A Ciência da Cultura? Computação Social, Humanidades Digitais e Analítica Cultural*

The Science of Culture? Social Computing, Digital Humanities and Cultural Analytics

LEV MANOVICH**

City University of New York. The Graduate Center. Nova York-NY, EUA

RESUMO

O artigo discute características das duas principais abordagens de pesquisa relacionadas com o estudo de amplas bases de dados: a Computação Social (*Social Computing*) e as Humanidades Digitais (*Digital Humanities*). Mostra como estas perspectivas desenvolveram-se até o momento, apontando as oportunidades e as ideias ainda não exploradas, que indicam outras dimensões a respeito do modo como essas abordagens podem ser válidas para a análise cultural e para a elaboração de uma *ciência da cultura*.

Palavras-chave: Epistemologia, Computação Social, Humanidades Digitais, Analítica Cultural, *big data*

ABSTRACT

This article discusses the characteristics of two of the main approaches of research related to the study of broad data bases: Social Computing and Digital Humanities. It shows how these perspectives developed so far, pointing out opportunities and ideas not yet explored, which indicate other dimensions concerning how these approaches can be valid for cultural analysis and for the development of a *culture science*.

Keywords: Epistemology, Social Computing, Digital Humanities, Cultural Analysis, big data

* Artigo a ser publicado no livro *The Datafied Society. Social Research in the Age of Big Data*, Amsterdam University Press, organizado por Mirko Tobias Schaefer e Karin van Es.

** Professor da City University of New York, diretor da Software Studies Initiative (softwarestudies.com). Autor e editor de oito livros, entre eles, *Data Drift* (RIXC, 2015), *Software Takes Command* (Bloomsbury Academic, 2013) e *The Language of New Media* (The MIT Press, 2001). E-mail: manovich.lev@gmail.com

ESTUDANDO O GRANDE VOLUME DE DADOS CULTURAIS: COMPUTAÇÃO SOCIAL E HUMANIDADES DIGITAIS

DEFINO A ANÁLISE Cultural como *a análise de grandes conjuntos de dados culturais e fluxos usando técnicas computacionais e de visualização*. Desenvolvi este conceito em 2005, e em 2007 nós estabelecemos um laboratório de pesquisa (Software Studies Initiatiwve, <softwarestudies.com>) para começar a trabalhar em projetos práticos. A seguir estão os exemplos de questões teóricas e práticas que estão direcionando nosso trabalho:

> O que significa representar a *cultura* por *dados*? Quais são as possibilidades especificas oferecidas pela análise computacional de grande volume de dados culturais em contraste com os métodos qualitativos usados nas humanidades e ciências sociais? Como usar técnicas quantitativas para estudar a principal forma cultural da nossa era – a mídia interativa? Como podemos combinar a análise computacional e a visualização do grande volume de dados culturais com métodos qualitativos, incluindo uma *leitura atenta*? (Em outras palavras, como combinar a análise de padrões maiores com a análise de artefatos individuais e seus detalhes?) Como a análise computacional pode fazer justiça à variabilidade e diversidade de artefatos culturais e processos, ao invés de focar no que é *típico* e *mais popular*?

Oito anos mais tarde, o trabalho de nosso laboratório se tornou apenas uma pequena parte do grande corpo de investigação. Muitos pesquisadores já publicaram milhares de documentos analisando padrões em grandes volumes de conjuntos de dados culturais. Em primeiro lugar, são os dados que descrevem a atividade nas redes sociais mais populares (Flickr, Instagram, YouTube, Twitter etc.), conteúdo criado pelo usuário e compartilhado nestas redes (tweets, imagens, vídeo etc.), e também as interações dos usuários com este conteúdo (curtidas, favoritos, compartilhamentos, comentários). Em segundo lugar, pesquisadores também já começaram a analisar determinadas áreas culturais profissionais e períodos históricos, tais como web design, moda, fotografia, música popular do século XX, literatura do século XIX etc. Esse trabalho é realizado em dois campos recém-desenvolvidos – Computação Social e Humanidades Digitais.

E onde é que isso coloca a Analítica Cultural? Acho que continua a ser relevante como programa intelectual. Como veremos, as Humanidades Digitais e a Computação Social configuram seus domínios específicos em relação aos tipos de dados culturais que estudam, mas a Analítica Cultural não tem essas limitações. Também não estamos interessados em escolher entre objetivos e

metodologia humanística vs científica ou subordinar um ao outro. Em vez disso, estamos interessados em combinar ambos nos estudos de culturas – foco no particular, na interpretação e no passado das ciências humanas e o foco nos modelos gerais, formais e predição do futuro pelas ciências. Neste artigo vou discutir essas e outras características de ambas as abordagens para o estudo de grandes conjuntos de dados culturais conforme elas se desenvolveram até agora, apontando possibilidades e ideias que ainda não haviam sido exploradas.

Estudiosos de Humanidades Digitais usam computadores para analisar principalmente os artefatos históricos criados por profissionais. Os exemplos são romances redigidos por escritores profissionais no século XIX. Com relação ao tempo, eles param nas fronteiras históricas definidas pelas leis de direitos autorais em seus países. Por exemplo, de acordo com a lei de direitos autorais (copyright) dos EUA, os trabalhos publicados nos últimos 95 anos são automaticamente protegidos por direitos autorais. (Então, por exemplo, a partir de 2015, tudo que fosse criado após 1920 seria protegido por direitos autorais, a menos que fosse de conteúdo digital recente utilizando licenças Creative Commons.) Eu entendo o respeito pelas leis de direitos autorais – mas isso também significa que os humanistas digitais se autoexcluíram de estudar o presente.

O campo da Computação Social é milhares de vezes maior. Aqui, os pesquisadores com formação avançada em Ciências da Computação estudam conteúdos criados por usuários e interações on-line com esse conteúdo. Observe que esta pesquisa é feita não somente por cientistas de computação e informação que profissionalmente se identificam com o campo da *Computação Social*[1], mas também por diversos pesquisadores de outros campos de ciência da computação, como Multimídia do Computador, Visão Computacional, Recuperação de Informações Musical, Processamento de Linguagem Natural e Ciência da Web. Portanto, a computação social pode ser também usada como um termo abrangente para toda pesquisa de ciência da computação que analisa o conteúdo e a atividade em redes sociais. Os pesquisadores lidam com dados posteriores a 2004, quando as redes sociais e serviços de compartilhamento de mídia começaram a se popularizar. (Visto que leva um a dois anos para a pesquisa e publicação de um livro, normalmente um artigo publicado em 2015 usará os dados coletados em 2012-2014.) Os conjuntos de dados são normalmente muito maiores do que os usados nas Humanidades Digitais. Dezenas ou milhares de publicações, fotos ou outros itens não são incomuns. Visto que a grande maioria do conteúdo gerado pelo usuário é criada por pessoas comuns, ao invés de profissionais, a Computação Social estuda a cultura de amadores, cultura popular por definição.

1. Veja os programas das conferências nesses campos para a variação de tópicos que abrangem: <http://cscw.acm.org/2016/submit/; http://www.www2015.it/accepted-papers/>.

A escala desta pesquisa pode ser surpreendente para praticantes das humanidades e das artes que podem não perceber quantas pessoas estão trabalhando em campos da ciência da computação e áreas relacionadas. Por exemplo, a busca no Google Scholar por *algoritmo de conjunto de dados do twitter* encontrou 102.000 artigos, a busca por *conjunto de dados de vídeo do YouTube* encontrou 27.800 artigos e a busca por *algoritmo de imagens do flickr* encontrou 17.400 artigos. Ao buscar por *conjunto de dados da estética computacional*, obtive 14.100 resultados. Mesmo os números reais sendo muito menores, ainda é impressionante. Obviamente nem todas essas publicações abordam questões culturais diretamente, mas muitas o fazem.

A tabela a seguir resume as diferenças entre os dois campos, da forma como as vejo:

Campos	A Computação Social e os vários campos da ciência da computação nas quais pesquisadores estudam as redes sociais e a mídia compartilhada	Humanidades Digitais (especificamente pesquisadores em HD que fazem análise quantitativa, utilizando técnicas de ciência da computação)
Número de publicações	Dezenas de milhares	Menos de 100
Período e material estudado	Sites e conteúdo de mídia social e atividade após 2004	Artefatos históricos até o início do século XX
Autores de artefatos estudados	Pessoas comuns, que compartilham conteúdo em redes sociais	Escritores profissionais, artistas, compositores etc.
Tamanho dos conjuntos de dados	De milhares a centenas de milhões de itens e bilhões de relações	Tipicamente centenas ou milhares de itens

Por que os cientistas da computação raramente trabalham com grandes conjuntos de dados históricos de qualquer tipo? Normalmente, eles justificam suas pesquisas por referência às aplicações industriais já existentes – por exemplo, sistemas de pesquisas ou recomendação para conteúdo on-line. O pressuposto geral é que a ciência da computação criará melhores algoritmos e outras tecnologias da computação úteis para organizações da indústria e do governo. A análise de artefatos históricos fica fora desse objetivo e, consequentemente, poucos cientistas da computação trabalham com dados históricos (o campo do Patrimônio Digital é uma exceção).

No entanto, ao olhar para muitos exemplos de artigos de ciência da computação, é claro que eles realmente estão fazendo estudos das Humanidades ou da Comunicação (em relação à mídia contemporânea) – mas em uma escala muito maior. Considere, por exemplo, estas publicações recentes: *Quantifying Visual Preferences Around the World* (Reinecke; Gajos, 2014) e *What We*

Instagram: A First Analysis of Instagram Photo Content and User Types (Hu; Manikonda; Kambhampati, 2014). O primeiro estudo analisa as preferências no mundo para o design de websites usando 2,4 milhões de classificações de 40.000 pessoas de 179 países. Obviamente, o estudo da estética e design tradicionalmente fazia parte das humanidades. O segundo estudo analisou os temas mais frequentes de fotos no Instagram – um tópico que pode ser comparado com estudos históricos de gêneros artísticos na arte holandesa do século XVII.

Outro exemplo é o livro chamado *What is Twitter, a Social Network or a News Media?* (Kwak; Lee; Park; Hosung; Moon, 2014). Publicado em 2010, foi desde então citado 3.284 vezes em outras publicações de ciência da computação[2]. Foi a primeira análise em larga escala da rede social Twitter, usando 106 milhões de tweets de 41,7 milhões de usuários. O estudo analisou especificamente os trending topics, mostrando "em quais categorias os trending topics estão classificados, quanto tempo eles duram, e quantos usuários participam". Esta é uma pergunta clássica dos estudos de comunicação, desde o trabalho pioneiro de Paul F. Lazarsfeld e seus colegas na década de 1940 que contaram manualmente os tópicos das transmissões de rádio. Porém, tendo em vista que o Twitter e outros serviços de microblogging representam uma nova forma de mídia (como a pintura a óleo, livros impressos e fotografia anteriormente), compreender a especificidade do Twitter como um meio também é um tópico das humanidades.

2. <https://scholar.google.com/citations?user=M6i3Be0AAAAJ&hl=en>.

Um pequeno número de publicações reside no cruzamento entre as Humanidades Digitais e a Computação Social. Aceitam métodos computacionais e algoritmos desenvolvidos por cientistas da computação para estudar a mídia e o conteúdo contemporâneo gerado pelo usuário e aplicá-los aos artefatos históricos criados por profissionais (ou seja, artistas profissionais, escritores, editores, músicos ou cineastas). Os exemplos proeminentes são *Toward Automated Discovery of Artistic Influence* (Saleh; Abe; Singh; Elgammal, 2014), *Infectious Texts: Modeling Text Reuse in Nineteenth-Century Newspapers* (Smith; Cordell; Dillon, 2013), *Measuring the Evolution of Contemporary Western Popular Music* (Serrà; Corral, Boguñá; Haro; Arcos, 2012), e *Quicker, faster, darker: Changes in Hollywood film over 75 years* (Cutting; Brunick; DeLong; Iricinschi; Candan, 2011).

Até poucos anos atrás, o único projeto que analisava a história cultural em uma grande escala de milhões de textos foi realizado por cientistas em vez de humanistas. Refiro-me ao *N-Gram Viewer*[3] criado em 2010 pelos cientistas do Google, Jon Orwant e Will Brockman seguindo o protótipo de dois estudantes de doutorado de Harvard em Biologia e Matemática Aplicada. No entanto, mais recentemente, vemos pessoas em Humanidades Digitais aumen-

3. <https://books.google.com/ngrams>.

tando o tamanho dos dados que estudam. Por exemplo, em *Mapping Mutable Genres in Structurally Complex Volumes* (Underwood; Black; Auvil; Capitanu, 2013), o acadêmico literário Ted Underwood e seus colaboradores analisaram 469.200 volumes da Trust Digital Library. O historiador da arte Maximilian Schich e seus colegas analisaram as trajetórias de vida de 120.000 indivíduos históricos notáveis – *A network framework of cultural history* (Schich; Song; Ahn; Mirsky; Martino; Barabási; Helbing, 2014). Conjuntos de dados históricos ainda maiores estão se tornando disponíveis nas áreas de literatura, fotografia, cinema e TV – embora continuem a ser analisados. Em 2012, os Arquivos Municipais de Nova York divulgaram 870.000 fotos históricas de NYC digitalizadas[4]. Em 2015, a HathiTrust disponibilizou dados de pesquisa extraídos de 4.801.237 volumes (contendo 1,8 bilhões de páginas)[5]. No mesmo ano, The Associated Press[6] e British Movietone[7] carregaram para o YouTube 550.000 notícias digitalizadas, cobrindo o período de 1895 até hoje[8].

Qual é a importância de ter esses grandes conjuntos de dados culturais? Não podemos simplesmente usar amostras menores? Eu acredito que existam justificativas. Em primeiro lugar, para ter uma amostra representativa, primeiro precisamos ter um conjunto muito maior de itens reais do que a amostra, ou pelo menos uma boa compreensão do que esse conjunto maior inclui. Então, por exemplo, se queremos criar uma amostra representativa de filmes do século XX, podemos usar o IMDb que contém informações sobre 3,4 milhões de filmes e programas de TV (incluindo episódios separados)[9]. Da mesma forma, podemos criar uma boa amostra de páginas históricas de jornal dos EUA usando a coleção do Historical American Newspaper de páginas digitalizadas da The Library of Congress (Biblioteca do Congresso)[10]. Mas em muitos outros domínios culturais, não existem tais conjuntos de dados maiores, e sem eles, seria impossível construir amostras representativas.

Aqui está a segunda razão. Supondo que podemos construir uma amostra representativa de um campo cultural, podemos usá-la para encontrar padrões e tendências gerais. Por exemplo, no já referido artigo *What We Instagram: A First Analysis of Instagram Photo Content and User Types* (Hu; Manikonda; Kambhampati, 2014), três cientistas da computação analisaram 1000 fotos do Instagram e criaram as oito categorias mais frequentes (selfie, amigos, moda, comida, gadget, atividade, fotos de animais, legendas). A amostra de 1000 fotos foi selecionada aleatoriamente entre um conjunto maior de fotografias compartilhadas por 95.343 usuários únicos. É possível que estas oito categorias também sejam as mais populares entre todas as fotos do Instagram compartilhadas em todo o mundo na época em que os cientistas realizaram seu estudo. No entanto, como vimos em nossos projetos nos quais foram ana-

4. <http://www.theatlantic.com/photo/2012/04/historic-photos-from-the-nyc-municipal-archives/100286/>.

5. <https://sharc.hathitrust.org/features, retreived 8/20/2015>.

6. <https://www.youtube.com/c/aparchive>.

7. <https://www.youtube.com/channel/UCHq777_waKMJw6SZdABmyaA>.

8. <http://www.ap.org/content/press-release/2015/ap-makes-one-million-minutes-of-history-available-on-youtube>.

9. IMDb, *Stats*: <http://www.imdb.com/stats>. Retirado em 10 de agosto de 2015.

10. <http://chroniclingamerica.loc.gov/about/>.

lisadas fotos do Instagram em diferentes cidades e suas regiões – por exemplo, o centro de Kiev, durante a Revolução Ucraniana de 2014 em *The Exceptional and the Everyday* (Manovich; Yazdani; Tifentale; Chow, 2014), as pessoas também compartilharam muitos outros tipos de imagens. Dependendo da área geográfica e do período de tempo, alguns destes tipos podem substituir os oito primeiros na popularidade. Em outras palavras, enquanto uma pequena amostra permite encontrar o *típico* ou o mais *popular*, ela não revela o que chamo de *ilhas de conteúdo* – tipos de conteúdo coerentes com características particulares semânticas e/ou estéticas, compartilhadas em quantidades moderadas.

PODEMOS ESTUDAR TUDO?

Quando comecei a pensar sobre Analítica Cultural em 2005, as Humanidades Digitais e a Computação Social estavam apenas começando como campos de pesquisa. Senti a necessidade de introduzir esse novo termo para sinalizar que o trabalho de nosso laboratório não será simplesmente uma parte das Humanidades Digitais ou da Computação Social, mas abrangerá o assunto estudado em ambos os campos. Como humanistas digitais, estamos interessados em analisar artefatos históricos – mas também estamos igualmente interessados na cultura visual contemporânea digital (por exemplo, o Instagram). Além disso, estamos igualmente interessados na cultura profissional, artefatos criados por amadores dedicados e artistas fora do mundo da arte (por exemplo, <deviantart.com>, "a maior rede social on-line para artistas e entusiastas da arte"[11]) e criadores ocasionais (por exemplo, pessoas que, de vez em quando, fazem o upload de suas fotos para redes de mídia social).

Assim como os cientistas de computação social e cientistas da computação, também somos atraídos para o estudo da sociedade, usando a mídia social e fenômenos sociais específicos para redes sociais. Um exemplo do primeiro é encontrar bairros semelhantes na cidade usando a atividade de mídia social, como o *The Livehoods Project*: *Utilizing Social Media to Understand the Dynamics of a City* (Cranshaw; Schwartz; Hong; Sadeh, 2012). Um exemplo do segundo é analisar padrões de difusão de informações on-line, como em *Delayed information cascades in Flickr: Measurement, analysis, and modeling* (Cha; Benevenuto; Ahn; Gummadi, 2012). No entanto, se a Computação Social enfatiza o *social* nas redes sociais, a Analítica Cultural prioriza o *cultural*. (Portanto, a dimensão mais relevante das ciências sociais para a Analítica Cultural é a sociologia da cultura e somente depois a sociologia e a economia).

Acreditamos que os conteúdos da web e das redes sociais, e as atividades do usuário nos dão oportunidade sem precedentes para descrever, modelar e

11. <http://about.deviantart.com/>, em 22/08/2015.

simular o universo cultural global enquanto questionamos e repensamos os conceitos básicos e ferramentas de humanidades que foram desenvolvidas para analisar *dados culturais menores* (ou seja, amostras culturais altamente seletivas e não representativas). Na definição muito influente do crítico cultural britânico Matthew Arnold (1869), a cultura é "o melhor que foi pensado e dito no mundo". As humanidades acadêmicas em grande parte têm seguido esta definição. E quando elas começaram a se revoltar contra seus cânones e incluir as obras de pessoas anteriormente excluídas (mulheres, pessoas que não fossem brancas, autores não ocidentais, homossexuais etc.), elas incluíam frequentemente apenas *o melhor* criado por aqueles que foram excluídos anteriormente.

A Analítica Cultural está interessada em *tudo que seja criado por todo mundo*. Com isso, estamos nos aproximando da cultura da mesma forma em que os linguistas estudam idiomas ou biólogos que estudam a vida na terra. Similarmente, nós queremos olhar para cada manifestação cultural, ao invés de amostras seletivas. (Esta perspectiva mais sistemática não é dissimilar daquela da antropologia cultural.) O maior escopo inclusivo combinando profissional e popular, histórico e contemporâneo, é exemplificado com a variedade de projetos que trabalhamos em nosso laboratório desde 2008. Analisamos o conteúdo cultural criado historicamente e profissionalmente em todas as capas da revista *Time* (1923-2009); pinturas de Vincent van Gogh, Piet Mondrian e Mark Rothko; 20.000 fotografias da coleção do Museu de Arte Moderna em Nova York (MoMA); 1 milhão de páginas de mangá das 883 séries de mangá publicadas nos últimos 30 anos. Nossa análise de conteúdo popular contemporâneo inclui *Phototrails* (Hochman; Manovich; Chow, 2013) (a comparação de assinaturas visuais de 13 cidades globais usando 2,3 milhões de fotos do Instagram), *The Exceptional and the Everyday: 144 Hours in Kyiv* (Manovich; Yazdani; Tifentale; Chow, 2014) (a análise de imagens do Instagram compartilhadas em Kiev, durante a Revolução Ucraniana de 2014) e *On Broadway* (Goddemeyer; Stefaner; Baur; Manovich, 2014) (a instalação interativa explorando a Broadway em Nova York usando 40 milhões de imagens e dados gerados pelo usuário). Nós também olhamos o conteúdo semiprofissional ou amador contemporâneo (1 milhão de obras de artes compartilhadas por 30.000 artistas semiprofissionais no <www.deviantart.com>). Atualmente estamos explorando um conjunto de dados de 265 milhões imagens publicadas no Twitter em todo o mundo durante 2011-2014. Em suma, no nosso trabalho, não traçamos um limite entre artefatos históricos profissionais (menores) e de conteúdo digital on-line (maiores), criados por não profissionais. Em vez disso, abordamos livremente ambos.

Obviamente, as redes sociais de hoje não incluem todos os seres humanos, e o conteúdo compartilhado é, por vezes, específico para essas redes (por exemplo, selfies do Instagram), em oposição a algo que existia anteriormente. Esse conteúdo também é moldado pelas ferramentas e interfaces das tecnologias utilizadas para sua criação, captura, edição e compartilhamento (por exemplo, filtros do Instagram, ou seus layouts de colagem oferecidos por outros aplicativos, como o InstaCollage). O tipo de ações culturais disponíveis também são definidas por essas tecnologias. Por exemplo, em redes sociais você pode *curtir*, compartilhar, ou comentar sobre uma parte do conteúdo. Em outras palavras, como na física quântica, aqui o instrumento pode influenciar os fenômenos que queremos estudar. Tudo isso precisa ser considerado com cuidado quando estudamos o conteúdo gerado pelo usuário e as atividades do usuário. Embora as APIs de redes sociais facilitem o acesso a grandes quantidades de conteúdo, não é *tudo* para *todos*. (API significa Interface do Usuário do Aplicativo, e é um mecanismo que permite fazer o download de grandes quantidades de conteúdo de usuário de todas as principais redes sociais. Todas as publicações de ciência do computador usam APIs para baixar os dados que analisam.)

O GERAL E O ESPECÍFICO

Quando as Humanidades estavam preocupadas com *pequenos dados* (o conteúdo criado por autores únicos ou pequenos grupos), a perspectiva sociológica era apenas uma das muitas opções de interpretação – a menos que você fosse um marxista. Mas assim que começamos a estudar o conteúdo on-line e as atividades de milhões de pessoas, essa perspectiva torna-se quase inevitável. No caso de um *grande volume de dados culturais*, o social e o cultural se sobrepõem estreitamente. Grandes grupos de pessoas de diferentes países e contextos socioeconômicos (perspectiva sociológica) compartilham imagens, vídeo, textos e fazem escolhas estéticas específicas ao fazer isto (perspectiva das humanidades). Por causa dessa sobreposição, os tipos de perguntas investigadas na *sociologia da cultura* do século XX – exemplificado pelo seu pesquisador mais influente, Pierre Bourdieu (1984), são diretamente relevantes para a Analítica Cultural.

Considerando que certas categorias demográficas foram tomadas como certas no nosso pensamento sobre a sociedade, parece natural atualmente agrupar pessoas nessas categorias e compará-las em relação aos indicadores sociais, econômicos ou culturais. Por exemplo, o Pew Research Center relata regularmente estatísticas de uso das plataformas sociais mais populares, classificando sua amos-

tra de usuário por dados demográficos, como gênero, etnia, idade, educação, renda e local de moradia – urbana, suburbana e rural (Pew Research Center, 2015), portanto, se nós estamos interessados em vários detalhes das atividades de mídia social, tais como os tipos de imagens compartilhadas e curtidas, filtros usados, ou poses de selfie, é lógico estudar as diferenças entre pessoas de diferentes países, etnias, origens socioeconômicas ou níveis de conhecimento técnico. A pesquisa inicial em computação social não o fez, e a maior parte do trabalho atual ainda não considera tais diferenças, tratando todos os usuários como um grupo indiferenciado da *humanidade* – mas mais recentemente começamos a ver publicações que dividem os usuários em grupos demográficos. Embora esta seja uma direção muito boa, também devemos ser cautelosos no quão longe queremos ir. A análise humanística de fenômenos e processos culturais usando métodos quantitativos não deve ser reduzida simplesmente à Sociologia, ou seja, considerar apenas características e comportamentos comuns dos grupos humanos.

A tradição sociológica está preocupada em descobrir e descrever os padrões *gerais* de comportamento humano, em vez de analisar ou prever os comportamentos de cada indivíduo. A Analítica Cultural está também interessada em padrões que podem ser derivados da análise de grandes conjuntos de dados culturais. No entanto, *a análise dos padrões mais amplos também nos levará a casos individuais*, ou seja, criadores individuais, suas criações particulares ou comportamentos culturais. Por exemplo, a análise computacional de todas as fotos feitas por um fotógrafo durante sua extensa carreira pode nos levar para o desviante – as fotos que são mais diferentes de todo o resto. Da mesma forma, nós podemos analisar milhões de imagens do Instagram compartilhadas em várias cidades para descobrir os tipos de imagens exclusivas para cada cidade (este exemplo vem da atual pesquisa em nosso laboratório).

Em outras palavras, podemos combinar a preocupação das ciências sociais e das ciências em geral, com o *geral* e o *regular* e a preocupação de Humanidades com o *individual* e o *particular*. (Afinal, todos os grandes artistas da história da arte situavam-se fora da curva em comparação com seus contemporâneos.) Os exemplos que acabamos de descrever de análise massiva de dados para buscar os elementos singulares ilustram uma forma de fazer isso, mas não é a única maneira.

A CIÊNCIA DA CULTURA?

O objetivo da ciência é explicar os fenômenos e chegando a modelos matemáticos sintéticos que descrevem como eles funcionam. As três leis da física de Newton são um exemplo perfeito de como a ciência clássica

aproximou-se dessa meta. Desde meados do século XIX, uma série de novos campos científicos adotou uma nova abordagem probabilística. O primeiro exemplo foi a distribuição estatística descrevendo prováveis velocidades de partículas de gás apresentada por Maxwell em 1860 (agora chama-se a distribuição de Maxwell-Boltzmann). E a ciência social? Ao longo dos séculos XVIII e XIX, muitos pensadores esperavam que, da mesma forma que a física, as leis quantitativas que regem as sociedades também seriam eventualmente encontradas (Ball, 2004). Isso nunca aconteceu. (O pensamento social do século XIX mais próximo de postular leis objetivas decorre das obras de Karl Marx.) Em vez disso, quando a ciência positivista social começou a se desenvolver no fim do século XIX e no início do século XX, ela adotou a abordagem probabilística. Então em vez de procurar as leis deterministas da sociedade, os cientistas sociais estudaram as correlações entre características mensuráveis e modelaram as relações entre as variáveis *dependentes* e *independentes*, utilizando várias técnicas estatísticas.

Após os paradigmas determinísticos e probabilísticos na ciência, o próximo paradigma foi a simulação computacional – modelos executados em computadores para simular o comportamento de sistemas. A primeira simulação de computador em grande escala foi criada em 1940 pelo Projeto de Manhattan para modelar uma explosão nuclear. Posteriormente, a simulação foi adaptada em muitas ciências duras, e na década de 1990, também foi retomada nas ciências sociais.

No início do século XXI, o volume de conteúdo digital on-line e interações com o usuário nos permitem pensar em uma possível *ciência da cultura*. Por exemplo, no verão de 2015, os usuários do Facebook compartilharam 400 milhões de fotos e enviaram 45 bilhões de mensagens diariamente[12]. Esta escala é ainda muito menor do que a dos átomos e moléculas – por exemplo, 1 cm^3 de água contém 3,33*10^22 moléculas. No entanto, já é maior que o número de neurônios em todo o sistema nervoso adulto, uma média estimada em 86 bilhões. Mas visto que a ciência agora inclui algumas abordagens fundamentais para estudar e compreender os fenômenos – leis deterministas, modelos estatísticos e simulação – em qual deles uma ciência hipotética da cultura deve se adequar?

Olhando para os artigos dos cientistas da computação que estão estudando os conjuntos de dados de mídias sociais, é claro que sua abordagem padrão é a estatística[13]. Eles descrevem o comportamento do usuário e dados de mídia social em termos de probabilidades. Isso inclui a criação de modelos estatísticos – equações matemáticas que especificam as relações entre as variáveis que podem ser descritas usando distribuições de probabilidade em vez de va-

12. <http://expandedramblings.com/index.php/by-the-numbers-17-amazing-facebook-stats/15/>, em 24/07/2015.

13. Cientistas da computação também usam muitos métodos recentemente desenvolvidos, incluindo técnicas de mineração de dados e aprendizado por máquina que não faziam parte da estatística do século XX. Discuto essas diferenças em Manovich (2015).

lores específicos. A maioria dos artigos atualmente também usam o aprendizado por máquina supervisionado – uma criação automática de modelos que podem classificar ou prever os valores dos novos dados usando exemplos já existentes. Em ambos os casos, um modelo pode representar apenas parte dos dados, e isto é típico da abordagem estatística.

Cientistas da computação que estudam os meios de comunicação sociais usam estatísticas, diferentemente dos cientistas sociais. Os últimos querem *explicar* fenômenos sociais, econômicos ou políticos (por exemplo, o efeito do contexto familiar sobre o desempenho educacional de crianças). Os cientistas da computação geralmente não estão preocupados em explicar os padrões que descobrem nas mídias sociais por referência a alguns fatores sociais, econômicos ou tecnológicos externos. Em vez disso, eles normalmente analisam fenômenos da mídia social internamente, ou tentam prever os fenômenos externos usando informações extraídas de conjuntos de dados de mídias sociais. Um exemplo do primeiro tipo é uma descrição estatística de quantas pessoas em média podem favoritar uma foto no Flickr após um determinado período de tempo[14]. O exemplo do último é o serviço Google Flu Trends que prevê a disseminação da gripe usando uma combinação de dados de pesquisa do Google e dados oficiais de gripe do CDC (Centros de Controle e Prevenção de Doenças dos EUA)[15].

14. Consultar Cha; Benevenuto; Ahn; Gummadi, 2012.

15. <http://googleresearch.blogspot.com/2014/10/google-flu-trends-gets-brand-new-engine.html>, 31/10/2014.

A diferença entre leis deterministas e modelos não determinísticos é que estes últimos apenas descrevem as probabilidades e não as certezas. As leis da mecânica clássica se aplicam a quaisquer objetos macroscópicos. Em contraste, um modelo probabilístico para prever o número de favoritos de uma foto do Flickr em função do tempo, desde que ela foi carregada, não consegue nos dizer exatamente o número de favoritos de uma foto específica. Apenas descreve a tendência global. Este parece ser o método apropriado para uma *ciência da cultura*. Se em vez disso começarmos a postular leis deterministas da atividade cultural humana, o que acontece com a ideia de livre arbítrio? Mesmo no caso de comportamento cultural aparentemente muito automático (pessoas favoritando fotos nas redes sociais com determinadas características como paisagens bonitas, bichinhos fofos ou mulheres jovens), não queremos reduzir os seres humanos a robôs mecânicos para o compartilhamento de memes.

O foco atual em modelos probabilísticos ao estudar a atividade on-line deixa de fora o terceiro paradigma científico – simulação. Até onde sei, a simulação ainda não foi explorada em Computação Social ou Humanidades Digitais como uma ferramenta para estudar o conteúdo gerado pelo usuário, seus temas, tipos de imagens, etc. Se em 2009 os cientistas do Centro de Pesquisa da IBM de Almaden simularam o córtex visual humano usando 1,6 bilhões com

9 trilhões de sinapses[16], por que não pensamos em simular, por exemplo, todo o conteúdo produzido anualmente pelos usuários do Instagram? Ou todo o conteúdo compartilhado por todos os usuários das principais redes sociais? Ou as categorias de imagens que as pessoas compartilham? O objetivo de tais simulações não será obter exatamente tudo ou prever com precisão o que as pessoas estarão compartilhando no próximo ano. Em vez disso, podemos seguir os autores dos influentes livros *Simulation for the Social Scientist* quando eles afirmam que um dos objetivos da simulação é "obter uma melhor *compreensão* de algumas características do mundo social" e essa simulação pode ser usada como "um método de *desenvolvimento da teoria*" (Gilbert; Troitzsch, 2005: 3-4, grifo meu – LM.) Uma vez que a simulação computacional requer desenvolver um modelo explícito e preciso dos fenômenos, pensar em como os processos culturais podem ser simulados, pode nos ajudar a desenvolver teorias mais explícitas e detalhadas do que as que usamos normalmente. (Para o exemplo de como a simulação baseada em agente pode ser usada para estudar a evolução das sociedades humanas, consulte *War, space, and the evolution of Old World complex societies* [Turchina; Currieb, Turnerc, Gavriletsd, 2013: 16384-16389].)

16. <http://www.popularmechanics.com/technology/a4948/4337190/, 12/17/2009>.

E o que dizer sobre o *big data*? Ele não representa um novo paradigma na ciência com seus próprios novos métodos de pesquisa? Isto é uma questão complexa que merece seu próprio artigo. (Se estamos falando sobre os métodos e técnicas de pesquisa, os desenvolvimentos no hardware do computador na década de 2000, incluindo a crescente velocidade do CPU e tamanho da RAM, e o uso de GPUs e clusters da computação, eram provavelmente mais importantes do que a disponibilidade de conjuntos de dados maiores. E embora o uso de aprendizado pela máquina, com grandes conjuntos de dados de treinamento, tenha alcançado um sucesso notável, na maioria dos casos, ele não fornece explicações dos fenômenos). No entanto, como forma de conclusão, quero mencionar um conceito interessante para as Humanidades que podemos pegar emprestado da analítica de *big data*, e em seguida, levá-lo em uma nova direção.

A ciência social do século XX estava trabalhando no que podemos chamar de *long data*[17]. Ou seja, o número de casos foi muitas vezes maior que o número de variáveis a serem analisadas. Por exemplo, imagine que pesquisamos 2000 pessoas perguntando a elas sobre sua renda, realização educacional da família e seus anos de educação. Como resultado, temos 2000 casos e três variáveis. Podemos então examinar as correlações entre estas variáveis, ou olhar para os clusters nos dados ou fazer outros tipos de análise estatística.

17. Eu estou usando este termo de forma diferente de Samuel Abresman (2013).

Os primórdios das ciências sociais caracterizam-se pelas assimetrias mais extremas deste tipo. O primeiro sociólogo positivista – Karl Marx – divide

toda a humanidade em apenas duas classes: as pessoas que possuem meios de produção e as pessoas que não, ou seja, os capitalistas e o proletariado. Sociólogos adicionaram posteriormente outras divisões. Atualmente, estas divisões estão presentes em inúmeras pesquisas, estudos e relatórios na mídia popular e publicações acadêmicas – normalmente, por gênero, raça, etnia, idade, formação escolar, renda, moradia, religião e entre outras (a lista de variáveis adicionais varia de estudo para estudo). Mas independentemente de detalhes, os dados coletados, analisados e interpretados ainda são muito *long*. As populações completas ou suas amostras são descritas usando um número muito menor de variáveis.

Mas por que este deve ser o caso? Nas áreas de análise de mídia por computador e a visualização por computador, os cientistas da computação usam algoritmos para extrair milhares de características de cada imagem, um vídeo, um tweet, um e-mail e assim por diante[18]. Por exemplo, Vincent van Gogh criou apenas cerca de 900 pinturas, essas pinturas podem ser descritas em milhares de dimensões separadas. Da mesma forma, podemos descrever todos que vivem em uma cidade de milhões de dimensões separadas extraindo todos os tipos de características de sua atividade de mídia social. Outro exemplo, o nosso próprio projeto *On Broadway*, onde representamos a Broadway em Manhattan com pontos de 40 milhões de dados e imagens, usando mensagens, imagens e check-in compartilhados ao longo desta rua no Twitter, Instagram e Foursquare, bem como dados de passeios de táxi e os indicadores do censo dos Estados Unidos para as áreas circundantes[19].

18. Eu explico o motivo para usar grande número de características em Manovich (2015).

Em outras palavras, em vez de *long data* podemos ter *wide data* – e números de variáveis muito grandes e potencialmente infinitos que descrevem um conjunto de casos. Observe que, se temos mais variáveis do que casos, tal representação iria contra o senso comum da ciência social e ciência de dados. Esta última refere-se ao processo de fazer um grande número de variáveis mais gerenciáveis como uma *redução de dimensão*. Mas para nós, os *wide data* oferecem uma oportunidade para repensar as suposições fundamentais sobre o que é a sociedade e como estudá-la; e da mesma forma, o que é cultura, uma carreira artística, um corpo de imagens, um grupo de pessoas com gosto estético semelhante e assim por diante. Ao invés de dividir a história cultural usando uma dimensão (tempo), ou duas (tempo e localização geográfica) ou um pouco mais (por exemplo, mídia, gênero), dimensões infinitas podem ser manipuladas. O objetivo da análise de *wide data* não será apenas descobrir novas semelhanças, afinidades e clusters no universo dos artefatos culturais, mas, antes de tudo, nos ajuda a questionar o nosso senso comum das coisas, onde certas dimensões são dadas como certas. E este é um exemplo do método

19. <http://www.on-broadway.nyc/>.

geral de Analítica Cultural: estranhamento (*ostranenie*)[20], tornando estranho nossos conceitos e modos culturais básicos ou organizando e entendendo conjuntos de dados culturais. Usando dados e técnicas para desenvolver o questionamento da forma como pensamos, vemos e, por fim, agimos sobre o nosso conhecimento.

20. O termo *ostranenie* foi introduzido pelo teórico literário russo Viktor Shklovsky em seu ensaio "Art as a Technique" em 1917. <http://www.vahidnab.com/defam.htm>.

AGRADECIMENTOS

Agradeço aos meus colegas dos campos da Ciência da Computação e Humanidades Digitais por muitas discussões ao longo dos anos. Minha gratidão também se estende aos alunos e pós-doutorados e cientistas pesquisadores que trabalharam em nosso laboratório desde 2007, que me ensinaram tanto. Nosso trabalho foi generosamente apoiado pela The Andrew Mellon Foundation, The National Endowment for the Humanities, The National Science Foundation, National Energy Research Scientific Computing Center (NERSC), The Graduate Center, City University of New York (CUNY), California Institute for Telecommunications and Information Technology (Calit2), University of California – San Diego (UCSD), California Humanities Research Institute, Singapore Ministry of Education e Museum of Modern Art (NYC). M

REFERÊNCIAS

ABRESMAN, S. Stop Hyping Big Data and Start Paying Attention to "Long Data". *wired.com*, 29/01/2013. Disponível em <http://www.wired.com/2013/01/forget-big-data-think-long-data/>. Acesso em 15 nov. 2015.

ARNOLD, M. *Culture and Anarchy*, London: 1869. Disponível em: <http://www.library.utoronto.ca/utel/nonfiction_u/arnoldm_ca/ca_all.html>. Acesso em 15 nov. 2015.

BALL, P. *Critical Mass*. London: Arrow Books, 2004.

BOURDIEU, P. *Distinctions. A Social Critique of the Judgment of Taste*. Cambridge, MA, Harvard University Press, 1984.

CHA, M.; BENEVENUTO, F.; AHN, Y.-Y.; GUMMADI, K. P. Delayed information cascades in Flickr: Measurement, analysis, and modeling, *Computer Networks* 56, p. 1066–1076, 2012. Disponível em: <http://200.131.208.43/bitstream/123456789/2022/1/ARTIGO_DelayedInformationCascades.pdf>. Acesso em 15 nov. 2015.

CRANSHAW, J.; SCHWARTZ, R.; HONG, J. I.; SADEH, N. The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City. *The*

6th International AAAI Conference on Weblogs and Social Media. Dublin, 2012. Disponível em: <https://s3.amazonaws.com/livehoods/livehoods_icwsm12.pdf>. Acesso em 15 nov. 2015.

CUTTING, J. E.; BRUNICK, K. L.; DELONG, J.; IRICINSCHI, C.; CANDAN, A. Quicker, faster, darker: Changes in Hollywood film over 75 years, *i-Perception*, vol. 2, p. 569- 576, 2011. Disponível em: <http://people.psych.cornell.edu/~jec7/pubs/iperception.pdf>. Acesso em 15 nov. 2015.

GILBERT, N.; TROITZSCH, K. G. *Simulation for the Social Scientist*, 2. ed., 2005.

GODDEMEYER, D.; STEFANER, M.; BAUR, D.; MANOVICH, L. *On Broadway*, 2014. Disponivel em: <http://on-broadway.net/>. Acesso em 15 nov. 2015.

HOCHMAN, N.; MANOVICH, L.; CHOW, J. *Phototrails*, 2013. Disponivel em: <http://phototrails.net/>. Acesso em 15 nov. 2015.

HU, Y.; MANIKONDA, L.; KAMBHAMPATI, S. What We Instagram: A First Analysis of Instagram Photo Content and User Types. *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media,* ICWSM, 2014. Disponível em: <http://rakaposhi.eas.asu.edu/instagram-icwsm.pdf>. Acesso em 15 nov. 2015.

KWAK, H.; LEE, C.; PARK, H.; MOON, S. What is Twitter, a Social Network or a News Media? *Proceedings of the 19th International World Wide Web (WWW) Conference*, ACM, p. 591-600, 2014. Disponível em: <http://www.eecs.wsu.edu/~assefaw/CptS580-06/papers/2010-www-twitter.pdf>. Acesso em 15 nov. 2015.

MANOVICH, L. Data Science and Digital Art History. *International Journal for Digital Art History*, issue 1, 2015. Disponível em: <https://journals.ub.uni-heidelberg.de/index.php/dah/article/view/21631>. Acesso em 15 nov. 2015.

MANOVICH, L.; YAZDANI, M.; TIFENTALE, A.; CHOW, J. *The Exceptional and the Everyday: 144 hours in Kyiv*, 2014. Disponível em: <http://www.the-everyday.net/>.

PEW RESEARCH CENTER, Demographics of Key Social Networking Platforms. January 9, 2015. Disponível em: <http://www.pewinternet.org/2015/01/09/demographics-of-key-social-networking-platforms-2/>. Acesso em 15 nov. 2015.

REINECKE, K.; GAJOS, K. Z. Quantifying Visual Preferences Around the World. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI'14, Nova York: ACM, 2014. Disponível em <http://www.eecs.harvard.edu/~kgajos/papers/2014/reinecke14visual.pdf; Yuheng>. Acesso em 15 nov. 2015.

SALEH, B.; ABE, K.; SINGH, R.; ELGAMMAL, A. A. Toward Automated Discovery of Artistic Influence. *Multimedia Tools and* Applications, Springler, p. 1-27, 19/08/2014. Disponível em: <http://arxiv.org/abs/1408.3218>. Acesso em 15 nov. 2015.

SCHICH, M.; SONG, C.; AHN, Y.-Y.; MIRSKY, A.; MARTINO, M.; BARABÁSI, A.-L.; HELBING, D. A network framework of cultural history. *Science*, vol. 345, n. 6196, p. 558-562, 1 August 2014. Disponível em: <http://www.uvm.edu/~cdanfort/csc-reading-group/schich-science-2014.pdf>. Acesso em 15 nov. 2015.

SERRÀ, J.; CORRAL, Á.; BOGUÑÁ, M.; HARO, M.; ARCOS, J. L. Measuring the Evolution of Contemporary Western Popular Music. *Nature Scientific Reports* 2, article number: 521, 2012. Disponível em: <http://www.nature.com/articles/srep00521>. Acesso em 15 nov. 2015.

SMITH, D. A.; CORDELL, R.; DILLON, E. M. Infectious texts: Modeling text reuse in nineteenth-century newspapers. *Proceedings of 2013 IEEE Conference on Big Data*, IEEE, p. 84-94, 2013. Disponível em: <http://www.ccs.neu.edu/home/dasmith/infect-bighum-2013.pdf>. Acesso em 15 nov. 2015.

TURCHINA, P.; CURRIEB, T. E.; TURNERC, E. A. L.; GAVRILETSD, S. War, space, and the evolution of Old World complex societies. *PNAS*, v. 110, n. 41, October 8, 2013.

UNDERWOOD, T.; BLACK, M. L.; AUVIL, L.; CAPITANU, B. Mapping Mutable Genres in Structurally Complex Volumes. *Proceedings of the 2013 IEEE Conference on Big Data*, IEEE, 2013. Disponível em <http://arxiv.org/abs/1309.3323>. Acesso em 15 nov. 2015.

Artigo recebido em 11 de setembro de 2015 e aprovado em 28 de setembro de 2015.