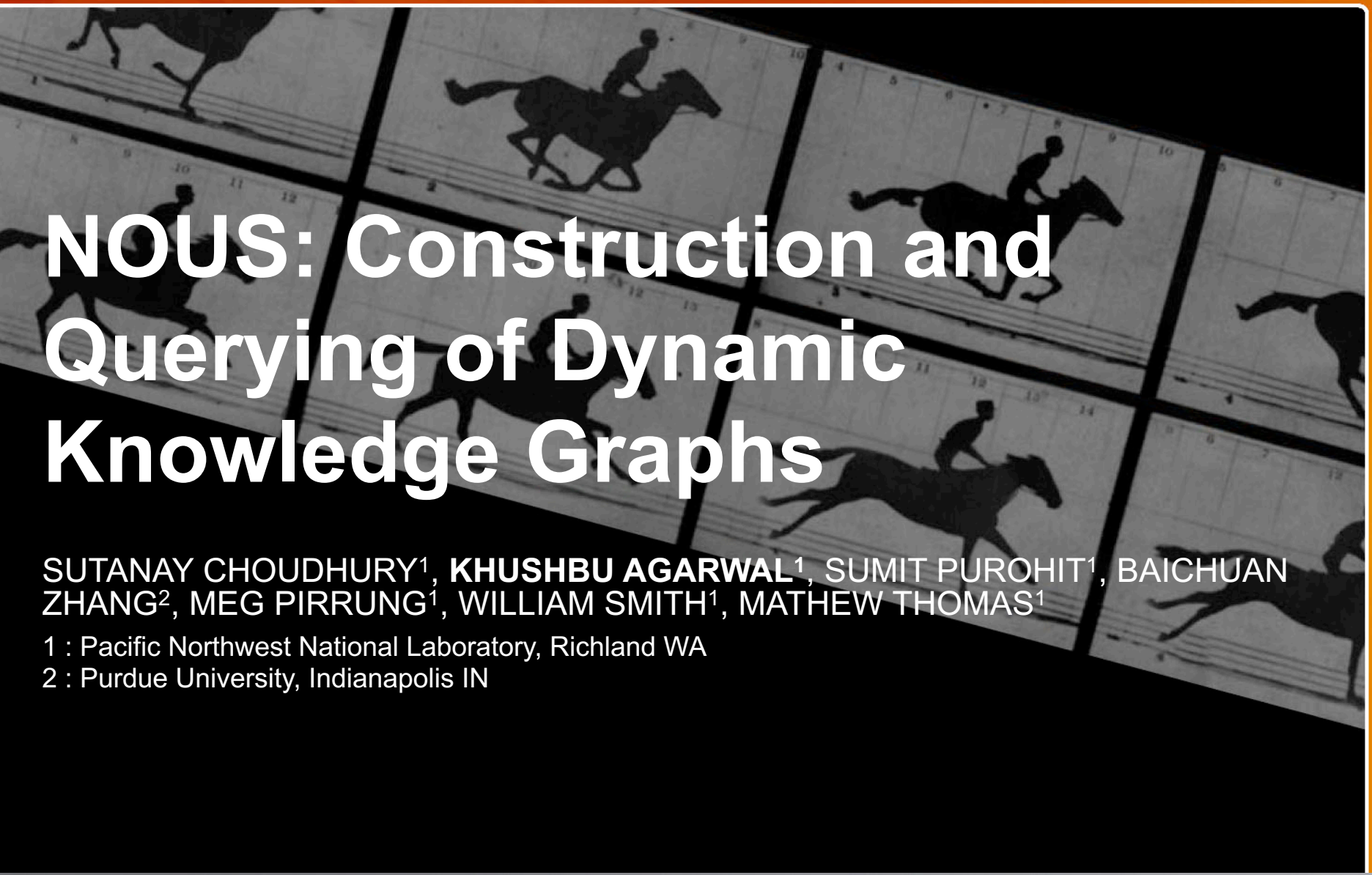
Pacific Northwest
NATIONAL LABORATORY
*Proudly Operated by* **Battelle** *Since 1965*



# NOUS: Construction and Querying of Dynamic Knowledge Graphs

SUTANAY CHOUDHURY[1], **KHUSHBU AGARWAL**[1], SUMIT PUROHIT[1], BAICHUAN ZHANG[2], MEG PIRRUNG[1], WILLIAM SMITH[1], MATHEW THOMAS[1]

1 : Pacific Northwest National Laboratory, Richland WA
2 : Purdue University, Indianapolis IN

# Outline



- **Introduction**
  - Knowledge Graphs
  - Challenges
- **Motivation (Use Cases)**
  - Domain Specific Querying of Web
  - Question-answering for Climate Science
- **Technical Approach**
  - NLP, Entity Disambiguation, Relation Learning
  - Frequent Pattern Mining, Question Answering
  - What's unique
- **Results**
- **Status and Future Work**

# Knowledge Graphs: Why do we care?



- A collection of facts about people, places, things and relationships between them, **in a given context**

I am a scientist. Read papers for me.

I am a doctor, show me latest breakthrough to consider for this patient.


I am an analyst drowning in data. Help me find interesting events

# KG: Construction and Analytical Challenges



- Construction
  - Natural Language Processing is inherently noisy
  - Unseen entities and relationships
    - How do we determine its class
    - Should we include every new relation  and entity in KB?
- Analysis
  - Identifying KB ontology needed to answer user questions.
  - Mapping user questions to graph analytical tasks
  - Executing analytical tasks at scale
  - Pattern discovery:  What are new emerging patterns?  What is fading away?
  - Question Answering:
    - Entity based Querying (What, Who, When, Where)
    - Hypothesis Generation (Why)

# Outline



- Introduction
  - Knowledge Graphs
  - Challenges
- **Motivation (Use Cases)**
  - Domain Specific Querying of Web
  - Question-answering for Climate Science
- Technical Approach
  - NLP, Entity Disambiguation, Relation Learning
  - Frequent Pattern Mining, Question Answering
  - What's unique
- Results
- Status and Future Work

# Use Case : Domain-specific Querying of the Web




- Input:  multi-month web crawls obtained using domain expert suggestions
- Analytic questions:
  - Find top manufacturers and model names
  - Find components and features
  - Who is popular?
  - What are new releases?
  - How are companies and components related?
  - How is country and product related?

# Question-answering for Climate Science



- DOE's ARM program wants to make focused investments on experimental campaigns
- Monitor scientific literature to find which campaigns or instruments or data products are being cited
- Our target problems:
  - "What are the papers written on __**aerosols**_____?"
  - "What datasets are used for ___**aerosols**_____ publications?"
  - "What primary measurements are represented in the subset of **aerosol** publications?"
  - "What instruments are represented in the **aerosol** publications?"
  - "What sites are most represented in the **aerosol** publications"?

# NOUS Motivation



- Tasks involved in building  KB are not domain specific
- Most questions can be mapped to a set of common graph analytical tasks.
  - Tell me about X
  - Tell me about X in context of Y
  - What are recent trends about X
  - How are X and Y related
  - Why did X do Y
- Users don’t care about a graph, they have questions
  - Graphs are our way to model the world
  - The ability to transform the data into a graph allows us to bring all the database/graph algorithm toolkits

# Outline

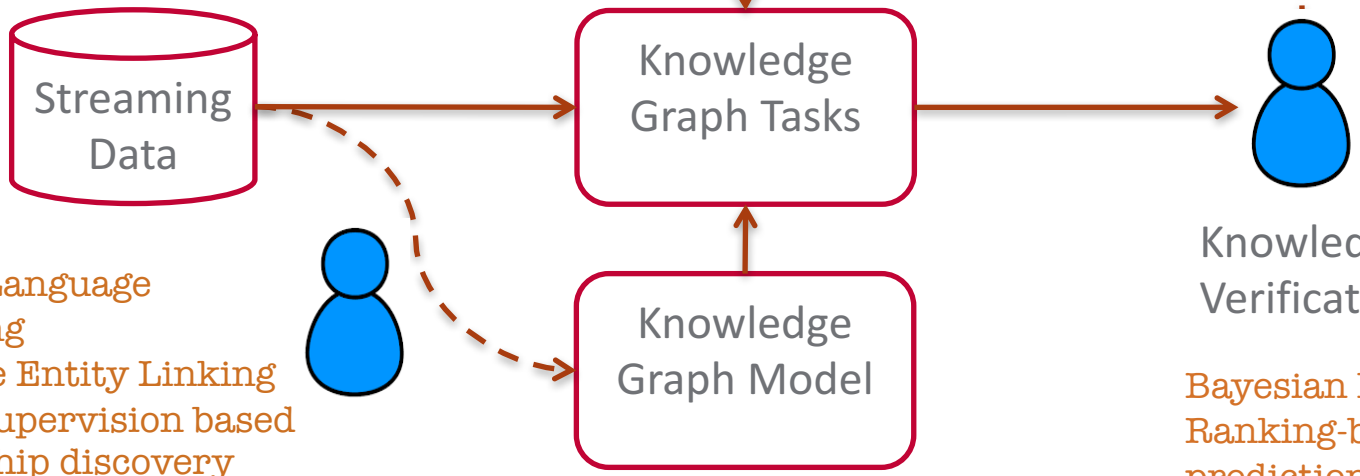


- Introduction
  - Knowledge Graphs
  - Challenges
- Motivation (Use Cases)
  - Domain Specific Querying of Web
  - Question-answering for Climate Science
- **Technical Approach**
  - NLP, Entity Disambiguation, Relation Learning
  - Frequent Pattern Mining, Question Answering
  - What's unique
- Results
- Status and Future Work

# NOUS Workflow



- Continuous pattern discovery.
- Question Answering




- Natural Language processing
- Collective Entity Linking
- Distant supervision based relationship discovery

Base KB + Streaming Data

Bayesian Personalized Ranking-based link prediction to assign confidence

Apache Spark

# Triple Extraction from Natural Language



- Use existing tools
  - Stanford Core NLP
  - Open IE

*Apple's New Challenge: Learning How the U.S. Cracked Its iPhone*

By KATIE BENNER, JOHN MARKOFF and NICOLE PERLROTH MARCH 29, 2016

A worker checking the innards of an iPhone at an electronics repair store in New York City last month. Eduardo Munoz/Reuters

SAN FRANCISCO — Now that the United States government has cracked open an iPhone that belonged to a gunman in the San Bernardino, Calif., mass shooting without Apple's help, the tech company is under pressure to find and fix the flaw.

But unlike other cases where security vulnerabilities have cropped up, Apple may face a higher set of hurdles in ferreting out and repairing the particular iPhone hole that the government hacked.

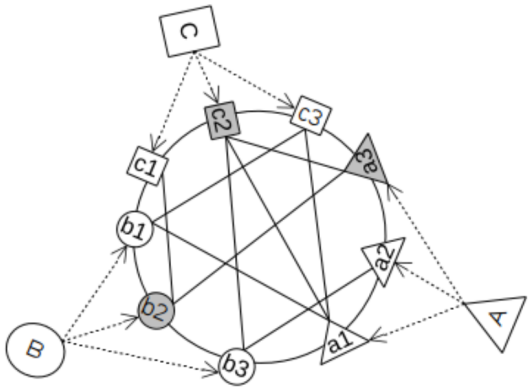
| | | |
|---|---|---|
| the United States government | crack | an iPhone that belonged to a gunman in the San Bernardino |
| Apple | repair | the particular iPhone hole that the government hacked. |
| Federal officials | specify | the procedure used to open the iPhone |
| Federal officials | deny | to specify the procedure used to open the iPhone. |
| Jay Kaplan, chief executive of the tech securAppley company Synack and a former National Security Agency analyst. | say | Apple has to earn the trust of Apples customers," |
| the F.B.I. | crack | Mr. Farook's |
| LegbaCore, which previously found and fixed flaws for Apple. | find | flaws for Apple. |
| LegbaCore, which previously found and fixed flaws for Apple. | fix | flaws for Apple. |

**The challenge: turning this into a high quality representation**

# Entity Disambiguation and Relation Learning



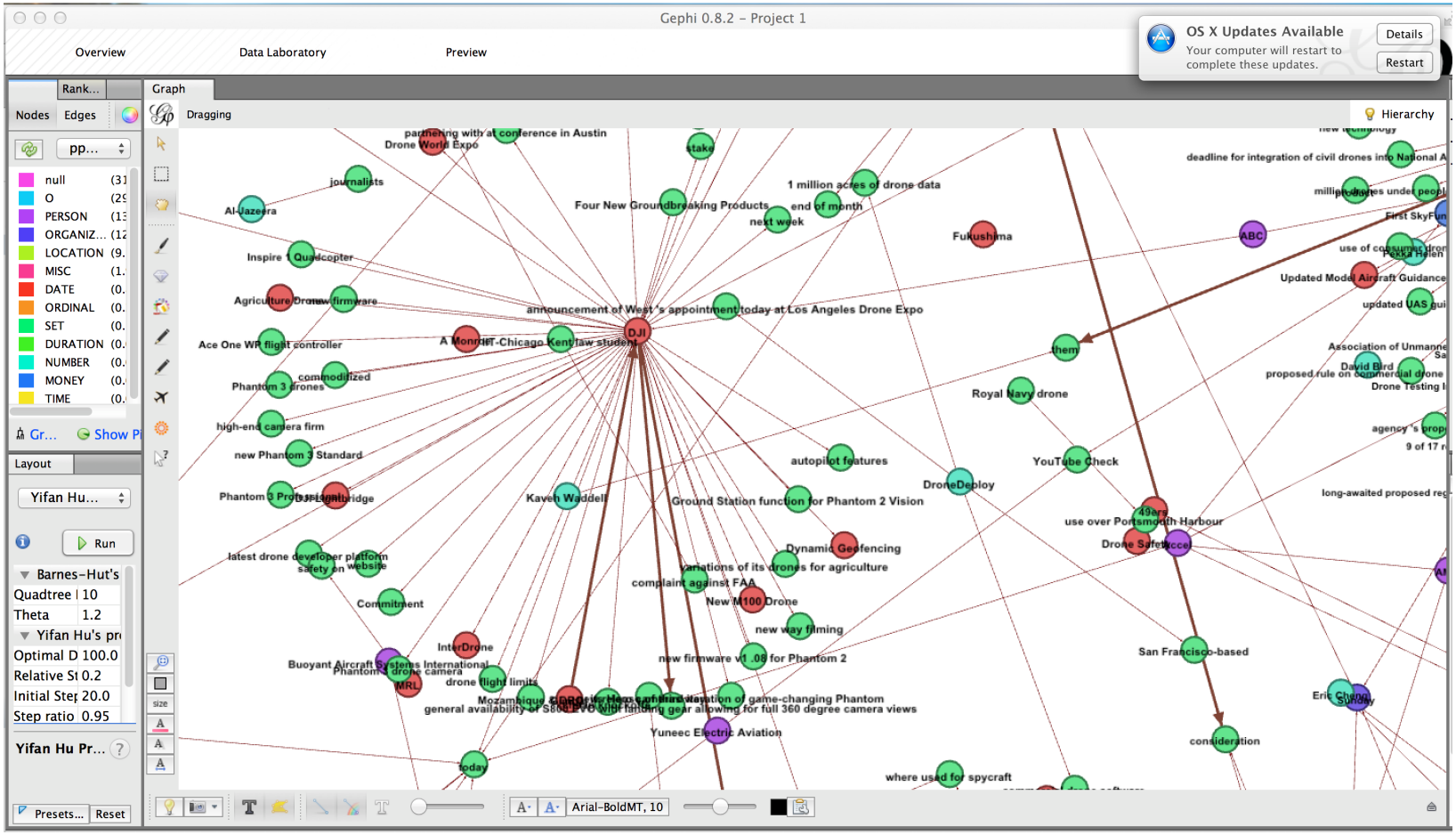
- Implements Collective Entity Linking from Han et al, SIGIR 2011
  - Key idea: Search the graph with matches for all the above terms, build a mesh of terms and their related terms and pick the most densely connected combination
  - Out of box performance low (higher 60%), with domain-specific rules (85-90%+)
- Relation Learning
  - Implements Distance Supervision




Han et al in "Collective Entity Linking in Web Text: A Graph-based Method, SIGIR 2011"

# Now we have a Graph!

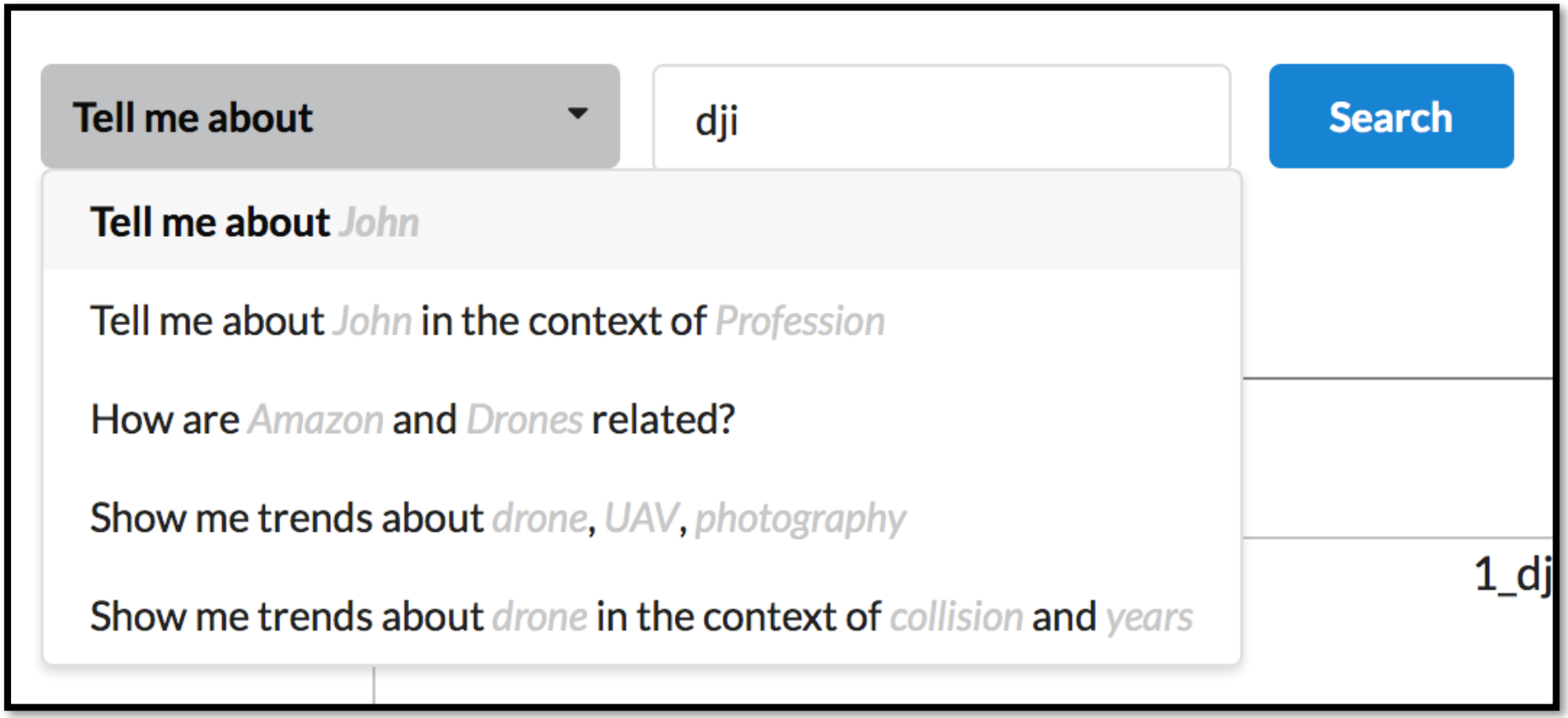
Pacific Northwest NATIONAL LABORATORY
*Proudly Operated by Battelle Since 1965*



# A Different Approach to Querying



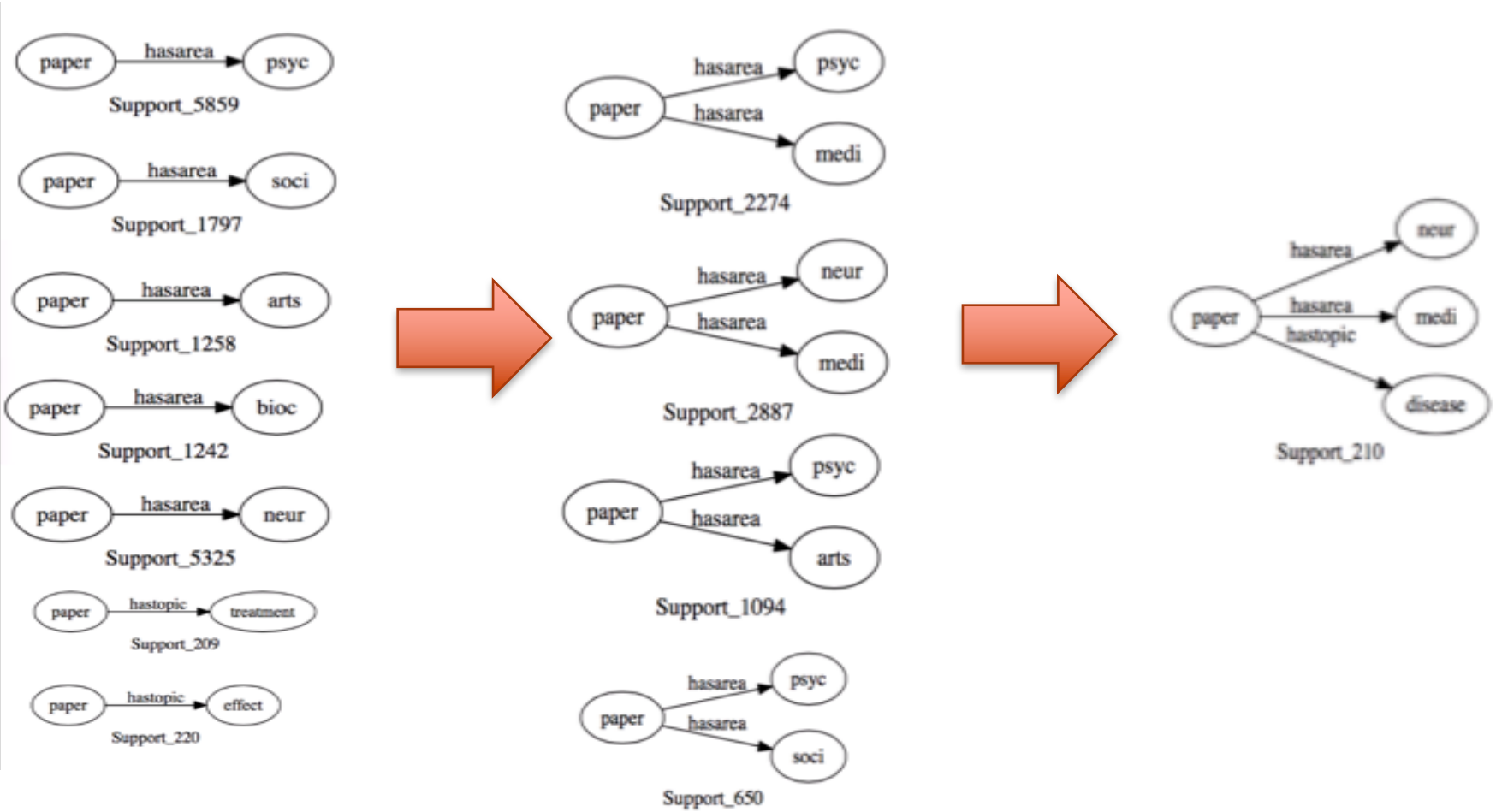
- Let's not demand users learn SQL or SPARQL
- Think in plain English, and we will transparently translate queries in background



# Task 1: Finding Patterns from Data Stream



- A Pattern Growth approach : VLDB 2017 [Under Review]

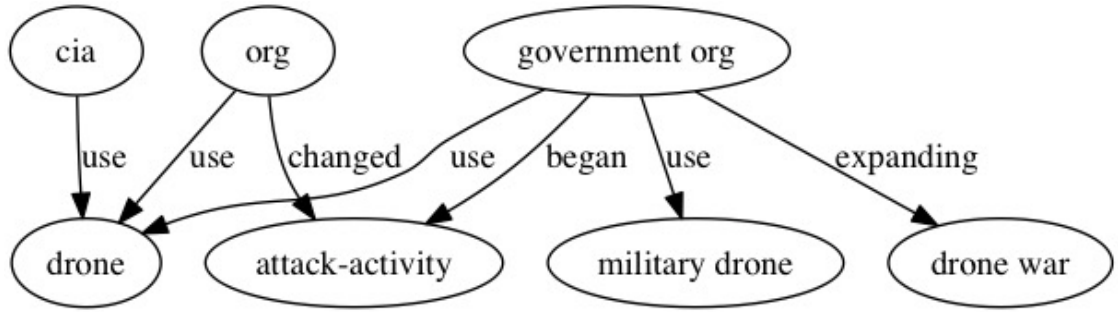


# Task 1: Finding Patterns from Data Stream (Contd.)

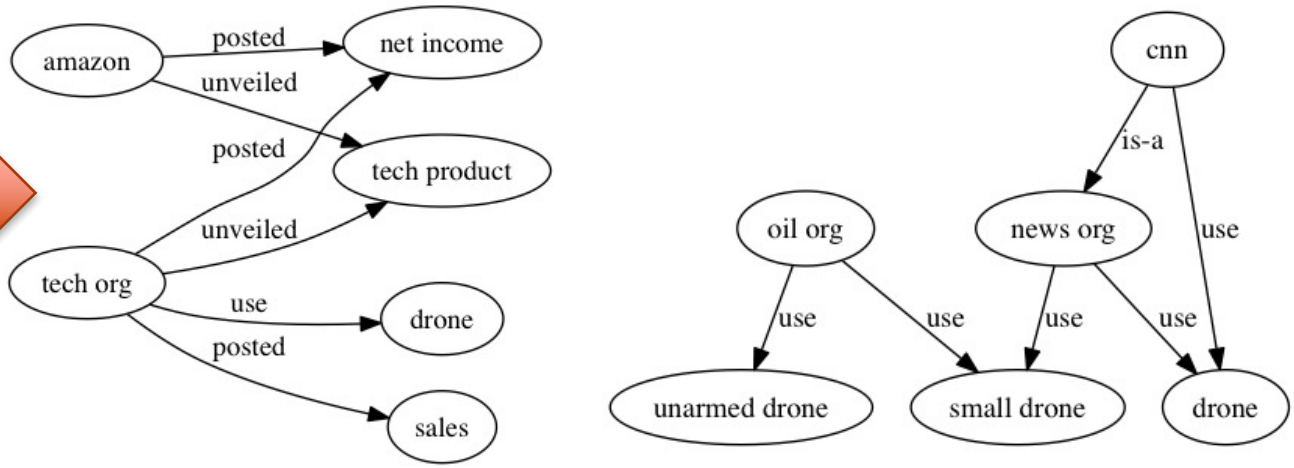


We discover behavioral patterns of drone related entities (WSJ, 2010-2015)

Early in the period




Towards the end




**Patterns help us understand the shift in a domain, and discover new trends**

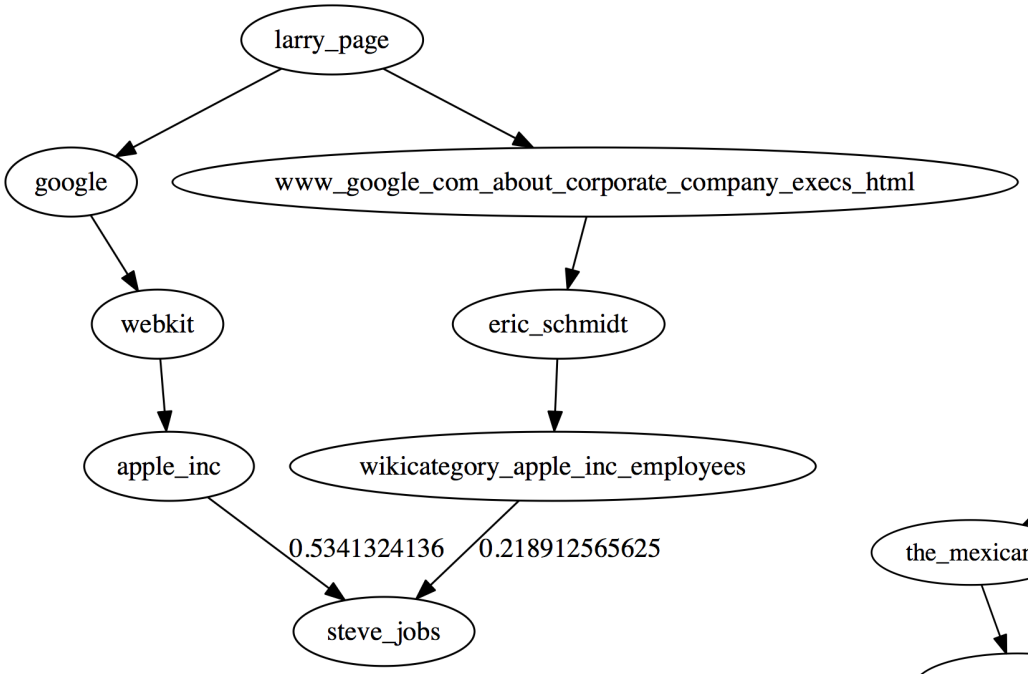
# Task 2: Question Answering for Relationship Explanations

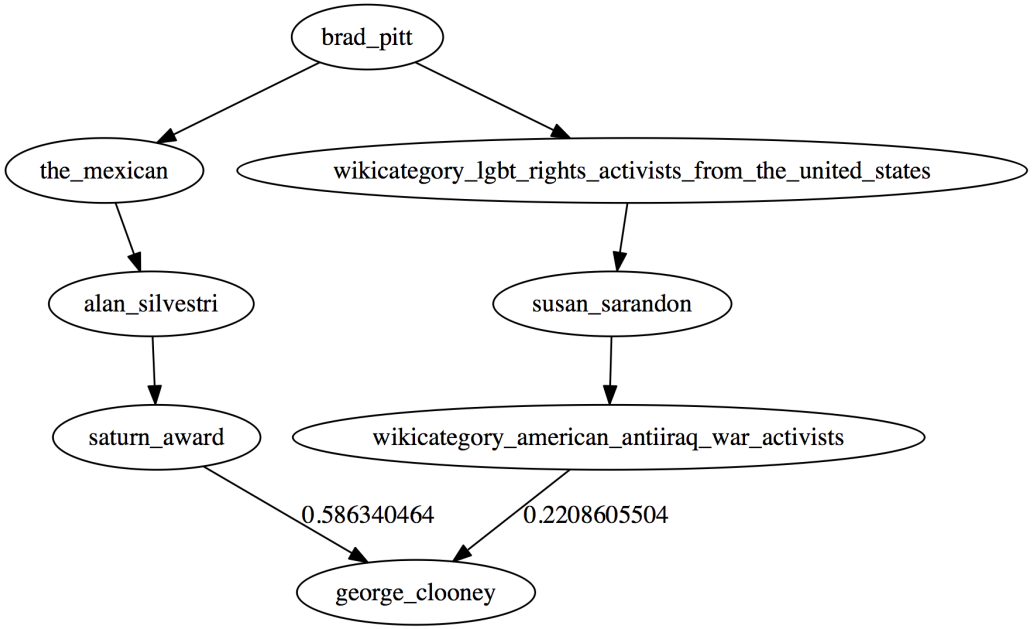


- Our goal is to learn “patterns of explanations” by walking the graph structure
- Given a new question and the knowledge of such patterns, we can generalize and answer questions about unseen entities
- Examples:
  - Why would Ford buy Palladium?
    - Ford is-a automotive company, automotive company has-part catalytic converters, catalytic converters has-material Palladium
    - Answer pattern:  A is-a company B, B has-part C, C has-dependency D
  - Why would Sutanay visit SFO in August 15?
    - Sutanay has-interest Machine-Learning, SIGKDD related-to Machine-Learning, SIGKDD has-location SFO, SIGKDD has-start-date August 13

# Results – Explanatory Paths using coherence





# Use Case Results: Domain-specific Querying of the Web



- Input Size:  2 million+ webpages
- Graph construction: 64-node cluster, each node with 16 cores
- Graph Analytics:  16-node Spark/Hadoop cluster, each with 16 cores

# Analytics step by step



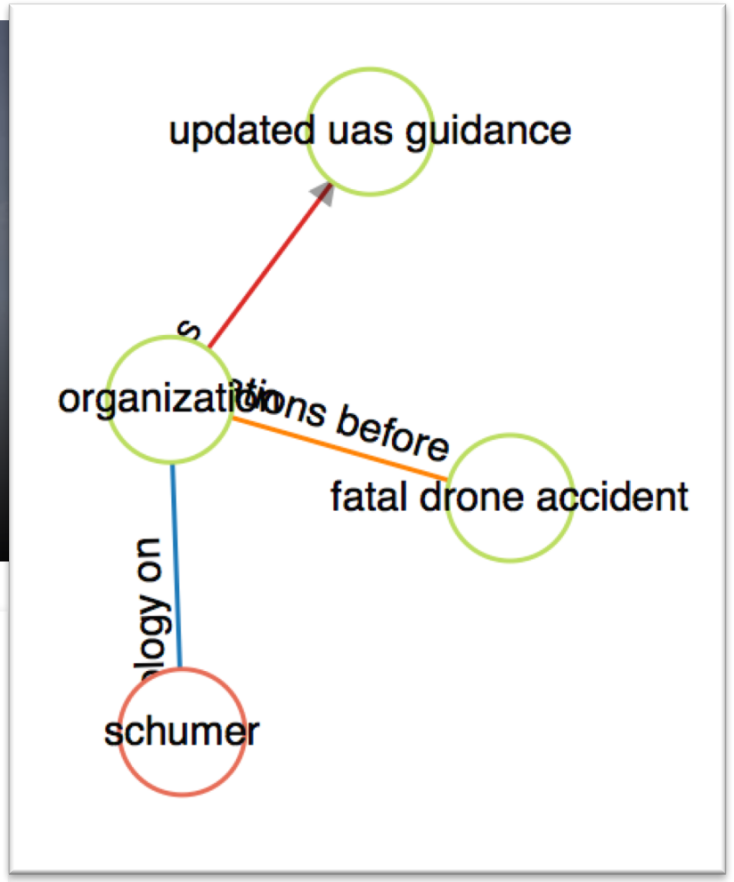
- Starting with analyzing popular entities in the extracted knowledge graph
- Extracted node labels from 2013-03 **Robotshop** and **HobbyKing** subgraph and found most popular entities
  - Arduino is a top entry with 49 mentions in one and 24 mentions in other
- Search triples from **Amazon.com** matching with Arduino, found
  - Tera-Controller-Side-Pin-Connectors-ARDUPILOT
  - MRM-Zeus-20-Amp-ESC
  - YKS-Upgraded-Controller-Absorber-Quadcopter
- Discovered the link between AudoPilot and Arduino from the **Wikipedia** page
- What was missing in the data was the **human guidance**: Tracking autonomous drones and their uses

# Picking up real events by Patterns




## Schumer presses FAA to require geo-fencing technology on drones

*Sen. Charles Schumer plans to introduce law aimed at keeping unmanned aircraft away from areas like airports*



# What's Phenomenal Here?



- **We are** data scientists, **not drone experts**
- We **processed** nearly **two million** web crawls **with minimal custom code** and built the **Knowledge Graph**
- Analyzed trends, connected dots across multiple source and **presented a hypothesis**:
  - Is tracking autonomous drones and their use important?
- Starting with the data analysis to coming up with the question : **took less than 1.5 hours**

# Key Takeaways



- Open Source KB Construction and Querying Pipeline https://github.com/streaming-graphs/NOUS
- Version 1.0:
  - Support for extracting standard entities, relation extraction via distant supervision,
  - Advanced trending and explanatory questions
  - The ability to answer queries where the answer is embedded across multiple data sources
- All algorithms implemented on top of Apache Spark and Tensorflow

# Status and Future Work



- We are committed towards developing an open source community
  - A MySQL for Knowledge Graphs!
- Preferred method of payment ☺
  - Add test code (in Scala or Python)
  - Break the code AND show an important class of problem

- Version 2.0:
  - Improved KB quality,
  - Information maintenance over time,
  - Initial support for human-computer interaction.
  - Incorporate algorithms for latest advances in NLP, ED

# Questions

# I. Event: news around autonomous drones

UAV EXPERTS HOM

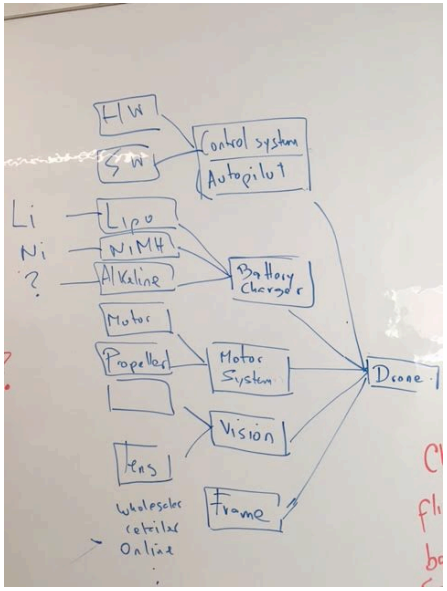

## DJI Phantom can now perform autonomous flight

Ben Coxworth | July 2, 2014


Thanks to an app update, the DJI Phantom 2 Vision and Vision+ are now able to follow pre-programmed flight paths . View gallery (2 images)

## II. What products they may use?

From Supply Chain Analysis Team



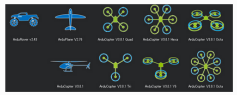

## III. Tell me about software components

Hardware - The embedded systems and peripheral sensors that act as the vehicle's brain, eyes, ears, etc.

Almost any mobile machine can be transformed into a robot, by simply integrating a small hardware package into it.

Firmware- The "skill set" code running on the hardware, which configures it for the kind of vehicle you've put it in. You choose the firmware and vehicle that match your mission: Plane, Copter, Rover...

The choice is yours – one autopilot for any mission. An easy firmware update is all it takes to repurpose your hardware into a different role.



Software- Your interface to the hardware.

Initial set-up, configuration, and testing. Mission-planning/operation, and post-mission analysis.

Point-and-click intuitive interaction with your hardware, or advanced custom scripting for niche mission profiles. Options are everything with ArduPilot.

## Ardupilot: An Open Source award winning platform