An Information Geometric Approach for Feature Selection

Lizhong Zheng

EECS, MIT

Huawei, January, 2017

in collaboration with Shao-Lun Huang, Anuran Makur, Greg Wornell

Who's NOT Doing BigData?

Who's NOT Doing BigData?

What Do We Hope from a Theory?

- More General: from CS/MRI to An Overall User Profile

What Do We Hope from a Theory?

- More General: from CS/MRI to An Overall User Profile
    - Many different types of data.

What Do We Hope from a Theory?

- More General: from CS/MRI to An Overall User Profile
  - Many different types of data.
  - Different time scales and qualities.

What Do We Hope from a Theory?

- More General: from CS/MRI to An Overall User Profile
    - Many different types of data.
    - Different time scales and qualities.
    - Not clear what we are looking for.

What Do We Hope from a Theory?

- More General: from CS/MRI to An Overall User Profile
  - Many different types of data.
  - Different time scales and qualities.
  - Not clear what we are looking for.
- More Flexible: General Purpose Processing and Information Market

What Do We Hope from a Theory?

- More General: from CS/MRI to An Overall User Profile
  - Many different types of data.
  - Different time scales and qualities.
  - Not clear what we are looking for.

- More Flexible: General Purpose Processing and Information Market
  - Labels, Experts, and Fake News

What Do We Hope from a Theory?

- More General: from CS/MRI to An Overall User Profile
  - Many different types of data.
  - Different time scales and qualities.
  - Not clear what we are looking for.
- More Flexible: General Purpose Processing and Information Market
  - Labels, Experts, and Fake News
  - Sensitive Information

What Do We Hope from a Theory?

- More General: from CS/MRI to An Overall User Profile
  - Many different types of data.
  - Different time scales and qualities.
  - Not clear what we are looking for.
- More Flexible: General Purpose Processing and Information Market
  - Labels, Experts, and Fake News
  - Sensitive Information
- More Guarantees:

What Do We Hope from a Theory?

- More General: from CS/MRI to An Overall User Profile
  - Many different types of data.
  - Different time scales and qualities.
  - Not clear what we are looking for.
- More Flexible: General Purpose Processing and Information Market
  - Labels, Experts, and Fake News
  - Sensitive Information
- More Guarantees:
  - How Good Is Your Data?

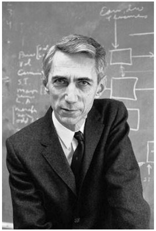
- More General: from CS/MRI to An Overall User Profile
  - Many different types of data.
  - Different time scales and qualities.
  - Not clear what we are looking for.
- More Flexible: General Purpose Processing and Information Market
  - Labels, Experts, and Fake News
  - Sensitive Information
- More Guarantees:
  - How Good Is Your Data?
  - Does It Solve My Problem?

What Do We Hope from a Theory?

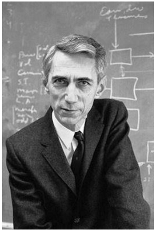
- More General: from CS/MRI to An Overall User Profile
  - Many different types of data.
  - Different time scales and qualities.
  - Not clear what we are looking for.
- More Flexible: General Purpose Processing and Information Market
  - Labels, Experts, and Fake News
  - Sensitive Information
- More Guarantees:
  - How Good Is Your Data?
  - Does It Solve My Problem?
  - Can Others Solve It Better?

Information Theory is the Right Tool

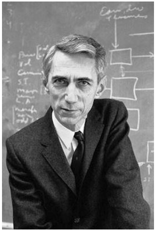
- It is ALWAYS the right tool.



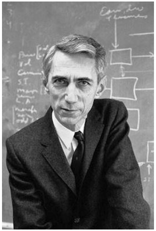
- It is ALWAYS the right tool.
- What is IT? Shannon:
    - When you make an observation, how much information you get?
    - How many bits?
    - Transform, Transmit, and Store information as bits.

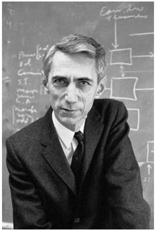


- It is ALWAYS the right tool.
- What is IT? Shannon:
  - When you make an observation, how much information you get?
  - How many bits?
  - Transform, Transmit, and Store information as bits.
- Bottomline: the more surprised you are, the more information you get.

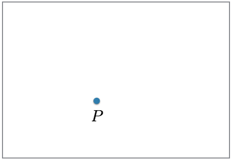


- It is ALWAYS the right tool.
- What is IT? Shannon:
  - When you make an observation, how much information you get?
  - How many bits?
  - Transform, Transmit, and Store information as bits.
- Bottomline: the more surprised you are, the more information you get.
- K-L divergence $D(P||Q)$, a distance measure between distributions P and Q

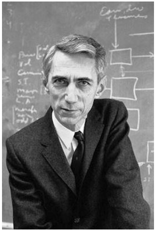


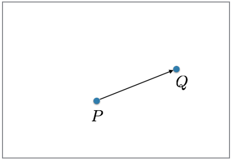


- It is ALWAYS the right tool.
- What is IT? Shannon:
  - When you make an observation, how much information you get?
  - How many bits?
  - Transform, Transmit, and Store information as bits.
- Bottomline: the more surprised you are, the more information you get.
- K-L divergence $D(P||Q)$, a distance measure between distributions P and Q

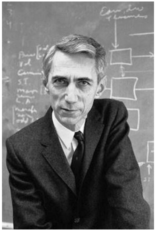


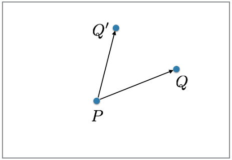
- It is ALWAYS the right tool.
- What is IT? Shannon:
  - When you make an observation, how much information you get?
  - How many bits?
  - Transform, Transmit, and Store information as bits.
- Bottomline: the more surprised you are, the more information you get.
- K-L divergence $D(P||Q)$, a distance measure between distributions P and Q





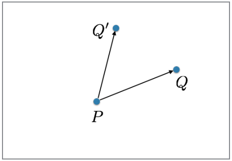
- It is ALWAYS the right tool.
- What is IT? Shannon:
  - When you make an observation, how much information you get?
  - How many bits?
  - Transform, Transmit, and Store information as bits.
- Bottomline: the more surprised you are, the more information you get.
- K-L divergence $D(P||Q)$, a distance measure between distributions P and Q





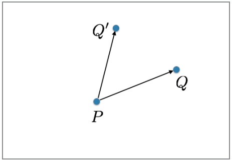
Did You See What Was Missing?

- put all information together
- never lose a bit
- only worry about the volume



Did You See What Was Missing?

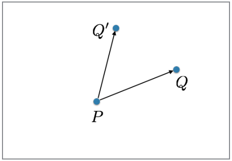
- Shannon didn't have Big Data
  - put all information together
  - never lose a bit
  - only worry about the volume



- What we need:
  - The data is TOO big, we can only take a part

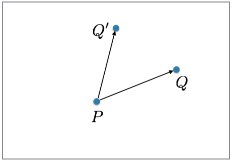
Did You See What Was Missing?

- Shannon didn't have Big Data
  - put all information together
  - never lose a bit
  - only worry about the volume



- What we need:
  - The data is TOO big, we can only take a part
  - Still have the issue of how to store and transmit, but we already know that.

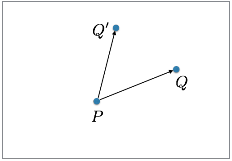
Did You See What Was Missing?



- Shannon didn't have Big Data
  - put all information together
  - never lose a bit
  - only worry about the volume

- What we need:
  - The data is TOO big, we can only take a part
  - Still have the issue of how to store and transmit, but we already know that.
  - The new question: Which part?

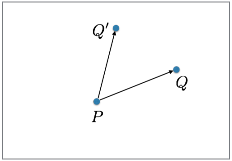
- Shannon didn't have Big Data
  - put all information together
  - never lose a bit
  - only worry about the volume



- What we need:
  - The data is TOO big, we can only take a part
  - Still have the issue of how to store and transmit, but we already know that.
  - The new question: Which part?
  - Need a new measure of relevance

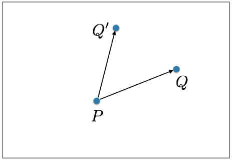
Did You See What Was Missing?

- Shannon didn't have Big Data
  - put all information together
  - never lose a bit
  - only worry about the volume



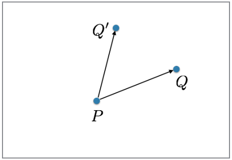
- What we need:
  - The data is TOO big, we can only take a part
  - Still have the issue of how to store and transmit, but we already know that.
  - The new question: Which part?
  - Need a new measure of relevance
  - Back to the picture: relevance and direction

- The geometry of probability distributions is complex.



- The geometry of probability distributions is complex.
- Focus on a small neighborhood around P_0

$$P(y) = P_0(y)(1 + \epsilon \cdot L_P(y)), \quad y \in \mathcal{Y}$$
$$Q(y) = P_0(y)(1 + \epsilon \cdot L_Q(y)), \quad y \in \mathcal{Y}$$



- The geometry of probability distributions is complex.
- Focus on a small neighborhood around P_0

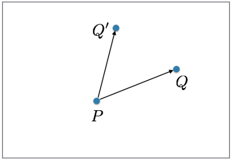
$$P(y) = P_0(y)(1 + \epsilon \cdot L_P(y)), \quad y \in \mathcal{Y}$$
$$Q(y) = P_0(y)(1 + \epsilon \cdot L_Q(y)), \quad y \in \mathcal{Y}$$

- Information Vector: 3 equivalent ways to write it.
  - Difference between two distributions $Q(y) - P_0(y)$
  - Log Likelihood functions

$$L_Q(y) = \log Q(y)/P_0(y)$$

  - Euclidean Vector form $\underline{\phi}$ with

$$\phi(y) = \frac{1}{\sqrt{P_0(y)}}(Q(y) - P_0(y))$$



Information Vector

- 3-way equivalence: distribution - feature function - information vector

$$P \leftrightarrow \underline{\phi} \qquad Q \leftrightarrow \underline{\nu}$$

Information Vector

- 3-way equivalence: distribution - feature function - information vector

$$P \leftrightarrow \underline{\phi} \qquad Q \leftrightarrow \underline{\nu}$$

- Shannon's notion of information volume

$$D(P||Q) = \|\underline{\phi} - \underline{\nu}\|^2 + o(\epsilon^2)$$

Information Vector

- 3-way equivalence: distribution - feature function - information vector

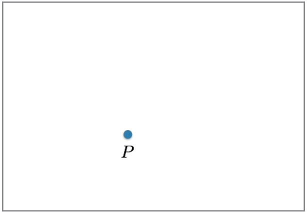
$$P \leftrightarrow \underline{\phi} \qquad Q \leftrightarrow \underline{\nu}$$

- Shannon's notion of information volume

$$D(P||Q) = ||\underline{\phi} - \underline{\nu}||^2 + o(\epsilon^2)$$

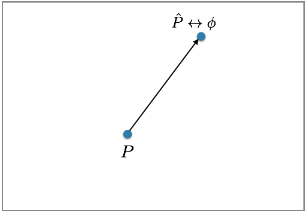
- Much more importantly, now information vector has directions

What is the Direction of Information Vectors?



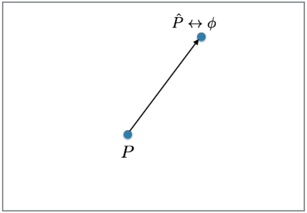
- Suppose we have some data Y with a long term average distribution P

What is the Direction of Information Vectors?



- Suppose we have some data Y with a long term average distribution P
- Now we observe a string of symbols $y_1, \ldots, y_n$, with empirical distribution $\hat{P}$

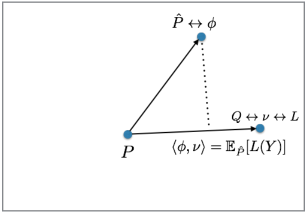
What is the Direction of Information Vectors?



- Suppose we have some data Y with a long term average distribution P
- Now we observe a string of symbols $y_1, \ldots, y_n$, with empirical distribution $\hat{P}$
- Now we are surprised, total information we get is

$$D(\hat{P}||P) \leftrightarrow \|\phi\|^2$$

What is the Direction of Information Vectors?



- Suppose we have some data Y with a long term average distribution P
- Now we observe a string of symbols $y_1, \dots, y_n$, with empirical distribution $\hat{P}$
- Now we are surprised, total information we get is

$$D(\hat{P}||P) \leftrightarrow \|\phi\|^2$$

- Suppose we evaluate a specific function with the data

$$\frac{1}{n} \sum_{i=1}^{n} L(y_i) \leftrightarrow \langle \underline{\phi}, \underline{\nu} \rangle \qquad \text{for some } Q \leftrightarrow L \leftarrow \underline{\nu}$$

- Correspondence : Distribution $\leftrightarrow$ feature function $\leftrightarrow$ information vector.

- Correspondence : Distribution $\leftrightarrow$ feature function $\leftrightarrow$ information vector.

- Evaluating a feature function = parameter estimation for a particular exponential family, hypothesis testing between two given distributions = Taking a specific component of the information vector.

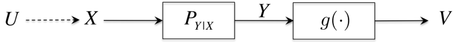
Inner Product for the Space of Distributions

- Correspondence : Distribution $\leftrightarrow$ feature function $\leftrightarrow$ information vector.
- Evaluating a feature function = parameter estimation for a particular exponential family, hypothesis testing between two given distributions = Taking a specific component of the information vector.
- If we knew the model, we of course pick the “right" feature function, i.e. pick the part of the information relevant to the problem;

Inner Product for the Space of Distributions

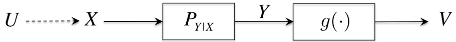
- Correspondence : Distribution $\leftrightarrow$ feature function $\leftrightarrow$ information vector.
- Evaluating a feature function = parameter estimation for a particular exponential family, hypothesis testing between two given distributions = Taking a specific component of the information vector.
- If we knew the model, we of course pick the “right" feature function, i.e. pick the part of the information relevant to the problem;
- Even if we don’t have the model, the picture still holds.

Let's Solve One Problem

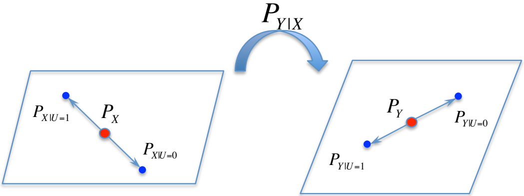


- Taking a feature of Y, $g(Y)$, so it carries information about X, or target U encoded in X.

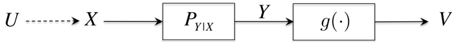
Let's Solve One Problem



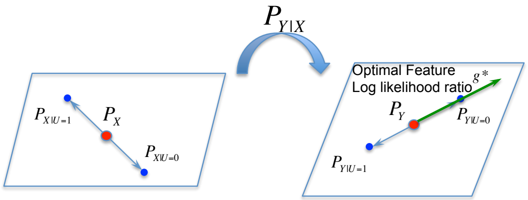
- Taking a feature of Y, $g(Y)$, so it carries information about X, or target U encoded in X.
- Encoding of the target: $P_{X|U}$ different from P_X, $\log \frac{P_{X|U}(\cdot|u)}{P_X(\cdot)} \leftrightarrow \underline{\phi}_X$,



Let's Solve One Problem



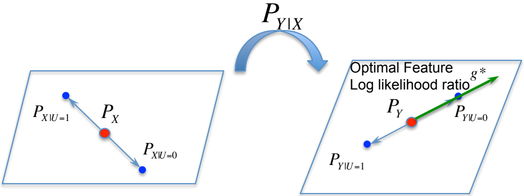
- Taking a feature of Y, $g(Y)$, so it carries information about X, or target U encoded in X.
- Encoding of the target: $P_{X|U}$ different from P_X, $\log \frac{P_{X|U}(\cdot|u)}{P_X(\cdot)} \leftrightarrow \underline{\phi}_X$,
- This gets mapped to the Y space, $\log \frac{P_{Y|U}(\cdot|u)}{P_Y(\cdot)} \leftrightarrow \underline{\phi}_Y$,



Let's Solve One Problem

$$U \dashrightarrow X \longrightarrow \boxed{P_{Y|X}} \xrightarrow{Y} \boxed{g(\cdot)} \longrightarrow V$$

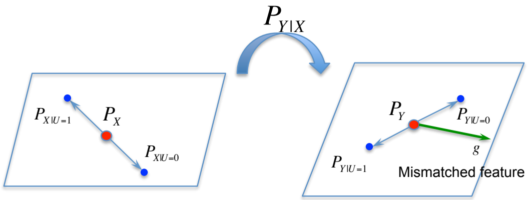
- Taking a feature of Y, $g(Y)$, so it carries information about X, or target U encoded in X.
- Encoding of the target: $P_{X|U}$ different from P_X, $\log \frac{P_{X|U}(\cdot|u)}{P_X(\cdot)} \leftrightarrow \underline{\phi}_X$,
- This gets mapped to the Y space, $\log \frac{P_{Y|U}(\cdot|u)}{P_Y(\cdot)} \leftrightarrow \underline{\phi}_Y$,
- If we know the model, always use $g^* \leftrightarrow \underline{\phi}_Y$ as the sufficient statistic,



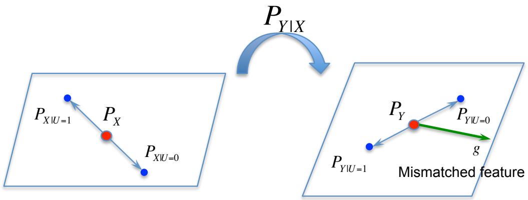
Let's Solve One Problem

$$U \dashrightarrow X \longrightarrow \boxed{P_{Y|X}} \xrightarrow{Y} \boxed{g(\cdot)} \longrightarrow V$$

- Taking a feature of Y, $g(Y)$, so it carries information about X, or target U encoded in X.
- Encoding of the target: $P_{X|U}$ different from P_X, $\log \frac{P_{X|U}(\cdot|u)}{P_X(\cdot)} \leftrightarrow \underline{\phi}_X$,
- This gets mapped to the Y space, $\log \frac{P_{Y|U}(\cdot|u)}{P_Y(\cdot)} \leftrightarrow \underline{\phi}_Y$,
- If we know the model, always use $g^* \leftrightarrow \underline{\phi}_Y$ as the sufficient statistic,
- If not, we get a mismatch,

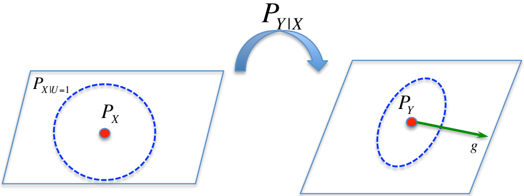


What If We Don't Know About the Target?



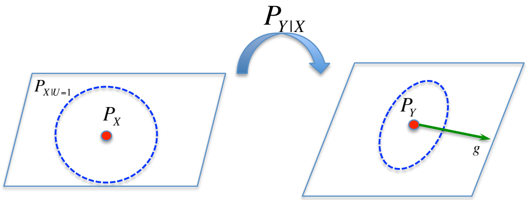
- When do we not know the model?
  - Data is too high-dimensional that we need to first reduce dimensionality before learning a model
  - We want to have a generic processing to serve multiple purposes
  - We don't have a good way to represent the data
- Linear map between $\underline{\phi}_X \mapsto \underline{\phi}_Y$;
- Norms don't matter

What If We Don't Know About the Target?



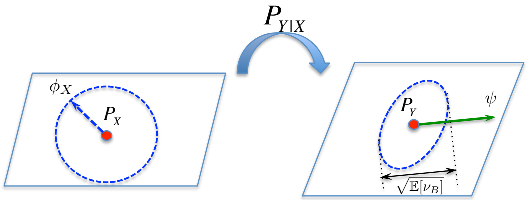
- When do we not know the model?
  - Data is too high-dimensional that we need to first reduce dimensionality before learning a model
  - We want to have a generic processing to serve multiple purposes
  - We don't have a good way to represent the data
- Linear map between $\underline{\phi}_X \mapsto \underline{\phi}_Y$;
- Norms don't matter
- Contraction divergence ball mapped into an ellipsoid;

What If We Don't Know About the Target?



- When do we not know the model?
  - Data is too high-dimensional that we need to first reduce dimensionality before learning a model
  - We want to have a generic processing to serve multiple purposes
  - We don't have a good way to represent the data
- Linear map between $\underline{\phi}_X \mapsto \underline{\phi}_Y$;
- Norms don't matter
- Contraction divergence ball mapped into an ellipsoid;

What If We Don't Know About the Target?



- When do we not know the model?
  - Data is too high-dimensional that we need to first reduce dimensionality before learning a model
  - We want to have a generic processing to serve multiple purposes
  - We don't have a good way to represent the data
- Linear map between $\underline{\phi}_X \mapsto \underline{\phi}_Y$;
- Norms don't matter
- Contraction divergence ball mapped into an ellipsoid;
- Average performance

$$\max_g \quad \mathbb{E}_{U-X}[D(P_{g(Y)|U=1}||P_{g(Y)|U=0})]$$

Equivalent Problems

Theorem

The following problems are equivalent (under local approximation)

- *Average inference performance over unknown models*

$$\max_{g} \quad \mathbb{E}_{U-X}[D(P_{g(Y)|U=1}||P_{g(Y)|U=0})]$$

- *Opportunistic formulation*

$$\max_{U-X}\max_{g} \quad \mathbb{E}_{U-X}[D(P_{g(Y)|U=1}||P_{g(Y)|U=0})]$$

- *PCA in the space of distributions*

$$\max_{\|\underline{\phi}_X\|^2=1} \quad \|\underline{\phi}_Y = B\underline{\phi}_X\|^2$$

- *Rényi maximal correlation (HGR)*

$$\rho = \max_{f,g:\mathbb{E}[f(X)]=\mathbb{E}[g(Y)]=1} \quad \mathbb{E}[f(X)\cdot g(Y)]$$

Many Good Things Happen at the Same Time

Many Good Things Happen at the Same Time

- There is an efficient algorithm: ACE algorithm (Brieman, Friedman' 85), power method for this SVD

Many Good Things Happen at the Same Time

- There is an efficient algorithm: ACE algorithm (Brieman, Friedman' 85), power method for this SVD
- Applies to any type of data, or combination of multi-modal data

Many Good Things Happen at the Same Time

- There is an efficient algorithm: ACE algorithm (Brieman, Friedman' 85), power method for this SVD
- Applies to any type of data, or combination of multi-modal data
- Choose the optimal non-linear features

Many Good Things Happen at the Same Time

- There is an efficient algorithm: ACE algorithm (Brieman, Friedman' 85), power method for this SVD
- Applies to any type of data, or combination of multi-modal data
- Choose the optimal non-linear features
- It comes with the guarantee of optimal performance (without knowledge of the model)

Many Good Things Happen at the Same Time

- There is an efficient algorithm: ACE algorithm (Brieman, Friedman' 85), power method for this SVD
- Applies to any type of data, or combination of multi-modal data
- Choose the optimal non-linear features
- It comes with the guarantee of optimal performance (without knowledge of the model)
- If we do know or partially know the model we can always incorporate that knowledge, optimally.

Many Good Things Happen at the Same Time

- There is an efficient algorithm: ACE algorithm (Brieman, Friedman' 85), power method for this SVD
- Applies to any type of data, or combination of multi-modal data
- Choose the optimal non-linear features
- It comes with the guarantee of optimal performance (without knowledge of the model)
- If we do know or partially know the model we can always incorporate that knowledge, optimally.
- It works well with data: provably lowest sample complexity: learn the most learnable things.

Many Good Things Happen at the Same Time

- There is an efficient algorithm: ACE algorithm (Brieman, Friedman' 85), power method for this SVD
- Applies to any type of data, or combination of multi-modal data
- Choose the optimal non-linear features
- It comes with the guarantee of optimal performance (without knowledge of the model)
- If we do know or partially know the model we can always incorporate that knowledge, optimally.
- It works well with data: provably lowest sample complexity: learn the most learnable things.
- One can always add operation constraints, such as linear feature functions (PCA), sparsity (CS), activation function (NN)?

Many Good Things Happen at the Same Time

- There is an efficient algorithm: ACE algorithm (Brieman, Friedman' 85), power method for this SVD
- Applies to any type of data, or combination of multi-modal data
- Choose the optimal non-linear features
- It comes with the guarantee of optimal performance (without knowledge of the model)
- If we do know or partially know the model we can always incorporate that knowledge, optimally.
- It works well with data: provably lowest sample complexity: learn the most learnable things.
- One can always add operation constraints, such as linear feature functions (PCA), sparsity (CS), activation function (NN)?
- Protection of sensitive information

The Problem I Played With On My Way

- MIT EECS has about 70-80 active upper level undergraduate courses

The Problem I Played With On My Way

- MIT EECS has about 70-80 active upper level undergraduate courses
- Each MEng student needs to pick a concentration of 4 courses.

The Problem I Played With On My Way

- MIT EECS has about 70-80 active upper level undergraduate courses
- Each MEng student needs to pick a concentration of 4 courses.
- But what and what make a concentration?

Applied Physics *AUS2:* 6.061, 6.602; *AAGS:* **6.621**, 6.630, **6.631**, **6.632**, **6.634**, 6.637, 6.638, **6.641**, 6.642, 6.644, 6.645, **6.685**, 6.690, 6.691, 6.695, 6.728, 6.730, 6.731, 6.732

Artificial Intelligence *AUS2:* 6.801, 6.802, 6.803, 6.804, 6.806, 6.813, 6.819, 6.905, IDS.012; *AAGS:* 6.345, **6.437**, **6.438**, **6.831**, **6.832**, **6.833**, 6.834, **6.860**, **6.861**, **6.862**, **6.863**, **6.864**, **6.866**, **6.867**, 6.868, 6.869, 6.872, **6.874**, 6.881, 6.882, 6.883, 6.884, 6.945, 6.946, **IDS.131**, MAS.S63

BioEECS *AUS2:* 6.022, 6.023, 6.025, 6.027, 6.047, 6.503, 6.580, 6.802; *AAGS:* **6.521**, **6.522**, 6.524, 6.525, 6.541, 6.542, 6.544, 6.545, **6.551**, 6.552, **6.555**, 6.556, 6.557, **6.561**, 6.580, 6.581, 6.582, 6.589, 6.872, 6.874, 6.877, 6.878

Circuits *AUS2:* 6.301, 6.302; *AAGS:* 6.331, 6.332, 6.333, **6.334**, **6.374**, **6.375**, **6.376**, **6.775**, 6.776

Communications *AAGS:* 6.231, **6.255**, 6.260, 6.261, **6.262**, 6.263, 6.264, 6.265, 6.266, **6.267**, 6.268, 6.434, 6.435, **6.436**, **6.437**, **6.438**, 6.440, 6.441, 6.442, 6.443, 6.450, 6.452, 6.453

Computer Systems *AUS2:* 6.035, 6.172, 6.175, 6.814, 6.816, 6.S062; *AAGS:* **6.820**, 6.821, **6.823**, **6.824**, **6.828**, **6.829**, **6.830**, 6.836, 6.846, **6.857**, **6.858**, 6.885, 6.886, 6.887, 6.888

Control *AUS2:* 6.302; *AAGS:* 6.231, **6.241**, 6.242, 6.243, 6.245, 6.246, 6.247

Graphics and Human-Computer Interfaces *AUS2:* 6.801, 6.807, 6.813, 6.815, 6.819, 6.837; *AAGS:* **6.345**, **6.831**, 6.835, 6.838, **6.839**, 6.865, 6.869, 6.870, 6.894, 6.895, 6.896

Materials, Devices and Nanotechnology *AUS2:* 6.701, 6.717, *AAGS:* **6.719**, **6.720**, **6.728**, **6.730**, 6.731, 6.732, 6.735, 6.736, 6.763, 6.772, **6.774**, **6.777**, 6.780J, 6.781, 6.789

The Problem I Played With On My Way

- MIT EECS has about 70-80 active upper level undergraduate courses
- Each MEng student needs to pick a concentration of 4 courses.
- But what and what make a concentration?
- How do we quantify the similarity of courses?

Applied Physics *AUS2:* 6.061, 6.602; *AAGS:* **6.621**, 6.630, **6.631**, **6.632**, **6.634**, 6.637, 6.638, **6.641**, 6.642, 6.644, 6.645, **6.685**, 6.690, 6.691, 6.695, 6.728, 6.730, 6.731, 6.732

Artificial Intelligence *AUS2:* 6.801, 6.802, 6.803, 6.804, 6.806, 6.813, 6.819, 6.905, IDS.012; *AAGS:* 6.345, **6.437**, **6.438**, **6.831**, **6.832**, **6.833**, 6.834, **6.860**, **6.861**, **6.862**, **6.863**, **6.864**, **6.866**, **6.867**, 6.868, 6.869, 6.872, **6.874**, 6.881, 6.882, 6.883, 6.884, 6.945, 6.946, **IDS.131**, MAS.S63

BioEECS *AUS2:* 6.022, 6.023, 6.025, 6.027, 6.047, 6.503, 6.580, 6.802; *AAGS:* **6.521**, **6.522**, 6.524, 6.525, 6.541, 6.542, 6.544, 6.545, **6.551**, 6.552, **6.555**, 6.556, 6.557, **6.561**, 6.580, 6.581, 6.582, 6.589, 6.872, 6.874, 6.877, 6.878

Circuits *AUS2:* 6.301, 6.302; *AAGS:* 6.331, 6.332, 6.333, **6.334**, **6.374**, **6.375**, **6.376**, **6.775**, 6.776

Communications *AAGS:* 6.231, **6.255**, 6.260, 6.261, **6.262**, 6.263, 6.264, 6.265, 6.266, **6.267**, 6.268, 6.434, 6.435, **6.436**, **6.437**, **6.438**, 6.440, 6.441, 6.442, 6.443, 6.450, 6.452, 6.453

Computer Systems *AUS2:* 6.035, 6.172, 6.175, 6.814, 6.816, 6.S062; *AAGS:* **6.820**, 6.821, **6.823**, **6.824**, **6.828**, **6.829**, **6.830**, 6.836, 6.846, **6.857**, **6.858**, 6.885, 6.886, 6.887, 6.888

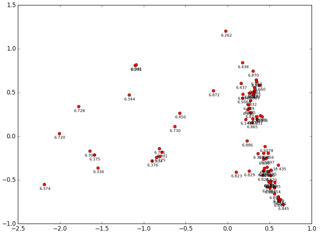
Control *AUS2:* 6.302; *AAGS:* 6.231, **6.241**, 6.242, 6.243, 6.245, 6.246, 6.247

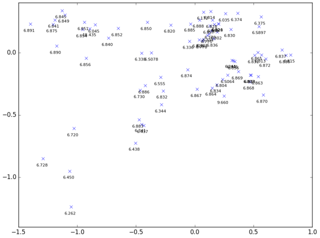
Graphics and Human-Computer Interfaces *AUS2:* 6.801, 6.807, 6.813, 6.815, 6.819, 6.837; *AAGS:* **6.345**, **6.831**, 6.835, 6.838, **6.839**, 6.865, 6.869, 6.870, 6.894, 6.895, 6.896

Materials, Devices and Nanotechnology *AUS2:* 6.701, 6.717, *AAGS:* **6.719**, **6.720**, **6.728**, **6.730**, 6.731, 6.732, 6.735, 6.736, 6.763, 6.772, **6.774**, **6.777**, 6.780J, 6.781, 6.789

The Problem I Played With On My Way

- MIT EECS has about 70-80 active upper level undergraduate courses
- Each MEng student needs to pick a concentration of 4 courses.
- But what and what make a concentration?
- How do we quantify the similarity of courses?





- Measurement of events occurrence at different places and different time

A More Serious Problem: Cyber-Security/Network Management

- Measurement of events occurrence at different places and different time
- Some particular patterns of sequences of events leads attacks/crash

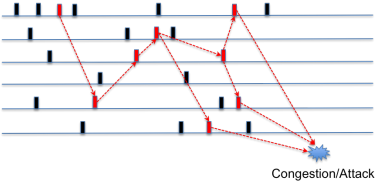
A More Serious Problem: Cyber-Security/Network Management

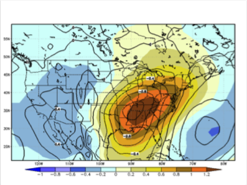
- Measurement of events occurrence at different places and different time
- Some particular patterns of sequences of events leads attacks/crash
- Typically many thousands of different types of events, over long time, and large areas

A More Serious Problem: Cyber-Security/Network Management

- Measurement of events occurrence at different places and different time
- Some particular patterns of sequences of events leads attacks/crash
- Typically many thousands of different types of events, over long time, and large areas

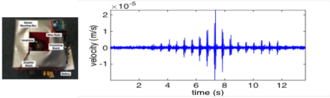
1438922809100	EVENT_LTE_RRC_PAGING_DRX_CYCLE	1614
1438922809100	RRC TIMER - DEADLOCK STOP	1605
1438922809100	EVENT_LTE_BSR_SR_REQUEST	1719
1438922809102	LTE_MAC_TIMER_ Start	1720
1438922809102	EVENT_LTE_TIMING_ADVANCE	1498
1438922809106	EVENT_LTE_BSR_SR_REQUEST	1719
1438922809110	EVENT_LTE_BSR_SR_REQUEST	1719
1438922809112	EVENT_LTE_REG_OUTGOING_MSG	1634
1438922809116	EVENT_LTE_BSR_SR_REQUEST	1719
1438922809126	EVENT_SD_EVENT_ACTION	621
1438922809129	EVENT_CM_SERVICE_CONFIRMED	558
1438922809135	EVENT_LTE_BSR_SR_REQUEST	1719
1438922809145	EVENT_LTE_BSR_SR_REQUEST	1719
1438922809153	LTE_MAC_TIMER_ Start	1720





Detection of Extreme Weather Pattern

The NetFlix Problem



User Recognition by Footsteps

MNIST Handwriting Recognition

Community Detection on Social Networks, Cyber-Security of Large Networks, Joint Video-Audio Recognition...

- The most important: pick a partial information

- The most important: pick a partial information
  - We know how to take a part of information: a function;

Summary: Concepts and Advantages

- The most important: pick a partial information
  - We know how to take a part of information: a function;
  - We know how good it is: inner product;

Summary: Concepts and Advantages

- The most important: pick a partial information
  - We know how to take a part of information: a function;
  - We know how good it is: inner product;
  - We know how to take it universally: detecting the divergence ball;

- The most important: pick a partial information
    - We know how to take a part of information: a function;
    - We know how good it is: inner product;
    - We know how to take it universally: detecting the divergence ball;
    - We know how to take it efficiently: the algorithm.

Summary: Concepts and Advantages

- The most important: pick a partial information
  - We know how to take a part of information: a function;
  - We know how good it is: inner product;
  - We know how to take it universally: detecting the divergence ball;
  - We know how to take it efficiently: the algorithm.

- Once we know what is a part of information:

Summary: Concepts and Advantages

- The most important: pick a partial information
  - We know how to take a part of information: a function;
  - We know how good it is: inner product;
  - We know how to take it universally: detecting the divergence ball;
  - We know how to take it efficiently: the algorithm.
- Once we know what is a part of information:
  - A Cross-Platform processing

Summary: Concepts and Advantages

- The most important: pick a partial information
  - We know how to take a part of information: a function;
  - We know how good it is: inner product;
  - We know how to take it universally: detecting the divergence ball;
  - We know how to take it efficiently: the algorithm.

- Once we know what is a part of information:
  - A Cross-Platform processing
  - A Universal interface between data and knowledge

- The most important: pick a partial information
  - We know how to take a part of information: a function;
  - We know how good it is: inner product;
  - We know how to take it universally: detecting the divergence ball;
  - We know how to take it efficiently: the algorithm.

- Once we know what is a part of information:
  - A Cross-Platform processing
  - A Universal interface between data and knowledge
  - A Secure Marketplace for data sharing