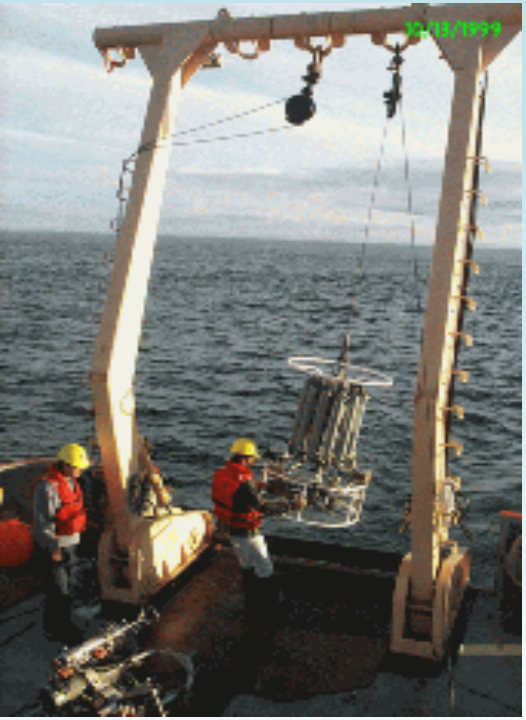
What do we do with observatory data? a user's perspective



Rich Pawlowicz*

Dept. of Earth, Ocean, and Atmospheric Sciences,

University of British Columbia

*Observatory supporter since 2001
Observatory user since 2009
Observatory advisor since 2011

UBC THE UNIVERSITY OF BRITISH COLUMBIA

ODL OCEAN DYNAMICS LABORATORY UNIVERSITY OF BRITISH COLUMBIA EARTH AND OCEAN SCIENCES

What do we do with observatory data?
A user's perspective is…**we fix it***.

*This is not primarily an Engineering Problem, although it involves engineering. I believe it is really a Sociological Problem – one we must solve.

Talk outline:

1. Observatories: Background and Context
2. Data and its problems!
3. How does science work?
4. How do they do this elsewhere?
5. Lessons

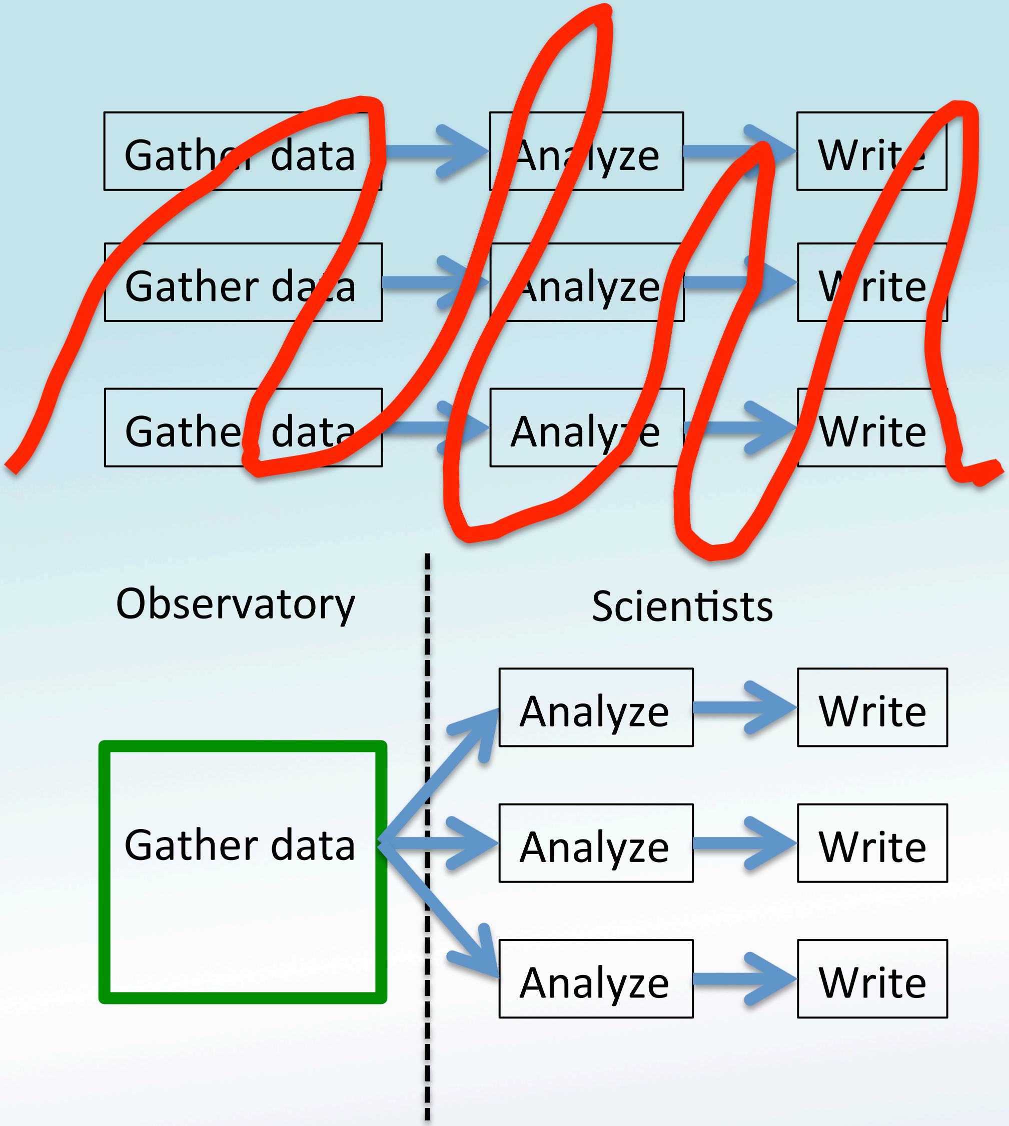
Background – the 1990s

- What was exciting? Internet (WWW anyway) was new(ish):
  - **What if** it went underwater to our instruments?
  - **What if** we weren’t limited by power and data storage limitations for our measurements?
  - **What if** we could get ocean data in real-time?
  - **What if** ANYONE could get the data online and see the “current state of the ocean”?

- In Canada at least, infrastructure was old, needed updating….and so the gov. decided to set up a new funding structure specifically for infrastructure.

NEW

A model for science



Gather data → Analyze → Write

Gather data → Analyze → Write

Gather data → Analyze → Write

Observatory

Scientists

Gather data

Analyze → Write

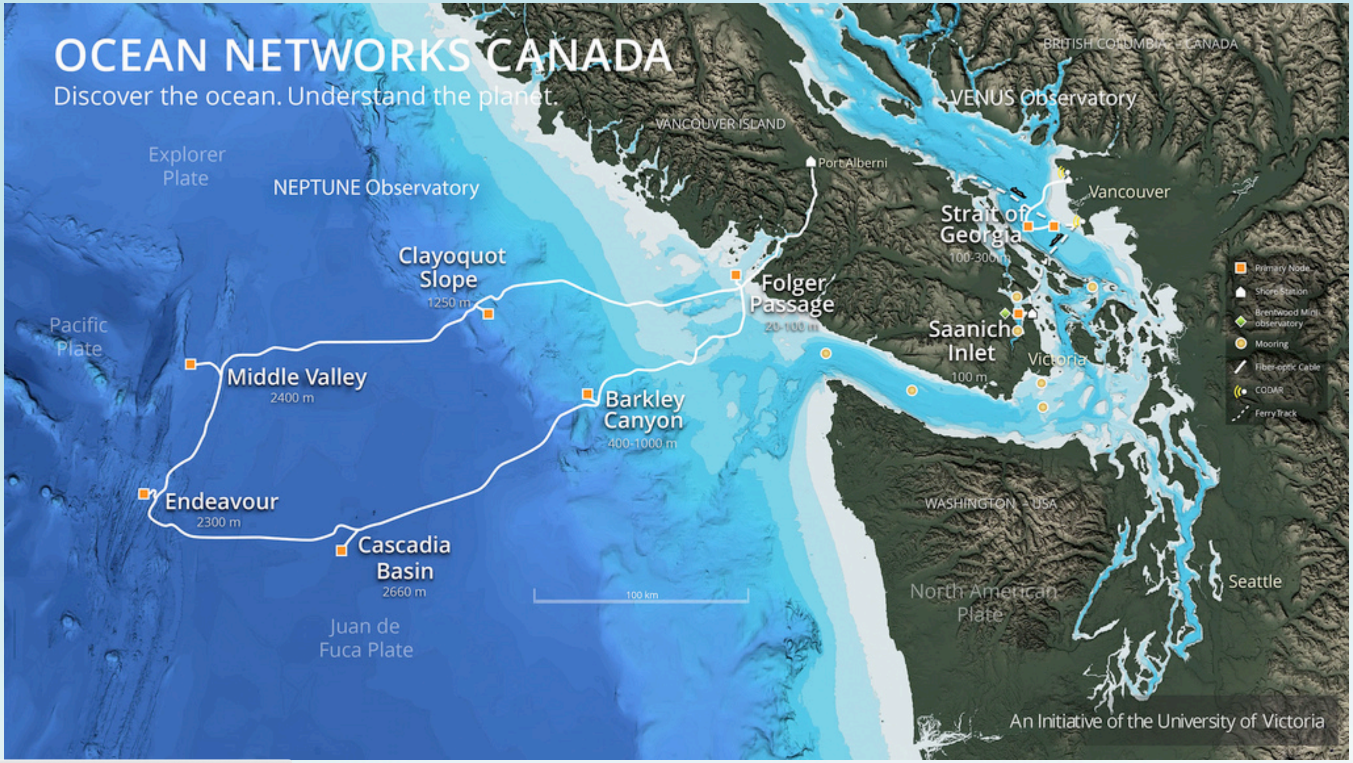
Analyze → Write

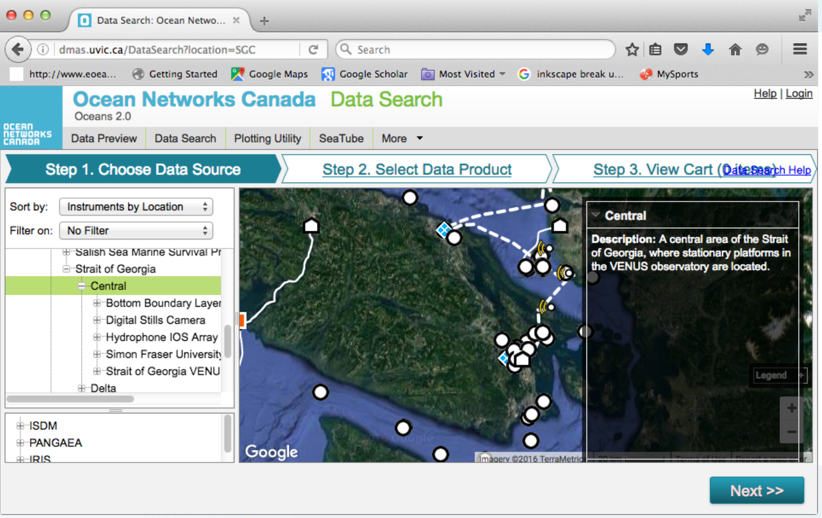
Analyze → Write

“Scientist” images courtesy of Times Higher Education

2016 - Internet under water!

- ≈ 400 instruments
- ≈ 5000 sensors
- ≈ 250 Gb/day of “data”




Data Search: Ocean Netwo...

dmas.uvic.ca/DataSearch?location=SGC

Ocean Networks Canada Data Search

Oceans 2.0

Step 1. Choose Data Source | Step 2. Select Data Product | Step 3. View Cart (Datasets) | Data Search Help

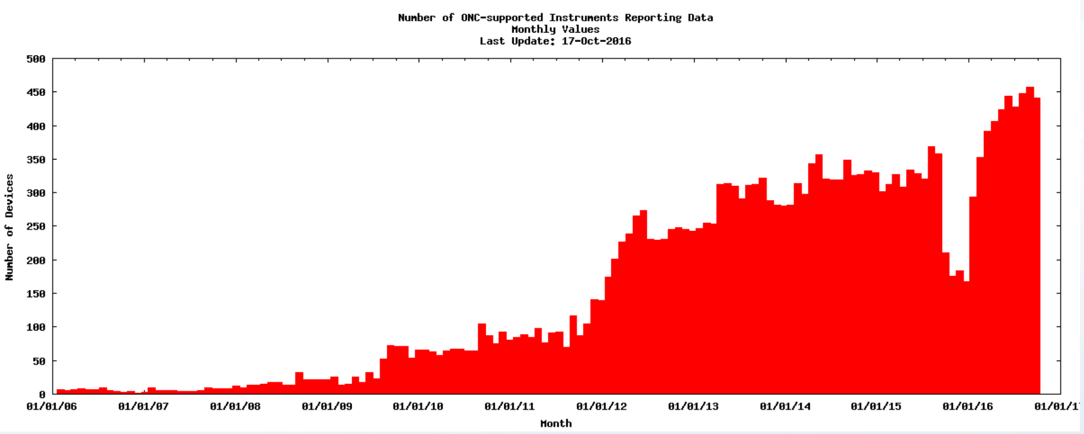
Sort by: Instruments by Location

Filter on: No Filter

- Salish Sea Marine Survival Pr
- Strait of Georgia
  - Central
    - Bottom Boundary Layer
    - Digital Stills Camera
    - Hydrophone IOS Array
    - Simon Fraser University
    - Strait of Georgia VENU
  - Delta
- ISDM
- PANGAEA
- IRIS

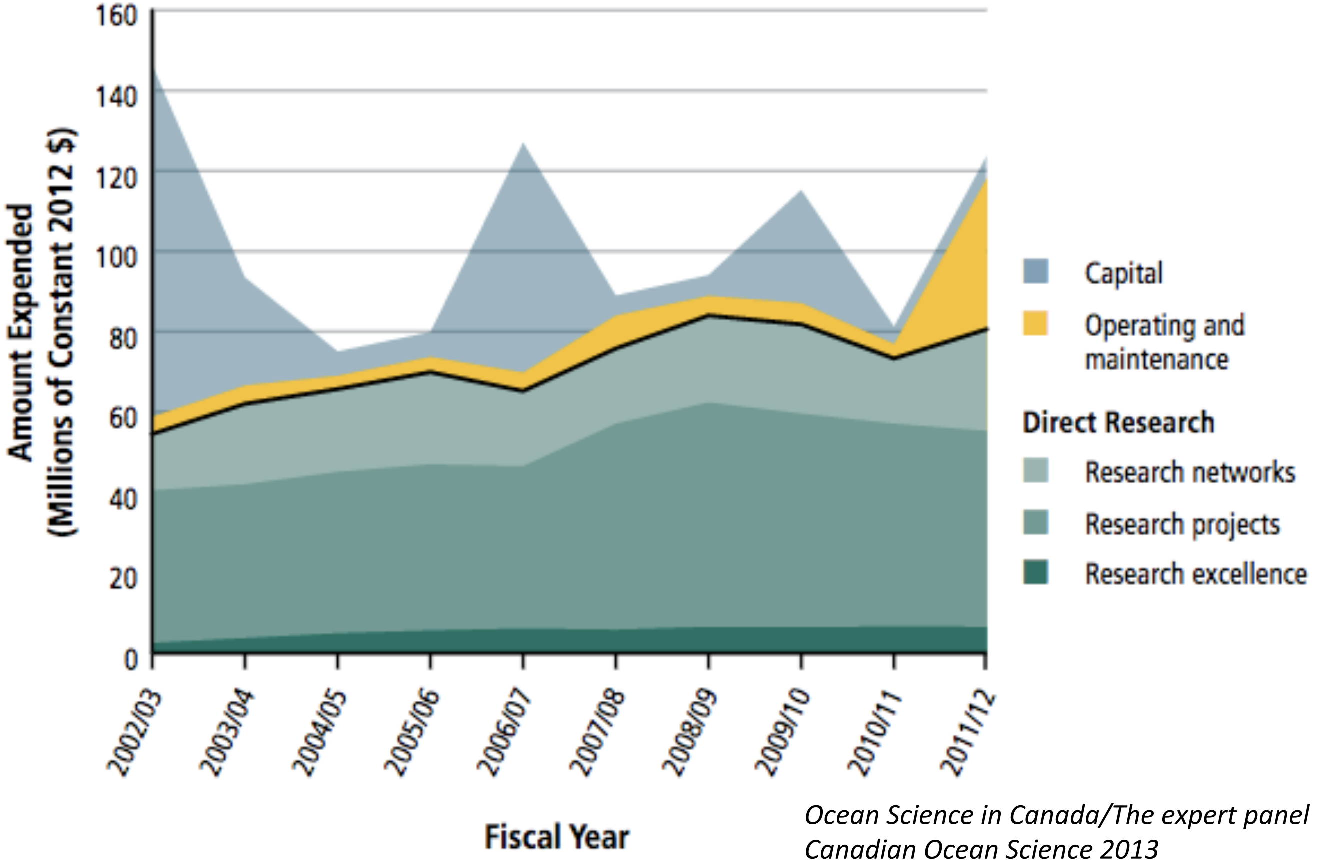
Central

Description: A central area of the Strait of Georgia, where stationary platforms in the VENUS observatory are located.


Number of ONC-supported Instruments Reporting Data
Monthly Values
Last Update: 17-Oct-2016

2006 2012 2017

Ocean Science Funding in Canada

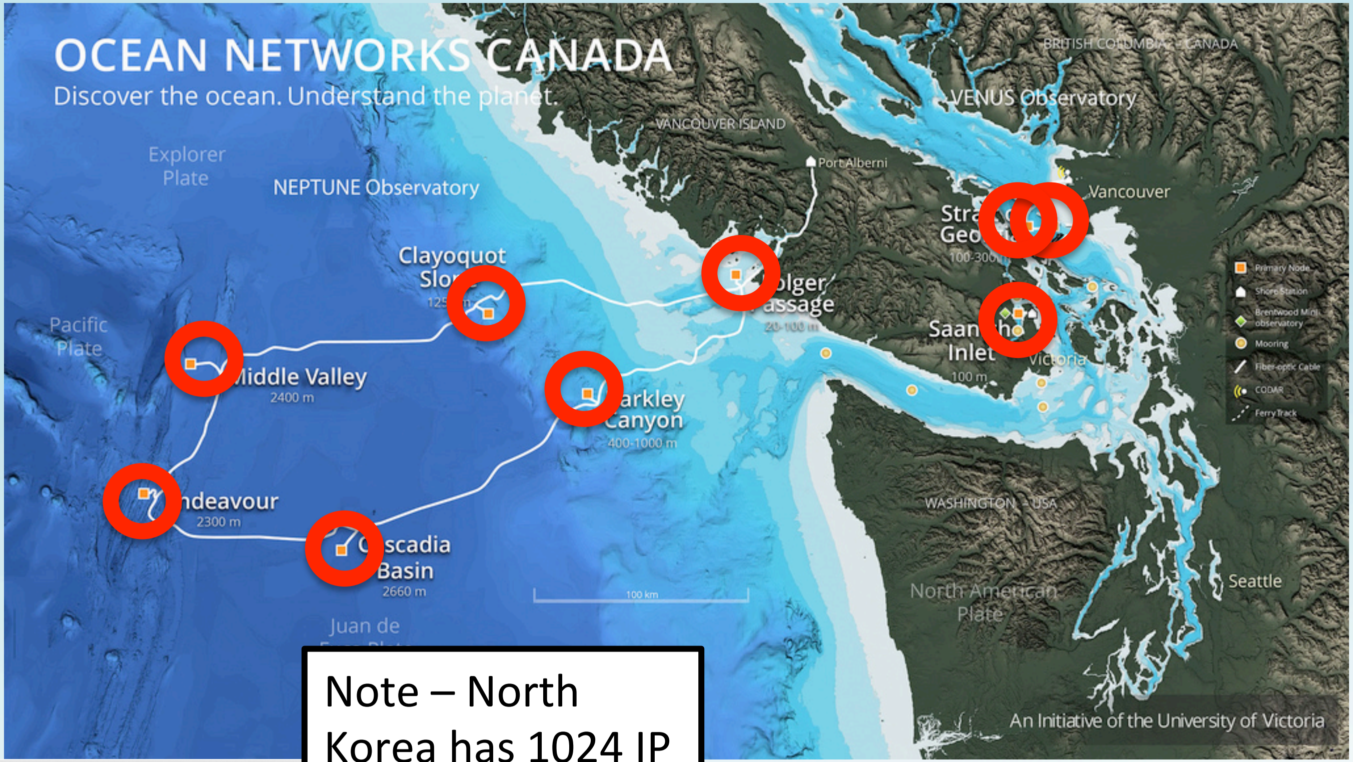


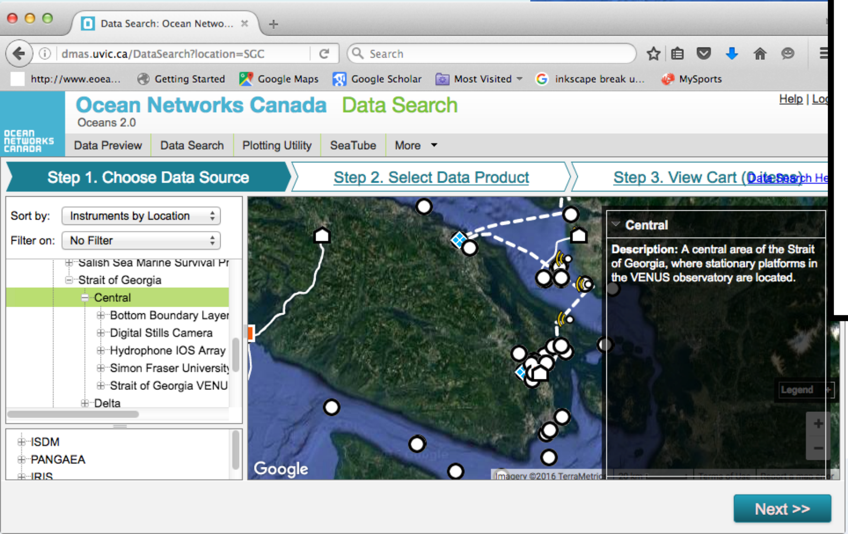
Ocean Science in Canada/The expert panel on Canadian Ocean Science 2013

- ONC Observatory Infrastructure:
  - Operating costs (all people/equip/ship) at ≈ $10-12M/year
- Comparison: a research icebreaker (CCGS Amundsen)
  - Shiptime costs ≈ $1.5M/month
- Comparison: a large-ish Earth Sciences Dept. (45 faculty, 50 staff, 200 grads, 900FTE undergrad):
  - Operating costs ≈ $10M/year, PLUS external research funding ≈$12M/year

2016 - Internet under water!

- 400 instruments
- 5000 sensors
- 250 Gb/day of “data”

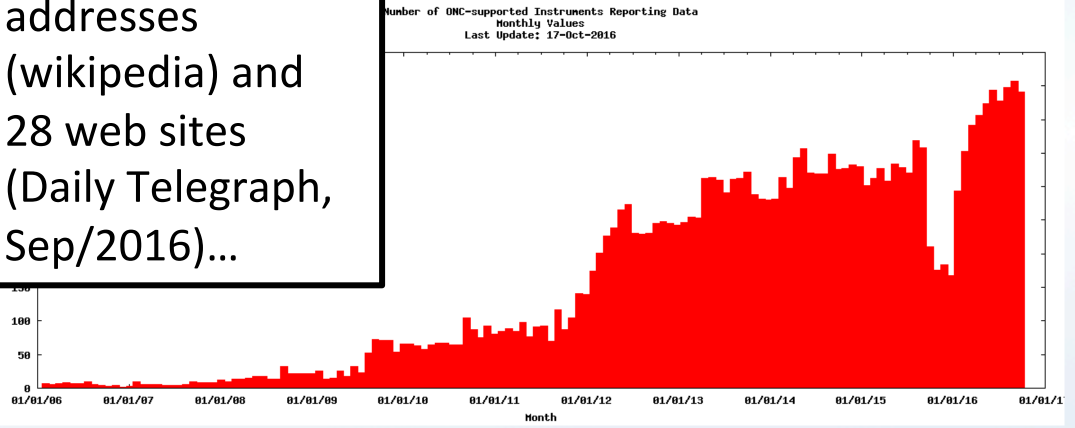




Central

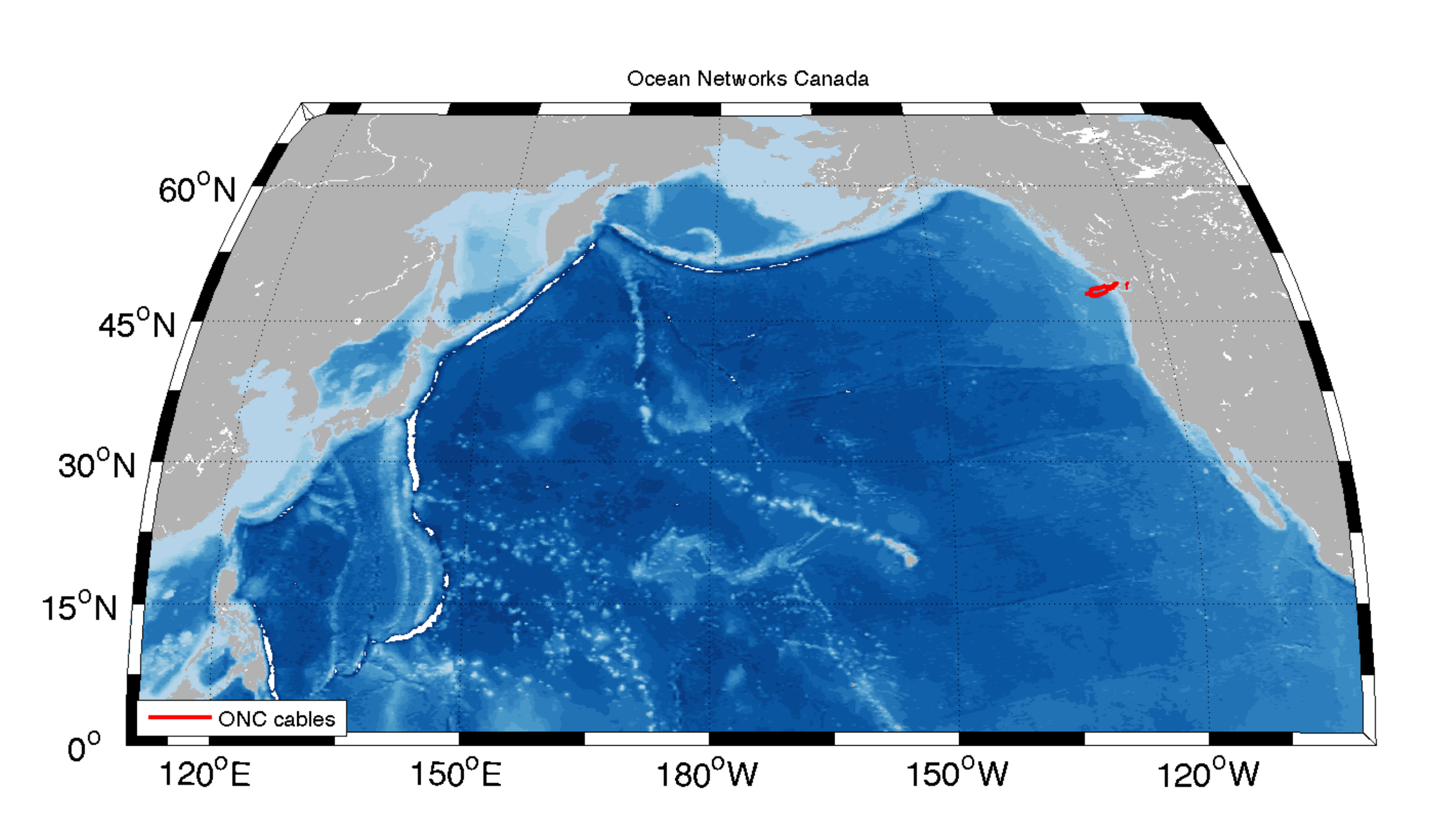
Description: A central area of the Strait of Georgia, where stationary platforms in the VENUS observatory are located.

Note – North Korea has 1024 IP addresses (wikipedia) and 28 web sites (Daily Telegraph, Sep/2016)…


Number of ONC-supported Instruments Reporting Data
Monthly Values
Last Update: 17-Oct-2016

2006 2012 2017

Wiring the ocean?

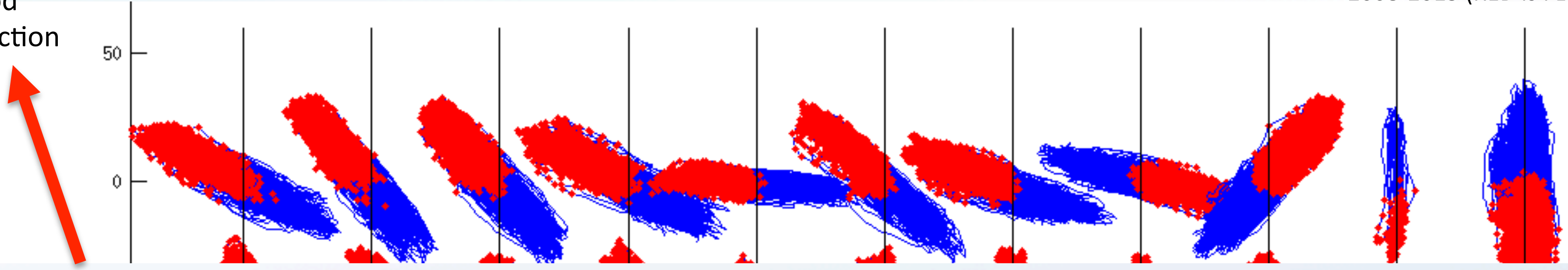


- OK, but not quite the "ocean"

  ….but now let us start on the science!

Case study 1: ADCP data

- Offshore node:
  - After 2 years of data gathering I discover:
    - Up and down are confused in archive processing. Currents are “mirrored” in a weird way (E is N, N is E) (only obvious when studying tidal ellipses in long-term harmonic analysis).
    - Fixed within a few months.
- Inshore nodes:
  - After 5 years of data gathering (22 separate deployments at 3 sites) I find that:
    - downloaded data can have deployment-dependent orientation errors of anywhere between -120 to +150 degrees!
    - Still not fixed after 8 years of data gathering.

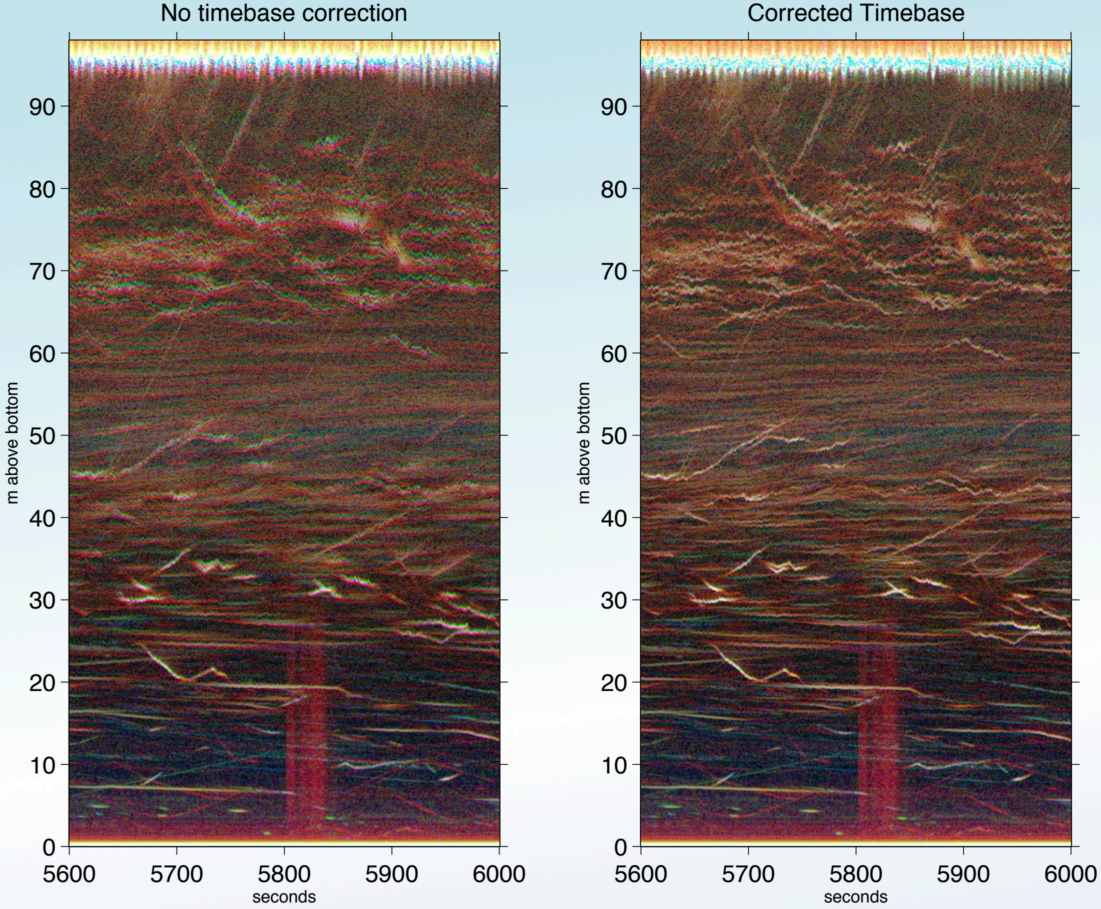


True
Flood
direction

Central Node ADCP:
Scattergram of depth-mean
flows for 11 deployments
2008-2015 (RED is FLOOD)

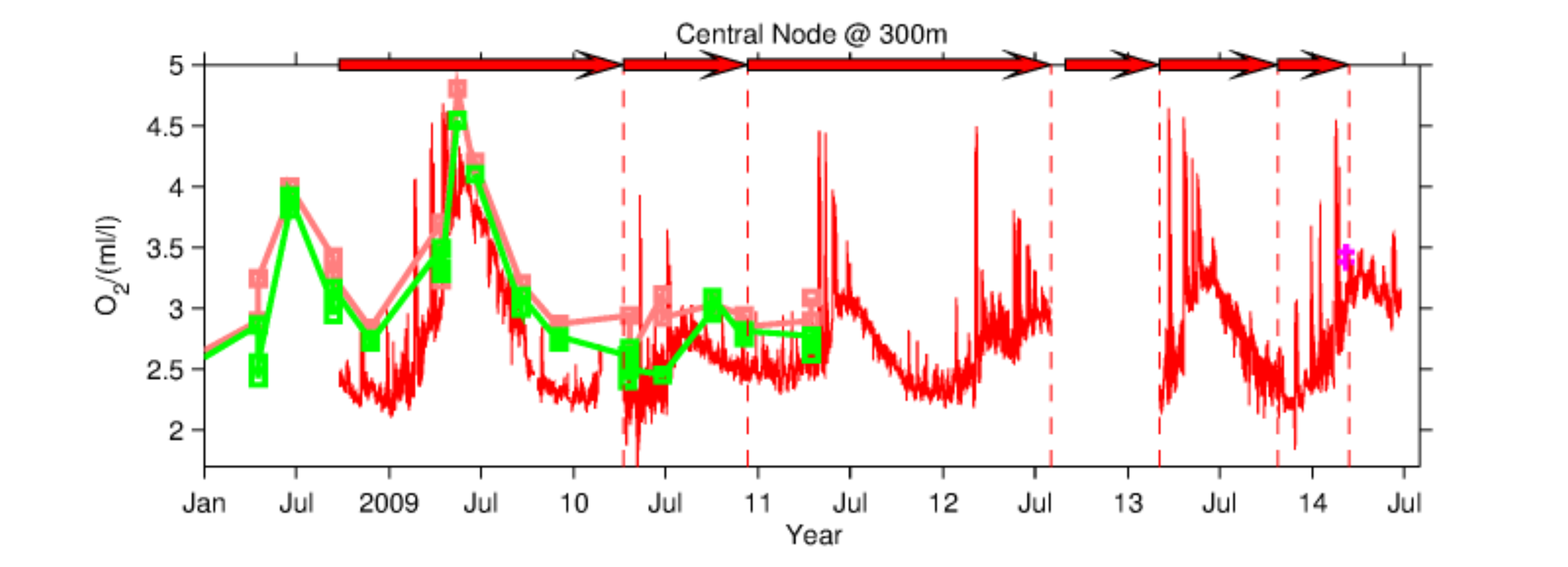
Case study 2: Scientific multi-frequency echo sounder

- After 1 year of data gathering
  - Firmware bug in data compression scheme means “ping” data for different frequencies appears unaligned by up to 5m in range…”sometimes” (only obvious when attempts made to carry out target strength calculations)
- After 3 years of data gathering
  - Firmware bug in quadrature demultiplexing sometime inadvertently “clips” surface return (noticed only after correlating apparent scattering strength against wind)



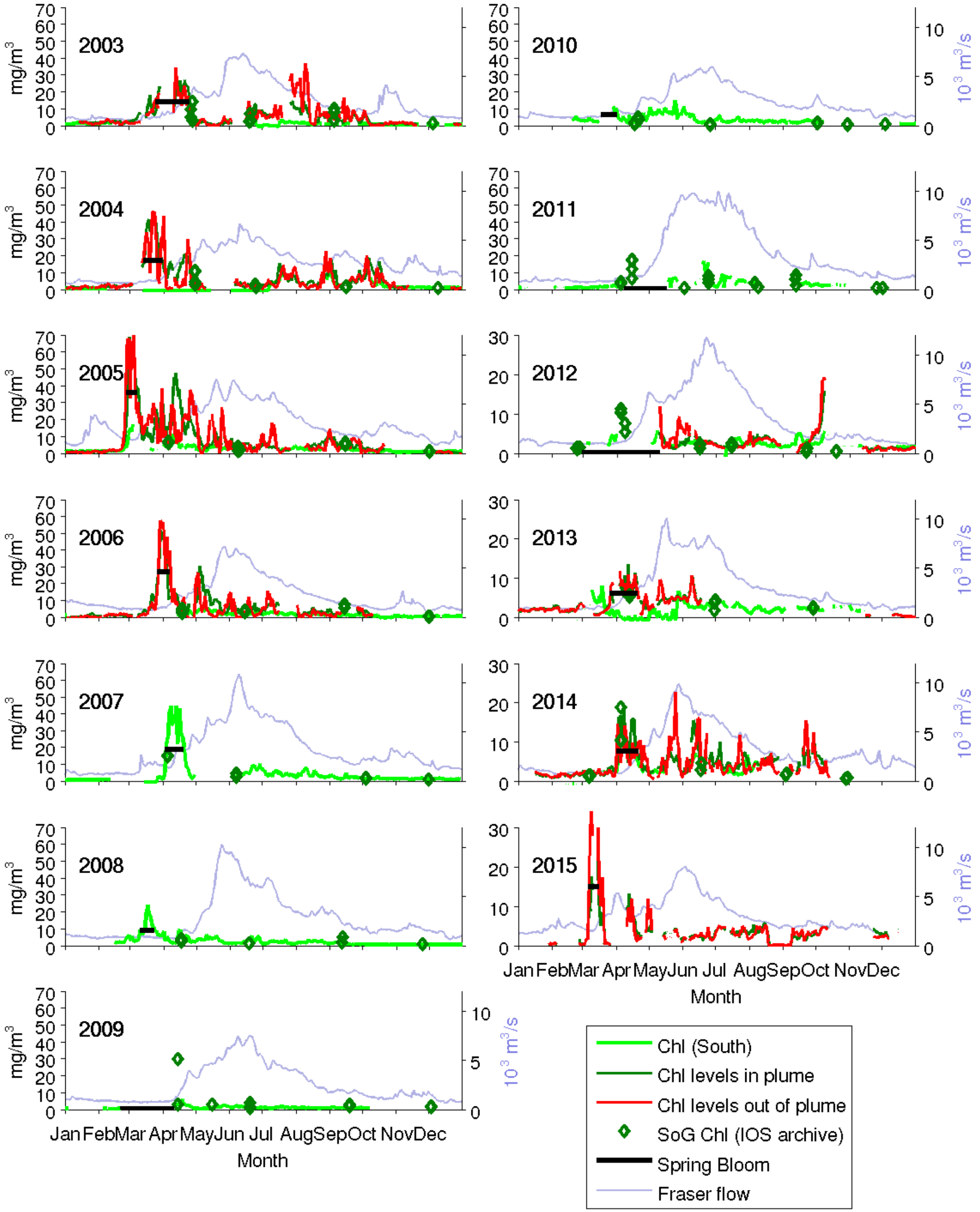
Case study 3: Bottom CTD/O_2

- After 5 years of data gathering
  - Calibration issues with T and S, especially in early deployments (identified as issue when looking at multi-year time series)
  - Vendor cal files are not stored in a systematic way
  - Optode O_2 data stored as mL/L although conversion has been made using inappropriate T and S (discovered after trying to understand exactly how data was gathered)
  - Optode data also has offset errors of up to 20% (discovered after comparing with other long-term O_2 datasets)
  - Calibrations have now been incorporated into maintenance procedures



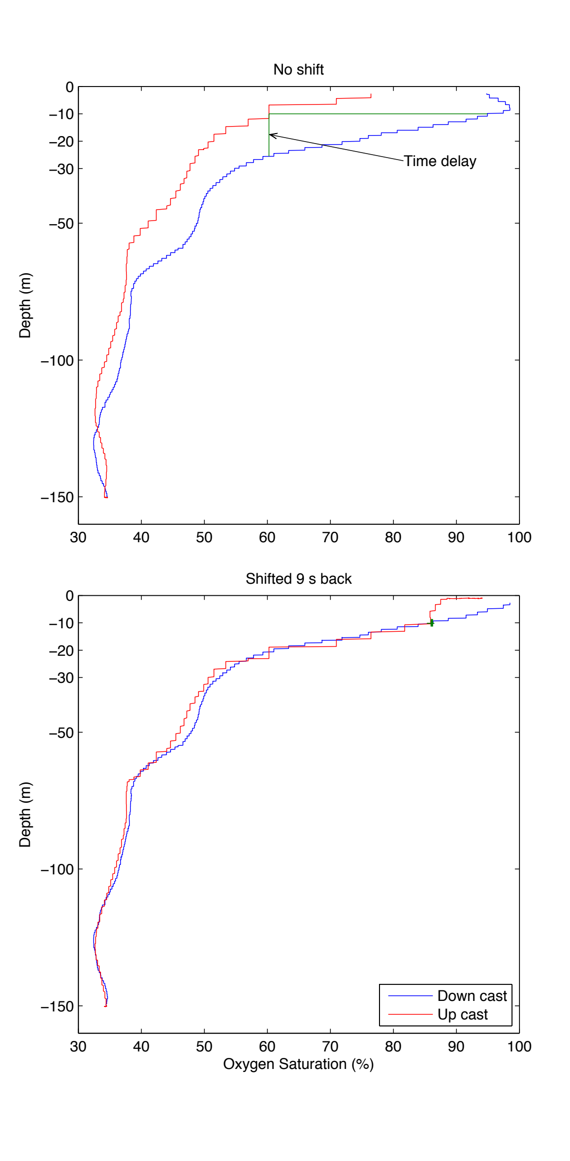
Case Study 4: Ferry monitoring systems

- After 2 years of data gathering
  - Chl and Turbidity time series transposed in data processing (suspicions raised after examining spatio-temporal variations in two year time series)
  - Shade problems with radiation sensors in met package
  - Compass correction for wind data added instead of subtracted (found during attempts to verify wind data against a shore station)
  - Fouling problems for optical sensors (require cleaning and a calibration procedure, not implemented until Y4



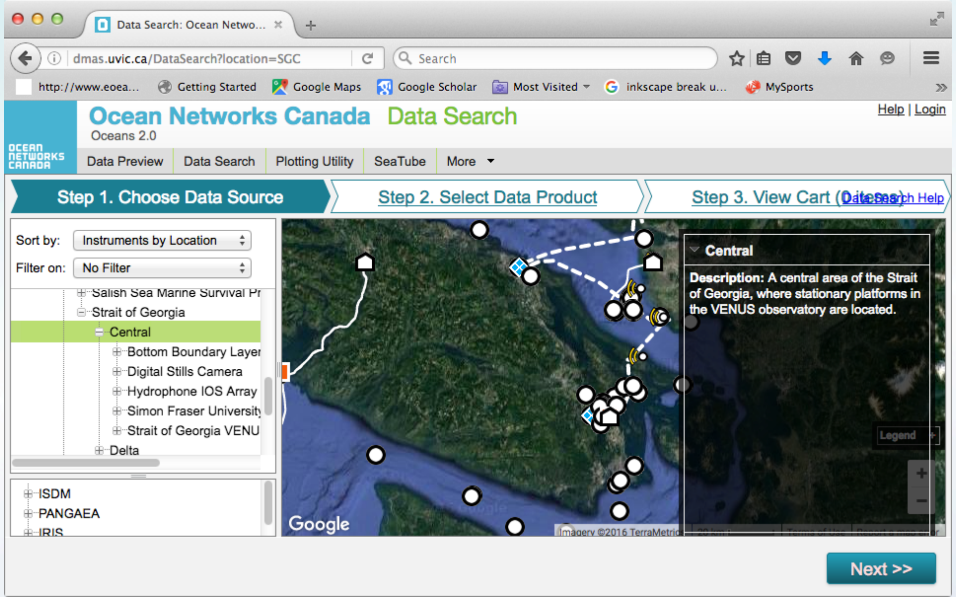
Case study 5: CTD/O_2/Fl Profiler dataset from a "Citizen Science" program

- After 18 months of data gathering
  - Deployment data files not correctly divided into casts
  - O_2 processing does not take into account 3 and 15 sec delays in sensor response
  - Sensor serial #-dependent offsets in Chl fluorescence profiles.



Case study 6: Data acquisition from archive – data for “anyone”?

- Various “click-ey” interfaces have been developed to allow “anyone” to access data.
  - These prove totally unsuitable for downloading anything more than a short period (one to a few days) of data.
  - And yet…almost anything I do requires analyzing many days of data just to make sure it is “right”
- ONC IT people develop “hack” Matlab scripts that can access data bypassing the interface (and possibly also ONC IT security!).
- That is, for all SERIOUS work data is either
  - bundled and provided by ONC staff, or
  - obtained from “bulk” downloads via the hack avoiding the “click-ey” interface.
- After 8 years of data gathering (14 years since funding began), SOME data is available online using an API that avoids the “click-ey” interface.



Sort by: Instruments by Location

Filter on: No Filter

Central

Description: A central area of the Strait of Georgia, where stationary platforms in the VENUS observatory are located.

On the other hand….
Examples of datasets I use that don't have problems (or at least not unusual problems)

- River flow datasets (Water Survey of Canada)
- Archived hydrographic profile data (NODC, CCHDO, OSD/MEDS, IOS data archive)
- Weather Buoy data (NDBC, OSD/MEDS)
- Fraser River Water Quality buoy (Environment Canada)
- Satellite imagery (NASA)
- River chemistry datasets (USGS, other)
- CODAR dataset (ONC, although we have suggested about a dozen minor "tweaks" to improve data quality and utility)

- Why the difference?
  - – most of this data has been collected routinely for a long time. Many of the bugs have already been worked out.

Summary and Conclusions?

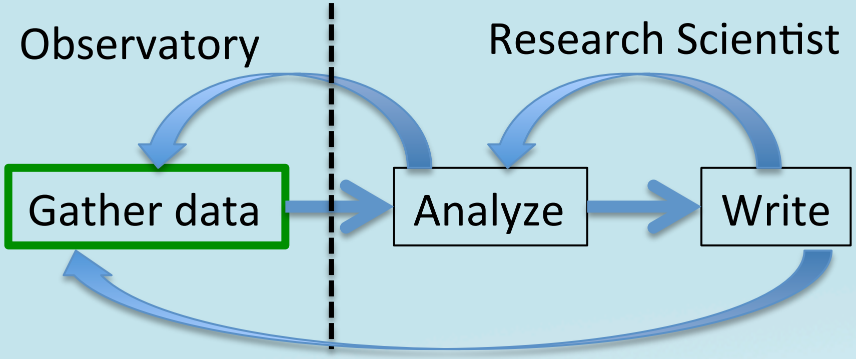
- Being able to STORE data without limitation does NOT mean that data size is no longer a problem.
  - I conclude that, to properly deal with large amounts of complex data, observatory "people" must be involved in a non-scalable way.

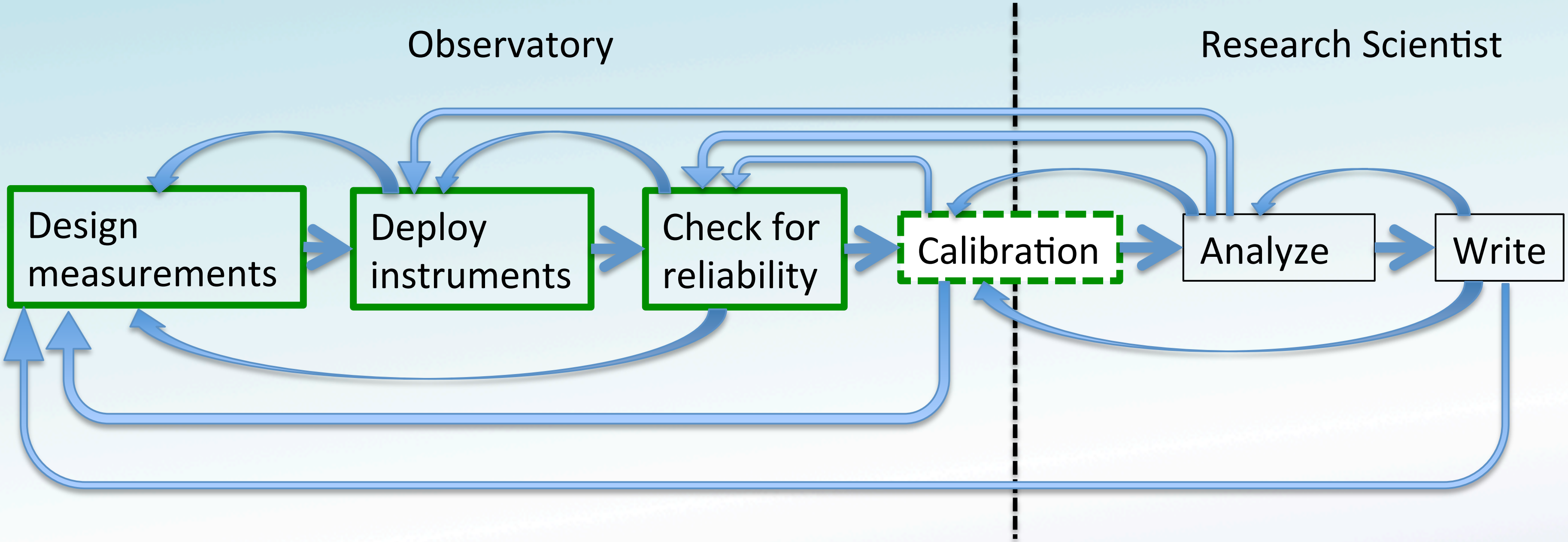
- Basically, **every** observatory dataset examined (usually after a few years of data gathering, so presumably after it had been proved "good" internally) had **severe** problems of some sort – enough to make any interpretation **extremely** suspect.
  - I conclude that significant data errors are **widespread** and likely UNAVOIDABLE in any complex system.

- These problems are often highly technical, and mostly not "findable" by "anyone", or (often) even by observatory staff, but are only found in the course of quantitative analysis.
  - I conclude that research scientists must be CLOSELY involved with data streams.

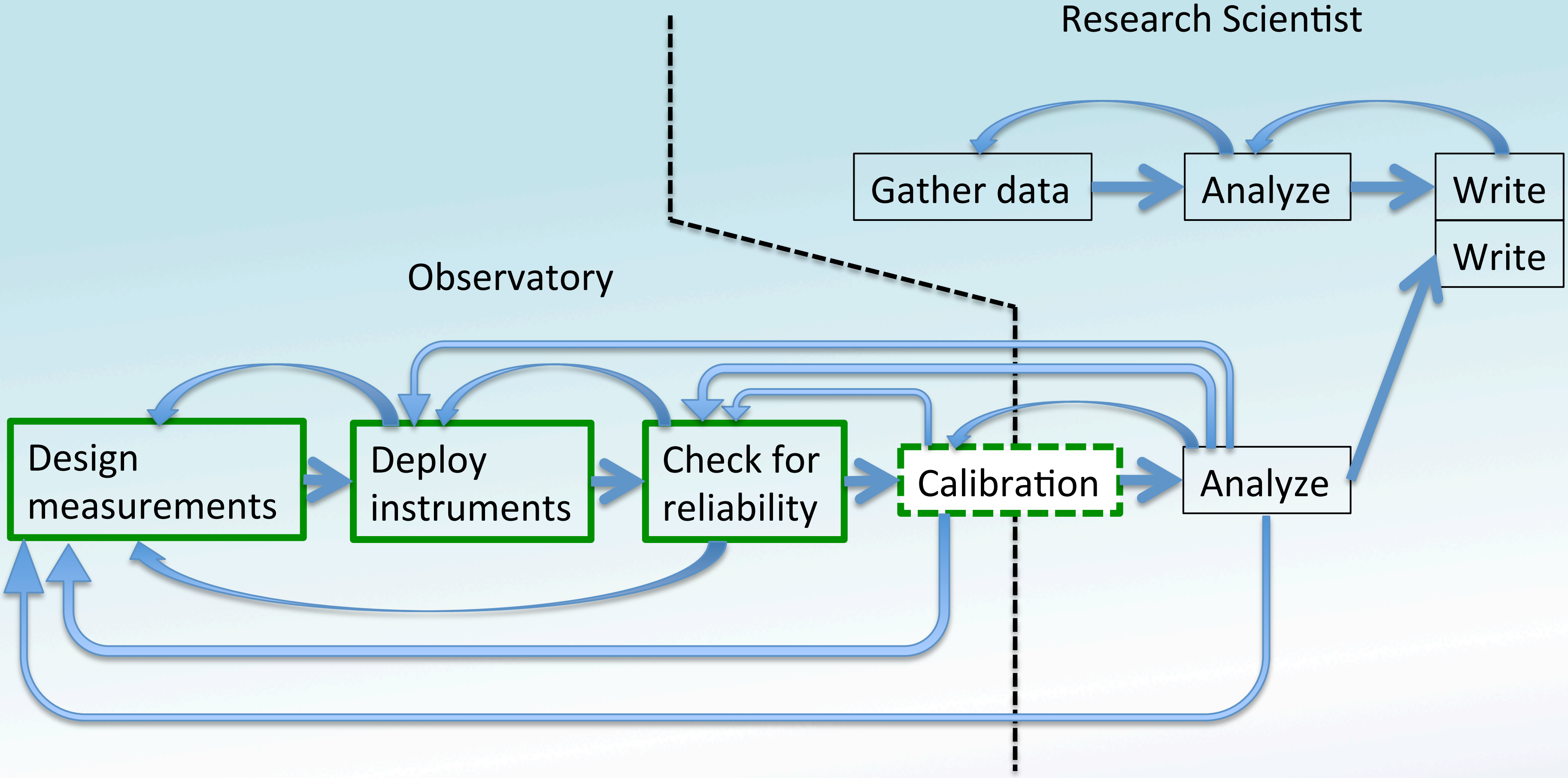
- Many of these problems are **still** in the archived datasets.
  - I conclude that there are problems with the idea of "data for anyone", and with ideas to widely distribute data in general.

A model for science – lets think about this some more…





But from the scientist point of view…(a cautionary)



Research Scientist

Gather data

Analyze

Write

Write

Observatory

Design measurements

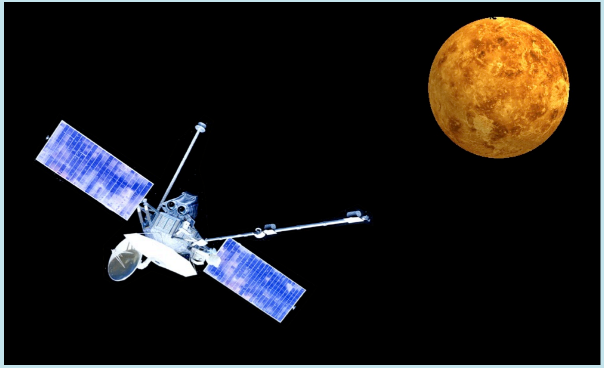
Deploy instruments

Check for reliability

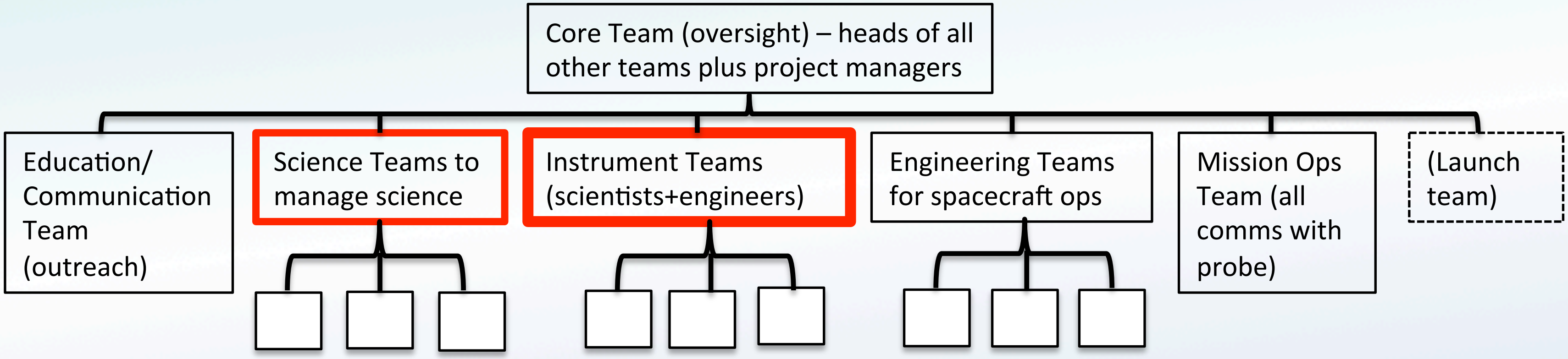
Calibration

Analyze

What about other communities (1): Is an ocean observatory like a space probe?



- Yes
  - Complex engineering problem.
    - This includes instrument design, and
    - Instrument deployment
    - Instrument configuration
  - Many cooperating science groups.
    - Not co-located; spread out over national and international boundaries
  - Potential for new results from new (more/better/different) measurements.
  - Long lead time for results.
- No
  - Many missions are carried out by a single (large) proposal that funds BOTH engineering AND science (multi-year, with mid-term reviews; later 'add-ons' are also possible).
  - Major science players are usually also heavily involved in engineering (instrument development).
  - Probes are the ONLY way to address many planetary science problems.
  - Data is not for "anyone" (*although a condition of funding is that mission data is publically available in 3 months)
  - Complex hierarchy of responsibility on both science and engineering sides – and LOTS of meetings by "Instrument Teams" involving both scientists and engineers, "Science Teams" for science, coordination meetings, etc.



Core Team (oversight) – heads of all other teams plus project managers

Education/ Communication Team (outreach)

Science Teams to manage science

Instrument Teams (scientists+engineers)

Engineering Teams for spacecraft ops

Mission Ops Team (all comms with probe)

(Launch team)

What about other communities (2): “Startups in 13 sentences”

Essay by Paul Graham, Y-combinator

1. Pick good cofounders
2. Launch fast

3. Let your ideas evolve

4. Understand your users
5. Better to make a few users love you than a lot ambivalent
6. Offer surprisingly good customer service

7. You make what you measure

8. Spend little
9. Get ramen profitable

10. Avoid distractions
11. Don’t get demoralized
12. Don’t give up

13. Deals fall through

So, what lessons are to be learned?

- "Good" data only comes from concentrated scientific attention – engineers and data specialists aren't enough.
- Scientists must be deeply involved to get that concentrated attention – communication must go both ways.
- Because this cooperation is complex, incentives are necessary to involve them - $$ is one, must think of others. Remember that USERS HAVE OTHER OPTIONS!
- "You make what you measure"!

My thanks to Ocean Networks Canada, for a) their wonderful cooperation over the years, and specifically for b) funding this talk