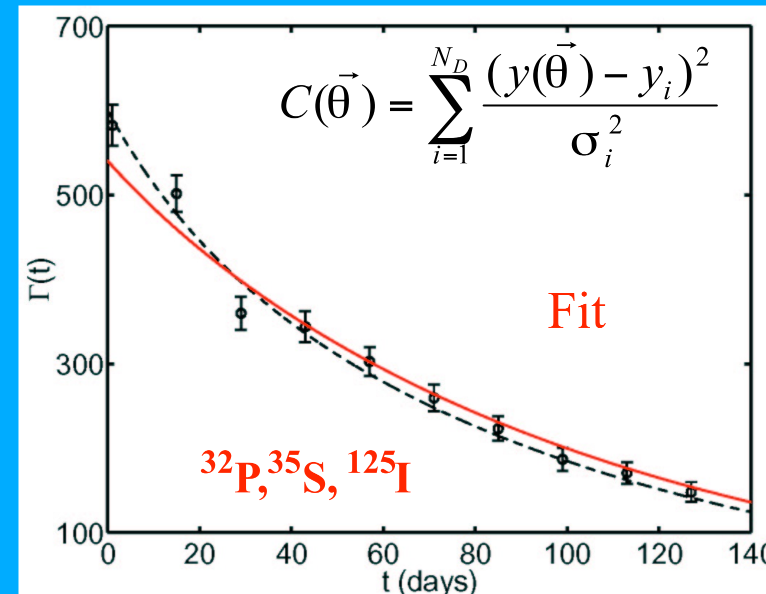
Model Manifolds: Hyperribbons and emergent simplicity

Today we move out of parameter space, to study the curved surface in behavior space that contains all the predictions of our model.

Fitting Decaying Exponentials

Classic ill-posed inverse problem

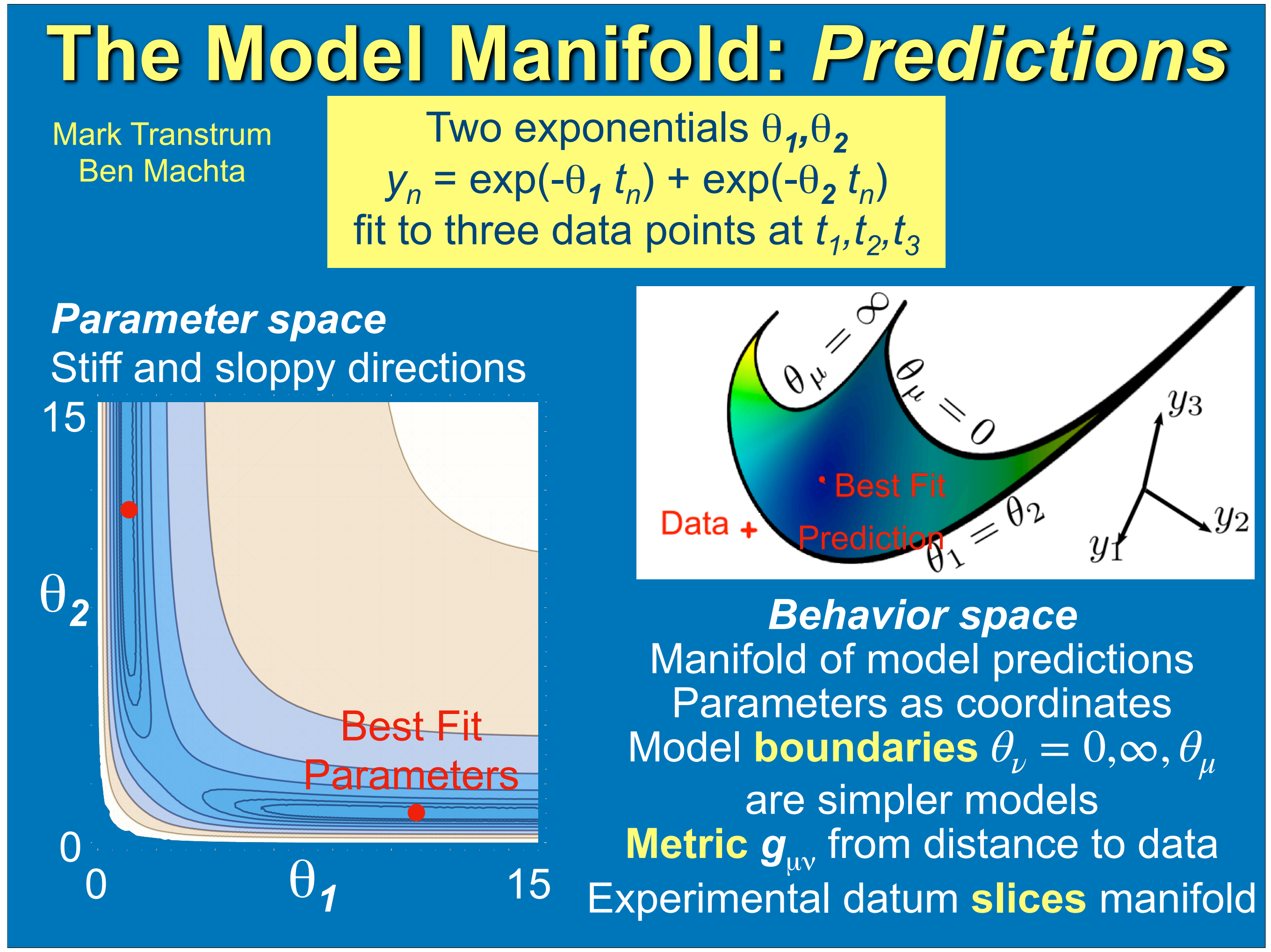
Given Geiger counter measurements from a radioactive pile, can we recover the identity of the elements and/or predict future radioactivity? Good fits with bad decay rates!



$$y(\mathbf{A},\boldsymbol{\gamma},t) = A_1 e^{-\gamma_1 t} + A_2 e^{-\gamma_2 t} + A_3 e^{-\gamma_3 t}$$

6 Parameter Fit

We'll illustrate our model manifold with one of the classic ill-posed problems: extracting the parameters from a sum of three exponentials.



It's useful to look at this 'model manifold' for a two parameter model, in a three-dimensional behavior space. Consider a model with two decay rates, evaluated at three times.

At left we have parameter space: the Hessian at the best fit has a stiff and sloppy direction.

At right we have the model manifold — the set of predictions forms a 2D manifold in the space (y1,y2,y3), with coordinates θ_1 and θ_2.

Later on, we shall be interested in
(1) the edges of the model manifold — simpler models with fewer parameters,
(2) the metric on the model manifold, which is given by the distance in data space (and equals the cost Hessian for a perfect model)
(3) slices of the model manifold given by fitting a data point, say d2 measured at t2. This is given by cutting along the plane y2=d2.

Hessian for perfect data is metric in behavior space

Metric Tensor defined from distance in behavior space

$$|\mathbf{y}(\theta') - \mathbf{y}(\theta)|^2 = \delta_\alpha \delta_\beta g_{\alpha\beta}$$

$$\sum_i (y_i(\theta') - y_i(\theta))^2 = \sum_i \left(\sum_\alpha \delta_\alpha \frac{\partial y_i}{\partial \delta_\alpha} \right)^2$$

For NLLS models

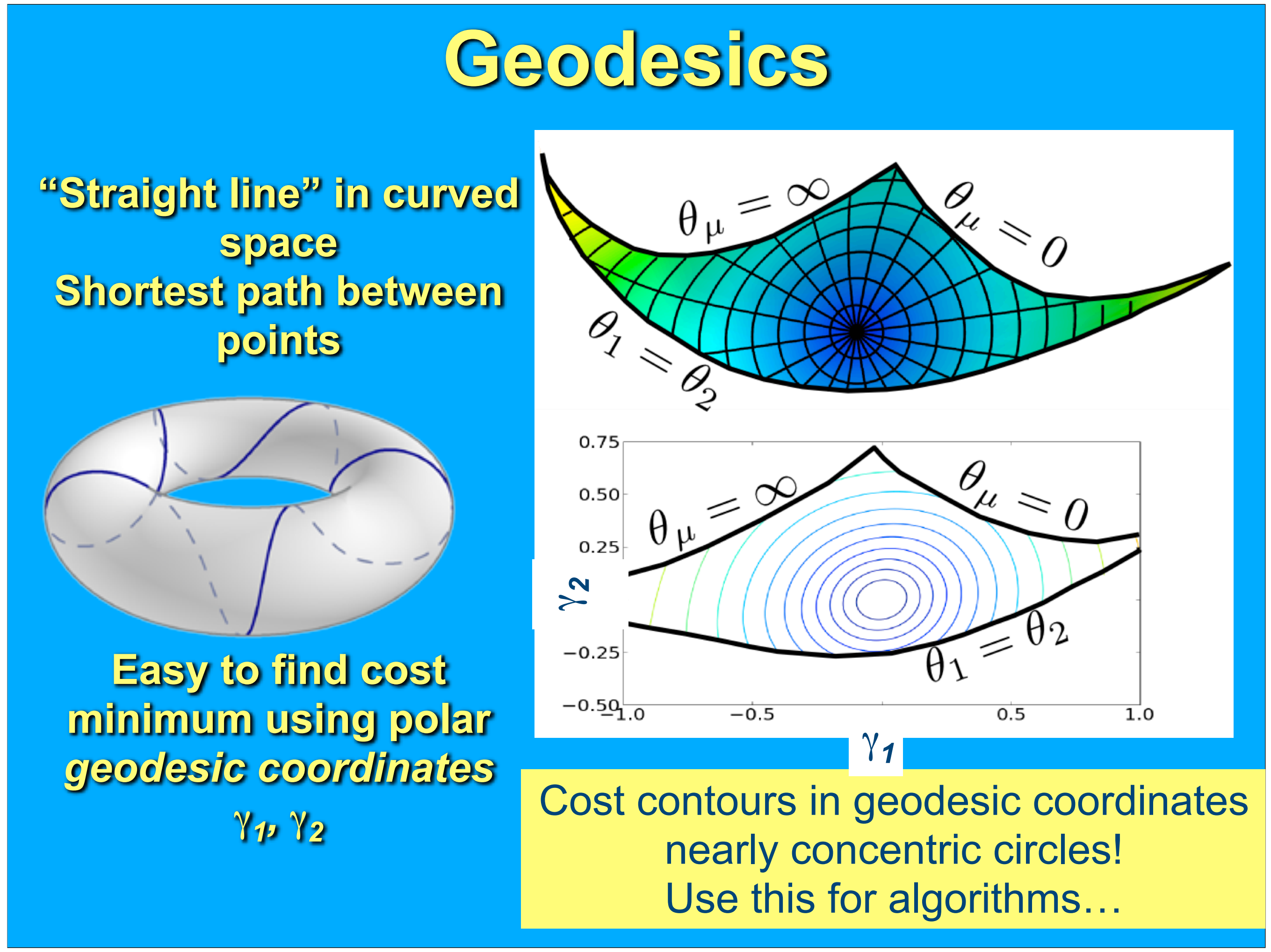
$$= \sum_i \left(\sum_\alpha \delta_\alpha \frac{\partial y_i}{\partial \delta_\alpha} \right) \left(\sum_\beta \delta_\beta \frac{\partial y_i}{\partial \delta_\beta} \right)$$

$$= \sum_{\alpha\beta} \delta_\alpha \delta_\beta J^T_{\alpha i} J_{i\beta}$$

Giving us the approximate cost Hessian: sloppy if J is skewed

$$g_{\alpha\beta} = (J^T J)_{\alpha\beta}$$

The metric tensor on the model manifold is the approximate Hessian we've seen before. The Jacobian is important: it maps perturbations in parameter space into behavior space. The skewness of this mapping gives the large range of eigenvalues of the cost Hessian (and metric tensor), and in the end will give us the hyperribbon structure of the model manifold.



How will we measure and characterize our model manifolds? An important tool is the geodesic — the analogue of straight lines when you live on a curved surface. The geodesic between two points is a shortest path. On the torus, we see a geodesic which isn’t globally the shortest, but locally is.

Given the metric tensor, how do we find the equations to solve for the geodesics?

Geodesics

Minimum length paths on the model manifold

Label path by arbitrary coordinate t: $\boldsymbol{y}(t)$

$$L = \int_a^b dt \sqrt{\Sigma_i \left(\frac{dy_i}{dt}\right)^2} = \int_a^b dt \sqrt{\Sigma_i \left(\frac{\partial y_i}{\partial \theta_\alpha}\dot{\theta}_\alpha\right)\left(\frac{\partial y_i}{\partial \theta_\beta}\dot{\theta}_\beta\right)}$$

$$= \int_a^b dt \sqrt{\Sigma_i J_{i\alpha} J_{i\beta} \dot{\theta}_\alpha \dot{\theta}_\beta} = \int_a^b dt \sqrt{g_{\alpha\beta} \dot{\theta}_\alpha \dot{\theta}_\beta}$$

Can't use calculus of variations! Path length invariant for any parameterization s(t): run a while, then dawdle, …

These notes were written up by Katherine Quinn last time I taught the course. The geodesic equations won't be used in these lectures, and so I include the derivation as an optional sideline. They are less straightforward to derive than one might guess, and are one of the significant results of differential geometry.

This slide writes down the length, but there are an infinite family of minimal paths once you allow the parameters to vary in speed: change coordinates to s(t) and get a new function.

Geodesic equation

At the minimum, the (functional) derivative is 0

Cauchy Schwartz inequality: $(\mathbf{U}\cdot\mathbf{V})^2 \le (\mathbf{U}\cdot\mathbf{U})(\mathbf{V}\cdot\mathbf{V})$.

Let $U(t) = \sqrt{g_{\alpha\beta}\dot{\theta}_\alpha\dot{\theta}_\beta}$, $V(t) \equiv 1$, and $\mathbf{U}\cdot\mathbf{V} = \int dt U(t)V(t)$.

Then $\mathbf{V}\cdot\mathbf{V} = (b-a)$ and $\mathbf{U}\cdot\mathbf{V} = \int dt U(t) \times 1 = L$. So

$$L^2 = (\mathbf{U}\cdot\mathbf{V})^2 \le (\mathbf{U}\cdot\mathbf{U})(b-a) = (b-a)\int_a^b dt g_{\alpha\beta}\dot{\theta}_\alpha\dot{\theta}_\beta$$

Minimizing $E[\boldsymbol{\Theta}(t)] = \int_a^b dt g_{\alpha\beta}\dot{\theta}_\alpha\dot{\theta}_\beta$ gives the minimum path $\boldsymbol{\Theta}(t)$, traversed at a constant speed.* Now we can use calculus of variations.

The inequality is an equality if $\mathbf{U}$ and $\mathbf{V}$ are parallel. Since $\mathbf{V}\equiv 1$, this means the speed, $\mathbf{U}$, is constant.

Instead, we use the Cauchy-Schwartz inequality to both get rid of the nasty square root and to force the minimum to have constant velocity in behavior space.

Geodesic trick

Calculus of variations

$$E[\boldsymbol{\Theta}] = \int_a^b dt g_{\alpha\beta}\dot{\theta}^\alpha\dot{\theta}^\beta$$

Add small perturbation to the path

$$0 = \Delta E = E[\boldsymbol{\Theta} + \boldsymbol{\Delta}] - E[\boldsymbol{\Theta}] = E[\boldsymbol{\Theta} + \boldsymbol{\Delta}] - E[\boldsymbol{\Theta}]$$

$$= \int_a^b dt\, \delta^\gamma \partial_\gamma g_{\alpha\beta}\dot{\theta}^\alpha\dot{\theta}^\beta + g_{\alpha\beta}\dot{\theta}^\alpha\dot{\delta}^\beta + g_{\alpha\beta}\dot{\delta}^\alpha\dot{\theta}^\beta + O(\delta^2)$$

$$= \ldots$$

Integrate last two by parts, factor out δ(t), set rest to zero

$$g_{\gamma\beta}\ddot{\theta}^\beta = -1/2\,(\partial_\alpha g_{\gamma\beta} + \partial_\beta g_{\alpha\gamma} - \partial_\gamma g_{\alpha\beta})\dot{\theta}^\alpha\dot{\theta}^\beta = -\Gamma_{\gamma\alpha\beta}\dot{\theta}^\alpha\dot{\theta}^\beta$$

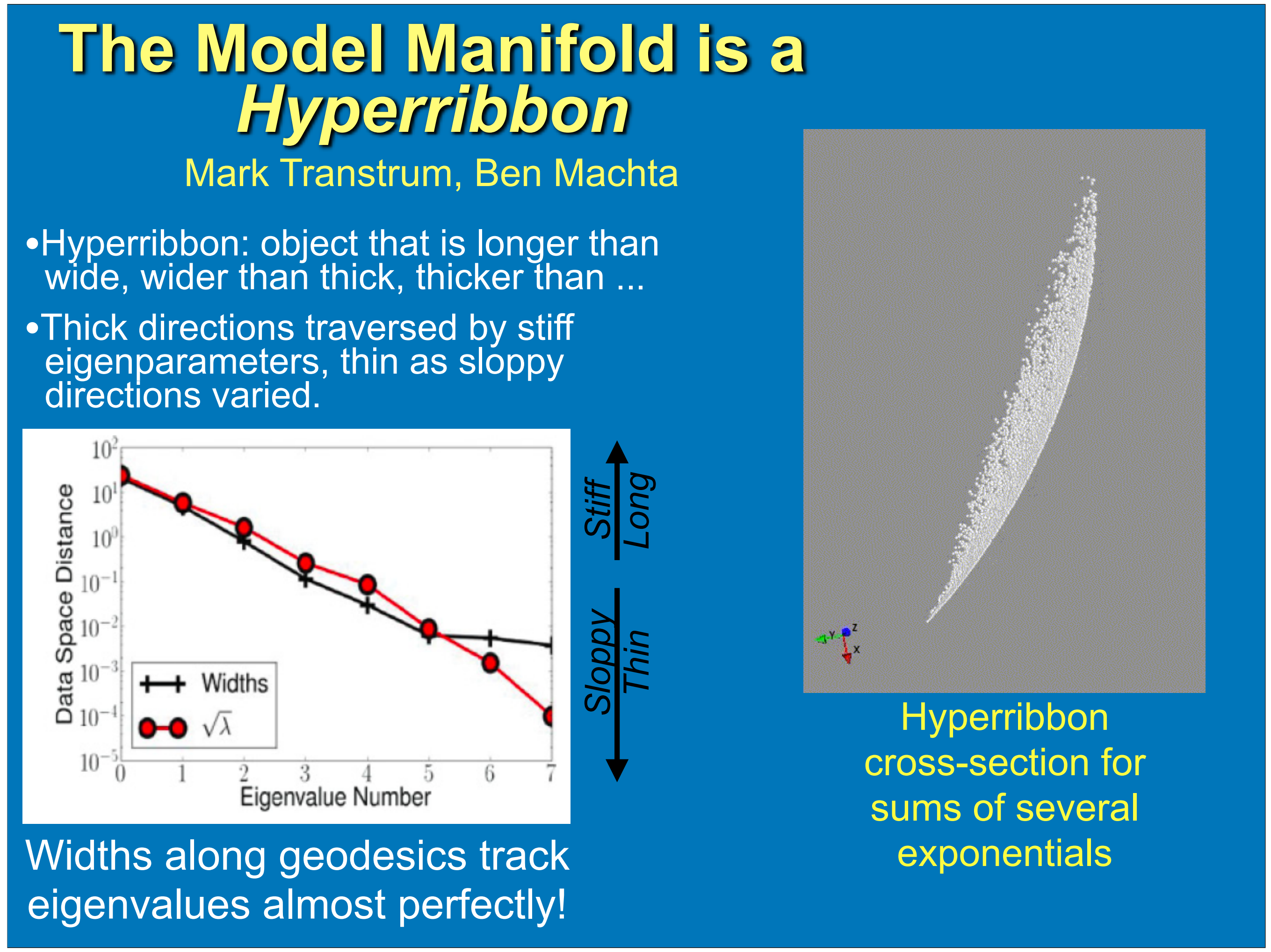
$$\ddot{\theta}^\mu = -g^{\mu\gamma}\Gamma_{\gamma\alpha\beta}\dot{\theta}^\alpha\dot{\theta}^\beta = -\Gamma^\mu_{\alpha\beta}\dot{\theta}^\alpha\dot{\theta}^\beta$$

Giving us the geodesic equation in terms of the Christoffel symbols $\Gamma^\mu_{\alpha\beta}$

This introduces the rather complicated formula for the Christoffel symbol in terms of derivatives of the metric tensor, which makes for a nice geodesic equation.

The Christoffel symbols are needed whenever one is comparing things at different points on the model manifold.

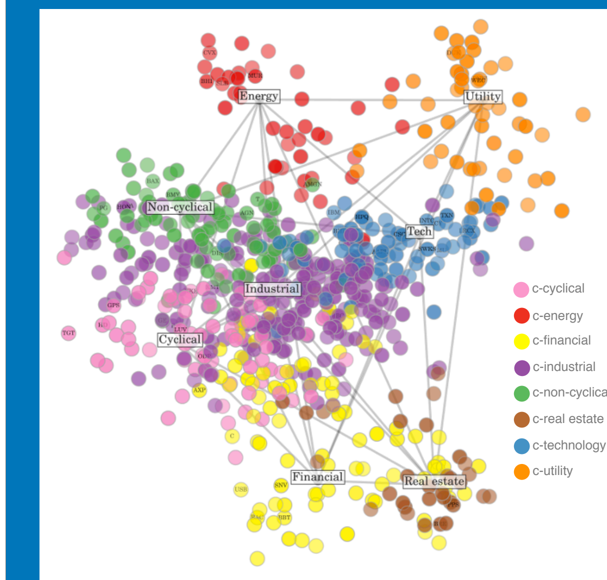
Note that all my thetas should have upper indices here, and

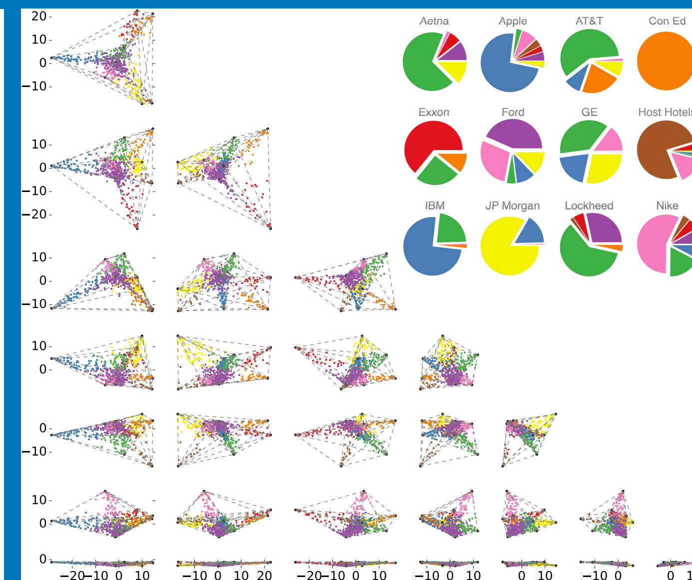


When we have many parameters fit to many data points, the model manifold becomes a hyperribbon. A hyperribbon is like a hypersphere or a hypercube, but longer than wide and wider than thick, … The long direction corresponds to the stiffest eigenvalue. If we start at the best fit, and measure the widths of the manifold using geodesics pointed along the eigendirections of the cost Hessian, we find that the widths nicely track the square roots of the eigenvalues. So the hierarchy of sloppiness in parameter space can be explained if we understand why the model forms a hyperribbon in behavior space.

Stock prices form a hyperribbon

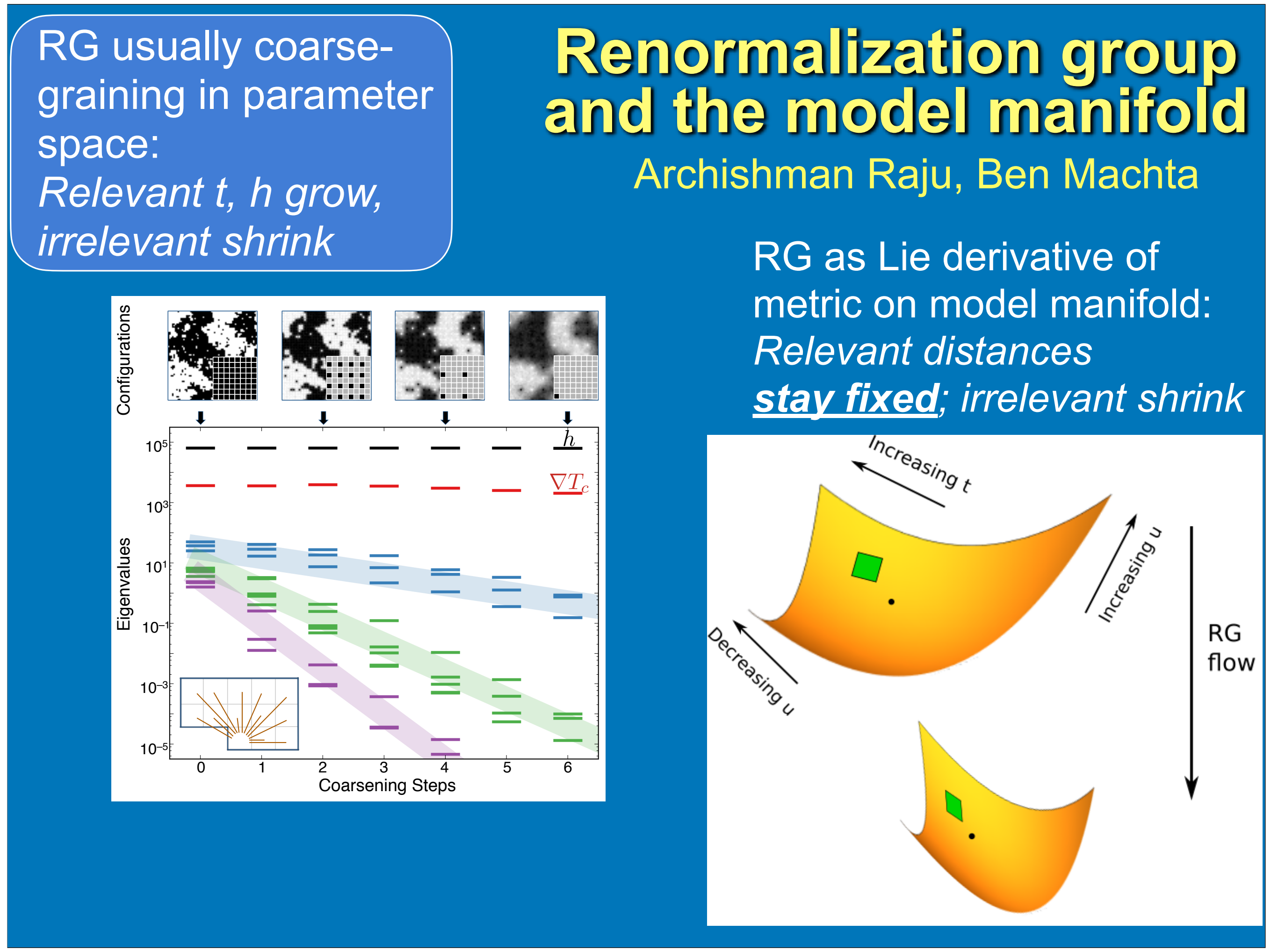
Lorien X. Hayden, Ricky Chachra, Alexander A. Alemi, Paul H. Ginsparg, Alen Senanian, Noé Beserman





Nine 'stiff' directions distinguishable from noise.
Sectors of the economy are vertices. Why low dimensional?
Why does PCA work?

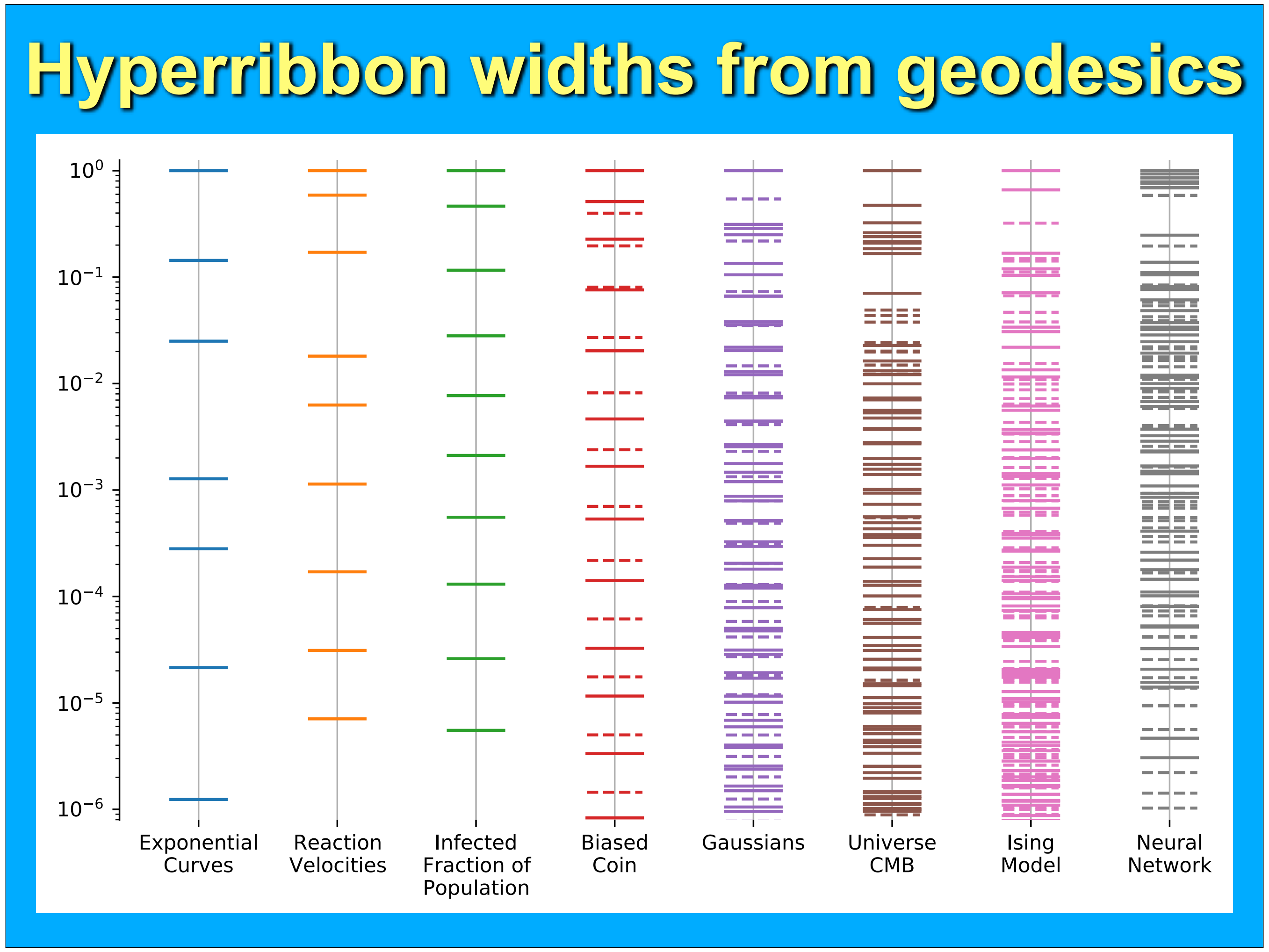
Marc Potters and JP Bouchaud, ages ago, wrote a paper noting that the main principal components of stock market prices had strong overlaps with the sectors of the economy. Inspired by this work, we found that a better description is given by a hyper-tetrahedron, allowing companies to be divided into percentages of, say tech and non-cyclical sectors (IBM). My emphasis here is that in a 50000 dimensional space of price trajectories, a nine-dimensional hyperribbon describes most of the explainable variation, allowing a useful emergent description into sectors.



What would the renormalization-group flow do to the model manifold?

* The Ising model manifold is known to have a cusp at the critical point, given by the curvature of $g_{\mu\nu}$.
* The intensive embedding for the Ising model can be derived, conveniently, from the free energy.
* The RG acts as a flow on the model manifold.
* Irrelevant perturbations shrink in behavior space, as they do in parameter space.
* Relevant distances stay the same under coarse graining, unlike they do in parameter space.
* Under coarse-graining, new interactions are formed in the RG. These could become subsumed into finding the coordinates on the model manifold after coarse-graining.

It would be fascinating to visualize this flow using the techniques of Lecture 6.



We’ve seen this figure before. The hierarchy of widths we see here is why we call the model manifold a hyperribbon. It should remind you of the hierarchy of metric eigenvalues. In most models with sensible definitions of the parameters, the square roots of the eigenvalues roughly correspond to the widths of the model manifold.

Hyperribbons = Emergent Simplicity

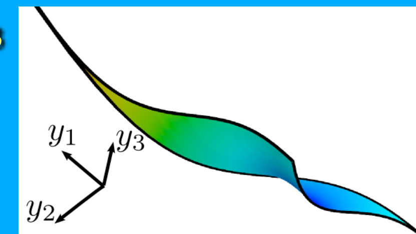
I'm going to speculate wildly here.

* Occam's razor: MBAM will derive a series of simpler models using the hyperribbon structure, allowing coarse-grained simpler models.
* How science can work: Explain the big picture, let others prove that more and more details are important.
* Wisdom: If only a few degrees of freedom matter, experience will guide you into what changes will help.
* Superstition: Often your intuition is fleshed out into a story, which may have predictive power without having any microscopic basis.

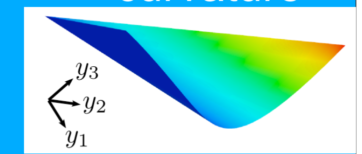
Curvatures

Intrinsic curvature $R^{\mu}{}_{\nu\alpha\beta}$

- determines geodesic shortest paths
- independent of embedding, parameters

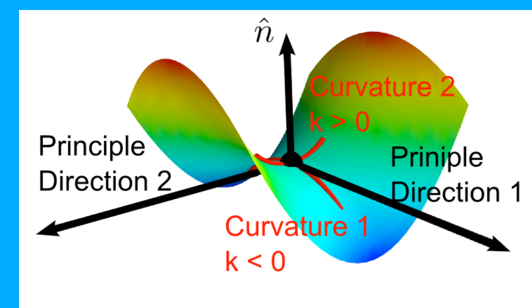


No intrinsic curvature



Extrinsic curvature

- also measures bending in embedding space (i.e., cylinder)
- independent of parameters
- Shape operator, geodesic curvature

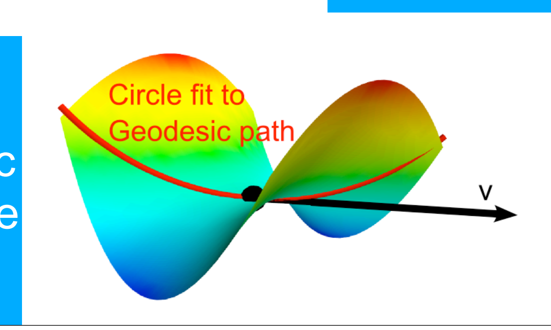


Shape Operator

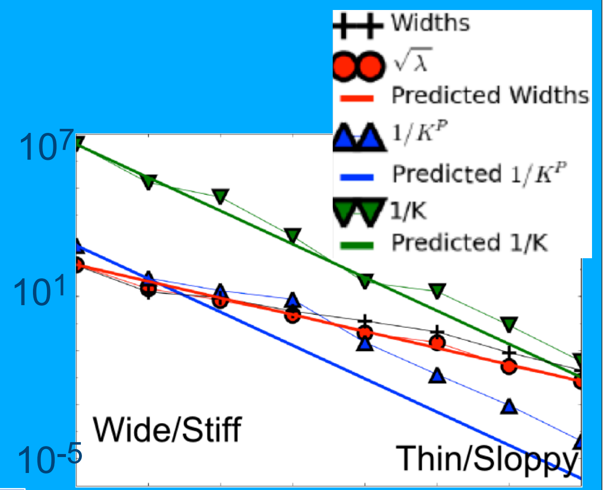
Parameter effects "curvature"

- Usually much the largest
- Defined in analogy to extrinsic curvature (projecting out of surface, rather than into)

Geodesic Curvature

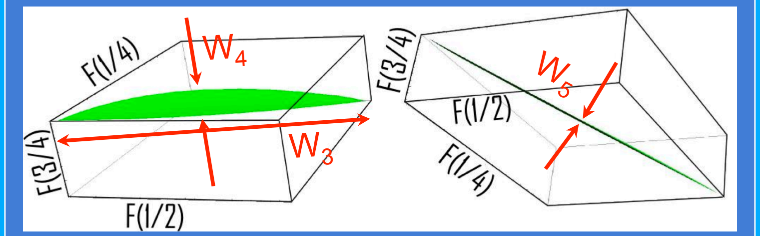


Hierarchy of widths and curvatures



Eigendirection at best fit

Hierarchy of widths



Cross sections: fixing f at 0, ½, 1

Multi-decade span of widths, curvatures, eigenvalues

Widths ~ $\sqrt{\lambda}$ sloppy eigs

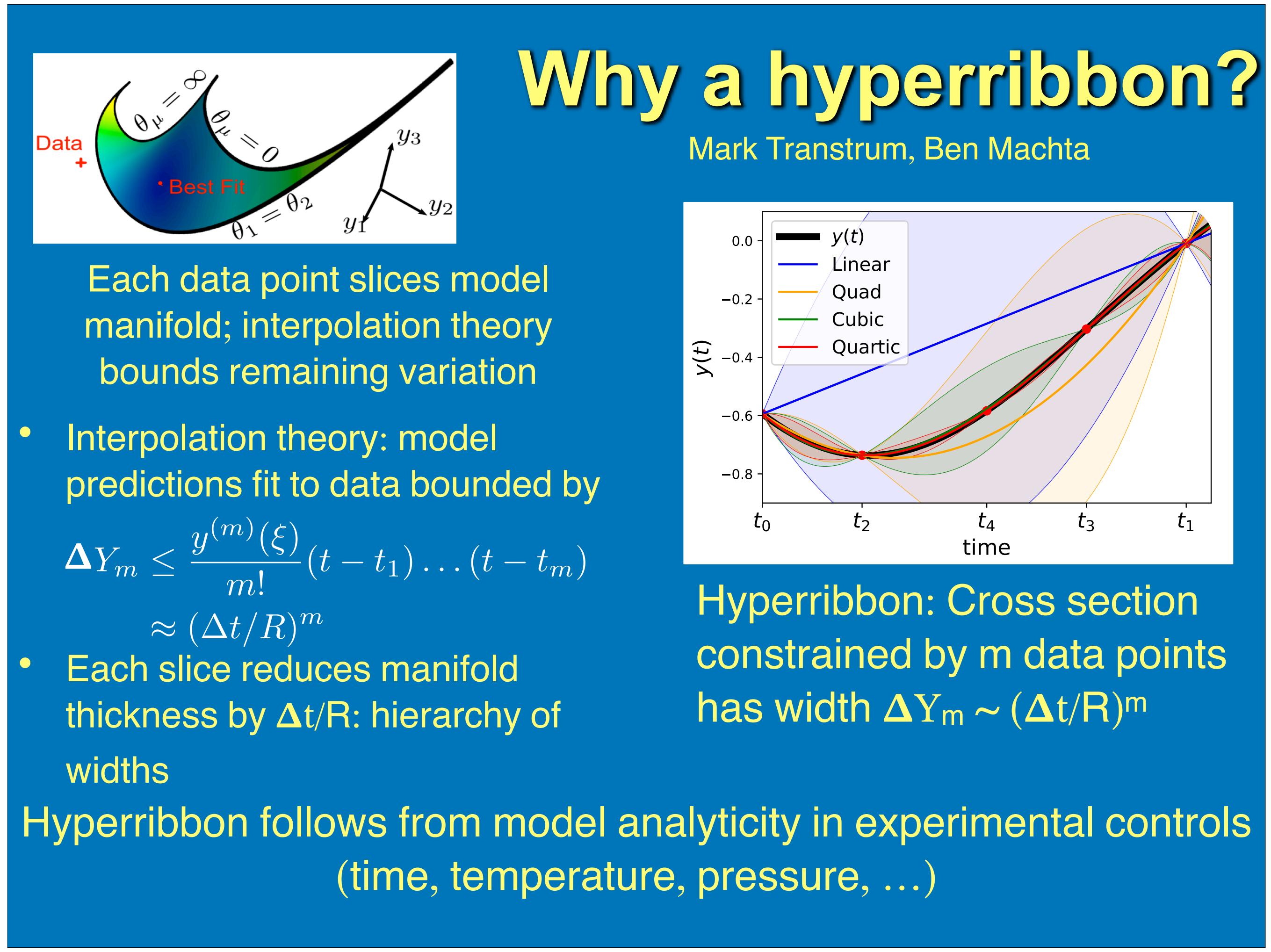
Parameter curvature
$K^P = 10^3 \times K$
>> extrinsic curvature

Why a hyperribbon?

Physical argument
and rigorous proof

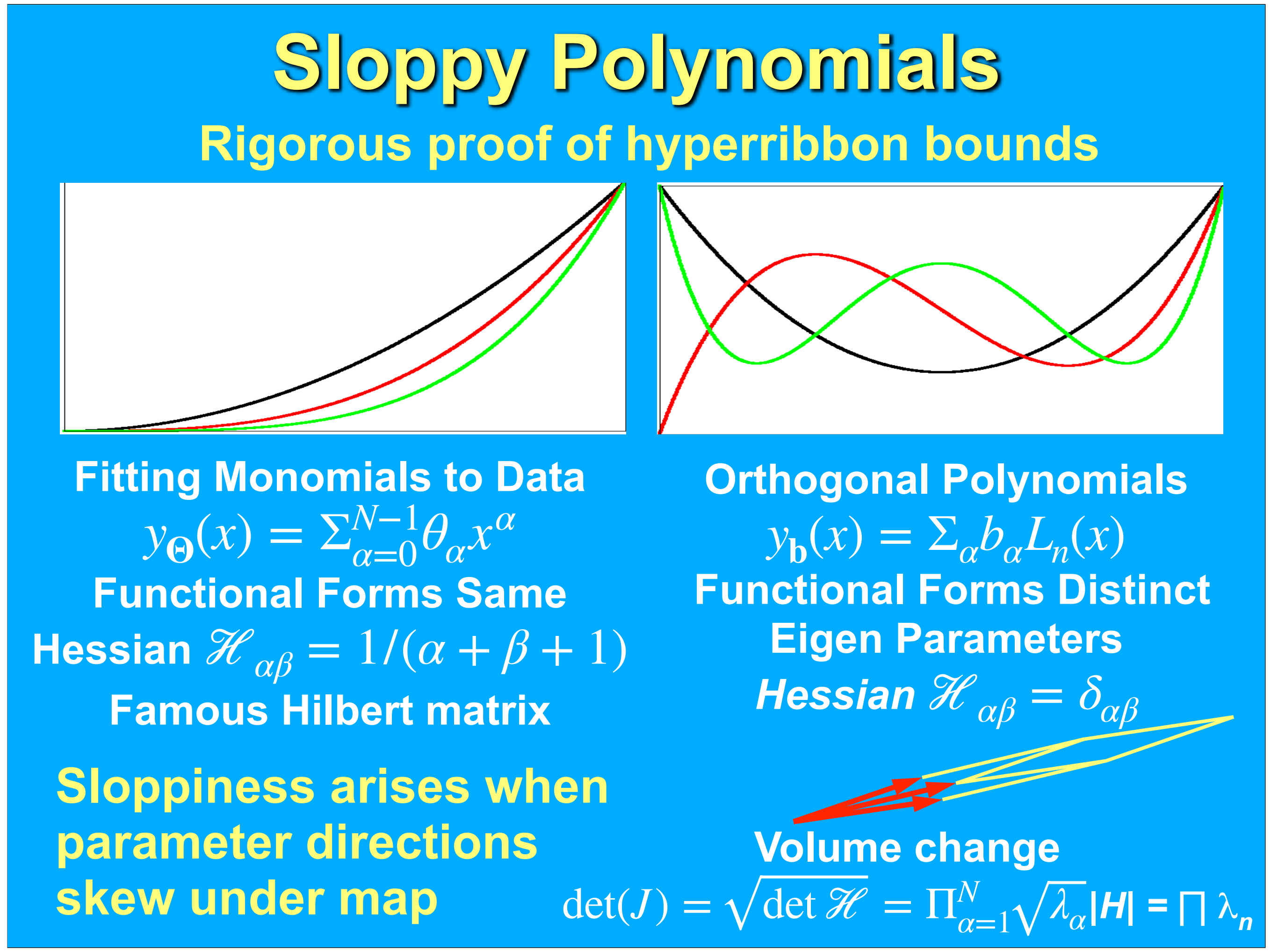
Katherine Quinn, Friday!

I may have convinced you that models have behaviors that form hyperribbons, and the hyperribbons explain the sloppy behavior of parameters. Now I need to explain why this occurs.



Three steps to understand the basic argument that multiparameter models have hyperribbons (due to Transtrum).
(1) Each experiment is a dimension of the model manifold. Adding a data point (say at t2), slices the model manifold along the plane y(t2) = d2. Slicing a hyperribbon cuts off the longest direction. Can we show adding a data point makes the remaining range of predictions shrink?
(2) Interpolation theory tells us that a model fit to m data points, that everywhere has a m^{th} derivative bounded by $m!\,R^{-m}$, will have interpolated values that differ no more than $\Delta Y_m = R^{-m}$ times a polynomial that vanishes at all of the fitted points.
(3) Thus adding another data point reduces the variations of all the other points, by an amount given by the range $\Delta t/R$.

Thus the widths of the model manifold decrease by a geometrical factor every time it is sliced — making it a hyperribbon.



Sloppiness also arises in linear fits. In exercise “Sloppy monomials”, you will see that the mapping between polynomial coefficients and the resulting behavior is sloppy. But the model manifold for a linear fit is a plane — as the monomial coefficients go to infinity, the behavior keeps changing.

Polynomials have biggest range of predictions given Taylor series bounds

Range $$Y_m \leq \frac{y^{(m)}(\xi)}{m!}(t-t_1)\dots(t-t_m)$$ Poly Range

Model manifold for polynomials with bounded gradients $y^{(m)}(t)$? Easier: spherical bound on coefficients $\theta_m = y^{(m)}(0)/m!$

Exercise 'Monomial Hyperribbons'

Jacobian $J_{t\alpha}$ is Vandermonde matrix;

Hamiltonian $\mathscr{H}_{\alpha\beta}$ is Hilbert matrix.

Tiny eigenvalues = Skew mapping, takes sphere $\Sigma\theta_\alpha^2 < R^2$ into hyper-ellipsoid

It turns out that this sloppiness is the key to our rigorous proof that nonlinear models are sloppy, and to understanding under what conditions we know that it will be so. This will be discussed in the exercise "Monomial hyperribbons".

Vandermond Matrix is Sloppy

Katherine Quinn, Heather Wilbur, Alex Townsend

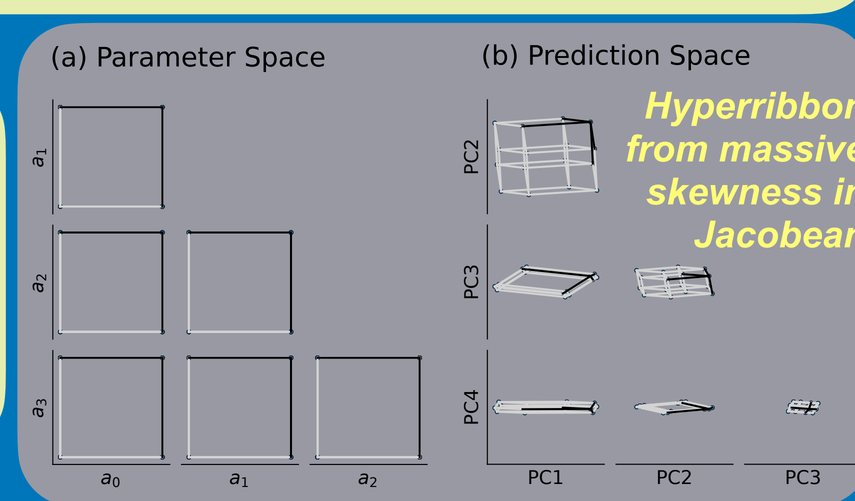
$$\begin{bmatrix} f_{N-1}(t_0) \\ f_{N-1}(t_1) \\ \dots \\ f_{N-1}(t_{N-1}) \end{bmatrix} = \begin{bmatrix} a_0 + a_1 t_0 + \dots + a_{N-1} t_0^{N-1} \\ a_0 + a_1 t_1 + \dots + a_{N-1} t_1^{N-1} \\ \dots \\ a_0 + a_1 t_{N-1} + \dots + a_{N-1} t_{N-1}^{N-1} \end{bmatrix} = \begin{bmatrix} 1 & t_0 & \dots & t_0^{N-1} \\ 1 & t_1 & \dots & t_1^{N-1} \\ \dots & \dots & \dots & \dots \\ 1 & t_{N-1} & \dots & t_{N-1}^{N-1} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \dots \\ a_{N-1} \end{bmatrix}$$

$$f_{N-1}(\mathbf{t}) = V(\mathbf{t})\mathbf{a}$$

$$\det(V) = \prod_{0 \le i \le j < N} (t_i - t_j)$$

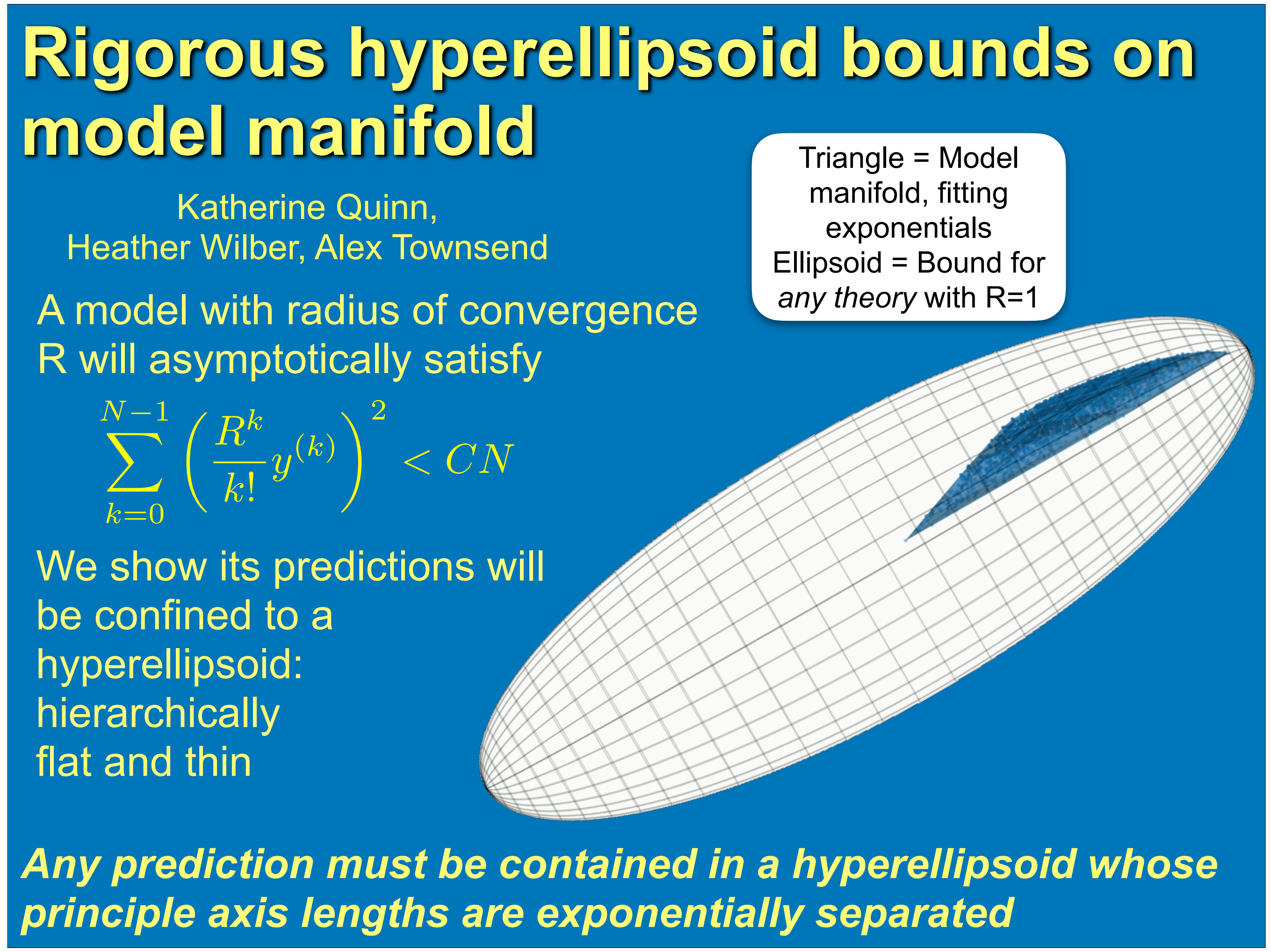
$$\sim (\Delta T)^{N(N-1)/2}$$

$$= \prod_{n=1}^{N} [\lambda_n \sim (\Delta T)^n]$$

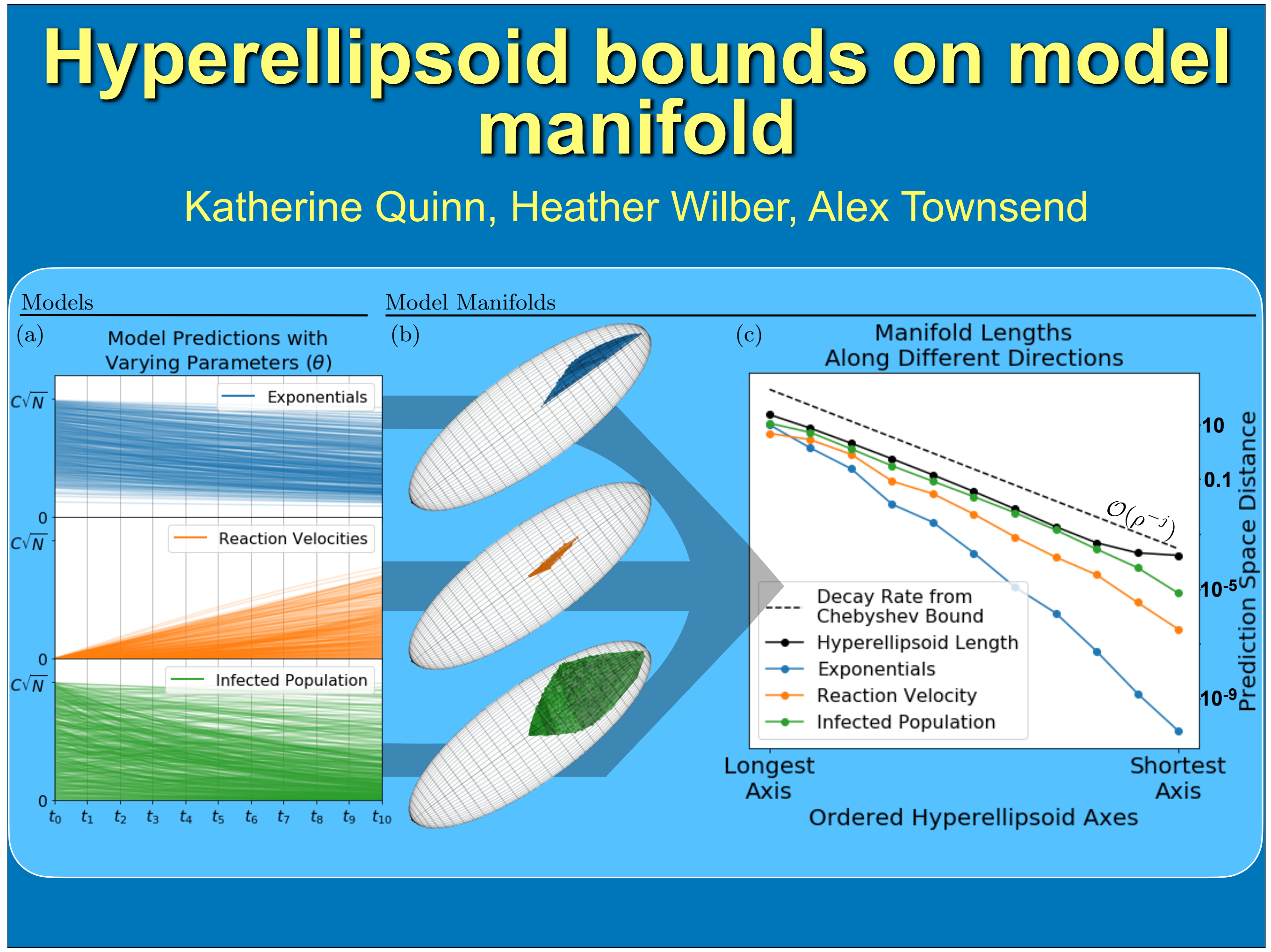


Taylor series R=1, approx function in range ΔT = 0.9

We can turn this into a rigorous proof by using the properties of the Vandermonde matrix. Consider a multiparameter function f(t), whose Taylor series has coefficients bounded by one (hence with a radius of convergence one). Suppose we fit f(t) to a function over a range 0.9. We can view the truncated Taylor expansion as a map of a hypercube of coefficients **a** into a space of function values **f**, whose expansion is given by a Vandermonde matrix. This matrix is famous for having a tiny determinant, and indeed can be proven to have eigenvalues roughly equally spaced in log — the image of the cube is an incredibly skewed parallelepiped. This, plus a small fuzz around it from the truncation, shows that the space of possible model predictions is a hyperribbon.



In collaboration with mathematicians, Katherine Quinn has shown that any model which has smooth derivatives in the experimental control variables (time, temperature, pressure, concentration) will have model manifolds that are enclosed in hyperellipsoids. The axes depend on the radius of convergence of the model and the experimental conditions being predicted. This theorem forces the model manifold (blue) to not only be a hyperribbon, but a flat hyperribbon.



Katherine Quinn, working with mathematicians Heather Wilber and Alex Townsend, proved that any model which has smooth derivatives in the experimental control variables (time, temperature, pressure, concentration) will have model manifolds that are enclosed in hyperellipsoids. The axes depend on the radius of convergence of the model and the experimental conditions being predicted. This theorem forces the model manifold (blue) to not only be a hyperribbon, but a flat hyperribbon.

Here are three different models (physics, chemistry, and epidemiology), each evaluated at ten equally spaced times. All of them share the same hyperellipsoid bounds. At right you see the exponential decays of the successive widths of their model manifolds, showing that they are all hyperribbons. You see that the hyperellipsoid widths and our rigorous bounds also decay geometrically.