Big Data Variety,
or I've got your number.

Tidewater Big Data Enthusiasts
Chuck Cartledge
Developer

22 March 2016

Contents

List of Figures

List of Tables

1 Introduction

Doug Laney has been credited with identifying the original Big Data 3Vs: volume, velocity, and variety[1]. He characterized these as being part of the 3D Data Management problem that was "breaking" traditional relational database management systems. His view of the 3Vs were from a business merger and acquision perspective (see Figure 1). His characterization of Big Data caught on and was expanded into 4Vs (now including veracity) (see Figure 2) [3], and then swelled to at least seven[2].

So depending on who you ask when, Big Data has a lot of Vs:

1. Volume – companies have more data and are reluctant to discard it[1]
2. Velocity – interaction with data happen faster and faster because of new technology[1]
3. Variety – more data sources lead to more different data formats[1]
4. Veracity – can the data be trusted [3]
5. Variability – the variety of data formats means that traditational relational database management systems may not be appropriate [2]
6. Visualization – the scope and size of the data requires new and novel ways of exploring and underestanding the data [2]
7. Value – proper use of Big Data can have direct economic impact (more targeted marketing, increased efficiencies, discovery of unknown relationships, etc.) [2]

We are going to explore Big Data a little

2 Discussion

Lets focus a little on different types of numbers that are requested and available on the Internet.

2.1 Phone numbers

We enter phone numbers all the time when we use the internet (see Figure 3). American phone numbers have a specific format known as the North American Numbering Plan (NANP). The NANP is a closed plan (meaning that all phone numbers are of a fixed format), and defines the three parts of a phone number (see Table 1). A simple thing like entering a phone number can have many different forms and formats. If a human is entering the phone number, the particular nuances of the form can be easily understood. Having a computer enter the data may be more problematic.

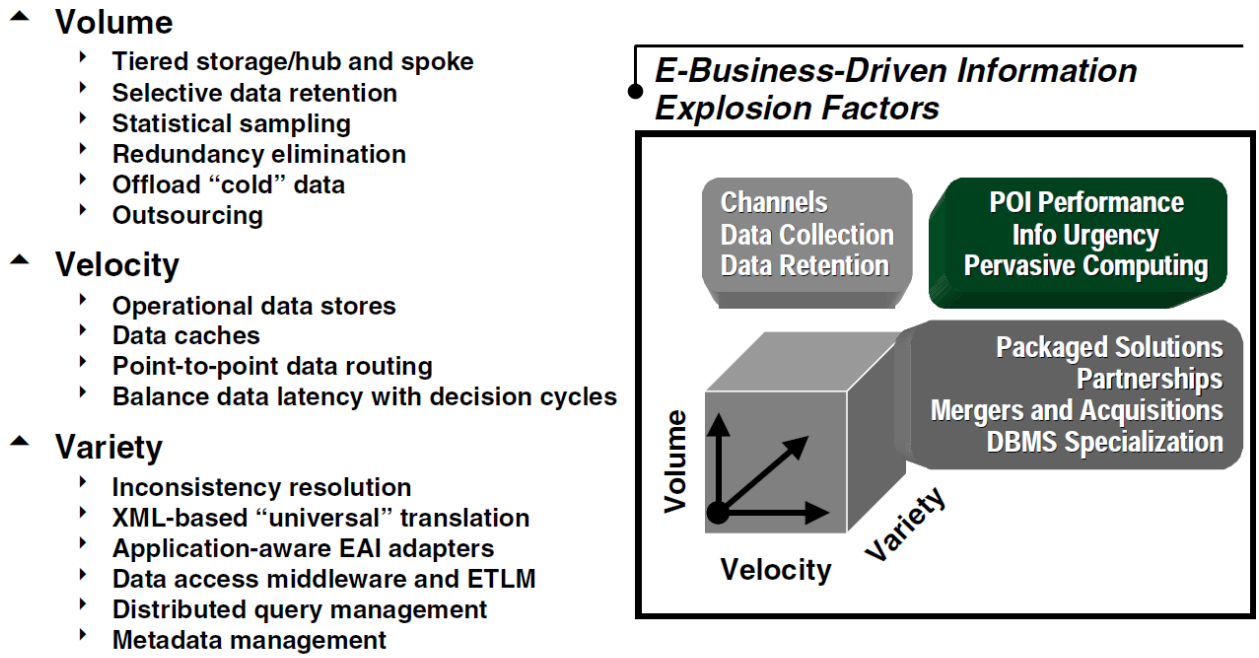


Figure 1: Doug Laney's original Big Data 3Vs. Original image from [1].

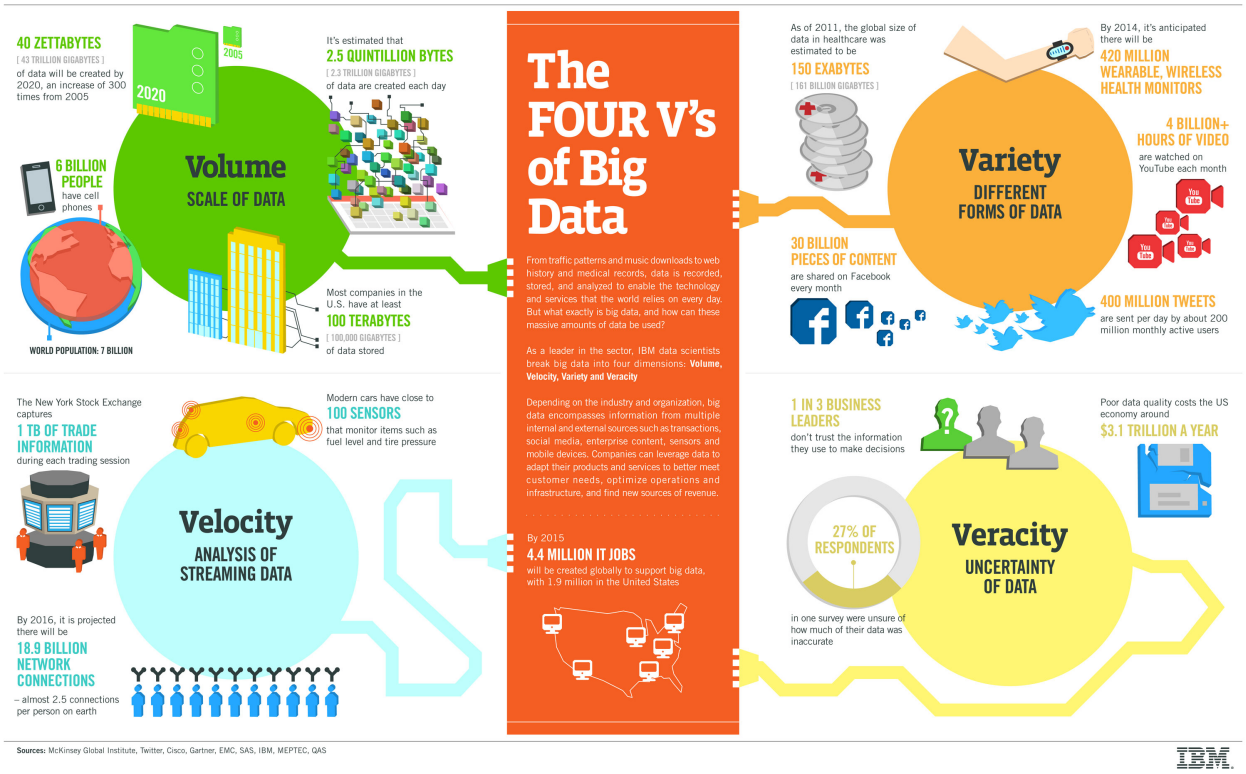


Figure 2: Big Data 4Vs.

Processing phone numbers from service providers may also present challenges. Phone numbers could be in a character delimited file (something like comma or tab separated values), or an XML file[1] (see Table 2). The data provider has complete control over the format of the data, and data from different providers may look entirely different.

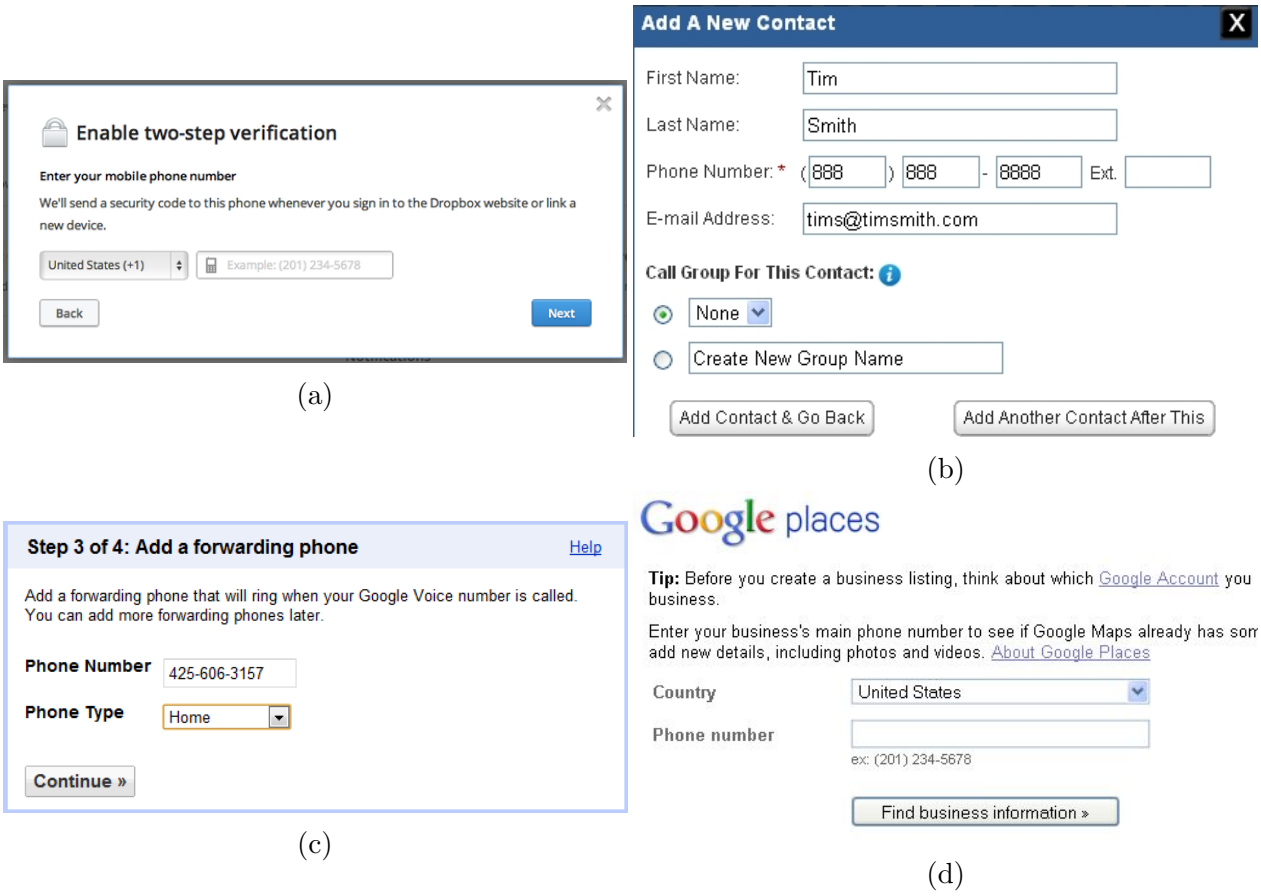


Figure 3: Representative data phone number entry fields. NANP formatted phone numbers do not actually have parens or dashes. Those characters are added for human readability, not telecommunication needs.

[1] `https://telemarketing.donotcall.gov/faq/faqbusiness.aspx`

Table 1: North American Numbering Plan (NANP) format.

Component	**Name**	**Number restrictions**
NPA	Numbering Plan Area Code	2–9 for the first digit, 0–9 for the second and third
NXX	Central Office (exchange) code	2–9 for the first digit, 0–9 for the second and third (additional restrictions if it is a geographic area code)
xxxx	Subscriber number	0–9 for all digits

Table 2: Sample phone numbers in an XML format. Example taken from the National Do Not Call Registry. The example represents phone numbers for the state of NY, area code 212, and select subscriber numbers.

```
<list type='full' level='state' val='NY' />
<ac val='212'>
    <ph val='4567890' />
    <ph val='4567890' />
    <ph val='4567890' />
</ac>
<ac val='xxx'>
    <ph val='xxxxxxx' />
    <ph val='xxxxxxx' />
    <ph val='xxxxxxx' />
    </ac>
    </list>
```

Table 3: Ways to interpret numerical dates. A date written as 10/31/94 is logical. A date written as 10/3/94 is not logical.

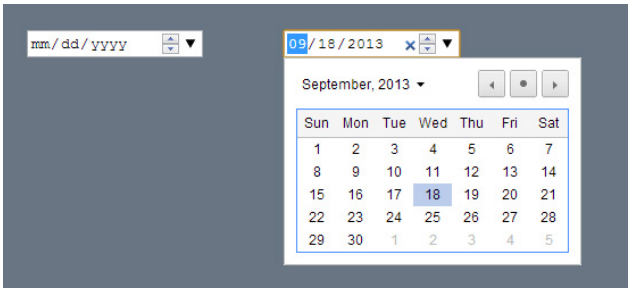
Magnitude	**Probable meaning**
> 31	A year. With only a 2 digit year, the century will be in question.
> 12	A month.
Otherwise	A day.

2.2 Dates

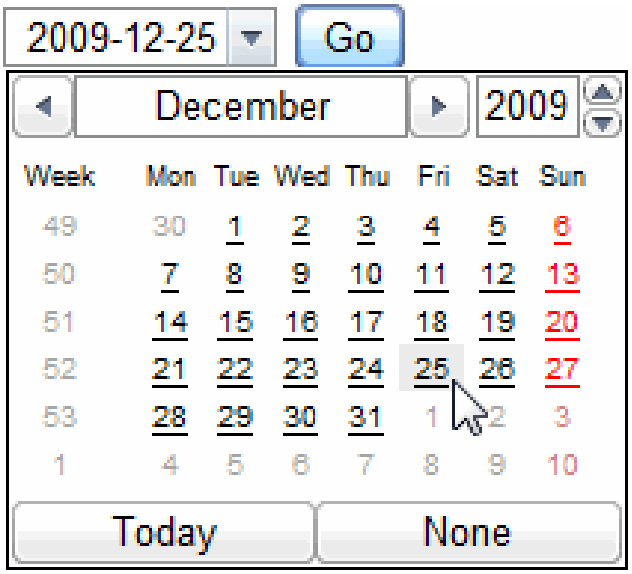
Numerical dates can be confusing. It seems like such a simple thing. A number for month, another for day, and a third for the year. How hard can dates be (see Figure 4)? It turns out that date entry is anything but standard. If you take a simple approach to handling the numbers, then you can make assumptions on what the numbers mean based on the size/magnitude of each (see Table 3). There are a number of standards promoted by International Organization for Standardization (ISO), American National Standards Institute (ANSI) InterNational Committee for Information Technology Standards (INCITS), Environmental Data Standards Council (EDSC), and the World Wide Web Consortium (W3C) that apply to the numeric representation of dates, whether as entered by humans or exchanged by computer systems. These standards include:

- ANSI INCITS 30-1997 Data elements and interchange formats – Information interchange – Representation of dates and times
- ANSI INCITS 310-1998 Data elements and interchange formats – Information interchange – Representation of dates and times
- EDSC Standard No.: EX000013.1, REPRESENTATION OF DATE AND TIME DATA STANDARD
- ISO 8601:2000 Data elements and interchange formats – Information interchange – Representation of dates and times
- W3C Recommendation XML Schema Part 2: Datatypes 02 May 2001

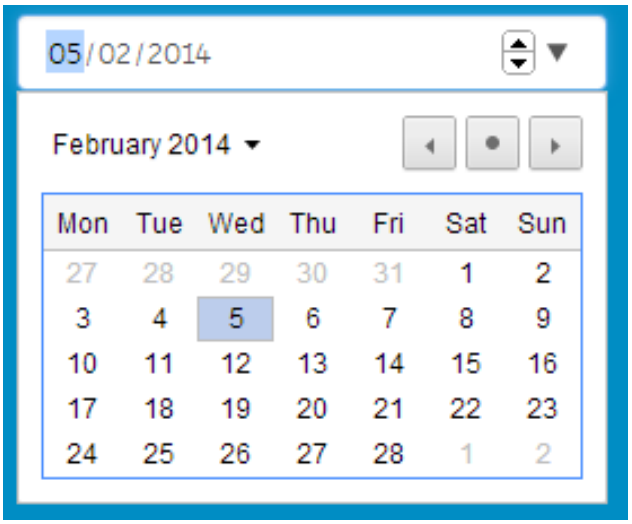
When in doubt, the safest thing to do is to write out the month and use a four digit year. Following that advice, 2 March 1952 would not be written as 3/2/1952 which is either 3 February 1952, or March 2, 1952. About one third of all numerically written dates can be ambiguous.



(a)



(b)



(c)

(d)

Figure 4: Representative date entry fields.

Table 4: The three parts of a credit card number.

Digits	Usage
1 – 6	Issuer identification number (IIN). The first digit is the major industry identifier (MII).
7 – ??	The individual account identification (variable length 6 – 12 digits).
Last	The last digit, a Luhn checksum.

2.3 Credit card numbers

Credit card numbers have a lot of information that is readily available, if you know how to read it. The numbers are formatted in a very specific way, that makes many of the "normal" ways of entering the associated information superfluous.

All credit cards numbers are formatted in accordance with International Organization for Standardization / International Electrotechnical Commission (ISO/IEC) standard 7812 Identification cards – Identification of issuers – Part 1: Numbering system[2]. The standard divides the number into three fields (see Table 4). The IIN contains information about the issuing organization (see Table 5). Some issuing networks and their IIN ranges are provided (see Table 6) There are web sites that offer to validate IINs on line, and provide information beyond that which is strictly part of the IIN[3].

Based on the numbers that a user enters, the type of issuing institution can be derived, so it shouldn't be necessary for the user to pick an institution from a list. If the organization does not have an arrangement in place to accept payment from the card's issuing company, then a message could be presented to the user that that type of card was accepted. The credit card numbers can be validated by recomputing the checksum. This validation only ensures that the numbers are correct (to a limited degree), and not that the numbers will be accepted by the institution. Most credit cards and many government identification numbers use the algorithm as a simple method of distinguishing valid numbers from collections of random digits.[4][5]

[2] `http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=39698`

[3] `https://www.binlist.net/`

[4] `http://www.worldlibrary.org/articles/Luhn_algorithm`

[5] A very readable explanation of the Luhn algorithm can be found here:`https://en.wikipedia.org/wiki/Luhn_algorithm`

Table 5: Details of a credit card IIN.

MII	Issuer category
0	ISO Technical Committee (TC) 068[6] "Financial services"
1	Airlines
2	Airlines, financial and other future industry assignments
3	Travel and entertainment
4	Banking and financial
5	Banking and financial
6	Merchandising and banking/financial
7	Petroleum and other future industry assignments
8	Healthcare, telecommunications and other future industry assignments
9	For assignment by national standards bodies

Table 6: Credit card issuing networks and IIN ranges. Data extracted from `https://en.wikipedia.org/wiki/Bank_card_number`

Issuing network	IIN ranges
American Express	34, 37
Diners Club Carte Blanche	300 - 305
Diners Club enRoute	2014, 2149
Diners Club International	300 - 305, 309, 36, 38 - 39
Diners Club	54, 55
Discover Card	6011, 622126 - 622925, 644 - 649, 65
MasterCard	2221 - 2720, 51 - 55
Visa	4

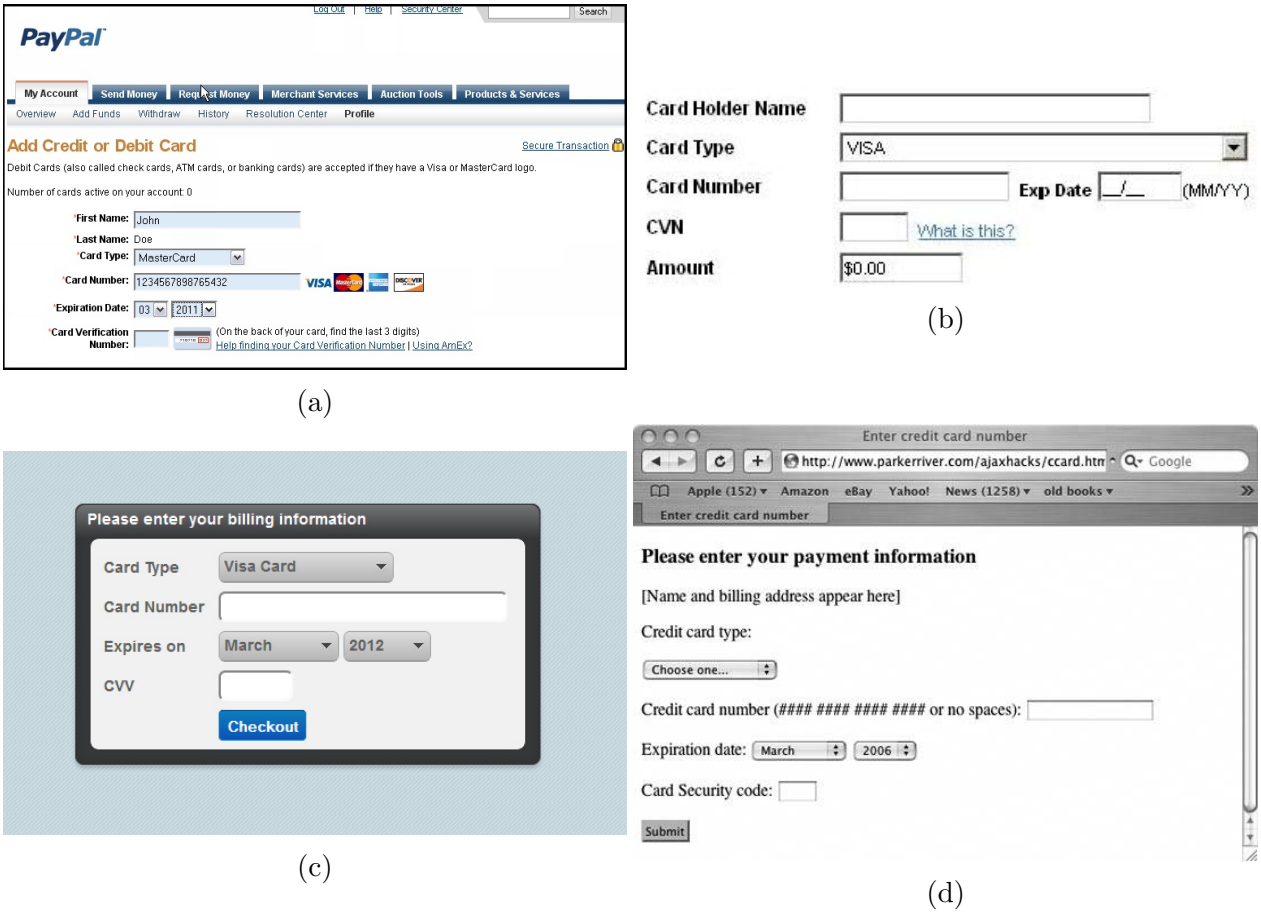


Figure 5: Representative credit card entry fields.

2.4 Thinking about Social Security Numbers (SSNs)

If you are old (as in you got your Social Security Card/number prior to 25 June 2011[7]), then there was intelligence in your SSN. If you are young, then there isn't. Like other numeric fields, there are lots of different formats for entering your SSN (see Figure 6).

In the olden days (pre 25 June 2011), the SSN was broken into three parts with meaning (see Table 7). Since 1972, when SSA began assigning SSNs and issuing cards centrally from Baltimore, the area number assigned has been based on the ZIP code in the mailing address provided on the application for the original Social Security card. The applicant's mailing address does not have to be the same as their place of residence. Thus, the Area Number does not necessarily represent the State of residence of the applicant, either prior to 1972 or since.

Starting 25 June 2011, the SSA changed from a geographically structured way of allocated SSNs to a randomized process.[8] The geographically structured way limited which numbers could be assigned within which area, but there were significant differences in the number of people in each number. As a result, some areas had an over abundance, while other areas were severely constrained. The randomization process affected the number allocation in the following ways:

- It eliminated the geographical significance of the first three digits of the SSN, referred to as the area number, by no longer allocating the area numbers for assignment to individuals in specific states.
- It eliminated the significance of the highest group number and, as a result, the High Group List is frozen in time and can only be used to see the area and group numbers SSA issued prior to the randomization implementation date.
- Previously unassigned area numbers were introduced for assignment excluding area numbers 000, 666 and 900-999.

The SSN model has some limitations:

1. It is a 9 digit number,
2. There are administrative limitations on which specific values can be issued,
3. Numbers can not be reused.

Eventually we will run out of available SSNs. When we will run out is unclear. By some estimates, we the current SSN scheme has approximately 750 million possible values[9]. If we

[7] `https://www.socialsecurity.gov/employer/randomizationfaqs.html`
[8] `https://www.socialsecurity.gov/employer/randomization.html`
[9] `http://www.mathiseasy.net/WillTheUSeverRunOutOfSocialSecurityNumbers.php`

have used approximately 600 million SSNs already[10], then there should be enough unused SSNs to last between 50 and 80 years.

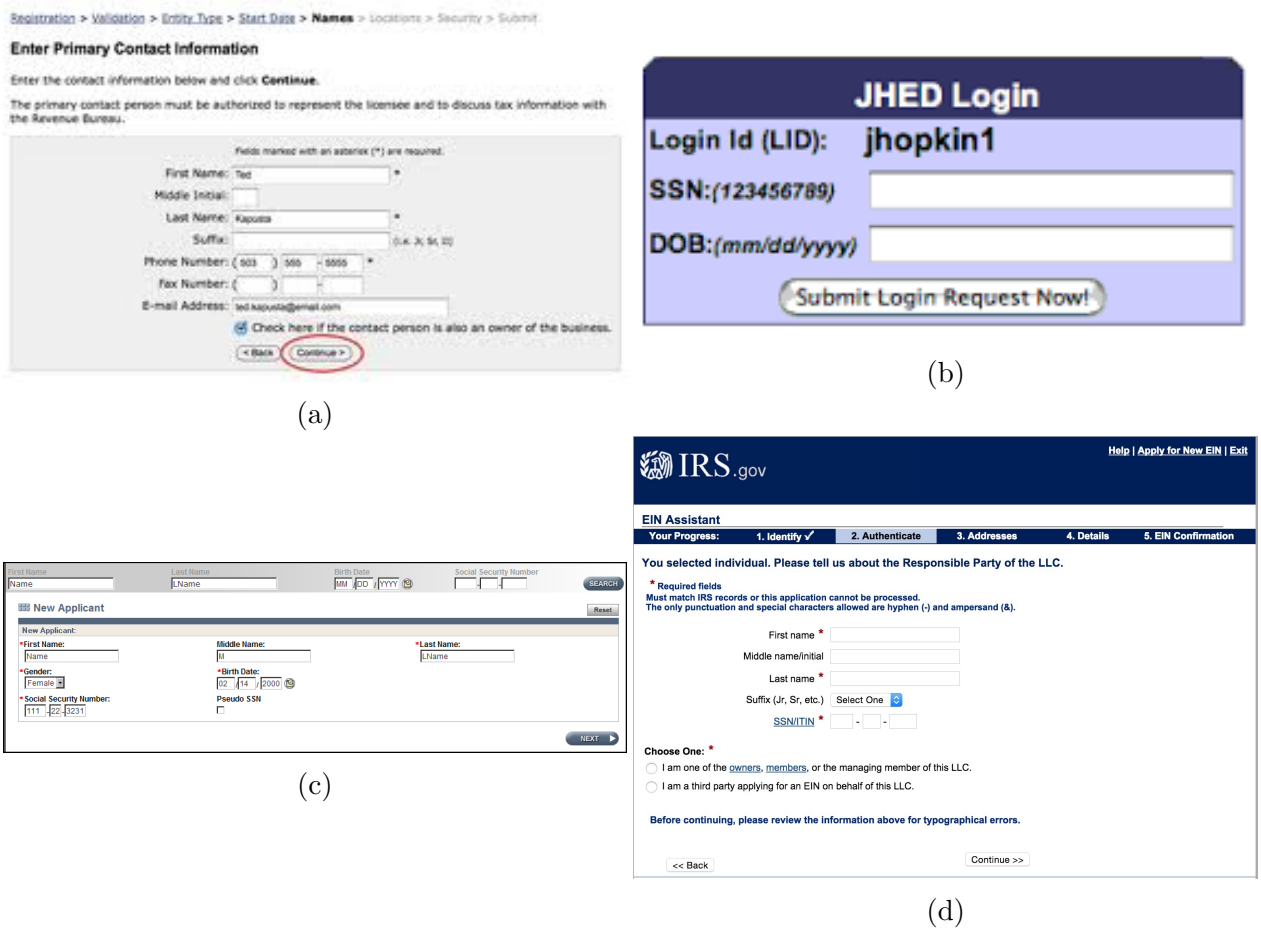


Figure 6: Representative SSN entry fields. Different applications appear to be interested in the three different parts of the SSN (area number, group number, and serial number). Or maybe it reflects the way that we think/sing our numbers.

[10]Current population of the US as approximately 325 million, and assuming that 100 million people have had SSNs and died between 1936 and 2015, and that the average growth rate remains constant

Table 7: Format and breakdown of pre 2012 SSNs. Much of this information is taken from the SSA historical web site: `https://www.ssa.gov/history/ssn/geocard.html`

Part	Length	Name	Meaning
1	3	Area Number	The Area Number is assigned by the geographical region. Prior to 1972, cards were issued in local Social Security offices around the country and the Area Number represented the State in which the card was issued. This did not necessarily have to be the State where the applicant lived, since a person could apply for their card in any Social Security office. Generally the numbers are assigned from the Northeast and then continue South and West. A complete list of SSN Area Numbers can be found at: `https://www.socialsecurity.gov/employer/stateweb.htm`
2	2	Group Number	Group numbers range from 01 to 99 but are not assigned in consecutive order. For administrative reasons, group numbers issued first consist of the ODD numbers from 01 through 09 and then EVEN numbers from 10 through 98, within each area number allocated to a State. After all numbers in group 98 of a particular area have been issued, the EVEN Groups 02 through 08 are used, followed by ODD Groups 11 through 99.
3	4	Serial number	Numbers 0001 to 9999 are valid. 0000 is not valid.

2.5 Thinking about addresses

The Centers for Medicare and Medicaid Services provides a wealth of data about the Medicare system[11], including individual records of Medicare payments to physicians. These data are available for download wit the click of a button. The files are large, and the number of rows of data in each file can be large (in excess of 9,000,000 records for CY 2013).

Each record contains the name and address of the physician or clinic that received a Medicare payment. The CY 2013 payment file was used as a source of addresses for processing. Initially 20,000,000 street addresses were requested from a Hive database and 9,287,878 were returned. There were many duplicate addresses. Ultimately there were 321,771 unique addresses. Each address was "tokenized" (i.e., each address was broken into pieces where ever there was a space), resulting in 1,221,229 tokens, or pieces. Of these, 349,395 were strictly numbers, with the remaining (871,834) contained text. These non-numeric tokens are what we are after.

We analyzed the tokens in a couple of different ways. The first was a simple listing of the tokens by how often they appear. This is shown in the following table. There are a number of things that are interesting in the table. They include:

1. The number of purely digital tokens (number 1) far out paces any other token.
2. There are number of ways to abbreviate "street" as shown in 2, 13, 41, and 174.
3. There cardinal directions (North, South, East, and West) can be spelled out or abbreviated, as shown in 7, 30, and 56.
4. Things like road, avenue, boulevard, parkway, and others are also treated creatively.

[11] https://www.cms.gov/research-statistics-data-and-systems/statistics-trends-and-reports/medicare-provider-charge-data/physician-and-other-supplier.html

1	NA	349,395	41	ST.	2,764	81	GRAND	1,025
2	ST	68,359	42	WASHINGTON	2,385	82	D	982
3	AVE	49,440	43	PL	2,327	83	MEMORIAL	974
4	RD	49,235	44	BROADWAY	2,281	84	JEFFERSON	968
5	DR	27,921	45	SUITE	2,247	85	COURT	931
6	W	23,902	46	US	2,230	86	LINCOLN	922
7	N	23,636	47	LAKE	2,096	87	UNION	915
8	STE	21,942	48	CIR	2,042	88	SAN	908
9	E	21,678	49	HOSPITAL	1,978	89	COLLEGE	880
10	S	21,257	50	OLD	1,908	90	SAINT	864
11	TH	21,047	51	HILL	1,800	91	MILL	859
12	BLVD	19,970	52	-	1,769	92	MILE	847
13	STREET	12,466	53	AVE.	1,726	93	CHURCH	825
14	ROAD	8,401	54	PIKE	1,687	94	MAPLE	817
15	MAIN	8,356	55	UNIVERSITY	1,684	95	FOREST	811
16	HWY	6,727	56	N.	1,668	96	PLZ	766
17	AVENUE	6,536	57	W.	1,628	97	JACKSON	754
18	HIGHWAY	6,321	58	CENTRAL	1,601	98	MADISON	752
19	DRIVE	5,256	59	S.	1,585	99	VILLAGE	746
20	PKWY	4,995	60	C	1,578	100	COUNTY	746
21	LN	4,949	61	RIDGE	1,548	101	PINE	730
22	PARK	4,923	62	CALLE	1,494	102	LOOP	728
23	NE	4,248	63	E.	1,421	103	HIGHLAND	725
24	NW	4,091	64	VALLEY	1,371	104	BEACH	723
25	A	4,072	65	RIVER	1,339	105	HIGH	711
26	#	4,072	66	OAK	1,329	106	DE	709
27	WEST	3,995	67	CREEK	1,319	107	CEDAR	704
28	SW	3,841	68	RD.	1,275	108	BAY	697
29	WAY	3,763	69	PARKWAY	1,214	109	FRANKLIN	696
30	NORTH	3,693	70	PLAZA	1,199	110	GROVE	691
31	SOUTH	3,633	71	LANE	1,186	111	WALNUT	685
32	EAST	3,630	72	OF	1,179	112	SPRING	685
33	CENTER	3,602	73	DR.	1,168	113	PLACE	675
34	CT	3,389	74	APT	1,136	114	CTR	667
35	STATE	3,181	75	BROAD	1,096	115	FL	657
36	SE	3,097	76	BOULEVARD	1,063	116	CIRCLE	646
37	MEDICAL	2,999	77	NEW	1,059	117	KING	644
38	B	2,971	78	MARKET	1,052	118	ELM	620
39	ND	2,804	79	TRL	1,042	119	BLDG	614
40	ROUTE	2,786	80	BLVD.	1,041	120	UNIT	591

121	PROFESSIONAL	591	161	HARRISON	432	201	COMMERCIAL	348
122	LA	589	162	RIVERSIDE	430	202	TAMIAMI	347
123	GREEN	582	163	POST	426	203	MONROE	347
124	POINT	581	164	.	425	204	CARR	344
125	BRIDGE	572	165	FIRST	415	205	MONTGOMERY	343
126	HEALTH	570	166	DIVISION	412	206	RM	340
127	HILLS	565	167	ROOSEVELT	410	207	ISLAND	338
128	SANTA	553	168	BOX	410	208	LOCUST	337
129	WHITE	552	169	FRONT	408	209	ORANGE	335
130	YORK	551	170	TOWN	406	210	COMMONS	335
131	SUNSET	551	171	JOHNSON	406	211	CLINIC	331
132	MOUNTAIN	540	172	DIXIE	403	212	VETERANS	330
133	TPKE	535	173	LEE	402	213	ROCK	319
134	JOHN	531	174	STREET,	399	214	INDUSTRIAL	317
135	F	527	175	FM	397	215	MERIDIAN	315
136	COUNTRY	517	176	SCHOOL	395	216	DAVIS	315
137	VIEW	516	177	COLUMBIA	394	217	LUTHER	314
138	SPRINGS	511	178	WILLOW	393	218	WATER	311
139	CHESTNUT	510	179	FERRY	392	219	NATIONAL	311
140	MALL	495	180	CHERRY	389	220	LONG	311
141	G	495	181	CANYON	389	221	SHORE	310
142	TER	490	182	PROSPECT	388	222	RUN	310
143	OCEAN	489	183	J	388	223	KINGS	309
144	MICHIGAN	488	184	PACIFIC	378	224	INDIAN	309
145	VISTA	486	185	LIBERTY	372	225	BELL	308
146	PLEASANT	486	186	ATLANTIC	370	226	CLARK	306
147	CENTRE	482	187	LINE	366	227	ONE	305
148	AND	468	188	FWY	365	228	&	305
149	EL	459	189	FEDERAL	365	229	MILITARY	304
150	MARTIN	457	190	M	363	230	SHERIDAN	303
151	COMMERCE	455	191	EUCLID	363	231	JAMES	303
152	WESTERN	448	192	CLEVELAND	362	232	L	302
153	SQ	447	193	AIRPORT	362	233	GRANT	302
154	CAMINO	446	194	TRAIL	360	234	DEPT	302
155	VAN	441	195	H	359	235	LAUREL	299
156	SQUARE	441	196	CITY	357	236	RED	297
157	OAKS	441	197	KM	353	237	PENNSYLVANIA	297
158	FORT	438	198	CLUB	353	238	RICHMOND	295
159	MOUNT	433	199	RT	350	239	VIRGINIA	291
160	DEL	433	200	EXECUTIVE	350	240	SUMMIT	287

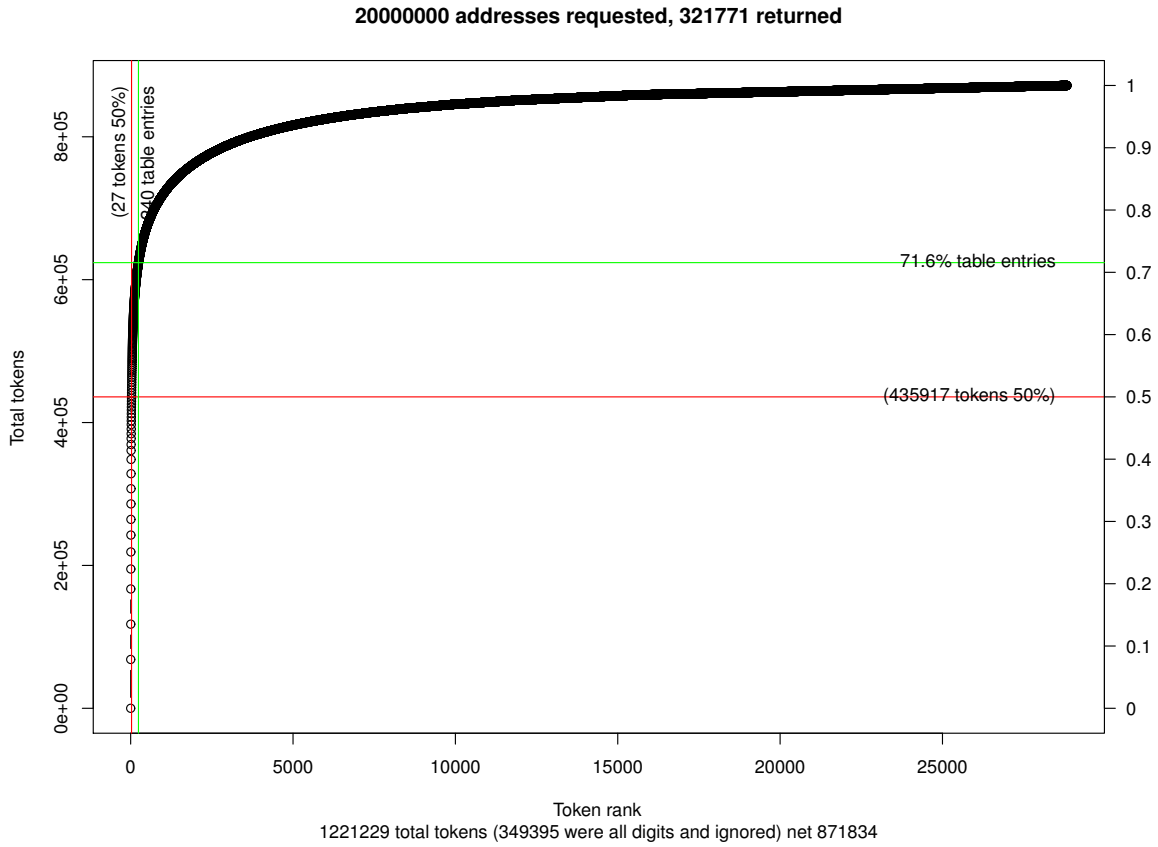


Figure 7: Cumulative distribution of address tokens. The red lines are where 50% of the tokens have been identified. The red lines are where all tokens in the table fall on the distribution curve.

The second was to show the cumulative distribution of the tokens (see Figure 7). A cumulative distribution curve shows where the greatest change is, and therefore where the greatest "bang for the buck" lies. For the address tokens, correctly processing the first 27 tokens will account for 50% of all cases. Processing the next 213 tokens will only buy an additional 22%. While processing any after the first 5,000 will only result in marginal improvements.

The takeaway from this simple analysis is that "normalizing" an address has to take into account a lot of variations in spelling, punctuation, order, and spacing. Normalizing is a non-trivial task to do correctly.

Various companies offer address normalization services, including:

1. ServiceObjects `https://www.serviceobjects.com/products/address-geocoding/usps-address-validation`

2. USgeocoder `https://www.usgeocoder.com/`

3. Cdyne `http://cdyne.com/api/address-verification/`

The thing to watch for when choosing a geocoding service or API is to ensure they are Coding Accuracy Support System (CASS) certified[12]. CASS software will correct and standardize addresses. It will also add missing address information, such as ZIP codes, cities, and states to ensure the address is complete[13]. To be certified as CASS compliant, the software must score at least 98% on a set of standardized tests[14].

Part of the CASS development process (known as Stage 1) is a self certification using a "well known" data file. The Stage I file contains approximately 150,000 test addresses extracted from the City State and ZIP + 4 files with samples of all types of addressing used around the country. Some test addresses have been changed for test purposes, and not all records have valid ZIP + 4 codes or valid addresses. Developers can evaluate the accuracy of their address-matching software by applying the:

- correct carrier route: a group of addresses that receive the same USPS code to aid in efficient mail delivery. Postal carrier route codes are 9 digits: 5 numbers for the ZIP Code, one letter for the carrier route type, and 3 numbers for the carrier route code;for example, "92019C005" or "84604R009." A carrier route is generally associated with where a particular mail carrier delivers.

- five-digit, ZIP + 4 codes

- DPV: Delivery Point Verification (DPV) is a USPS validation process that confirms the existence of a specific address (down to the apartment or suite number) and whether or not it can be delivered to.

- DSF2: Delivery Sequence File (DSF2) identifies whether a ZIP + 4 coded address is currently represented in the USPS delivery file as a known address record.

- LACS Link: Locatable Address Conversion System (LACS) is a system that matches US addresses against a list of rural route, highway route, and box number addresses that have been renumbered or renamed due to 911 address conversions. These conversions involve changing rural-style addresses to city-style addresses for 911 emergency system implementations.

- Suite Link: is designed add suite information to business addresses in large buildings.

[12] A partial list of CASS certified software can be found at `https://ribbs.usps.gov/files/vendors/cassn01d.TXT`

[13] `https://en.wikipedia.org/wiki/Coding_Accuracy_Support_System`

[14] `http://pe.usps.com/text/dmm/A950.htm`

and by comparing the applied codes with the correct codes provided by the Postal Service.

The US Postal Service (USPS) offers an application program interface (API) to validate addresses[15].

Pathological addresses problems are easy to create (see Table 8).

Table 8: A small collection of pathological address resolution problems. Most of the examples are based on relatively simple token parsing. More difficult problems require additional processing beyond string matching. Techniques such as soundex come into play.

#	**Input**	**Possible resolution**
1	110 Main St.	110 Main Street
2	110 Main St	110 Main Street
3	110 Main Str	110 Main Street
4	110 St Main Blvd.	110 Saint Main Boulevard
5	110 St. Main St.	110 Saint Main Street
6	Mikrsft	Microsoft

3 Conclusion

The internet is a wild and woolly place. When it comes to standardizing how data should be represented, people are free to do almost anything they want because most of the times, humans are in the loop and can sort things out. It becomes much more challenging when a piece of software has to try and sort things out.

We looked at a few relatively simple numeric data types, that are well understood by humans. The types are:

- Phone numbers that are formatted in multiple and inconsistent ways,
- Dates when entered numerically can have very different formats and orders,
- Credit card numbers entry fields that ask for redundant information, and
- Social Security Numbers that had intelligence at one time, but have now lost it.

Using the multitude of different ways simple numbers can be confusing, we extrapolated the same types of problems and confusing formats to US addresses. There are companies that will "normalize" a USPS address in order to improve delivery times. And the USPS provides an API to provide low volume support.

[15] `https://www.usps.com/business/web-tools-apis/welcome.htm`

Simple numbers can be difficult to manage, and simple mail addresses can be very difficult to handle correctly.

4 References

[1] Doug Laney, *3D Data Management: Controlling Data Volume, Velocity and Variety*, META Group Research Note **6** (2001).

[2] Eileen McNulty, *Understanding big data: The seven vs*, `http://dataconomy.com/seven-vs-big-data/`, 2014.

[3] IBM Staff, *The Four V's of Big Data*, `http://www.ibmbigdatahub.com/infographic/four-vs-big-data`, 2016.