Bringing the Anna Karenina Principle to Data Quality Monitoring
A Multi-Tiered Approach

Andy Walsh, Health Monitoring

June 14, 2019

The Anna Karenina Principle

"All happy families are alike;
each unhappy family is unhappy in its own way."

– Leo Tolstoy, *Anna Karenina*

Not All Data Quality Problems are Identical

- Are We Getting Anything?
  - Connection Disruption
  - Broken Interface
- Are We Getting Enough for Surveillance?
  - Interface Upgrade
  - Reorganization
- Are We Getting Everything We Expect?
  - Process Change
  - Interface Tweak

Guiding Considerations

- Cross products get big quick
  - 600 facilities x 11 analyses x 24 hours = Frenzied staff
- Different issues require different responses
  - If there is no action to take, don't alert
  - Incorporate decision logic into monitoring
- Customize to service level
  - Don't check for race if a hospital never sends it
  - Don't check for inpatient admissions from an urgent care

How Long Do We Wait?

Geometric Distribution

How many failures before a success?

$$Pr(X = k) = p(1 - p)^k$$

k failures
p probability of success

Geometric Distribution

How many minutes without a message?

$$Pr(X = k) = p(1 - p)^k$$

k minutes with no messages
p probability of any message in a minute

Geometric Distribution

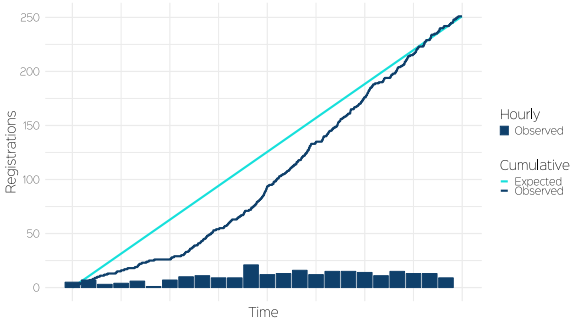
How many minutes without a message?

$$Pr(X = k) = p(1 - p)^k$$

k minutes with no messages
p probability of any message in a minute
But p varies throughout the day!

“Continuous” data feeds are not so continuous

- Thirty minutes w/o messages from an urban hospital during peak hours is unusual
- Twelve hours w/o messages from a rural hospital overnight is not unusual



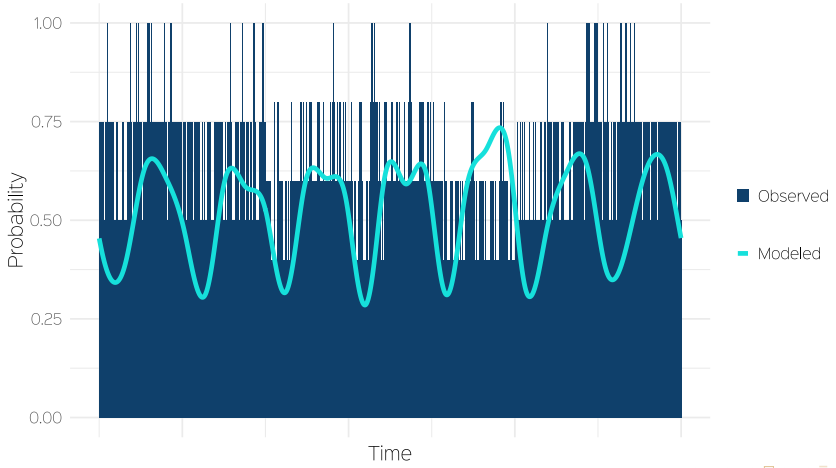
Model Probability of Receiving Messages

Nonparametric regression

```
prob.model <- mgcv::gam(messages ~ s(t, bs = "cc", k = 30),
family = binomial, data = messages.binary)
```

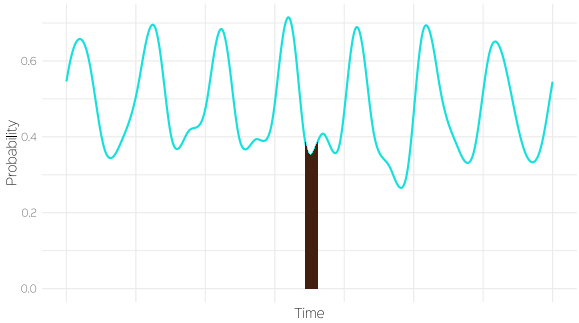
$$P_{\text{wait}} = \prod_{t_{\text{last message}}}^{t_{\text{now}}} 1 - p(t)$$

Model Probability of Receiving Messages

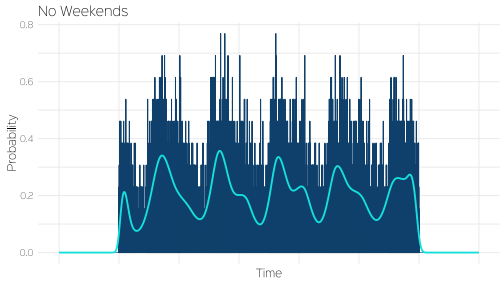


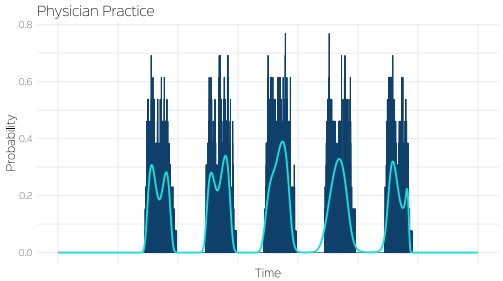
Model Probability of Receiving Messages

If last message was received at 10:32am and it is now 2:45pm, calculate the probability of waiting that long as the product of the probability of not receiving a message each intervening minute

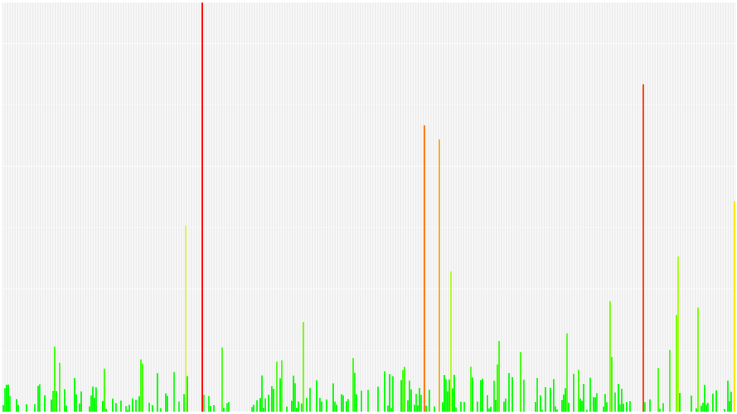


Model Can Accomodate Many Sending Patterns

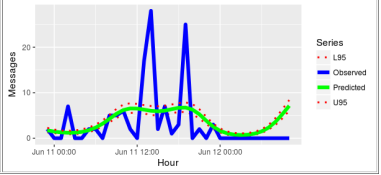




Dashboard



Email

Outage ID	Aspect	Target (Sender)	Facility Type	Facility Name	Outage Start	Outage Close (If Available)	Outage Plot
6345	message arrival	AN_Hospital	Hospital	Anonymous General	06/11/2019 23:17:12		

Sent after outage persists for multiple monitoring intervals

Is a facility sending a consistent number of

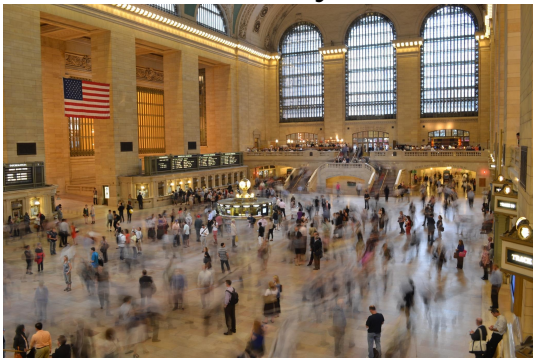
- ED Registrations
- Discharge Diagnosis Codes
- Clinical Notes

with sufficient elements to create a record?

Consistently Inconsistent

Sunday

Monday



Model Counts by Day of Week

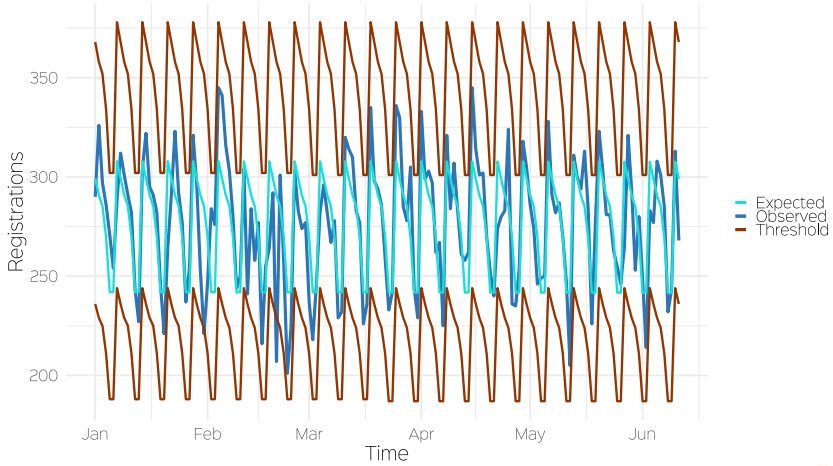
If $\sigma^2 > \mu$, negative binomial regression

```
count.model <- MASS::glm.nb(admits ~ dow + trend,
data = admit.counts)
```

Else Poisson regression

```
count.model <- glm(admits ~ dow + trend, data = admit.counts,
family = poisson)
```

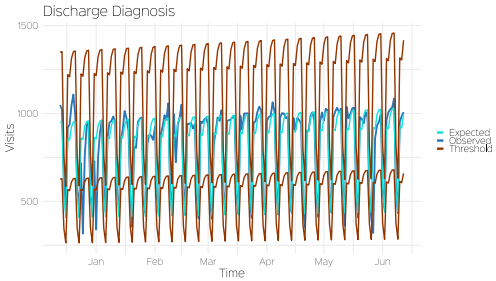
Model Counts by Day of Week

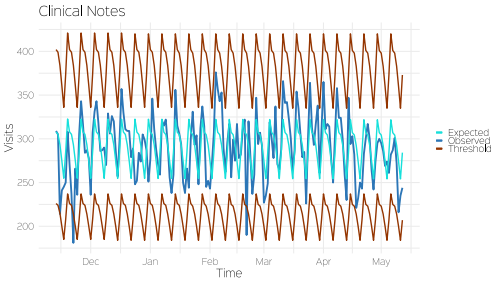


Report on Issues

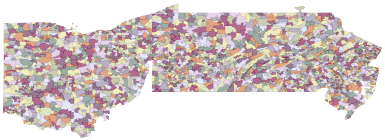
Facility	Aspect	Mean	Current	P	Status
884	ED Registration	114	75	1e-02	Lowest Total
3154	ED Registration	37	0	0e+00	Too Low
11826	ED Registration	12	0	0e+00	Too Low
4562	ED Registration	25	0	0e+00	Too Low
880	ED Registration	144	73	7e-06	Too Low
1120	ED Registration	6	0	1e-02	Zero Count

Monitoring Multiple Data Types

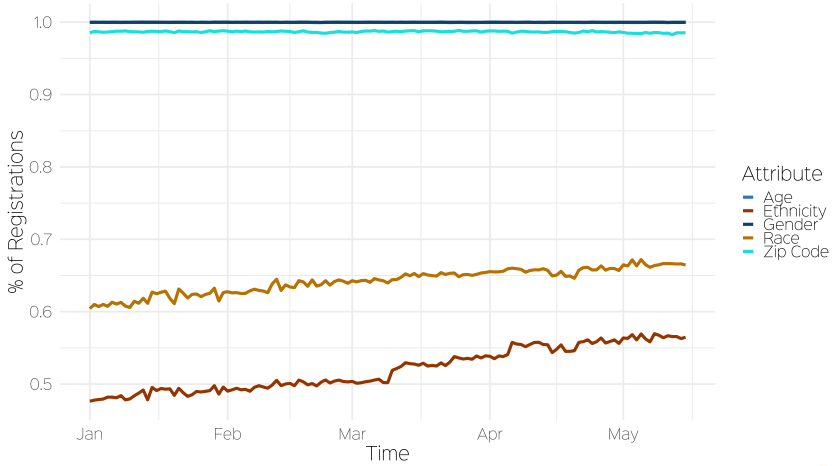




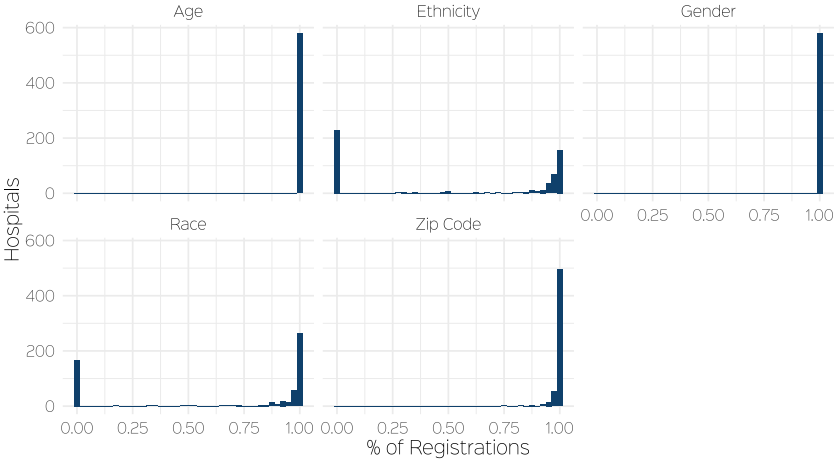
Tell Me About Yourself



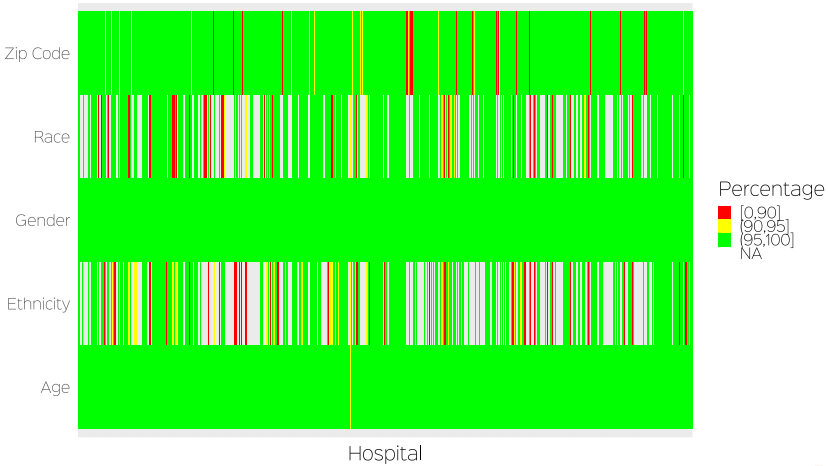
Consistency of Demographics



Facility Differences



Demographics Review



Discussion

- Created distinct monitoring techniques for different potential data quality issues
- Implemented monitoring to minimize need for human review
- TODO: Review performance in real usage
- TODO: Expand to batch senders
- TODO: Validate specific elements (e.g. diagnosis codes)

Questions?

andy.walsh@hmsinc.com