Lev Manovich

Date culturale

Posibilități și limitări ale universului datelor digitale

http://manovich.net/content/04-projects/102-cultural-data/cultural_data_article.pdf

Traducere de Miruna Pleșoianu

Digitalizarea patrimoniului cultural în ultimii 20 de ani a deschis posibilități foarte interesante pentru studiul trecutului nostru cultural folosind metode computerizate „big data". Astăzi, pentru că peste 2 miliarde de oameni creează global „cultură digitală" prin distribuirea fotografiilor, a clipurilor video, a link-urior scriind postări, comentarii, referințe, etc, putem, de asemenea, folosi aceleași metode pentru a studia acest univers al culturii digitale contemporane.

În acest capitol voi discuta un număr de probleme referitor la „forma" colecțiilor digitale vizuale pe care le avem din punctul de vedere al cercetătorilor care folosesc metode computerizate. Aceștia lucrează astăzi în multe domenii inclusiv în automatică și informatică, sociologie computerizată, istoria artei digitale, științele umaniste digitale, patrimoniul digital și Cultural Analytics – care este termenul pe care l-am introdus în 2007 referitor la toată această cercetare și, de asemenea, referitor la un anumit program de cercetare al propriului nostru laborator care s-a concentrat pe explorarea colecțiilor vizuale mari.

Indiferent de metodele analitice folosite în această cercetare, analiza trebuie să înceapă cu date concrete deja existente. „Formele" colecțiilor digitale deja existente pot ușura anumite direcții de cercetare și le pot complica pe altele. Deci, ce este universul de date creat de digitalizare, ce face el posibil sau imposibil?

Insulele și oceanul

Înainte de a se fi născut conținutul digital, creatorii de media foloseau, mai întâi, mijloace fizice și, mai târziu, electronice (video și audio). Începând de la jumătatea anilor 90, gradual, din ce în ce mai mult conținut a fost digitalizat. Putem să numim acest conținut *born analog* (născut analog).

Primul proiect care a digitalizat texte culturale și le-a făcut accesibile a fost Proiectul Gutenberg care a început în 1970. Astăzi, cele mai mari site-uri pentru conținut digitalizat sunt Europeana (peste 53 de milioane de „opere de artă, artefacte, cărți, clipuri video și înregistrări audio din întreaga Europă" la nivelul anului 2016), Digital Public Library of America (peste 13 milioane articole la nivelul anului 2016), HathiTrust (13 milioane de volume la nivelul anului 2015), Digital Collections at the Library of Congress and Internet Archive. Cel din urmă conține colecții digitale ale diferitelor tipuri de media, pornind de la cea mai mare colecție de software istorice la 10,7 miliarde de texte istorice (la momentul 12/2016).

Site-urile oferă de obicei o serie de modalități utile de a naviga în aceste colecții masive. De exemplu, Digital Public Library of America (DPLA) permite căutarea directă, vizualizarea cronologică, vizualizarea hărții site-ului și a expozițiilor tematice. Atât DPLA cât și Europeana încurajează și ajută dezvoltatorii să creeze interfețe experimentale și aplicații care extind modul în care artefactele lor pot fi vizualizate și utilizate. Dar, în ceea ce privește utilizarea acestora pentru Cultural Analytics, ele au o singură limitare. Deși lucrările din aceste colecții și din alte colecții pot fi vizualizate online oricând , nu toate pot fi descărcate din cauza restricțiilor impuse de proprietarii lucrărilor.

Site-ul care, după părerea mea, este cel mai interesant din acest punct de vedere este Google Arts & Culture. El are mai puține lucrări dar are cea mai fluidă interfață. Acest site a evoluat de la proiectul anterior Google Art care a lucrat cu multe muzee pentru a scana opere de artă pe care apoi le-a prezentat online pe o interfață numita „muzeu virtual". Astăzi, acesta oferă tururi virtuale ale multor muzee cu milioane de

opere de artă și fotografii din trecut digitalizate și, inclusiv, artă contemporană. De asemenea, sunt create proiecte media și povești fotografice. Interfețele includ instrumente de mărire-micșorare, cronologie, căutare după culoare, expoziții tematice precum și sortare (artiști, tehici, curente artistice, parteneri, nume de obiecte și locuri). Când exploram website-ul (iulie 2016), acesta oferea 3 mii de expoziții tematice cu tot felul de subiecte culturale. Când am început propriul nostru proiect Cultural Analytics Lab (culturalanalytics.info) în 2007 a fost o provocare. În timp ce cultura contemporană era deja bine reprezentată pe web, nu existau încă tipuri de colecții digitale online la scară largă cu funcții multiple de navigație și nici aplicații dedicate, cum ar fi Europeana sau DPLA. Cu toate acestea, am presupus că în următorii câțiva ani milioane de imagini digitale ale artei din toate epocile, fotografii și alte mijloace media vor deveni disponibile dar nu era clar în acel moment cât de incluzive aveau să devină aceste rețele.

În articolul pe care l-am scris despre Cultural Analytics în martie 2009 am descris experiența mea de a încerca să folosesc colecțiile de imagini digitale disponibile în acel moment. M-a interesat următoarea întrebare: Ce au pictat oamenii din întreaga lume în 1930 - în afara seriei de „ismi" moderniști care cuprindeau cel mult 150 de artiști (care lucrau la Paris, Amsterdam, Berlin și alte câteva orașe) care sunt acum incluși în canonul istoric al artei occidentale? Nu mă gândeam la „picturi în zeci de mii de muzee mici în orașe mici", mai degrabă la picturi ale unor artiști „importanți" la nivel național care au intrat în canoanele istoriei artei din țările lor.

Am făcut o căutare pe artstor.org – un serviciu comercial de vârf pentru imagini digitale de artă utilizate în majoritatea orelor de istoria artei din SUA și din alte țări. În 2009 acesta conținea deja aproape 1 milion de imagini digitale de artă, arhitectură și design. Aceste imagini au provenit din numeroase muzee importante din SUA, colecții de artă și biblioteci universitare[1]. Pentru a colecta imagini ale operelor de

[1] Prima mare colecție instituțională care a format nucleul Artstor a fost biblioteca de diapozitive a Universității din California, San Diego (UCSD) – aceeași universitate în care predam din 1996. Biblioteca avea peste 200 de mii de diapozitive care au fost toate digitalizate și incluse în Artstor. În

artă care se află în afara canonului istoric obișnuit al artei occidentale de pe Artstor, am exclus Europa de Vest și America de Nord de la căutare. Ca urmare, căutarea s-a efectuat în restul lumii: Europa de Est, Asia de Sud-Est, Asia de Est, Asia de Vest, Oceania, America Centrală, America de Sud și Africa. O suprafață mare, după cum se observă! Când am căutat în Artstor picturi realizate în 1930 în aceste părți ale lumii, am găsit doar câteva zeci de imagini. Deși a existat un număr foarte mare de imagini ale picturilor unor artiști canonici din Europa și SUA pictate în același an, au existat doar câteva imagini pentru un întreg continent precum Asia de Est.

Această distribuție extrem de inegală a artefactelor culturale digitalizate nu se datorează alegerilor lui Artstor. Acesta nu digitalizează imaginile în sine dar pune la dispoziție imagini care au fost trimise de muzee și alte instituții culturale. Rezultatele căutării noastre reflectă ceea ce colectează muzeele participante și ceea ce cred că ar trebui să fie digitalizat mai întâi. Cu alte cuvinte, o serie de mari colecții americane și o bibliotecă de diapozitive a unei universități majore de cercetare (unde până în 2007 proporția studenților asiatici era de 45%) conțineau împreună doar câteva zeci de tablouri create în 1930 în afara Occidentului care au fost digitalizate. În schimb, căutarea lui Picasso a generat aproximativ 700 de imagini. Descriind acest exemplu am scris în acest articol din 2009:

Dacă acest exemplu este o indicație, arhivarea digitală ar putea amplifica diferențele deja existente din cauza filtrelor canoanelor culturale moderne. În loc să-i transferăm pe primii 40 la coada căutărilor (top forty into the long tail), digitalizarea poate produce efectul opus.

Ceea ce rămâne în afara colecțiilor digitalizate este: ziare provinciale din sec al XIX-lea așezate undeva într-o bibliotecă; milioane de picturi în zeci de mii de muzee mici

2009 aceasta era cea mai mare colecție unică din Artstor. Diapozitivele au fost create direct de către profesorii de istoria artei care predau în cadrul Departamentului de Artă Vizuală sau de către personalul bibliotecii de artă în urma listelor de imagini oferite de facultate. Această colecție este foarte interesantă pentru că reflectă prejudecățile istoriei artei așa cum a fost ea predată de-a lungul câtorva decenii când diapozitivele color erau principalul suport pentru predarea și studierea artei.

din orașe mici din întreaga lume; milioane de mii de reviste specializate în tot felul de domenii care nici nu mai există; milioane de filme și fotografii casnice... Acest lucru creează o problemă pentru Cultural Analytics care are potențialul de a cartografia tot ceea ce rămâne în afara canonului – și de a începe să scrie o istorie culturală mai incluzivă, fără „nume mari". Vrem să înțelegem nu numai excepționalul ci și tipicul; nu doar puținele „propoziții culturale rostite de câțiva oameni mari", ci și modelele din toate propozițiile culturale rostite de toți ceilalți; ceea ce se află în afara câtorva muzee mari mai degrabă decât ceea ce se află în interior și ceea ce a fost deja discutat pe larg de prea multe ori.

M-am îngrijorat că ceea ce a fost digitalizat este doar o „insulă" și că un „ocean" cultural masiv rămâne inaccesibil pentru analize cantitative. Din fericire, o astfel de amplificare a prejudecăților și concentrarea pe „ceea ce este important" nu s-a întâmplat. Explorând bibliotecile online de artefacte culturale digitalizate, 7 ani mai târziu, sunt uimit de bogăția și varietatea lor. Motivul este că Europeana, DPLA, Library of Congress, NYPL, Internet Archive sau Google Arts & Cultures nu ne oferă doar imagini de artă de înaltă clasă precum muzeele de artă. Acestea sunt extinderi ale bibliotecilor tradiționale. Bibliotecile din timpurile moderne au o funcție importantă în afară de a oferi cititorilor cărți și periodice – sunt locuri în care numeroși oameni și organizații își donează arhivele. Pe măsură ce aceste arhive au început să fie digitalizate, un peisaj cultural istoric uimitor de bogat și variat a început să apară online.

Iată 3 exemple dintre sutele de colecții de imagini digitale de la New York Public Library (NYPL):

„Fotografii ale sistemului de alimentare cu apă Catskill în proces de construcție" - 55 de albume cu fotografii tipărite create între 1906 și 1915.

„Buttolph Collection of Menus" – o colecție a domnișoarei Frank E. Buttolph (1850 – 1924), o figură misterioasă și pasională a cărei misiune în viață a fost de a colecționa meniuri donate către NYPL în 1899, 18.964 de articole digitalizate.

„Catalogul Chiropterelor a lui G.E. Dobson” – 31 de tipărituri digitalizate dintr-o carte din 1878.

Urmează exemplele enumerate în postarea de pe blogul europeana.eu denumită „Punctele culminante ale noilor seturi de date adăugate în ultimele luni”:

Aproape 100 de artefacte (desene, picturi, fotografii) de la Telegraph Museum din Marea Britanie

Peste 3.000 de fotografii din secolele al XIX-lea și al XX-lea, în principal reprezentând clădiri de la Culture Center din Helsingborg.

O colecție de 620 de desene botanice de Georg Schweinfurth din Grădina Botanică și Muzeul Botanic Berlin – Dahlem.

Comparând aceste colecții cu cele ale ofertelor de imagini digitale ale celor mai mari muzee de artă constatăm că acestea sunt opuse între ele. Deși colecțiile muzeelor de artă modernă precum și cele ale bibliotecilor s-au dezvoltat atât prin programele lor de achiziții cât și prin donații, ceea ce le-a fost donat – sau ceea ce muzeele au ales să accepte, a fost destul de diferit. Bibliotecile au ajuns să găzduiască milioane de tot felul de articole eterogene, puține dintre ele valoroase din punct de vedere financiar. În contrast, muzeele de artă modernă s-au concentrat, în mod tradițional, pe ceea ce a fost recunoscut ca fiind foarte valoros. Într-adevăr, „muzeele” europene originare s-au constituit din averile oamenilor foarte bogați, erau părți ale palatelor regale sau ale catedralelor și bisericilor. De exemplu, Muzeele Vaticanului au apărut în 1506 când Papa Iulius al II-lea a achiziționat sculptura antică reprezentându-i pe Laocoon și pe fii săi și a expus-o publicului. (Ar trebui reținut că există colecții digitalizate de muzee de design și meșteșuguri precum Victoria & Albert din Londra sau Cooper – Hewitt din New York care sunt mai apropiate de cele ale bibliotecilor – fondurile lor sunt mai variate și, de asemenea, sunt organizate în mai multe categorii decât cele ale muzeelor de artă).

Biblioteci versus muzee

Cu toate acestea, există și un alt aspect în istoria muzeelor. Unele dintre muzeele europene originare nu conțineau artă ci „curiozități”. Un astfel de muzeu celebru este Kunstkamera care a fost înființat la Sankt Petersburg în 1716 de Petru cel Mare pentru a prezenta „curiozități și rarități naturale și umane”. Un altul este British Museum deschis la Londra în 1759 care a prezentat inițial o colecție privată a medicului și omului de știință Sir Hans Sloane.

Istoria artei, începând cu secolul al XX-lea, a creat un sistem extrem de controlat care împarte moștenirea noastră vizuală în „artă” și orice altceva și organizează „arta” în funcție de artiști (originea lor națională, perioada de timp, tehnica și stilul). Colecțiile digitale ale muzeelor de artă de astăzi par, de asemenea, ordonate și sistematizate.

Suntem obișnuiți cu clasificările lor ordonate. În comparație, meta-colecțiile de artefacte vizuale digitalizate de Europeana, DPLA și altele ne pot aminti de cabinetele de curiozități. În loc de „parade” militare de istoria artei jucate în muzeele fizice sau pe site-urile lor găsim „trivia” și „ephemera”. (Ultimul cuvânt provine din limba greacă și latină nouă care se referea la insecte sau flori ce erau vii uneori mai puțin de o zi.)

Răsfoind pagină după pagină, vom descoperi colecții nesfârșite care conțin adesea câteva zeci sau doar câteva articole – cum ar fi cele din exemplele de mai sus. În această perspectivă, trecutul pare neperiodic și nesistematizat. „Depozite” nesfârșite de material cultural uman au rămas în biblioteci, au fost apoi digitalizate și sunt acum conectate prin standarde comune de metadate, protocoale web, cod Javascript, API-uri.

Labirint, caleidoscop, Kunstkamera, hipertext Memex, memorie cu acces aleatoriu, baze de date relaționale – niciunul dintre aceste modele nu descrie experiența mea de navigare în colecțiile culturale digitale. De exemplu, luați în considerare Europeana cu cele 53 de milioane de articole. Ideea din spatele acestui proiect masiv

multianual a fost de a conecta artefacte digitalizate de la mii de muzee europene și arhive regionale. Mai degrabă decât să căutați pe site-urile lor individuale, puteți să utilizați platforma Europeana ca punct unic de acces. Platforma oferă o interfață comună tuturor articolelor dar nu le stochează. Acestea sunt stocate în muzee și arhive individuale. European Film Gateway, unul dintre proiectele Europeana, face același lucru pentru zeci de arhive de film european.

Din punct de vedere tehnic și conceptual, acest lucru funcționează strălucit dar, din punct de vedere experiențial, rezultatul are unele consecințe neintenționate. În loc să creeze un fel de „Europă unită” – un spațiu paneuropean unic pentru patrimoniul cultural – Europeana e posibil să o fragmenteze. Pe măsură ce navighez prin interminabile colecții separate sau articole individuale din aceste colecții care se potrivesc termenilor mei de căutare, țările, relațiile geografice și perioadele de timp sunt dizolvate. În loc de un continent „european”, simt că mă uit la dosarele aleatorii supraviețuitoare ale multor civilizații extraterestre care s-au amestecat.

Acest sentiment este creat atât de subiecte eterogene cât și de stiluri la fel de eterogene. Fotografii create în tot felul de tehnici, gravuri, ilustrații din ziare, desene ale pachetelor de țigări, fotografii timpurii colorate manual, picturi... imaginile sunt în format dreptunghiular, rotunde, parte a unui text, desenate într-un colț pe o scrisoare scrisă de mână... texte scrise la mașină, seturi de fonturi scrise de mână, tipărite cu imprimante matriceale timpurii sau desenate cu atenție cu pensula... fiecare subiect și formă posibilă de comunicare vizuală este aici. (Dacă platforma Instagram, în perioada 2010 – 2015, poate fi considerată un exemplu extrem de constrângeri vizuale, toate imaginile având aceeași dimensiune și aparținând aceluiași mediu, o colecție istorică digitală este la cealaltă extremă.)

Această eterogenitate, bogăție și varietate este, de fapt, un lucru bun. Ele ne fac conștienți de cât de rigide și limitate sunt astăzi conceptele noastre de „imagine” – câteva medii clar separate, formate dreptunghiulare și, de asemenea, diferențiere între imagini și texte. În timp ce abundența „speciilor” de comunicare în bibliotecile

digitale este dezorientantă la prima vedere – și este, cu siguranță, o provocare pentru analiza pe scară largă folosind sisteme Computer Vision dezvoltate inițial pentru fotografii contemporane – pe termen lung este cea mai bună pentru noi. Ne obligă să confruntăm cultura vizuală umană așa cum există într-adevăr din punct de vedere istoric – mii de variații și combinațiile lor, mai degrabă decât un set mic de categorii.

Eșantionare culturală

„Insulele" conținutului istoric digitalizat sunt în continuă creștere dar vor fi vreodată suficient de mari pentru a ne permite să înțelegem „oceanul" – adică să construim o hartă suficient de detaliată a istoriei vizuale umane din ultimele secole? Bogăția și varietatea nu înseamnă comprehensivitate. Cu alte cuvinte, în timp ce digitalizarea și organizarea articolelor digitalizate de către Europeana, DPLA și alte proiecte continuă, întrebarea de bază pentru orice studiu cantitativ de istorie culturală rămâne neadresată. Această întrebare este: Cum putem compila eșantioane reprezentative care acoperă în mod sistematic tot ceea ce a fost creat într-o anumită perioadă, zonă geografică și suport media – sau în multe astfel de perioade și zone împreună.

Antropologii folosesc metode de eșantionare în cercetările lor atunci când excavează situri sau studiază grupuri de oameni (cum ar fi în antropologia urbană care privește orașele contemporane). Există o întrebare de bază care este mai dificil de abordat: Pentru că tipurile și cantitățile de artefacte care au rămas din diferite civilizații antice variază semnificativ, acestea se pot adăuga împreună la același eșantion reprezentativ? (Desigur, pe măsură ce săpăturile de situri și analiza noilor artefacte continuă, acest eșantion este continuu rafinat.)

Întrucât sunt istoric al culturii vizuale moderne și al mediei din ultimii 200 de ani, sunt sigur că pentru această perioadă nu aveam niciun eșantion reprezentativ de cultură vizuală înainte de apariția rețelelor sociale. În timp ce „insulele" cresc ca mărime și număr, reconstrucția întregului ocean poate deveni foarte dificilă. Folosesc

termenul „eșantion” în sensul în care este folosit în statistici: un subset mai mic de date dintr-o bază mai mare de date. Construirea de eșantioane adecvate și determinarea validității predicțiilor pe baza acestor eșantioane este unul dintre principalele domenii ale statisticii. În toate științele sociale, inclusiv în sociologie, demografie, psihologie și științe politice, aceste întrebări sunt deosebit de importante deoarece aceste discipline folosesc adesea mici grupuri umane pentru sondaje sau observații. Construcția de probe adecvate este, de asemenea, crucială pentru cercetarea de marketing, cercetarea interacțiunii om – computer și pentru toate celelalte domenii aplicate în care cercetătorii doresc să găsească atitudinea oamenilor cu privire la produsele existente, interesul pentru produsele noi și noile lor caracteristici, aspirațiile stilului lor de viață, etc. Apariția cantităților mari de date din rețelele sociale în a doua parte a anilor 2000 a schimbat situația în mod semnificativ pentru că, acum, companiile pot să urmărească online milioane de persoane înregistrând paginile vizitate, click-urile, reclamele urmărite, produsele achiziționate. Grupurile mici sunt folosite astfel pe scară largă pentru cercetare. (Puteți face un mic sondaj punând diferite întrebări persoanelor care și-au dat acordul să participe sau punându-i în diferite situații și observând alegerile – lucru care nu este mereu posibil online).

Nu avem mostre sistematice de cultură vizuală și media modernă dar avem numeroase colecții și arhive separate care sunt digitalizate. Prin urmare, genul de întrebare pe care l-am pus în 2009 – Ce au pictat oamenii în întreaga lume în 1930? – este încă de nerezolvat. Pentru multe alte întrebări, situația este și mai gravă. Luați în considerare, de exemplu, istoria fotografiei. În timp ce lucram la o carte despre estetica Instagram în contextul design-ului artei și fotografiei moderne, am avut un eșantion destul de mare de imagini Instagram: 16 milioane de fotografii distribuite în 17 orașe din toată lumea între 2012 și 2016. Este important de reținut că acestea nu sunt fotografii cu anumite etichete. În schimb, toate sunt fotografii codificate geografic, distribuite în zone mai mari ale orașelor într-o anumită perioadă. Potrivit câtorva publicații de informatică care au analizat eșantioane mari de postări

Instagram în 2014, în acea perioadă utilizatorii aplicației au distribuit locații pentru 20% dintre fotografiile lor. Aceasta înseamnă că seturile noastre de date reprezintă, de asemenea, aproximativ 20% din toate fotografiile Instagram distribuite într-o anumită zonă și perioadă. Din punct de vedere al eșantionării, acestea sunt exemple foarte bune. Nu numai că sunt destul de substanțiale dar știm și ce parte a unei „populații" este reprezentată. („Populația" în statistică este un termen tehnic care se referă la datele întregi care, din motive practice, nu ne sunt accesibile. În schimb, putem folosi eșantioane mici din care putem deduce probabilistic caracteristicile tuturor datelor.)

Cu siguranță, nu m-am așteptat să găsesc ceva asemănător acestor mostre pentru fotografii vernaculare din secolul al XX-lea dar am presupus că după toate activitățile de digitalizare din ultimii 20 de ani pot găsi cu ușurință eșantioane din cel puțin câteva mii de fotografii digitalizate specifice pentru anumite țări și perioade. Se pare că nu am găsit așa ceva.

Ceea ce a fost digitalizat și pus la dispoziție online este compus din diverse colecții de fotografii vernaculare din anumite colecții private. Colecționarii au adăugat anumite fotografii la colecția lor pentru că acestea au părut interesante din anumite motive. Expozițiile muzeale de fotografie vernaculară pe care le-am consultat au fost, în mod similar, „nonobiective" – au fost asamblate de curatori care aveau propriile idei. Am găsit și câteva grupuri de utilizatori pe Flickr cu „fotografii găsite" cu care contribuiau membrii grupului. Fiecare colecție pe care am consultat-o a fost rezultatul gustului fiecărui grup în parte și a ideilor acestora despre ce ar trebui adăugat. Adesea oamenii erau interesați doar de exemple „mai artistice" și „avantgardiste" ale fotografiei vernaculare, mai degrabă decât cele tipice.

Din câte știu, nimeni nu s-a gândit vreodată să creeze un eșantion reprezentativ care să conțină caracteristicile domeniului fotografiei vernaculare în ansamblu, în anumite perioade istorice, tipuri de camere foto și tipuri de imprimare, șamd (de exemplu fotografii realizate cu camere Kodak Brownie din 1900 sau cu prima Leica portabilă

de 35 mm în 1925 sau tipărituri folosind Kodacolor după 1942 sau tipărituri Polaroid dupa 1972). Acum, pentru că am învățat din studiile de informatică pe eșantioane masive din rețelele sociale, putem să ne uităm la orice cultură ca la o pupulație statistică întrebându-ne despre distribuții, medii, variațe, clustere șamd . Dorim eșantioane istorice similare dar acestea nu există.

De exemplu, National Gallery of Art din Washington a prezentat în 2010 o expoziție numită „The Art of American Snapshot 1888-1978: din colecția lui Robert E. Jackson". Potrivit curatorilor, „organizată cronologic, expoziția se concentrează pe schimbările culturale și tehnologice care au permis și determinat aspectul instantaneelor. Aceasta examinează influența imaginilor populare precum și utilizarea posturilor definitorii, a încadrării, a specificului camerelor și a subiectelor observând modul în care acestea se schimbă în timp.

Catalogul online al expoziției arată că organizatorii au făcut o treabă excelentă în a surprinde o serie de aspecte ale fotografiei vernaculare și ale evoluției sale. Cu toate acestea, din moment ce expoziția avea doar 200 de fotografii pentru o perioadă de 90 de ani ne arată că harta istorică a expoziției avea o „rezoluție redusă" (pentru a utiliza metafora spațială) și nu era completă. Dacă vrem să înțelegem diferențele dintre fotografia instantanee din diferite țări sau să găsim schimbări treptate în stil și subiect care nu sunt legate doar de introducerea noilor tehnologii de fotografiere sau să vedem dacă pot exista unele diferențe regionale sau demografice, nu putem realiza acest lucru cu 200 de fotografii.

Pentru o comparație, luați în considerare sondajul Gallup SUA Daily. Pentru acest sondaj Gallup intervievează (la telefon) 500 de persoane din SUA în fiecare zi. Pentru o țară de 300 milioane de oameni acesta arată ca un eșantion mic dar, pentru că Gallup selectează oamenii la întâmplare și realizează aceste interviuri în fiecare zi, acumulează 15.000 de răspunsuri pe lună și 175.000 pe an. De asemenea, aflăm că „Gallup își alege eșantioanele finale pentru a se potrivi cu populația SUA în funcție de sex, vârstă, rasă, etnie hispanică, educație, regiune, densitate a populației și stare a

telefonului.” Această ponderare se face folosind datele din alte studii. De exemplu, pentru a pondera în funcție de densitatea populației Gallup folosește rapoartele recensământului SUA. Această abordare sistematică a eșantionării și analizei rezultatelor este tipică pentru toate științele naturale și sociale, administrația publică, demografia, sondajele publice, cercetările de marketing și nenumărate alte domenii. De fapt, singura zonă în care aceasta lipsește este în științele umaniste.

Întrebarea pe care și-au pus-o umaniștii este despre canoane și despre cum să le facă pe acestea mai reprezentative în domeniul lor. Există o paralelă aici cu tipul de ponderare pe care îl fac Gallup și alte organizații care colectează date demografice. Cu toate acestea, uneori, în încercările de a compensa lipsa de reprezentare a canoanelor mai vechi, noile canoane sunt „ponderate” mai mult asupra grupurilor care anterior nu erau reprezentate. Ca rezultat, obținem din nou ceva complet condus de ideologii, mai degrabă decât un eșantion echilibrat.

Un „eșantion cultural echilibrat” poate fi definit în mai multe moduri, toate la fel de informative și complementare între ele. De exemplu, putem include o proporție din toate lucrările produse în anumite medii, perioade și locuri sau ne putem concentra nu pe ceea ce a fost produs ci pe ceea ce publicul a citit, a urmărit sau a ascultat de fapt. Putem decide să selectăm numai lucrări care au obținut o anumită recunoaștere (care ar fi echivalentul aprecierilor și preferințelor din rețelele sociale contemporane) sau să ignorăm aceste informații dar, orice facem, avem nevoie de o procedură sistematică nu doar de o judecată de gust. Statistica a dezvoltat o teorie sofisticată a eșantionării care include multe metode și, deoarece aceste metode sunt folosite astăzi de toate științele, acestea ar trebui adoptate și pentru analiza artefactelor culturale istorice – dacă suntem interesați să le înțelegem ca pe un sistem ecologic sau geologic, unde toți participanții și artefactele sunt importante – spre deosebire de observarea unui singur set de „capodopere”.

Ideea creării unor mostre sistematice și reprezentative de cultură este interesantă de la sine deoarece duce la tot felul de întrebări de urmărit. Întrucât manualele,

muzeele, portalurile culturale, cursurile și documentarele noastre reprezintă întotdeauna arta și cultura umană folosind doar anumite exemple, întrebările despre eșantionarea culturală sunt importante în general chiar dacă nu efectuăm analize cantitative. Acestea se referă la modul în care înțelegem, reprezentăm și predăm istoria culturală umană – și, de asemenea, modul în care ne gândim la prezentul nostru cultural cu noua sa scară de evaluare a numărului de participanți, interacțiunile și experiențele lor culturale.

De exemplu, imaginați-vă un scenariu ipotetic în care să putem include în eșantion orice pictură creată în Franța în secolul al XIX-lea. Acum, imaginați-vă că dorim să creăm un eșantion reprezentativ, așa că, selectăm în mod aleatoriu un număr X de picturi. Un astfel de eșantion va include mai multe tablouri academice de salon, picturi realiste, portrete, etc și nu ar include acea artă din secolul al XIX-lea pe care acum o recunoaștem ca fiind cea mai importantă – lucrări ale impresioniștilor și ale postimpresioniștilor. De ce? S-a estimat că 13 artiști impresioniști francezi importanți au creat împreună 13.000 de picturi și pasteluri în timpul vieții lor dar acesta este un număr foarte mic în comparație cu toate picturile create de artiștii care trăiesc în Franța pe parcursul întregului secol al XIX-lea. Un eșantion aleatoriu probabil nu le-ar include.

Aceasta este exact aceeași problemă care însoțește o mulțime de cercetări cantitative ale rețelelor sociale în domeniul informaticii. În multe articole autorii explică modul în care construiesc cu atenție un eșantion aleatoriu extras de la toți utilizatorii Pinterest, Instagram sau Twitter. Folosind astfel de eșantioane, cercetătorii dezvoltă modele statistice care să țină seama de unele caracteristici ale comportamentului și postărilor utilizatorilor. Această cercetare este foarte interesantă și importantă dar, utilizarea unui singur eșantion global dintr-o rețea cu sute de milioane de oameni din majoritatea țărilor din lume care împărtășesc miliarde de mesaje text, imagini și videoclipuri zilnic, are limitări serioase. Nu putem vedea decât „tipicul”. Prin urmare, pierdem din vedere tot felul de variații regionale

și prezența și activitatea utilizatorilor care nu au comportamente și postări tipice. Cu alte cuvinte, dacă oricare dintre aceste rețele are proprii „impresioniști”, aceștia nu sunt vizibili în analiza care utilizează eșantioane unice.

Uneori, procedurile de eșantionare utilizate ajung să includă doar anumite tipuri de utilizatori. De exemplu, în lucrarea „Analizând activitățile utilizatorilor, demografia, structura rețelelor sociale și conținutul generat de utilizatori pe Instagram” (2014), cercetătorii afirmă: „Din câte știm, credem că aceasta este prima lucrare care realizează o analiză extinsă și profundă a rețelei sociale Instagram, a activităților utilizatorilor, a datelor demografice și a conținutului postat de utilizatori pe Instagram.” Acesta este modul în care ei descriu metoda pe care au folosit-o pentru a crea un eșantion de utilizatori pentru studiul lor:

În primul rând, am recuperat ID-urile unice ale utilizatorilor care aveau imagini ce apăreau pe cronologia publică a Instagramului folosind Instagram API care afișează un subset de media Instagram care era cel mai popular în acel moment. Acest proces a dus la un eșantion de utilizatori unici. Cu toate acestea, după o examinare atentă a fiecărui utilizator din acest eșantion, am constatat că aceștia erau, în mare parte, vedete (ceea ce explică de ce postările lor erau atât de populare). Pentru a evita părtinirea eșantionării, pentru fiecare utilizator din acest eșantion am accesat ID-urile atât ale urmăritorilor lor cât și ale prietenilor lor iar ulterior am combinat două liste pentru a forma o listă unică de utilizatori ce cuprindea un milion de utilizatori unici.

Setul de date final are 5.659.795 de imagini pentru 369.828 de utilizatori (restul aveau conturi private). Dintre aceste imagini 1.064.041 au locații geografice. Cât de bine reprezintă acești utilizatori universul Instagram? Majoritatea oamenilor urmăresc alți oameni spre deosebire de celebrități. Persoanele care urmăresc celebritățile și prietenii acestora reprezintă, probabil, un singur tip de utilizator Instagram. În plus, având în vedere că numărul de utilizatori Instagram din fiecare țară diferă, cele mai mari țări având adesea un număr mai mare de utilizatori, un

astfel de eșantion „aleatoriu” reprezintă, probabil, mai bine unele țări decât pe altele.

Aceste considerații nu invalidează rezultatele din această lucrare și din toate celelalte lucrări care utilizează un singur eșantion mare din rețelele sociale globale. Constatările lor sunt valide. Este posibil să nu se aplice fiecărui tip de utilizator sau de postare pe astfel de rețele. (Rețineți că nu vorbim despre utilizatori individuali ci despre grupări, fiecare cu propriile caracteristici. Cu alte cuvinte, aceștia sunt ca impresioniștii secolului al XIX-lea care aveau caracteristici comune.)

Trebuie să ne amintim aici poate cel mai fundamental „călcâi al lui Ahile” al statisticilor. „Scopul statisticii este de a reprezenta faptele în modul cel mai condensat” (1833). Plătim un preț mare pentru o astfel de compresie. Măsurile utilizate în statisticile descriptive rezumă o anumită populație (adică un set de elemente) dar acestea nu pot corespunde cu niciun membru concret al acestei populații. De exemplu, să luăm o serie de numere: 1,1,2,3,2,9,9,10,11,11,11. Media (numită „medie” în statistici) a acestei serii este de 6,36 dar nu avem niciun număr real apropiat de această medie! În schimb, avem două „clustere”: de la 1 la 3 și altul de la 9 la 11. (Aceasta se numește distribuție bimodală).

Cu alte cuvinte, măsurile statistice standard ale unei populații mari pot rata cu ușurință prezența diferitelor grupări în această populație. Dacă reprezentăm o „populație culturală” – fie că este vorba despre picturi din sec. al XIX-lea sau cinematografie din sec. al XX-lea, despre Instagram de astăzi sau videoclipuri muzicale globale – cu un singur eșantion aleatoriu, putem rata toate tipurile de grupări (New Wave din anii 1960 sau școala de montaj sovietic din anii 1920 din istoria cinematografiei; videoclipuri muzicale contemporane din India, Coreea, Vietnam, Thailanda sau Kazakstan care au propriile diferențe în ciuda asemănării generale șamd). Caracteristicile pe care le vom găsi pot descrie „media” care nu a existat niciodată în realitate iar aceasta s-ar putea să nu corespundă unui grup real.

Mai degrabă, decît să surprindă prezența mai multor grupuri distincte, le poate ascunde de vedere.

De fapt, aș dori să susțin că în societățile și culturile umane nu există „medii”. Cu siguranță, îl putem urmări pe Adolphe Quetelet care, la începutul anilor 1830, a fost primul care a început să măsoare caracteristicile fizice ale oamenilor, cum ar fi înălțimea și greutatea, și a constatat că distribuțiile lor au urmat „curbele normale”. Dacă vom efectua astfel de măsurători, astăzi vom găsi distribuții similare. Într-un eșantion de 1 milion de oameni, cu siguranță mulți ar avea înălțimea exactă specificată de medie. În același mod, dacă măsurăm, de exemplu, lungimea a zeci de mii de romane moderne vom descoperi că unele au exact aceeași lungime ca media.

Astfel de rezultate sunt valide numai dacă limităm studiul artefactelor culturale, al interacțiunilor și experiențelor la o singură caracteristică la un moment dat. Dacă ne uităm la mai multe selfie-uri eșantionate de pe Instagram putem calcula gradul mediu de zâmbet, dimensiunea unei fețe din fotografie și poziția acesteia. Dacă dimensiunea eșantionului este suficient de mare, unele selfie-uri reale vor avea exact aceleași cifre ca media dar, la fel cum fața fiecărei persoane este unică precum amprentele sale, fotografiile lor sunt, de asemenea, unice. Dacă înmulțim numărul de caracteristici, în cele din urmă nu vom găsi niciun selfie real care să se potrivească cu media eșantionului. Același lucru se aplică oricărui alt tip de expresie culturală din trecut sau din prezent.

Există un domeniu care se gândește la eșantionarea culturală și folosește metode statistice pentru a crea și apoi a analiza eșantioane. Acest domeniu este sociologia culturii. Cea mai cunoscută carte din acest domeniu rămâne faimoasa „Distincție: o critică socială a judecății gustului” de sociologul francez Pierre Bourdieu. Publicată în 1979, a fost recunoscută ca una dintre cele mai importante zece cărți de sociologie din secolul al XX-lea. Bourdieu a oferit idei și teorii intelectuale puternice care conectau gusturile culturale ale oamenilor și statutul lor socio-economic. Aceste teorii s-au bazat pe analiza statistică a doua anchete mari ale gusturilor publicului

francez efectuate în anii 1960. Bourdieu a colaborat cu „data scientists” francezi (pentru a utiliza termenul contemporan) care au dezvoltat o nouă metodă analitică și de vizualizare pentru a reprezenta relațiile dintre multele elemente și a folosit această metodă în toate studiile sale ulterioare inclusiv în „Distincție”.

În prezent, sociologii culturii continuă să folosească anchete ale unor grupuri de oameni dar folosesc și mostre din publicații culturale. Un exemplu, ar fi un studiu în care cercetătorii „au cerut celor 1.544 de participanți vorbitori de limbă germană să enumere adjective pe care le folosesc pentru a eticheta dimensiunile estetice ale literaturii în general și ale formelor și genurilor literare în special (romane, nuvele, poezii, piese de teatru, comedii). Un alt exemplu este un studiu numit „Recunoaștere instituțională în domeniul literar transnațional 1955-2005”. Acesta folosește „un eșantion de articole din 1955, 1975, 1995, 2005 din lucrări de elită franceze, germane, olandeze și americane (N=2.419).” Iată un alt exemplu: o analiză a discursului modei în perioada 1949-2010 care folosește 1.301 recenzii de modă de la New York Times și The International Herald Tribune. Deși astfel de eșantioane sunt destul de mici în comparație cu scara rețelelor sociale acestea sunt suficiente pentru a răspunde la întrebări specifice pe care cercetătorii le-au pus în aceste studii.

Când m-am gândit pentru prima dată la Cultural Analytics în 2005 mi-am imaginat că pot construi hărți detaliate la nivel mondial ale unor domenii particulare – cum ar fi pictura, cinematografia, design-ul grafic sau videoclipul muzical – pentru perioade istorice îndelungate. De vreme ce mi-am dat seama că eforturile de digitalizare nu creează eșantioane sistematice de care ar fi nevoie pentru astfel de hărți, a trebuit să renunț la aceste idei pentru moment. În schimb, m-am concentrat pe un alt tip de eșantionare pe care l-aș putea face având în vedere ceea ce a fost digitalizat – după tipul de media. Începând din 2008, în laboratorul nostru am lucrat la peste 40 de seturi de date care acoperă aproape toate tipurile majore de medii vizuale de astăzi. Am analizat benzi desenate și serii Manga, jocuri video, lungmetraje, documentare, grafică video, videoclipuri muzicale, reclame video politice, reviste tipărite, fotografii

istorice, fotografii digitale și alte imagini și lumi virtuale interactive. De asemenea, am inclus, în mod deliberat, seturi de date care se află la extreme (înalt – scăzut, profesional – neprofesional): de la picturile lui Van Gogh, Mondrian și Rothko la 10 milioane de fotografii Instagram distribuite în New York de 5 milioane de oameni. Am echilibrat în mod deliberat surse culturale occidentale și nonoccidentale. Acestea din urmă includ jocuri video japoneze, videoclipuri muzicale din toată Coreea, fotografii pe Instagram distribuite în 17 orașe globale care acoperă 4 continente. Am publicat analize folosind fotografii de Instagram distribuite în Tel Aviv, Israel, în timpul Zilei Comemorării Soldaților Căzuți și Victimelor Terorismului și o altă analiză a fotografiilor de Instagram distribuite în timpul revoluției Maidan din februarie 2014 Kiev, Ucraina.

De fapt, avantajul utilizării datelor din rețelele sociale este că acestea nu sunt „canonice" sau „naționale". Rețelele populare precum Facebook, Instagram și altele sunt utilizate în fiecare țară cu excepția câtorva în care sunt/au fost blocate pentru perioade de timp (în cazul Facebook Bangladesh, China, Iran, Coreea de Nord, Siria). În mai 2016 aplicația de mesagerie WhatsApp care a apărut în China era utilizată în 109 țări cu 1 miliard de utilizatori care trimiteau zilnic 42 miliarde de mesaje. În același timp, 80% din cei 500 milioane utilizatori activi Instagram se aflau în afara SUA.

De exemplu, atunci când cream seturile de date din Instagram între 2012 și 2016, Instagram API a permis oricui să descarce toate fotografiile geo-etichetate distribuite într-o anumită zonă dreptunghiulară definită de latitudine și longitudine. Fiecare zonă ar putea avea o dimensiune de 5 km x 5 km iar colectarea dintr-o serie de zone nu a fost complicată. Așadar, a fost la fel de ușor să descărcați imagini din părți din Manhattan, Moscova, Bangkok sau Kiev, șamd. (Pentru a descărca toate imaginile geo-etichetate distribuite pe parcursul a 5 luni în Manhattan am combinat o serie de zone pentru a închide insula într-un dreptunghi mare apoi am scos datele provenind din afara Manhattan-ului.)

Aceasta înseamnă că, în practică, compararea multor zone din întreaga lume este la fel de ușoară ca și compararea zonelor apropiate din același oraș – întrucât oamenii distribuie cantități suficiente pe rețelele sociale în aceste zone globale. Perspectiva globală este „încorporată” în rețelele sociale. Acest lucru se aplică, desigur, și formatelor standard, constrângerilor și accesibilităților pe care rețelele și aplicațiile le oferă utilizatorilor lor. Toți cei care au folosit Twitter între 2007 și 2017 au fost nevoiți să-și încadreze mesajele în aceleași 140 caractere. Toți cei care foloseau Instagram între 2010 și 2015 au trebuit să se restrângă la formatul său de imagine pătrat cu aceleași dimensiuni 640x640. Toată lumea are acces exact la aceleași funcții (adăugarea de hashtag-uri, etichetare geografică opțională, etc) și la aceeași interfață de utilizare. Acest lucru ridică de la sine o întrebare importantă: Soft-ware-ul rețelelor sociale duce la o mai mică diversitate a conținutului generat de utilizatori? Aceasta a fost una dintre întrebările cheie pentru mine în cei 8 ani de cercetare.

Reprezentarea datelor

La fel ca orice alt tip de date despre societate, datele din rețelele sociale au propriile limitări și acestea sunt semnificative. Voi discuta pe scurt 5 probleme care se referă la reprezentare – ce este reprezentat (și disponibil pentru cercetare) și ce este absent. În timp ce utilizarea rețelelor sociale și a internetului continuă să crească în jurul lumii, miliarde de oameni totuși nu le folosesc. Iată un exemplu concret din propria noastră cercetare a modului în care această situație limitează ceea ce putem „vedea” folosind datele lor. În 2014 Twitter a fost de acord să ofere cercetătorilor selectați acces la orice parte a datelor dacă erau folosite în noi feluri interesante. 13 mii de laboratoare din întreaga lume au aplicat și noi am fost unul dintre cele 6 laboratoare care au câștigat. I-am rugat pe cei de la Twitter să ne ofere toate tweet-urile cu imagini geo-localizate incluse. Twitter a adăugat funcția de a distribui imagini în 2011 și ni s-a dat acces la toate tweet-urile cu imagini geografice distribuite în întreaga lume între 2011 și 2014. Când am trasat locațiile unui eșantion aleatoriu de

100 milioane de tweet-uri am constatat că aproximativ jumătate din suprafața populată a globului nu era acoperită.

A doua problemă are legătură cu datele demografice ale utilizatorilor care folosesc rețelele sociale. În țările „dezvoltate” și în marile metropole globale, oamenii din toate grupurile demografice folosesc rețelele sociale. Într-o țară precum SUA nu există diferențe semnificative în ceea ce privește utilizarea rețelelor sociale între femei și bărbați sau rase diferite sau persoane cu nivel de educație diferit – dar există încă diferențe mari între grupurile de vârstă. Acest lucru este valabil și la nivel global – deși diferențele devin din ce în ce mai mici cu timpul. Un raport privind utilizarea rețelelor sociale în rândul persoanelor care erau online în 34 de țări în primul trimestru din 2016 a constatat că 92% dintre cei care se află în grupul de vârstă 45-54 ani au conturi de socializare; pentru persoanele din grupa de vârstă 55-65 cifra este de 82%.

În multe țări în curs de dezvoltare, proporțiile persoanelor care utilizează rețelele sociale dintre cele care folosesc internetul sunt mai mici decât în țările dezvoltate. La început, aceasta pare o veste bună deoarece ar putea însemna că obținem date despre activitățile culturale ale unei proporții mai mari din populația acelor țări. Cu toate acestea, realitatea este diferită. După cum se explică în raport, „98% dintre utilizatorii de internet din țări precum Malaezia, Brazilia, Indonezia și Vietnam se află pe cel puțin o rețea. În parte, acesta este rezultatul nivelurilor lor mai scăzute de utilizare a internetului, ceea ce înseamnă că adulții online din aceste regiuni au mai multe șanse decât omologii lor din Europa sau America de Nord să provină din segmente tinere urbane și relativ bogate.

A treia problemă este distribuția spațială inegală a activității și a conținutului rețelelor sociale chiar și în zonele urbane mari unde vedem o utilizare foarte mare până când ne uităm mai atent. Am strâns și analizat 7.442.454 imagini publice Instagram etichetate geografic distribuite în Manhattan pe parcursul a 5 luni. Inegalitatea pe care am găsit-o între părțile mai populate și mai puțin populate din

Manhattan a fost uluitoare. Am constatat că raportul dintre o suprafață de km pătrați cu cele mai multe imagini și aria cu cele mai puține imagini a fost de 250.000:1. Conform analizei noastre 50% din toate imaginile distribuite de rezidenții locali se află în numai 21% din zona Manhattan. Pentru vizitatori, această diferență este aproape de 2 ori mai mare: 50% din imaginile lor au fost distribuite în doar 12% din zona Manhattan-ului. Pe scurt, chiar și pentru o zonă urbană atât de dens populată ca Manhattan, imaginea sa colectivă din Instagram reflectă doar o parte din realitate.

A patra problemă se referă la ce conținut distribuie oamenii, ce comentarii fac și ce sunt dispuși să spună online. Rețelele sociale nu sunt o oglindă a societății. Așa cum oamenii, în unele zone din viața lor, joacă roluri, respectă norme, prezintă identități particulare și se comportă în modurile așteptate de la ei, ei fac acest lucru și online pentru că postările și comentariile lor pot fi văzute de toți ceilalți utilizatori ai rețelei (cu excepția cazului în care fac postări private sau au întregul cont privat), apar în căutarea Google și sunt salvate de rețele, împărtășite cu specialiști în marketing, etc, este probabil să fie foarte atenți la ce postează. La fel ca în cazul produselor culturale profesionale, o parte din conținutul generat de utilizatori este condus de convenții, stereotipuri și modele pe care oamenii le văd în jurul lor. De exemplu, găsim foarte multe fotografii în genul „table top” pe Instagram create de utilizatori obișnuiți, proporții copleșitoare de zâmbete în selfie-uri, iar fotografiile de călătorie își respectă propriile convenții. Cu alte cuvinte, „cultura” pe care o putem analiza folosind rețelele sociale are propriul ei univers și nu este un eșantion de activități culturale, gust și opinii ale oamenilor din afara rețelelor.

În cele din urmă, a cincea problemă se referă la accesul la datele din rețelele sociale. La jumătatea anilor 2000, toate rețelele sociale mari au creat API-uri care permit oamenilor să descarce liber eșantioane mari de date care conțin postări ale utilizatorilor și toate informațiile publice despre ele vizibile online – data și ora în care a fost distribuită o postare, locația (dacă utilizatorul a distribuit aceste

informații), numele de utilizator, etichetele, comentariile și numărul de aprecieri și redistribuiri. În cazul rețelelor vizuale precum Instagram și Flickr, imaginea și videoclipul împreună cu descrierile utilizatorilor și toate celelalte informații au fost, de asemenea, disponibile pentru descărcare. Flickr și-a lansat API-ul în 2004 iar Facebook și Twitter în 2006.

În timp ce aceste API-uri au fost destinate dezvoltatorilor care construiesc aplicații ce utilizează date de pe platforme și, de asemenea, au fost destinate utilizatorilor pentru ca ei să poată distribui conținut între rețele și bloguri, cercetătorii din domeniul informaticii, artiștii de vizualizare a datelor și alți tehnologi creativi și-au dat seama că pot accesa și ei în mod liber aceste date. Astfel, numeroase studii și proiecte au fost create. Sute de mii de oameni de știință în computere și științe sociale și studenți au folosit aceste API-uri pentru a descărca date, a le analiza și a publica lucrări.

Cu toate acestea, au existat întotdeauna limite cu privire la cât de multe date pot fi descărcate. De exemplu, în perioada în care descărcam activ datele Instagram (2012-2016), această rețea avea o limită de 3.000 de imagini pe oră și erau disponibile doar imaginile din ultimele zile. Cu toate acestea, am reușit să adunăm 16 milioane de fotografii Instagram distribuite în 17 orașe globale în diferite perioade între 2012 și 2016. Având în vedere că în 2016 oamenii distribuie 80 milioane de imagini pe Instagram pe zi, ceea ce am reușit să strângem a fost o mică porțiune.

Din cauza preocupărilor legate de confidențialitate și de utilizarea neautorizată a postărilor, unele dintre cele mai mari rețele au limitat sau au închis treptat accesul API la datele în bloc ale utilizatorilor. Facebook a limitat utilizarea API-ului său la 30 aprilie 2015 iar Instagram a încetat să mai permită descărcări în bloc pe 1 iunie 2016. În acest moment (sfârșitul anului 2016), Twitter este încă accesibil împreună cu unele rețele populare din anumite zone geografice precum VK din Rusia.

Pe scurt, știm că rețelele sociale și internetul nu sunt utilizate de toată lumea; proporțiile și demografia celor care utilizează rețelele sociale variază de la un loc la

altul iar ceea ce publică și distribuie oamenii constituie o realitate culturală proprie și nu este o fereastră transparentă către realitățile din afară. Ar trebui să avem întotdeauna în vedere aceste limitări. În același timp, folosind datele web, rețelele sociale și tehnologiile contemporane de urmărire și analiză a cestora, se pune sub semnul întrebării chiar ideea de reprezentare. Aceasta privește chiar fundamentul metodelor moderne de cercetare bazate pe eșantionare.

Aceste metode presupun că, din motive practice, nu putem avea acces la „populația" completă (adică date complete). Putem accesa și analiza doar unul sau mai multe eșantioane ale populației. În consecință, statisticile moderne sunt împărțite în două categorii. Statistica inferențială este un set de metode pentru estimarea caracteristicilor populației pe baza eșantioanelor acesteia. Statisticile descriptive descriu doar proprietățile oricăror date pe care le avem și nu presupune că aceste date provin dintr-o populație mai mare.

Cu toate acestea, atunci când analizăm conținutul și interacțiunile de pe internet și de pe rețelele sociale, de multe ori putem avea date complete. Cu siguranță, companiile care gestionează rețelele sociale, site-urile de partajare media sau platformele de publicare pot înregistra toate interacțiunile care au loc pe platformele lor. Acest lucru este valabil pentru Facebook, YouTube, Twitter, Pinterest, Spotify, Amazon, Scribd, Shutterstock, Behance, academia.edu și alte servicii de socializare și publicare. Acest lucru nu înseamnă că o companie își va analiza toate datele sau le va păstra pentru totdeauna sau chiar va avea proprii cercetători care lucrează la acestea – deoarece companiile nu vor să fie date în judecată, să aibă publicitate proastă sau să aibă probleme cu guvernele. Așadar, datele sunt anonimizate, eșantionate la nevoie și numai anumite părți ale datelor sunt puse la dispoziția cercetătorilor interni în funcție de laboratorul în care lucrează. Cu toate acestea, marile companii profită cu siguranță de faptul că au date masive despre interacțiunile utilizatorilor pe platformele lor folosindu-le pentru a instrui sisteme care recomandă altor utilizatori pe cine să urmărească sau ce videoclipuri să vizioneze și decid ce postări de la

prieteni să fie afișate, selectează subiecte populare, etc. Datele sunt, de asemenea, principala sursă de venit pentru marile companii ale rețelelor sociale – de exemplu, sisteme automate de publicitate precum Google AdWords și Facebook Ads.

Deși cercetătorii universitari nu au acces direct la date complete de la aceste companii este posibil să utilizeze API-urile lor pentru a descărca date complete care îndeplinesc criterii specifice cum ar fi toată activitatea pe o anumită platformă într-o anumită perioadă de timp. Multe lucrări folosesc astfel de seturi de date. În propria noastră lucrare am folosit această abordare. Foloseam API-ul Instagram pentru a descărca toate imaginile geocodate distribuite public pe o anumită perioadă de timp. De fapt, fiecare set de date Instagram pe care l-am folosit a fost generat în acest fel. De exemplu, pentru a crea un set de date de 7.442.454 imagini Instagram publice distribuite în Manhattan pe parcursul a 5 luni, am folosit un singur Mac pentru a rula programul nostru personalizat de descărcare 24/7 în toată această perioadă. Din câte știm, imaginile pe care le-am descărcat sunt toate imaginile pe care oamenii le-au partajat în această zonă și timp cu geolocație (care reprezintă aproximativ 20% din ceea ce a fost partajat în total.)

De ce am dori să folosim date culturale complete? Dacă ne interesează doar extragerea tiparelor, caracteristicilor și tipurilor generale – de exemplu, cele mai frecvente 10 tipuri de imagini pe Instagram – cu siguranță nu avem nevoie de toate datele. O astfel de rezumare și agregare comună a utilizării metodelor statistice în secolele XIX și XX este doar o modalitate de a utiliza datele culturale. Așa cum am explicat mai sus, utilizarea unor eșantioane mici din diverse „populații” culturale (cum ar fi trilioane de imagini Instagram) pot dezvălui doar „tipicul” și „cel mai popular” și pot rata „variațiile regionale” și „prezența și activitatea utilizatorilor care nu au comportamente și postări tipice”. Prin urmare, în mod ideal, Cultural Analytics ar trebui să încerce să obțină și să analizeze date complete generate de un anumit proces cultural (fie că este vorba despre cariera unui singur fotograf sau toate fotografiile distribuie pe Instagram).

În loc să trateze cultura doar ca „puncte de date” care împreună creează modele pe care dorim să le descoperim, ignorând punctele individuale, Cultural Analytics ar trebui să acorde o atenție egală atât modelelor cât și artefactelor individuale, experiențelor și interacțiunilor. Fiind creatori și membrii ai publicului ne angajăm și ne bucurăm de artefacte și de experiențe concrete și nu de „modele”. Un artefact deosebit de reușit este adesea descris ca „unic” – adică nu poate fi redus la modele deja existente. Ca subiecte estetice căutăm și ne bucurăm de o asemenea unicitate. Unul dintre obiectivele Cultural Analytics este de a ne ajuta să găsim artefacte cu adevărat unice în universurile infinite de media care sunt acum create. Chiar dacă alte artefacte nu sunt unice în cele mai multe moduri, ele pot avea, totuși, ceva unic în alte moduri care se pot pierde dacă le reducem la modele. De exemplu, fiecare față umană este unică și, prin urmare, chiar și cea mai convențională fotografie a acestei fețe va fi specială pentru noi. (În acest aspect Cultural Analytics ar trebui să combine o perspectivă specială a științelor și a științelor umaniste – preocuparea primelor cu legile și regularitățile generale și preocuparea celor din urmă cu obiecte culturale unice.)

În concluzie, aș dori să remarc dezvoltarea tehnico-culturală din ultimii 20 de ani care leagă multe probleme despre care am discutat – creșterea căutării ca un nou mod dominant de interacțiune cu informațiile. Această evoluție este doar una dintre numeroasele consecințe ale dezvoltării dramatice și rapide a informațiilor și a conținutului pe care le-am experimentat de la mijlocul anilor 1990. Pentru a difuza rezultatele căutării, Google, Bing, Baidu, Yandex și alte motoare de căutare analizează mai multe tipuri diferite de date – inclusiv metadatele anumitor pagini web (așa numitele „meta elements”) și conținutul acestora. De exemplu, potrivit Google, algoritmul motorului său de căutare folosește mai mult de 200 de tipuri de intrări.

Cu toate acestea Google, Yandex sau Bing nu dezvăluie măsurătorile paginilor web pe care le analizează – servesc doar la propriile concluzii adică ce web-site-uri se

potrivesc cel mai bine temei de căutare introdusă de utilizator determinată de algoritmii proprii care combină aceste măsuri. În schimb, scopul Cultural Analytics este de a permite ceea ce am putea numi „căutare culturală profundă” – de a oferi utilizatorilor instrumente open-source, astfel încât aceștia să poată analiza în detaliu orice tip de conținut cultural și să poată utiliza rezultatele acestei analize în moduri noi.