Text analytics
and
new visualization techniques

Richard Brath  rbrath@unchartedsoftware.com
               brathr@lsbu.ac.uk
               @rkbrath
               richardbrath.wordpress.com

Scott Langevin slangevin@unchartedsoftware.com
                   blog.unchartedsoftware.com
                   @slangevi
Tag cloud via wordle.com. Note color is random, location is random and size is ambiguous – does Alice occur 10x or 100x more than Queen?
1. THE 500 YEAR CONSPIRACY against text visualization
Genealogical tree from late 1400’s. Note graph nodes use of image (people, shield) or text, where text may be black, red or start with a red initial. The nodes can vary in size, color, or shape (circle, crescent, shield). Textual commentary is intertwined throughout.

1439
Johannes Gutenberg changes everything

1573: Image separates from text. From William Bullein's A Dialogue... Against the Fever Pestilence. Author photo from Bodleian exhibition "Shakespeare's Dead".
Diderot's Encyclopedia has great illustrations of various occupations—all neatly labeled, but the viewer has to cross-reference the text to understand.

The Encyclopédie ou Dictionnaire raisonné des sciences, des arts et des métiers, par une Société de Gens de lettres was published under the direction of Diderot and d'Alembert, with 17 volumes of text and 11 volumes of plates between 1751 and 1772. Containing 74,000 articles written by more than 130 contributors, the Encyclopédie was a massive reference work for the arts and sciences, as well as a propaganda piece. Its authors, Diderot and d'Alembert, were involved in the French Enlightenment, which sought to promote the ideas of reason and progress. Through its attempt to classify learning and to open all domains of human activity to its readers, the Encyclopédie gave expression to many of the most important intellectual and social developments of its time.

more than would be required for the tabulated figures, and if there is any serious limitation on space, Fig. 97 could be much reduced in size without detracting from its clearness.

Health-department reports are not usually of interest to the layman. Yet health-department reports, well presented, may be of as much popular interest as a well-written magazine article. Fig. 98 is taken from a commendable report by the health department of the city of Boston. In the report itself, different colors of ink were used for the various curves, with the effect of emphasizing the contrast.

Though the colored inks assisted in catching the eye of the reader, the chart with curves designated by letters is usually sufficient for all practical purposes whenever the curves do not cross over each other in such manner as to be confusing. As mentioned elsewhere, a certain slope of a curve plotted on rectangular co-ordinate paper does not in itself indicate a greater or less amount of increase or decrease than holds true for some other curve having a different slope. The slope of a curve plotted on paper with ordinary co-ordinate ruling depends largely on whether the data of the curve are in large figures, so as to bring the curve near the top of the chart, or in small figures, bringing the curve near the bottom of the chart. In Fig. 98 the fact that curve A slopes more than curve C is due to the fact that curve A is placed higher on the vertical scale of the chart than curve C. A little study will show that the reduction in mortality portrayed by curve A is much less on a percentage basis than that depicted by curve C, yet curve A has the steeper slope. The slope of these two curves can best be compared if a pencil line is drawn in such manner that the peaks above the pencil line are approximately equal to the valleys below the pencil line for each curve.

Fig. 99 contains some interesting information. Though the chart proves fairly well the close dependence of the price of cast-iron pipe upon the price of the pig iron from which it is made, the chart is nevertheless misleading in that the first glance would indicate a much greater fluctuation in the price of pig iron than actually occurred.

The reader is apt to overlook the fact that the vertical scale of the chart does not extend below 811 per ton. He is quite likely to think that the price of pig iron had all the rapid fluctuations which would be indicated by the changing vertical distances between the pig-iron curve and the bottom line of the chart itself. The amount of fluctuation would look much less if the chart extended to the zero line of the vertical scale.
Modern visualizations in news. Note how there is a lot of text – but moved out and round the charts.


OUTSIDE THE LABORATORY...

DER BOSS SAID THE LABORATORY IS DEVELOPING AN IMPORTANT WAR SECRET. HEINRICH IT'S UP TO US TO FIND OUT MORE ABOUT IT!

JAI! WE WILL WAIT! HERE COME TWO OF THE LABORATORY WORKERS!

SURE...THE NEW RADIO ACTIVATOR TURNS ANY GAS INTO A SOLID INSTANTANEOUSLY!

WE DO WHATEVER WE WANT! NAZI GERMANY WILL NOT BE STOPPED!

WHAT THE...

THAT EVENING...

UP WITH YOUR HANDS QUICK! HANG, GRAB THE RADIO-ACTIVATOR!

YOU CAN'T DO THIS!

HERE'S WHERE WE MAKE A TRY AT IT ANYWAY!

WHAT THE!!

IN THE CENTRAL TECH LABORATORY...

IT IS A GREAT ACHIEVEMENT, DICK! WE CAN NOW ACTUALLY SWING INTO BATTLE AGAINST THE FORCES OF EVIL!

IT WILL SPEED UP WAR CONSTRUCTION TREMENDOUSLY, BY ELIMINATING DANGEROUS EXPLOSIONS!

AT THAT MOMENT...

VERDAMMT! WE WILL HAVE TO GET AWAY FAST! COME!

DICK WHAT?

ды!

BOSS! GERMAN POLICE! THEY MUST HAVE HEARD OF OUR FIGHTING!

POW!
2. Bringing **TEXT** back into **VISUALIZATION**
Plain Dots

- Simple scatterplot
- Macro pattern e.g. crescent shape
- Also color and size set to population

Image: Birth rate vs. Death rate per country, with size and color indicating population; and diagonal lines indicating birth to death ratio. Note visible crescent shape. Image by author.
Alphanumeric Points

- Same scatterplot
- 2-letter ISO codes added
- More information, more insight

Image: 2 letter country ISO codes act as mnemonics. Note codes upper left: LV, UA, RU, EE – Latvia, Ukraine, Russia, Estonia, etc., all below population replacement rate. Image by author.
Bar Chart of Text
Better than label and bar: See artist, title and first line.

Simple label has problems: Artist as label doesn’t show the song. Song as label is ambiguous (e.g. who’s version of Silent Night?). Label using artist + song is very long and leaves little space for bar (e.g. see entry for Elton John).

First line provides rich context: maybe you don’t know the Wabash Cannonball, but with the first line you know what it’s about. Images by author.

**Best-Selling Singles** more than 10 million copies, sorted by release date

<table>
<thead>
<tr>
<th>Artist</th>
<th>Title</th>
<th>Sales (m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bing Crosby</td>
<td>&quot;Silent Night&quot;</td>
<td>95.1</td>
</tr>
<tr>
<td>Roy Acuff</td>
<td>&quot;Silent Night&quot;</td>
<td>83.9</td>
</tr>
<tr>
<td>Elvis Presley</td>
<td>&quot;Hound Dog&quot;</td>
<td>82.5</td>
</tr>
<tr>
<td>The Monkees</td>
<td>&quot;I'm a Believer&quot;</td>
<td>81.3</td>
</tr>
<tr>
<td>Joni Mitchell</td>
<td>&quot;Woodstock&quot;</td>
<td>80.2</td>
</tr>
<tr>
<td>Garth Brooks</td>
<td>&quot;The Dance&quot;</td>
<td>79.6</td>
</tr>
</tbody>
</table>

Bar Chart of Text

Plus content analysis:

Is “love” the top word in the top 40?
Bar Chart of Text

No: “Christmas” beats love (but occurs in only 2 songs)

Best-Selling Singles more than 10 million copies, sorted by release date

<table>
<thead>
<tr>
<th>No.</th>
<th>Artist</th>
<th>Title</th>
<th>Sales (in millions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bing Crosby</td>
<td>&quot;Silent Night&quot;</td>
<td>28</td>
</tr>
<tr>
<td>2</td>
<td>Bing Crosby</td>
<td>&quot;White Christmas&quot;</td>
<td>22</td>
</tr>
<tr>
<td>3</td>
<td>Dolly Parton</td>
<td>&quot;I Will Always Love You&quot;</td>
<td>19</td>
</tr>
<tr>
<td>4</td>
<td>Whitney Houston</td>
<td>&quot;I Will Always Love You&quot;</td>
<td>19</td>
</tr>
<tr>
<td>5</td>
<td>Elton John</td>
<td>&quot;Candle in the Wind&quot;</td>
<td>18</td>
</tr>
<tr>
<td>6</td>
<td>Celine Dion</td>
<td>&quot;My Heart Will Go On&quot;</td>
<td>17</td>
</tr>
<tr>
<td>7</td>
<td>George Michael</td>
<td>&quot;I Want to Break Free&quot;</td>
<td>17</td>
</tr>
<tr>
<td>8</td>
<td>Barbra Streisand</td>
<td>&quot;Evergreen&quot;</td>
<td>16</td>
</tr>
<tr>
<td>9</td>
<td>Bing Crosby</td>
<td>&quot;White Christmas&quot;</td>
<td>15</td>
</tr>
<tr>
<td>10</td>
<td>Bing Crosby</td>
<td>&quot;Silent Night&quot;</td>
<td>14</td>
</tr>
<tr>
<td>11</td>
<td>Barbra Streisand</td>
<td>&quot;Evergreen&quot;</td>
<td>14</td>
</tr>
<tr>
<td>12</td>
<td>Carole King</td>
<td>&quot;You Got Me&quot;</td>
<td>13</td>
</tr>
<tr>
<td>13</td>
<td>Carole King</td>
<td>&quot;You Got Me&quot;</td>
<td>13</td>
</tr>
<tr>
<td>14</td>
<td>Bing Crosby</td>
<td>&quot;White Christmas&quot;</td>
<td>12</td>
</tr>
<tr>
<td>15</td>
<td>Bing Crosby</td>
<td>&quot;Silent Night&quot;</td>
<td>12</td>
</tr>
<tr>
<td>16</td>
<td>Bing Crosby</td>
<td>&quot;White Christmas&quot;</td>
<td>11</td>
</tr>
<tr>
<td>17</td>
<td>Bing Crosby</td>
<td>&quot;Silent Night&quot;</td>
<td>11</td>
</tr>
<tr>
<td>18</td>
<td>Bing Crosby</td>
<td>&quot;White Christmas&quot;</td>
<td>10</td>
</tr>
</tbody>
</table>


Bar Chart of Text

And “baby” beats both in counts and songs

Best-Selling Singles more than 10 million copies, sorted by release date

<table>
<thead>
<tr>
<th>Artist/Group</th>
<th>Title</th>
<th>Release Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bing Crosby</td>
<td>&quot;Silent Night&quot;</td>
<td>1943</td>
</tr>
<tr>
<td>Bing Crosby</td>
<td>&quot;White Christmas&quot;</td>
<td>1942</td>
</tr>
<tr>
<td>Roy Acuff</td>
<td>&quot;Nashville Cats&quot;</td>
<td>1950</td>
</tr>
<tr>
<td>Mills Brothers</td>
<td>&quot;Paper Doll&quot;</td>
<td>1946</td>
</tr>
<tr>
<td>Gene Autry</td>
<td>&quot;Red-Nosed Reindeer&quot;</td>
<td>1949</td>
</tr>
<tr>
<td>Patti Page</td>
<td>&quot;Tennessee Waltz&quot;</td>
<td>1953</td>
</tr>
<tr>
<td>Bill Haley &amp; His Comets</td>
<td>&quot;Rock Around the Clock&quot;</td>
<td>1955</td>
</tr>
<tr>
<td>The Penguins</td>
<td>&quot;Earth Angel&quot;</td>
<td>1957</td>
</tr>
<tr>
<td>Elvis Presley</td>
<td>&quot;Hound Dog&quot;</td>
<td>1956</td>
</tr>
<tr>
<td>Tony Orlando</td>
<td>&quot;Moonlight Bay&quot;</td>
<td>1973</td>
</tr>
<tr>
<td>Welcome</td>
<td>&quot;Domenico Modugno&quot;</td>
<td>1962</td>
</tr>
<tr>
<td>Elvis Presley</td>
<td>&quot;It's Now or Never&quot;</td>
<td>1960</td>
</tr>
<tr>
<td>Kyu Sakamoto</td>
<td>&quot;Sukiyaki&quot;</td>
<td>1963</td>
</tr>
<tr>
<td>The Beatles</td>
<td>&quot;I Want to Hold Your Hand&quot;</td>
<td>1964</td>
</tr>
<tr>
<td>The Monkees</td>
<td>&quot;I'm a Believer&quot;</td>
<td>1965</td>
</tr>
<tr>
<td>Procol Harum</td>
<td>&quot;A Whiter Shawl&quot;</td>
<td>1975</td>
</tr>
<tr>
<td>Mungo Jerry</td>
<td>&quot;In the Summertime&quot;</td>
<td>1972</td>
</tr>
<tr>
<td>Middle of the Road</td>
<td>&quot;Chirpy Chirpy Cheep Cheep&quot;</td>
<td>1972</td>
</tr>
<tr>
<td>Carl Douglas</td>
<td>&quot;Kung Fu Fighting&quot;</td>
<td>1974</td>
</tr>
<tr>
<td>George McCrae</td>
<td>&quot;Rock Your Baby&quot;</td>
<td>1974</td>
</tr>
<tr>
<td>Roger Whittaker</td>
<td>&quot;The Last Resort&quot;</td>
<td>1970</td>
</tr>
<tr>
<td>ABBA</td>
<td>&quot;Fernando&quot;</td>
<td>1979</td>
</tr>
<tr>
<td>John Travolta &amp; Olivia Newton-John</td>
<td>&quot;You're the One That I Want&quot;</td>
<td>1978</td>
</tr>
<tr>
<td>Gloria Gaynor</td>
<td>&quot;I Will Survive&quot;</td>
<td>1979</td>
</tr>
<tr>
<td>The Three Degrees</td>
<td>&quot;Da Da Da&quot;</td>
<td>1974</td>
</tr>
<tr>
<td>Brian Adams</td>
<td>&quot;Do They Know It's Christmas?&quot;</td>
<td>1984</td>
</tr>
<tr>
<td>Whitney Houston</td>
<td>&quot;I Will Always Love You&quot;</td>
<td>1992</td>
</tr>
<tr>
<td>Andrea Bocelli &amp; Sarah Brightman</td>
<td>&quot;Time to Say Goodbye&quot;</td>
<td>1994</td>
</tr>
<tr>
<td>Toni Braxton</td>
<td>&quot;Un-Break My Heart&quot;</td>
<td>1993</td>
</tr>
<tr>
<td>Elton John</td>
<td>&quot;Candle in the Wind&quot;</td>
<td>1997</td>
</tr>
<tr>
<td>Celine Dion</td>
<td>&quot;My Heart Will Go On&quot;</td>
<td>1997</td>
</tr>
<tr>
<td>Cher</td>
<td>&quot;Believe&quot;</td>
<td>1998</td>
</tr>
<tr>
<td>Britney Spears</td>
<td>&quot;Baby One More Time&quot;</td>
<td>1999</td>
</tr>
<tr>
<td>Celine Dion &amp; Andrea Bocelli</td>
<td>&quot;The Prayer&quot;</td>
<td>1998</td>
</tr>
</tbody>
</table>

3. But what about BIG DATA TEXT ANALYTICS?
Text Analytics 101

Typical “Natural Language Processing” Pipeline

Raw Text → Preprocessing → Numerical Representation → Analysis

- Raw Text: John hit the ball.
- Preprocessing: S → NP → VP
- Numerical Representation:
  - Term(s) 1: 10 0 1 0 0 0 2
  - Term(s) 2: 0 2 0 0 18 0 2
  - Term(s) 3: 0 0 0 0 0 0 2
  - Term(s) 4: 0 0 0 6 0 0 0
  - Term(s) 5: 0 0 0 0 0 0 2
  - Term(s) 6: 0 0 0 0 1 0 0
  - Term(s) 7: 0 0 1 0 0
  - Term(s) 8: 0 0 0 0 0 3 0

References:
- https://en.wikipedia.org/wiki/Parse_tree
- http://brandonrose.org/clustering
Preprocessing Raw Text

- **Tokenize**
  - words
  - sentences
  - “Hello, I’m Dr. Jones.”
  - ['Hello', 'I', 'm', 'Dr', 'Jones']

- **Normalize**
  - stem
  - lemmatize
  - lowercase
  - ponies
  - poni
  - ponies
  - pony
  - Ponies
  - ponies

- **Remove Noise**
  - stop words
  - punctuation
  - John’s car is red, right?
  - John’s car is red, right?
Document Representation

Counting
- word count
- sentence count
- syllable count

TFIDF
\[
\frac{1}{\# \text{ docs term } t \text{ appears in}} \times \# \text{ times term } t \text{ is in doc } d
\]

Word Embedding Text Vectors

<table>
<thead>
<tr>
<th></th>
<th>king</th>
<th>queen</th>
<th>man</th>
<th>woman</th>
</tr>
</thead>
<tbody>
<tr>
<td>count</td>
<td>32</td>
<td>147</td>
<td>14</td>
<td>83</td>
</tr>
<tr>
<td>TFIDF</td>
<td>0.001</td>
<td>0.097</td>
<td>0.051</td>
<td>0.147</td>
</tr>
</tbody>
</table>

Document Representation

Part-of-Speech tagging

['And', 'now', 'for', 'something', 'completely', 'different']

'CC' 'RB' 'IN' 'NN' 'RB' 'JJ'

Chunking / Phrase extraction

![Phrase Chunking Example](https://en.wikibooks.org/wiki/LaTeX/Linguistics#media/File:Latex-dependency-parse-example-with-tikz-dependency.png)

Dependency parse tree

![Dependency Parse Tree](https://en.wikibooks.org/wiki/LaTeX/Linguistics#media/File:Latex-dependency-parse-example-with-tikz-dependency.png)

Text Analysis

Topic Modeling
Using statistical features to discover abstract ‘topics’ that occur in a collection of documents

‘topic’ is a probability distribution of words

Topic 1: algorithm, computer, data, ...
Topic 2: gene, data, bio, ...
Topic 3: math, equation, matrix, ...
Topic 4: chemical, carbon, water, ...


Other NLP Tasks

<table>
<thead>
<tr>
<th>Task</th>
<th>Example/Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coreference resolution</td>
<td>&quot;Scott took a flight. He said it was very crowded&quot;</td>
</tr>
<tr>
<td>Text Classification</td>
<td>Spam / Not Spam</td>
</tr>
<tr>
<td>Text Similarity</td>
<td>(definitely, definately) = 1</td>
</tr>
<tr>
<td>Sentiment Analysis</td>
<td>Strata was great this year! :) = POSITIVE</td>
</tr>
<tr>
<td>Entity extraction</td>
<td>New York is a metropolis → (‘New York’, ‘city’)</td>
</tr>
<tr>
<td>Text Summarization</td>
<td>Mostly harmless</td>
</tr>
<tr>
<td>Machine Translation</td>
<td>Mostly harmless -&gt; ほとんど無害</td>
</tr>
<tr>
<td>Natural Language Generation</td>
<td>&quot;Sounds good, talk to you then&quot;</td>
</tr>
</tbody>
</table>
4. TEXT ANALYTICS & VIZ together!
Character Traits

Who are the characters?
What are their traits?

Tag cloud showing frequency of words in *Grimms’ Fairy Tales*, a collection of 50+ short stories. Tag clouds reveals some characters, e.g. king, princess, fox; but doesn’t provide any insight into those characters. (via wordle.net)
NLP to find Traits

1. Extract the characters
2. Extract the adjectives near the characters
3. Visualize

e.g. This example, one sentence from Grimms’ Fairy Tales.

When you come to the old king, and he asks for the beautiful princess, you must say, “Here she is!”

When you come to the old king, and he asks for the beautiful princess, you must say, “Here she is!”

When you come to the old king, and he asks for the beautiful princess, you must say, “Here she is!”
## Characters from Grimms’ Fairy Tales with Associated Adjectives Weighted by Frequency

<table>
<thead>
<tr>
<th>Character</th>
<th>List of adjectives, weighted by frequency: 2 3 4-5 6-9 10+</th>
</tr>
</thead>
<tbody>
<tr>
<td>bird</td>
<td>beautiful splendid open wooden like hanging</td>
</tr>
<tr>
<td>cat</td>
<td>little one long</td>
</tr>
<tr>
<td>fox</td>
<td>old dead young first fast</td>
</tr>
<tr>
<td>gretel</td>
<td>little poor good</td>
</tr>
<tr>
<td>hansel</td>
<td>little like fat</td>
</tr>
<tr>
<td>bride</td>
<td>false true first right real</td>
</tr>
<tr>
<td>king</td>
<td>old great one young three angry beautiful married like third</td>
</tr>
<tr>
<td>princess</td>
<td>beautiful young last dear enchanted strange true third free</td>
</tr>
<tr>
<td>girl</td>
<td>little poor lazy pretty young dead beautiful silly</td>
</tr>
<tr>
<td>queen</td>
<td>beautiful late little far</td>
</tr>
<tr>
<td>wife</td>
<td>standing married next poor two beautiful new one dear true</td>
</tr>
<tr>
<td>witch</td>
<td>old wicked</td>
</tr>
</tbody>
</table>
# CEO’s Use of Emotion Words During Earnings Calls, weighted by frequency

<table>
<thead>
<tr>
<th>CEO</th>
<th>List of adjectives, weighted by frequency: 2 3-4 5-7 8-11 12+</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carl Bass</td>
<td><strong>pleased</strong> good</td>
</tr>
<tr>
<td>Carl Trowell</td>
<td><strong>demand lower</strong> good debt improve cash older grow tender retirement late termination</td>
</tr>
<tr>
<td>Carlos A. Rodriguez</td>
<td><strong>grow confidence pleased success excited proud progress</strong></td>
</tr>
<tr>
<td>Carol M. Meyrowitz</td>
<td><strong>pleased confident share margin grow powerful organization excellent shopping terrific passionate gain</strong></td>
</tr>
<tr>
<td>Charles Alutto</td>
<td><strong>good kind waste cash excited lines share safe hazardous versus fee pleased resources tough respect</strong></td>
</tr>
<tr>
<td>Charles B. Stanley</td>
<td><strong>completion improvement lower gross progress reject rejection</strong></td>
</tr>
<tr>
<td>Charles Bunch</td>
<td><strong>improvement lower share cash versus margin unfavorable gross good aggressive tax income kind in</strong></td>
</tr>
<tr>
<td>Charlie Scharf</td>
<td><strong>kind good share change lower grow pay pleased cross case versus helpful terrific progress excited</strong></td>
</tr>
<tr>
<td>Chris Connor</td>
<td><strong>share kind margin good pretty demand gross john confidence versus income cash bad interior rem</strong></td>
</tr>
<tr>
<td>Chris Sinclair</td>
<td><strong>good progress pretty hot content grow share change organization improvement gross finally kind e</strong></td>
</tr>
<tr>
<td>Christopher Crane</td>
<td><strong>good create late hearing grow successful pleased share advocacy regulatory clean majority kind debt</strong></td>
</tr>
<tr>
<td>Chuck Bunch</td>
<td><strong>good improvement share unfavorable versus cash lower excellent demand respect difficult debt income</strong></td>
</tr>
<tr>
<td>Chuck Jones</td>
<td><strong>case cash good kind improvement progress finally cutting inflation court improve share risk electric</strong></td>
</tr>
<tr>
<td>Clay Williams</td>
<td><strong>lower demand pretty kind good excluding improve completion versus share grow deflation tough dial</strong></td>
</tr>
<tr>
<td>Clifton Penble</td>
<td><strong>margin lower income gross finally excited share pretty</strong></td>
</tr>
<tr>
<td>Craig A. Meneer</td>
<td>share pleased</td>
</tr>
</tbody>
</table>

Approximate number of unique adjectives

Most CEO’s tend to use the same words – not much insight.
CEO’s Use of Emotion Words During Earnings Calls, weighted by frequency

<table>
<thead>
<tr>
<th>CEO</th>
<th>List of adjectives, weighted by frequency:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Andrew P. Wilson</td>
<td>battlefield star console entertainment</td>
</tr>
<tr>
<td></td>
<td>dragon fun ultimate</td>
</tr>
<tr>
<td>Bruce Broussard</td>
<td>star tax enforce good progress successful</td>
</tr>
<tr>
<td>David Hallal</td>
<td>disease complement devastating</td>
</tr>
<tr>
<td>Fabrizio Freda</td>
<td>progress therapeutic success</td>
</tr>
<tr>
<td>Gracia Martore</td>
<td>majority hospital</td>
</tr>
<tr>
<td>J.P. Billrey</td>
<td>beauty exciting prestige successful</td>
</tr>
<tr>
<td>Jim Tisch</td>
<td>love content respect terrific apologize lagging</td>
</tr>
<tr>
<td>John D. Idol</td>
<td>chocolate shanghai lower unfavorable force</td>
</tr>
<tr>
<td>John Ferriola</td>
<td>diamond cash income special pay doubt lose</td>
</tr>
<tr>
<td>John Lechleiter</td>
<td>luxury demand shopping watch bottom</td>
</tr>
<tr>
<td>Kelly King</td>
<td>pig improvement cash lower challenge</td>
</tr>
<tr>
<td>Larry D. Young</td>
<td>cancer disease deal regulatory tax therapeutic</td>
</tr>
<tr>
<td>Les Moonves</td>
<td>mortgage risk income tough money fee</td>
</tr>
<tr>
<td>Mark Parker</td>
<td>punch case music football love passion</td>
</tr>
<tr>
<td>Mark R. Hunter</td>
<td>content deal bang success terrific confident scorpion late</td>
</tr>
<tr>
<td>Omid Kordestani</td>
<td>basketball demand excited elite offense discipline</td>
</tr>
<tr>
<td>Richard J. Harshman</td>
<td>shopping holiday progress pay music watch friction</td>
</tr>
<tr>
<td>Steven Temares</td>
<td>demand fell defect excellent labor wages aggressive falling</td>
</tr>
<tr>
<td>Thomas Kennedy</td>
<td>sue shopping baby improve excited exciting share success pleased fulfillment</td>
</tr>
<tr>
<td></td>
<td>missile defense demand government award bottom threat</td>
</tr>
</tbody>
</table>

Focusing on unique terms (TFIDF) characterizes the companies.

Electronic Arts

Estée Lauder

Alexion Pharmaceuticals

Loews

Approximate number of unique adjectives: 5
5. STICKY social media
Uncharted Flare Chart. Vertical axis is number of retweets. Lines are individual tweets. Some tweets flare up quickly then flatten, some go up, stutter, then go up more, etc.
Superbowl 50 Retweets

Note: some retweet values are offscreen

Uncharted Flare Chart. Betty White goes up flattens a bit, then up more, repeats. Live version: http://unchartedsoftware.github.io/salt-core/demos/superbowl-retweets/
Superbowl 50 Retweets

Uncharted Flare Chart. Coldplay has a popular tweet at 8:45 pm which then flattens out.
Superbowl 50 Retweets

Uncharted Flare Chart. Esurance has a retweet campaign which pops up quickly but starts to flatten after 20 minutes.
Uncharted Flare Chart. This tweet from the International Space Station doesn’t flare up, but keeps getting more and more retweets. Note how only one tweet can be examined at a time.
Unemployment Rate as Percent Civilian Labour Force

Similar problem: many lines on a line chart are difficult to disambiguate. Only way to disambiguate is to interact with each line individually. Slow!

Unemployment Rate as Percent Civilian Labour Force

2000-2014 by Country. Each line is labeled with microtext in multiple languages.

Top Hashtags

Sept 2017

Flare Chart of Hashtags
Top Hashtags

Sept 2017

Apple event generates flares.

#iPhoneX

#FaceID
Hurricanes are popular hashtags, even misspellings.
Top Hurricane Retweets

Sep. 6-14, 2017

Detailed hurricane tweet content on flare chart.
@PopeQuanPaul – Oh we good. Trump won't let Jose in the US.
@jdrudd - Hurricane #Irma now a category 5 and it's a beast. It's larger than the state of Ohio, if that helps put it in perspective.
@PascoSheriff_ To clarify, DO NOT shoot weapons @ #Irma. You won't make it turn around; it will have very dangerous side effects.
So What

Sticky is about a lot of things
• number retweets
• longevity
• reach across communities
• who’s saying it
• what they’re saying

And...
• geographic reach
• topics
• demographic segmentation
• message impact
• drift over time
Salt

TILE-BASED VISUAL ANALYTICS
• Hierarchical data tiling using cluster computing.
• Interactive on-demand image tile generation.
• Layers of raw data and derivative analytics.

OPEN SOURCE
• Saltlib.com
• Built on Apache Spark, Hadoop and ElasticSearch.
  o Cross Plots, Geospatial, Time-series, Graph analysis
Geo-temporal Movement and Events View

People Trails

- By connecting users geo-coded social media posts ordered temporally aggregate “people trails” can be extracted and aggregated to see movement patterns
- User provides selection criteria to specify population segment to generate trails
  - Based upon topic or keywords, time and location or start with seed social media accounts and generate trails for community
  - E.g. generate trails for all people going through Kennedy Airport June 1 – June 20
Raw Data Drilldown

Longitude: -73.81° to -73.81°
Latitude: 40.70° to 40.70°
Total Tweets: 104

Top 10 user distribution (86 total)

Top 25 hashtag distribution (29 total)

Users

User: nytokenblackguy
Total Tweets: 5
Most Recent Tweets

September 13, 2013

NYTokenblackguy

Even as a passenger I hate the Airtrain. (@ JFK AirTrain - Jamaica Station) http://t.co/UB81Fes57D

8:38:30 am
Geo-temporal Movement and Events View

Delta-plots and heatmaps

- Use heatmaps of activity in NYC taxi or social media to summarize overall activity and hot spots
- Delta plots show differences from average behavior or previous time periods
  - See increases or decreases in typical behavior for selected time period
- Supports identifying abnormal patterns and spatial extents
- Filtering by frequency allows focusing in on areas of largest activity or change, e.g. traffic problems during construction at the airport or change in usage during a street festival

Polar color ramps show positive / negative difference from “normal” or previous time periods

Heatmaps with a perceptually correct color ramp summarize overall activity
The delta plot data is noisy, but still a red stripe is clearly visible in this plot indicating a big change such as an event occurring.
The most frequent hashtags (TFIDF) can be used to quickly characterize the event – in this case a street festival.
6. LITIGIOUS TOPICS
Intellectual Property

- Where are the hot spots?
  - What skills and expertise do we need?

- Where are the problem spots?
  - More diligent in crafting patent
    (more time, more resources)
Plot showing every patent since 1982, grouped into communities based on each patent’s citation of earlier relevant patents.
Labels, extracted using TFIDF, show the most unique terms on the most heavily cited patent in each community. These terms give an indication of the topics in each community.
And focus on this community and sub-communicates. Labels indicate topics related to telephones.
Italic angle indicates litigation ratio in that community. Steep italic indicates high litigation, reverse italic indicates low litigation.
See citations to other communities.
Method and system for preventing illicit use of a telephony platform
US 8737962 B2

ABSTRACT
A system and method for preventing illicit use of a telephony platform that includes enrolling a plurality of accounts on a telecommunications platform, wherein an account includes account configuration; at a fraud detection system of the telecommunications platform, receiving account usage data, wherein the usage data includes at least communication configuration data and billing configuration data of account configuration and further includes communication history of the plurality of accounts; calculating fraud scores of a set of fraud rules from the usage data, wherein at least a sub-set of the fraud rules include conditions of usage data patterns between at least two accounts; detecting when the fraud scores of an account satisfy a fraud threshold; and initiating an action response when a fraud score satisfies the fraud threshold.

Drill to patent publication.
6. SO WHAT?
1. Visualization isn’t just about numbers. Text repositories may be 4x or more > numeric data.

2. The next big leap in analytics and visualization will be much deeper integration of content throughout the full analytics + visualization analysis cycle.

3. Real-world applications
   - Tone analysis of your (external) communications
   - Ongoing stickiness of (marketing) messages
   - Local topics and event detection
   - Litigious communities
Text Analytics and Visualization is Much Broader

Letters

Words

Phrases

Sentences

Paragraphs

Documents

Letters

Categories

Topics

Sentiment & Emotion

Tone characterization

Prosody Display

Words

Social Stickiness

Classifiers & Sets

Graph Analysis

Sentence Analysis

Phrases

Sentimental

Entity Recognition & Tagging

Style Analysis

Semantics/Summarization

Sentence

Analysis

Sentences

Social Stickiness

Classifiers & Sets

Graph Analysis

Sentence Analysis

Proximity Analysis

Paragraphs

Documents

Entity Recognition & Tagging

Style Analysis

Semantics/Summarization

Opinion Analysis
What can you do?

- Understand what text data you have
- Define what business insights you want to extract
- Use open source tools for analytics and visualization (e.g. Saltlib, D3js, NLTK, Stanford Core NLP)
- Incrementally experiment towards a solution
- Productionize the parts that are successful
THANK YOU

Richard Brath
rbrath@unchartedsoftware.com
@rkbrath
richardbrath.wordpress.com

Scott Langevin
slangevin@unchartedsoftware.com
@slangevi
blog.uncharted.software
About Uncharted: Profoundly visual software for improved awareness, analysis and decision-making, used by hundreds of thousands in government, law enforcement, finance, and more.