INCORPORATING WEATHER DATA INTO CAPACITY PLANNING ANALYSIS: AN EFFORT TO MAKE IT RESILIENT DURING ALL SEASONS

Andrea Vasco, Movìri Inc. and Ben Davies, Movìri Inc.

Incorporating 'non-capacity' data into capacity planning efforts expands the capacity analysts view and understanding. Business metrics vs. hardware metrics make compelling capacity models, but external forces may have a measurable effect as well. This is a conversation about Incorporating weather data to determine if and how weather conditions affect customer interaction, employee behavior, and infrastructure utilization. Weather conditions did not have the impact we expected, but severe weather and cultural events are indeed a key driver for VPN (virtual private network) and VDI (virtual desktop infrastructure) demand.

Introduction

Not so very long ago, as part of a capacity team, we were mostly successful at identifying major events impacting the business and therefore the IT infrastructure, but needed to convince data owners to let the capacity team have access to 'non-capacity data', to expand and improve analysis efforts.

What we needed was to demonstrate that a 'non-capacity' metric could be effectively used in a capacity context. We chose weather data, partly because it was available. The thought was, "If we can get weather data, data that has nothing to do with capacity, and show that it can be incorporated into our forecasts in a meaningful way, then we can make a case for other 'non-capacity data' that we also wanted to incorporate." As the company was based in Chicago, getting the national weather service data from Chicago seemed like a logical place to start.

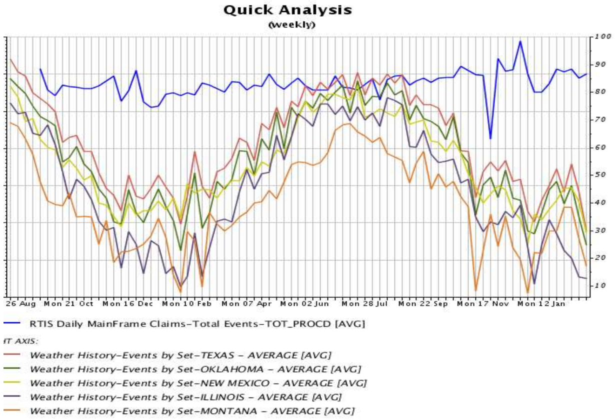
Method

Collect data from a free site[i]. Finding a site to get weather data at the time was easy as I remember. Looking now yields paid sites[ii], and a lot of projects that have compared business metrics vs the weather. [iii] [iv] [v]

Weather data was in CSV format. While other formats were available, this was the easiest for us to use. The overall method was one of iterative, bruit force. Compare metrics, make observations, thesis statements then try to prove or disprove. Sometimes this caused us to find new data, clarify the definition and meaning of existing data, alter our understandings of the systems and sub systems involved, and revise the thesis statements.

Results

Results were not what we expected. Our thesis statements were all proven wrong on normal scales. Daily weather had no impact on overall usage of IT resources, and while there are several reasons for this, the largest reason was our geographic dispersity; five states, and associated weather has zero measurable impact on the count of global claims processed.



However, in our original goal of increasing the awareness of our capacity efforts and gathering 'non-capacity' metrics, we were successful. At highest count, we were allowed to collect nearly one hundred "non-capacity" metrics and sub metrics. These metrics did allow for better capacity analysis.

Discussion

With Chicago weather in hand, we compared weather data with claims processed, web page hits, and a number of other business metrics, but quickly realized that our internal data did not lend itself to comparing to the weather at a Chicago Airport. None of our business data was broken down by geographic location, and very little by region. In addition, which weather metric was the most significant? The fact that it rained or snowed, the temperature or humidity or barometric pressure? It was not clear that a direct, meaningful, correlation could be made. Undeterred, we gathered weather data from other major cities in our business areas, and continued the analysis.

We had two specific ‘easy’ questions: Is weather a key driver for insurance claim volumes? And is weather a key driver in work from home infrastructure utilization?

As we will see these are not easy questions at all, but the investigation better informed our understanding of the systems, and metrics, allowing for better questions, different ways to look at the data, and gave useful results, even if completely different than we originally expected.

The problem statement is “determine if and how weather conditions affect customer interaction, employee behavior, and infrastructure utilization”. This was interpreted to two thesis statements / questions:
 Is weather a key driver for insurance claim volumes?
 Is weather a key driver in work from home infrastructure utilization?

To answer the questions, it is clear that we need the business metrics of ‘insurance claim volumes”, which were already had in our capacity tool, and work from home metrics which were defined as VPN use counts, VDI use counts and the equipment metrics for these systems, which were also in our capacity tool. That left only the “weather data”. We quickly found out that ‘the weather data’ is an incompletely defined term.

The hard problem: defining weather by location temperature is not the only useful weather metric, as precipitation, humidity, ice or other measures may be significant. In addition, the combination of conditions could be important and the definition of ‘nice’ changes with the seasons / location. We attempted to score the weather. A hot day with high humidity is different than a hot day that has low humidity. And a nice day in the winter is different from a nice day in the summer. And does a nice day in the morning followed by a thunderstorm in the evening count as a good day or a stormy day? Does a bad day in Chicago count as a bad day in all of Illinois as a region? These questions are magnified by 5 states, including Texas which can have completely different weather scores in Amarillo in the north, vs. Dallas in the east, vs. Corpus Christi in the south, vs. El Paso in the West. See Table 1 for a sample of weather in one city.

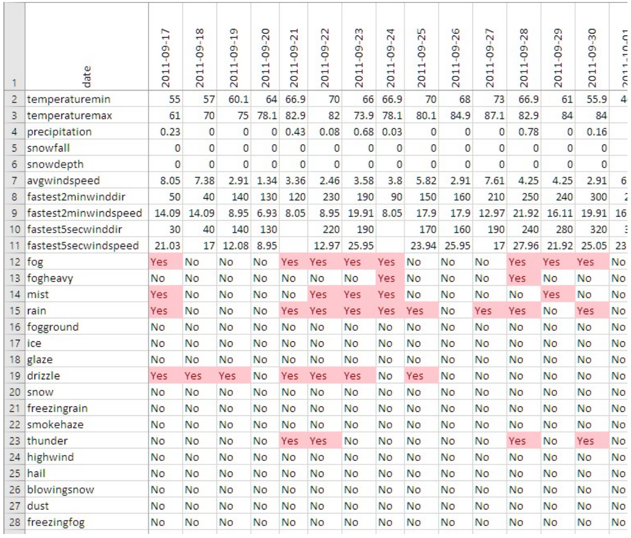
Breaking down the business metrics by location or region so that these can be compared to the local weather metrics is also difficult. Our business metrics were divided by business unit, not location. Estimating location even to the region was problematic, and ultimately not useful as business loads shift internally based on code releases, maintenance, and marketing efforts.

The lesson of this effort is that data should be grouped together in similar ways to be compared. To the extent that precision of the groupings matches the precision of the metrics, useful comparisons can be done. When grouping is imprecise, data will also be imprecise. Massive winter storm vs global daily counts with has a chance to be helpful, but was not for us. Hour by hour temperature in one city vs. hour by hour global claim counts were not useful.

 Significant weather events correlate easier to global total daily counts. If our business data allowed for location by location comparisons then useful observations may have been found. Essentially smaller geographic locations need to be compared to corresponding geographic location weather data.

A further complication is the correlation delays. Today’s weather vs todays VPN usage is appropriate, however, todays weather should be compared to a few days hence claim counts, as there is a systemic delay between doctor visit and claim submission.

Table 1 – Sample weather data for one site



1	date	2011-09-17	2011-09-18	2011-09-19	2011-09-20	2011-09-21	2011-09-22	2011-09-23	2011-09-24	2011-09-25	2011-09-26	2011-09-27	2011-09-28	2011-09-29	2011-09-30	2011-10-01
2	temperaturemin	55	57	60.1	64	66.9	70	66	66.9	70	68	73	66.9	61	55.9	4
3	temperaturemax	61	70	75	78.1	82.9	82	73.9	78.1	80.1	84.9	87.1	82.9	84	84	
4	precipitation	0.23	0	0	0	0.43	0.08	0.68	0.03	0	0	0	0.78	0	0.16	
5	snowfall	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
6	snowdepth	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
7	avgwindspeed	8.05	7.38	2.91	1.34	3.36	2.46	3.58	3.8	5.82	2.91	7.61	4.25	4.25	2.91	6
8	fastest2minwinddir	50	40	140	130	120	230	190	90	150	160	210	250	240	300	2
9	fastest2minwindspeed	14.09	14.09	8.95	6.93	8.05	8.95	19.91	8.05	17.9	17.9	12.97	21.92	16.11	19.91	16
10	fastest5secwinddir	30	40	140	130		220	190		170	160	190	240	280	320	3
11	fastest5secwindspeed	21.03	17	12.08	8.95		12.97	25.95		23.94	25.95	17	27.96	21.92	25.05	23
12	fog	Yes	No	No	No	Yes	Yes	Yes	Yes	No	No	No	Yes	Yes	Yes	No
13	fogheavy	No	No	No	No	No	No	No	Yes	No	No	No	Yes	No	No	No
14	mist	Yes	No	No	No	No	Yes	Yes	Yes	No	No	No	No	Yes	No	No
15	rain	Yes	No	No	No	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	No	Yes	No
16	fogground	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
17	ice	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
18	glaze	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
19	drizzle	Yes	Yes	Yes	No	Yes	Yes	Yes	No	Yes	No	No	No	No	No	No
20	snow	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
21	freezingrain	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
22	smokehaze	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
23	thunder	No	No	No	No	Yes	Yes	No	No	No	No	No	Yes	No	Yes	No
24	highwind	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
25	hail	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
26	blowingsnow	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
27	dust	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
28	freezingfog	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No

Rdu-weather-history.csv Sample data

The problem of False Flags and Metric Dilution

Once we had weather data, we compared to internal metrics. This forced a ‘better understanding’ of what the metric measures and what, exactly it means. Web ‘hits’ are recorded from the users IP address which is local to that user. Comparing weather data to the web log should show correlation but this did not happen. The IP is not of the user but rather, the user of record, which is the proxy or firewall of that user. This is a false flag problem. Our analysis could make observations, but not strong correlations. For example, when I connected from my work computer from Tulsa or Dallas, my IP address of record was from Spring Field, Illinois. Where my corporate firewall IP address was registered to our ISP (Internet Service Provider). Many of our B2B customers have geographically dispersed workers, but all come from the same IP address associated with their corporate datacenter ISP. Correlating their activity to the weather where their ISP registered the firewalls IP does not work.

Other metrics quickly became diluted to the point of not meaning what we thought. Claims processed are not original claims but a mixture of original claims (which happen within a day or three of the original 'event'), reprocessed claims (that could be processed dozens of times and can be up to 20% of all claims in a given day), follow up service claims (which happen way after the original 'event', which may also be reprocessed) and duplicate claims (claims resubmitted by a service provider 'just to be safe'). A very small slice of original claims could be attributed to the weather, per our actuaries (who would not share their data).

There were dozens of other examples where the metric with seemingly obvious meanings, were completely different than expected.

Is weather a key driver for insurance claim volumes?

The first question we intended to answer is if the weather would drive behavior that caused claim numbers to increase. Good weather drives people outside where they are active and get hurt vs bad weather keeping people inside where they are inactive and do not get hurt.

We found that the insurance claim lifecycle is not consistent enough to correlate claim count to the weather. The time between being hurt (weather related or not), to filing a claim is haphazard, and one event may create dozens of claims over extended periods of time. Claims that are not weather related overwhelm any weather induced claims. While the actuaries could tell the cost of the claims that were 'weather related', the counts of claims on a given day, did not correlate to the weather.

We also found that the weather driving people inside or outside did not diminish a person's propensity to hurt themselves, as measured by claim counts. Dumb luck and dumbassery are equal opportunity causations for injury. Inside, outside, good weather or bad. One can rock climb indoors and even sky dive indoors. We have even combined sports that can be played anywhere. Cycle Ball,[vi] and Unicycle Hockey.[vii] Suggesting we are quite creative at inventing potentially hazardous activities.

We did not find an 'at the time of service' metric available to compare with the weather. The daily counts are not consistent enough from the injury to the claim activity for a correlation to be established. Background noise of 'normal claims' and random delays between events, systemic dilution, and the small slice of population of incidents that may be weather related, conspire to inhibit correlation. We did see a yearly cycle where the last quarter claim rates increased but this was attributed to beating the reset of deductibles, and the first quarter when newly insured used new to them policies and exercising once a year benefits, but alas, no correlation with the weather.

Other research in this area, focused on admissions for the purpose of staffing decisions.
-Predicting Trauma Admissions, The Effect of Weather, Weekday, and Other Variables[viii]
-Effect of weather on attendance with injury at a pediatric emergency department[ix]
Both of these studies, and many others referenced, delivered mixed correlation. Some studies show weather was a strong predictor sometimes not. In some studies, day of the week showed strong correlation, while other studies did not. Results seemed locally dependent, and our five states presented way too many locations for an overall trend.

While these efforts helped us to understand the claim data better, we were not able to correlate weather to the submission of claims.

Determining the impact of weather on Clients and Employees activities

Measuring VDI / VPN access

VPN or Virtual Private Network, establish a connection from a users laptop to the corporate network using an encryption key. The company had several end points for these connections and dedicated internet "pipes". This was mostly outbound traffic from our datacenter perspective. Mouse clicks from external users into the data centers, big chunks of data out (to the remote user). This is opposite of normal internet traffic. Mouse clicks from internal users out to the internet, and big chunks of data back to the internal users.

There are a few key metrics for VPN connections. Total authorized VPN users, connected users, bandwidth in available, bandwidth out available, bandwidth used in, and bandwidth used out. Comparing these will suggest the total number of users that given bandwidth will support. The count of users could also be compared against the equipment that services VPN connections. Encryption is somewhat CPU expensive (vs. no encryption) and number of concurrent connected sessions may be limited by configuration or license.

VDI has the same sort of metrics, with similar comparisons. However, as the entire desktop is virtual there is a limit to the number of concurrent sessions one 'server' could hosts, and the limiting factor tends not to be

CPU, but rather memory, network and disk IO. Most of this traffic stays on the Corporate network with only 'screen scrapes' being sent back to the user.

Previous studies of VPN resources suggested that the hardware was 'enough' for 90% of the authorized user count, but the bandwidth would not 'handle the load'. The solution to this was cleaver in that when there was 90% of the authorized users on VPN they would NOT be 'in the office', so the outbound internet connection would be leveraged via load managers and clever network addressing to allow for inbound VPN traffic.

VDI was a different story. At the time, building servers took several days and half a day to incorporate it into the VDI pool, it was decided that 'in a pinch', disaster recovery equipment would be commandeered as part of the 'business continuity plan' and the build process would be worked on to shorten deployment cycles. These efforts would prove most helpful.

Is weather as a key driver in 'Work from Home'?

The second question we intended to answer is if the weather would drive the work from home resources. The thesis is that on especially good days, people would not go to the office but rather work from home using the VPN and VDI resources. And that bad weather would also drive people to use work from home resources.

What we found is that our employees were completely opposite of what we expected, much to their credit. Good days did not induce statistically significant changes in work from home resources, nor did bad days. What did register was a push for work from home programs (which caused a steady increase in VPN and VDI utilization) and the closing of small offices, then progressively larger offices. Any weather induced fluctuation were obscured by normal fluctuation and this new growth. Establishing business practices that were tolerant of the work from home environment, for conducting all phases of business helped with business continuity posture, as we will see below. As a side note, this effort also caused IT costs to increase (unfunded growth) and business cost to drop (closing offices). This, in turn, caused some budget and chargeback conversations that are outside the scope of this paper, but none the less, an interesting effect.

While looking at VPN and VDI use data, we found interesting anomalies in the internal count data NOT explained by the weather which caused us to ask different questions. What is driving these spikes? It turns out they are event driven.

Presidential election politics, sporting championship parades, and city-wide block parties have a more dramatic effect on VPN and VDI resources, than the temperature and relative humidity. The resulting employee behavior seems to be that when going to work was made especially difficult, like road closures for parades that bring an extra million people to down town, employees would opt for work from home options rather than 'call in' or 'brave the conditions'. This behavior was encouraged by Corporate management for these disruptive events.

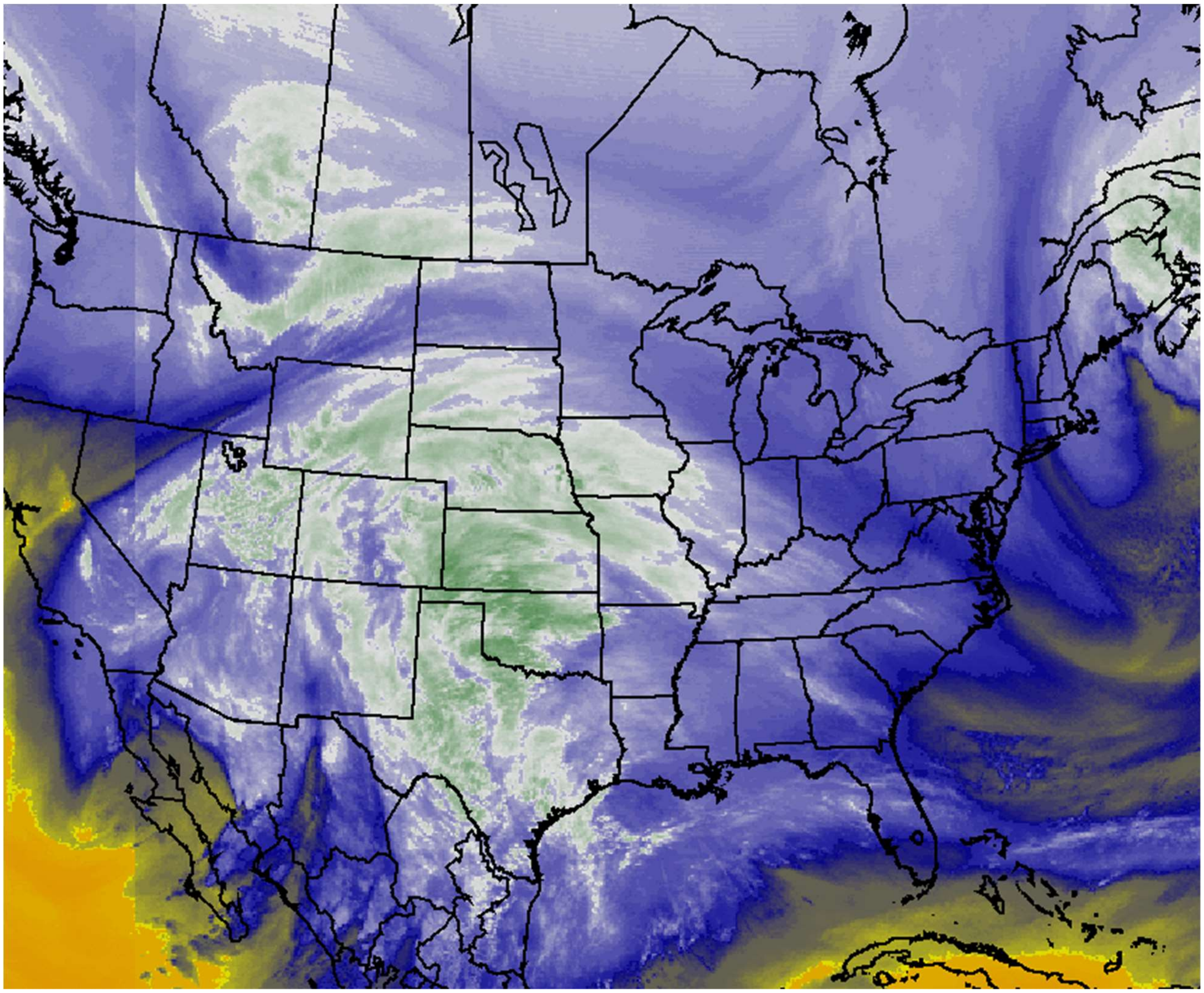
The Chicago Corporate office of about three thousand sits right next to the AON Building that was President Obamas Chicago Office, which was the center of considerable disruption when President Obama was in town. It is also next to Millennium park, which is the end of the championship parades for the Chicago Black Hawks winning Sir Stanley's Cup (three times in six years), The Chicago Cubs breaking the hundred eight year drought between World Series Wins (Cubs Win. Cubs Win), and hosts the taste of Chicago food week. Granted having a corporate office in close proximity to these very disruptive events put us in a unique position, and we did see spikes in VPN and VDI use as employees stayed home rather than deal with the crowds. [x]

While these events showed notable utilization increases in work from home technologies and resources, the effect was not sufficient to require additional resources. Overall, **we determined local ROUTINE weather is NOT a key driver in work from home resources.**

Except for big weather.

We compared isolated events and determined that the VPN and VDI environments saw upward demand pressure from regional events and big bad weather, but the environment absorbed the additional loads when generated one region at a time. However, while doing our investigations, a large winter storm was going to impact all of our states over a week.

The monster storm of Feb 2015 would combine these regional impacts in not trivial ways. Our models predicted a five to ten fold increase in utilization of the work from home technologies and infrastructure. The various technology owners exercised the business continuity plan (which is distinct from a disaster recovery plan) and quickly reinforced the VPN and VDI capacity to service 8 times current capacity, to accommodate capacity spike seen for the week of the weather event.



GEOS Water Vapor Imagery from January 31st. https://www.weather.gov/lot/2015_Feb01_Snow

This is a significant result as these employees, working from home, **allowed the company to function uninterrupted**. While many other companies were closed (losing revenue and disappointing customers during a stressful time) we remained open and processing claims, answering provider and user questions, and reacting to the local demands of the storm as it affected our locations. As the storm had an edge, the area impacted was "sent home" and allowed to deal with the impacts, (poor travel conditions, school closings, power outage and the like) while other areas of the company took over business function. Then as the storm moved to a different area the next area would be "Sent Home" while the first area resumed the work routine 'but from home'. After the storm, the aftermath made travel hazardous, but still we remained essentially fully staffed using the work from home resources, and business process that made working from home a 'normal business activity'. Yes, many people did take PTO (Personal Time Off), but quite a few did not, specifically because business processes were modified to use the work from home technology, and the technology was expandable to meet this demand.

Conclusion

Adding non-capacity data to the Capacity Optimization tools have non-obvious benefits, and to the extent that the capacity planners and their customers are comfortable with adding and removing metrics for analysis, unexpected observations, insights and benefits are delivered. Specifically, that the data enables analysis, which answers some questions **but raises different ones.** This causes more data to be added, manipulated and analyzed in a different way. Again, changing understandings of 'well understood metrics', answering some questions but raising even more questions.

To the extent that capacity operations can easily and freely add and analyze data, even data that is not overtly 'capacity data', then observations and insights abound. Playing with the tools and techniques of adding, manipulating, analyzing and otherwise attempting to extract information from data, delivers value in direct insights and observations, but more importantly in the mental exercise of understanding what the metrics actually mean about the business processes.

In this particular case, we set out to determine two very specific questions, but quickly realized the limits of our data, and how that data could be broken out. Making observations and insights that change the premise of the questions, which in turn lead to observations and insights that allowed us to recognize a significant event, quantify the assumptions and confidently act upon our assessments.

We are confident that by liberating ourselves from the artificial constraint of 'only capacity data belongs in a capacity tool', we deliver better capacity insights based on better understandings of the metrics and systems we evaluate.

The business adopting work from home technologies into their normal business practices allowed the userbase to easily shift work modes as external conditions dictated. This is just as important to successfully handling this situation as being able to predict demand and add capacity.

[i] https://www.ncdc.noaa.gov/cdo-web/#t=secondTabLink NOAA Climate Data Online
[ii] https://darksky.net/dev Dark Sky API
[iii] Weather vs. crime - http://crime.static-eric.com/
[iv] Weather and sex - http://www.smh.com.au/comment/how-climate-change-could-ruin-your-sex-life-20151105-gkrc67.html
[v] Weather determines sex https://www.sciencealert.com/scientists-find-australian-lizards-that-swap-gender-in-hot-weather
[vi] https://en.wikipedia.org/wiki/Cycle_ball
[vii] https://en.wikipedia.org/wiki/Unicycle_hockey

[viii] Predicting Trauma Admissions, The Effect of Weather, Weekday, and Other Variables
https://www.researchgate.net/profile/Jon_Roesler/publication/41011428_Predicting_trauma_admissions_the_effect_of_weather_weekday_and_other_variables/links/555d08a608ae6f4dcc8bd32b.pdf
The forecasting model was successful in reflecting the pattern of trauma admissions; however, its usefulness was limited in that the predicted range of daily trauma admissions was much narrower than the observed number of admissions.

[ix] Effect of weather on attendance with injury at a pediatric emergency department
http://emj.bmj.com/content/20/2/204

[x] 2010 Chicago Blackhawks win Stanley Cup (June 2010)
2012 NATO Summit Chicago (May 2012).
2013 Chicago Blackhawks win Stanley Cup (June 2013)
2015 Winter Storm (Feb 2015)
2015 Chicago Blackhawks win Stanley Cup (June 2015)
2016 Presidential trips (April, October)
2016 Chicago Cubs win World Series (October 2016)
https://en.wikipedia.org/wiki/January_31_%E2%80%93_February_2,_2015_North_American_blizzard
https://en.wikipedia.org/wiki/2012_Chicago_summit
Note BCO was installed Sep 2013