Building Blocks For Analytics Common Sense

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

September 2017  |  Washington, DC
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.
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
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
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 → missed quarterly results
- One-month comparison of channel effectiveness → abandoned healthy channels
- Track “problem size” without normalizing “baseline size” → overinvest in solution

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

QUICK WINS!!!
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
How To Actually Get (Hopefully Quick) Wins

Why? For whom? When? Possible actions?
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?
Example 2: Bad Instrumentation

Customer engagement (relative)

OMG OMG OMG
What is happening?!?!
Example 2: Bad Instrumentation

Unrelated software change screwed up the monitoring system...
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.
Identify Outliers And Make Sense Of Them

Outliers are as important to understand as “the average”
Graphs have same numerical results. Left graph has overlapping margin of error.

Ignoring margin of error could have led to wrong investment decisions.
Check For Trend vs Noise vs Seasonality

Are we losing customers?!?
Check For Trend vs Noise vs Seasonality

Same data, vertical axis not truncated
Check For Trend vs Noise vs Seasonality

Same data, 12 months to check for noise
Check For Trend vs Noise vs Seasonality

Same data, 24 months to check for seasonality

Proxy for size of customer base

![Graph showing proxy for size of customer base over time](image-url)
Check For Trend vs Noise vs 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 ...
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.”

- 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

Visualize

Explore

Analyze (SPL & MLTK)

Ingest
Splunk Is Actually A Pretty Powerful Calculator!

Ingest
Explore
Analyze (SPL & MLTK)
Visualize
Save & share results
Splunk Is Actually A Pretty Powerful Calculator!

Ingest

Explore

Analyze (SPL & MLTK)

Visualize

Save & share results
Splunk Is Actually A Pretty Powerful Calculator!

Ingest
Explore
Analyze (SPL & MLTK)
Visualize

Save & share results
Alternative: Assembling Way-Too-Many Parts

Save & share results

Ingest

Analyze (SPL & MLTK)

Visualize

Explore

Hacked script for ingest

DFS

Data cleaning thing

KV store

OLAP

Stream engine

ML engine

Visualization

OLTP

MapReduce Spark

Stats engine

Al engine

2x Parts if you also use Cloud

splunk>
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
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

![Graph showing proxy for size of customer base with annotations:](#)

- New product released
- New product causes customer loss???
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

Also, umbrellas cause rain ☺️
Infrastructure

Technical, cultural, organizational
Data Is A Corporate Asset

Enterprise-wide Splunk deployment

- Web logs
- Downloads
- Cloud ops
- Product telemetry
- Eng. Jiras

- Sales
- Support cases
- Contacts

Downloads

Cloud ops

Product telemetry

Eng. Jiras
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
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