the CDC can gather data and give an estimate. Sounds amazing unless you check the premise that the theory is based on, which is that those searched “flu” queries are searched only when the flu is around the person typing that query. With news reports educating far more people than before, what happened was that people started hearing about the flu season much before the spread of it among them and started typing lots of queries into Google to learn more about it even if they were not ill
[61]. These often typed queries also included the “flu” query subset even though the people typing it did not have the flu but how are Google’s servers and the calculating algorithms to know about this particular context of the situation? They took in the numbers anyway and churned out a trend that reflected that the flu virus had spread double, 11% the estimates that the CDC put out, 6% for 2013 [59]. Other times, the Flu Trends has underestimated a H1N1 outbreak in 2009 and had to tweak its algorithms [62].

Till recent times, journalists have always practiced their profession with their own intuition and knowledge to guide them and this cannot be discarded just because the data says otherwise. One of the pitfalls of a lot of data or Big Data is that there might be many variables at play that you discover in a situation but you also might have too little data per variable [63]. With such a menu before you, you can get data to match your theory, any theory and stop as soon as you have collected enough data to match your theory, and fail to report or even notice that there is data that does not agree with your conclusion. This is another instance of too much noise, willful ignorance of signal or not even knowing that there is other set of data that does not agree with your model. What to choose here then? The story or the statistic? “The focus of stories is on individuals rather than averages or sets, on motives rather than movements or deviations, on context rather than raw data” [64]. And here lies the dilemma, how do use the data to report on a story rather than make the data the story? Data journalism is still about the story so the instincts that the journalist has cultivated to write good stories should not be thrown away at the first sight of data; in fact, they are more relevant now than ever.

It isn’t the case that computing and number crunching are new found discoveries in Journalism. Computer Assisted Reporting – CAR, has been a companion of journalism ever since desktop computers were seen as more than a personal luxury and recognized for their computing advantages. But CAR has been usually associated with spreadsheets, stats, maps and a sprinkle of SQL for backing up the validity of stories. Now it is the opposite role that computing devices perform, that of sniffing at and finding stories via stats, maps and effective scripts, exhausting the usage of Excel and in turn inventing new tools to make sense of the data inundation. But all this is still for the same cause, stories [64].

From the words of John Allen Paulos, Professor of mathematics at Temple University, “Stories are open-ended and metaphorical rather than determinate and literal”
Journalists like to answer and need to answer very open questions. The five Ws (What, When, Why, Where, Who and How) still govern the main facets of reporting regardless of the changing labels of journalism, may it be investigative, explanatory or data journalism. Many times the answers to these questions are not definitive or precise. The answers might lead to a sense of understanding to a user or might open up new questions. With the advent of putting data as part of the journalistic equation, it does not change these open ended and inviting questions and answers to a bunch of precise figures or percentages. In fact good stories that make use of extensive amounts of data utilize data to lay a stable premise for conclusions reached, use data to be clear about the impact of the elements and factors in the story and help to establish how the news being reported hits the user at a more personal level by being fine grained than reading off a headlines detachedly. If statistics and other number crunching methodologies are used to work on data purely to get a set of figures that is more “accurate” rather than to tell the background domain of the figures and to weave a good story which helps in discovering new angles of perception, it is not going to be more than a stock market report sheet.

And the angle with which you view any set of information is vital to what you perceive the information to contain. There will be little packets of information that may not seem relevant at one instance but very important when perceived from the right angle [20]. And all this data wrangling is required to either find that angle or find facts to support an angle that you already have. Or, discover facts that do not support your angle and, not discarding them, change your theory to the facts, what can only be described as doing good science and in the same breath, good journalism.

**Recipe for a Story**

Now that we have touched upon the tools that are used and are necessary to make help new stories or data stories as news is packaged data in some sense, let us look at the process of building a story. In the newsgathering stage, the aim is to gather as much information and facts as possible. This may consist of interviewing sources, looking up information online, going through various paper's archives to look at past stories on an issue, compiling necessary documents or reports, mainly foraging for as much information as possible about an issue at hand. There may also be instances where the data is plonked in
front of a journalist with no idea about what they will find in it beforehand, something like the Wikileaks reports, but the journalist knows that there are nuggets buried in the heap.

Then the reporter has to weave these pieces together to start to make a story, from the most important to the least important, a technique called the inverted pyramid style. If they are weaving in multiple sources - people, documents, reports or data - they have to attribute each in the story and clearly show how this source information validates or substantiates a point they are trying to make.

There are a lot of skills and know how being integrated here, organization, being detail oriented, recognizing relevant information, discerning the information that can be used and cannot be. And finally the able to use all these bits and pieces of information come together in a seamless way. And all this is not just for traditional sources, when it comes to data, don't just worship it but look at it skeptically, considering your sources, reviewing it carefully - making sure that it is accurate and correct - interviewing the data per se and then running analyses on the data to showcase the patterns and trends found.

**THE NEWSMAKING PROCESS**

At a glance, these phases of news making may appear as ticking off tasks in sequence but in fact, they are assortments of tasks that make sense and get structure over time. The tasks of knowing what to interview, whom, what information to look up, consists of lots of cognitive processes and they are constantly boiling and competing with each other and different concepts which finally make a meaningful thread.

Along with content management systems for organization and presentation, search engines and statistical tools for searching and numerical analysis, there exist computational tools and software for helping the journalistic user with the separate tasks. But when it comes to integration of all the results and forging a path where the results and analysis can make sense with relevance to the context at hand, it is the responsibility of the journalist to use his own intuition and experience to make sense of the facts and interpretations. So eventually, it is the complex working of the brain that has to make sense, no matter what tools are used to provide the information for processing, the actual processing is guided by intuition, experience, interpretations and other ambiguous behavior.
COMPUTATION IN LIVING SYSTEMS

Computation, for the layman, has a reputation to be a precision driven mechanism. A fixed set of rules, an algorithm that anticipates the input and knows how to get the output, a structured standardized format or language to feed the data and request the processor for compiling the results as per the rules given are the foundations on which the definition of computation is understood. Computation or the field of computer science has the words testing, debugging, documentation and maintenance of software as tedious mechanical activities but precise ones in the context in which we are speaking [65]. But there exist other, imprecise forms and definitions of computation too.

Living systems perform a lot of computation and their architecture and processes to do so are a lot different from the computing interfaces that are considered the norm for such purposes. Any information processing is but a form of computation and there are living systems which need data crunching for survival.

In traditional computers defined by Turing machines and von Neumann - style computers, information processing involves having straightforward answers. Programs are written at machine code level and programming level (human understandable). The meaning of the input and output is interpreted by humans as is needed.

But how about living systems, what type of information processing happens in them? What computations do living systems actually need? Let us have a quick look at some living systems. With the help of Melanie Mitchell's narration of naturally occurring information processing systems, we would like to narrate the similarities and tenets of computation as seen in living systems.

**THE IMMUNE SYSTEM**

Of the many cells in the immune systems, the focus here is on the lymphocytes. They are created in the bone marrow and are of two types, B cells which release antibodies to fight foreign pathogens and T cells which regulate response of other cells and also kill foreign invaders.

Each cell in the body has receptor molecules which are used for binding with surface molecules of other cells. A lymphocyte's surface also has identical receptors that bind to specific ranges of molecular shapes. If the lymphocyte encounters a pathogen whose shape
fits its receptors, that pathogen is called as an antigen and the lymphocyte has recognized that antigen.

The problem with foreign pathogens is that they are so numerous that it is not possible to produce enough lymphocytes with correct receptors to cover their range; the diversity is astronomical to counter.

The immune system counters this problem by producing very diverse lymphocytes such that each individual lymphocyte can handle a range of shapes that are unlike its neighbor. A lot of randomness is used to obtain this result. With this shuffling process, it will only be a matter of time that a pathogen entering the body will get recognized by a lymphocyte set which is similar to its receptors. Once a binding takes place, even if it’s weak, the immune system has to identify the threat level of the bound pathogen since activating a widespread attack in the body costs the body resources and can even harm the body.

B and T cells jointly work to take the decision to launch an immune system response or not. If, in the same time frame, the number of receptors to the B cell exceed a threshold and the T cells give a go ahead signal (via signaling molecules called cytokines), the B cell gets activated. This means that the immune system recognizes a threat to the body and the B cells reach lymph nodes where they undergo rapid division to daughter cells with mutations that alter the copy receptor shapes. These copies are tested on antigens brought to the lymph nodes and the unbinding ones die.

The surviving copies enter the blood stream and bind vigorously to the antigens; the better binding ones make more mutated copies of themselves in other lymph nodes. This is a process of natural selection in which better antigen binding b cells are evolved. Thus B cells get designed to attack the recognized antigen.

**ANT COLONIES**

Information processing in ant colonies involves optimally and adaptively foraging for food and the ability to allocate ants to the diverse tasks needed for the maintenance and growth of the colony. These behaviors are accomplished without a central arbiter or controller making the decisions.
In foraging, the main communicative factor in use are pheromones deposited by ants, a chemical trail that varies in intensity depending on the reinforcement it gets from further deposits or is diminished from evaporation. The greater the concentration of the trail, the more likely will an ant follows it. So the concentrations of the trails are temporal representations of food availability discovered by the foraging ants.

Task allocation can be divided into any of foraging, nest maintenance, patrolling and refuse sorting (This is from ecologist Deborah Gordon's study of Red Harvester ant colonies). The number of workers allotting themselves to any of these tasks is a function of the activity intensity that is read and interpreted by the ants. The ants switch tasks based on which pressing activity they encounter in the environment and the rate at which other ants they encounter other ants performing those tasks. Also via direct communication by antennae and other pheromone trails specific to the tasks, ants are able to gauge what tasks their neighbors are engaged in.

**ASPECTS OF INFORMATION PROCESSING**

Now that we have given a brief idea of living systems involved in information processing, let us look at components of these systems relevant to us:

- **Information processors:** Processors in the above contexts are not individual cells or ants but collective actions of groups of these elements.
- **Information:** Unlike traditional systems, information is not located in any specific place nor fed via specified slots that deal with input and output but takes the form of statistical numbers and dynamic patterns that play over the system components.

In the immune system, the distribution and concentration of the lymphocytes over time and space in the body is interpreted as information about the dynamically changing pathogen concentrations in the body. In ant colonies, the distribution of ants observed statistically over various trails give food information at any point in time and space. The various numbers of ants distributed over different tasks give information about the colony's overall state at any given time.

**Processing:**

- **Sampling:** Since the information is perceived as temporal and spatial patterns of low lying components, individual components like a particular ant or lymphocyte has no idea about the overall state of the system or big picture. The information is mainly communicated via sampling over space and time.
In the immune system, lymphocytes sample their ambient environment via receptors for antigen concentrations. The decision to stay dormant or get active is governed by the cytokine concentration in the system and their sampling guides the pathogen killers to specific regions in the body. In ant colonies, individual ants use sampling of the pheromones and their concentrations to base its decisions on. By random encounters with other ants and this kind of sampling, ants allot themselves to other tasks if inactive or if the tasks need more workers.

- Randomness: Given the diverse dispersion of information and the processors, a certain amount of randomness or uncertainty has to be introduced into the system. Probabilities are the quantifications of analysis of the actions of such systems, not precise numbers.

The shapes of the receptors in lymphocytes have the randomness component to allow for a range of possible shapes to latch onto. The distribution of lymphocytes in the body has to be random to allow for sampling of various pathogens present in the body. The activation thresholds, division rates and mutations in daughter cells all involve randomness. And in the world of ants, foraging ants encountering food in the environment, switching tasks and following pheromone trails are random and probabilistic.

- Granularity: These systems consist of many simple elements that work together in parallel. The advantage of this is that many venues towards solving the issue at hand can be pursued simultaneously and resources can be allotted or deallotted by considering the success or failure of those venues. Continuous reassessment happens about the different explorations [9].

In the immune system, to solve the huge problem of which shape to explore in the countless shapes of receptors that the pathogens can be in, each of the lymphocytes have a range of shapes to explore themselves. The shapes that are successful are given more resources in the form of more mutations for their daughter cells than the shapes which are less successful. However, even with successful shapes being encouraged, novel shapes are generated thus exploring further potential successful receptor shapes. In ant colonies, many ants initially explore many directions for foraging food. If food is discovered, then by the process of pheromone trail concentration, more ants are diverted towards the successful direction. However other paths are explored as well even when a food source is found so as not to remove better sources from being discovered. With this simultaneous encouraging and discouraging forces, the systems perform focusing and unfocusing techniques directed
towards the system's needs. Thus the system explores and exploits information to adapt successfully and intelligently.

**Complex Systems**

The above scenarios of living systems have an underlying theme of multiple small cogs working together to form a bigger system. We would like to define this phenomenon in reverse, so to speak, by naming their properties.

- **Collective (complex) behavior**: These systems consist of large networks of individual components that together, produce complex, not easily predictable behavior by following simple rules with no central controller.

- **Signaling and processing**: These systems use and produce both information and signals in their internal and external operating environments.

- **Adaptation**: Via processes of learning, these systems adapt to survive and grow in their respective environments [9].

These naturally occurring living systems which engage in information processing, which seem to have many parts yet function as a whole, are termed as complex systems. Melanie Mitchell (Portland State University and emeritus faculty of the Santa Fe Institute) defines a complex system as “a system in which large networks of components with no central control and simple rules of operation give rise to complex collective behavior, sophisticated information processing, and adaptation via learning or evolution” [9].

Complexity as seen in these systems is not chaos, but not order either. It is at the edge of chaos, and the system must not be in equilibrium because if it is, the pushing and prodding, the pressures that drive the interpretation of information and processing will be less at work. There has to be a characteristic of constant turmoil not spilling into utter chaos that contains the signature of complexity in complex systems.

From the above examples, we can posture that complex system can be seen as a non linear system. When we say that the sum of the whole is more than that of the parts, we mean that the result of the interactions between the components is more than their individual contributions [66]. Also, models built to predict behavior of such interactions cannot be precise since they are studying the dynamics of a constantly evolving and adapting system that both changes the environment and changes with the environment which is the problem space.
THE MYTH OF PRECISION

It is a bit of a consternation to title a section that so challenges at least three sacred beliefs in the worlds of the communities interested in this work. Precision is no myth to a journalist. As mentioned in the introduction, it is something to which each journalist makes a vital commitment to preserve in the greatest fidelity, through careful discovery, unrelenting validation of information sources and so on. It may be that we live in a fuzzy, gray rather than black-and-white world that invites the word myth to be useful with regards to precision, but no journalist would consider the effort to pierce that fuzziness with the greatest possible precision a myth. To a computer scientist thinking only in terms of the evolution of our machinery and its capabilities, of course, precision is one thing that has improved by orders of magnitude decade by decade—from the number of bits representing the precision of numerical representation to the size and speed of memory and therefore the effectiveness and indeed precision of the algorithms that run on them. Lastly, and mostly historically profound for computer scientists, there is the adage that computers are only able to produce results of quality matched by the precision, correctness, fidelity, etc. of the information they operate on. This is true, and also in more modern artificial intelligence and complex systems endeavors, the combinatorics (for example genomics) are so huge that uncertainty or lack of precision becomes an issue as a result of the sheer magnitude of mostly useless information that must be sifted for that of value. All three of these could be entailed in one’s response to the “myth of precision”. We are here concerned with a version of the third one. Human input to systems such as we describe is itself blurry and overlapping and imprecise and defines its meaning from relationship to a world ever-changing in social, economic, cultural, political, linguistic, technological and even biological capacities. Because of this it is a myth to think that just because a new tool is based in a powerful computational paradigm it will necessarily be more precise than before or more precise than a human. The two most important lessons from this inquiry are first to look for ways to bound the uncertainty or lack of precision through metadata constraints that eliminate many branches of related ideas and secondly to leverage the liability as an asset. By that we mean to use the uncertainty, the apparent randomness and the lack of precision at times as a means of introducing novelty that may not have ever been discovered. This is not a new concept in AI and in fact is the basis of the
undeniable success of evolution. And it puts us back on the map of our goal, to aid in the
discovery of even novel connections for the users of a tool suggesting connections of interest.

The problem space presented here is one of cognitive recommendation, not just data
recommendation. For a data handling system that has to be useful to a journalist so that he
can get leads out of the data present in the system, the system needs to understand the
concepts that the journalist is searching to match, not just the syntax of the keywords for the
system to work on. Because complex systems have features of deep interaction and
imprecision as a part of their mechanism, algorithmic methods, which require a unique
output to be delivered for each input set themselves up for failure [66]. One cannot take into
account all the factors and their interaction and map them in precise pathways to predict
possible behaviors but only guess at it, many times wrongly.

The cognitive mechanism that helps humans think, solve and navigate through
problem spaces acts as a complex system that produces behavior. Sometimes the behavior
will lead to answers and by correct prodding in the right direction among the myriad
variables that are taken into consideration or “thought” of and their interaction, the behavior
of thinking can be trained to solve problems more often. Note that by changing the cognitive
process from a mechanism to a behavior, the problem space now grows in size to solving any
general problem by the right behavioral techniques instead of having to generate specific
mechanisms for every single problem space.

Problem solving mechanisms, including those of artificial intelligence systems should
thus explore the domain of multiple variables and their properties, interactions and life cycles
along with the behaviors that make use of them. Using multiple agents to model such systems
help us explain such approaches [66]. Multiple agent models also give us the advantage of
introducing a vital ingredient in behavior development – Randomness. By making the
behaviors of the agents imprecise, by adding a dose of randomness in their functions,
decisions, interactions and life span, we are creating a dynamic, non equilibrium environment
that will force diversity and unpredictability in the system.

This may seem like a paradox to some; if the point of the system is to solve a problem
space, why introduce unpredictability or randomness in the mixture? Won’t it come in the
way of solving the problem at hand if you cannot determine what the system or mechanism
that runs the system does? But why does computation mean that you should control all
aspects of the computing system? Is it hard to believe that an answer or many answers can be generated out of the system simply due to the architecture and interactive complexity of the system?

**EMERGENCE**

Intelligent behavior, or more correctly, the behavior of computation or information processing is an emergent phenomenon observed in naturally occurring complex systems. Apart from ant colonies and immune systems, flocks of birds, shoals of fish and even civilizations show a semblance of organization and collective behavior arising out of simple structures exchanging information among themselves, in other words, behavior emerges out of the rules. This phenomenon of global behavior arising out of local structures due to the reactions and communication among local structures is called as emergence.

More eloquently put, emergent phenomena are characterized by globally coordinated behavior arising out of the activities of many local agents acting on local information [1]. An architecture is described as an emergent architecture if its top level behavior emerges as a consequence of many small computational actions [67].

You cannot build emergence into a system, rather, it’s a global view of the system and not explicitly implemented [68]. However, there are conditions that are necessary among the lower level structures like response and communication protocols, presence of multiple agents and different orders of structures, though not in a hierarchy that are helpful for emergence to permeate and arise from the system.

However this does not mean that if we build a complex system that satisfies certain conditions, emergence will spring intelligence out of the system when you turn it on. That would be too convenient and we would have lots of intelligent company on this planet and others if it were true of so formed intelligent beings! This, for want of a better phrase, handicap of obviousness, is the allure to emergent complexity sciences.

**OUR PROPOSAL FOR INCORPORATING CONCEPTUAL LINKAGES**

We propose to build an artificial intelligence system that incorporates the architecture of a complex system. We plan to deploy multiple agents that work on the metadata fed along with the data as input into the system. By setting up the conditions for and encouraging
emergence in the system, we hope that the sediment of the interactions between the multiple agents ends in formation of concepts and sets of concepts. More importantly, when these concepts are connected to each other to form higher level structures or hierarchies, the trail they leave behind by linking with each other forms a breadcrumb trail of concepts after which the system is named.

The metadata fed into the system will be in the form of sets of strings that are recognized as related to the data they are tagged with. These metadata labels are user filled derivations and interpretations of data that are tagged to each piece of data fed into the system. Fed into the ruleset of the system are instructions of how these strings relate to each other and with respect to the contexts from which they are obtained. The many agents, which are snippets of code will engage themselves with each metadata label and follow the rules presented to them while interacting with these labels. The intention of these rules and behavior protocols is for an emergent behavior to arise as a global phenomenon which helps pool similar metadata labels and the attached datasets together along with leaving a trail of logically progressive linked metadata to follow among the dataset for the user to investigate and get leads for a story. When these metadata clump together owing to their similarities in interpretation, these clumps can be regarded as the birth of concepts. However, the individual metadata tags can be considered as whole concepts too. In the above explanation, replacing the word metadata with concept will still hold good for the processing within the system. Thus, a further clumping of concepts will represent a step further in the complexity of the concept definition or interpretation thus leading to deeper concept levels that the system will discover for the user.

But what is the basis for making these assumptions of conceptual linkages in the first place? The human cognitive phenomena are emergent statistical effects of a large number of small, local, and distributed subcognitive events with no global executive says Hofstadter [69]. This interpretation gives us the idea that instead of modeling higher order semantic elements directly on the elements of an AI network, there is a possibility that clumping of syntactic elements together might derive these semantic elements, simply put, they are higher level semantic interpretations developed from lower level syntactic processes in the system.
EMBODIMENT AND USER FEEDBACK

User feedback is both a control mechanism to direct the information processing in the system to the right areas to be explored in the problem space along with being an extension of the cognitive mechanism of the user that the system is hoping to supplement. The intuitive experience of the user by virtue of his real life knowledge and experience can be transferred to the system via the window of User Interface (UI). The UI should not just be seen as a combination of buttons being pushed or information read off the screen. In fact in a transparent system that reveals its working mechanism to be manipulated in order to suit the solution the user is trying to look for, the UI can be seen as a part of both the mental activity sphere of the user along with a control mechanism for the system.

In our artificial intelligence system via the UI, the system experiences pressures from the outside environment and reacts to it with changes in it’s processing. This in turn leads to pressures on the outside environment to react to the changes and it pushes the system. This cyclic process is vital in creating and retaining the dynamic environment in and around the system that will lead to changes in the information processing. The external pressure here is mainly from the user and data fed into the system. This changes the context or priorities of the agents working in the system and shows different trails and this in turn leads to different reactions from the user.

METADATA

Simply put, metadata is information about information or data about data. When we talk about making use of data exhaust to help us be more knowledgeable about data, we actually mean that better observation skills about metadata reveals a lot more about the properties of the data other than its contents. Metadata is a vital component of the Semantic Web [70], which is a point of view about the World Wide Web that it is more than a set of documents or web pages that make up the internet. By using additional information about the nature and source of the data present in the web pages, by understanding the relationships between the links that connect the different datasets and web pages in the web, the agents and processes that interact with the web can comprehend such data and react to it better rather than just be a passive super library of documents and data that people can look up to for information.
Metadata in the context of our system is user-curated data about data that defines the relationships between the context and the data. These relationships determine how the agents react to the data and what scenarios of clumping are possible to generate concepts as the result of those clumping. In our system, metadata are the primary factors that are worked on and utilized and the data from which they are drawn are just placeholders for the contextual linkage that is present. Since concepts differ with contexts, working on metadata gives a sense of the context for any intuitive experience to derive an advantage out of the linkages. If there is one idea we would like to implant firmly, it is this nexus between metadata and concept realization. That concepts are the sediments of relationships between data, metadata and contexts that will emerge when agents, be it pieces of code or largely networked neurons, work on them.

It is now quite clear, in the course of our investigation and explanation that these essential relationships need a significant amount of abstract and interdisciplinary research. And that is about refining iteratively till we are satisfied about the definitions of these terms and relationships with sufficient precision. This is needed, we would argue, either simultaneously or before more work is carried on in unraveling the complexities of Big Data, understanding the specific needs of journalists in news making and bringing out computational tools that can be tested on further iterations. We look forward to continuing our work on this test bed of ideas and ultimately realize a sophisticated tool as implied by our investigation.
CHAPTER 3

DESIGN AND TESTING METHODOLOGY

BREADCRUMBS COMPUTATIONAL OVERVIEW

The essential point of this research is to clearly identify three areas of previously disconnected endeavor that desperately need each other for real progress to be made. In the preceding we have argued that this fine-grained level of precision in articulating the issues, definitions, requirements, technologies and limitations were necessary to even consider constructing a computational artifact to test for its merit in addressing the needs of journalists in the face of Big Data using a novel computational paradigm such as emergence. Yet it is helpful at minimum to articulate some of what might comprise a software architecture, some of its necessary instantiated components and the tests that validate it. This chapter attempts to provide such an articulation.

The basic idea is that “agents” must be interacting with only limited pieces of a bigger picture, yet somehow either through modifications to their environment or through direct communication, manage to give rise to the identification of documents (here used as a general term) that have some useful connection to one another. The cogent difference, of course as already stated, is that the connection is not one of statistical co-occurrence of specific tokens but rather of degree of relatedness of concepts underlying the semantics of the specific tokens. This is what we have been referring to as conceptual linkage rather than statistical linkage. It is simple enough on the surface to imagine a collection of objects (data and behavior) each with its own persistently and periodically activated thread of execution engaging in those behaviors taking on the task of encountering documents and assessing the utility of suggesting connections between them on the basis of categorical or other conceptual relationships rather than simply on statistics of occurrence. The real difficulty in leaving the statement of the implementation at this level is that we have no sense of what environment these agents move in nor how they assess the relatedness of the concepts that apparently link the documents.
There is a framework for managing emergent perception in software that addresses some of these issues. The use of such a framework could go a long way toward realizing for these simple agents in an environment like the one they would encounter in a world full of documents and concepts and metadata about relationships between these concepts. In the remainder of this chapter we describe this framework and some ideas about how it might be brought to bear on the peculiar trifecta of domains and requirements we have so far described.

ANCHORING THE BREADCRUMBS IMPLEMENTATION WITH STARCAT

Starcat is a framework which was created by Dr. Joseph Lewis to provide the basic functionality of the Copycat program; it allows other domains to utilize the essential functionality of Copycat. “Emergent representation is a key feature of Starcat’s functioning and adaptive behavior is the desired result that it supports” [1]. Unlike Copycat, the Starcat framework is not specific to any domain; it allows custom domain-specific modules to be added on top of the main Copycat engines. Lewis states that the Madcat and Copycat systems have “two important limitations… their inability to adapt to novel situations and their lack of true autonomy” [1]. Aside from not having domain-specificity, Starcat also adds value to existing CAS in that it “is an open-ended architecture for computational systems that autonomously adapt their behavior to continuously changing environments” [1]. Starcat’s ability to allow for autonomous systems presents a new niche in the field of AI.

Machines that were limited both by their domain and their ability to learn and adapt are now capable of exhibiting emergent behavior in the face of an ever-changing environment. This revelation allows a system to develop adequate enough representation to understand the scenarios that it comes across while interacting with the environment, making decisions and learning from experiences, much as a child does. This novel ability provides the potential platform for the AI field to expand in a direction that lacks human guidance and control. This may be the basis for machines to exhibit intelligence without the underlying regulation of programmed control.

The main structures of the Starcat Framework are the Workspace, Slipnet, and Coderack. The Workspace is the environment where objects are observed and analyzed. The Slipnet is a collection of connected nodes, which represent concepts pertaining to the
environment. The Coderack is a collection of Codelets (the agents of Starcat, small executable builders and assessors of structures in the Workspace emitted from the Slipnet), which are to be performed on objects in the Workspace. It ensures the parallel exploration of several possible interpretations at rates roughly proportional to their currently perceived utility. All three of these components are continuously working together in order for the framework to allow the domain to which it is applied to approximate an understanding of its environment.

The Slipnet is a network of nodes that collectively transmit via links between the nodes something called activation, a kind of “currency” for assessments such as relevance, utility or evaluation. Each node in the Slipnet represents a concept that is specific to the environment of each unique domain that the Starcat framework is utilized for. Examples of some concepts that may occur in environments typical of our domains of interest. In journalism, concepts that occur which require identifying single or multiple elements as a lead, for example the “who” of a story or the “when”. These notions come from the well-established inverted pyramid paradigm for story-construction. The notions just described, i.e. “what is a lead”, is precisely what is meant to be included in the metadata-specific network. The actual “facts” that answer such a question, like “Roger Johnson on May 7”, contain specific concepts, which would be instantiated in the node(s) of the network instantiated specific to a particular linkage search. The linkage search term reifies to some degree the project title, “Breadcrumbs”.

An environment of documents among which linkages are sought requires certain concepts to be assessed as verified or unverified (these being metadata nodes that pertain during a particular linkage search to specific concepts from the actual documents). Activation values of nodes in the Slipnet range from 0 to 100. As a concept is increasingly realized by the system its activation value(s) will be increased as well. High levels of activation of a concept mean that in the Workspace (the environment where both documents of interest and the Breadcrumbs are manipulated by agents) that particular concept has an increased likelihood of being useful in further activity in the Workspace. If the environment is tested to see if a lead is present in the most important or least important position, then the concept for lead as most important is realized and will receive a large portion of the activation. When the activation value of a Slipnet node reaches a high enough level, it will
produce Codelets (agents) that will be placed in the Coderack to perpetuate this discovery cycle. Actually two things are happening here: one is the creation of a new agent to pursue the newly active concepts and the other is, via activation in the Slipnet, and in analogy with traditional blackboard architectures, the concept is essentially added to the blackboard, namely it is instantiated in the Workspace as a first class object with which agents interact.

The Coderack is a stochastic queue of codelets; it is used to ensure that numerous possibilities as manifest by the agents interacting with the metadata/instance-data collection are explored with fair effort given to each in an interleaved, parallel fashion. That is why the term stochastic is used to refer to the placement of the Codelets into the Coderack and subsequently emitted for behavior in the environment. Codelets are placed onto the Coderack using a priority value; higher priority codelets have a better chance of being chosen over the other codelets in the Coderack, though all will generally get some chance. This is where the term randomness or unpredictability is sometimes, without quite the necessary fidelity, inserted in this discussion. The real point is that no Breadcrumb trail is sacred; at any moment a “surprise” might gain enough activation to motivate heretofore unexpected recommendations whose novelty is just what the user of the system might have hoped for. A Codelet priority value is determined by the amount of activation that it receives from its corresponding Slipnet node. When a Codelet is chosen this means that it will be placed into the Workspace (basically its parameters bound and its method executed).

A Codelet is placed into the Workspace in order to interact with the environment. Codelets could be placed into the environment in order to detect such things as leads (metadata), specific tokens with particular semantics in the context of some document (instance data), and other self-regulating behaviors within the system. The information that is returned to the system from the Codelets activates a chain of events, which involve the two kinds of Slipnet and the Workspace.

What we see as the best utilization of this framework in the context of the proposed recommendation system is as follows. The agents previously described are akin to the Codelets of the Starcat framework. They engage in small-scale, simple interactions with an environment. They may interact with one another but more likely they leave behind modifications to the environment that other agents react to. This is the source of the metaphor of “Breadcrumbs” in the project title. The modifications are more formally known as
stigmergy, “signaling modifications in the environment to other agents in the system”. The Workspace of Starcat is not only the site of interaction and stigmergy among agents or Codelets but also the locale of the documents that these agents encounter. Moreover the Workspace is the arena in which instances of concepts that provide relevant linkage between documents are left—on the basis of semantic content as assessed in the small by one or more agents and their collective behaviors. The instantiation of these concepts in turn provide activation to the nodes in the Starcat Slipnet. The current state of the Slipnet is what provides the time-varying sense of “context” for the overall system (the autonomous recommender agent, so to speak). As new relationships become useful either by repetition or by feedback from the user, certain collections of related nodes become more dominating in directing the instantiations that occur in the Workspace.

A new piece to the story not previously explicit in uses of the Starcat architecture is the heavy reliance on the use of metadata, as investigated extensively in the journalism domain that is central to this thesis. Here most likely we would see the fleeting instantiations of currently relevant concept halos, (collections of nodes and activation levels) that pertain to the set of documents and topics contained therein; and then a more persistent set of nodes and relationships among them ,that drive the choices made by the agents or Codelets in trying to determine which particular notions (token-semantic bindings) matter, for the documents under consideration by some journalist using this tool to explore possibly usefully related documents from a very large set.

With these basic mappings and one slight elaboration (the two styles of Slipnet, one for instantiation per linkage search and one common and persistent for almost all metadata) we now need to describe what kinds of test scenarios could be engaged to address this proof of concept.

Let us start with one simple scenario. A journalist seeks connections of interest and sources to substantiate links in an urban region between the presence of vacant lots, say because of pending zoning approvals, and crime rates. On the surface this may seem somewhat obvious. Such locations attract folks engaged in nefarious activity they can sometimes keep out of sight; and this further attracts satellite activity that is also possibly deleterious to the quality of life in the area, if not downright illegal and dangerous. But a traditional search for published information linking the ideas on both sides of this equation
would currently be most likely pursued using a tool like the page-rank algorithm of Google. Tokens chosen by the journalist for the semantics relevant to their perception of the players in the problem space are identified and statistics suggests “most likely” associated documents. It is all too familiar how often such statistical linkages provide only surface associations that are at best a distraction and occasionally misleading and time-wasting. Associations or linkages suggested on the basis of relationships between concepts both given by the user and found in the high-utility documents (a place for user feedback in the system to play a major role) are not only more likely to avoid those pitfalls, but also because any system taking on such functionality would almost certainly have its own set of relationship among instances of concepts known to be connected to those already found. Moreover a set of meta-relationship about how to look for certain “categories” of relationships between instances of concepts can typically be provided to guide the process. As previously identified this is one of the major areas identified in the triad of journalism, Big Data and emergent computation that is ripe for additional research efforts. Now we add to this simple picture, already hopefully more capable of supporting the story-construction effort, the fact that journalists are often searching through databases of massive amounts of occasionally poorly qualified or undifferentiated data and their confidence in any tool that relies heavily on their intuition about the tool’s functioning drops dramatically. We believe conceptual linkages can be provide through software adhering to the paradigm of complex systems or emergent computation that will not only work well but garner increasing faith on the part of the journalists who use it. So this is the story we wish to tell with the Starcat inspired rapid-prototyped Python-agent Breadcrumb world. Let’s take one last trip through the details to see what kinds of test scenarios could give us a sense of the merit or utility of this approach.

Through a combination of “pre-population” by us the experimenters and automated extraction of keywords and lookup procedures for their semantics both of which can be mapped into something like the familiar Copycat variant of the “semantic network” of early NLP research, we arrive at partially filled instance-level Slipnet(s) of a running Starcat-inspired Breadcrumb system. These nets can also be amplified via the circular process of Breadcrumb agents (Starcat codelets) circulating through the system. Some will look for specific and currently cogent token-semantic bindings in potential documents (concepts). Others will use metadata to identify parts of documents potentially containing useful
concepts and simply note what is there. These get added to the Workspace for evaluation, much as in familiar expert system or blackboard systems. They also increase the likelihood by activation level in the nets for those instance concepts and those meta-relationships to continue to play important roles, at least for a while. The metadata network is of course mostly pre-populated by the developers. The agents, as just implied, come in two flavors: those that seek already “hot” items and those that seek what is there, possibly new, to drive more discovery. And the internal Breadcrumbs, or communication markers, between agents allow for threads of exploration to grow and wane as more and more information synergy leads to suggested linkages, on the basis of identified concepts and their relationships, rather than just on counting the number of times the same words show up. Other advantages not cited here but which will come in later phases, such as learning that can be applied to later runs on different journalistic questions (or whatever the domain may be). Implicit here is that the Workspace be populated with candidate documents. More accurately, agents are also available to seek new elements (documents, items) to add to the Workspace for evaluation and to provide pressure on the concepts that are relevant.

So in sum, a developer builds a metadata Slipnet and a primer set of both kinds of agents and provides links for the document finder agents to seek new material and possibly agents that seek user feedback. The initial instance-level Slipnet(s) are populated to some degree that must be determined by empirical experiment. Then the system is allowed to run long enough to make a number of recommendations of related documents that meets some (predetermined or dynamically measured) criteria from the user. For the exact same set of journalistic questions and possible world of documents, we repeat this experiment a “large” (statistically relevant) number of times, say a thousand. The “peak” recommendation sets are then compared with what some number of trained journalists would have suggested using any of several standard techniques. We expect to see that the conceptual linkage not only hits well within the human-selected peak, but also that humans are often more pleased with the coherence of these suggestions, the fidelity of the connections and occasionally the surprise of unexpected but highly valuable connections. In order to test the abstraction possible by a system of this nature (the power of emergence), we repeat the entire scenario for as many different journalistic questions and domains of potential data sources as is feasible.
In the second phase of this project, where some rapid-prototyped version is actually tested, possibly without the full machinery of the existing Starcat framework, we would quantify not only the parameters for running the system and the test setup, but the relative “bandwidth” of the subsumed Breadcrumbs recommendation set peak to the human-selected peak that is acceptable and also some metric for the amount of surprise that is good and not detrimental. This thesis answered the first questions to know even what kind of recommendation system is worth contemplating here. The first of several next steps is this quantification process for thorough testing of the rapidly prototyped system. Following that is a full version, learning capabilities and a number of other features described in some more detail in the remainder of this chapter.

What has been done by initiating this proof of concept initiative is opening a door towards asking questions and debating on answers obtained. This triage of community, challenge and approach to solve a problem yield a rich opportunity for collaboration across the many disciplines of data sciences, computation, human computer interface, journalism, the artificial intelligence community, to name a few, to ask fundamental questions of the working of their own domain and other domains, to stand in different shoes and look at what they take for granted about assumptions and definitions of their own work, to learn innovative methods from other communities to apply to their own and contribute in the understanding of the mystery of cognition.

**Future Phases**

We have envisioned multiple phases to take the Breadcrumb idea forward to a full blown product and would like to end our discussion by presenting these phases to you

**Phase 1**

What we have done with this project of capturing the characters needed for an active system that will respond to situations. In other words, a worldly problem to solve for instead of a “let’s see what we can do here with X system” approach. This also means asking the right questions, getting lost down confusing answers, coming up with a testable architecture and putting forward test scenarios for the claims to be tested.
Phase 2

Rapid prototyping in Python to test the practicality of the relationships mentioned but not starting from scratch. We can take the existing Starcat architecture and hack it to test the scenarios and support the architectural claims from phase 1.

Phase 3

Full blown version of the architecture including any new mechanisms that have to be placed if the rapid prototype shows discrepancies. Feeding uncurated data to the system, in other words, letting it run in the real world.

Phase 4

Look at the HCI part of the architecture and tweak it to better user design so that the UI helps in the cognitive tasks too along with the internal architecture of the system. Asking the agents to generate the metadata themselves by training them and again let it loose in the real world as a fully functional product.
REFERENCES


APPENDIX

GLOSSARY
Agent: An autonomous operator with simple instructions on how to react to and operate on objects or other agents in its local environment. In Breadcrumbs/Starcat context, agents are pieces of code fed with specific instructions/protocols.

Token: An object that represents something else, either abstract or in the physical world. Essentially any kind of sign or symbol often encapsulated in a sequence of alphanumeric characters.

Document: A generic term to describe different file structures in the software world. Ranging from spreadsheets, pictures, audio/video snippets, textual documents, etc.

Concept: In the Starcat/Breadcrumb context, clumps of similarly related metadata are called concepts. However if the metadata is semantically strong enough, a single unit is also considered as a concept.

Semantics: The meaning of the syntactical string as defined by the underlying programming language or used human language.

Linkage-association: The clump of metadata or concepts that are linked together by their similarity or relatedness in semantic meaning.

Environment: An underlying framework that holds the users, operating system, variables and applications to interact with each other. The variables hold the current environment configuration.

Framework: A layered structure with a set of rules that determine how programs, interfaces, tools inter relate and communicate to each other.

Metadata: Data about data. Additional information describing the characteristics and various identifying features of the data.

Adapt, evolve, learn: Change in behavior during momentous changes in environment is adapting, abstracting useful information from that behavior to use in future scenarios is learning, evolving comprises both but over generations of life spans

Stygmergy: A trail of tiny decisions or markers that are left in response to encountering the external world by an agent that in turn change the external world which leads to a cycle of trails and changes.