# splunk> conf2017 Building Blocks For Analytics Common Sense

Yanpei Chen | Sr. Product Manager, Product Analytics Archana Ganapathi | Director, Data Strategy

September 2017 | Washington, DC

#### **PRESENTATIONS.** Forward-Looking Statements

During the course of this presentation, we may make forward-looking statements regarding future events or the expected performance of the company. We caution you that such statements reflect our current expectations and estimates based on factors currently known to us and that actual events or results could differ materially. For important factors that may cause actual results to differ from those contained in our forward-looking statements, please review our filings with the SEC.

The forward-looking statements made in this presentation are being made as of the time and date of its live presentation. If reviewed after its live presentation, this presentation may not contain current or accurate information. We do not assume any obligation to update any forward looking statements we may make. In addition, any information about our roadmap outlines our general product direction and is subject to change at any time without notice. It is for informational purposes only and shall not be incorporated into any contract or other commitment. Splunk undertakes no obligation either to develop the features or functionality described or to include any such feature or functionality in a future release.

Splunk, Splunk>, Listen to Your Data, The Engine for Machine Data, Splunk Cloud, Splunk Light and SPL are trademarks and registered trademarks of Splunk Inc. in the United States and other countries. All other brand names, product names, or trademarks belong to their respective owners. © 2017 Splunk Inc. All rights reserved.



splun

## **About Your Speakers**

- Both have worked at the bleeding edge of data science
  - UC Berkeley Computer Science PhD, RAD Lab/AMP Lab alums, hands-on experience with a variety of data
- Both have experience translating data into tangible business value
  - Archana: Wide experience of how customers use Splunk on their data
  - Yanpei: Internal experience of how Splunk uses its own data
- We start to see patterns in success factors, hence want to share



## Context

splunk>



### **Data Has Gone Mainstream**

- Data is a corporate-level asset and key differentiator
- Impacts the majority of the economy, majority of a business, not just tech!
- There are common pitfalls -- and common building blocks for success
- Success depends on people + process + technology





### Organizations Have A Spectrum Of Data Maturity

- Data and analytics savvy
  - Already making data-driven business decisions based on rigorous analysis
  - Challenges in connecting data observations (past and present) and business vision (future)
- Data literate but not analytics savvy
  - Need help with analytic rigor (e.g. count vs distinct count, trend vs noise vs seasonality)
  - Some risk of business decisions based on incomplete or insufficiently rigorous data
- Data illiterate
  - Critical data not collected or not visible
  - Need help seeding data-driven DNA





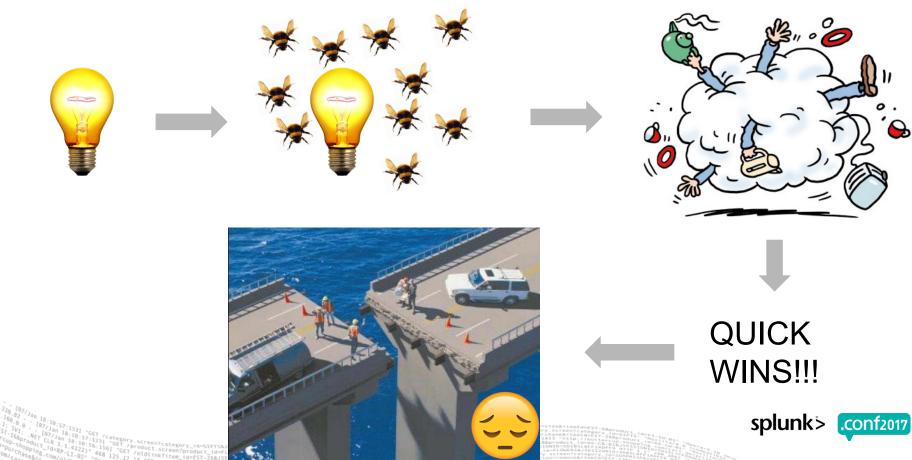
#### Failures Threaten The Data-driven Mindset

- Machine learning on 100 deals to predict sales  $\rightarrow$  missed quarterly results
- One-month comparison of channel effectiveness  $\rightarrow$  abandoned healthy channels
- Track "problem size" without normalizing "baseline size"  $\rightarrow$  overinvest in solution

...all these can and have led to high profile missteps



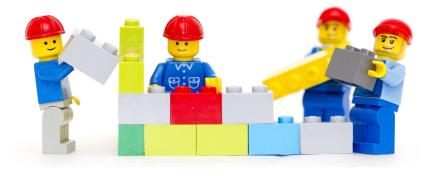
#### **Common Pitfall**



#### Building Blocks For Data-driven Decision Making

- Clarity of purpose and context
- Data worth a damn
- Numerical discipline
- An appropriate calculator
- Sanity checked results

Strong infrastructure - technical and non-technical





# Clarity Of Purpose And Context

splunk> .conf2017

us hopping con/cort.d07action=view6item1ad=3r.d00; hopping con/cort.d07action=view6it

### How To Actually Get (Hopefully Quick) Wins

## Why? For whom? When? Possible actions?





sopping converte (d) factory tends tendings terrors-bound in a converte tend tending in the tending of the tending terrors-bound in the tending of tending of the tending of the tending of the tending of the tending of tending of the tending of the tending of the tending of the tending of tending of the tending of the tending of the tending of tending of the tending of te



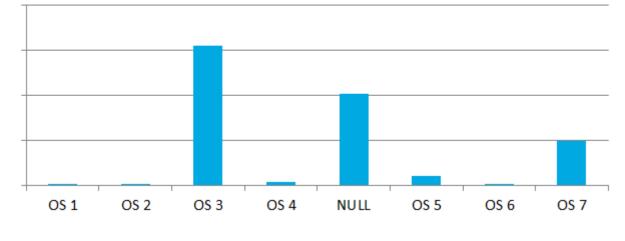
## Data Worth A Damn

Bad data means garbage in, garbage out



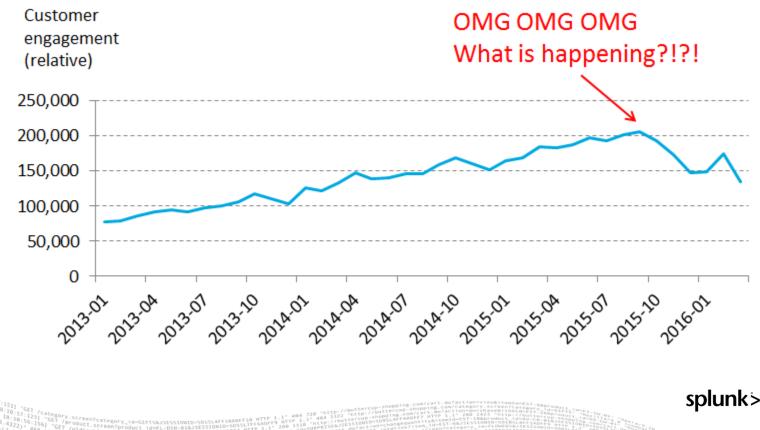
### **Example 1: Nulls**

- Real NULLs or missing values?
- Presence of NULLs relevant or invalidates decision at hand? Sample customer tickets



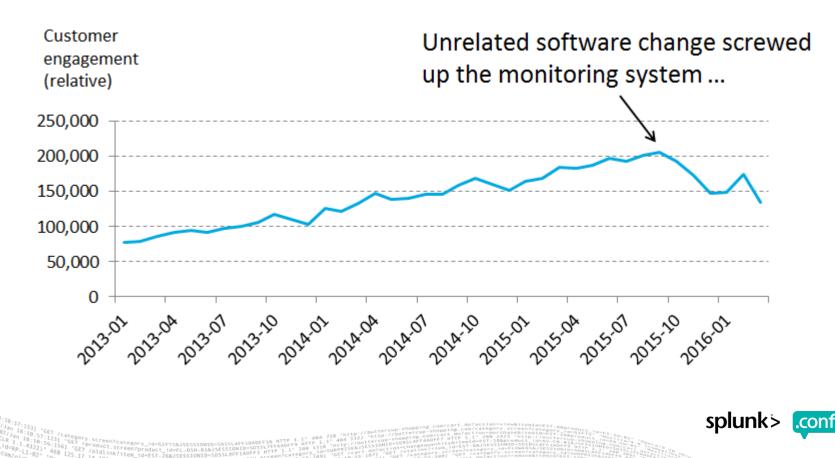


## **Example 2: Bad Instrumentation**





## **Example 2: Bad Instrumentation**



## **Check List For Data Quality**

#### Coverage

- How much data? What time frame?
- Who/what is covered? How much of "total universe" is visible?
- Granularity
  - Frequency (hourly, weekly)?
  - Level of detail (entire system, per component, per session/user)?
- Semantics
  - Direct vs proxy instrumentation?
  - Reference vs divergent interpretations?

#### Dedicated talk on data quality - see schedule!



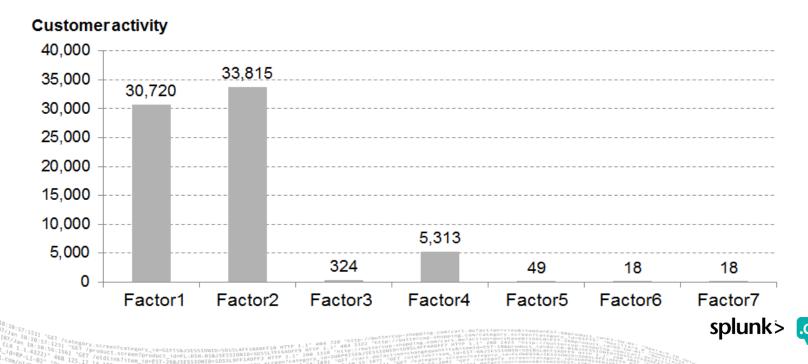
# **Numerical Discipline**

Non-scary techniques that help you be rigorous with data



#### Count Interesting Things By Interesting Factors

This often reveals valuable info, and everyone understands it!



## Look At The Distribution Of Values

Starting to dig deeper. Everyone still understands this. Statisticians call these PDFs and CDFs.

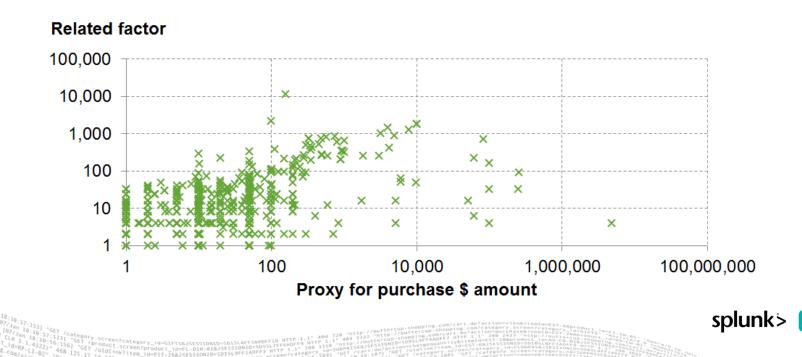


# Customers



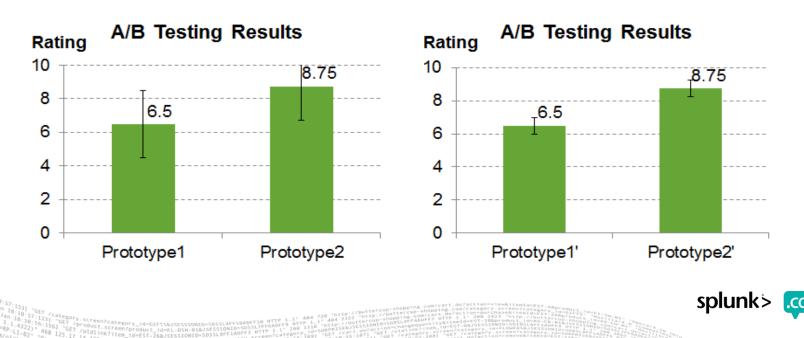
#### Identify Outliers And Make Sense Of Them

Outliers are as important to understand as "the average"

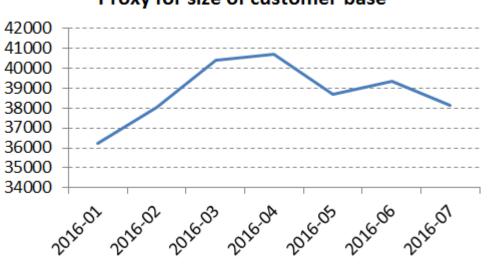


## **Calculate The Margin Of Error In Results**

- Graphs have same numerical results. Left graph has overlapping margin of error
- Ignoring margin of error could have led to wrong investment decisions

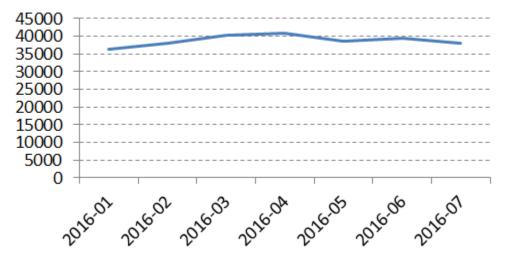


Are we losing customers?!?



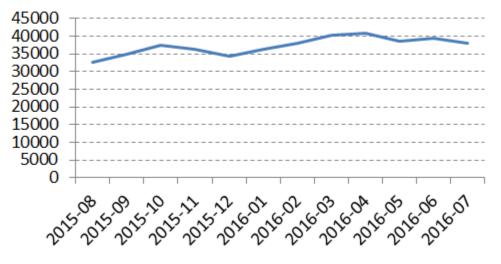


Same data, vertical axis not truncated



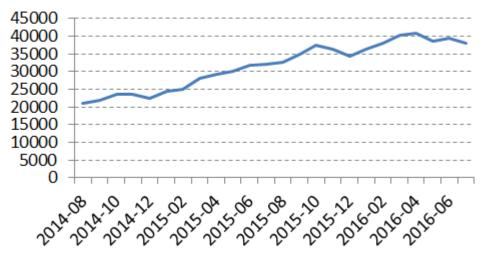


Same data, 12 months to check for noise



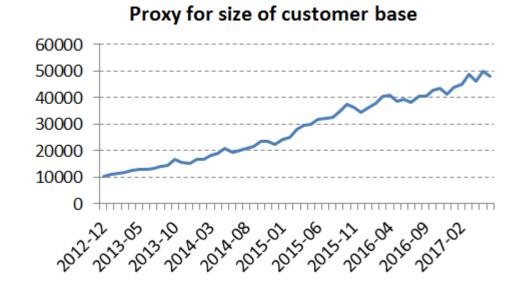


Same data, 24 months to check for seasonality





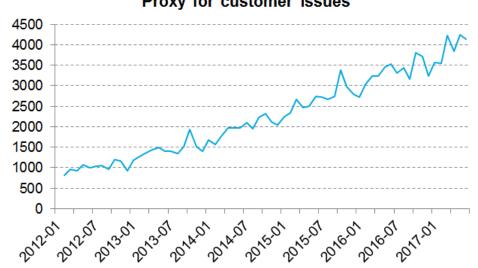
Same data, historical trend





#### **Check If Any Trends Need To Be Normalized**

Customer issues growing unbounded?!?

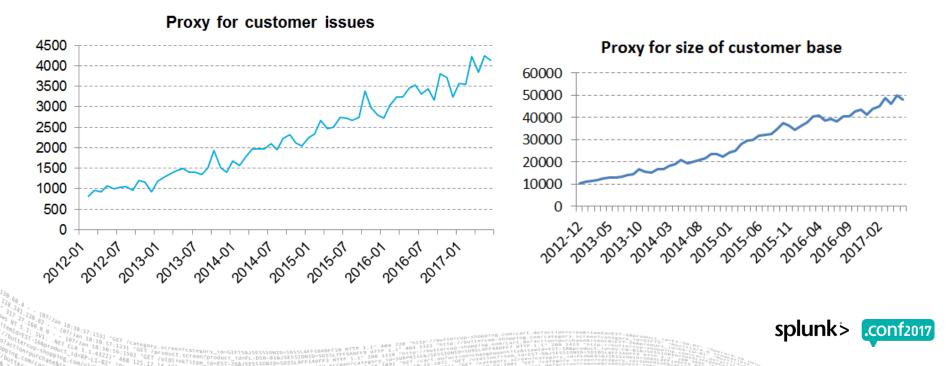






#### Check If Any Trends Need To Be Normalized

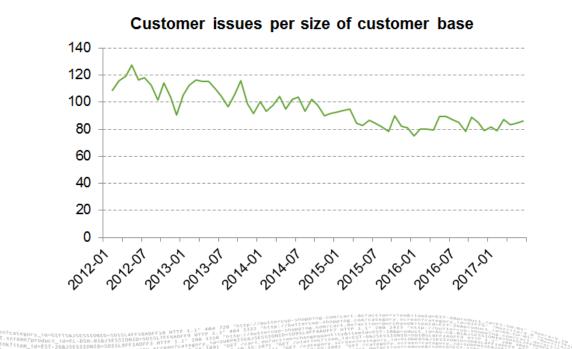
But customer base is also growing ...



splun

#### Check If Any Trends Need To Be Normalized

Let's normalize customer issues by size of customer base Normalized data hugely informative

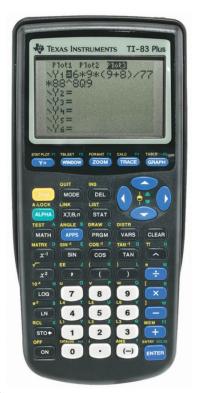


# An Appropriate Calculator

Knowing what technique to use when and how far to go



# *"Everything should be made as simple as possible, but not simpler."*



- How many buttons do you need to
  - Compute the dinner bill?
  - Do rocket science?
- What % of the world's problems are similar to
  - Computing the dinner bill?
  - Doing rocket science?



# *"Everything should be made as simple as possible, but not simpler."*

- Machine learning! TI-83 Plu **Big Data!** Al! Deep Learning!
  - How many buttons do you need to:
    - Compute the dinner bill?
    - Do rocket science?
  - What % of the world's problems are similar to:
    - Computing the dinner bill?
    - Doing rocket science?



#### Splunk Is Actually A Pretty Powerful Calculator!

#### Save & share results

XT.O.n

LOG

STO

TEXAS INSTRUMENTS TI-83 Plus Plot1 Plot2 2022 \V186+9+(9+8)/77

#### Visualize

Analyze (SPL & MLTK) Explore

CIEAR

Ingest

TAN

COS

splunk> .conf2017

\* .

2m ago

count

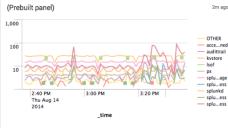
#### Splunk Is Actually A Pretty **Powerful Calculator!**

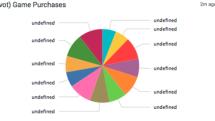
#### Save & share results



Analyze

(Inl	ine Search) Top Source Types		2m ago	(Report) Top Se	ource Ty	pes			
	sourcetype 0	count 0	percent o	splunkd					
- 1	splunkd	24775	73.43	splunkdaccess					
2	splunkd_ui_access	6207	18.40	splunkd_access					
3	splunkd_access	1612	4.78	splunk_wservice splunkaccess mongod scheduler					
4	splunk_web_service	553	1.64	mongod					
5	splunk_web_access	388	1.15						
6	mongod	169	0.50	first_inoo_small splunkd_stderr					
7	scheduler	25	0.07	splunkd_conf					
8	first_install-too_small	4	0.01		0	10	100	1,000	10
9	splunkd_stderr	3	0.01				count		
10	splunkd_conf	2	0.01						
		« prev	1 2 next »						





splunk> .conf2017

#### Splunk Is Actually A Pretty Powerful Calculator!

#### Save & share results



#### Analyze (SPL & MLTK)

Texas Instruments         TI-83 Plus           Plot1         Plot2         2028           V166+9*(9+8)/77         *88^809           V2=         V3=           V4=         V5=           V5=         V5=           V5=         V5=	
Image: Street of the street	Explore
Ing	est

Show the various searches to powe	er a panel.						
lunk> App: Enterprise Security ~			A	dministrator 🗸 🦷 Messages 🗸 🖇	Settings $\sim$ Activity $\sim$ Help $\sim$	Find	
curity Posture Incident Review My Investigations Advanced	i Threat 🗸 Security Domains 🗸	Audit 🗸 Search 🗸	Configure 🗸			Enterprise 8	Secu
ecurity Posture					Edit 🗸	More Info ~	¥
Excess notables Table Count     ENDPOINT Notables Total Count       1k +375     753 +532	Total Count	Total Count	Total Count	THREAT NOTABLES	Total Count		
Notable Events By Urgency			Notable Events Over Time				
Cuandan and a second se		unknown informational	200 B 100	٨	$\wedge M$		access audit endpo
0 250 500 750 1,000 1,250 1,50 count	0 1,750 2,000 2,250	nedium high critical	6:00 PM Tue Oct 6 2015	12:00 AM Wed Oct 7 time	6:00 AM 12:0	=	identi netwo
	00 1,750 2,000 2,250	high	Tue Oct 6	Wed Oct 7	6:00 AM 12:0	=	identit
count	00 1,750 2,000 2,250	high	Tue Oct 5 2015	Wed Oct 7 time		=	identi netwo threat
count count de_name o		high critical	Tue Oct 6 2015 Top Notable Event Sources src 0 sparkline 0 10.11.3620	Wed Oct 7 time		DO PM	coun
count op Notable Events de.name :: izithfisited Event Observed		high critical count 0	Top Notable Event Sources src 0 sparkline 0 10.11.3620^	Wed Oct 7 time correlation	n_search_count 0 security_t	domain_count 0	coun
count op Notable Events de_parme : stathland Event Observed consult Fälled Logina	sparkline :	high critical count 0 1459	Tue Oct 6 2015 Top Notable Event Sources src 0 sparkline 0 10.11.3620	Wed Oct 7 time correlation	n_search_count o security_c	domain_count 0	coun
count op Notable Events Ae_name : istelbiand Event Observed consisier Failed Event Observed consister Failed Event Observed co	sparkline :	count o	Tue Oct 6           2013           Top Notable Event Sources           src 0         sparkline 0           10.11.36.5	wed Oct 7 time	n_search_count o security_s 6 3	domain_count 0	coun
count Cop Notable Events de_rame = extensioned Event Observed extensioned E	spartline : 	count 0 1459 599 305	Tue for 6           2015           Top Notable Event Sources           sre 0         sparkine 0           10.11.36.20	wed Oct 7 time Correlation	n_search_count ≎ security_o 6 3 3	domain_count o	coun
count op Notable Events &	sparkline :	critical	Tue for 6           2015           Top Notable Event Sources           sre 0         sparkine 0           10.11.36.20	wed Oct 7           time	n_search_count o security_ct 6 3 3 3	domain_count 0 4 2 2 2	coun
count Cop Notable Events Ade_name 1 Ade_name 2 Ade_name 2 Ade_name 2 Ade_name 3 Ade_name	sparkline :	count 0 1459 305 286 282	Top Notable Event Sources           sp:rkine 0           10113620	wed Oct 7           time	Lsearch_count c security_r 6 3 3 3 3 3	domain_count 0 4 2 2 2 2	coun
count Cop Notable Events Ac.asme 1 Ac.asme 2 Ac.asme 3 Ac.asme 4 A	sparkline :	count 0 1499 305 286 282 284	Top Notable Event Sources are : sparkline : 10113620 - M 10113620 - M 10113620 - M 10113610 - M 10113610 - M 10113610 - M	Wed Oct 7         time	Learch_court : security_ 6 3 3 3 3 3 3	domain_count 0 4 2 2 2 2 2	coun
count	spestine ::	ecunt 0 control 1459 305 285 285 242 140	Tue for 6 2015           Top Notable Event Sources           src :         sparkine 0           1011 36.0         -//	Wed Oct 7           time           M	search_count : seouthy.e 6 3 3 3 3 3 3 3 3	domain_count = 4 2 2 2 2 2 2 2	coun
count op Notable Events &c.anne :  Ac.name :	sparkline :	ecurt : critical court : 1459 300 286 282 244 142 128	Top Notable Event Sources are c aparkline c 10113620	Wed Oct 7           time           M	1_64970h_count : 86047by,6 6 3 3 3 3 3 3 3 3 3 3 3 3 3 3	domain_count 0 4 2 2 2 2 2 2 2 2 2 2	coun

28 "http://buttercup-shopping.com/cart.id/?action=vtenkitemia=Esr-6Aproduc 28 "http://buttercup-shopping.com/cart.id/?action=purchasekitemid=Eprov\_tateCite%\_tate\_\_\_\_ 404\_3322 "http://buttercup-shopping.com/cartry 1.1" 200 2423 "http://butterEdbarout\_tate intercupiesesitemid=Estimated and the shopping and the sh



Edit 🗸

More Info V

#### Splunk Is Actually A Pretty **Powerful Calculator!**

Show the various searches to power a pan

#### Save & share results

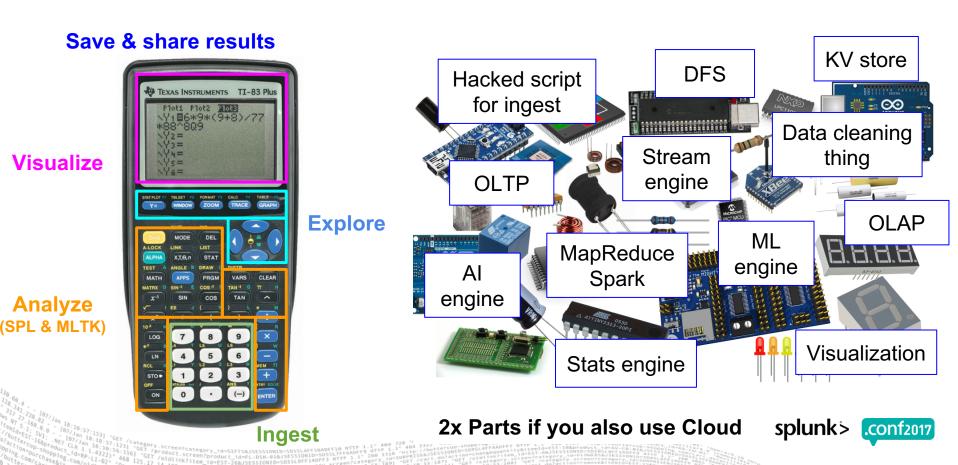


Analyze

	curity ~				Admi	nistrator v Messages v Se	ttings $\sim$ Activity $\sim$ Help $\sim$	
Security Posture Incident Rev	view My Investigations Adv	anced Threat 🗸 Sec	urity Domains 🗸 🛛 Aud	dit ∽ Search ∽	Configure 🗸			Enterprise Secu
Security Posture							Edit 🗸	More Info 🗸
✓Edit ACCESS NOTABLES Total Count	ENDPOINT NOTABLES	NETWORK Total Count		DENTITY NOTABLES	AUDIT NOTABLES	THREAT NOTABLES	UEBA NOTABLES Total Count Save As ~	<ul> <li>New Table</li> <li>Close</li> </ul>
<pre>dexeprod_analytics where component="licensing. evalisenseType=if(compone eventstats latest(execution where executionDPlatestExe eval operatingSystemcase(component eval operatingSystemcase(component idata.os'==null* and 'da rename data.splumkVersion a fields * where component="deployment</pre>	<pre>nt=="licensing.stack", 'get ID) as latestExecutionID = cutionID AND (1==100 OR li !="deployment.node", NULL, omponent!="deployment.node ta.disk.fileSystem"=="ext4 s splunkVersion</pre>	_license_value(dep in(licenseType) as censeType==1) AND ( 'data.cpu.architec ", NULL, 'data.os' ","Linux", 'data.os	loymentID,'data.typ licenseType max(da quota>=1 AND (20000 :ture'="x86_64" or =="null" and 'data. ;'=="null", "Unknow	<pre>we','data.subgroup','' ita.quota) as quota by 0000000000000000=0 06 'data.cpu.architectur disk.fileSystem'=="hg</pre>	deploymentID quota<20000000000000000000 e'="x64", "x86 64-bit", 1=	1, 'data.cpu.architecture	°)	revious week Y
8,315 events (7/23/17 12:00:00.000 A		No Event Sampling 🗸				u u	± ♦ ♦ ■ II ∨ do	📍 Smart Mode 🗸
ents (13,315) Patterns	Statistics Visualization							
rmat Timeline ~ - Zoom Out	+ Zoom to Selection × Deselec							1 hour per column
mat Timeline - Zoom Out	+ Zoom to Selection × Deselec	_			_			1 hour per column
mat Timeline ∽ — Zoom Out	_	t □ D Per Page ∽	_		_	< Prev 1	2 3 4 5 6 7	
-	_	-	-		-	CPrev 1	2 3 4 5 6 7	
de Fields ≣ All Fields	- List → ZFormat 24	-	-	×	_	C Prev 1	2 3 4 5 6 7	
de Fields III All Fields setted Fields out 1	List V ZFormat 21	-	- S	kielected Yes No	_	C Prov 1	2 3 4 5 6 7	
ide Fields III Fields ected Fields out 0 uutor 1	List V Pormat 20 I Time Event data.cpu.architecture 3 Values, 100% of events Reports	_ D Per Page ∨		Selected Yes No	-	- CPuv 1	2 3 4 5 6 7	
ide Fields III All Fields ected Fields out 1 outce 1 outce 1 outcetype 2 vesting Fields	List V Pormat 20 I Time Event data.cpu.architecture 3 Values, 100% of events Reports	-			-	< Pter 1	2 3 4 5 6 7	
lide Fields	List V Format 21 1 Time Event data.cpu.architecture 3 Values, 100% of events Reports Top values To Events with this field	□ Per Page ~	Rare	Selected Yes No	-	< Prev 1	2 3 4 5 6 7	
de Fields ected Fields ource 1 ource 1 ource 1 ourcetype 2 resting Fields orgonert 1 pucht 2	List v Format 21 1 Time Event data.cpu.architecture 3 Values, 100% of events Reports Top values To Events with this field Values	□ Per Page ∨ p values by time Count	Rare	Selected Yes No	-	CPrev 1	2 3 4 5 6 7	
	List V Format 21 1 Time Event data.cpu.architecture 3 Values, 100% of events Reports Top values To Events with this field	p values by time Count 12,369	Rare	Selected Yes No	-	CPrev 1	2 3 4 5 6 7	



### **Alternative: Assembling Way-Too-Many Parts**



### Easy To Get Answers, Easy To Go Wrong

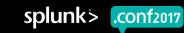


For problems that are rocket science:

- Did you punch in the right numbers?
- Do you actually understand the numbers?
- Did you press the right buttons?
- Do you actually understand the buttons?
- Can you explain what you did to stakeholders?
- Will an astronaut place their life in your hands?
- What checks did you make on the data?



Sanity Checking Results



 $\begin{array}{c} u_{2} = \operatorname{Ropp}(n_{2}, \operatorname{Com}/\operatorname{cart}, \operatorname{do}/\operatorname{cart}, \operatorname{do}/\operatorname{cart}, \operatorname{com}/\operatorname{cart}, \operatorname{cart}, \operatorname{cart},$ 

### List Of Sanity Checks

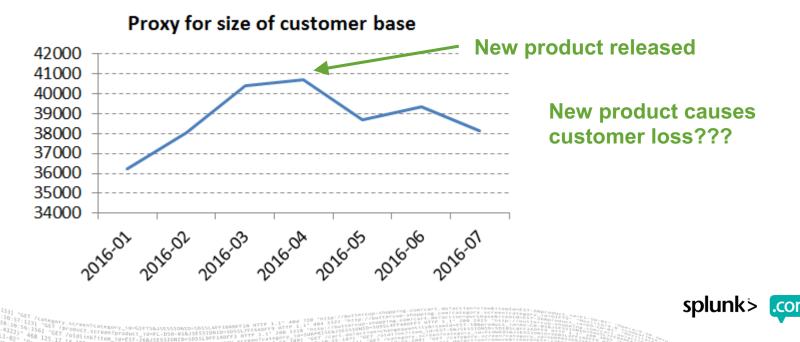
- Proxy measurement bias
- Instrumentation coverage bias
- Data quality issues
- Cross-check for well-known behavior
- Confirmation bias
- Confusing correlation with causation





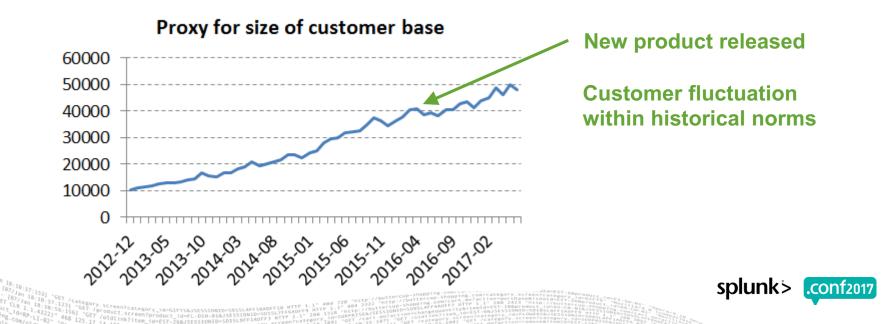
### **Confirmation Bias**

- Data often gives a signal to confirm your gut-feel
- Need to check whether other signals are stronger



### **Confirmation Bias**

- Data often gives a signal to confirm your gut-feel
- Need to check whether other signals are stronger

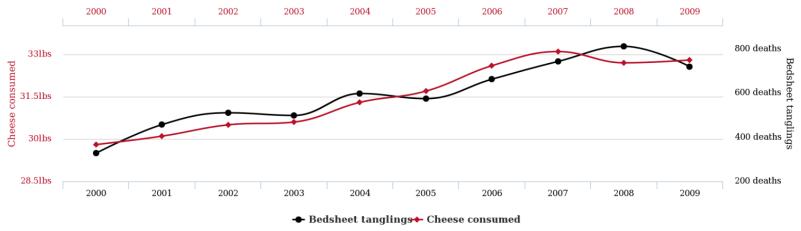


### **Confusing Correlation With Causation**

#### Per capita cheese consumption

correlates with

Number of people who died by becoming tangled in their bedsheets



tylervigen.com

#### Also, umbrellas cause rain 🙂

(/category.screen?category\_id=GIFTSEJSESSIONID=SDISLAFFI0ADFF10 HTTP 1.1° 404 720 "http://buttercup=shooping.com/category.screen?category.scre

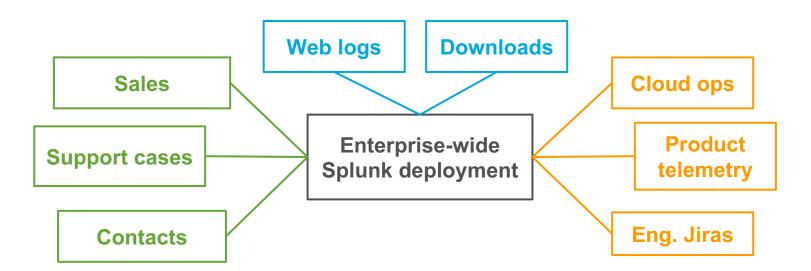


### Infrastructure

Technical, cultural, organizational



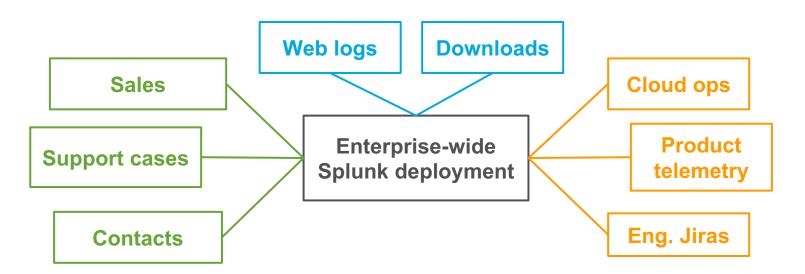
### **Data Is A Corporate Asset**





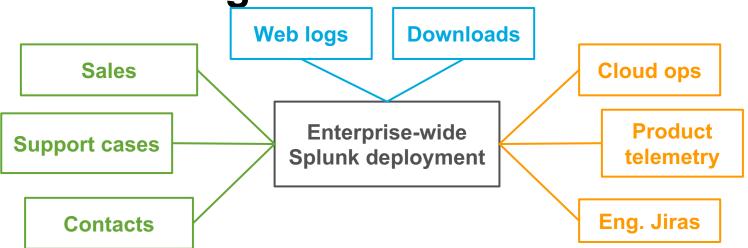
splunk

### **Data Is A Corporate Asset**



This entails data governance, data custodians, knowledge sharing, dedicated infra and ops teams, corporate-wide data security, connectors to previously siloed tools

#### Translation: Technical, Cultural, Organizational Infra



- Technical infra: Splunk must scale, especially for cloud ops and product telemetry
- Cultural infra: To understand ever changing market, learn and adapt continuously
- Organizational infra: Collaborate across org silos to create much higher value

# **Closing Thoughts**

splunk> .conf2017

university is com/cart.do/action=viewkitemid=51.60product viewsity is shoping com/cart.do/actegory.argentations/viewsity/isolation/ /buckets/viewsity/isolation/ /buckets/viewsity/isolation/ /buckets/viewsity/isolation/ /buckets/viewsity/isolation/ /buckets/ /buck

### You Will Be "Data Champions"

- Your greatest assets: perception of scientific objectivity and neutrality
- You will play many roles as leaders, individual contributors, and/or influencers
  - Translate between data semantics and business semantics
  - Clarify limits and decision boundaries with existing data
  - Nurture data literacy, advocate for long-term investment
  - Create stop gap "data plumbing" while "things improve"
  - Nurture and exemplify openness and transparency
- You now have more tools to be better data champions
  - PS: See you again for talk on Data Quality!





# Thank You

## Don't forget to rate this session in the .conf2017 mobile app

