# 2
# *BIG DATA*

## 2.1.
## Conceito

Após enviar um *e-mail* para seus amigos combinando uma viagem, a publicidade que aparece na tela do seu correio eletrônico é a de um *site* de agendamento de hotéis. Pouco tempo depois de comprar um livro em uma grande empresa de *e-commerce*, você recebe um *e-mail* com ofertas de outros livros que muito lhe interessariam ler em sequência. Nas sugestões de amizade de sua rede social favorita, você encontra um grande número de rostos conhecidos. Estes são apenas alguns exemplos do que o aumento da capacidade de armazenamento, processamento e análise de dados ocorrido nas últimas décadas é capaz. A este conjunto de soluções tecnológicas dá-se o nome de *big data*.

A tradução para o português - "grandes dados" - raramente é usada. E uma busca na internet sobre o surgimento da expressão é capaz de retornar as mais diferentes versões e origens. O termo é novo, mas o conceito data de milênios antes de Cristo. Por volta de 3500 a.C., os burocratas da antiga Mesopotâmia precisavam de ferramentas capazes de registrar e manter o controle das transações comerciais, para isso criaram a escrita. "A linguagem escrita permitiu que as primeiras civilizações medissem a realidade, a gravassem e, mais tarde, a evocassem. Juntas, a medição e a gravação facilitaram a criação dos dados. Elas são as bases da *dataficação*." (MAYER-SCHONBERGER & CUKIER, 2013). Entende-se dataficar como a ação de adequar um fenômeno a um "formato quantificado de modo que possa ser tabulado e analisado." (ibidem).

O desejo por medir e analisar sempre esteve relacionado ao registro e ao conhecimento. Flusser (2007) relata que na Mesopotâmia as pessoas se dirigiam para o alto das montanhas, "olhavam em direção à nascente dos rios e podiam prever secas e inundações, e depois traçavam linhas em plaquetas de argila para representar os canais que deveriam ser cavados futuramente." (FLUSSER, 2007).

PUC-Rio - Certificação Digital Nº 1312280/CA

Na época, tais pessoas eram consideradas profetas, porém o que elas faziam era analisar os dados existentes para entender a natureza, e deduzir como ela se comportaria no futuro.

A palavra "dado", derivada do Latim *data*, significa aquilo que se concede, aquilo que é dado. Neste sentido, os dados são valores, quantitativos ou qualitativos, pertencentes a um objeto ou grupo de objetos. Os dados podem ser analógicos ou digitais. Quando um dado analógico é convertido em digital, diz-se que este foi digitalizado. Além disso, os dados podem possuir os mais variados níveis de estruturação, podendo ser: brutos, semiestruturados ou estruturados. Um dado é dito estruturado quando o mesmo encontra-se formatado para ser inserido em um banco de dados. Um numeral pode ser um dado, por exemplo o número 273,6 milhões; por si só ele nada significa, no entanto, se associado a outros valores, como: linhas ativas de celulares, Brasil, e Fevereiro de 2014, estes dados juntos tornam-se uma informação. Mais adiante, no capítulo 3, serão tratados os conceitos de dado, informação e conhecimento, bem como o processo de transformação entre eles.

Ainda sobre o termo *big data*, sua definição não é precisa, mas algo com que muitos parecem concordar é que não se trata apenas uma questão de uma certa quantidade de *bytes*. Três características o distinguem: volume, variedade e velocidade. A Gartner, empresa de pesquisa em tecnologia da informação, define *big data* como: "Grande volume, alta velocidade, e/ou grande variedade de dados que exigem novas formas de processamento para permitir melhores tomadas de decisões, descobertas de conhecimento e otimização de processos". (BEYER & LANEY, 2012).

A IBM (2012) ilustra os três *V*'s do *big data* na figura adiante:

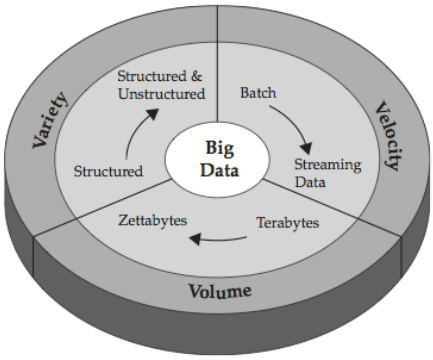


Figura 1 – A IBM caracteriza *big data* por seu volume, velocidade e variedade ou simplesmente os três V's. (IBM, 2012)

PUC-Rio - Certificação Digital Nº 1312280/CA

### 2.1.1. Volume

Uma quantidade imensurável de dados é gerada e capturada diariamente, e a previsão é de que, até 2020, cerca de 35 Zettabytes (ZB) estejam armazenados no mundo (IBM, 2012). Os números são gigantescos; Jay Parikh, vice-presidente de infraestrutura do Facebook, revelou em entrevistas (CONSTINE, 2012) que a empresa insere mais de 500 Terabytes por dia em seu banco de dados. Este é apenas um único exemplo, e é muito provável que os números citados acima sejam ainda muito maiores no momento da publicação desta dissertação.

Para alguns, o volume está ligado à quantidade de dados. "*Big data* são dados que excedem a capacidade de processamento dos sistemas convencionais" (tradução livre – EDD DUMBILL, 2012). No entanto, não é necessariamente uma questão de quantidade. Segundo Lev Manovich (2011), grandes volumes de dados podem ser analisados em computadores pessoais usando *softwares* convencionais sem a necessidade de supercomputadores. Logo, a questão do volume é mais uma questão de proporção do que de quantidade. Com o *big data*, não é mais necessário delimitar amostras ou grupo de testes. "*N = all"* (N é igual ao todo). (MAYER-SCHONBERGER & CUKIER, 2013). "Apoiando-se na totalidade dos dados é possível ter uma visão mais clara dos detalhes e explorar novas análises." (tradução livre – ibidem).

PUC-Rio - Certificação Digital Nº 1312280/CA

### 2.1.2. Variedade

O grande volume de dados traz à tona a questão da variedade dos mesmos. Como a análise é feita a partir da maior parte ou da totalidade dos dados, eles podem ser gerados a partir das mais diversas fontes e estruturados da forma mais variada possível. Segundo a IBM: "O sucesso de uma empresa vai depender na sua habilidade em gerar conhecimento a partir dos vários tipos de dados disponíveis." (IBM, 2012).

Os dados podem ser: textos de uma rede social, metadados de imagens, dados de um sensor, GPS de um celular, vídeos de uma câmera de segurança, entre outros. Tais dados integrados podem ser capazes de gerar descobertas e o reconhecimento de padrões que antes eram mais difíceis de serem percebidos. No

entanto, para que os dados façam parte de uma única aplicação, ou para serem interpretados por uma pessoa ou sistema, eles precisam ser estruturados. Neste processo pode ocorrer perda de informações. "Por isso, a importância de armazenar tudo que for possível na forma ainda bruta." (STRATA, 2012).

Sendo assim, quanto maior o volume de dados, menor a precisão da informação. Ou seja, considerando apenas os dados estruturados, aqueles que se enquadram perfeitamente em um banco de dados, perde-se a informação do todo, e assim valiosas descobertas são ignoradas. “Claro que o dado não pode ser completamente incorreto, mas o sacrifício de um pouco de precisão em troca do conhecimento de uma tendência geral é aceitável” (tradução livre – MAYER-SCHONBERGER & CUKIER, 2013).

O *big data* acarretou uma mudança de mentalidade. Anteriormente os dados eram escassos e custosos de serem coletados, armazenados e interpretados. Não havia espaço para a imprecisão, cada informação precisava ser exata. No momento em que tornou-se possível lidar com o volume total, ou quase total, a tolerância por ambiguidades aumentou. O plano de observação se distanciou do detalhe e foi para uma perspectiva mais distante e abrangente. A consequência disso é uma imagem mais ampla e que permite maiores descobertas. “Ao permitir a imprecisão, uma janela para um universo inexplorado de descobertas é aberta.” (tradução livre – MAYER-SCHONBERGER & CUKIER, 2013).

PUC-Rio - Certificação Digital Nº 1312280/CA

### 2.1.3.
### Velocidade

A velocidade com que os dados são gerados, armazenados e interpretados também é uma característica do *big data*. Ser mais veloz pode significar estar à frente dos concorrentes, identificar padrões, uma importante tendência, um problema ou uma oportunidade antes de alguém. Além disso, dependendo do que se está analisando, as respostas precisam ser geradas em tempo real, como por exemplo, recomendações baseadas em histórico de consumo, detecção de fraudes, informações sobre o trânsito, entre outros.

O diferencial existe quando o processo de gerar conhecimento através do dado coletado é rápido o suficiente para ser insumo de tomadas de decisões, ou para fazer parte do serviço. Por exemplo, empresas como a Amazon.com e o Netflix exibem recomendações de conteúdo baseadas em interações do usuário em

seu *site* e aplicativos. Além disso, muitas empresas utilizam dados em tempo real para construir *dashboards*[1] com informações que apoiem as tomadas de decisão da diretoria.

## 2.2. *Data scientist*

A partir do cenário descrito acima, tornou-se necessário o surgimento de profissionais capazes de dar assistência às questões técnicas e operacionais do armazenamento, processamento e análise dos dados: o analista de dados.

Para D.J. Patil e Jeff Hammerbacher, responsáveis pelas equipes de análise de dados do LinkedIn e Facebook, respectivamente, o termo que melhor descreve tais funções é *data scientist* (cientista de dados), pois estes profissionais geralmente possuem experiência em engenharia e não trabalham em laboratórios de pesquisa em projetos futuristas distantes do dia-dia das empresas, mas, ao contrário, usam os dados e a ciência para desenvolver aplicações com resultados imediatos.

Os cientistas são responsáveis por definir a questão a ser descoberta, os dados a serem colhidos, e determinam quem poderá ter acesso aos mesmos. Eles obtêm, tratam, analisam e interpretam o que foi coletado. E ainda, sintetizam os resultados e disseminam o conhecimento adquirido através de relatórios ou *dashboards*.

Segundo Patil (2011), um bom cientista de dados deve possuir conhecimentos técnicos em alguma disciplina científica. Ser curioso e capaz de dissecar um problema em várias hipóteses que possam ser testadas. Deve ter a habilidade de olhar um problema de diferentes ângulos de forma criativa e a capacidade de comunicar os dados de forma eficiente, seja através de histórias ou visualmente.

Não é raro encontrar um *designer* com essas características. Por definição, *designers* são profissionais que exercem atividades de caráter técnico-científico,

[1] Na Computação, um *dashboard* é uma aplicação utilizada para demonstrar indicadores de desempenho de um determinado produto ou empresa. Para tal, faz uso de elementos como gráficos, números e ilustrações, por exemplo.

PUC-Rio - Certificação Digital Nº 1312280/CA

criativo e artístico para elaboração de projetos industriais. É exatamente a união entre *Data Science* e Design que esta pesquisa pretende investigar.

Para Jeffrey Leek, professor da Universidade John Hopkins, mais importante que os dados são as questões que se deseja descobrir. Muitas vezes os dados limitarão as perguntas, mas elas determinarão o conhecimento a ser adquirido. Ele enumerou os diferentes tipos de perguntas:

**Descritiva**: tem como objetivo descrever um conjunto de dados e é comumente usada em censos. Os dados são apresentados mas não são interpretados. Exemplos: Censo para descrever a população de um país ou Google Ngram[2], serviço que exibe a quantidade de vezes que um termo foi citado em publicações. Estes serviços apenas exibem as informações sem interpretar os motivos que levaram a elas.

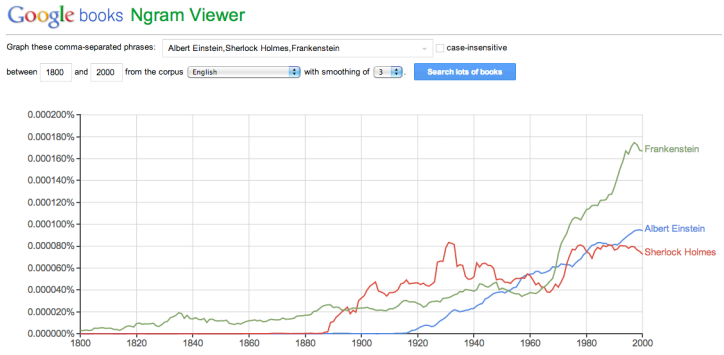


Figura 2 – Neste exemplo, o serviço compara as palavras: Albert Einstein, Sherlock Holmes e Frankstein. Acesso em: 21 de janeiro de 2014.

**Exploratória**: o objetivo deste tipo de questão é encontrar relações previamente desconhecidas, mas que não são necessariamente confirmadas. A finalidade é explorar e identificar novas relações, no entanto, é importante lembrar que correlação não implica causa. São úteis para definir novos estudos. Exemplo: pesquisa exploratória para descobrir que áreas do cérebro reagem a determinados estímulos sem necessariamente entender o que as reações significam.

**Inferencial**: questões desse tipo usam pequenas amostras de dados para inferir descobertas sobre uma população maior. É preciso estimar a quantidade desejada para a pesquisa. Exemplo: pesquisas com grupo de testes.

[2] https://books.google.com/ngrams

PUC-Rio - Certificação Digital Nº 1312280/CA

**Preditiva**: usa os dados de um objeto para prever valores em outro. No entanto, se X ocorre frequentemente em conjunto com Y, não significa necessariamente que X causa Y. Através de Y é possível entender X, mesmo que X não seja observado diretamente. Exemplos: previsão da vida útil de peças de certos equipamentos e uso de dados para prever o vencedor de uma eleição.

**Causal:** o objetivo é descobrir o que acontece com uma variável quando se faz outra variável mudar. Este tipo de estudo geralmente é feito com itens randômicos para garantir a veracidade da pesquisa. De uma forma geral os resultados são aplicados a uma média, mas não são necessariamente aplicados a cada indivíduo.

**Mecânica**: é o mais raro tipo de análise, pois é a mais difícil. Tem como objetivo entender as mudanças nas variáveis que levam a mudanças em outras variáveis. Geralmente usada em questões de Física, onde as equações são conhecidas, mas os parâmetros não o são. Exemplo: uso de modelos matemáticos para descrever fenômenos físicos e poder estudar seus comportamentos em função de gráficos e outras ferramentas matemáticas.

Para responder a estas questões, diferentes técnicas podem ser utilizadas, seja para a coleta ou análise dos dados. Tais técnicas derivam de inúmeras áreas do saber: Estatística, Marketing, Design, Ciência da Computação (Mineração de Dados, Aprendizado da Máquina, Inteligência Artificial, Processamento de Linguagem Natural), entre outros. O relatório McKinsey (2011) sobre *big data* lista algumas práticas herdadas da análise de dados convencionais e outras criadas especialmente para os dados grandes. Seguem alguns exemplos:

**Testes A/B:** técnica onde um grupo de controle é comparado com uma variedade de grupos de teste a fim de determinar quais mudanças nas variáveis possuem as melhores taxas de resposta. O *big data* possibilita um número maior de testes, assegurando que os grupos são de tamanho suficiente para detectar diferenças significativas entre eles. Google e Amazon.com foram os pioneiros no uso desta técnica com o objetivo de otimizar a interface de seus *websites*.

**Aprendizado de regras de associação** (*association rule learning*): conjunto de técnicas para descobrir relacionamentos interessantes entre variáveis em grandes bancos de dados. Uma das possíveis aplicações é a análise de carrinhos de compra para determinar quais produtos são comprados juntos com maior frequência e usar esta informação para aumentar o número de vendas.

PUC-Rio - Certificação Digital Nº 1312280/CA

**Classificação:** conjunto de técnicas para identificar e classificar novas unidades de dados em grupos previamente definidos baseados em características especificadas inicialmente. É usado para prever eventos futuros e comportamentos.

**Análise de agrupamento** (*cluster analysis*): este método é usado para classificar uma coleção de objetos em segmentos com características em comum cujos atributos não são previamente conhecidos. Um exemplo de análise de agrupamento é segmentar consumidores em grupos para iniciativas de marketing direcionadas.

***Crowdsourcing*:** técnica utilizada para coletar dados de um grande grupo de pessoas através de um convite aberto, geralmente via rede social. Este é um tipo de colaboração em massa.

**Fusão de dados e integração de dados** (*data fusion* e *data integration*): conjunto de técnicas para integrar e analisar dados de fontes variadas com o objetivo de gerar conhecimento de forma mais eficiente do que seria se uma única fonte de dados fosse analisada.

**Redes neurais:** modelos computacionais são construídos para identificar padrões nos dados inspirados nas redes neurais da biologia. Ideais para identificar padrões não lineares, como por exemplo, encontrar fraudes em sistemas financeiros.

**Análise de redes:** conjunto de técnicas usadas para caracterizar relações entre nós de uma rede. Em análises de redes sociais os nós (ou nodos) são representações dos atores envolvidos enquanto as conexões podem ser as interações entre os atores ou a interação com o sistema.

**Regressão:** conjunto de métodos usados para determinar a forma que a variável dependente se modifica quando uma ou mais variáveis independentes são alteradas. Geralmente usados para previsão de fenômenos, como por exemplo, para prever o volume de vendas de uma empresa baseado em dados de mercado e variáveis na economia.

**Análise de sentimentos:** conjunto de técnicas usadas para identificar e extrair informações subjetivas a partir de textos não estruturados. Os objetivos principais são: identificar o item que está sendo falado e a polaridade do sentimento expresso (positivo, negativo ou neutro). Usado principalmente para identificar a percepção de produtos e marcas em redes sociais.

PUC-Rio - Certificação Digital Nº 1312280/CA

**Visualização:** uso de técnicas de design da informação para criar imagens, gráficos, diagramas e animações para promover um melhor entendimento das análises de *big data*. No capítulo 3 esta técnica será mais aprofundada.

## 2.3. Aplicações

O aumento na demanda de *data scientists* se deu pelo sucesso das empresas pioneiras no uso dos grandes dados como Amazon.com, Google, LinkedIn, Facebook e Walmart. Essas empresas identificaram formas de extrair valor dos dados disponíveis, seja utilizando-os para uma nova funcionalidade no seu serviço ou simplesmente para otimizar processos internos.

O uso mais popular do *big data* é a recomendação de conteúdo. A Amazon.com foi a primeira a ter a funcionalidade "*quem viu isso, viu também...*" em seu comércio eletrônico. Ao sugerir itens de forma bem sucedida, a empresa consegue transformar a compra de um único item em uma compra maior. Outra empresa que conseguiu aumentar o tempo de uso do seu *site* foi o Netflix, ao recomendar conteúdos relevantes, o usuário tende a assistir vários vídeos em sequencia. Os casos da Amazon.com, Netflix e os sistemas de recomendação serão tratados com mais detalhes no capítulo 4.

Seguindo a mesma linha de raciocínio, redes sociais como o LinkedIn e Facebook, criaram a funcionalidade "*People You May Know"* (Pessoas que você talvez conheça) para ajudar os usuários a encontrarem pessoas conhecidas. Se Maria conhece João e João conhece Pedro, a probabilidade de Maria conhecer Pedro é grande. Além disso, ao busca um nome comum como João da Silva, por exemplo, as redes identificam qual João é o mais provável de ser o que se está buscando.

Com o objetivo de oferecer produtos para serem consumidos, a empresa americana Target consegue até mesmo prever, através do histórico de compras de seus clientes, uma possível gravidez. Duas dúzias de produtos quando associados são capazes, ainda, de indicar uma provável data de nascimento. Sendo assim, a empresa conseguia enviar para suas clientes ofertas para cada fase do período de gestação.

PUC-Rio - Certificação Digital Nº 1312280/CA

Logo, é possível observar que o aumento da capacidade de armazenamento, processamento e análise de dados está permitindo que muitas empresas criem serviços e funcionalidades baseadas em informações geradas por seus usuários. O conhecimento extraído desses dados é muito valioso e, se bem aproveitado, ajuda a promover a fidelização de clientes e torna-se um diferencial ao escolher um serviço.

Mas não é apenas o meio comercial que vem se beneficiando dos avanços tecnológicos. Há pouco tempo seria impossível imaginar um serviço que pudesse prever surtos de gripe pelo mundo em tempo real, como faz hoje o Google Flu Trends[3] rastreando termos de buscas relacionados à gripe.

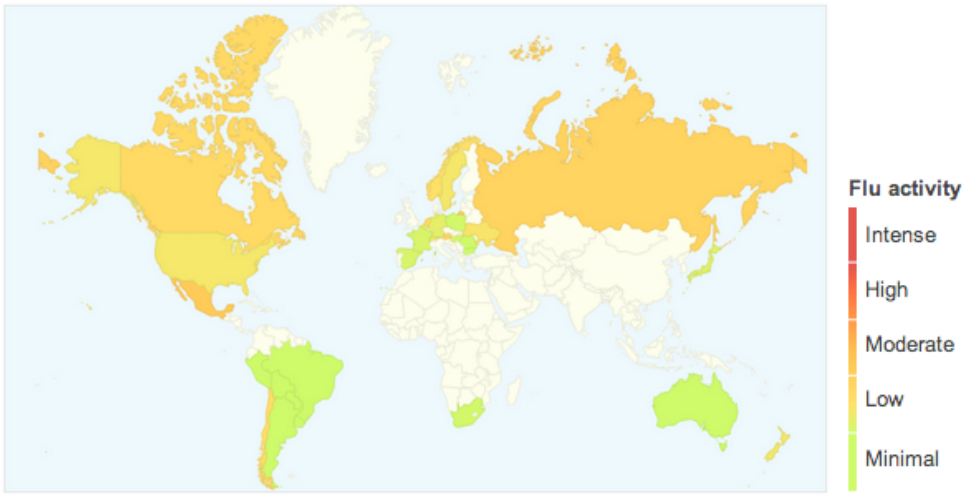


Figura 3 – Google Flu Trends - acesso em 20 de novembro de 2013.

Na área da saúde, muitas iniciativas e pesquisas estão em andamento. Dr. Miguel Luengo-Oroz e sua equipe, pesquisadores na Universidade Politécnica de Madrid, elaboraram uma prova de conceito que se constitui em um jogo onde as pessoas fazem diagnóstico de malária a baixo custo e com rapidez. Tal jogo encontra-se disponível *online*[4] e, após um breve tutorial, as pessoas são incentivadas a identificar, em imagens ampliadas de microscópio, se o sangue está ou não infectado pela doença. Como muitas pessoas fazem a análise de uma mesma imagem, a probabilidade de um diagnóstico correto aumenta.

[3] http://www.google.org/flutrends/
[4] http://malariaspot.com/

PUC-Rio - Certificação Digital Nº 1312280/CA

O governo e as cidades também estão usando o *big data* para promover melhorias em suas estruturas. A polícia de Los Angeles está usando a tecnologia para prever crimes antes mesmo que eles aconteçam. O sistema[5] analisa a hora, a data e o local onde crimes aconteceram e marca as regiões mais atingidas, onde o policiamento é então reforçado. Desde que o sistema foi introduzido os roubos caíram em 33%, e crimes violentos também foram reduzidos em 21%.

Pesquisadores do MIT (*Massachussets Institute of Technology*) estão usando dados dos celulares para entender como motoristas se locomovem, com o objetivo de reduzir o tráfego e melhorar o planejamento urbano.[6]

No Brasil, o governo de Minas Gerais tornou públicos os dados sobre exportações, atividades econômicas, localidades e ocupações de todo o Brasil através de uma rica ferramenta *online* de visualização de dados chamada Data Viva[7]. O objetivo é tornar tais dados públicos e promover informações valiosas para decisões de investimento e formulação de políticas públicas.

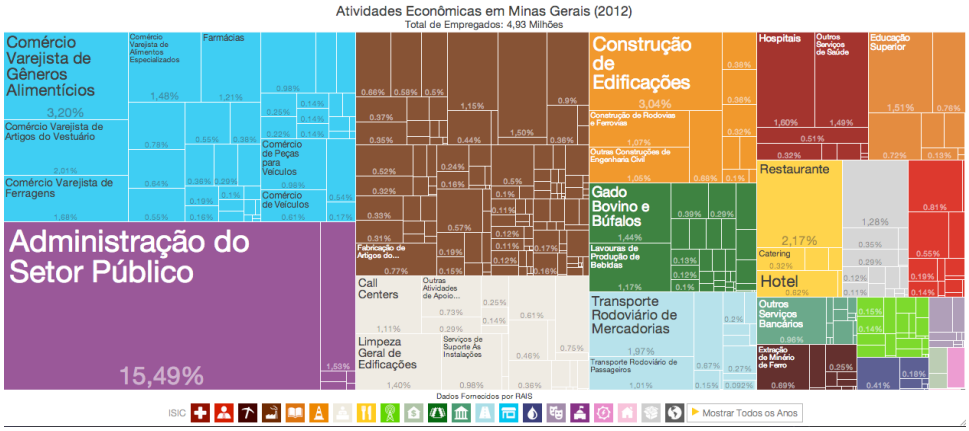


Figura 4 – Visualização das atividades econômicas em Minas Gerais em 2012. Acesso em 30 de Novembro de 2013.

Outro movimento que tornou-se possível tem o nome de "o ser quantificado" (*quantified self*), no qual as pessoas medem seus próprios dados, geralmente com o intuito de melhorar seus hábitos e viver melhor, ou simplesmente para se conhecerem mais a fundo. Muitas empresas já se

[5] http://www.dailymail.co.uk/news/article-2128666/The-REAL-Minority-Report-LA-cops-using-computers-pred ict-crimes-happen.html#ixzz2lziCPQJg

[6] http://web.mit.edu/newsoffice/2012/cellphone-data-helps-pinpoint-source-of-traffic-tie-ups-1220.html

[7] http://www.dataviva.info/

PUC-Rio - Certificação Digital Nº 1312280/CA

beneficiaram deste desejo de autoconhecimento. A Nike, por exemplo, declarou que o seu sistema chamado Nike+, um dos primeiros medidores de velocidade pessoal lançados no mercado em 2006, atualmente é usado por mais de 18 milhões de pessoas.


Figura 5 – Imagens da tela do aplicativo Nike+ para Android.

Outra forma de se obter dados para análise é através das redes sociais. Segundo Mayer-Schonberger e Cukier (2013), elas não são simplesmente ferramentas que nos mantem em contato com nossos amigos, mas sim meios que transformam em dados aspectos intangíveis das nossas vidas. O Facebook transformou em dados nossos gostos e relações. O Twitter é uma ferramenta onde as pessoas compartilham seus pensamentos, permitindo assim transformá-los em dados. Não trata-se apenas da análise de números, mas, sobretudo do comportamento humano e da sociedade.

Para Lev Manovich (2011), as redes sociais e o *big data* mudam o cenário das pesquisas sobre a humanidade e a sociedade. Até então, era preciso escolher entre analisar dados superficiais sobre muitas pessoas (estatística e sociologia) ou dados aprofundados sobre poucos indivíduos (psicologia, psicanálise, antropologia, história da arte). Agora, não é preciso escolher e nem apoiar-se em pequenas amostras que na maioria das vezes não representam o todo. “Pela primeira vez, podemos acompanhar imaginações, opiniões, ideias, e sentimentos

PUC-Rio - Certificação Digital Nº 1312280/CA

de milhões de pessoas." (MANOVICH, 2011). "E não precisamos pedir permissão, já que eles mesmos nos encorajam a fazer isso ao tornarem todos esses dados públicos." (ibidem). No entanto, o próprio Manovich levanta algumas provocações sobre esta nova condição. São elas:

- Apenas as grandes redes sociais como Google e Facebook têm acesso aos dados sociais em sua totalidade. Muitas vezes, as *APIs*[8] fornecidas por essas empresas não disponibilizam todos os dados obtidos sobre os usuários.

- É preciso ter cuidado com a autenticidade das informações publicadas em redes sociais. As pessoas se apresentam publicamente da forma como querem ser e não como de fato o são, sendo assim, muitos dos dados exibidos podem ser ficcionais, criados para representar a imagem que elas gostariam de exibir para o mundo.

- Dados superficiais sobre muitos e dados aprofundados sobre poucos podem ser combinados, resultando em diferentes descobertas. Um tipo de pesquisa não anula a outra e elas podem ser complementares: o ideal é combinar a habilidade humana em entender e interpretar à capacidade analítica do computador.

- O *big data* requer conhecimentos na ciência da computação, estatística e em mineração de dados[9] que muitas vezes pesquisadores das áreas sociais não possuem.

Chris Anderson, editor chefe da revista *Wired*, também discute a questão dos impactos do *big data* nas ciências. Segundo ele, durante anos o método científico foi baseado em hipóteses, as quais eram confirmadas ou negadas pelos cientistas. No entanto, atualmente é possível analisar os dados sem hipóteses pré-formuladas e descobrir padrões antes impossíveis de serem descobertos. Anderson (2008) descreve esta nova realidade como a "Era do Petabyte":

[8] API (*Application Programming Interface*) é um conjunto de comandos e padrões usados para acessar, via programação, dados e funcionalidades de um sistema. Por exemplo: o Twitter disponibiliza em sua API o conteúdo dos *tweets* publicados por seus usuários para que outras companhias possam desenvolver produtos utilizando seu serviço.

[9] Mineração de dados (*data mining*) é uma área dentro da Ciência da Computação responsável pelo processo de explorar grandes quantidades de dados com o objetivo de encontrar padrões.

PUC-Rio - Certificação Digital Nº 1312280/CA

> Este é um mundo onde grandes quantidades de dados e a matemática aplicada substituem qualquer outra ferramenta que possa existir. Fora todas as teorias do comportamento humano, da linguística à sociologia. Esqueça a taxonomia, a ontologia e a psicologia. Quem sabe por que as pessoas fazem o que fazem? O ponto é que eles fazem, e podemos rastrear e medir com alta fidelidade. Com dados suficientes, os números falam por si. (tradução livre – CHRIS ANDERSON, 2008)

Exageros à parte, o fato é que o *big data* permitiu uma série de descobertas e conexões inimagináveis. E se antes estávamos limitados a testar algumas hipóteses pré-definidas, agora o poder de processamento dos computadores permite que uma série de dados sejam analisados sem muitos custos adicionais. Muitos serviços tornaram-se viáveis nos últimos anos a partir do uso desta tecnologia, e nos capítulos a seguir novos exemplos serão apresentados. No entanto, há um amplo debate em curso a respeito dos aspectos éticos no que se refere ao uso (autorizado ou não) de dados pessoais.

PUC-Rio - Certificação Digital Nº 1312280/CA

## 2.4. Ética e privacidade

O volume, a variedade e a velocidade dos dados disponíveis aumentam a complexidade da informação e promovem discussões sobre a questão ética do uso de dados dos usuários. Além de fazer emergir sentimentos de vigilância e de invasão de privacidade muitos sentem-se incomodados quando um sistema parece saber demais sobre si.

Assim como o *big data* pode ser utilizado para promover melhorias na vida das pessoas, empresas podem lucrar de forma questionável com o grande conhecimento sobre as atividades e o comportamento dos indivíduos e da sociedade. "Algoritmos irão prever a probabilidade de alguém ter um ataque cardíaco (e estes pagarão mais caro em seus planos de saúde), dever uma hipoteca (e empréstimos serão negados), ou cometer um crime (e talvez sejam presos antecipadamente). Tais situações nos levam à discussão ética do papel do livre arbítrio *versus* a ditadura dos dados. A era do *big data* vai exigir novas regras para resguardar o indivíduo." (tradução livre – MAYER-SCHONBERGER & CUKIER, 2013).

DAVIS (2012) enumera quatro aspectos que devem ser considerados em relação à ética no uso de big *data*. São eles:

**- Identidade**: Qual a relação entre a identidade *offline* e a identidade *online* das pessoas?

**- Privacidade**: Quem deve controlar o acesso aos dados?

**- Autoria**: a quem pertencem os dados? Os direitos podem ser transferidos? E quais são as obrigações das pessoas que geram e usam esses dados?

**- Reputação**: Como determinar quais dados são confiáveis?

Em 2008, um estudo publicado analisou o perfil do Facebook de 1700 estudantes de uma universidade americana com o objetivo de estudar as mudanças nas amizades e interesses ao longo do tempo destes indivíduos. Depois de publicado, apesar do anonimato dos usuários, a partir das informações fornecidas pelo estudo, descobriu-se facilmente qual havia sido a instituição pesquisada, bem como a identidade dos sujeitos. (ZIMMER, 2010). O uso que vem sendo feito de dados públicos em redes sociais levanta a questão do público *versus* o privado. Dados de redes sociais podem ser usados para pesquisas simplesmente por serem acessíveis e visíveis a qualquer um? Os dados criados em um contexto podem ser utilizados em outro sem permissão de quem os criou? Os usuários estão cientes dos múltiplos usos de suas informações e publicações?

Muitas perguntas ainda não foram respondidas e ainda há muito a se discutir sobre as implicações éticas e a responsabilidade no uso de dados. Jonathan Harris, artista digital que desenvolve projetos ligados à forma como as pessoas se relacionam com a tecnologia e umas com as outras, publicou um manifesto no jornal New York Times (LOHR, 2013) que questiona o uso do *big data*, com o intuito de gerar uma reflexão a respeito do uso desta tecnologia. Ele escreve:

> Os dados vão nos ajudar a lembrar, mas irão nos deixar esquecer? (...) Vão ajudar empresas a fazerem produtos viciantes, mas irão nos ajudar a largá-los uma vez que estamos viciados? (...) Vão ajudar soldados a matarem inimigos remotamente com *drones*, mas irão nos ajudar a ver a guerra como algo maior do que um jogo? Vão ajudar urbanistas a criarem 'cidades inteligentes', mas o que serão das nossas cidades? (...) Vão ajudar a polícia a triangular a localização de tiroteios, mas irão nos ajudar a resolver às causas mais profundas da violência? (...) Vão ajudar geneticistas a mapearem nosso genoma, mas irão nos ajudar a entender quem somos? (...) Vão nos ajudar a descobrir os fatos, mas irão nos ajudar a sermos sábios? (...) Vão nos ajudar a ver o mundo como ele é, mas irão nos ajudar a ver o mundo como ele poderia ser? (tradução livre - HARRIS, 2013)

PUC-Rio - Certificação Digital Nº 1312280/CA

Mayer-Schonberger & Cukier (2013) também questionam: “na era do *big data*, qual papel restou para a intuição, a fé, a incerteza e o aprendizado por experiências pessoais?”. Segundo Harris, o *big data*, como qualquer outra tecnologia, tem o potencial para o bem e para o mal, mas cabe a nós decidir como usá-lo.

O *designer* da experiência do usuário tem um papel crucial no que se refere a estes questionamentos. Para Flusser (2007), "no processo de criação dos objetos, faz-se presente a questão da responsabilidade". Se o *designer* é capaz de projetar sistemas de visualização de dados, ele é capaz também de interpretar e manipular tais informações. Além disso, ao projetar novos produtos e funcionalidades baseadas na tecnologia do *big data*, ele possui acesso aos dados de usuários e pode utilizá-los para fins maliciosos. Flusser ratifica a importância de tal discussão e teme que o desinteresse por essas questões possa levar à total ausência de responsabilidades.

Finalmente, uma das principais características do uso do *big data* é a possibilidade de personalização de produtos e serviços. Segundo Foster (2002), tal característica é um dos fatores responsáveis pela inflação do design, permitindo que um produto de massa por essência, mas personalizado e preciso em seu direcionamento, estimule o desejo desenfreado de consumo. Quanto mais se sabe sobre as pessoas, mais se pode ofertar e maiores são as chances de acertar.

PUC-Rio - Certificação Digital Nº 1312280/CA

## 2.5. Conclusão do capítulo

Muitas funcionalidades e serviços vêm surgindo a partir de descobertas oriundas de análise de dados, e, consequentemente o *big data* é um grande gerador de oportunidades e propicia a solução de diversos problemas. As descobertas abrem portas para melhorias e para a inovação. Muitas áreas podem se beneficiar através do uso desta tecnologia, não apenas as empresas e indústrias, mas também as instituições governamentais. Cidades inteligentes, controle de tráfego, melhorias na educação, saúde e segurança pública podem ser vislumbradas a partir da análise de dados da população.

Para que isso aconteça, os dados devem ser considerados como algo vivo, em constante mutação e análise. A construção do conhecimento a partir desse grande volume de dados variados deve ser um processo permanente.

Para os *designers*, as oportunidades são inúmeras, desde projetar sistemas que permitam a visualização de dados de forma eficiente, até projetar funcionalidades ou novos serviços baseados nessa tecnologia. Vale ressaltar a importância do diálogo com outros campos do saber, e da ética e implicações provenientes desse novo cenário em que vivemos.

Neste panorama, o conjunto de soluções tecnológicas que configuram o *big data* está cada vez mais acessível. Saiu exclusivamente da área de atuários e migrou para as ciências sociais, governo, educação, e ainda é usado para o autoconhecimento de indivíduos.

Assim, o *big data* vem transformando a forma como entendemos o indivíduo, a sociedade e como próprio se percebe. As novas tecnologias estão transformando as relações humanas e a relação do homem com a tecnologia. Esta pesquisa visa incentivar o uso do *big data* como ferramenta de construção do conhecimento e entendimento do cenário atual em que vivemos. A transformação do dado em conhecimento será tratada no capítulo a seguir.

PUC-Rio - Certificação Digital Nº 1312280/CA