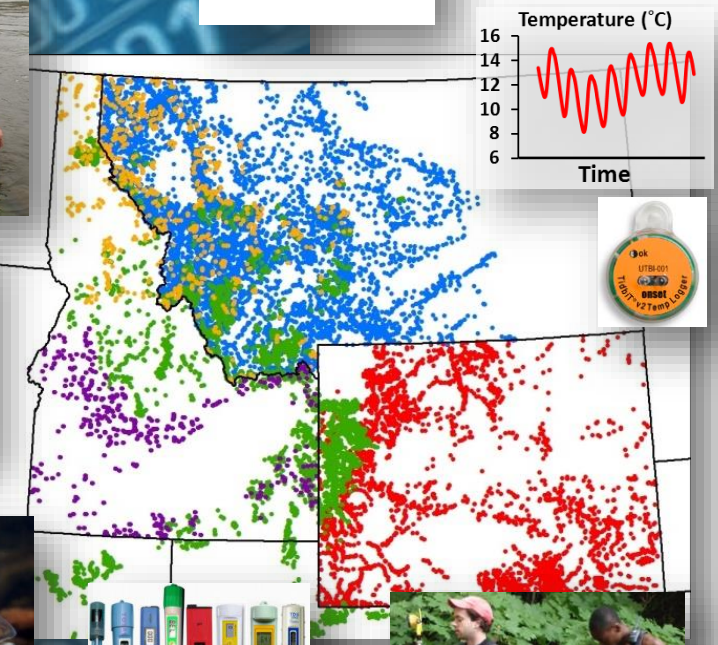
BIG DATA Gets Wet: How Stream Geostatistics, eDNA, and Crowd-Sourcing Massive Datasets are Revolutionizing Aquatic Science

Dan Isaak, Mike Young, Erin Peterson,
Jay Ver Hoef, Seth Wenger, Dave Nagel



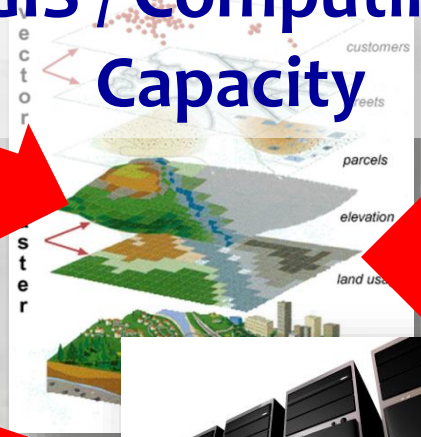
BIG DATA = BIG POSSIBILITIES

A Revolution is Happening

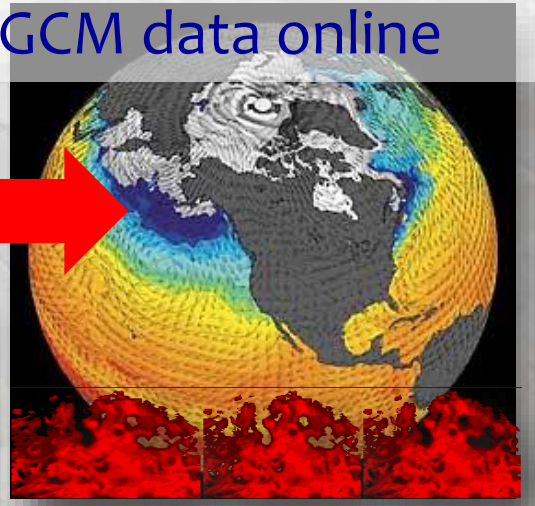
Geospatial Technologies & Computing Horsepower

Remote Sensing

GIS / Computing Capacity

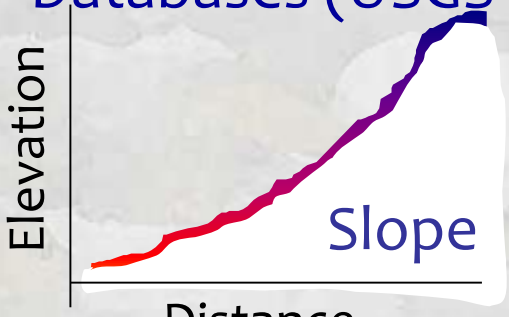


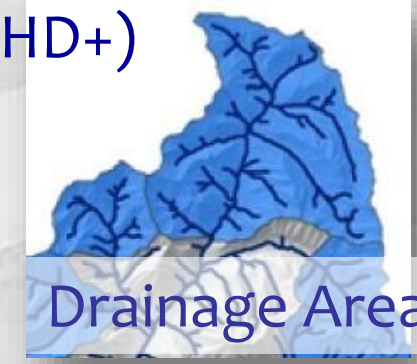
Climate, weather, GCM data online



Visualization

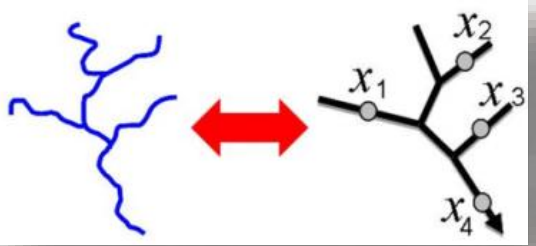
Nationally Consistent Hydrology Databases (USGS NHD+)

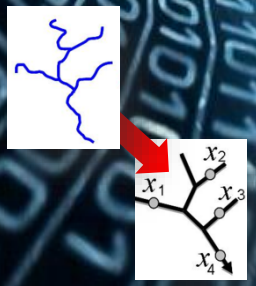




Drainage Area

Spatial analyses

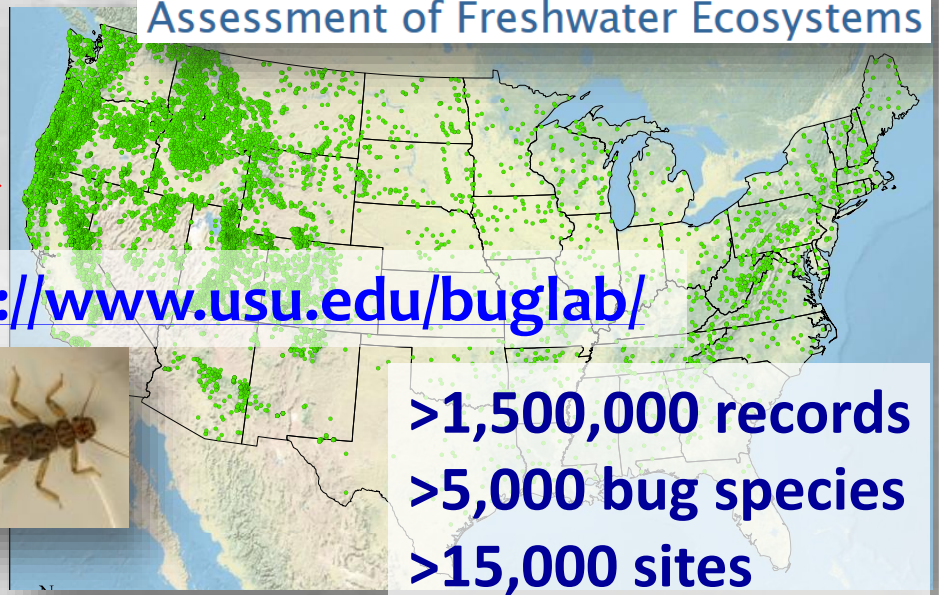




Mountains of Aquatic Data Already Exist

Databases of biological measurements

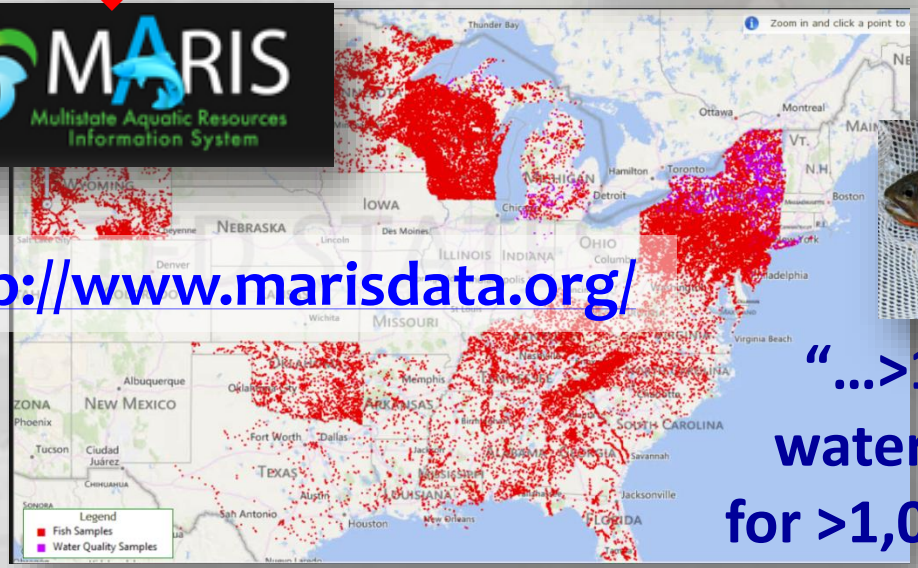
Western Center for Monitoring & Assessment of Freshwater Ecosystems



http://www.usu.edu/buglab/

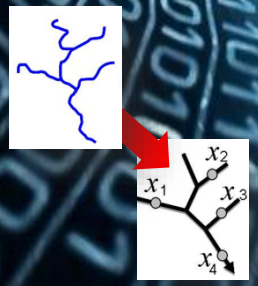
>1,500,000 records
>5,000 bug species
>15,000 sites

MARIS
Multistate Aquatic Resources Information System

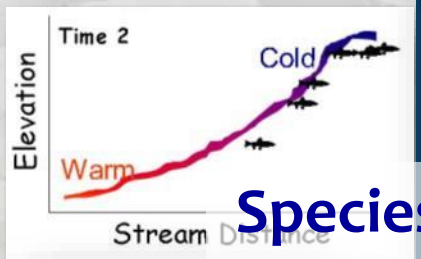


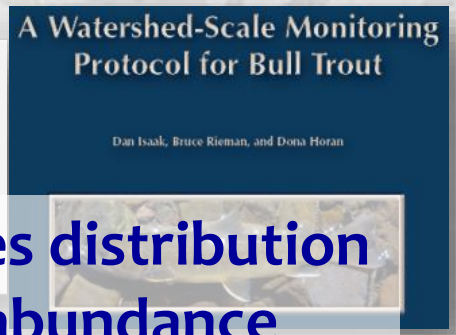
http://www.marisdata.org/

"...>1,000,000 fish & water quality records for >1,000 fish species"



Standard Protocols & Inexpensive Technology

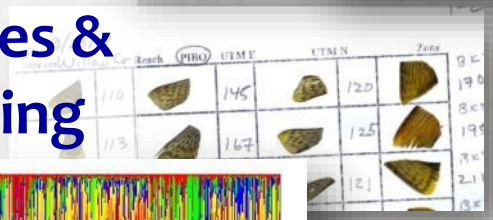


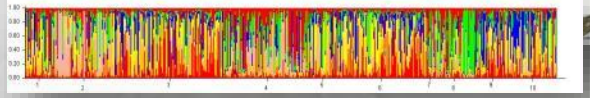


Species distribution & abundance

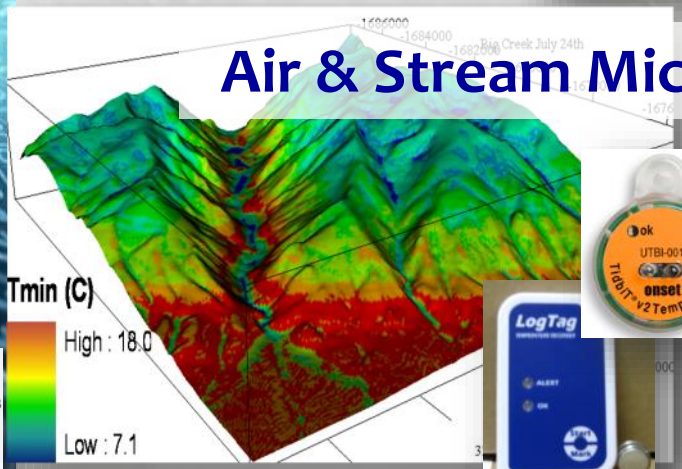
Stream discharge

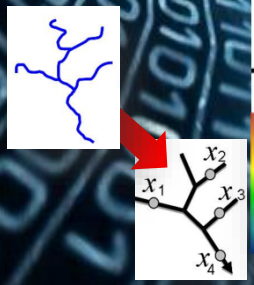
Tissue Samples & DNA barcoding





Air & Stream Microclimates



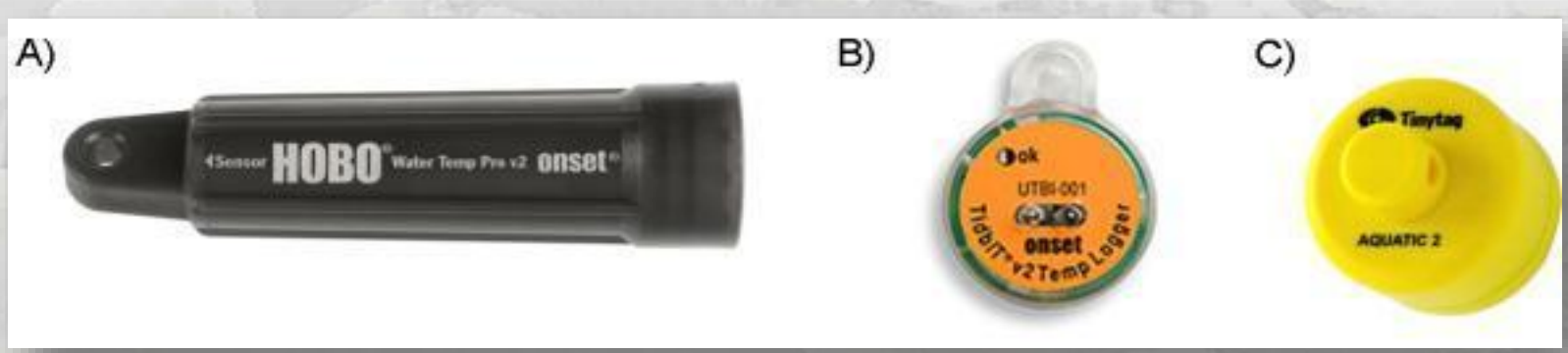


Short communication

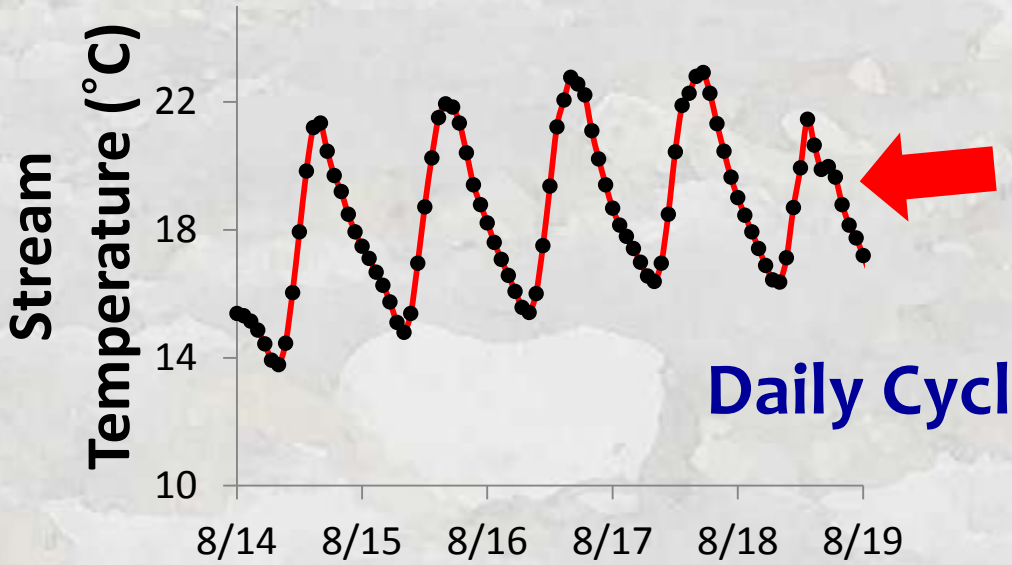
Design and evaluation of an inexpensive radiation shield for monitoring surface air temperatures

Zachary A. Holden[a,*], Anna E. Klene[b], Robert F. Keefe[c], Gretchen G. Moisen[d]

Miniature Digital Sensors Make Temperature Data Collection Easy…

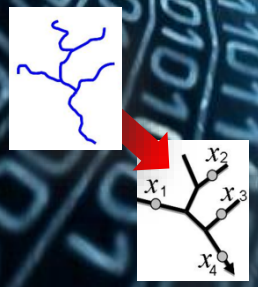


$100 = 5 years of data



Many measurements each day (12 – 48)

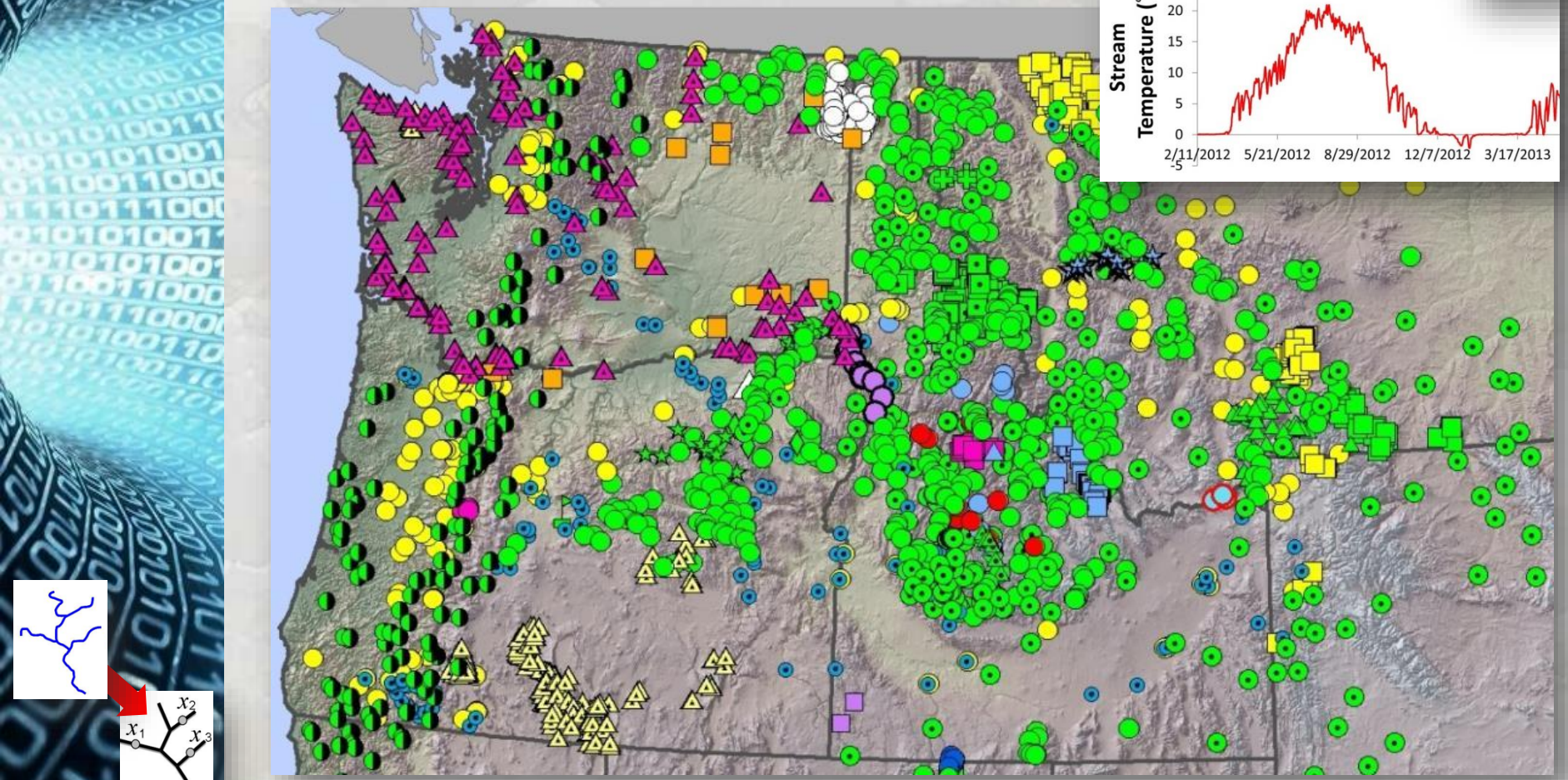
Daily Cycles



Rates of Data Acquisition are Accelerating

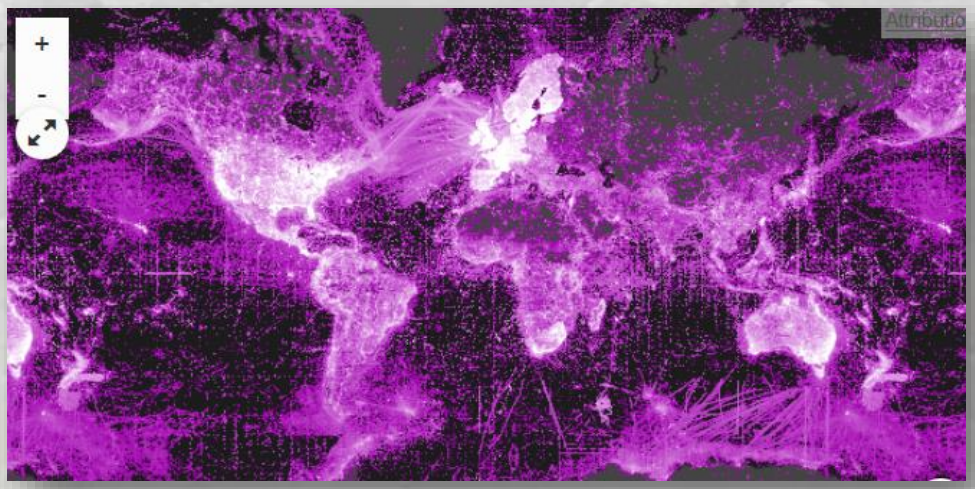
~4,000 annual temperature monitoring sites

35,000,000 hourly records annually!

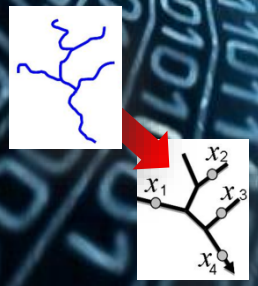


Rates of Data Acquisition are Accelerating

eDNA puts occurrence sampling on steroids

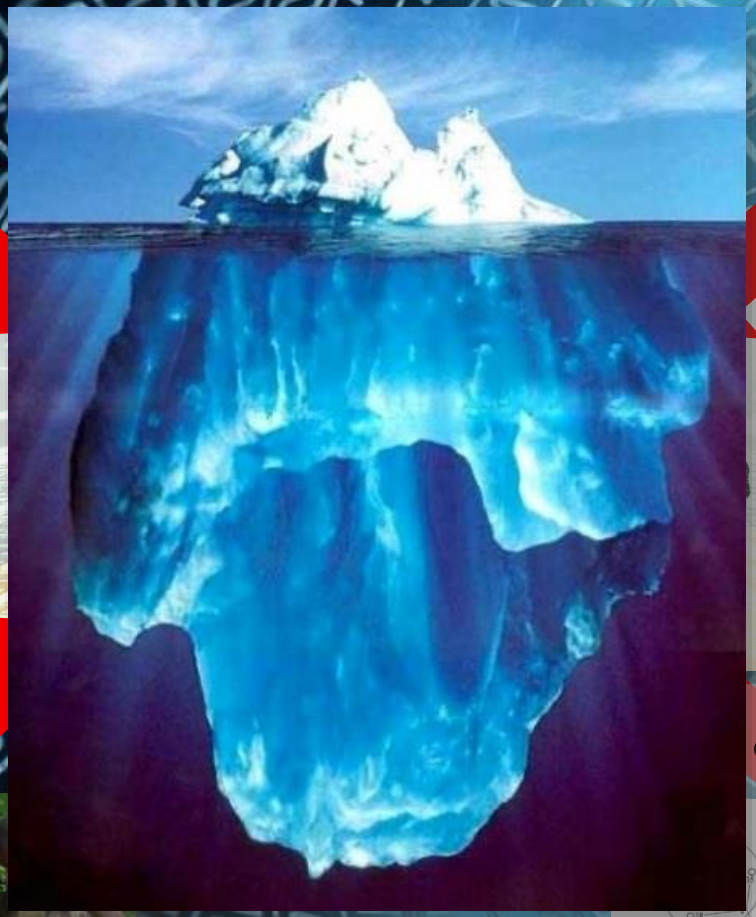


GBIF Database >600,000,000 species occurrence records

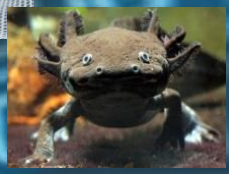


We're Being Buried Alive

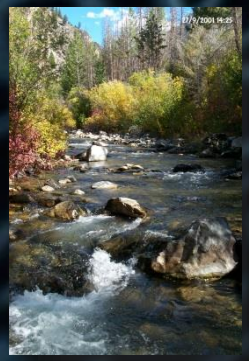
- Water Quality



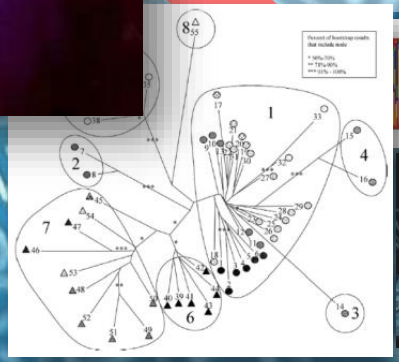
- Distribution & Abundance

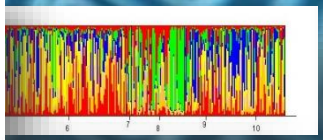


- Habitat Condition



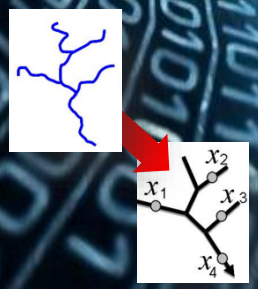
- Genetic Attributes





Data Needs to be Organized & Accessible to be Useful

Data In → Information Out



A BIG DATA Crowd-Sourcing Example with Stream Temperature Data

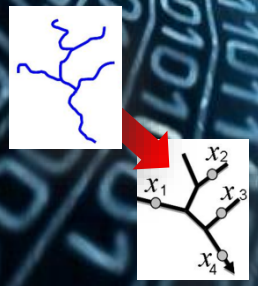
Steps in the Database Creation Process:

1) Database team cleans/organizes data into an Oracle database
   a. temperature sites linked to NHDPlus stream reaches & unique COMID field
   b. temperature data are passed through a cleaning macro so that anomalous records are flagged
   c. database team contacts data providers to resolve discrepancies in a or b

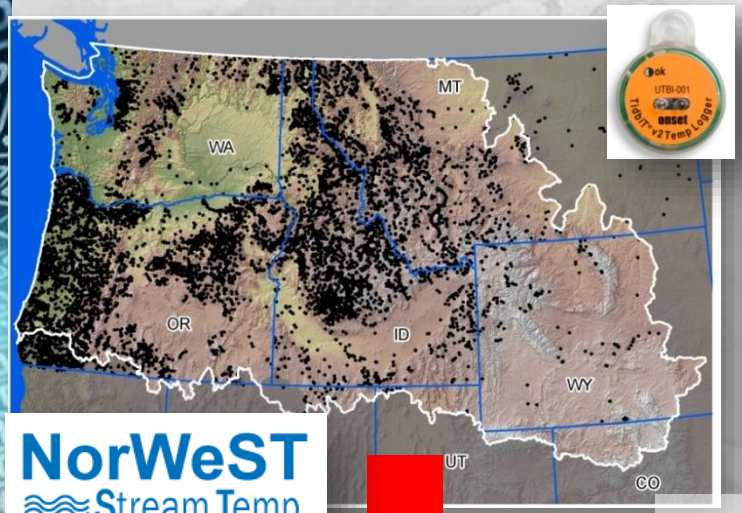
2) Summary metrics calculated (daily min/max/mean & many others) using custom scripts

3) Meta-data describing procedures are developed & linked to data

4) Data are packaged in user-friendly digital file formats & posted to website for distribution

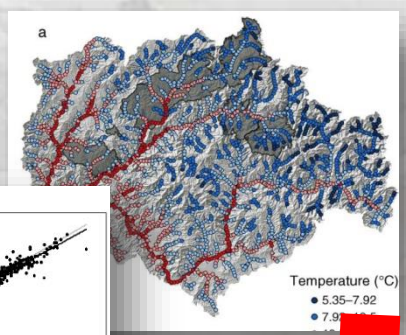


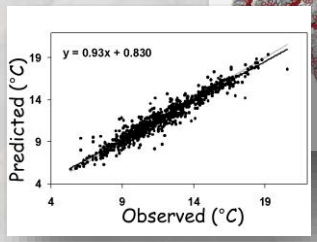
Regional Temperature Model



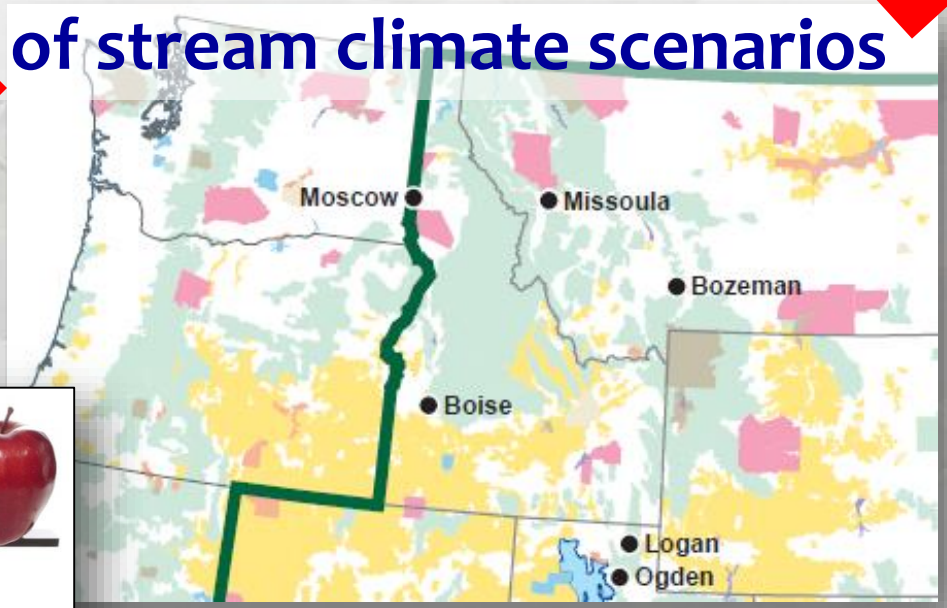
NorWeST
Stream Temp

Accurate stream temp model

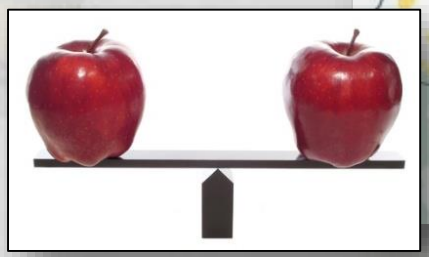


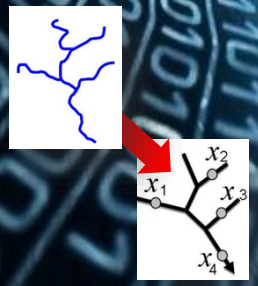


Cross-jurisdictional "maps" of stream climate scenarios



Consistent datum for strategic planning across all streams



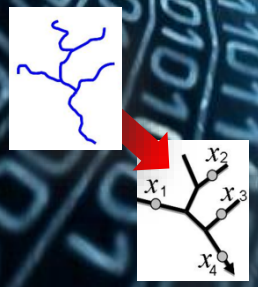


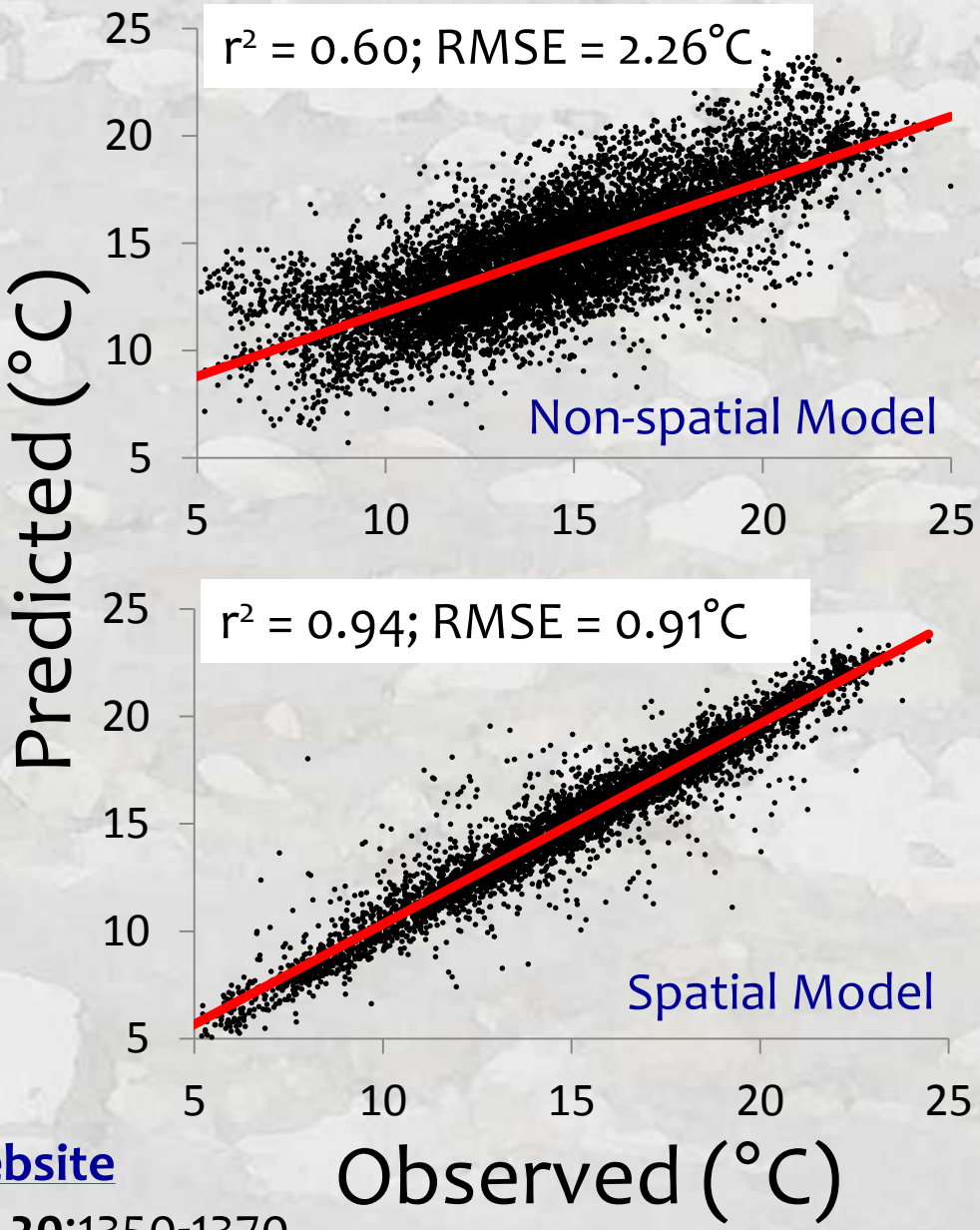
Accurate GeoStatistical Stream Models

Covariate Predictors

1. Elevation (m)
2. Canopy (%)
3. Stream slope (%)
4. Ave Precipitation (mm)
5. Latitude (km)
6. Lakes upstream (%)
7. Baseflow Index
8. Watershed size (km^2)
9. Glacier (%)

10. Discharge (m^3/s)
    USGS gage data
11. Air Temperature (˚C)
    RegCM3 NCEP reanalysis
    Hostetler et al. 2011

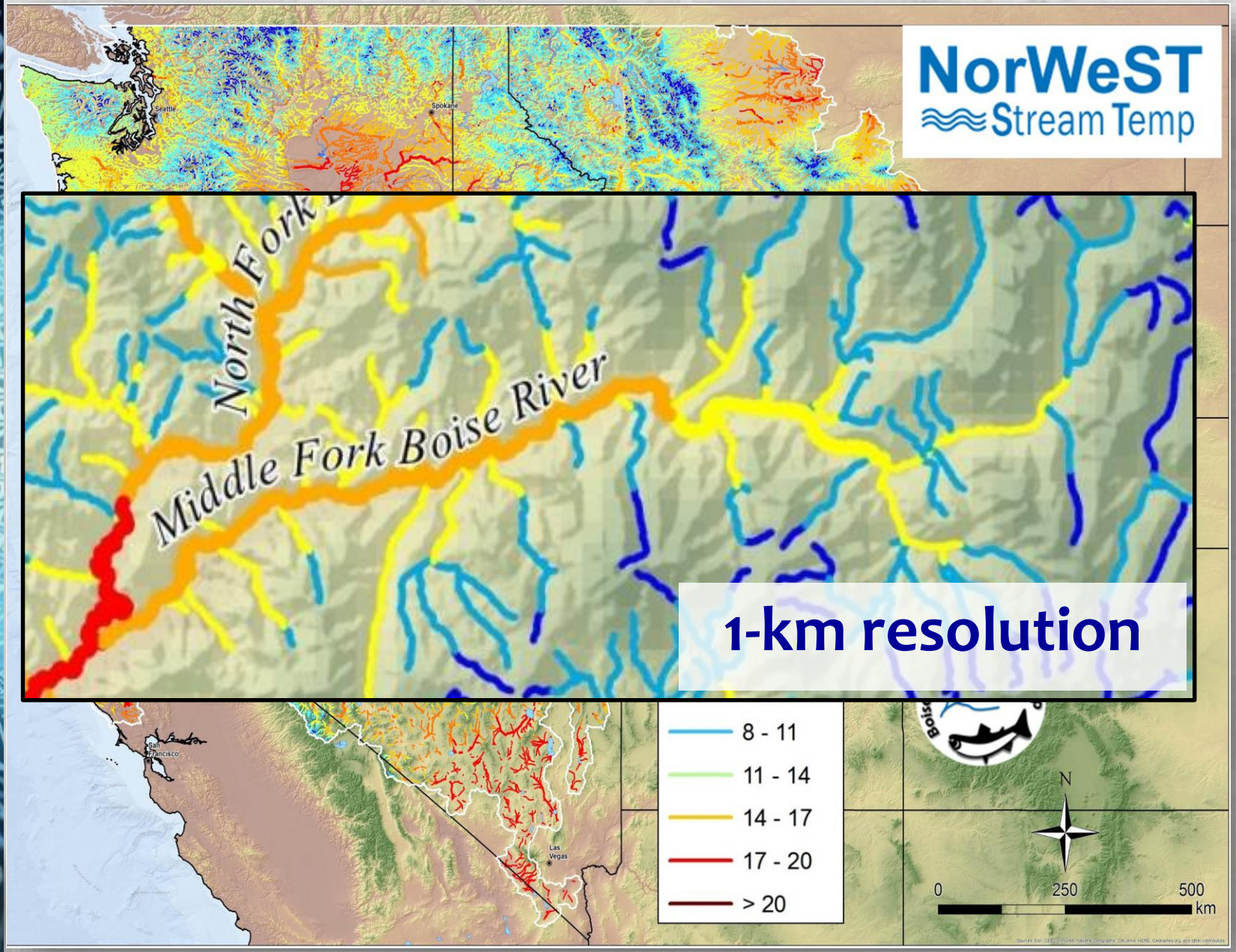


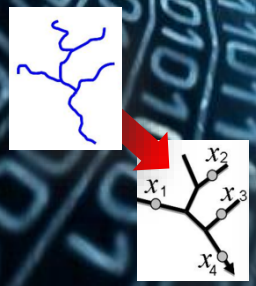


More details: **NorWeST website**

Isaak et al. 2010. *Ecol. Apps* **20**:1350-1370.

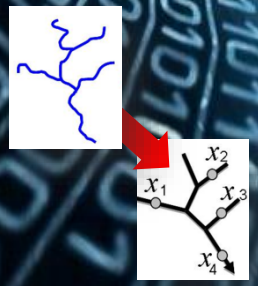
Model Enables Accurate Prediction Maps





30 NorWeST Climate Scenarios

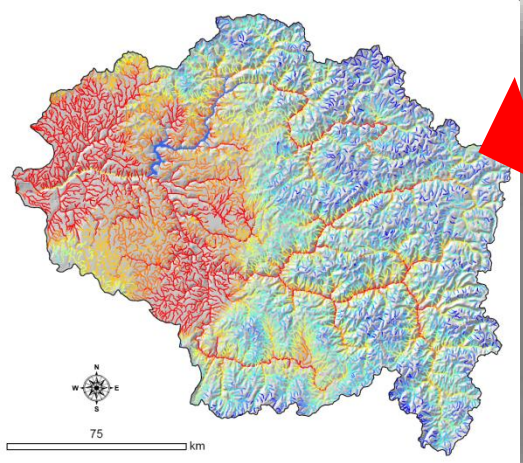
Scenario	Description
S1_93_11	Historical scenario representing 19 year average August mean stream temperatures for 1993-2011
S2_02_11	Historical scenario representing 10 year average August mean stream temperatures for 2002-2011
S3_1993	Historical scenario representing August mean stream temperatures for 1993
S4_1994	Historical scenario representing August mean stream temperatures for 1994
Etc…	
S23-33	10 Future scenarios…



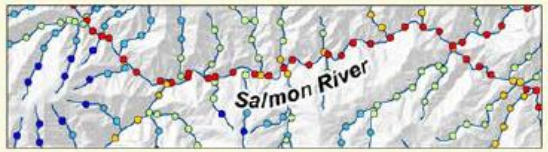
***Extensive metadata on website**

Website Distributes Raw Data & BLOB Scenarios as GIS Layers

1) GIS shapefiles of stream temperature scenarios

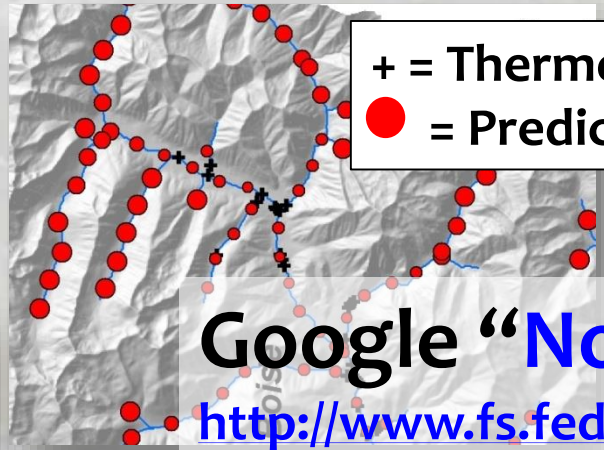


NorWeST Stream Temp

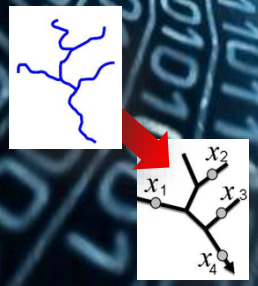


Regional Database and Modeled Stream Temperatures

2) GIS shapefiles of stream temperature model prediction precision



3) Temperature data summaries

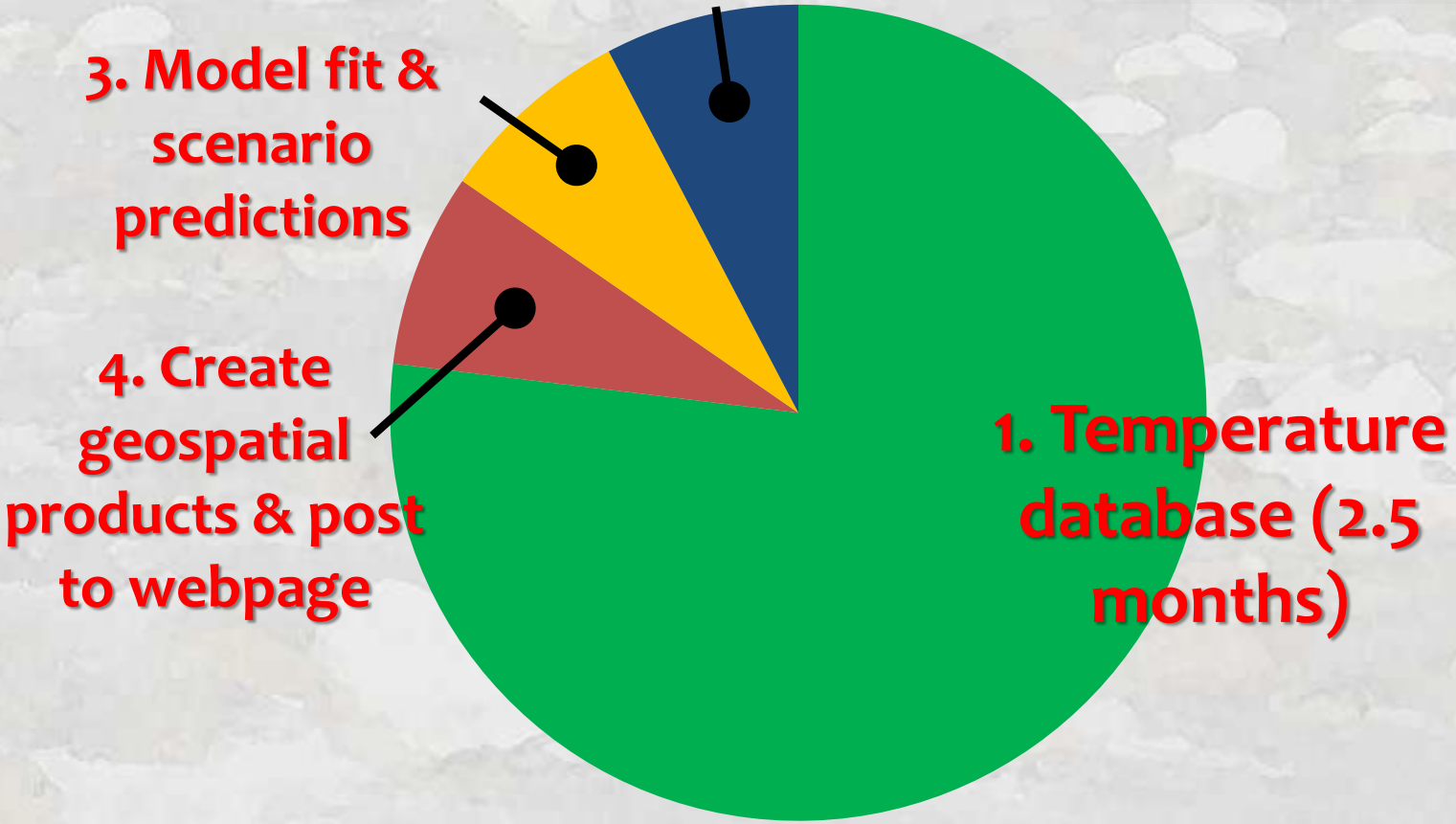


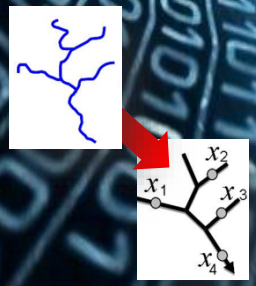
Google "NorWeST" or go here…

http://www.fs.fed.us/rm/boise/AWAE/projects/NorWeST.shtml

Potato Baking Time…

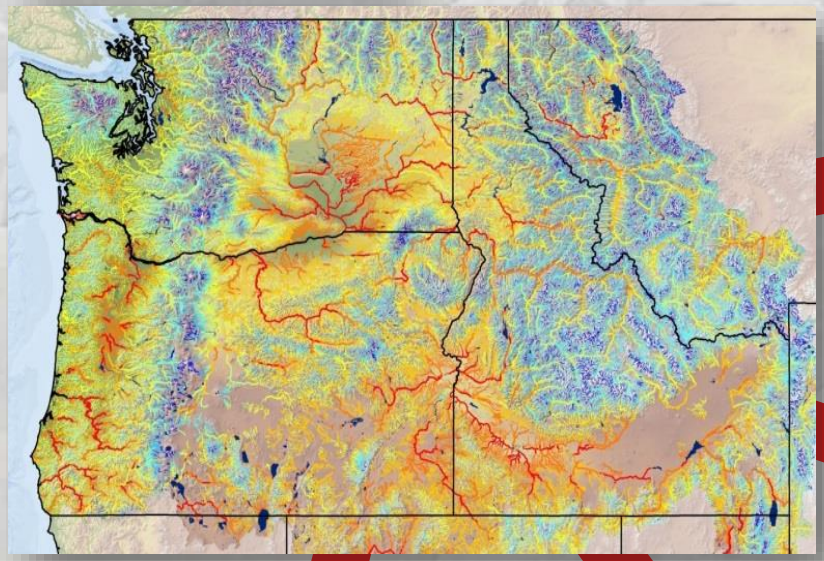
3 Months Per River Basin



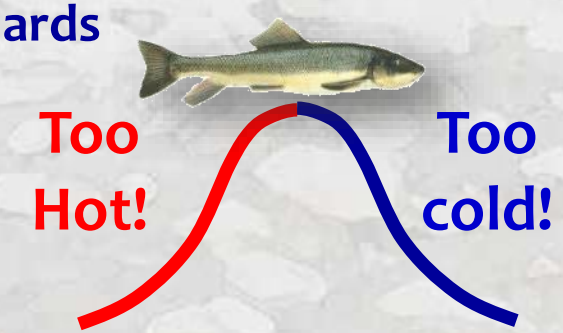


NorWeST Stream Temp

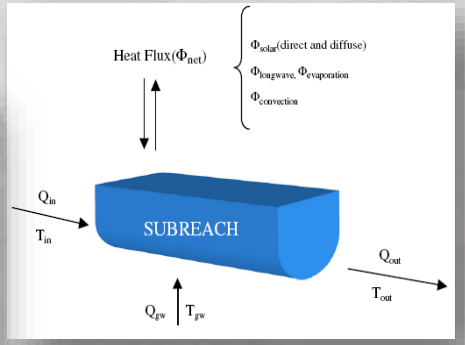
Temperature Applications

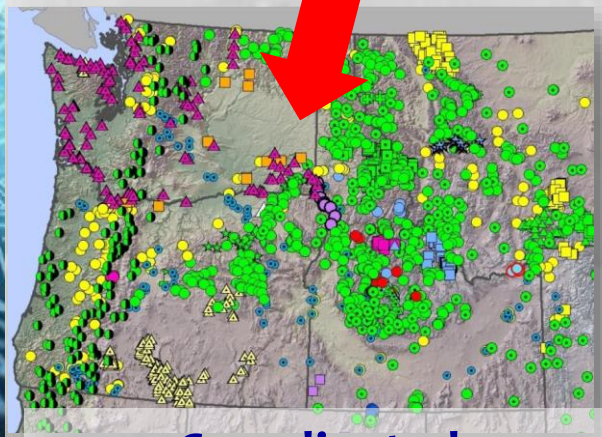


Regulatory temperature standards

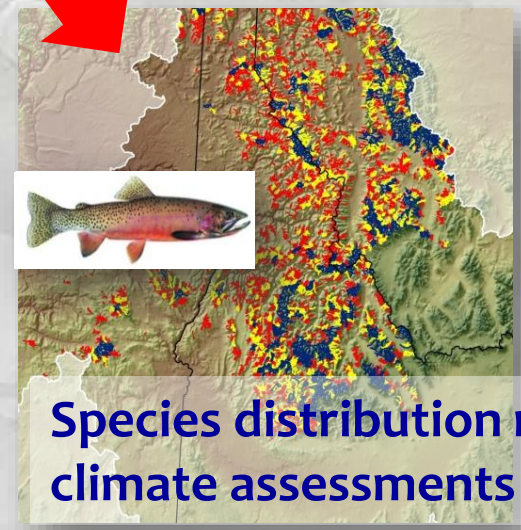


Data access accelerates temperature research

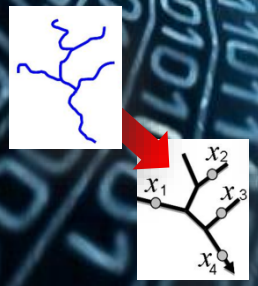




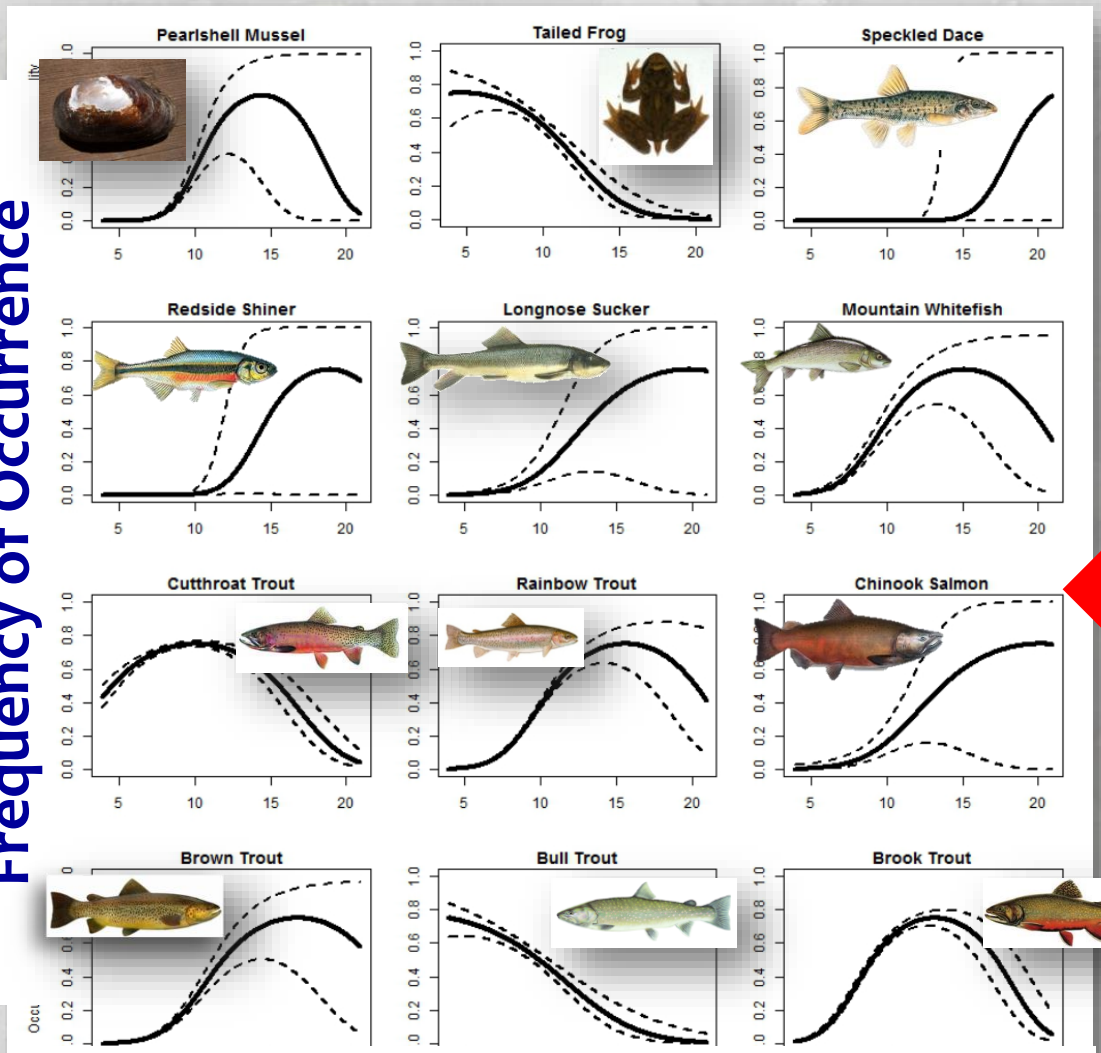
Coordinated Interagency monitoring



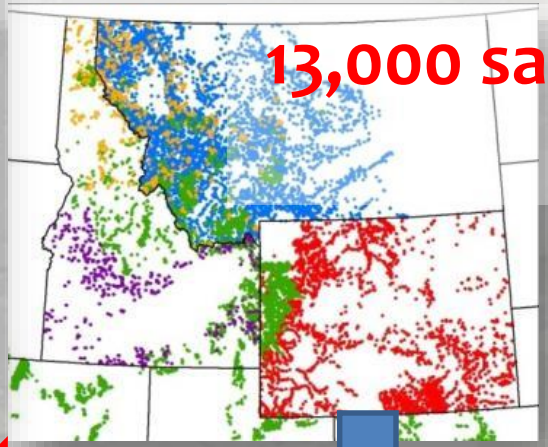
Species distribution models & climate assessments

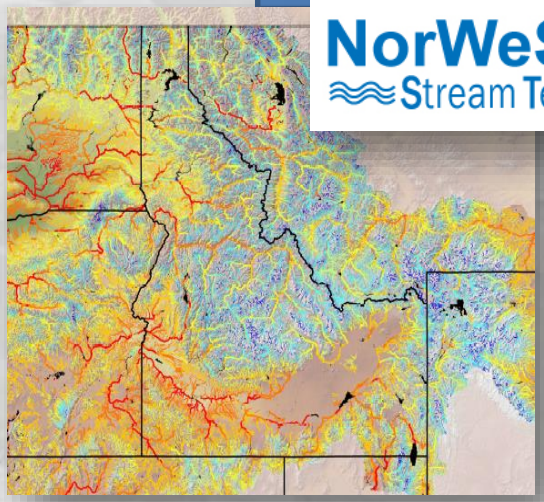


BIG DATA Thermal Criteria For Dozens of Species



BIG FISH Databases

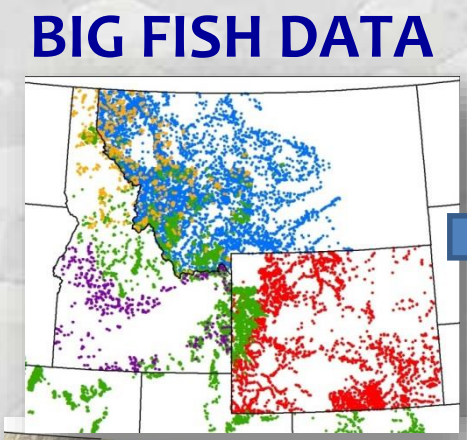


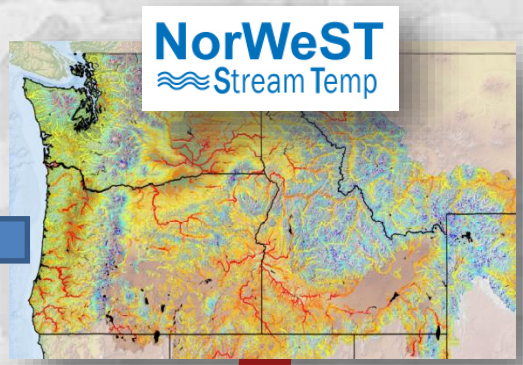


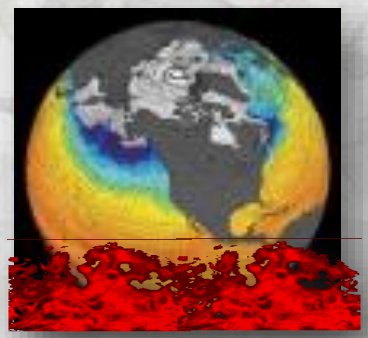
Wenger et al. *In Review.* Description of realized thermal niches from massive biological & temperature databases. *EcoSphere*

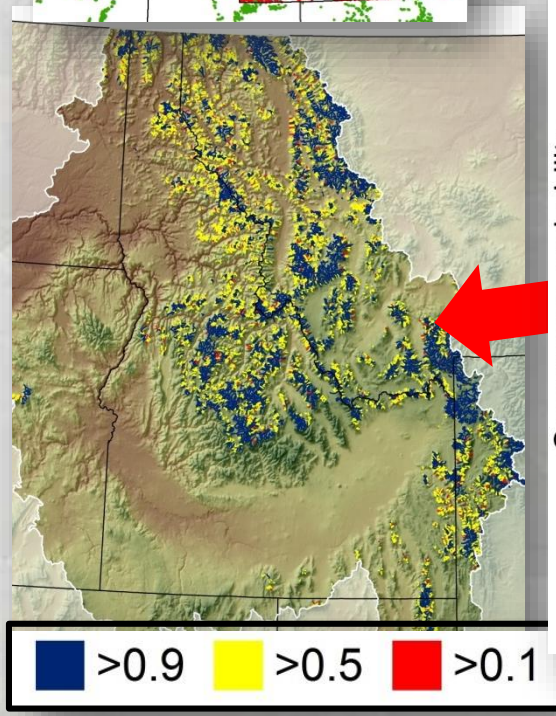
Accurate Species Distribution Models

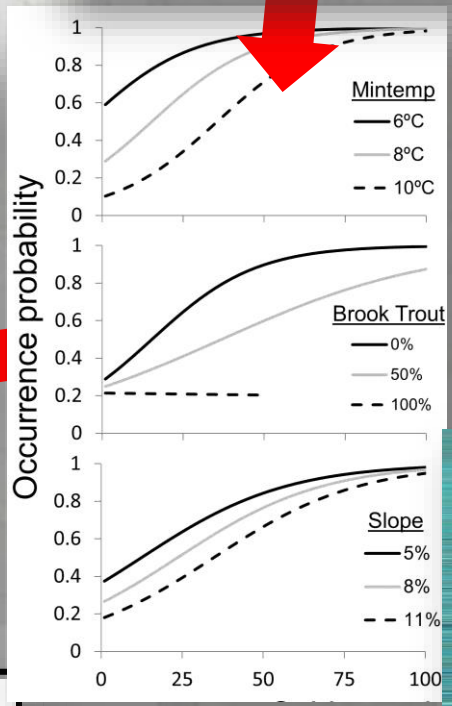
BIG FISH DATA

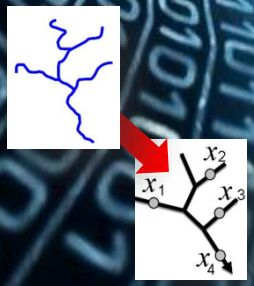








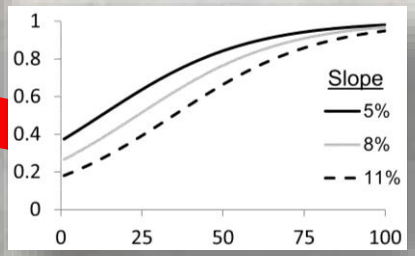




Isaak et al. 2015. The cold-water climate shield: Delineating refugia for preserving native trout through the 21st Century. *Global Change Biology* **21:**2540-2553.

Forecast Specific Climate Refugia

2080s



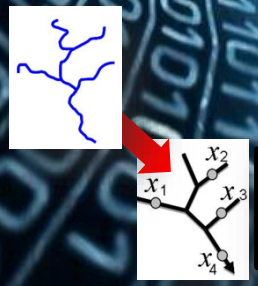
North Cascades

Flathead

Walla Walla

Metolius

Upper Salmon

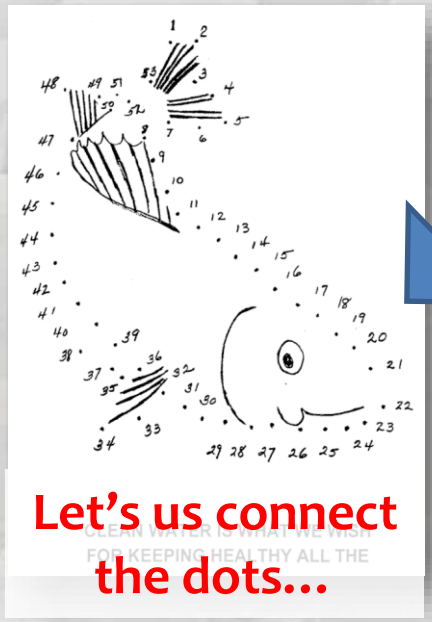


>0.9 >0.5 >0.1

62 Worst case "Bomb-shelters"

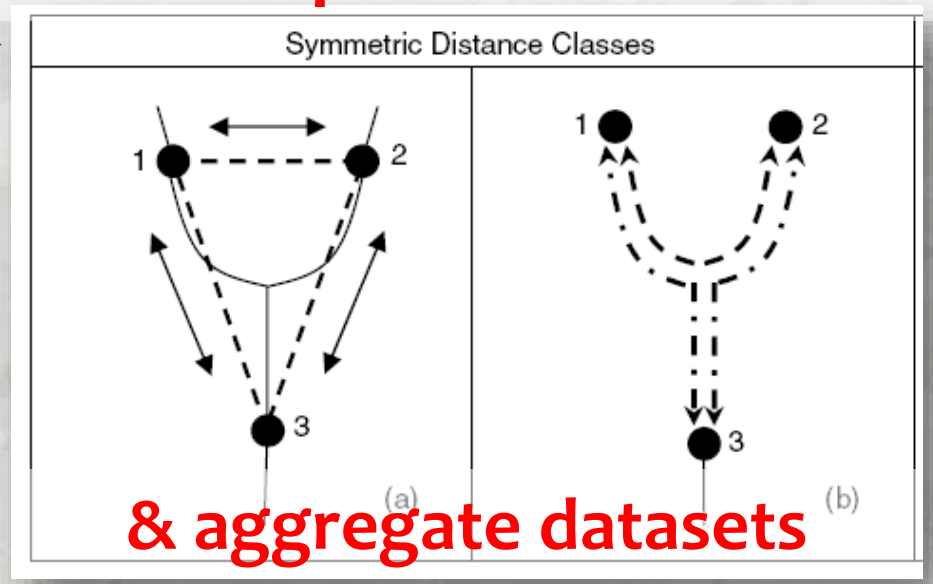
BIG DATA are often Autocorrelated

Spatial Statistical Network Models



Let's us connect the dots…

Valid interpolation on networks



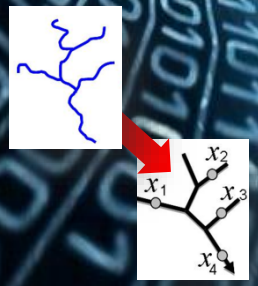
& aggregate datasets

Advantages:

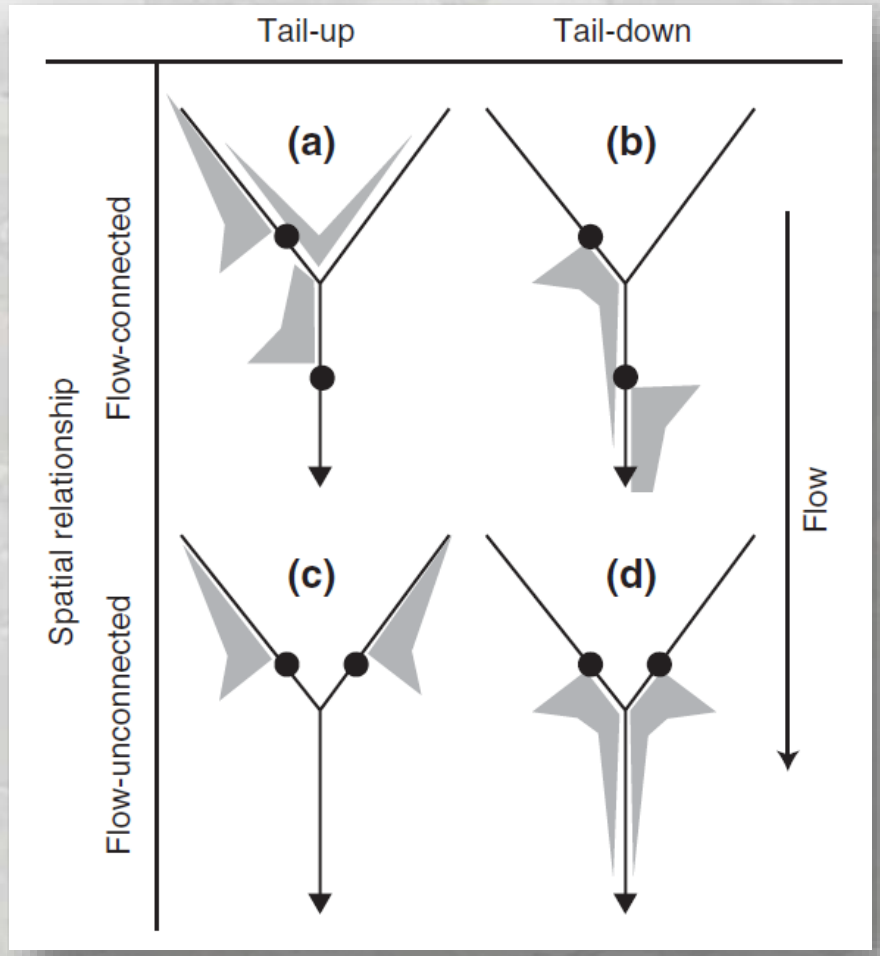
-flexible & valid autocovariance structures that accommodate network topology & non-independence among observations

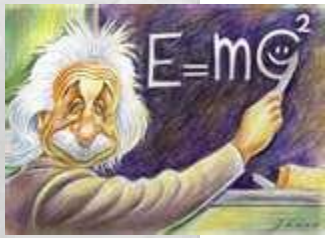
-improved predictive ability & parameter estimates relative to non-spatial models

Ver Hoef et al. 2006; Ver Hoef & Peterson 2010; Peterson & Ver Hoef 2013

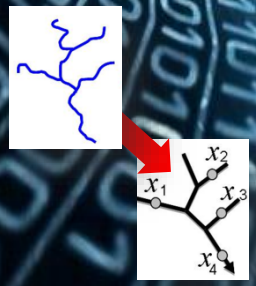


Key Innovation is Covariance Structure Based On Network Structure





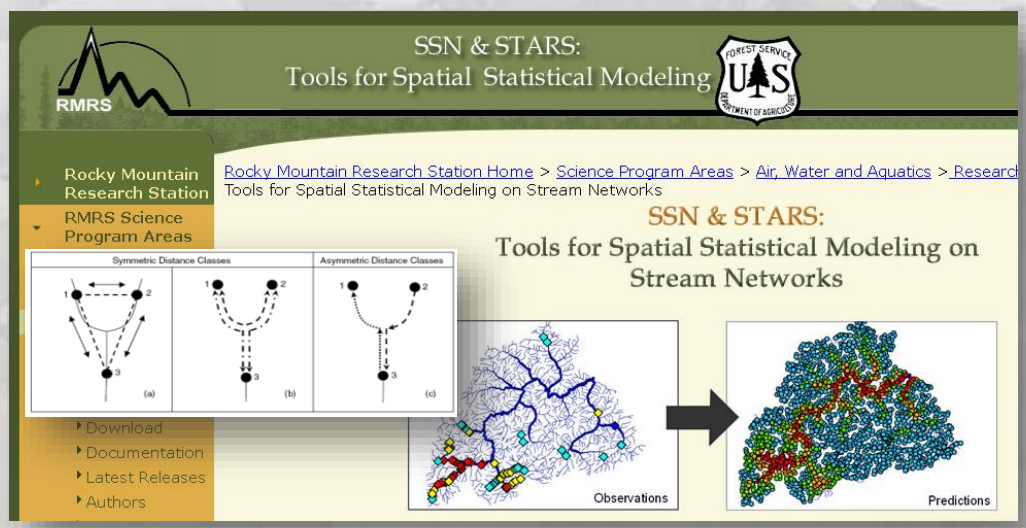
- **Models "understand" how information moves among locations**
- **Models account for spatial autocorrelation among observations**



Peterson et al. 2007. *Freshwater Biology* **52**:267-279;
Peterson & Ver Hoef. 2010. *Ecology* **91**:644-651.

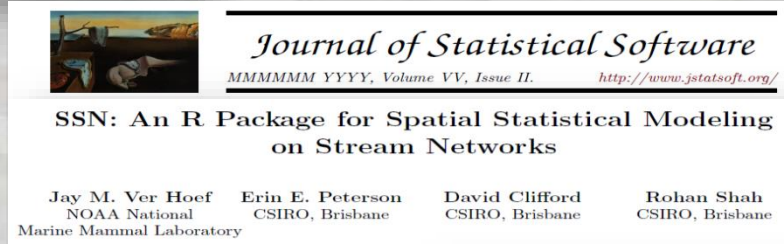
Geostatistical Stream Software is Free

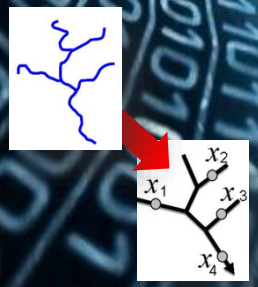
SSN/STARS Website



Spatial Stream Networks (SSN) Package for R

- **Software**
- **Example Datasets**
- **Documentation**

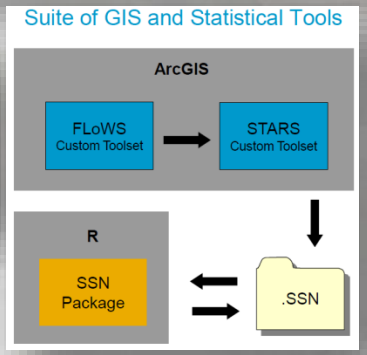




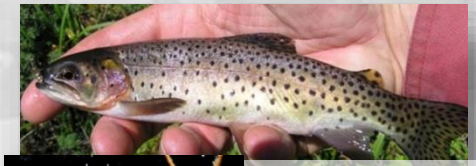
A Moving Average Approach for Spatial Statistical Models of Stream Networks

Jay M. VER HOEF and Erin E. PETERSON

STARS: An ArcGIS toolset used to calculate the spatial data needed to fit spatial statistical models to stream network data



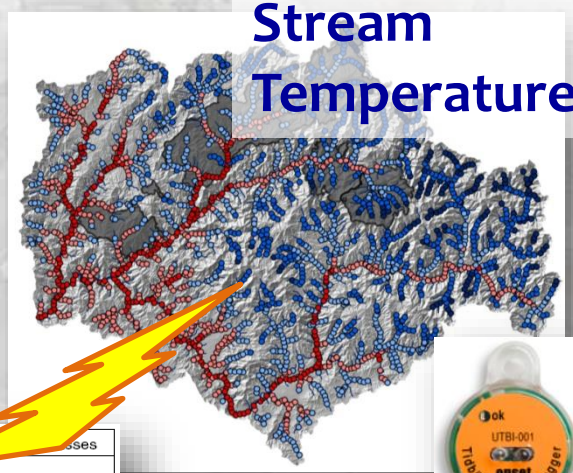
Stream Models are Generalizable…

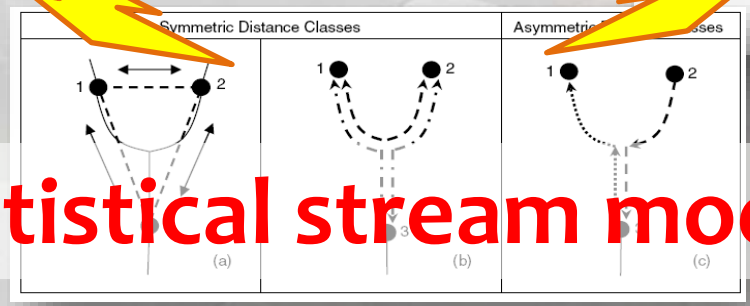


Distribution & abundance

Response Metrics
- Gaussian
- Poissan
- Binomial

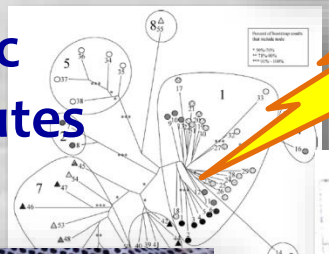
Stream Temperature

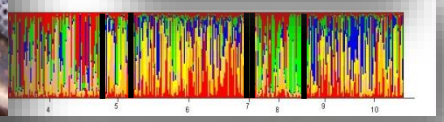




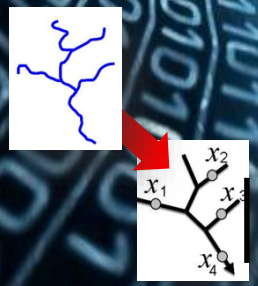
Statistical stream models

Genetic Attributes





Water Quality Parameters



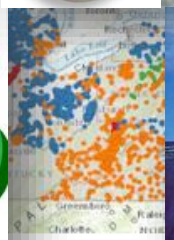
Mountains of Stream Data Can be Mined for Valuable "Information"

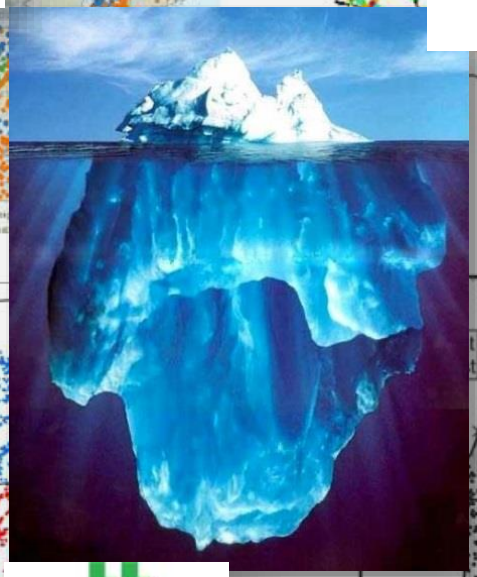
Free millions!

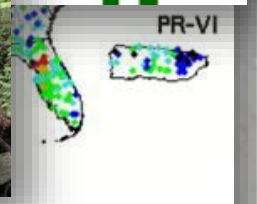
Temperature

Discharge – U

Free millions!

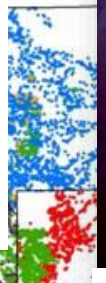






Free millions!

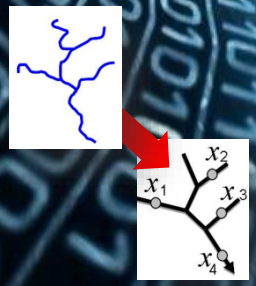
Sp
di



Water Quality

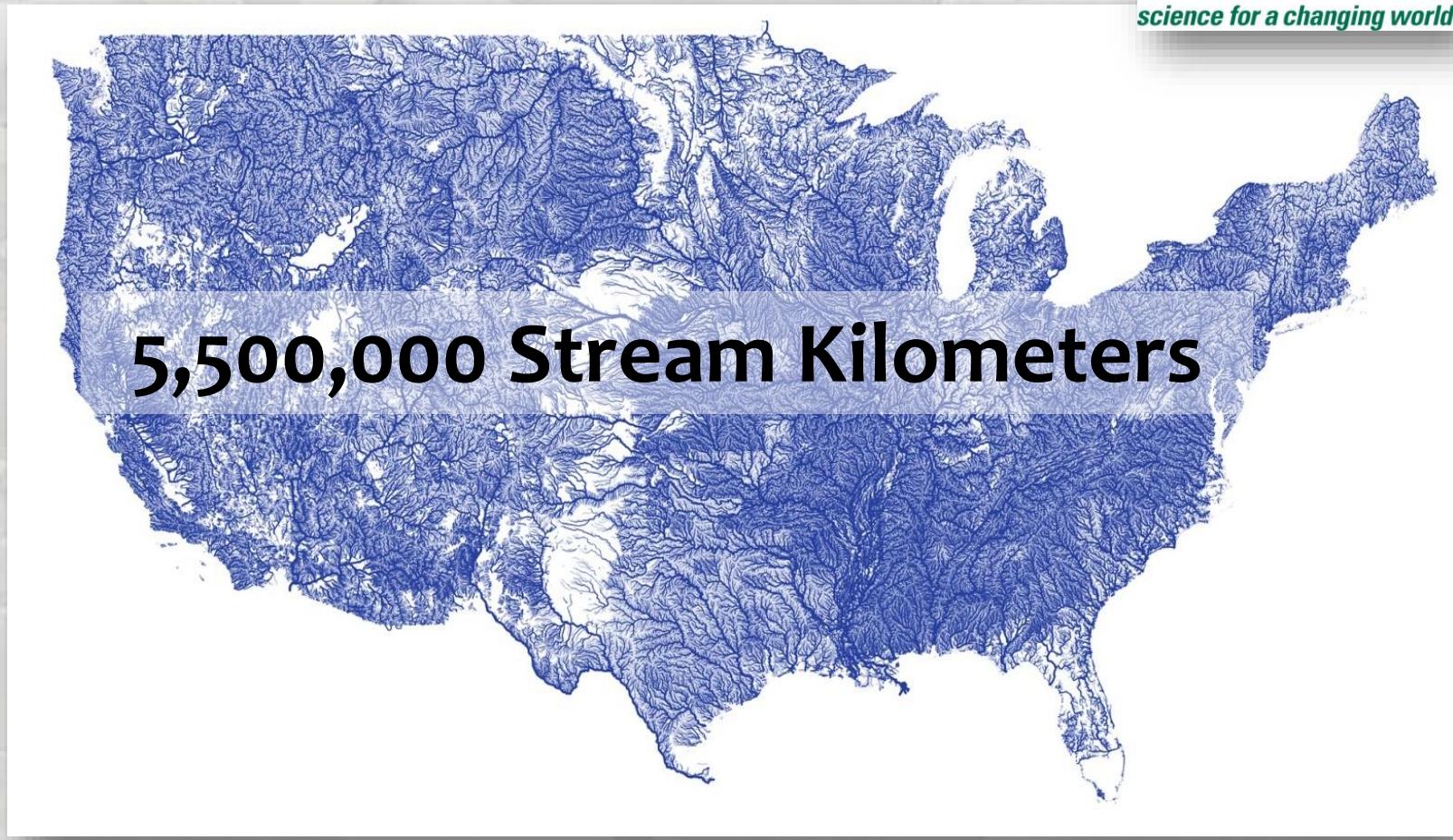
Genetic Samples

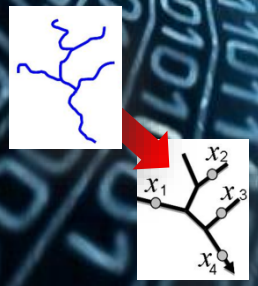
Free millions!



NHD Digital Stream Network

Nationally consistent geospatial database



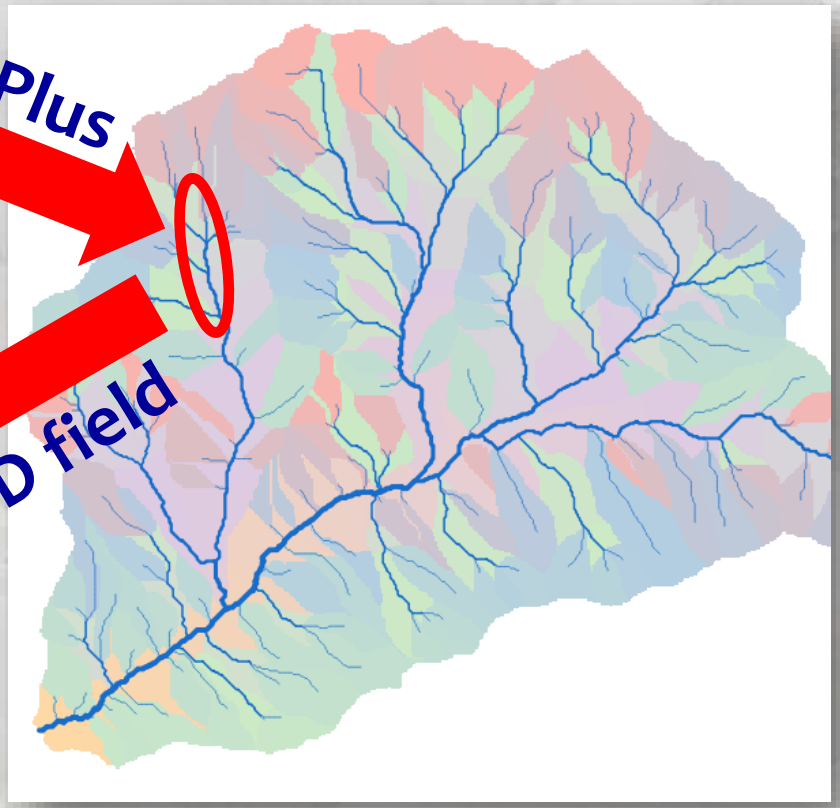


Cooter et al. 2010. A nationally consistent NHDPlus framework for identifying interstate waters: Implications for integrated assessments and interjurisdictional TMDLs. *Environmental Management* **46**:510-524.

"PLUS" part of NHDPlus (Stream Reach Predictors/Descriptors)

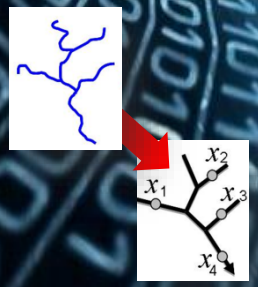
NHDPlus

COMID field



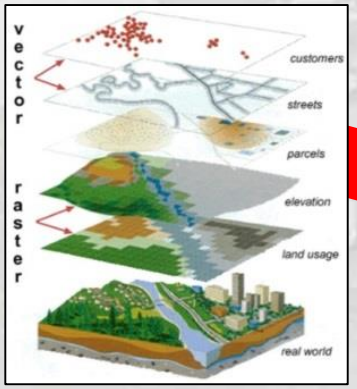
- Elevation
- Slope
- %Landuse
- Precipitation

10's more…



Wang et al. 2011. A Hierarchical Spatial Framework and Database for the National River Fish Habitat Condition Assessment. *Fisheries* **36**:436-449.

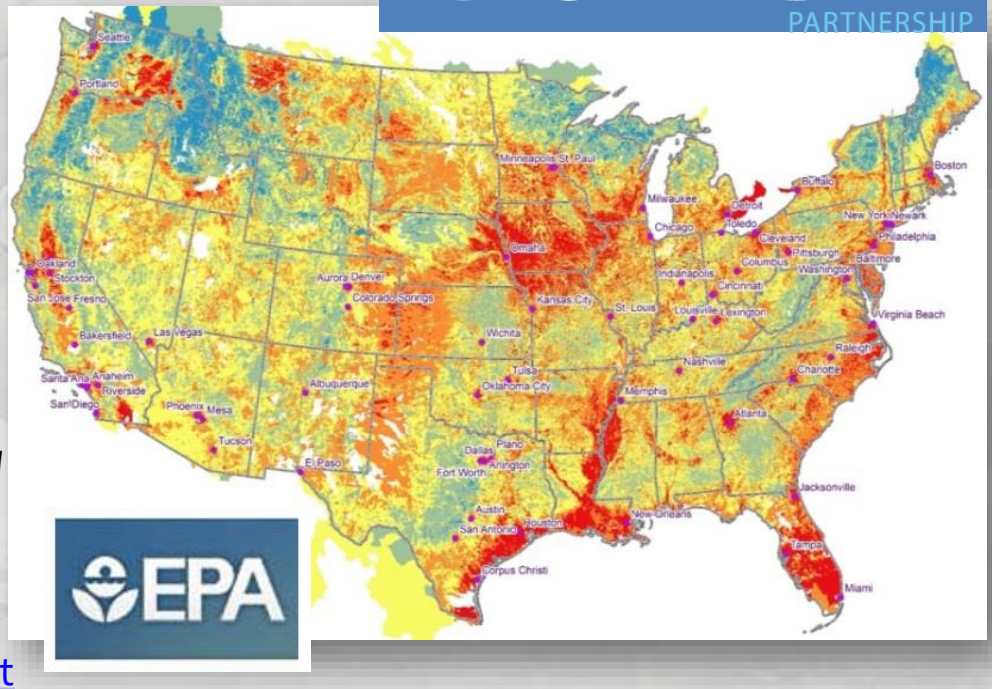
More Stream Reach Predictors/Descriptors in Nationally Available GeoDatabases

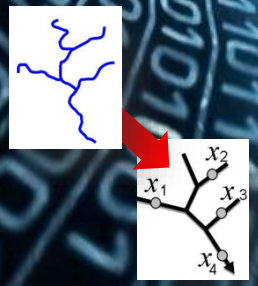


Wang et al. 2011. A hierarchical spatial framework and database for the national river fish habitat condition assessment. *Fisheries* **36**: 436-449. https://www.researchgate.net/profile/Lizhu_Wang2

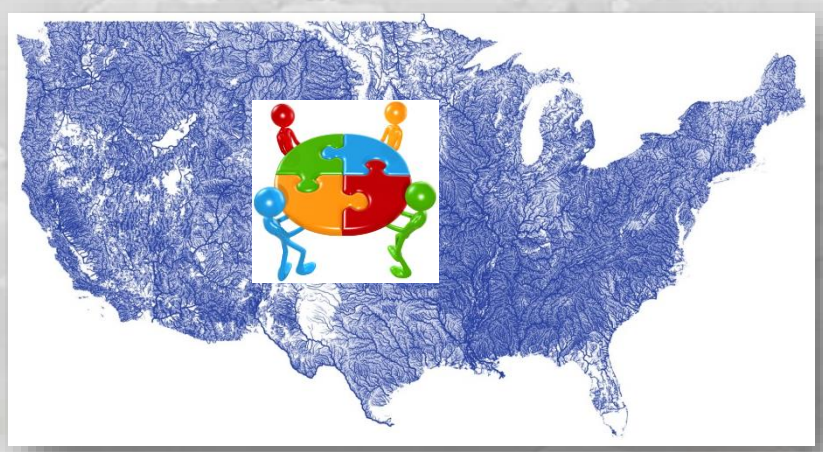
Databases of stream reach descriptors

Hill et al. In Press. The stream-catchment (StreamCat) dataset: A database of watershed metrics for the conterminous USA. *The Journal of the American Water Resources Association*. http://www2.epa.gov/national-aquatic-resource-surveys/streamcat





Website Hub: The National Stream Internet



LANDSCAPE CONSERVATION COOPERATIVES

NOAA

ACEMS

Boise Spatial Streams Group

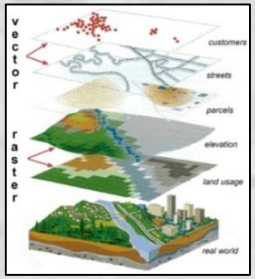
FOREST SERVICE US DEPARTMENT OF AGRICULTURE

USGS

NSI Resources

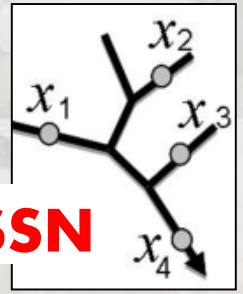
Workshop & presentations

NSI hydrography network (shapefiles)



Databases of stream reach descriptors

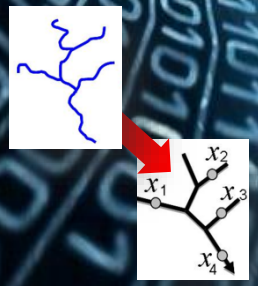
Databases of stream measurements



Spatial stream-network models

Ideas → Data → Analysis → Information

One Last Thing—Who All Lives Here?



Aquatic eDNA frontier

USFS National Genomics Center for Wildlife & Fish Conservation

- Pioneered the technique for salmonids
- Species specific, highly reliable (1 trout / 100 m = 85% detection)
- Field-proven protocol
- Cost: $65 sample

USDA Forest Service
National Genomics
Center for Wildlife
and Fish Conservation

Mike Schwartz
Mike Young
Kevin McKelvey

Google the website:

http://www.fs.fed.us/research/genomics-center/

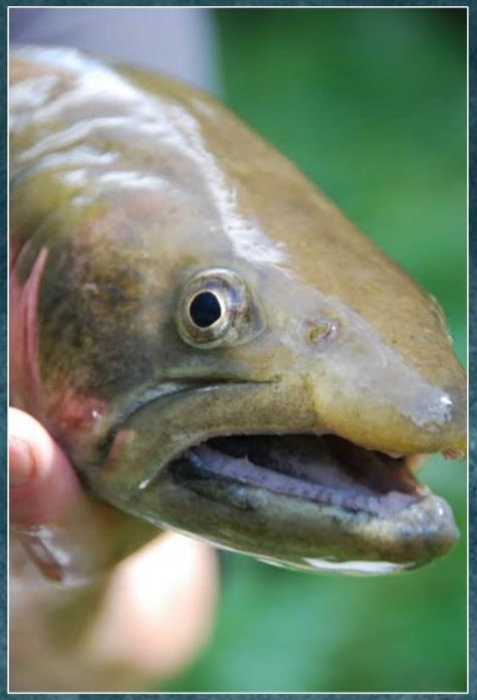
eDNA project to census Bull Trout streams for regional status assessment (2015-2018)

The rapid, range-wide inventory of bull trout: a crowd-sourced, eDNA-based approach with application to many aquatic species

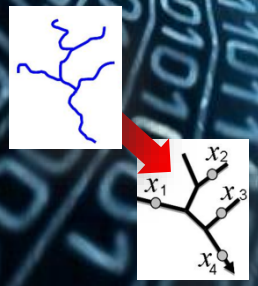
Michael Young, Kevin McKelvey, Michael Schwartz, Dan Isaak, Kellie Carim, Taylor Wilcox, Katie Zarn, Kristy Pilgrim, Dona Horan, Sherry Wollrab

Collaborators

Bureau of Reclamation
Clark Fork Coalition
Clearwater Resource Council
Coeur d'Alene Tribes
Idaho Department of Fish and Game
Idaho Power Company
Montana Department of Natural Resources Conservation
Montana Fish, Wildlife & Parks
National Fish and Wildlife Foundation
The Nature Conservancy
Nez Perce Tribes
Oregon Department of Fish and Wildlife



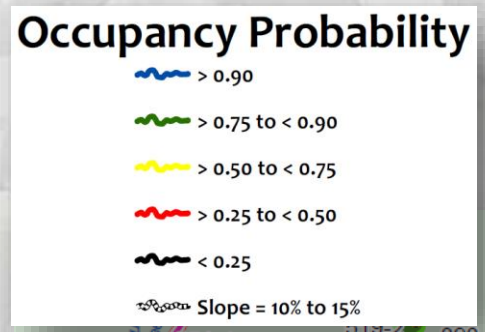
Trout Unlimited
U.S. Fish and Wildlife Service
USFS Beaverhead-Deer Lodge NF
USFS Boise NF
USFS Helena NF
USFS Idaho Panhandle NF
USFS Lolo NF
USFS Region 1
USFS Region 4
USFS Region 6
USFS Sawtooth NF
Washington Department of Fish and Wildlife
Yakima Nation

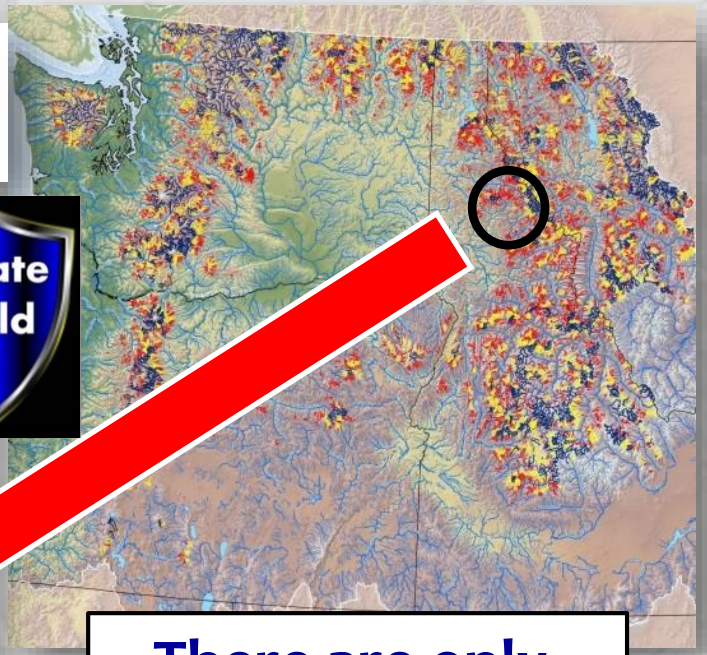


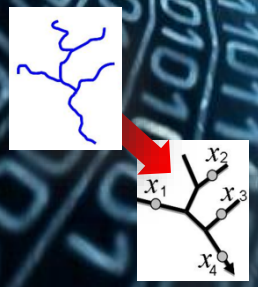
Forest Service Research & Development
National Genomics Center for Wildlife & Fish Conservation

Rocky Mountain Research Station

Combine eDNA sampling with Precise Predictions from Climate Shield Model







There are only…

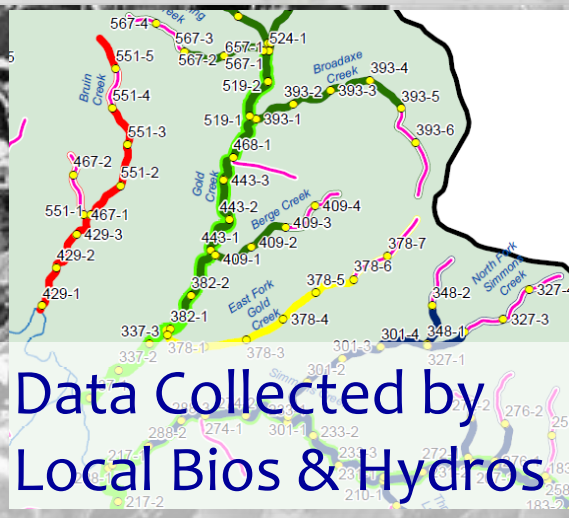
5,332 >0.1 habitats

1,325 >0.5 habitats

348 >0.9 habitats

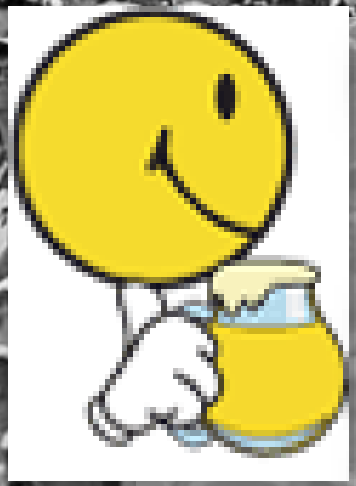
Target samples to resolve greatest uncertainty

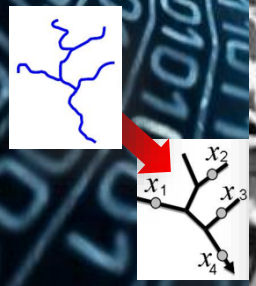
Data Primarily Collected by Crowd-Sourcing



Data Collected by Local Bios & Hydros

High-quality data developed collaboratively

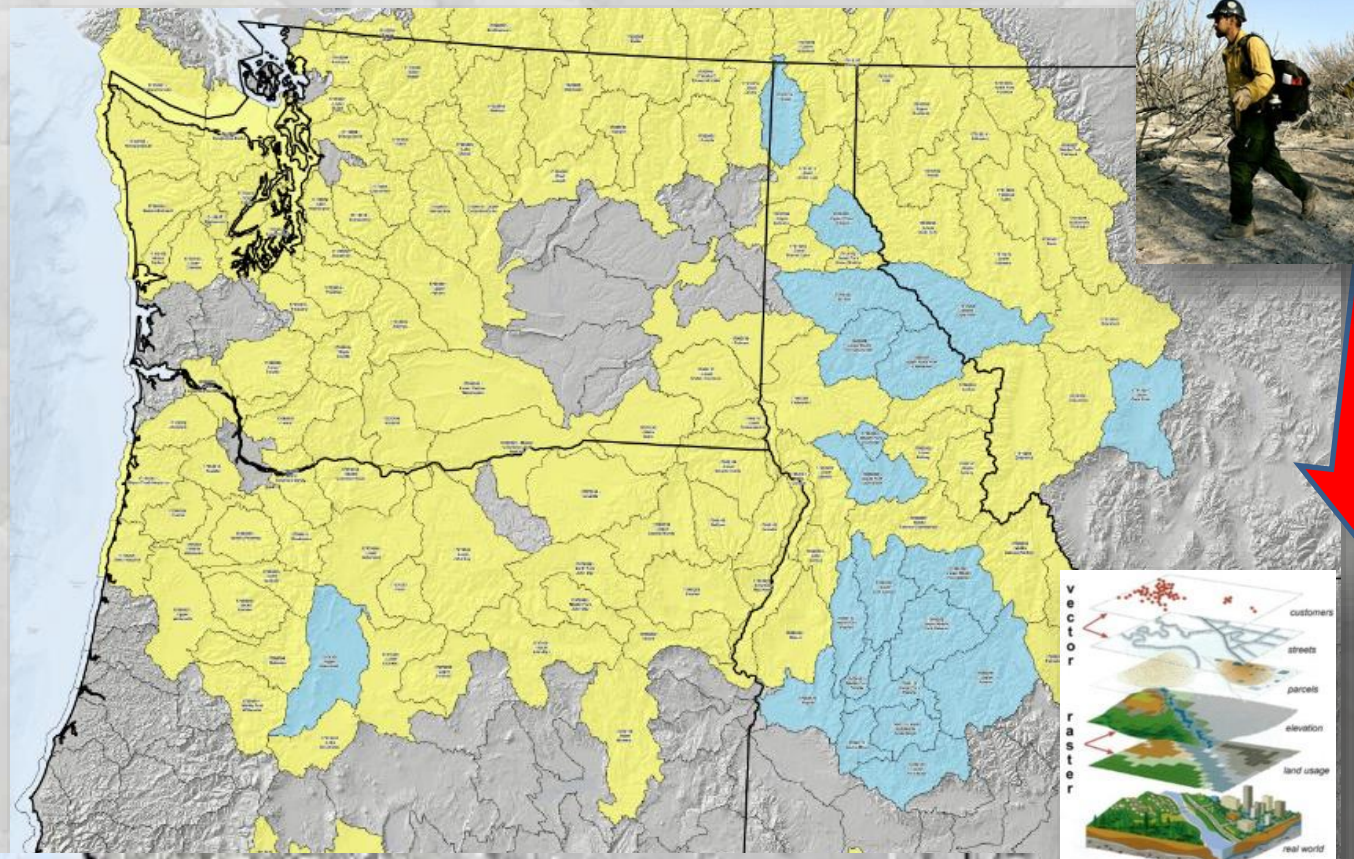


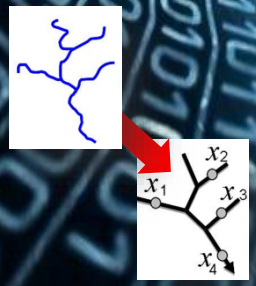


Management & regulatory actions

2016-2018: Industrial scale field campaigns

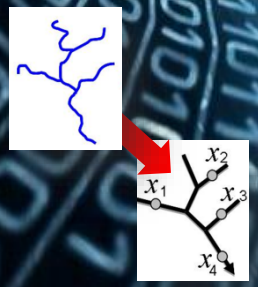
Everything is a digital database from day 1!

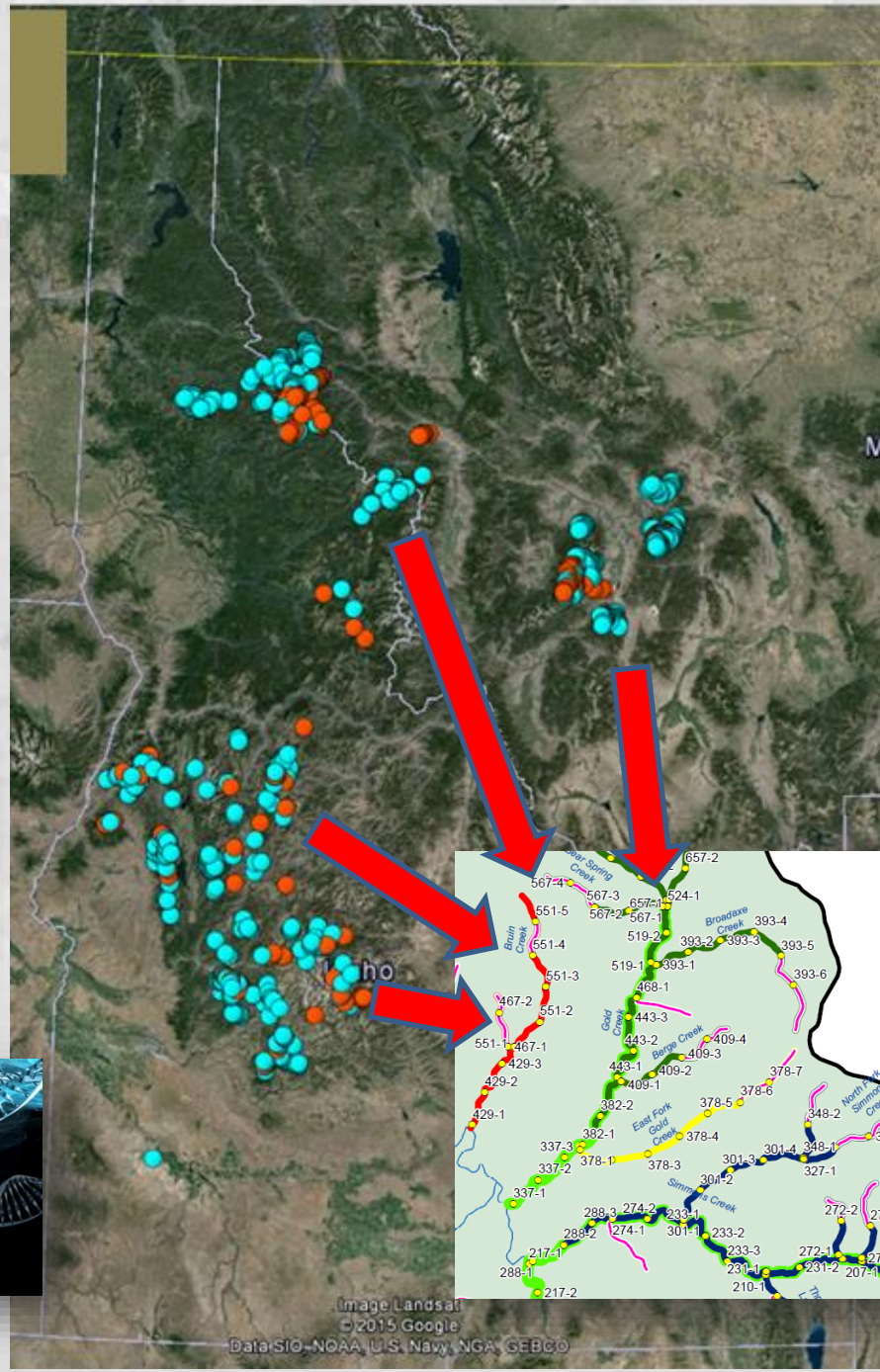




2015 “Pilot” Year:

- eDNA samples collected at 833 stream sites
- A few new populations discovered
- A few old populations “found again”





Website for Bull Trout Information Updated with eDNA Sample Results…

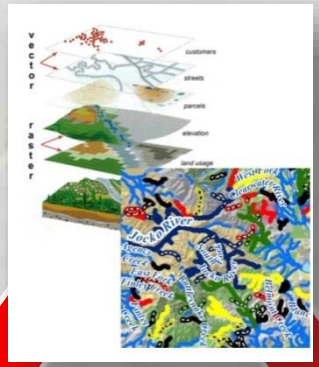
Climate Shield website:

http://www.fs.fed.us/rm/boise/AWAE/projects/ClimateShield.html

Presentations & Publications

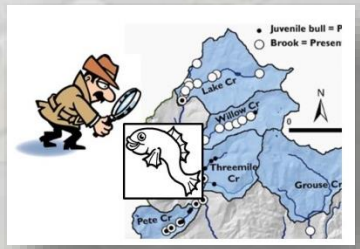
COLDWATER AS A CLIMATE SHIELD TO PRESERVE NATIVE TROUT THROUGH THE 21ST CENTURY

Digital Maps & ArcGIS Shapefiles



Fish Data Sources

Distribution Monitoring



File formats:

- ArcGIS files
- pdf files

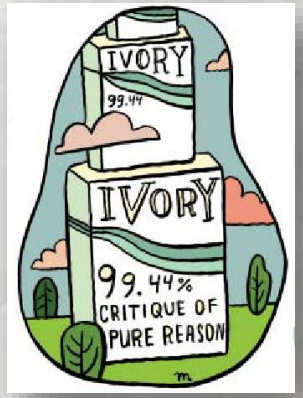
15 Scenarios:

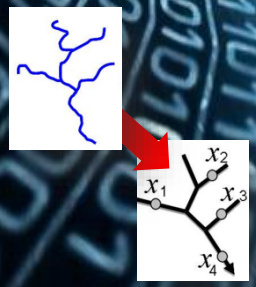
- 3 climate periods
- 5 Brook invasion levels

Create an Efficient Cycle of Information

Many stakeholders

"Boots-on-the-Ground"



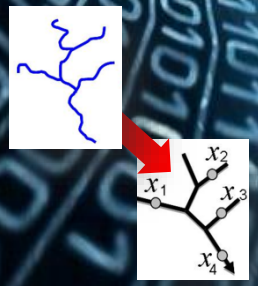


Research develops databases & relevant information

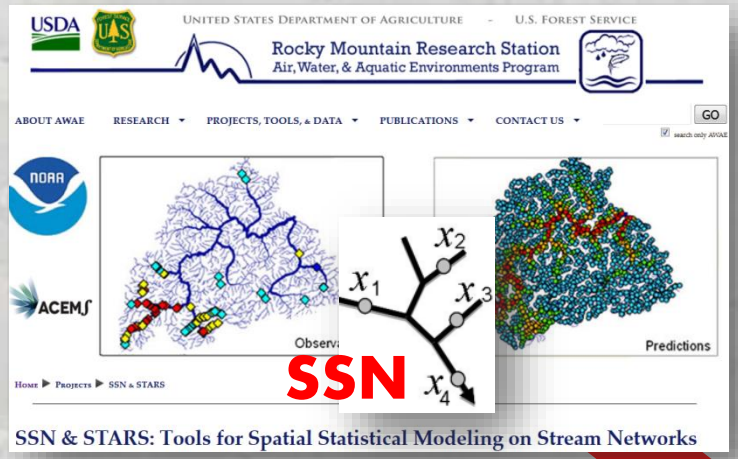
Mountains of data

Samples contain eDNA for all Critters!

A biodiversity archive as side benefit

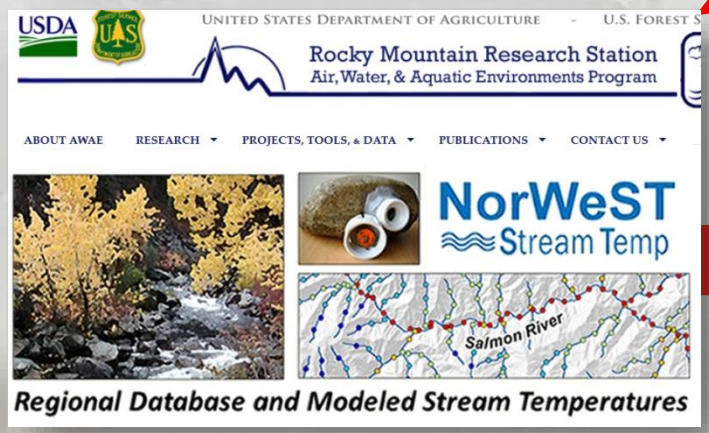


Websites Distribute Information in User-friendly Formats (GIS databases, software, digital maps, manuscripts, videos, etc.)



SSN & STARS: Tools for Spatial Statistical Modeling on Stream Networks

Climate Shield Cold-Water Refuge Streams for Native Trout



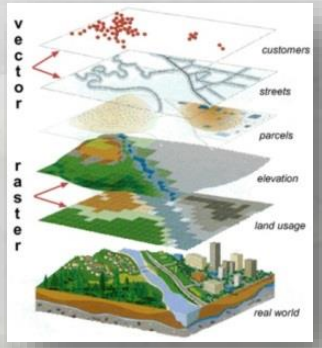
Regional Database and Modeled Stream Temperatures

Locations of recent website visits

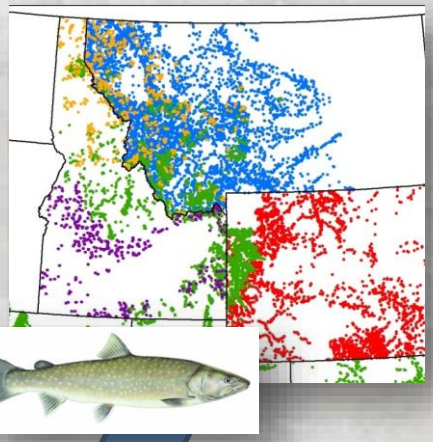
Components Exist for Creating Massive Amounts of Stream Information

Hypotheses & Ideas

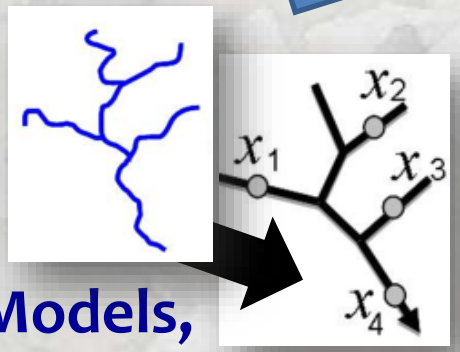
Environmental Predictors



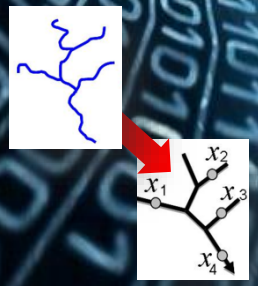
BIG DATA



Models, parameters & tests

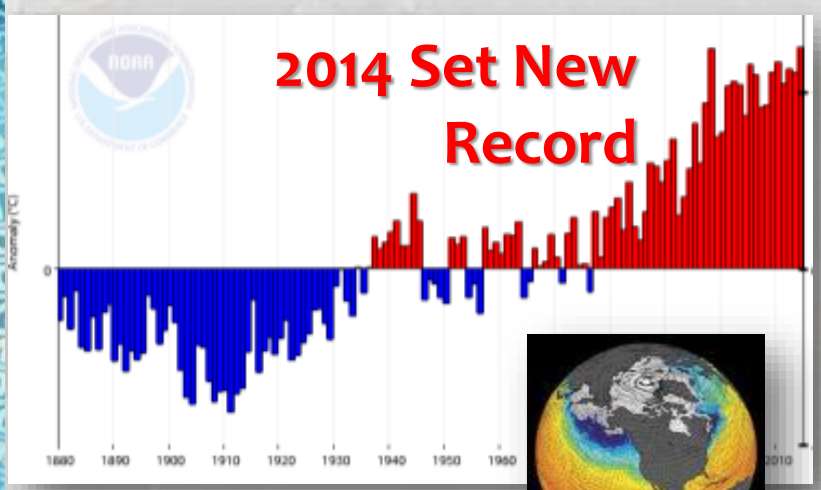


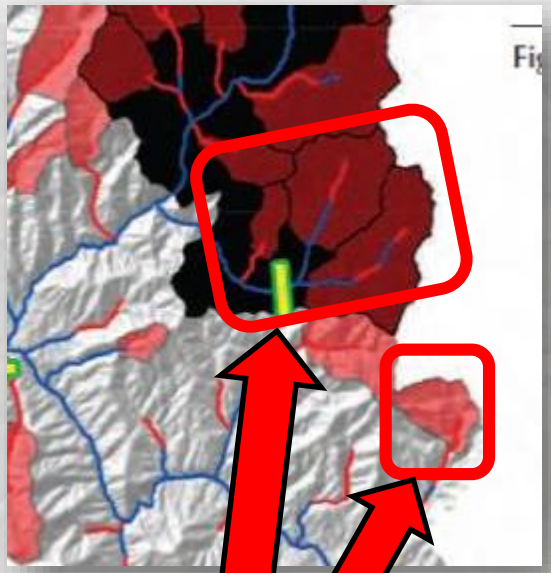
New Information!



Good Information for Strategic Decision Making Will Be Critical

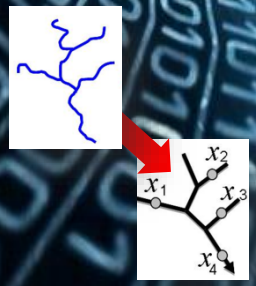
The 21st-Century will Be a Transitional One





I'm going to invest here…

…not here



Challenge 1. organizing it all – norwest example 4 years

- **Before collect new, organize old – millions$**
- **Database team**

Challenge 2. autocorrelation

Once you do it can be transformative – norwest example

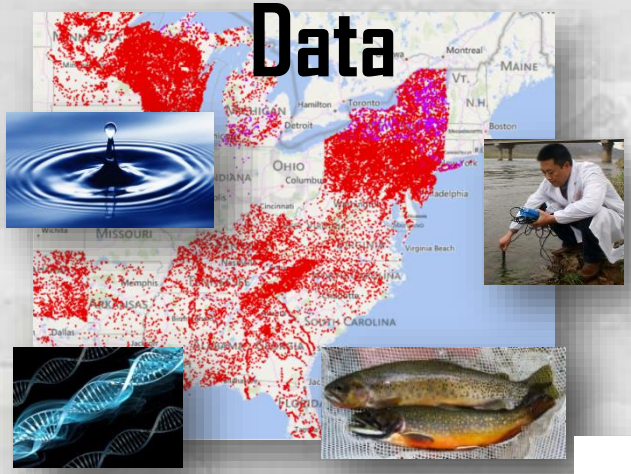
- **examples**
- **potato baking time**

Next generation synergies – grows together. Crowdsourcing, standard protocol, eDNA & Oracle & bull trout

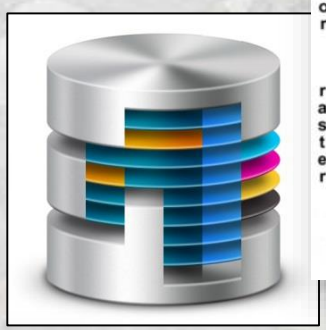
Information creation easy

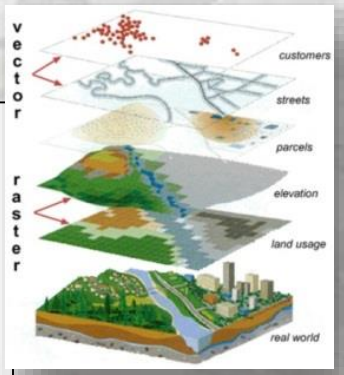
NSI – makes possible everywhere

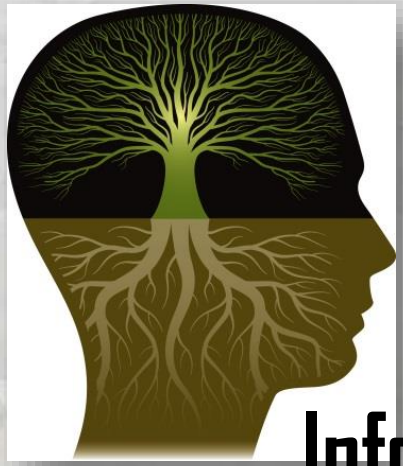
Data



Database

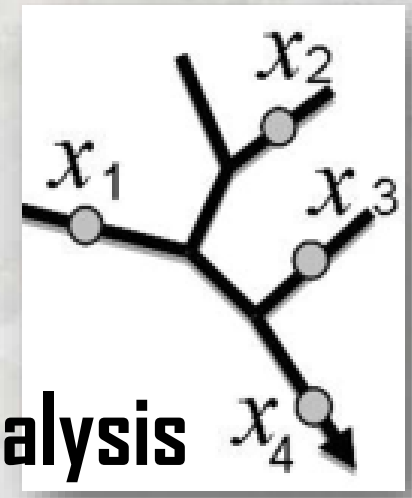


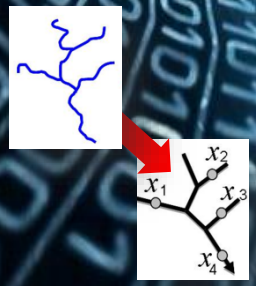




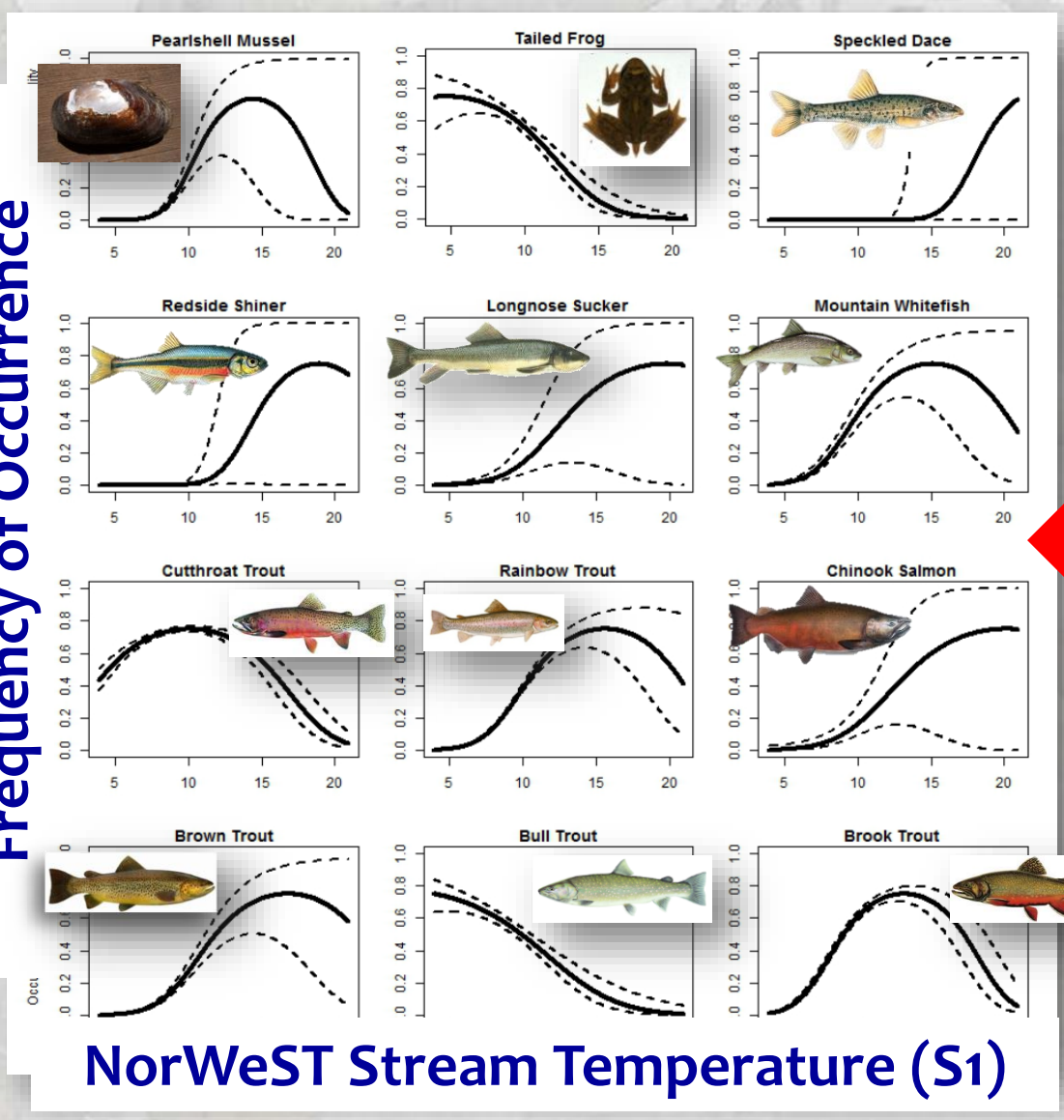
Information

Analysis

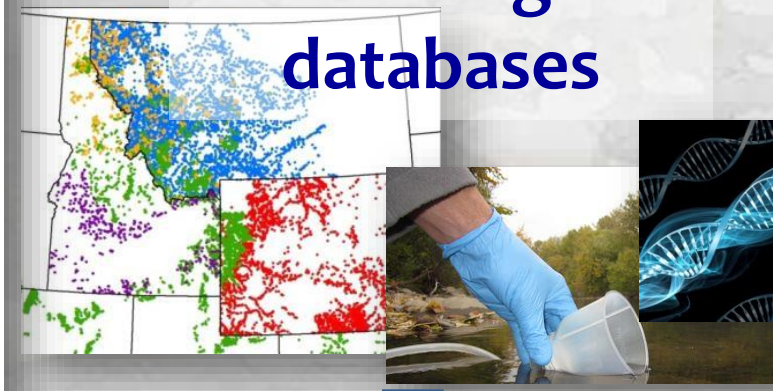




Goal: Empower people with tools to develop high-resolution distribution/abundance/genetic information for all aquatic life…

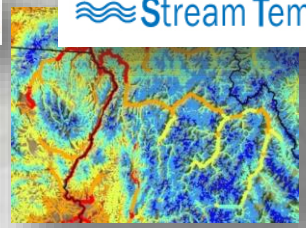


BIG biological databases



Geotechnical & analytical processing

NorWeST Stream Temp



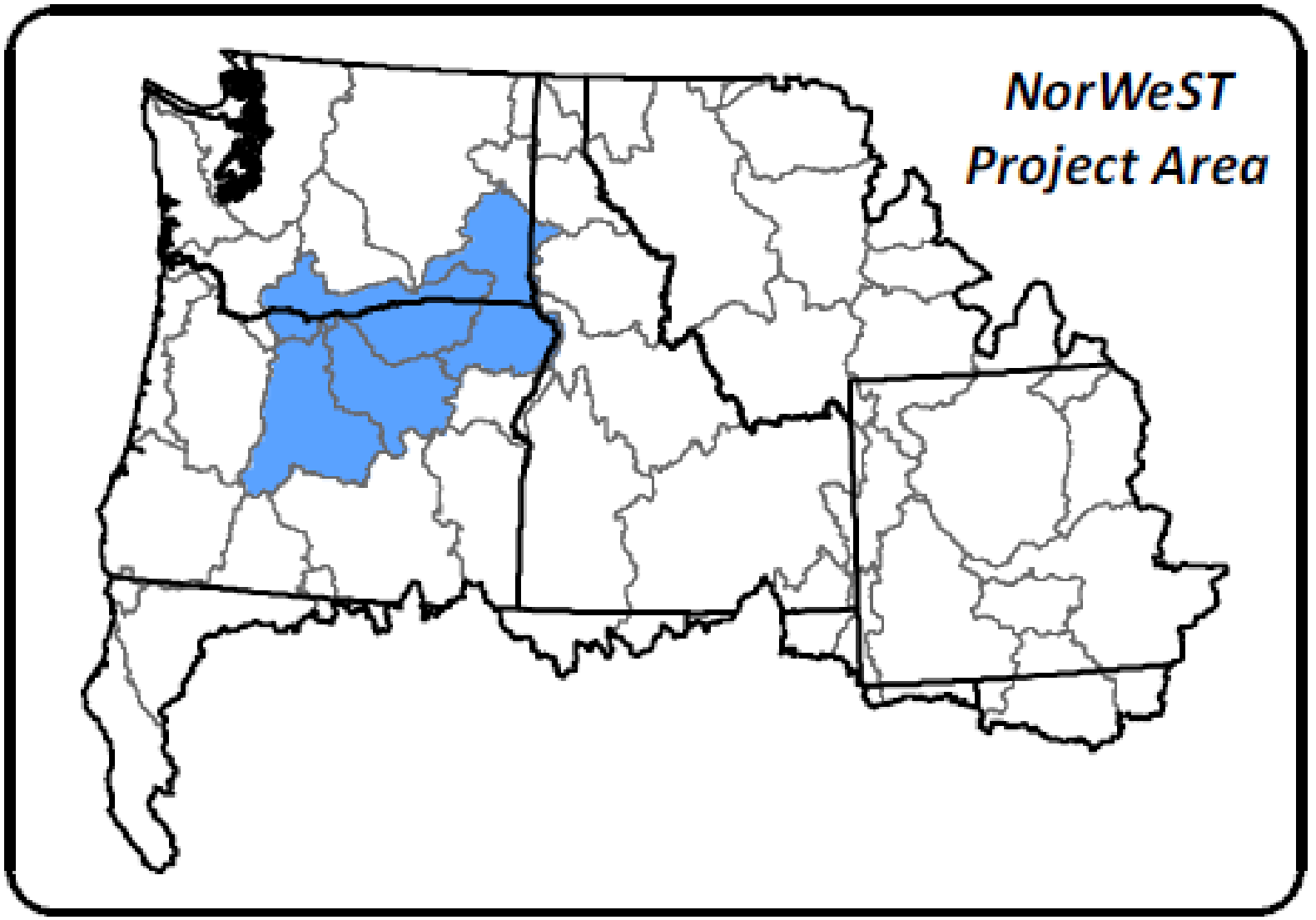
SSN

Climate Shield

e.g., Wenger et al. *In Review.* ***Diversity & Distributions***

Creating "Information" from the Database

An Example for the Mid-Columbia



ee NSI talk

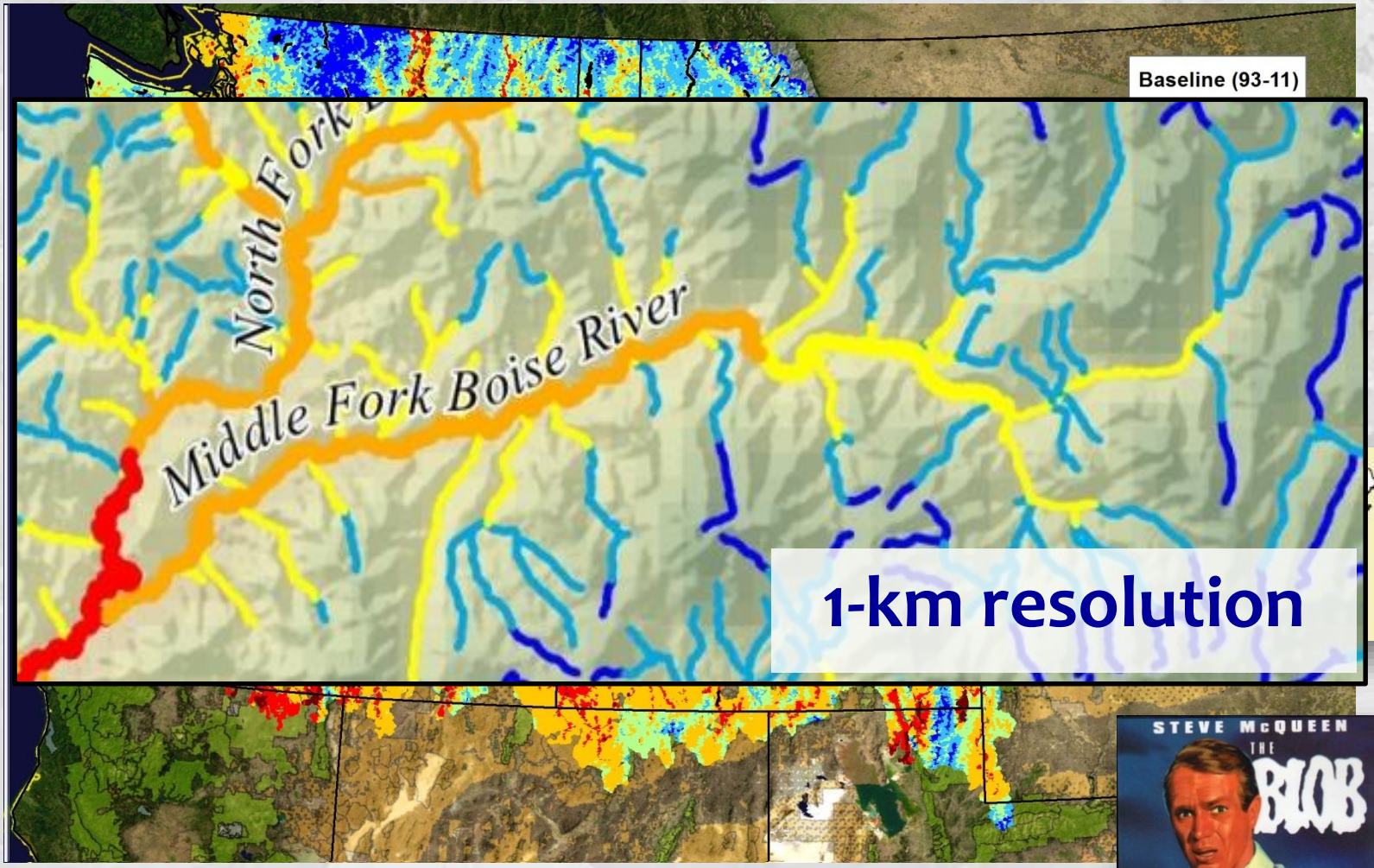
Ve live in a world dripping with data – show GBIF occurrence map examp
00,000,000 occurrence locations

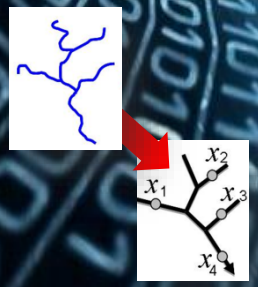
Here locally in the PNW & aquatic datasets the same is true…examples of
Norwest temp, Wenger fish. This country spends 100s million dollars to co

The aquatics world verging on the era of 'big data'. I say on the verge of be
ve can't yet fully tap the wealth of information offered by these big datas
ecause we very rarely can yet access those data in any semblance of a
unctional database.

o do that we need to create dedicated, expert database teams that aggre
rganize, and document data to build open access databases. Once that's
oing next step of extracting lots of useful information is possible and muc
asier than database building. here's a few examples of what can be done
Norwest & SSN, thermal criteria, CS.

High-Resolution Stream Temp Scenarios





The BLOB… it just keeps growing…

- 46,674 summers of data swallowed
- 222,000 stream kilometers of thermal ooze