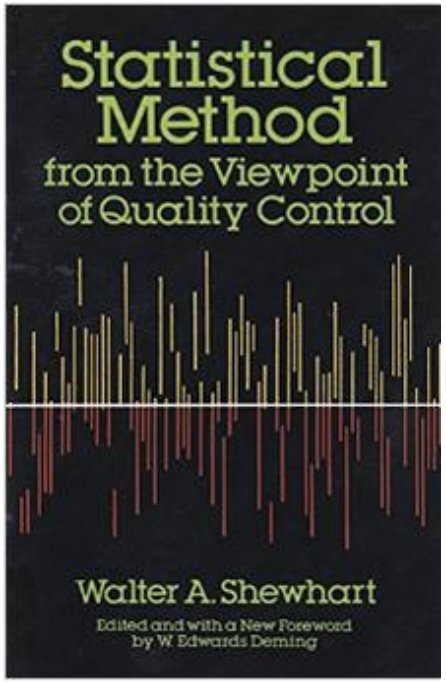
Practical Methods for Identifying Anomalies That Matter in Large Datasets

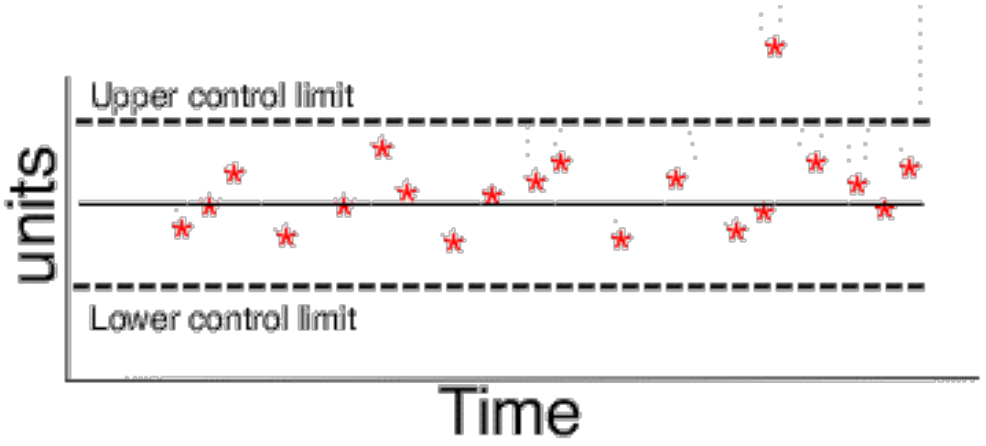
Robert L. Grossman
University of Chicago
and
Open Data Group

O'Reilly Strata Conference
February 20, 2015



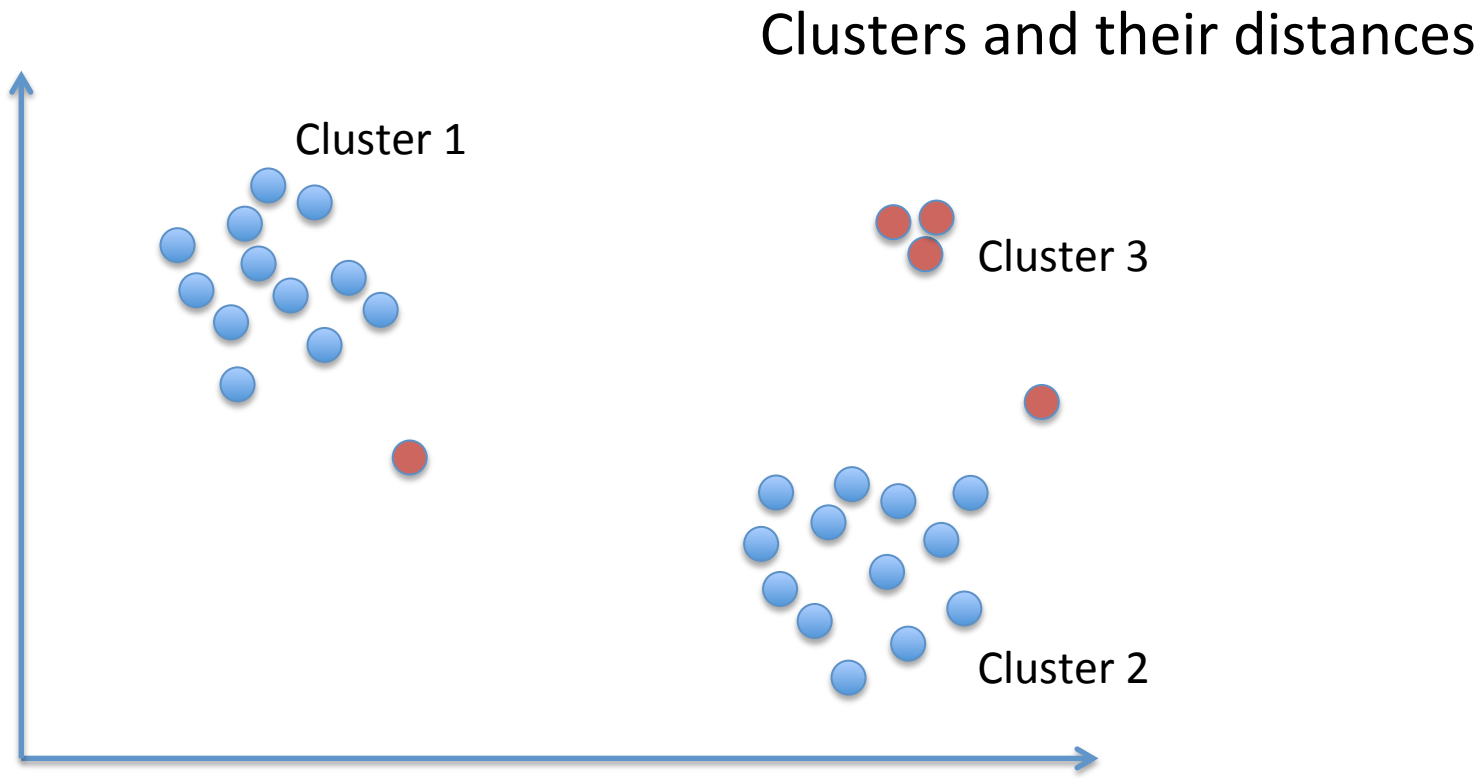
Originally published in 1939.

Modeling populations



- Walter A. Shewhart of the Bell Telephone Laboratories wrote a memorandum on May 16, 1924 that included a diagram of a control chart.
- Upper and lower control limits were defined by k*standard deviation of the observations in a test set.
- Methods, such as CUSUM, developed for quickest detection of changes.

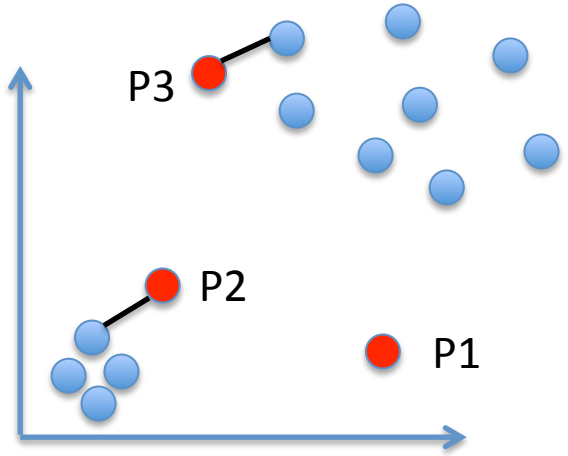
Source of figure: http://www.itl.nist.gov/div898/handbook/mpc/section2/mpc22.htm



- In this toy example, blue points are normal behavior, red points are anomalies.
- There are many way to measure distance of potential anomalies to clusters.
- Each new clustering algorithm potentially introduces a new method of defining anomalies.

Neighbors and their densities

- Neighborhoods, density and context matter**
- Local Outlier Factor (LOF)
  - P1 and P2 outliers
- Nearest Neighbor
  - P1 outlier



Source: *V. Chandola, A. Banerjee and V. Kumar, Anomaly detection: A survey, ACM Computing Surveys Volume 41, Number 3, 2009.
** R. Grossman and C. Gupta, Outlier Detection with Streaming Dyadic Decomposition, Industrial Conference on Data Mining, 2007, pages 77-91.

- An outlier is an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism.*
- Anomaly detection is the search for items or events which do not conform to an expected pattern.
- As a working definition, we view anomalies as arising from a different population or process.

*D. Hawkins. Identification of Outliers. Chapman and Hall, 1980.

**V. Chandola, A. Banerjee and V. Kumar, Anomaly detection: A survey, ACM Computing Surveys Volume 41, Number 3, 2009.

- Three common cases:
  - Supervised: We have two classes of data, one normal and one consisting of anomalies. This is a standard classification problem, perhaps with imbalanced classes. Well understood problem.
  - Unsupervised: We have no labeled data representing anomalies or normal data.
  - Semi-supervised: We have some labeled data available for training, but not very much; perhaps just for normal data.

The core of most large scale systems that process anomalies is a ranking and packaging of candidate anomalies so that they may be carefully investigated by subject matter experts.

Case Study 1: Active Voxels in FMRI

Source: Salmo salar, (Atlantic Salmon), wikipedia.org

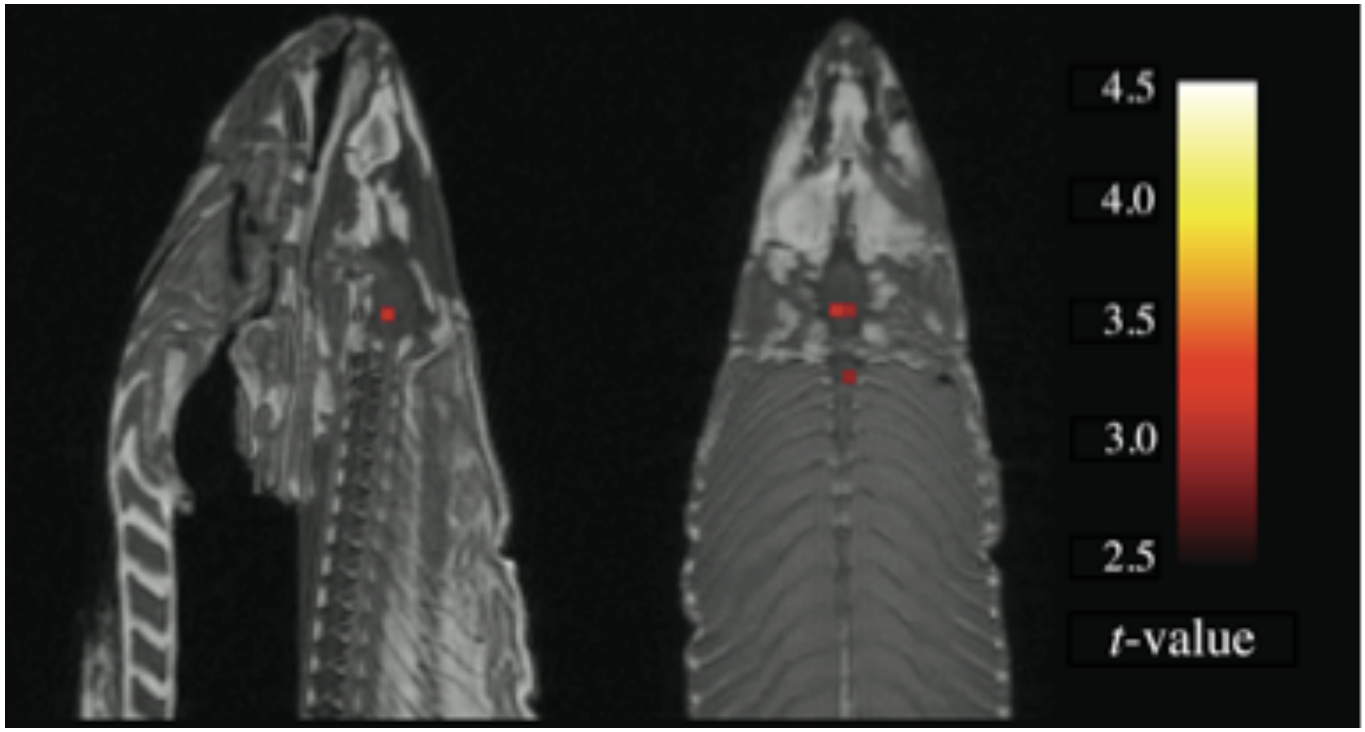
Methods

Subject. One mature Atlantic Salmon (Salmo salar) participated in the fMRI study. The salmon was approximately 18 inches long, weighed 3.8 lbs, and was not alive at the time of scanning.

Task. The task administered to the salmon involved completing an open-ended mentalizing task. The salmon was shown a series of photographs depicting human individuals in social situations with a specified emotional valence. The salmon was asked to determine what emotion the individual in the photo must have been experiencing.

Design. Stimuli were presented in a block design with each photo presented for 10 seconds followed by 12 seconds of rest. A total of 15 photos were displayed. Total scan time was 5.5 minutes.

Source: Craig M. Bennett, Abigail A. Baird, Michael B. Miller, and George L. Wolford, Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: An argument for multiple comparisons correction, retrieved from http://prefrontal.org/files/posters/Bennett-Salmon-2009.pdf.



Several active voxels were discovered in a cluster located within the salmon's brain cavity. The size of this cluster was 81 mm^3 with a cluster-level significance of p = 0.001. Due to the coarse resolution of the echo-planar image acquisition and the relatively small size of the salmon brain further discrimination between brain regions could not be completed. Out of a search volume of 8064 voxels a total of 16 voxels were significant.

Source: ibid.

Rule of Thumb:
Beware of Multiple Tests

- Beware of drawing any conclusions from many independent tests (asking whether each voxel is “on” is an independent test) unless the analysis is properly done.

What Happened: Toy Example
Flipping a coin 10 times

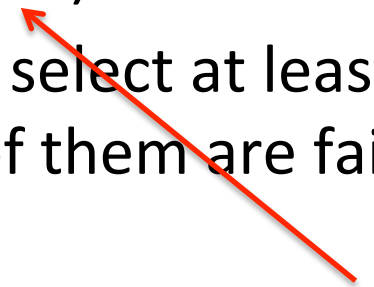
- How likely is it if you flip a fair coin 10 times that you will get 9 or 10 heads?

HHHHH HHHHT	HHHHT HHHHH
HHHHH HHHTH	HHHTH HHHHH
HHHHH HHTHH	HHTHH HHHHH
HHHHH HTHHH	HTHHH HHHHH
HHHHH THHHH	THHHH HHHHH
HHHHH HHHHH	

Likelihood
$= 11 / (1/2)^{10}$
$= 11 / 1024$
$= .01074$
$= 1.074\%$

Bonferroni Correction for Multiple Tests

- What is the probability if a repeat this experiment 125 times that I will never a get a run of 9 or 10 heads?
- The probability is $(1 - 0.01074)^{125} = 25.9\%$
- So almost 75% of the time I will select at least one of the coins as biased when in fact all of them are fair.
- To be more conservative, divide the significance of coming up with 9 or 10 heads by the number of tests.
- The probability is $(1 - 0.01074/125)^{125} = 98.2\%$
- So less than 2% of the time will I select at least one of the coins as biased when in fact all of them are fair.

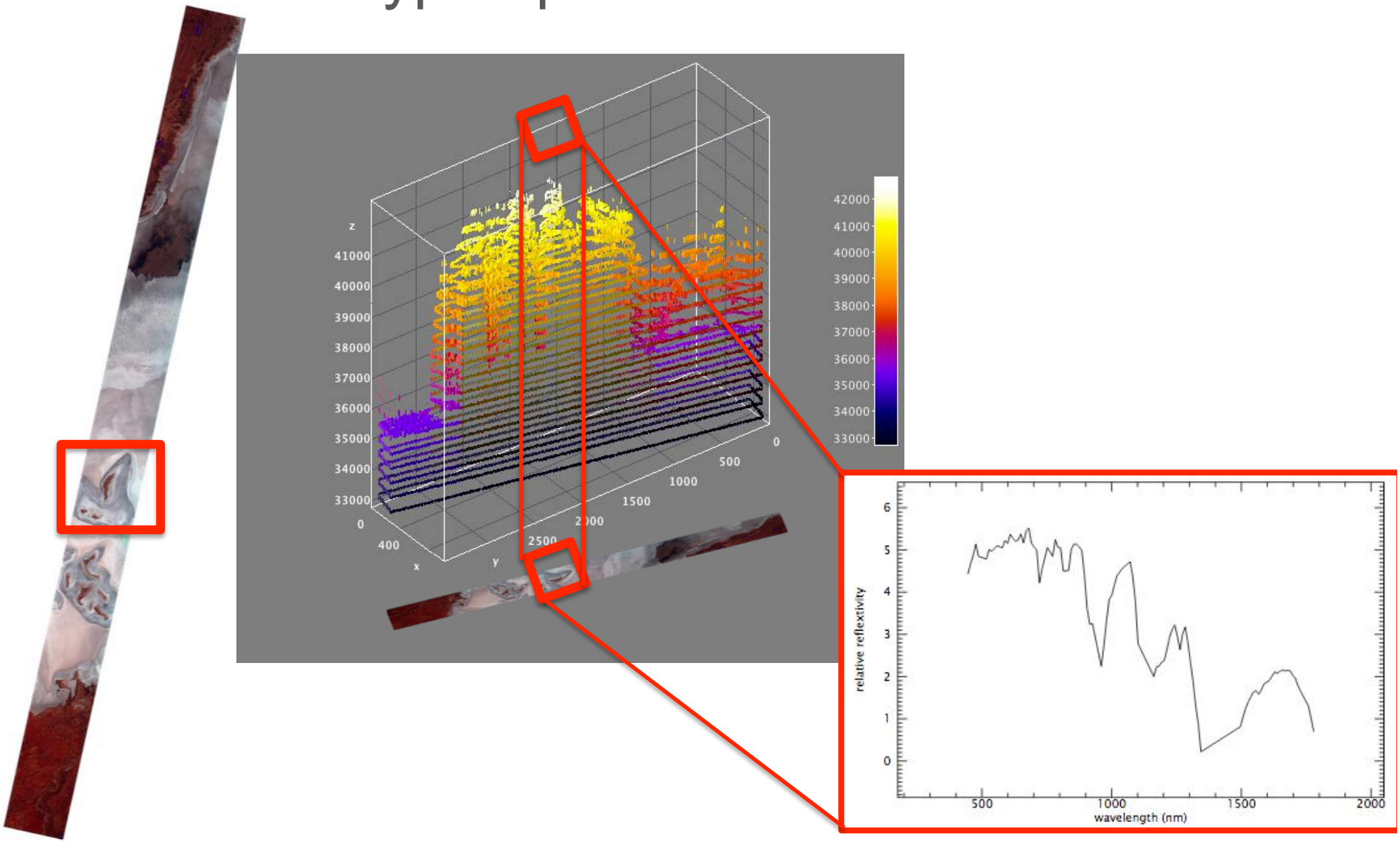


Bonferroni Correction

Case Study 2: Hyperspectral Anomalies from NASA's EO-1 Satellite

Source: Project Matsu, matsu.opensciencedatacloud.org

NASA's EO-1 Hyperion and ALI Hyperspectral Instruments



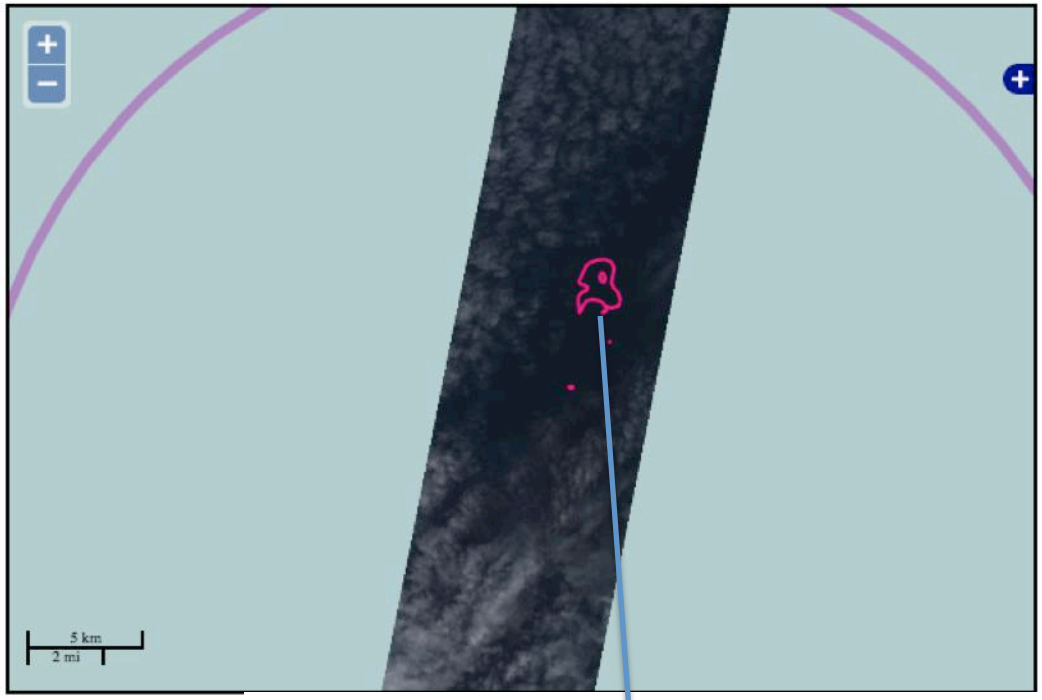
Matsu “Wheel” Spectral Anomaly Detector

- “Contours and Clusters” – looks for physical contours around spectral clusters
  - PCA analysis applied to the set of reflectivity values (spectra) for every pixel, and the top 5 components are extracted for further analysis.
  - Pixels are clustered in the transformed 5-D spectral space using a k-means clustering algorithm.
  - For each image, k = 50 spectral clusters are formed and ranked from most to least extreme using the Mahalanobis distance of the cluster from the spectral center.
  - For each spectral cluster, adjacent pixels are grouped together into contiguous objects.
- Returns geographic regions of spectral anomalies that are scored again as anomalous (0 least , 1000 most) compared to a set of “normal” spectra, constructed for comparison over a baseline of time

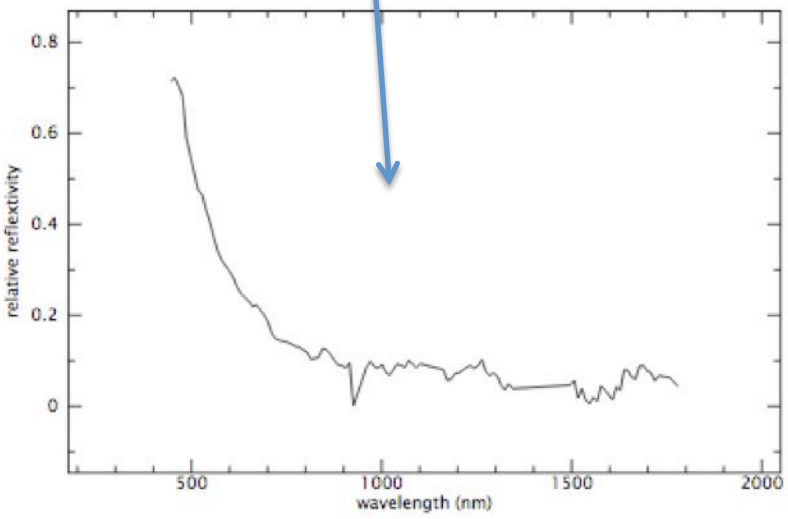
Spectral anomaly detected: Nishinoshima active volcano, Dec, 2014

Matsu Analytic Image Report

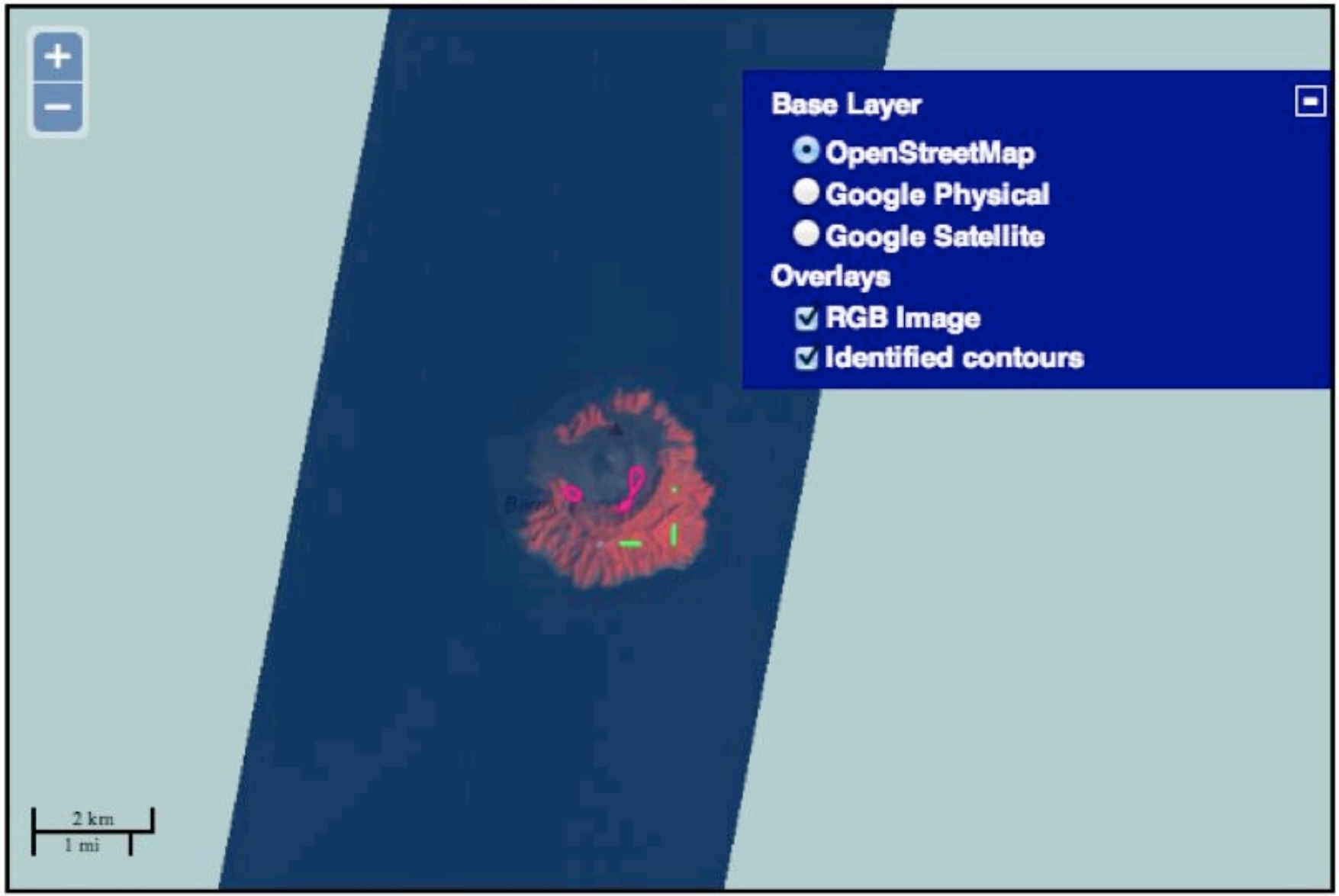
Collection Date	2014-12-02 (day 336)
Analysis Date	Wed Dec 17 12:27:25 2014
Analytic Environment	
Analytic	Contours-2013-12-r4
Noise Correction Enabled	False
Summary Stats	ss-2013-12-r1
Data Ingest	populateHDFS-2013-11-r1
Report Format	reportContoursR4
Hyperspectral Image	
Image	EO1H1050412014336110KF_HYP_L1G
Number of Bands	242



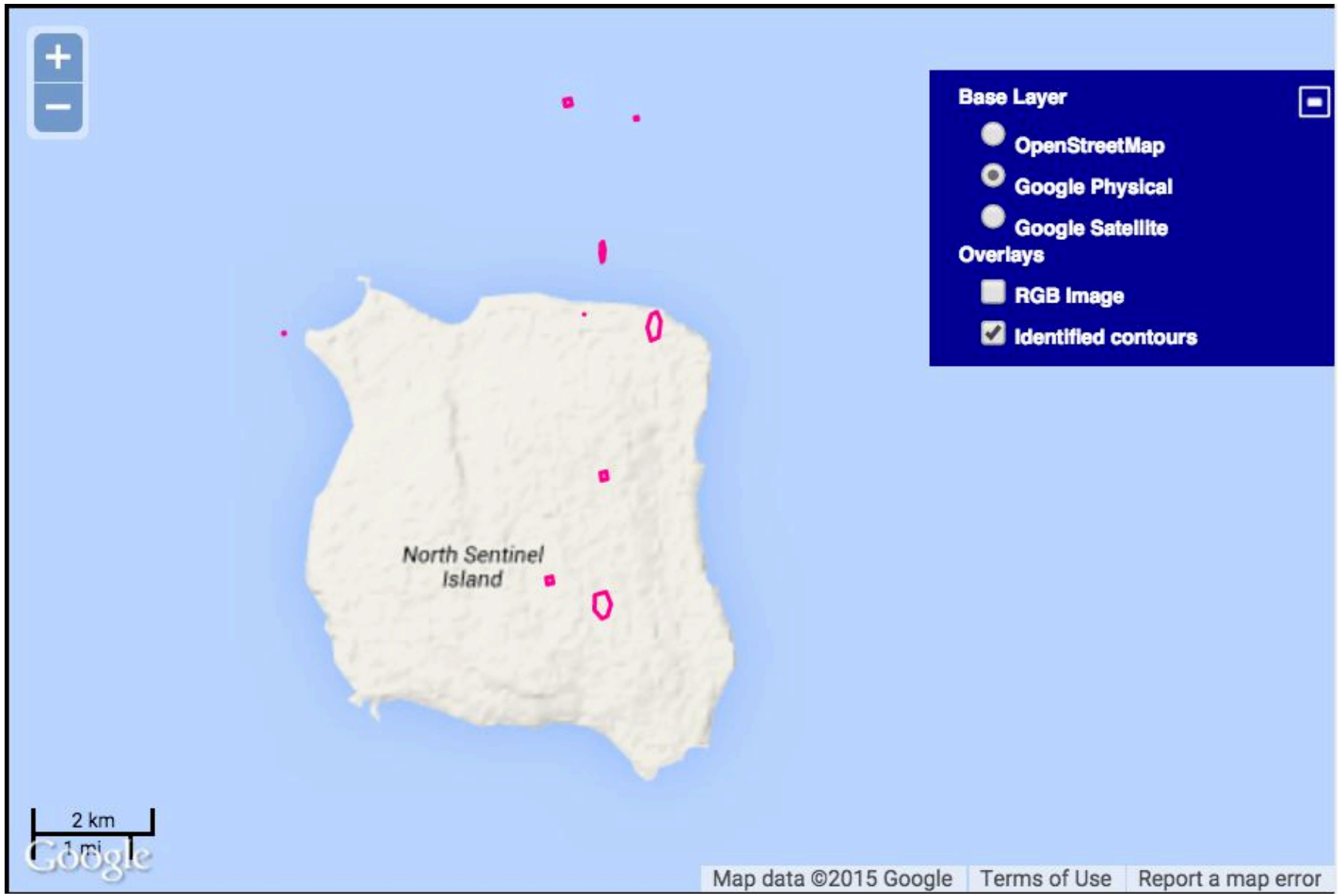
Contour ID	Cluster Score	Contour Score	lat,lng	Area (Pixels)	Area (Meters)	color
C1-05041-0KF	351	0.9719	140.886733625,27.2918559268	7.9589	6259.0137	COLOR
C1-05041-0KF	351	1.0807	140.897972808,27.3285963336	2447.4154	1925311.5337	COLOR
C1-05041-0KF	351	1.1266	140.899385769,27.3310296144	66.3332	52183.5335	COLOR
C1-05041-0KF	351	1.4893	140.900233529,27.3190516554	8.5744	6744.6581	COLOR
C1-05041-0KF	351	0.9264	140.902293378,27.3081518463	0.6165	484.8863	COLOR



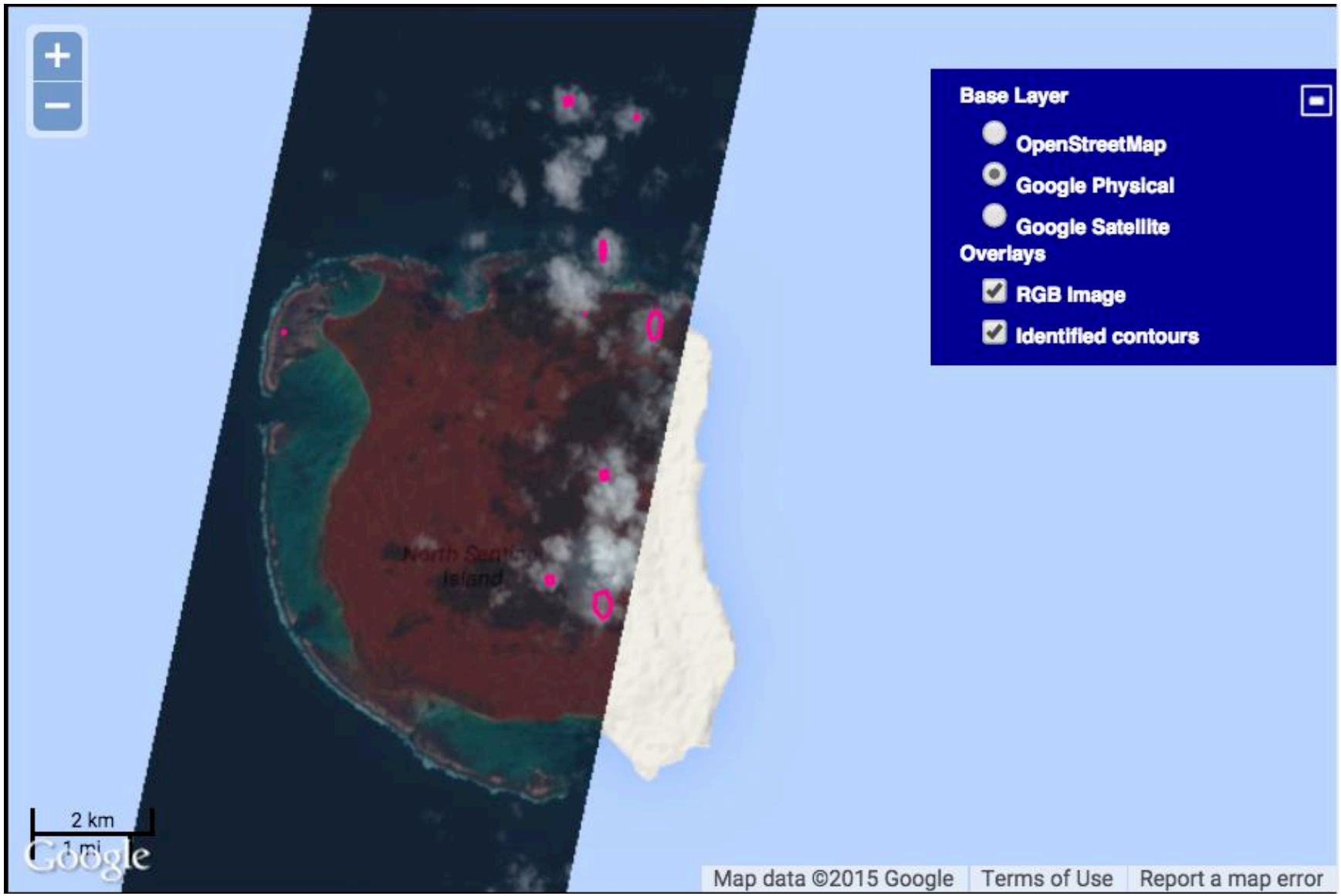
Spectral anomaly detected:
Barren Island active volcano, Feb, 2014



Spectral anomaly detected:
North Sentinel Island fires, May, 2014



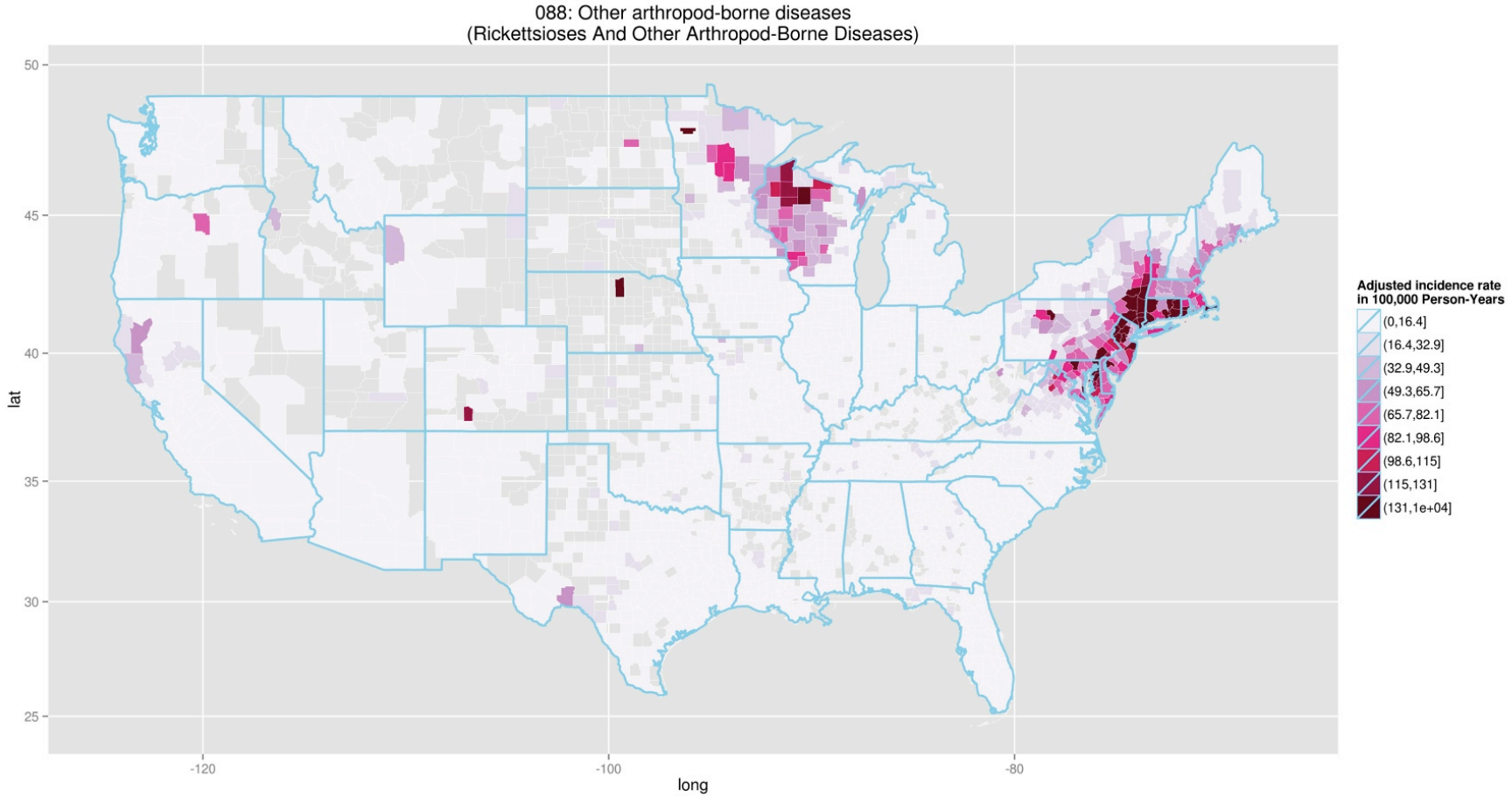
Spectral anomaly detected:
North Sentinel Island fires, May, 2014



Summary of Matsu

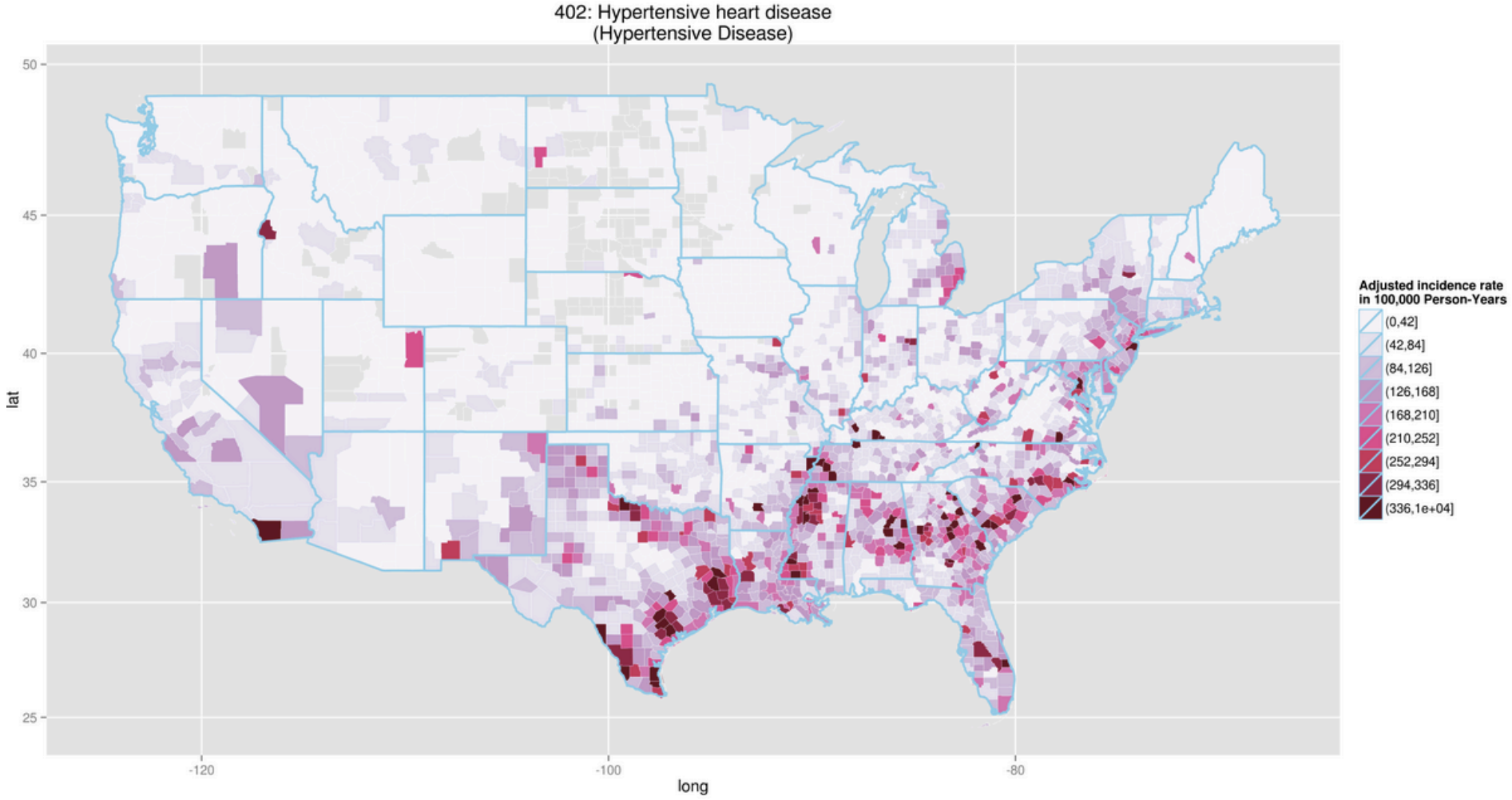
Analytic Infrastructure	Hadoop & Accumulo
Analytic algorithms	Clustering, PCA, custom distance function
Analytic operations	Daily list of candidate anomalies organized so that it can be scanned quickly by domain experts
Noteworthy	Special support for scanning queries allowing algorithms to be loaded and run automatically against all the data each day (“Matsu Wheel”)

Case Study 3: Geospatial Variation of Disease Incidence



Source: Neighbor Based Bootstrap (NB2) applied to 99.1 million electronic million medical records. Joint work with Maria Patterson and Andrey Rzhetsky, to appear.

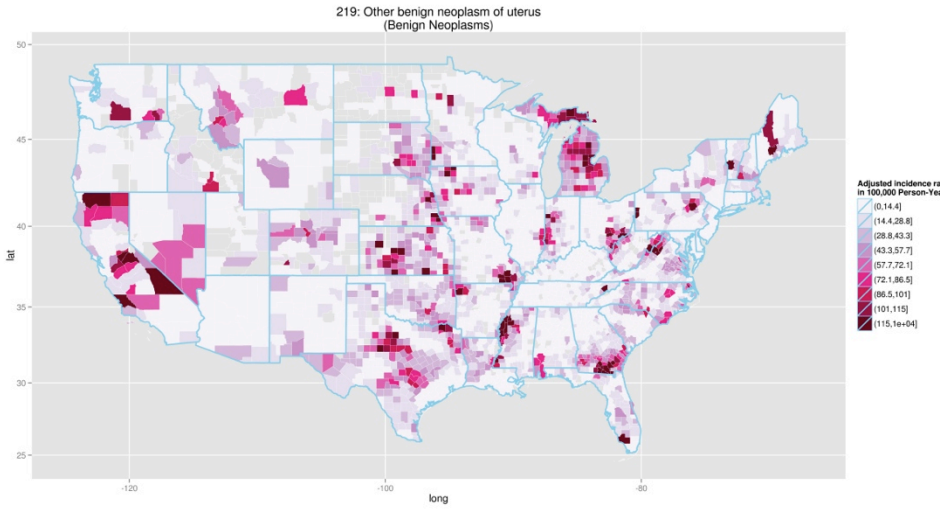
Hypertensive Heart Disease

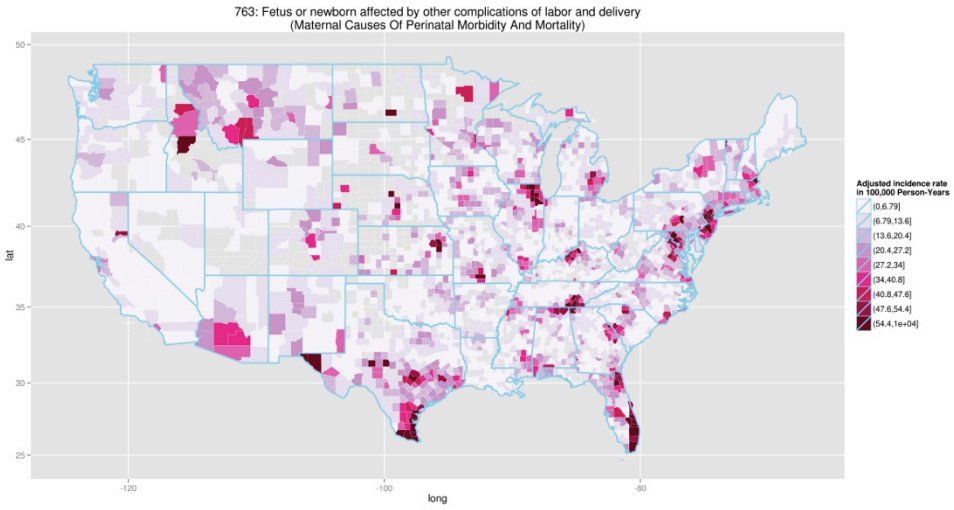


NB2 identifies small peaked clusters

Table 1. Top 25 ICD-9 codes as ranked by NB2

NB2	Moran	ICD-9 diagnosis name	ICD-9	Range	Sill
1	1.5	Hypertensive heart disease	402	1900	2.32
2	3	Trichomoniasis	131	610	1.33
3	5	Legally induced abortion	635	330	1.13
4	4	Other arthropod-borne diseases	088	580	1.41
5	7	Histoplasmosis	115	830	1.55
6	17	Other benign neoplasm of uterus	219	87	0.95
7	6	Angina pectoris	413	790	0.7
8	1.5	Nonallopathic lesions not elsewhere classified	739	490	0.59
9	26	Other disorders of prostate	602	120	0.81
10.5	21	Other venereal diseases	099	280	0.66
10.5	31	Disorders of tooth development and eruption	520	130	0.45
12	21	Ill-defined intestinal infections	009	190	0.64
13	10	Other acute and subacute forms of ischemic heart disease	411	1000	0.6
14	51.5	Fetus or newborn affected by other complications of labor and delivery	763	59	0.78
15	14.5	Other deficiency anemias	281	1100	0.69
16	44	Human immunodeficiency virus (HIV) infection	042	410	0.72
17	37	Long labor	662	98	1.01
18	21	Pulmonary congestion and hypostasis	514	480	0.58
19	26	Vitamin D deficiency	268	310	0.54
21	74.5	Other arthropod-borne viral diseases	066	370	1.34
21	9	Other diseases of endocardium	424	570	0.4
21	11	Influenza	487	830	0.46
23	44	Sarcoidosis	135	540	0.49
24	26	Other endocrine disorders	259	200	0.49
25	88	Chronic laryngitis and laryngotracheitis	476	54	0.62





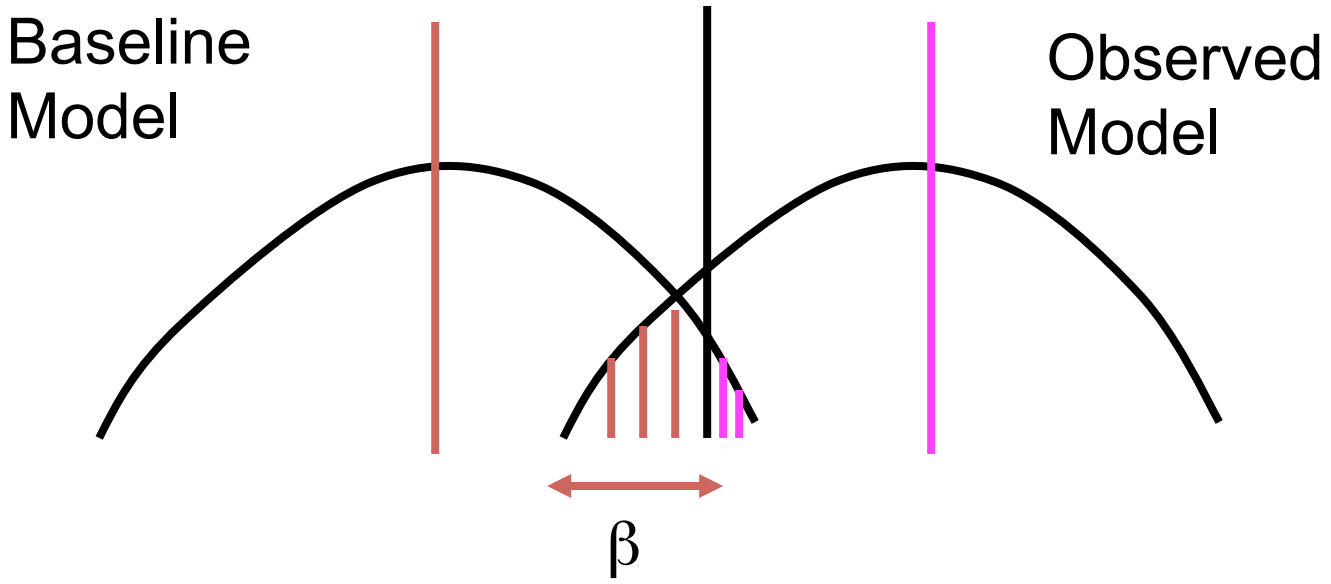
Summary of Disease Incidence Variation

Analytic Infrastructure	Private secure OpenStack cloud designed to hold biomedical data*
Analytic algorithms	Neighbor based bootstrap (NB2) method with significance checks using randomization. Comparison with other methods for detecting statistically significant geospatial variation.
Analytic operations	Visualization of results. Currently working out best way to review with subject matter experts.

*For more information, see Allison P. Heath, et. al. , Bionimbus: A Cloud for Managing, Analyzing and Sharing Large Genomics Datasets, Journal of the American Medical Informatics Association, 2014.

Case Study 4: Millions of POS Devices Sending Payment Information

If POS Data Were Homogeneous, We Could Use Change Detection Model

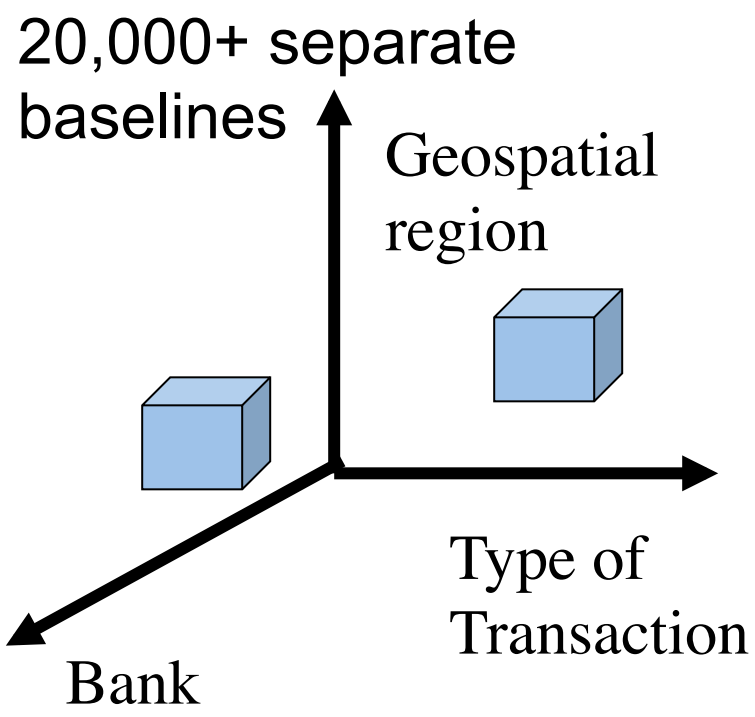


- Sequence of events x[1], x[2], x[3], …
- Question: is the observed distribution different than the baseline distribution?
- Use simple CUSUM & Generalized Likelihood Ratio (GLR) tests*

*H. Vincent Poor and Olympia Hadjiliadis, Quickest Detection, Cambridge University Press, 2008.

Build Thousands of Baseline Models, One for Each Cell in Data Cube

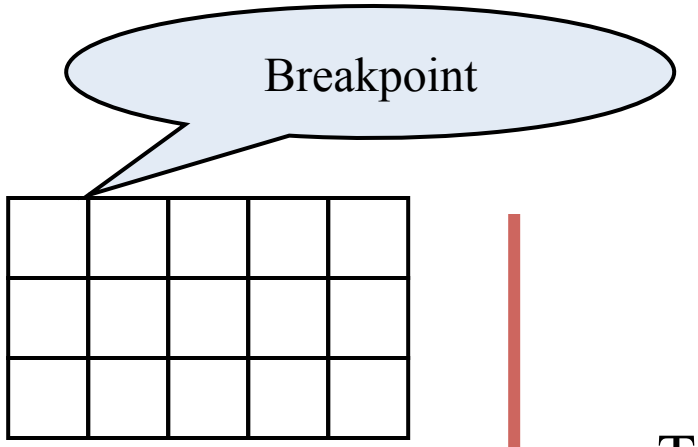
- Build separate model for each bank (1000+)
- Build separate model for each geographical region (6 regions)
- Build separate model for each different type of merchant (c. 800 types of merchants)
- For each distinct cell in cube, establish separate baselines for each metric of interest (declines, etc.)
- Detect changes from baselines



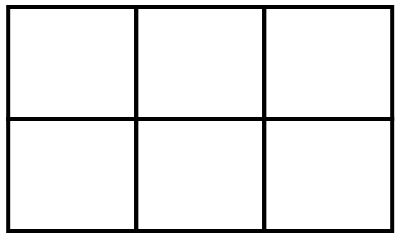
Source: Chris Curry, Robert L. Grossman, David Locke, Steve Vejcik, and Joseph Bugajski, Detecting changes in large data sets of payment card data: a case study, Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '07).

Balance Meaningful vs Manageable Number of Alerts

- More alerts
- Alerts more ***meaningful***
- To increase alerts, add breakpoint to split cubes, order by number of new alerts, & select one or more new breakpoints



One model for each cell in data cube

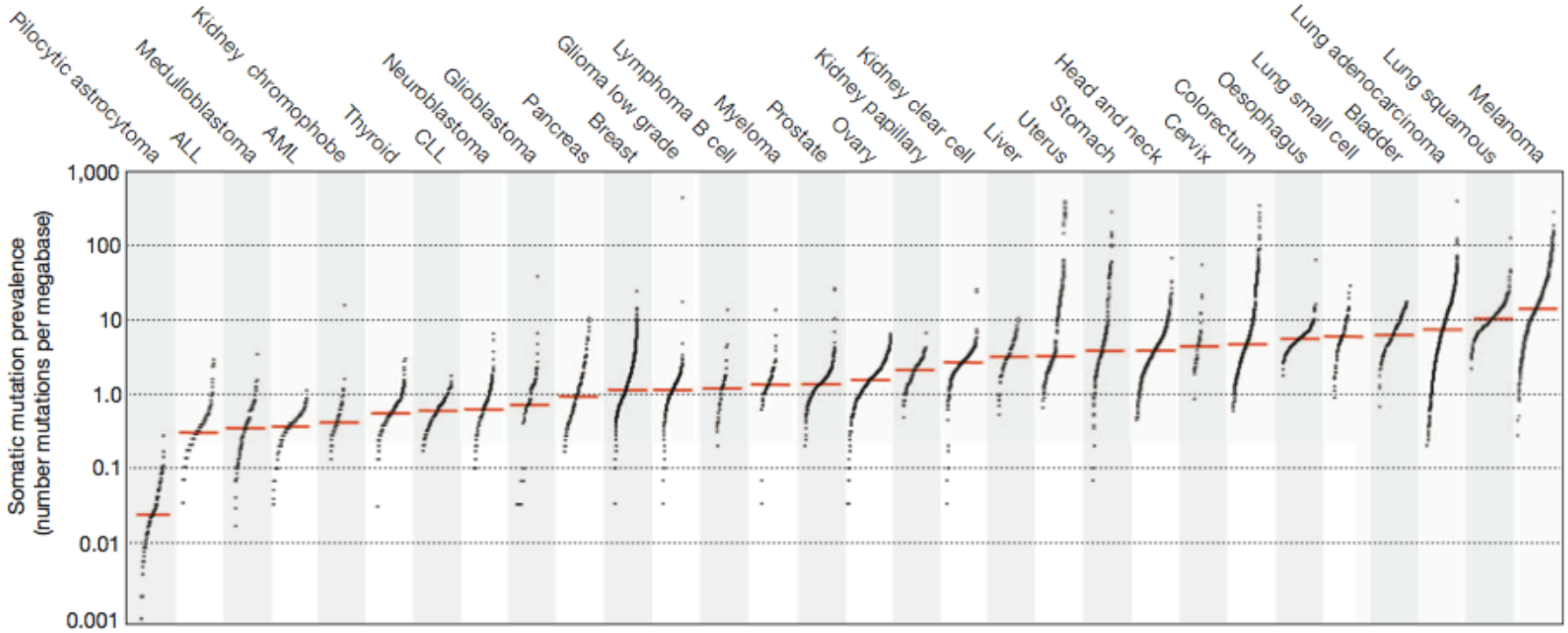


- Fewer alerts
- Alerts more ***manageable***
- To decrease alerts, remove breakpoint, order by number of decreased alerts, & select one or more breakpoints to remove

Summary of Cubes of Baseline Models

Analytic Infrastructure	Models were built in data warehouse and exported in the Predictive Model Markup Language (PMML). PMML models imported into real time scoring engine.
Analytic algorithms	Many segmented baseline models, one for each cell in a multi-dimensional data cube.
Analytic operations	Order most meaningful alerts each day, create custom visual report, and present to subject matter experts each day
Noteworthy	There is a successor to PMML called the Portable Format for Analytics (PFA) designed for today's big data environments.

Case Study 5: Mutations in Cancer and Their Variation



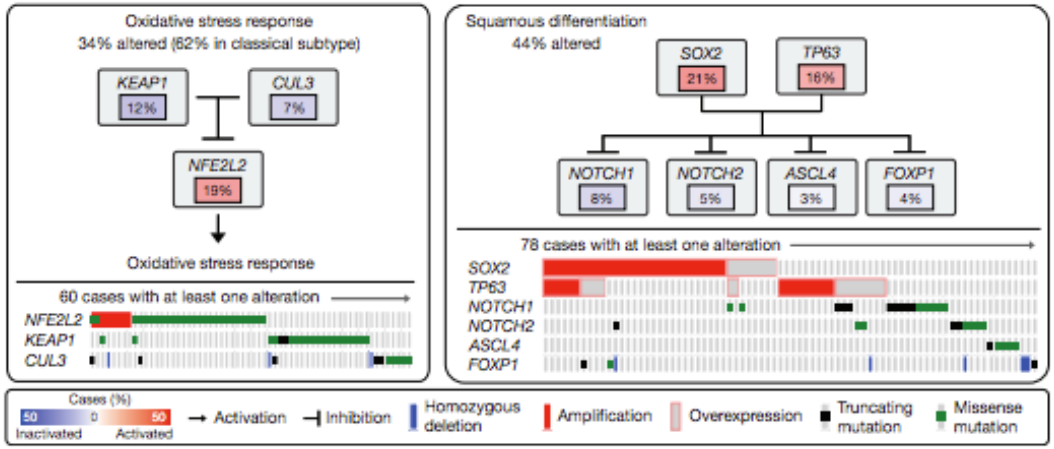
Source: Michael S. Lawrence, Petar Stojanov, Paz Polak, et. al., Mutational heterogeneity in cancer and the search for new cancer-associated genes, Nature 449, pages 214-218, 2013.

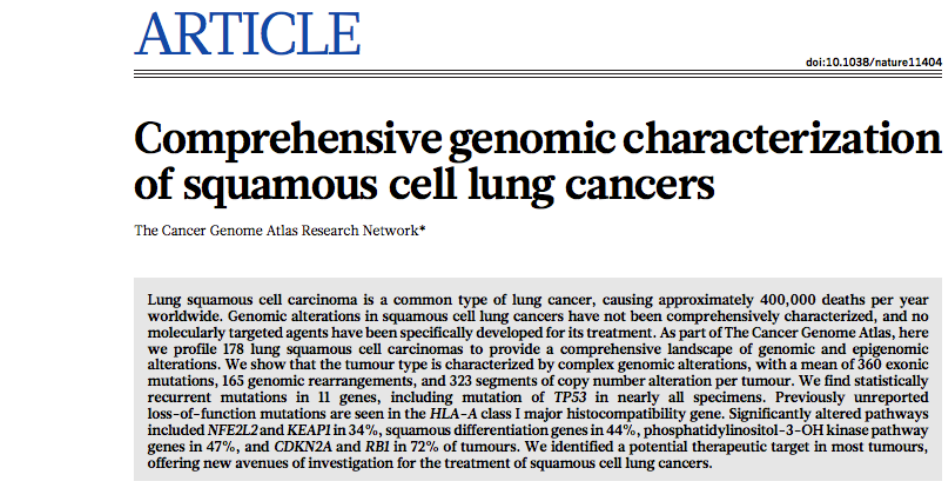
The Cancer Genome Atlas

Understanding genomics to improve cancer care

- The Cancer Genome Atlas (TCGA) is a large scale NIH funded project that is collecting and sequencing disease and normal tissue from 500 or more patients x 20 cancers.
- Currently about 12,000 patients are available.
- There is about 4 PB of research data today and growing.

TCGA Analysis of Lung Cancer





ARTICLE

doi:10.1038/nature11404

Comprehensive genomic characterization of squamous cell lung cancers

The Cancer Genome Atlas Research Network*

Lung squamous cell carcinoma is a common type of lung cancer, causing approximately 400,000 deaths per year worldwide. Genomic alterations in squamous cell lung cancers have not been comprehensively characterized, and no molecularly targeted agents have been specifically developed for its treatment. As part of The Cancer Genome Atlas, here we profile 178 lung squamous cell carcinomas to provide a comprehensive landscape of genomic and epigenomic alterations. We show that the tumour type is characterized by complex genomic alterations, with a mean of 360 exonic mutations, 165 genomic rearrangements, and 323 segments of copy number alteration per tumour. We find statistically recurrent mutations in 11 genes, including mutation of *TP53* in nearly all specimens. Previously unreported loss-of-function mutations are seen in the *HLA-A* class I major histocompatibility gene. Significantly altered pathways included *NFE2L2* and *KEAP1* in 34%, squamous differentiation genes in 44%, phosphatidylinositol-3-OH kinase pathway genes in 47%, and *CDKN2A* and *RB1* in 72% of tumours. We identified a potential therapeutic target in most tumours, offering new avenues of investigation for the treatment of squamous cell lung cancers.

- 178 cases of SQCC (lung cancer)
- Matched tumor & normal
- Mean of 360 exonic mutations, 323 CNV, & 165 rearrangements per tumor
- Tumors also vary spatially and temporally.

Source: The Cancer Genome Atlas Research Network, Comprehensive genomic characterization of squamous cell lung cancers, Nature, 2012, doi:10.1038/nature11404.

How Can This Be Operationalized?

- Given a sequenced tumor, nearby tumors with similar genomic signatures can be returned along with the associated clinical data, including treatments and survival curves.

Summary of Tumor Signatures

Analytic Infrastructure	Large scale computation to “harmonize” existing disparate datasets so that resulting dataset can be analyzed more easily by the research community.
Analytic algorithms	Number of specialized algorithms developed by different research groups for identifying mutations from tumor-normal pairs of sequenced data.
Analytic operations	Still being developed.
Noteworthy	An example of a PB-scale “data commons” that is being developed for the biomedical research community.

Summary

When You Have No Labeled Data and Lots of It – Approach 1

1. Create clusters or micro-clusters
2. Manually curate some of the clusters with tags or labels (often with "orthogonal" data)
3. Produce scores from 0 to 1000 based upon distances to interesting clusters
4. Visualize
5. Discuss weekly with subject matter experts and use these sessions to improve the model.

Case Study 2: Detecting anomalies in NASA's EO-1 hyperspectral data.

When You Have No Labeled Data and Lots of It – Approach 2

1. Create clusters or micro-clusters
2. Rank order the findings from most significant to least significant.
3. Compare your findings with other ranking methods. Enrich with other covariates.
4. Visualize
5. Discuss weekly with subject matter experts and use these sessions to improve the model.

Case Study 3: Ranking incidence levels of diseases by their geo-spatial variation.

When You Have No Labeled Data and Lots of It – Approach 3

1. Build baseline models for *each entity* of interest, even if thousands of millions of models.
2. Produce scores from 0 to 1000 based upon deviation of observed behavior from baseline
3. Visualize
4. Discuss weekly with subject matter experts and use these sessions to improve the model.

Case Study 4: Detecting anomalies in payment data using baseline models for millions POS devices.

Questions?

home page: rgrossman.com