BETTER TOGETHER

USING SPARK AND REDSHIFT TO COMBINE YOUR DATA WITH PUBLIC DATASETS

EUGENE MANDEL (@EUGMANDEL)
JAWBONE
QCON SF 2014

JAWBONE DATA

MOVEMENT

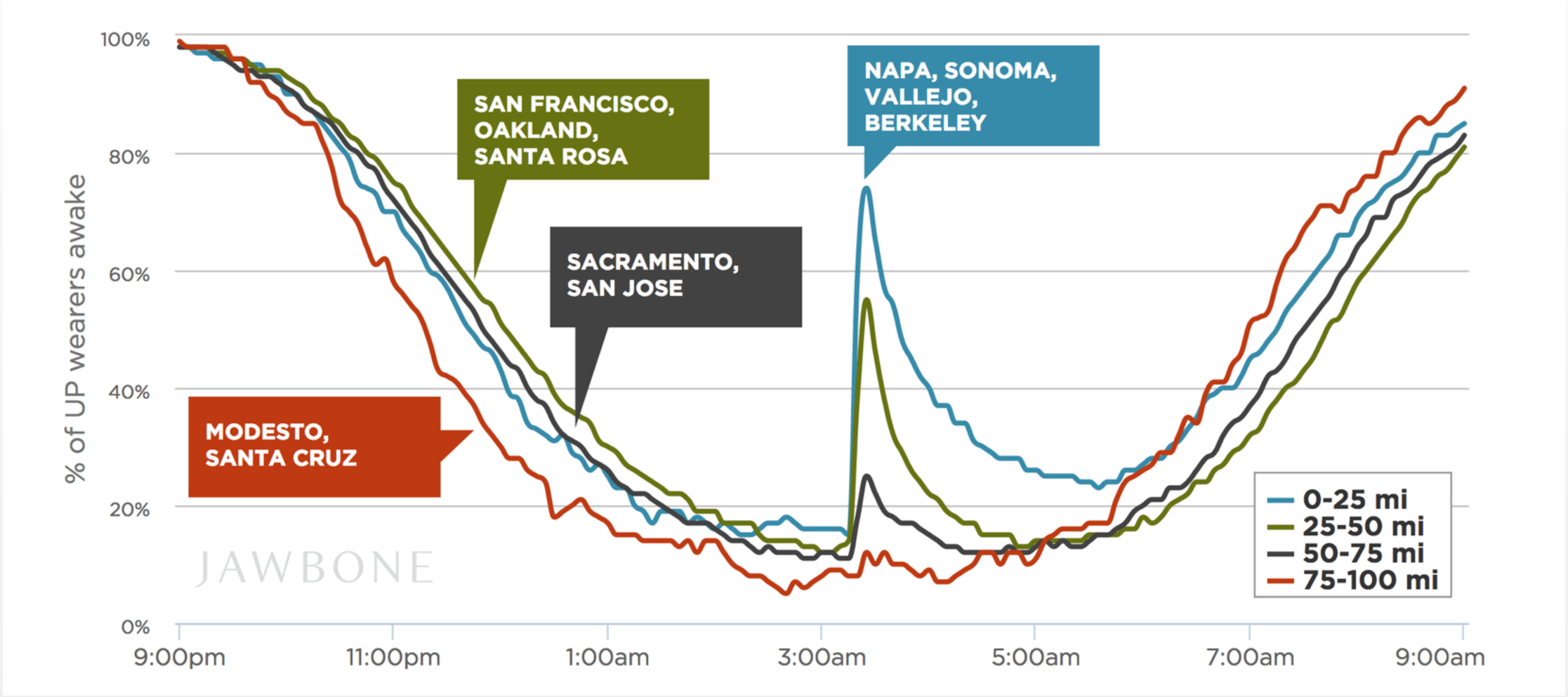
SLEEP

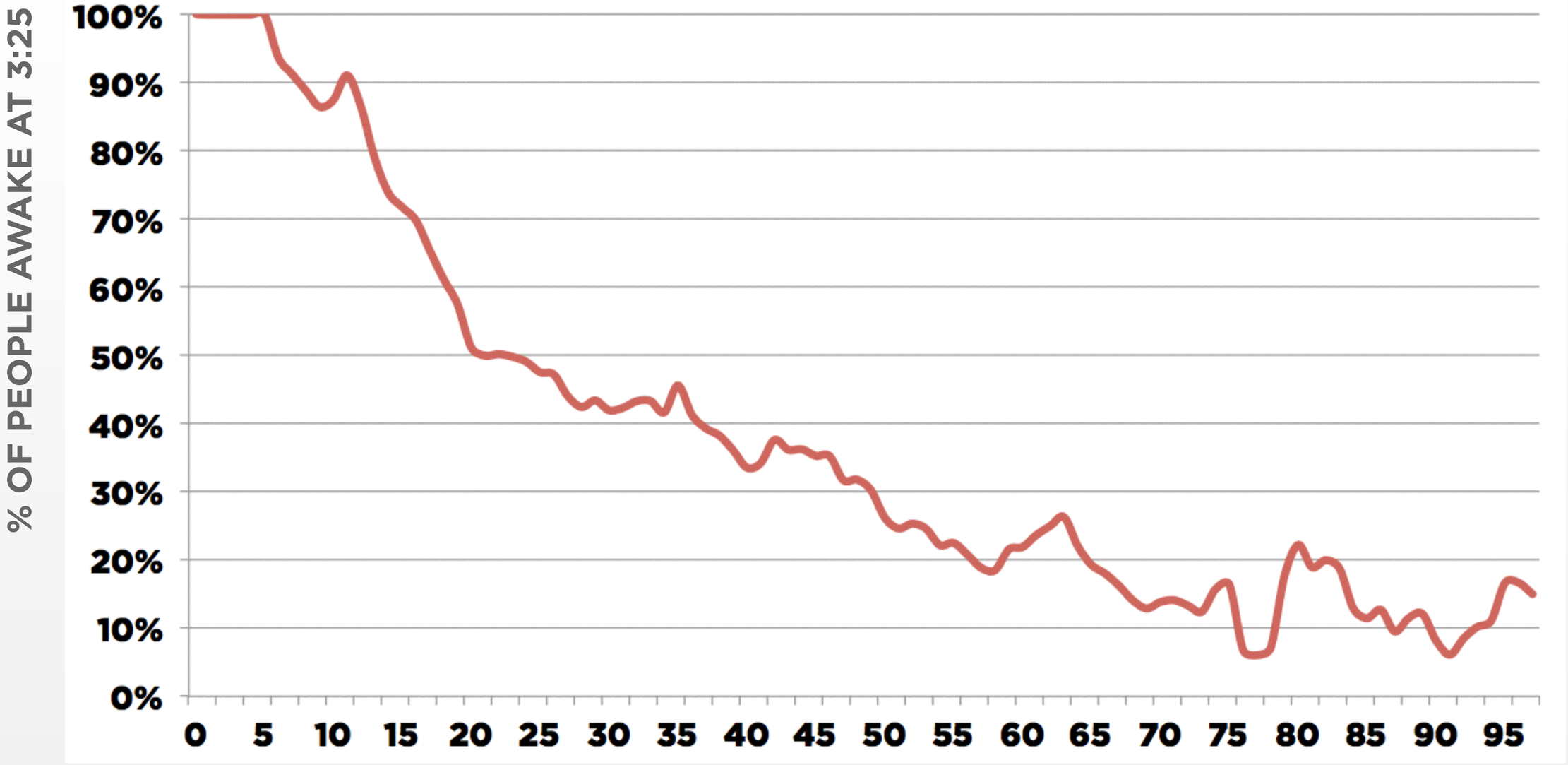
WORKOUTS

MEALS

MOOD

SOUTH NAPA EARTHQUAKE 2014




% OF PEOPLE AWAKE AT 3:25
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
DISTANCE FROM EPICENTER (MILES)

users were disturbed. All very predictable, yet this is a powerful demonstration of how big data can allow anyone from emergency relief workers to governments to measure the effects of a disaster.

Earthquakes And Fitness Monitors

Want to know how a new freeway is affecting the sleep of local residents? Interested in the comparative fitness levels of the neighborhoods around your area? Anonymized data from the likes of Jawbone would be able to tell you. As these always-on wearable devices get smaller, smarter and more comprehensive, the potential uses go even further

DATA FUSION IS THE PROCESS OF INTEGRATION OF MULTIPLE DATA AND KNOWLEDGE REPRESENTING THE SAME REAL-WORLD OBJECT INTO A CONSISTENT, ACCURATE, AND USEFUL REPRESENTATION.

(WIKIPEDIA)

DATA FUSION - HOW TO FIND THE ELEPHANT

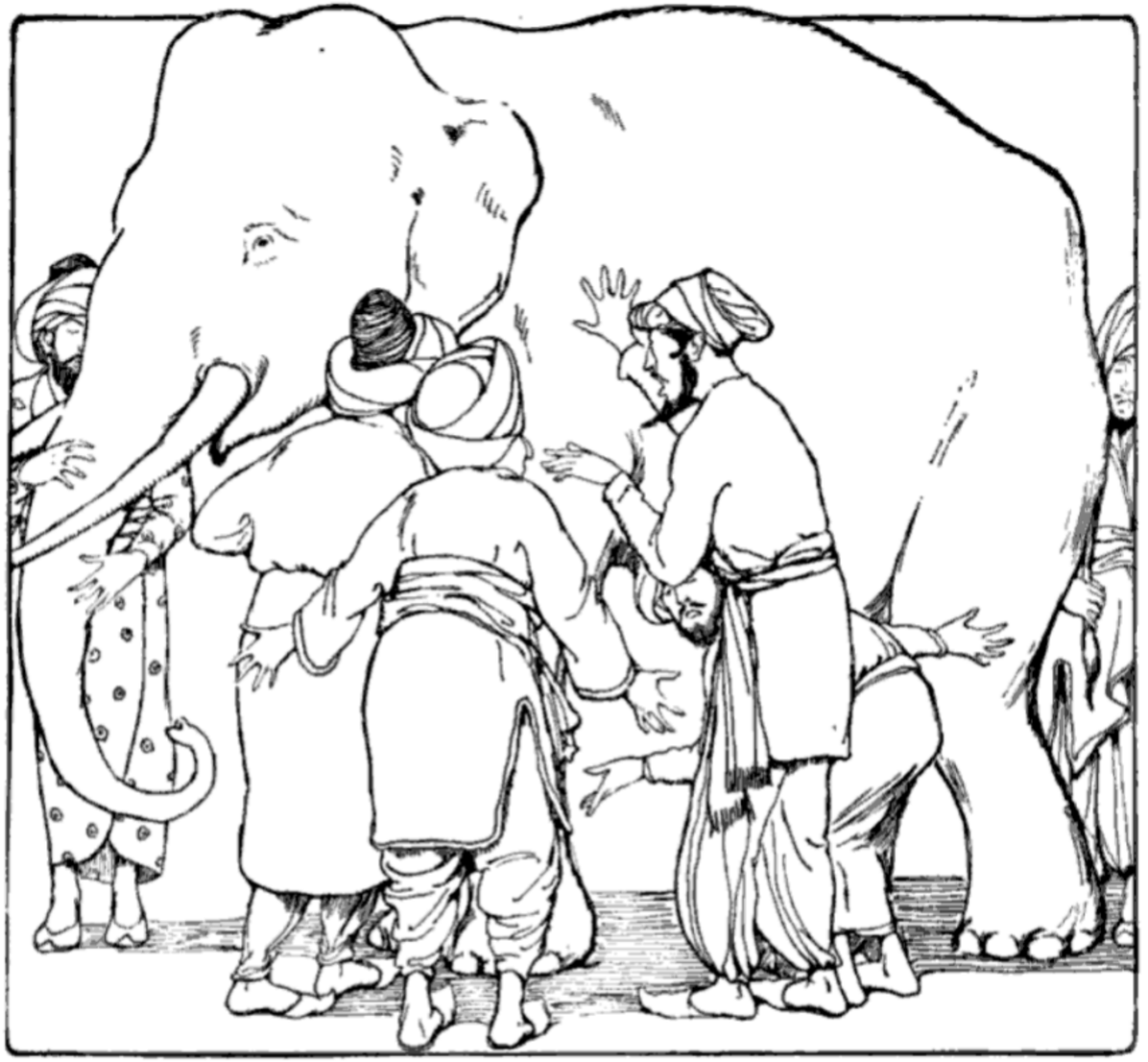


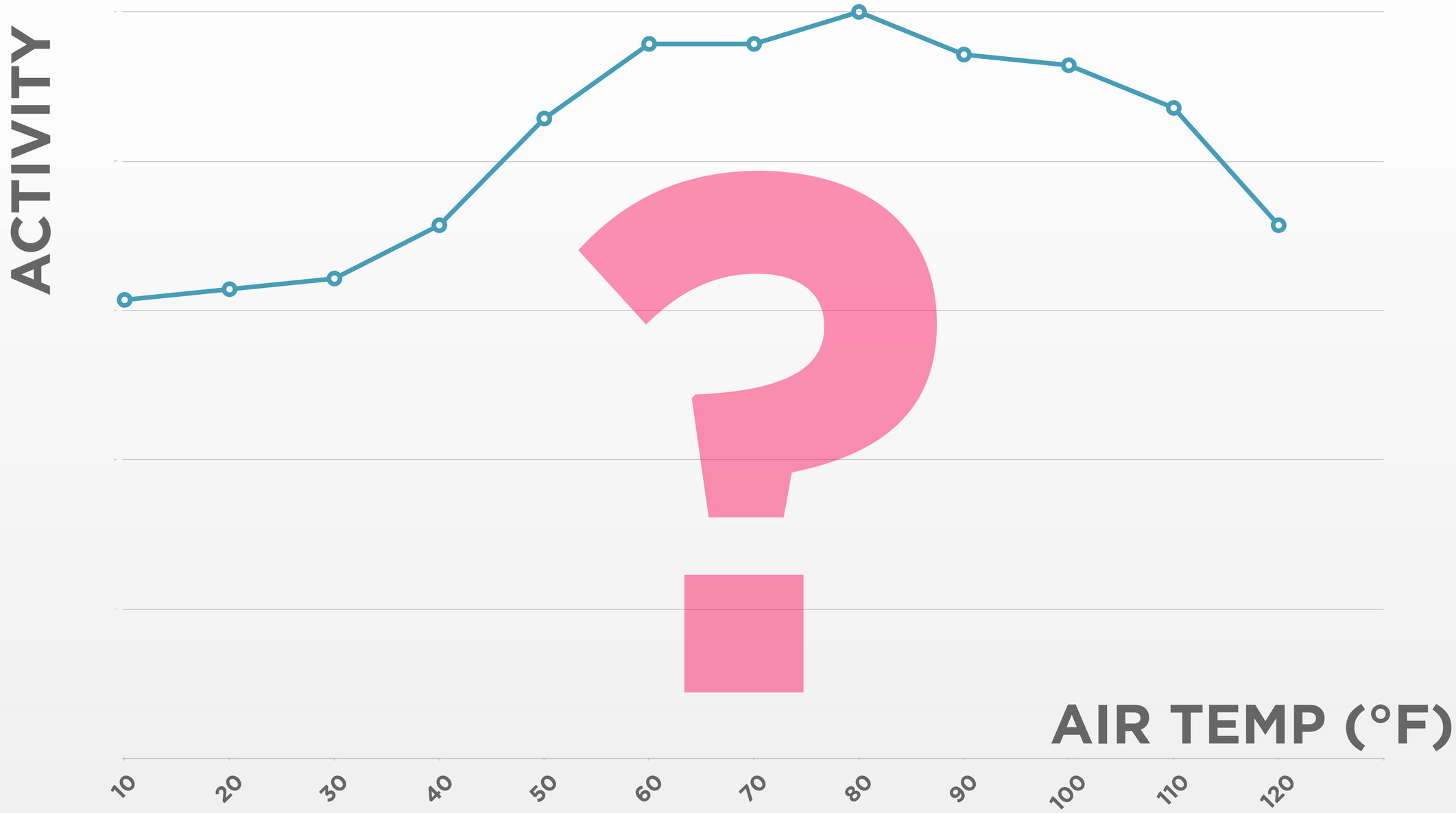
IMAGE SOURCE: HTTP://COMMONS.WIKIMEDIA.ORG/WIKI/FILE%3ABLIND_MEN_AND_ELEPHANT.PNG

DATA FUSION

POWERFUL BUT HARD
DATA IS NOISY
DOMAIN UNDERSTANDING IS KEY

LET'S TALK ABOUT THE WEATHER

MODEL THE PROBLEM


ACTIVITY
AIR TEMP (°F)
10
20
30
40
50
60
70
80
90
100
110
120

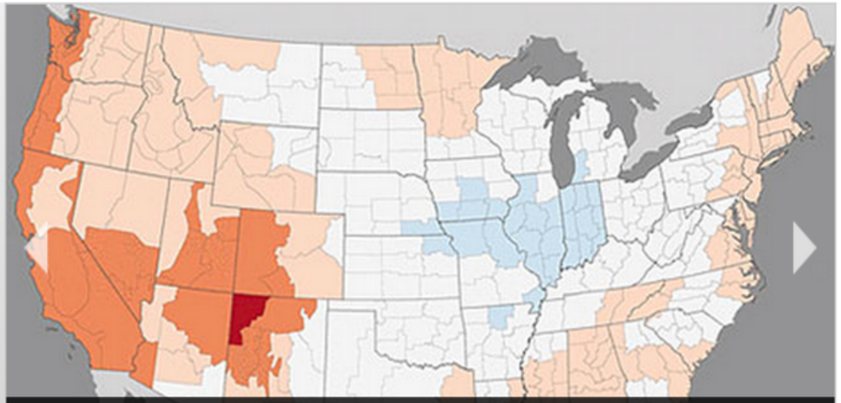
FIND THE DATA

NOAA NATIONAL CLIMATIC DATA CENTER
NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION

NOAA's National Climatic Data Center (NCDC) is responsible for preserving, monitoring, assessing, and providing public access to the Nation's treasure of climate and historical weather data and information.
Learn more about NCDC »

How may we assist you?

- I want to search for data at a particular location.
- I want quick access to your products.
- I want to see your monthly climate reports.
- I want to find a specific dataset.
- I want to know about climate change and variability.



NCDC Releases September 2014 U.S. Climate Report

The average temperature for the contiguous U.S. during September was 66.2°F, 1.3°F above the 20th century average.

1 2 3 4 5

Highlights

Upcoming Events, Products, and Services
View a complete listing of the upcoming products and services.

State of the Climate in 2013 Report Release
NCDC is announcing the release of the State of the Climate in 2013 report, an assessment of the

Newsroom

NCDC Insider: Meet Meteorologist, Mike Squires
As a meteorologist, Mike Squires develops new ways to look at climate data using geographical information systems and statistical analyses.

This Month in Climate History: Hurricane Ivan 2004
Ten years ago, on September 16, 2004, Hurricane Ivan slammed into the United States near Gulf Shores, Alabama.

State Annual and Seasonal Time Series
NCDC is announcing the release of a new tool that allows users to

NCDC Partners

Climate.gov
science & information for a climate-smart nation

weather.gov
THE NATIONAL WEATHER SERVICE

NIDIS drought.gov
U.S. Drought Portal

UNDERSTAND THE DATA

NOAA NATIONAL CLIMATIC DATA CENTER
NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION

HOURLY


```
0247010280999992014010100004+74517+019000FM-12+001899999V0200801N006019999999N030000199-00441-00891102181ADDAA106000091AY101021AY201021GA1021+999991061GE19MSL
+99999+99999GF102991021990999990991991MA199999010197 1MD1210021+9999MW1021OD139900901999REMSYN07601028 11780 20806 11044 21089 30197 40218 52002 69901 70200 82500 333
91109=EQDQ01  00002PRCP06
0187010280999992014010101004+74517+019000FM-12+001899999V0200801N006019999999N030000199-00441-00911102171ADDAY101021AY201021GA1011+999991061GE19MSL
+99999+99999GF101991011990999990991991MA1999990101961MD1710011+9999MW1011REMSYN06001028 41780 10806 11044 21091 30196 40217 57001 70100 81500=
0187010280999992014010102004+74517+019000FM-12+001899999V0200901N006019999999N030000199-00481-00951102151ADDAY101021AY201021GA1011+999991061GE19MSL
+99999+99999GF101991011990999990991991MA1999990101941MD1710031+9999MW1021REMSYN06001028 41780 10906 11048 21095 30194 40215 57003 70200 81500=
0187010280999992014010104004+74517+019000FM-12+001899999V0200901N007019999999N020000199-00441-00921102071ADDAY101021AY201021GA1021+999991061GE19MSL
+99999+99999GF102991021990999990991991MA1999990101861MD1810091+9999MW1031REMSYN06001028 41670 20907 11044 21092 30186 40207 58009 70300 82500=
0187010280999992014010105004+74517+019000FM-12+001899999V0201001N006019999999N020000199-00471-00921102071ADDAY101021AY201021GA1021+999991061GE19MSL
+99999+99999GF102991021990999990991991MA1999990101861MD1510081+9999MW1021REMSYN06001028 41670 21006 11047 21092 30186 40207 55008 70200 82500=
0283010280999992014010106004+74517+019000FM-12+001899999V0201001N007019999999N020000199-00421-00921102081ADDAA112000091AA224000431AY101021AY201021GA1021+999991061GE19MSL
+99999+99999GF102991021990999990991991KA1240N-00541MA1999990101871MD1510041+9999MW1021OD139900901999REMSYN08801028 11670 21007 11042 21092 30187 40208 55004 69902 70200 82500
333 21054 70004 91109=EQDQ01  00002PRCP12
0187010280999992014010107004+74517+019000FM-12+001899999V0201001N008019999999N020000199-00401-00921102041ADDAY101021AY201021GA1021+999991061GE19MSL
+99999+99999GF102991021990999990991991MA1999990101831MD1810031+9999MW1021REMSYN06001028 41670 21008 11040 21092 30183 40204 58003 70200 82500=
0187010280999992014010108004+74517+019000FM-12+001899999V0201001N007019999999N020000199-00371-00821102061ADDAY101021AY201021GA1031+999991061GE19MSL
+99999+99999GF103991031990999990991991MA1999990101851MD1510021+9999MW1031REMSYN06001028 41670 31007 11037 21082 30185 40206 55002 70300 83500=
0187010280999992014010109004+74517+019000FM-12+001899999V0201001N008019999999N020000199-00351-00831102061ADDAY101021AY201021GA1031+999991061GE19MSL
+99999+99999GF103991031990999990991991MA1999990101851MD1510011+9999MW1021REMSYN06001028 41670 31008 11035 21083 30185 40206 55001 70200 83500=
0187010280999992014010110004+74517+019000FM-12+001899999V0200901N008019999999N020000199-00381-00841102061ADDAY101021AY201021GA1021+999991061GE19MSL
+99999+99999GF102991021990999990991991MA1999990101851MD1210021+9999MW1021REMSYN06001028 41670 20908 11038 21084 30185 40206 52002 70200 82500=
0187010280999992014010111004+74517+019000FM-12+001899999V0201001N007019999999N020000199-00331-00791102041ADDAY101021AY201021GA1031+999991061GE19MSL
+99999+99999GF103991031990999990991991MA1999990101831MD1710021+9999MW1021REMSYN06001028 41670 31007 11033 21079 30183 40204 57002 70200 83500=
0248010280999992014010112004+74517+019000FM-12+001899999V0201001N008019999999N030000199-00361-00791102021ADDAA106000091AY101021AY201021GA1031+999991081GE19MSL
+99999+99999GF103991031990999990991991MA1999990101811MD1710041+9999MW1021OD139901201999SA1-0131REMSYN08801028 11780 31008 11036 21079 30181 40202 57004 60001 70200 83100
222// 01013 333 91112=
0187010280999992014010113004+74517+019000FM-12+001899999V0200901N009019999999N030000199-00321-00771102001ADDAY101021AY201021GA1031+999991081GE19MSL
+99999+99999GF103991031990999990991991MA1999990101791MD1710061+9999MW1021REMSYN06001028 41680 30909 11032 21077 30179 40200 57006 70200 83100=
0187010280999992014010114004+74517+019000FM-12+001899999V0200901N009019999999N030000199-00271-00651101961ADDAY101021AY201021GA1031+999991091GE19MSL
+99999+99999GF103991031990999990991991MA1999990101761MD1710071+9999MW1021REMSYN06001028 41680 30909 11027 21065 30176 40196 57007 70200 83300=
0187010280999992014010115004+74517+019000FM-12+001899999V0200901N008019999999N030000199-00301-00701101941ADDAY101021AY201021GA1031+999991081GE19MSL
+99999+99999GF103991031990999990991991MA1999990101741MD1710071+9999MW1021REMSYN06001028 41680 30908 11030 21070 30174 40194 57007 70200 83100=
0187010280999992014010116004+74517+019000FM-12+001899999V0200901N009019999999N030000199-00261-00641101891ADDAY101021AY201021GA1031+999991081GE19MSL
+99999+99999GF103991031990999990991991MA1999990101691MD1810111+9999MW1021REMSYN06001028 41680 30909 11026 21064 30169 40189 58011 70200 83100=
0187010280999992014010117004+74517+019000FM-12+001899999V0201001N010019999999N030000199-00221-00491101871ADDAY101021AY201021GA1021+999991061GE19MSL
+99999+99999GF102991021990999990991991MA1999990101671MD1710091+9999MW1021REMSYN06001028 41680 21010 11022 21049 30167 40187 57009 70200 82500=
0266010280999992014010118004+74517+019000FM-12+001899999V0200901N010019999999N030000199-00231-00451101851ADDAA112000091AY101021AY201021GA1021+999991081GE19MSL
+99999+99999GF102991021990999990991991KA1120M-00221MA1999990101651MD1710091+9999MW1021OD139901401999REMSYN08201028 11680 20910 11023 21045 30165 40185 57009 69902 70200 82100
333 11022 91114=EQDQ01  00002PRCP12
0187010280999992014010119004+74517+019000FM-12+001899999V0200801N009019999999N025000199-00211-00411101821ADDAY101021AY201021GA1021+999991081GE19MSL
+99999+99999GF102991021990999990991991MA1999990101621MD1710071+9999MW1021REMSYN06001028 41675 20809 11021 21041 30162 40182 57007 70200 82100=
0203010280999992014010120004+74517+019000FM-12+001899999V0201001N008019999999N020000199-00211-00441101831ADDAY101021AY201021GA1991+999991001GA2021+999991061GE19MSL
+99999+99999GF103991990990999990991991MA1999990101631MD1510041+9999MW1031REMSYN06001028 41670 31008 11021 21044 30163 40183 55004 70300 82502=
0187010280999992014010121004+74517+019000FM-12+001899999V0201001N008019999999N020000199-00181-00541101851ADDAY101021AY201021GA1021+999991061GE19MSL
+99999+99999GF102991021990999990991991MA1999990101651MD1510011+9999MW1011REMSYN06001028 41670 21008 11018 21054 30165 40185 55001 70100 82500=
0187010280999992014010122004+74517+019000FM-12+001899999V0200901N009019999999N020000199-00151-00351101821ADDAY101021AY201021GA1041+999991061GE19MSL
+99999+99999GF104991041990999990991991MA1999990101621MD1900001+9999MW1031REMSYN06001028 41670 40909 11015 21035 30162 40182 50000 70300 84500=
0187010280999992014010123004+74517+019000FM-12+001899999V0201201N008019999999N020000199-00231-00501101791ADDAY101021AY201021GA1031+999991061GE19MSL
+99999+99999GF103991031990999990991991MA1999990101591MD1810041+9999MW1011REMSYN06001028 41770 31208 11023 21050 30159 40179 58004 70100 83500=
0228010280999992014010200004+74517+019000FM-12+001899999V0201001N006019999999N020000199-00291-00451101781ADDAA106000131AY101021AY201021GA1041+999991081GE19MSL
+99999+99999GF104991041990999990991991MA1999990101581MD1710071+9999MW1261OD139901301999REMSYN07601028 11670 41006 11029 21045 30158 40178 57007 69911 72600 84800 333 91113=
0187010280999992014010201004+74517+019000FM-12+001899999V0201001N007019999999N020000199-00221-00441101751ADDAY111021AY201021GA1041+999991061GE19MSL
```


umn 1 Tab Siz

DAILY


```
US1FLSL0019,20130101,PRCP,0,,,N,
US1FLSL0019,20130101,SNOW,0,,,N,
US1TXTV0133,20130101,PRCP,30,,,N,
USC00178998,20130101,TMAX,-22,,,7,0700
USC00178998,20130101,TMIN,-117,,,7,0700
USC00178998,20130101,TOBS,-28,,,7,0700
USC00178998,20130101,PRCP,0,T,,7,0700
USC00178998,20130101,SNOW,0,T,,7,
USC00178998,20130101,SNWD,0,,,7,
USC00242347,20130101,TMAX,6,,,7,0800
USC00242347,20130101,TMIN,-139,,,7,0800
USC00242347,20130101,TOBS,6,,,7,0800
USC00242347,20130101,PRCP,0,,,7,0800
USC00242347,20130101,SNOW,0,,,7,
USC00242347,20130101,SNWD,76,,,7,
NOE00133566,20130101,TMAX,62,,,E,
NOE00133566,20130101,TMIN,9,,,E,
NOE00133566,20130101,PRCP,193,,,E,
NOE00133566,20130101,SNWD,0,,,E,
USC00141761,20130101,TMAX,-50,,,7,0700
USC00141761,20130101,TMIN,-100,,,7,0700
USC00141761,20130101,TOBS,-100,,,7,0700
USC00141761,20130101,PRCP,135,,,7,0700
USC00141761,20130101,SNOW,170,,,7,
USC00141761,20130101,SNWD,178,,,7,0700
```

DATA GENERATION PROCESS

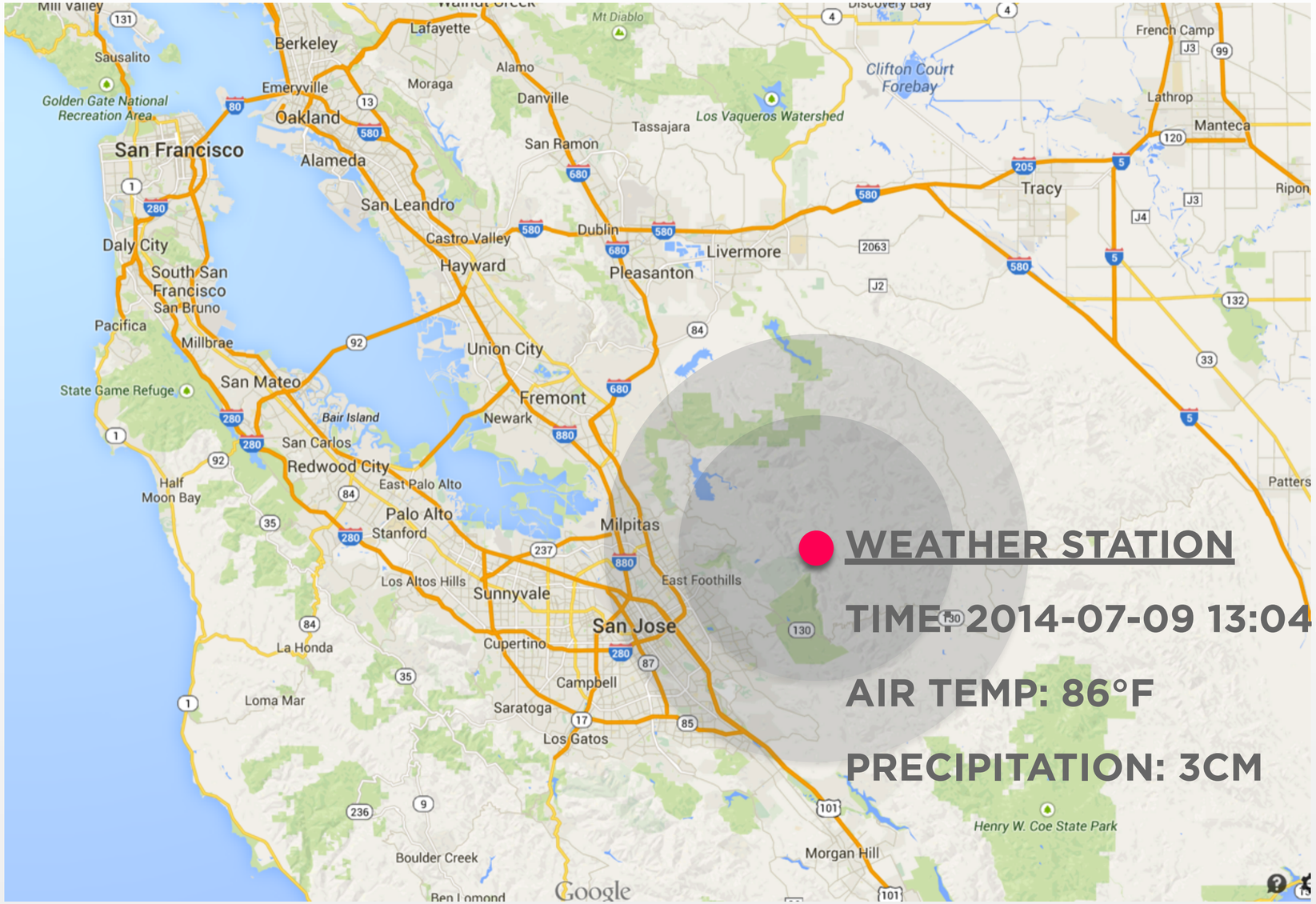
NETWORK OF WEATHER STATIONS

FREQUENCY OF MEASUREMENTS - HOURLY TO DAILY

COLLABORATION WITH INTERNATIONAL AGENCIES

AGGREGATION AND QA BY NCDC

UNDERSTAND THE DOMAIN



QA THE DATA

BUT ISN'T IT DONE?

FEDERAL CLIMATE COMPLEX

DATA DOCUMENTATION

FOR

INTEGRATED SURFACE DATA

September 4, 2014

National Climatic Data Center
14th Weather Squadron
Fleet Numerical Meteorology and Oceanography Detachment
151 Patton Avenue
Asheville, NC 28801-5001 USA

```
0133999999940822013010103101+40245-108968CRN05+1848999
```

POS: 93-93

AIR-TEMPERATURE-OBSERVATION air temperature quality code

The code that denotes a quality status of an AIR-TEMPERATURE-OBSERVATION.

DOM: A specific domain comprised of the characters in the ASCII character set.

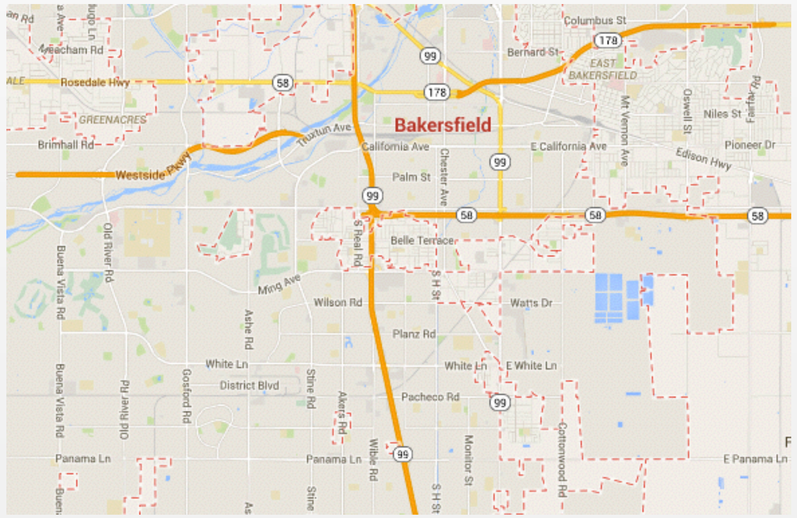
- 0 = Passed gross limits check
- 1 = Passed all quality control checks
- 2 = Suspect
- 3 = Erroneous
- 4 = Passed gross limits check , data originate from an NCDC data source
- 5 = Passed all quality control checks, data originate from an NCDC data source
- 6 = Suspect, data originate from an NCDC data source
- 7 = Erroneous, data originate from an NCDC data source
- 9 = Passed gross limits check if element is present
- A = Data value flagged as suspect, but accepted as a good value
- C = Temperature and dew point received from Automated Weather Observing System (AWOS) are reported in whole degrees Celsius. Automated QC flags these values, but they are accepted as valid.
- I = Data value not originally in data, but inserted by validator
- M = Manual changes made to value based on information provided by NWS or FAA
- P = Data value not originally flagged as suspect, but replaced by validator
- R = Data value replaced with value computed by NCDC software
- U = Data value replaced with edited value

...MAYBE NOT!

AIR TEMP:
105°F

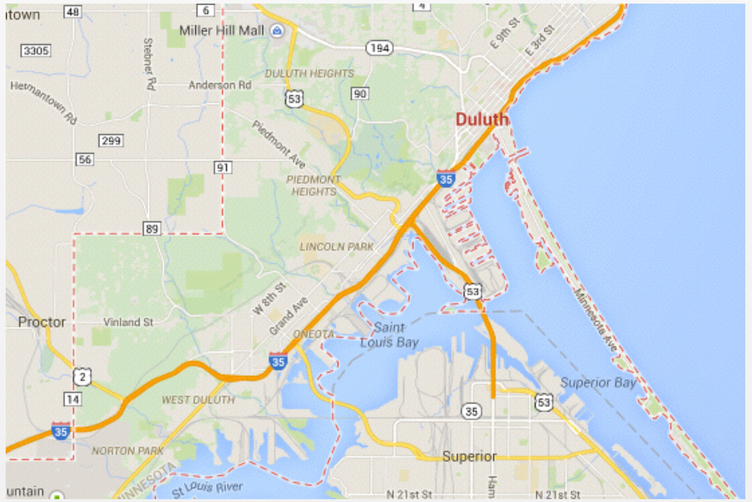
BAKERSFIELD, CA

JULY 17, 15:00



DULUTH, MN

JAN 12, 05:00



DATA VALIDATION

DOMAIN KNOWLEDGE

COMPARE MULTIPLE SOURCES - E.G. CLIMATE

MANUAL REVIEW OF FLAGGED DATA POINTS

JOIN

HOW?

DOMAIN SPECIFIC

WEATHER STATION A

LAT: 39.36

LON: -74.45

TIME: 2014-07-09 13:04:00

AIR TEMP: 74°F

ELEVATION: 30FT

WEATHER STATION B

LAT: 39.35

LON: -74.44

TIME: 2014-07-09 13:00:00

AIR TEMP: 60°F

ELEVATION: 120FT

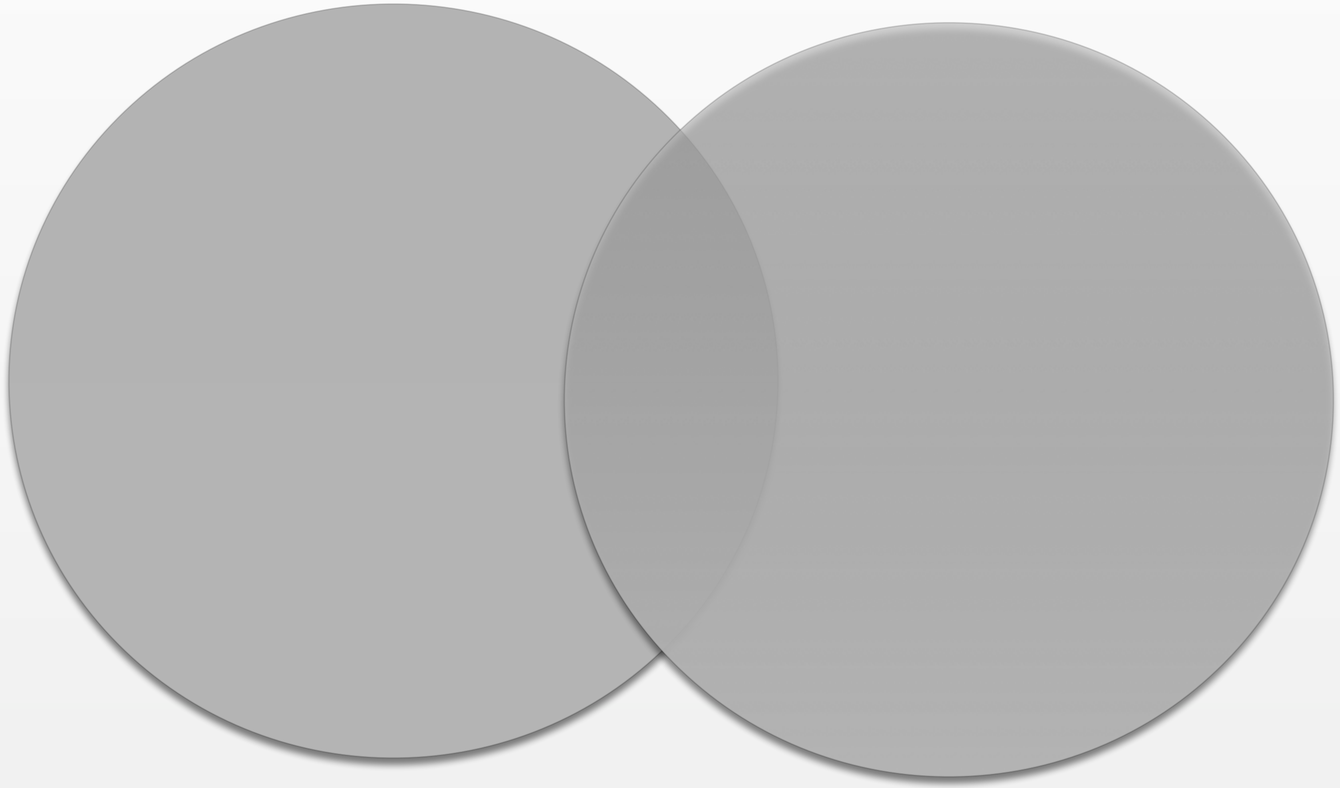
COVERAGE

DO THE DATASETS INTERSECT ENOUGH?

PLACES

TIMES

USERS



ISOLATE THE EFFECT

CONFOUNDING VARIABLES

WHAT ELSE AFFECTS ACTIVITY?

WEEKDAYS/WEEKENDS

DAYLIGHT

RAIN/SNOW

REDSHIFT VS SPARK

AMAZON REDSHIFT

RELATIONAL ANALYTICAL DATABASE BY AMAZON

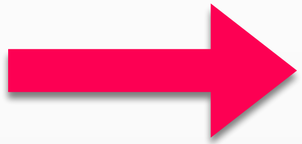
COMPLEX QUERIES ON LARGE DATASETS IN SECONDS

SQL INTERFACE (POSTGRES)

MANAGED CLUSTER

EXAMPLE: DAYLIGHT

PYTHON



REDSHIFT

```python
def compute_sunrise_utc(lat, lon, date):
    o=ephem.Observer()
    o.lat=lat
    o.long=lon
    o.date = date
    s=ephem.Sun()
    s.compute()
    return o.previous_rising(s).datetime()

def compute_sunset_utc(lat, lon, date):
    o=ephem.Observer()
    o.lat=lat
    o.long=lon
    o.date = date
    s=ephem.Sun()
    s.compute()
    return o.next_setting(s).datetime()
```

uszipcode	date	sunrise_localized	sunset_localized
94014	2013-04-09	2013-04-09 06:42:33.337692	2013-04-09 19:40:28.568333
94014	2013-04-21	2013-04-21 06:25:55.117312	2013-04-21 19:51:26.171288
94014	2013-05-03	2013-05-03 06:11:18.985279	2013-05-03 20:02:24.563496
94014	2013-05-15	2013-05-15 05:59:37.767078	2013-05-15 20:13:05.965446
94014	2013-05-27	2013-05-27 05:51:33.742782	2013-05-27 20:22:44.060134
94014	2013-06-08	2013-06-08 05:47:39.783952	2013-06-08 20:30:21.368739
94014	2013-06-20	2013-06-20 05:48:01.497239	2013-06-20 20:34:51.956094
94014	2013-07-02	2013-07-02 05:52:13.523309	2013-07-02 20:35:23.766219
94014	2013-07-14	2013-07-14 05:59:27.116127	2013-07-14 20:31:35.201628
94014	2013-07-26	2013-07-26 06:08:34.284014	2013-07-26 20:23:30.722577
94014	2013-08-07	2013-08-07 06:18:38.73022	2013-08-07 20:11:43.878524
94014	2013-08-19	2013-08-19 06:28:56.422508	2013-08-19 19:56:58.625318
94014	2013-08-31	2013-08-31 06:39:06.747193	2013-08-31 19:40:05.091958
94014	2013-09-12	2013-09-12 06:49:10.766203	2013-09-12 19:21:55.583515
94014	2013-09-24	2013-09-24 06:59:17.295704	2013-09-24 19:03:18.416829
94014	2013-10-06	2013-10-06 07:09:48.461751	2013-10-06 18:45:06.036396
94014	2013-10-18	2013-10-18 07:20:59.246095	2013-10-18 18:28:09.03394
94014	2013-10-30	2013-10-30 07:32:57.98174	2013-10-30 18:13:22.558828
94014	2013-11-11	2013-11-11 06:45:35.548386	2013-11-11 17:01:44.527159
94014	2013-11-23	2013-11-23 06:58:11.695529	2013-11-23 16:54:07.402144
94014	2013-12-05	2013-12-05 07:09:44.65474	2013-12-05 16:51:16.437288
94014	2013-12-17	2013-12-17 07:18:47.818248	2013-12-17 16:53:24.766744

IN-MEMORY DATA PROCESSING FRAMEWORK

MODELS COMPUTATION AS A GRAPH OF RDDS (RESILIENT DISTRIBUTED DATASETS)

FUNCTIONAL PROGRAMMING MODEL (SCALA, PYTHON)

SQL

CAN READ FROM SAME SOURCES AS HADOOP

EXAMPLE: DAYLIGHT

SPARK

```
> import ephem
  from datetime import datetime, timedelta
  from pytz import timezone

  #Returns true if the sun was up during the whole hour
  def is_daylight_hour(lat, lon, datetime_with_tz):
    utc_datetime = datetime_with_tz.astimezone(timezone('UTC'))
    o=ephem.Observer()
    o.lat=str(lat)
    o.long=str(lon)
    o.date = utc_datetime.strftime("%Y-%m-%d")
    s=ephem.Sun()
    s.compute()
    prev_sunrise_utc_datetime = o.previous_rising(s).datetime()
    next_sunset_utc_datetime = o.next_setting(s).datetime()
    return prev_sunrise_utc_datetime.hour < utc_datetime.hour and next_sunset_utc_datetime.hour > utc_datetime.hour
```

```
df_hourly_steps_by_temp_us_weekends = [[hour, \\
  the_master_set_steps_location_air_temp.filter(lambda record: record[0]['hour'] == hour \\
                                                and is_daylight_hour(record[0]['lat'], record[0]['lon'], record[1]['local_datetime']) \\
                                                and (record[0]['date'].weekday() == 5 or record[0]['date'].weekday() == 6))\\
  .groupBy(lambda record: int(round(float(record[1]['air_temp'])/10 * 9 / 5 + 32))) \\
  .map(compute_mean).collect()] for hour in range(5, 21)]
```

SILVER BULLET?

PICK YOUR OWN ADVENTURE

SPARK

PROGRAMMER-FRIENDLY

END-TO-END SOLUTION

SELF-DOCUMENTING

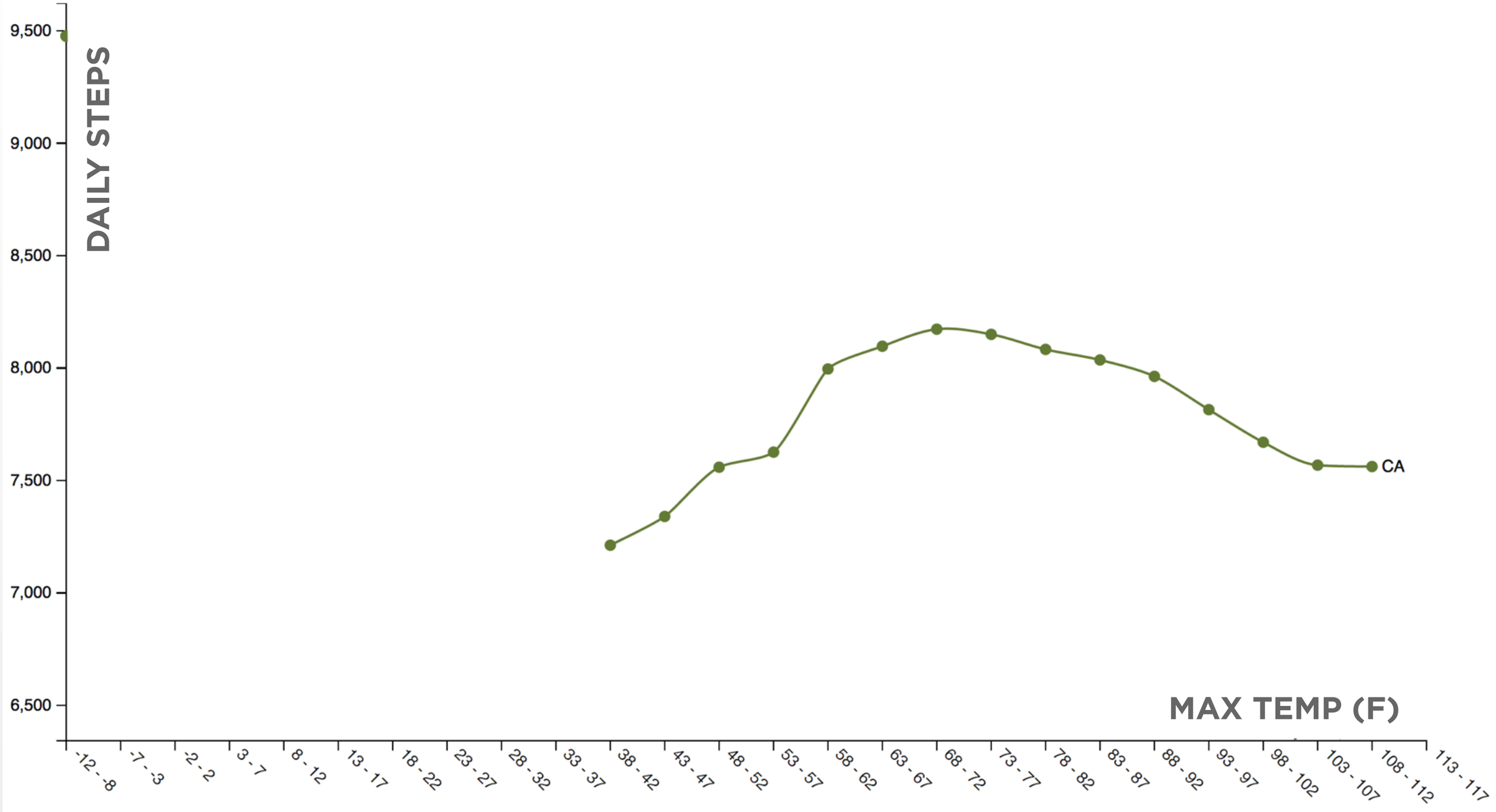
REDSHIFT

EASY TO SHARE DATA WITH NON-DEVELOPERS

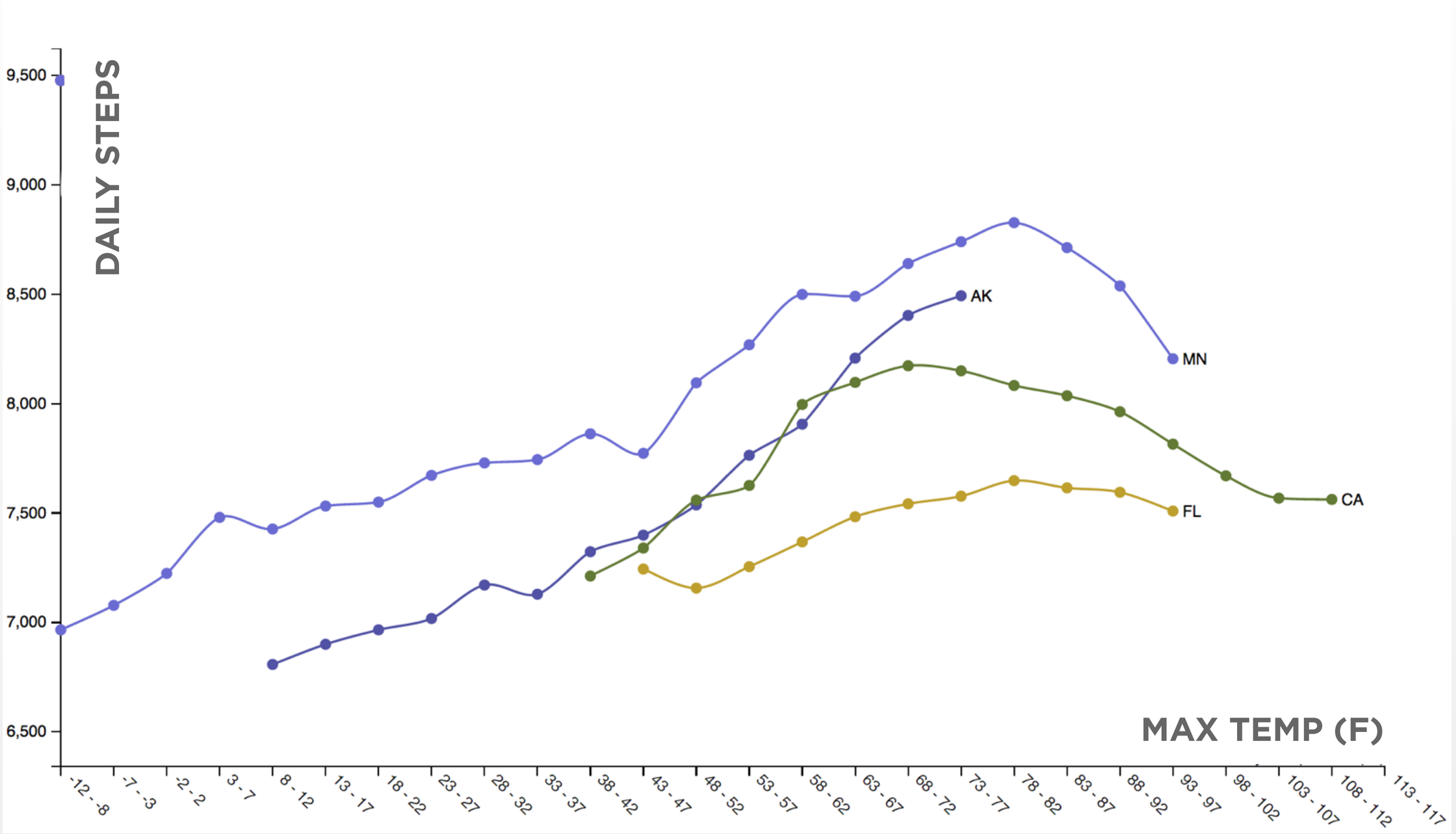
MANAGED - EASY SCALING

WHAT DID WE FIND?

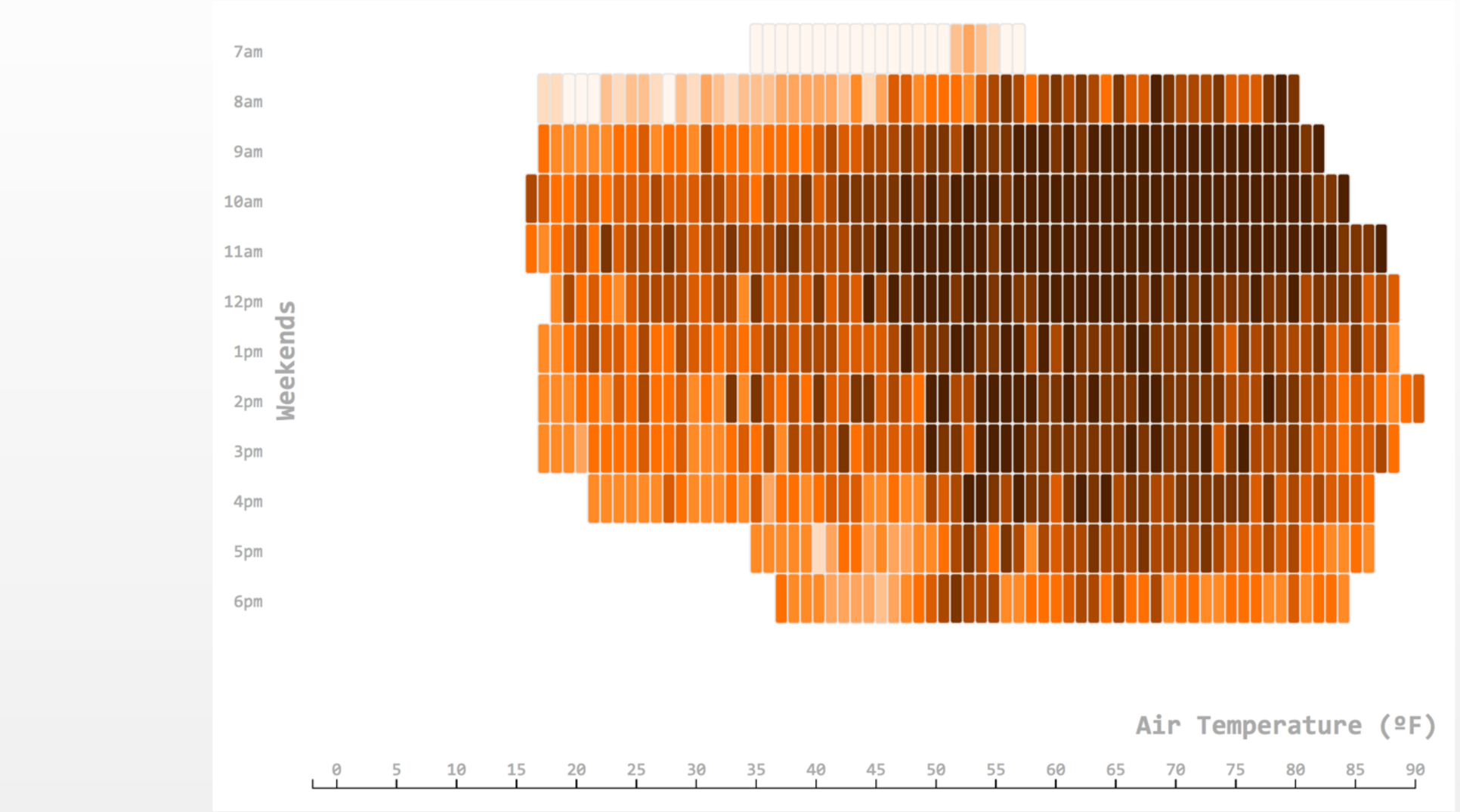
IDEAL TEMP FOR MOVEMENT



AND NOW BY STATE...

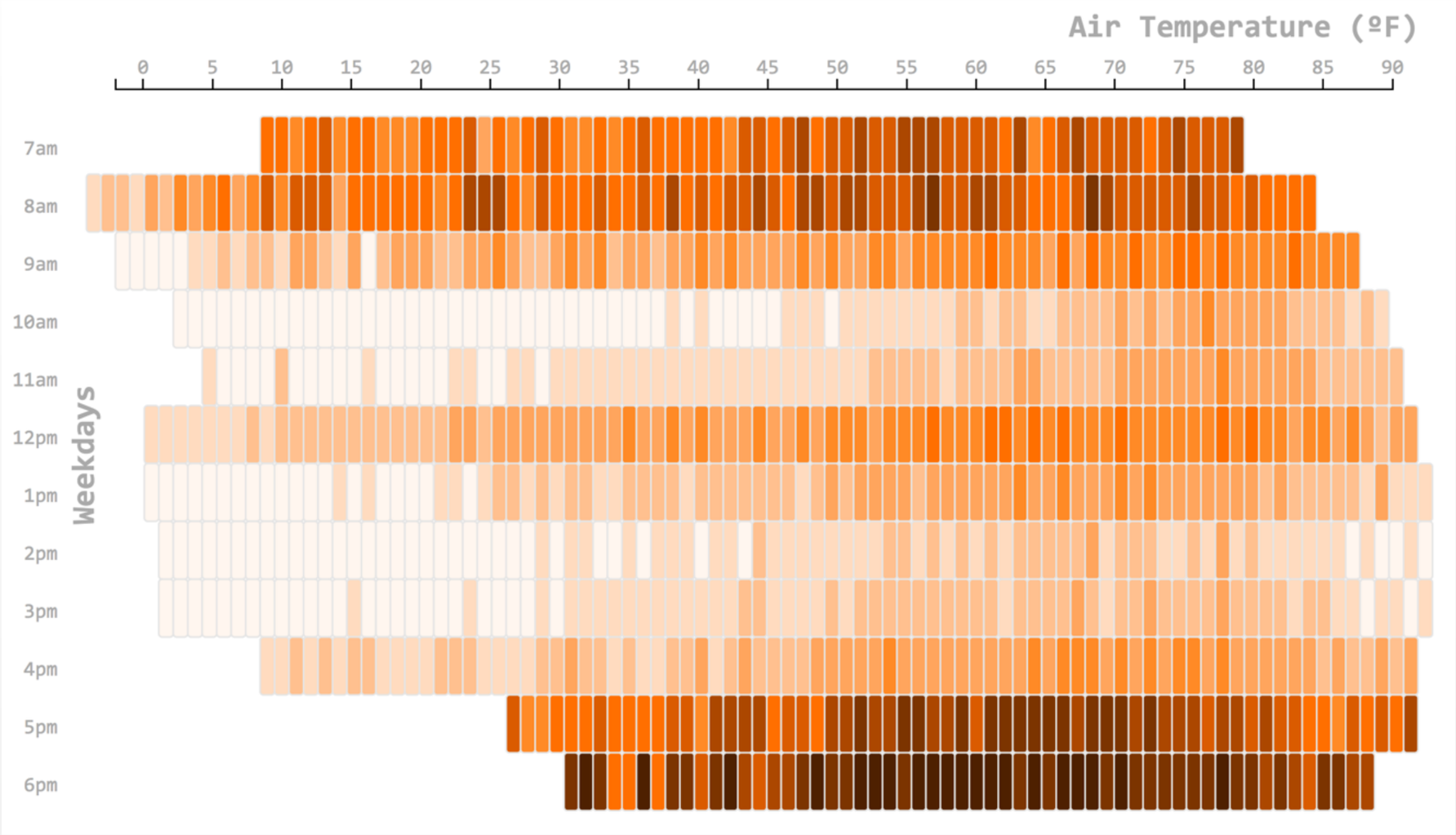


HOURLY STEPS BY AIR TEMP



WEEKENDS

LESS CHOICE = SMALLER EFFECT



WEEKDAYS

DATA FUSION

POWERFUL BUT HARD
DATA IS NOISY
DOMAIN UNDERSTANDING IS KEY

THANK YOU!

@EUGMANDEL

WWW.LINKEDIN.COM/IN/EUGENEMANDEL