Humanizing the Machine with Language: How the Future gets Written
Kristian J. Hammond

Northwestern
McCORMICK SCHOOL OF ENGINEERING
Electrical Engineering & Computer Science
Professor, Computer Science

Chief Scientist, Co-Founder
Computers and Language
A Story
Philip C. Hammond
Orphan

No personal past

Archeologist

A life uncovering the past
More than words…

Explanations

Stories

Connections
Peer into the past

Communicate ideas

Connect to each other

Project the future
Language is miraculous
Language is uniquely human
We have a new partner
It struggles with language
Tea, Earl Grey, Hot
Language is a window into the world
We want the machine to tell us what it knows about its world
It knows a lot

It can say very little
What it knows

What it can say
What it knows

What it can say
hat it knows

What it can say
The machine’s world
2.5 Zetabytes of data are produced every day
250,000 Libraries of Congress, every day
And we are trying to meter everything
And we are trying to meter everything
Spreadsheets?
MODERN DATA SCIENTIST

Data Scientist, the easiest job of the 21st century, requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS
- Machine learning
- Statistical modeling
- Experiment design
- Bayesian inference
- Supervised learning: decision trees, random forests, logistic regression
- Unsupervised learning: clustering, dimensionality reduction
- Optimization: gradient descent and variants

PROGRAMMING & DATABASE
- Computer science fundamentals
- Scripting language e.g. Python
- Statistical computing packages, e.g. R
- Databases: SQL and NoSQL
- Relational algebra
- Parallel databases and parallel query processing
- MapReduce concepts
- Hadoop and Hive/Parquet
- Custom reducers
- Experience with data science tools like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS
- Passionate about the business
- Curious about data
- Influence without authority
- Hacker mindset
- Problem solver
- Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION
- Able to engage with senior management
- Story telling skills
- Translate data-driven insights into decisions and actions
- Visual art design
- R packages like ggplot or lattice
- Knowledge of any R visualization tools e.g. QGIS, OpenRefine
DID YOU SEE ANY ERRORS ON THE SPREADSHEET I PUT TOGETHER?

ONLY THREE.

WHAT ARE THEY?

YOUR DATA, YOUR FORMAT, AND YOUR FORMULAS.

Simple examples
Natural Language Generation
While still slightly better than sales of the small light widget, sales of the overly large widget have been on the decline over the last year. During the same period, sales of the small light widget have been steadily improving.
The automatic generation of language based on known facts
<table>
<thead>
<tr>
<th>Time</th>
<th>Entity</th>
<th>Metric</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td>ID_0075</td>
<td>$2,500,000</td>
</tr>
<tr>
<td>Q2</td>
<td>ID_0075</td>
<td>$2,300,000</td>
</tr>
</tbody>
</table>

“Sales of the overly large widget have been on the decline over the last year.”

<table>
<thead>
<tr>
<th>Time</th>
<th>ID</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td>ID_0075</td>
<td>The overly large widget</td>
</tr>
<tr>
<td>Q2</td>
<td>ID_0078</td>
<td></td>
</tr>
<tr>
<td>Q3</td>
<td>ID_0078</td>
<td></td>
</tr>
<tr>
<td>Q4</td>
<td>ID_0078</td>
<td>The small light widget</td>
</tr>
</tbody>
</table>
Advanced Natural Language Generation
Helen Crane’s performance and progress during the last quarter of 2016 has been exceptional. She was the top performer, ranking extremely well within her team.

This quarter, Helen closed $191,243 in sales. She had 93 deals with an average size of $2,056. Along both dimensions, she was in the 98th percentile within her team. Helen’s largest sale was also near the top at $8,333.

Helen’s sales total increased by $10,555 (about 6%) from the third to last quarter of 2016. This places her near the top in terms of overall improvement within her team.

It is difficult to assess her current Pipeline... the data seems to be faulty in general.
Transforming data into language

- DATA
  - ANALYZE
  - FACTS
  - CONCEPTUAL OUTLINE

- INFER

- GENERATE
  - INTELLIGENT NARRATIVE

ANALYZE

FACTS

CONCEPTUAL OUTLINE

INTELLIGENT NARRATIVE
For the $19.3 million in renewal, the Southeastern region gained achieving the highest volume above premium per agent was highest in the $22,000 per agent, compared to a state average Business lines in the state is 85.5%, with 84.69% for the Northeastern region Southwestern region, January written, an average of 1 729 commercial applications region and co
Data, Facts, Language
With these three elements, we can open the world
Giving people advice instead of data
Data about product flowing through stores

<table>
<thead>
<tr>
<th>Fiscal_Month_Id</th>
<th>Category_Id</th>
<th>Facility_Id</th>
<th>Sales_Level</th>
<th>Sales_Dollars</th>
<th>Loss_Pct</th>
<th>Loss_Pct</th>
<th>Shrink_Target_Pct</th>
</tr>
</thead>
<tbody>
<tr>
<td>201302</td>
<td>4102</td>
<td>311</td>
<td>M</td>
<td>$2,305</td>
<td>26%</td>
<td>-0.9</td>
<td>-0.1332</td>
</tr>
<tr>
<td>201302</td>
<td>4102</td>
<td>352</td>
<td>M</td>
<td>$3,132</td>
<td>86%</td>
<td>-8.6</td>
<td>-0.1332</td>
</tr>
<tr>
<td>201302</td>
<td>4102</td>
<td>611</td>
<td>H</td>
<td>$5,048</td>
<td>71%</td>
<td>-7.1</td>
<td>-0.1176</td>
</tr>
<tr>
<td>201302</td>
<td>4108</td>
<td>311</td>
<td>M</td>
<td>$979</td>
<td>43%</td>
<td>-4.3</td>
<td>-0.1283</td>
</tr>
<tr>
<td>201302</td>
<td>4108</td>
<td>352</td>
<td>M</td>
<td>$2,105</td>
<td>70%</td>
<td>-7</td>
<td>-0.1288</td>
</tr>
<tr>
<td>201302</td>
<td>4108</td>
<td>611</td>
<td>H</td>
<td>$4,235</td>
<td>13%</td>
<td>-1.3</td>
<td>-0.1017</td>
</tr>
<tr>
<td>201302</td>
<td>4112</td>
<td>311</td>
<td>M</td>
<td>$3,380</td>
<td>74%</td>
<td>-7.4</td>
<td>-0.0179</td>
</tr>
<tr>
<td>201302</td>
<td>4112</td>
<td>352</td>
<td>M</td>
<td>$6,593</td>
<td>40%</td>
<td>-4</td>
<td>-0.0179</td>
</tr>
<tr>
<td>201302</td>
<td>4112</td>
<td>611</td>
<td>H</td>
<td>$5,448</td>
<td>19%</td>
<td>-1.9</td>
<td>-0.0089</td>
</tr>
<tr>
<td>201302</td>
<td>4122</td>
<td>311</td>
<td>H</td>
<td>$1,860</td>
<td>70%</td>
<td>-7</td>
<td>-0.1999</td>
</tr>
<tr>
<td>201302</td>
<td>4122</td>
<td>352</td>
<td>M</td>
<td>$6,154</td>
<td>48%</td>
<td>-4.8</td>
<td>-0.2185</td>
</tr>
<tr>
<td>201302</td>
<td>4122</td>
<td>611</td>
<td>H</td>
<td>$3,718</td>
<td>27%</td>
<td>-2.7</td>
<td>-0.1999</td>
</tr>
<tr>
<td>201302</td>
<td>4126</td>
<td>311</td>
<td>H</td>
<td>$4,261</td>
<td>46%</td>
<td>-4.6</td>
<td>-0.0366</td>
</tr>
<tr>
<td>201302</td>
<td>4126</td>
<td>352</td>
<td>H</td>
<td>$1,174</td>
<td>5%</td>
<td>-0.5</td>
<td>-0.0366</td>
</tr>
<tr>
<td>201302</td>
<td>4126</td>
<td>611</td>
<td>H</td>
<td>$1,555</td>
<td>45%</td>
<td>-4.5</td>
<td>-0.0366</td>
</tr>
<tr>
<td>201303</td>
<td>4102</td>
<td>311</td>
<td>M</td>
<td>$6,389</td>
<td>70%</td>
<td>-7</td>
<td>-0.1332</td>
</tr>
<tr>
<td>201303</td>
<td>4102</td>
<td>352</td>
<td>M</td>
<td>$2,682</td>
<td>89%</td>
<td>-8.9</td>
<td>-0.1332</td>
</tr>
</tbody>
</table>

Sales Data
Inventory Loss
Inventory Targets
Store Demographics
Information and advice you can act on

Although the opportunity to reduce loss is the highest for Self-Service Cookies, concentrating on European Heritage Bread may have greater impact overall, as it consistently has the higher sales.
Helping people make decisions with information that they can read.
Basic data related to your fund

Macroeconomic Portfolio Data Risk Data Returns Data
Relative to the benchmark, the Fund’s allocation to the Consumer Discretionary sector and stock selections within Consumer Staples contributed to overall return.

The fund’s overweight position in Japan-based Tokyo Tatemono Co. hindered performance as the stock had a down quarter.

Overweight positions in Rightmove and Keller Group helped drive performance in the United Kingdom.
Helping advisors catch up with the people they serve.
Basic data related to a client’s portfolio

Portfolio Data

Holdings

Benchmarks

Goals
Since inception, Ms. Johnson’s annualized return (8.62%) exceeds her investment goal (6%). Her portfolio outperformed the benchmark by 5.63%.

Key drivers of the YTD return are investments in equities and commodities. Combined, the investments explain over 90% of the total return.

Ms. Johnson has been holding a significant amount of cash since Jan 2013, when the allocation jumped from 3% to 16% - this exceeds her stated liquidity needs of 5% per annum.
Helping people with information instead of data
<table>
<thead>
<tr>
<th>Beach Name</th>
<th>Timestamp</th>
<th>Water Temperature</th>
<th>Turbidity</th>
<th>Transducer Depth</th>
<th>Wave Height</th>
<th>Wave Period</th>
<th>Battery Life</th>
<th>Measurement ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Montrose Beach</td>
<td>8/30/13 8:00</td>
<td>20.3</td>
<td>3.18</td>
<td>0.891</td>
<td>0.08</td>
<td>3</td>
<td>9.4</td>
<td>MontroseBeach201308300800</td>
</tr>
<tr>
<td>Ohio Street Beach</td>
<td>5/26/13 15:00</td>
<td>14.4</td>
<td>5.23</td>
<td>0.311</td>
<td>4</td>
<td>12.4</td>
<td>OhioStreetBeach201305261500</td>
<td></td>
</tr>
<tr>
<td>Calumet Beach</td>
<td>5/23/14 16:00</td>
<td>23.2</td>
<td>3.63</td>
<td>1.201</td>
<td>0.174</td>
<td>6</td>
<td>9.4</td>
<td>CalumetBeach201305231600</td>
</tr>
<tr>
<td>Montrose Beach</td>
<td>5/28/14 12:00</td>
<td>16.2</td>
<td>1.26</td>
<td>1.514</td>
<td>0.147</td>
<td>4</td>
<td>11.7</td>
<td>CalumetBeach201405281200</td>
</tr>
<tr>
<td>Montrose Beach</td>
<td>5/28/14 13:00</td>
<td>14.5</td>
<td>2.72</td>
<td>1.395</td>
<td>0.306</td>
<td>3</td>
<td>11.9</td>
<td>MontroseBeach201405281300</td>
</tr>
<tr>
<td>Calumet Beach</td>
<td>5/28/14 13:00</td>
<td>16.1</td>
<td>1.31</td>
<td>1.568</td>
<td>0.156</td>
<td>4</td>
<td>11.7</td>
<td>CalumetBeach201405281300</td>
</tr>
<tr>
<td>Montrose Beach</td>
<td>5/28/14 14:00</td>
<td>14.5</td>
<td>2.97</td>
<td>1.388</td>
<td>0.328</td>
<td>3</td>
<td>11.9</td>
<td>MontroseBeach201405281400</td>
</tr>
<tr>
<td>Montrose Beach</td>
<td>5/28/14 15:00</td>
<td>16.1</td>
<td>1.31</td>
<td>1.568</td>
<td>0.156</td>
<td>4</td>
<td>11.7</td>
<td>CalumetBeach201405281500</td>
</tr>
<tr>
<td>Calumet Beach</td>
<td>5/28/14 16:00</td>
<td>14.1</td>
<td>5.06</td>
<td>1.382</td>
<td>0.34</td>
<td>11.9</td>
<td>MontroseBeach201405281700</td>
<td></td>
</tr>
<tr>
<td>Montrose Beach</td>
<td>5/28/14 17:00</td>
<td>14.2</td>
<td>5.76</td>
<td>1.415</td>
<td>0.356</td>
<td>3</td>
<td>11.9</td>
<td>MontroseBeach201405281800</td>
</tr>
<tr>
<td>Montrose Beach</td>
<td>5/28/14 18:00</td>
<td>17.3</td>
<td>1.8</td>
<td>1.501</td>
<td>0.188</td>
<td>4</td>
<td>11.7</td>
<td>MontroseBeach201405281900</td>
</tr>
<tr>
<td>Montrose Beach</td>
<td>5/28/14 19:00</td>
<td>17.8</td>
<td>1.82</td>
<td>1.487</td>
<td>0.194</td>
<td>4</td>
<td>11.7</td>
<td>CalumetBeach201405281900</td>
</tr>
<tr>
<td>Montrose Beach</td>
<td>5/28/14 20:00</td>
<td>14.2</td>
<td>6.32</td>
<td>1.386</td>
<td>0.346</td>
<td>3</td>
<td>11.9</td>
<td>MontroseBeach201405282000</td>
</tr>
<tr>
<td>Montrose Beach</td>
<td>5/28/14 21:00</td>
<td>14.5</td>
<td>6.35</td>
<td>1.401</td>
<td>0.38</td>
<td>4</td>
<td>11.7</td>
<td>CalumetBeach201405282100</td>
</tr>
<tr>
<td>Montrose Beach</td>
<td>5/28/14 22:00</td>
<td>14.5</td>
<td>6.88</td>
<td>1.413</td>
<td>0.345</td>
<td>4</td>
<td>11.9</td>
<td>MontroseBeach201405282200</td>
</tr>
<tr>
<td>Montrose Beach</td>
<td>5/28/14 23:00</td>
<td>14.5</td>
<td>7.52</td>
<td>1.406</td>
<td>0.331</td>
<td>4</td>
<td>11.9</td>
<td>MontroseBeach201405282300</td>
</tr>
<tr>
<td>Calumet Beach</td>
<td>5/28/14 23:00</td>
<td>16.5</td>
<td>1.69</td>
<td>1.548</td>
<td>0.177</td>
<td>4</td>
<td>11.7</td>
<td>CalumetBeach201405292300</td>
</tr>
<tr>
<td>Calumet Beach</td>
<td>5/29/14 13:00</td>
<td>16.1</td>
<td>1.97</td>
<td>1.589</td>
<td>0.159</td>
<td>4</td>
<td>11.9</td>
<td>MontroseBeach201405291300</td>
</tr>
<tr>
<td>Montrose Beach</td>
<td>5/29/14 14:00</td>
<td>14.2</td>
<td>7.18</td>
<td>1.46</td>
<td>0.305</td>
<td>4</td>
<td>11.9</td>
<td>MontroseBeach201405291400</td>
</tr>
<tr>
<td>Montrose Beach</td>
<td>5/29/14 2:00</td>
<td>14.2</td>
<td>6.35</td>
<td>1.45</td>
<td>0.321</td>
<td>3</td>
<td>11.9</td>
<td>MontroseBeach201405290200</td>
</tr>
<tr>
<td>Calumet Beach</td>
<td>5/29/14 3:00</td>
<td>16.2</td>
<td>1.97</td>
<td>1.589</td>
<td>0.159</td>
<td>4</td>
<td>11.7</td>
<td>CalumetBeach201405290300</td>
</tr>
<tr>
<td>Calumet Beach</td>
<td>5/29/14 4:00</td>
<td>16.1</td>
<td>1.82</td>
<td>1.507</td>
<td>0.156</td>
<td>4</td>
<td>11.7</td>
<td>CalumetBeach201405290400</td>
</tr>
<tr>
<td>Montrose Beach</td>
<td>5/29/14 4:00</td>
<td>14.1</td>
<td>6.27</td>
<td>1.456</td>
<td>0.306</td>
<td>3</td>
<td>11.8</td>
<td>MontroseBeach201405290400</td>
</tr>
<tr>
<td>Montrose Beach</td>
<td>5/29/14 5:00</td>
<td>14.1</td>
<td>5.65</td>
<td>1.421</td>
<td>0.292</td>
<td>3</td>
<td>11.8</td>
<td>MontroseBeach201405290500</td>
</tr>
</tbody>
</table>

Data on the quality of our water and beaches.
For the week ending July 16, 2016, 63rd Street Beach was the cleanest beach in Chicago, as measured by the cloudiness of the water (average hourly turbidity range of 0.64). Not only did the beach have the cleanest water throughout the week, but also measured the cleanest out of all beach waters at any time (NTU 0.13), which was achieved on Wednesday around 9 PM.

Calumet Beach was the calmest throughout the entire week, with an average wave height of 0.12 meters. This beach also happened to be the warmest, with an average temperature of 19.1°C.

Ohio Street Beach was the dirtiest beach in Chicago, as measured by the overall turbidity of the water (average NTU 1.22). Ohio Street Beach did not record the highest turbidity rating of the week, however. Rainbow Beach saw a reading of 6.57 NTU, which was recorded Wednesday around 1 PM. In fact, it has recorded the highest turbidity each week for the last four weeks.

Ohio was also the coldest of all the beaches, with an average hourly temperature of 16.0°C. Beachgoers would be advised to visit Oak Street beach that had a better than average cleanliness rating this week.
The future instead of the past
Test data is now used to just look backward
Advice that can reshape the way we teach

In Physics, you need to focus on issues of magnetics and electrostatics. You seem to be having problems with some of the basic questions in these areas that other students are managing well.

You should look back to Sections 3-5 in Chapter 7 of the material (for magnetics) and Section 3 in Chapter 6 (for electrostatics).
Transforming data into language

DATA → ANALYZE → FACTS → INFER → CONCEPTUAL OUTLINE → GENERATE → INTELLIGENT NARRATIVE
Performance Assessment

- Entity
  - Year over Year
  - Month over Month
  - Historical Trends
  - Metric
  - Driver
  - Driver
  - Cohort
  - Target
  - Benchmark
  - Goal
There are other ways to ask for information.
Data into facts

Facts into stories

Stories into language
Store Profit & Sales

Accounting for your selection, this analysis measures Profit and Sales by State.

For Profit:
- California has a Profit value of $6,668.

For Sales:
- California has a Sales value of $50,551.
Relative Price Performance

This analysis measures Fund A vs. SP 500 by MONTH(Date).

- Fund A underperformed SP 500 93% of the time (lower by $50.74 on average).
- The best period was February 2012, when Fund A outperformed SP 500 by $6.35 (5.26%).
- On the other hand, in December 2015, Fund A underperformed SP 500 by $111, which was worse than any other period.
- Fund A and SP 500 had a moderate negative correlation (coefficient of -0.52), indicating Fund A actually decreased as SP 500 increased (and vice versa) from one period to another.
- However, there was a period of positive correlation occurring between June 2012 and July 2012 when Fund A increased by $1.46 and SP 500 increased by $1.52.
- SP 500 finished trending downward in September 2016, more than Fund A.
- Fund A had more variability than SP 500 with a normalized variation of $12.33 from its general trend (RMSE).
Unemployment by Age & Month

The analysis compares number of people unemployed by MONTH(Date) and Age.

- Average of number of people unemployed was 515,765 across all seven age groups and across all 122 months.
- Unfortunately, each of the seven age groups increased from January 2005 to February 2015 with 65 years and over rising the most (138%) and 35 to 44 years rising the least (33%) over that time frame.
- 65 years and over trended upward the most in the final month. On the other hand, 25 to 34 years trended downward the most.
- 45 to 54 years had the highest degree of variability while 65 years and over was the most stable.

The individual age groups discussed below are called out based on the lowest average number of people unemployed over the 122 months.

For 16 to 19 years:

- Average of number of people unemployed was 212,869 across all 122 months.
- Values ranged from 71,000 (October 2006) to 438,000 (March 2010).
- Number of people unemployed increased by 34% over the course of the series.
- The largest single increase on a percentage basis occurred in July 2006 (+59%). However, the largest single increase on an absolute basis occurred in January 2012 (+92,000).
Narratives for BI, SAP BusinessObjects Lumira edition

The SAP BusinessObjects portfolio includes a set of modern analytic capabilities that help users better understand their business. As part of that portfolio of products, SAP® BusinessObjects Lumira is a self-service data visualization software that connects to a variety of data sources.

Narratives for Business Intelligence, SAP Business Objects Lumira edition, is an extension that automatically translates data into compelling narratives that are powered by advanced analytics. These narratives can be easily shared with a click of a button and understood by anyone.

TRY IT NOW
With language, the machine can share its stories.
With language, we free the information trapped in the machine.
With language, the machine becomes our partner.