Big Data:

Estudo do Ambiente, Desafios e Analise Estratégica para o Brasil

Daniella Zuppo
Luciana Sodré Costa
Soraya Fernandes

Projeto Final apresentado ao Programa de Pós-graduação em Engenharia de Produção, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Especialista em Gestão do Conhecimento.

Orientador: Marconi Ponte Pereira

Rio de Janeiro
Setembro / 2013

BIG DATA:

ESTUDO DO AMBIENTE, DESAFIOS E

ANÁLISE ESTRATÉGICA PARA O BRASIL

Daniella Zuppo

Luciana Sodré Costa

Soraya Fernandes

PROJETO FINAL SUBMETIDO AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃOE PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO TÍTULO DE ESPECIALISTA EM GESTAO DO CONHECIENTO

Aprovado por:

_____

Luis Claudio Mangi, M.Sc.

_____

Marconi Ponte Pereira, Esp

_____

Marcos do Couto Bezerra Cavalcanti, D.Sc.

RIO DE JANEIRO, RJ - BRASIL
SETEMBRO DE 2013

Costa, Luciana Sodré; Fernandes, Soraya; Zuppo, Daniella

Big Data: **Estudo do Ambiente, Desafios e Análise Estratégica para o Brasil** / Daniella Zuppo; Luciana Sodré Costa; Soraya Fernandes. – Rio de Janeiro: UFRJ/COPPE, 2013

vi, 56 p

Orientador: Marconi Ponte Pereira

Projeto Final (MBA) – UFRJ/ COPPE/ Programa de Engenharia de Produção, 2013.

Referencias Bibliográficas: p. 54-56.

1. *Big Data*. 2. Gestão do conhecimento. 3. Ciência das redes. I. Pereira, Marconi Ponte. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia de produção. III. *Big Data*: Estudo do Ambiente, Desafios e Análise Estratégica para o Brasil

Resumo do Projeto apresentado à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do título de Especialista em Gestão do Conhecimento.

BIG DATA:
ESTUDO DO AMBIENTE, DESAFIOS E ANÁLISE ESTRATÉGICA PARA O BRASIL

Daniella Zuppo
Luciana Sodré Costa
Soraya Fernandes

Setembro/2013

Orientador: Marconi Ponte pereira

Programa: Engenharia de Produção

Este trabalho apresenta um estudo do ambiente *big data*, os principais desafios na apropriação do conceito e uma breve análise estratégica do Brasil em relação ao aproveitamento dessa oportunidade. Foi considerado ambiente *big data* todo o contexto do processo produtivo de conhecimento a partir de dados digitais gerados, armazenados e analisados em rede. Os desafios encontrados situam-se na ordem dos aspectos éticos, culturais e legais que podem tornar-se gargalos desse processo produtivo. A análise estratégica foi feita através de uma adaptação do modelo SWOT.

Abstract of Final Project presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree Knowledge Management Specialist.

BIG DATA:
STUDY ON THE ENVIRONMENT, CHALLENGES AND STRATEGIC ANALYSIS FOR BRAZIL

Daniella Zuppo
Luciana Sodré Costa
Soraya Fernandes

September/2013

Advisor: Marconi Ponte pereira

Department: Industrial Engineering

This work presents a study of the Big Data environment, the main challenges in appropriating the concept and a brief strategic analysis of Brazil to embrace this opportunity. The Big Data environment considered was the entire context of knowledge production from digital data generated by the use of the internet. The challenges refer to ethical, cultural and legal issues that can become bottlenecks in this process of producing knowledge. The strategic analysis was performed using an adaptation of the SWOT model.

Sumário

Introdução **7**

1. Gestão do Conhecimento em Ambiente *Big Data* **9**

2. Ciência das Redes **12**

3. *Big Data* **20**

3.1. Principais conceitos 20

3.2. Ecossistema 27

3.3. Ciclo de Adoção de Tecnologia 31

3.4. Curva de Aprendizagem 34

4. O Mercado para *Big Data* **38**

4.1. Global 38

4.2. Brasil 39

4.3. Análise Estratégica para o Brasil 41

5. Conclusão **44**

5.1. Desafios Estruturais 44

5.2. Desafios Culturais 45

5.3. Desafios Estratégicos 48

5.4. Quatro Pilares para Continuação dos Estudos 51

Bibliografia **54**

Big Data:

Estudo do Ambiente, Desafios e Análise Estratégica para o Brasil

Introdução

Pela primeira vez na história nossa capacidade está começando a alcançar nossas ambições

Peter Diamandis – Xprize

Uma das questões mais importantes da atualidade diz respeito à utilização dos dados digitais gerados pelo uso intensivo da internet nos últimos 20 anos. É impossível estimar o volume de bytes acumulados, mas especialistas arriscam-se a fazer estimativas em bases diárias.

Todos os dias aproximadamente 2,5 bilhões[1] de gigabytes de dados são criados a partir de postagens em redes sociais, *upload* de fotos e vídeos, registros de transações comerciais, sinais de GPS, rastros de navegação e sensores dos mais diversos tipos.

A mais recente fonte geradora de dados digitais é a Internet das Coisas (IoT)[2] da qual falamos ainda muito pouco, mas que promete ser um divisor de águas não só em quantidade de dados, mas também em diversidade. Trata-se de uma rede sem fio de sensores[3] colocados nos mais diferentes tipos de objetos, animais e plantas que geram, trocam e armazenam dados na internet a partir de um sistema eletrônico de informação.

Os registros gerados abrangem desde dados estáticos inerentes ao objeto, como descrição, modelo, características físicas, local e data de fabricação, até dados dinâmicos sobre sua utilização, relação com seres humanos, com o meio ambiente ou com outros objetos. Entre os dados mais comuns coletados atualmente pela IoT estão a localização, deslocamento, temperatura, padrão de uso, desgaste de materiais, troca de peças etc.

[1] Estimativa da IBM publicada em http://www-01.ibm.com/software/data/bigdata/ (página visitada em 05/07/2013)

[2] Em inglês: *Internet of Things* (IoT). A rede que interliga objetos no mundo inteiro. Algumas empresas como a CISCO já utilizam a nomenclatura Internet of Everything (IoE) - Internet de Todas as Coisas - para evitar a setorização da internet.

[3] *Wireless Sensor Network*, um protocolo de comunicação criado pelo Auto-ID Laboratory do MIT.

Em 2008 um relatório da CISCO[4] afirmou que o número de coisas conectadas à internet já ultrapassava o número de humanos na terra (em torno de 7 bilhões).

Em 2011[5] a empresa estimou que em 2020 um trilhão de objetos estariam conectados à internet. A mesma estimativa foi compartilhada pela IBM[6].

Confirmadas as previsões, em breve teremos um trilhão de objetos se comunicando de modo contínuo, gerando uma quantidade de dados incalculável.

Não existe consenso sobre a definição de *big data*. Na maioria das vezes a definição é feita pela enumeração das suas características como veremos no capítulo 3, item 3.1. O fato é que, convencionou chamar de *big data* todo esse conjunto de dados digitais e a expectativa que existe em torno dele justifica-se pela suposição de que essa grande quantidade de dados possa nos trazer informações inéditas e relevantes sobre fenômenos ainda pouco explicados.

Apesar de ser um conceito relativamente novo, já existem diversos casos de utilização dos dados digitais resignificando redes, teorias e processos nos mais diferentes domínios do conhecimento.

O objetivo desse trabalho é oferecer uma introdução ao assunto *big data*, identificar os principais desafios para sua efetiva utilização e apresentar uma breve análise de como o Brasil se encontra estrategicamente posicionado para a absorção dessa oportunidade.

A primeira parte desse trabalho, procura contextualizar o assunto dentro do campo de atuação da Gestão do Conhecimento. Esse posicionamento corresponde à intenção de dar ao assunto um foco mais estratégico que técnico, aproximando *big data* mais do tomador de decisão que do profissional de tecnologia da informação.

No capítulo 2, será feita uma breve exposição da Ciência das Redes, um campo de estudo interdisciplinar das redes complexas que legitima as promessas revolucionárias do *big data*. Toda linha de argumentação deste trabalho será baseada nos fundamentos dessa ciência.

[4] http://www.techenet.com/2011/07/cisco-net-coisas-infografico/ (página visitada em 25/02/2013)
[5] http://www.cisco.com/web/about/ac79/docs/innov/IoT_IBSG_0411FINAL.pdf (página visitada em 25/02/2013)
[6] http://www.techenet.com/2011/07/cisco-net-coisas-infografico/ (página visitada em 25/02/2013)

O capítulo 3 apresentar *big data* a partir de suas dimensões críticas, as diferentes utilizações do termo, comparação com outros tipos de base de dados, o seu ecossistema e uma apreciação do ciclo que se espera percorrer para a adoção da tecnologia com a produtividade desejada.

Por fim faremos uma breve análise do mercado brasileiro através da identificação de forças, fraquezas, ameaças e oportunidades a partir de uma adaptação do modelo SWOT.

1. Gestão do Conhecimento em Ambiente *Big Data*

O conhecimento é a única forma de garantir a sobrevivência. Isso quer dizer que aprendemos para sobreviver[7] às ameaças do ambiente. Isso é igualmente verdadeiro quer estejamos falando do aprendizado das espécies para sobreviver às ameaças do seu ambiente natural ou do aprendizado de membros de uma organização para perpetuar no seu ambiente de atuação. Todas as espécies fazem isso, através da adaptação que é a forma básica do aprendizado, mas a espécie humana é a única que além de se adaptar, faz previsões e tenta controlar seu ambiente. Nesse contexto, controlar será sempre a ação de resolver problemas que ameaçam a sobrevivência.

Na tentativa de melhorar a eficiência dessa habilidade natural da espécie humana, muitas organizações praticam a chamada Gestão do Conhecimento (GC).

Resumidamente a Gestão do Conhecimento é um conjunto de processos organizacionais que consistem em obter, acumular e interpretar informações para construir, disseminar e integrar conhecimentos que possam ser aplicados na solução de problemas[8].

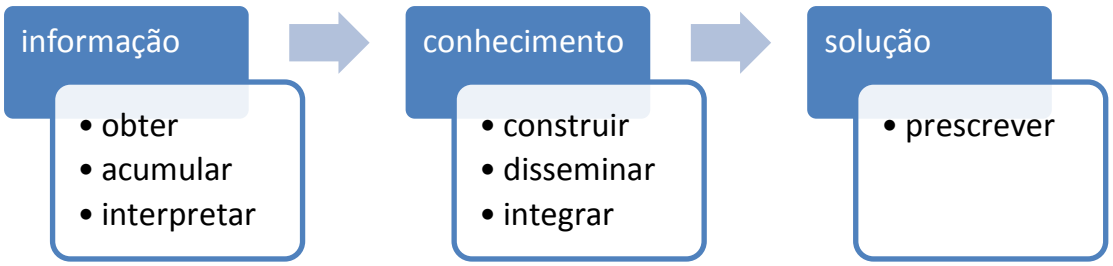
Vista como um processo produtivo, a Gestão do Conhecimento transforma informações em soluções.

A *figura 1* mostra de maneira esquemática esse processo.

[7] Nesse trabalho sobreviver terá o sentido amplo de perpetuar uma situação desejável, inclusive a vida, mas não somente ela.

[8] Adaptação das autoras a partir da definição de Johnson, 2011, p. 22: *A Gestão do Conhecimento (GC) pode ser definida como um conjunto de práticas organizacionais que visam construir, obter, acumular, disseminar, interpretar, integrar e aplicar conhecimento*.

Figura 1. Gestão do Conhecimento[9]



A informação é um conteúdo que foi organizado a partir do discernimento de padrões no que é observado[10]. O primeiro esforço da GC para agregar valor a esse conteúdo é garantir o acesso a ele (obter a informação) e incrementar sua quantidade e diversidade (acumular). Em seguida a informação ganha significado e relevância[11] a partir da interpretação que se dá a ela. Nesse ponto teremos uma informação qualificada.

Para transformar informações em conhecimento, o esforço exercido é o de relacioná-las com a experiência das pessoas. Nas organizações o conhecimento é algo que pode ser aplicado em algum tipo de sistema produtivo[12]. Portanto ele só existe quando gera prescrições que possam transformar uma situação existente em uma situação desejada. Essas prescrições visam dar apoio a outro processo muito relevante nas organizações: a tomada e decisão.

Podemos então dizer que a informação é a base do conhecimento, e que este é a base da ação. Portanto, o maior indicador de sucesso de um projeto de GC está nos objetivos alcançados pela organização.

Qual a importância de se pensar a Gestão do Conhecimento em ambiente *big data*? Inicialmente é forçoso reconhecer que a identificação de padrões no que é observado ao nosso redor (informação) está cada vez mais sendo feita de modo digital, gerando uma grande quantidade de dados. Ou seja, estamos vivenciando um processo progressivo de dataficação[13] da sociedade.

[9] Quadro das autoras
[10] Johnson, p. 32
[11] Davenport e Prusak
[12] Johnson, p. 33
[13] Neologismo: do inglês *datafication* que é o processo de transformar informações de qualquer tipo em dados digitais (Schönberger e Cukier, 2013).

O impacto desse fenômeno começou a ser percebido na última década e a relevância desse impacto pode ser explicada resumidamente da seguinte maneira: Dados digitais só podem integrar uma única cadeia produtiva: a de conhecimento. Isso significa que é esperado uma explosão na quantidade e efetividade do principal produto do conhecimento: solução.

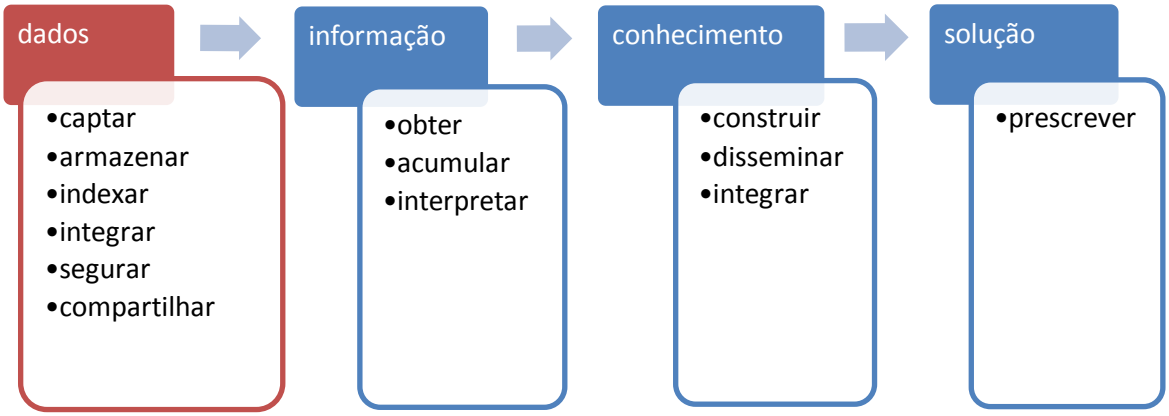
Por tudo o que foi dito acima, acreditamos que a GC continuará sendo a principal força atuante em um projeto bem sucedido em *big data*. A importância de pensar a gestão do conhecimento especificamente em ambiente *big* data é que esse processo apresenta algumas especificidades nesse ambiente que não devem ser desprezadas.

A primeira, e mais evidente, é que GC em *big data* terá suas ferramentas fortemente apoiadas em Tecnologia da Informação (TI).

A segunda é que, enquanto a GC tradicional atuava a partir das informações como mostra a *figura 1*, em ambiente *big data* essa gestão deverá ampliar seu escopo, incorporando uma participação ativa numa etapa anterior: a gestão de dados.

A *figura 2* mostra como fica o processo produtivo de soluções quando iniciado sempre a partir de dados digitais.

Figura 2. Gestão do Conhecimento em ambiente big data [14]



Em TI a gestão de dados é chamada de governança de dados e inclui todas as atividades necessárias para gerar e disponibilizar os dados para utilização: captação, armazenagem, indexação, integração, segurança e compartilhamento[15].

[14] Quadro das autoras

O profissional de GC não terá que se apropriar dos conhecimentos de TI que asseguram a governança de dados, mas terá que compreender que, no ambiente *big data*, a qualidade da informação que ele terá dependerá diretamente do modelo de governança utilizado.

A sua atuação nas decisões sobre como obter base de dados e como definir a plataforma de *hardware* e *software* será um diferencial não só para as competências individuais do profissional, mas também para a organização que pretende utilizar *big data* como um dos sistemas de apoio à decisão.

Apesar dessas duas especificidades intrinsecamente ligadas à tecnologia, a GC em ambiente *big data* é semelhante à já praticada em outros ambientes. O último valor gerado é a solução de um problema e isso só é feito a partir de uma tomada de decisão que é uma competência humana indelegável.

O desafio da gestão do conhecimento de transformar as melhores informações nas melhores decisões receberá um reforço incontestável da tecnologia da informação, mas em última instância continua sendo um desafio humano.

2. Ciência das Redes

A pobreza é uma condição artificial e externa de uma pessoa. E já que é externa, pode ser removida.

Muhammad Yunus – Economista, Nobel da Paz

A informação de que existe um gigantesco banco de dados inéditos[16] já é suficiente para revolucionar qualquer área do conhecimento porque aumenta exponencialmente as possibilidades de descoberta de correlações inusitadas.

No entanto, a promessa inovadora do *big data* é a sua capacidade de tornar as redes complexas[17] não apenas mais compreensíveis, mas também mais previsíveis e mais controláveis. A afirmação vem dos pesquisadores da novíssima Ciência das Redes[18].

[15] O capítulo 3 (no item 3.2 Ecossistema) explica essas atividades.
[16] Estima-se que mais de 90% dos dados gerados ainda não são utilizados. Estes dados são conhecidos como *dark data.*

Em 2000, após diversas experiências com grande quantidade de dados digitais (documentos e dados de navegação) o cientista Albert-lazló Barabási publica[19] o que pode ser reconhecido como o surgimento da Ciência das Redes.

No século XVIII, as redes eram estudadas pela Teoria dos Grafos. Para os matemáticos da época: as redes, estruturas formadas por nós e ligações entre eles, possuíam propriedades fixas.

No século XX os cientistas começaram a enxergar outras redes que não podiam ser explicadas em termos matemáticos simples. Como desconheciam as propriedades que governavam essas redes, acreditaram que elas se formavam e se comportavam de modo randômico. Eram redes randômicas**:** a sociedade, a célula, o organismo, o ecossistema e todas as demais redes naturais e sociais.

Einstein destoava da comunidade científica e, convencido da existência de leis e princípios que organizavam e governavam o que os demais cientistas chamavam de redes randômicas, afirmou intuitivamente que "*Deus não jogava dados com o universo".*

A Teoria do Caos contribuiu identificando a interconectividade entre diversos sistemas complexos, mas ainda atribuía a eles um comportamento imprevisível e caótico. Sabia-se que as redes reais eram totalmente interconectadas, mas ainda se ignorava suas propriedades.

A Teoria dos Fractais, considerada a geometria da Teoria do Caos, identificou a existência das características e propriedades de uma rede complexa em qualquer subparte dela. Isso quer dizer que mesmo que esse sistema seja subdividido, suas partes conservarão as características do todo.

A publicação de um trabalho de Stanley Milgran[20] sobre a interconectividade social é uma das primeiras tentativas de encontrar provas para a intuição de Einstein. Através de

---

[17] Exemplos de redes complexas: código genético, rede neural, corpo humano, ecossistema, grupos sociais etc. Como são todas formadas por elementos naturais também são chamadas de redes naturais ou reais.

[18] As ideias gerais da ciência foram apresentadas ao grande público em 2002, no livro *Linked: The New Science of Network*, do cientista Albert-László Barabasi

[19] Trabalho publicado em 2000: *Scale-free characteristics of random networks*: *the topology of the world-wide web* (Albert-László Barabási, Réka Albert e Hawoong Jeong) http://www.barabasilab.com/pubs/CCNR-ALB_Publications/200006-15_PhysA-ScalefreeRandom/200006-15_PhysA-ScalefreeRandom.pdfLinked: The New Science of Networks.

uma experiência social utilizando troca de cartas entre os membros de uma população, o cientista concluiu que se cada pessoa conhecesse apenas outras duas pessoas, toda a população estaria conectada, ou seja, existiria um caminho entre quaisquer duas pessoas de uma sociedade.

Especulações a partir dessa experiência levaram à teoria dos seis graus de separação que afirmou que não só todas as pessoas de uma determinada população estão de alguma forma interligadas, mas também que o número máximo de caminhos necessários para relacionar quaisquer duas pessoas era seis.

No início da década de 80, Tim Berners-Lee criou um programa que permitia que computadores compartilhassem informações. Dez anos depois essa experiência muito bem sucedida seria conhecida como WWW[21], a rede mundial de computadores.

Em 2008 a Microsoft repetiu a experiência de Milgran utilizando 30 bilhões de mensagens eletrônicas de 180 milhões de usuários do MSN Messenger[22] chegando surpreendentemente ao mesmo resultado[23]. Se a experiência de Milgran tinha sido desacreditada pela comunidade cientifica[24] os resultados da Microsoft parecem endossá-la.

A rede de comunicação, postal ou eletrônica, parece ter sido uma boa representação do sistema social. De fato, durante a década de 90 e início dos anos 2000, a obsessão pela teoria dos 6 graus de separação inspirou diversos experimentos semelhantes, com populações distintas e amostras diferentes, que obtiveram como resultado sempre a mesma ordem de grandeza. Por isso, além de definir o grau de separação dos nós do sistema social, as experiências parecem também comprovar a propriedade fractal nesse sistema.

Apesar das críticas a Milgran, a teoria dos 6 graus de separação nunca foi refutada o que sugere, segundo os pesquisadores da Ciência das Redes, algum tipo de princípio envolvido.

---

[20] Trabalho publicado em 1967: *The Small-World Problem*. Stanley Milgran.
[21] *World Wide Web*
[22] Programa de comunicação via internet da Microsoft muito popular entre 1999 e 2008. Em 27/05/13 o serviço foi definitivamente descontinuado.
http://pt.wikipedia.org/wiki/MSN_Messenger
http://pt.wikipedia.org/wiki/Facebook
[23] http://www.theguardian.com/technology/2008/aug/03/internet.email
[24] *Could it be a Big World?* http://www.judithkleinfeld.com/ar_bigworld.html

Em 1999 dois pesquisadores[25] divulgaram que a web já contabilizava 1 bilhão de documentos.

Barabási e seus colaboradores quiseram saber se, apesar do seu imenso tamanho, a web era completamente interconectada, ou seja, se comportava como uma rede real. Partindo de dados de navegação cedidos por amigos que administravam sites de grande tráfego, Barabási não só conclui que a web era totalmente interconectada, mas também que entre quaisquer dois documentos, o grau de separação era no máximo 11. Ou seja, para chegar a qualquer documento da internet, a partir de um ponto qualquer, não seriam necessários mais que 11 links. A semelhança com o número sugerido pela experiência de Milgran realizada trinta anos antes levou ao surgimento da teoria dos "poucos graus de separação"[26].

Como a internet é uma representação de diversos sistemas complexos, Barabási trabalha com a hipótese de que a rede digital possa ser uma abstração ideal de todos esses sistemas.

O cientista chega a afirmar que a web é a única rede artificialmente construída que se comporta como uma rede real. Isso abriu um mundo de possibilidades para o estudo de redes complexas naturais, já que resultados obtidos facilmente em estudos de dados digitais poderiam ser analogamente aplicados à outras redes que apresentam naturalmente maior dificuldade na coleta e análise de dados, facilitando a compreensão do seu surgimento, organização e comportamento.

Os avanços da Ciência das Redes podem revolucionar outras ciências que estudam sistemas complexos específicos. Como desconheciam as propriedades que governavam as redes complexas, essas ciências no século XIX e grande parte do século XX reduziram os sistemas estudados às suas mínimas partes na tentativa de compreendê-los. Quase todo sistema complexo conhecido já foi descrito através da enumeração e codificação das suas partes. Até hoje, esse método é utilizado e tem grande importância na ciência, mas não é suficiente para os problemas complexos[27].

[25] Steve Lawrence e Lee Giles, citados em *Linked*, pág. 31.
[26] Em inglês: *small degrees of separation*
[27] Apresentam causa e efeito separados no tempo e no espaço, possuem desdobramentos ainda desconhecidos e imprevisíveis e são vistos de modo diferente por diferentes pessoas (Kahane, 2008)

Cada mapeamento que a ciência faz de um sistema complexo cria uma expectativa na sociedade de que todos os problemas que envolvem esse sistema podem ser resolvidos. Mas há muito os cientistas sabem que esse reducionismo não é suficiente, porque alguns problemas e consequentemente suas soluções, estão nas relações que as coisas estabelecem, e não em suas partes isoladas.

Um bom exemplo disso é o recente Projeto Genoma Humano. Ao identificar todos os genes humanos[28] a ciência e a sociedade esperavam um súbito aumento na eficácia dos tratamentos de algumas doenças. No entanto, o que se viu no ano seguinte foi uma redução de mais de 80% na quantidade de registros de medicamentos comparados com anos anteriores. Essa redução é um alerta de que soluções não foram encontradas nas partes (Barabási, 2012).

O resultado do projeto não é para ser desprezado, mas não é suficiente. Um modo de analisar positivamente a redução dos lançamentos de remédios é que, por causa das informações que ficaram conhecidas pelo sequenciamento genético, a ineficácia de muitas drogas pôde ser apontada antes delas alcançarem o mercado.

Uma doença é um excelente exemplo de problema complexo. Ela pode ser visível em uma determinada parte do sistema, mas a sua causa pode estar em qualquer outra parte não diretamente relacionada porque os *poucos graus de separação* permitem esse fenômeno. Além disso, pode existir mais de uma causa, já que os sistemas reais são totalmente interconectados.

De tanto agir pontualmente em nós que pareciam conter a raiz de um problema e não verem os resultados surgirem como esperados, os cientistas começaram a entender que olhar partes isoladas de um sistema não traz a solução para tudo. Os problemas complexos precisam de uma nova abordagem para serem compreendidos.

Quando as experiências de Milgran e Barabási apontam um número máximo de caminhos para ligar dois nós de uma rede complexa, está implícito o caráter não randômico dessas redes. Ao contrário, o menor caminho jamais seria encontrado se a escolha e a ordem dos nós intermediários fossem aleatórias.

[28] Algo em torno de três bilhões.

A descoberta de que as redes naturais[29] não eram randômicas era o início de uma nova era para a ciência.

A palavra randômica desapareceu da definição dessas redes. Barabási (2002), por exemplo, define redes complexas como aquelas que se reconfiguram continuamente não só através do surgimento e desaparecimento de nós, mas também pela reorganização de links. O cientista percebeu que essa reconfiguração é feita através de duas capacidades que se complementam: a capacidade preferencial e a capacidade competitiva.

A capacidade preferencial de certa forma explica porque *sair na frente* atribui uma vantagem ao nó. Muitas das concentrações de links em um determinado nó de um sistema são explicadas por essa capacidade de ser eleito como preferencial por ter se estabelecido primeiro.

No entanto, Barabási percebeu que novos nós, não raramente, *roubam a cena* e passam a atrair mais links que os antigos. A essa capacidade o cientista deu o nome de capacidade competitiva. Essa característica de um nó explica a dominação de um novo entrante em um ambiente aparentemente estável.

As duas capacidades, preferencial e competitiva, atendem à necessidade de adaptação dos sistemas complexos visando a própria sobrevivência e elevam o nível das incertezas da ciência para um novo patamar: como se dá esse alinhamento entre a necessidade de sobrevivência de um sistema e o comportamento da sua estrutura?

A ciência das redes encontrou ainda outras coisas interessantes sobre os sistemas complexos. Eles estão interligados com outros sistemas complexos, formando em sua representação uma rede de redes. Muitas dessas interconexões ainda são completamente desconhecidas, mas existem indícios de que temos um universo totalmente conectado, mesmo que ainda não existam provas concretas disso.

As conexões inter-redes são feitas por links específicos que chamamos de links pontes ou links fracos. O termo "fraco" pode nos iludir e fazer pensar que sejam pouco importantes, no entanto esses links são fundamentais no estudo das redes. Eles garantem a oxigenação dos sistemas trazendo conteúdo de outros.

[29] Redes naturais (ou reais) é uma expressão utilizada na literatura da Ciência das Redes como sinônimo de sistemas complexos.

Num sistema social, o link fraco é aquele que permite, por exemplo, que encontremos oportunidades (pessoais e profissionais) fora do nosso círculo menor. Num sistema de conhecimento, o link fraco nos garante maiores possibilidades de correlação entre informações de diferentes domínios. Em última instancia, os links fracos aumentam a chance de sobrevivência de um sistema já que são as ligações que garantem a conexão vital com todo ambiente do qual faz parte.

Experiências da internet mostraram também que os links de uma rede complexa além de conteúdo, possuem direção e peso.

O sociólogo Peter Marsden utilizou a expressão *core discussion network*[30] para referir-se ao conteúdo ou tipo de relação que se estabelece entre os nós de uma determinada rede. O conteúdo das relações indica o tipo de rede que estamos analisando. Ou seja, uma rede social pode ser definida pelas relações profissionais, de amizade, de parentesco etc.

Essa dimensão das relações já era bem conhecida mesmo antes das redes digitais. O que os estudos digitais evidenciaram e comprovaram foi a importância da direção e do peso das relações.

A direção refere-se ao caminho que pode ser percorrido na rede. Em uma determinada rede pode-se partir de um nó para outro, mas pode ser impossível fazer o caminho inverso. Exemplos disso são páginas na internet que possuem *links* que levam a *sites* maiores cujo caminho inverso inexiste. Cadeia de predadores, onde o teor alimentar segue numa só direção, e rede de fãs de um determinado artista, em que um fã sabe tudo do seu ídolo e no sentido contrário a relação quase inexiste.

O peso de um link refere-se à intensidade da relação. Numa rede de amizade, o peso se dará pela quantidade e qualidade de ocorrências que fortalecem a relação.

A rigor, uma rede complexa bem representada deve identificar não apenas as quantidades de links, mas também sua direção (geralmente representada através de setas) e peso (representado pela espessura da linha ou números indicativos). Para estudos específicos essa diferenciação pode ser fundamental para se compreender a dinâmica de uma rede.

[30] Apud Caldarelli e Catanzaro (2012)

No entanto, para grande parte dos estudos de redes complexas, a visualização da existência dos links, sem conteúdo, direção ou peso, é suficiente. Nesse caso a representação é simplificada com a utilização de nós de igual tamanho e de linhas de igual espessura e sem setas direcionais.

Embora usemos os termos redes e sistemas como sinônimos, aqui é importante que uma distinção seja feita. O pesquisador Daniel Figueiredo define rede como uma "*abstração que permite codificar relacionamentos entre pares de objetos*[31]" de um sistema real.

Os registros digitais representam uma grande oportunidade para a ciência, porque oferecem acesso inédito aos dados de estrutura e dinâmica de uma rede complexa por excelência, a internet, que codifica os relacionamentos entre pares de objetos de um sistema real, a sociedade.

Além disso, as características e propriedades inferidas desses estudos podem ser de grande utilidade para compreender outras redes complexas conhecidas.

Admitindo a existência de padrões na dinâmica humana, com dados suficientes é possível descrever e quantificar um problema e formular hipóteses. Um problema matematicamente formulado ganha poder preditivo e, a partir daí é possível desenvolver mecanismos de controle que é o objetivo maior de toda ciência (Barabási 2012).

"A ultima prova do nosso entendimento de um determinado sistema, seja ele tecnológico ou natural, está refletida na nossa habilidade em controlá-lo" (Liu, Slotine e Barabási, 2011). *Big data* será capaz de tornar as redes complexas não apenas mais compreendidas, mas também mais previsíveis e mais controláveis.

[31] Essa definição foi retirada do material de aula de Daniel Figueiredo, professor e pesquisador em Redes Complexas da Coppe-UFRJ.

3. Big Data

Você pode chamar de Ciência das Redes, dinâmica humana, computação social ou big data. Não importa o nome que você dê, existe uma quantidade imensa de informações coletadas sobre nós e temos a obrigação de dar um sentido a tudo isso.

Albert-László Barabasi – cientista

3.1. Principais conceitos

O termo *big data* tem sido utilizado para designar diversas coisas. O único ponto em comum entre essas diferentes designações é que todas tratam de grande quantidade de dados digitais.

Entendemos que não existe certo ou errado e que cada definição reflete a abordagem ou o uso que se faz dos dados digitais.

As principais utilizações do termo são:

Big data como banco de dados:

Nessa abordagem identificamos as quatro dimensões críticas do *big data*: volume, variedade, veracidade e velocidade.

A primeira utilização do termo foi feita dentro da comunidade científica, mais precisamente entre os astrônomos. Cientistas da NASA[32] denominaram *big data* os grandes conjuntos de dados complexos que desafiavam os limites tradicionais de captura, análise e armazenagem.

Nessa época, inicio da década de 90, ainda não existiam os dados de dinâmica social, não se falava em rastros digitais nem em internet das coisas. Os cientistas lidavam com os dados gerados nas pesquisas específicas do seu domínio e que eram livremente compartilhadas, o que gerava conhecimentos e novos dados de forma exponencial.

[32] Sigla em inglês de National Aeronautics and Space Administration (Administração Nacional da Aeronáutica e do Espaço). Agência do Governo dos Estados Unidos da América, responsável pela pesquisa e desenvolvimento de tecnologias e programas de exploração espacial.

Outros domínios do conhecimento também acumulavam seus dados coletados no campo de pesquisa.

Estava definida a primeira dimensão crítica do *big data*: o **volume**. Inicialmente esse volume era totalmente representado por dados estruturados que são os dados organizados em estruturas pré-estabelecidas. Por isso possuem o mesmo formato, o mesmo tamanho e seguem a mesma ordem descritiva.

Conforme foram sendo adicionados dados semi-estruturados (textos como um catálogo de produtos ou a descrição de um objeto) e não estruturados (fotos, vídeos, sons), já citados na introdução, surgiu a segunda dimensão crítica: a **variedade**.

Os dados semi-estruturados e não estruturados exigem um esforço maior na descrição porque não possuem unidade de formato e tamanho. Além disso,pelo mesmo motivo, o processo de descrição passa a ser feito a *posteriori*, ou seja após a sua captura.

A dimensão conhecida como **veracidade** é mais fácil de ser verificada quando pensamos nos dados de dinâmica humana que são registrados quando interagimos em redes sociais ou deixamos nossos rastros de navegação. Os dados registrados são fidedignos porque representam interações reais. Outros métodos de captação de dados de dinâmica humana como entrevistas e observação, podem conter desvios próprios do método como ruídos na comunicação, parcialidade, interferência no fenômeno observado, etc. Nem sempre dados digitais de dinâmica humana contemplam o escopo do que se planeja compreender, e outros métodos de coleta são mais pertinentes ou necessários de modo complementar, mas isso não invalida o caráter de veracidade que eles apresentam.

O mesmo raciocínio vale para dados captados por sensores e armazenados automaticamente. Nesse caso, a veracidade estará não apenas no método de coleta, mas também na possibilidade de se captar mais dados, tornando a base mais representativa do fenômeno estudado. Um exemplo clássico é o registro de condições climáticas. Se uma pessoa consulta regularmente os equipamentos de medição (temperatura, umidade, etc) e registra os dados encontrados, jamais conseguirá a quantidade de dados que um sensor digital com armazenagem automática de dados é capaz de captar. Entende-se que nesse caso, uma maior quantidade de dados representa uma maior fidedignidade ao fenômeno.

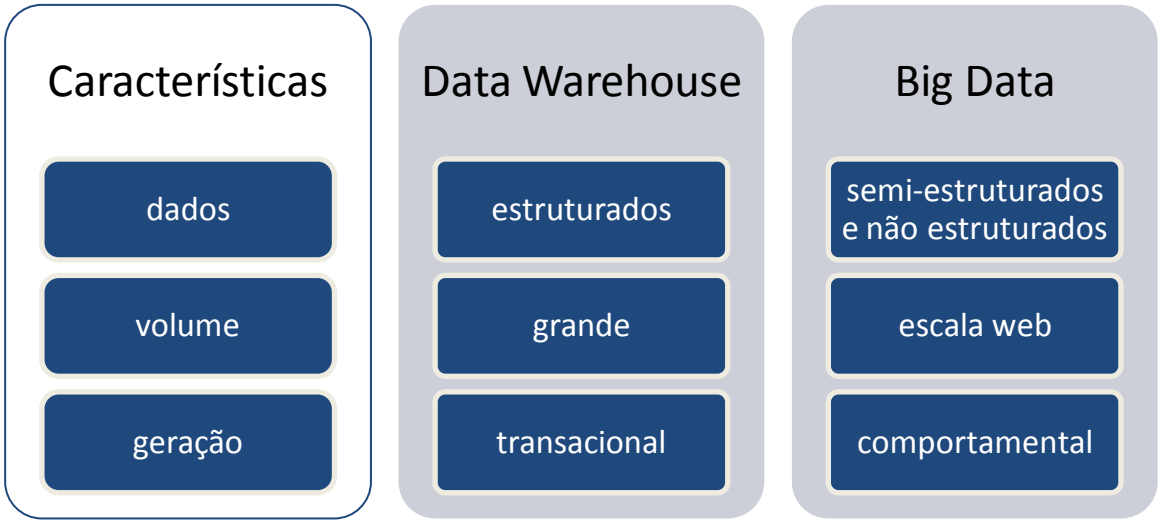
A dimensão mais característica do *big data* é a que a literatura convencionou chamar de **velocidade**[33]**.** Trata-se da possibilidade de se obter dados sobre um determinado fenômeno em tempo real, ou seja, no mesmo momento em que ele ocorre. Isso é fundamental para o valor das informações que podem ser geradas a partir desses dados, já que podem reduzir drasticamente o tempo entre a obtenção dessa informação e a tomada de decisão. Essa dimensão também define o processo de governança de dados, onde novas tecnologias são desenvolvidas para que essa característica fundamental do *big data* não seja perdida. Gerar dados em tempo real só tem valor se a tecnologia permitir sua utilização imediata da cadeia produtiva de soluções.

Essas quatro dimensões ou características (volume, variedade, veracidade e velocidade) formam o **Quatro Vs** do *big data* visto como um banco de dados.

O risco de referir-se a *big data* como um grande banco de dados é confundi-lo com outros bancos de dados digitais de grande porte como *data warehouse* por exemplo. As duas plataformas apresentam características distintas e fundamentais que definem a importância de cada uma na geração de conhecimento e no planejamento das estruturas e processos necessários ao seu pleno funcionamento.

Apresentamos na *figura 3* um resumo esquemático das particularidades dessas plataformas.

Figura 3. Data Warehouse X Big Data [34]



[33] Acreditamos que o termo velocidade (*velocity*) não é o mais adequado porque dá a ideia de rapidez quando se quer destacar a dinâmica (*stream*). Neste trabalho velocidade significa *stream* apenas trazer a referência dos 4 Vs, citados em quase todos textos de introdução a *big data.*

[34] Quadro de nossa autoria

Toda cadeia produtiva de conhecimento a partir de dados digitais se diferencia de uma plataforma e outra. Desde as estruturas de geração, armazenagem, segurança e compartilhamento, até os processos de indexação e análise.

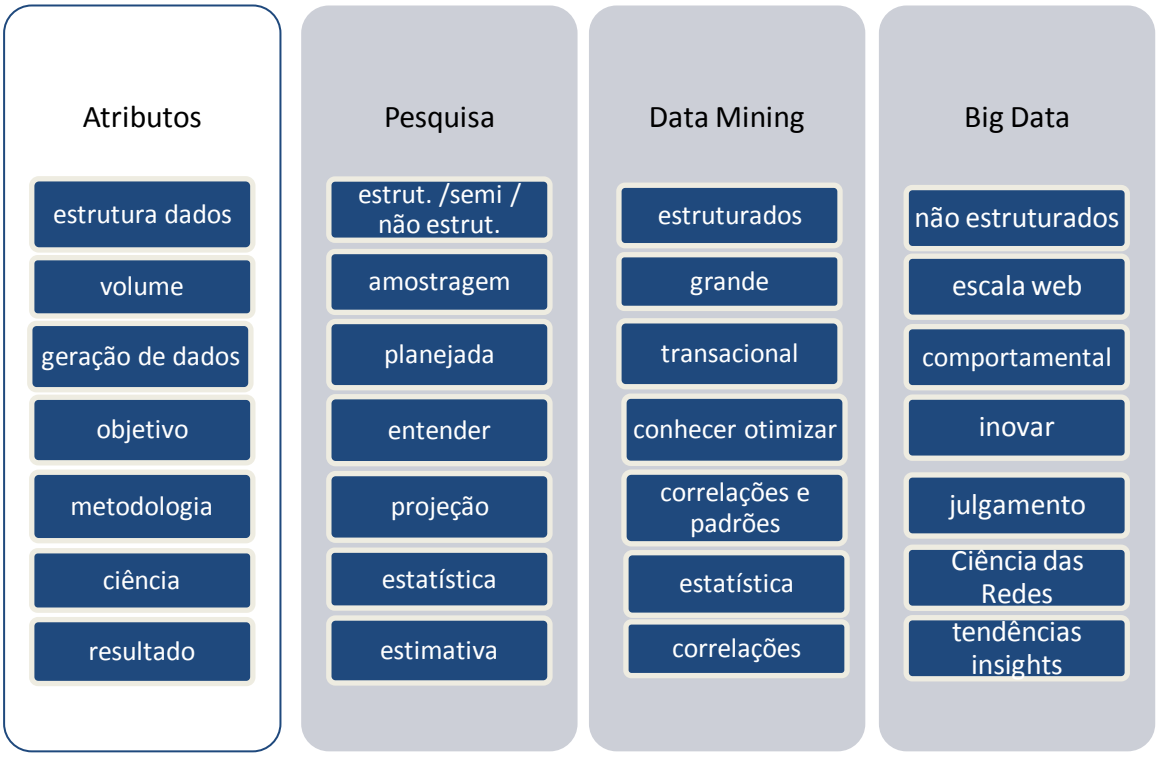
Big data como disciplina

Quando é utilizado para se referir a uma disciplina, o termo *big data* corresponde ao campo empírico de estudo de uma ciência.

A ciência que utiliza dados digitais em suas pesquisas é a Ciência das Redes, que surgiu há pouco mais de 10 anos, conforme explanado capítulo 2 desse trabalho.

A *figura 4* apresenta o resumo dos principais conceitos dessa sessão. O objetivo é destacar o *big data* não para consagrá-lo superior aos demais, mas para deixar claro que o que está em questão não é simplesmente uma nova metodologia ou uma tecnologia mais eficiente e sim uma nova essência que dados digitais passam a apresentar.

Figura 4. Pesquisa X Data Mining X Big Data [35]



[35] Quadro de nossa autoria

A importância de citar a Ciência das Redes nesse trabalho é a de dar ao assunto a seriedade que ele merece. Visto de maneira superficial, sem os fundamentos científicos, *big data* pode se manifestar em modismos, desencadear ações vazias e causar frustração e descrédito.

Por isso acreditamos que teorias, metodologias e ferramentas em *big data* devem estar estritamente fundamentadas na ciência. Aspectos culturais como leis, princípios éticos e posicionamentos estratégicos também encontram na ciência orientação e direcionamento.

Big data como instrumento

Muitas vezes ouvimos falar de *big data* como uma metodologia, tecnologia ou ferramenta capaz de extrair valor[36] de uma grande quantidade de dados digitais.

A melhor maneira para falar de metodologias e tecnologias em *big data* é utilizar a expressão **analítica de dados**[37] que pode ser definida como o uso da tecnologia para a compreensão, comunicação e utilização inteligente dos dados digitais[38].

Existem basicamente quatro tipos de Analítica[39]:

Analítica Descritiva: se encarrega de analisar o que aconteceu;

Analítica Diagnóstica: analisa porque determinado evento aconteceu;

Analítica Preditiva: aponta o que irá (ou poderá) acontecer (predição);

Analítica Prescritiva: identifica o que se deseja que aconteça e o que pode ser feito para alcançar esse objetivo (reforçar, modificar ou evitar uma predição).

A rigor, analítica de dados pode ser feita com qualquer tipo de dados. Como já foi dito antes controlar o ambiente é fundamental para a sobrevivência e todas as analíticas acima contribuem para alcançar esse objetivo. Então desde o principio a humanidade já

[36] Alguns autores consideram o *valor* como mais uma dimensão crítica do *big data,* o quinto **V**. Nesse trabalho optamos por desconsiderar essa dimensão por não ser uma características intrínseca dos dados digitais e sim uma dimensão que se alcança após o esforço de inteligência exercido sobre eles.
[37] Em inglês, *data analytics* ou simplesmente *analytics*
[38] Definição nossa.
[39] Professor José Carlos Cavalcanti (UFPE) http://creativante.com/new/index.php/2013-02-03-19-36-05/2013-02-04-18-19-49/155-big-data-e-analitica-ferramentas-e-modelos-de-negocios-revolucionarios

fazia analítica utilizando o cérebro como ferramenta e a experiência e intuição como metodologias.

De modo igual, nas organizações, mesmo antes da existência dos computadores os gestores já praticavam os quatro tipos de analíticas citados utilizando tabelas, análise de cenários, simulações, *brainstorms* e também muita experiência e intuição.

O processo é o mesmo em todos os contextos, só mudam as ferramentas e a metodologia.

Em *big data* os quatro tipos de analíticas são possíveis embora os maiores valores agregados aos dados estejam nas duas últimas (preditiva e prescritiva) já que as duas primeiras (descritiva e diagnóstica) já eram amplamente utilizadas e com sucesso antes do fenômeno *big data.*

Nesse ponto, aparece a primeira das importantes diferenças básicas entre os dados em *big data* e bancos de dados estruturados de grande porte.

Quando falamos em analítica preditiva, estamos falando de uma metodologia completamente diferente das previsões que estamos acostumados a ver a partir de *data warehouses* ou de pesquisas de opinião. O que determina a mudança na metodologia é a natureza dos dados utilizados.

A previsão (ou *forecast*) é feita a partir da utilização de grandes quantidades de dados de mesma natureza. A metodologia consiste em uma projeção, em que os dados viajam para dimensões maiores de tempo ou espaço.

O exemplo mais conhecido disso é a previsão do tempo. Dados de um mesmo domínio, que têm uma relação direta de causa e efeito com o fenômeno, são transferidos para o futuro (viagem no tempo) gerando um resultado com grau de incerteza conhecido.

Outro exemplo é a pesquisa de opinião. Dados de uma mesma natureza, coletados em uma amostra, são transferidos para o total da população (viagem no espaço) gerando também um resultado confiável.

O resultado de uma previsão é uma estimativa e a ciência que fundamenta a metodologia é a Estatística.

A predição (analítica preditiva ou *predictive analytics*) é feita a partir de uma grande quantidade de dados de domínios diferentes (*big data*). A metodologia é um julgamento baseado na experiência e no aprendizado[40] que se dá quando dados de um domínio do conhecimento viaja para outros domínios.

Em 2008, por exemplo, o Google conseguiu antecipar em 7 a 10 dias, a dinâmica da contaminação pelo vírus da gripe H1N1, baseado nos dados de utilização da ferramenta de busca dos seus usuários. Rastreando o conteúdo das buscas, a empresa conseguiu identificar correlações inusitadas entre as pesquisas e a contração da doença em um determinado grupo social. É importante entender que o modelo antecipava não apenas a quantidade de casos, mas também o deslocamento do vírus. O modelo foi tão bem sucedido que fundamentou as ações preventivas do governo americano para impedir uma pandemia da doença.

Diferente da previsão, a analítica preditiva é capaz de antecipar as mudanças nas tendências porque considera fatores que não têm uma relação causa-efeito imediata com o fenômeno.

O resultado de uma analítica preditiva é um *insight* e a ciência que fundamenta a metodologia é a Ciência das Redes.

Afinal, o que é *big data*?

Para nos referir aos instrumentos, às metodologias de geração de valor a partir do *big data*, utilizaremos a expressão **analítica de dados.**

A utilização do método científico para gerar conhecimento a partir dos dados digitais será chamada **Ciência das Redes.**

Considerar *big data como* um grande banco de dados digital não alcança sua verdadeira essência, conforme podemos apreender nas palavras de Peter Norvig, pesquisador[41] de inteligência artificial na empresa Google.

[40] Predição pode ser feita com dados digitais ou não. No contexto de dados digitais (*big data*), experiência e aprendizado correspondem ao auto ajuste do algoritmo feito por inteligência artificial.
[41] apud Schönberger e Cukier, 2013

Na impossibilidade de definir *big data*, Norvig faz uma analogia com imagens. Para ele, um desenho na caverna, uma pintura ou uma fotografia, embora sejam obtidos através de diferentes técnicas, são essencialmente a mesma coisa: uma imagem. No entanto, diversas imagens capturadas e exibidas em velocidade transformam-se em um filme. E um filme é essencialmente diferente de uma fotografia.

A conclusão é que sempre que algo muda significativamente em suas dimensões, muda também em sua essência. Se fosse apenas uma coleção de dados digitais, *big data* seria como uma coleção de fotografias. Mas as tecnologias de comunicação, armazenagem e processamento mudam as dimensões dos dados de modo tão desproporcional a tudo que já havia antes, que inevitavelmente gera algo de outra essência.

O potencial revolucionário do *big data* é a sua própria definição. A dimensão real desse potencial nunca será conhecida porque seria o equivalente a colocar um limite nas capacidades humanas de conhecer e inovar.

O que nos leva à resposta de Bill Franks, CAO (*Chief Analytics Officer*) de uma grande empresa de analítica de dados digitais, que respondeu da seguinte maneira quando perguntado sobre qual era a definição de *big data:"Definition of big data? Who cares? It is what we are doing with it".*[42]

Embora não seja uma definição, a resposta de Franks é de extrema sabedoria porque capta exatamente a essência do *big data* como desencadeador de ações. Afirma que *big data* só tem importância quando usado e como só pode ser usado para gerar conhecimento, coloca o ser humano onde ele deve estar nessa discussão: no comando. A resposta tangencia ainda a questão da ética na utilização desses dados. Bom ou ruim, para o bem ou para o mal, *big data* de fato será o que nós fizermos com ele.

3.2. Ecossistema

Uma das principais colaborações da Ciência das Redes para as organizações que desejam desenvolver metodologias e ferramentas em *big data* vem da afirmação de que devemos ver os dados como parte de um ecossistema. Isso quer dizer que não deve existir uma abordagem a partir de uma única perspectiva e que todos os participantes da

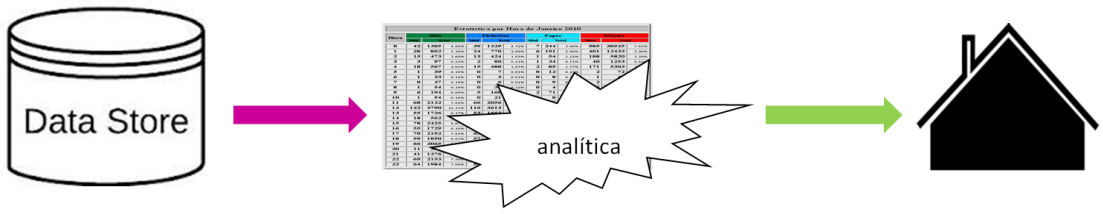
[42] www.opentracker.net/article/definitions-big-data

rede deverão ser levados em conta nos estudos das ameaças e oportunidades na modelagem de novos empreendimentos.

Em consistência com a diversidade dos dados existentes é necessário que a abordagem metodológica não seja prerrogativa de uma única ciência. Os dados que temos são universais, gerados por todos (usuários ou não da internet) e com um potencial de inovação que não deve respeitar fronteiras geográficas ou de domínio do conhecimento.

Diante dessas afirmações, e concordando com elas, aceitamos o desafio de esquematizar o que seria um ecossistema *big data*. A descrição mais básica de um sistema de geração de conhecimento a partir de dados digitais seria: a partir de dados armazenados, aplicamos tecnologias de processamento cujas informações geram insights para os tomadores de decisão. O esquema ilustrativo está representado na *figura 5*.

Figura 5. Ecossistema: descrição básica

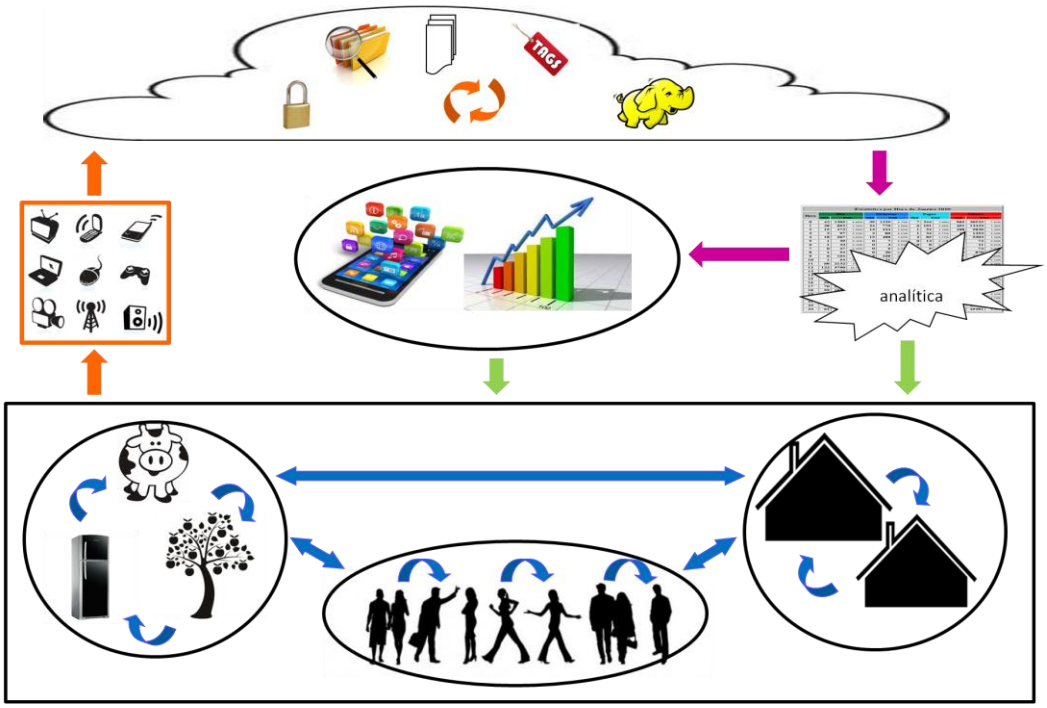


No entanto, se fizermos perguntas simples para compreendê-lo um pouco mais, vamos agregando diversos outros agentes que evidenciam maior complexidade ao sistema. Quem gera os dados do sistema? Como eles são gerados? Quais os seus conteúdos? Como são enviados para armazenagem? Quem decide o que tem valor? Que valor os dados geram e para quem?

A *figura 6* mostra apresenta o ecossistema *big data* com um maior grau de complexidade[43]. É importante ressaltar que, como se trata de ambiente *big data*, estão representadas na figura exclusivamente as relações digitais entre os agentes.

[43] Por se tratar de um trabalho original e subjetivo, acreditamos que outras contribuições deverão ser feitas para melhorar seu conteúdo e visualização.

Figura 6. Ecossistema *big data*



No esquema acima estão representados os seguintes agentes e relações:

- Geração de dados: representados na parte baixa do desenho pelas silhuetas pretas. Os agentes geradores de dados são pessoas, organizações, objetos, animais e plantas. Os dados são gerados a partir das relações (setas azuis) que os agentes estabelecem uns com os outros. Os dados podem ser comportamentais (da relação entre pessoas), transacionais (das relações entre empresas e entre elas e pessoas) e de localização, de utilização, de condição física, química e funcional etc (gerados por objetos, animais e plantas e suas relações através da internet das coisas).

- Codificação de dados: são os agentes responsáveis pela modificação da informação com o objetivo de deixá-la em condições de transmissão e armazenamento digital.

- Captação de dados: são os computadores, *tablets*, celulares, GPS, radares e sensores de todos os tipos. Esses agentes dataficam as informações do ambiente a partir de códigos previamente definidos.

- Transmissão de dados digitais: esses agentes são responsáveis pela mobilidade dos dados.

  Nesse trabalho, para feito de melhor visualização do ecossistema, agentes de codificação, captação e transmissão de dados estão representados juntos na caixa com contorno cor de laranja. As setas laranja representam relações de mobilidade, que são parte das relações de governança dos dados.

- Armazenagem de dados: o ecossistema *big data* pressupõe armazenagem em nuvem para atender às necessidades da dimensão "velocidade" dos dados, apresentada no capítulo 3, página 18. Estão representados no desenho pela nuvem.

- Apoio ao processo de armazenagem: representados pelos ícones dentro da nuvem, são agentes que prestam serviços de *tagging*, indexação, cópia, integração, segurança e compartilhamento. Usamos duas setas formando um círculo para indicar interelação entre todos esses agentes de governança.

- Processamento de dados: representado pela planilha de dados. No caso do *big data* são as ferramentas de analítica de dados que extraem informações dos dados digitais. As setas verdes indicam relações de entrega dos resultados da analítica para os tomadores de decisão. Essa entrega pode passar ainda por outros agentes que agregam mais valor à informação gerada. São eles os designers e os desenvolvedores de aplicativos.

- Design de resultados de analítica: criadores de soluções para visualização de dados (representados pelo gráfico colorido no meio do desenho). Esses agentes dão um *aproach* mais concreto e intuitivo à analítica aumentando a sua acessibilidade por parte de profissionais que não são da área de TI.

- Desenvolvimento de interface para analítica: representados pelo celular com os ícones de aplicativos, esses agentes colocam todo processo analítico nas mãos de usuários finais (pessoas comuns ou gestores fora da seara da tecnologia da informação). As setas rosa indicam relações de agregação de valor aos dados digitais.

- Utilização de dados: consideraremos como tomadores de decisão não apenas gestores de organizações, mas também qualquer pessoa, institucionalizada ou não, que utilize a analítica de dados para gerar soluções que transformem uma situação real em uma situação desejada. Ampliando um pouco mais o conceito, podemos considerar a Internet das Coisas também como agentes de utilização de dados. Nesse caso, se não há o elemento humano agindo na transformação, consideramos o conhecimento como o produto da inteligência artificial e a tomada de decisão como resultado de um processo de automação.

Os nós e links acrescentados nessa *figura 5* já existiam implicitamente na figura anterior, mas a representação explicita nos obriga e vê-los e estudá-los. Em todo ecossistema existem oportunidades.

3.3. Ciclo de adoção de tecnologia[44]

Como vimos, os primeiros estudos que despertaram o mundo para a utilização dos dados gerados em rede foram feitos há mais de 10 anos. De lá para cá muita pesquisa foi feita, muita tecnologia foi desenvolvida e várias iniciativas foram tomadas para gerar soluções a partir de *big data*. No entanto os resultados, embora alguns muito bons, ainda não representam todo potencial da tecnologia.

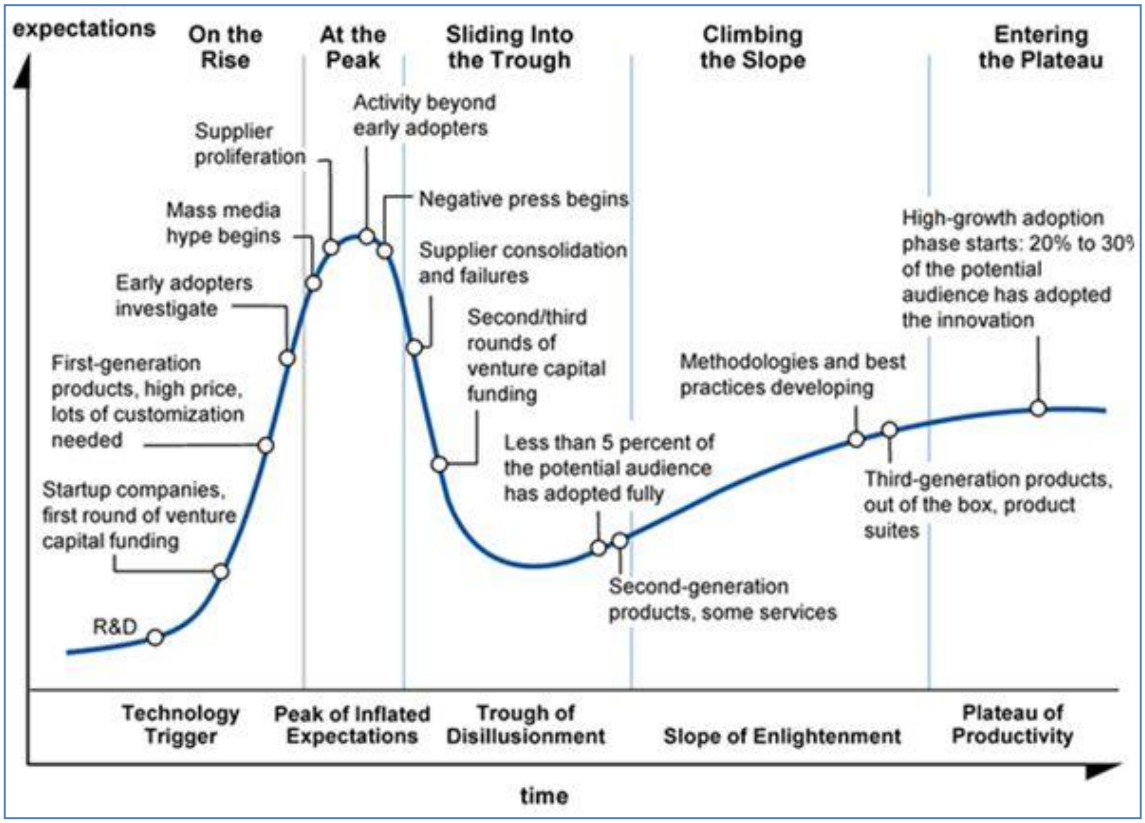
Segundo o Instituto Gartner toda tecnologia percorre um ciclo semelhante desde que sai dos laboratórios de pesquisa e desenvolvimento (P&D) até atingir a maturidade identificada como a fase em que a tecnologia atende plenamente à sociedade.

Para facilitar o posicionamento e permitir a visualização da dinâmica de cada tecnologia em direção à maturidade, o Gartner criou o *Technology Hype Cycle Graph*. Como

[44] Hype Cicle Graph do Gartner Institute

podemos ver na *figura 7*, ao percorrer esse ciclo uma nova tecnologia passa pelas seguintes fases:

Figura 7. Technology Hype Cycle Graph (Gartner Institute)



Inflação das expectativas: nessa fase, justificada pela novidade e pelo desejo de soluções inovadoras, a nova tecnologia recebe os primeiros financiamentos de investidores visionários, consegue ser adotada pelas empresas que lideram a inovação no seu setor de atuação e recebem uma atenção incentivadora da mídia.

Pico das expectativas: nesse ponto conquista industrias dispostas a suprir a tecnologia propiciando a sua adoção por seguidores próximos dos inovadores. Com casos para serem analisados e massa crítica formada dá-se início a fase de análise dos primeiros resultados e ponderação das promessas. Nesse ponto a mídia coloca em pauta o lado negativo da tecnologia.

Poço da desilusão: na verdade essa fase pode ser um ajuste entre expectativa e realidade, mas sempre arrasta junto os mais desavisados que pensam tratar-se de uma

desqualificação da tecnologia. Ao atingir o ponto mais baixo do ciclo a tecnologia conquistou apenas 5% do potencial inicialmente projetado. Insatisfeitos com os resultados, entusiastas da tecnologia buscam corrigir os rumos tomados. Surge a segunda geração de aplicações e a possibilidade de fazer diferente oferece novas oportunidades de negócio. Então a tecnologia entra na fase de escalada em direção à maturidade.

Platô de produtividade: nessa fase a tecnologia estará na terceira geração de aplicações, já possui maturidade de metodologias, processos e indicadores de sucesso e atinge em torno de 30% do seu mercado potencial.

O Instituto Gartner sugere que toda tecnologia[45] percorre esse caminho teórico. O que diferencia é o tempo decorrido nesse processo.

A *figura 8* mostra a evolução das tecnologias em *big data*. Os pontos marcados na curva para os anos de 2011, 2012 e 2013 foram retirados dos relatórios do instituto. Segundo o Gartner, *big data* chegou em 2013 no pico das expectativas, o que significa dizer que 2014 será o ano de ajuste à realidade.

Figura 8. Big Data Hype Cycle Graph

[45] Com exceção das que entram em obsolescência antes de atingir a maturidade

3.4. Curva de Aprendizagem

Uma vez adotada, uma nova tecnologia apresenta uma curva de aprendizagem que pode representar uma barreira para o alcance do potencial previsto.

Curva de aprendizagem é um conceito que representa o nível médio cognitivo de aprendizagem para uma determinada atividade ou ferramenta[46], e maneira mais simples de entender isso é através da metáfora de um jogo.

Quando decide participar de um jogo novo, quanto tempo vai levar até que o jogador aprenda as regras, os movimentos básicos, domine as sutilezas, tenha interesse de interagir com outros jogadores, desenvolva suas próprias táticas, para começar a ficar satisfeito com o seu desempenho?

Neste trabalho, consideraremos na curva de aprendizagem todo o caminho percorrido, de modo consciente ou não, pelas organizações que já obtiveram retorno satisfatório sobre o investimento em projetos *big data*. O objetivo é identificar quais as características comuns das organizações que já atingiram o ponto da curva onde o desempenho compensa o investimento.

Para isso utilizaremos os resultados de duas pesquisas recém conduzidas para identificar características comuns às organizações que haviam incorporado *big data* em suas estratégias competitivas.

Uma delas foi uma pesquisa feita pela Tata Consultancy Services[47] em janeiro de 2013.

Desta pesquisa apresentaremos a análise das 276 empresas que estimaram um retorno de mais de 25% no investimento em *big data* em 2012 que aponta os seguintes fatores críticos da curva de aprendizagem:

Internet Centric. Quanto uma organização esteja centrada na internet, mais adiantada ela estará na curva. Existem quatro grupos de empresas que lideram nesse item:

- organizações que se digitalizaram por imposição transacional (como bancos e grandes varejistas);

[46] Tools and Techniques for Strategic Management. www.tools-and-techniques.24xls.com/pt104

[47] TCS (Tata Consultancy Services) entrevistou 1,217 empresas em nove países (EUA e países da Europa, Ásia e América Latina) em dezembro de 2012 e janeiro de 2013. 643 (53%) afirmaram que já haviam feito investimentos em Big Data, sendo que 276 delas já esperavam retorno sobre esses investimentos. Disponível em http://blogs.hbr.org/cs/2013/06/what_the_companies_winning_at.html

- organizações cujo *core business*[48] evoluiu naturalmente para tecnologias digitais (como as empresas de telecomunicação);
- organizações cujo modelo de negócio incluía desde o inicio a internet (como a Amazon);
- e finalmente aquelas que tem o digital e a internet como *core business* (Google, Facebook e Twitter).

Embora a longevidade de atuação na rede seja importante para esse fator, não é uma variável decisiva. O que define a ordem das organizações melhores colocadas na curva é o tipo de relação que elas têm com a internet. Sendo assim, a ordem dos tipos de organizações que apresentarão melhores resultados utilizando *big data* é, naturalmente, a ordem inversa da citada acima.

Abrangência de domínios na busca de dados: Esse fator é a comprovação da importância da diversidade de dados no valor do *big data*. As organizações que lideram nesse ambiente são aquelas que conseguem gerar conhecimento a partir de dados das mais diferentes origens, coletados para os mais diferentes fins. Assim, dados que aparentemente não se relacionam com o negócio da empresa podem ser definitivos para o sucesso dos projetos. Dentre os mais utilizados podemos citar os dados de redes sociais e os de geolocalização. Essa agressividade na abrangência de domínios implica também na capacidade de lidar com uma quantidade cada vez maior de dados não estruturados.

Abrangência na aplicação: Em última análise *big data* é utilizado para ajudar na tomada de decisões. No entanto isso é muito abrangente porque decisões podem ser tomadas nas mais diversas áreas de atuação de uma empresa. A pesquisa detectou que empresas líderes utilizam *big data* para aumento das vendas, reposicionamento de produtos, pesquisa e desenvolvimento, prospecção de novos negócios, melhor experiência do cliente em lojas físicas, monitoramento da concorrência, monitoramento da percepção da marca, monitoramento do desempenho de seus produtos (durabilidade, eficiência etc), melhorias na logística etc.

***Big data* em "casa própria":** Uma das conclusões mais interessantes da pesquisa é a de que 80% das empresas líderes em *big data* empoderam um profissional específico para esse fim dando assim uma "casa própria" ao projeto. Na prática isso significa dizer que

[48] Principal atuação de uma organização.

essas organizações têm como responsáveis pelo projeto um profissional que não pertence à área de TI e tão pouco é o tomador de decisão. De fato, profissionais de TI, embora dominem as ferramentas para analítica de dados não colocam projetos de *big data* em perspectiva nem em consonância com a estratégia organizacional. Em poucas palavras, não sabem o que procuram, portanto não entendem o que encontram[49]. Por outro lado, na seara do tomador de decisão as iniciativas em *big data* são pressionadas a confirmar as ações que já estão sendo feitas ou as decisões que se deseja tomar. O objetivo desse profissional é fazer *big data* ser um ativo nuclear em toda a organização gerando valor em toda cadeia da informação.

A segunda pesquisa foi feita pelo MIT Sloan[50] para conhecer melhor as empresas que estão conduzindo a revolução em analítica de dados.

Entre outras informações e análises trazidas pela pesquisa, destacamos a que foi feita com as organizações que já estão adiantadas na curva de aprendizagem da tecnologia, A partir da abordagem que têm com *big data* os autores da pesquisa hierarquizaram essas organizações por sua *sofisticação analítica*.

Innovators: A mais sofisticada categoria é a dos **inovadores** que usam analítica de dados para inovação e ganham vantagem competitiva com isso. Correspondem a 11% dos pesquisados. Esse grupo possui 3 características chave:

- *Mindset*: veem dados digitais como ativo fundamental, utilizam analítica em diversas atividades e estão abertos a novas maneiras de pensar que desafiem o *status quo*.
- Ação: são mais efetivos em toda cadeia de valor de *big data (*capturam mais dados de forma inteligente, integram melhor os dados coletados, usam maior quantidade dos dados coletados e usam os resultados para gerar insights para tomada de decisão).
- Resultados: analítica de dados tem forte impacto na estrutura de poder da organização. Aqueles que têm habilidade de gerar valor a partir de dados participam de certa dinâmica no poder de inovar e decidir.

[49] Referência à frase célebre do pesquisador canadense (de origem húngara) Hans Selye sobre a importância do método: "Quem não sabe o que procura não entende o que encontra".
[50] MIT Sloan em parceria com a SAS Institute Inc. Publicado em MIT Sloan Management Review, Research Report, Spring 2013. Disponível em http://sloanreview.mit.edu/reports/analytics-innovation/

Practitioners: São os **pragmáticos** e representam a maioria dos entrevistados (60%). Inclui aqueles que já fizeram um significante progresso no uso de analítica e veem os resultados disso. No entanto, não usam dados digitais para inovar e e sim para ganho de eficiência o que não se traduz em grande vantagem competitiva. As principais características são:

- Baixa ambição na coleta de dados: esse grupo possui uma série histórica relevante, acessível e confiável, mas restringem-se basicamente a dados transacionais.
- Analítica focada no operacional: utilizam mais a analítica descritiva para conhecer melhor o negócio e aumentar a eficiência de práticas do dia-a-dia como reduzir custos, melhorar a alocação de recursos, melhorar fluxo da cadeia de suprimentos etc. A analítica preditiva e prescritiva continua se apoiando exclusivamente na capacidade intuitiva dos gestores.
- Ecossistema analítico fragmentado: a analítica tem valor diferente em diferentes pontos da cadeia de valor da informação. Muitas vezes a ponta que tem maior potencial de utilização dos insights da analítica não recebe a informação gerada.

Challenged**:** a terceira categoria de sofisticação analítica é aquela dos que ainda são **desafiados** pelo ambiente de inovação em *big data*. Possuem menos maturidade no uso de dados digitais o que significa dizer que não os utiliza para inovação. Para os pesquisadores essa atitude explica o fato desse grupo não ter alcançado vantagem competitiva com o uso do *big data*. Esse grupo comporta 29% dos entrevistados. As principais características desse grupo são:

- Deficiência de dados: essa característica significa baixo aproveitamento dos dados por sua inadequação ou por falta de habilidade para utilizá-los da melhor maneira
- Cadeia de valor deficiente: capturam e dados em grande quantidade sem um plano específico para eles; armazenam dados em silos que não se comunicam.
- Falta de colaboração: baixa propensão em compartilhar dados, pelo desconhecimento da sua importância. Essa característica se manifesta na tecnologia e nas pessoas desse grupo, ambas despreparadas para o compartilhamento (reflete a cultura da empresa)

- Falta de incentivo: resistência cultural à mudança pelo sucesso de métodos mais tradicionais ou baixa competitividade criam um ambiente que não incentiva a inovação.

Comparando os resultados das duas pesquisas podemos identificar que os principais desafios estão em barreiras culturais, mais que em barreiras tecnológicas ou financeiras.

As empresas que já estão numa posição confortável na curva de aprendizagem já quebraram o paradigma de abordagem da analítica digital, ou já nasceram no novo paradigma.

A nossa conclusão sobre os resultados dessas pesquisas é que a curva de aprendizagem em *big data* é o processo de sofisticação da compreensão dessa tecnologia que, de modo consciente ou não, vai se aproximando dos princípios da Ciência das Redes.

4. O Mercado para *Big Data*

Não é escassez, é falta de acessibilidade. Com a lente da tecnologia poucos recursos são realmente escassos

Peter Diamandis - XPrize

4.1. Global

O mercado global de *big data* em 2012 foi de US$ 6,3 bilhões, segundo um relatório da TRM[51] publicado em janeiro de 2013. Para chegar a esse número o estudo incluiu o volume movimentado por empresas demandadas para atuar em gestão de grande quantidade de dados digitais (armazenagem, *hardware*, *software* e serviços) e considerou os setores de serviços financeiros, saúde, governo, indústria, telecomunicações, varejo e serviços de mídia e entretenimento. Segundo os números apurados os três primeiros responderam por 55% do mercado de *big data* em 2012, mas

[51] *Big Data Market – Global Scenario, Trends, Industry Analysis, Size, Share and Forecast, 2012 – 2018* da Transparency Market Research (TMR), empresa de inteligência em marketing global.

mídia e entretenimento e saúde serão os líderes da taxa de crescimento nos próximos anos.

Sob a ótica geográfica, o estudo dividiu o mercado em quatro grandes áreas: América do Norte, Europa, Ásia-Pacífico e Resto do Mundo.

A América do Norte concentra mais da metade do mercado e deverá continuar dominando o *share* dos próximos anos. No entanto, a Ásia-Pacífico apresentará o crescimento mais acelerado no mercado, em torno de 43% ao ano, nos próximos 5 anos.

O relatório ainda estima que até 2018 o mercado deverá atingir o volume total de US$ 48,3 bilhões. Isso significa uma taxa de crescimento de quase 40% ao ano.

Ainda segundo a TRM, o crescimento desse mercado só não será maior porque haverá um gargalo significativo na oferta de cientistas de dados e outros profissionais com habilidade para trabalhar em ambiente *big data*.

Para se ter uma ideia, em um estudo publicado em 2011 a McKinsey estimou que em 2018 somente os Estados Unidos teriam uma demanda reprimida por 140 a 190 mil profissionais especializados em analítica de dados e por 1,5 milhão de gestores com habilidade para tomar decisões e gerir projetos em *big data*. Claro que a projeção foi feita considerando inalteradas as atuais taxas de formação desses profissionais.

4.2. Brasil

No Brasil, alguns setores da economia apresentam uma boa prontidão para utilização de *big data* como, por exemplo, as redes de varejo, instituições financeiras, companhias de telecomunicações, fornecedoras de energia e serviços públicos. Esses setores tem em comum um longo tempo de operação do seu *core business* em bases digitais e por isso tem uma grande quantidade de dados transacionais em seus bancos de dados.

Em relação às setores corporativos, o marketing está entre os primeiros interessados. Isso se justifica pelo fácil entendimento e rápida aplicação dos modelos preditivos gerados pelo *big data* em ações mercadológicas mais eficazes.

Apesar da boa base de dados existente, identificamos que o Brasil vem seguindo os Estados Unidos a uma distância de pelo menos cinco anos. Uma simples pesquisa

mostra que em 2004 já havia interesse suficiente nos Estados Unidos pelo termo *big data* para justificar o registro pela ferramenta Google Trends. No Brasil esse interesse geral só foi captado pela ferramenta em 2009.

Quanto à adoção da tecnologia, inferimos um gap ainda maior utilizando a mesma ferramenta do Google. Enquanto nos Estados Unidos as pesquisas pelos termos *big data* e *Hadoop* (maior e mais importante plataforma em nuvem para analítica de dados existente) apresentam números colados desde 2007, indicando um interesse pela aplicação da tecnologia, no Brasil, mesmo em 2013, o interesse pelo termo *Hadoop* corresponde a menos de 20% do interesse pelo termo *big data*. Associado ao crescimento das buscas pelo termo *big data* no Brasil encontramos fortes picos de interesse pelos termos "privacidade" e "espionagem".

Embora não haja prova da correlação nos índices apurados, não deixa de ser curioso notar que a quantidade de buscas por *big data* no Brasil esteja acompanhada de perto pela quantidade de buscas por "privacidade" e "espionagem" e bem de longe pela quantidade de buscas sobre a plataforma de compartilhamento e analítica. Isso nos parece ser um indicativo de que o conceito *big data* já desperta bastante interesse, no entanto poucos estão de fato buscando meios de colocá-lo em prática.

Outra grande defasagem brasileira em relação ao mercado americano está no que diz respeito a políticas públicas.

Em 2012, o governo americano criou o programa *Big Data Research and Development Initiative* com o objetivo de incentivar pesquisas em tecnologia e formação de profissionais para utilização de *big data*.

No Brasil não existe nenhum sinal de política pública para incentivo do tipo. Todas as iniciativas restringem-se à computação em nuvem dentro da administração publica e mesmo assim ainda esbarram em questões burocráticas como a inadequação das regras de compras publicas que impedem os investimentos necessários.

De qualquer modo, o mercado brasileiro de *big data* movimentará em 2013 em torno de US$ 285 milhões segundo a IDC[52]. Isso significa 3% do mercado global. Portanto, mantendo-se todas as previsões da TRM e da IDC, e mantendo sua participação no mercado global, em 2018 o mercado brasileiro deverá estar próximo de US$ 1,4 bilhão.

[52] *International Data Corporation*, empresa de inteligência em marketing global.

A taxa de crescimento deverá ser de 46% ao ano, bem superior à taxa global que foi estimada em 40%.

A consultoria faz a ressalva de que os projetos em *big data* no Brasil ainda restringem-se mineração de dados transacionais para melhorar a eficiência das organizações. Pouca coisa ainda é feita em inovação e com dados de diferentes domínios.

Nos setores com maior prontidão já estão atentos para a necessidade de investimentos em analítica de dados. No entanto, mesmo nesses casos existe uma concentração de esforço em ações na seara tecnológica.

Provavelmente, o gap que se formará entre a tecnologia e a capacidade de transformar analítica digital em soluções responderá por boa parte da força que puxará os pioneiros em projetos *big data* para o poço da desilusão.

4.3. Análise Estratégica para o Brasil - Matriz SWOT

Para compreender e visualizar melhor os desafios do Brasil na utilização efetiva dos dados digitais utilizamos a ferramenta SWOT.

A análise SWOT[53] é amplamente utilizada em processos de planejamento estratégico e análise de ambientes de negócios.

A ferramenta consiste na elaboração de uma matriz de fatores que podem interferir nos resultados de um determinado empreendimento. Esses fatores são identificados em dois ambientes:

- interno (forças e fraquezas), que são aspectos positivos ou negativos, vantagens e desvantagens de um projeto ou uma empresa;
- externo (oportunidades e ameaças), que são perspectivas de mercado, do ambiente no qual o projeto ou empresa estão inseridos.

[53] Sigla em inglês das palavras Strengths, Weaknesses, Opportunities e Threats (Forças, Fraquezas, Oportunidades e Ameaças), que formam a matriz de análise estratégica.

A *figura 9* a seguir representa uma proposta de matriz para analise estratégica de *big data* para o Brasil.

Figura 9. Matriz SWOT

MATRIZ SWOT BIG DATA[54]		INFLUÊNCIAS	
		POSITIVAS	**NEGATIVAS**
ORIGEM DO FATOR	**INTERNA**	**FORÇAS** • Qualidade da base de dados (1) • Capacidade de armazenagem e segurança de dados (1) • Tecnologia da Informação e Comunicação bem desenvolvida (2) • Profissionais com alto nível técnico (2) • Capacidade de criação de modelos de negócio e de reestruturação de processos(2) • Tomadores de decisão interessados em *big data* • Aumento dos investimentos em TIC (4)	**FRAQUEZAS** • Dificuldade de descrever e analisar dados não estruturados (1) • Grande parte do database em silos (1) • Baixa propensão em compartilhar dados (1) • Baixa especialização profissional (4) • Pouca massa crítica para discutir o tema (1) • Tomadores de decisão pouco familiarizados com o potencial do *big data* • *Big data* ainda muito centralizado em TI (7)
	EXTERNA	**OPORTUNIDADES** • Alta penetração da telefonia móvel (3) • Forte adesão às Redes Sociais (5) • Efetiva participação on-line (2) • Cultura favorável à inovação (2) • Empreendedorismo (6) • Boa taxa de adesão à tecnologias digitais (1) • Serviços digitais móveis são considerados essenciais para a inclusão social (3) • Bons Institutos de Pesquisa e Desenvolvimento (2) • Boa taxa de importação de alta tecnologia (2) • Boa taxa de crescimento da penetração da rede móvel 3G (3)	**AMEAÇAS** • Baixa taxa de penetração da rede móvel 3G (60 milhões) (3) • Custo da comunicação móvel • Falta política pública p/ investimento e normatização em *big data* • Preocupação excessiva com a privacidade na utilização dos dados • Dificuldade de abertura de pequenas empresas (2) • Complexa estrutura de impostos (2) • Burocracia na regulamentação para instalação de redes (3) • Parcerias governo X universidades X empresas(2) • Poucas publicações em português (1) • Baixa proficiência em português (leitura), matemática e engenharia (2)

[54] **Fontes**: **1**. Dedução do grupo a partir da bibliografia lida / **2**. Global Index Innovation / **3**. Observatorio Movel Brasil / **4**. Instituto Gartner / **5**. eMarketer / **6**. IBOPE

Como o objetivo desse trabalho é identificar os desafios do Brasil para a utilização de *big data* como tecnologia cognitiva de inovação e apoio à decisão, consideraremos nessa matriz SWOT:

- ambiente interno: fatores diretamente envolvidos na cadeia produtiva de conhecimento a partir de dados digitais, ou seja os fatores orgânicos e estruturais do ecossistema *big data*.
- ambiente externo: fatores sociais, culturais, políticos e econômicos que podem impactar o ambiente de negócio em *big data*.

Figura 10. Matriz SWOT Cruzada

SWOT CRUZADA		**AMBIENTE INTERNO**	
		FORÇAS	**FRAQUEZAS**
AMBIENTE EXTERNO	**OPORTUNIDADES**	• Tecnologia para captação de dados; • Tecnologia para analítica de dados não estruturados; • Modelos de negócios em ambiente *Big Data;* • Metodologias; • Indicadores.	• Desenvolvimento de tecnologia para compartilhamento de dados; • Desenvolvimento de ferramentas para tomadores de decisão; • Eventos sobre o tema.
	AMEAÇAS	• Novas tecnologias para ampliação da rede; • Especialização de pessoas; • Publicações / traduções; • Consultoria para políticas públicas, legislação, códigos de ética, normatizações etc.; • Parcerias com universidades no Brasil e no exterior; • *Big Data* nas mãos do tomador de decisão.	• Identificação da posição na curva de aprendizagem e realização de planejamento para investimentos, estabelecimento de parcerias e incentivo para especialização.

Acreditamos que o diagnóstico dessas variáveis é imprescindível para a definição estratégica de cada agente econômico. Após o diagnóstico, é possível valorizar ainda mais os pontos fortes, minimizar os pontos fracos, aproveitar as oportunidades e reduzir os impactos negativos das ameaças.

Essa análise é feita através da ferramenta SWOT Cruzada e é apresentada na *figura 10* na página anterior,

Os comentários dessa análise serão apresentados na conclusão desse trabalho, no item 5.3., intitulado **Desafios estratégicos.**

5. Conclusão

O maior inimigo da criatividade e
da inovação é o senso comum.

Ken Robinson - educador

O objetivo desse trabalho é o de identificar quais os desafios que já se impõem ou os que podemos vislumbrar num futuro próximo, para que a sociedade brasileira, governo, empresas e indivíduos, se beneficiem dos valores que a analítica de dados digitais pode oferecer.

Após o estudo do ecossistema *big data* e das especificidades do Brasil para o desenvolvimento desse ecossistema podemos enumerar algumas conclusões.

Nesse trabalho classificamos os desafios do *big data* em estruturais, culturais e estratégicos.

5.1. Desafios estruturais

No primeiro grupo, estão os gargalos de tecnologia. Esses desafios serão apenas citados, pois acreditamos que as soluções aparecerão a contento conforme a necessidade e a disponibilidade de investimento.

Codificação e dataficação: Será necessário o desenvolvimento ou aprimoramento de metodologia para codificação de informações e de tecnologia para captação de novos dados. *Big data* fundamenta-se em grande quantidade e variedade de dados e tudo que puder ser dataficado ajudará no aprimoramento dos modelos analíticos.

Rede wi-fi: A disponibilidade e qualidade da rede de transmissão de dados digitais ainda não consegue acompanhar a velocidade da oferta de serviços móveis fundamentais na captação de dados.

Armazenagem: Embora poucos apontem a armazenagem como gargalo da cadeia produtiva do *big data*, destacamos o desafio da constante ampliação do serviço já que não existe nenhuma expectativa de desaceleração na geração de dados.

Compartilhamento: Plataformas para compartilhamento de dados serão cada vez mais comum. Vemos como desafio o desenvolvimento de linguagens e interface mais acessíveis.

Analítica: Nesse ponto da cadeia está o maior gargalo pela escassez dos chamados cientistas de dados. Tecnicamente falando, desenvolver modelos analíticos é um desafio porque cada modelo é adequado a um problema em um determinado contexto e porque precisa ser dinâmico para manter-se atualizado. Cada modelo analítico pode ser considerado uma peça única e sua efetividade tem que ser colocada à prova continuamente.

Visualização de dados: Se o objetivo da analítica de dados é empoderar o tomador de decisão, será fundamental o desenvolvimento de soluções que deem aos dados uma forma de apreensão fácil e adequada ao problema que se quer resolver. A visualização eficaz das informações geradas pela analítica será fator determinante para melhores decisões em menos tempo.

5.2. Desafios culturais

No segundo grupo, estão os desafios que transcendem as capacidades técnicas da humanidade. O esforço para enfrentá-los inclui revisita à filosofia e à ciência, quebra de tabus e preconceitos, reconsiderações de processos etc. Coisas que a humanidade em geral só se dispõe a fazer quando está confrontada com problemas muito complexos que

ameacem diretamente sua sobrevivência. Acreditamos que associar dados digitais a um código que precisa ser decifrado para enfrentar doenças sem cura, epidemias, terrorismo, escassez de recursos, mobilidade, segurança entre outros podem ser fatores de mobilização social para as mudanças necessárias.

Nenhum esforço em resolver gargalos tecnológicos (desafios estruturais) ou de superar fraquezas e defender-se de ameaças (desafios estratégicos) será suficiente se não lidarmos antes com as grandes questões culturais que se apresentam no ecossistema *big data*.

Apropriação do conceito: Em primeiro lugar, não apenas em importância, mas também como pré-requisito para todos os demais desafios que seguirão, está a compreensão de *big data*. O desafio é percebê-lo não como uma solução em tecnologia da informação, mas como um ativo que deve ser acionado para encontrar as melhores soluções dentro do planejamento estratégico de uma instituição.

Compartilhamento de dados: A partir dos fundamentos da Ciência das Redes percebemos que os modelos preditivos mais efetivos são aqueles que utilizam dados dos mais diferentes domínios de conhecimento. Sabemos que esses dados diversificados são captados e controlados por diferentes agentes do ecossistema e que seria de extrema ineficiência captar os mesmos dados todas as vezes que eles fossem necessários. Diante dessas constatações podemos afirmar que o compartilhamento de dados é fundamental para a utilização efetiva de dados digitais. O grande desafio do compartilhamento é mais que um problema de tecnologia. O modelo econômico atual pressupõe assimetria de informação para obtenção de vantagem competitiva. Portanto compartilhar dados ainda é visto como perda de poder de barganha, exposição de fraquezas ou divulgação de informações estratégicas. O desafio será a compreensão de dados como um ativo de valor potencial, sobre o qual ainda será preciso exercer a inteligência para extrair seu valor fato. Só através dessa mudança de postura e da constatação de que dados digitais não se esgotam nem depreciam com o uso é que o compartilhamento de dados será visto como uma oportunidade e nãocomo uma ameaça.

Propriedade de dados: Importância do reconhecimento de que o valor do *big data* não está em possuir ou controlar os dados, mais no valor que extrai da correlação entre diferentes dados.

Colocar o ser humano no comando: Em concordância com a curva de Gartner de adoção de tecnologia, destacamos como grande desafio a formação de massa crítica para participar do debate sobre a utilização de dados digitais como novo paradigma de conhecimento. O desconhecimento faz com que *big data* seja, na maioria das vezes, exageradamente enaltecido ou injustamente desqualificado. Acreditamos que as duas posturas extremistas cometem o mesmo erro na origem: não colocam o ser humano no comando. A primeira será a responsável pela escalada vertiginosa ao *pico das expectativas* porque cria a expectativa de que a tecnologia sozinha trará todas soluções para nossos problemas. A segunda postura comandará a queda livre em direção ao *poço da desilusão* atribuindo à tecnologia todos os impactos negativos de sua adoção. Colocar o ser humano no comando de qualquer projeto em *big data* é reconhecer que nenhuma tecnologia pode prescindir da inteligência que deve agir não apenas para alcançar os desejados fins, mas também para escolher com responsabilidade os adequados meios. Nesse ponto estaremos aptos para subir a *ladeira do esclarecimento.*

Design de redes: A falta de cientistas de dados e de profissionais para tomar decisões a partir de modelos preditivos tem sido destacada como gargalo por todas as empresas de consultoria no assunto. No entanto identificamos outro desafio na formação profissional para trabalhar com *big data*: a habilidade de ver e desenhar redes complexas. *Big data* é um código digital de relações que são os links de uma rede. Decifrar e utilizar as informações contidas nesses códigos passa necessariamente por entender as redes complexas onde estão os problemas que desejamos resolver. Acreditamos que *design de redes complexas* será uma competência necessária e valorizada para se trabalhar com *big data*.

Privacidade: A polêmica questão da privacidade na utilização de dados digitais precisará evoluir para o campo da ética para ser resolvida. Antes, porém será preciso ficar clara a diferença entre dados e informação. **Dado** é a representação física de um evento no tempo e espaço que não agrega fundamento para quem o recebe, não podendo ser possível entender o que ele representa ou para que ele existe[55] (...) porém, no momento que existir uma agregação com outro dado ele passa a ser uma **informação**"[56]. O pressuposto do *big data* é que dados em grande quantidade de um determinado domínio quando transportado para outro, gera informações (dados agregados) e insights

[55] http://pt.wikipedia.org/wiki/Dados
[56] http://pt.wikipedia.org/wiki/Dados

relevantes para a compreensão de fenômenos que não podem ser explicados numa relação causa-efeito linear. Esse tipo de analítica não trás nenhum prejuízo à privacidade e tem sido utilizado nos mais diversos estudos científicos. No entanto, o mesmo acesso que se tem a dados isolados é possível se ter a dados associados à identidades pessoais. Nesse caso saímos da esfera *big data*-privacidade e entramos na esfera espionagem-ética. Afirmar que *big data* é sinônimo de perda de privacidade e desconhecer o assunto ou tratá-lo de maneira antiética. Impedir que a utilização inadequada dos dados digitais seja considerada a prática corrente de projetos *big data* é o maior desafio quando se fala em privacidade.

Metodologias: Gerar conhecimento sobre ambientes complexos a partir de dados digitais exigirá novos processos. Acreditamos que, independente da prontidão que setores e áreas corporativas tenham para *big data*, as instituições aumentarão seus investimentos em tecnologia da informação nos próximos anos para avançar do estágio de onde se encontram em relação ao *big data*. O primeiro desafio metodológico será o de medir esse **grau de prontidão**. Ao contrário do que se pensa, estar pronto para analítica de dados é muito mais que ter dados, hardware, software e cientistas de dados. Além das condições ligadas à tecnologia será preciso identificar os aspectos culturais que podem impulsionar ou atravancar a adoção de *big data* em uma determinada instituição. Outro desafio metodológico será o de **definir o problema** em um sistema complexo. Em qualquer contexto definir adequadamente o problema é essencial para encontrar a melhor solução que responda às condições humanas (desejo), técnicas (possibilidade) e econômicas (viabilidade). Num projeto *big data* o custo da coleta-processamento-analitica de dados trona-se fator de pressão para a identificação correta do problema. Ainda identificamos a necessidade de se definir **indicadores de sucesso** para diferentes tipos de modelos preditivos. Nem todo modelo poderá ser avaliado por acertar sua predição. Modelos preditivos que visam desencadear ações para reverter ou reforçar as predições certamente serão bem sucedidos se a curva prevista não coincidir com o que de fato acontecer.

5.3. Desafios estratégicos

Chamamos de desafios estratégicos aqueles extraídos das informações de mercado e da análise SWOT mostrada na sessão anterior. Esses desafios, embora não possam

prescindir de fundamentação teórica para serem enfrentados, são basicamente de posicionamento e alinhamento de ações entre o ambiente externo, a sociedade como um todo, e o ambiente interno, o ecossistema *big data*.

Do cruzamento dos pontos fortes e fracos dos dois ambientes identificamos os principais desafios:

Fraquezas-ameaças

Os maiores desafios estratégicos para o Brasil está no quadrante da análise SWOT cruzada onde as fraquezas do ambiente *big data* são reforçadas pelas ameaças do ambiente externo.

O principal obstáculo tanto no ambiente externo quanto no interno é a baixa compreensão do assunto. Dentro do ecossistema essa deficiência tem como consequências o desconhecimento da importância de compartilhar dados e o baixo investimento em ferramentas e metodologias especificam para analítica de dados. No ambiente externo isso se manifesta na falta de pressa da sociedade em exigir uma política pública de normatização e investimento em infraestrutura e educação para *big data*.

Forças-ameaças

Outra ameaça à *big data* no Brasil é a baixa qualidade e o alto custo da rede móvel de comunicação. Sendo a mobilidade a principal força motora do *big* data pela diversidade de dados de dinâmica humana que ela propicia, essa ameaça terá forte impacto na expansão da base de dados. Por outro lado, a base de dados existente e qualidade do setor de TI são forças que podem ser usadas para reverter esse quadro. O desafio é o desenvolvimento de projetos em analítica de dados de grande visibilidade e relevância para justificar o investimento na melhoria da rede e a queda no custo melhorando assim a penetração da rede em regiões e segmentos da população onde o acesso ainda é restrito.

A pouca articulação do governo com as instituições de pesquisa no país e a inexistência de publicações sobre o tema em língua portuguesa são ameaças que podem ser compensadas pelo forte interesse dos gestores brasileiros em *big data*, que pode levar ao estabelecimento de parcerias diretas com instituições dentro e fora do país.

Fraquezas-oportunidades

Apesar de apresentar muitas ameaças ao ecossistema *big data*, o Brasil oferece também condições para o desenvolvimento de boas vantagens competitivas. A principal delas diz respeito a rápida adesão da população a novas tecnologias e processos fortemente dependentes dela. Esse fator, somado à forte adesão às redes sociais e à participação *on line* e ao conceito de que inclusão digital é importante para a inclusão social, geram as condições ideais para a geração espontânea de dados de alta qualidade para analítica preditiva de comportamento humano e dinâmica social. O desafio aqui é o desenvolvimento de plataformas para compartilhamento desses dados, já que a armazenagem em silos e a baixa propensão em compartilhar dados foram citadas como fraqueza do nosso ecossistema.

A existência de bons institutos de pesquisa e desenvolvimento no país podem garantir a oferta de tecnologias e metodologias para criar um ambiente tecnológico e cultural favorável ao compartilhamento de dados.

Forças-oportunidade

O quadrante onde forças e oportunidades se encontram só são desafios até o momento do seu reconhecimento. A partir daí torna-se natural o caminho de associar esses fatores positivos para ganhar vantagem competitiva em algumas áreas específicas do ecossistema *big data*.

As condições de geração espontânea de dados digitais citadas no item anterior, associadas à alta capacidade de armazenagem e segurança dos dados e ao bom desenvolvimento do setor de TI colocam o Brasil em posição privilegiada na parte inicial da cadeia de valor de *big data*: desenvolvimento de métodos de codificação, de dispositivos, sensores e serviços para dataficação de informações importantes.

O desafio na parte alta da cadeia de valor é transformar o interesse dos tomadores de decisão e capacidade de articulação para o desenvolvimento de novos modelos de negócio, metodologias e indicadores de sucesso para ambiente *big data*.

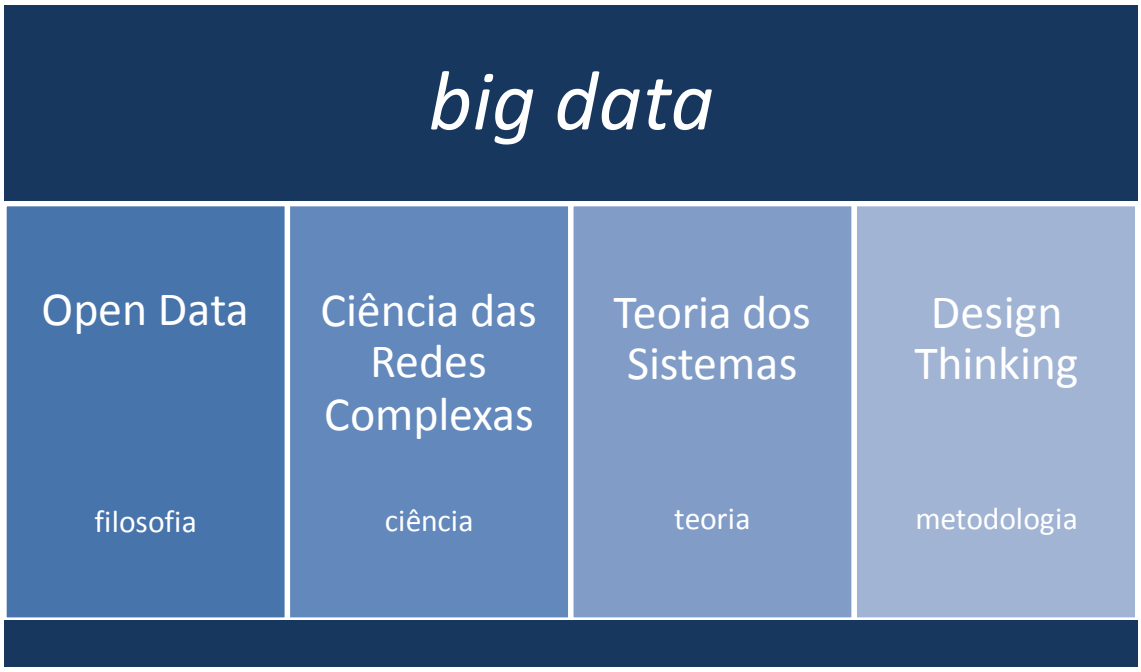
5.4. Quatro pilares para continuação dos estudos

Por fim, e como contribuição final desse trabalho, sugerimos que a continuação dos estudos sobre os assunto seja feita com base em quatro pilares que acreditamos serem fundamentais para compreensão e uso de dados digitais, e sobre os quais um projeto para *big data* deverá ser concebido.

A *figura 11* apresenta esquematicamente a ideia dos quatro pilares fundamentais.

Open Data: A filosofia *Open Data* deverá nortear a compreensão do potencial do *big data*. *Open Data* é a ideia que certos dados deveriam estar disponíveis para todos para serem usados e republicados, sem restrições de direitos autorais, patentes ou qualquer outro mecanismo de controle[57]. Acreditamos que o acesso livre aos dados digitais, embora seja uma utopia, deve estar na base das discussões sobre privacidade, abuso de poder pela concentração da informação e má gestão da informação gerada pelas analíticas de dados. A filosofia *Open Data* parte do principio de que dados não têm valor intrínseco e a eles não podem ser atribuídos a priori ganhos ou perdas com sua utilização.

Figura 11. Os quatro pilares para novos estudos



[57] http://opendefinition.org/okd/

Ciencia das Redes: Dados digitais são códigos de relações entre nós de redes complexas. Qualquer rede formada por seres humanos e suas relações entre si ou com o meio em que vivem é uma rede complexa e a nossa utilização de computadores e internet registram os rastros dessas relações. A Ciência das Redes nos ensina que a topologia de uma rede determina o que podemos fazer com ela. Conhecer a topologia e as propriedades das redes que pretendemos compreender é importante para utilizarmos com sabedoria os dados digitais disponíveis.

Teoria dos Sistema: A abordagem sistêmica estuda os sistemas como abertos, vivos e altamente interconectados. Permite a busca de correlações de modo mais complexo porque não estabelece limites entre os sistemas estudados. Acreditamos que essa deva ser a abordagem no desenvolvimento de metodologias para uso de analíticas, desde as descritivas até as prescritivas. O exercício de ver o todo em movimento é algo que pode ser feito com os dados digitais.

Design Thinking: Trata-se de métodos para abordar problemas, coletar e analisar informações, utilizar conhecimento e propor soluções a partir da empatia. Isso significa uma mudança muito grande no processo porque necessariamente coloca o tomador de decisão dentro do problema que pretende resolver. O *design thinking* é a única maneira legítima de abordagem de um problema complexo. A metodologia pressupõe ainda trabalho colaborativo e multidisciplinar.

Open Data, Ciência das Redes, Abordagem Sistêmica e Design Thinking são as lentes filosóficas, científicas, teóricas e metodológicas através das quais devemos estudar o fenômeno *big data*, elaborar normas, códigos de ética, modelos de negócio, de processos e indicadores dentro do seu ecossistema.

Os quatro pilares são fortemente humanistas o que é importante em *big data* porque, apesar de tanta tecnologia envolvida, a analítica digital não neutraliza a importância humana na tomada de decisão e na inovação. Ao contrário. A profusão de modelos preditivos e correlações de dados demandarão como nunca a capacidade humana de gerar insights, discernir e decidir.

Se precisamos definir *big data*[58], diremos que é a representação digital da vida. Dados digitais iluminam as pessoas e a rede vital que elas estabelecem entre si, com a natureza e com os objetos que estão à sua volta[59].

Acreditamos que lidar com esses dados é uma missão que todo ser humano deverá abraçar com cuidado porque eles são o código que nos mostrará como dar equilíbrio ao nosso planeta.

[58] O site www.opentracker.net/article/25-definitions-big-data apresenta mais de 30 definições de *big data* coletadas nas mais diversas publicações sobre assunto. Não é uma falta de consenso, mas definições que se complementam e diferentes visões para o fenômeno, sintoma de mudança de paradigma.

[59] Nossa definição é uma variação da definição do jornalista Rick Smolan no livro *The Human Face of Big Data*: "*the process of helping the planet grow a nervous system, one in which we are just another, human, type of sensor*."

Bibliografia

BARABÁSI, Albert-László. *Linked: the new science of networks.* Kindle Edition. Cambridge: Perseu Books Group, 2002.

BARABÁSI, Albert-László. *Burst: the hidden patterns behind everything we do.* Kindle Edition. London: Dutton (Penguin Books Ltd), 2010. BARABÁSI, Albert-László. Thinking in network terms. *Edge Conversation.* Set. 2012. Disponível em http://edge.org/conversation/thinking-in-network-terms. Acesso em 29 de set. 2012.

BERTOLUCCI, Jeff. Big Data´s New Buzzword: Datafication. *Infomation Week.* Fev.2013. Disponível em http://www.informationweek.com/big-data/news/big-data-analytics/big-datas-new-buzzword-datafication/240149288. Acesso em 08 de jan. 2013.

CHESBROUGHT, Henry W. Top 10 Lessons on the New Business of Innovation. *MIT Sloan Management Review.* North Hollywood, Winter 2011, 35-41 (The Era of Open Innovation)

EVANS, Dave. *The Internet of Things: How the Next Evolution of the Internet Is Changing Everything.* Cisco Internet Business Group. Abr. 2011.Disponível em http://www.cisco.com/web/about/ac79/docs/innov/IoT_IBSG_0411FINAL.pdf. Acesso 25 de fev. 2013.

GONZÁLEZ, Marta C.; BARABÁSI, Albert-László. From data to models. *Nature.* V.3, abr. 2007. Disponível em http://www.barabasilab.com/pubs/CCNR-ALB_Publications/200704-05_NaturePhys-DataModels/200704-05_NaturePhys-DataModels.pdf. Acesso em 15 de dez. 2012.

JOHNSON, David J. *Gestão de redes de conhecimento.* São Paulo: Editora Senac, 2011.

KAHANE, Adam. *Como resolver problemas complexos.* Ed. 1. São Paulo: Editora Senac, 2008.

MILGRAM, Stanley; JEFFREY, Travers. The Small-World Problem. 1967. Disponível em http://measure.igpp.ucla.edu/GK12-SEE-LA/Lesson_Files_09/Tina_Wey/TW_social_networks_Milgram_1967_small_world_problem.pdf. Acesso em 08 de jan. 2013

SANTOS, Augusto P.de M. *Análise dos Impactos da Tecnologia da Informação nas Organizações: Um ensaio à luz da Teoria da Agência e da Teoria dos Custos de Transação*. 2003. Dissertação (Mestrado em administração Pública e de Empresas) – Fundação Getúlio Vargas FGV, Rio de Janeiro. 2003. P. 23-32.

SCHÖNBERGER, Viktor M. ; CUKIER, Kenneth. *Big Data: a revolution that Will transform how we live, work and think.* Kindle Edition. New York: Houghton Mifflin Harcourt Publishing Company, 2013.

SLOTINE, Jean-Jacque; LIU, Yang-Yu; BARABÁSI, Albert-László. Controllability of complex networks. *Nature*. V.473, mai. 2011. Disponível em http://www.uvm.edu/~cmplxsys/newsevents/pdfs/2011/barabasi-network-controllability-nature-2011.pdf. Acesso em 15 de dez. 2012.

Sites visitados

Computer World UK. http://www.computerworlduk.com/news/public-sector/3400618/maude-introduces-independent-reviews-for-open-data-countries/. Acesso em 25 de mar. 2013.

Google Inc. www.google.com/intl/en/policies/privacy/ e http://support.google.com/a/bin/answer.py?hl=pt-BR&answer=60762. Acesso em 13 de mar. 2013.

IBM http://www-01.ibm.com/software/data/bigdata/ Acesso em 13/01/2013

Instituto do Governo Eletrônico http://igov.com.br/tigov/?p=564p%C3%A1gina. Acesso em 13/01/2013

Massachusetts Institute of Technology (MIT). Auto –ID LabsMIT. Grupo de pesquisa em indentificação por radiofrequência (RFID). http://web.mit.edu/newsoffice/2012/auto-id-cloud-of-things-big-data.html. Acesso em 17 de fev. 2013.

NASDAQ. http://www.nasdaq.com. Acesso em 27 de mar. 2013

Omidyar Network. Investimento Filantrópico. http://www.omidyar.com. Acesso em 20 de mar. 2013.

Open Data Institute. http://www.theodi.org. Acesso em 20 de mar. 2013.

Open Definition. http://support.google.com/a/bin/answer.py?hl=pt-BR&answer=60762. Acesso em 13 de mar. 2013.

Royal Pingdom. http://royal.pingdom.com/2013/01/16/internet-2012-in-numbers/. Acesso em 13 de mar. 2013.

Tech & Net. Divulgação de notícias de tecnologia e internet. http://www.techenet.com/2011/07/cisco-net-coisas-infografico/. Acesso em 25 de fev. 2013.

The Guardian. http://www.guardian.co.uk/technology/free-our-data. Acesso em 25 de mar. 2013.

Trendwatching (setembro de 2012) http://www.trendwatching.com/trends/pdf/2012-08%20MINITRENDS.pdf. Acesso em 08 de jan. 2013

Wikipedia http://en.wikipedia.org/wiki/Main_Page. Diversos acessos.