Northwestern

Public Health Informatics

Nicholas Soulakis, PhD
Northwestern University
Feinberg School of Medicine
Department of Preventive Medicine
Division of Health and Biomedical Informatics

About Me

Training

- **PhD**, Epidemiology
  @University of Pittsburgh Graduate School of Public Health

Business as Usual:

- **Assistant Professor**, Department of Preventive Medicine, Division(s) of Biomedical Informatics & Epidemiology @Northwestern Feinberg School of Medicine
- **Director, Data Science Services** @NUCATS Center for Data Science & Informatics

These Days:

- **Chief Public Health Informatics Officer and Deputy Commissioner, Bureau of Informatics and IT** @Chicago Department of Public Health

Objectives

- Understand and discuss public health informatics
- Understand public health surveillance
- Identify outbreaks in time-series data
- Discuss clustering, classification, and outbreak detection approaches

Informatics

What is Informatics?

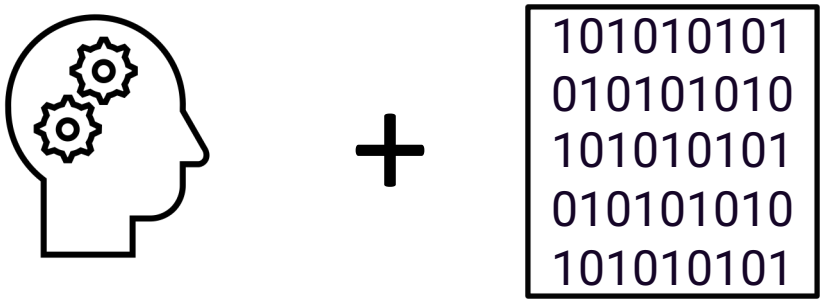
The science of how we collect, organize and use information.

The kind of informatics depends on the domain user.

Justin B. Starren, MD PhD. Personal Communication

A Single Unit of Informatics

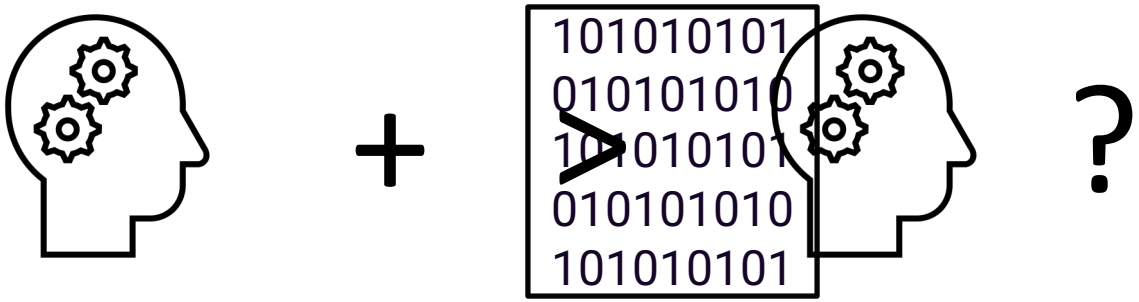
People Using Information



From: Friedman CP. A “fundamental theorem” of biomedical informatics. Journal of the American Medical Informatics Association. 2009 Mar 1;16(2):169-70.

Asking a Good Informatics Question

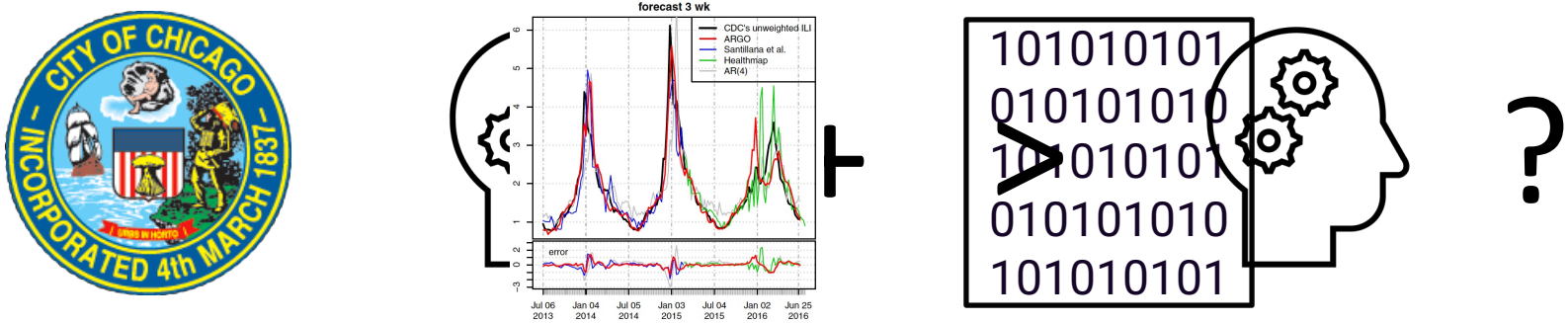
Is a user, working in concert with an information resource, measurably better at achieving their objectives than without the resource?



From: Friedman CP. A “fundamental theorem” of biomedical informatics. Journal of the American Medical Informatics Association. 2009 Mar 1;16(2):169-70.

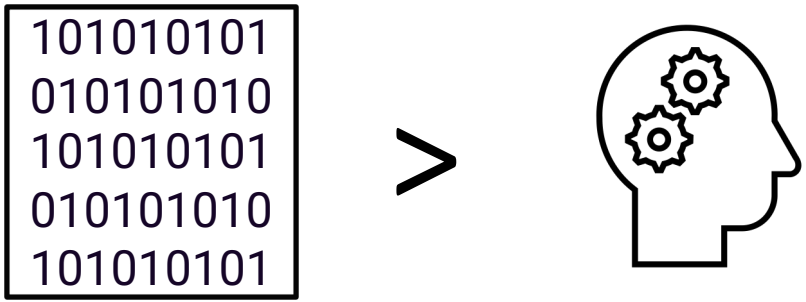
Information for Action

Public Health Information for Public Health Actions



From: Friedman CP. A “fundamental theorem” of biomedical informatics. Journal of the American Medical Informatics Association. 2009 Mar 1;16(2):169-70.

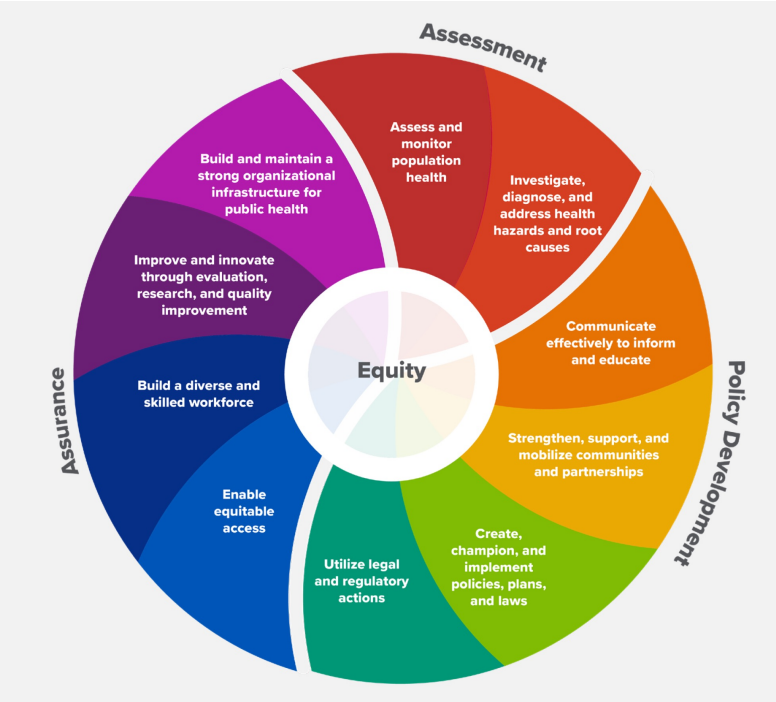
Informatics – What it isn't



From: Friedman CP. A "fundamental theorem" of biomedical informatics. Journal of the American Medical Informatics Association. 2009 Mar 1;16(2):169-70.

Public Health

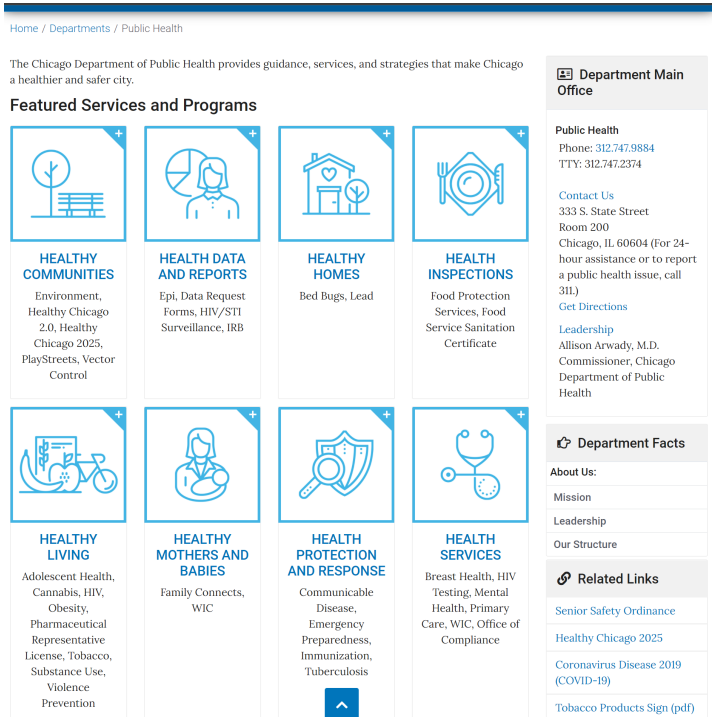
Essential Public Health Services



Framework to protect and promote the health of all people in all communities.

CDC. 10 Essential Public Health Services. https://www.cdc.gov/publichealthgateway/publichealthservices/essentialhealthservices.html

Chicago Public Health

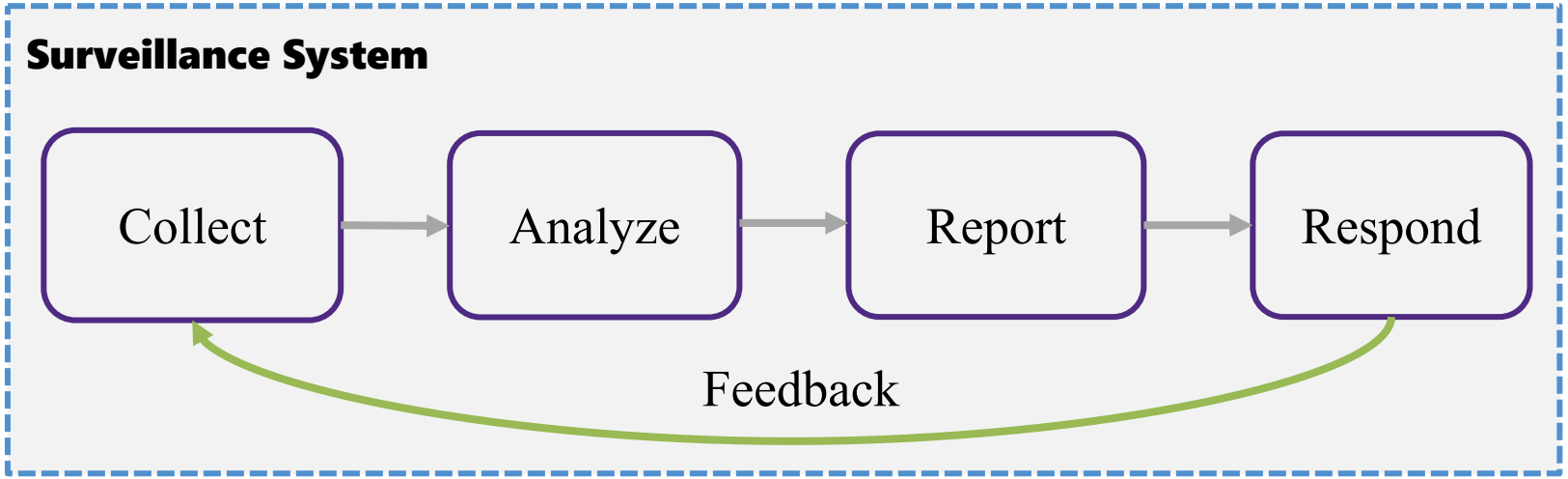


https://www.chicago.gov/city/en/depts/cdph.html

Public Health Surveillance

Public Health Workflows

Traditional public health practice use of information reflects many informatics use cases.



Thacker, Stephen B., and Ruth L. Berkelman. "Public health surveillance in the United States." Epidemiologic reviews 10.1 (1988): 164-190.

Clusters

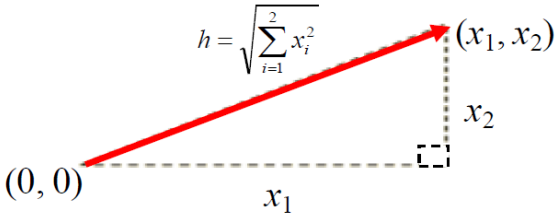
Measuring Distances

PROC DISTANCE METHOD=EUCLID PROC DISTANCE METHOD=CITYBLOCK

Euclidean Distance

$$D_E = \|\mathbf{x} - \mathbf{w}\| = \sqrt{\sum_{i=1}^{k} (x_i - w_i)^2}$$

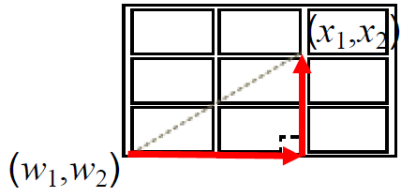
- Euclidean distance gives the linear distance between any two points in *n*-dimensional space.
- It is a generalization of the Pythagorean theorem.



City Block Distance

$$D_{M_1} = \sum_{i=1}^{d} |x_i - w_i|$$

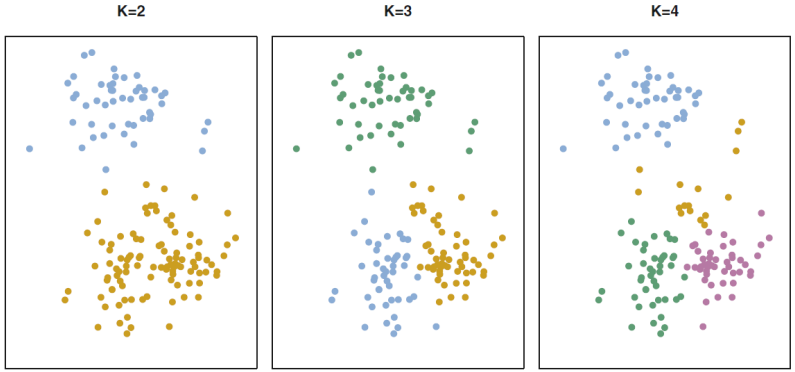
- The distance between two points is measured along the sides of a right triangle.
- It is the distance that you would travel if you had to walk along the streets of a right-angled city.



https://documentation.sas.com/?docsetId=statug&docsetTarget=statug_introclus_sect004.htm&docsetVersion=14.3&locale=en

K-means Clustering

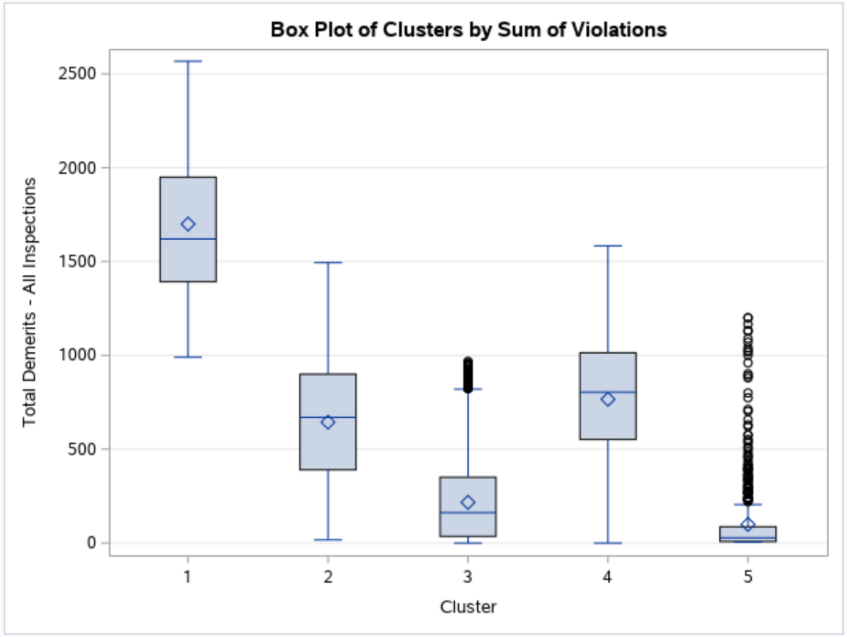
- “Partive” clustering
- Specify number of clusters
- K-means algorithm assigns exactly each observation to exactly one cluster

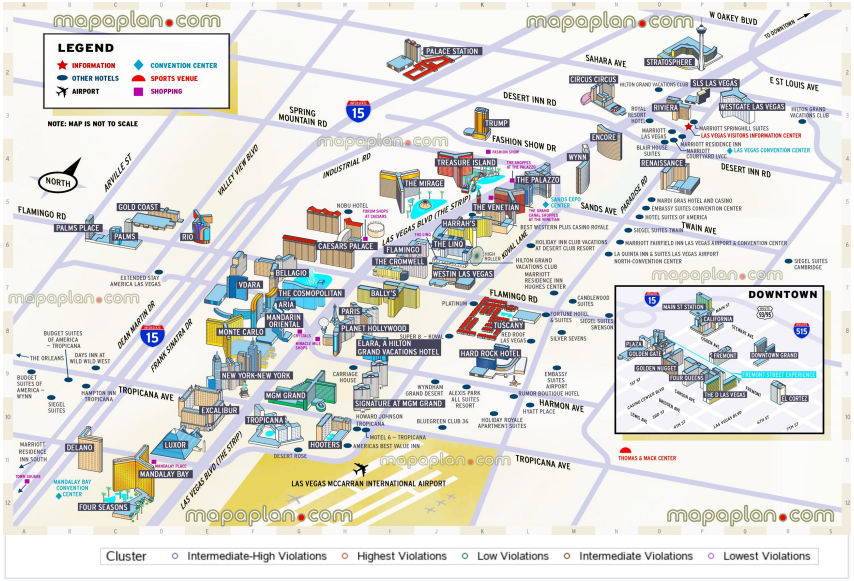


May be the only practical choice when the data set is large.

James, Gareth, et al. An introduction to statistical learning. Vol. 112. New York: springer, 2013.

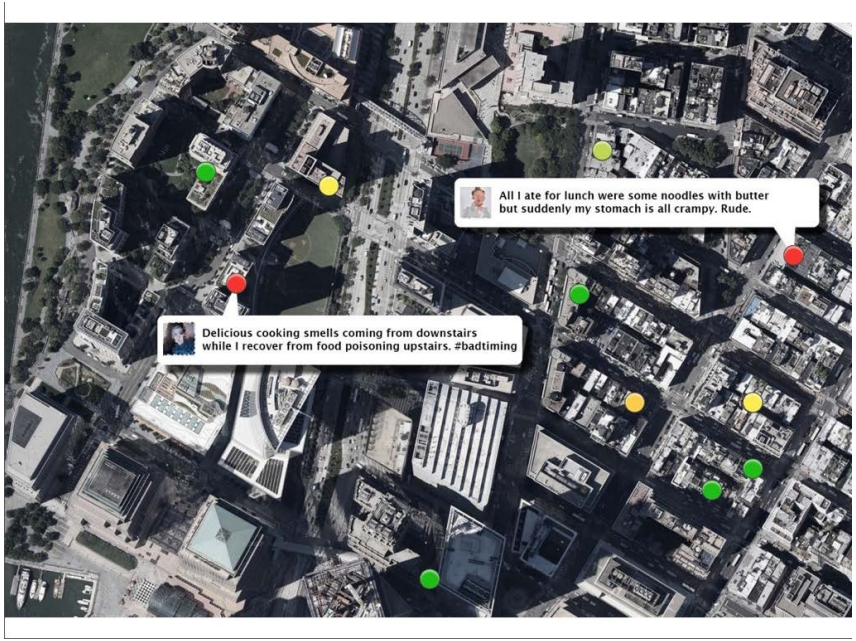
Plot by Coordinates

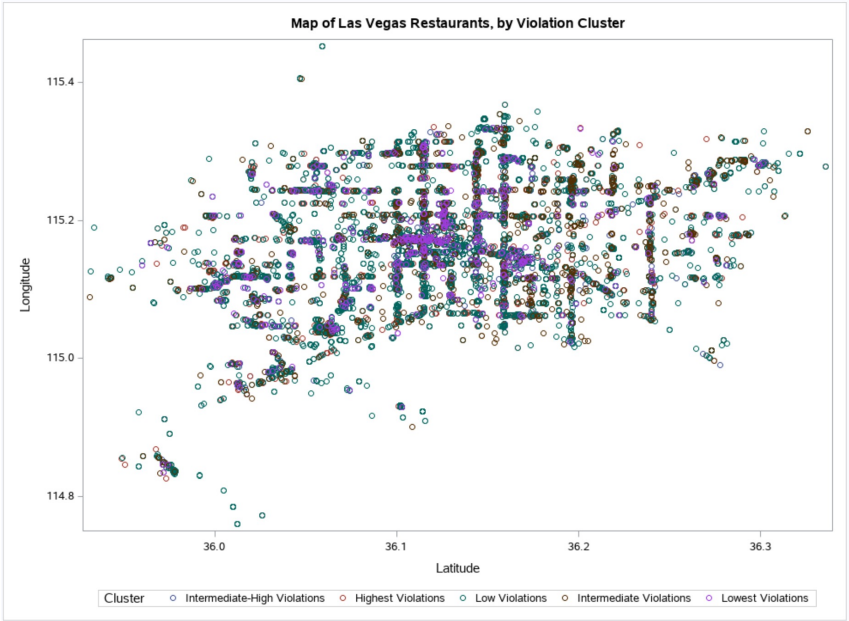




See: Course shared files > Clustering > sas > lv_inspections_clustering_190522_1434.sas

Refined Social Listening





See: Course shared files > Clustering > sas > lv_inspections_clustering_190522_1434.sas

Classification

Surveillance Case Definitions

A surveillance case definition is a set of uniform criteria used to define a disease for public health surveillance. Surveillance case definitions:

- Enable public health officials to classify and count cases consistently across reporting jurisdictions.

  Not intended to be used by healthcare providers for making a clinical diagnosis or determining how to meet an individual patient's health needs.
- The Council of State and Territorial Epidemiologists recommends that state health departments report cases of selected diseases to CDC's National Notifiable Diseases Surveillance System (NNDSS).
- CSTE issues Position Statements, providing uniform criteria of national notifiable infectious and non-infectious conditions for reporting purposes.

CDC. "Surveillance Case Definitions for Current and Historical Conditions." https://ndc.services.cdc.gov/

Terms Used in Case Classification

- **Confirmed case:** a case that is classified as confirmed for reporting purposes.
- **Probable case:** a case that is classified as probable for reporting purposes.
- **Suspected case:** a case that is classified as suspected for reporting purposes.

From: CDC. Case definitions for infectious conditions under public health surveillance. MMWR

Terms Used in Case Classification

- **Clinically compatible case:** a clinical syndrome generally compatible with the disease, as described in the clinical description.
- **Epidemiologically linked case:** a case in which a) the patient has had contact with one or more persons who either have/had the disease or have been exposed to a point source of infection (i.e., a single source of infection, foodborne-disease outbreak) and b) transmission of the agent by the usual modes of transmission is plausible. A case may be considered epidemiologically linked to a laboratory confirmed case if at least one case in the chain of transmission is laboratory confirmed.

From: CDC. Case definitions for infectious conditions under public health surveillance. MMWR

Terms Used in Case Classification

- **Laboratory-confirmed case:** a case that is confirmed by one or more of the laboratory methods listed in the case definition under Laboratory Criteria for Diagnosis. Although other laboratory methods can be used in clinical diagnosis, only those listed are accepted as laboratory confirmation for national reporting purposes.
- **Supportive or presumptive laboratory results:** specified laboratory results that are consistent with the diagnosis, yet do not meet the criteria for laboratory confirmation.

From: CDC. Case definitions for infectious conditions under public health surveillance. MMWR

Ebola

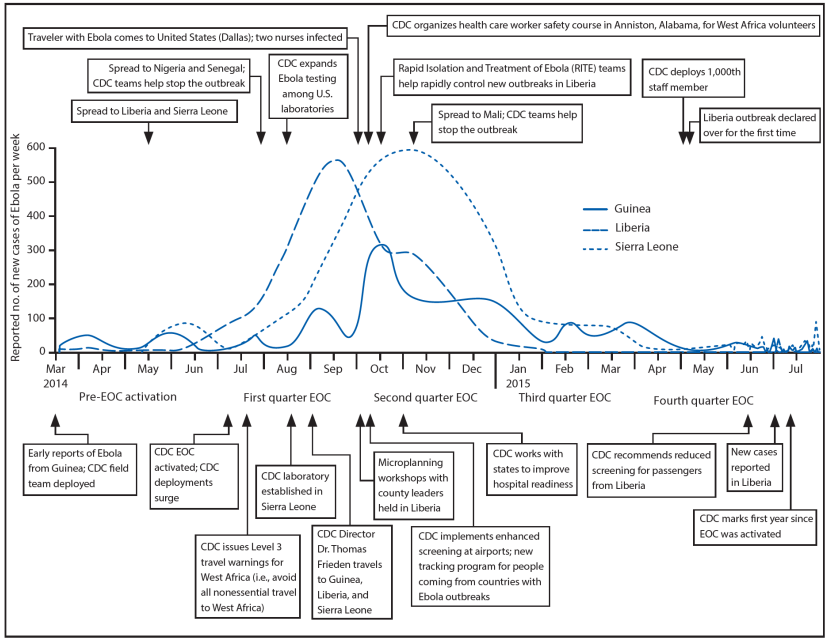
Ebola Virus Disease (EVD)

- Rare and deadly disease in people and nonhuman primates.
- The viruses that cause EVD are located mainly in sub-Saharan Africa.
- People can get EVD through direct contact with an infected animal (bat or nonhuman primate) or a sick or dead person infected with Ebola virus.

CDC. “Ebola (Ebola Virus Disease)” https://www.cdc.gov/vhf/ebola/index.html

2014–2016 Ebola Epidemic

FIGURE 1. CDC's response to the Ebola epidemic, from the first reported cases through the first year after CDC's EOC was activated, and approximate number of reported new cases of Ebola per week — Guinea, Liberia, and Sierra Leone, March 2014–July 2015



Abbreviations: Ebola = Ebola virus disease; EOC = Emergency Operations Center.

Bell BP, Damon IK, Jernigan DB, Kenyon TA, Nichol ST, O'Connor JP, Tappero JW. Overview, Control Strategies, and Lessons Learned in the CDC Response to the 2014-2016 Ebola Epidemic. MMWR Suppl. 2016 Jul 8;65(3):4-11. doi: 10.15585/mmwr.su6503a2. PMID: 27389903.

WHO Ebola ’14 Case Definition

Standard

Suspected: Illness with onset of fever and no response to treatment for usual causes of fever in the area, and at least one of the following signs: bloody diarrhoea, bleeding from gums, bleeding into skin (purpura), bleeding into eyes and urine.

Confirmed: A suspected case with laboratory confirmation (positive IgM antibody, positive PCR or viral isolation)

World Health Organization. ‘Case definition recommendations for Ebola or Marburg Virus Diseases.” https://www.who.int/csr/resources/publications/ebola/ebola-case-definition-contact-en.pdf?ua=1

Ebola Outbreak West Africa ‘14

“WHO and partners respond to the outbreak of Ebola virus disease in Guinea.” Posted 08/11/2014. https://youtu.be/uDicsP3femI

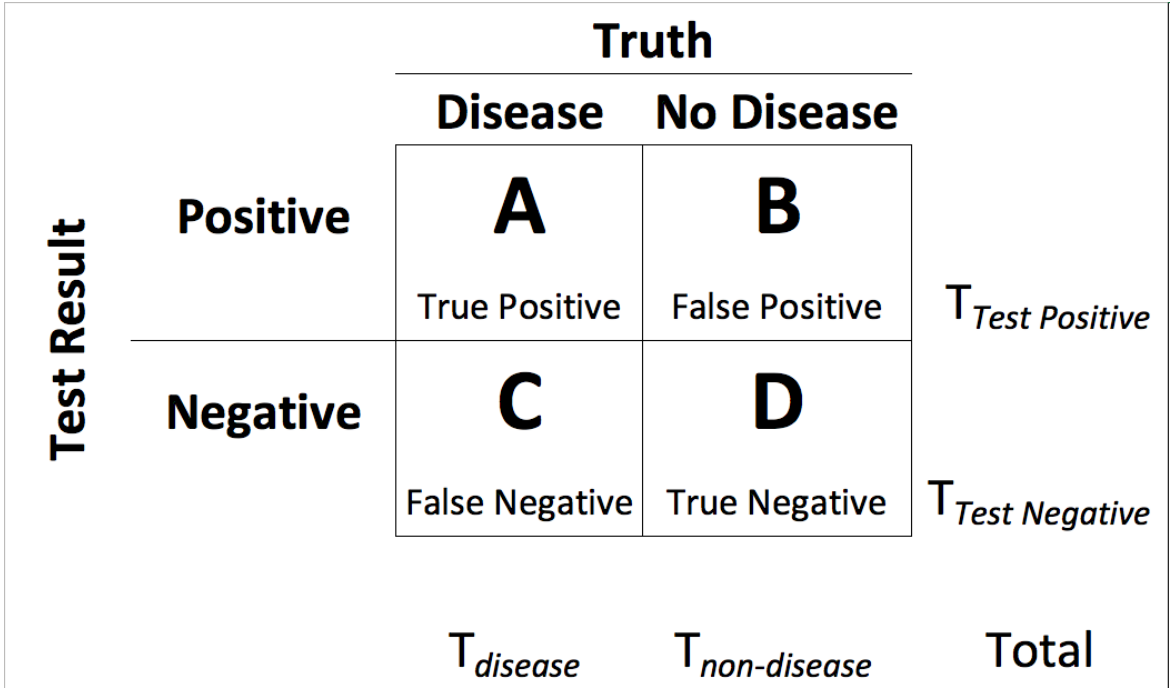
Ebola Outbreak West Africa ‘14

- How do barriers to specimen collection affect case reporting?
- How do low resources affect laboratory confirmation?
- How does ‘social mobilization’ improve case detection in villages?

“WHO and partners respond to the outbreak of Ebola virus disease in Guinea.” Posted 08/11/2014. https://youtu.be/uDicsP3femI

Evaluating Definitions

Evaluating Case Definitions

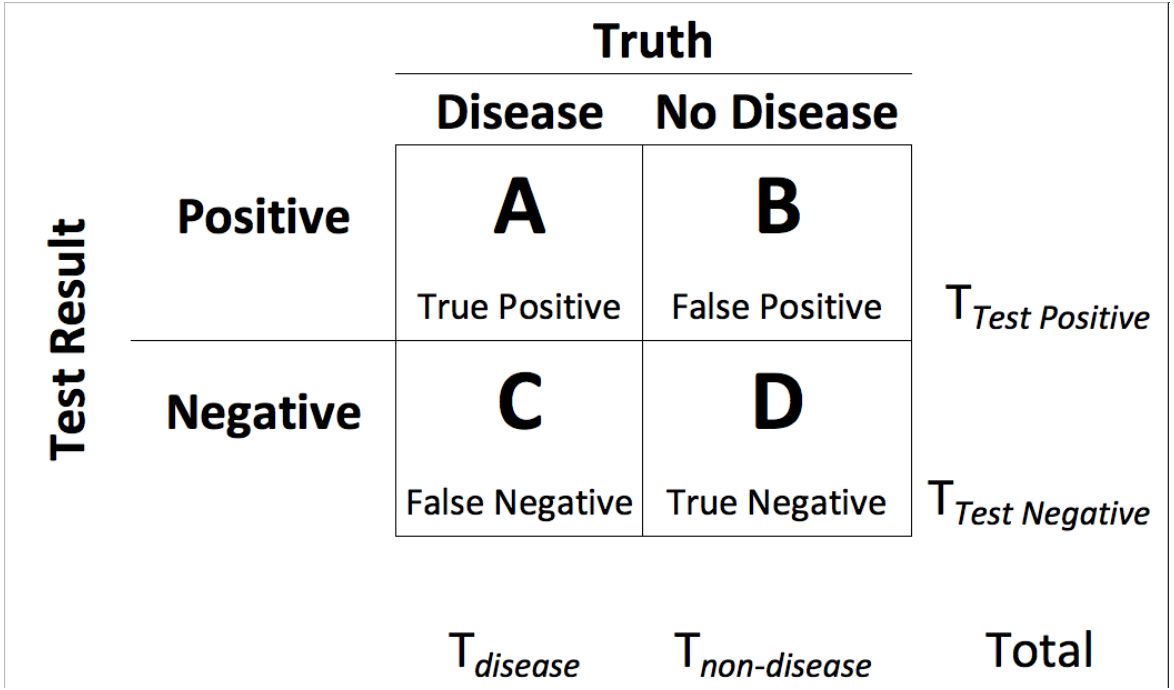


Test Result	Truth: Disease	Truth: No Disease	
Positive	A True Positive	B False Positive	$T_{Test\ Positive}$
Negative	C False Negative	D True Negative	$T_{Test\ Negative}$
	$T_{disease}$	$T_{non\text{-}disease}$	Total

- Prevalence of disease
  - $T_{disease}$/ **Total** × **100**
- Sensitivity - Probability that a test will indicate 'disease' among those with the disease
  - **A/(A+C)** × **100**
- Specificity - The fraction of those without disease who will have a negative test result
  - **D/(D+B)** × **100**

See: Doi SA, Williams GM, editors. Methods of clinical epidemiology. Berlin: Springer; 2013 Jun 1.

Evaluating Case Definitions



Test Result	Truth: Disease	Truth: No Disease	
Positive	A True Positive	B False Positive	$T_{Test\ Positive}$
Negative	C False Negative	D True Negative	$T_{Test\ Negative}$
	$T_{disease}$	$T_{non\text{-}disease}$	Total

- Positive Predictive Value: what is the chance that a person with a positive test truly has the disease?
  - **A/(A+B) × 100**
- Negative Predictive Value: what is the chance that a person with a negative test truly does not have the disease?
  - **D/(D+C) × 100**

See: Doi SA, Williams GM, editors. Methods of clinical epidemiology. Berlin: Springer; 2013 Jun 1.

Classifiers in Practice

Syndromic Surveillance

- Monitor emergency department visits to detect disease outbreaks early
- Collect chief complaint information transmitted electronically to the health department daily
- Analyze for temporal and spatial aberrations

Heffernan R et al. Syndromic surveillance in public health practice, New York City. Emerg Infect Dis. 2004 May;10(5):858-64. doi: 10.3201/eid1005.030646.

Syndrome Coding

Table 1. Syndrome coding and hierarchy

Syndrome	Includes	Excludes
Common cold	Nasal drip, congestion, stuffiness	Chest congestion, sore throat
Sepsis	Sepsis, cardiac arrest, unresponsive, unconscious, dead on arrival	
Respiratory	Cough, shortness of breath, difficulty breathing, croup, dyspnea, bronchitis, pneumonia, hypoxia, upper respiratory illness, chest congestion	Cold
Diarrhea	Diarrhea, enteritis, gastroenteritis, stomach virus	
Fever	Fever, chills, flu, viral syndrome, body ache and pain, malaise	Hay fever
Rash	Vesicles, chicken pox, folliculitis, herpes, shingles	Thrush, diaper and genital rash
Asthma	Asthma, wheezing, reactive airway, chronic obstructive airway disease	
Vomiting	Vomiting, food poisoning	

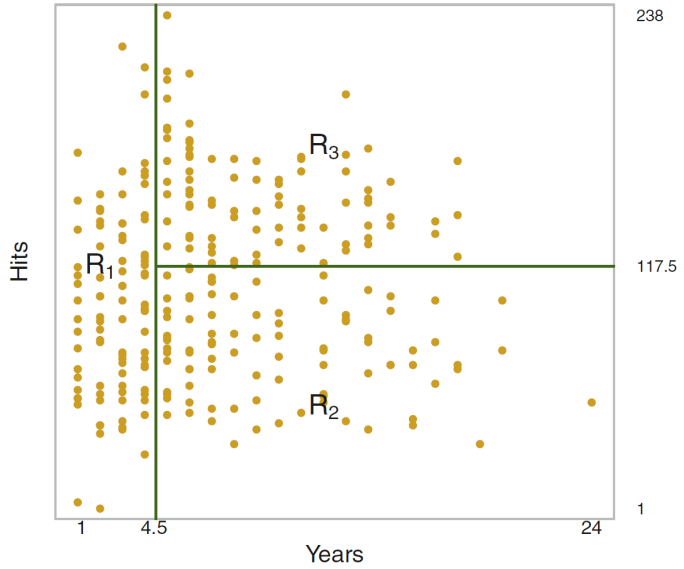
Heffernan R et al. Syndromic surveillance in public health practice, New York City. Emerg Infect Dis. 2004 May;10(5):858-64. doi: 10.3201/eid1005.030646.

Introduction to Trees

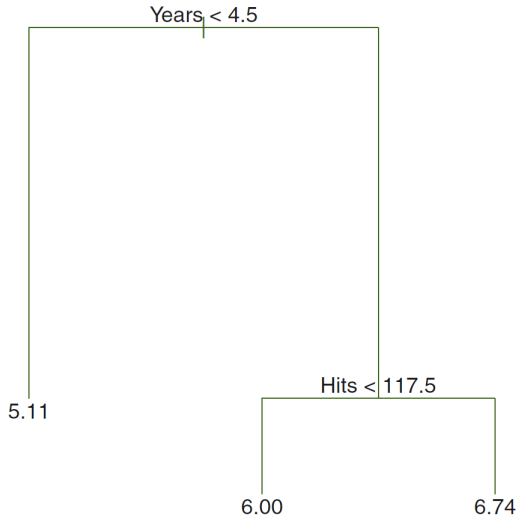
- Trees segment a predictor space into a number of simple regions.
- The set of splitting rules segmenting the predictor space can be summarized in a tree.
- This approach often called 'decision tree' methods.
- Trees predicting numeric outcomes are regression trees.
- Trees predicting nominal outcomes are classification trees.
- Trees offer high interpretability
- Trees often overfit, resulting in high bias

Partitioning

Three-region partition regression tree



Tree for predicting the log salary of a baseball player, based on the number of years that he has played in the major leagues and the number of hits that he made in the previous year.



See Figure 8.1, 8.2. Chapter 08: James, Gareth, et al. An introduction to statistical learning. Vol. 112. New York: springer, 2013.

Tree terminology

Terminal node: final segment for an observation. Can also be called leaves.

Leaves are at the bottom of the **upside down** tree

Internal nodes: The points along the tree where the predictor space is split.

Branches emanate from an internal node split.

See Chapter 08: James, Gareth, et al. An introduction to statistical learning. Vol. 112. New York: springer, 2013.

Classification Trees

- For a classification tree, we predict that each observation belongs to the most commonly occurring class of training observations in the region to which it belongs.
- Use recursive binary splitting to grow a classification tree.

  Classification error rate: Fraction of the training observations in that region that do not belong to the most common class.
- Classification error is not sufficiently sensitive for tree-growing.

See Chapter 08: James, Gareth, et al. An introduction to statistical learning. Vol. 112. New York: springer, 2013.

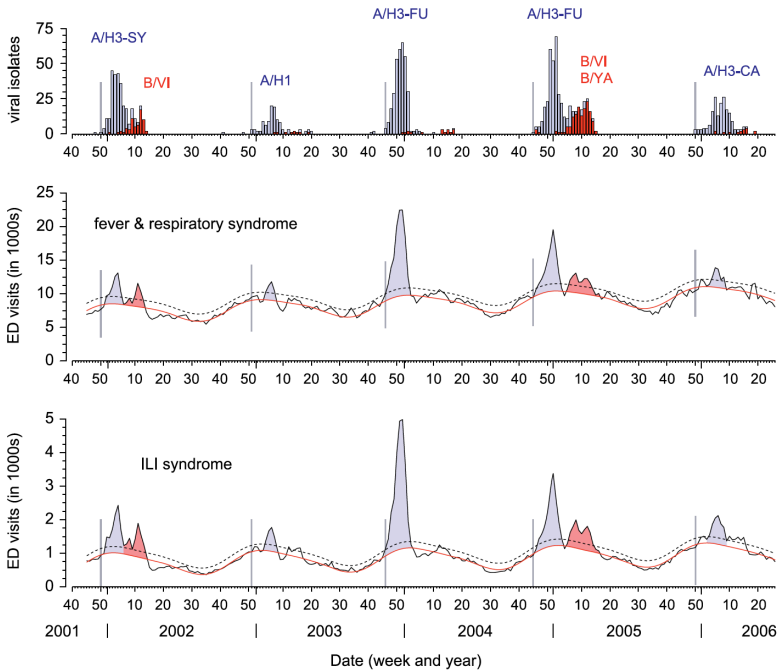
Pruning a Classification Tree

- Gini index is a measure of purity; a small value indicates that a node contains predominantly observations from a single class.
- Gini measures total variance of the misclassification rate across all classes.
- The Gini index takes on a small value if all misclassification rates are close to zero or one.

See Chapter 08: James, Gareth, et al. An introduction to statistical learning. Vol. 112. New York: springer, 2013.

Baseline Simulations

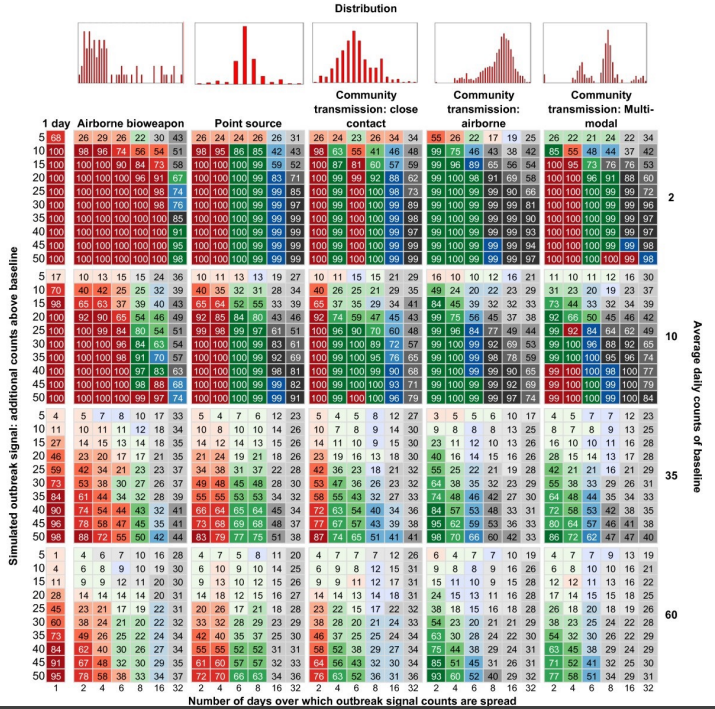
Baselines



- Plot for trends – seasonal trend, secular trend (regression)
- Describe min, max, standard deviation.
  - How strong is the seasonal affect?
  - Are there essentially no cases in the off-season?
- Does a secondary data source exist (e.g. google trends, CDC data) to correlate with your baseline?

Olson, Donald R., et al. "Monitoring the impact of influenza by age: emergency department fever and respiratory complaint surveillance in New York City." PLoS Medicine 4.8 (2007): e247.

Determinants of Detectability



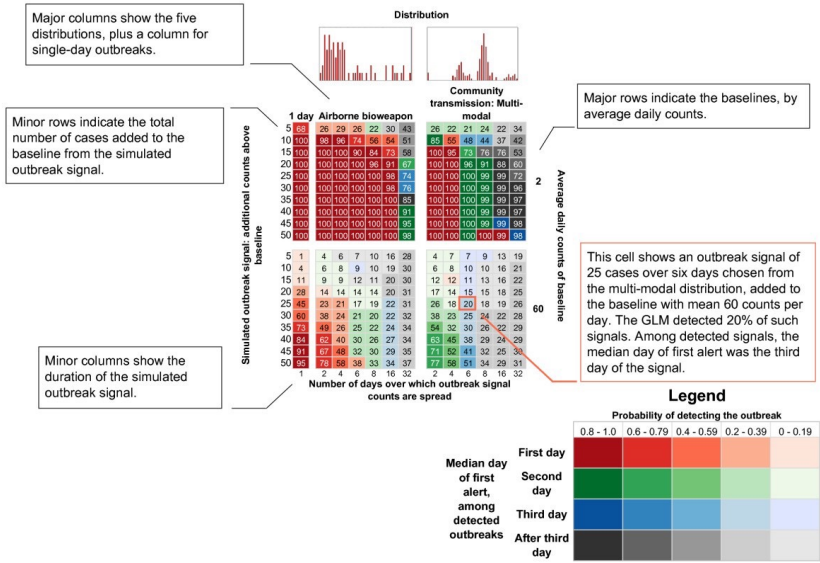
Visualize the effects of baseline and:

- outbreak temporal distribution,
- size, and
- duration

on relative algorithm performance

Figure 3. Jackson ML et al. A simulation study comparing aberration detection algorithms for syndromic surveillance. BMC Med Inform Decis Mak. 2007 Mar 1;7:6.

Surveillance Dynamics



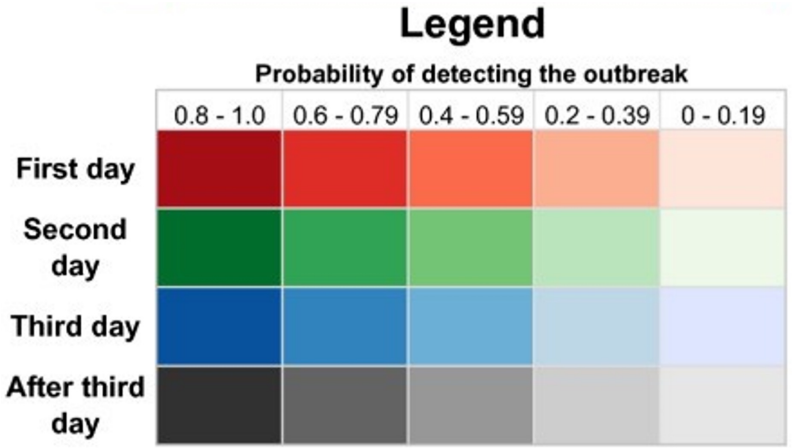
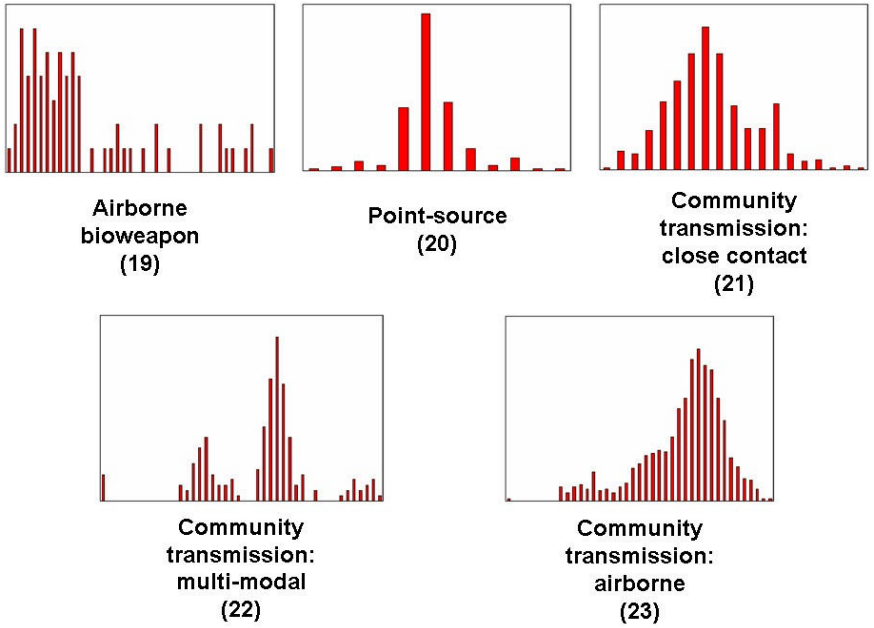


Figure 2. Jackson ML et al. A simulation study comparing aberration detection algorithms for syndromic surveillance. BMC Med Inform Decis Mak. 2007 Mar 1;7:6.

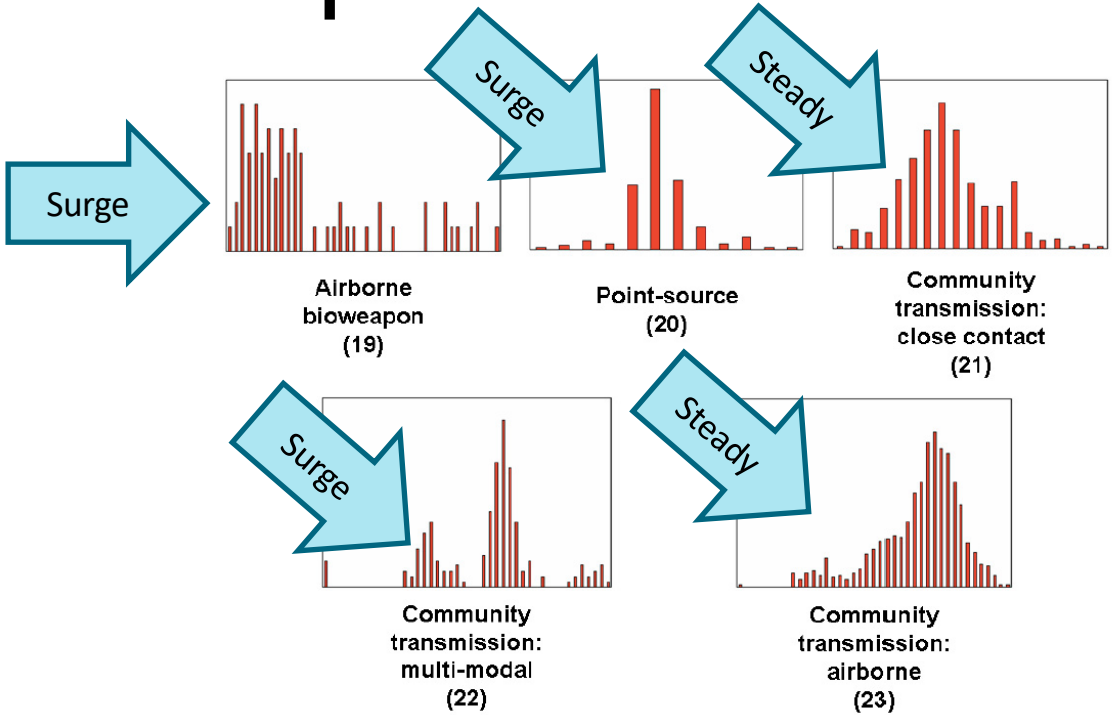
What does it look like?

- **S**hape
- **S**pread
- **S**cenario

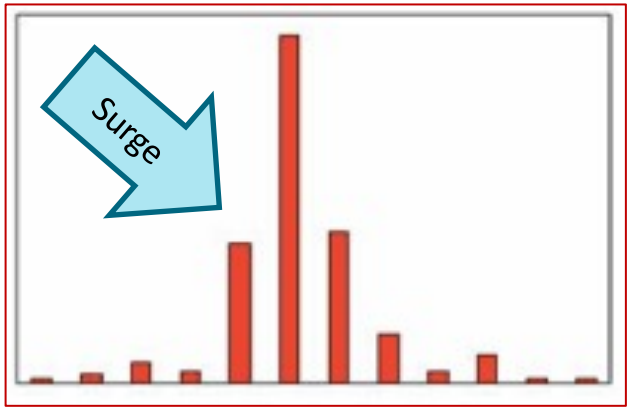


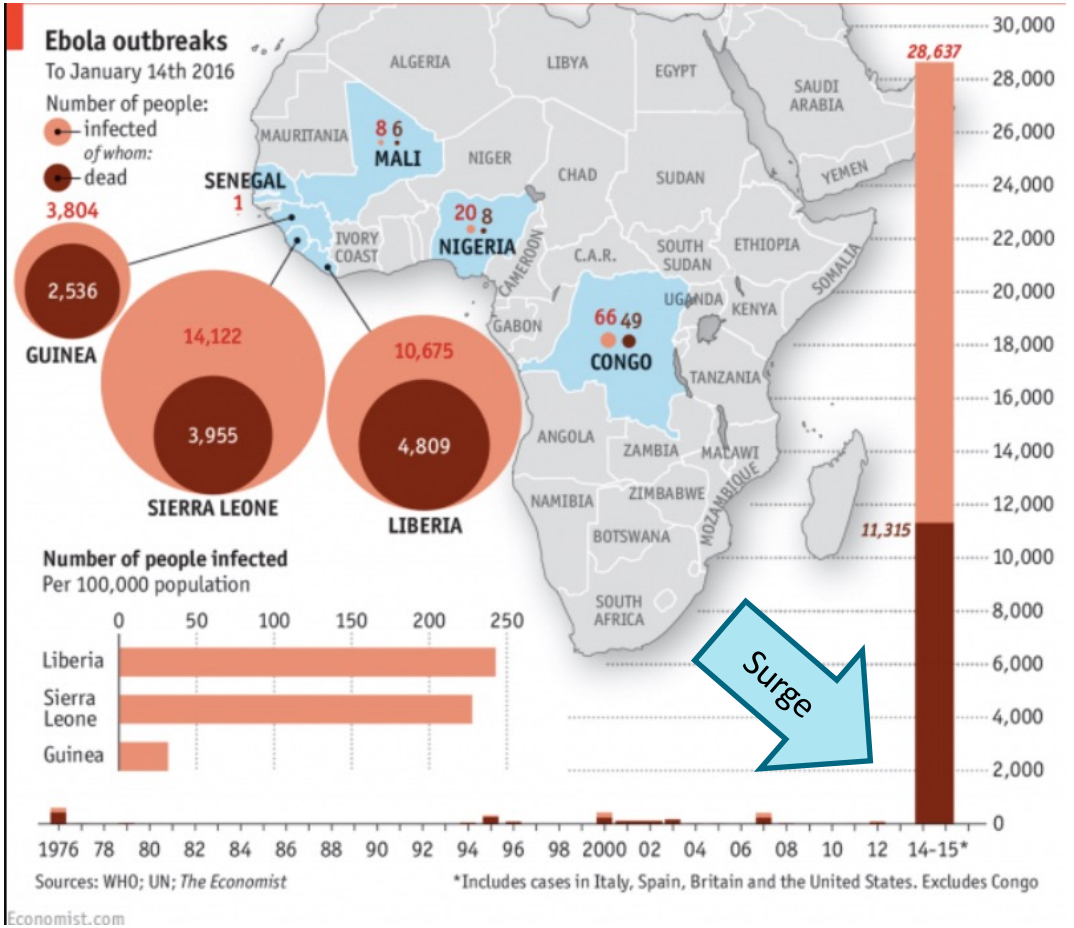
Jackson, Michael L., et al. "A simulation study comparing aberration detection algorithms for syndromic surveillance." BMC Medical Informatics and Decision Making 7.1 (2007): 6.

Epidemic Curves



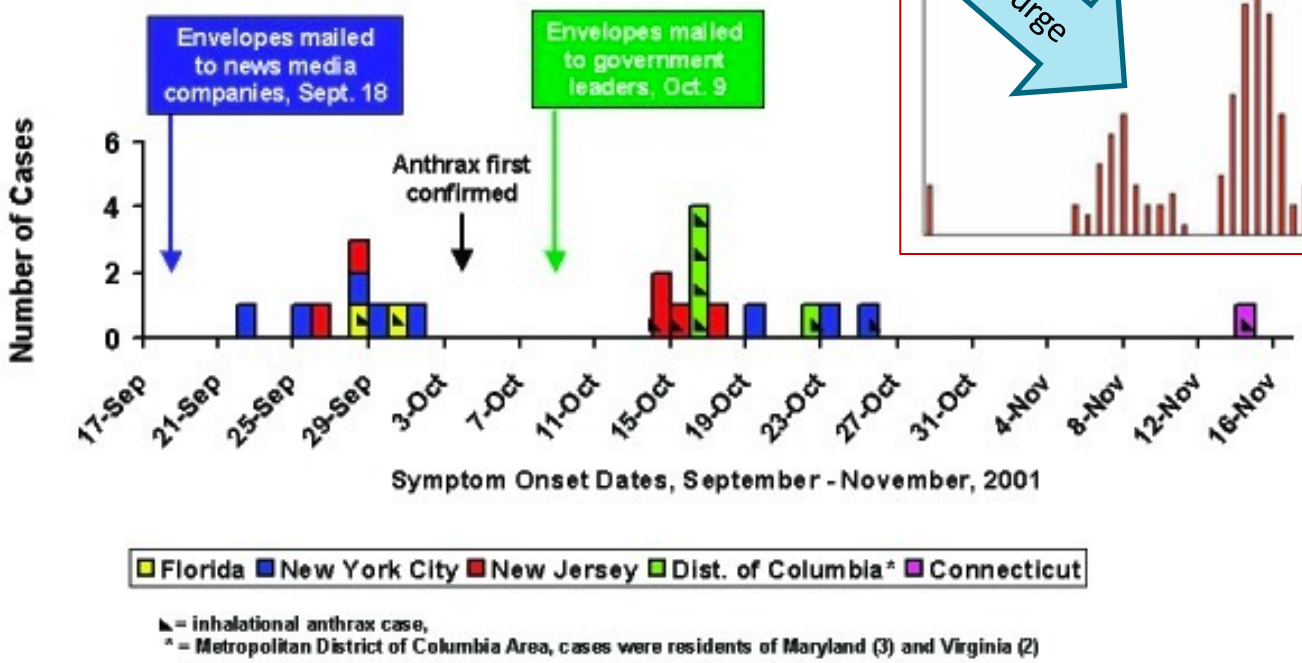
Jackson, Michael L., et al. "A simulation study comparing aberration detection algorithms for syndromic surveillance." BMC Medical Informatics and Decision Making 7.1 (2007): 6.

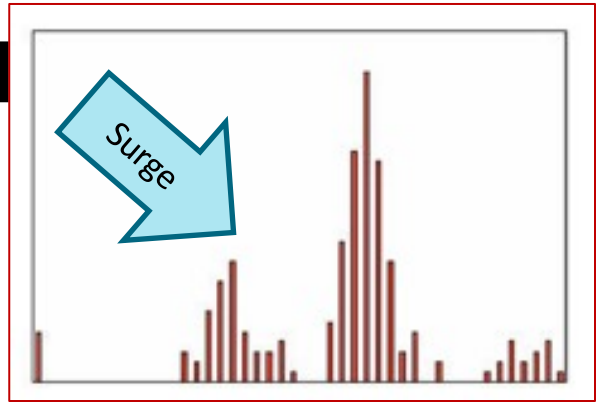




From the Economist: [Ebola in Africa: the end of a tragedy?]

Epidemic Cu





From: Jernigan, Daniel B., et al. "Investigation of bioterrorism-related anthrax, United States, 2001: epidemiologic findings." Emerging infectious diseases 8.10 (2002): 1019.

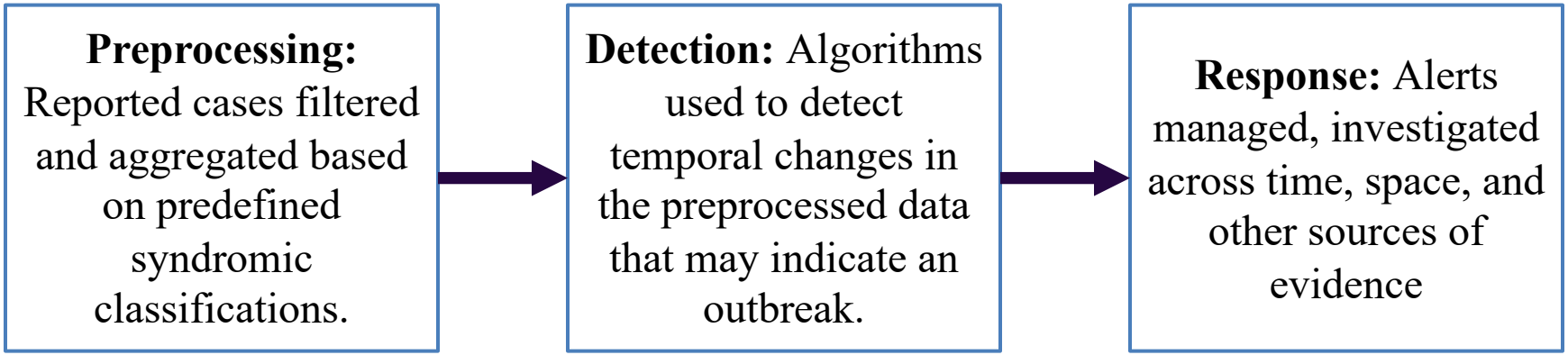
Detecting Outbreaks

What's An Outbreak

- **Epidemic:** refers to an increase, often sudden, in the number of cases of a disease above what is normally expected in that population in that area.
- **Outbreak** carries the same definition of epidemic, but is often used for a more limited geographic area.
- **Cluster** refers to an aggregation of cases grouped in place and time that are suspected to be greater than the number expected, even though the expected number may not be known.
- **Pandemic** refers to an epidemic that has spread over several countries or continents, usually affecting a large number of people.

Dicker, Richard, et al. "Principles of epidemiology in public health practice." Atlanta GA: US Department of Health and Human Services (2006).

Algorithmic Surveillance



Murphy SP, Burkom H. Recombinant temporal aberration detection algorithms for enhanced biosurveillance. J Am Med Inform Assoc. 2008 Jan-Feb;15(1):77-86.

Which algorithm is most effective at detecting potential outbreaks?

Challenging Factors

- Size with median from 0 to 100s
- Cyclic behaviors including day-of-week (DOW) effects and seasonal fluctuations
- Pathogen, its infectivity, the method of introduction to the susceptible population, and the population's underlying social network
- Unknown shape of the resulting time series

Murphy SP, Burkom H. Recombinant temporal aberration detection algorithms for enhanced biosurveillance. J Am Med Inform Assoc. 2008 Jan-Feb;15(1):77-86.

Two steps

ONE

Attempt to capture the expected behavior of the time series in the absence of an outbreak signal.

TWO

Produce a test statistic based on the difference between the observed and expected data.

Murphy SP, Burkom H. Recombinant temporal aberration detection algorithms for enhanced biosurveillance. J Am Med Inform Assoc. 2008 Jan-Feb;15(1):77-86.

Stages of Detection

Table 1 ■ Decomposition of Several Existing Aberration Detection Algorithms from the Literature into Data Forecast and Anomaly Measure Stages

Algorithm	Data Forecasting	Anomaly Measure	Threshold
Fixed Data Threshold	Not applicable	Not applicable	Predetermined number of counts
C1 (EARS)	Moving average with 7-Day window	Z-Score	Three standard deviations above the mean (7-day)
C2 (EARS)	Moving average with 7-day window and 2-day guardband	Z-Score	Three standard deviations above the mean (7-day)
C3 (EARS)	Moving average with 7-day window and 2-day guardband	Sum of last three Z-Scores	Three standard deviations above the mean (7-day)
Reis [3]	Auto regressive moving average applied to trimmed seasonal model	7-day filter applied to residuals	Empirically derived based on desired sensitivity
GScan [5]	Spline fit to historic data	GLRT applied to sum of values in fixed-size moving window	Empirically derived based on simulation studies
Brillman [4]	Loglinear regression model with a fixed baseline	Page's test applied to the residuals in log space	Empirically derived

Murphy SP, Burkom H. Recombinant temporal aberration detection algorithms for enhanced biosurveillance. J Am Med Inform Assoc. 2008 Jan-Feb;15(1):77-86.

Forecasting

- Simplest forecaster would simply use the previous day's value or the value from a week ago without further modification.
- The EARS family of algorithms (C1, C2, C3) all use a 7-day moving average to generate expected values.
- C1 uses the last 7 days of data
- C2 and C3 use a 2-day buffer data from 3 to 9 days in the past which ignores the current and two most recent days of data.
  - This guardband prevents outbreak effects from contaminating the predictions, due to presence of the early part of the outbreak signal in the baseline.

Murphy SP, Burkom H. Recombinant temporal aberration detection algorithms for enhanced biosurveillance. J Am Med Inform Assoc. 2008 Jan-Feb;15(1):77-86.

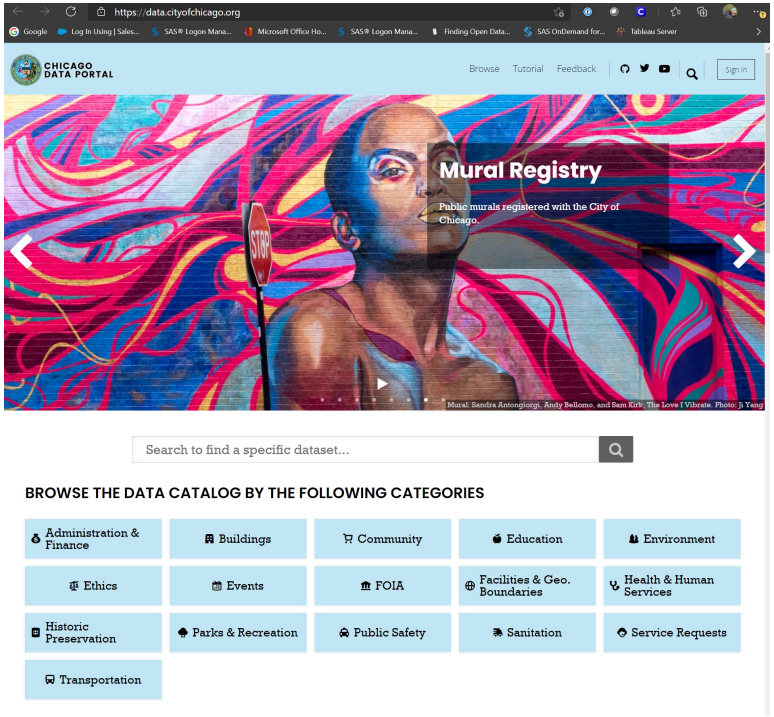
Anomaly Detection

- Calculate a numeric value to quantify the positive differences between observation and prediction.
- Control alert rates for sensitivity to high values such overpredictions not tested for anomalies.
- Use calculated numeric value or test statistic to decide whether to investigate the cases causing the anomaly for evidence of a public health event.
- Calculate difference between observed and forecast, resulting in a time series of residuals that are then used to form a test statistic as the basis for an alerting decision.
- May normalize the residuals to account for natural variance in the baseline data.

Murphy SP, Burkom H. Recombinant temporal aberration detection algorithms for enhanced biosurveillance. J Am Med Inform Assoc. 2008 Jan-Feb;15(1):77-86.

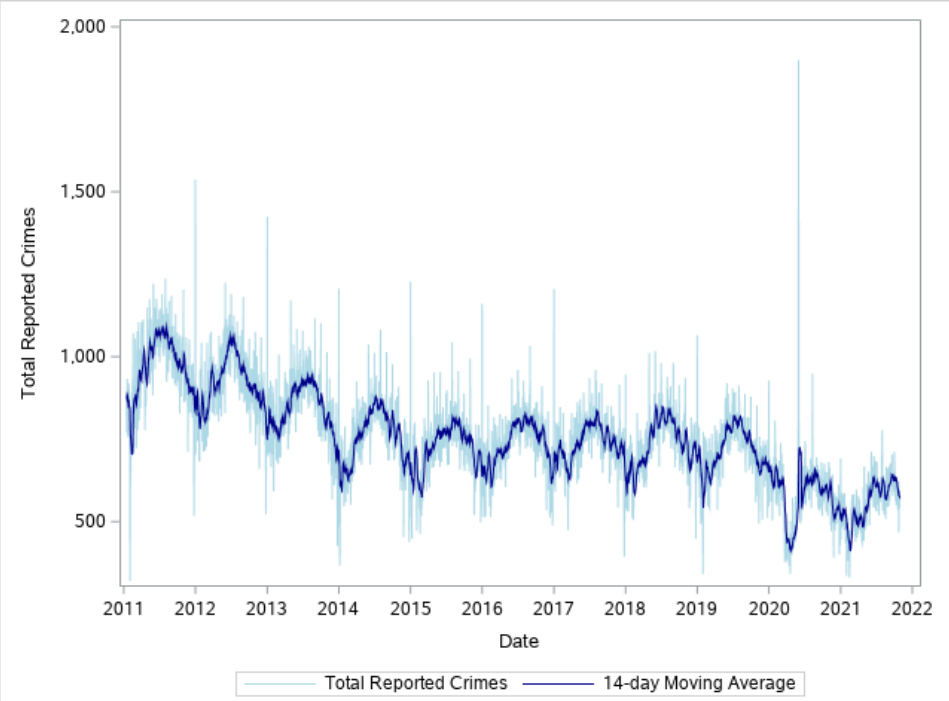
Outbreak Example

Chicago Open Data Portal



https://data.cityofchicago.org/

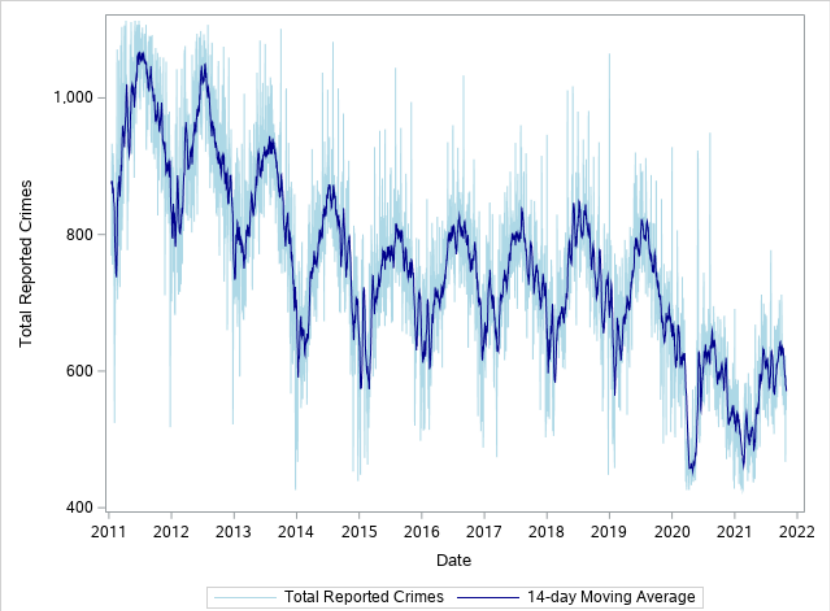
Reported Crimes 2011 - Present



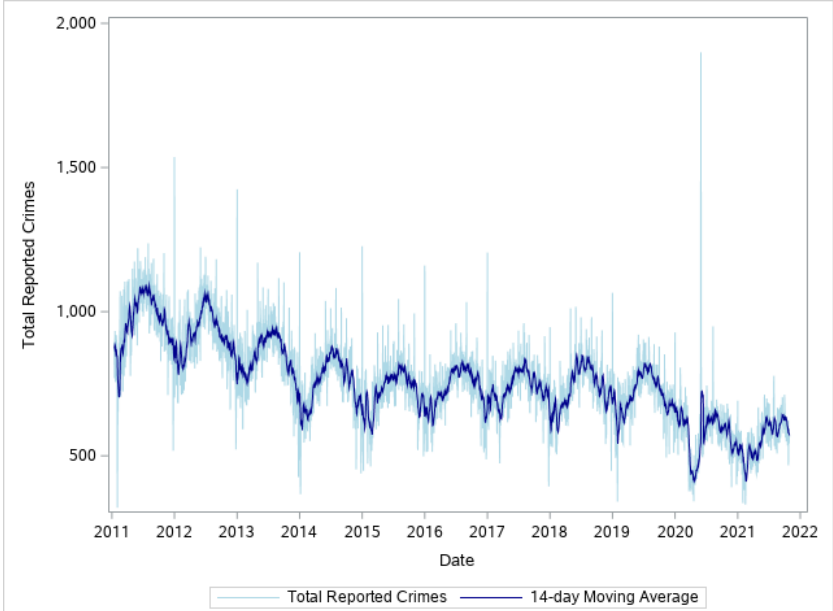
https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2

Outliers

Winsorized Observations – 1%

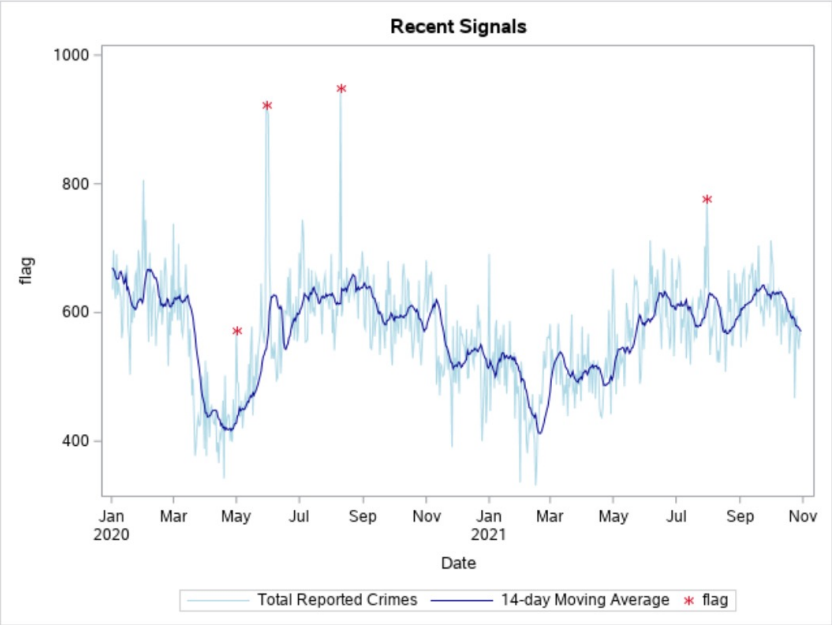


All Data



https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2

Recent Signals – All Crime



Number of Flags

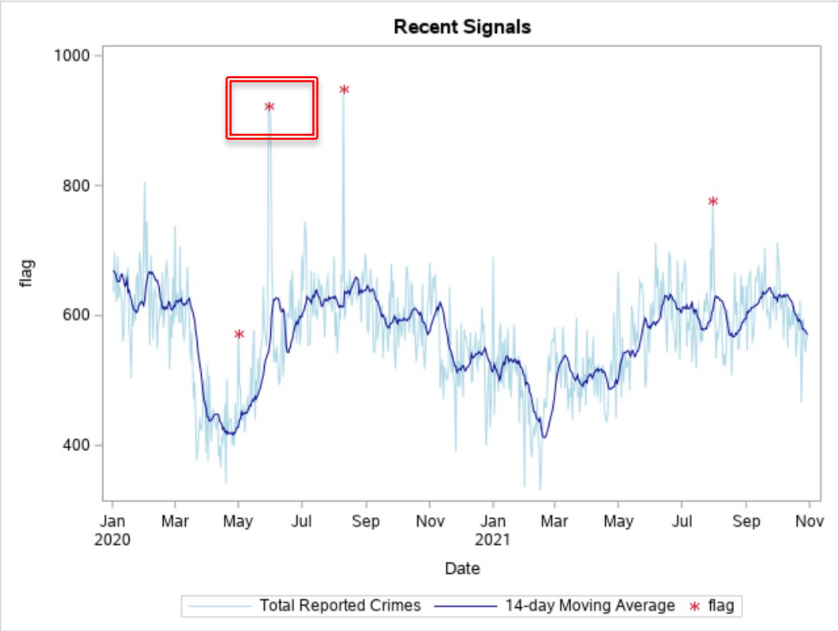
Total Intervals	Flags	Flags Per Hundred
3866	35	0.91

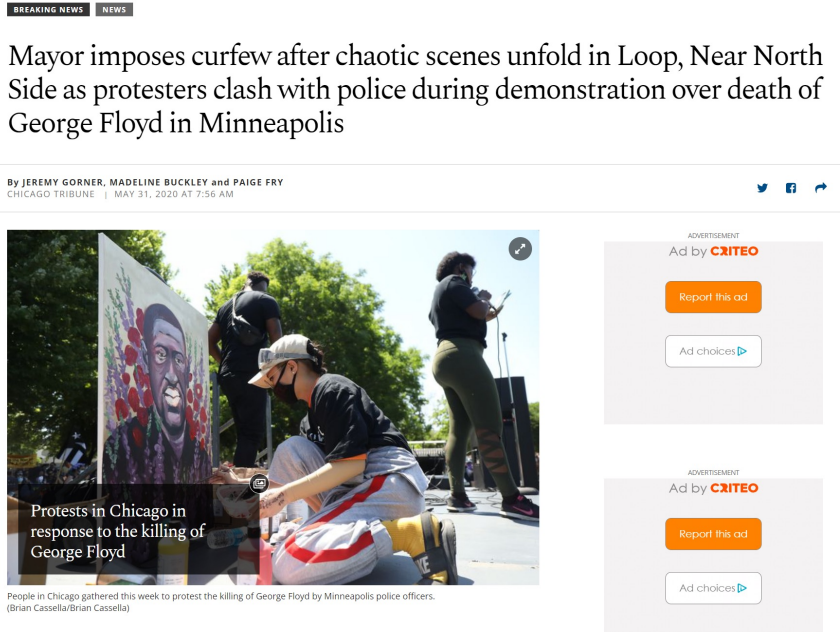
Top 20 Recent Signals

Date	Total Reported Crimes	Upper CI
07/31/2021	777	748.18
08/10/2020	949	760.56
05/30/2020	923	738.59
05/01/2020	573	546.02
01/01/2020	928	894.83

https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2

Recent Signals – All Crime





BREAKING NEWS | NEWS

Mayor imposes curfew after chaotic scenes unfold in Loop, Near North Side as protesters clash with police during demonstration over death of George Floyd in Minneapolis

By JEREMY GORNER, MADELINE BUCKLEY and PAIGE FRY
CHICAGO TRIBUNE | MAY 31, 2020 AT 7:56 AM

Protests in Chicago in response to the killing of George Floyd

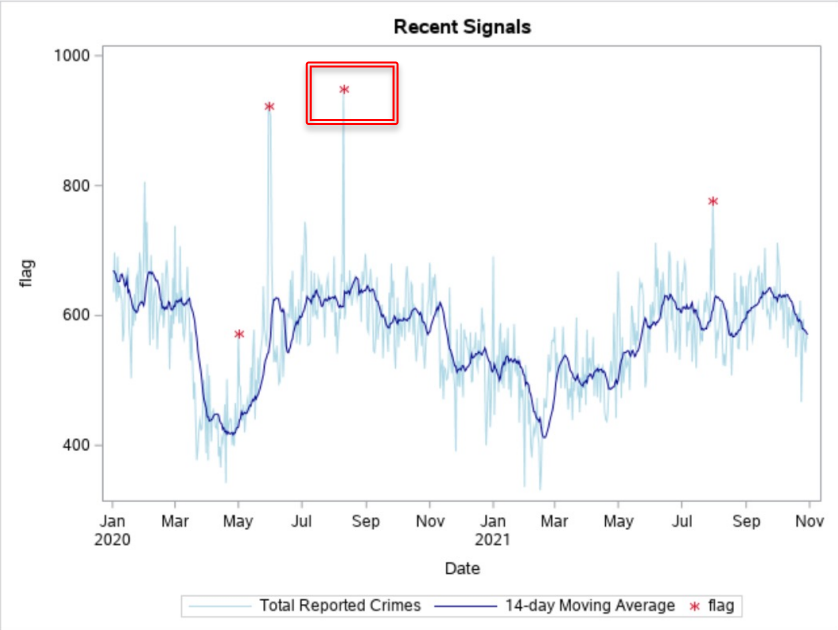
People in Chicago gathered this week to protest the killing of George Floyd by Minneapolis police officers. (Brian Cassella/Brian Cassella)

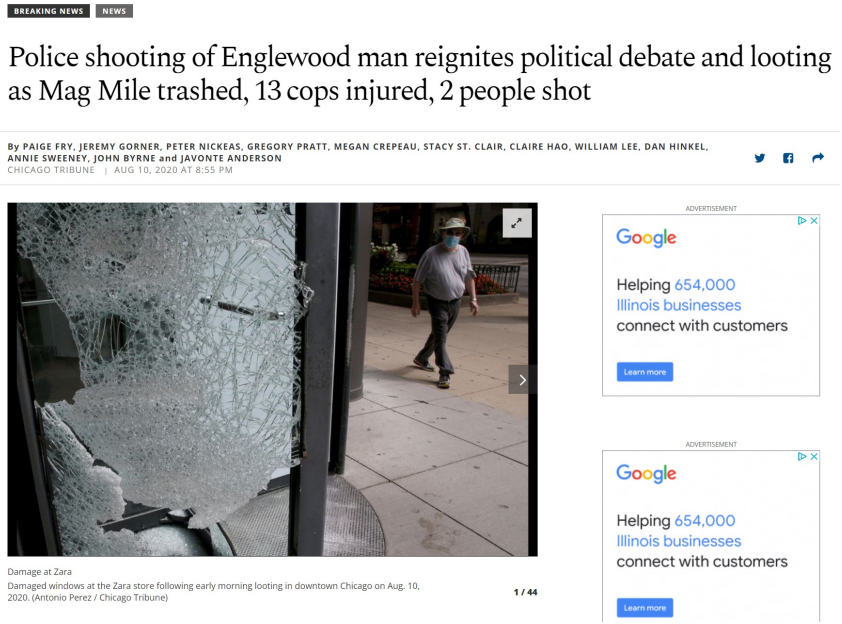
ADVERTISEMENT
Ad by CRITEO
Report this ad
Ad choices

ADVERTISEMENT
Ad by CRITEO
Report this ad
Ad choices

https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2

Recent Signals – All Crime





BREAKING NEWS | NEWS

Police shooting of Englewood man reignites political debate and looting as Mag Mile trashed, 13 cops injured, 2 people shot

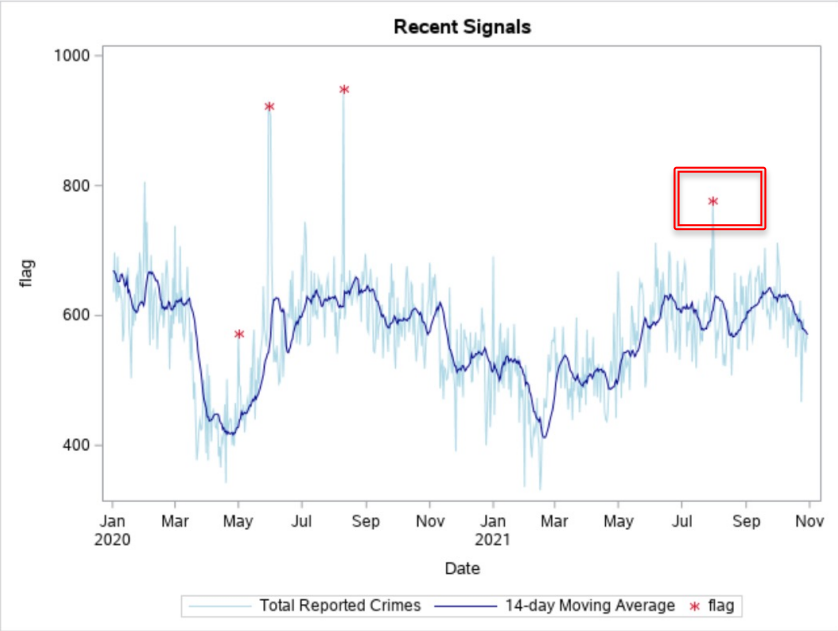
By PAIGE FRY, JEREMY GORNER, PETER NICKEAS, GREGORY PRATT, MEGAN CREPEAU, STACY ST. CLAIR, CLAIRE HAO, WILLIAM LEE, DAN HINKEL, ANNIE SWEENEY, JOHN BYRNE and JAVONTE ANDERSON
CHICAGO TRIBUNE | AUG 10, 2020 AT 8:55 PM

Damage at Zara
Damaged windows at the Zara store following early morning looting in downtown Chicago on Aug. 10, 2020. (Antonio Perez / Chicago Tribune)

1 / 44

https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2

Recent Signals – All Crime



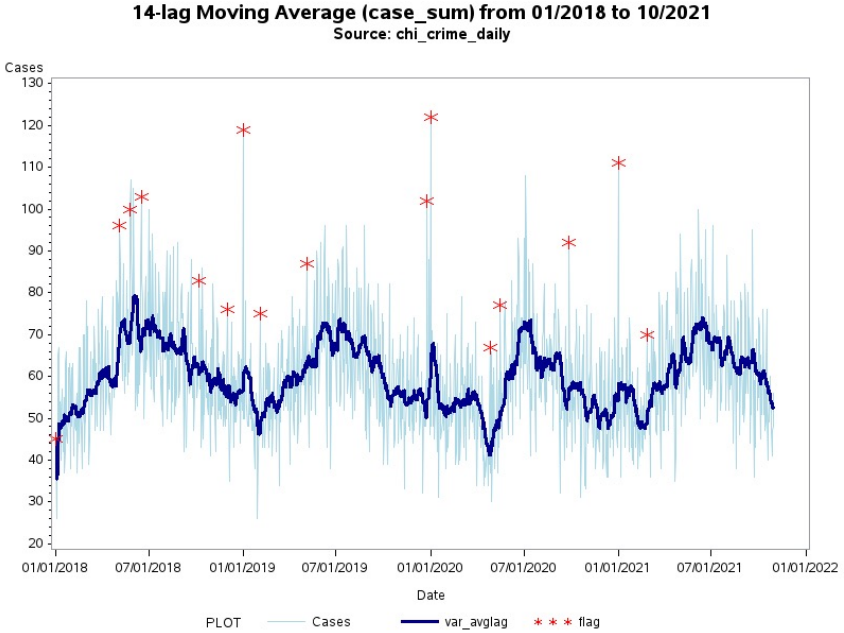
https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2

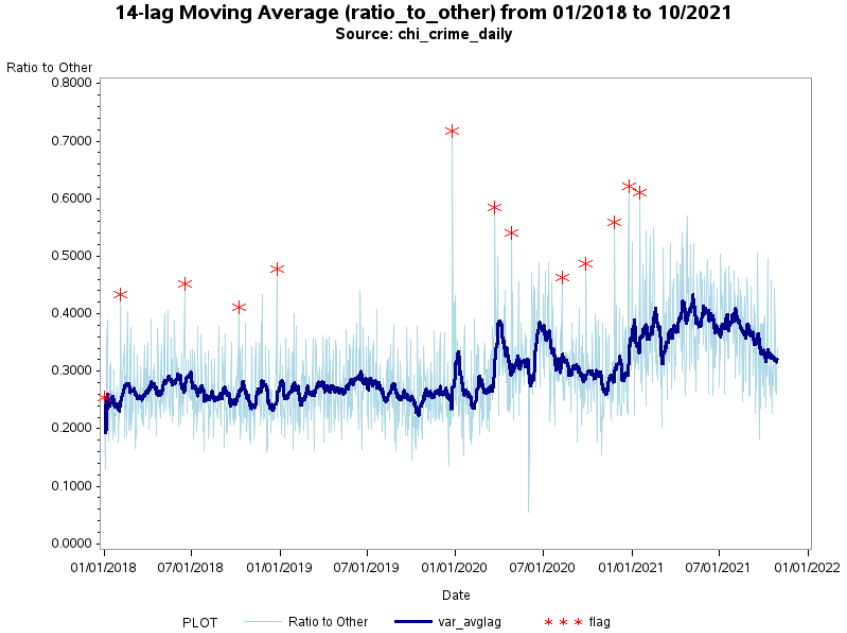
CCVI vs Domestic Crime

Domestic Crime	Frequency	High CCVI	Pct High CCVI	Medium CCVI	Pct Medium CCVI	Low CCVI	Pct Low CCVI
false	2,508,091	1,031,268	41.12%	636,132	25.36%	840,691	33.52%
true	476,328	244,876	51.41%	151,909	31.89%	79,543	16.70%

https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2

High CCVI Domestic Crime





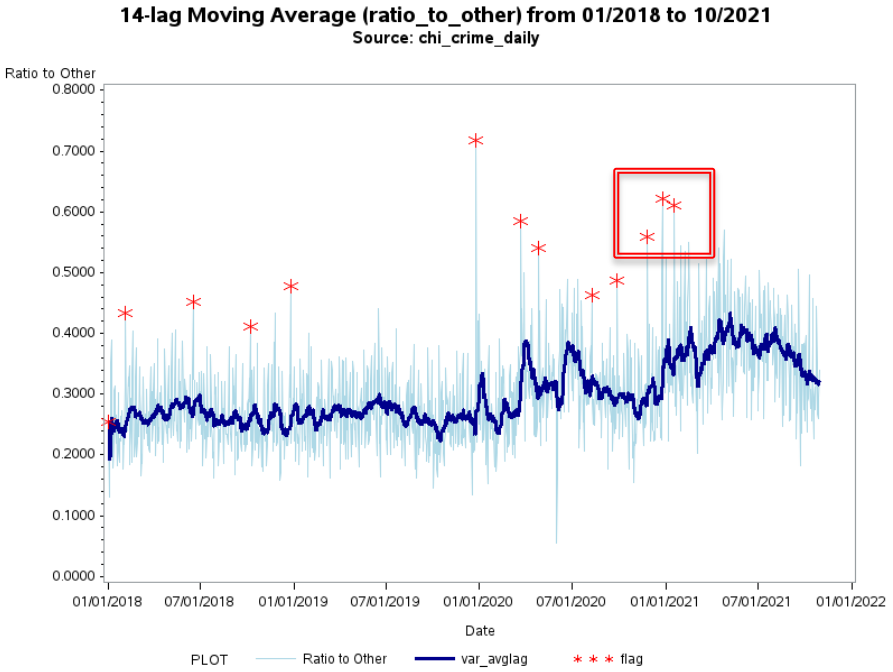
https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2

High CCVI Domestic Crime

Total Intervals	Flags	Flags Per Hundred
1398	13	0.93

Top 20 Recent Signals

Date	Ratio to Other	Upper CI	Cases
01/18/2021	0.6106	0.54	69
12/25/2020	0.6218	0.60	74
11/26/2020	0.5583	0.41	67
09/27/2020	0.4868	0.48	92
08/09/2020	0.4633	0.45	82
04/26/2020	0.5403	0.52	67
03/22/2020	0.5856	0.49	65
12/25/2019	0.7183	0.41	102
12/25/2018	0.4779	0.37	65
10/07/2018	0.4109	0.38	83
06/17/2018	0.4518	0.41	103
02/04/2018	0.4326	0.38	61



https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2

Ask Questions

nicholas.soulakis@northwestern.edu