Lecture 17 – Data-streaming in Real time: Survey Analysis and Sliding Time Series Windows

Linkup – Take Control of Your Job Search

Every day we index employer websites for real job listings. We're committed to providing an accurate, high-quality job search so you won't find old, duplicated, or spammy listings here. Our job seeker tools will help you organize and automate your search. Create an account on LinkUp for additional tools and features.

Our jobs data is the most comprehensive, detailed, timely, and accurate labor dataset in the market.

- Jobs indexed from 35,000 companies and growing
- 120 million jobs collected to date
- 4 million jobs published daily
- Represents entire economy
- Over 10 years of history
- Index updated and delivered daily
- Global coverage

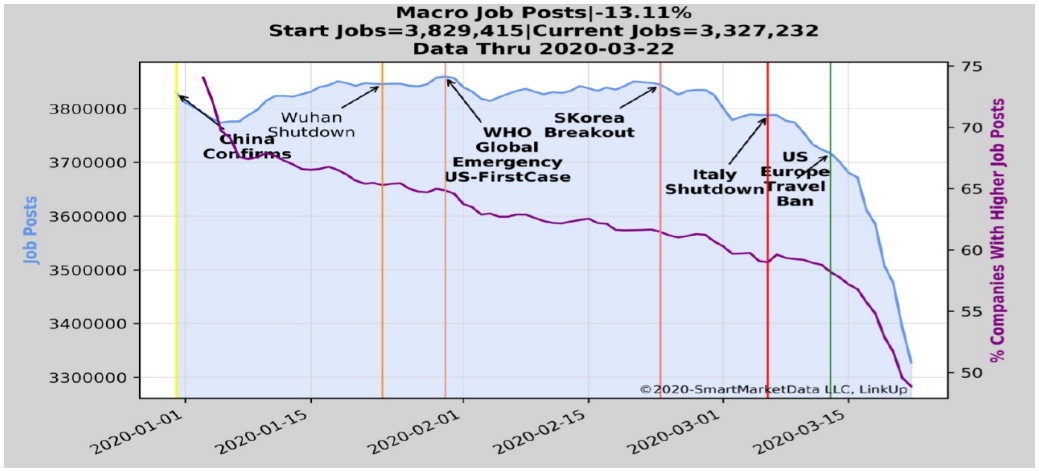
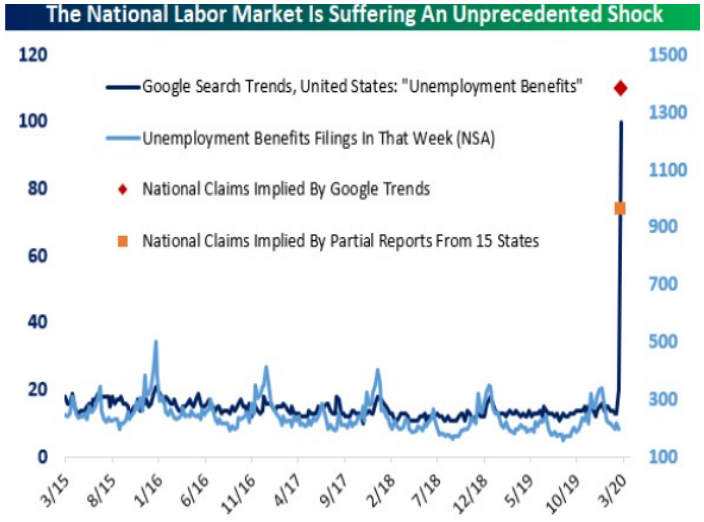


Table 1-Travel and Leisure % Change in Job Posts and Stock Correlation

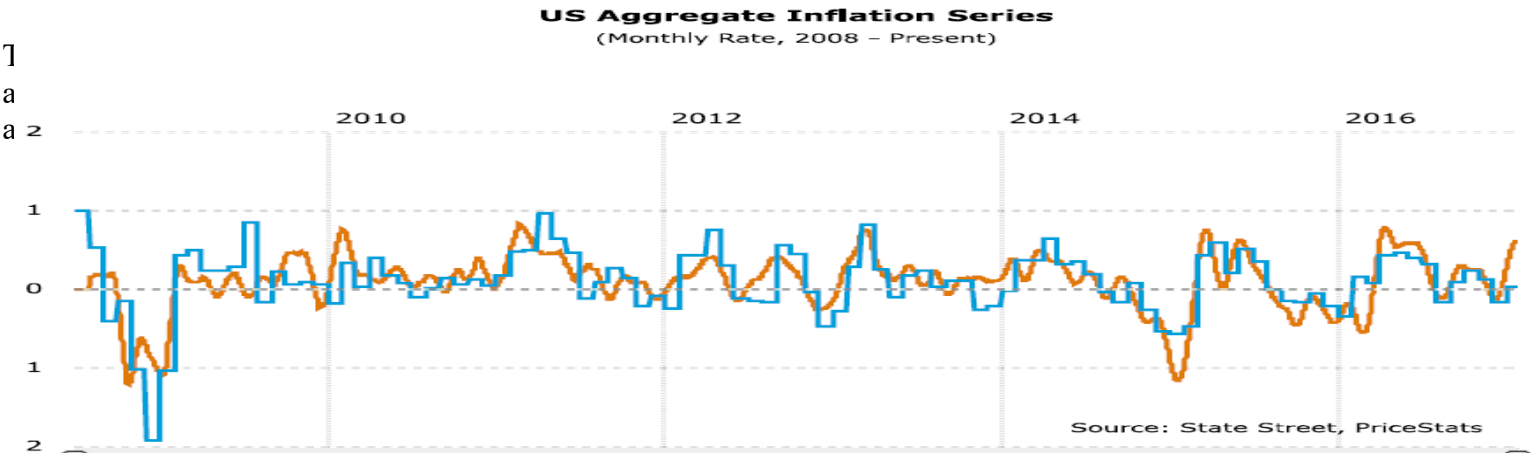
company	ticker	jobs change	stock change	corr
United	UAL.OQ	-100%	-73%	0.93
Southwest	LUV.N	-97%	-29%	0.39
JetBlue	JBLU.OQ	-94%	-67%	0.76
Uber	UBER.N	-93%	-48%	0.56
Hawaiian	HA.OQ	-93%	-72%	0.76
Qantas	QAN.AX	-92%	-61%	0.40
International Air Group	ICAG.L	-91%		
Hertz	HTZ.N	-89%	-75%	0.62
American	AAL.OQ	-87%	-66%	0.30
Expedia	EXPE.OQ	-83%	-60%	0.72
Easyjet	EZJ.L	-82%	-61%	0.47
Hilton	HLT.N	-81%	-46%	0.88
Norwegian	NCLH.N	-78%	-85%	0.59
Marriott	MAR.OQ	-76%	-51%	0.81
Allegiant	ALGT.OQ	-71%	-59%	0.66
Booking	BKNG.OQ	-70%	-44%	0.67
Air France KLM	AIRF.PA	-67%		



The Billion Prices Project @MIT:collects daily prices on products sold by online retailers using a software that scans the underlying code in public webpages and stores retailers... includes information on product descriptions, package sizes, brands, special characteristics (e.g. "organic"), and whether the item is on sale or price control. They construct daily inflation indexes and study their ability to match official statistics. Interested in:

Pricing Behavior: What drives price stickiness around the world? How much can be explained by current inflation, and inflation histories? How much by competition and industries' structure? Are prices synchronized?

Daily Inflation and Asset Prices:. **Pass-Through**: How much do prices adjust when the exchange rate, or the international price of commodities change?



Cavallo (2017) "Are Online and Offline Prices Similar? Evidence from Large Multi-Channel Retailers" - American Economic Review - Vol. 107(1), p.283–303

Online prices are increasingly being used for measurement and research applications, yet little is known about their relation to prices collected offline, where most retail transactions take place. I conduct the first large-scale comparison of prices simultaneously collected from the websites and physical stores of 56 large multi-channel retailers in 10 countries. I find that price levels are identical about 72 percent of the time. Price changes are not synchronized but have similar frequencies and average sizes. These results have implications for National Statistical Offices, researchers using online data, and anyone interested in the effect of the Internet on retail prices.

TABLE 3—COUNTRY: PRICE-LEVEL DIFFERENCES (*Percent*)

Country	Retailers (1)	Observations (2)	Identical (3)	Higher online (4)	Lower online (5)	Online markup (6)	Online difference (7)
Argentina	5	3,699	60	27	13	3	1
Australia	4	3,797	74	20	5	5	1
Brazil	5	1,915	42	18	40	−7	−4
Canada	5	4,031	91	3	5	−5	0
China	2	513	87	7	6	3	0
Germany	5	1,604	74	4	23	−8	−2
Japan	4	2,186	48	7	45	−13	−7
South Africa	5	3,212	85	6	9	−3	−1
United Kingdom	4	2,094	91	2	7	−8	−1
United States	17	15,332	69	8	22	−5	−1
All countries	56	38,383	72	11	18	−4	−1

Notes: Column 3 shows the percentage of observations that have identical online and offline prices. Column 4 shows the percent of observation where prices are higher online and column 5 the percentage of prices that are lower online. Column 6 shows the online markup, defined as the average price difference excluding cases that are identical. Column 7 shows the average price difference including identical prices.

TABLE 6—SYNCHRONIZED PRICE CHANGES

	Observations (1)	Price changes (2)	Synchronized price changes (percent) (3)	Unconditional probability (percent) (4)
Argentina	1,392	245	35	2.0
Australia	759	72	22	0.5
Brazil	483	85	18	2.3
Canada	1,427	120	32	0.5
Germany	419	16	31	0.1
Japan	1,071	98	1	0.1
South Africa	882	109	15	0.8
United Kingdom	429	25	44	0.3
United States	7,505	563	11	0.2
All countries	14,367	1,328	19	0.5

Notes: China is excluded due to lack of price change data. Column 3 reports the percentage of price changes for a given product that occur both online and offline at the same time, which I refer to as "synchronized." The unconditional probability of a synchronized price change in column 4 is obtained by multiplying the frequencies of price change in Table 5.

TABLE 7—ONLINE-OFFLINE PRICE-LEVEL DIFFERENCES FOR MULTIPLE ZIP CODES (*Percent*)

Country	Retailers (1)	Observations (2)	Identical (3)	Higher online (4)	Lower online (5)	Online markup (6)	Online difference (7)
United States	9	406	60	11	29	−4	−2
Different offline	7	85	35	16	48	−5	−3
Identical offline	8	316	67	9	24	−3	−1

Notes: Column 3 shows the percentage of observations that have identical online and offline prices. Column 4 has the percent of observations where prices are higher online and column 5 the percentage of prices that are lower online. Column 6, is the online markup, defined as the average price difference excluding cases that are identical. Column 7 is the average price difference including identical prices.

Data-mining real time data – aka **data streaming** – Needs on-Line learning algorithms to reflect/respond to changes in variables/structure in short time. Since decision-makers want to respond quickly, tools require ability to switch models and variables rapidly. This often requires high level representation of the data rather than the original raw data. – reducing time series into fewer dimensions by variable selection/clustering – and extracting and interpreting information quickly subject to constraint of memory and time.

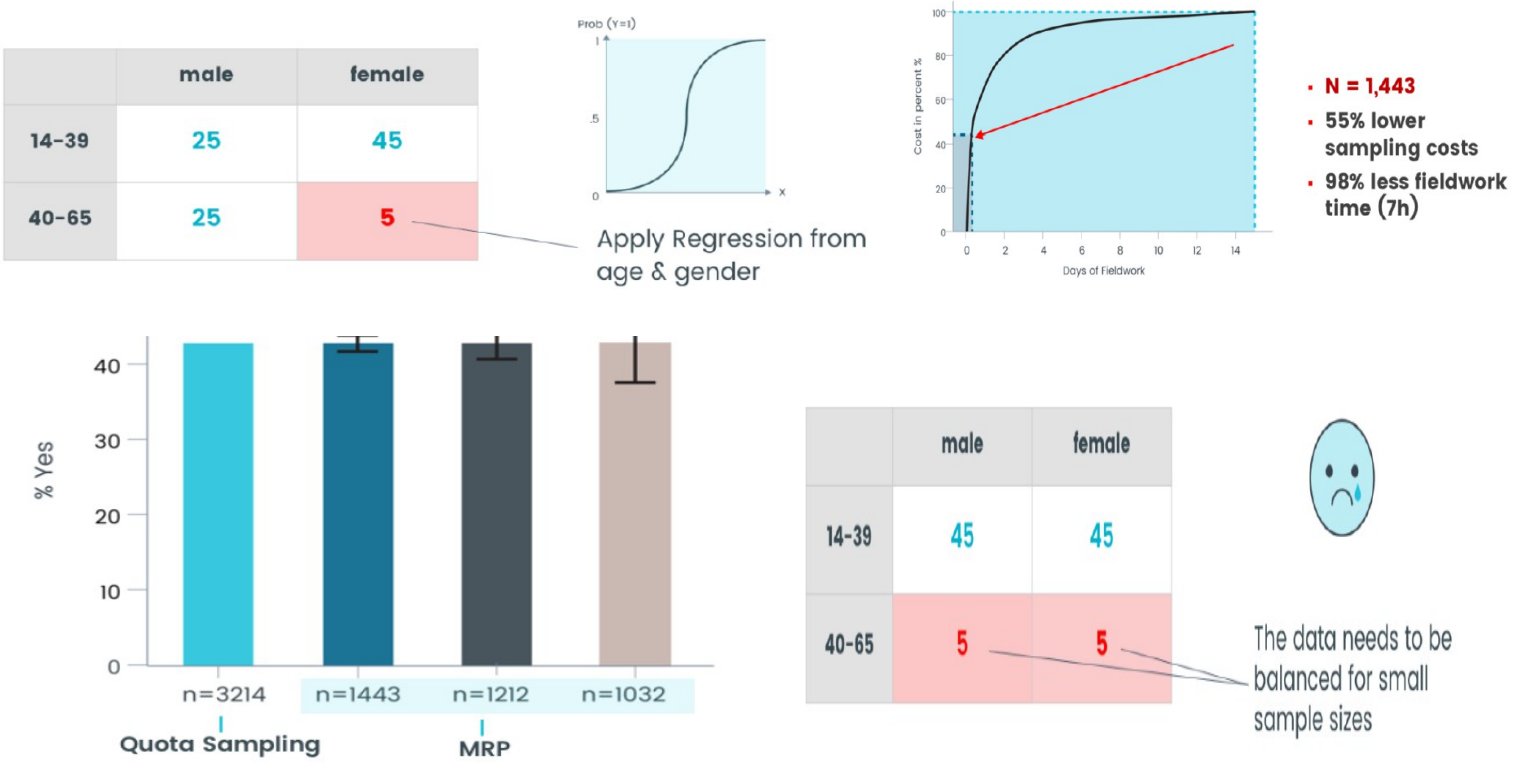
You believe the world operates according to model A but the data are changing to suggest the system is moving in a different direction – **a discontinuous change in a short time period**, which will out-mode your current way of operating. You must deal with the actual situation not the situation that existed before.

Changes in model are called **concept drift** … the properties of the target variable (concept) have changed (drift). Change could be sharp or gradual and could easily be confused with random fluctuation. Problem is big in stock market forecasting, consumer fashion, in climatology. Does recent weather/variability evidence of global warming or just normal variability? And in polling.

Cell phone Surveys as way to track sentiment.

Standard technology is through phone surveys often through a fixed panel or quota sampling. A fixed panel are set of people who agree to respond to questions over time. Quota sampling is to decide on representative characteristics and use some on-line sampling to fill the quota. Problem in both is selectivity, especially for groups that your design has trouble capturing.

Dalia, German start-up: "Instead of using a panel with a fixed number of respondents, Dalia continuously sources new survey respondents from tens of thousands of online apps and mobile websites. This enables Dalia to reach hundreds of millions of people ...in real-time. But this has problems --> **quota sampling**."Quota sampling builds a representative sample by dividing up the population into groups (usually based on demographics of age, gender, geography, income and education). Survey researchers need to reach enough respondents in each group before they stop fieldwork to obtain completed sample distribution matches the general population's distribution based on census statistics. DeVeaux &Oswald "How to cut fieldwork time and costs in half while increasing data quality" claims that Multilevel Regression and Poststratification **MRP** gets comparable survey estimates to quota sampling, but drastically cuts costs by 55% and fieldwork from 15 days to 7 hours, by estimating the cells with just a few people from regression rather than keeping surveying at big cost to reach quota in block.



This is related to James-Stein estimator in statistics – a biased estimator that has **lower mean squared error** than "ordinary" least squares. Intuition – in first ten games red sox won n 2019. This is unbiased but if you want to predict win % over year, more likely would be weighted average of 10% and … 50% for all games …. 65% for Red sox last year … etc. Hhrink toward mean from larger group.

Problem: Can Dalia be sure that its MRP works as well for **all** surveys? No. Depends on smoothness of sample space – VC dimension and concentration of measure type analysis. If hard-to-survey group is very different from others, regression analysis will not work well.

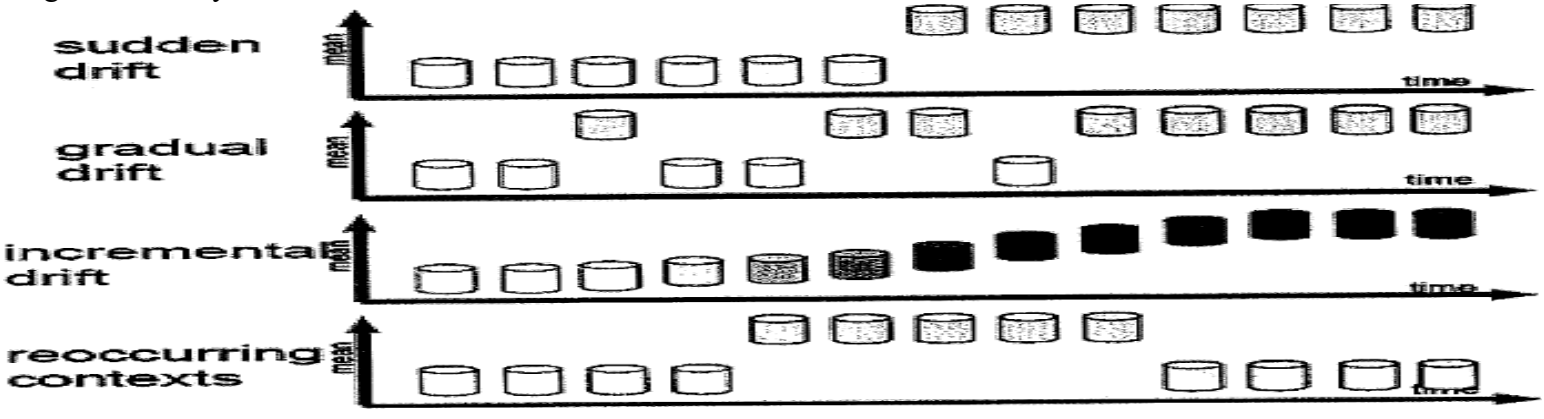


Figure 4: Illustration of the four structural types of the drift.

In **sudden drift**, you should change your model fast; **gradual drift** stick with old model and maybe have both operating in transition period. **Incremental** – have model gradually change; **reoccurring** is cycle. Similar problem with non-time series data with natural ordering: Take income for persons ranked from lowest to highest. As you move along the data you find that at lower earnings the distribution fits a log normal but at higher incomes, find more than lognormal -->power law/Pareto distributed. You want to estimate when lognormal ended and power law began.

Update/refresh models in real time or change to new model? There is a trade-off between using data to get a better fix on current model (if it is right, as N--> ∞ the variance of estimate falls and parameters get closer to truth, which makes model better; vs using data to shift to a new model to capture change in the world. New model often based on small number of data points. Do you check athletic scores every day or every minute as games go on and adjust your bets accordingly? To learn from streaming data, need to decide if/when to shift and/or how much of past to include in your model if you shift.

Detecting when a change occurs requires a trigger that says "world is different"

Example 1: – you smooch with favorite person at 3 after classes but one day instead of smooch you get a shove or slap. World had changed. Is it because HHS has issued SMOOCH alert about new VIRUS? Or did you forget Daylight Savings Time? Or … Did you confuse favorite person with identical twin look-alike? All you know is that something is different. Your best response will differ with your theory of what is different.

Example 2: You are watching corona-virus cases to decide when to call off health emergency and tell people to go back to work. The virus cases are bouncing around more than normal. If this is because the testing is reaching riskier groups. Or maybe it is just a random variation of extra variability.

Ways to deal with this POSSIBLE change in universe.

Fixed period sample weighting with greater weights to most recent events. Could use most recent (week, months, years) or take a weighted average that gives lots of weight to current period and little weight to past. Thinking about next recession, analysts don't go back to study the PANIC of 1819, the first major financial crisis in the US, for guidance (unlike climatology studies which go back to ice age to try to understand what may happen to us). Before Great Recession many dismissed the Great Depression as irrelevant on the notion that good macro-economic performance and policies – the Great Moderation – and modern financial tools that spread risk widely effectively eliminated it (Shiller, The New Financial Order). From search theory we know that fixed sample design is not right. From Minority game we know that length of memory in decisions affects outcomes.

But … From search theory we know that **fixed sample design is not right**. From Minority game we know that length of memory in decisions affects outcomes.

So, let's try **sliding windows.** We determine the size of the window – length of time – that we use to estimate a model based on the variation in the data. There are two errors we can make: The world has changed and we reject that in favor of the null hypothesis that world has not changed (Type I error in statistics where null is nothing new under sun). Big blip in today's cases of corona-virus is just usual noise in the data and tomorrow will be like all other days. OR we might claim the world has changed when it has not (Type II error – in which we reject the null). Chicken Little.

Sliding window is a sequential sample. Window increases when the data stream is stationary to get more accurate measure of the current model but shrinks when the data is jumpy on notion that a period of abnormal high variation is sign that world is changing. You forget earlier events when you shrink a window. "decremental non-learning". The algorithm automatically increases window when data stream is stationary to get more accurate estimates of current model but shrinks window when things change so you ignore 1819 recession. You do not specify a fixed window/sample size but let the algorithm make sequential decision.

Given the uncertainty, it is often useful to apply an ensemble of models based on different features of the world to try to minimize generalization error on the notion that if you diverse models/experts you have a better chance of dealing with change. **Ensemble methods** are popular in data mining due in part to their empirical effectiveness.

ENSEMBLE LEARNING uses an ensemble of models based on different features of world to try to minimize generalization error on the notion that if diverse models/experts has better chance of dealing with change. You pick your team to include some aberrant views. Diversity of views has an advantage in that if you shift to "new world" you have a model that predicts well for it.

How an ensemble works better:

						#Wrong
REAL	1	0	1	1	0	
Expert 1	1	0	1	0	1	2
Expert 2	0	0	0	1	0	2
Expert 3	0	1	1	1	0	2
ENSEMBLE	0	0	1	1	0	1

You can do a lot with this type of example. Note that experts 1 and 3 differed in how they erred. Expert 1 and 2 predicted the right number of events but got the timing wrong. Expert 2 predicted just one event but got it right and therefore failed by not seeing two of the events.

You could play with adding a 4th expert, who would have a different error structure. Note that a ZERO INTELLIGENCE agent that predicted a 1 in all periods would also make 2 errors!

But the issue of streaming data is that the real model for this example would change so maybe the next 5 data points come from a world with 0 REAL events + some error so you might see 1 event by chance.

Ensembles of classifiers based on different feature subsets that seek not only to get accuracy of individual error but disagree on some parts of the world do better. The idea is that model A is best in some world and model B is best in some other world. YOU PICK YOUR TEAM NOT ONLY TO GIVE BEST PREDICTIONS BUT ALSO TO HAVE SOME DISAGREEMENT. **DIVERSITY OF VIEWS PAYS OFF.**

The most common source of diversity in machine learning is to train different experts/predictors on different parts of the training data. This is often done by bootstrapping (sampling with replacement) different training sets. The main theme of analyzes is that good results require disagreement. If all classifiers/expert systems make the same mistakes there is no value to an ensemble. You want to keep the odd voice on your team so that if it turns out that the world changes in the way the odd person sees it, you can adjust quickly.

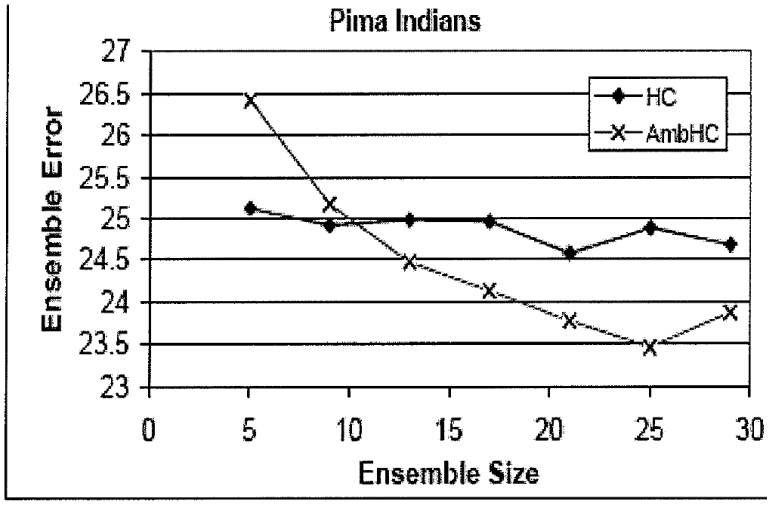
This is mindful of simulated annealing. You go in the wrong direction – keep the model that has failed in the past – because in a new environment that model may save the day.

Using Diversity in Preparing Ensembles of Classifiers Based on Different Feature Subsets to Minimize Generalization Error Gabriele Zenobi, Pádraig Cunningham

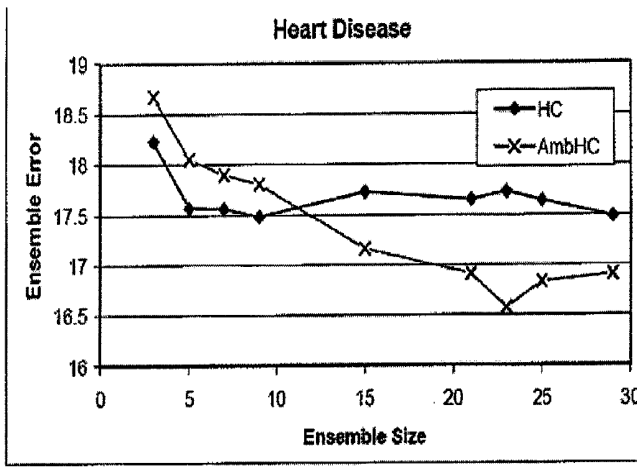
They define an Ambiguity algorithm which is based on how far a particular model deviates from the others. Instead of choosing models that work best, which come to resemble each other, they keep in the ensemble of expert models, some that diverge.

Two examples – PIMA Indians Diabetes dataset
Cleveland Heart Disease Database

Hill climbing strategy (HC) is normal iteration with maximand to get the best you can with the group-- think of going straight up some mountain and getting caught on a local maximum, say in genetic algorithm
Ambiguity hill climbing (AHC) you give reward for keeping some aberrant views in your ensemble.



Generalisaton error of different ensemble sizes on the Pima data..



Generalization error of different ensemble sizes on the Heart data.

" It shows that it can be a good thing to have a committee of experts consistently voting 5 : 4 in favour of a prediction rather than 8 : 1. In fact, we are proposing selecting experts in a manner that will push down consensus in the committee. Intuitively, this is not what you want from a committee of physicians discussing your particular illness! You would like the committee of physicians to agree. A resolution of this paradox is as follows. If the committee members are very accurate there is little benefit in diversity; indeed there is little benefit in ensembles in classification tasks where accuracies of >93% are achievable with a single classifier. However, ensembles make sense where individual classifiers have significant errors (say > 15%). In such cases, instead of adding a new very accurate committee member that makes the same errors as existing members in the ensemble it is sensible to add a member that makes different errors, one that has a different set of competences. There is no benefit in adding members that will change votes of 8:1 to 9:1.

NOTE: IMPLICATION FOR DECISION-MAKERS

You want a committee/cabinet/advisors with diverse views.
You do not want people who agree with you on almost all points
or who disagree only in their heads because if they speak up they will get fired/arrested/banished.

Or you can have it "The Company Way" from How to succeed in business without Really Trying
MR. TWIMBLE: When I joined this firm As a brash young man.
Well I said to myself not brash young man. Don't get any ideas.
Well, I stuck to that And haven't had one in years. FINCH: You play it safe!

MR. TWIMBLE:I play it the company way. Wherever the company puts me. There I stay. FINCH:But what's your point of view?

MR. TWIMBLE:I have no point of view! FINCH: Supposing the company thinks that...
MR. TWIMBLE:I think so too! FINCH:Now, what would you say...

MR. TWIMBLE: I wouldn't say! FINCH: Your face is a company face...
MR. TWIMBLE:It smiles at executives then goes back in place!

FINCH:The company furniture...MR. TWIMBLE::it suits me fine!
FINCH:The company letterheader... MR. TWIMBLE:A Valentine!
...
FINCH:So, You play it the company way. MR. TWIMBLE: Oh, company policy is by me ok.
FINCH: Your brain is a company brain. MR. TWIMBLE: The company washed it,
MR. TWIMBLE:'Cause I play it the company way.

FishEye Patents Real-Time Processing of Streaming Data

Posted by John R Crowley Jr on Aug 24, 2017 in About FishEye, News, Real-Time Platform (RTTK)

IMMEDIATE RELEASE

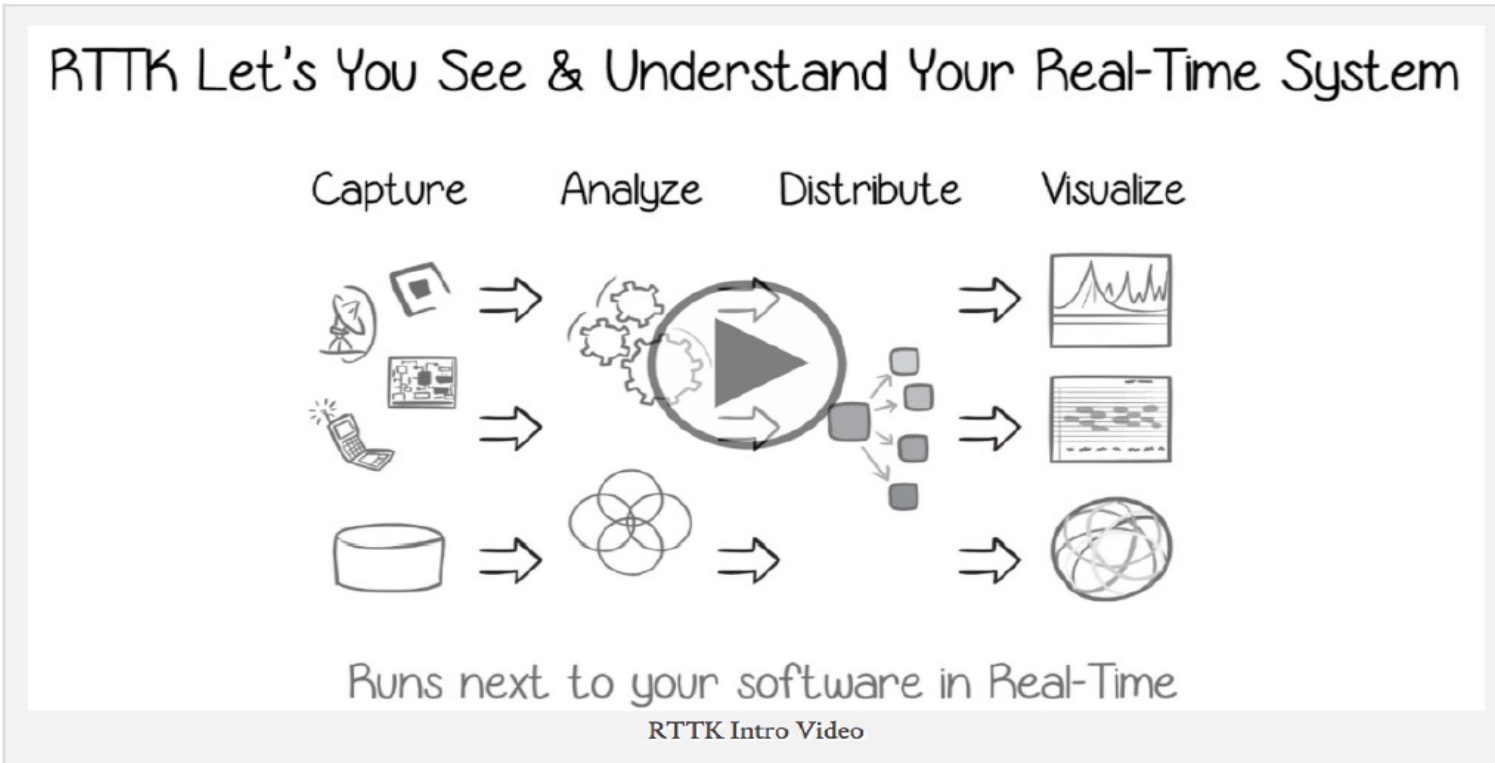
MAYNARD, MA — (NASDAQ GLOBENEWSWIRE) — Sept. 12, 2017 — FishEye Products, LLC, a leading provider of real-time systems, has been awarded a patent for their Real-Time Platform (RTTK). The invention defines a process that looks inside applications to extract knowledge to unlock the fastest access to real-time machine data and real-time analytics.

Real-time embedded systems have hard deadlines, are complex, and must share machine data across varying hardware. These systems are hard to understand making them difficult and expensive to make, debug, validate, and keep operational. The patent radically simplifies the process of understanding fast and complex systems and does it in real-time. Just like a cardiologist assesses and analyze heart performance with an EKG and a blood pressure monitor, engineers can see inside and understand what is going on while complex systems run.

The innovative approach radically reduces systems full-life cycle costs and risk by simplifying how machine data is accessed and shared. The patent enables real-time systems to operate safer and faster, while producing new insight instantly. Systems that keep our society functioning benefit from real-time data analysis to predict, manage and troubleshoot anomalies before they have an impact. Power generation, railway, manufacturing, and defense are just a few of the core industries in which the patent can drive major improvements.

BOUT FISHEYE FishEye Software, Inc. is a leading provider of real-time system products and software engineering headquartered in the greater Boston, Massachusetts area. Since 1997, the company has been developing, integrating and testing mission-critical and operational software for government and commercial customers in systems like phased-array radars, air-traffic control, missile defense, and command and control. The FishEye Real-Time Platform (RTTK) is a real-time system which provides real-time data capture, real-time analytics and more. The FishEye Radar Simulator (VREX) simulates complex radar in real-time. Stay informed with FishEye News.

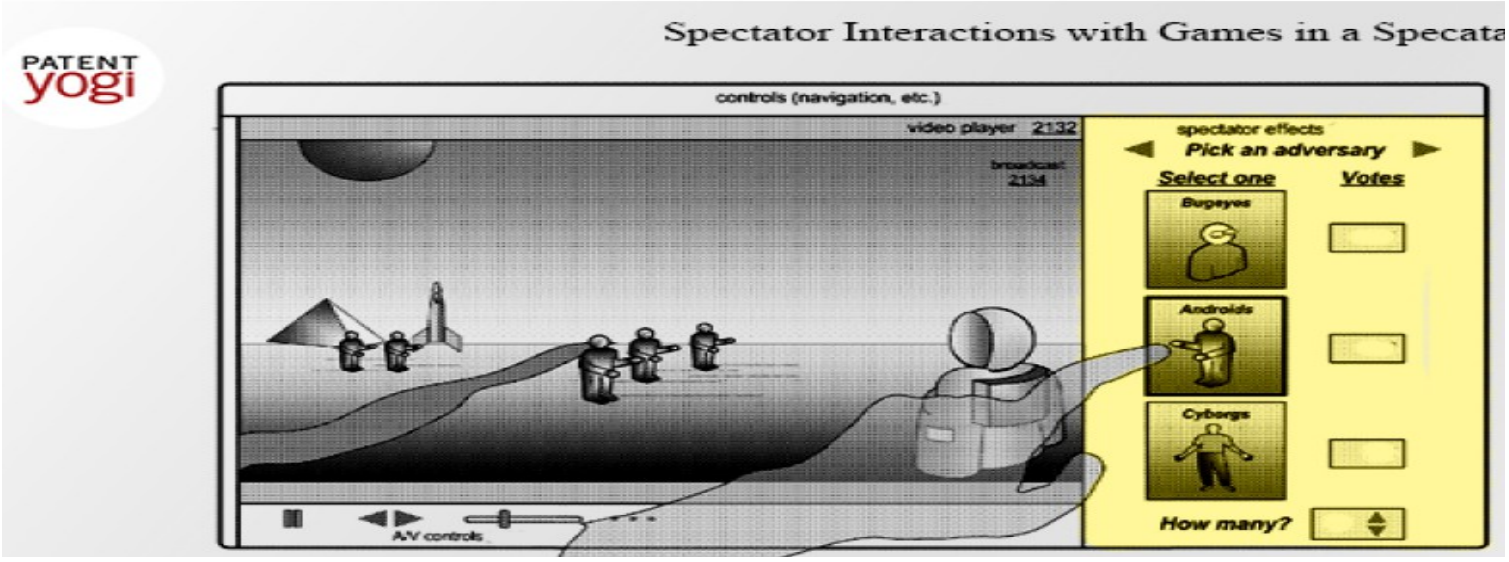
The U.S. Patent and Trademark Office awarded FishEye Products, LLC US 9,652,312 in May of 2017.



RTTK Intro Video

Amazon files 7 patent applications in a week related to its live streaming video platform “Twitch”. This is unusual!

The patent application reveals (20170003740) that Twitch will allow spectators to interact with and a□ect a game being broadcast via inputs to and interactions with user nterface elements presented on the spectating user interface. The spectators may a□ect or in□uence the game,objects within the game universe, events within the game,or the players in the game via the user interface elements on the spectating user interface. The spectators may become involved in the games being broadcast by in□uencing game play via the spectating inputs. Game play for the players may be enhanced by providing interesting variations in game play based on the spectating inputs. For example, as shown in the image below, the spectating community is allowed to select what type of adversaries (“Bugeyes”, “Androids”, or “Cyborgs”) are to appear at a level of the game based on community votes as shown in “Votes” column.



Patent Information Publication number: US20170003740 A1
Patent Title: Spectator interactions with games in a specatating system
Publication date: 5 Jan 2017 Filing date: 30 Jun 2015 Inventors: David Hendrik Verfaillie, Hok Peng Leung,Patrick Gilmore, Ethan Zane Evans, Michael Anthony Willette, Christopher Paul Dury, Collin Charles Davis,Richard Bantegui, Francis Xavier Surjo-Subagio, Michael Anthony Frazzini, Michael Martin George Original Assignee: Amazon Technologies, Inc.

US 20170001122A1

(19) **United States**
(12) **Patent Application Publication**
LEUNG et al.

(10) **Pub. No.: US 2017/0001122 A1**
(43) **Pub. Date: Jan. 5, 2017**

(54) **INTEGRATING GAMES SYSTEMS WITH A SPECTATING SYSTEM**

(71) Applicant: **Amazon Technologies, Inc.**, Seattle, WA (US)

(72) Inventors: **HOK PENG LEUNG**, REDMOND, WA (US); **DAVID HENDRIK VERFAILLIE**, LAGUNA BEACH, CA (US); **PATRICK GILMORE**, AGOURA HILLS, CA (US); **ETHAN ZANE EVANS**, SNOQUALMIE, WA (US); **MICHAEL ANTHONY WILLETTE**, LAKE FOREST, CA (US); **CHRISTOPHER PAUL DURY**, BELLEVUE, WA (US); **COLLIN CHARLES DAVIS**, SEATTLE, WA (US); **RICHARD BANTEGUI**, IRVINE, CA (US); **FRANCIS XAVIER SURJO-SUBAGIO**, IRVINE, CA (US); **MICHAEL ANTHONY FRAZZINI**, SEATTLE, WA (US); **MICHAEL MARTIN GEORGE**, MERCER ISLAND, WA (US)

(73) Assignee: **AMAZON TECHNOLOGIES, INC.**, Seattle, WA (US)

(21) Appl. No.: **14/755,974**

(22) Filed: **Jun. 30, 2015**

Publication Classification

(51) **Int. Cl.**
A63F 13/86 (2006.01)
G06F 3/0484 (2006.01)
G06F 3/0482 (2006.01)

(52) **U.S. Cl.**
CPC *A63F 13/86* (2014.09); *G06F 3/0482* (2013.01); *G06F 3/04842* (2013.01)

(57) **ABSTRACT**

A spectating system that exposes an application programming interface (API) to game systems. The spectating system obtains game metadata from the game systems for games being broadcast by the spectating system according to the API, and generates content for the broadcasts based at least in part on the game metadata. The broadcast content is provided to spectator devices with the broadcasts. The spectating system receives indications of spectators' interactions with the broadcast content from the spectator devices, and provides indications of the interactions to the game systems according to the API.