Approximate Bayesian inference for neural networks: a case study in analysis of spectral data

Natalie Klein
Statistical Sciences Group
Los Alamos National Laboratory
neklein@lanl.gov

DATAWorks 2025
Session 1B: Uncertainty Quantification
23 April 2025

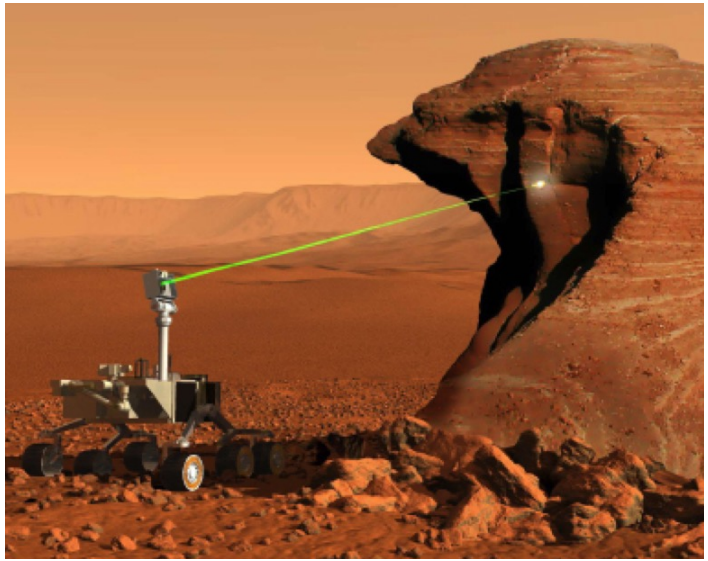
LA-UR-25-23448

Talk outline

- **Introduction**
- Methodology overview
- Evaluation metrics and framework
- Results and findings
- Takeaways and discussion

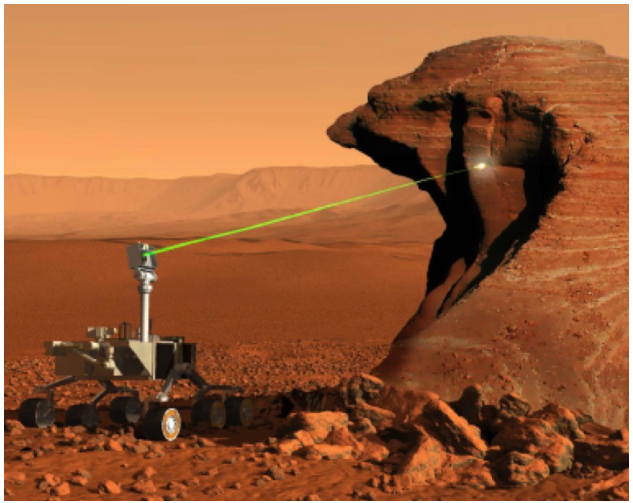
Can we trust neural network predictions in new situations?

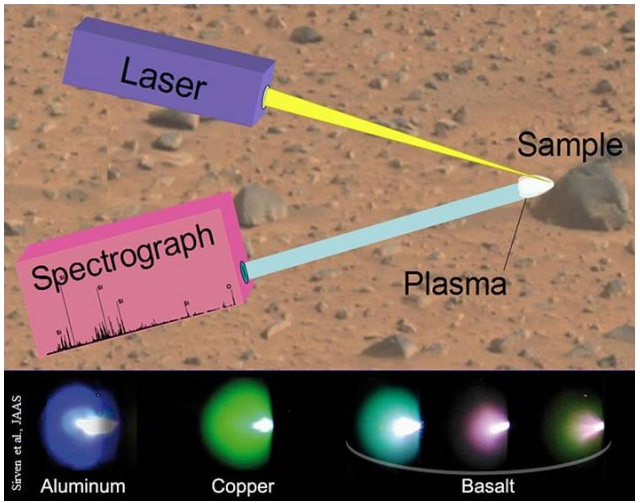
- We consider data from the Mars rover
  - Models are trained on laboratory data
  - Models are applied to data collected on Mars
- Standard neural networks can perform well, but don't accurately quantify uncertainty
- We evaluate three approaches to Bayesian UQ in neural networks
  - How much trust can we place in predictions?
  - We evaluate these methods with respect to *calibration* (more than just accuracy of predictions)



ChemCam spectroscopic data for remote analysis

- ChemCam measures atomic emission spectroscopy
- Elements emit light at characteristic frequencies
- Enables standoff quantification of specific elements

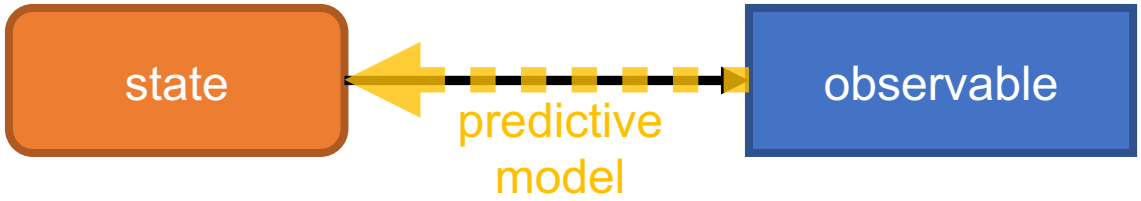




[Images: NASA/JPL-Caltech/LANL; Sirven et al 2007]

Chemical analysis via regression modeling

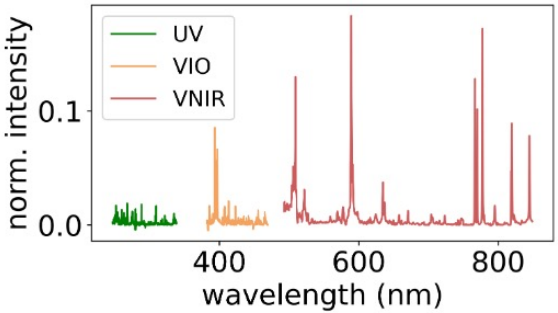
The data analysis task has been traditionally treated as a regression problem.



Chemical Composition

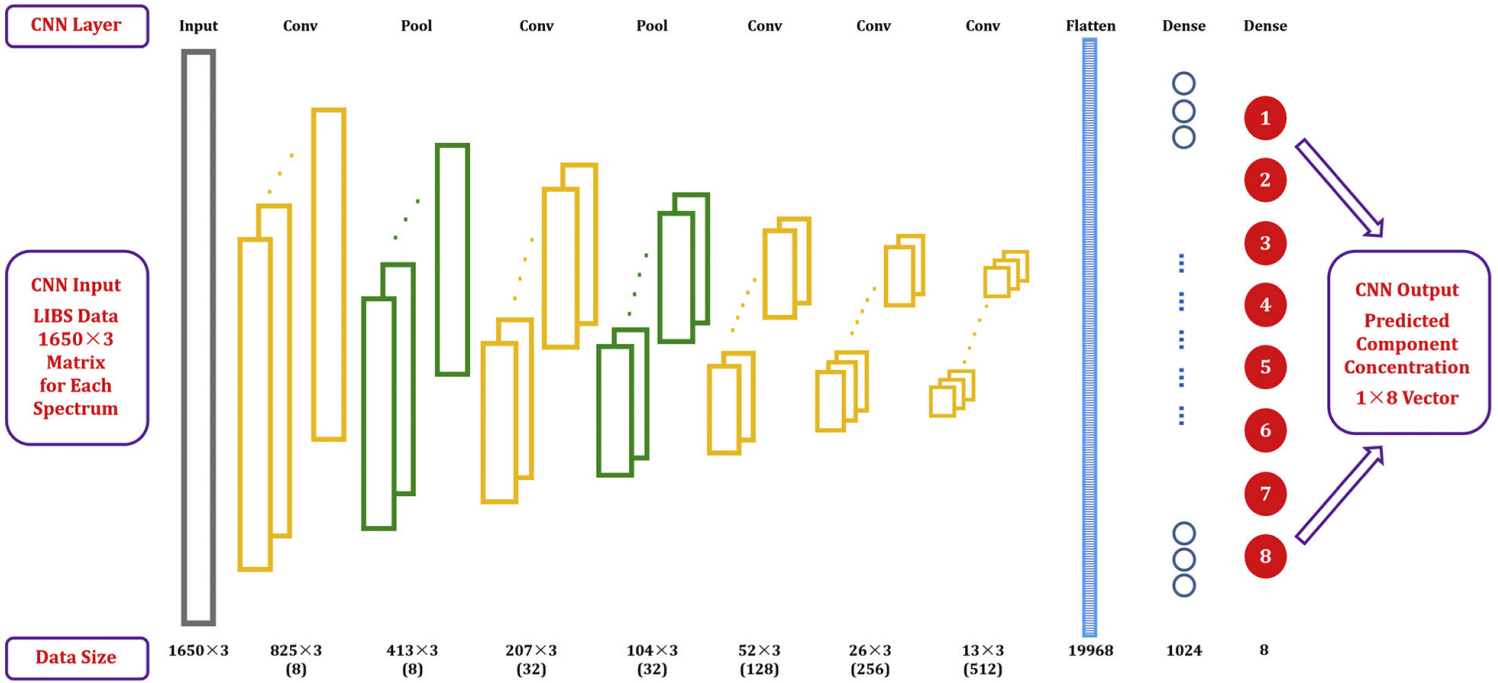
Oxide	Weight %
SiO_2	58.2
MgO	9.8
CaO	5.1
…	…

ChemCam measurement



Neural network regression for ChemCam data

Convolutional neural networks (CNNs) have been shown to work well.

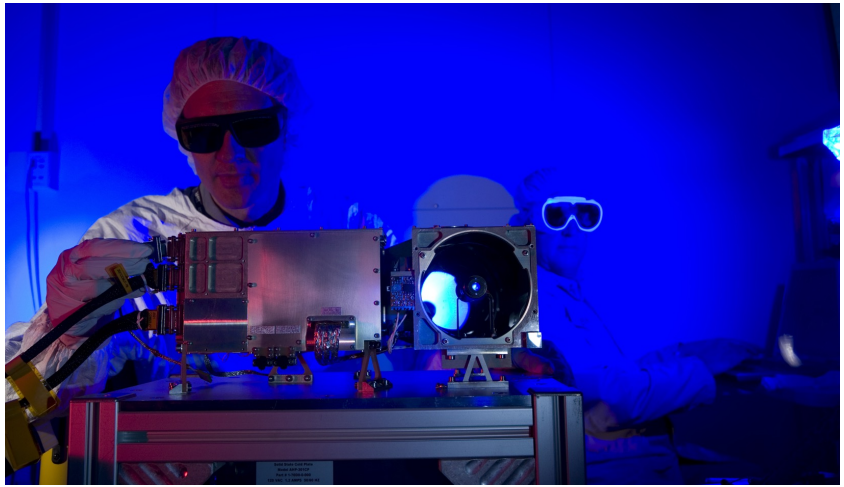


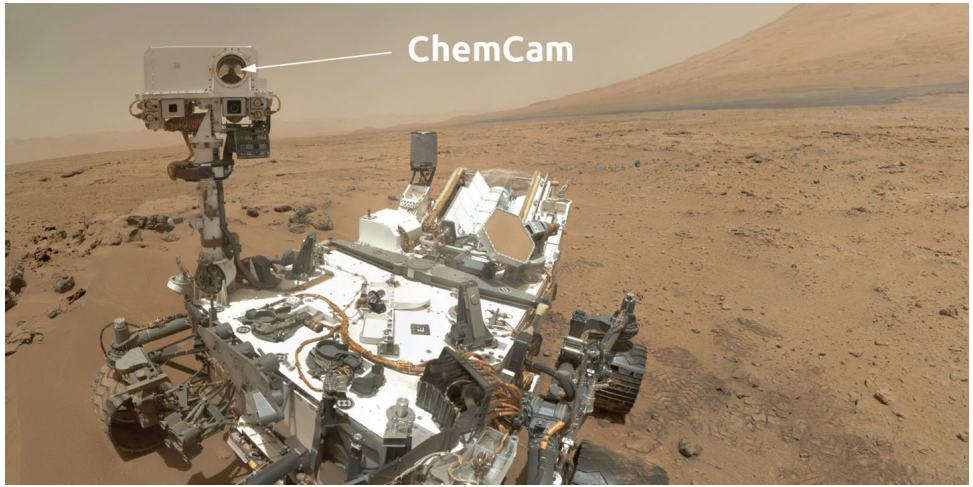
[Li et al 2020]

Quantified uncertainty is crucial for scientific insights

With great power comes great responsibility

- More complex models == overconfidence and overfitting?
- Trained on laboratory data, but deployed on Mars…





[Images: NASA/JPL-Caltech/LANL]

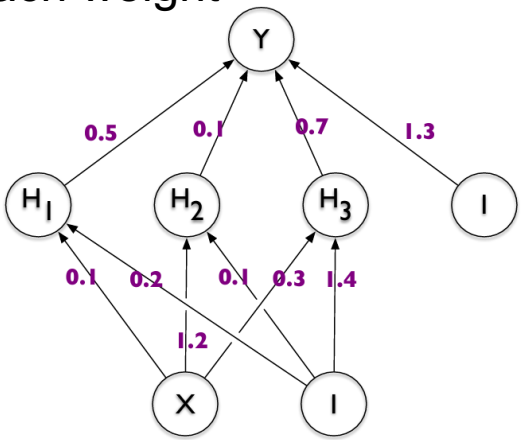
Natalie Klein (neklein@lanl.gov)

Talk outline

- Introduction
- **Methodology overview**
- Evaluation metrics and framework
- Results and findings
- Takeaways and discussion

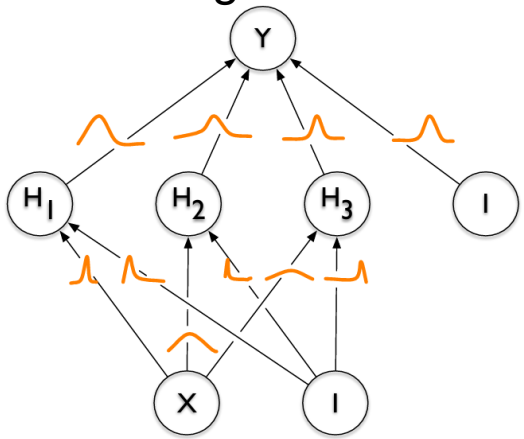
Uncertainty quantification for neural networks

Standard neural network: find a good *single value* for each weight



Result: one prediction for each input

UQ-aware neural network: find a plausible *distribution* for each weight



Result: predictive uncertainty (error bars)

[Figures adapted from Blundell et al 2015]

Bayesian inference for uncertainty quantification

- Bayesian inference gives a probability distribution over possible fitted models, enabling (among other things) predictive uncertainty quantification (UQ)

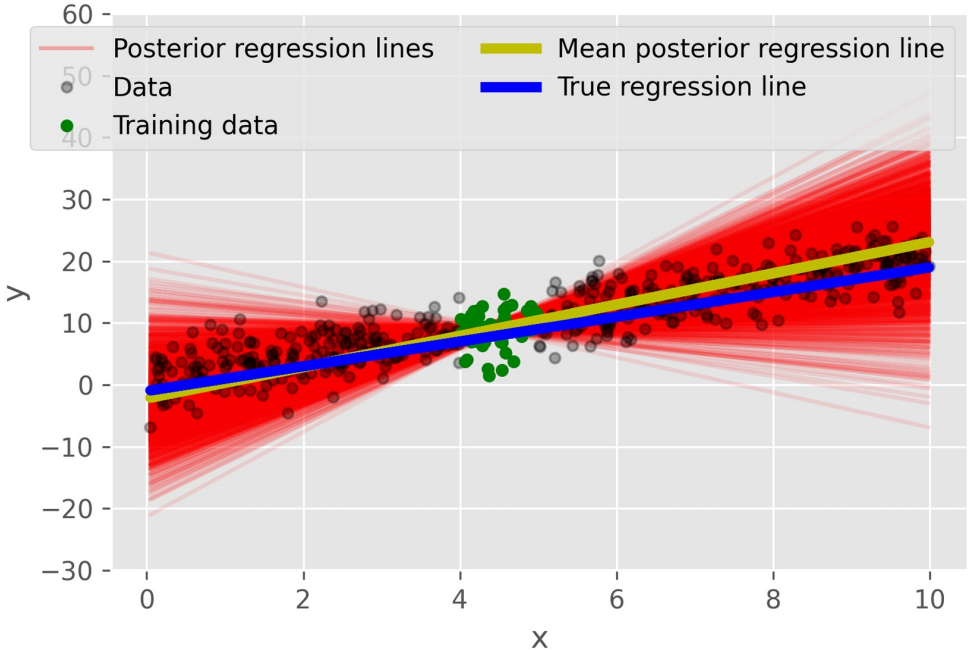
Likelihood	$p(Y \mid X, \theta)$	Describes likelihood of regression output *Y* conditional on inputs *X* and parameters *θ*
Prior	$p(\theta)$	Describes prior probability of parameters *θ*

- The posterior distribution describes the probability over parameters after seeing the data:

$$p(\theta \mid X, Y) = \frac{p(Y \mid X, \theta)p(\theta)}{\int p(Y \mid X, \theta)p(\theta)\, d\theta}$$

Bayesian inference for uncertainty quantification

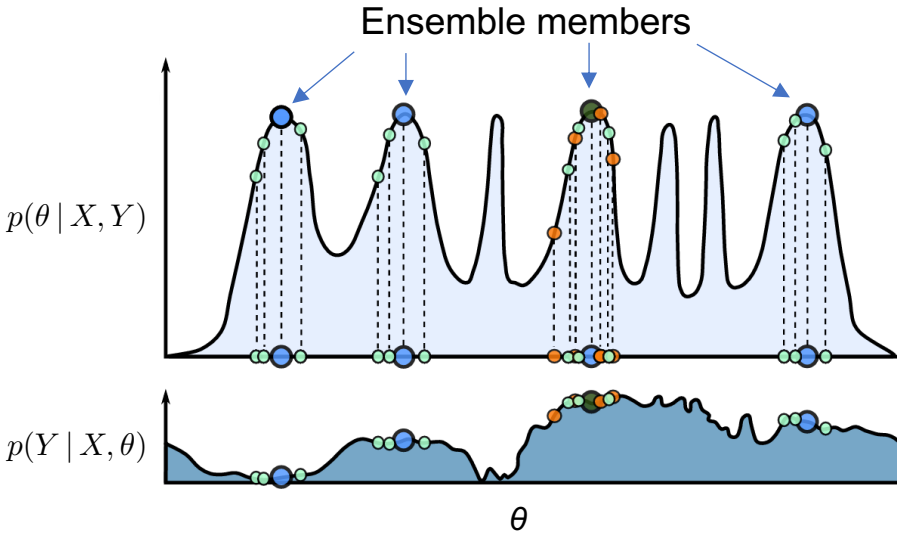
- The posterior predictive distribution gives uncertainty over a new prediction by *sampling* likely parameter settings



[Botsas et al 2022]

Deep ensembles offer a simple, intuitive way to get distributions over neural network weights

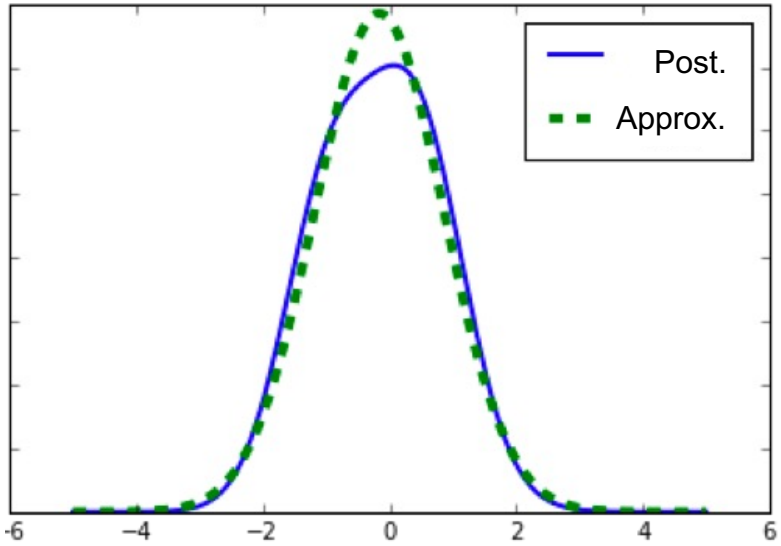
- Key idea: train multiple neural networks from different initial values
- Local maxima should correspond to modes of the Bayesian posterior



Adapted from [Wilson and Izmailov 2020]

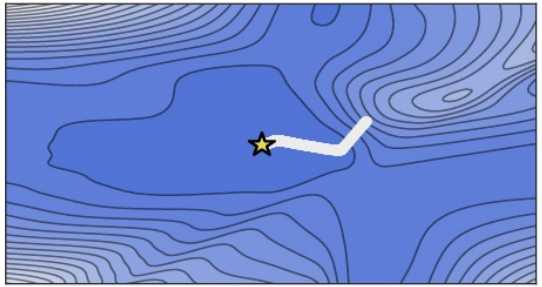
Variational inference optimizes over an approximation to the true Bayesian posterior

- Goal: find an easy distribution (e.g., Gaussian) that approximates the posterior
- Can formulate a tractable objective function to optimize
- Can sample from approximate posterior for UQ

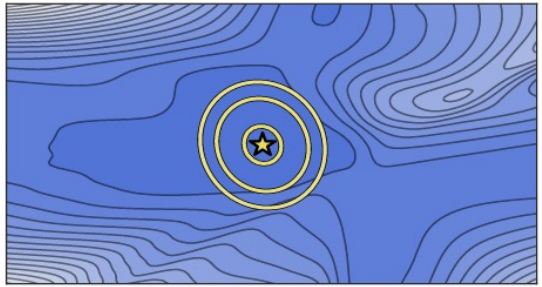


The Laplace approximation uses a Gaussian centered at a good parameter value for UQ

Step 1: optimize loss function (standard neural network training) to find good parameter values



Step 2: center a Gaussian distribution around these parameters for UQ



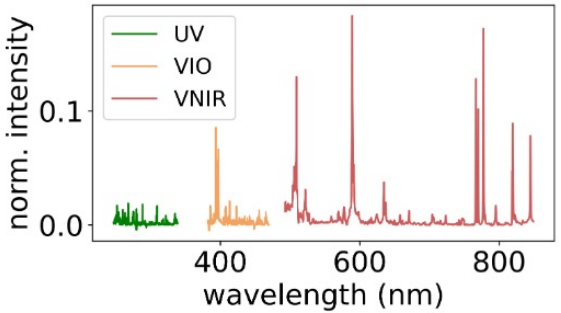
[Figure adapted from Daxberger et al 2021]

Talk outline

- Introduction
- Methodology overview
- **Evaluation metrics and framework**
- Results and findings
- Takeaways and discussion

Evaluating regression models with UQ

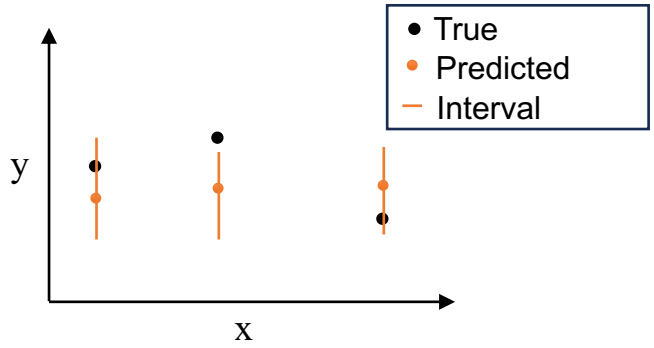
- Performance of regression models is often evaluated with accuracy measures (such as root mean squared error)
- How to evaluate the uncertainty quantification?



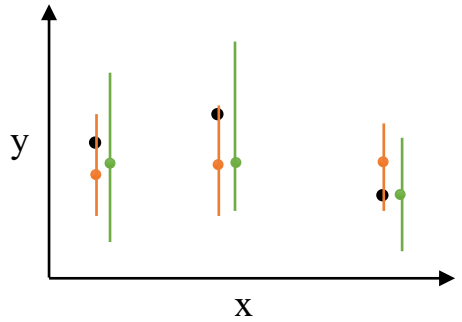
Oxide	Actual Weight %	Predicted Weight %
SiO_2	58.2	59.1
MgO	9.8	10.1
CaO	5.1	4.6
…	…	…

Evaluating predictive uncertainty intervals

- Coverage: how often does the credible prediction interval contain the true value?
  - Assessed on held-out test data



- BUT: intervals could be made infinitely wide to achieve coverage
- Tighter intervals are more accurate (if coverage is maintained)



Sensitivity and robustness

- To evaluate sensitivity to hyperparameter settings, we use **interval score**

If…	**Score is…**
True value inside interval	Interval width
True value outside interval	Interval width *plus* penalty (based on how far below or above interval)

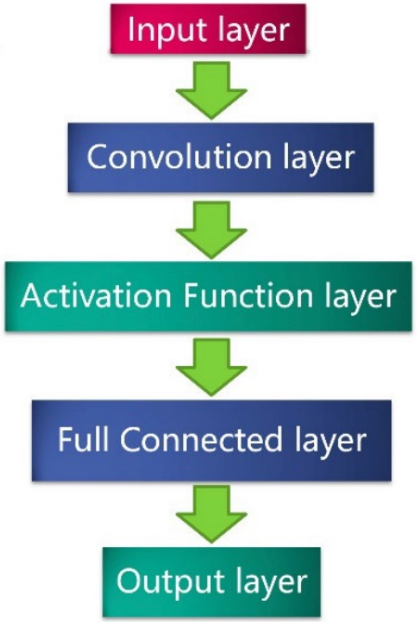
- Lower values are better
- We estimate sensitivity of interval score to variational inference hyperparameters

Talk outline

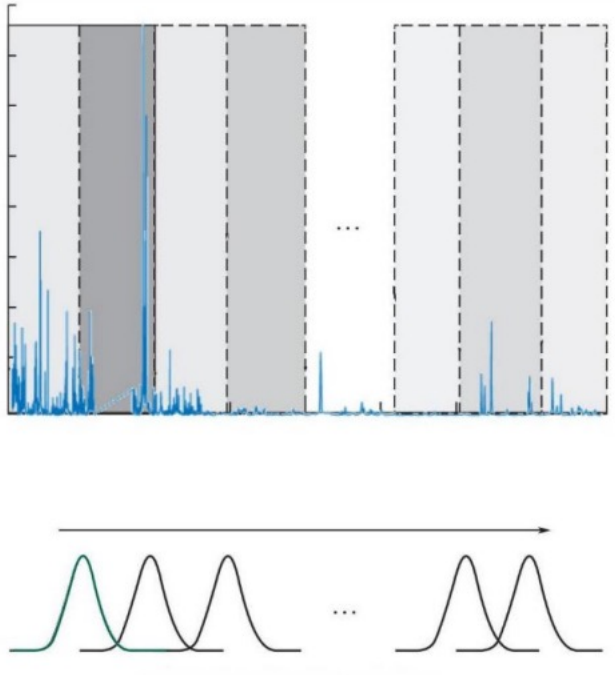
- Introduction
- Methodology overview
- Evaluation metrics and framework
- **Results and findings**
- Takeaways and discussion

Convolutional neural network (CNN) for ChemCam

- First, we developed a non-Bayesian CNN architecture and validated it on ChemCam data



- Convolution layers learn filters to extract features from spectra
- Fully connected layers map features to the output (oxide weight percent)

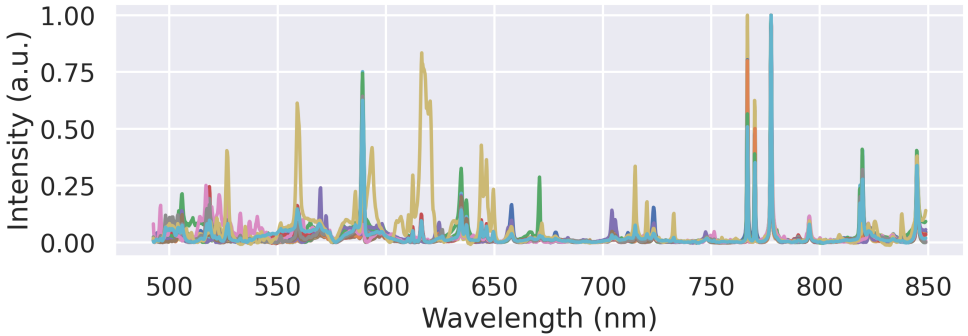


[Yu and Yao 2023]

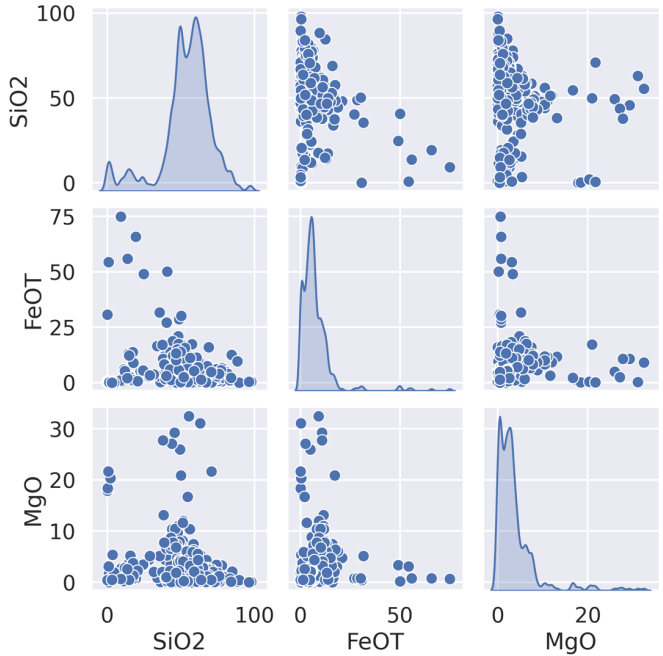
ChemCam data set details

- All data from NASA PDS*
- Spectral measurements at 5,605 wavelengths

	Number of targets	Total spectra
Training	452	90,000
Validation	55	11,000
Test	55	11,000



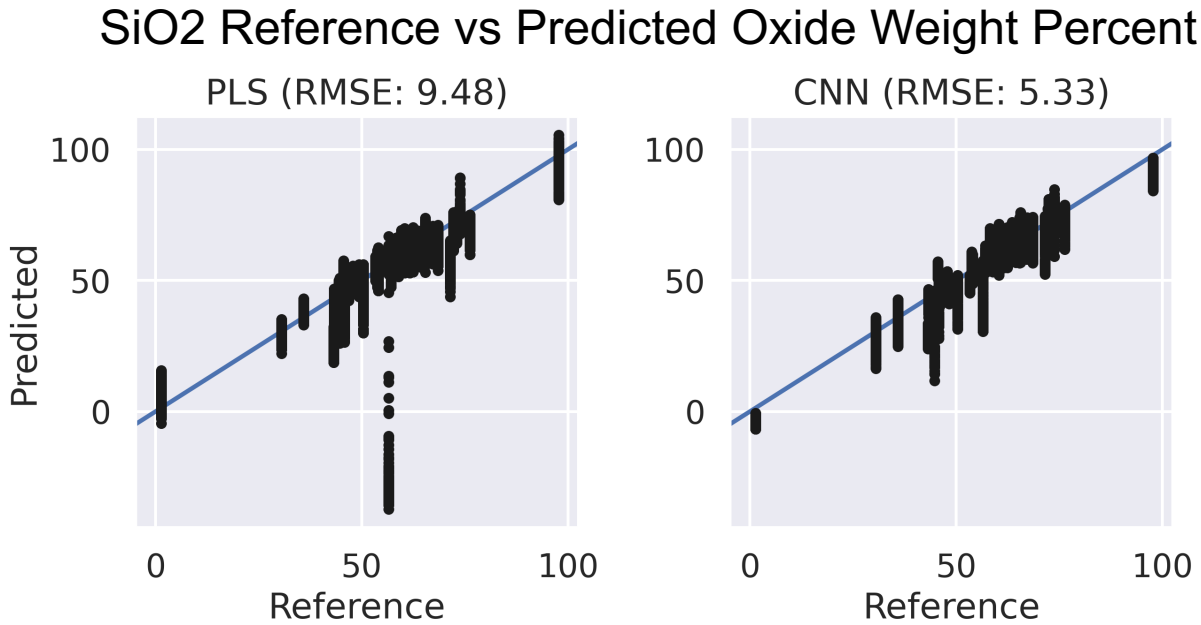
Chemical composition
(oxide weight percent)



* https://pds-geosciences.wustl.edu/msl/msl-m-chemcam-libs-4_5-rdr-v1/mslccm_1xxx/

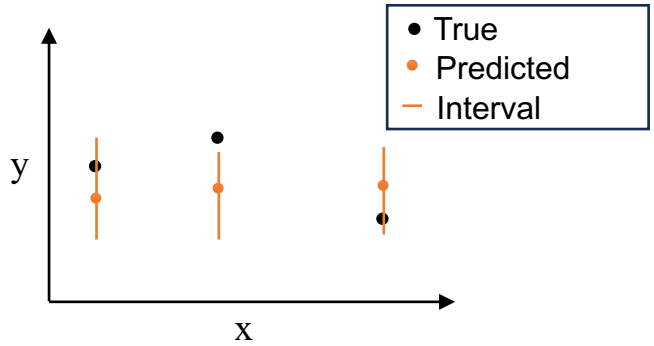
CNN predictive accuracy

- Our CNN architecture (non-Bayesian) can predict more accurately on a held-out test set than a baseline linear method (partial least squares)



Predictive uncertainty interval coverage

- Coverage: how often does the credible prediction interval contain the true value?
  - Assessed on held-out test data

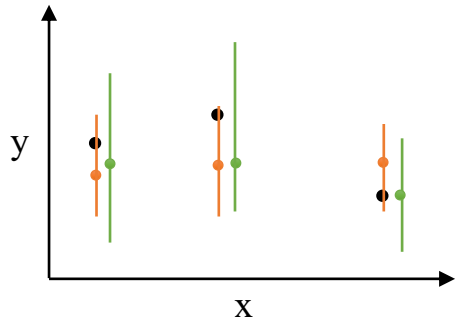


Observed coverage of 95% credible prediction interval

	SiO_2	TiO_2	Al_2O_3	FeOT	MnO	MgO	CaO	Na_2O	K_2O
Deep Ensemble	0.89	0.90	0.87	0.82	1.00	0.91	0.96	0.91	0.93
Laplace	0.94	0.92	0.93	0.89	0.95	0.97	0.95	0.93	0.93
VI	0.93	0.93	0.84	0.88	1.00	0.93	0.97	0.91	0.85

Predictive uncertainty interval width

- Intervals could be made infinitely wide to achieve coverage
- Tighter intervals are more accurate (if coverage is maintained)



Average width of 95% credible prediction interval (oxide weight percent)

	SiO_2	TiO_2	Al_2O_3	FeOT	MnO	MgO	CaO	Na_2O	K_2O
Deep Ensemble	10.89	**0.68**	5.22	**3.76**	**2.44**	2.86	5.61	1.54	1.57
Laplace	**10.36**	1.18	**4.33**	5.02	4.39	**2.85**	**4.79**	**1.48**	**1.39**
VI	11.49	1.19	4.80	5.20	4.43	3.02	5.00	1.60	1.51

Application to Mars data

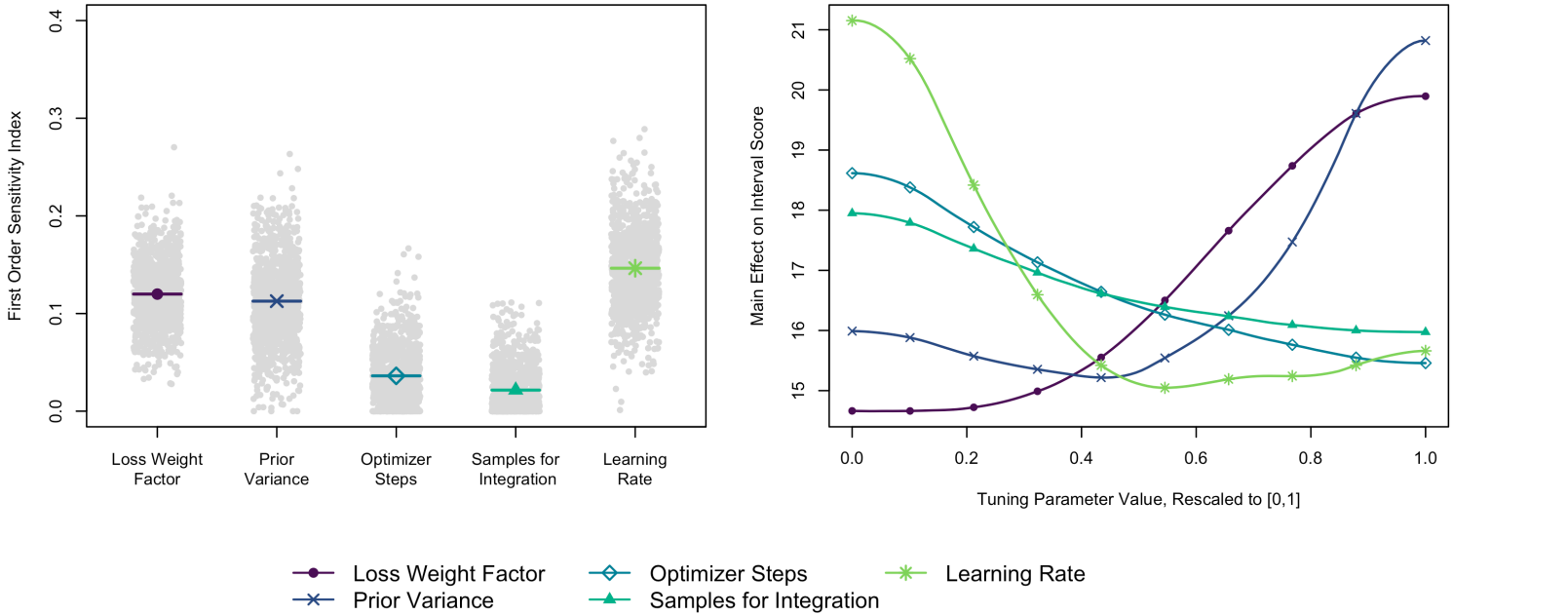
- Mars data may exhibit distribution shift due to variable distance, plasma temperature, viewing angle, etc.
- **How do the uncertainty intervals change on Mars compared to Earth data?**
- We computed relative percent change from Earth to Mars interval width:

$$\frac{|\text{Width}_{Earth} - \text{Width}_{Mars}|}{\text{Width}_{Earth}} \cdot 100$$

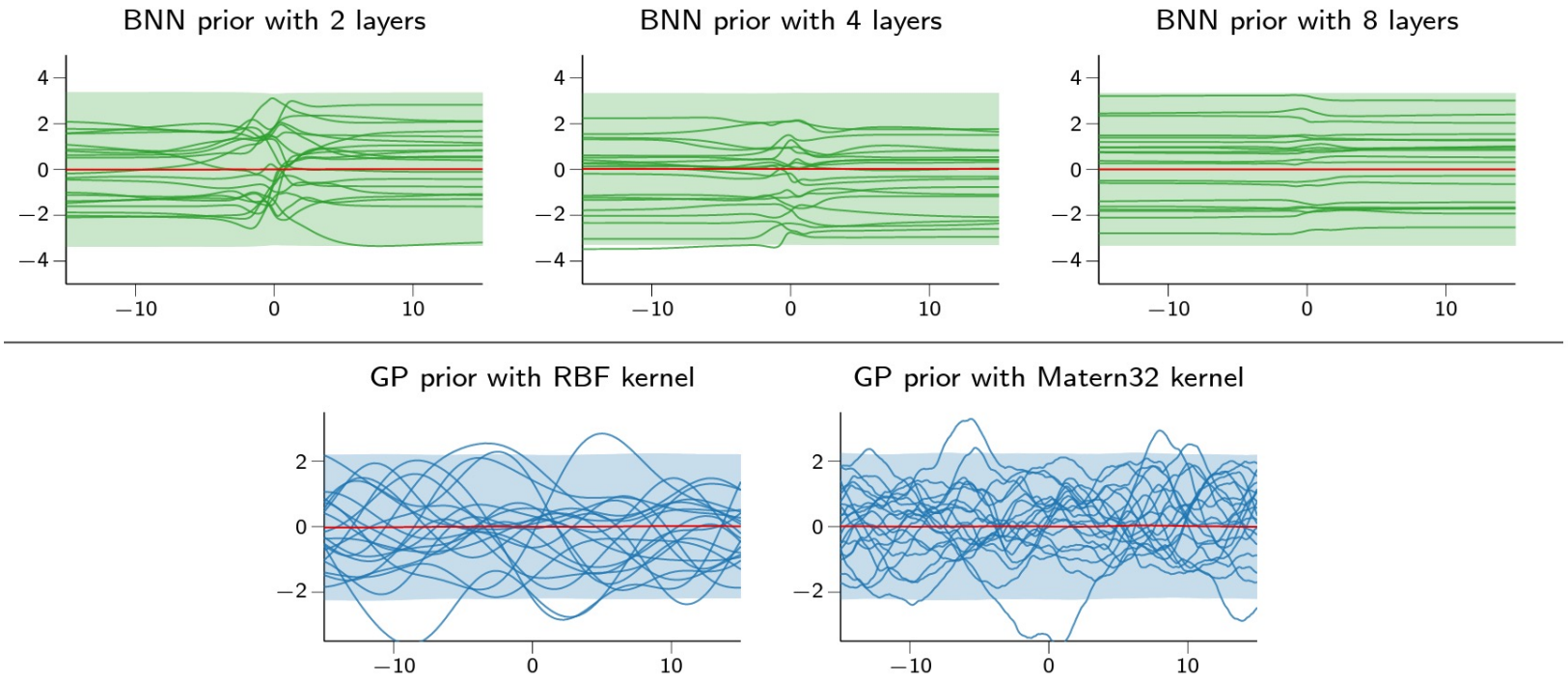
	SiO_2	TiO_2	Al_2O_3	**FeOT**	MnO	**MgO**	CaO	**Na_2O**	K_2O
Deep Ensemble	7.56%	16.94%	2.17%	**148.87%**	13.32	**62.30%**	29.70%	**21.45%**	2.78%
Laplace	0.20%	0.32%	0.58%	**0.52%**	0.55	**0.5%**	0.39%	**0.57%**	0.54%
VI	6.67%	13.50%	6.13%	**42.51%**	0.01	**60.83%**	13.43%	**0.85%**	1.02%

Sensitivity of variational inference to hyperparameters

The quality of variational inference is primarily impacted by prior settings, loss weighting, and learning rate.



Impact of priors in neural networks



[Tran et al 2022]

How to select the prior?

- Actual prior knowledge (not possible to elicit)
- “Uninformative” (not computable for NNs, or not actually uninformative)

Uniform prior
(Bayes and Laplace, 18th and 19th century)

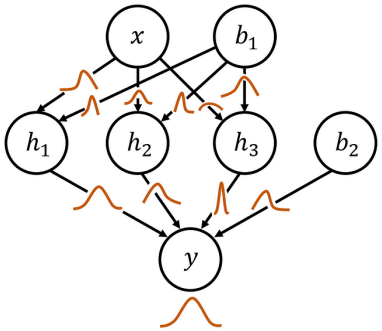
Jeffreys’ prior
(Jeffreys, 1946)

Maximum entropy prior
(Jaynes, 1968)

Reference prior
(Bernardo, 1979)

https://www.statlect.com/fundamentals-of-statistics/uninformative-prior

Bayesian Neural Network



Standard “uninformative” priors

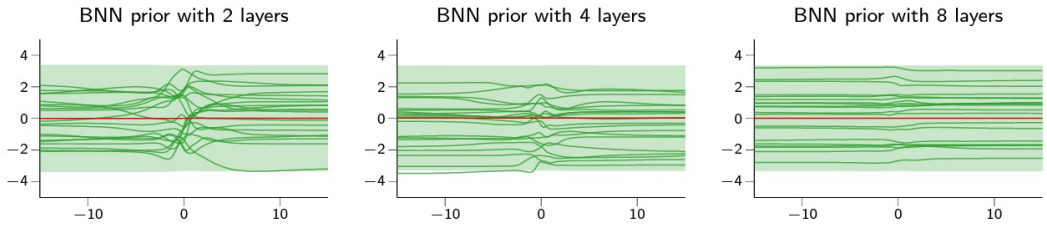


Image from [Tran et al 2022]

- Data-driven prior selection

Talk outline

- Introduction
- Methodology overview
- Evaluation metrics and framework
- Results and findings
- **Takeaways and discussion**

Challenges encountered in this analysis

- Data limitations
  - While we have many example spectra, we have a limited number of targets
  - Spectra can be noisy
- Computational complexity
  - Ensemble: can be parallelized, but requires retraining CNN repeatedly
  - VI: for mean-field VI, twice as many parameters and more expensive loss function
  - Laplace: we used subnetwork (fixed some parameters), otherwise gets very costly
- Sensitivity to hyperparameters
  - **VI and Laplace showed high sensitivity to prior, model hyperparameters**
  - Laplace can be numerically unstable

Summary

Uncertainty quantification for neural networks is hard.

Bayesian methods are attractive, but…

- Challenging computationally
- Difficult to choose model and prior
- Often see poor performance in practice

We are investigating sensitivity to the prior and alternate inference methods.

- Initial code base (in progress): www.github.com/lanl/multiverse

Approximate Bayesian inference for neural networks: a case study in analysis of spectral data

Natalie Klein
Statistical Sciences Group
Los Alamos National Laboratory
neklein@lanl.gov

23 April 2025

Postdocs
- Weizhi Li
- Scott Koermer

Students
- Giosue Migliorini (UCI)
- Tom Winckelman (TAMU)
- Mark Hinds (Georgia Tech)

References

- Blundell, C., Cornebise, J., Kavukcuoglu, K., & Wierstra, D. (2015, June). Weight uncertainty in neural network. In International Conference on Machine Learning (pp. 1613-1622). PMLR.
- Botsas, T., Mason, L. R., Matar, O. K., & Pan, I. (2022). Rule-based Evolutionary Bayesian Learning. arXiv preprint arXiv:2202.13778.
- Daxberger, E., Kristiadi, A., Immer, A., Eschenhagen, R., Bauer, M., & Hennig, P. (2021). Laplace redux-effortless Bayesian deep learning. Advances in Neural Information Processing Systems, 34, 20089-20103.
- Li, L. N., Liu, X. F., Xu, W. M., Wang, J. Y., & Shu, R. (2020). A laser-induced breakdown spectroscopy multi-component quantitative analytical method based on a deep convolutional neural network. Spectrochimica Acta Part B: Atomic Spectroscopy, 169, 105850.
- Tran, B. H., Rossi, S., Milios, D., & Filippone, M. (2022). All you need is a good functional prior for Bayesian deep learning. The Journal of Machine Learning Research, 23(1), 3210-3265
- Wilson, A. G., & Izmailov, P. (2020). Bayesian deep learning and a probabilistic perspective of generalization. Advances in Neural Information Processing Systems, 33, 4697-4708.
- Yu, Y., & Yao, M. (2023). When Convolutional Neural Networks Meet Laser-Induced Breakdown Spectroscopy: End-to-End Quantitative Analysis Modeling of ChemCam Spectral Data for Major Elements Based on Ensemble Convolutional Neural Networks. Remote Sensing, 15(13), 3422.