Diving In: A Behavioral Scientist's Guide to Analyzing xAPI-Based Data Lakes

Gregory A. Goodwin and Nick Washburn

RIPTIDE
An Employee Owned Company

US ARMY RDECOM

US ARMY NATICK SOLDIER RD&E CENTER

Overview: Experimentation in a Big Data Environment

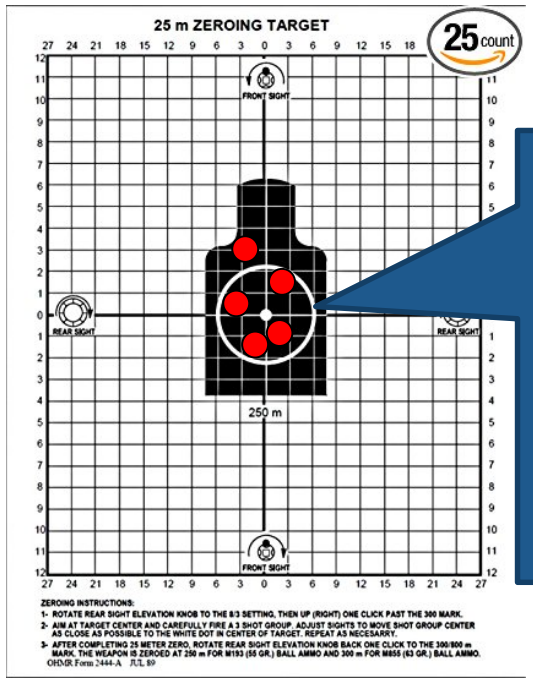
- Marksmanship Experiment
- Big data approach to storing/visualizing marksmanship training data (REAPER*)
- Demonstration:
  - Explore and analyze an xAPI data set.

*Range Experience Access Portal for Experimentation and Research (REAPER)

Marksmanship Use Case

Experimental Setup: Two groups, one gets our experimental training method and the other gets standard training. We want to compare performance of the two groups during two rifle marksmanship training events.

Confirm Zero at 75, 175, and 300m



Must get 4 of 5 in two consecutive groups inside white circle

Record Fire

Position	No. Targets	Range m
Prone Supported	20	50-300
Prone Unsupported	10	150-300
Kneeling	10	50-150

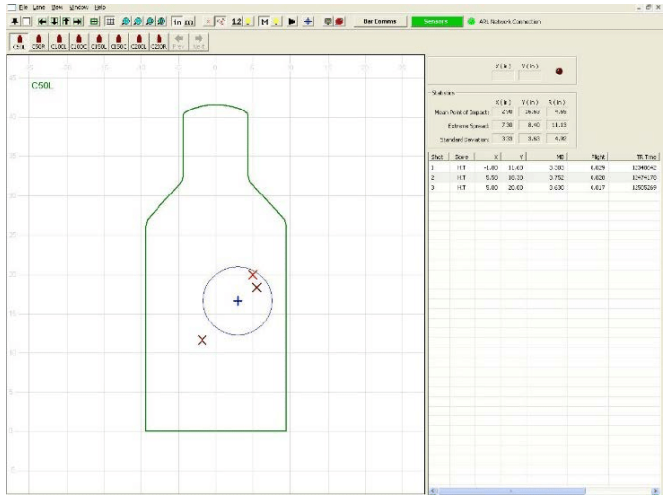
23-29 hits = marksman
30-35 hits = sharpshooter
36-40 hits = expert

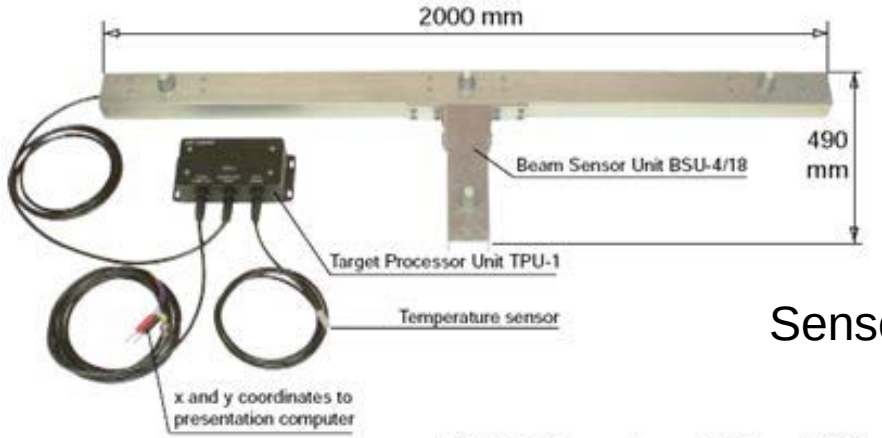
LOMAH Range

Location of Hits and Misses (LOMAH)

Target

Display

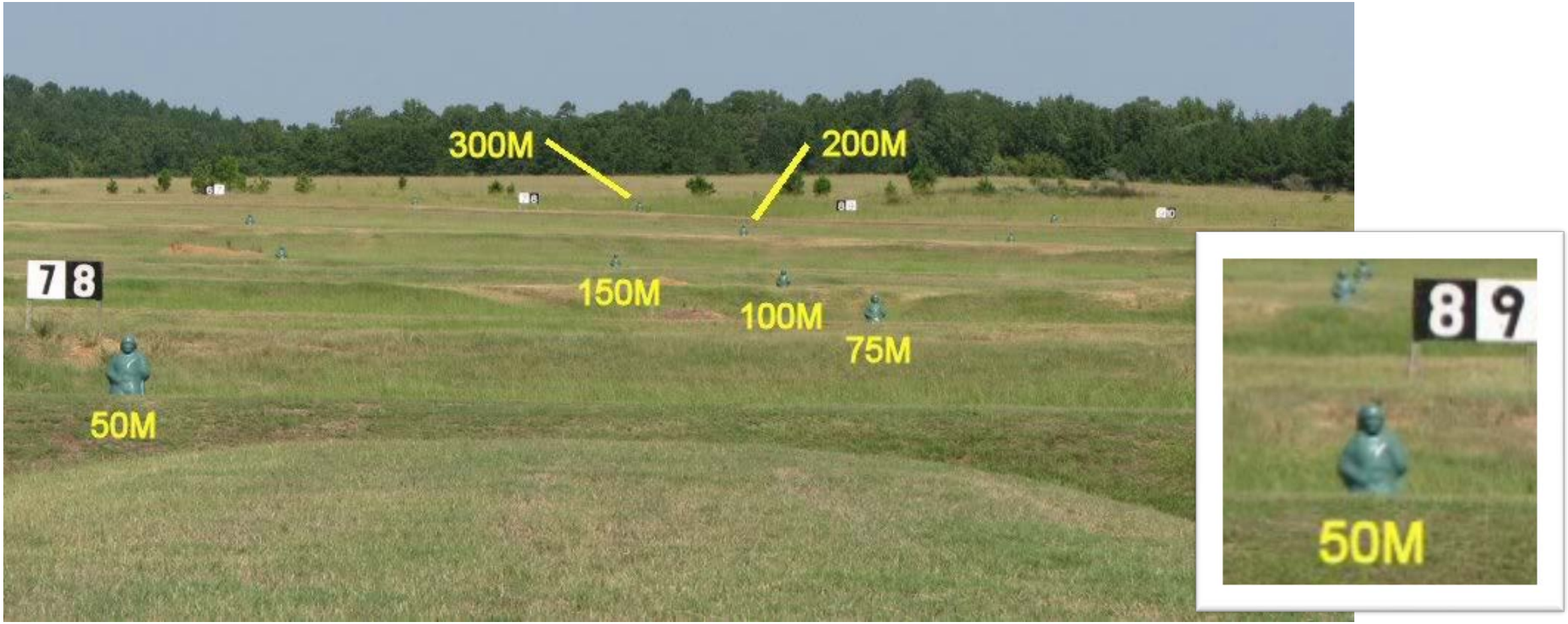




Sensor

(NOTE: Picture shows 1100 mm BSU)

Record Fire Automated Range



Timed, pop-up targets (single and multiple)
Automated detection of hits (not misses)

xAPI Activity Stream

TRACR

Range Sensors

REAPER

xAPI: Actor – Verb - Object

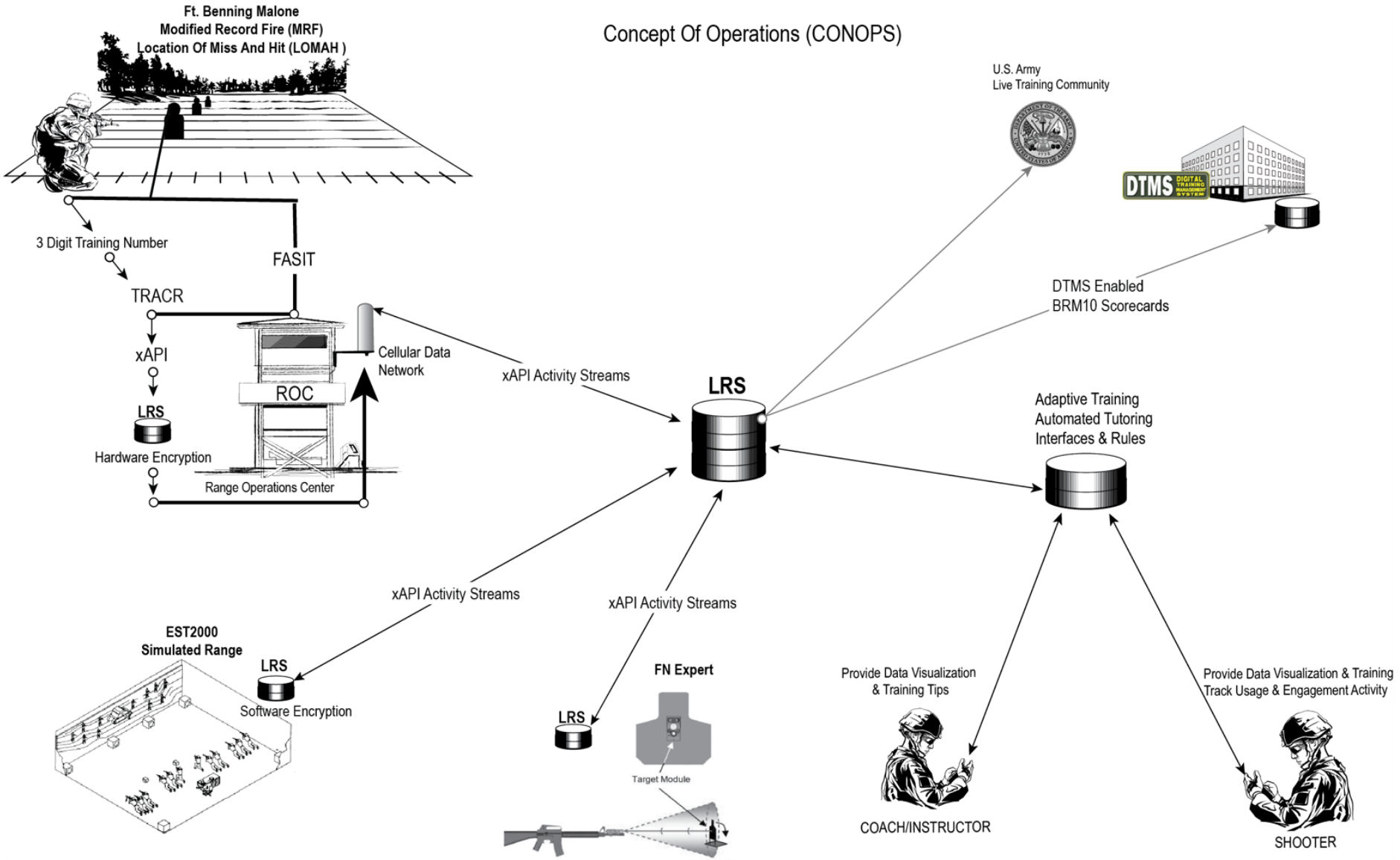
PVT Jones – Began – BRM5

PVT Jones – Hit – 75m target

PVT Jones – Missed – 300m target

PVT Jones – Passed – BRM5

REAPER Overview



Relational and Non-Relational Databases

- Relational – store data in tables with columns and rows
  - SQL or think spreadsheet
  - Good for data analysis
  - Challenge of referential integrity
- Non-relational – store data in a collection or pool
  - JSON (e.g., xAPI)
  - Good for Big Data, flexible
  - Challenge for data analysis

Experimental Data in a Relational Database

Independent Variable

Contextual Variables

Dependent Variables

Participant #	Group	Target Type	Training Event	Target Dist	X_{SHOT}	Y_{SHOT}
1	Experimental	E	BRM5	175	5	-2
2	Experimental	E	BRM5	175	7	3
3	Control	E	BRM5	175	8	-7
4	Control	E	BRM5	175	5	11
…						

Individual Participants

Easily analyzed in statistical analysis software.

Experimental Data in a Non-Relational Database - xAPI


```
{
    "id": "73fe78c5-8c8e-43bd-93f7-7814d5af2ec2",
    "actor": {
     "account": {
      "name": "249",
      "homePage": "http://www.benning.army.mil/"
     }
    },
    "verb": {
     "id": "http://www.riptidesoftware.com/products/soldier_tracking_system/hit"
    },
    "object": {
     "id": "http://www.riptidesoftware.com/products/soldier_tracking_system/BRM5",
     "objectType": "Activity"
    },
    "result": {
     "success": true,
     "extensions": {
      "http://www.riptidesoftware.com/products/soldier_tracking_system/shot": {
       "number": 1,
       "group": 1,
       "xLocation": -66,
       "yLocation": 806
```

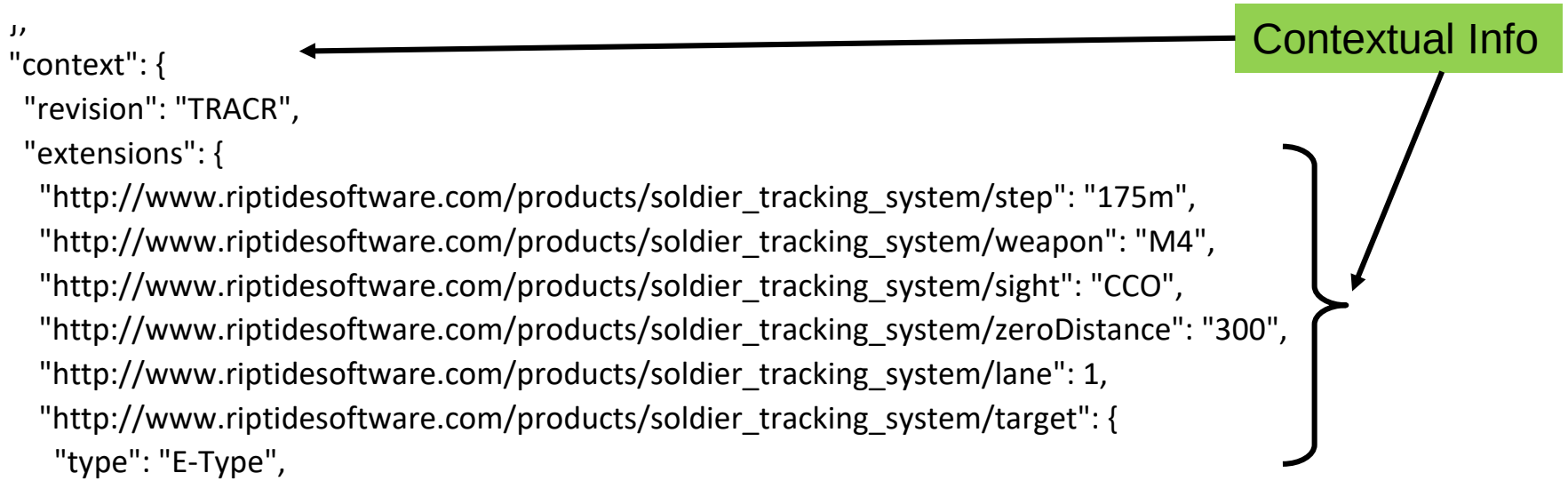

Participant Roster #

Hit/Miss

Training Event

Hit Coordinates

Non-Relational Database – Cont.




```
},
"context": {
 "revision": "TRACR",
 "extensions": {
  "http://www.riptidesoftware.com/products/soldier_tracking_system/step": "175m",
  "http://www.riptidesoftware.com/products/soldier_tracking_system/weapon": "M4",
  "http://www.riptidesoftware.com/products/soldier_tracking_system/sight": "CCO",
  "http://www.riptidesoftware.com/products/soldier_tracking_system/zeroDistance": "300",
  "http://www.riptidesoftware.com/products/soldier_tracking_system/lane": 1,
  "http://www.riptidesoftware.com/products/soldier_tracking_system/target": {
   "type": "E-Type",
   "width": 488.95,
   "height": 1022.35,
   "distance": "175m"
  },
  "http://www.riptidesoftware.com/products/soldier_tracking_system/killCircle": {
   "diameter": 304.8,
   "offset": 678.1800000000001
  }
```

Difficult to analyze in statistical analysis software.

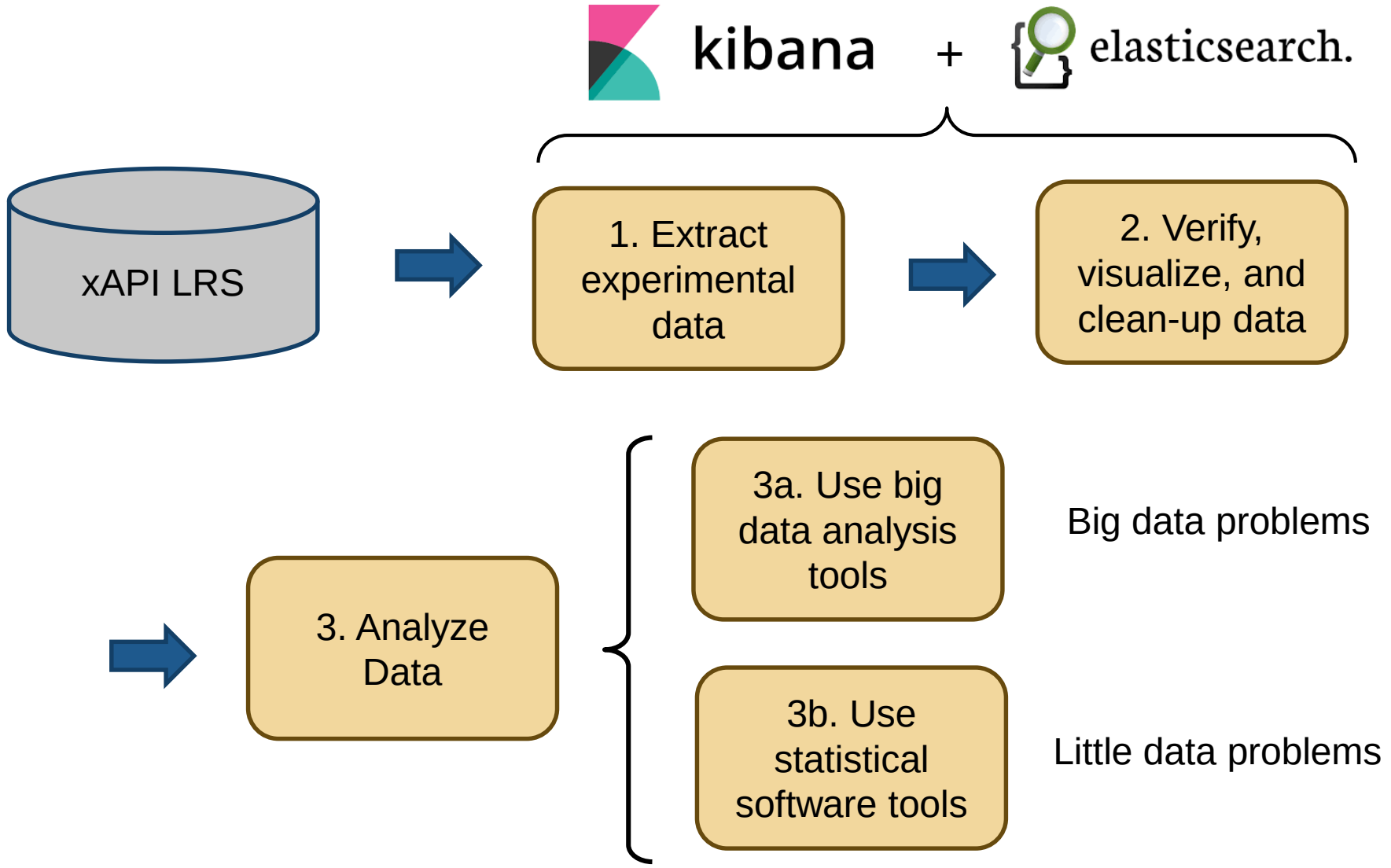
The Challenge

Accessing and Understanding our Experimental Data in this Data Lake

What Tools Can we Behavioral Scientists Use?

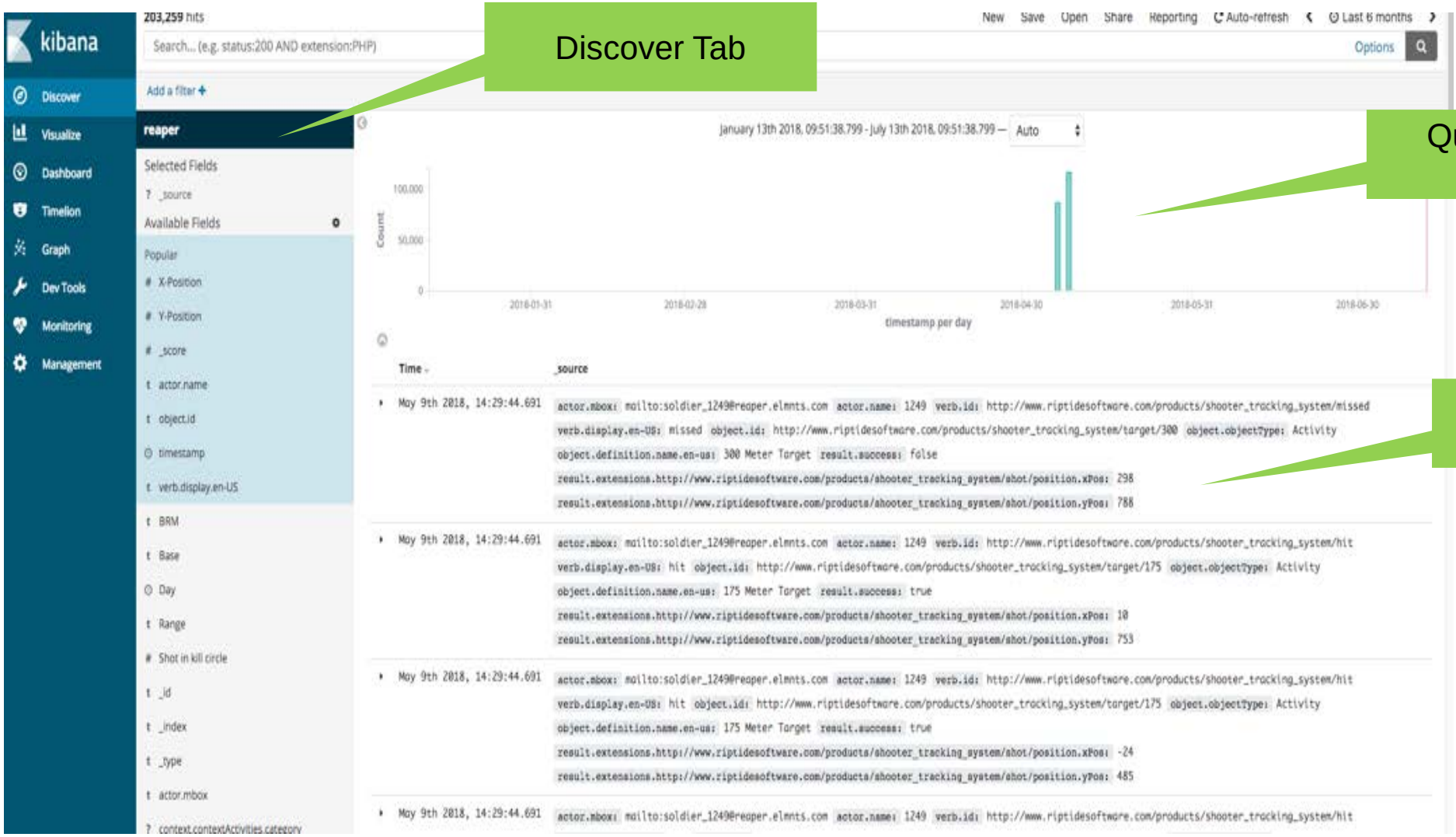
- Statistical Software (e.g., SPSS, SAS, Statistica, R, and even Excel)
  - Familiar to behavioral researchers
  - Prefer relational databases
  - Are not optimized to handle big data (yet)
- Big Data Tools (e.g., Mongo DB, Google Cloud Platform, Elasticsearch)
  - Unfamiliar to behavioral researchers
  - Don’t perform the range of lineal and non-linear inferential statistics that statistical software does
  - Excellent for combining, analyzing, and visualizing big data.

Demo Process Flow

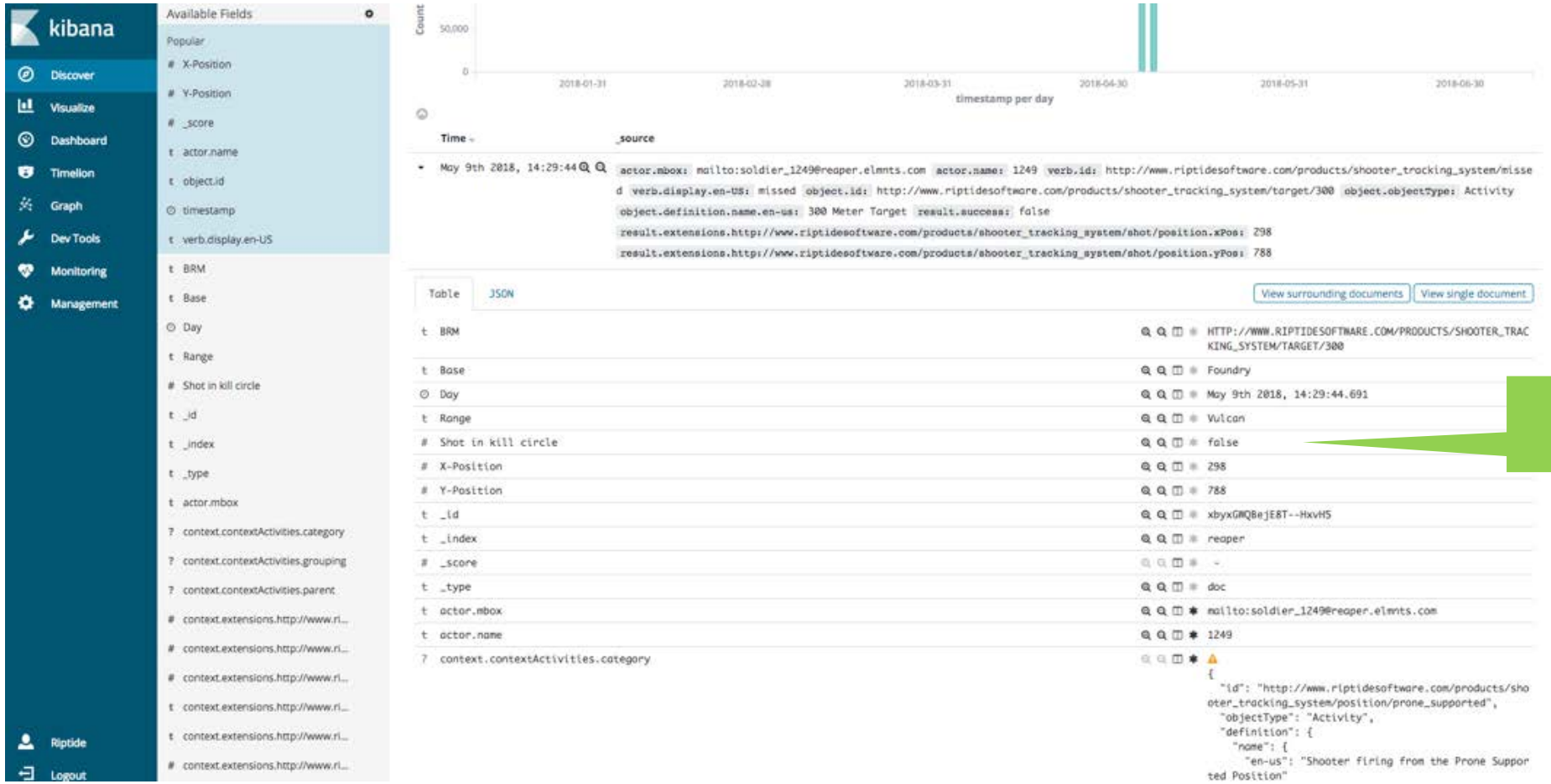


Demo Screenshots

Filter Data for Experiment



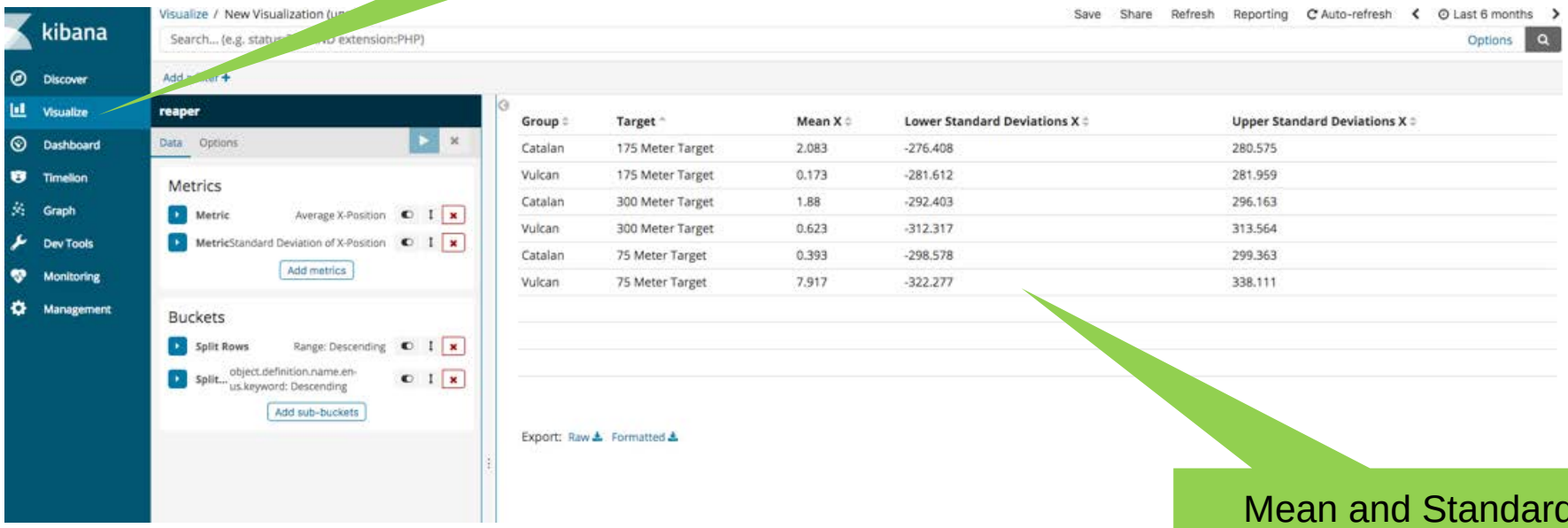
View relevant data points



xAPI Data Fields Expanded

Check ranges and means

Visualize Tab



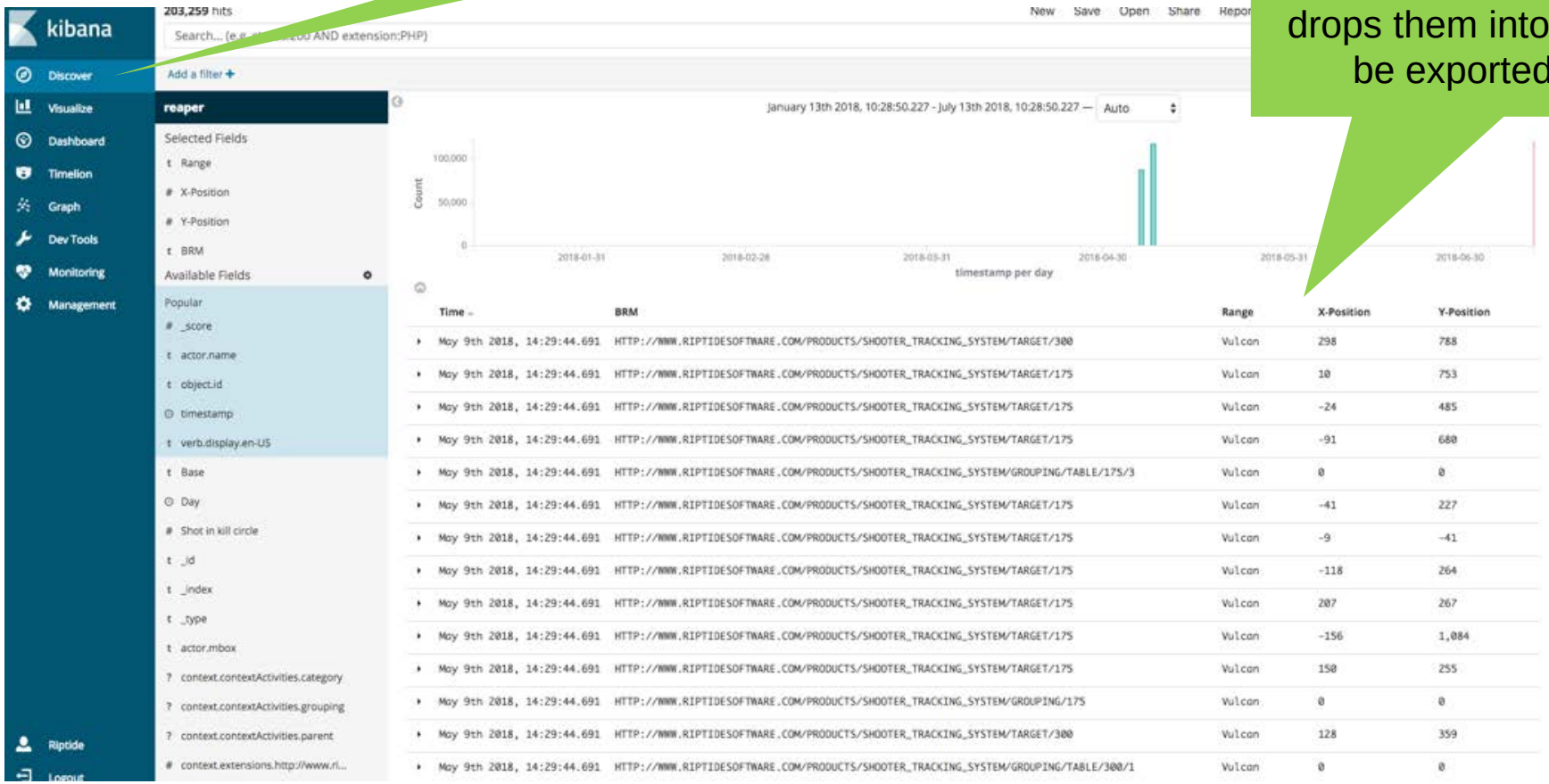
Group	Target	Mean X	Lower Standard Deviations X	Upper Standard Deviations X
Catalan	175 Meter Target	2.083	-276.408	280.575
Vulcan	175 Meter Target	0.173	-281.612	281.959
Catalan	300 Meter Target	1.88	-292.403	296.163
Vulcan	300 Meter Target	0.623	-312.317	313.564
Catalan	75 Meter Target	0.393	-298.578	299.363
Vulcan	75 Meter Target	7.917	-322.277	338.111

Mean and Standard Deviations of x Axis values

Select items for csv file

Discover Tab

Choosing data fields on left panel drops them into a table that can be exported to a csv file

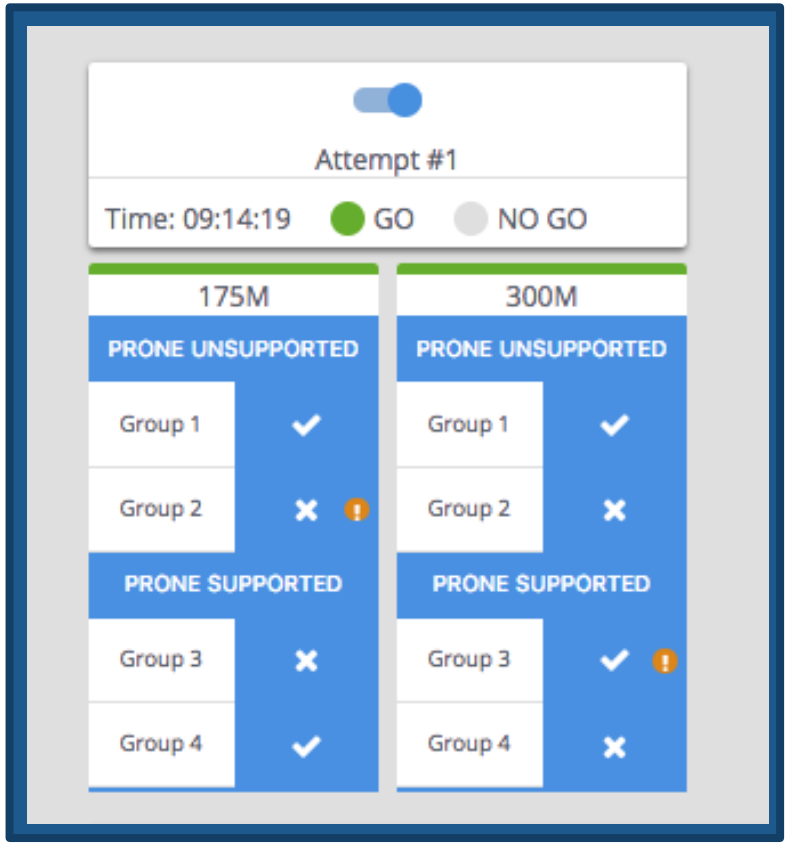


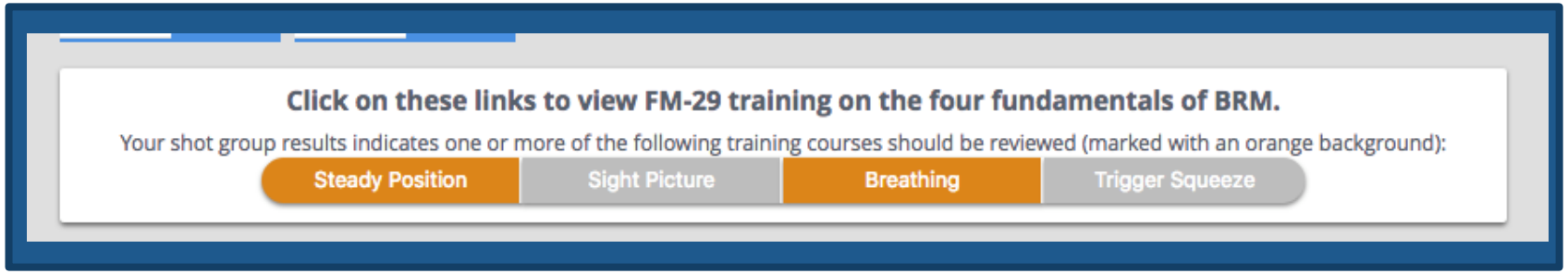
Use statistical software to analyze data…

REAPER Visualizations

REAPER Dashboards

BRM5 Soldier Dashboard

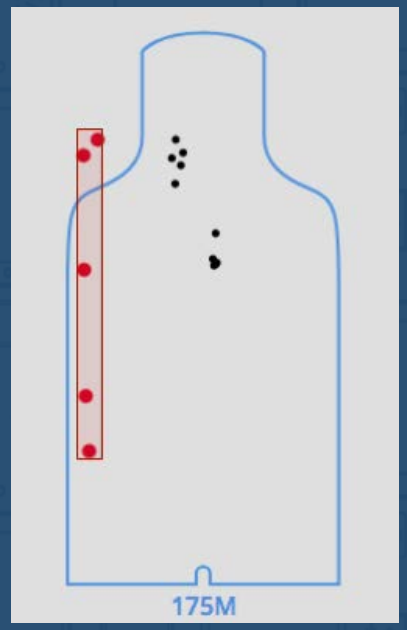


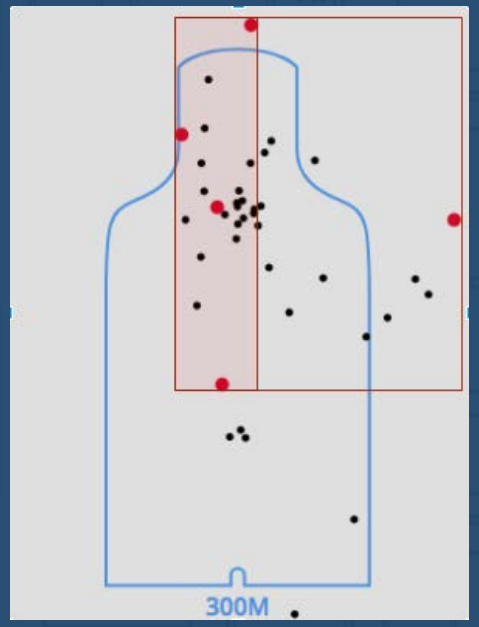


REAPER Dashboard – Deficiency Analysis

Improved Shot Pattern Analysis - Breathing

- Can be marked on any Grouping or Untimed scenario
- Pattern identified when the bounding box for the shots is largely vertical (3.5 x taller than wide)
- Accounts for “outliers” to mark breathing problem when 80% of the shots fall in the vertical pattern

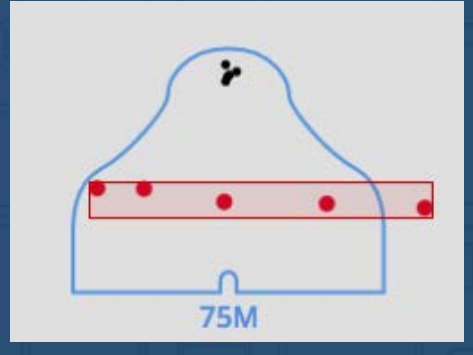


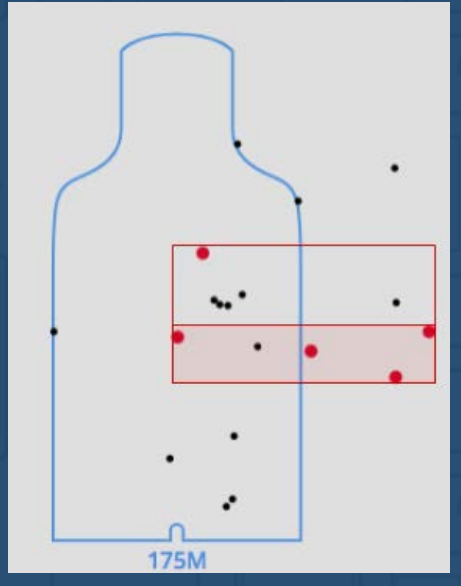


REAPER Dashboard – Deficiency Analysis

Improved Shot Pattern Analysis – Trigger Squeeze

- Can be marked on any Grouping or Untimed scenario
- Pattern identified when the bounding box for the shots is largely horizontal (3.5 x wider than tall)
- Accounts for “outliers” to mark breathing problem when 80% of the shots fall in the vertical pattern





Discussion

- **Experiments/Research are “little data” by design**
  - IRB restricts participant counts to the minimum needed for inference
  - Inferential stats were developed for small samples
- **Will Statistical software like SAS, SPSS, Statistica, and R eventually be able to use web services to analyze petabytes of big data quickly?**
  - Web services for business intelligence/big data are developing rapidly
- **Is there a fundamental philosophical difference between small data research and big data research?**
  - Naturalistic vs. laboratory experiments
  - Correlational vs. causal attribution