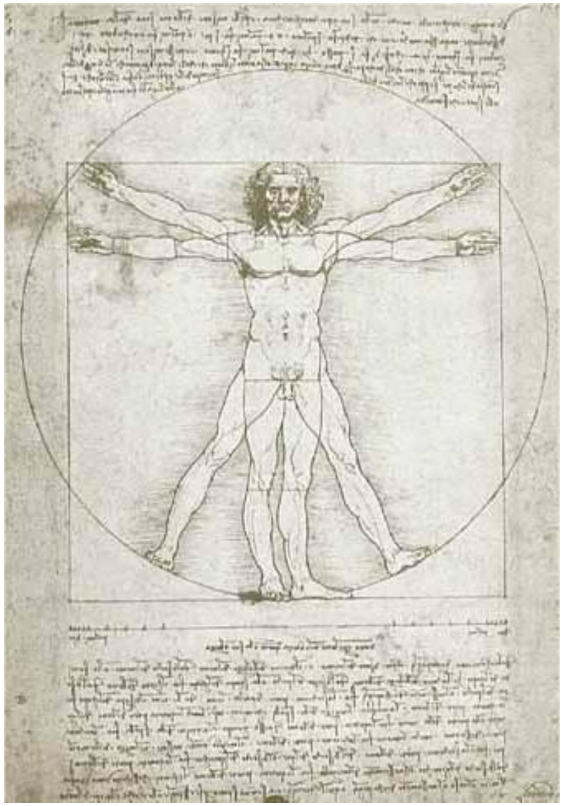
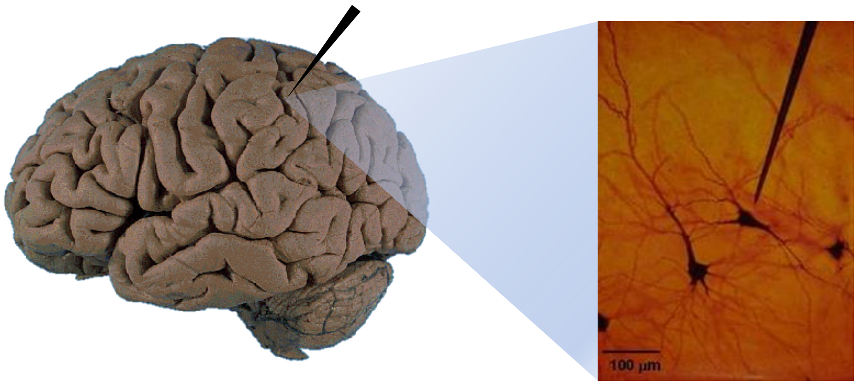
A Class of Efficiently Constructable Minimax Optimal Sequential Predictors

Todd P. Coleman
(with Sanggyun Kim and Diego Mesa)
Department of Bioengineering, UCSD

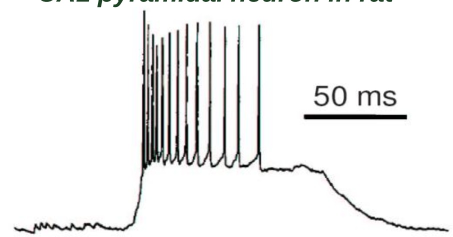


Multiple Neural Data

- Single electrode

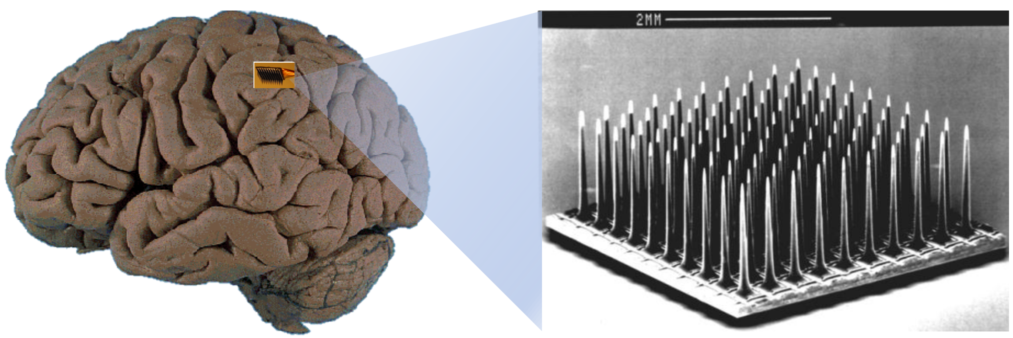


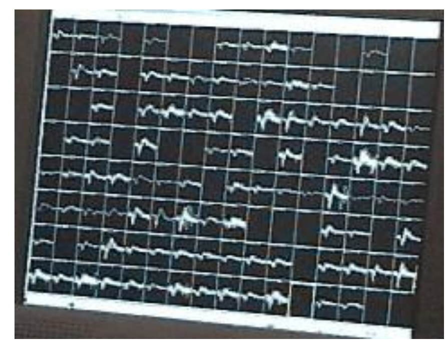
CA1 pyramidal neuron in rat



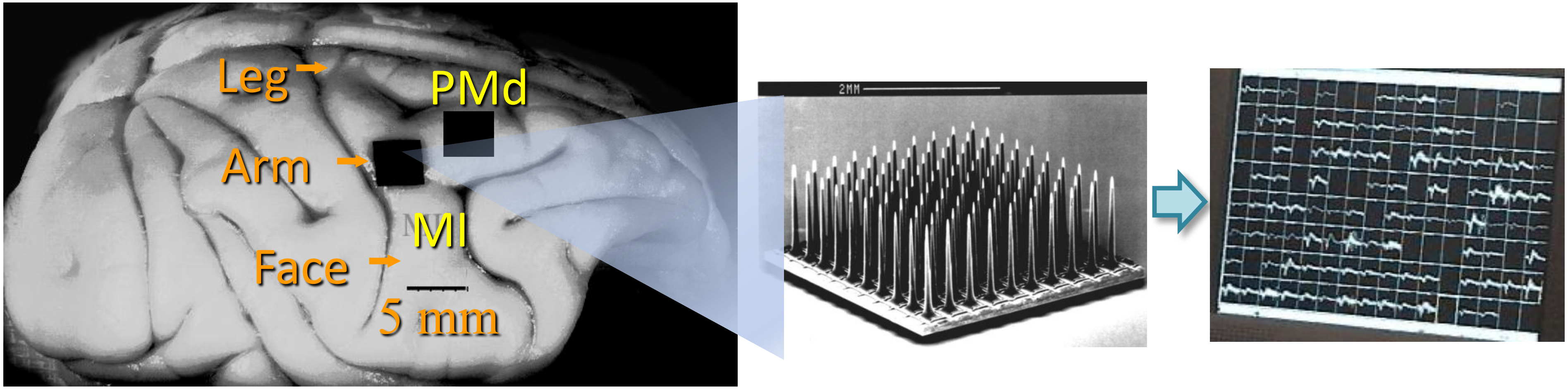
From (Su et al., J. Neurosci., 21: 4173-4182, 2001)

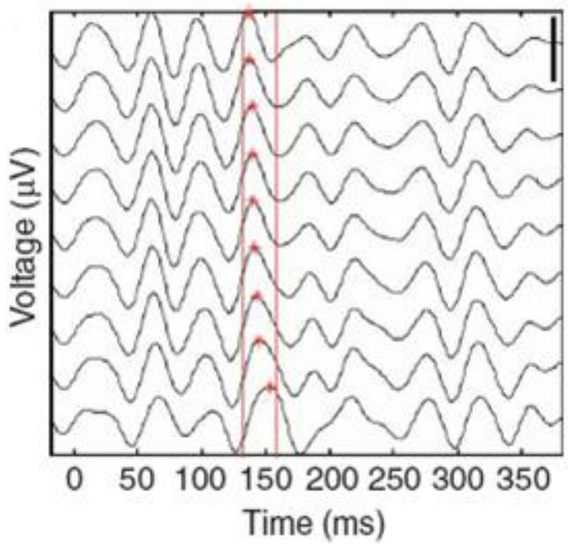
- Multiple electrodes

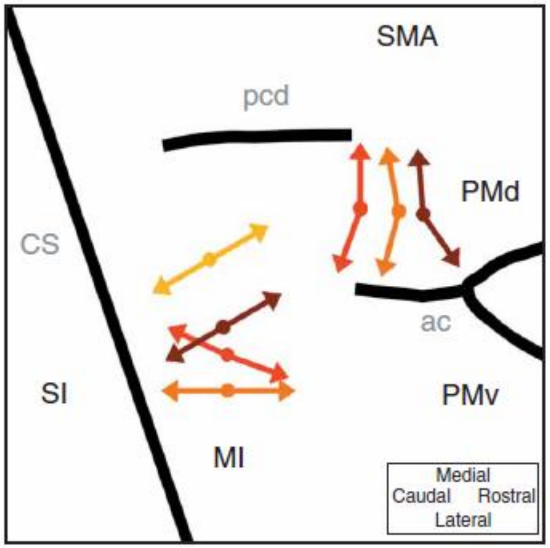




Wave Propagation: LFPs







beta frequency (15-40 Hz) oscillations in subthreshold activity

Rubino, Hatsopoulos, "Propagating waves Mediate Info Transmission in Motor Cortex", *Nature Neuroscience*, 2006

Takahashi, Saleh, Penn, Hatsopoulos, "Propagating waves in human motor cortex", *Frontiers in Human Neuroscience*, 2011

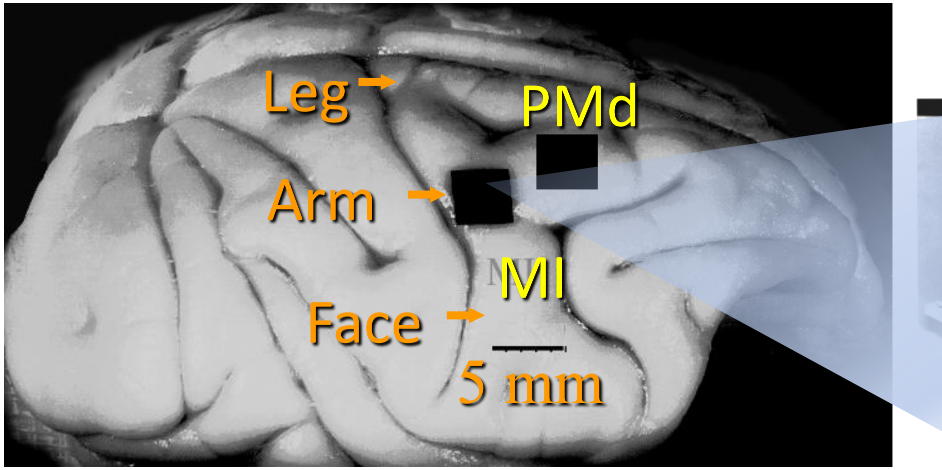
Wave Propagation: LFPs

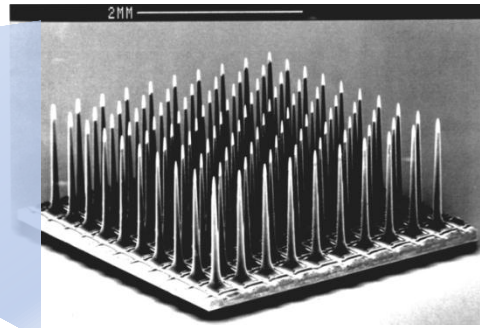
beta frequency (15-40 Hz) oscillations in subthreshold activity

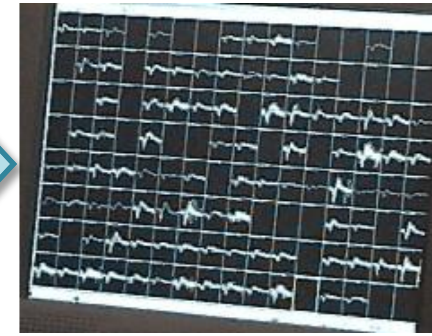
Rubino, Hatsopoulos, “Propagating waves Mediate Info Transmission in Motor Cortex”, *Nature Neuroscience*, 2006

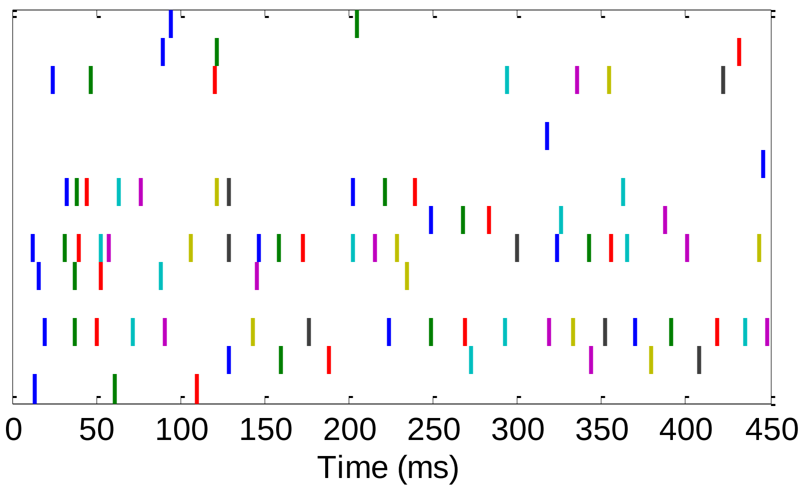
Takahashi, Saleh, Penn, Hatsopoulos, “Propagating waves in human motor cortex”, *Frontiers in Human Neuroscience*, 2011

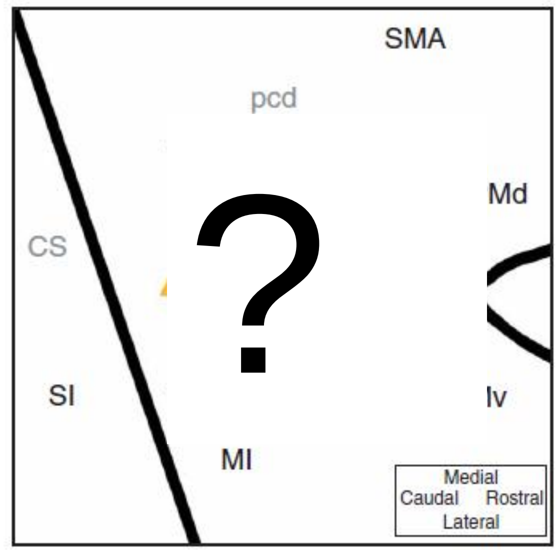
Wave Propagation: **Spikes**





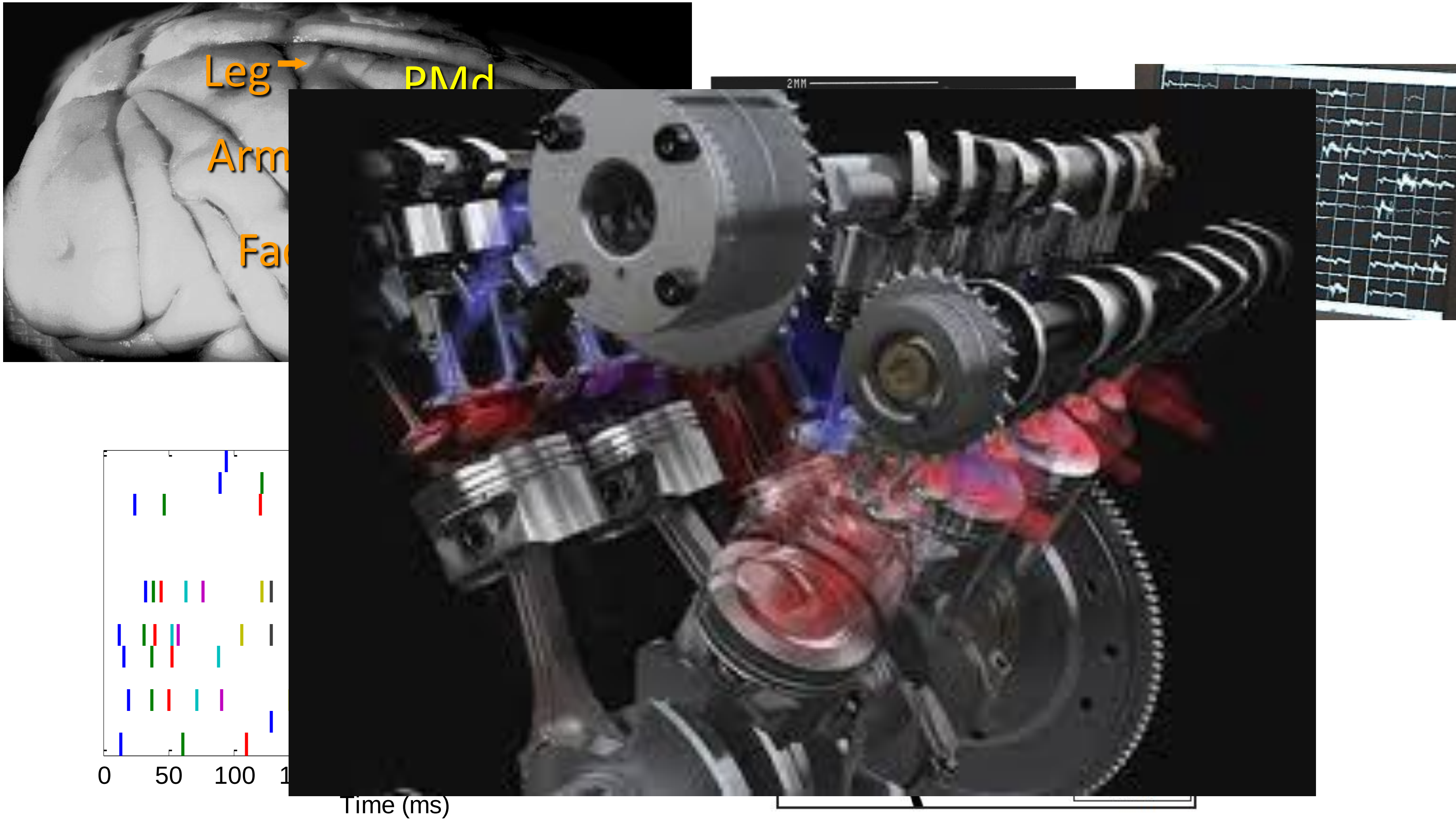






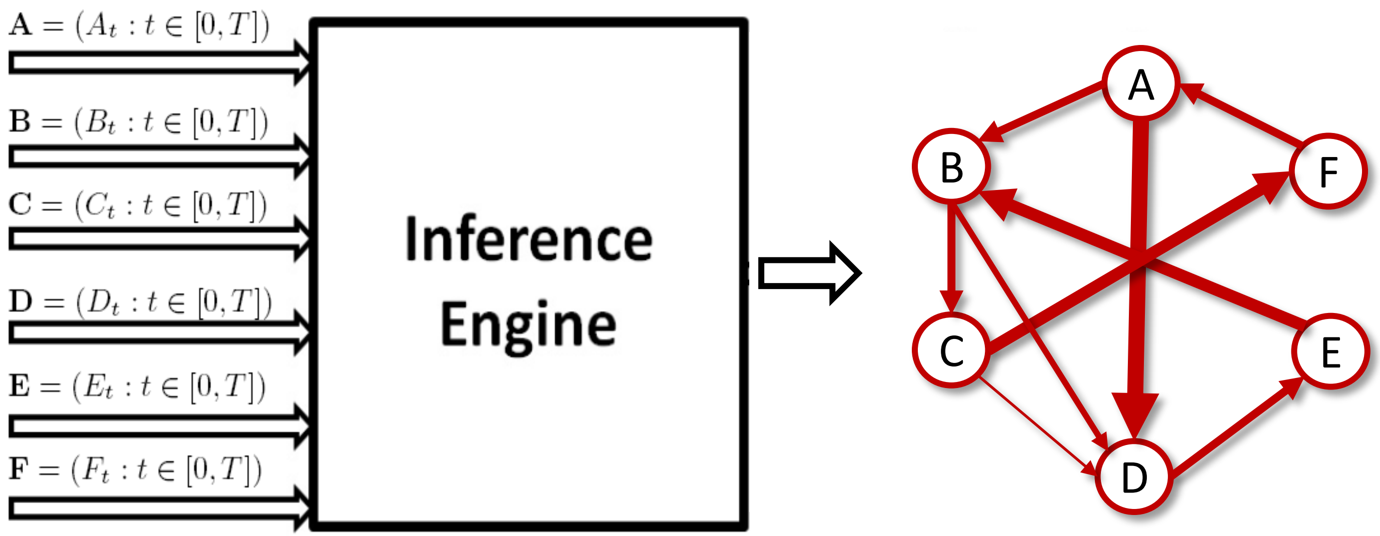
Q: Do action potentials mediating inter-cellular communication demonstrate spatio-temporal patterning consistent with wave propagation?

Wave Propagation: **Spikes**

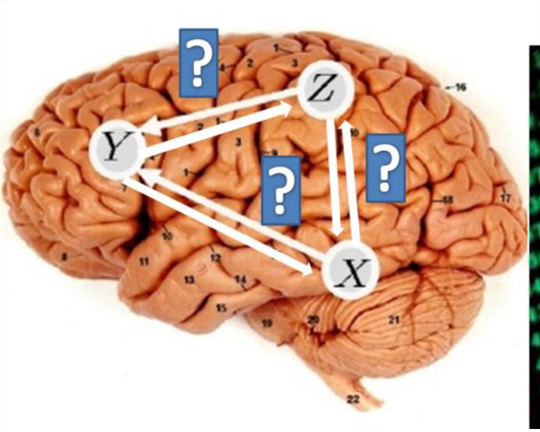


Q: Do action potentials mediating inter-cellular communication demonstrate spatio-temporal patterning consistent with wave propagation?

Beyond Correlation: Causation



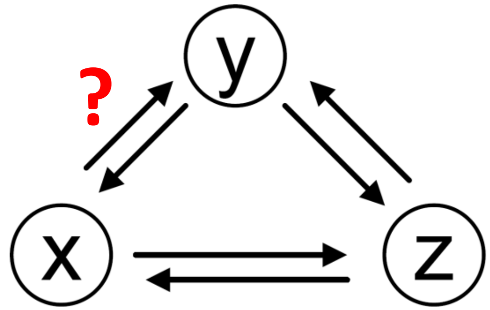
Idea: map a set of M time series to a directed graph with M nodes where an edge is placed from a to b if the past of a has an impact on the future of b

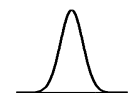


How do we quantitatively do this in a general purpose manner?

Revisiting Granger's *viewpoint* from 30,000 ft

We say that X is causing Y if we are **better** able to **predict** the **future** of Y using all available information than if the information apart from the **past** of X had been used



1. Sequential predictors: *beliefs* 
2. Log loss: $l(b,y) = -\log b(y)$
3. Better: $b_t^* = P_{Y_t|Y^{t-1},Z^{t-1}}$, $\tilde{b}_t^* = P_{Y_t|X^{t-1},Y^{t-1},Z^{t-1}}$

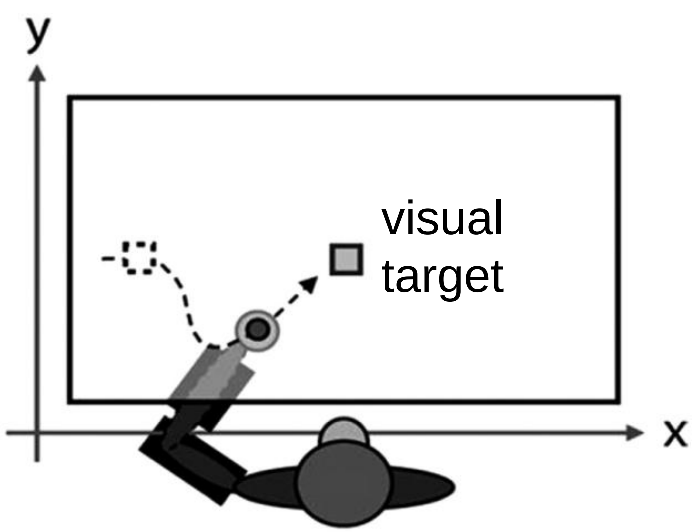
$$\bar{R}(b^*,\tilde{b}^*) = \sum_{t=1}^{T} \mathbb{E}\left[D\left(\tilde{b}_t \| b_t\right)\right] = I(X^T \to Y^T \| Z^T) > 0?$$

- *Granger causality is a special case*
- *Applicable to any modality (point processes)*
- *Can build efficient estimators w/ standard statistical assumptions*

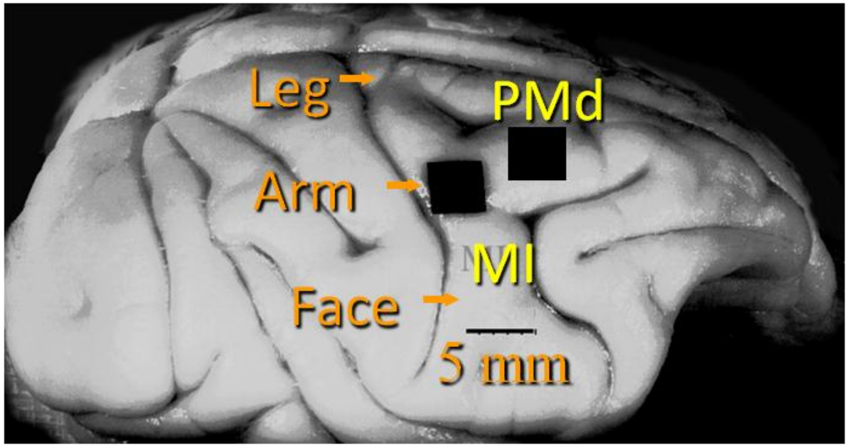
Quinn, Kiyavash, N. Hatsopoulos, **TPC**, *J. Computational Neuroscience, 2010*

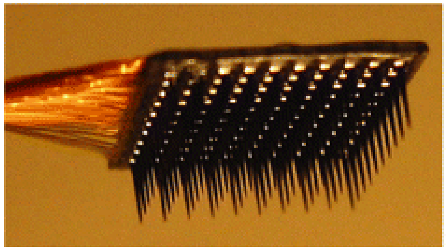
Experimental Paradigm

- Shoulder and elbow movements in horizontal plane
- Monkey: random-target pursuit task.

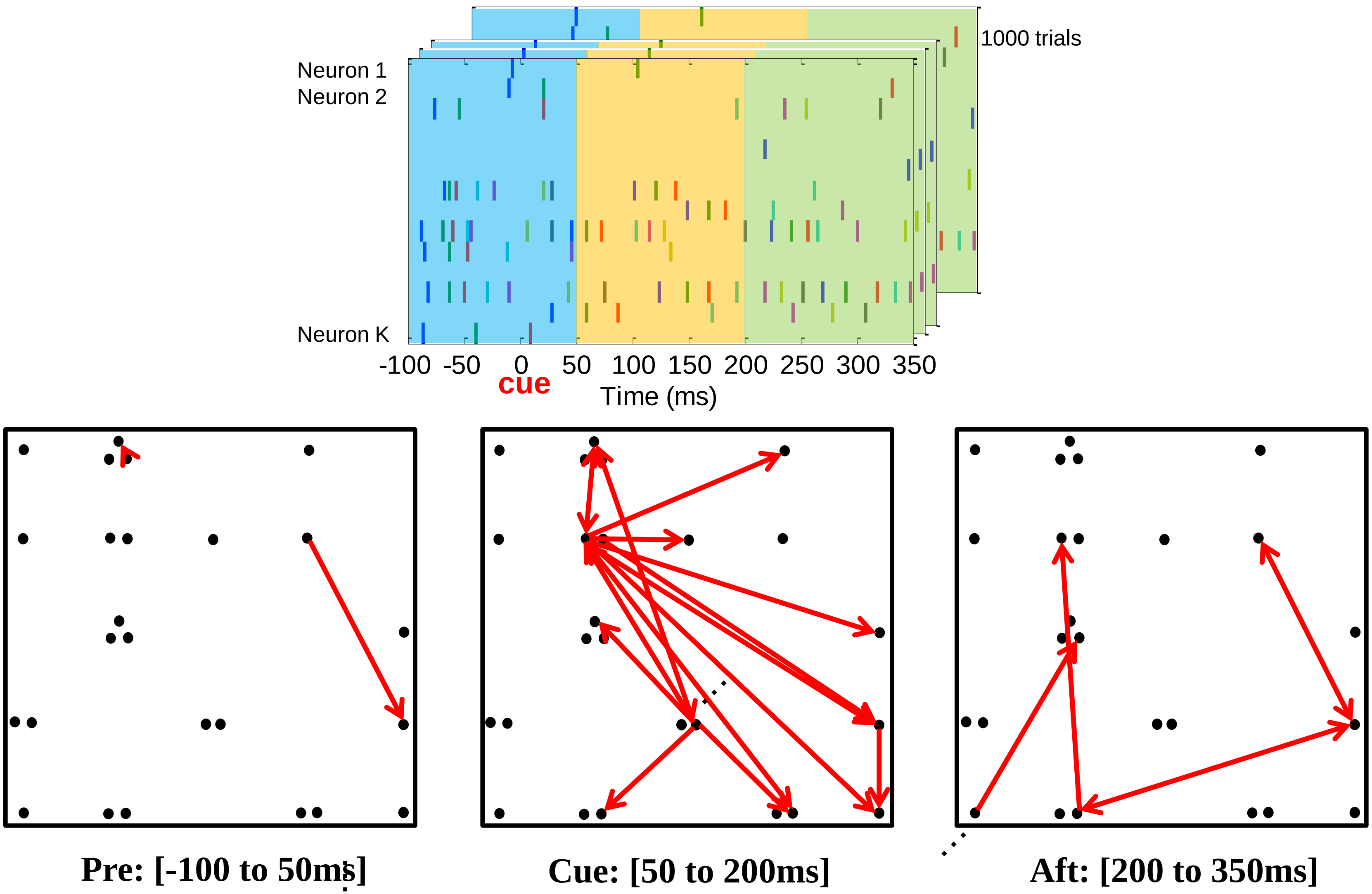


- Recorded ensemble neural processes in MI
- Spacing: 400 um
- ~96 LFPs, ~110 neurons





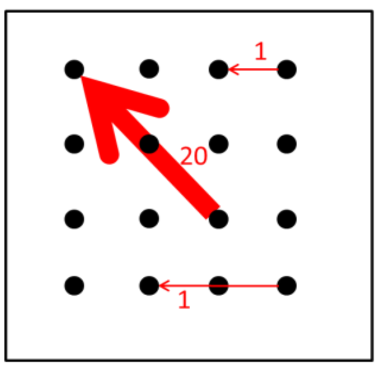
Causal Networks Before/After Cues

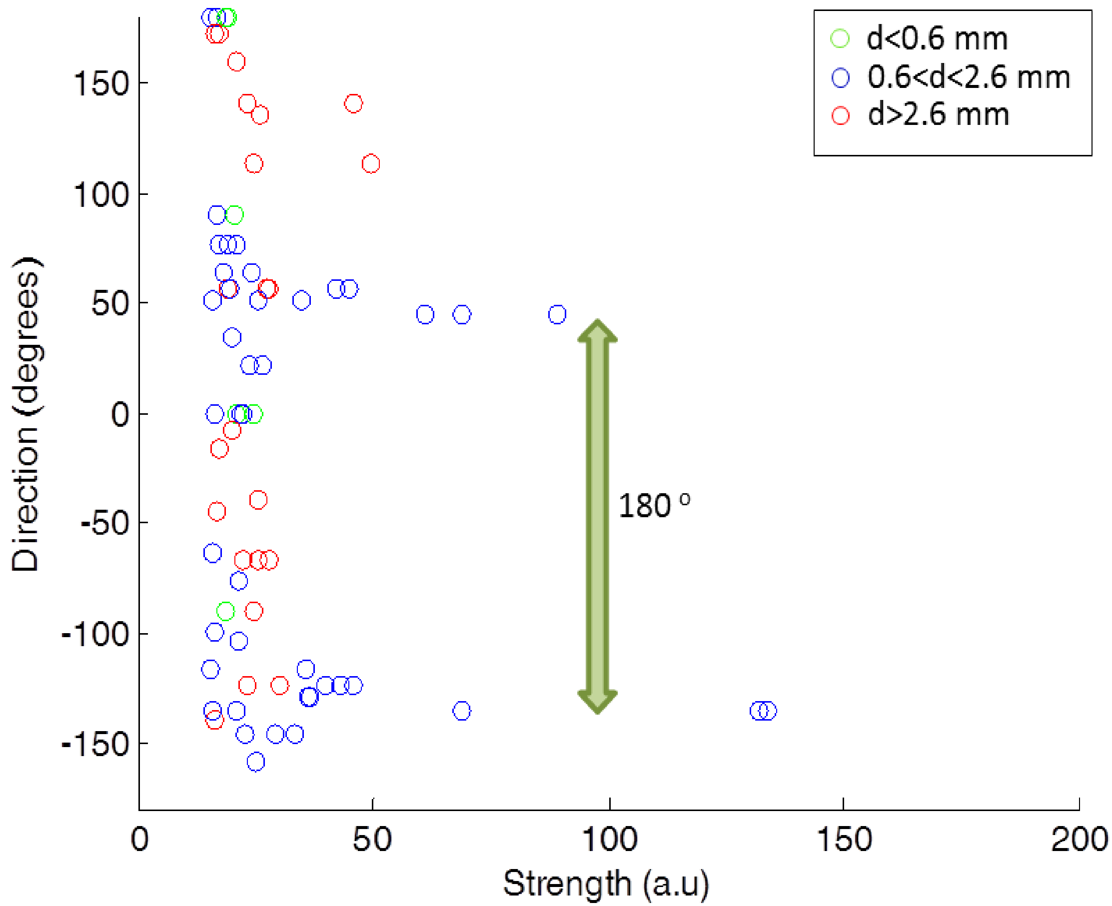


S. Kim, K. Takahashi, N. Hatsopoulos, and **TPC**, "Information Transfer Between Neurons in the Motor Cortex Triggered by Visual Cues", *IEEE Engineering in Medicine and Biology Society Annual Conference*, Sep 2011.

Causal Interaction: Strength, Direction, Distance

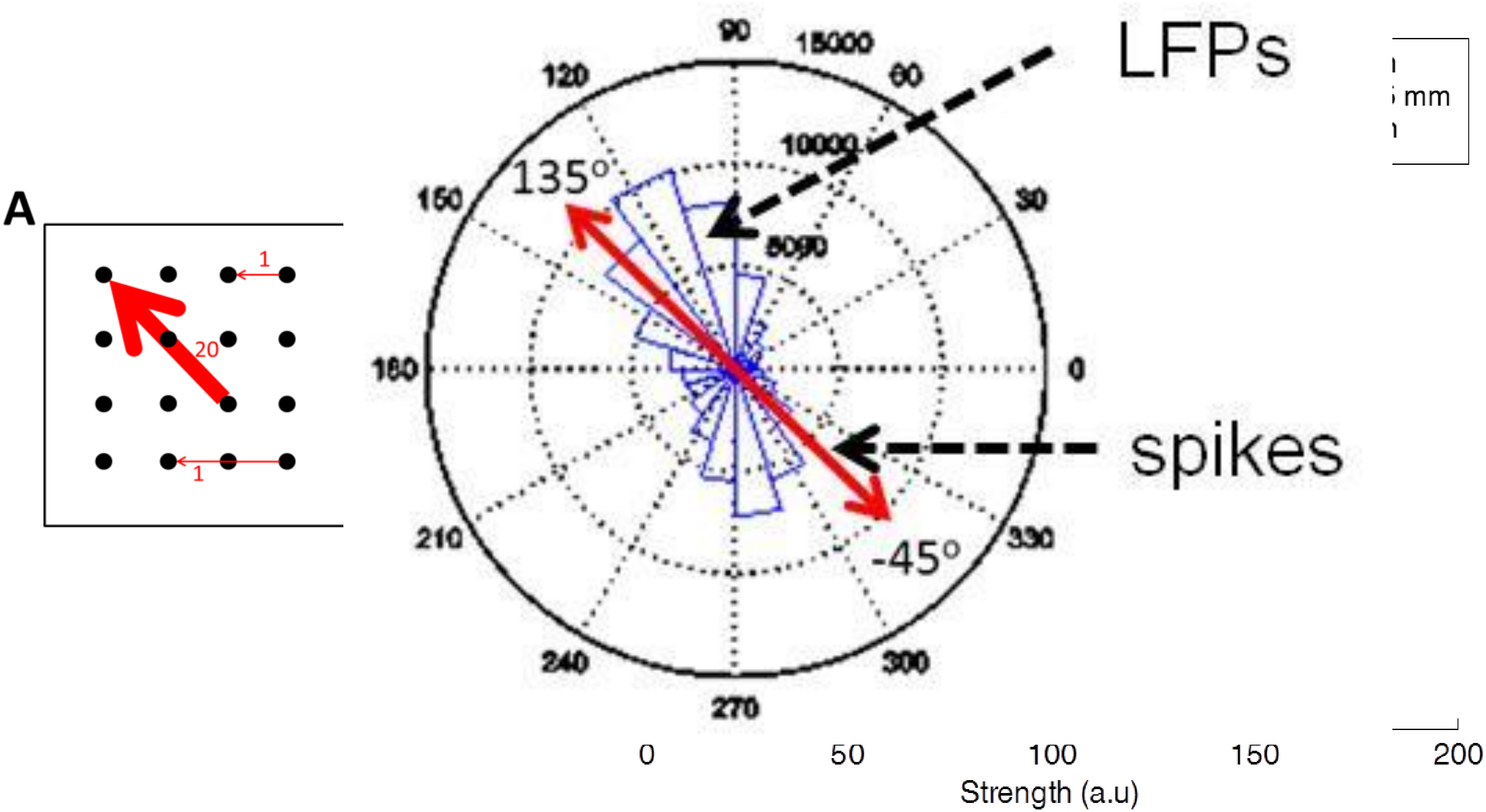
A





K. Takahashi, S. Kim, **TPC**, and N. Hatsopoulos, "Large-scale spike sequencing associated with wave propagation in motor cortex ", in preparation.

Causal Interaction: Strength, Direction, Distance



K. Takahashi, S. Kim, **TPC**, and N. Hatsopoulos, "Large-scale spike sequencing associated with wave propagation in motor cortex ", in preparation.

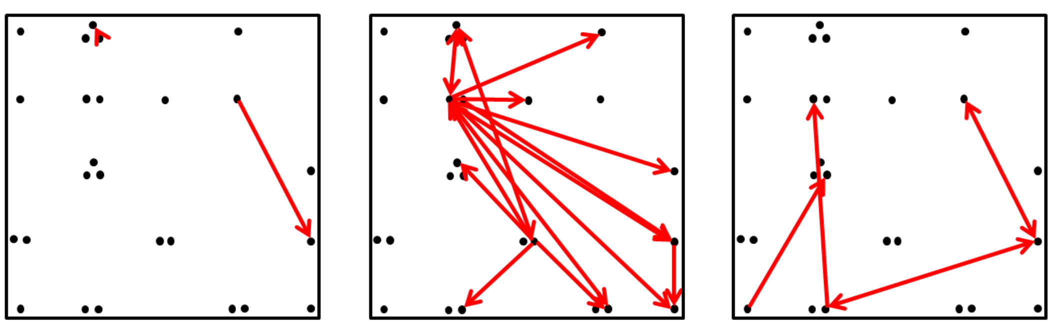
Dynamics of the Dynamics?

(Proposed)

non-static model, any modality

(Conventional)

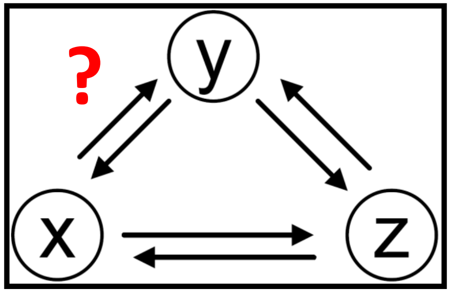
static model,
continuous-valued data



Pre: [-100 to 50ms] Cue: [50 to 200ms] Aft: [200 to 350ms]

$$p_t = P_{Y_t|Y^{t-1},Z^{t-1}} \qquad \tilde{p}_t = P_{Y_t|Y^{t-1},Z^{t-1},X^{t-1}}$$

Old



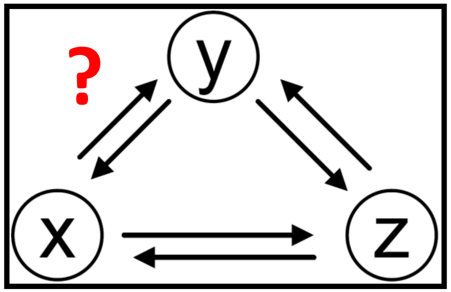
$$\boxed{\frac{1}{T} I(X \to Y \| Z) = \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[D(\tilde{p}_t \| p_t)]}$$

Pros: Finite-dimensional learning problem

Cons: a static graph

$$p_t = P_{Y_t|Y^{t-1},Z^{t-1}} \qquad \tilde{p}_t = P_{Y_t|Y^{t-1},Z^{t-1},X^{t-1}}$$

Old

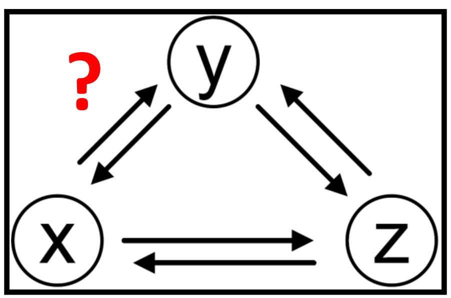


$$\boxed{\frac{1}{T} I(X \to Y \| Z) = \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[D(\tilde{p}_t \| p_t)]}$$

Pros: Finite-dimensional learning problem

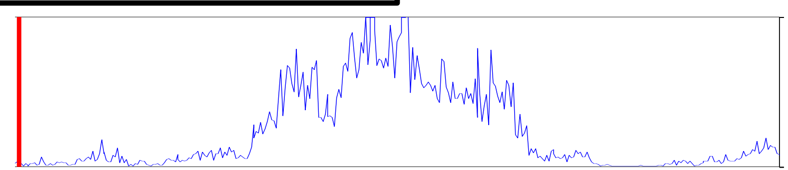
Cons: a static graph

New



$$\boxed{G_{X\to Y}(t) \quad = \quad D\left(\tilde{b}_t \| b_t\right)}$$

$G_{X\to Y}(t)$



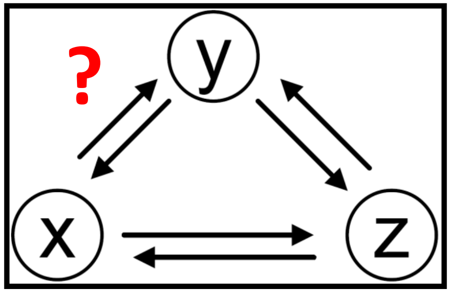
Pros: a dynamic graph

Cons: ∞-dimensional learning problem

$$\boxed{p_t = P_{Y_t|Y^{t-1},Z^{t-1}}}$$

$$\tilde{p}_t = P_{Y_t|Y^{t-1},Z^{t-1},X^{t-1}}$$

Old

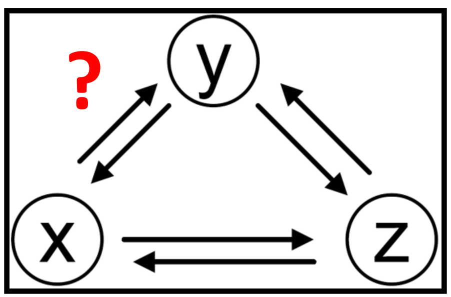


$$\boxed{\frac{1}{T} I(X \to Y \| Z) = \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[D(\tilde{p}_t \| p_t)]}$$

Pros: Finite-dimensional learning problem

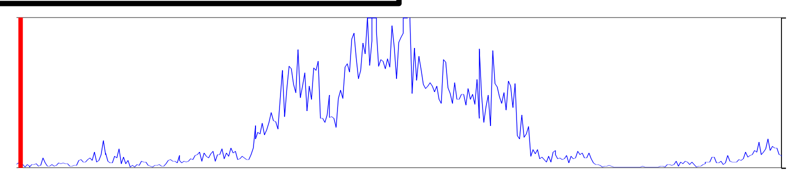
Cons: a static graph

New



$$\boxed{G_{X \to Y}(t) \quad = \quad D\left(\tilde{b}_t \| b_t\right)}$$

$G_{X \to Y}(t)$



Pros: a dynamic graph

Cons: ∞-dimensional learning problem

Prediction w Expert Advice

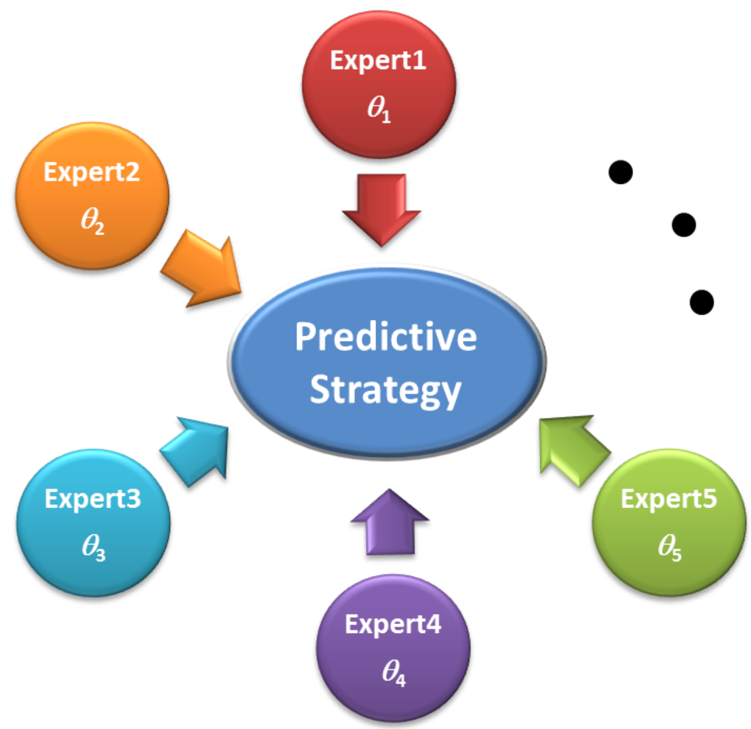
$$p_t = P_{Y_t|Y^{t-1},Z^{t-1}}$$

Experts

$$\{p_t^u \equiv P_{Y_t|Y^{t-1},Z^{t-1},\theta=u}\}_{u\in E}$$

For each round t = 1, 2,…, T

1. Expert specified by θ -> p_t^θ
2. Predictor combines advice p_t^θ along with $y_1, \ldots, y_{t-1}$ to choose p_t
3. Environment reveals y_t
4. Predictor p incurs loss $l(p_t,y_t)$, Expert θ incurs loss $l(p_t^\theta,y_t)$

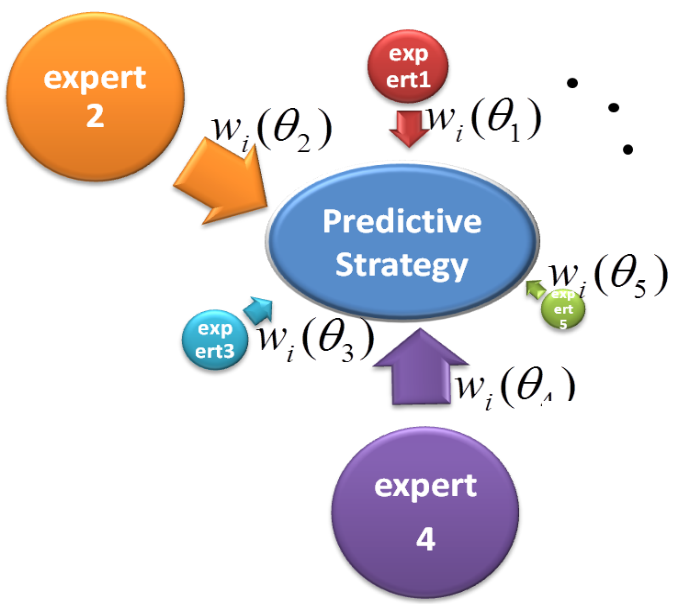


Good Sequential Predictors

- Regret $$R_n(p) = \sup_{y^n} \left(\underbrace{\sum_{i=1}^{n} l(y_i, p_i)}_{\text{my predictor}} - \underbrace{\min_{\theta} \sum_{k=1}^{n} l(y_k, p_k^{\theta})}_{\text{best predictor by nature in hindsight}} \right),$$

- p is minimax if $\frac{1}{n} R_n(p) \to \min_q \frac{1}{n} R_n(q)$

- "Good" predictor: if 2 and 4 recently did well, their advice is emphasized

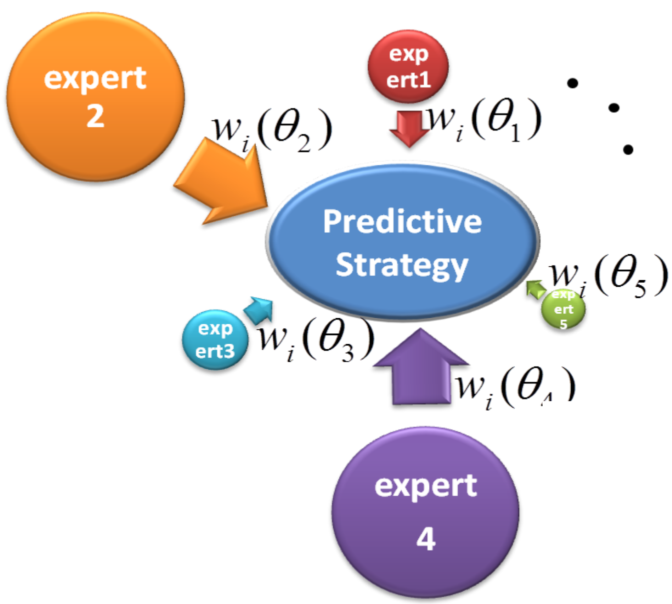


Good Sequential Predictors

- Regret $$R_n(p) = \sup_{y^n} \Bigg(\underbrace{\sum_{i=1}^{n} l(y_i, p_i)}_{\text{my predictor}} - \underbrace{\min_{\theta} \sum_{k=1}^{n} l(y_k, p_k^{\theta})}_{\text{best predictor by nature in hindsight}} \Bigg),$$

- p is minimax if $\frac{1}{n} R_n(p) \to \min_q \frac{1}{n} R_n(q)$

- "Good" predictor: if 2 and 4 recently did well, their advice is emphasized



$$\boxed{p_t(y) = \int_{u \in E} w_t(u) p_t^u(y) du}$$

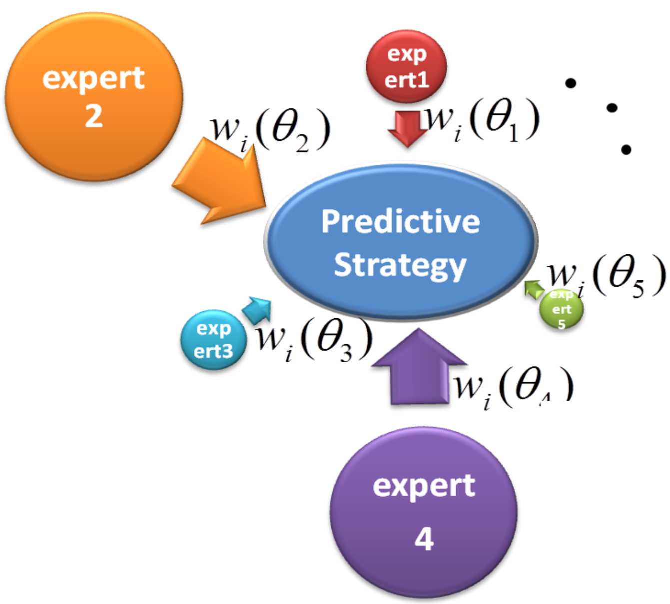
$$w_t(\theta) \triangleq f(\theta | y^{t-1}) = \frac{f(\theta) f(y^{t-1} | \theta)}{\beta_{y^{t-1}}}$$ $\leftarrow$ **hard**

Good Sequential Predictors

- Regret $$R_n(p) = \sup_{y^n}\Bigg(\underbrace{\sum_{i=1}^{n} l(y_i, p_i)}_{\text{my predictor}} - \underbrace{\min_{\theta} \sum_{k=1}^{n} l(y_k, p_k^{\theta})}_{\text{best predictor by nature in hindsight}} \Bigg),$$

- p is minimax if $\frac{1}{n} R_n(p) \to \min_{q} \frac{1}{n} R_n(q)$

- "Good" predictor: if 2 and 4 recently did well, their advice is emphasized



$$\boxed{p_t(y) = \int_{u \in E} w_t(u) p_t^u(y) du}$$

$$w_t(\theta) \triangleq f(\theta | y^{t-1}) = \frac{f(\theta) f(y^{t-1} | \theta)}{\beta_{y^{t-1}}} \longleftarrow \textbf{hard}$$

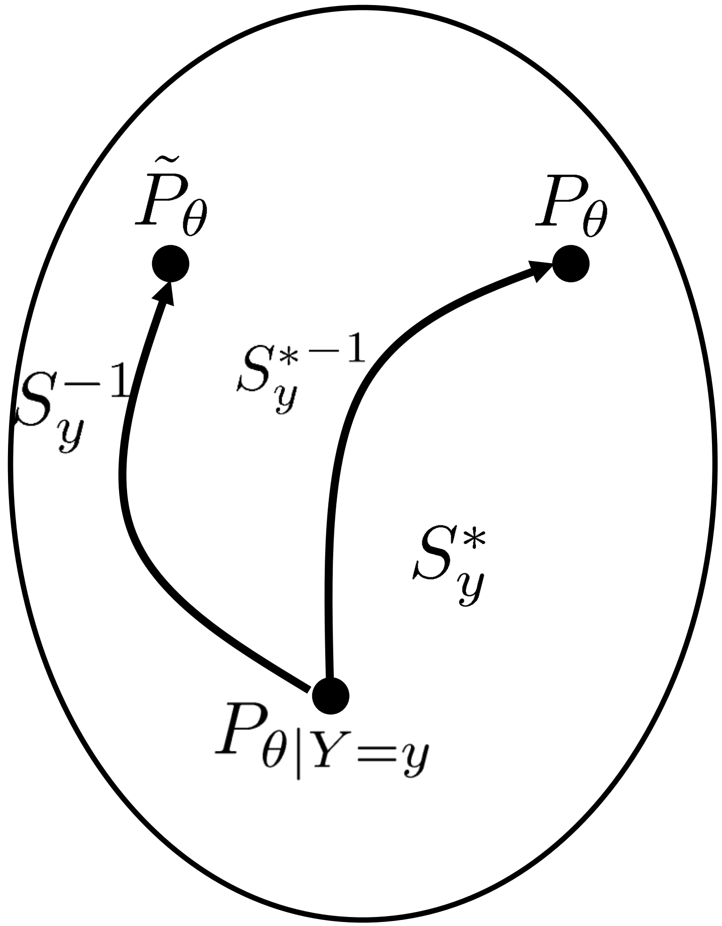
Attains minimax regret under Jeffrey's prior

$$f_\theta^*(u) = \frac{\sqrt{\det(J(u))}}{C}$$

Bayesian Inference with Optimal Maps

- A map S_y "pushes forward" prior to posterior

$$\text{if } U \sim P_\theta, \text{ then } Z = S_y(U) \sim P_{\theta|Y=y}$$



$$\min_{S_y} D(P_\theta \| \tilde{P}_\theta)$$

$$p(\theta) = q(S(\theta))|J_S(\theta)|$$

Finding the optimal map

$$\min_{S_y} D(P_\theta \| \tilde{P}_\theta)$$

$$\Updownarrow \quad \log \frac{f_X(x)}{\tilde{f}_X(x)} = \log \beta_y - T(S_y, x)$$

$$\boxed{\begin{aligned} (P2) \qquad S_y^* &= \arg\max_{S_y \in \mathcal{S}(\mathsf{X})} V_2(S_y), \\ V_2(S_y) &\triangleq \int_{x \in \mathsf{X}} f_X(x) T(S_y, x) dx. \end{aligned}}$$

Finding the optimal map

$$\min_{S_y} D(P_\theta \| \tilde{P}_\theta)$$

$$\Updownarrow \quad \log \frac{f_X(x)}{\tilde{f}_X(x)} = \log \beta_y - T(S_y, x)$$

$$\boxed{\begin{aligned} (P2) \qquad S_y^* &= \arg\max_{S_y \in \mathcal{S}(\mathsf{X})} V_2(S_y), \\ V_2(S_y) &\triangleq \int_{x \in \mathsf{X}} f_X(x) T(S_y, x) dx. \end{aligned}}$$

$$\begin{aligned} T(S, x) &\triangleq \log f_{Y|X}(y|S(x)) + \log f_X(S(x)) \\ &+ \log |\det(J_S(x))| - \log f_X(x) \end{aligned}$$

Finding the optimal map

$$\min_{S_y} D(P_\theta \| \tilde{P}_\theta)$$

$$\Updownarrow \quad \log \frac{f_X(x)}{\tilde{f}_X(x)} = \log \beta_y - T(S_y, x)$$

$$\boxed{\begin{aligned} (P2) \qquad S_y^* &= \arg\max_{S_y \in \mathcal{S}(\mathsf{X})} V_2(S_y), \\ V_2(S_y) &\triangleq \int_{x \in \mathsf{X}} f_X(x) T(S_y, x) dx. \end{aligned}}$$

Hard (non-convex)

$$\begin{aligned} T(S, x) &\triangleq \log f_{Y|X}(y|S(x)) + \log f_X(S(x)) \\ &+ \underline{\log |\det(J_S(x))|} - \log f_X(x) \end{aligned}$$

Theorem [Kim, Mesa, TPC '12]. If prior $p(\theta)$ and likelihood $p(y|\theta)$ are log-concave in θ, then calculating posterior $p(\theta|y)$ is a **convex optimization** problem

$$(P3) \qquad S_y^* = \underset{S_y \in \mathcal{S}(\mathsf{X}),\ J_{S_y}(x) \succ 0\ \forall x \in \mathsf{X}}{\arg\max} V_3(S_y),$$

$$V_3(S_y) \triangleq \int_{x \in \mathsf{X}} f_X(x) \tilde{T}(S_y, x) dx$$

$$\begin{aligned} \tilde{T}(S,x) &\triangleq \log f_{Y|X}(y|S(x)) + \log f_X(S(x)) \\ &+ \underline{\log \det (J_S(x))} - \log f_X(x) \end{aligned}$$

Theorem [Kim, Mesa, TPC '12]. If prior $p(\theta)$ and likelihood $p(y|\theta)$ are log-concave in θ, then calculating posterior $p(\theta|y)$ is a ∞-dimensional **convex optimization** problem

$$(P3) \quad S_y^* = \underset{S_y \in \mathcal{S}(\mathsf{X}),\ J_{S_y}(x) \succ 0\ \forall x \in \mathsf{X}}{\arg\max} V_3(S_y),$$

$$V_3(S_y) \triangleq \int_{x \in \mathsf{X}} f_X(x) \tilde{T}(S_y, x) dx$$

Easy (convex, but ∞-*dim*)

$$\tilde{T}(S,x) \triangleq \log f_{Y|X}(y|S(x)) + \log f_X(S(x)) + \log\det(J_S(x)) - \log f_X(x)$$

Finding the optimal map

$$\min_{S_y} D(P_\theta \| \tilde{P}_\theta)$$

$$\log \frac{f_X(x)}{\tilde{f}_X(x)} = \log \beta_y - T(S_y, x)$$

$$\boxed{\begin{aligned} (P4) \qquad F^* &= \underset{F \in \mathbb{R}^{d \times K} : F J_A(X_1) \succ 0, \ldots, F J_A(X_N) \succ 0}{\arg\max} V_4(F), \\ V_4(F) &\triangleq \frac{1}{N} \sum_{i=1}^{N} \tilde{T}(F, X_i) \qquad (25) \end{aligned}}$$

$$\begin{aligned} A(x) &= [\phi^{(1)}(x), \ldots, \phi^{(K)}(x)]^T, & , K \times 1 \\ S(x) &= F A(x), & d \times 1 \\ J_S(x) &= F J_A(x) & d \times d. \end{aligned}$$

Finding the optimal map

$$\min_{S_y} D(P_\theta \| \tilde{P}_\theta)$$

$$\Updownarrow \quad \log \frac{f_X(x)}{\tilde{f}_X(x)} = \log \beta_y - T(S_y, x)$$

$$\boxed{\begin{aligned} (P4) \qquad F^* &= \underset{F \in \mathbb{R}^{d\times K}: FJ_A(X_1)\succ 0, \dots, FJ_A(X_N)\succ 0}{\arg\max} V_4(F), \\ V_4(F) &\triangleq \frac{1}{N}\sum_{i=1}^{N} \tilde{T}(F, X_i) \qquad (25) \end{aligned}}$$

Easy (convex)

$$\begin{aligned} A(x) &= [\phi^{(1)}(x), \dots, \phi^{(K)}(x)]^T, \qquad , K \times 1 \\ S(x) &= FA(x), \qquad d \times 1 \\ J_S(x) &= FJ_A(x) \qquad d \times d. \end{aligned}$$

Polynomial Chaos expansion

Theorem [Kim, Mesa, TPC '12]. If prior $p(\theta)$ and likelihood $p(y|\theta)$ are log-concave in θ, then calculating posterior $p(\theta\,|\,y)$ is an easy (finite-dimensional) **convex optimization** problem

$$(P4) \qquad F^* = \underset{F\in\mathbb{R}^{d\times K}:FJ_A(X_1)\succ 0,\ldots,FJ_A(X_N)\succ 0}{\arg\max} V_4(F),$$

$$V_4(F) \triangleq \frac{1}{N}\sum_{i=1}^{N}\tilde{T}(F,X_i) \qquad (25)$$

Easy (convex)

$$A(x) = [\phi^{(1)}(x),\ldots,\phi^{(K)}(x)]^T, \qquad , K\times 1$$

$$S(x) = FA(x), \qquad d\times 1$$

$$J_S(x) = FJ_A(x) \qquad d\times d.$$

Polynomial Chaos expansion

Punchline

- KL divergence minimization:

$$\min_{S_y} D(P_\theta \| \tilde{P}_\theta) \leftrightarrow \max_{S_y} \int_\theta f_\Theta(\theta) T(S_y, x)$$

- Monotonicity of optimal solution to optimal transport problem

$$\min_{S_y} \int_\Theta \|\theta - S_y(\theta)\|^2 f(\theta) d\theta$$
$$s.t. \quad S_y \text{ pushes } f(\theta) \text{ to } f(\theta|y)$$

$$p(\theta) = q(S(\theta))|J_S(\theta)|$$

$$\Big\downarrow$$

$$p(\theta) = q(S(\theta))J_S(\theta)$$

- Wiener-Askey polynomial chaos expansion:

$$S_y(\theta) = \sum_{j \in \mathcal{J}} g_j \phi_j(\theta), \qquad \int_\Theta \phi_i(\theta)\phi_j(\theta) f(\theta) d\theta = C 1_{\{i \neq j\}}$$

Interesting Connection to Big Data

$$(P4) \quad F^* = \underset{F\in\mathbb{R}^{d\times K}: FJ_A(X_1)\succ 0,\ldots,FJ_A(X_N)\succ 0}{\arg\max} V_4(F),$$

$$V_4(F) \triangleq \frac{1}{N}\sum_{i=1}^{N}\tilde{T}(F, X_i) \tag{25}$$

$$\tilde{T}(S,x) \triangleq \log f_{Y|X}(y|S(x)) + \log f_X(S(x)) + \log\det(J_S(x)) - \log f_X(x)$$

$$\hat{J} \triangleq \underset{J\succ 0}{\arg\min}\langle\hat{\Sigma}, J\rangle - \log\det J + \gamma\|J\|_{1,\text{off}}$$

Highly scalable algorithms for large-scale covariance estimation

Discussion

- Theorem applicable to any Bayesian inference problem where log-concavity assumptions hold
  - all exponential families
  - GLMs (*used extensively in neuroscience*): **Jeffrey's prior is log-concave**
- Polynomial chaos expansion wrt prior enabls closed form computation of posterior moments

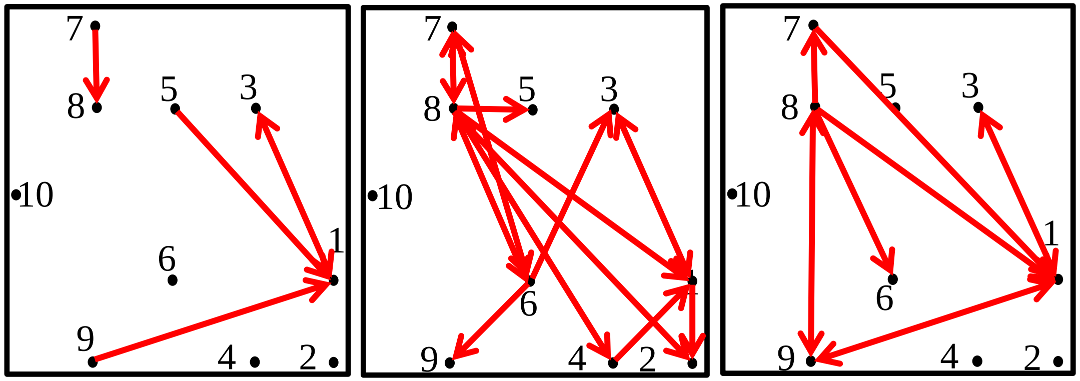
$$\int_{\Theta} \phi_i(\theta)\phi_j(\theta)f(\theta)d\theta = C1_{\{i\neq j\}}$$

- Comparison to Markov chain Monte carlo
  - Only sampling points from the prior!

Back to Neuroscience

Time-Varying Causality

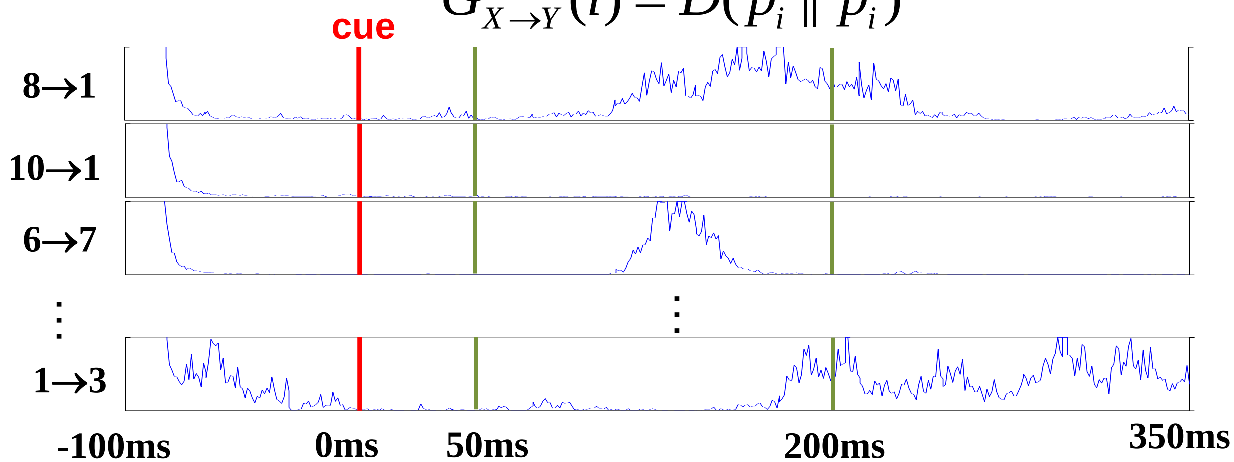
- Time-invariant case



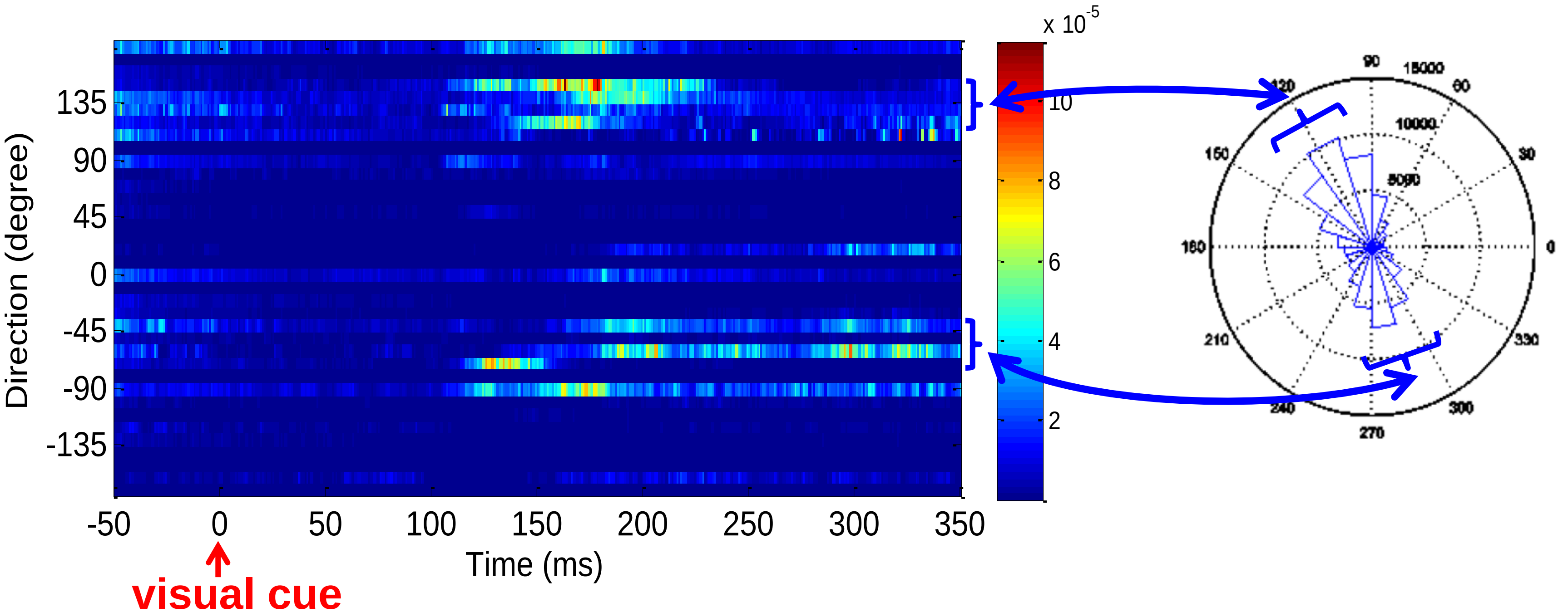
Pre: [-100 to 50ms] **Cue: [50 to 200ms]** **Aft: [200 to 350ms]**

- Time-varying case

$$G_{X \to Y}(i) = D(\tilde{p}_i \parallel p_i)$$



Direction of Time-Varying Causality



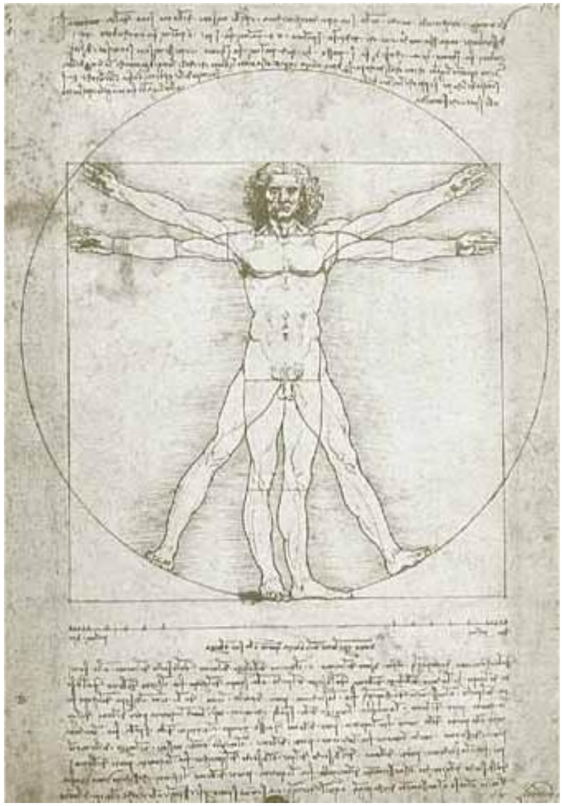
[Direction of time-varying causal interactions] [direction of β wave propagation]

Neural Interaction Laboratory

Sanggyun Kim

Diego Mesa

Thank You!



References

- K. Takahashi, S. Kim, TPC, and N. Hatsopoulos, "Large-scale spike sequencing associated with wave propagation in motor cortex ", in preparation.
- C. Quinn, N. Kiyavash, and T. P. Coleman, “Efficient Algorithms for Directed Tree Approximations to Minimal Generative Model Graphs ", to appear, *IEEE Trans Sig Proc*, May 2012.
- Y. Xu, S. Kim, S. Salapaka, C. Beck, and T. P. Coleman, "Clustering Large Networks of Parametric Dynamic Generative Models", submitted to CDC, Mar 2012.
- C. Quinn, N. Kiyavash, and T. P. Coleman, "A Minimal Approach to Causal Inference on Topologies with Bounded In-degree“, *IEEE Conference on Decision and Control* , Dec 2011.
- S. Kim, K. Takahashi, and N. Hatsopoulos, T. P. Coleman, "Information Transfer Between Neurons in the Motor Cortex Triggered by Visual Cues", *IEEE Engineering in Medicine and Biology Society Annual Conference*, Sep 2011.
- C. Quinn, T. P. Coleman, N. Kiyavash, and N. G. Hatsopoulos, "Estimating the directed information to infer causal relationships in ensemble neural spike train recordings", *Journal of Computational Neuroscience*, January 2011