MIT EARTH RESOURCES LABORATORY
ANNUAL FOUNDING MEMBERS MEETING 2020

Earth
Resources
Laboratory

Machine Learning for Natural Resource Assessment

An application to the Blind Geothermal Systems Of Nevada

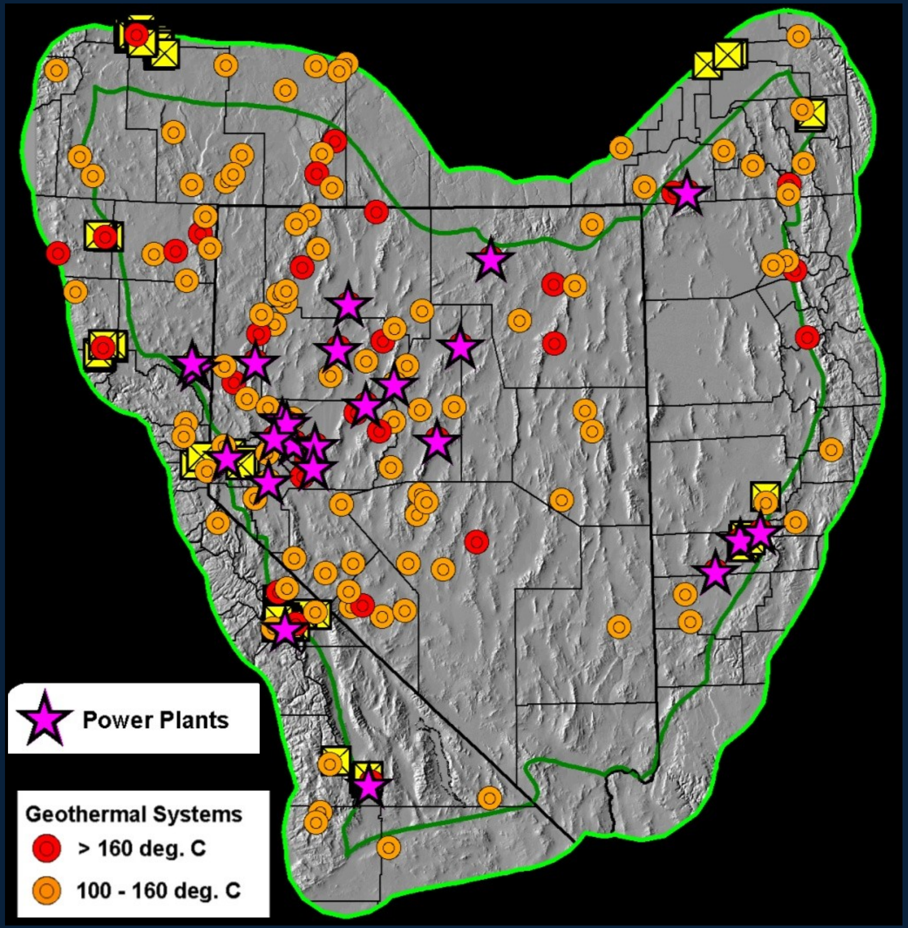
Stephen R. Brown, Ph.D.

RESEARCH SCIENTIST 浪人

In collaboration with James Faulds, Mark Coolbaugh, Jake DeAngelo, John Queen, Sven Treitel, Mike Fehler, Eli Mlawsky, Jonathan Glen, Cary Lindsey, Erick Burns, Connor Smith, Chen Gu, Bridget Ayling

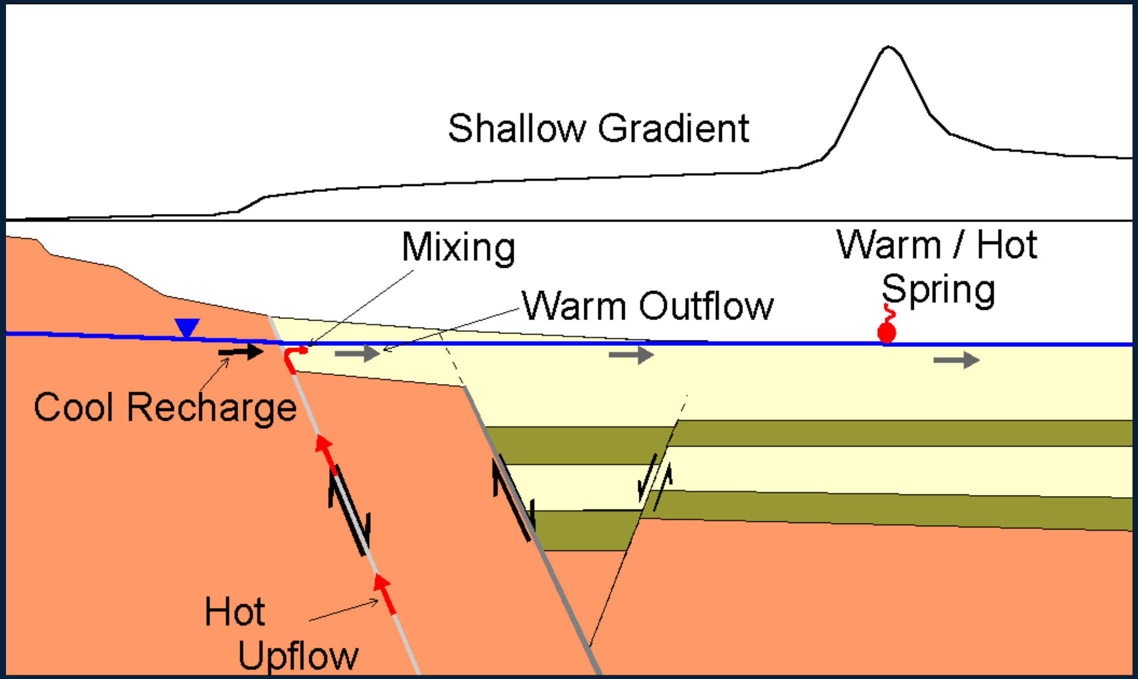
Great Basin Geothermal Production

- **Nearly 1 GW capacity in region**
- **Typical system produces 10 to 300 MW**
- **1 MW enough energy for 750-1,000 homes**
- **Region has much greater potential**

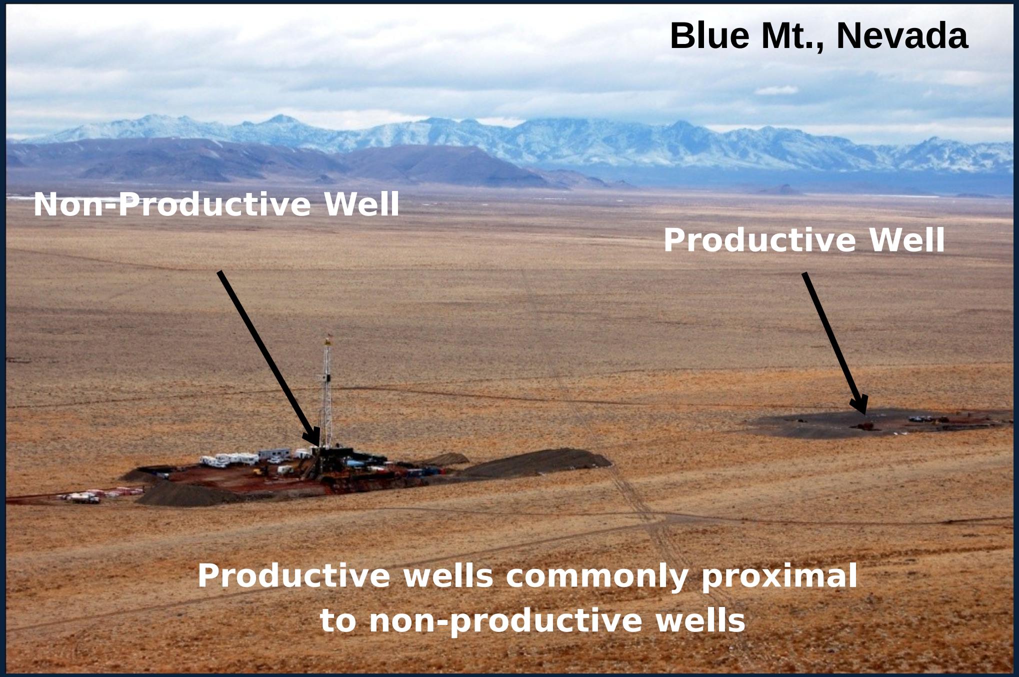


Distribution of Known Systems

Major Challenge – Blind Systems

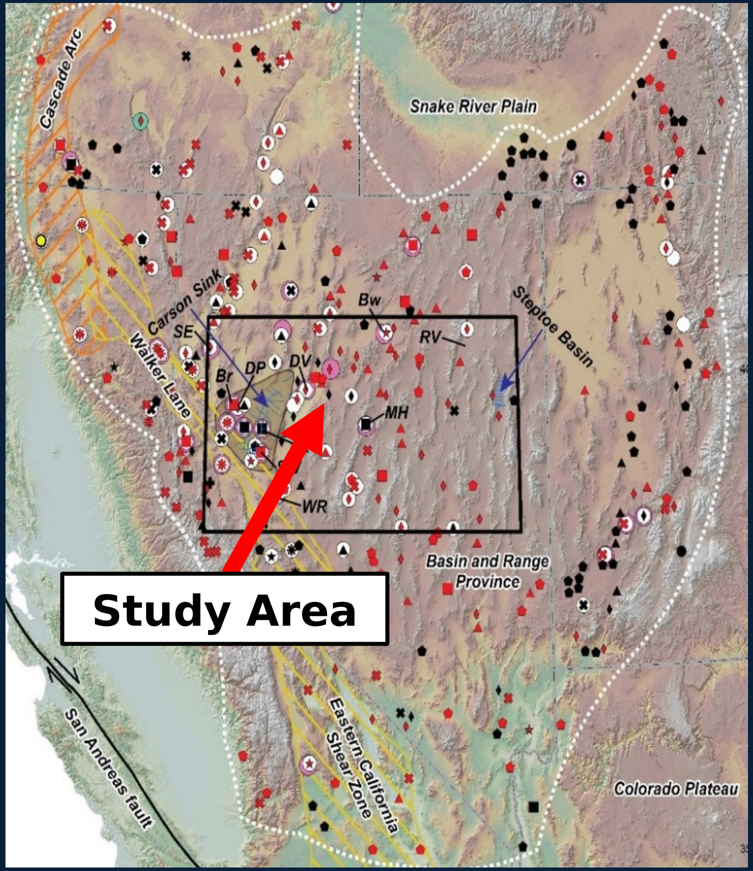


- 40% of known systems are blind
- Estimated 75% of all systems are blind
- Significant drilling / economic risk



- No surface expression
- Need to look for evidence elsewhere

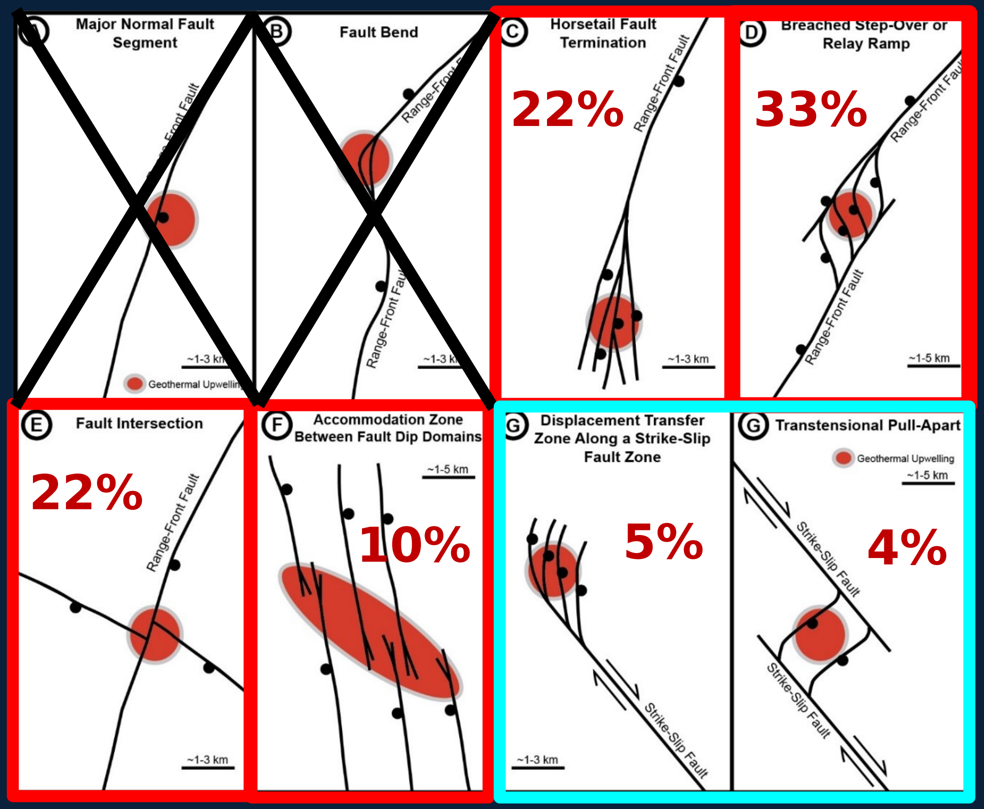
Synthesis of Multiple Parameters



Relevant Geological and Geophysical parameters

- Fault patterns and structural setting
- Age of faulting (lidar data)
- Fault slip rate
- Regional strain rate
- Slip and dilation tendency of faults
- Temperatures of springs and wells (geochemistry)
- Temperature at 3 km depth
- Paleo-geothermal features
- Temperatures at 2 meters depth
- Earthquake density
- Gravity data – horizontal gradient
- Magnetic data
- MT data

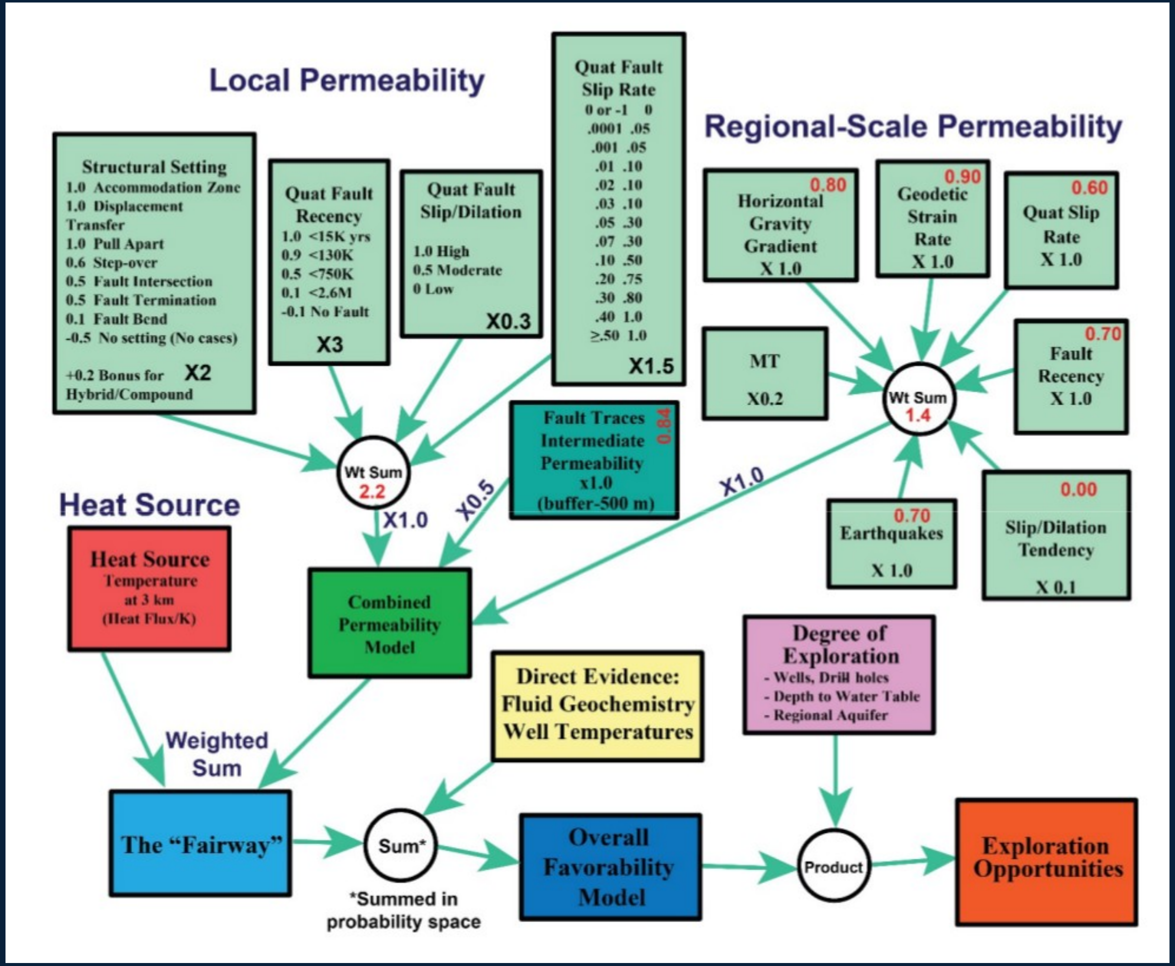
Favorable Structural Settings



Survey of Structure and Geothermal Systems

- 450 systems analyzed; ~250 cataloged
- Most fields not on mid-segments of major faults
- Most on less conspicuous Quaternary normal faults
- Higher temp systems generally on faults <750 ka
- Hybrid settings most productive

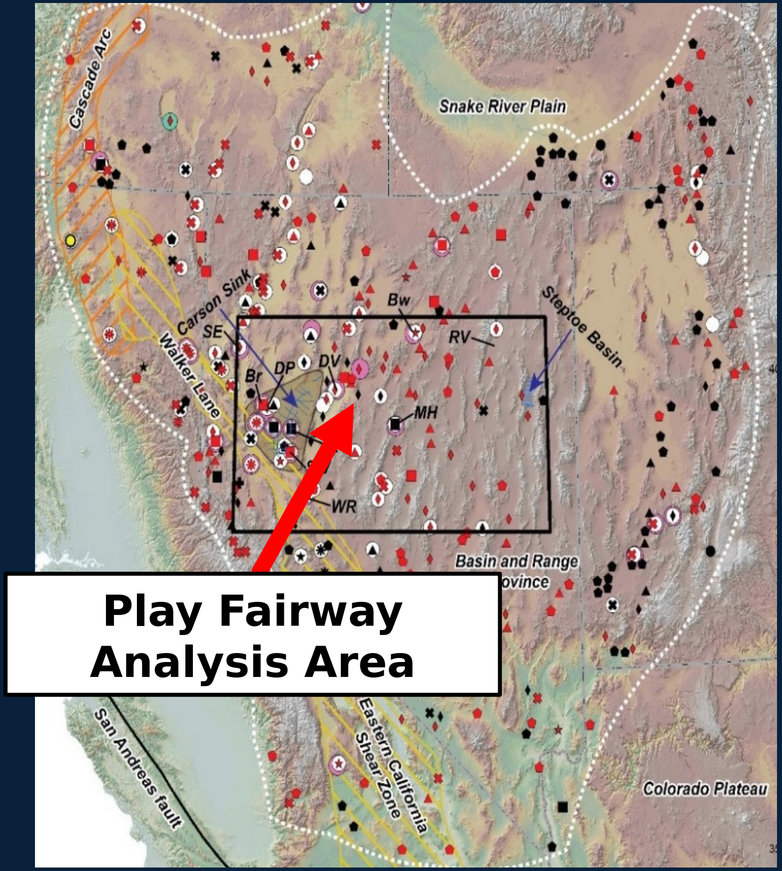
Play Fairway Analysis



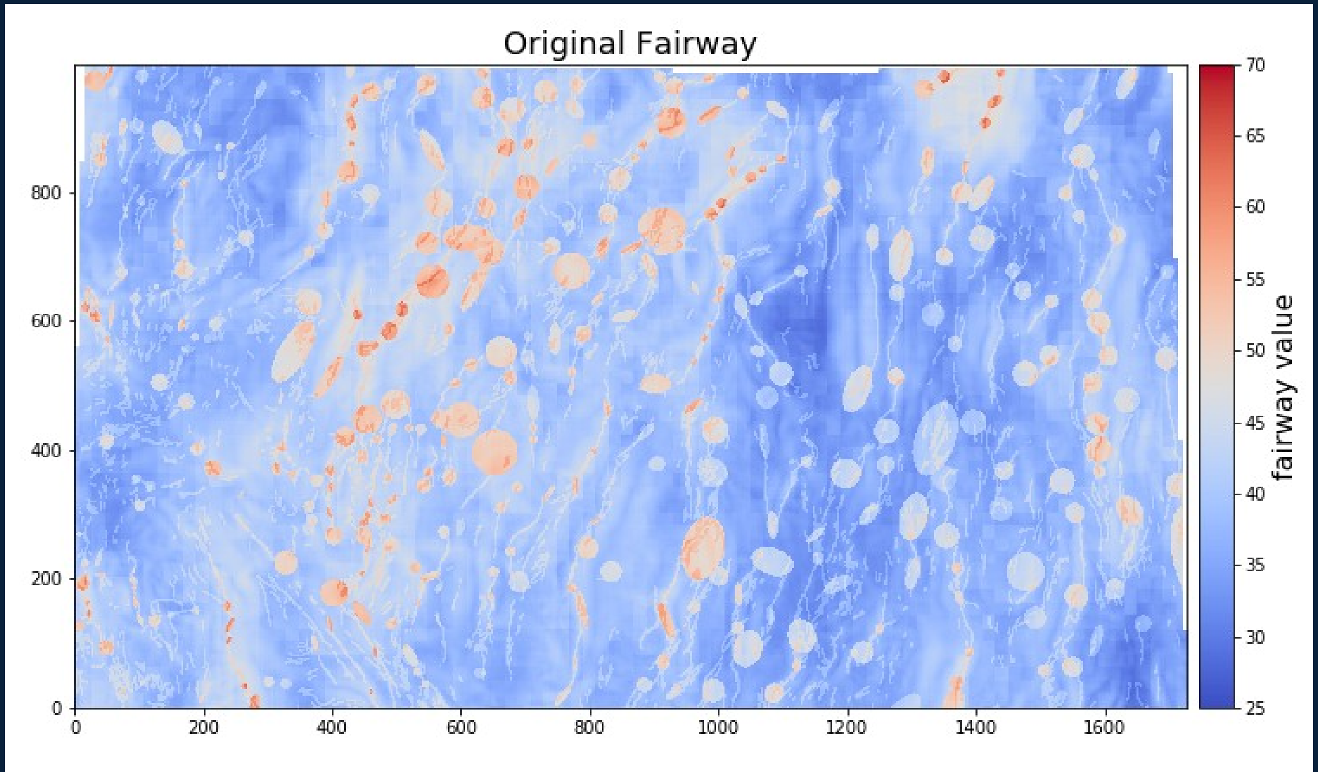
Nevada Geothermal Play Fairway Project

- Expert-derived workflow incorporating:
  - geology and geophysics parameters
  - permeability
  - sources of heat
- Constrained by known geothermal systems
- "Weights of Evidence" statistical analysis to derive sensitivities

Play Fairway Analysis

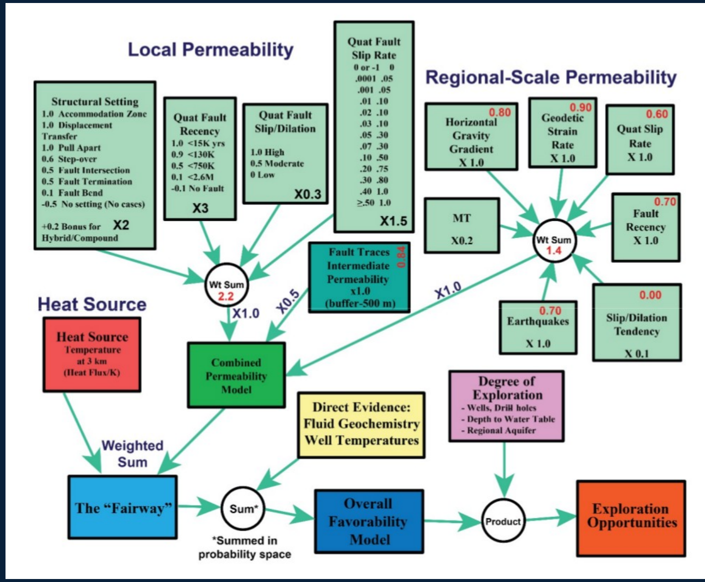


Nevada Geothermal Play Fairway Project

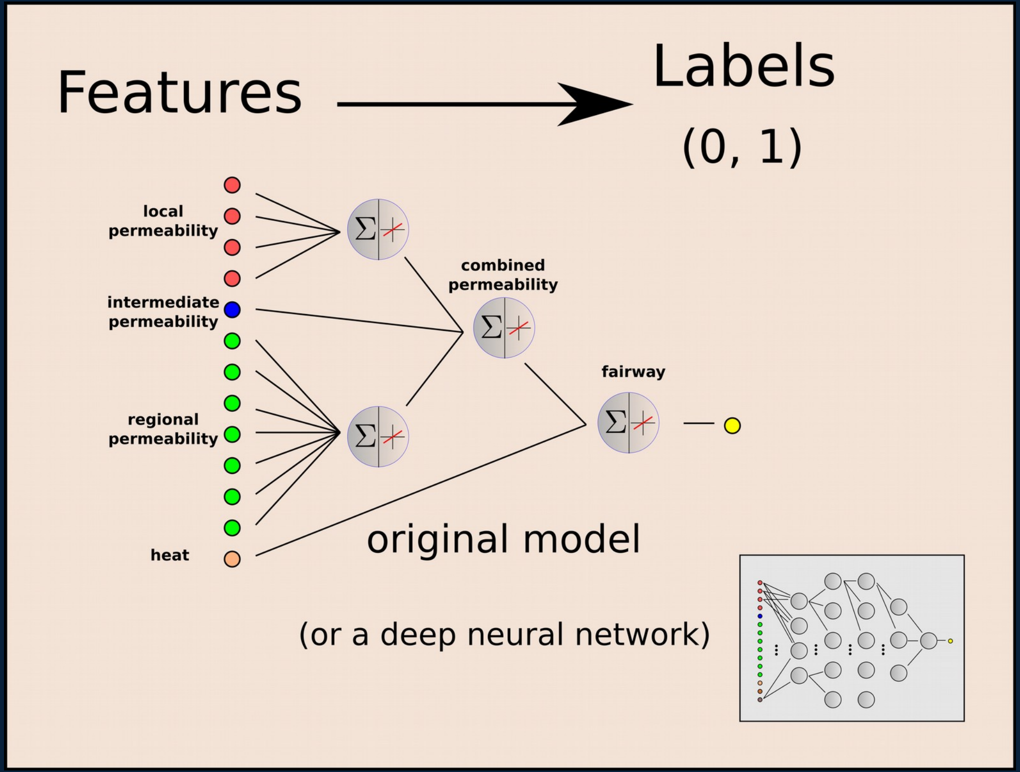


Later drilling suggests this has some predictive power!

Define a Supervised Learning Problem



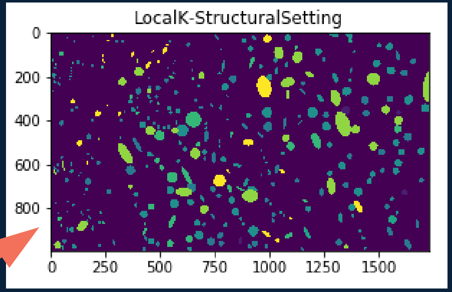
we redraw the PFA workflow and
find a highly-engineered neural network

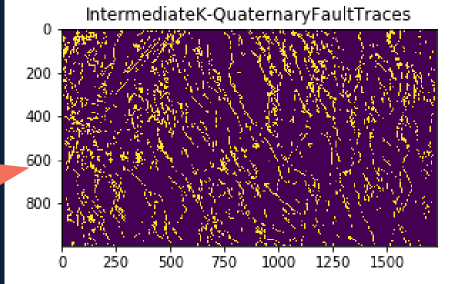


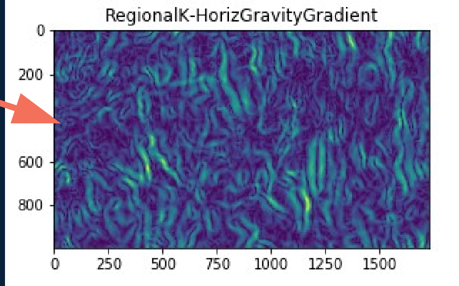
We wish to have unbiased prediction of exploration opportunity as a probability

Features and Labels

```
newfeatureNames = \
['FID',
 'pointid',
 'row',
 'column',
 'NAME',
 'Distance',
 'TrainCode',
 'NullInfo',

 'LocalK-StructuralSetting',
 'LocalK-QuaternaryFaultRecency',
 'LocalK-QuaternaryFaultSlipDilation',
 'LocalK-QuaternaryFaultSlipRate',

 'IntermediateK-QuaternaryFaultTraces',

 'RegionalK-HorizGravityGradient',
 'RegionalK-GeodeticStrainRate',
 'RegionalK-QuaternarySlipRate',
 'RegionalK-FaultRecency',
 'RegionalK-FaultSlipDilationTendency',
 'RegionalK-Earthquakes',

 'HeatSource-T@3km'
]
```







Features:

- Continuous numerical values
- Categorical geologic parameters
- > 1.6 million grid blocks within the study area

Labels:

- Initially 34 positive and zero negative training examples
- Now approximately 100 each positive and negative examples

Issues we have confronted for ML

- Small numbers of examples (initially only 34 positive benchmarks)
  - Can lead to over-fitting
    - Acquire data, data augmentation, regularization, dropout, transfer learning
- Few negative sites (initially none)
  - Imbalanced training data leads to bias
    - Acquire data, simulate negative sites
- Some features are not continuous (categorical)
  - Requires special treatment to prevent bias
    - Embeddings / smoothing / filtering / weighting / reassess data
- Which model architecture and parameters are the best ones?
  - Want as few parameters as possible
    - Optimize using genetic algorithms

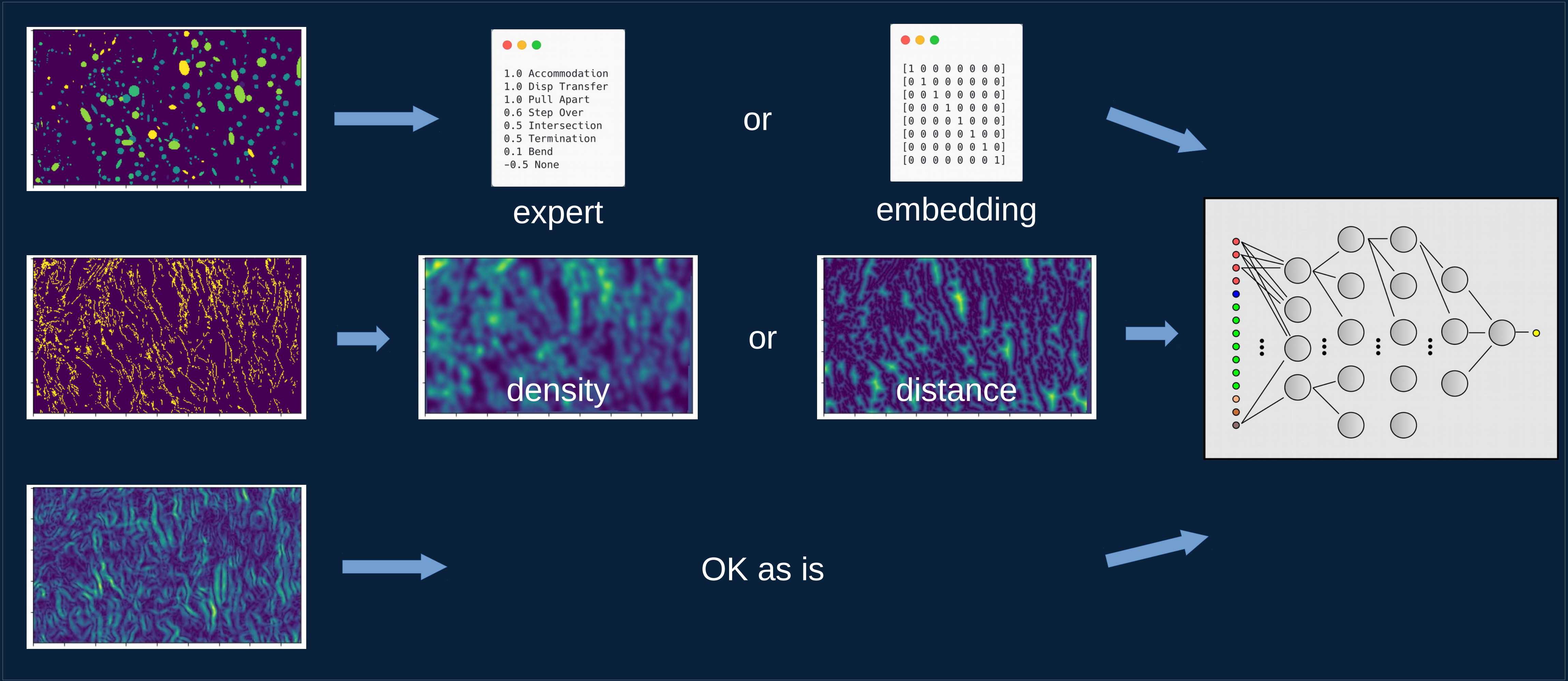
Highlight one problem and a solution

Earth Resources Laboratory

Consideration of categorical data – structural setting features

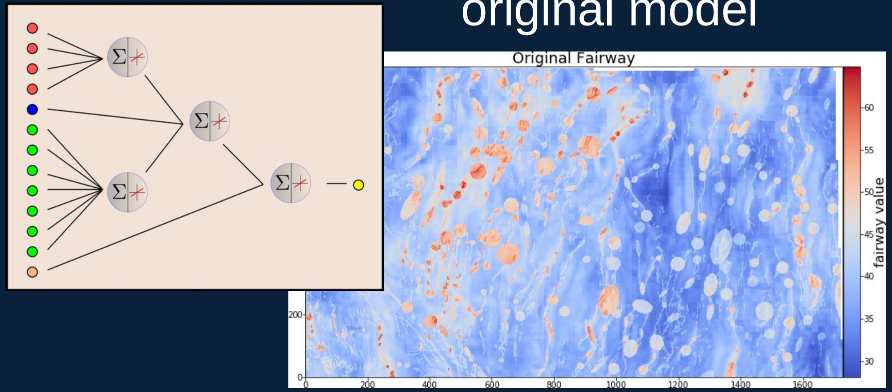
- Categories were pre-ranked by experts on a numerical scale in terms of importance lending possibility of bias.
- Thought to be extremely important features for discrimination, yet they are poorly sampled.
- All positive examples exist where these features are known.
- Direct use of “experts” raw category values leads to poor results as + and- sites are widely separated that it is too easy to divide them … basically an extreme over-fitting problem or becoming stuck in a local minimum results.

Categorical Features

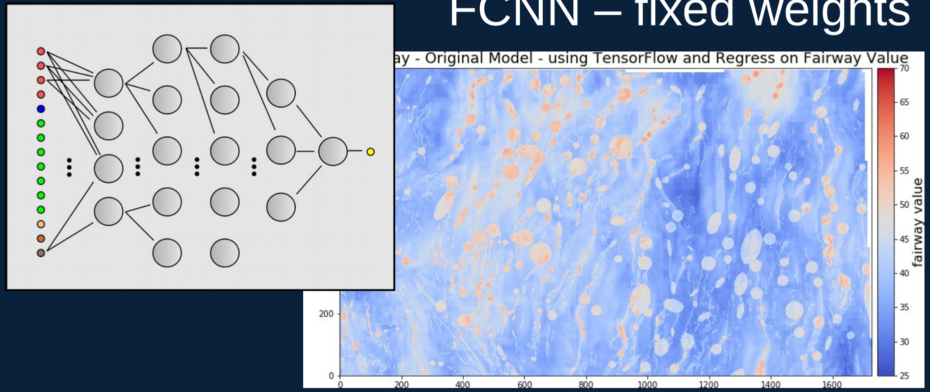


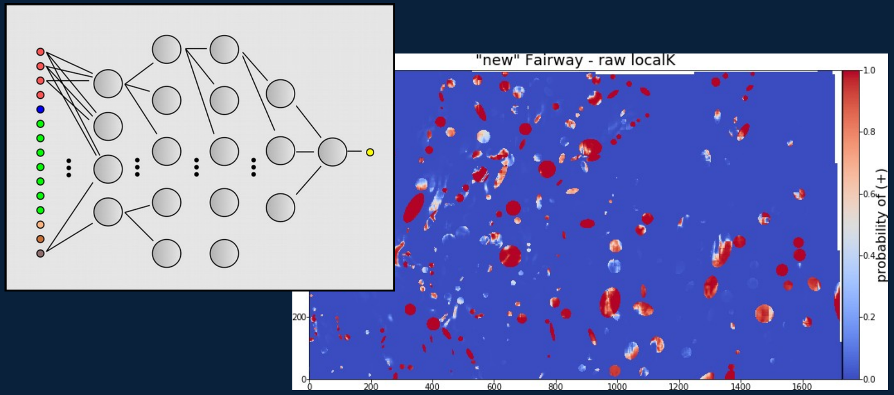
What is the Problem?

original model

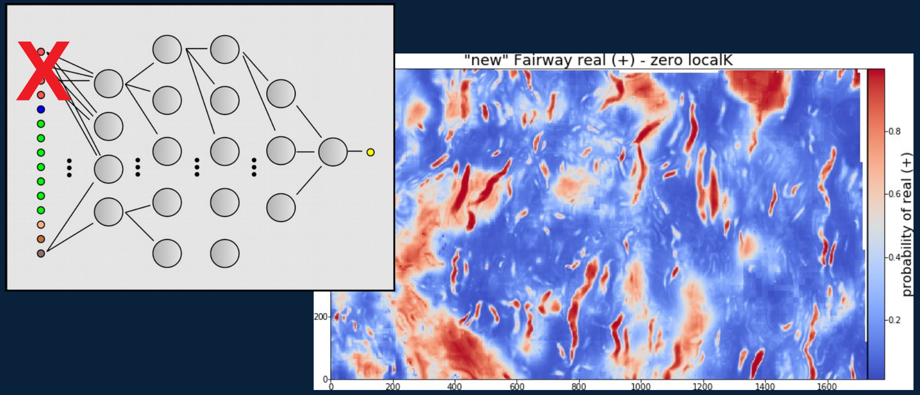


FCNN – fixed weights





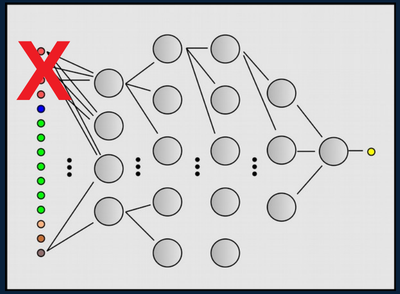
FCNN – trained with categorical



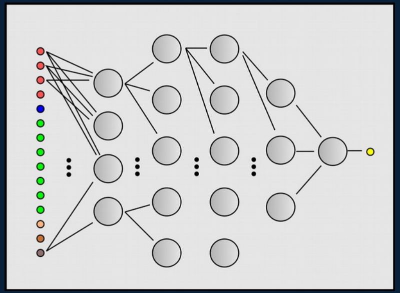
FCNN – trained without categorical

Solution

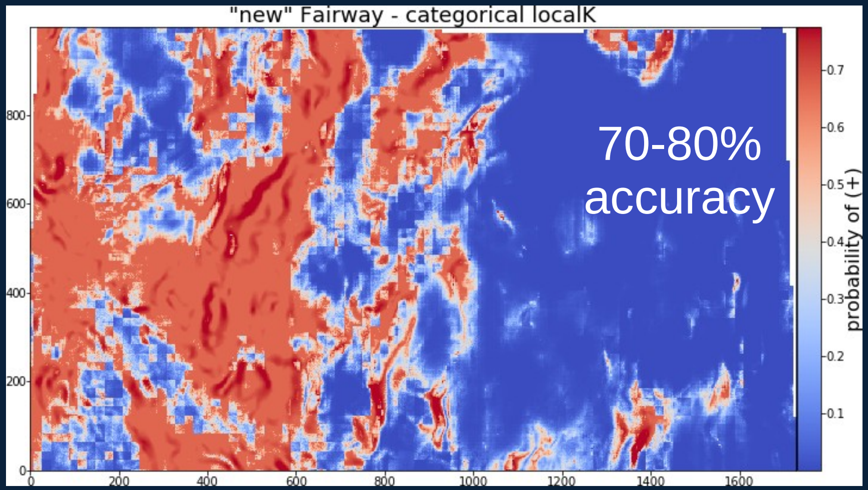
A Transfer Learning Approach

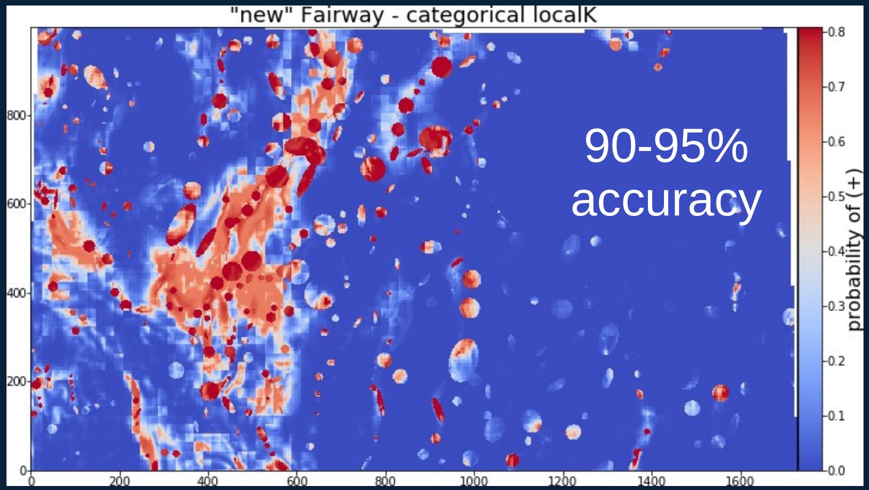


(1) augment data set and pre-train network with categorical features de-emphasized



(2) fine tune same network using all features and the real data set only





A Promising Workflow

1) Augment real (+) and (-) training sites by neighbors on the map

2) Use genetic algorithms to find ‘best’ starting model architecture and hyper-parameters

3) Use best model as basis for GAN and / or noisy student data augmentation to create large simulated data set for transfer learning

4) Pre-train best model on this master data set with de-emphasis of categorical features

5) Fine tune network on all real training sites and all feature sets within our study area

EXTRA SLIDES

Things we have accomplished

In confronting these issues we have:

- explored data augmentation sampling directly from the PFA study area grids using teacher / noisy- student networks
- used generative adversarial networks (GANs) to create “simulated” data sets for training and transfer learning
- considered the extreme of imbalance through outlier/novelty detection approaches
- used genetic algorithms to find “optimal” networks and parameters
- created “simulated” negative sites by sampling the study area at large
- explored various means to use categorical and numerical data together