# HamSci Data Plane + Satellite

Michelle Thompson

Open Research Institute

For HamSci

January 17, 2019

## Abstract

HamSci, or Ham Radio Science Citizen Investigation, advances scientific research and understanding through amateur radio activities. Primary cultural benefits include the development of new technologies along with providing excellent educational opportunities for both the amateur community and the general public.

The HamSci Space Weather System is a HamSci project. HamSci Space Weather Stations form a distributed radio network dedicated to space weather research. HamSci Space Weather Stations produce receiver data from transmitters associated with coordinated observations. Sensors range from ground magnetometers, to ionospheric sounders, to lightning detectors and more. The diversity of sensor types means a wide variety of radios can participate.

A collaboration between HamSci and Tucson Amateur Packet Radio (TAPR) was proposed at the Digital Communications Conference (DCC) on 14-16 September 2018 in Albuquerque, New Mexico. Discussions about custom software-defined radio hardware designed, built, and sold by TAPR as HamSci Space Weather Stations began at the conference and continued though a Google Group.

HamSci presented at the TAPR DCC Sunday Seminar. Slides introducing possible sensor types from that presentation are reproduced below and throughout this document.

# Personal Terrestrial WX Station

- Multi-instrument
- Internet Connected
- Easy Set-Up
- Reasonable Cost


Ambient Weather WS-2902

# Personal Terrestrial WX Station

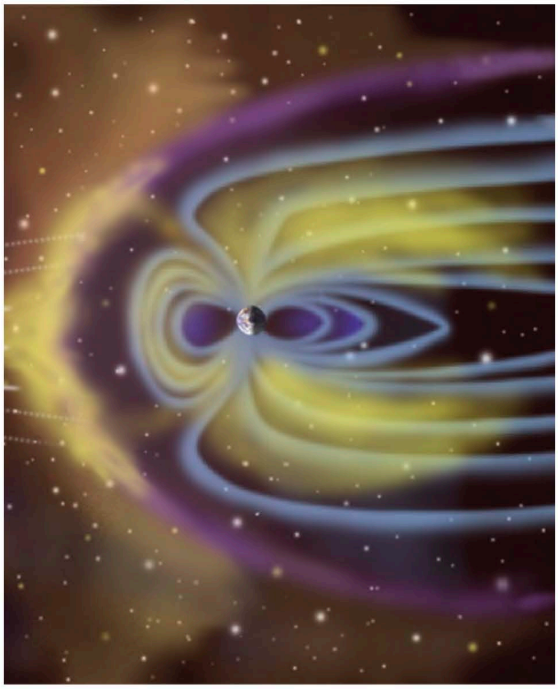
# Instrument Possibilities

- Ground Magnetometer?
- GPS-TEC Receiver?
- Ionosonde?
- Riometer?
- WWV/Standards Station Monitor?
- RBN/PSKReporter/WSPR Receiver?
- Lightning Detector?
- Others?

*What makes sense for a personal, ground-based local station?*

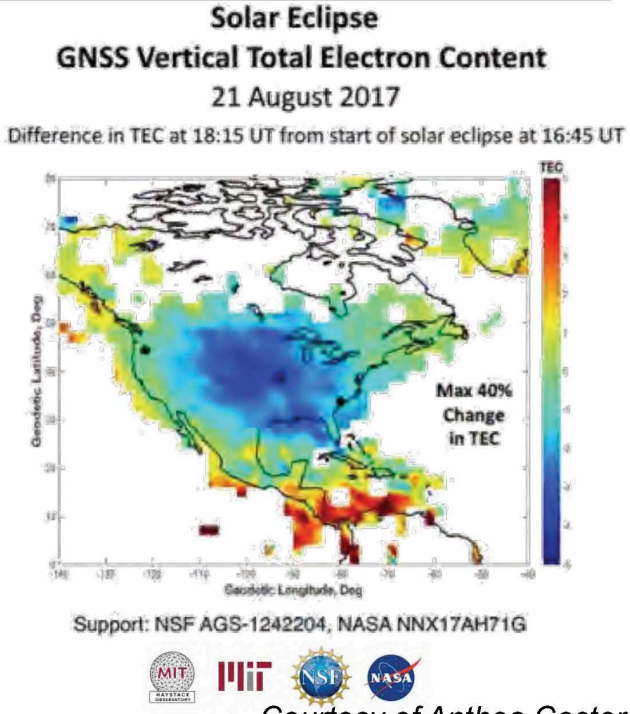
# Ground Magnetometers

- Detect Ionospheric & Space Currents
- Geomagnetic Storms
- Geomagnetic Substorms
- Kp and Ap are derived from GMAGs data.



# GPS Total Electron Content

- Total Number of electrons between ground and GPS Satellite
- Measured by examining delay between two GPS Frequencies
- Traveling Ionospheric Disturbances
- Storm Effects
- Ionospheric Scintillations



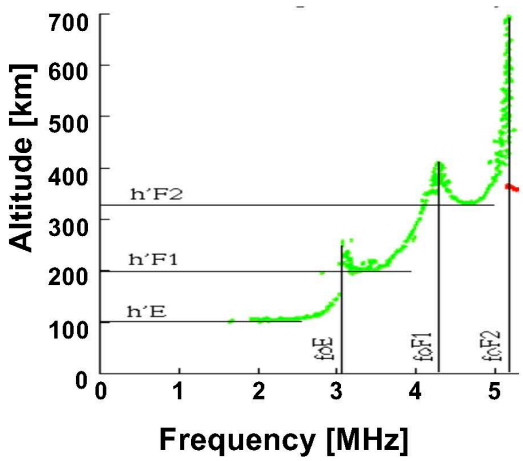

*Courtesy of Anthea Coster*

# Ionosondes

- Vertical Incidence HF Radar
- Measure Plasma Density for bottomside Ionosphere

$$f_{pe} \approx 9\sqrt{n_e}$$


[Dr. Terry Bullett, W0ASP, U of Colorado]



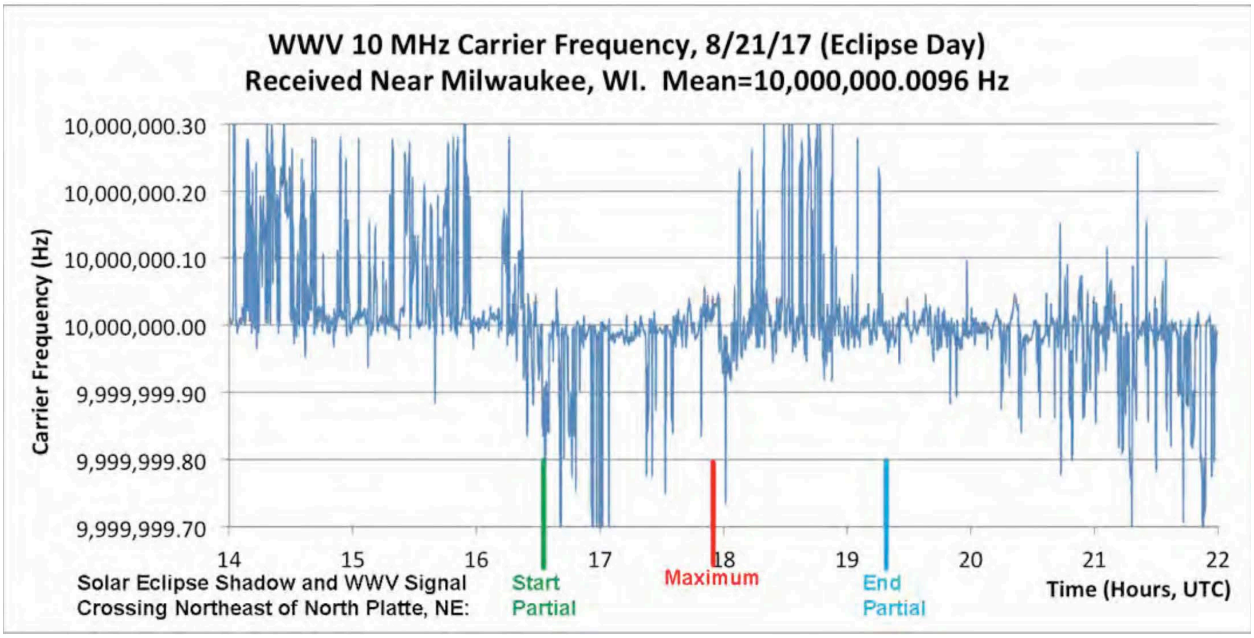
# Riometer

- **R**elative **I**onopheric **O**pacity **M**eter
- Directly measures absorption of cosmic rays
- Indirectly measures electron density, particle precipitation
- Typically passive instrument 30-50 MHz

IRIS - Imaging Riometer for Ionospheric Studies in Finland (http://kaira.sgo.fi/)

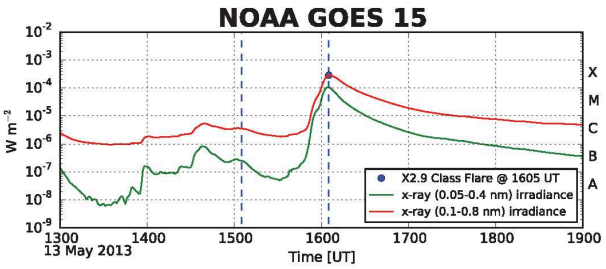
Photo: Derek McKay

# WWV/CHU Standards Monitor

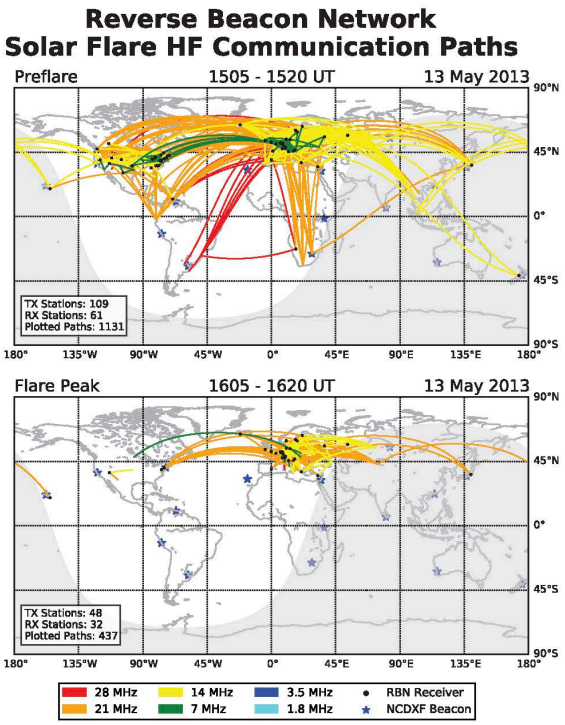



Steve Reyer, WA9VNJ

# RBN/PSKReporter/WSPRNet RX




[*Frissell et al.*, 2014, *Space Weather*]



# Lightning Detector

- Signatures from LF to VHF/UHF
- On HF, lightning noise can propagate long distances and disrupt communications

Photo: Jessie Eastland
(https://en.wikipedia.org/wiki/File:Desert_Electric.jpg)

The receiver network employs a wide variety of sensor types. Combining sensor data from disparate sources, when the end result has greater certainty, accuracy, or quality than if the data was used individually, is called sensor fusion. The HamSci Space Weather System, as proposed above, can be affordably accomplished through sensor fusion.

For example, a $150 dedicated lightning detector on a Raspberry Pi in Florida, USA can participate in this network with a $6331 USRP X310 station sampling at highest rate and bandwidth in Madrid, Spain. The inexpensive data from the lightning detector may enhance the data from the expensive radio and increase scientific knowledge. Another example is a set of five inexpensive radios configured as ionosondes. The data combined is better than any one station's individual contribution.

Open Research Institute (ORI) proposed an open source cubesat as part of the network. Observing from ground and space simultaneously provides substantial additional scientific value. The receiver network can be coordinated to make scheduled observations that align with satellite passes. This can be enabled with SatNOGS open source software. See https://satnogs.org/ for more information about this open source satellite network on the ground.

The central challenge of the HamSci Space Weather Station project is not the radio hardware. It is how the radios are interconnected, what metadata is accepted, how observations are scheduled, how the interactions between different sensor data is modeled, and how the large quantity of data is handled, organized, and re-used over time. This is the Data Plane.

## Introduction

HamSci Space Weather Systems produce radio receiver data. In some cases, the data stream may be small in size or volume. Data on lightning strikes or for particular, narrowly-defined, or infrequent atmospheric observations can be transmitted over normal retail internet channels without requiring any special equipment or services. In other cases, the data stream may be very large in size and volume.

The bandwidth of interest is up to 60MHz in frequency. Sampling spectrum 60MHz wide with any degree of precision in resolution and time will require substantial resources in data transport, storage, and processing. Slides 77-82 from the TAPR DCC Sunday Seminar propose a ring buffer to manage the observations. These slides are reproduced below. Using a ring buffer means that there's a window of opportunity to identify a set of desired observations before they circulate out of the ring buffer and are lost. Rapidly identifying interesting "unknown unknowns" is a requirement for successful ring buffer performance. Ring buffers can be local, in the cloud, or a combination.

# HF Receiver Instrument

# Where do we start?

- **General purpose HF Receiving Instrument.**
- **Why?**
  - Few networks of widespread scientific HF radio receivers currently exist.
  - "Signals of opportunity" available.
  - Extremely flexible research tool.
  - Directly applicable to ham radio.
  - Radio is TAPR's Bread and Butter ☺

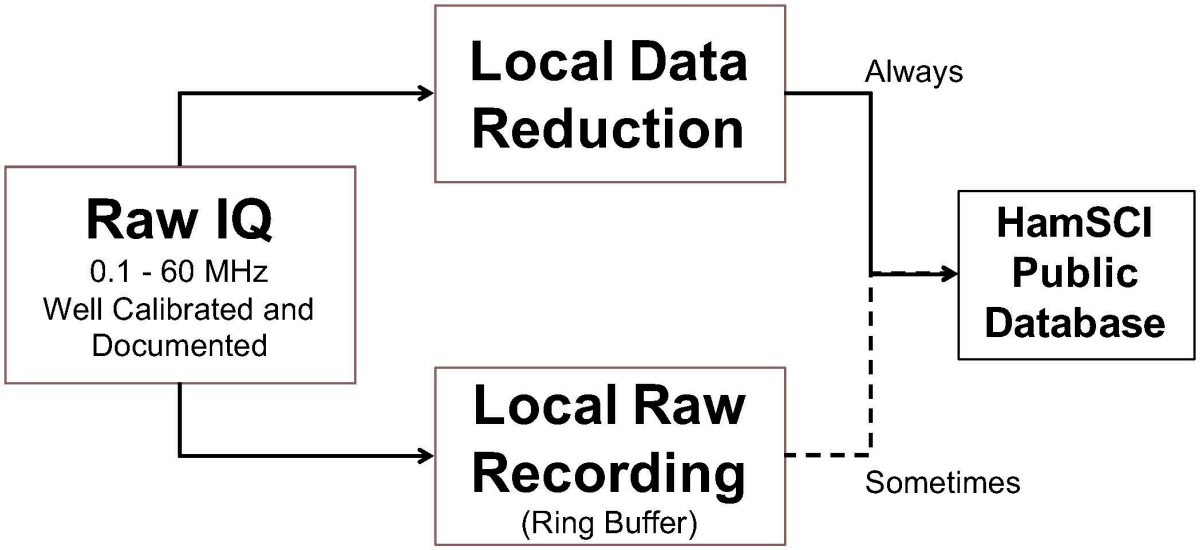
# Where do we start?

- General purpose HF Receiving Instrument.

**Raw IQ**
0.1 - 60 MHz
Well Calibrated and Documented

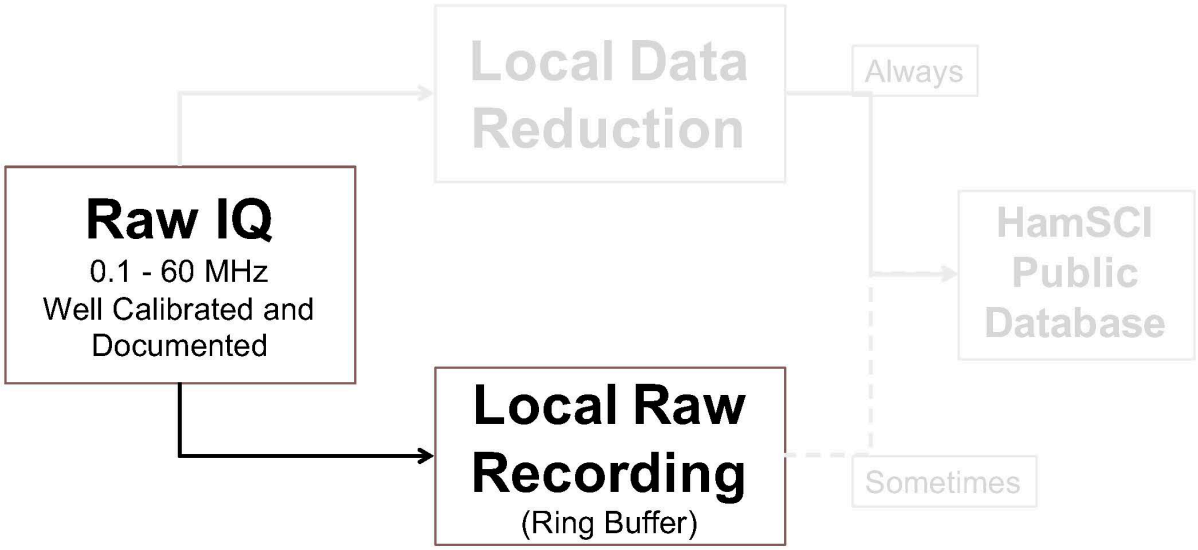
# Where does this go?

- General purpose HF Receiving Instrument.



# Where does this go?

•General purpose HF Receiving Instrument.



# Quality Raw IQ is the Foundation

- Quality HF raw IQ → all downstream research and operational products.

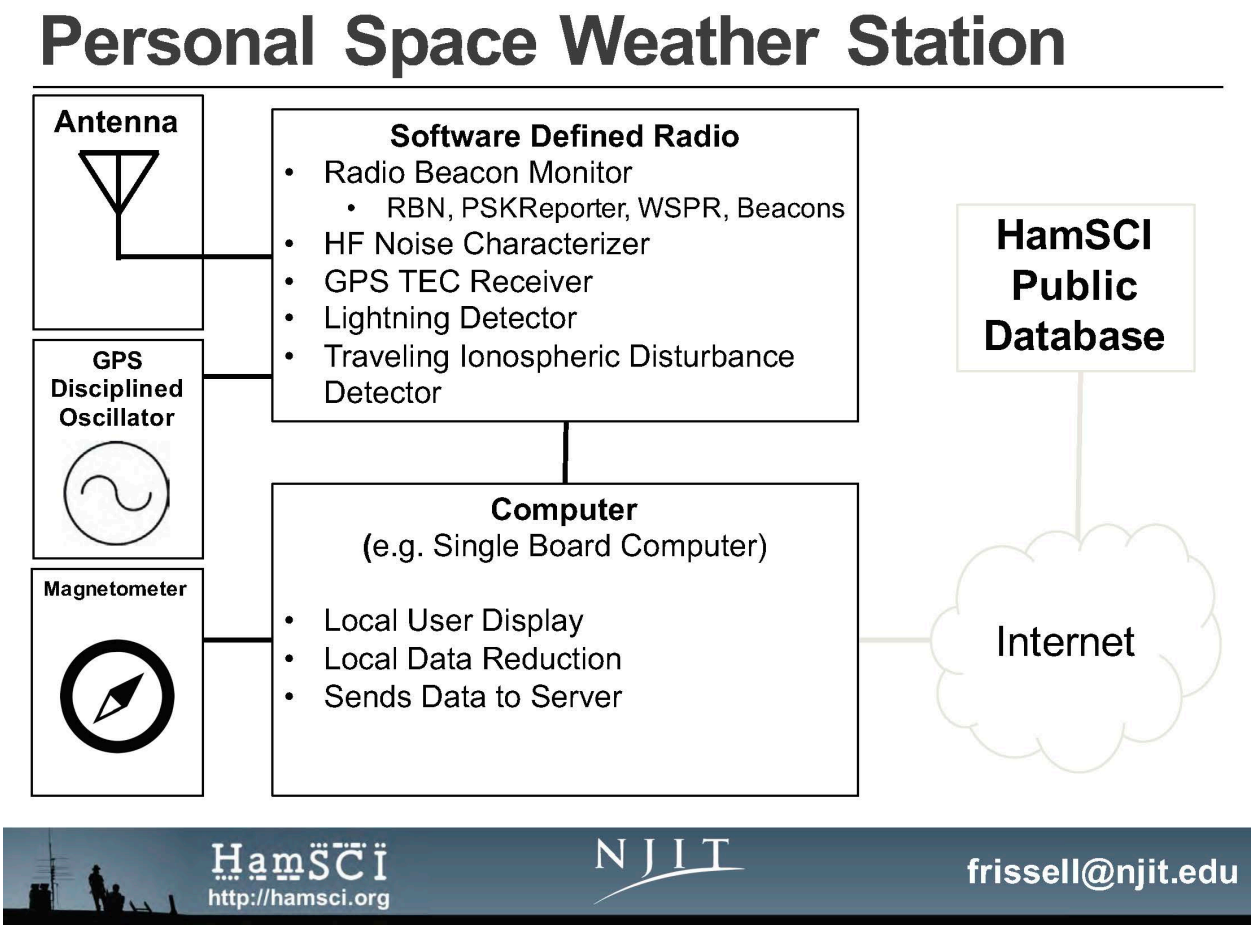
**Big Data**

The size of the data that we are dealing with means we very likely have a Big Data situation. Big Data groups together processing, collection, storage and visualization of large quantities of data. Data Science is the process of extracting knowledge from data. Data is collected, information is derived, and knowledge is gained. Knowledge comes from analysis and synthesis of information. Information comes from classifying, organizing, and interpreting the data.

HamSci expects to be dealing with some variety of structured digital samples and information from the radio spectrum. The amount of spectrum, the geographical location, antenna gain, antenna pattern, the resolution, the time accuracy, and the signal to noise ratio are the essential characteristics of these radio measurements. The characteristics will vary for each sensor type. The characteristics will vary within a population.

Successful data fusion reduces the cost of the receiver network while increasing participation.

Slide 69 is a high-level view of the network. It is reproduced below.


Personal Space Weather Station
Antenna
GPS Disciplined Oscillator
Magnetometer
Software Defined Radio
• Radio Beacon Monitor
• RBN, PSKReporter, WSPR, Beacons
• HF Noise Characterizer
• GPS TEC Receiver
• Lightning Detector
• Traveling Ionospheric Disturbance Detector
Computer
(e.g. Single Board Computer)
• Local User Display
• Local Data Reduction
• Sends Data to Server
HamSCI Public Database
Internet
HamSCI
http://hamsci.org
NJIT
frissell@njit.edu

The "Internet" and "HamSci Public Database" correspond to what ORI calls a Data Plane.

**Internet**

While the internet is the most obvious way to network a distributed receiver system together, very remote stations or stations that produce very large amounts of data could ship their data on physical media. There is a point where the cost to exfiltrate the data exceeds the cost of saving it to a local hard drive and periodically shipping that hard drive to the location where the data is processed. Shipping physical media introduces substantial potential latency. Stations located in areas without reliable internet access will require alternate interconnections and data transportation.

**HamSci Public Database**

In order to get the full potential out of data, it must be available to anyone that has an interest in using it. The data must be available without unreasonable obstacles and at affordable prices. Data must be provided along with with models and equations so that interested users do not have to reinvent well-understood processes of deriving information from the data. If a user is interested in deriving new information from the data, then the new models and equations can be published alongside the existing ones. Over time, this increases the value of what we call a dataset. Data by itself isn't a dataset. A dataset is a documented searchable set of data that includes models, tools, and equations. A database is an excellent template for a dataset.

A relational database is structured to recognize relationships among stored items of information. A relational database may meet HamSci data needs. If the data has clearly documented relationships, is less than or equal to single digit petabytes in size, has fewer than 300k writes per second, then a relational database is affordable and works well.

Hadoop is an open source framework that enables distributed processing of large data sets across clusters of computers. The data does not have to be structured by

relationships. Something like Hadoop should be considered or added if HamSci has a lot of data that users do not know what to do with. In other words, if the relationship between the data is unknown. If relationships between the data cannot easily be modeled, then a relational database is not a good fit.

From the HamSci Sunday Seminar presentation, it seems that finding new relationships between the disparate sources of sensor data is a fundamental goal of the project. Many potential relationships between atmospheric data are not yet known. Therefore a relational database would not be the best fit.

**What isn't yet represented?**

The network connections (internet, transport of physical media) and the public database are required elements. The models, equations, and computing resources for sensor fusion are not yet represented or defined.

**Metadata**

From slide 85, a list of RF Instrument Metadata is given. Metadata is information added to samples to record sample characteristics, increase findability, and define station identity, configuration, and location. The slide is reproduced below.

# Importance of Metadata

- RF Instrument Metadata
  - Center Frequency
  - Bandwidth
  - Impulse Response
  - Sampling Fidelity (e.g. # of bits)
  - Voltage to ADC Calibration Number
  - Timestamp (UTC Locked)
- Station Metadata
  - Station ID
  - Station Configuration
  - Geographic Location

Metadata is required for successful sensor fusion. The identification and integration of an open source metadata protocol is one of the first steps in the Work Plan. The selection of any particular metadata protocol has significant performance repercussions.

Is there an existing open source RF metadata protocol that fits our needs? Proposed options include but are not limited to:

Haystack
https://github.com/MITHaystack/digital_rf

GNU Radio SigMF
https://github.com/gnuradio/SigMF

Ion Metadata Working Group
https://github.com/IonMetadataWorkingGroup/GNSS-Metadata-Standard

**Visualization**

Visualization is the art of representing data in a visual manner. Graphs, diagrams, maps, animations and videos rapidly communicate essential knowledge from the dataset. The users of the datasets, visualizations, and knowledge are at the opposite end of the ecosystem from the station operators, who are producing and shipping data. However, beautiful visualizations can be a powerful additional motivation for station adoption. The dominant design pattern for software defined radio is a spectrum display plus waterfall. Visualization at any level, from Space Weather Station to the HamSci Public Database, may require additional hardware and software development.

**Accessibility**

Will the stations provide access and equal opportunity to people with diverse abilities? Can a low vision individual participate? Can someone on limited income afford to participate? Can the stations be deployed to areas with limited infrastructure and communications?

**Data Plane**

The Data Plane is how the radios are interconnected, what metadata is accepted, how observations are scheduled, how the interactions between different sensor data is modeled, and how the large quantity of data is handled, organized, and re-used over time. The outcome of a Data Plane research and development effort is a network that performs sensor fusion and delivers scientifically useful datasets. The Data Plan specification contains the requirements for building this system. This paper lays the groundwork for developing that specification.

What is the minimum level of radio performance required for each sensor? Radio specifications need to be derived and published for each sensor type. This is required in order for them to be built or included.

The output of each sensor has a structure and metadata associated with it. Some sensors have simpler or smaller data than others, which means some metadata may not exist for some samples. Some data will represent important singular events. Other data is useful only as a stream, as the statistics that emerge over time provide the essential features. Nothing in the specification should unnecessarily limit, impede, or exclude data.

Exactly what information needs to be derived from the data in order for the radios to work together as a group at a performance level higher than their individual specifications?

If digital signal processing techniques are used to create a body of high-quality data from less-capable receivers, exactly what are the inputs and outputs of those digital signal processing functions? What digital signal processing functions need to be defined in the Data Plane for successful sensor fusion?

There are many ways to process multi-sensor data. Proposed approaches and

techniques include but are not limited to:

Evolutionary algorithms
Classical optimization theory
Bayesian probability
Information theoretic functions
Force aggregation
Dempster-Shafer theory
Machine learning
Statistical models
Complementary, redundant, and cooperative data fusion

Multiple techniques should be simultaneously supported. A properly designed dataset and data flow will allow a wide variety of processing.

Information is derived from disparate sets of raw data. Information leads to feature identification. Features generally describe an entity in the environment under study. In atmospheric science, this could be something like a particular type of structure occurring at a particular time in the atmosphere. Being able to predict the emergence or occurrence of features, discovering connections between features, and establishing causes and effects between features is the goal of the project. New discoveries at the feature level may cause blocks of raw data to be fetched and re-processed based on new hypotheses or capabilities. With multi-sensor data, this iterative process has a lot of dimensions to explore.

Nothing in the specification should unnecessarily limit, impede, or exclude functions across the data set. Digital signal processing techniques must not remove information that machine and deep learning may need when the datasets are studied.

Machine and deep learning keep coming up with surprising new ways to use data. A

diverse body of measurements with enough metadata to discover new relationships is well within the combined abilities of the amateur and academic communities.

The Data Plane specification needs participation, input and review from data science, machine learning, and deep learning subject matter experts to fully live up to the potential.

## Research Questions

This is a summary of the research questions raised in this paper.

1) What kind of metadata is needed to enable the best possible scientific research?

2) If digital signal processing techniques are used to create a body of high-quality data from less-capable receivers, exactly what is required as the inputs to those digital signal processing functions? What digital signal processing functions need to be defined in the Data Plane for successful sensor fusion?

3) How is scheduling triggered? How are triggers updated?

# Budget

## Human Resources

Engineering and support costs for research and development of a sensor fusion big data system are considerable. They are listed here in order to provide some estimate of the value of the donated, volunteer, or granted time. It is understood that the project will be pursuing grants and institutional support and that there are other ways to express the time value of human resource money.

Big data and sensor fusion engineering salary range is US$90,000-$130,000, according the the IEEE.
Customer support salary range is US$35,000-$43,000 a year, according to World at Work.
Technical support salary range is US$50,000-$65,000 a year, according to World at Work.
Components engineering salary range is US$86,000-$106,000 a year, according to World at Work.
Software engineering salary range is US$80,000-$115,000 a year, according to World at Work.
RF engineering salary range is US$79,000-$92,000 a year, according to World at Work.
Data scientist salary range is US$50,000-$200,000 a year, according to World at Work.

Depending on how many human resources are required for the project, and how long they are needed, the donated time could easily exceed US$1,000,000 per year.

## Data Storage

Big data storage expenses go up dramatically if the data needs to be retrieved quickly. It's not possible to scope this cost at this time due to the very large ranges of estimates and the number of unknowns.
Relational database costs vary widely.
Hosted Hadoop nodes cost approximately $4000 a year.

## Station Cost

Costing a complex electronics project such as a software defined radio generally requires components engineering, development of a bill of materials, management of the logistics of a complex parts order, and working with a manufacturer throughout the process to contain and reduce costs. While a rough estimate can be made looking at comparable radio systems and drawing on experience in the field, it's not possible to accurately predict the cost for something that has not yet been engineered.

The most expensive sensor is the most capable. This assumes the maximum RF bandwidth requirement of 60MHz (from TAPR DCC Sunday Seminar HamSci presentation), an ADC resolution of ~16bits, GPSDO timing, fast and reliable data storage, fast and reliable remote control, durable RF interface, power supply, cooling, enclosure, a broadband high-performance antenna, and a high performance low

noise amplifier. If local visualization is required, then an additional processor and graphics driving capability would be added. As of January 2019, cost estimates range from $1500 to $2500 for the highest performance station.

In the relatively low volumes quoted in the slide deck, a radio with the highest performance sensor is substantially more expensive than the target range of US$100 to US$500 presented at DCC.

The budget depends on the mix of sensors that are chosen for manufacture. Sensors range from the highest performance (and most expensive) to very modest (and still very useful) sensors. This mix is unknown at this time. The total number of stations given in the DCC presentation (1000) is much higher than any coordinated amateur or open source distributed network with similar capabilities in existence to date. As of January 2019, SatNOGS had 300 coordinated satellite observing stations after two years of growth. All SatNOGS stations are built or integrated by the operator.

Some simple budget models for the cost of 1000 stations are listed below.

**100% High Performance**
1000 of the highest performing stations, at a cost of US$1500, is US$1,500,000.
1000 of the highest peforming stations, if they were supplied as USRP X310s with GPSDOs, would retail for US$6331 each. This does not include antenna or enclosure or any additional RF circuitry. Purchasing 1000 would be US$6,331,000.

**Uniform Distribution Low-Medium Performance**
A mix of ten sensor types, ranging from US$100 to US$1000, with uniform distribution among 1000 stations, is US$550,000

**Bimodal Favoring Less Expensive**
A mix of two sensor types, where 90% are inexpensive multi-sensor station at US$300, and 10% are the more expensive high performance type at US$1500, is US$420,000

**Heterogenous Computing Saves Lots of Money**
Open specifications allow for operators to construct their own stations. Relaxed requirements due to sensor fusion and Data Plane algorithms reduce the cost of the custom hardware and allow for off-the-shelf options, as long as the radio complies with whatever is necessary to access the Data Plane. If half of the stations were constructed by the operator or purchased off the shelf by the operator, the cost goes down. It won't go down by half, due to volume pricing for some of the radio components. However, the cost for a DIY-friendly heterogenous receiver network could be low six figures. Including off-the-shelf less-

capable gear through the mathematics of sensor fusion can reduce the cost from astronomical to affordable. ***The reduction in cost comes with getting the math and the metadata right.***

## Maintenance Costs

Data scientist as manager salary range is US$110,000-130,000 a year, according to the IEEE.
Network engineer salary range is US$78,000-$88,000 a year, according to the IEEE.

# Work Plan

**This is a work plan proposed for a project with many unanswered questions. This plan has been written with input from competent and experienced people. All of them were willing to take on the risk of defining, describing, and proposing action under conditions of uncertainty. The project should evaluate, adapt, and incorporate findings to adjust the work plan as needed to support the scientific mission. Furthering our understanding of atmospheric science through productive collaboration between science and amateur activities is a goal worthy of our best efforts.**

There are many tasks that can be done in parallel. One of the biggest advantages to using software defined radio hardware for Space Weather Stations is that maximum performance stations can be designed and built starting today, with the assumption that they can be successfully configured and the metadata added later. High performance stations can provide streams of data well after the fact, as particular types of data can be produced from raw samples. Storing all raw samples produced by all radios may be probibitively expensive in the short term. Tasks 4 and 5 address the question of cost. Sensor fusion is a way to reduce the need for massive data handling and storage and to increase participation in the system. Successful sensor fusion does require a large research and development effort, which begins in earnest by Task 7. If any traction can be derived from multi-sensor fusion, it will be a permanent part of the landcape and will be part of the dataset definition. If sensor fusion research and development reveals nothing, then the resulting datasets are somewhat simplified, with data and associated metadata and fewer equations and models.

**February - March**

1. Metadata protocol research, development, integration, test, and documentation using off the shelf radio gear. Getting firsthand experience with each protocol's API is very important in order to make the best possible selection.
2. Individual sensor identification. Each sensor type is listed with the minimum required individual performance specifications. These include but are not limited to bandwidth, resolution, timing, frequency range, signal to noise radio, transmit power, exfiltration bandwidth (internet upload speed), uptime, power consumption, and cost.

**February - April**

3. Prototype individual examples of each sensor. Integrate metadata. Connect each of them to the internet. Measure upload and download (for the command link) bandwidth requirements, reliability (uptime), and latency in standalone operation.
4. Prototype a variety of local storage needs and options for collected data (SD card, hard disk, tape drive). Measure performance (latency, size, reliability, cost, maintenance interventions). Scope the cost.

5. Prototype cloud storage needs and options for collected data (AWS, dropbox, scaled NAS, Hadoop nodes, cloudera, etc.). Measure performance. Getting firsthand experience with each service's API is very important in order to make the best possible selection. Scope the cost.

**May - August**

6. Design review for metadata protocol selection or definition. Output of design review is a metadata protocol.
7. Identify and implement algorithms that enable sensor fusion. Measure performance. Make adjustments. The expectation is that sensor fusion will allow for relaxing requirements on the radios. This allows a much wider population of individual radio to participate. Improving coverage while preserving the quality of the aggregated data maximizes the potential science.
8. Design review for local and cloud storage options. Output of the design review is rank-ordered list of recommended options.

**August**

9. Begin the command and control functional study. The project proposes coordinated observations. Coordinated observation may require remote control. Remote control may require a central authority. Functions of the central authority must be defined and documented. Central authority must have reasonable security, authorization, and authentication. Output of the study is to decide whether or not remote control is required to produce desired results.
10. Incorporate satellite inputs into the model. Sensors in space have unique scheduling.
11. Define the schedule model. The schedule model comes from the sensor fusion feature set. As observations are made, the sensor fusion feature set is modified. Models, equations, and data are all recorded as datasets in the HamSci Public Database. The schedule model produces the commands to the receivers. Example commands include: Stations are ordered to report particular tranches of data from their local ring buffer storage. Stations will point a directional antenna in a particular direction at a particular time and upload the received data. Stations will change frequency or polarization. Stations will transmit a probe. Etc.

**September**

12. Design Review for the Space Weather System Data Plane specification. Output of the design review is a document that can be used by and individual or group to understand, participate with, or build their own Data Plane. Metadata protocol, database specification, command and control model, schedule model, and dataset definition are collected together, reviewed, and revised for clarity.
13. Begin policy development for marketing, sales, customer service, technical support, delivery, and returns (TAPR).
14. Begin policy development for sustaining engineering and data security/ governance (HamSci).

**October**

15. Publish version 1.0 of individual Space Weather Station Data Plane documentation.
16. Commence productizing for custom Space Weather Station hardware (BOM, layout, SDR architecture, RF, antenna, networking, software, user guide, technical documentation, software development manual, regulatory compliance documentation, accessibility and accommodations).
17. Design Review for marketing, sales, customer service, technical support, delivery, and returns policy. Outcome of review are published policies from TAPR. Design Review for sustaining engineering and data security policy. Outcome of review is sustaining engineering plan and a data security/governance policy from HamSci.
18. Identify candidates for manufacturing prototype build.

**November**

19. Manufacturing review for Bill of Materials. Outcome of the review is a decision on manufacturability and a cost estimate.

**December**

20. Design review for layout. Outcome of the review is prototype layout.
21. Decide who will prototype the various sensors and how many will be made.
22. Place order for prototypes.

**January - March**

23. Prototype build, test, delivery.
24. Verify and validate Space Weather Station specification.
25. Prototype Design Review. Outcome of this review is manufacturing approval for Friends build.
26. Identify candidates for manufacturing Friends build.

**April - July**

27. Friends build. Limited sales to Friends.
28. Friends deployment. System is operational. Performance measured.
29. Friends Design Review. Outcome of the review is gating item for mass manufacturing. Did it work?
30. Marketing, sales, customer service, technical support, delivery, returns, sustaining engineering commences.

**August**

31. Mass manufacturing.
32. Marketing, sales, customer service, technical support, delivery, returns, sustaining engineering continues.

## References

S. Sedkaoui & JL Monino. (2016). *Big Data, Open Data and Data Development.* Wiley-ISTE.

https://github.com/MITHaystack/digital_rf

https://github.com/gnuradio/SigMF

https://github.com/IonMetadataWorkingGroup/GNSS-Metadata-Standard

https://tapr.org/dcc

https://satnogs.org/

https://www.worldatwork.org/

https://www.ieee.org/