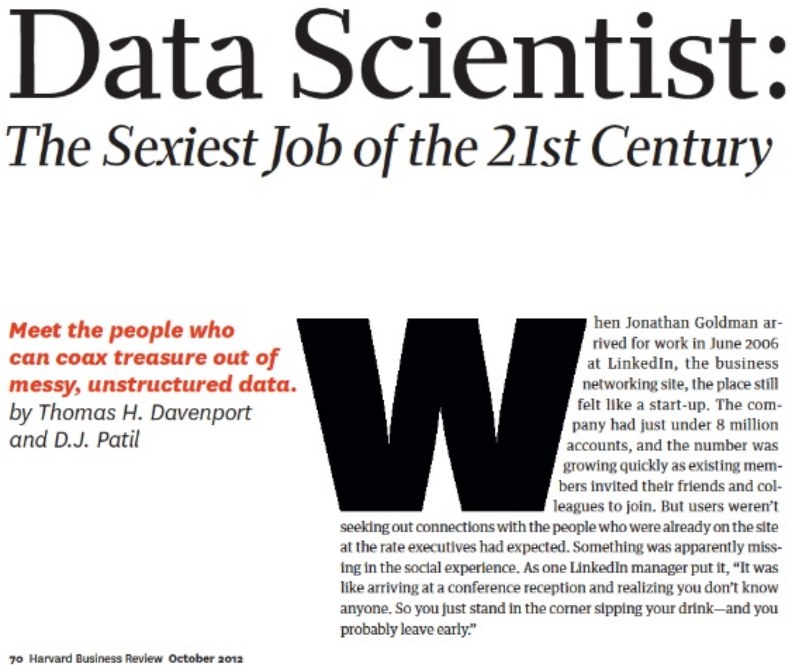
Data Science and Machine learning

Applications to data processing task, signal identification, classification, surrogate modelling and inversion.

Jobs for everyone



Data Scientist:

The Sexiest Job of the 21st Century

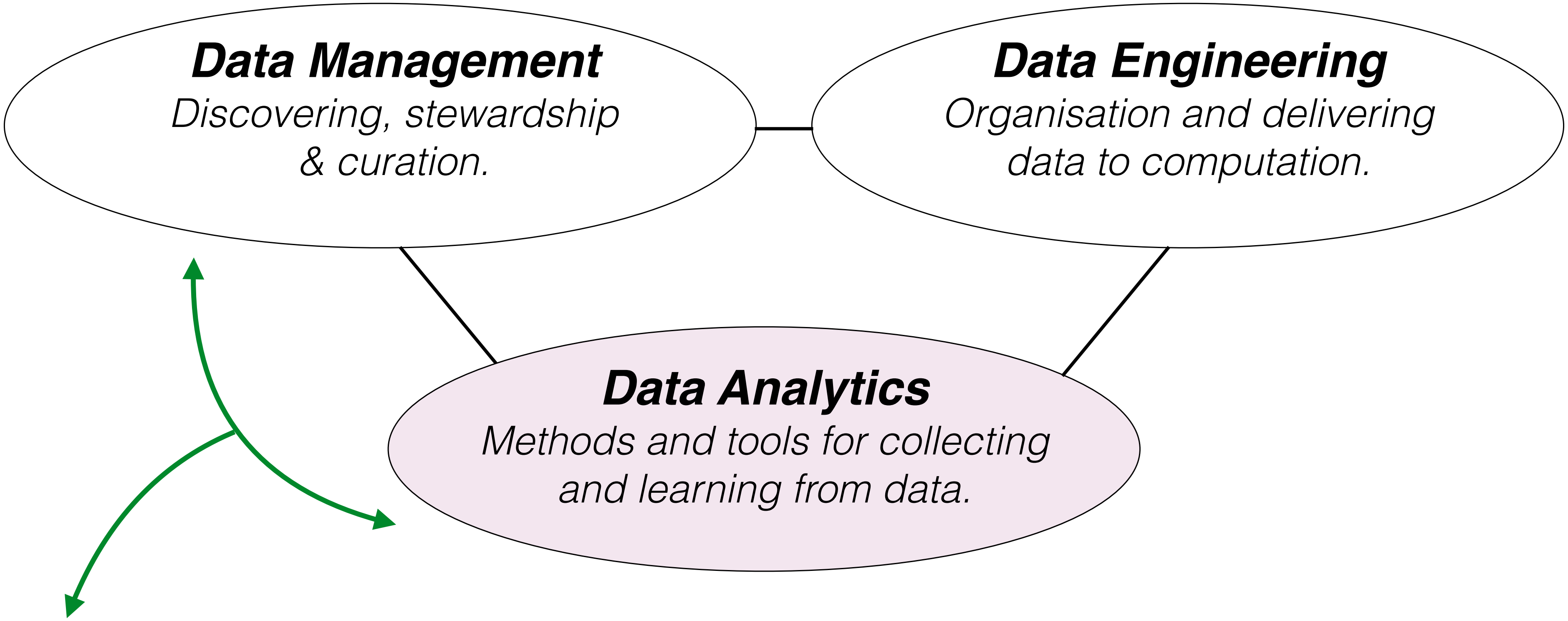
Meet the people who can coax treasure out of messy, unstructured data.

by Thomas H. Davenport and D.J. Patil

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—and you probably leave early."

70 Harvard Business Review **October 2012**

The three pillars of Data Science



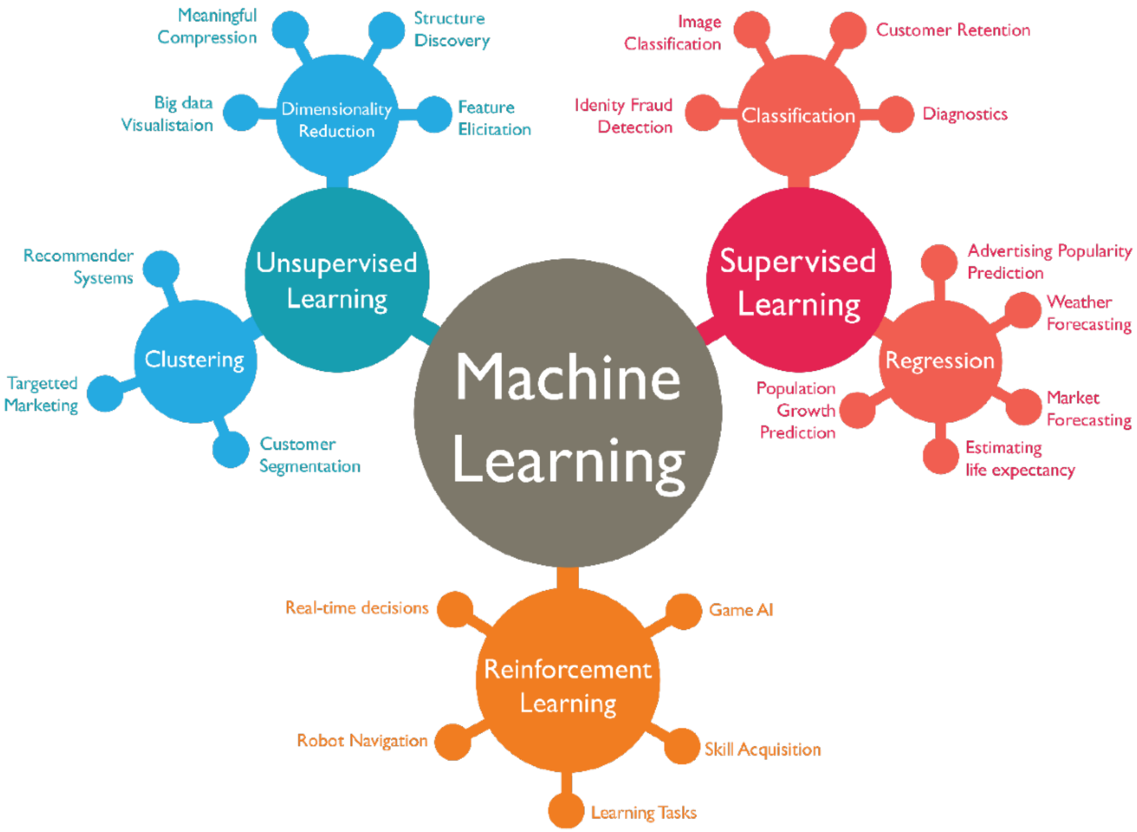
Data Management
Discovering, stewardship & curation.

Data Engineering
Organisation and delivering data to computation.

Data Analytics
Methods and tools for collecting and learning from data.

Geoscientists have been in this game for decades.

Machine Learning

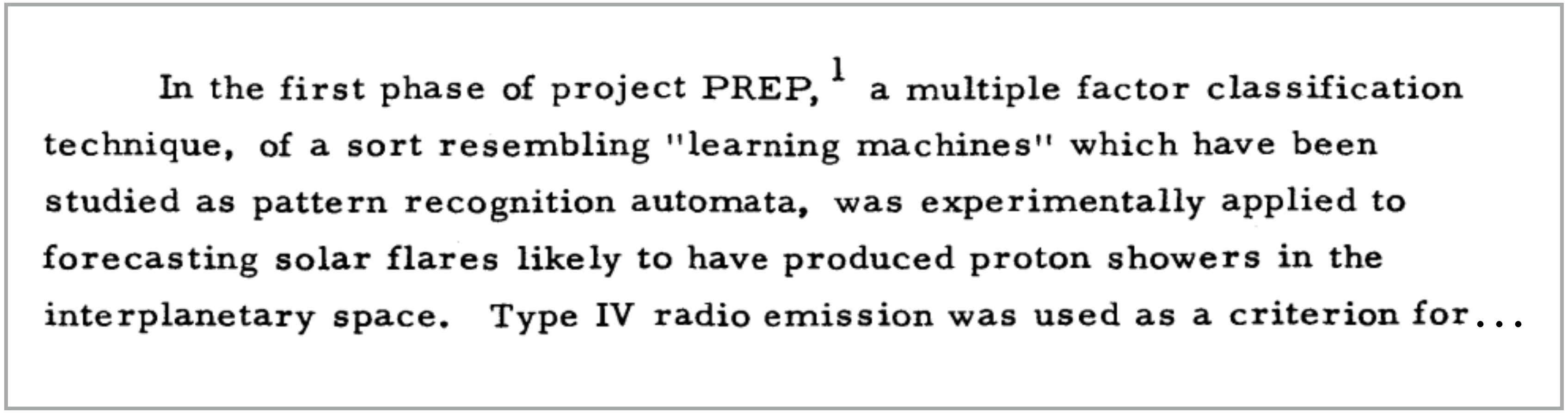


Machine Learning (ML): A branch of artificial intelligence in which a computer progressively improves its performance on a specific task by "learning" from data, without being explicitly programmed.

Deep Learning (DL): An extension of Machine Learning that uses the mathematical concept of a neural network (NN) to loosely simulate information processing and adaptation patterns seen in biological nervous systems.

The `new’ field of Machine Learning

We think of Machine Learning as a new field, but is it?



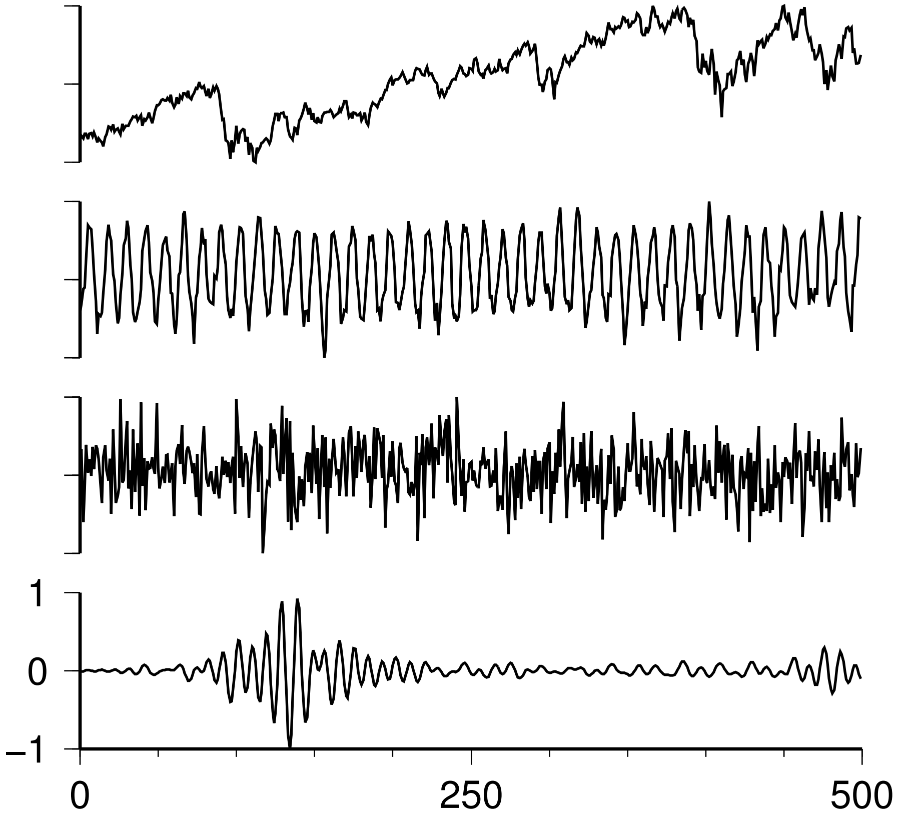
In the first phase of project PREP,[1] a multiple factor classification technique, of a sort resembling "learning machines" which have been studied as pattern recognition automata, was experimentally applied to forecasting solar flares likely to have produced proton showers in the interplanetary space. Type IV radio emission was used as a criterion for...

“Study of the application of perceptrons for prediction of Solar Flares,
Solar flare forecasting with a recognized automation”

`Final report for Phase II Project PREP. Prepared for NASA
by C.M. Theiss and A.E. Murray, Cornell Aeronautical Laboratory, Inc. (February 1965)

Human and Machine Learning

Classification



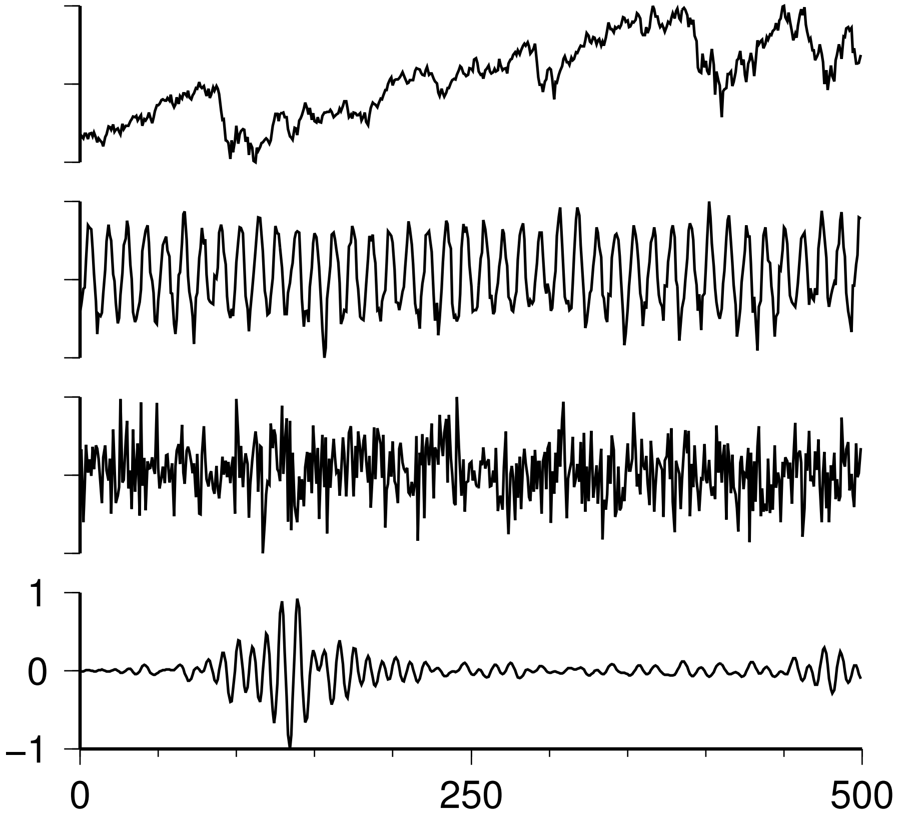
Seismogram, Temperature in Birmingham over 50 years, White noise Stock prices

Humans can recognise seismic signals. The way computers do it implicitly defines some misfit criteria.

From Valentine & Trampert (2012)

Human and Machine Learning

Classification



Stock prices

Temperature in Birmingham over 50 years

White noise

Seismogram

Seismogram,   Temperature in Birmingham over 50 years,   White noise   Stock prices

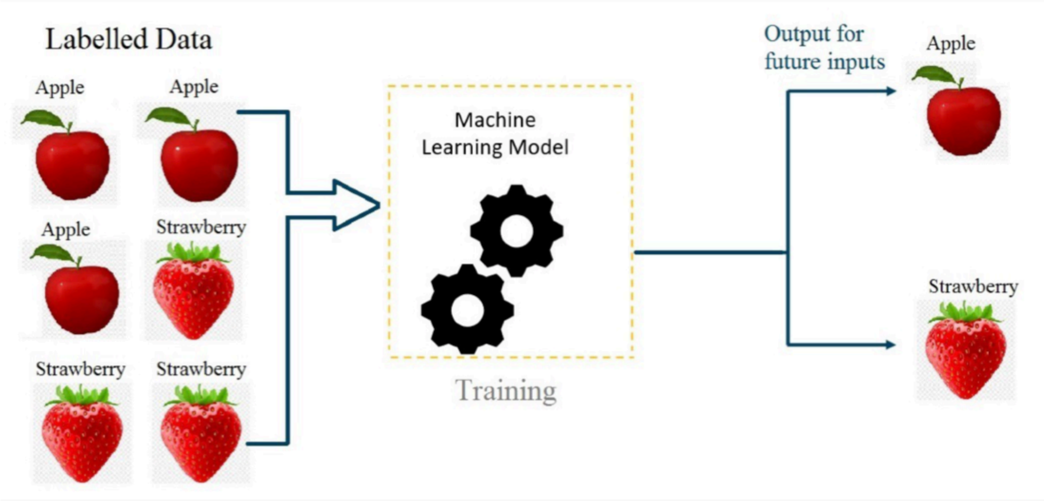
Humans can recognise seismic signals. The way computers do it implicitly defines some misfit criteria.

From Valentine & Trampert (2012)

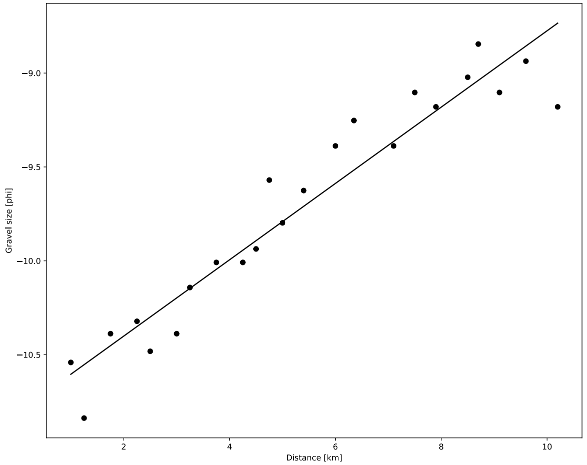
Supervised Learning

Given training data of inputs and outputs

Classification



Regression



High dimensional inputs require large volumes of data to detect correlations.

Make predictions of outputs for future inputs,. . ., generate new outputs `in a similar style’

Machine Learning in Seismology

First data centric applications is seismology from 1990s

- Earthquake first-arrival picking (McCormack et al. 1993; Veezhinathan et al, 1991)
- Deconvolution of seismic traces (Wang and Mendel, 1992)
- Discrimination between earthquakes and artificial sources (Dowla et al., 1990; Dysart and Pulli, 1990)

Recent applications use deep learning, but applied to similar applications

- First-arrival picking (Ross et al. 2018a,b; Zhu and Beroza, 2019; Mousavi et al. 2020)
- earthquake detection (Perol et al. 2018; Mousavi et al. 2020)
- Discrimination between signal and noise (Meier et al. 2019)
- Signal denoising (Zhu et al., 2019)

Machine Learning in Seismology

Improved computational capacity has led to vastly increased sophistication of Neural Networks in seismic applications.

Early and modern use of Neural Networks applied to first arrival picking.

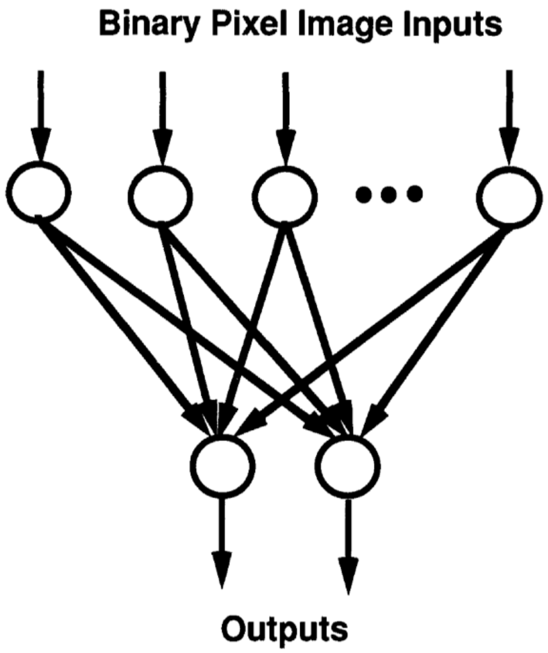
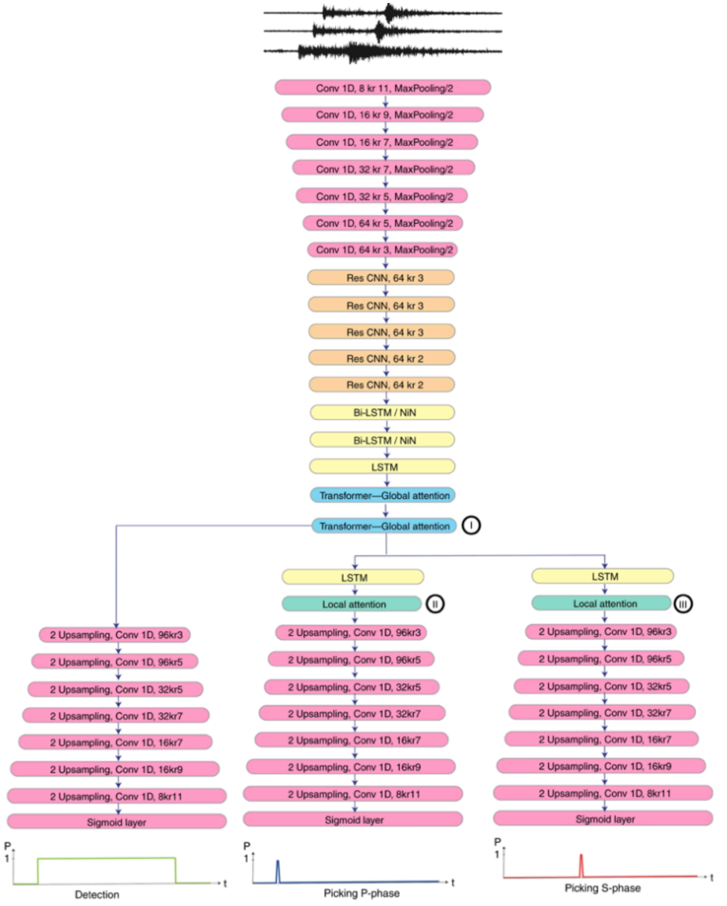


FIG. 2. The structure of the first-break picker neural network. The inputs to the network are pixel images of the seismic data about the peak currently being examined. Network outputs indicate whether the current peak is earlier or later in time than the first break on the trace.

The two layer model of McCormack et al. (1993)


Conv 1D, 8 kr 11, MaxPooling/2
Conv 1D, 16 kr 9, MaxPooling/2
Conv 1D, 16 kr 7, MaxPooling/2
Conv 1D, 32 kr 7, MaxPooling/2
Conv 1D, 32 kr 5, MaxPooling/2
Conv 1D, 64 kr 5, MaxPooling/2
Conv 1D, 64 kr 3, MaxPooling/2
Res CNN, 64 kr 3
Res CNN, 64 kr 3
Res CNN, 64 kr 3
Res CNN, 64 kr 2
Res CNN, 64 kr 2
Bi-LSTM / NiN
Bi-LSTM / NiN
LSTM
Transformer—Global attention
Transformer—Global attention
I
LSTM
Local attention
II
LSTM
Local attention
III
2 Upsampling, Conv 1D, 96kr3
2 Upsampling, Conv 1D, 96kr5
2 Upsampling, Conv 1D, 32kr5
2 Upsampling, Conv 1D, 32kr7
2 Upsampling, Conv 1D, 16kr7
2 Upsampling, Conv 1D, 16kr9
2 Upsampling, Conv 1D, 8kr11
Sigmoid layer
P
1
t
Detection
Picking P-phase
Picking S-phase

70 layer model of Mousavi et al. (2020) involving a mixture of convolutional layers using 372,000 tunable parameters.

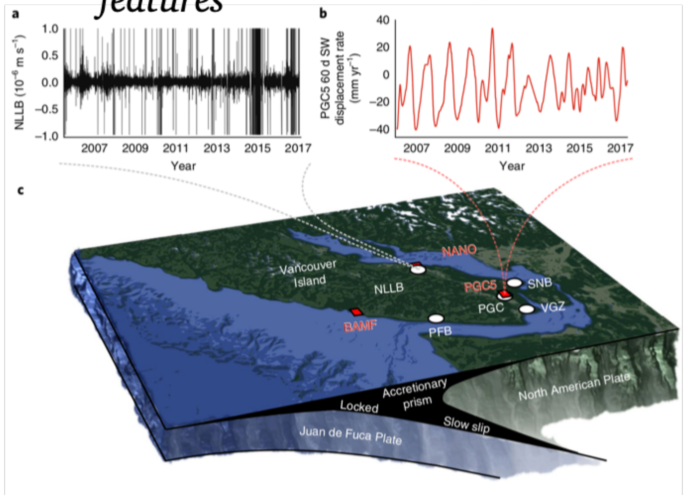
Detecting new signals in data

More exciting applications would be to detect new signals in seismic data….

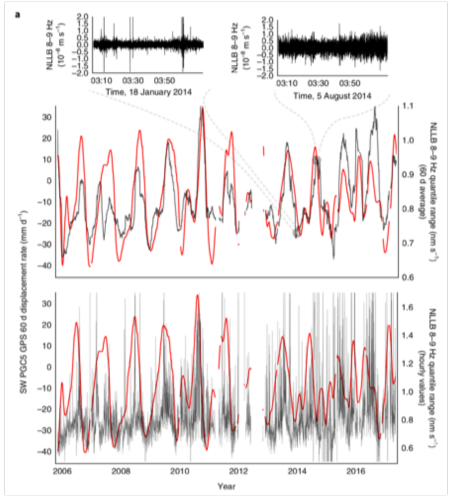
Detecting correlations between seismic noise features and GPS signals shows how `noise' coming from the slab relates to slab movement.

Seismic “noise” features

GPS displacement



Machine learning



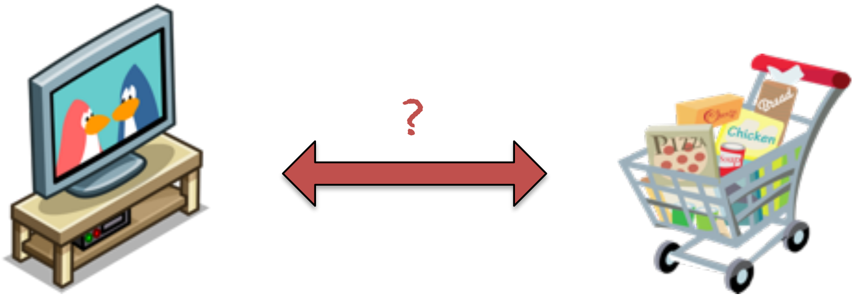
Seismic features
∼
GPS displacement

“Continuous chatter of the cascadia sub-duction zone revealed by machine learning” - Rouet-Leduc et al., 2019.

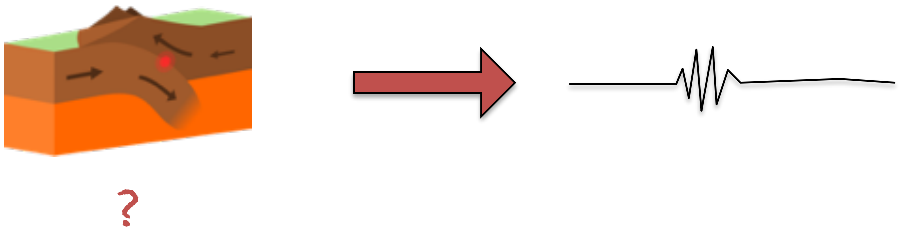
“An exponential build-up in seismic energy suggests a months-long nucleation of slow slip in Cascadia” - Hulbert et al., 2020.

Surrogate models

Commercial data science



Predictive inference in the absence of a forward problem.



Geophysical inverse problem

But we have physics and would not want to be parted from it!

But there are cases where we might want a faster approximate forward theory, one that can be automatically constructed.

Jonathan Tompson, Kristofer Schlachter, Pablo Sprechmann, Ken Perlin, Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017.

arXiv: 1607.03597v5 [cs.CV] 3 Mar 2017

https://www.youtube.com/watch?v=iOWamCtnwTc

Physics informed Surrogate models

An emerging trend is to go beyond correlations and include physical laws in Neural Networks (e.g. Li et al., 2021; Raissi, 2018; Raissi et al., 2019).

A simple ODE example

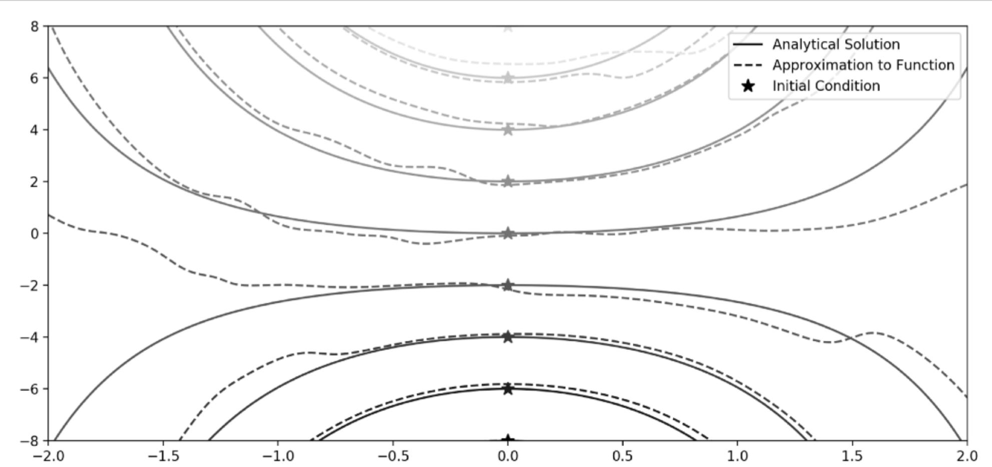
$$\frac{\partial f}{\partial x} = x + xy, \quad y(0) = y_0$$

with solution

$$y(x) = (y_0 + 1)\ e^{\frac{1}{2}x^2} - 1$$

A neural network can be set up to represent the function, $f(x)$, which minimizes

$$\int_a^b \left(\frac{\partial f}{\partial x} - x - xf(x) \right)^2 dx + \alpha(f(0) - y_o)^2$$



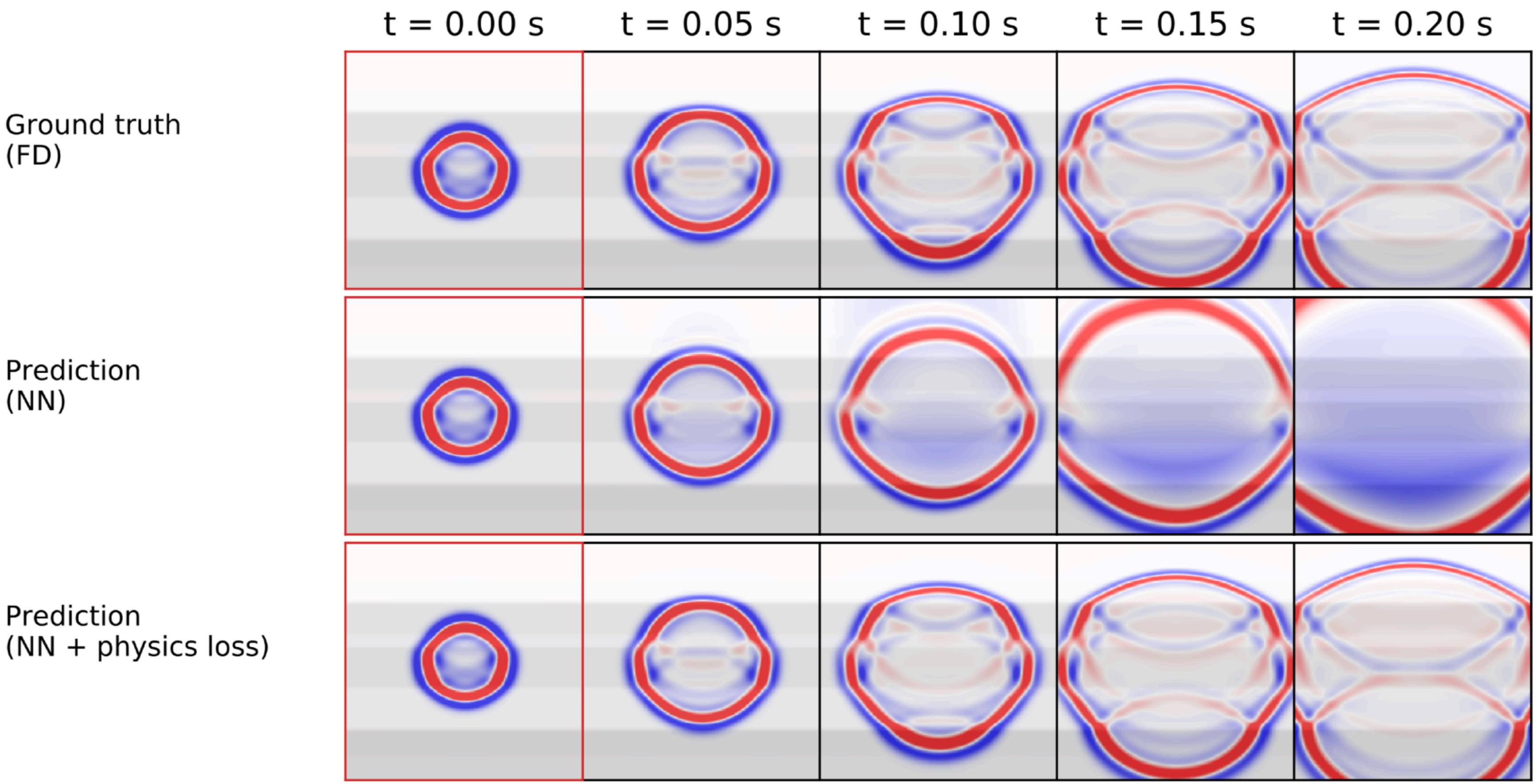
Solving a differential equation with a simple neural network.

Courtesy M. Scheiter.

Physics informed Surrogate models

An emerging trend is to go beyond correlations and include physical laws in Neural Networks (e.g. Li et al., 2021; Raissi, 2018; Raissi et al., 2019).

Comparing standard Neural network with Physics informed Neural Network.



, iter.

A wave equation example by Moseley et al. (2020)

Generative models

Another growing trend in Machine Learning is use of Generative models

- Variational autoencoders (Kingma and Welling, 2014)
- Generative adversarial networks (GANS) (Goodfellow et al., 2014)
- Diffusion models (Sohl-Dickstein et al., 2015)
- Flow based models (Rezende and Mohamed, 2016)

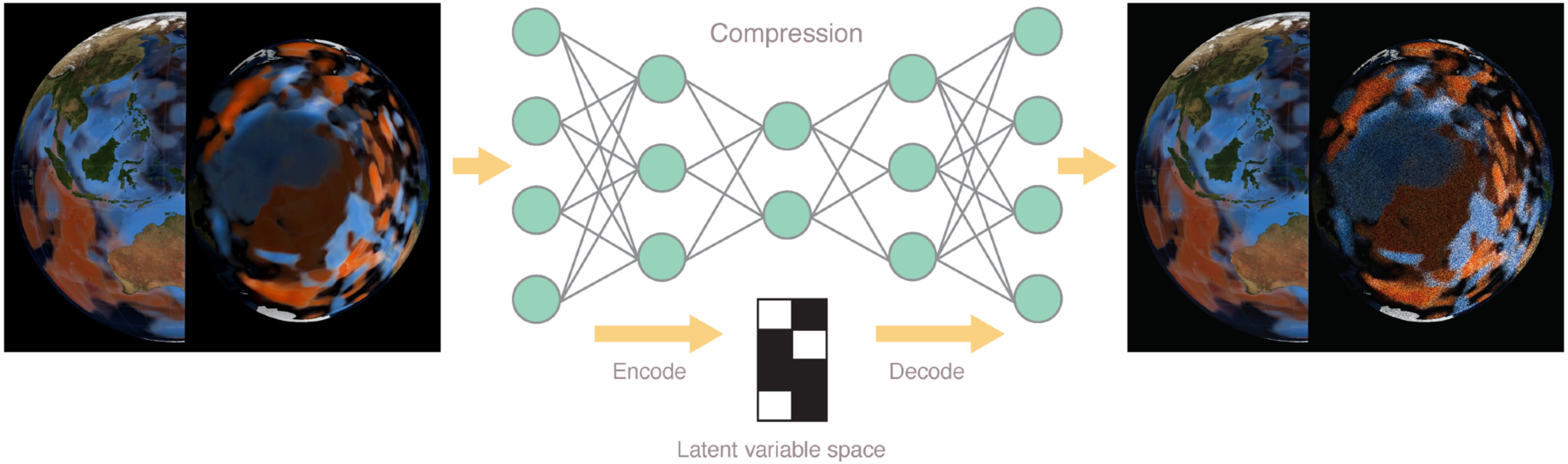
They are responsible for DeepFakes. Typical use is training a NN to mimic the features in training data and then generate new outputs (e.g. images) in the the same style.
Geophysical applications include:

- first arrival picking (Zhang and Sheng, 2020)
- Earthquake/noise discrimination (Li et al., 2018)
- Seismic data interpolation (Oliveria et al, 2018), data augmentation, (Wang et al., 2019) and data reconstruction (Siahkoohi et al., 2018).

Direct Inversion applications of generative models:

- Dimensionality reduction (Laloy et al., 2018,2019; Moser et al., 2020; Lopez-Alzis et al. 2021.)
- Model space samplers (Siahkoohi and Hermann (2021); Zhang and Curtis (2021); Zhao et al. (2021).

Latent variable inversion



Benefits: **Drastic reduction** in numbers of unknowns for little to no loss in representation.

But is there a **price to be paid** in that the inverse problem may be more difficult,

The optimisation function may be more complex, in the latent space than in the larger space.

There is some evidence for this, e.g. Laloy et al. (2019).

Generative models and Bayesian inference

Here a WGAN has been used to replicate the ensemble of shear wave velocity Earth models at the CMB produced by a large Trans-D MCMC Bayesian sampling algorithm.

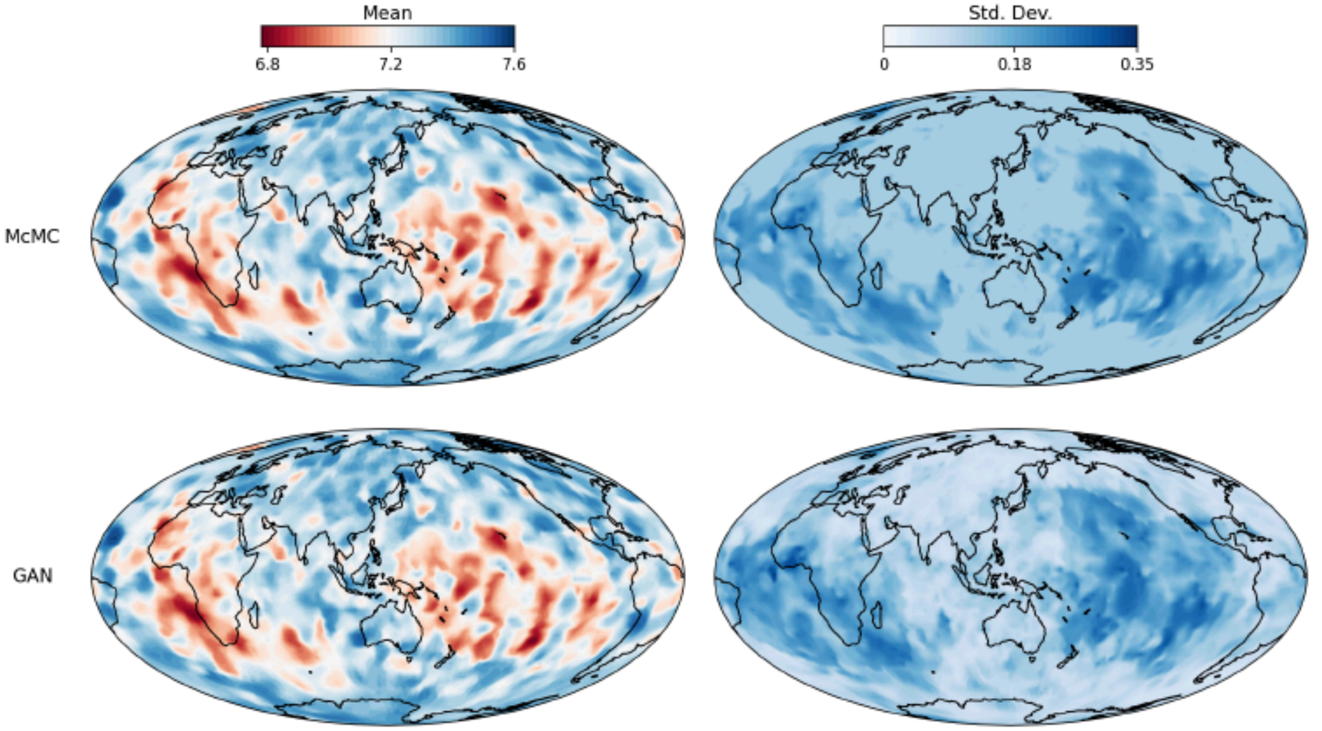


Figure 8. Maps of mean and standard deviation for the full model ensemble, divided into 16 regions; comparison between McMC samples and GAN samples. The GANs can recreate the full velocity model ensemble from Mousavi et al. (2021) in high resolution.

- Gives 95-99% reduction in digital volumes -> gives ease of distribution of more than mean and std.
- Allows any third party to generate new ensembles with dramatically increased computational speed.

From Scheiter. et al. (2021, in prep)

Generative models and Bayesian inference

Here a WGAN has been used to replicate the ensemble of shear wave velocity Earth models at the CMB produced by a large Trans-D MCMC Bayesian sampling algorithm.

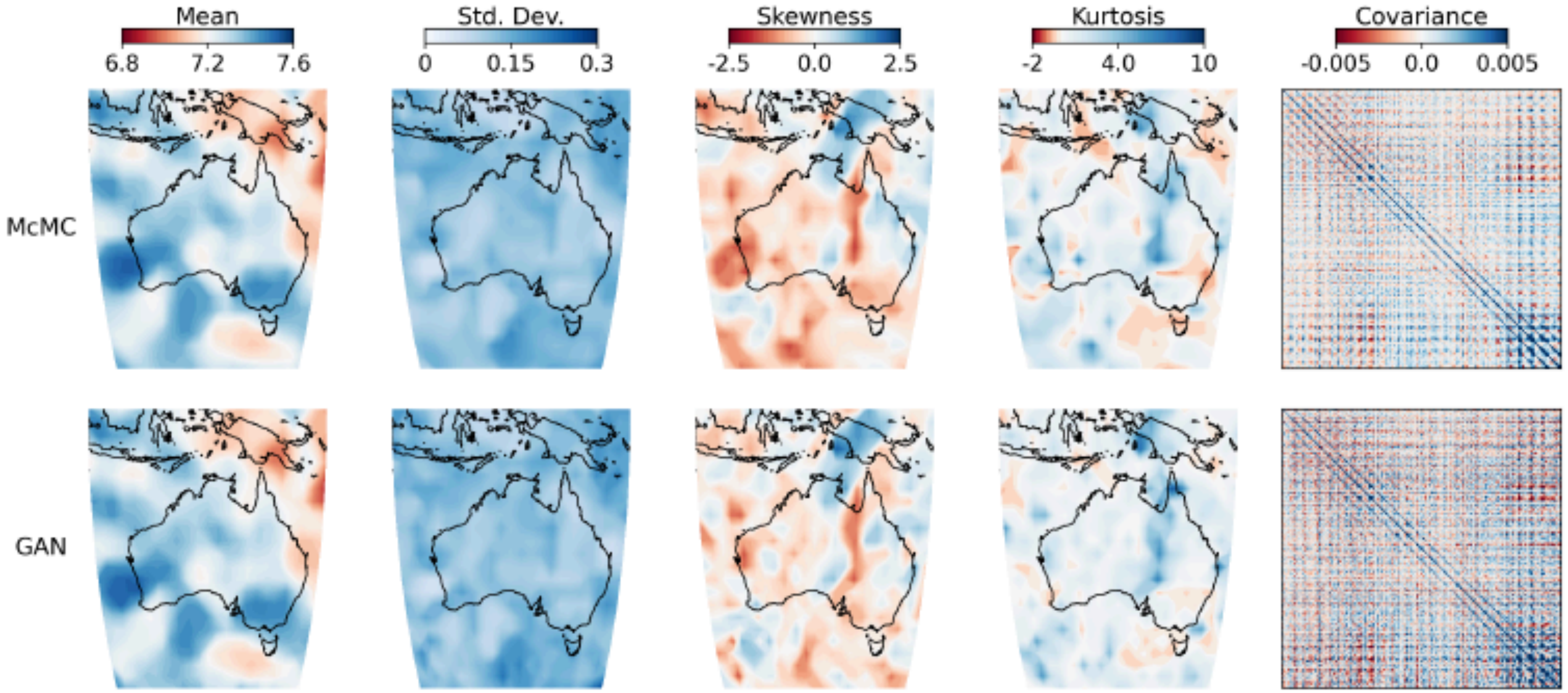


Figure 5. Comparison between the original McMC samples and an equal amount sampled from the GAN. Displayed are the first four moments of the distribution and the covariance matrix. Overall, the GAN learns the original ensemble in great detail.

The WGAN is able to capture higher order moments than just mean and standard deviation.

From Scheiter. et al. (2021, in prep)

Optimal Transport

Applications to misfit functions, optimisation and Bayesian Inference.

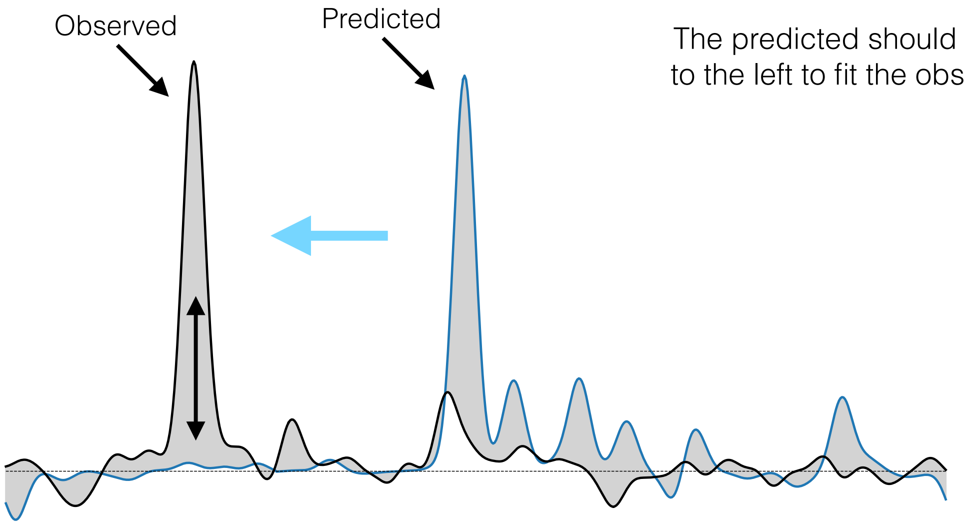
Optimal Transport

Inversion makes a lot of use of misfit functions

Least squares misfit is the sum of the amplitude separation between two signals.

$$\int (u_{obs}(t) - u_{pred}(t))^2 dt$$

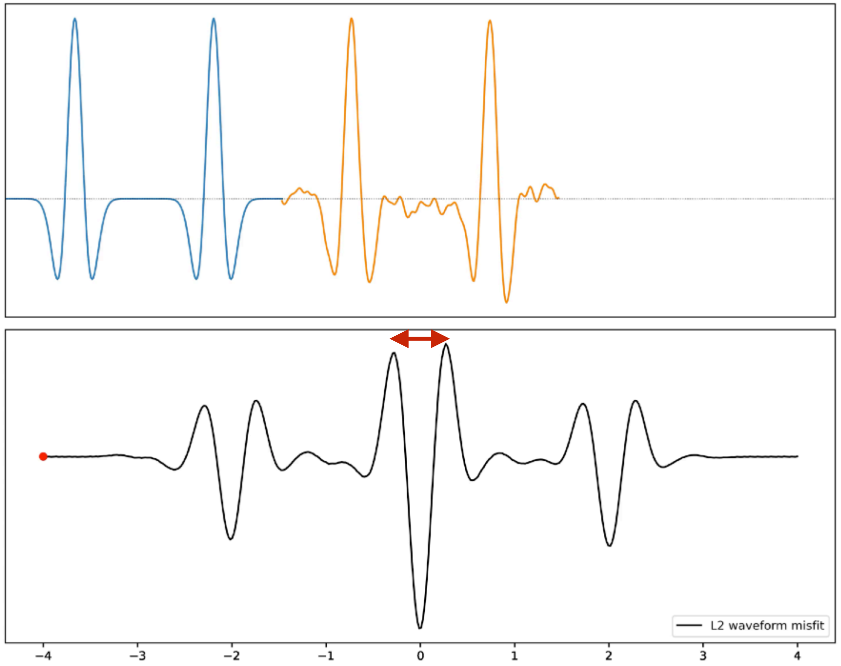
Amplitude, u changes at fixed time, t



The predicted should move to the left to fit the observed

In inversion gradient based optimisation will only converge within a narrow range of time shifts

Ideally we would like a misfit function which had `long range' information built into it.



An animation of least squares misfit between blue and orange waveforms as a function of phase shift

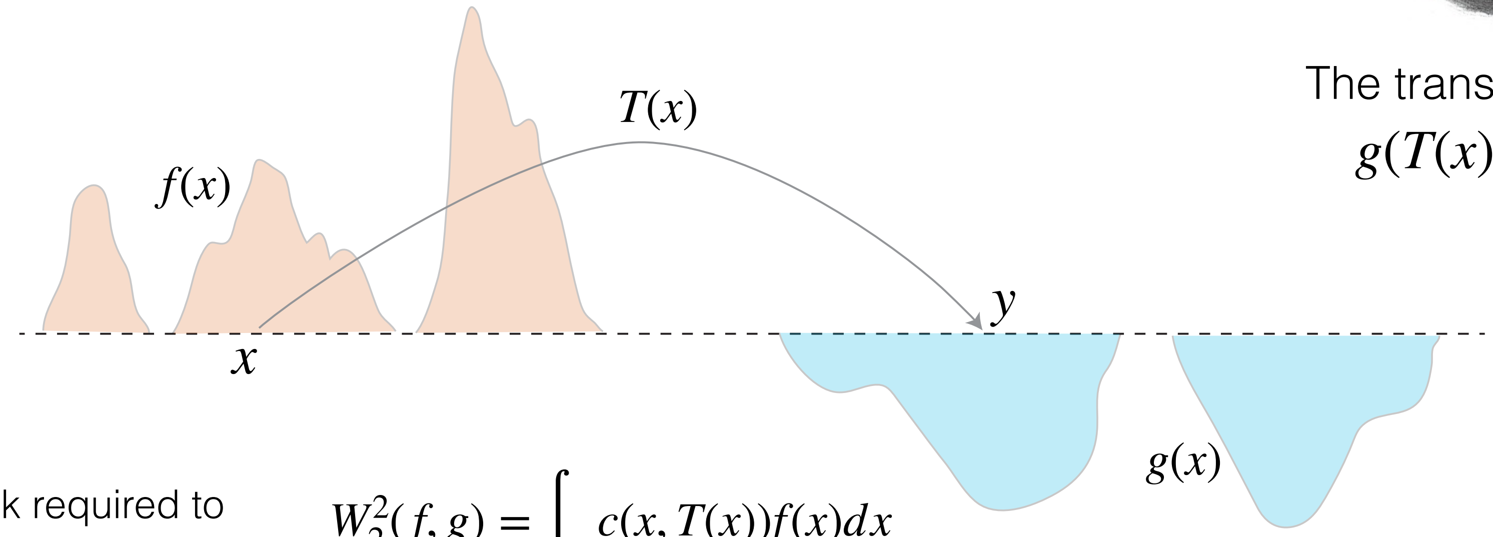
Cycle skipping

Following earlier work by Engquist and Froese (2014)

Napoleons problem

Optimal transport traces its roots back to the tail of Napoleon and his Gaspard Monge.

How to optimally transport the sand $f(x)$ to the holes $g(x)$?



The transport map

$$g(T(x)) = f(x)$$

The work required to complete the task

$$W_2^2(f,g) = \int_X c(x,T(x))f(x)dx$$

$W_p = (\sum_i d_i^p m_i)^{1/p}$ $p =$ 1 or 2

$W_2^2 \propto$ (distance)^2 x mass $\propto$ Energy

$c(x,y)$ is the distance between x and y

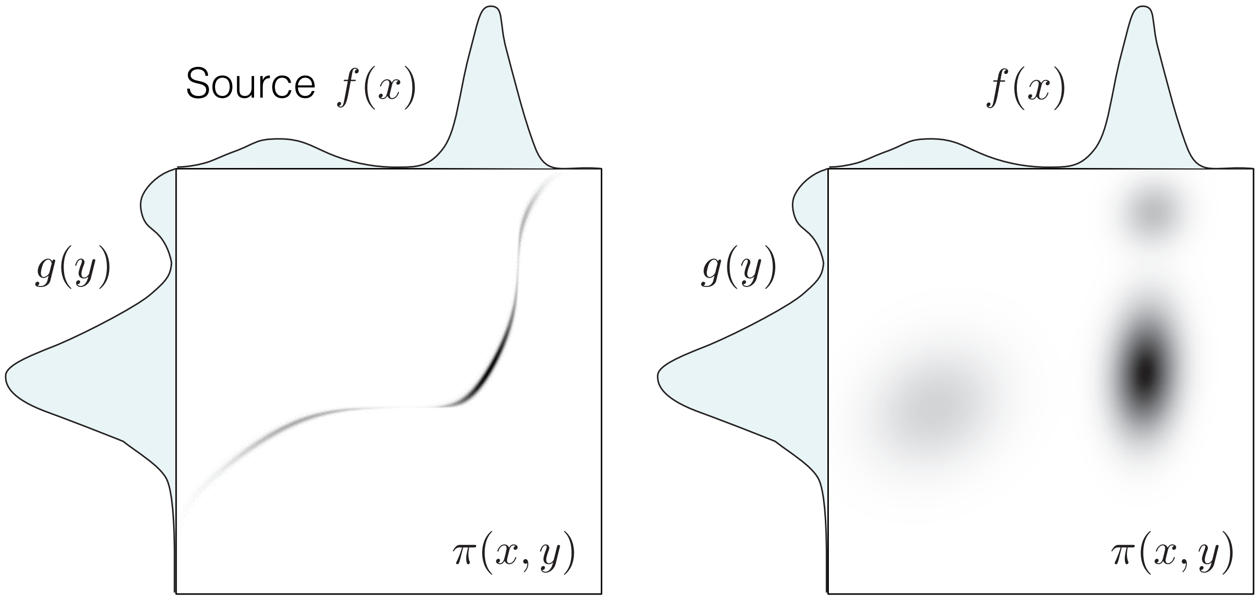
$$(f \cdot g)(\tau) = \int_{-\infty}^{\infty} f(t)g(t+\tau)dt$$

The modern field derives from the work of Kantorovich (1942); Villani (2003, 2008).

Wasserstein distances and Transport plans

Optimal transport traces its roots back to the tail of Napoleon and his Gaspard Monge.

Linear programming formulation of Kantorovich (1942). Solve for transport plan $\pi_{i,j}$



$$\min_{\pi(x_i,y_j)} W_p^p = \sum_{i,j} c_{i,j}\pi_{i,j}, \quad \sum_i \pi_{i,j} = f(x_i), \quad \sum_j \pi_{i,j} = g(y_i)$$

W_1 = distance x mass

W_2^2 = (distance)^2 x mass

$$W_p^p = (\text{Distance})^{\text{p}} \times (\text{mass})$$

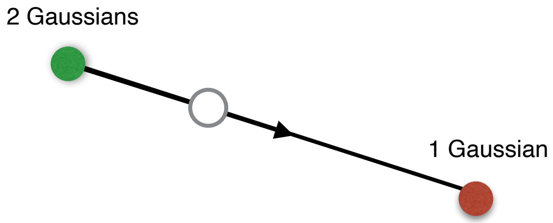
$\pi(x,y)$ = Transport plan

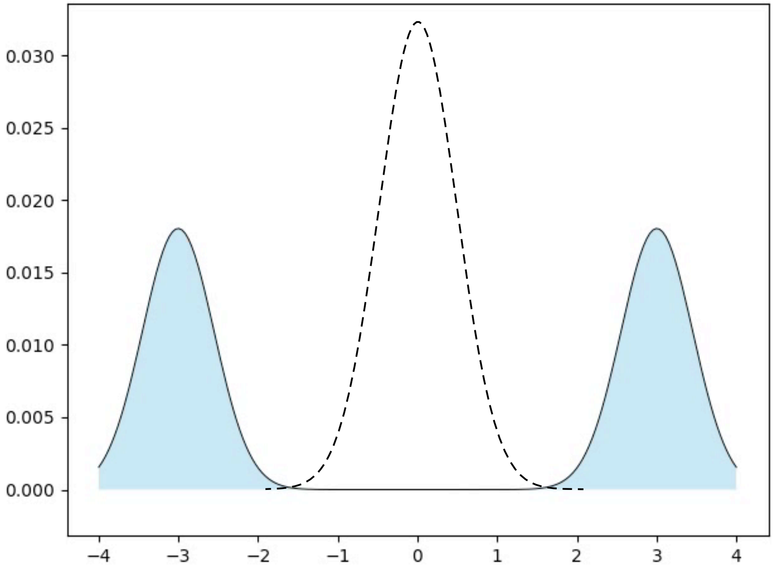
$c(x,y)$ = distance between x and y

Figure adapted from Solomon et al. (2015)

Distances and transformations

Equal steps along the Euclidean path between endpoints

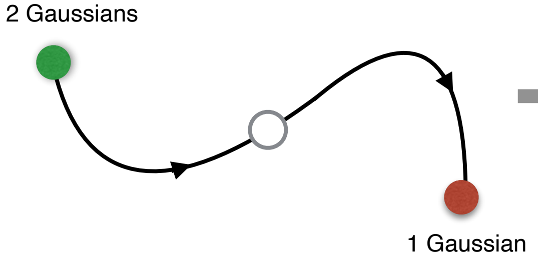


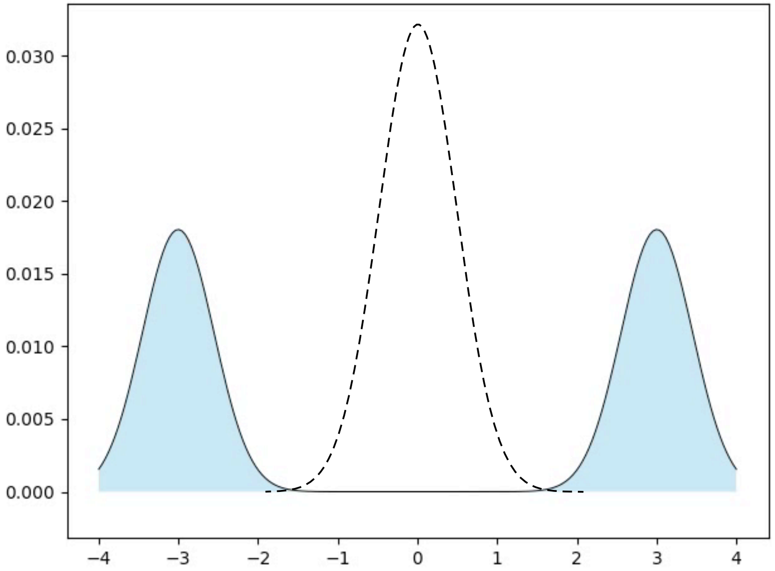


Animation of the linear (**least squares**) path between the start and end distributions

Only amplitude changes

Equal steps along the Optimal Transport path between endpoints



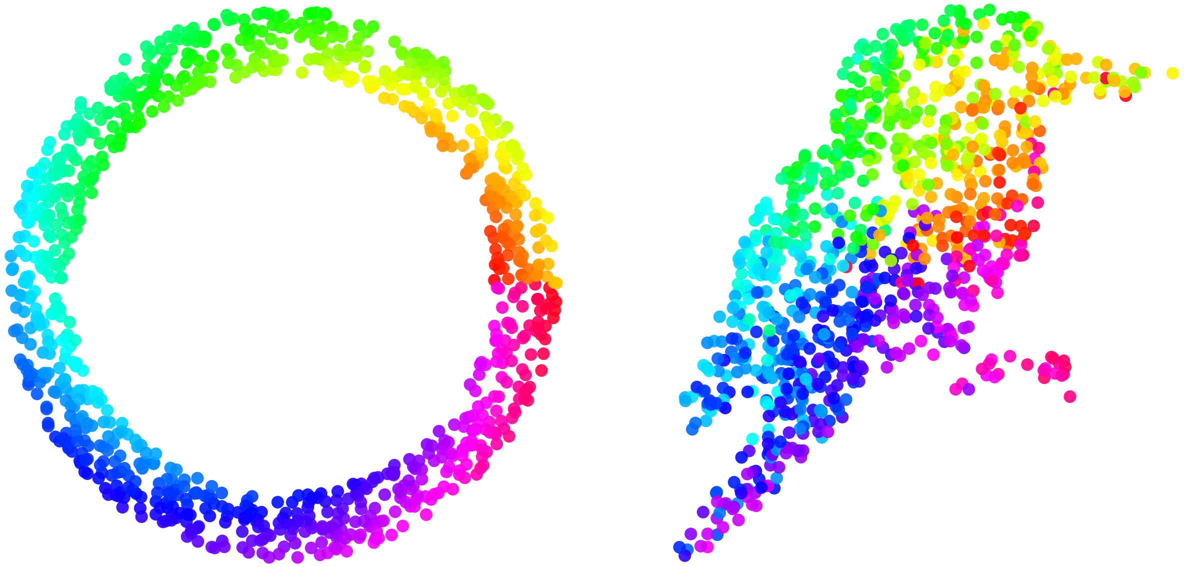


Animation of the **optimal transport** path between the start and end distributions

Amplitude and position changes

Optimal Transport: 2D point cloud

Optimal transport of a 1000 point cloud with uniform amplitude.



Initial

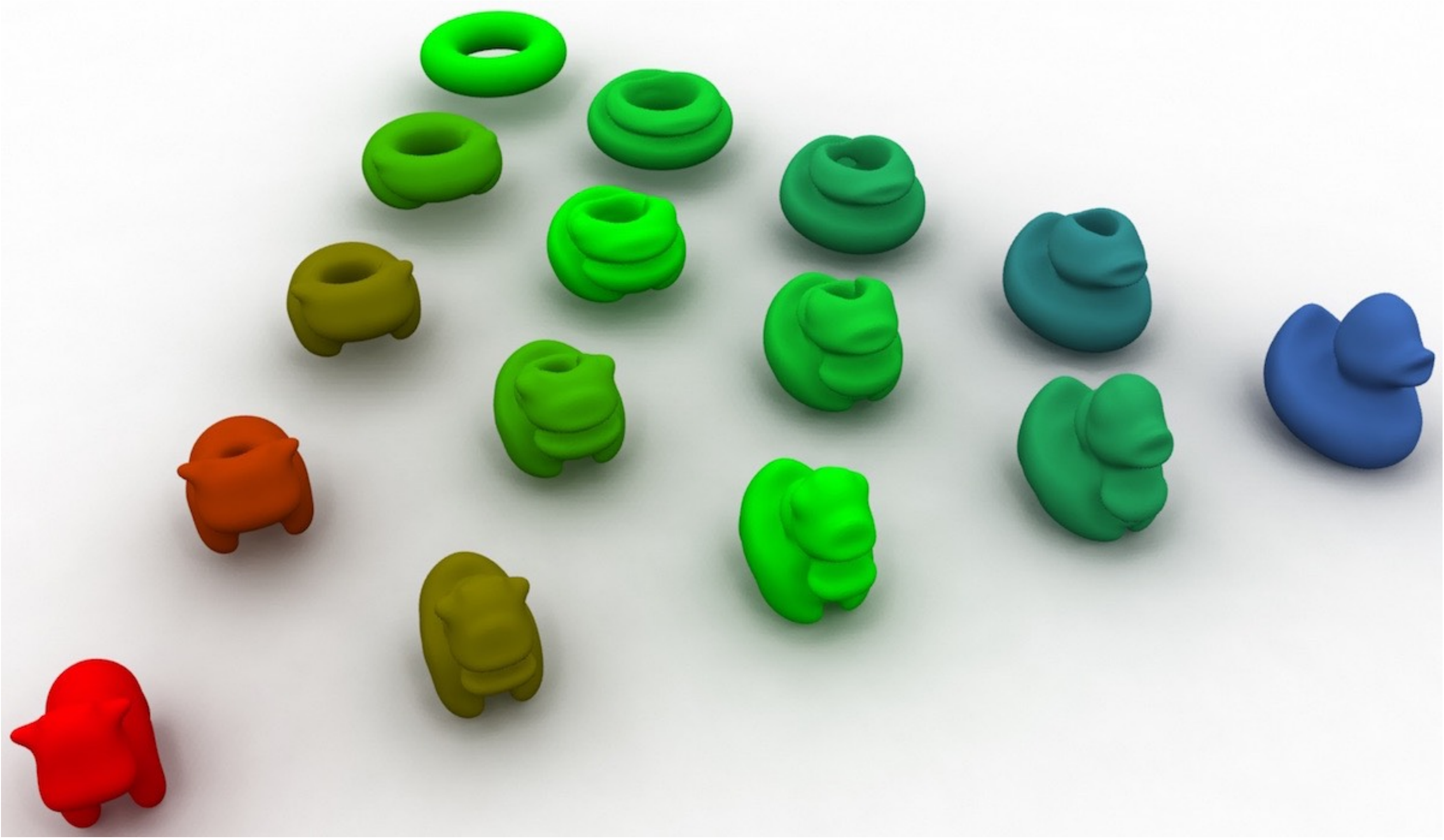
Final

Colours represent point index and allow to visualise Transport map.

Sliced Wasserstein algorithm

Optimal Transport of 3D shapes

From **Cow** to **Duck** to **Torus**



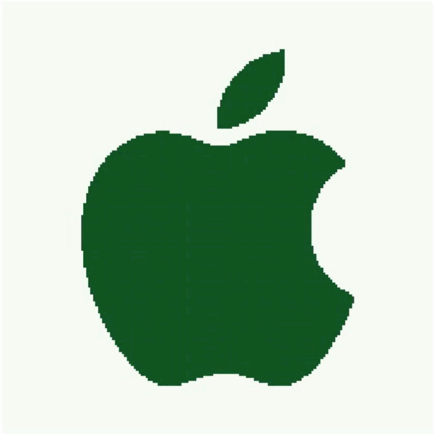
Convolutional Wasserstein Barycentres

From Solomon et al. (2015)

Optimal Transport Animations

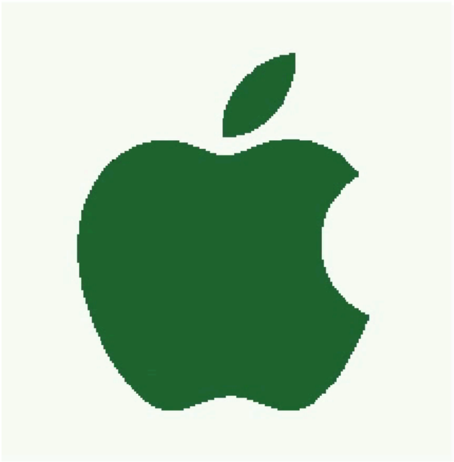
Transporting an apple to an orange

Source



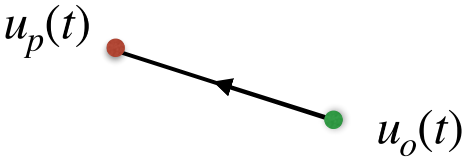
Linear

Source



Optimal transport

Target



Makes use of Sinkhorn Convolutional Wasserstein algorithm of Solomon et al. (2015)
Rémi Flamary and Nicolas Courty, POT Python Optimal Transport library, 2017. https://pythonot.github.io/

Optimal Transport in exploration seismic

A number of groups have applied variants of OT in geophysics, primarily to full waveform inversion (FWI) in exploration seismology.

Engquist and Froese (2014); Engquist et al. (2016); - Monge-Ampere PDE solver (p=2)
Yang and Engquist (2018); Yang et al. (2018);

Me´tivier et al. (2016 a,b,c,d); Me´tivier et al. (2018 a,b); - Dual formulation optimisation (p=1)
Me´tivier et al. (2019); Yong et al. (2018)

Hedjazian et al. (2019) - Seismic receiver functions; *Huang et al. (2019)* - Gravity inversion.

Books and lecture notes:

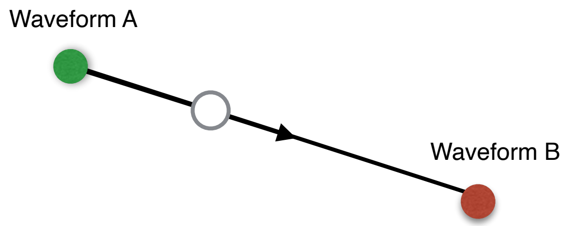
Villani (2003, 2008); Ambrosio (2003); Santambrogio (2015).

Approaches differ between studies:

- Solution method for Wasserstein distance, W_p, and also p value.
- Transform of seismic trace to a Probability Density Function (PDF).
- 1D OT Trace by trace or 2D reflection image.

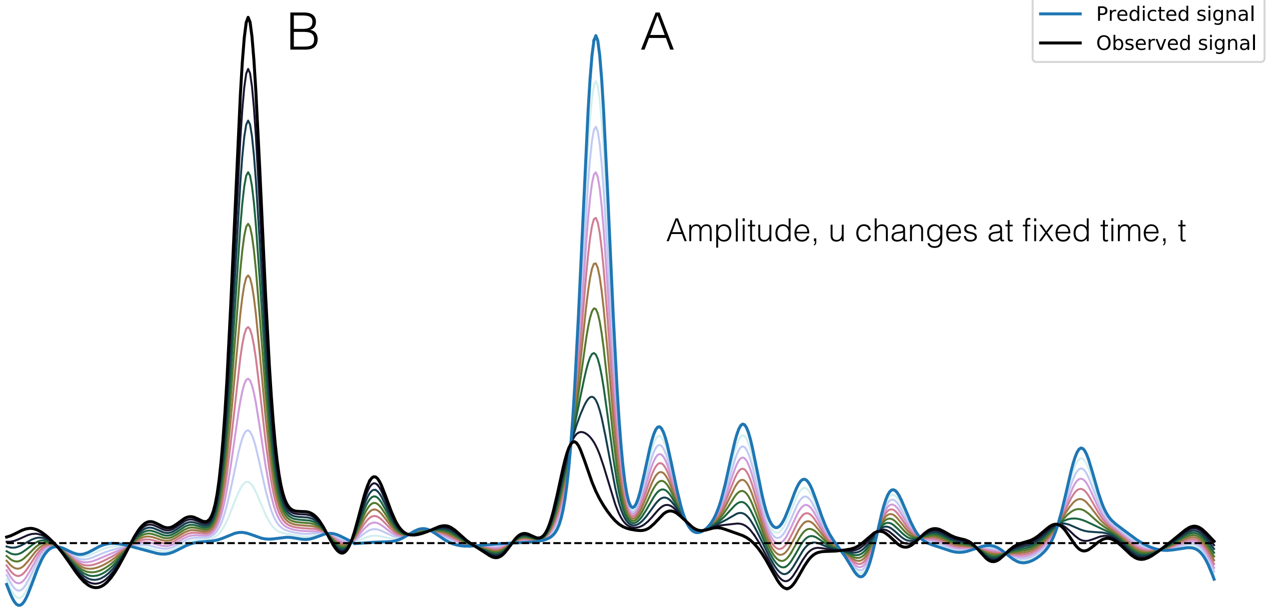
These are all **open** issues. It is an evolving field.

Transport between waveforms



Least squares

Uses the vertical separation as a measure of distance between two signals.



Amplitude, u changes at fixed time, t

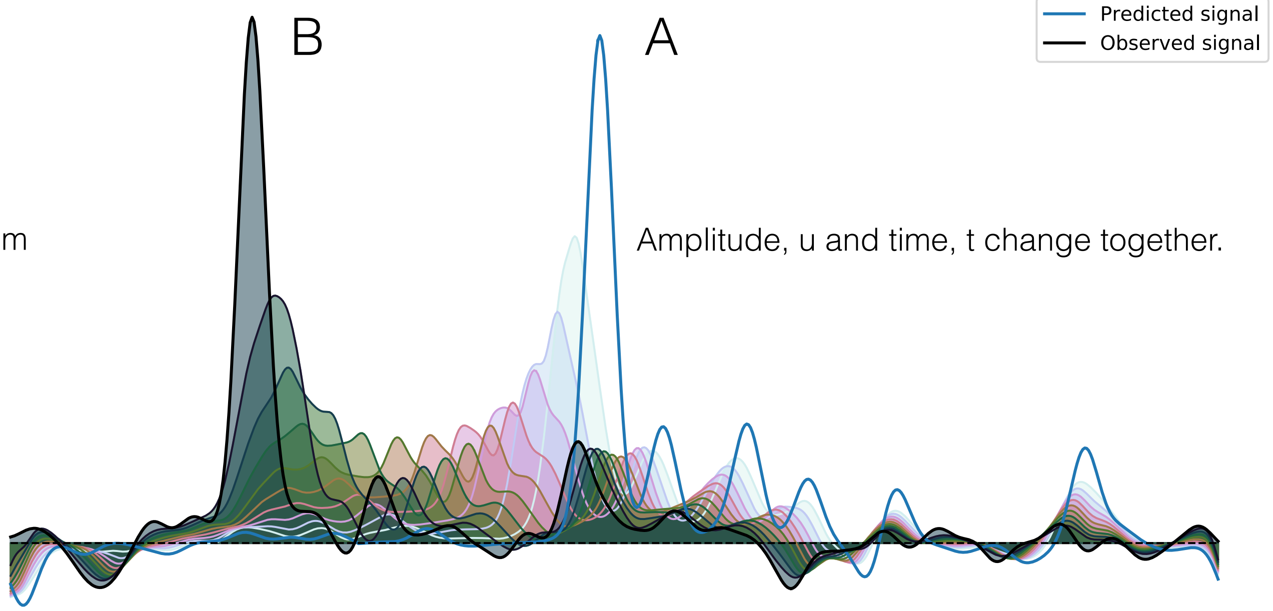
Optimal transport

Uses the minimum work required to transform one signal onto another.

W_1 = distance x mass

W_2^2 = (distance)^2 x mass

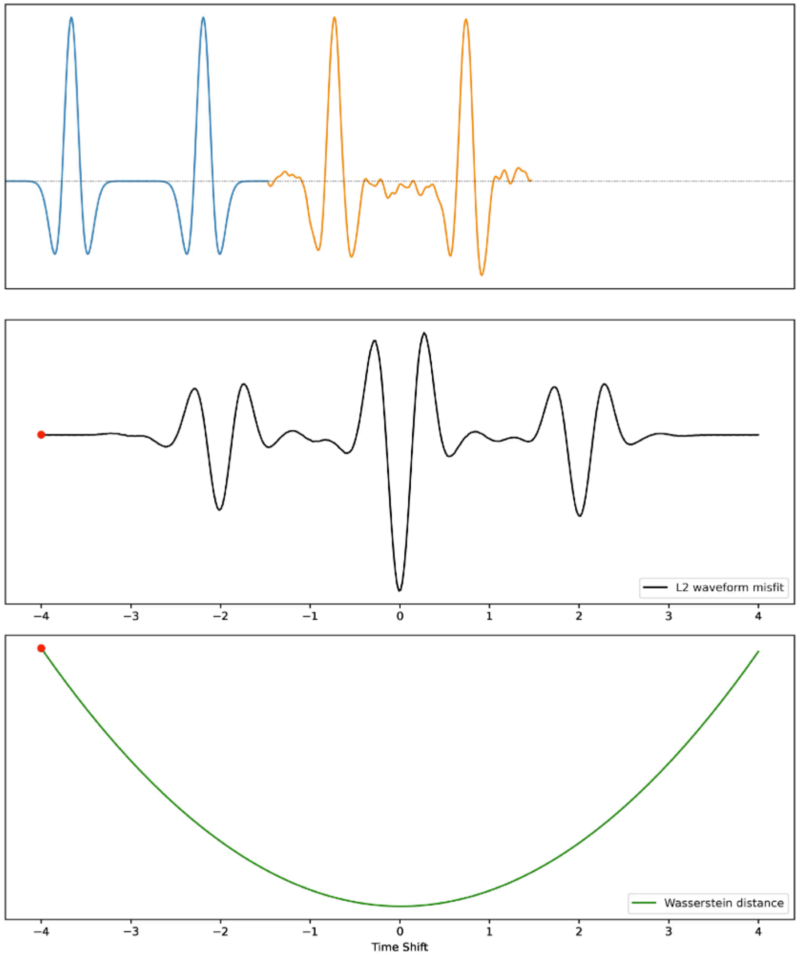
W_p^p = (distance)^p x mass



Amplitude, u and time, t change together.

OT Waveform misfit

Least squares misfit and Wasserstein distance between a pair of double Ricker wavelets

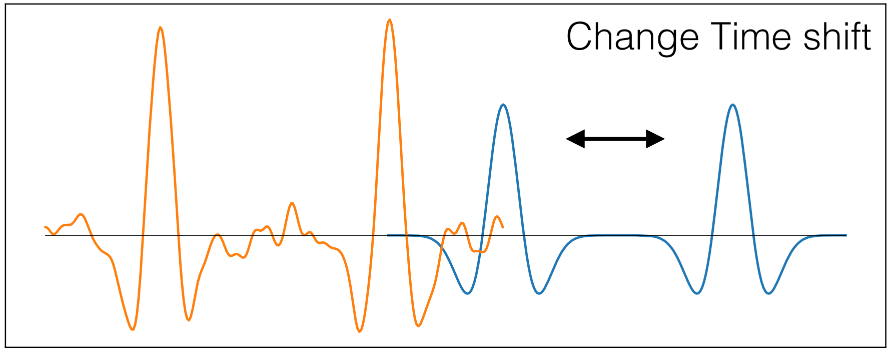


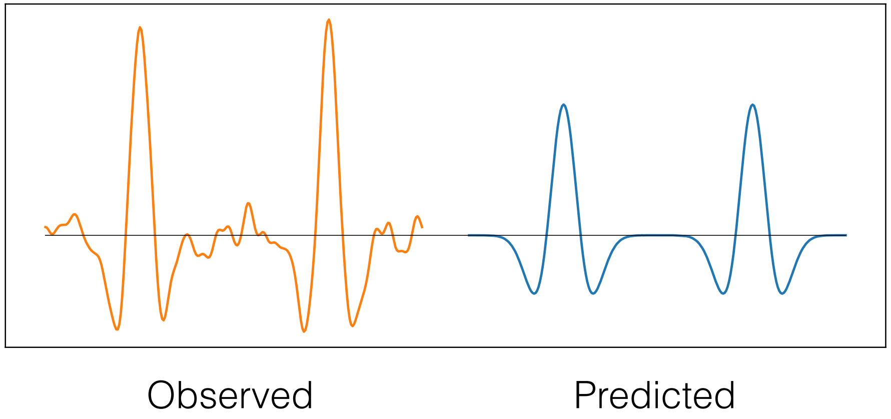
Using Wasserstein misfit formulation of Sambridge et al. (2021, under review)

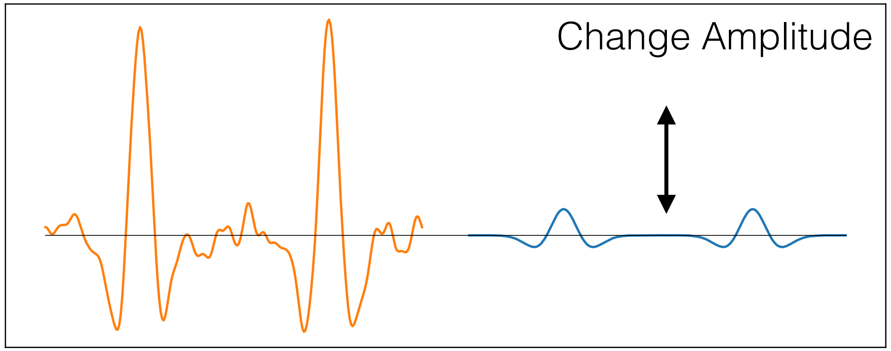
A variant of an experiment performed by Engquist and Froese (2014)

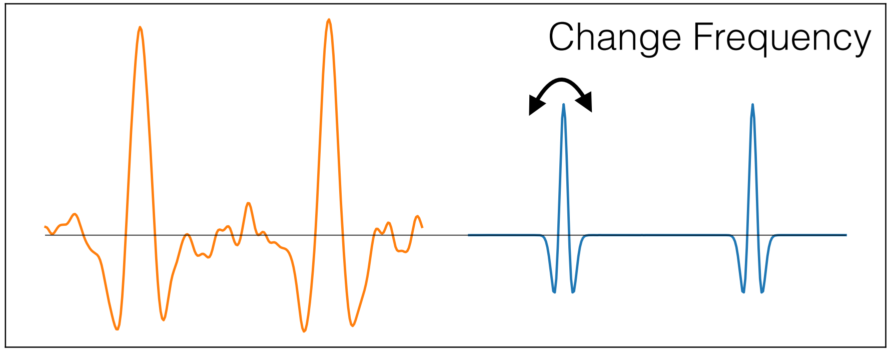
Minimizing the Wasserstein distance

Fit the noisy waveform by adjusting three parameters









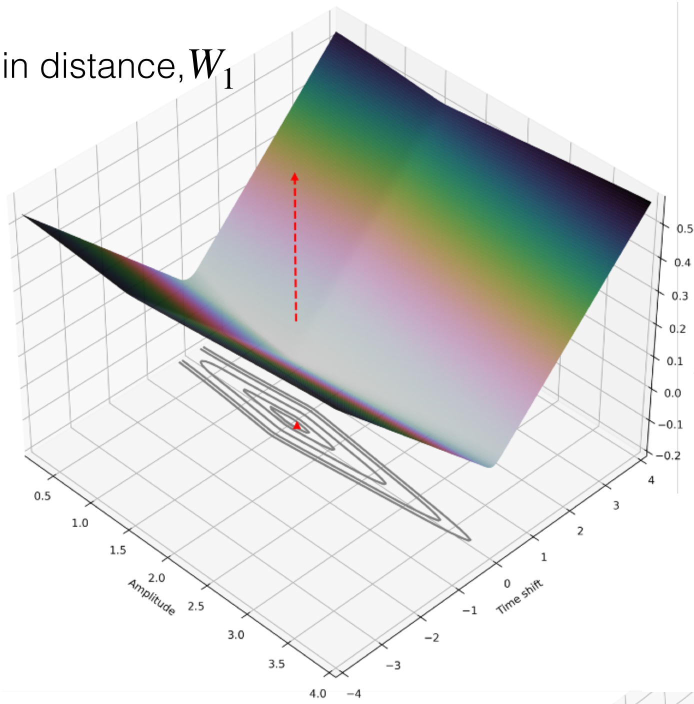
Noise is $N(\mu, \sigma^2)$

$\mu =$ 5% of maximum Ricker amplitude
$\sigma =$ 50% of maximum Ricker period

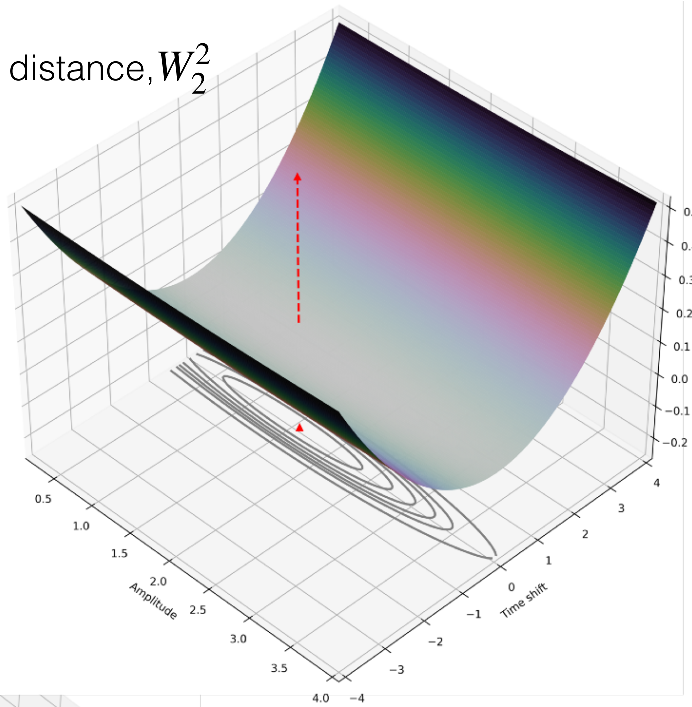
Sambridge et al. (2021, under review)

Wasserstein vs L_2 optimisation

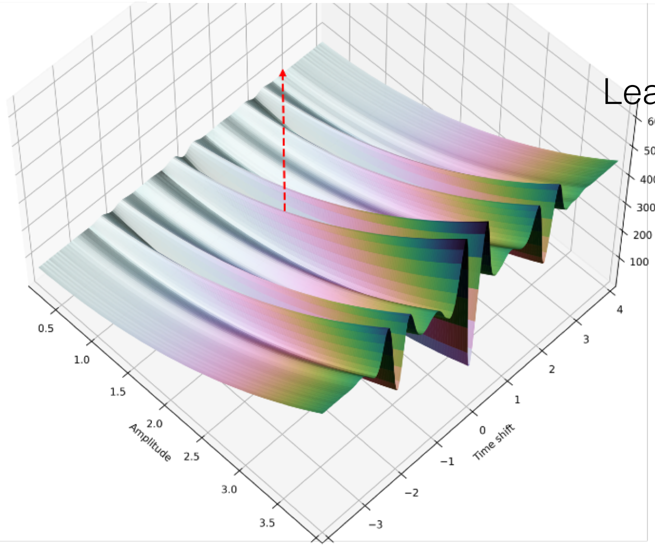
Wasserstein distance, W_1



Wasserstein distance, W_2^2



Least squares misfit, L_2



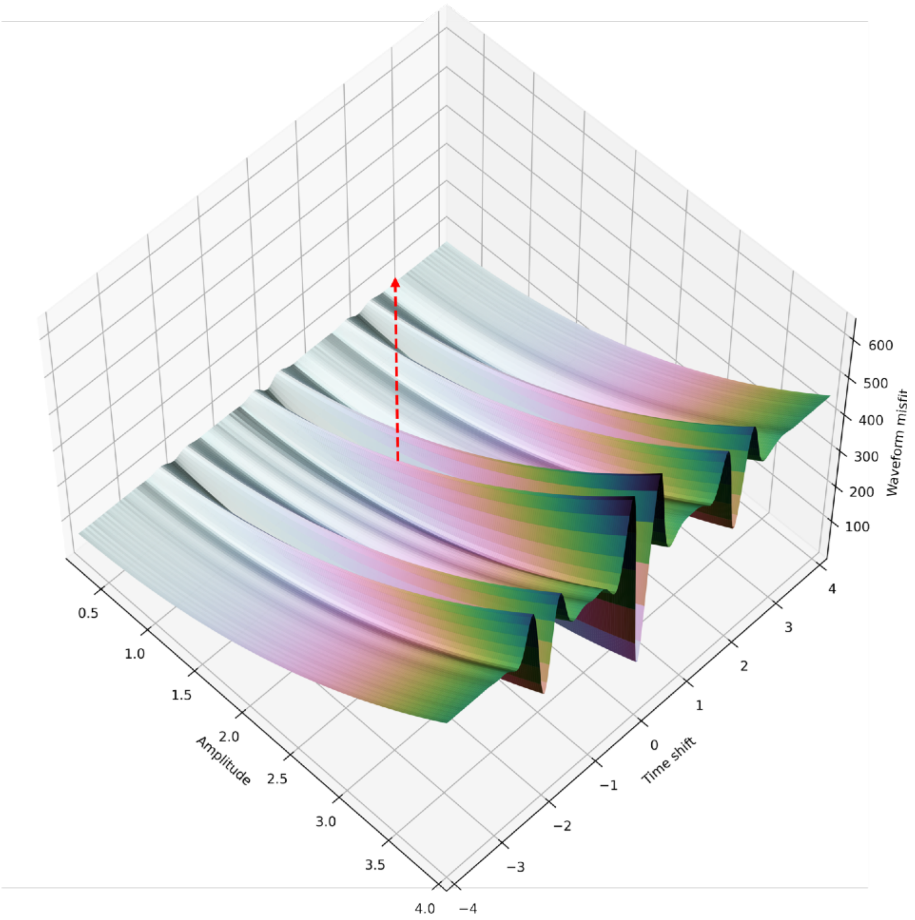
Wasserstein based on 2D fingerprint PDF

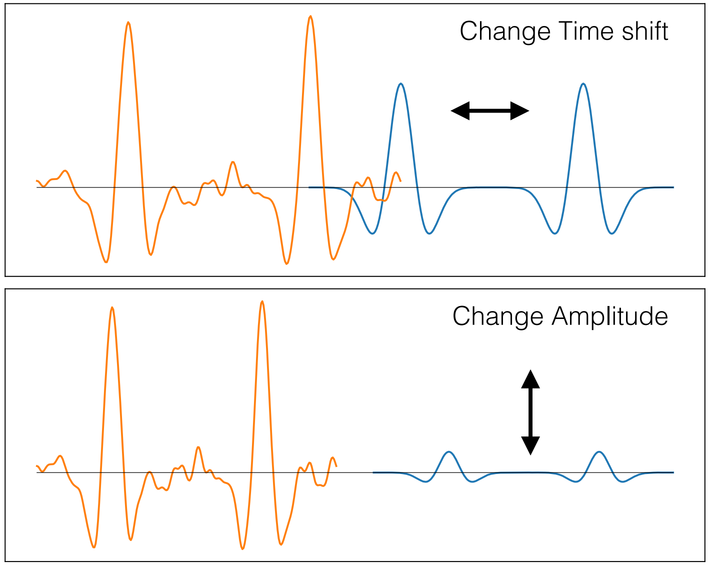
Red line indicates global minimum

Sambridge et al. (2021, under review)

L_2 optimisation misfit surface

Least squares waveform misfit as a function of
Time shift and Amplitude parameters



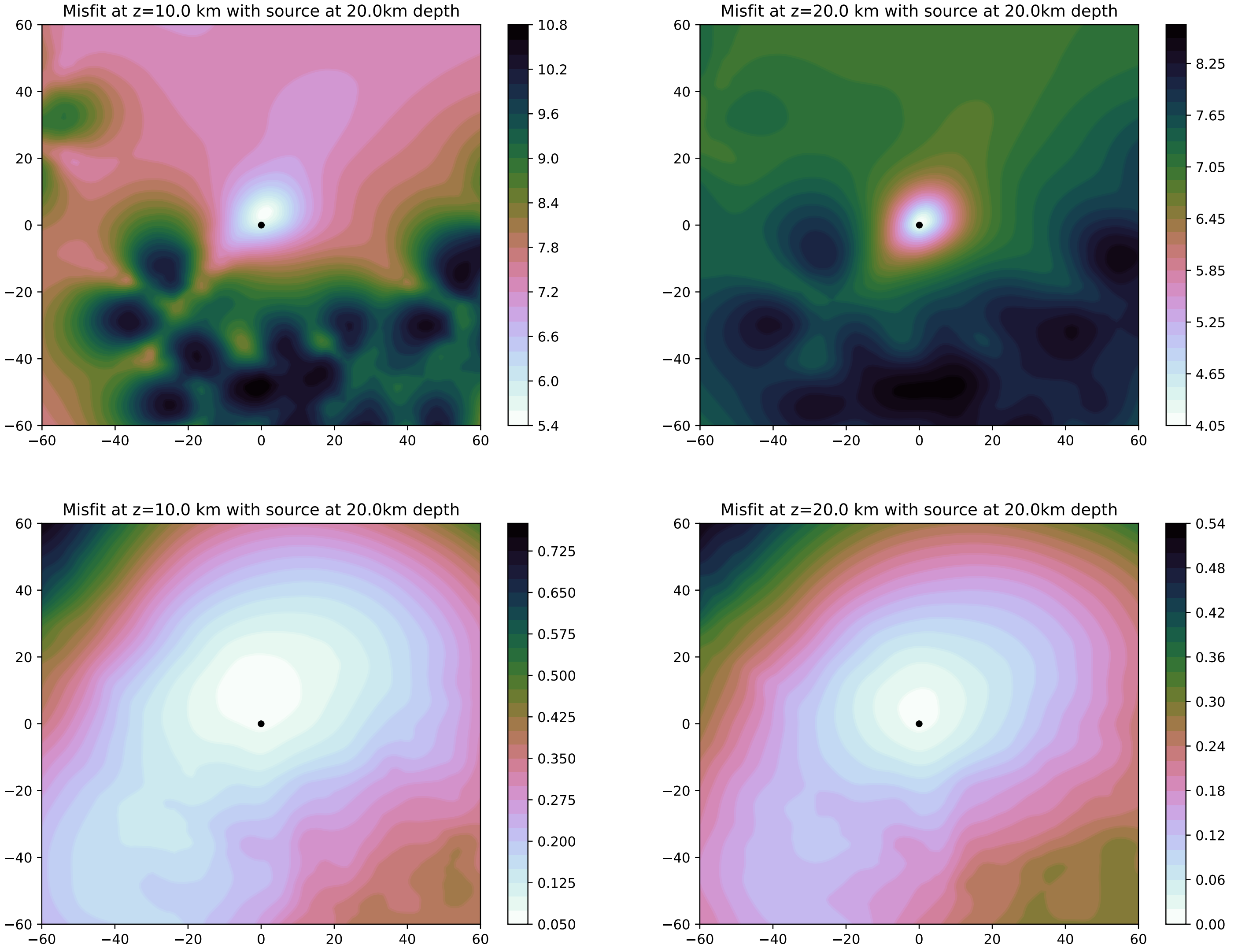


Red line indicates global minimum

Sambridge et al. (2021, under review)

A comparison of misfit surfaces

Seismic waveform misfit surfaces as a function of source position



Misfit at z=10.0 km with source at 20.0km depth

Misfit at z=20.0 km with source at 20.0km depth

Misfit at z=10.0 km with source at 20.0km depth

Misfit at z=20.0 km with source at 20.0km depth

Sum of
squared waveform
differences, L_2

Marginal
Wasserstein
algorithm, W_2^2

Sambridge et al. (2021, under review)

Some conclusions

- Many new developments in inversion of geophysical data translated from other fields
- We can expect new types of signal to be found in geophysical data
- We can expect new ways of performing inversion, but the principles of inversion are unchanged.
- An exciting time for applications of new mathematical and computational tools such as Sparsity, Machine Learning and Optimal Transport.
- This will require multi-skilled people like never before.

We will learn new things by

….doing things in new ways and

…asking new types of question!