perception, action and the information knot that ties them

stefano soatto

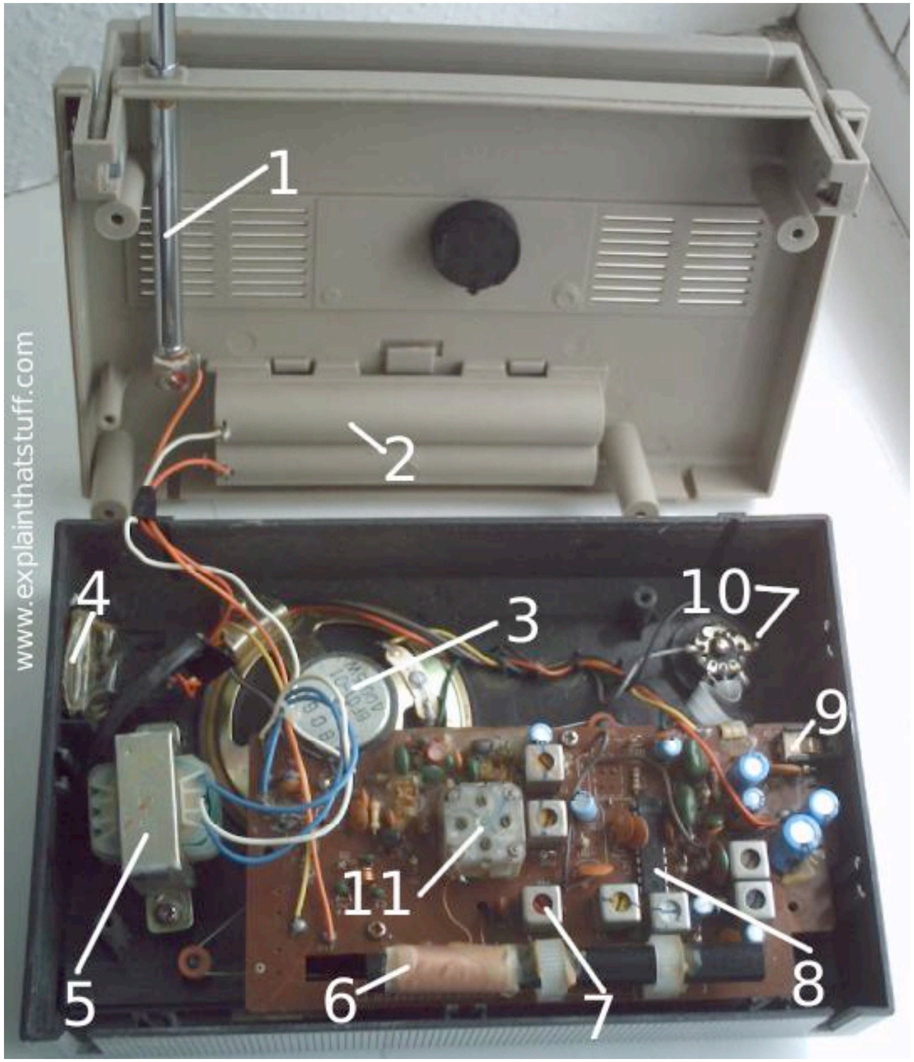
ucla

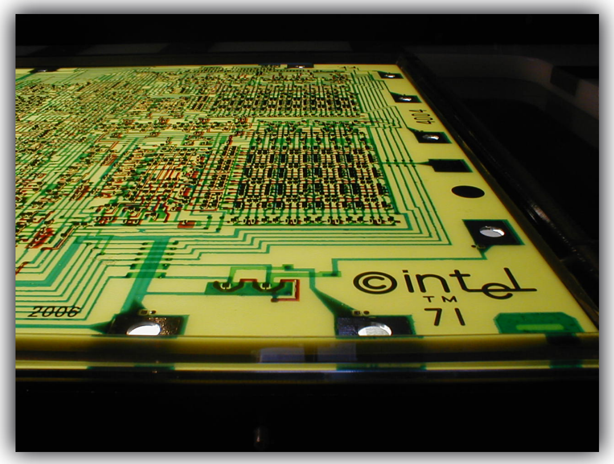
part I
why?

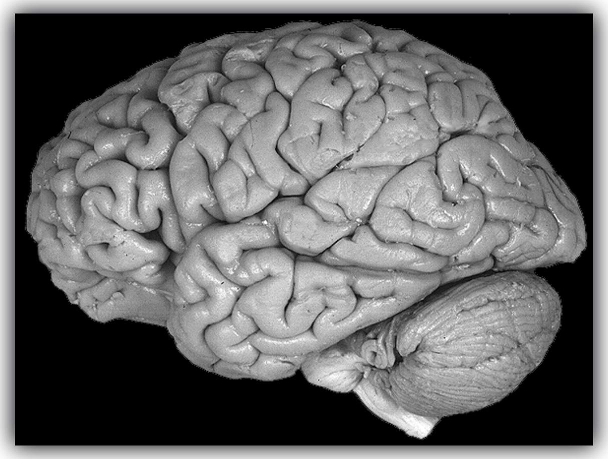
information? knowledge? intelligence?

data vs. symbols

how does a radio work?







 data vs. information

is data analysis necessary for intelligent behavior?

- rao & blackwell say **no**
- data compression vs. data analysis
- wiener & shannon: "semantic aspect of information is irrelevant" [to communications]

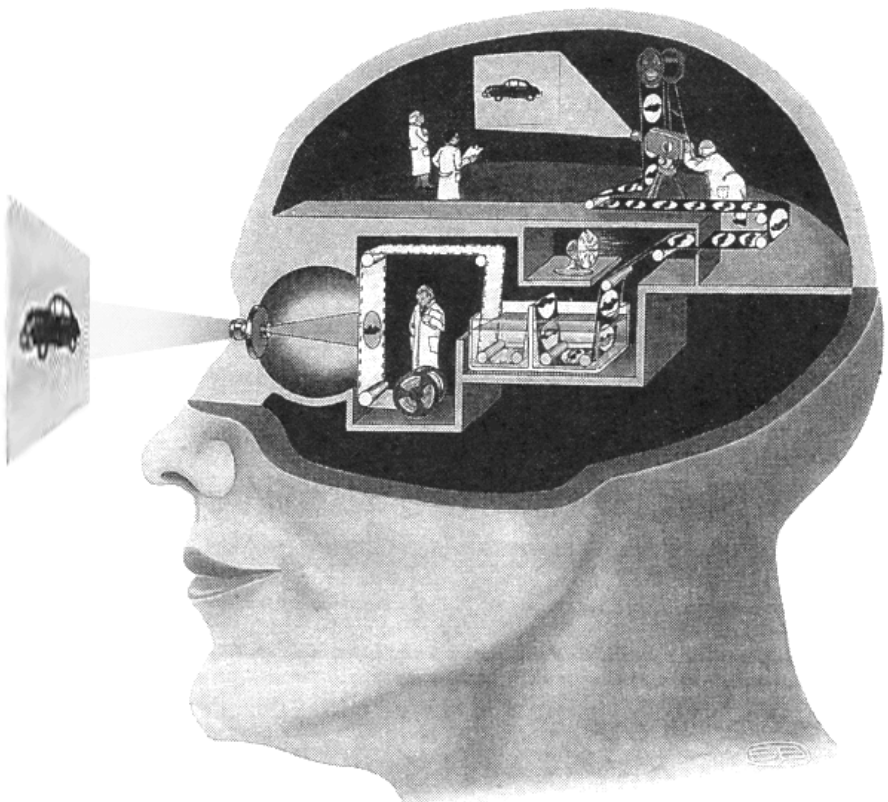
why ?

- why perform segmentation, edge detection, feature selection, clustering, “primal sketch” etc? what about falsifiability?
- why would the brain do so?
- is it better to just train an uber-classifier with the raw images?
- is “learning away” possible?

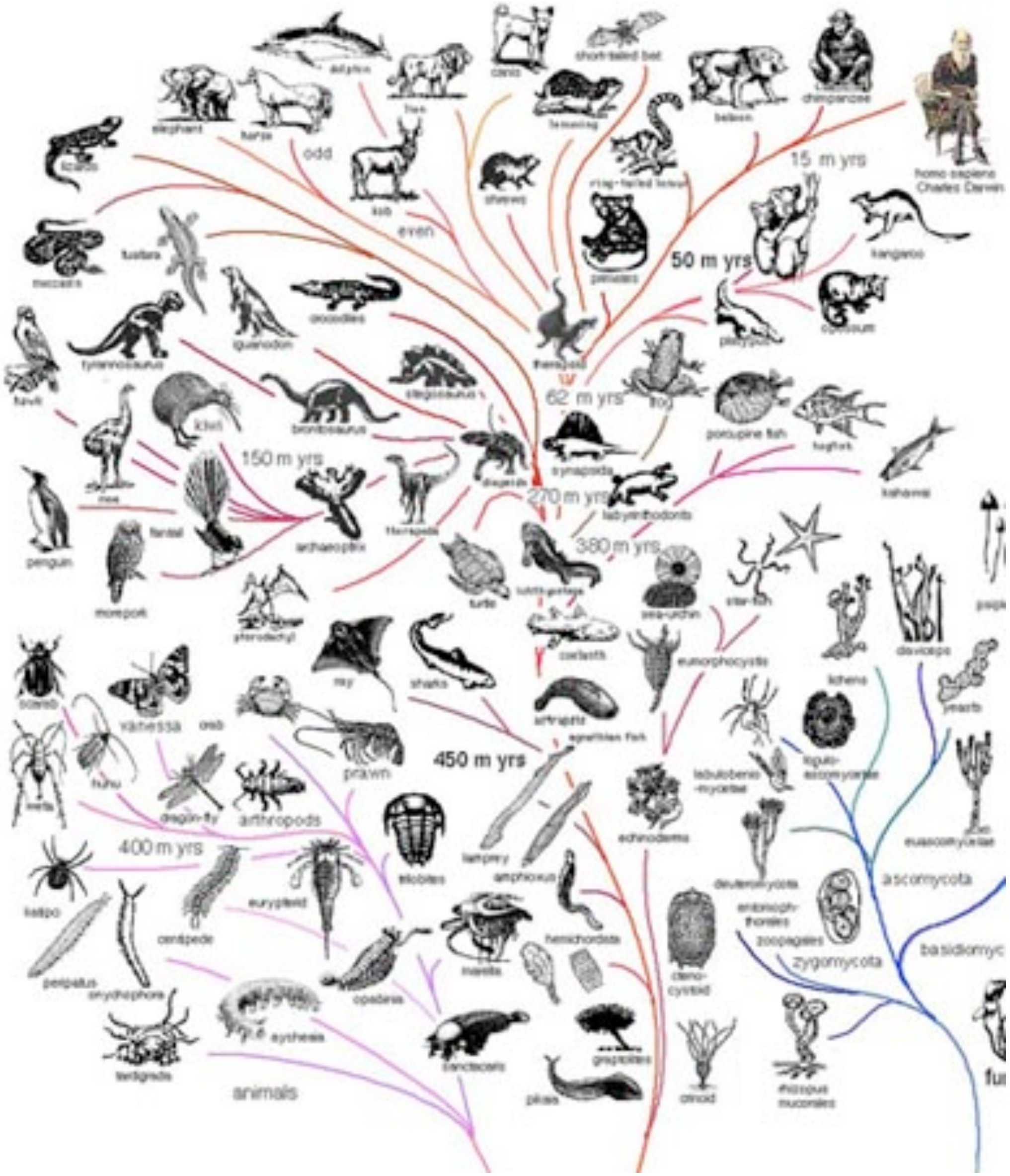
the epistemological gap

- primary epistemics/cognitive "science" starts from "discrete" tokens/atoms/symbols. how do we get there from data? and why?
- data-processing inequality: no advantage in breaking data into pieces (descartes)
- how do we reconcile?

signal-symbol barrier



no natural discretization

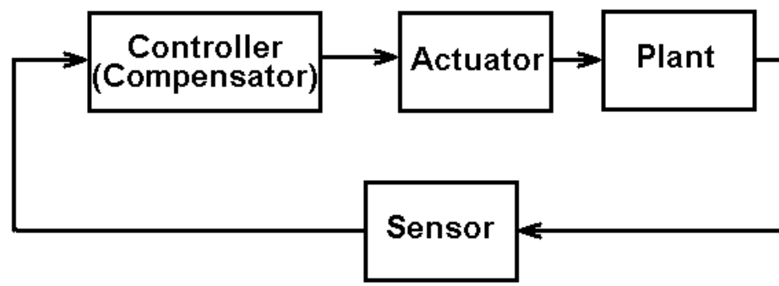


intelligence

many tasks

- measurable action performed by an agent (human or machine)
- most general: survival
- simplest: a **binary decision**

plants?



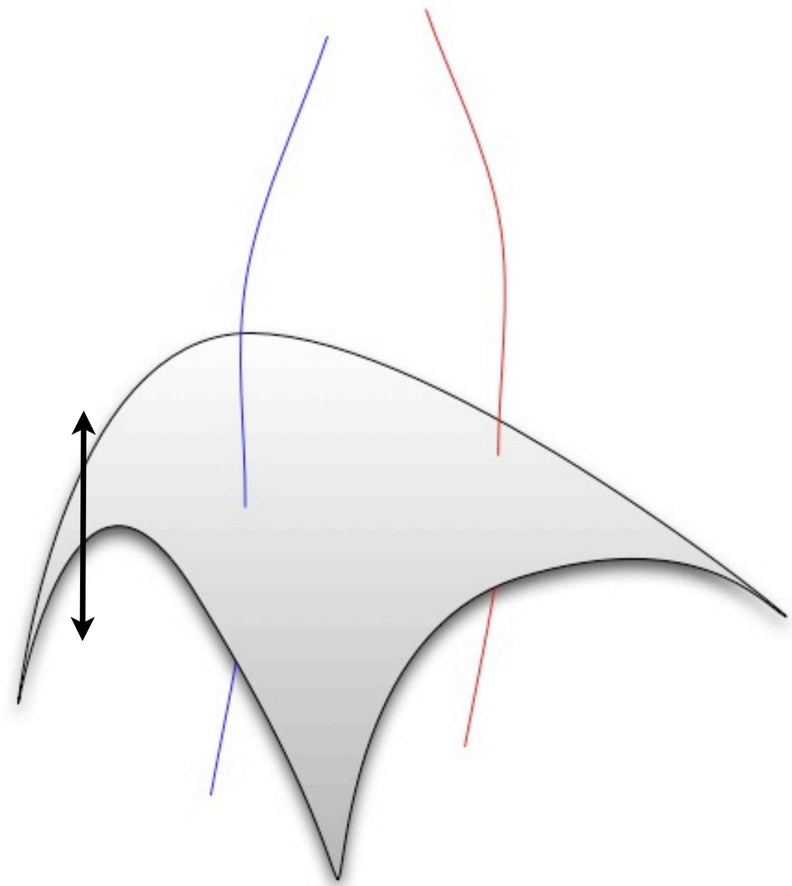
groups

Definition A.13 (Group). *A* group *is a set G with an operation "$\circ$" on the elements of G that:*

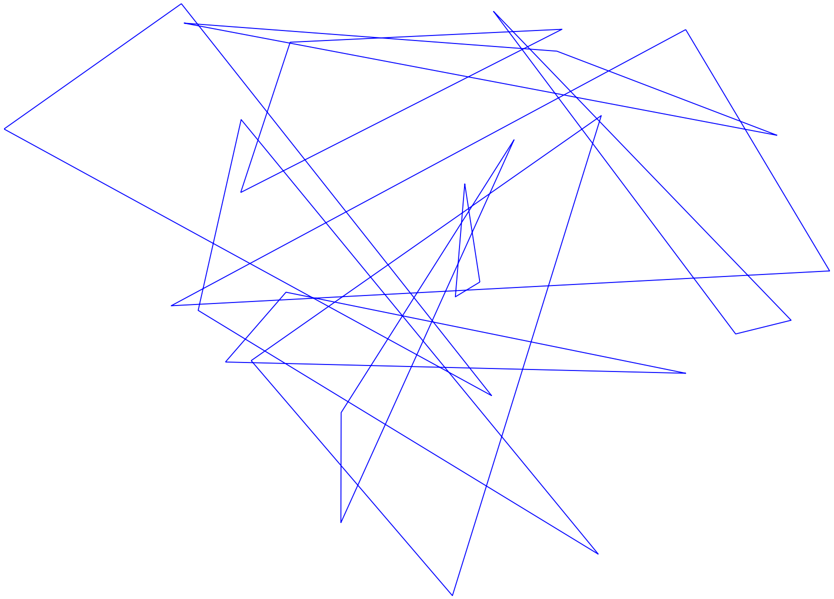
- is closed *: If $g_1, g_2 \in G$, then also $g_1 \circ g_2 \in G$;*
- is associative*: $(g_1 \circ g_2) \circ g_3 = g_1 \circ (g_2 \circ g_3)$, for all $g_1, g_2, g_3 \in G$;*
- has a unit element *e: $e \circ g = g \circ e = g$, for all $g \in G$;*
- is invertible: *For every element $g \in G$, there exists an element $g^{-1} \in G$ such that $g \circ g^{-1} = g^{-1} \circ g = e$.*

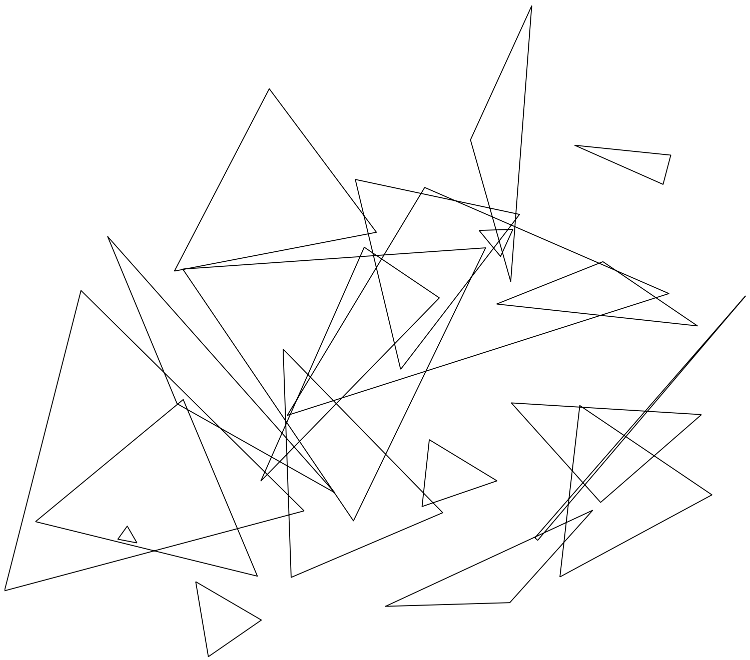
- e.g., translation, rotation (isometry, rigid motion SE (N)), scaling (similarity), affine, projective ... diffeomorphism; contrast
- groups "act"

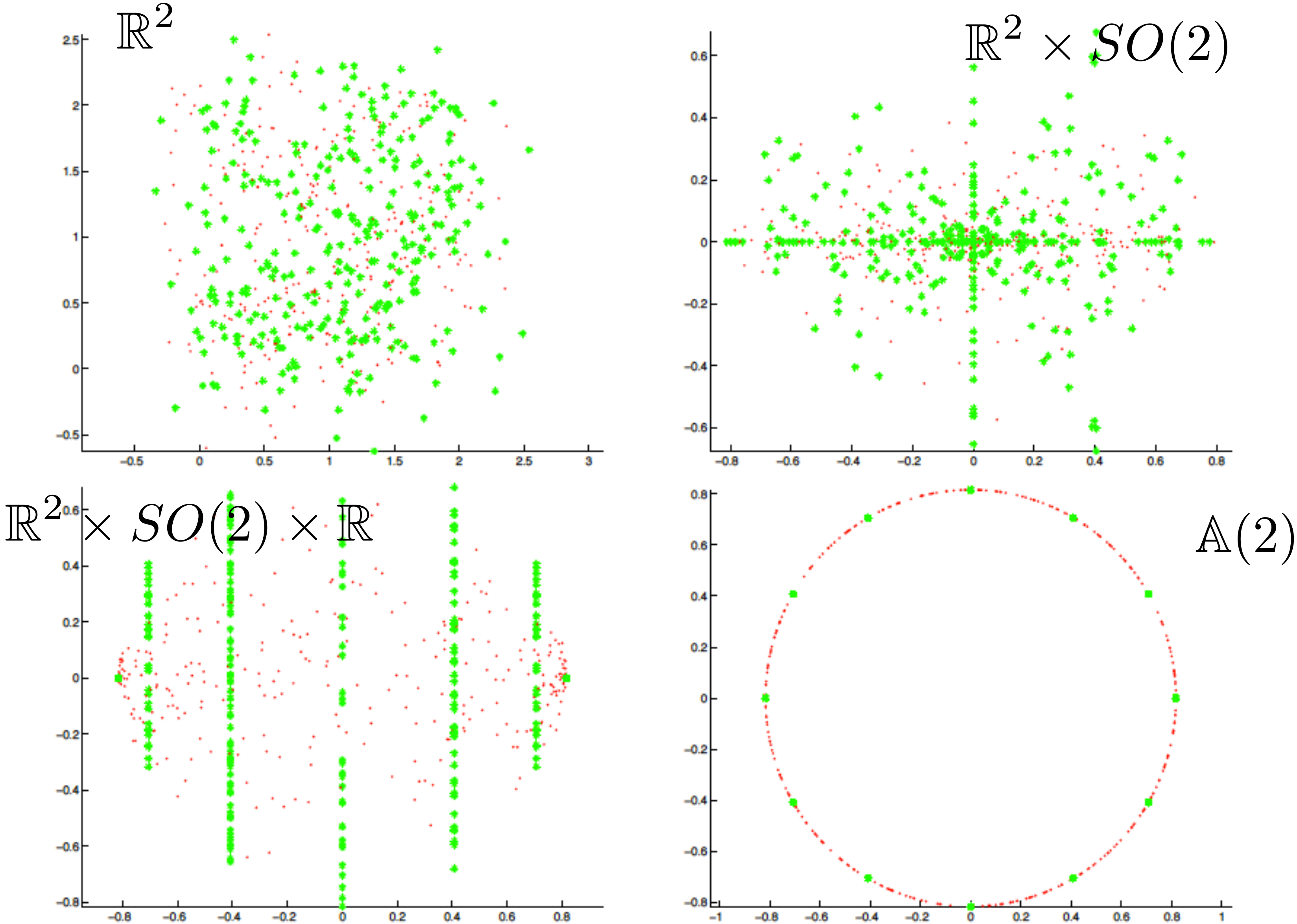
orbits
equivalence classes
base/quotient



e.g. shape space $\mathbb{R}^{M\times N}/SE(N)$

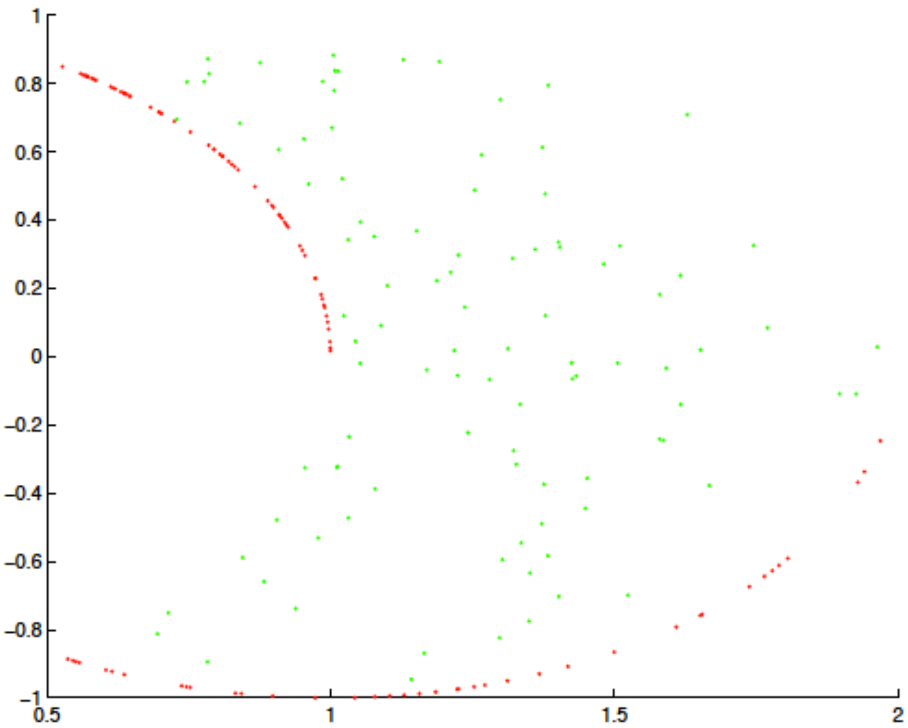




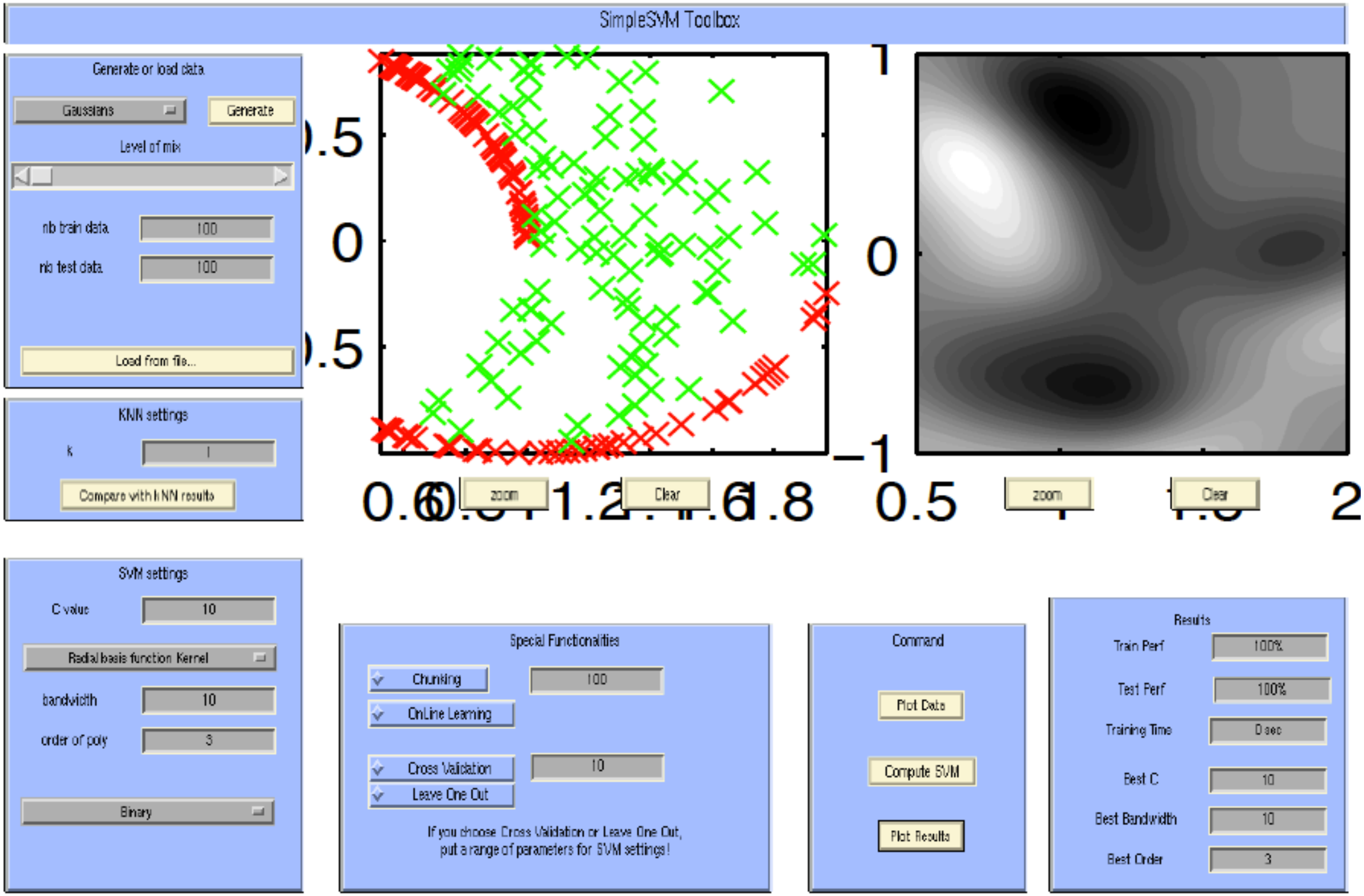


$\mathbb{R}^2$
2.5
2
1.5
1
0.5
0
-0.5
-0.5 0 0.5 1 1.5 2 2.5 3
$\mathbb{R}^2 \times SO(2)$
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8 -0.6 -0.4 -0.2 0 0.2 0.4 0.6 0.8
$\mathbb{R}^2 \times SO(2) \times \mathbb{R}$
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-0.8 -0.6 -0.4 -0.2 0 0.2 0.4 0.6 0.8
$\mathbb{A}(2)$
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1 -0.8 -0.6 -0.4 -0.2 0 0.2 0.4 0.6 0.8 1

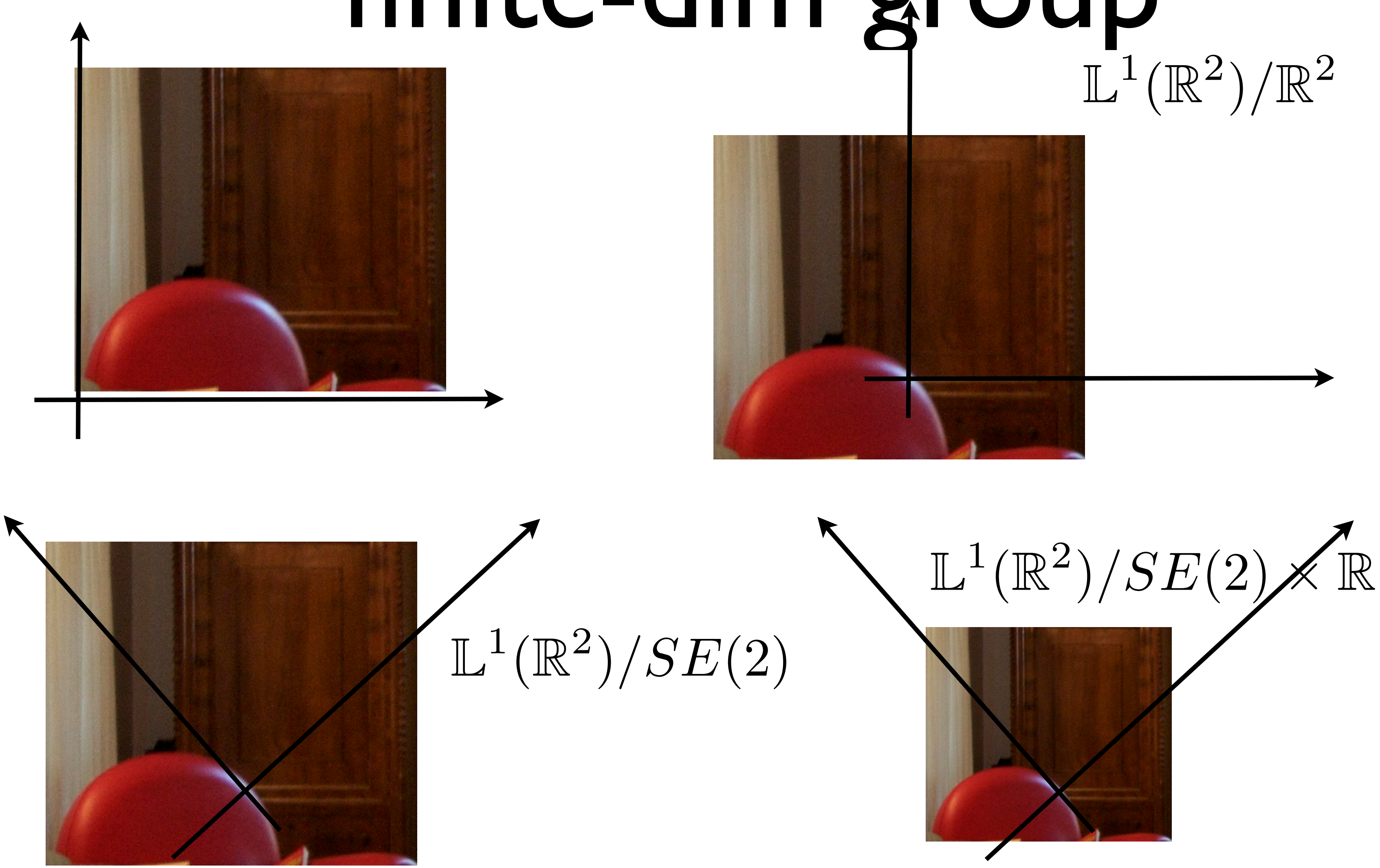
singular perturbations



procrustes


SimpleSVM Toolbox
Generate or load data
Gaussians
Generate
Level of mix
nb train data
100
nb test data
100
Load from file...
KNN settings
k
1
Compare with kNN results
zoom
Clear
zoom
Clear
SVM settings
C value
10
Radial basis function Kernel
bandwidth
10
order of poly
3
Binary
Special Functionalities
Chunking
100
OnLine Learning
Cross Validation
10
Leave One Out
If you choose Cross Validation or Leave One Out,
put a range of parameters for SVM settings!
Command
Plot Data
Compute SVM
Plot Results
Results
Train Perf
100%
Test Perf
100%
Training Time
0 sec
Best C
10
Best Bandwidth
10
Best Order
3

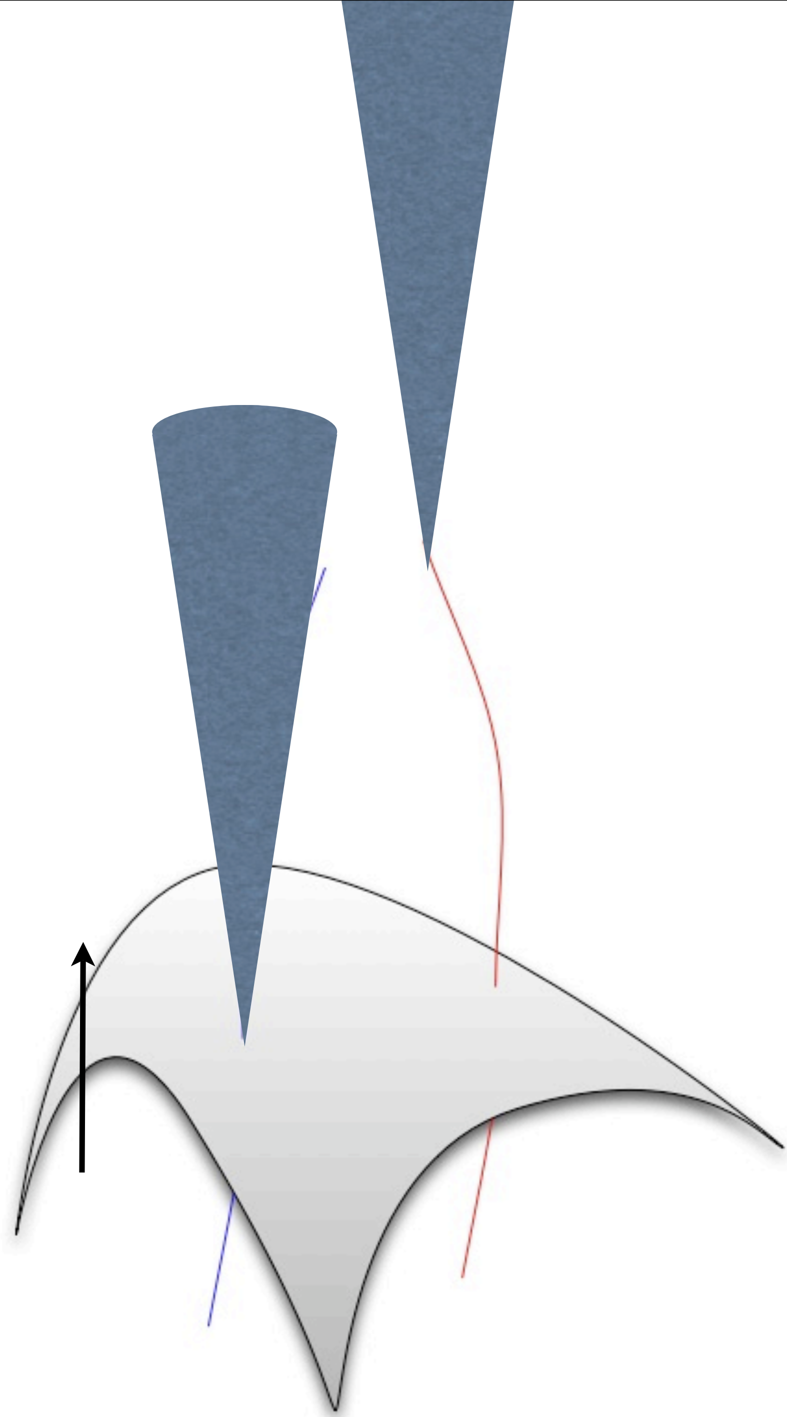
infinite-dim space, finite-dim group

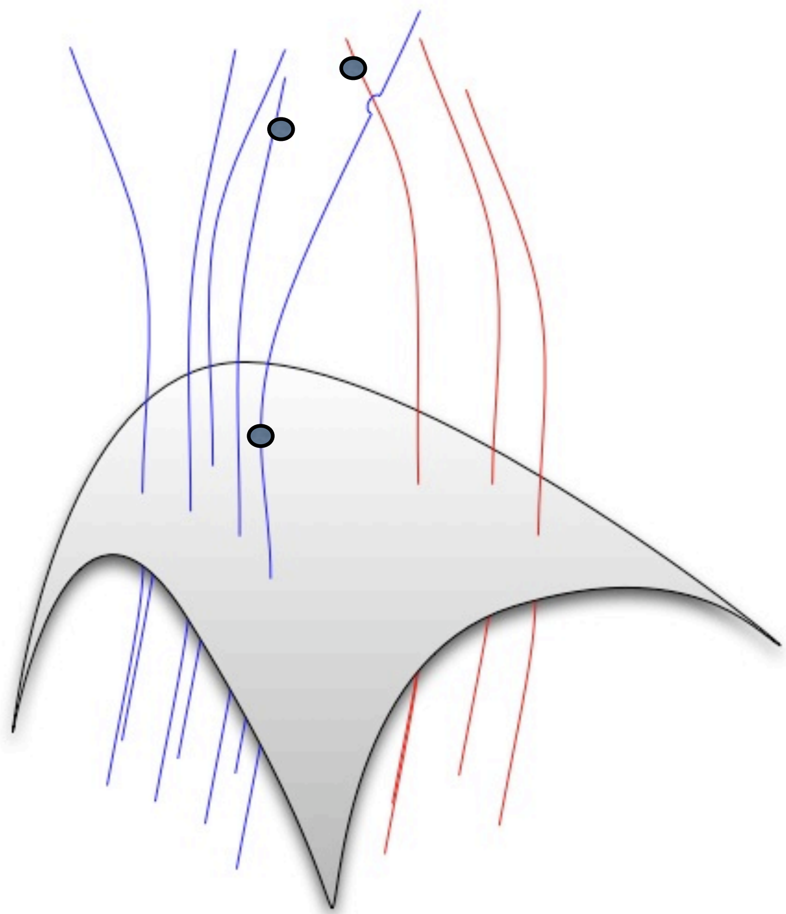


infinite-dim space, infinite-dim group?

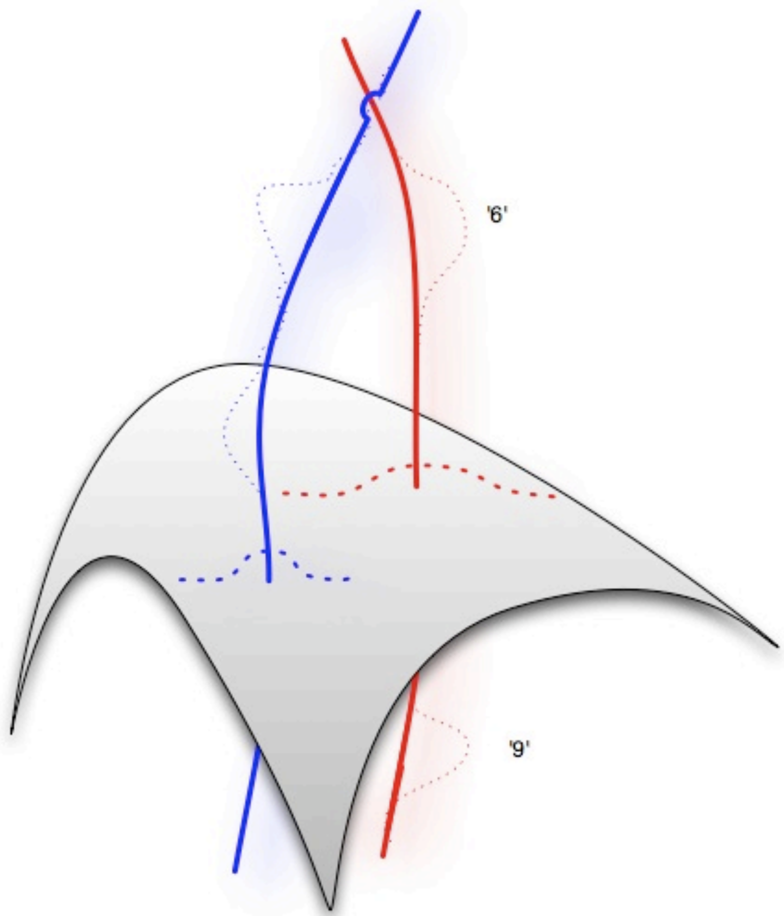
- symbols ...

semi-orbits





- marginalization, max-out, canonization

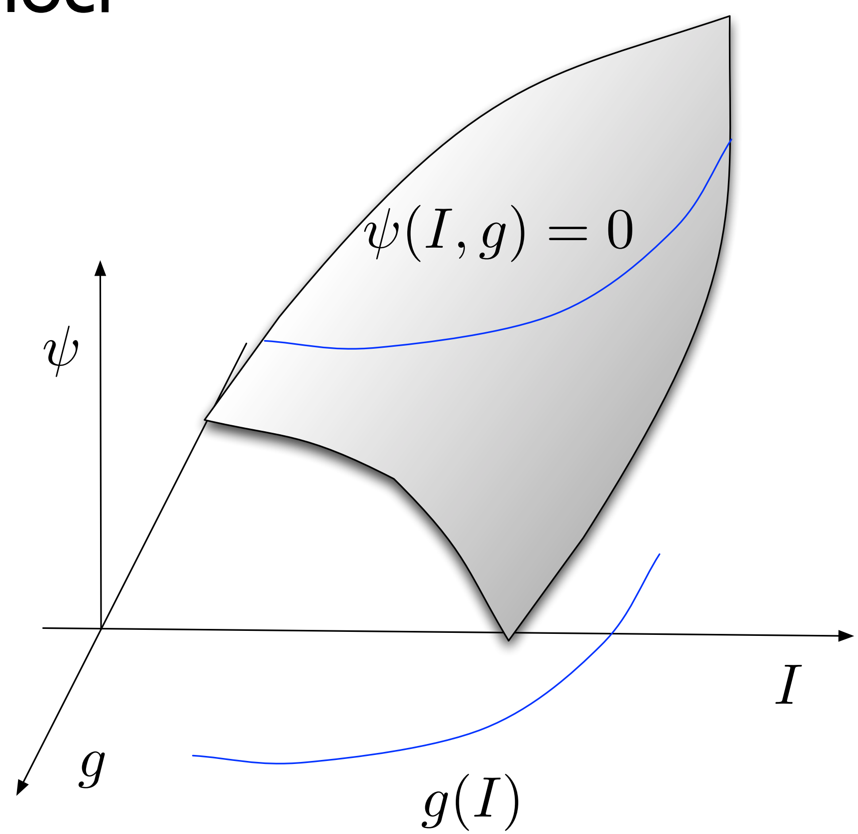


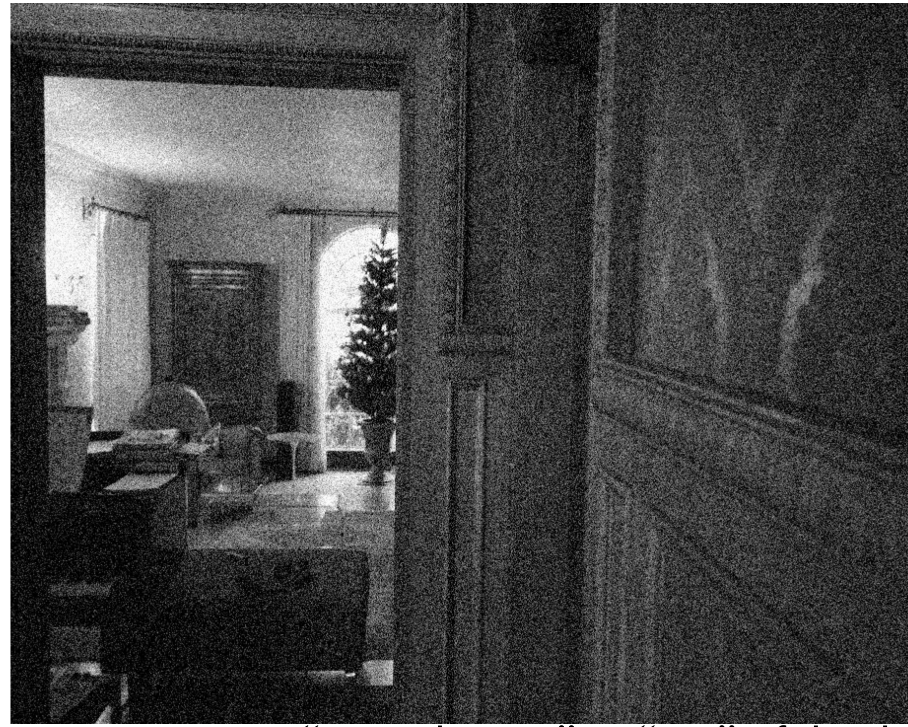
basic diff. topology

- transversality, critical loci

$$\psi(I, g) = 0 \Rightarrow g = g(I)$$

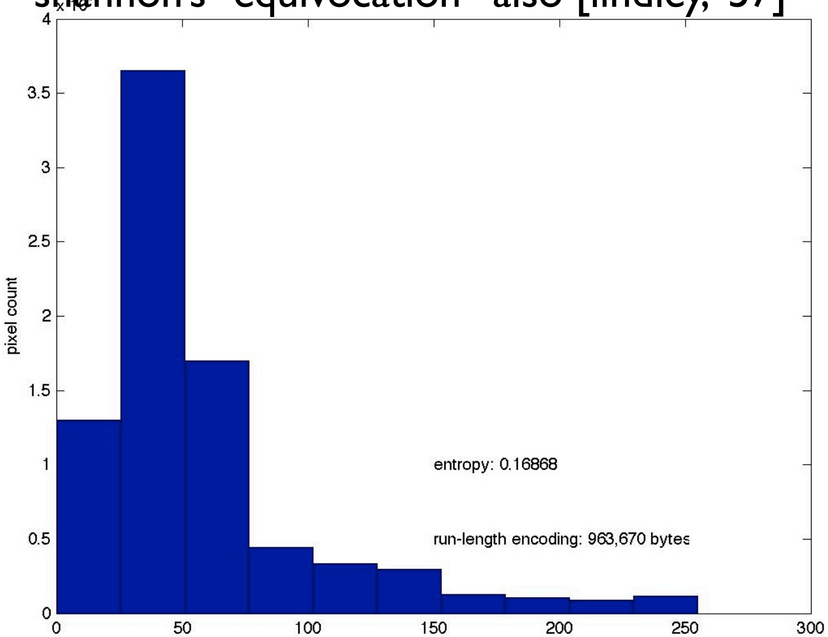
$$\det\left(\frac{\partial \psi}{\partial g}\right) \neq 0$$

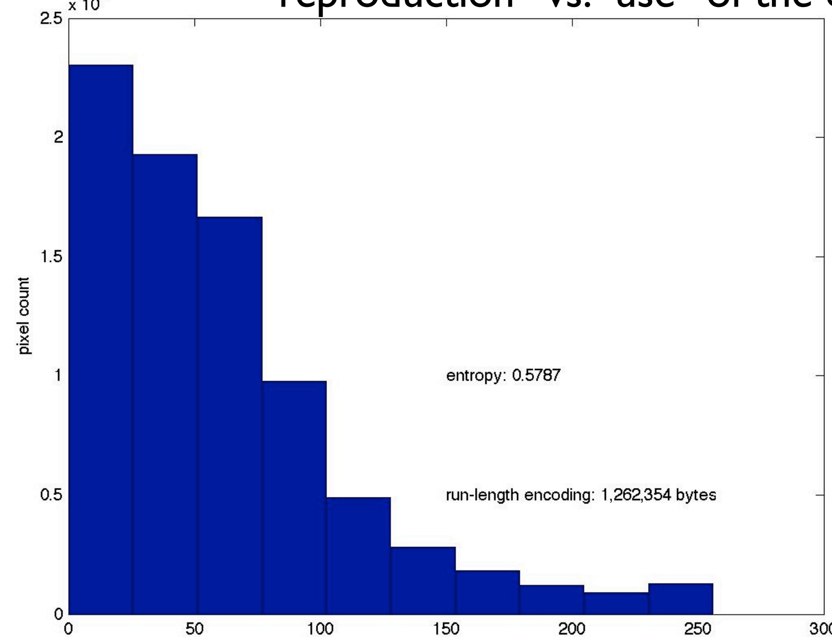




shannon's "equivocation" also [lindley, '57]

"reproduction" vs. "use" of the data





gibson's information

- task ⇨ data = "information" & (structured) "nuisance"
- information = complexity of the data after the effects of nuisances has been discounted
    - nuisances in vision:
        - viewpoint
        - illumination
        - visibility (occlusion, cast shadows)
        - quantization/noise

gibson: *"my notion is that information consists of invariants underlying change [...] of illumination, point of observation, overlapping samples [...] and disturbance of structure"*

is a "gibsonian information theory" viable? (take I)

- general-case viewpoint invariants do not exist [burns et al., '92]
- non-trivial illumination invariants do not exist [chen et al., '00]

is a “gibsonian information theory” viable? (take II)

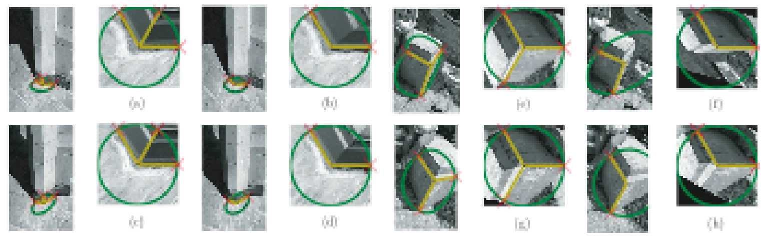
- ☑ general-case viewpoint invariants do exist, and are non-trivial, for lambertian scenes in ambient light [vedaldi-soatto ‘05-’06]
- ☑ non-trivial contrast invariants do exist, and are sufficient statistics [morel & c., ’93-’05]
- ☐ what is invariant to contrast (geometry of the level lines) is not invariant to viewpoint
- ☐ what is invariant to viewpoint (image range in a canonized domain) is not invariant to contrast

is a “gibsonian information theory” viable? (take III)

- ☑ general-case viewpoint invariants exist, and are non-trivial, for lambertian scenes in ambient light [vedaldi-soatto ‘05-’06]
- ☑ non-trivial contrast invariants exist, and are sufficient statistics [morel & c., ’93-’05]
- ☑ viewpoint-illumination invariants exist (ambient-lambert)
- ☑ they are “discrete” structures (attributed reeb tree, ART), supported on a thin set
- ☑ they are sufficient statistics! (equivalent to the image up to changes of viewpoint and contrast) [sundaramoorthi et al., ’09]

"the set of images modulo viewpoint and contrast changes"

[sundaramoorthi-petersen-varadarajan-soatto '09]



- viewpoint changes induce (epipolar-homeomorphic) deformations of the image domain; diffeomorphic closure (general non-planar surfaces)
- viewpoint-contrast invariants exists
- they are (supported on) a zero-measure subset of the image domain (attributed reeb tree)
- they are sufficient statistics! (equivalent to the image up to contrast and viewpoint transformations)

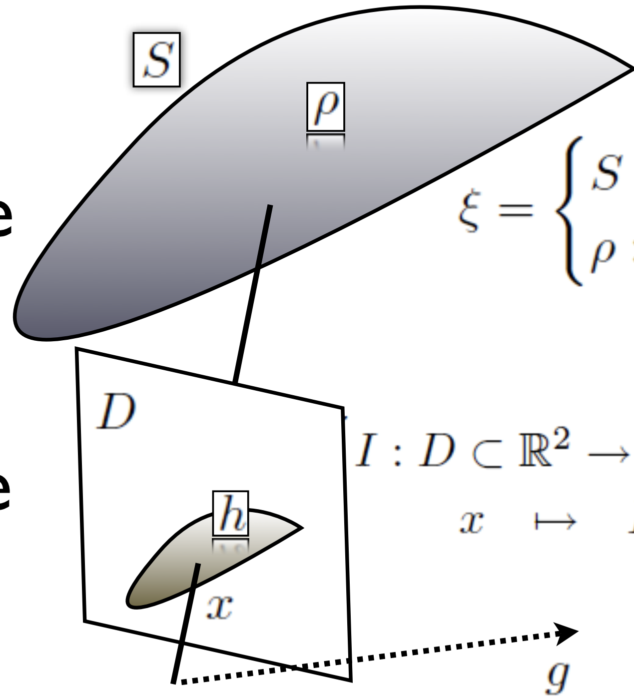
the ART

- infinite-dimensional space, infinite-dimensional group
- quotient of morse functions of the plane (dense in L1) modulo domain diffeomorphisms
- closure of epipolar domain deformations is the entire group of diffeomorphisms

is a “gibsonian information theory” viable? (take III)

- ☑ general-case viewpoint invariants exist, and are non-trivial, for lambertian scenes in ambient light [vedaldi-soatto ‘05-’06]
- ☑ non-trivial contrast invariants exist, and are sufficient statistics [morel & c., ’93-’05]
- ☑ viewpoint-illumination invariants exist (ambient-lambert)
- ☑ they are “discrete” structures (attributed reeb tree, ART), supported on a thin set
- ☑ they are sufficient statistics! (equivalent to the image up to changes of viewpoint and contrast) [sundaramoorthi et al., ’09]
- ☐ occlusions and quantization admit no invariants!

some notation



scene

$$\xi = \begin{cases} S \subset \mathbb{R}^3 & \text{shape (geometry)} \\ \rho : S \to \mathbb{R}^k & \text{radiance (reflectance)} \end{cases}$$

image

$$\begin{array}{rcl} I : D \subset \mathbb{R}^2 & \to & \mathbb{R}^k_+ \\ x & \mapsto & I(x) \end{array}$$

$\mathbb{L}^2$ integrable $\overset{\text{dense}}{\to}$ morse

$k = 1, 3, N$ spectral bands

nuisance

$$\nu = \begin{cases} h : \mathbb{R}^k \to \mathbb{R}^k & \text{contrast } (\simeq \text{illumination}) \\ g : \mathcal{Z}_+ \to SE(3) & \text{motion (viewpoint)} \end{cases}$$

lambert-ambient

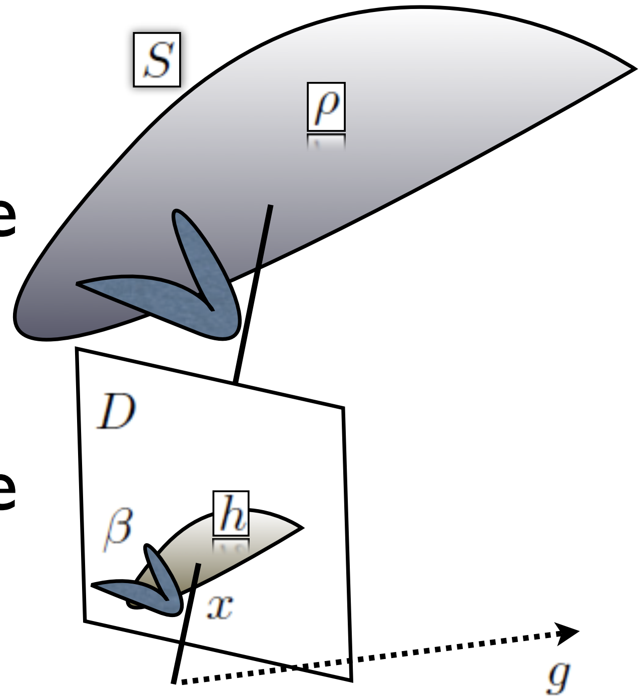
$$\boxed{\begin{cases} I(x,t) = h(t) \circ \rho(p) + n(x,t) \\ x = \pi(g(t)p) \end{cases}}$$

$$I(x,t) = f(\underbrace{\rho, S}_{\xi}; \underbrace{g, h, n}_{\nu})$$

some notation

scene

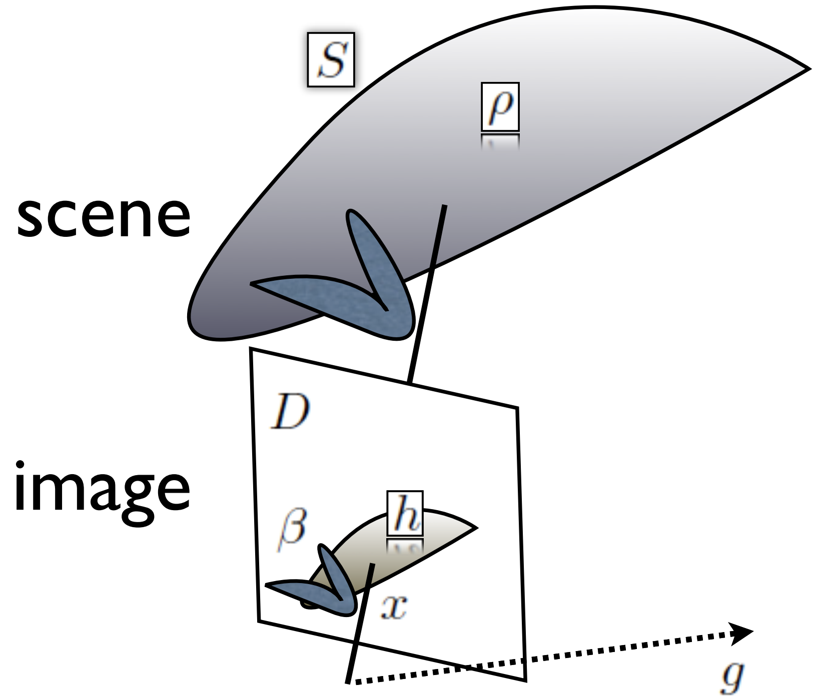
image



occlusions
$$I(x) = \begin{cases} f(\rho, S; g, h, n) & x \in D\backslash\Omega \\ \beta(x) & x \in \Omega \end{cases}$$

lambert-ambient
$$\boxed{\begin{cases} I(x,t) = h(t) \circ \rho(p) + n(x,t) \\ x = \pi(g(t)p) \end{cases}} \quad I(x,t) = f(\underbrace{\rho, S}_{\xi}; \underbrace{g, h, n}_{\nu})$$

some notation

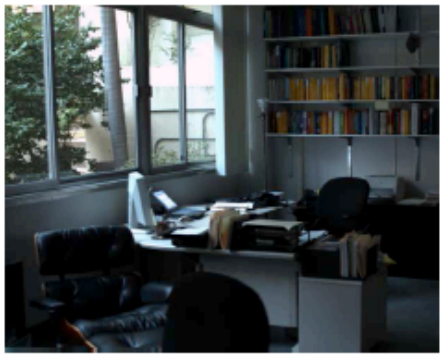


nuisance $\nu = g \quad h \quad \beta \quad n$

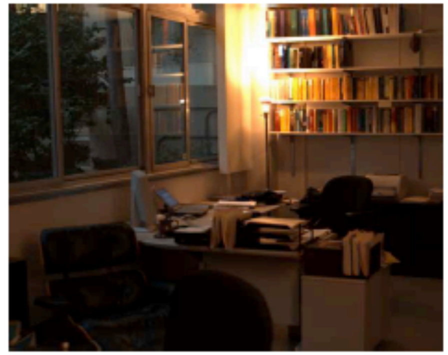
image formation model
(formal notation)

$$\boxed{I = f(\xi, \nu)}$$

$$I = f(g\xi, \nu) + n$$

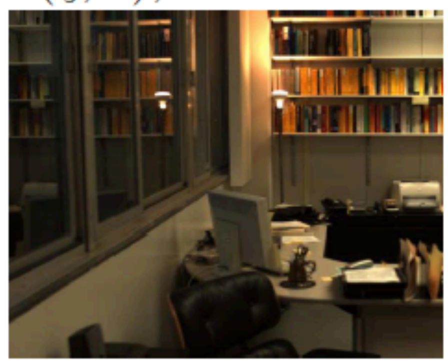


$I = h(\xi, \nu)$



$\tilde{I} = h(\xi, \tilde{\nu})$, $\tilde{\nu}$ = illumination

$\tilde{\nu}$ = visibility



$\tilde{\nu}$ = viewpoint

$\tilde{I} = h(\tilde{\xi}, \tilde{\nu})$, $\tilde{\xi} \neq \xi$

some definitions

feature $\begin{array}{rcl} \phi : \{I(x), x \in D\} & \rightarrow & \mathbb{R}^K \\ I & \mapsto & \phi(I) \end{array}$

invariant $\boxed{\phi \circ f(\xi, \nu) = \phi \circ f(\xi, \bar{\nu}) \quad \forall\, \nu, \bar{\nu};\ \forall\, \xi}$

maximal invariant $\boxed{\phi^\wedge(I)}$

sufficient statistic $\boxed{\phi \mid R(u|I) = R(u|\phi(I))}$

conditional risk $R(u|I) \doteq \int L(u, \bar{u}) dP(\bar{u}|I)$

loss function L decision/control policy u

minimal sufficient statistic $\boxed{\phi^\vee_\xi(I)}$

representation and hallucination

given one or more images $\{I\}$ a representation

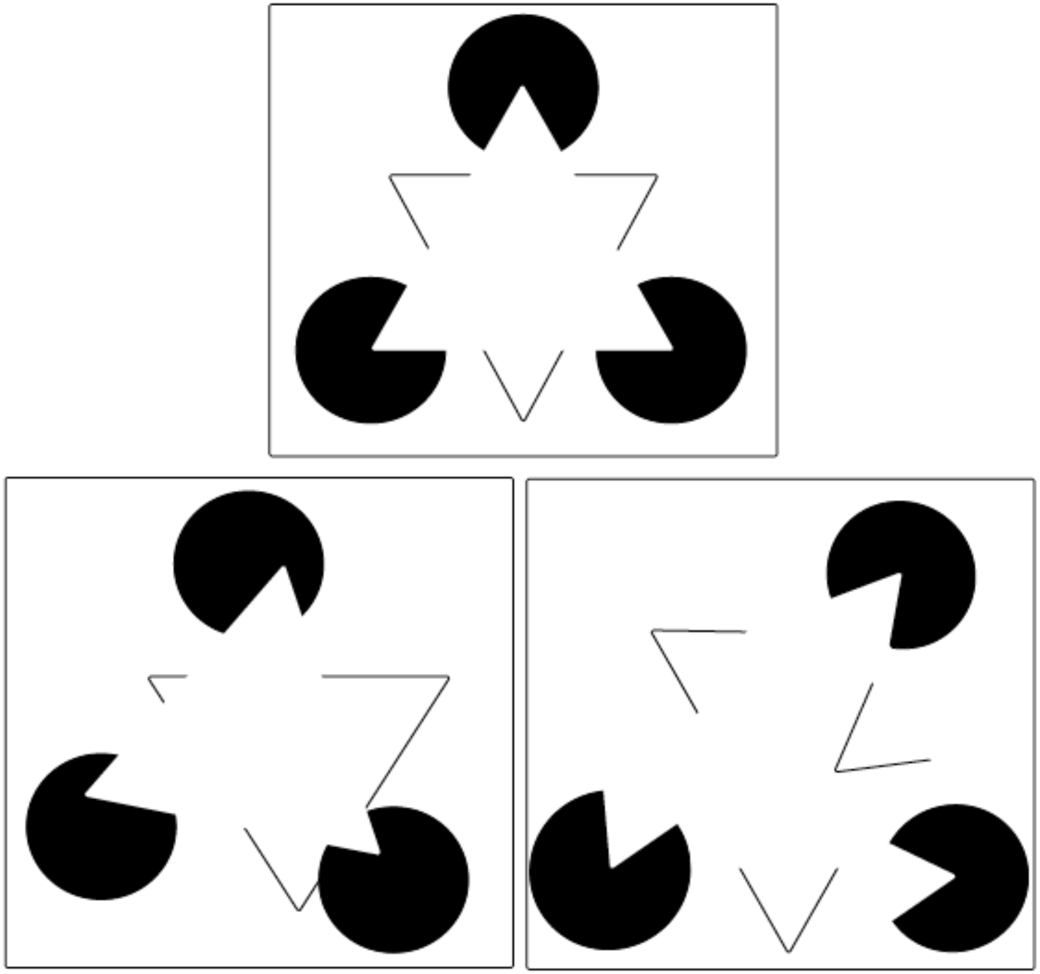
$\hat{\xi}$ is a statistic $\hat{\xi} = \phi(\{I\})$ such that

$$\boxed{\{I\} \in \{f(g\hat{\xi}, \nu), \quad g \in G, \nu \in \mathcal{V}\} \doteq \mathcal{L}(\hat{\xi})}$$

i.e., it is a statistic from which
the images can be hallucinated

$$\mathcal{L}(\hat{\xi}) = \mathcal{L}(\xi)$$

complete representation
minimal complete representation
(note it is invariant to G)



information gap

- actionable information: coding length of a maximal invariant statistic; can be computed from an image.

$$\boxed{\mathcal{H}(I) \doteq H(\phi^\wedge(I))}$$

- complete information: coding length of a minimal sufficient statistic of a (complete) representation

$$\boxed{\mathcal{I} = H(\phi^\vee(\hat{\xi}))}$$

- actionable information gap (AIG)

$$\boxed{\mathcal{G}(I) \doteq \mathcal{I} - \mathcal{H}(I)}$$

invertible nuisances

invertible nuisance $f(\xi, \emptyset) \mapsto f(\xi, \nu)$ injective $\boxed{\mathcal{G} = 0}$

contrast $\nu = h$ $\boxed{\phi^\wedge(I) = \frac{\nabla I(x)}{\|\nabla I(x)\|} \quad (\equiv \text{geom. level curves})}$

viewpoint
$$\nu = \begin{cases} w : D \subset \mathbb{R}^2 \to \mathbb{R}^2 \\ \qquad x \mapsto w(x) = \pi \circ g^{-1} \circ \pi^{-1}(x) \end{cases}$$

$\boxed{\phi^\wedge(I) = ART}$ away from occlusions

(non)invertible nuisances

- visibility (occlusions, cast shadows); quantization
- invertibility depends on the sensing process: control authority
- j. j. gibson: *"the occluded becomes unoccluded"* in the process of "information pickup"

is a “gibsonian information theory” viable? (take IV)

- ☑ general-case viewpoint invariants exist, and are non-trivial, for lambertian scenes in ambient light [vedaldi-soatto ‘05-’06]
- ☑ non-trivial contrast invariants exist, and are sufficient statistics [morel & c., ’93-’05]
- ☑ viewpoint-illumination invariants exist (ambient-lambert)
- ☑ they are “discrete” structures (attributed reeb tree, ART), supported on a thin set
- ☑ they are sufficient statistics! (equivalent to the image up to changes of viewpoint and contrast) [sundaramoorthi et al., ’09]
- ☑ occlusions and quantization are invertible! [gibson ’50s]

part II
how?

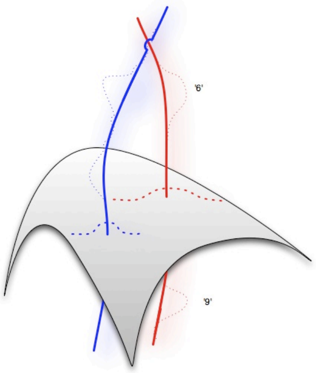
canonization, commutativity, structural stability,
proper sampling, exploration

how to deal with nuisances (aside)

- marginalization (bayes)
- extremization/max-out
- canonization

1. marginalization (bayes)

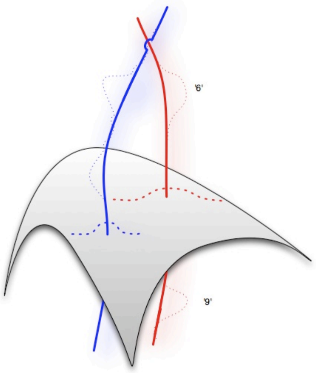
$$p(I|\xi) = \int p(I|\xi, \nu) dP(\nu)$$



- average over all possible nuisances, weighted by their own pdf (complex integration at run-time)
- can be learned (approximate w/vicinal risk)

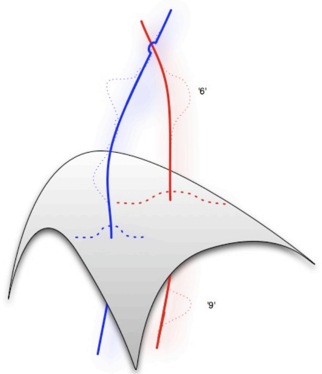
2. registration (maximum-likelihood)

$$\tilde{p}(I|\xi) = \sup_{\nu} p(I|\xi, \nu)$$



- find the nuisance together with the variable of interest (solve optimization (search) at run-time)

3. canonization



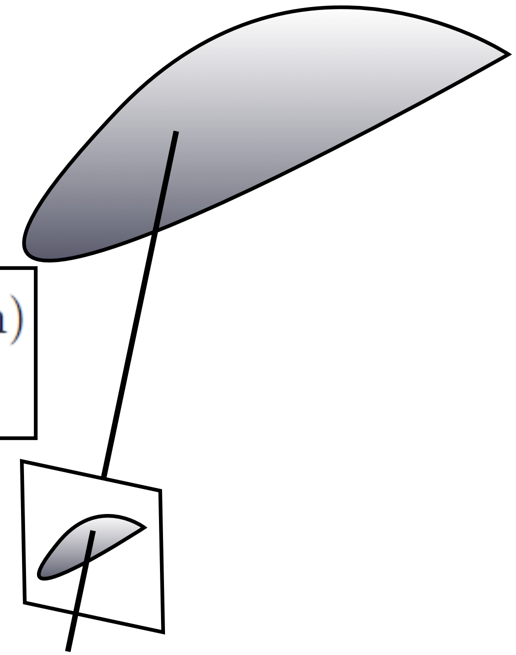
$$\phi(I)$$

- can we find a representation of the data that "does not depend on the nuisance (invariant) and yet "contains all the information" (sufficient statistic)?

which to use?

some notation

image $\begin{cases} I : D \subset \mathbb{R}^2 \to \mathbb{R}^k_+ & \mathbb{L}^2 \text{ integrable} \overset{\text{dense}}{\to} \text{ morse} \\ \quad x \mapsto I(x) & k = 1, 3, N \text{ spectral bands} \end{cases}$ $\{I(x,t)\}_{t=0}^T$, or $\{I\}$



scene $\xi = \begin{cases} S \subset \mathbb{R}^3 & \text{shape (geometry)} \\ \rho : S \to \mathbb{R}^k & \text{radiance (reflectance)} \end{cases}$

nuisance $\nu = \begin{cases} h : \mathbb{R}^k \to \mathbb{R}^k & \text{contrast } (\simeq \text{ illumination}) \\ g : \mathcal{Z}_+ \to SE(3) & \text{motion (viewpoint)} \end{cases}$

lambert-ambient $\begin{cases} I(x,t) = h(t) \circ \rho(p) + n(x,t) \\ x = \pi(g(t)p) \end{cases}$

visibility $I(x) = \begin{cases} f(\rho, S; g, h, n) & x \in D \backslash \Omega \\ \beta(x) & x \in \Omega \end{cases}$ $I(x,t) = f(\underbrace{\rho, S}_{\xi}; \underbrace{g, h, n}_{\nu})$

image formation model (formal notation) $\boxed{I = f(\xi, \nu)}$

actionable information increment

- must act to "invert occlusions" (optical flow):

$$\Omega(t, dt) = \arg\min_{\Omega, w} \int_{D\setminus\Omega} (I(w(x,t),t) - I(x, t+dt))^2 dx + \int_D \|\nabla w\|_1 dx + \int_\Omega dx$$

- innovation and Actionable Information Increment (AIN)

$$\boxed{\epsilon(I, t+dt) \doteq \phi^\wedge(I_{t+dt\,|_\Omega})} \qquad \boxed{AIN = H(\epsilon(I, t+dt)) = \mathcal{H}(I_{t+dt\,|_\Omega})}$$

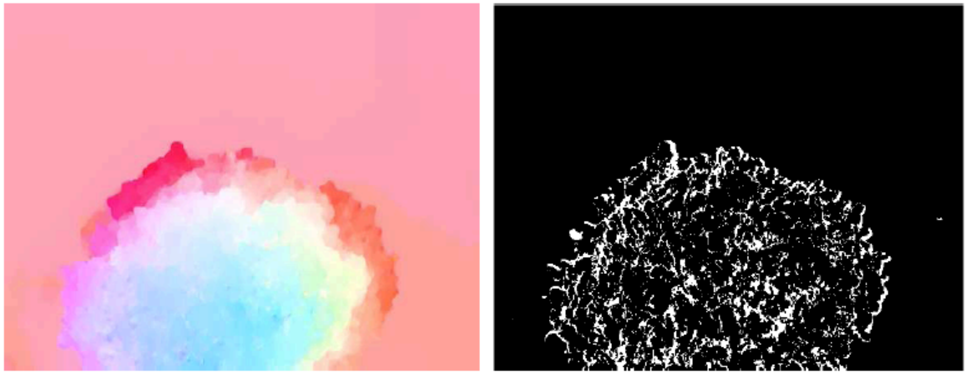
- (memoryless) perceptual exploration: value-of-information, next-best-view, actionableive vision etc.

$$\hat{u}_t = \arg\max_u AIN(I, t; u)$$

optimal occlusion detection

$$\Omega(t, dt) = \arg\min_{\Omega, w} \int_{D \setminus \Omega} (I(w(x,t),t) - I(x, t+dt))^2 dx + \int_D \|\nabla w\|_1 dx + \int_\Omega dx$$

- most optical flow literature neglects occlusions
- motion at occluded regions is not discontinuous, it does not exist
- difficult optimization problem, can't use trivial regularizers



$$\boxed{I(x,t) = \begin{cases} I(w(x,t), t+dt) + n(x,t), & x \in D\backslash\Omega(t;dt) \\ \rho(x,t), & x \in \Omega(t;dt) \end{cases}}$$

$$\text{(i)} \quad \lim_{dt \to 0} \Omega(t;dt) = \emptyset, \quad \text{and} \quad \text{(ii)} \;\; n \overset{IID}{\sim} \mathcal{N}(0,\lambda)$$

$$\begin{cases} e_1(x,t;dt) \doteq \rho(x,t) - I(w(x,t), t+dt), & x \in \Omega \\ e_2(x,t;dt) \doteq n(x,t), & x \in D\backslash\Omega. \end{cases}$$

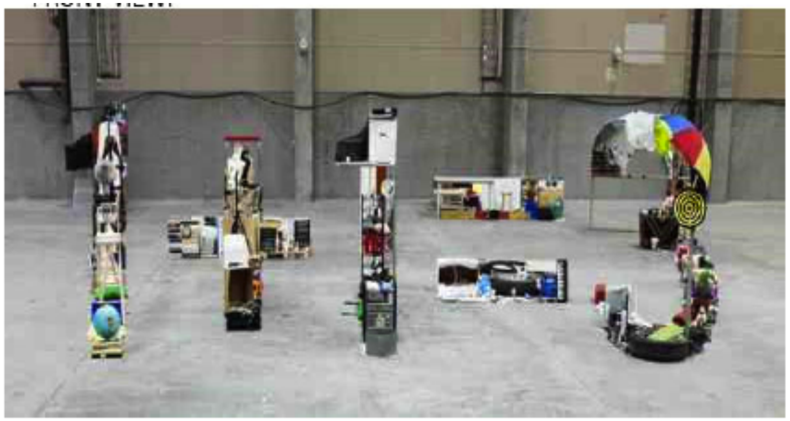
large but sparse (i)

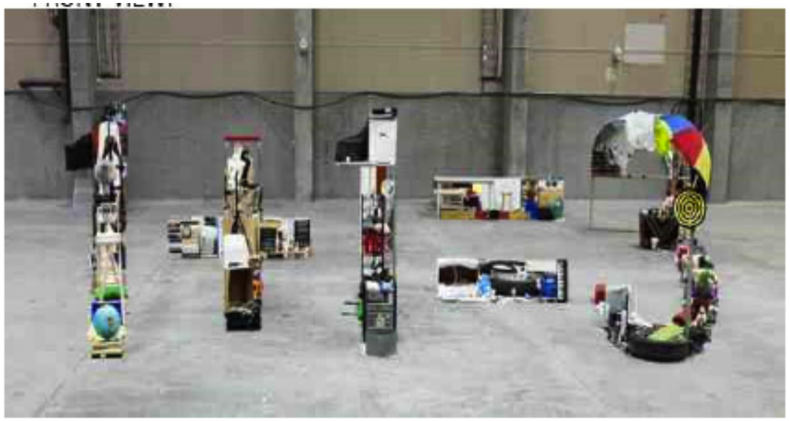
dense but small (ii)

$$I(x,t) = I(w(x,t), t+dt) + e_1(x,t;dt) + e_2(x,t;dt)$$

$$\boxed{\psi_{\text{data}}(v, e_1) = \|\nabla I v + I_t - e_1\|_{\mathbb{L}^2(D)} + \lambda \|e_1\|_{\mathbb{L}^0(D)}.}$$

relax to convex optimization (nesterov)

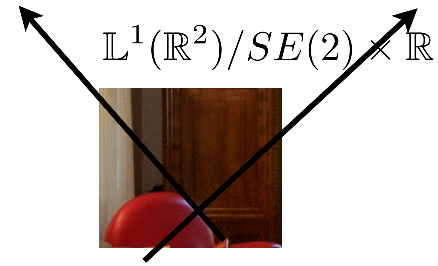



how to build representations?

1. canonizability (sparse yet lossless)
2. commutativity (beyond existing local descriptors)
3. structural stability (BIBO vs. structural stability)
4. proper sampling (beyond nyquist)
5. exploration (gibson)

canonizability



co-variant detector: a functional $\psi : \mathcal{I} \times G \rightarrow \mathbb{R}^{dim(G)}; (I, g) \mapsto \psi(I, g)$

I. the zero-level set $\psi(I, g) = 0$ uniquely determines $\hat{g} = \hat{g}(I)$

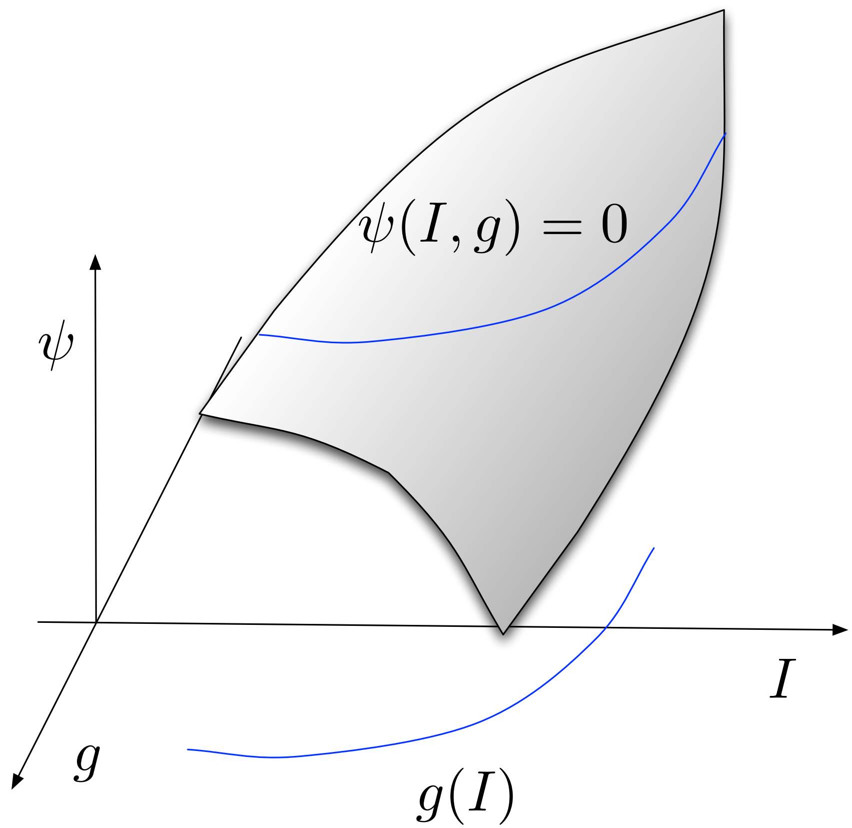
II. if $\psi(I, \hat{g}) = 0$ then $\psi(I \circ g, \hat{g} \circ g) = 0 \quad \forall\, g \in G$

canonizable: an image region is canonizable if it admits at least one co-variant detector

canonized descriptor: $\phi(I) \doteq I \circ \hat{g}^{-1}(I) \ \mid\ \psi(I, \hat{g}(I)) = 0$

transversality

$$\det\left(\frac{\partial\psi}{\partial g}\right) \doteq |\nabla\psi| \neq 0.$$



examples

- harris: bad (non-commutative) $\psi(I, g) = \det\left(\int_{\mathcal{B}_g(\sigma)} \nabla I^T \nabla I dx\right)$
- LoG: good (linear) $\psi(I, g) = \nabla^2 \mathcal{N}(Rx + T; \sigma)$
- HoG: better (monge-ampere) $\psi = \nabla|\nabla\psi|$
  - under wiener's illumination model: $\mathcal{G} * I = \nabla|\nabla\mathcal{G} * I|$
- TST: best (demo later)
- moments of the superpixel tree (quickshift)

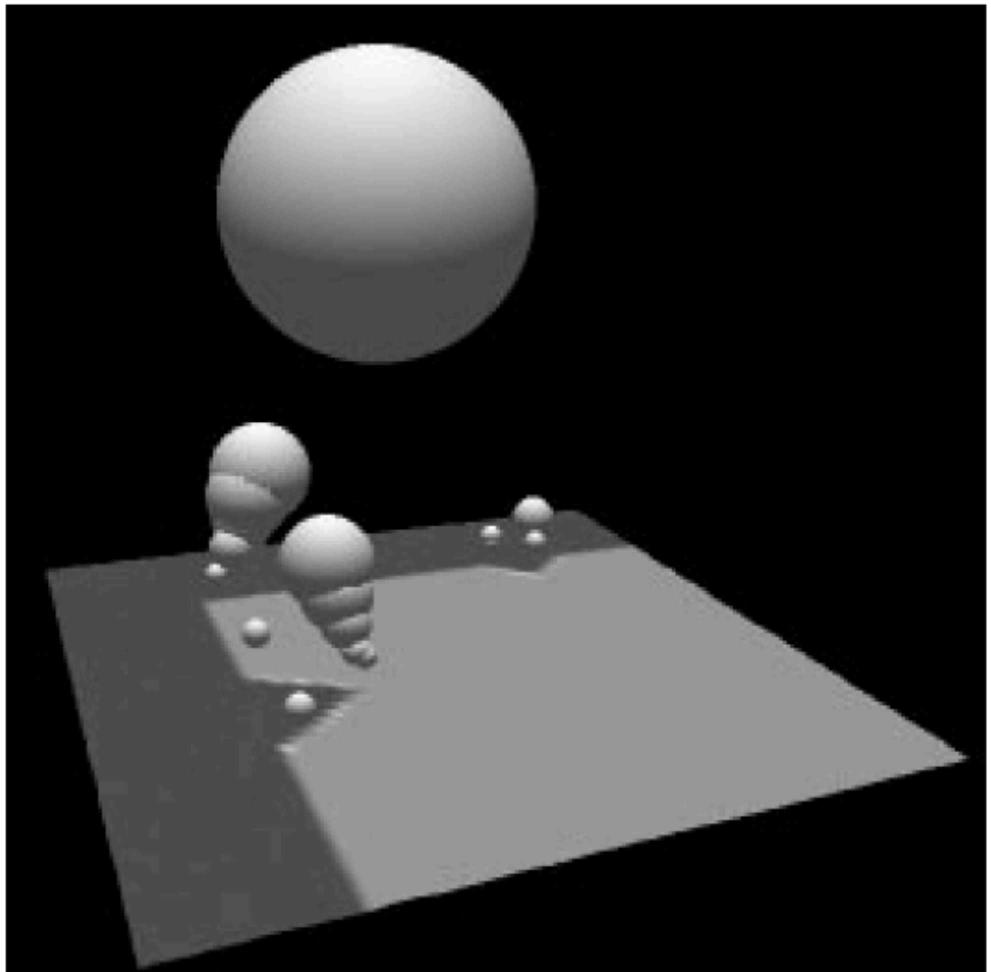
what is the “best” descriptor? when is it optimal? 1. canonizability

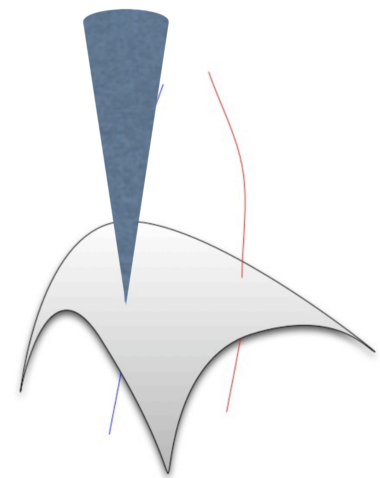
- Thm 1: canonized descriptors are complete invariant statistics (wrt canonized group)
- Thm 2: if a complete invariant descriptor can be constructed, an equi-variant classifier can be designed that attains the Bayes’ risk
- the best descriptor can be derived analytically (BTD)
- What about non-group nuisances?

2. commutativity

- commutative nuisance: $I \circ g \circ \nu = I \circ \nu \circ g$
- Thm 3: the only nuisances that are invertible and commutative are the isometric group of the plane and contrast range transformations
- Corollary: do not canonize scale (nor affine/ projective transformations)
- (Thm 5: an image region is a texture if and only if it is not canonizable)

e.g. canonize vs. sample





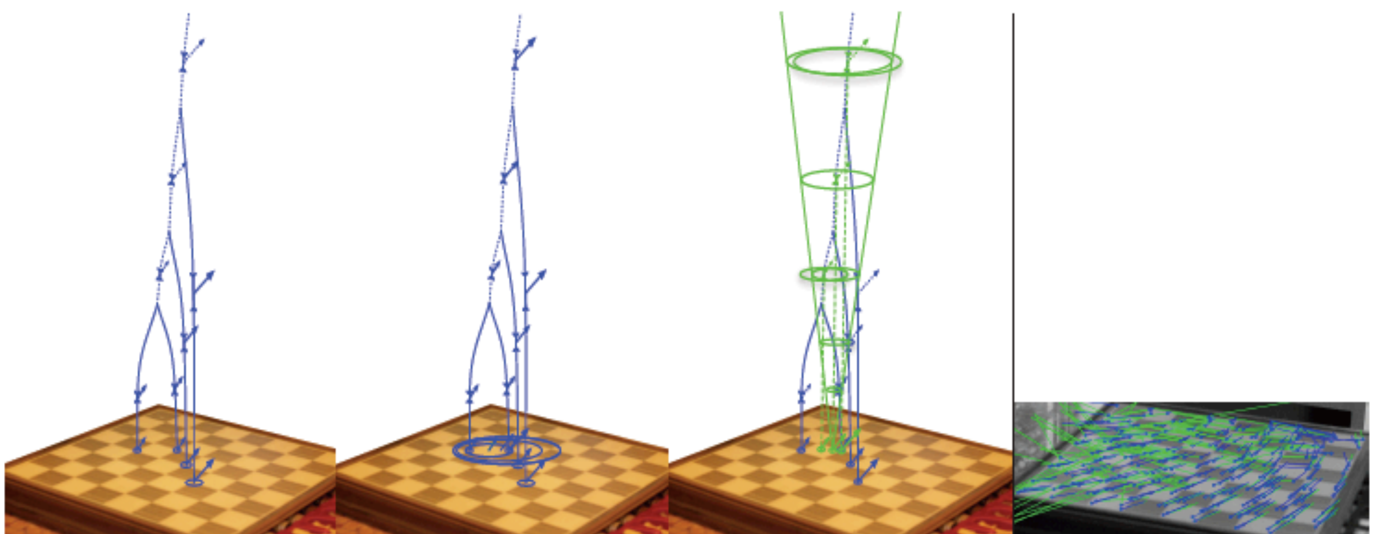
t. lindeberg

3. BIBO stability (sensitivity)

- BIBO sensitivity: a detector is BIBO insensitive (stable) if small nuisance variations cause small changes in the canonical element.
- Thm 6: any co-variant detector is BIBO stable
- BIBO stability is irrelevant for visual decisions!

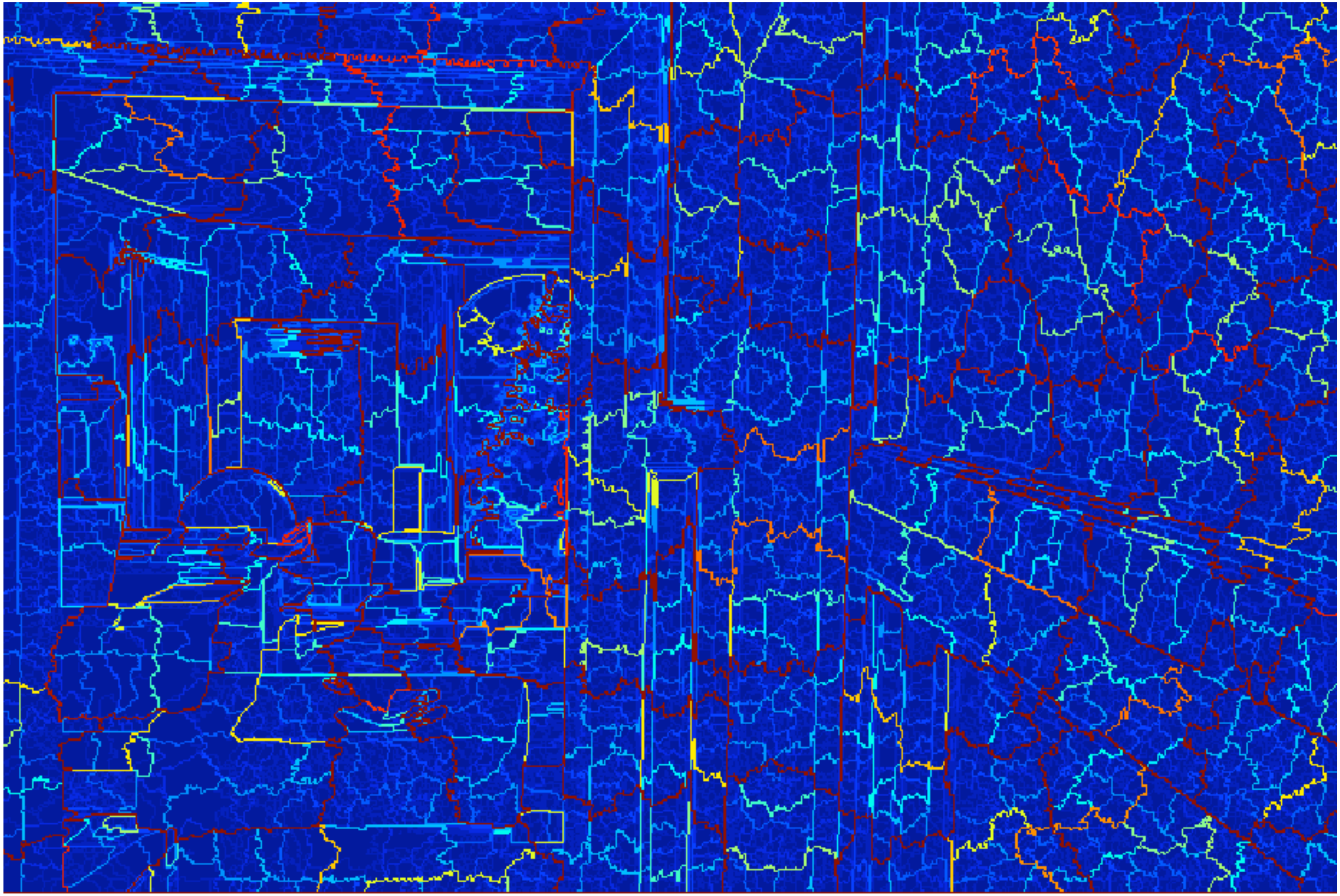
3. structural stability

- structural stability: small changes in the nuisance do not cause catastrophic (singular) perturbations in the detector
- design detectors by maximizing structural stability margins: the selection tree



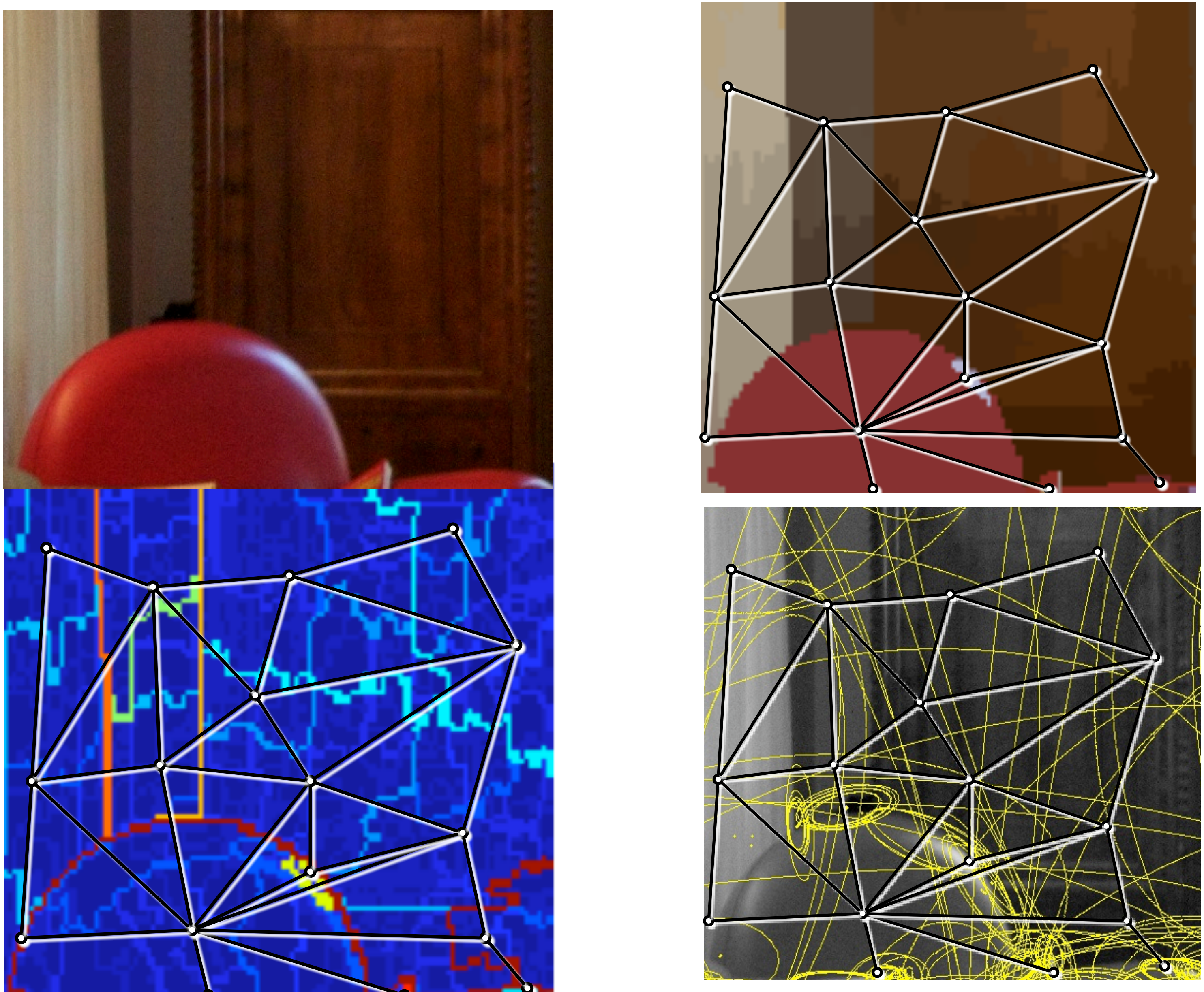
representational structures

- 2-d: regions and their texture/color description and smooth variability (ART)
- 1-d: boundaries/transitions between these descriptors
- 0-d: attributed points/junctions and their descriptors

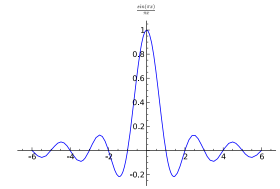


quickshift [vedaldi-soatto ’08]
(non-iterative, constant-time, returns entire segmentation tree)

representational (hyper)graph



4. proper sampling



- discretization “equivalent” to “true signal”, as good as the raw data
- topological equivalence of detector functionals between the sampled image and the “ideal image” (scene radiance)
- scene radiance unknown: under lambertian reflection and co-visibility assumption = topological equivalence across different images of the same scene
- trackability, TST/BTD/time HOG

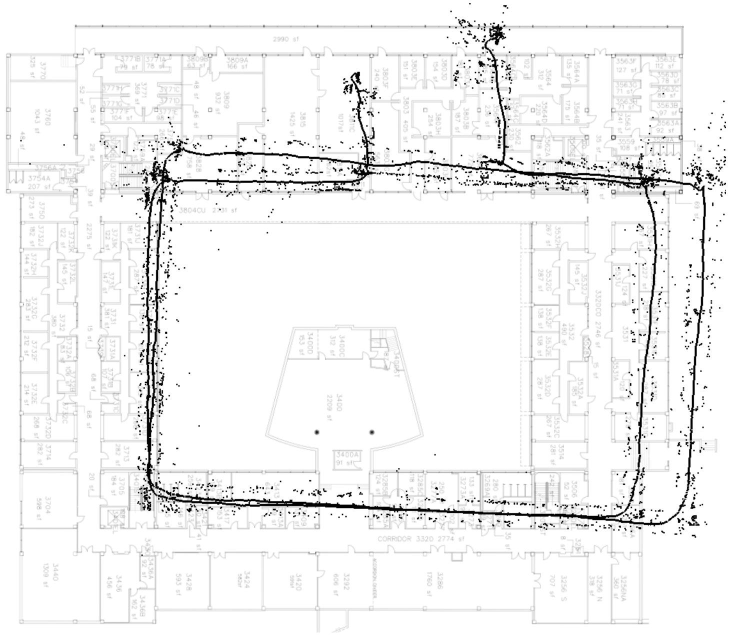
http://www.youtube.com/watch?v=cMv-McHw660

5. visual exploration

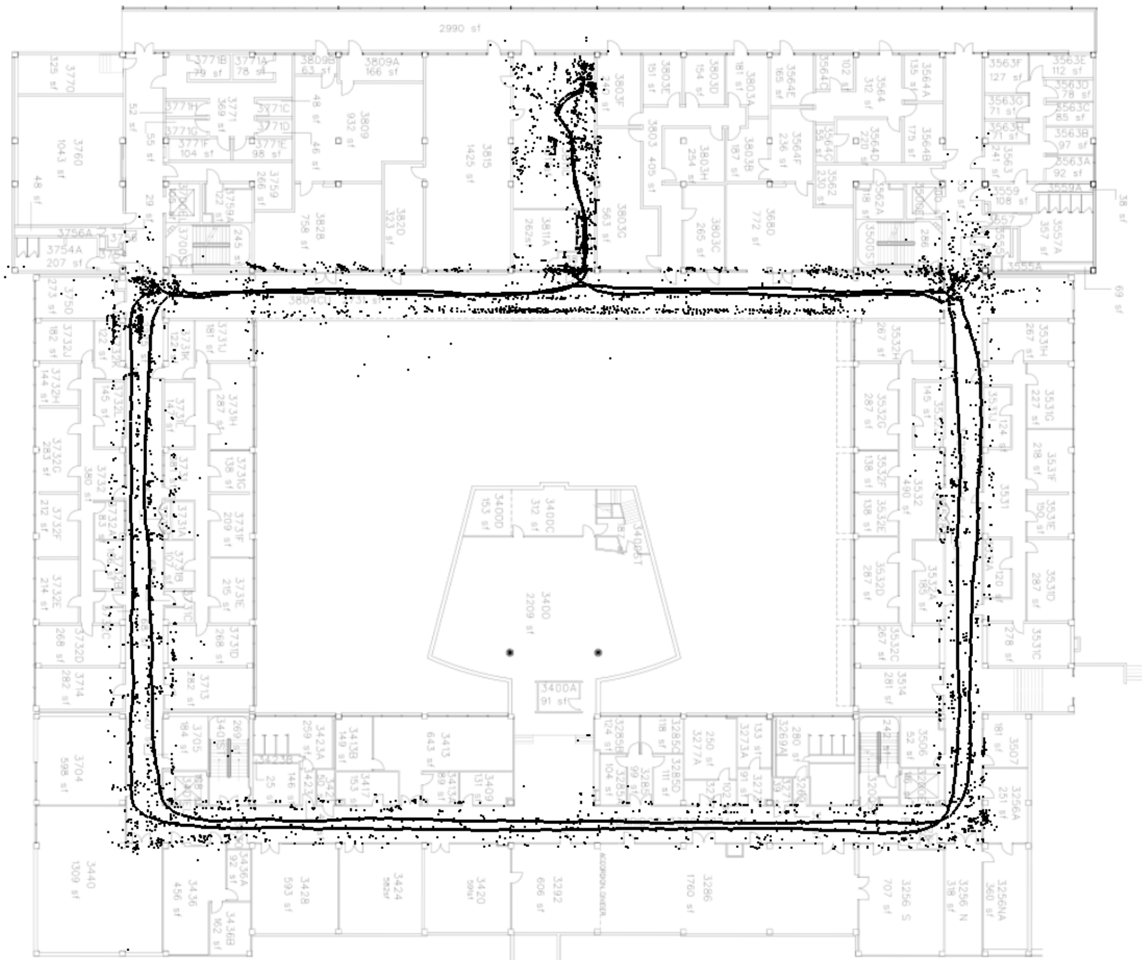
- Exploit gravity (but don't assume you know it!)
- Visual-Inertial navigation + Community Map Building

Inertial Only

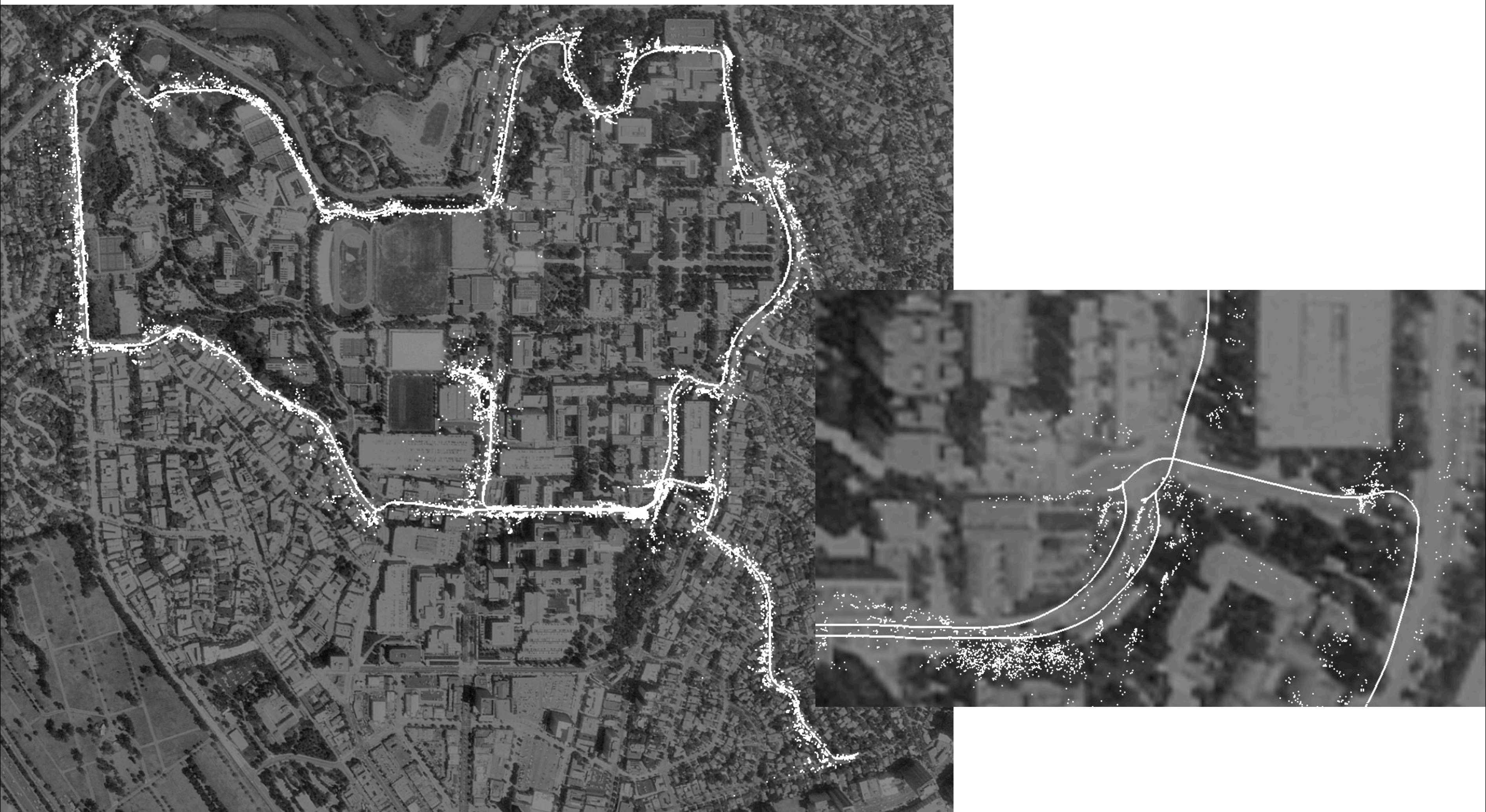
Vision Only



Drift: 0.19% (500 m)



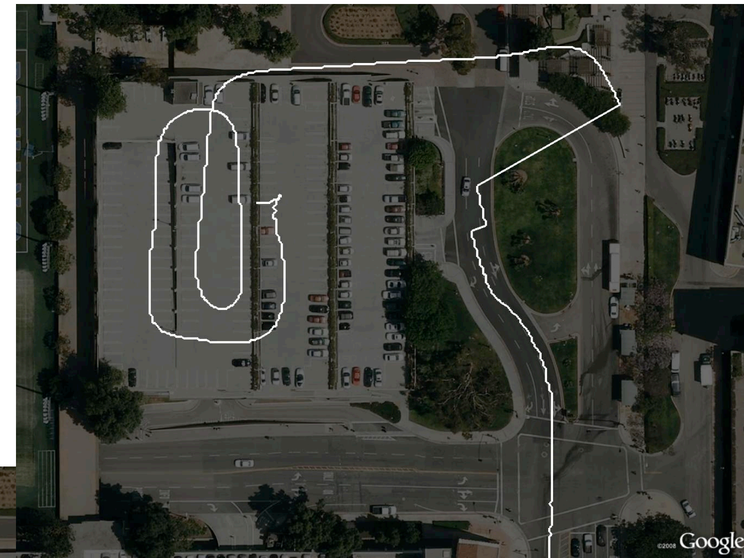
Drift: 0.27% (8 km)



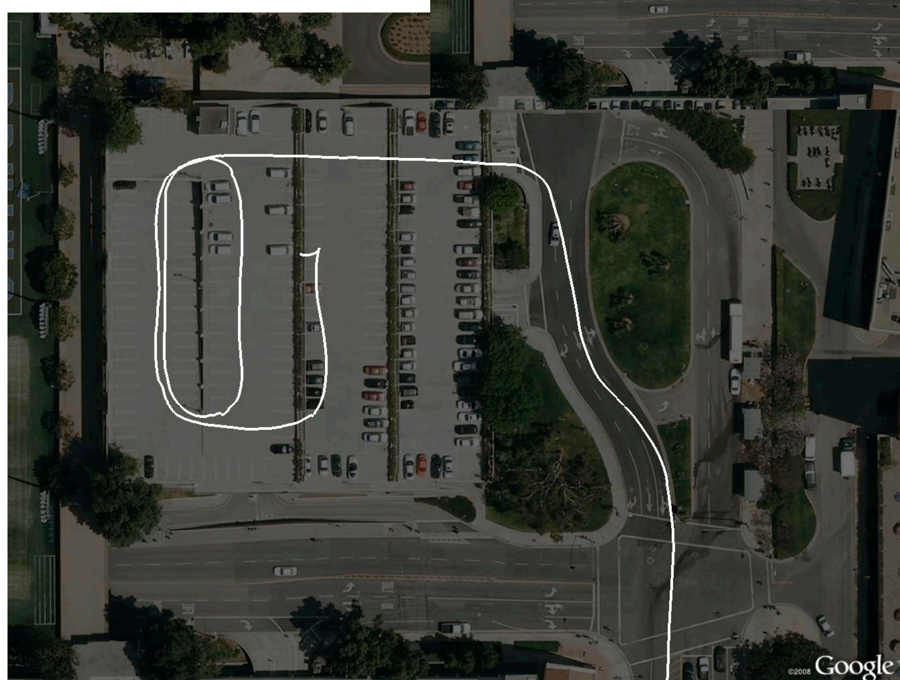
Drift: 0.5% (30km)

vs GPS+IMU

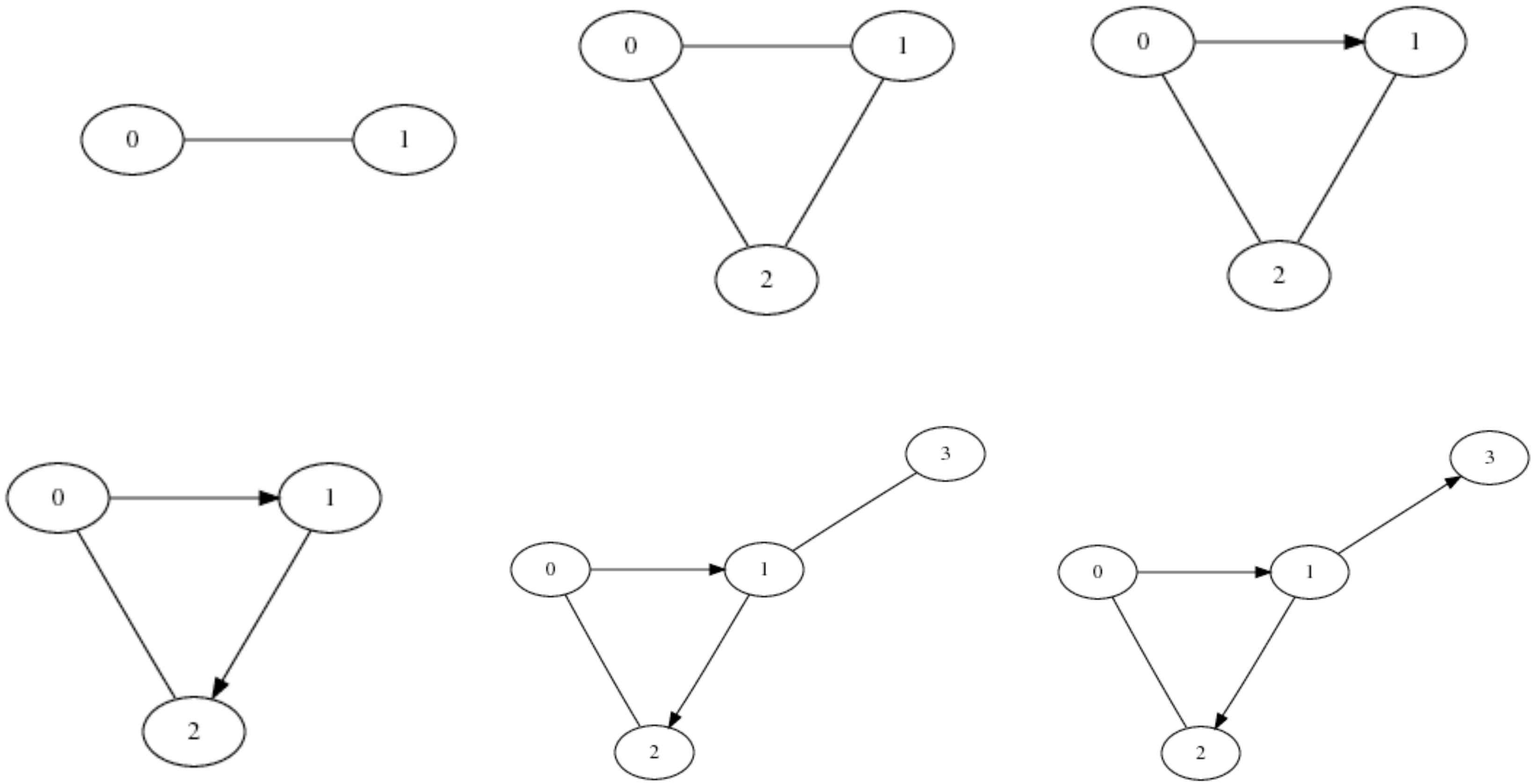
GPS+Inertial



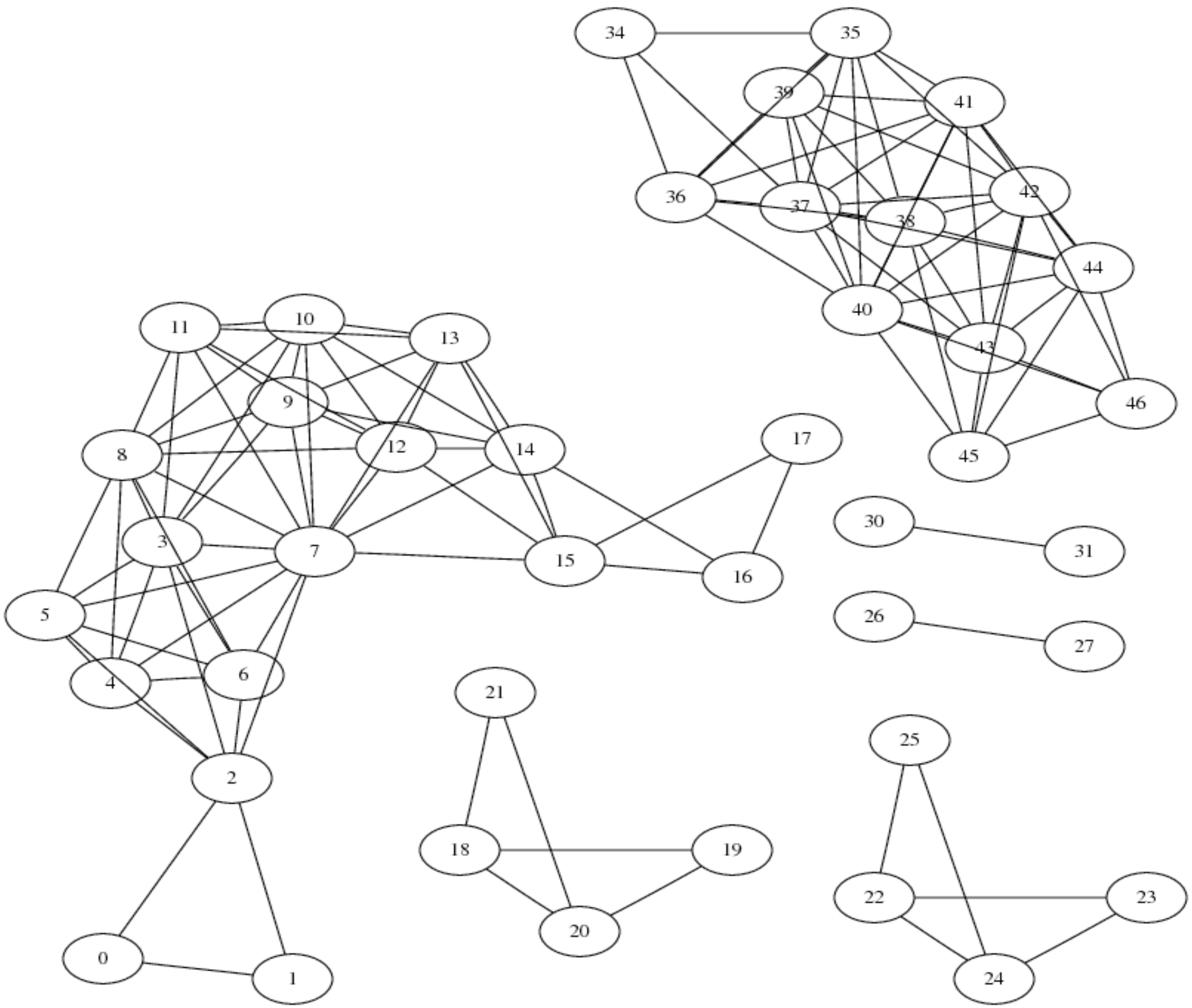
Vision+Inertial



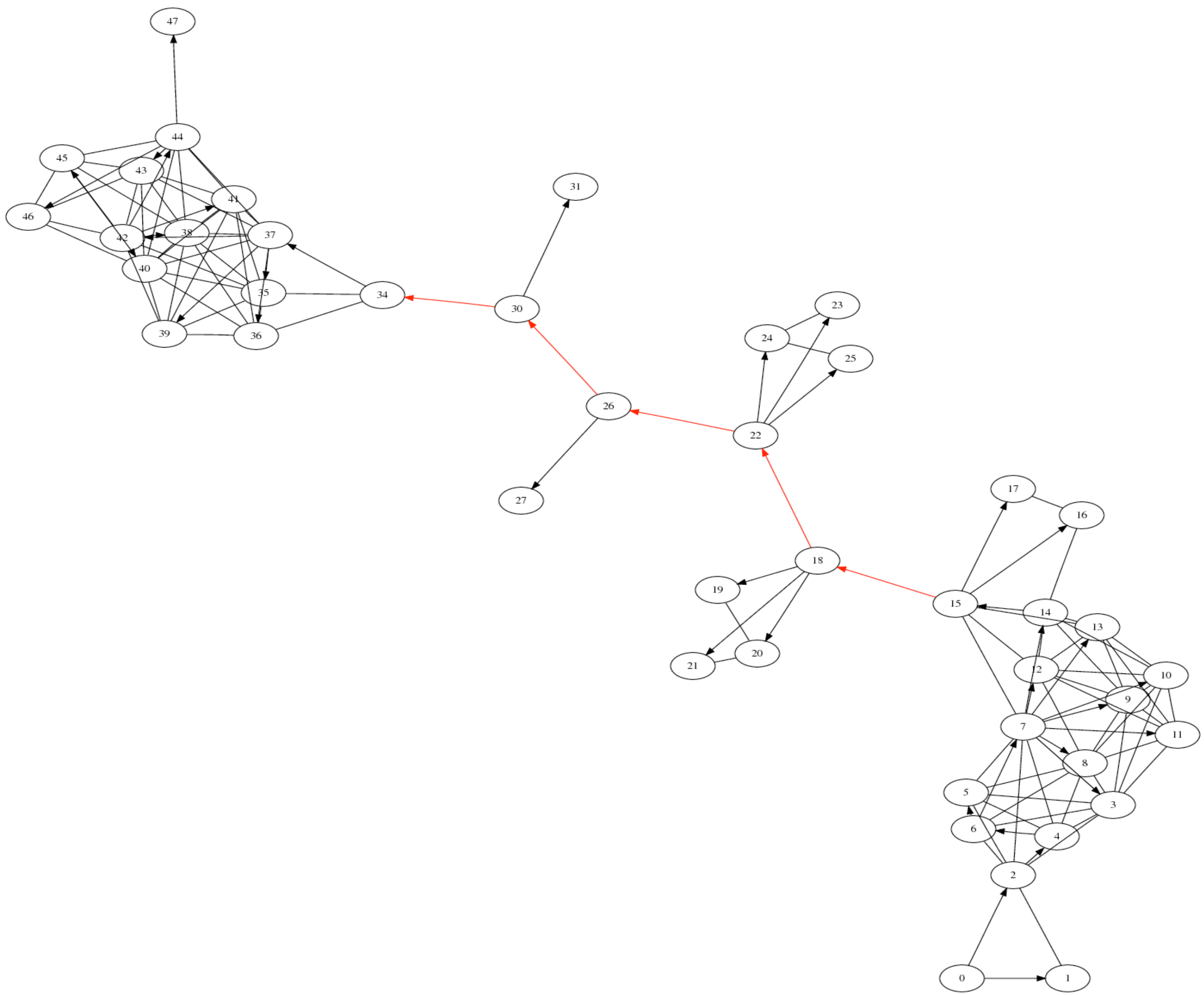
“location”, topology and co-visibility



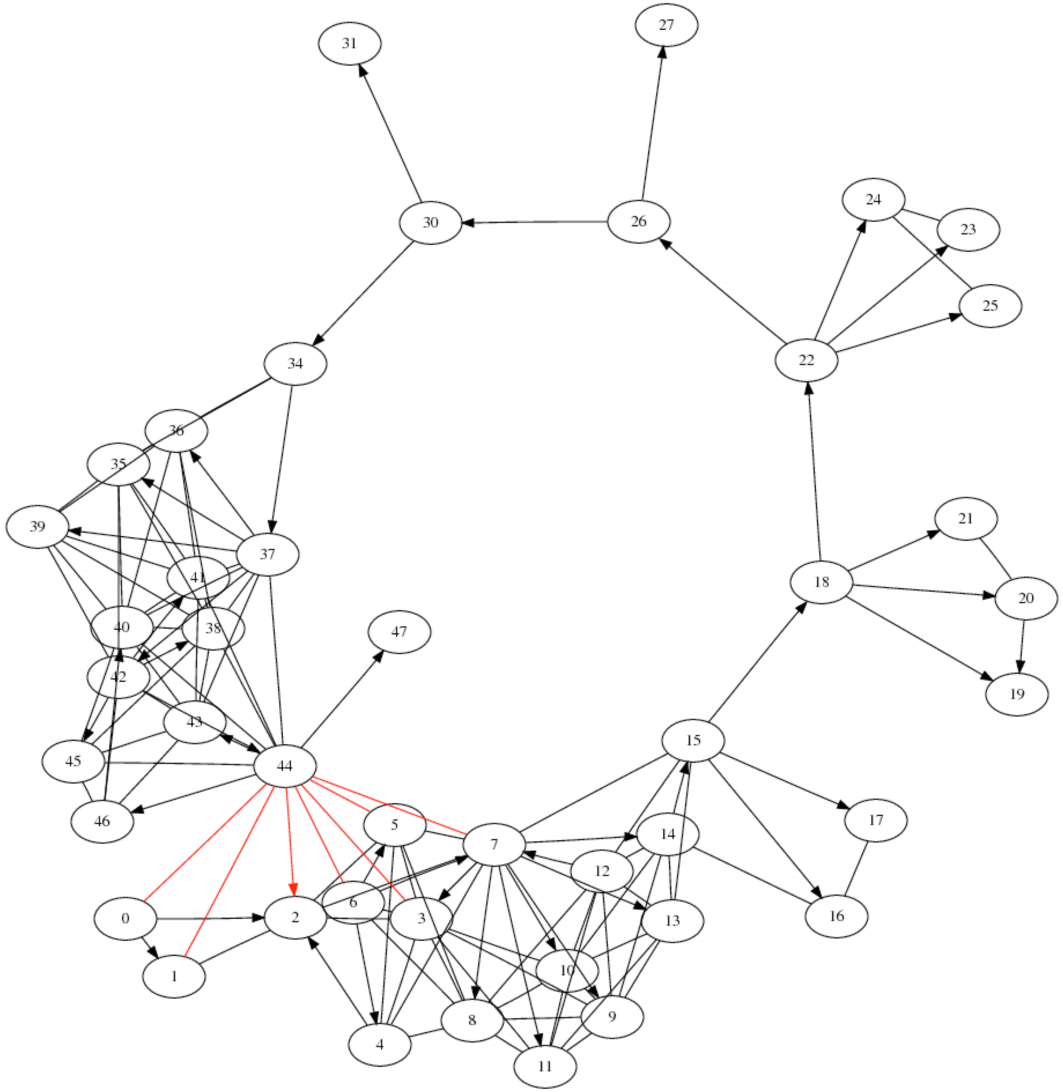
Covisibility Graph

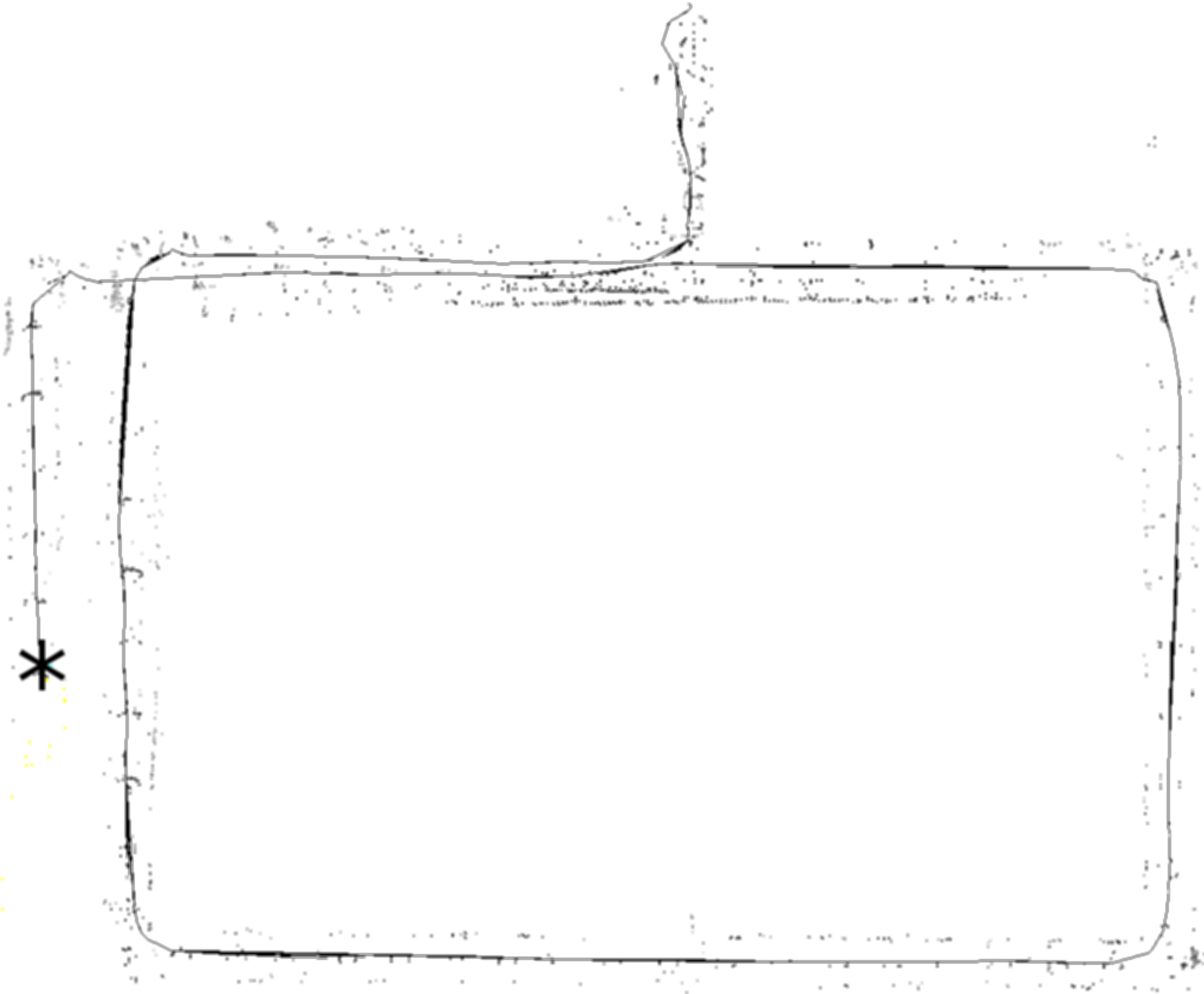


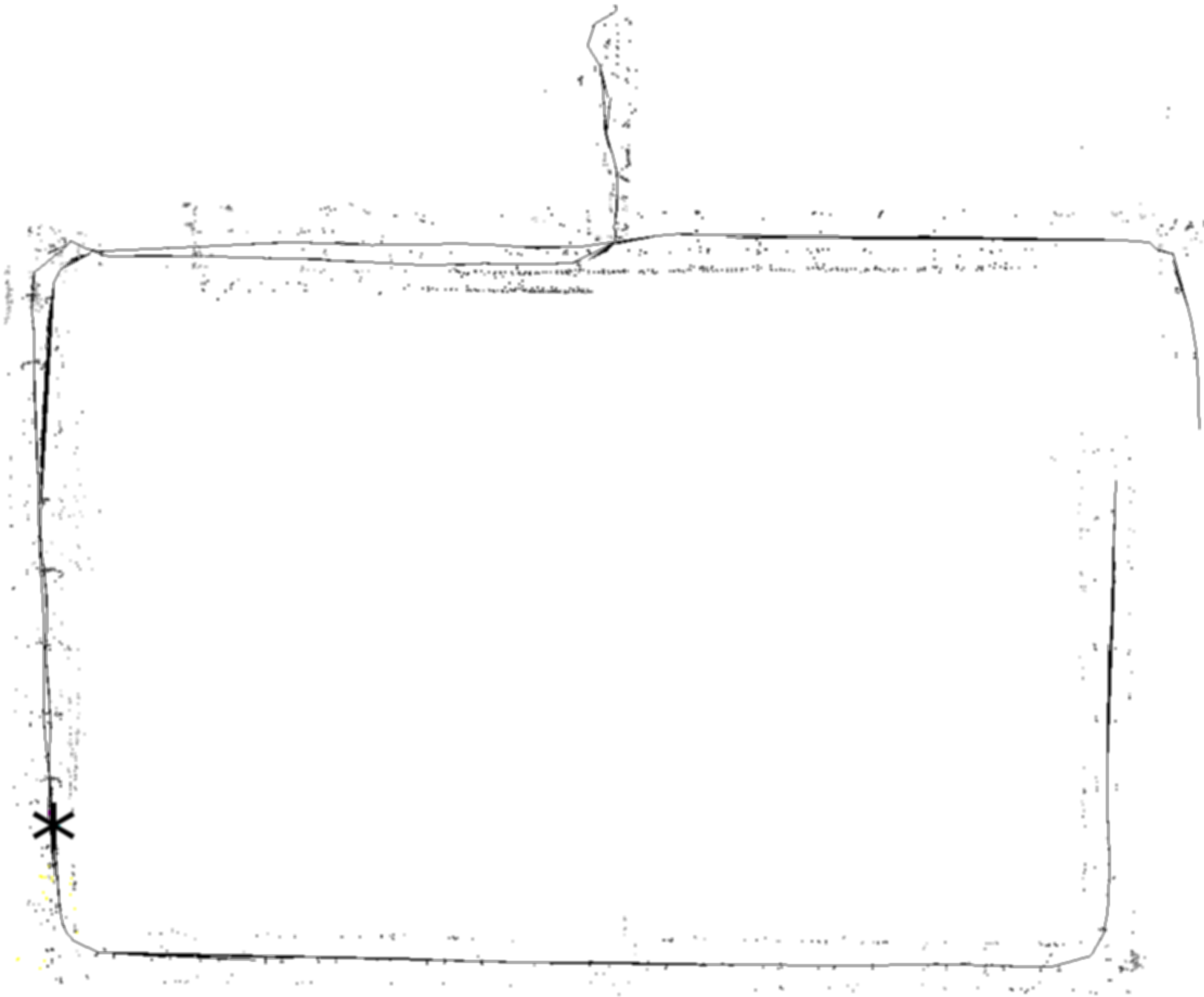
Adding Geometry

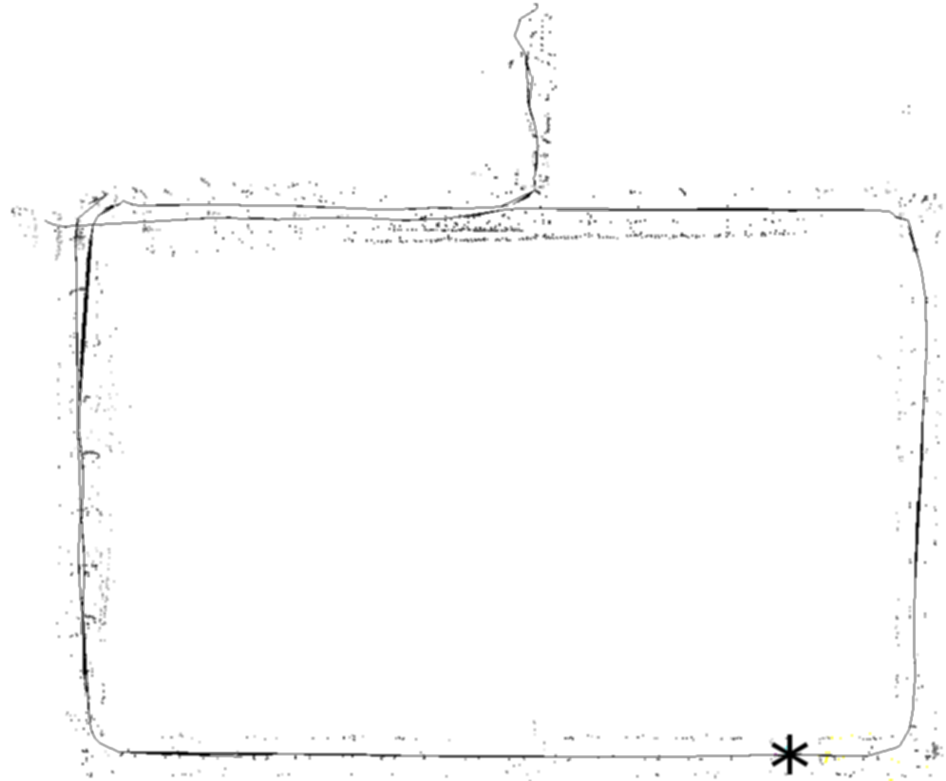


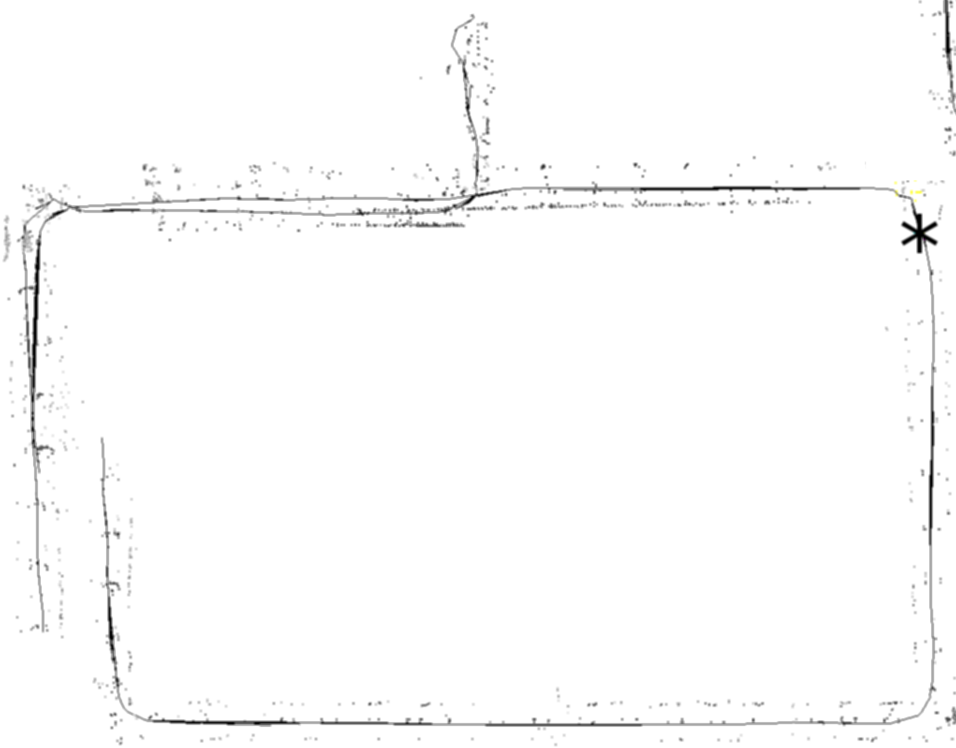
Loop Closing

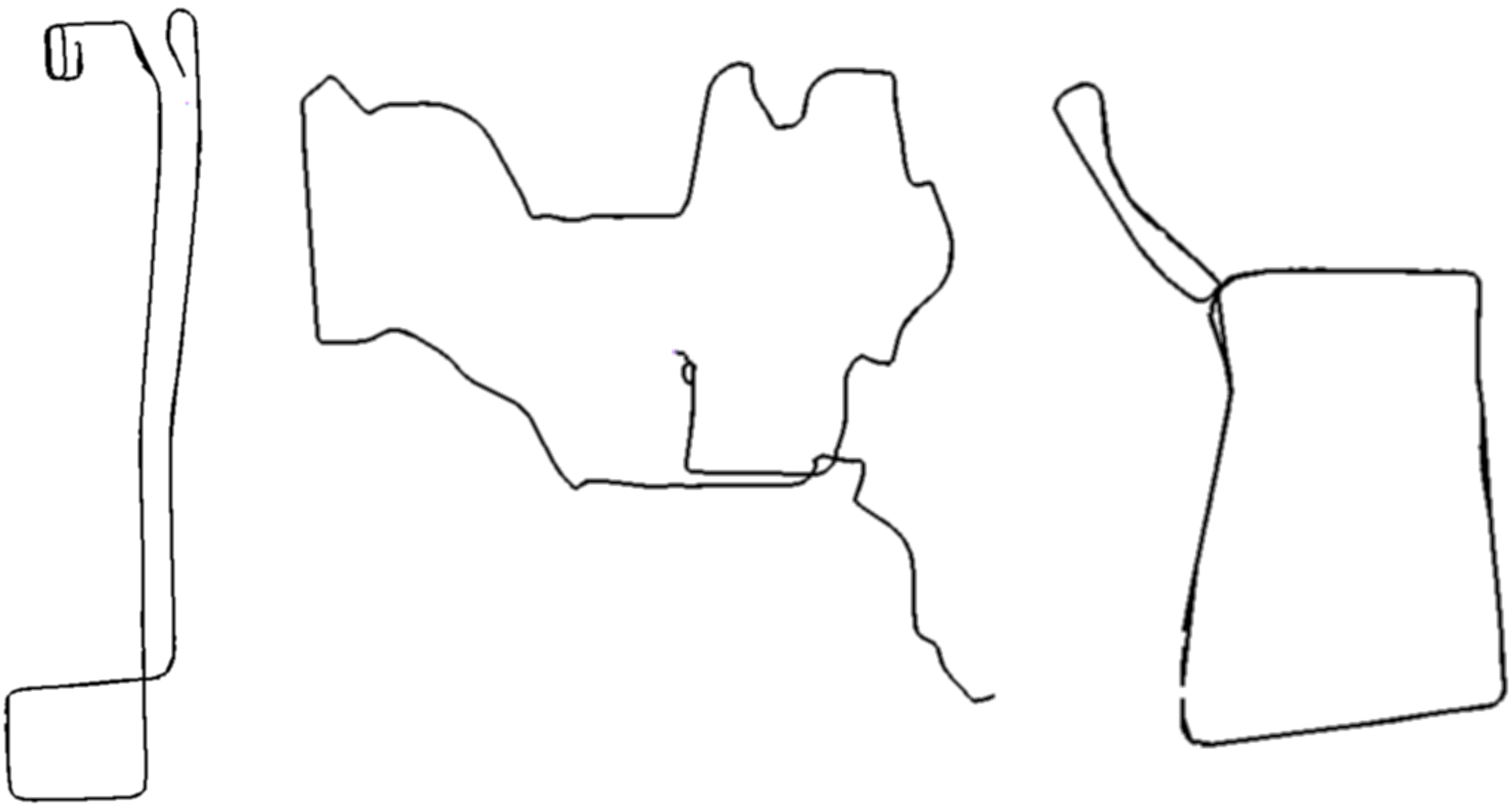


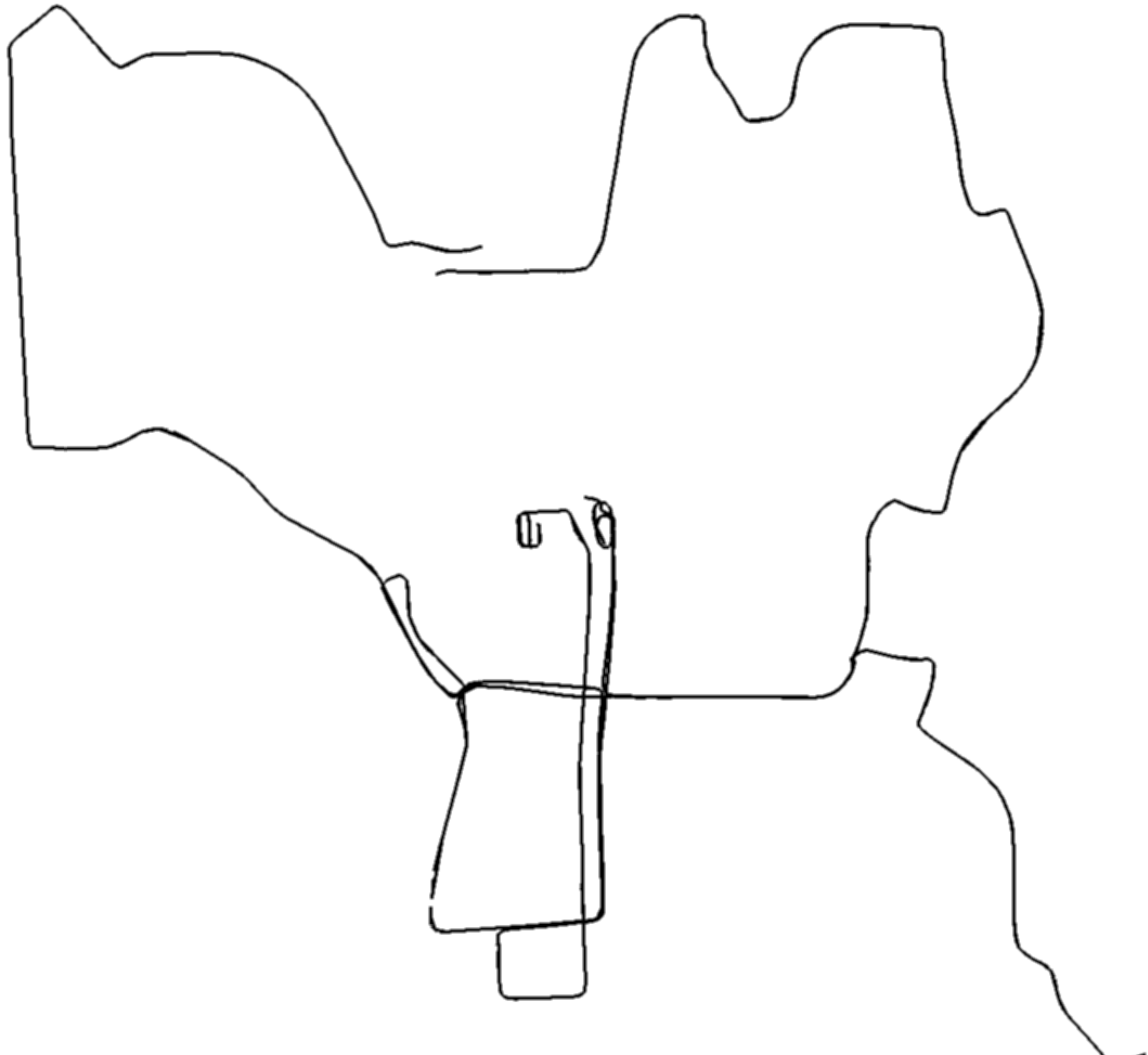












“The Black Box”

- Sensor Platform
  - Battery
  - Computation
  - D-GPS
  - Stereo, Omni Cameras
  - LADAR
  - IMU
- Portable
  - Wheels
  - Vehicle
  - Human

information pickup

- must move to "invert occlusions" (convex optimization!)

$$\Omega(t, dt) = \arg\min_{\Omega, w} \int_{D\backslash\Omega} (I(w(x,t),t) - I(x, t+dt))^2 dx + \int_D \|\nabla w\|_1 dx + \int_\Omega dx$$

- innovation and Actionable Information Increment

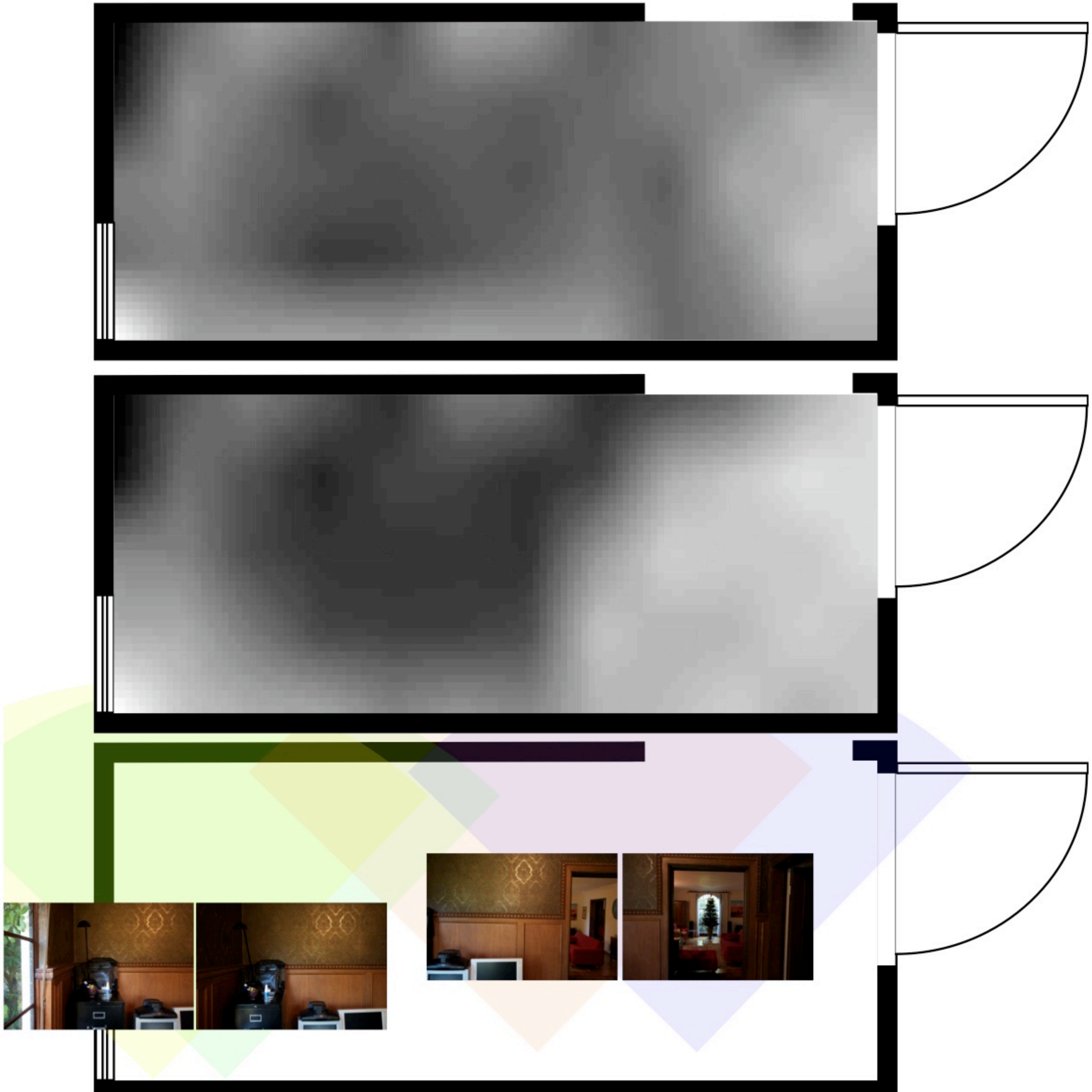
$$\boxed{\epsilon(I, t+dt) \doteq \phi^\wedge(I_{t+dt|_\Omega})} \qquad \boxed{AIN = H(\epsilon(I, t+dt)) = \mathcal{H}(I_{t+dt|_\Omega})}$$

- (memoryless) perceptual exploration:

$$\hat{u}_t = \arg\max_u AIN(I, t; u)$$

building a representation: perceptual explorers

$$\begin{cases} \hat{\xi}_{t+dt} = \hat{\xi}_t \oplus \epsilon(I_{t+dt}, t+dt; \hat{u}_t, \hat{\xi}_t) \\ \hat{u}_t = \arg\max_u H(\epsilon(I_t, t; u, \hat{\xi}_t)) \\ \hat{\xi}_0 = h^{-1}(I_0) \end{cases}$$

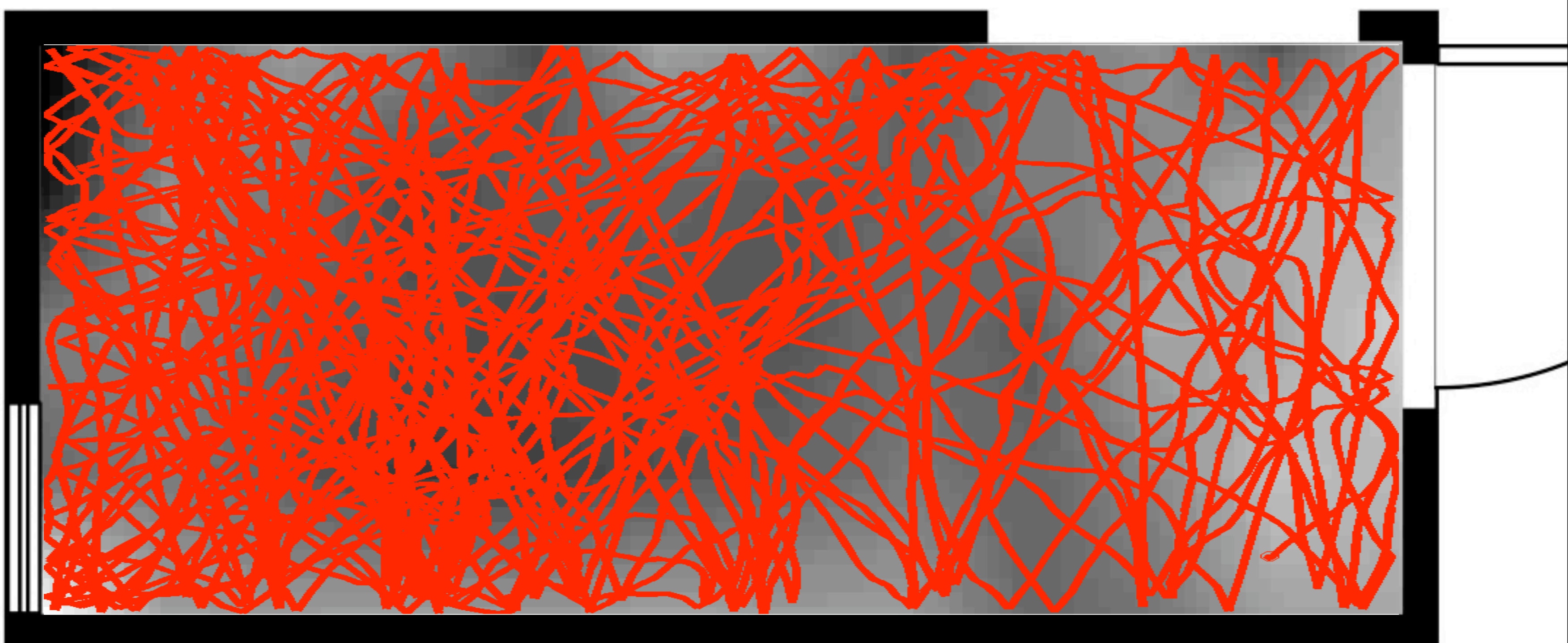


brownian explorer

$$\begin{cases} dg = \widehat{u} g dt \\ du = dW \quad \text{a wiener process} \end{cases}$$

brownian explorer

$$\begin{cases} dg = \widehat{u}g\,dt \\ du = dW \quad \text{a wiener process} \end{cases}$$

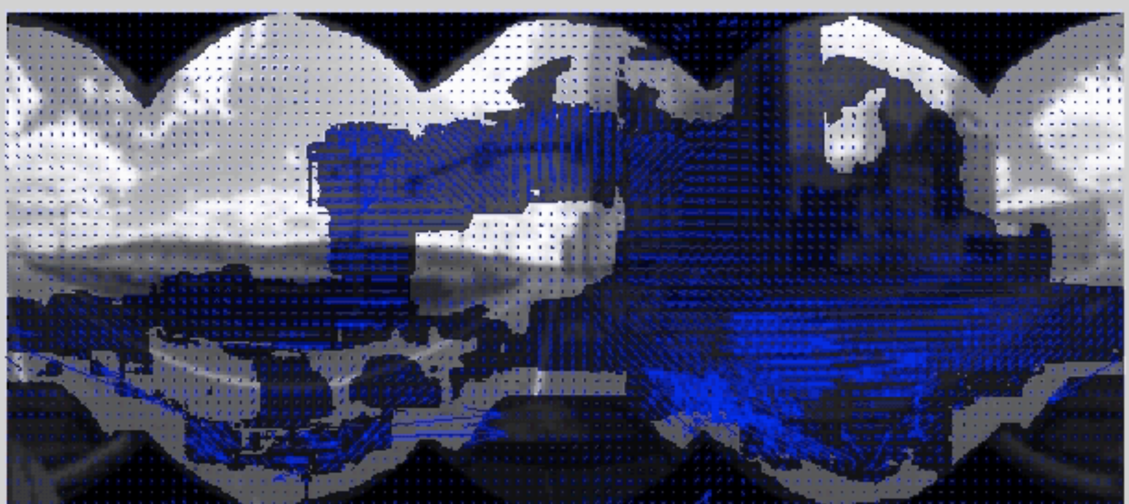


reflections/shadow-paths

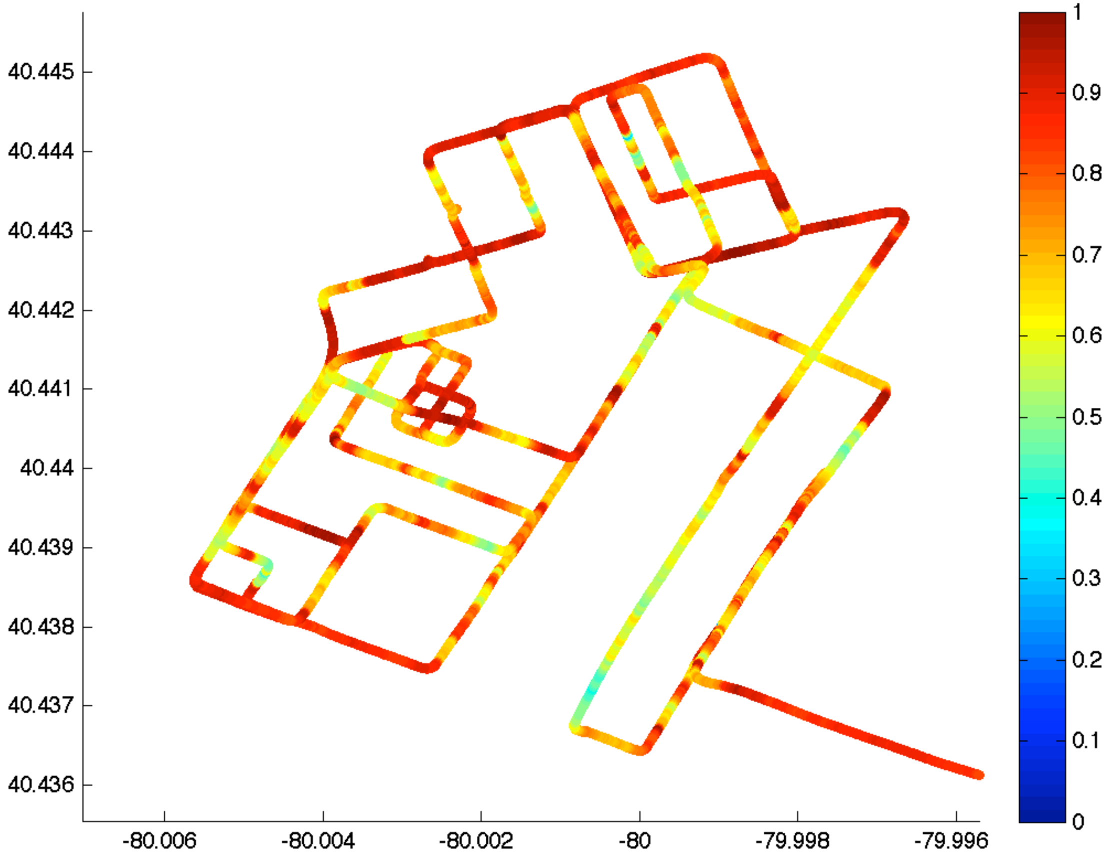
shannonian explorer

gibsonian explorer

googleonian explorer

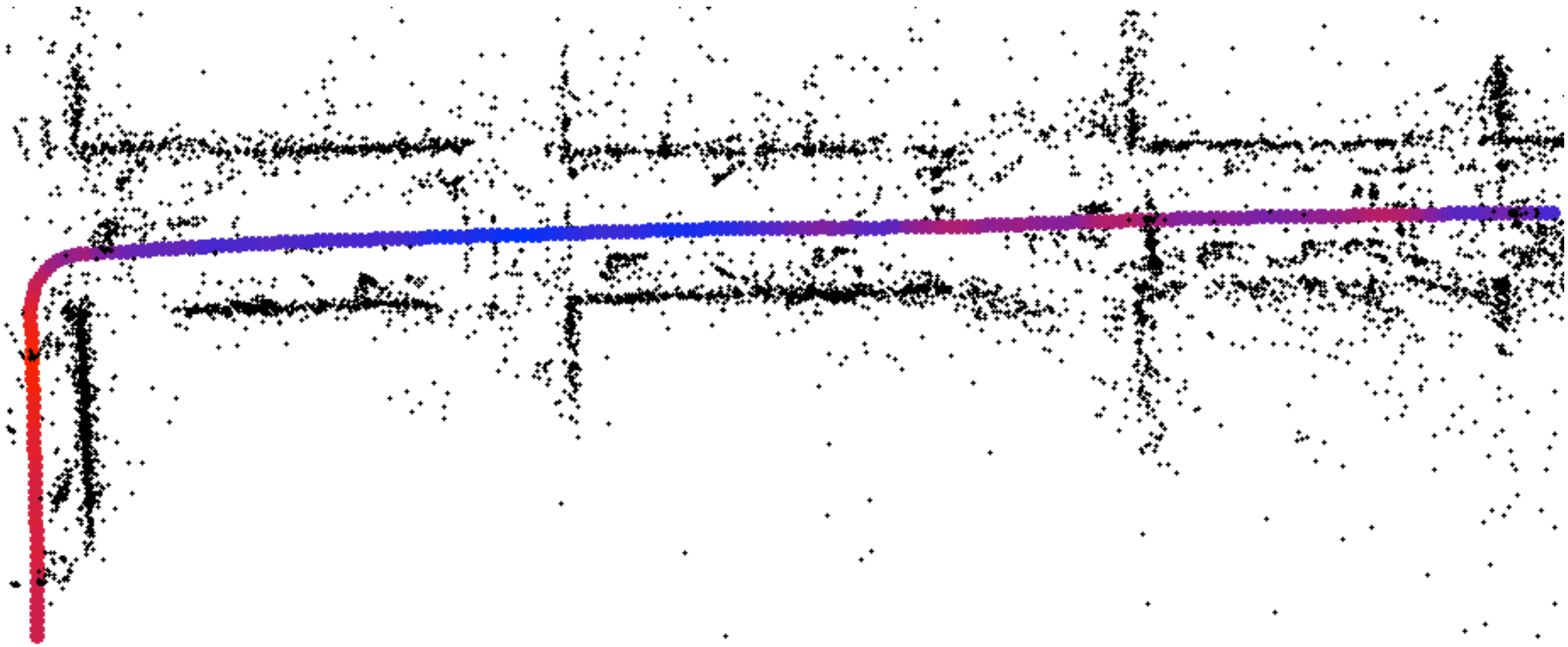


google street view dataset

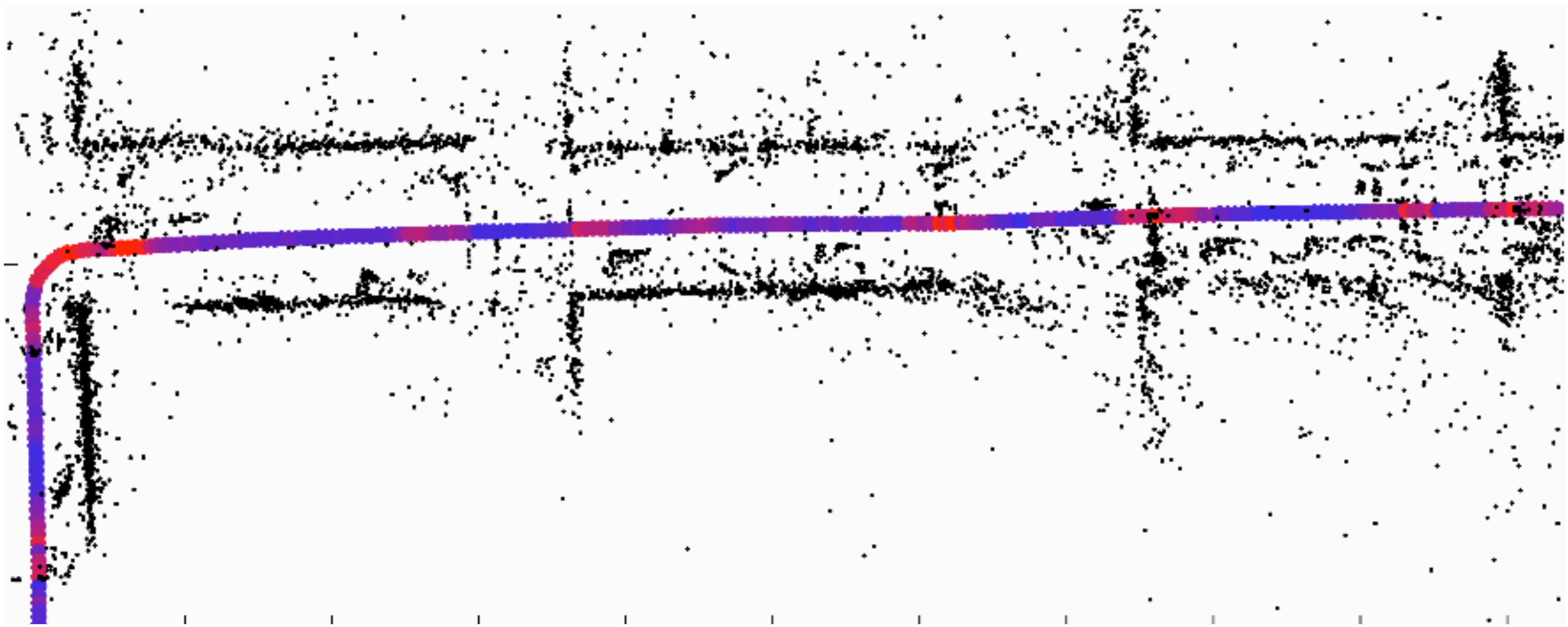


Courtesy of Taehee Lee

shannon in google's car seat



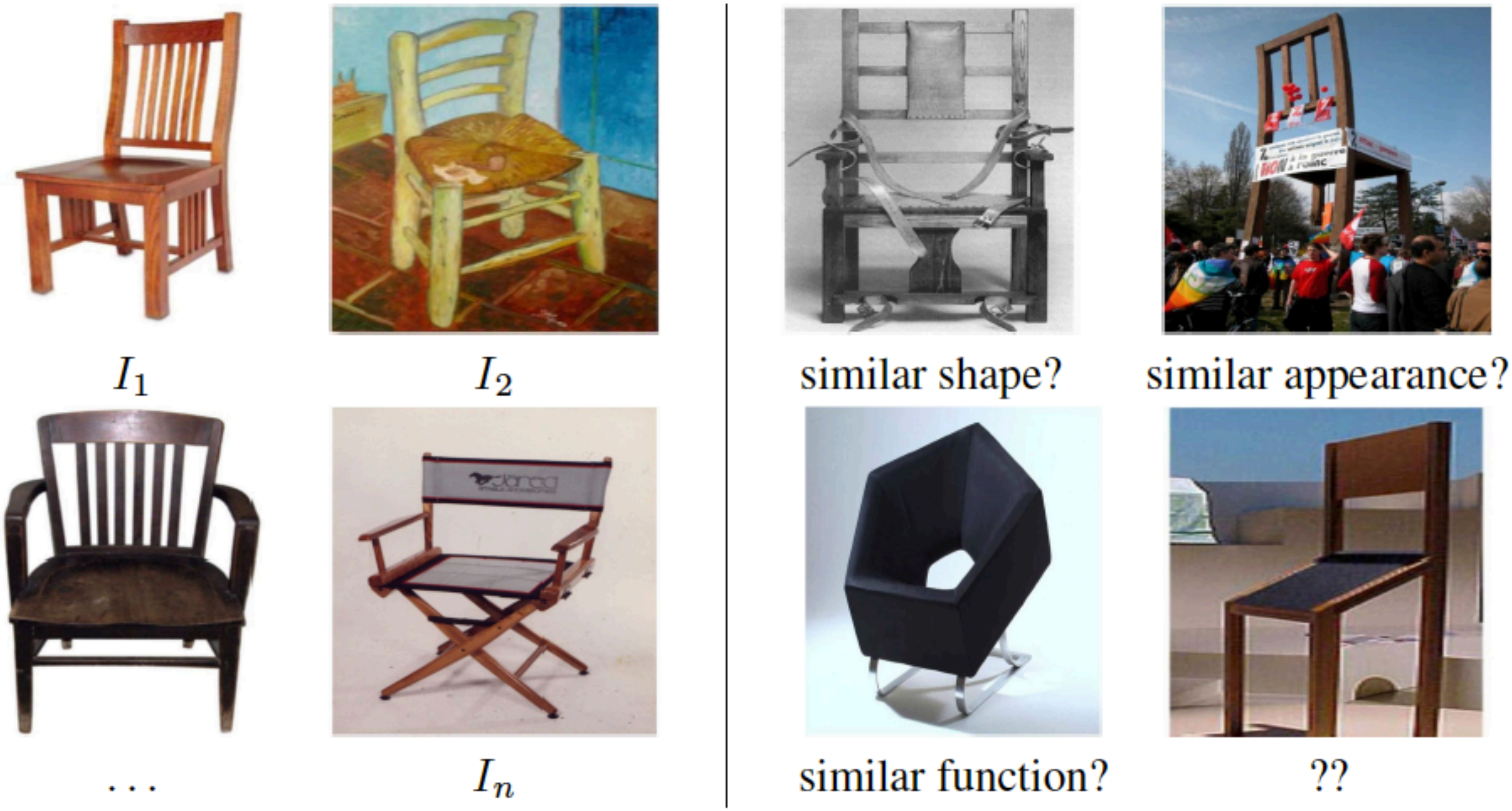
gibson in google's car seat



accommodation

part III
asides

learning priors and categories
texture
actions, events


I_1
I_2
similar shape?
similar appearance?
...
I_n
similar function?
??

learning priors

$$\hat{\xi}, \hat{g}_k, \hat{\nu}_k = \arg \min_{\xi, g_k, \nu_k} \|I_k - h(g_k \xi, \nu_k)\|_*$$

$$dP(\nu) = \sum_i \kappa_\nu(\nu - \hat{\nu}_i) d\mu(\nu); \; dP(g) = \sum_i \kappa_g(g - \hat{g}_i) d\mu(g)$$

category

$$dQ_c(\xi) = \sum_{i=1}^{M} \kappa_\xi(\xi - \hat{\xi}_i) d\mu(\xi)$$

part IV
time

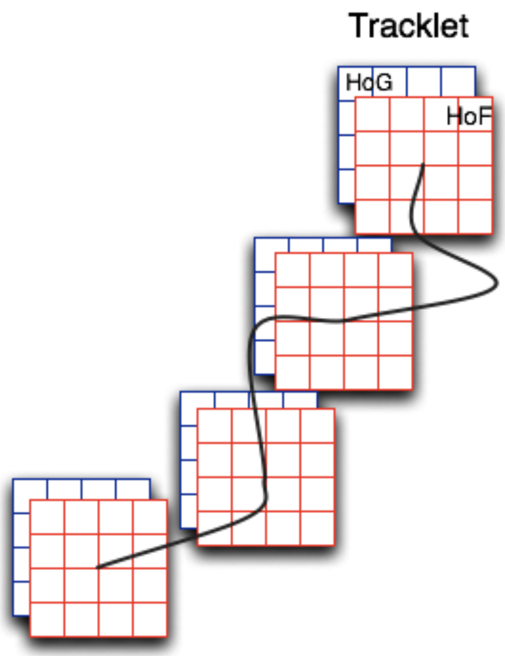
marginalizing time

- Tracklet Descriptor
- Time-warping under dynamic constraints

- # Tracklet Descriptor:

$$\pi_i(t|I) \doteq \{HoG_i(t), HoF_i(t)\}_{t=\tau_i}^{T_i}$$

The normalized histograms are concatenated and stacked sequentially building a time series $X \in \mathbb{R}^{256 \times N}$ where N is the temporal range of the trajectory.

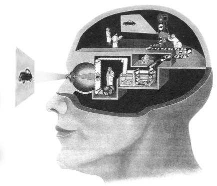


HoG (blue) and HoF (red) along a Trajectory

Examples of Actions in HOHA dataset

The color of the extracted tracks indicates their label based on the tracklet descriptor dictionary

epistemological fallout



- signal-to-symbol barrier: *is data analysis* (breaking down the data into pieces) *necessary for cognition?* an "analog car mechanic"?
- rao & blackwell: *no advantage in internal representation* (complexity calls for compression, not analysis)
- I. viewpoint/illumination invariants exist, they are "*discrete*" *sufficient statistics;* no harm done in discrete internal representation; benefits at run-time; still no analysis (locality)
- II. occlusions and mobility are key
- can they be "learned away"?

$$R(u|I) = R(u|\phi(I))$$

precursors

- alan turing: symbolization by *morphogenesis* (reaction-diffusion) specific to biological systems
- david marr: *"our view is that vision goes symbolic almost immediately, at the level of zero crossings, and the beauty of this is that the transition ... is probably accomplished without loss of information"* (without underlying task, remains self-referential)
- james gibson: missed discriminative component of the problem (sufficiency)
- norbert wiener: *"first moment* [integral wrt a group measure] *is invariant statistic ('gestalt')"*