mission-focused
- actions/reactions based on -
- system integration of information derived from -
complex real-world data

stefano soatto
ucla

machine reasoning workshop
ipam, october 20, 2010

reasoning vs. complex real-world data

- "reasoning" presumes "discrete entities" (symbols, concepts, objects, segments, ...) - acted upon by logic & probabilistic inference
- object of inference in "complex real world data" is not meaningfully discretizable at the outset (scale-invariant measures, occlusion)
- now: features/pre-processing (mostly not task-specific)
- task provides falsifiability mechanism (signal-symbol barrier)

- need a new "information theory" in support of decision and control tasks (as opposed to transmission and storage of data)
- lossless symbolization is possible (task-specific representations)
- requires exercising control on the sensing process
- enables controlled recognition bounds
- "proper sampling" conditions

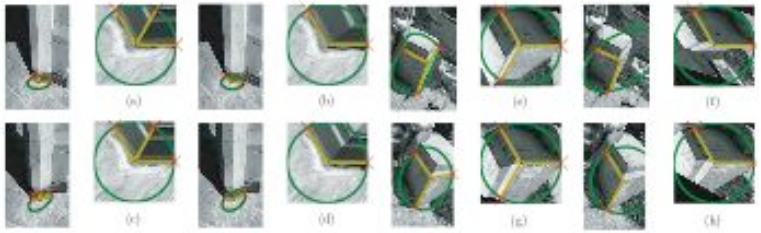
gibson's information

- 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"*

"the set of images modulo viewpoint and contrast changes"

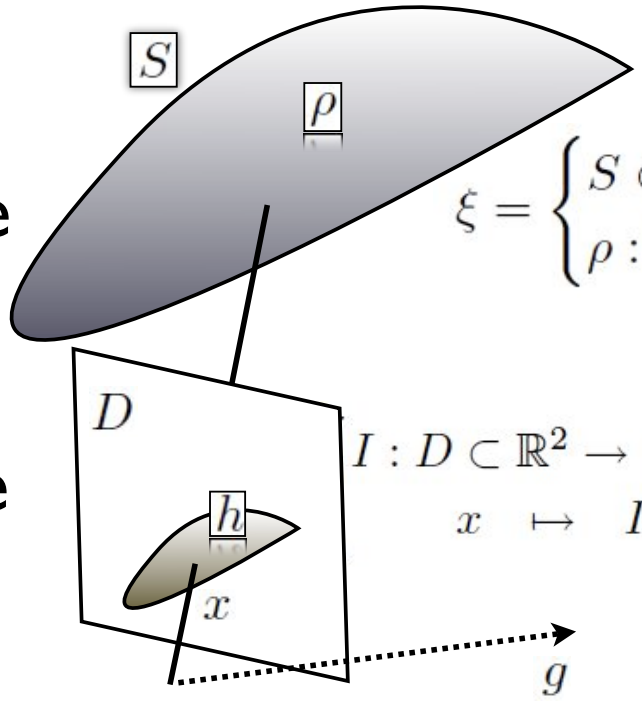
[sundaramoorthi-petersen-varadarajan-soatto '09]



- 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)

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

$$I : D \subset \mathbb{R}^2 \to \mathbb{R}^k_+ \qquad x \mapsto I(x)$$

$\mathbb{L}^2$ 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

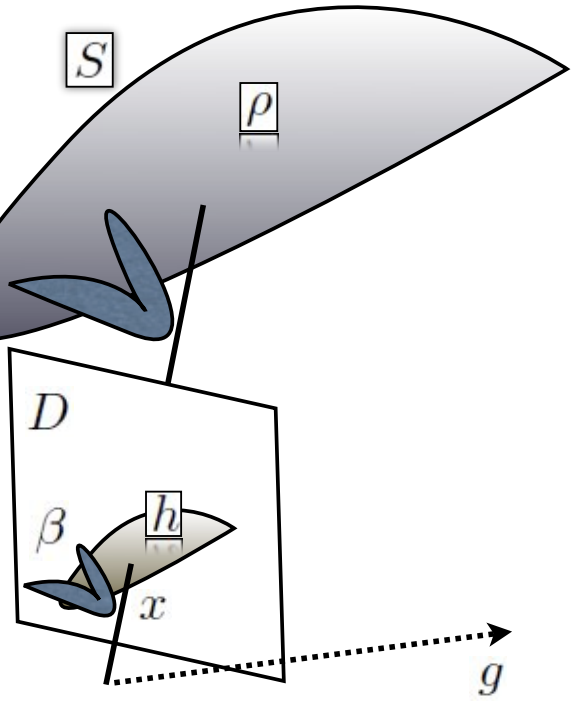
$$\begin{cases} I(x,t) = h(t) \circ \rho(p) + n(x,t) \\ x = \pi(g(t)p) \end{cases}$$

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

some notation

scene

image



occlusions

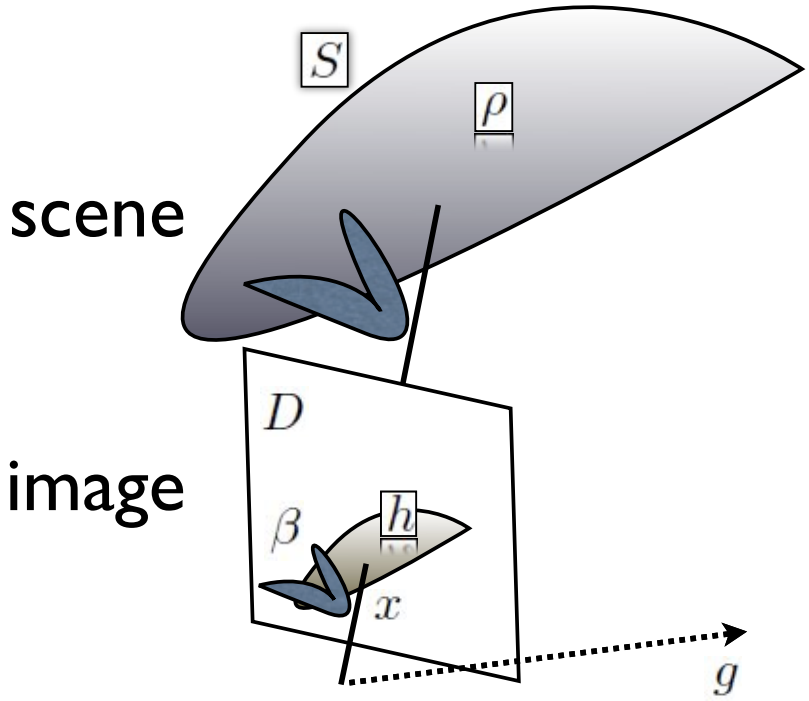
$$I(x) = \begin{cases} f(\rho, S; g, h, n) & x \in D \backslash \Omega \\ \beta(x) & x \in \Omega \end{cases}$$

lambert-ambient

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

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

some notation



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

image formation model
(formal notation)

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

some definitions

feature $\phi : \{I(x), x \in D\} \rightarrow \mathbb{R}^K$

$I \mapsto \phi(I)$

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

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

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

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 $\phi^{\vee}_{\xi}(I)$

representation

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

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

$$\{I\} \in \{\ h(g\hat{\xi}, \nu), \quad g \in G,\ \nu \in \mathcal{V}\}$$

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

information gap

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

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

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

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

- actionable information gap (AIG)

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

invertible nuisances

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

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

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 *"ecological integration"* ("information pickup")
- CAN compute representations that are sparse and yet lossless. how?

how to build (lossless) representations?

1. canonizability (optimality by design)
2. commutativity (no SIFT)
3. structural stability (no BIBO)
4. proper sampling (no nyquist)
5. exploration (gibson)

how to build representations? feature optimality by design

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

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

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)

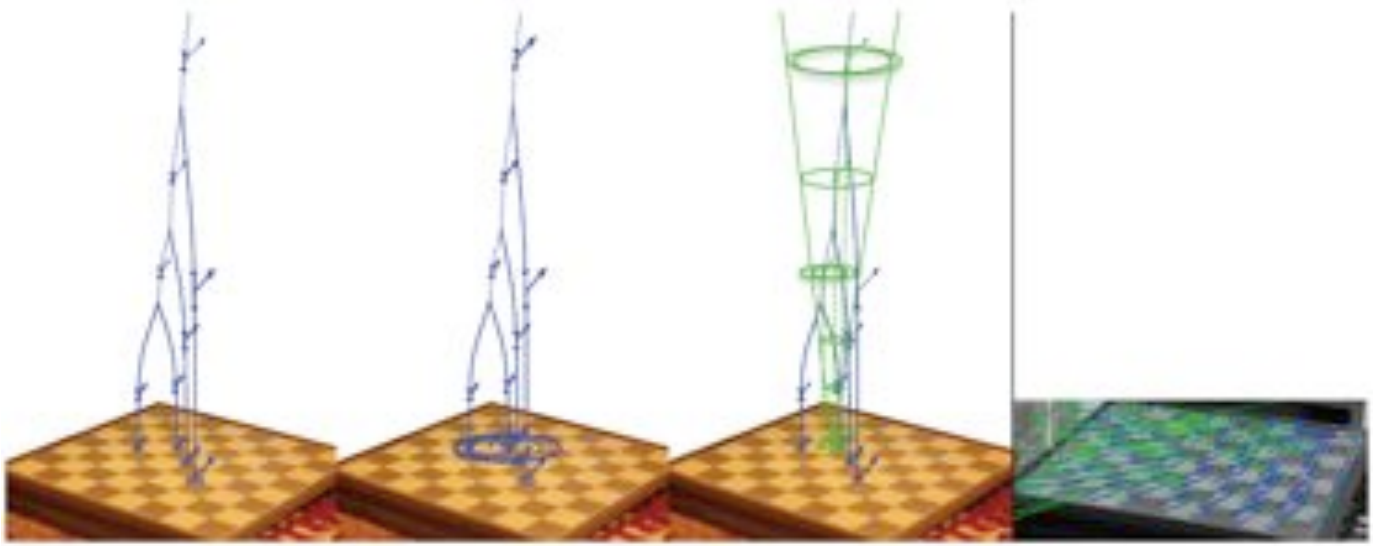
- (Thm 5: an image region is a texture if and only if it is not canonizable)

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

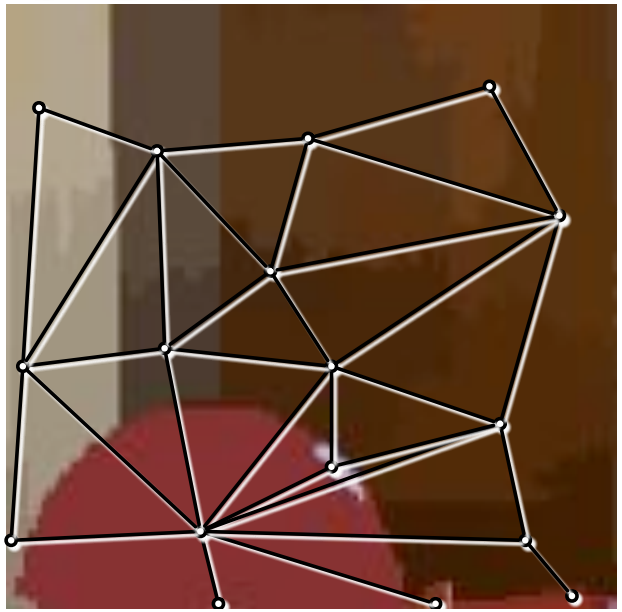
- structural stability: small changes in the nuisance do not cause catastrophic (singular) perturbations in the detector
- design detectors by maximizing structural stability margins: the selection tree

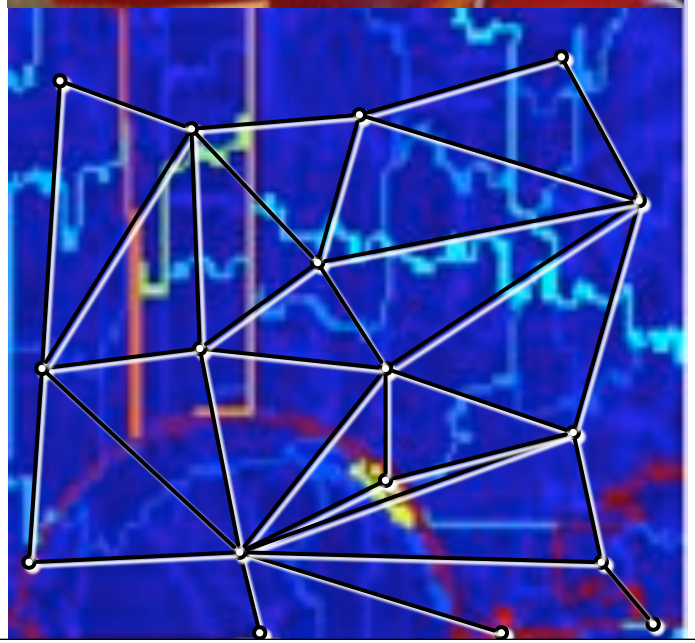


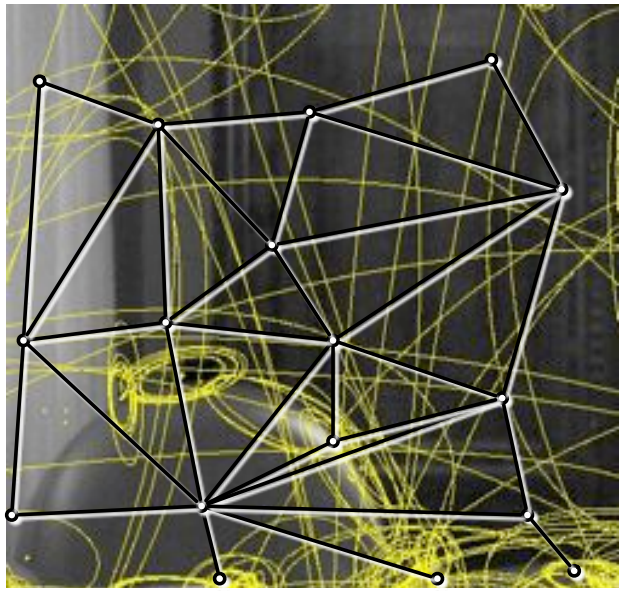
representational structures

- 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

representational (hyper)graph







4. proper sampling

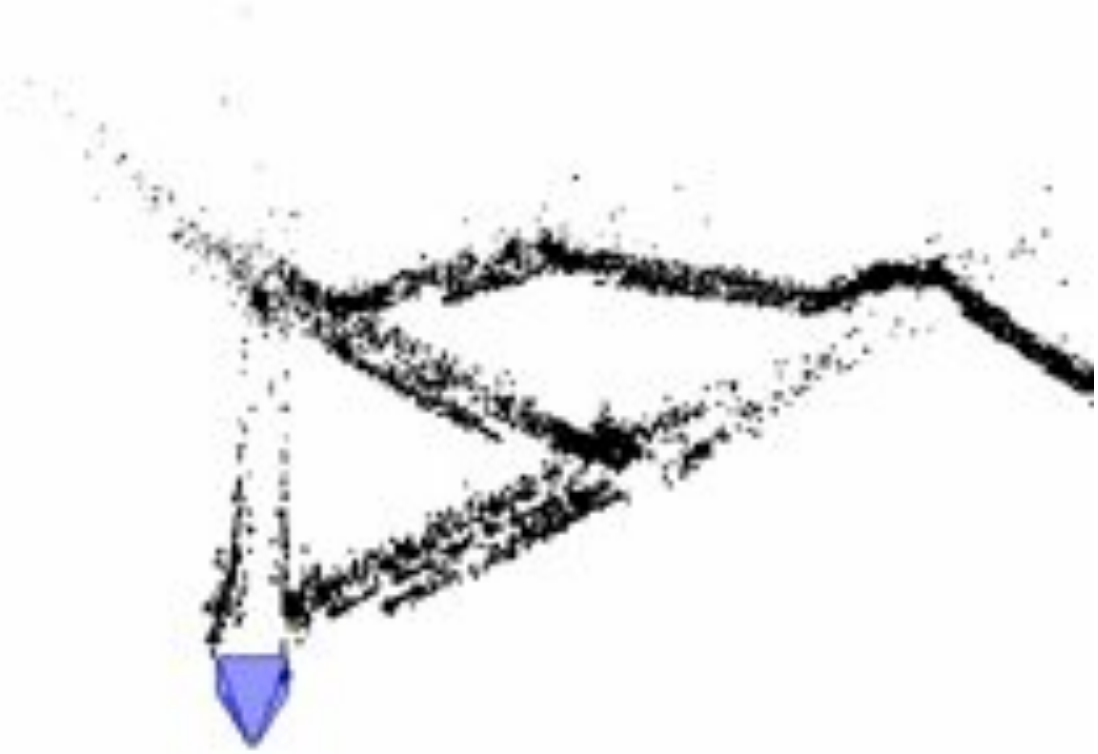
- topological equivalence of detector functionals between the sampled image and the "ideal image" (scene radiance)
- scene radiance unknown: under lambertian reflection and co-visibility assumption = topological equivalence across different images of the same scene
- trackability, TST/BTD/time HOG

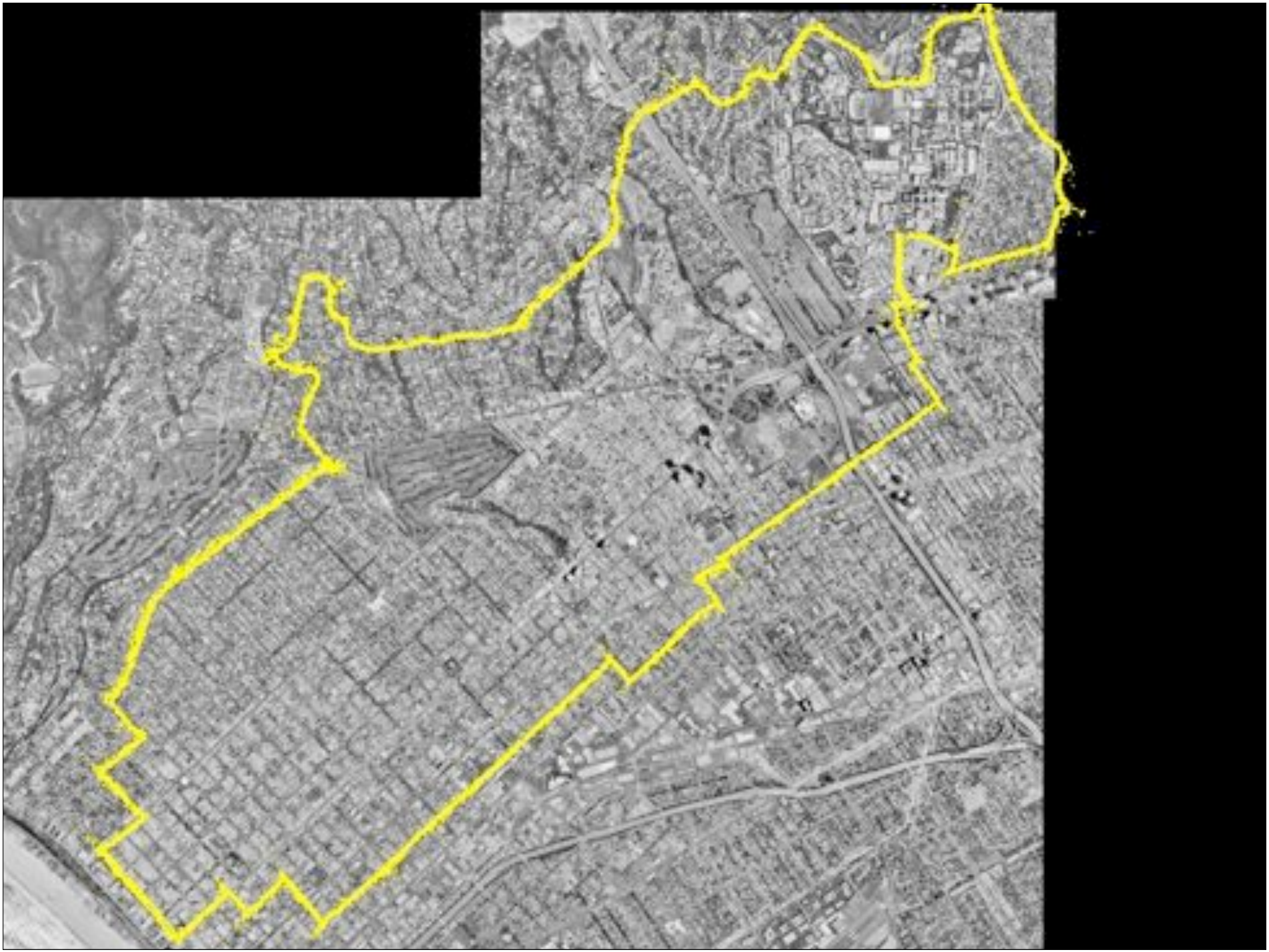
iphone demo

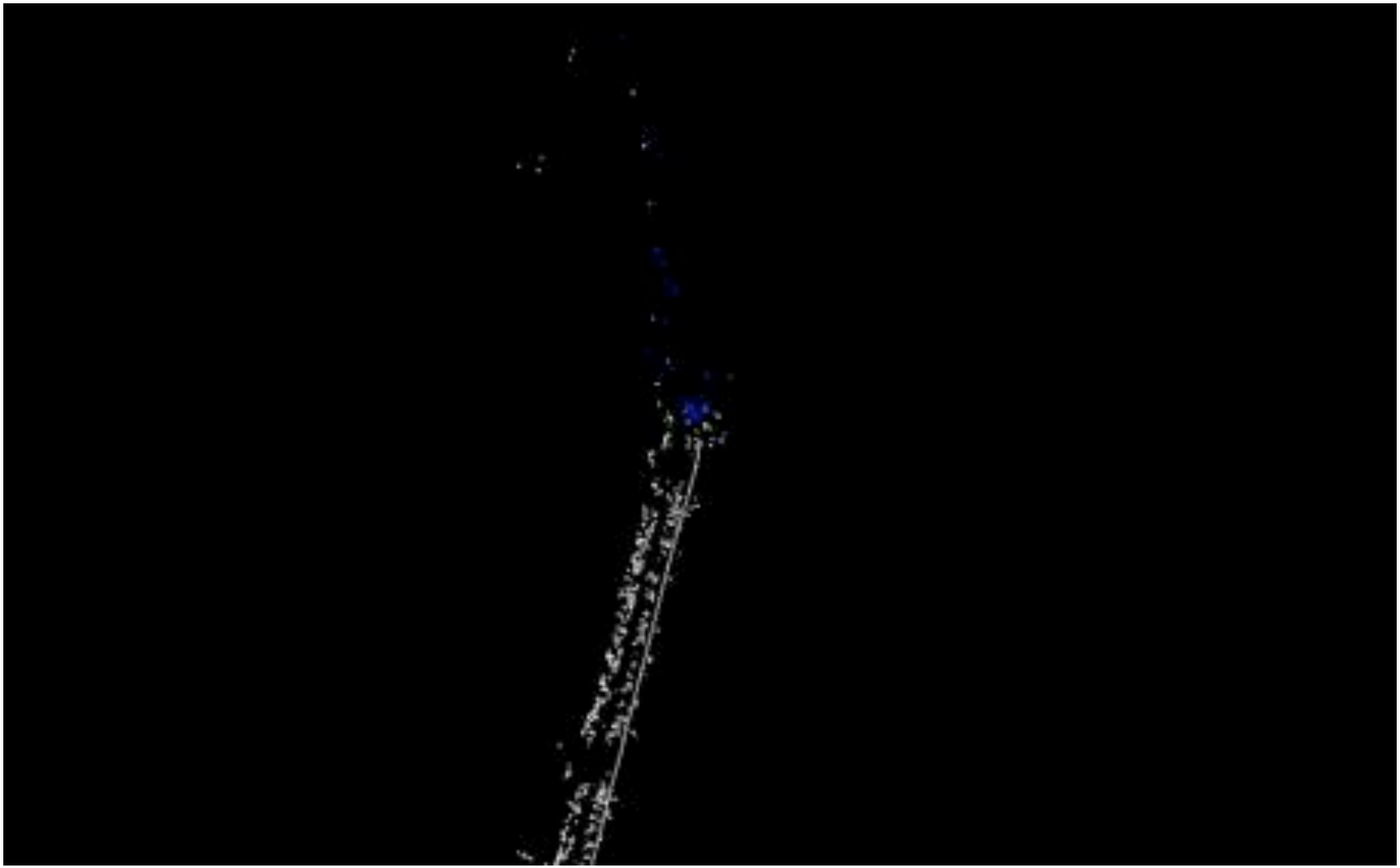
- TST/BTD

5. visual exploration

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

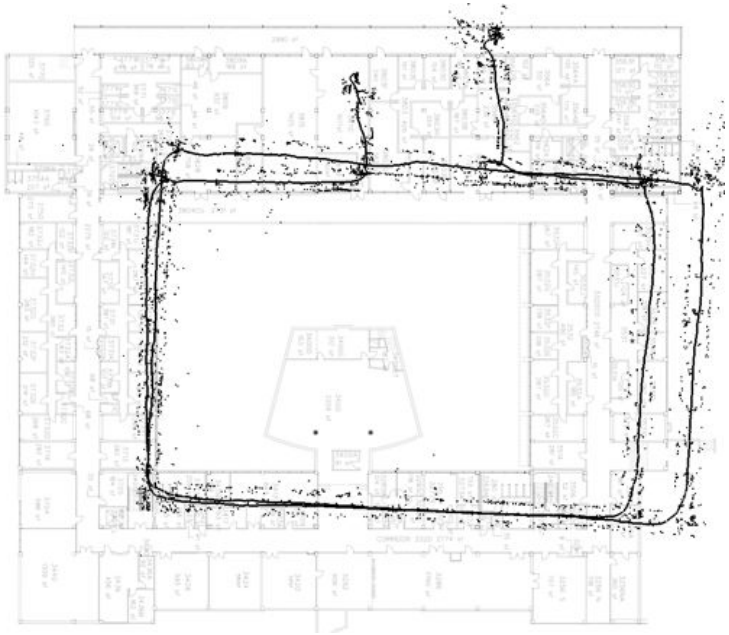




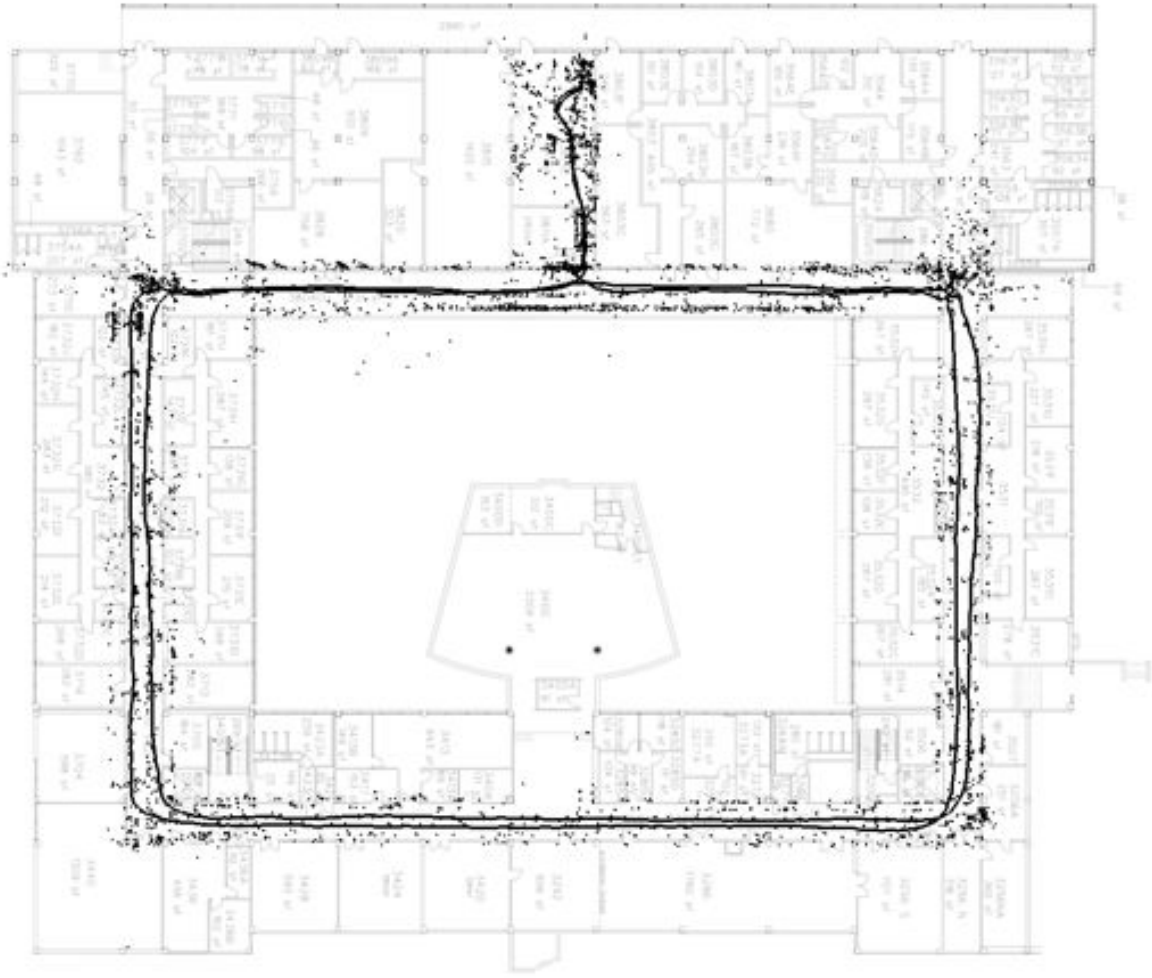


Inertial Only

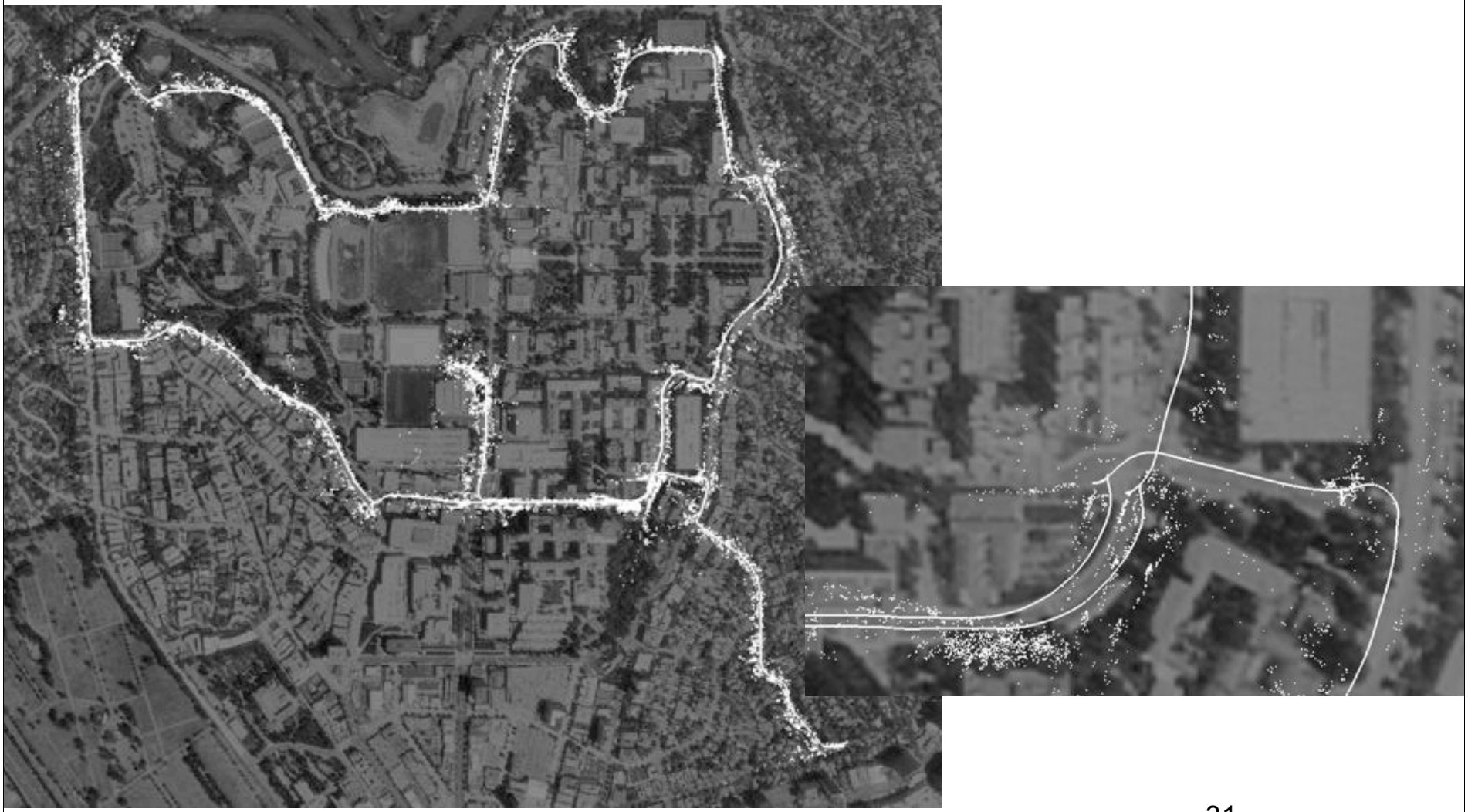
Vision Only



Drift: 0.19% (500 m)



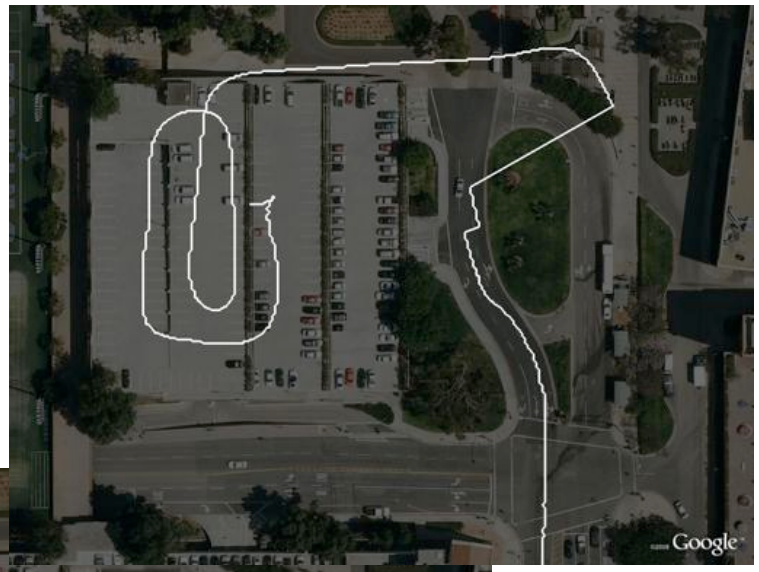
Drift: 0.27% (8 km)



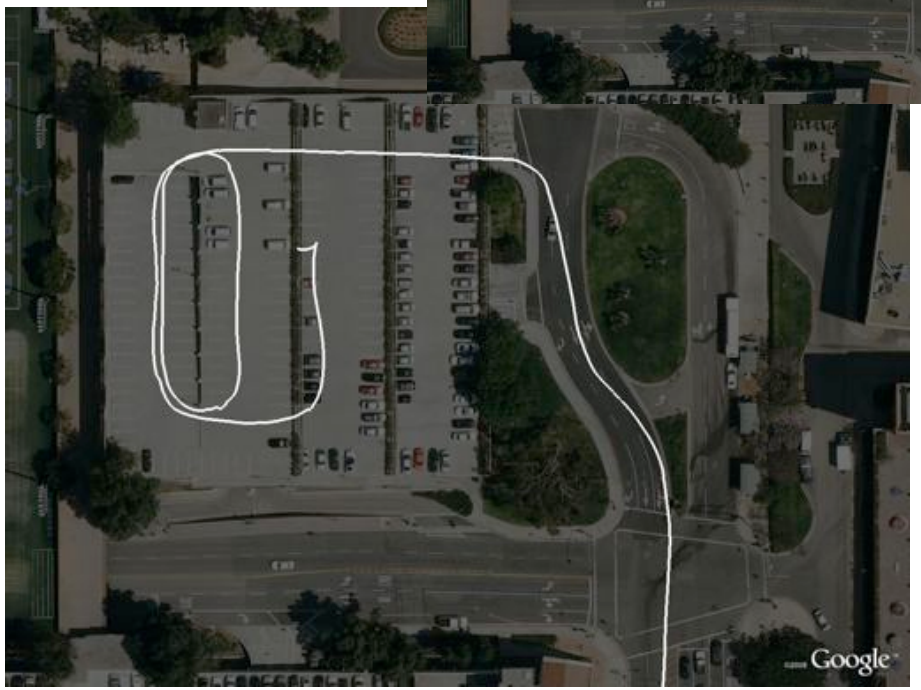
Drift: 0.5% (30km)

vs GPS+IMU

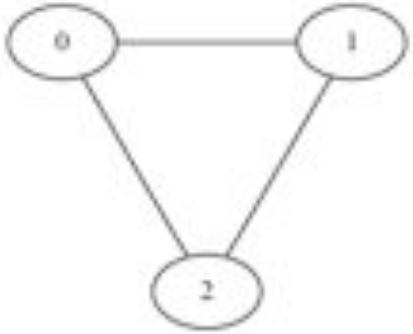
GPS+Inertial

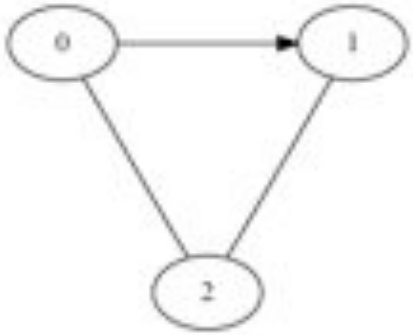


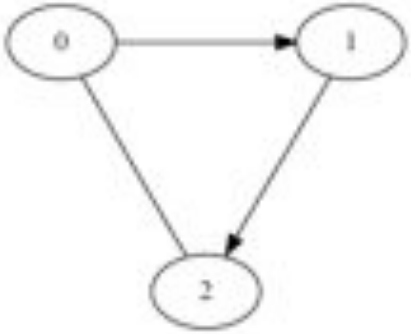
Vision+Inertial

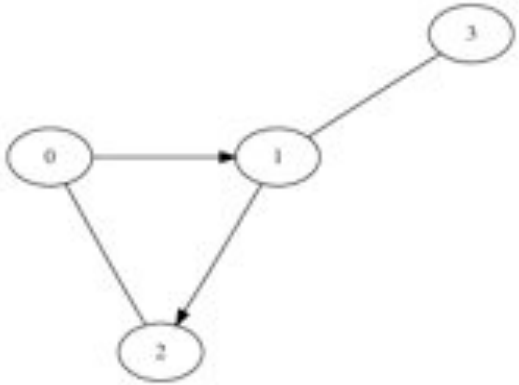


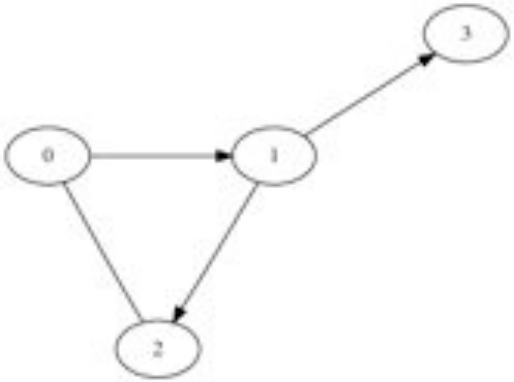
“location”, topology and co-visibility



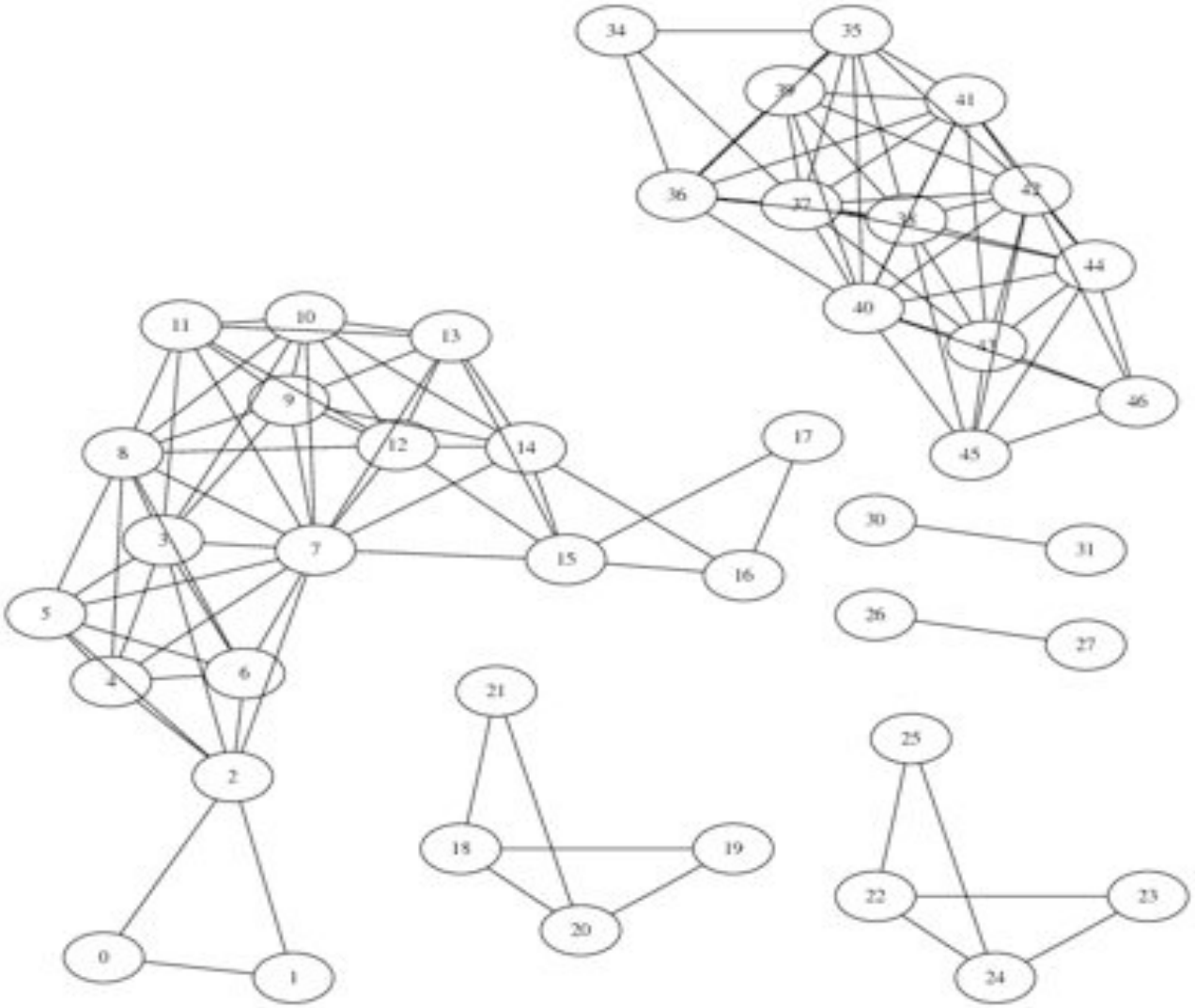




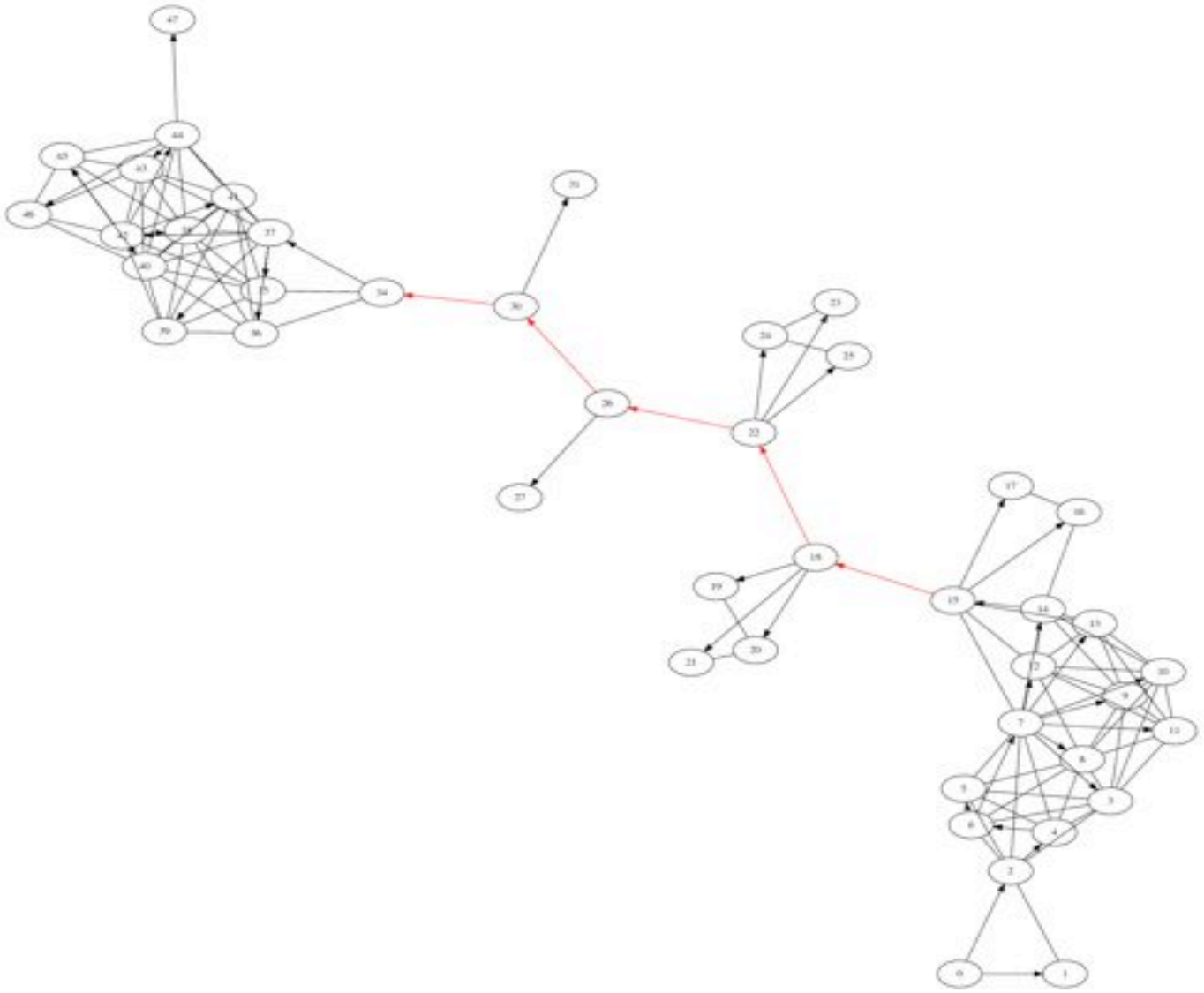




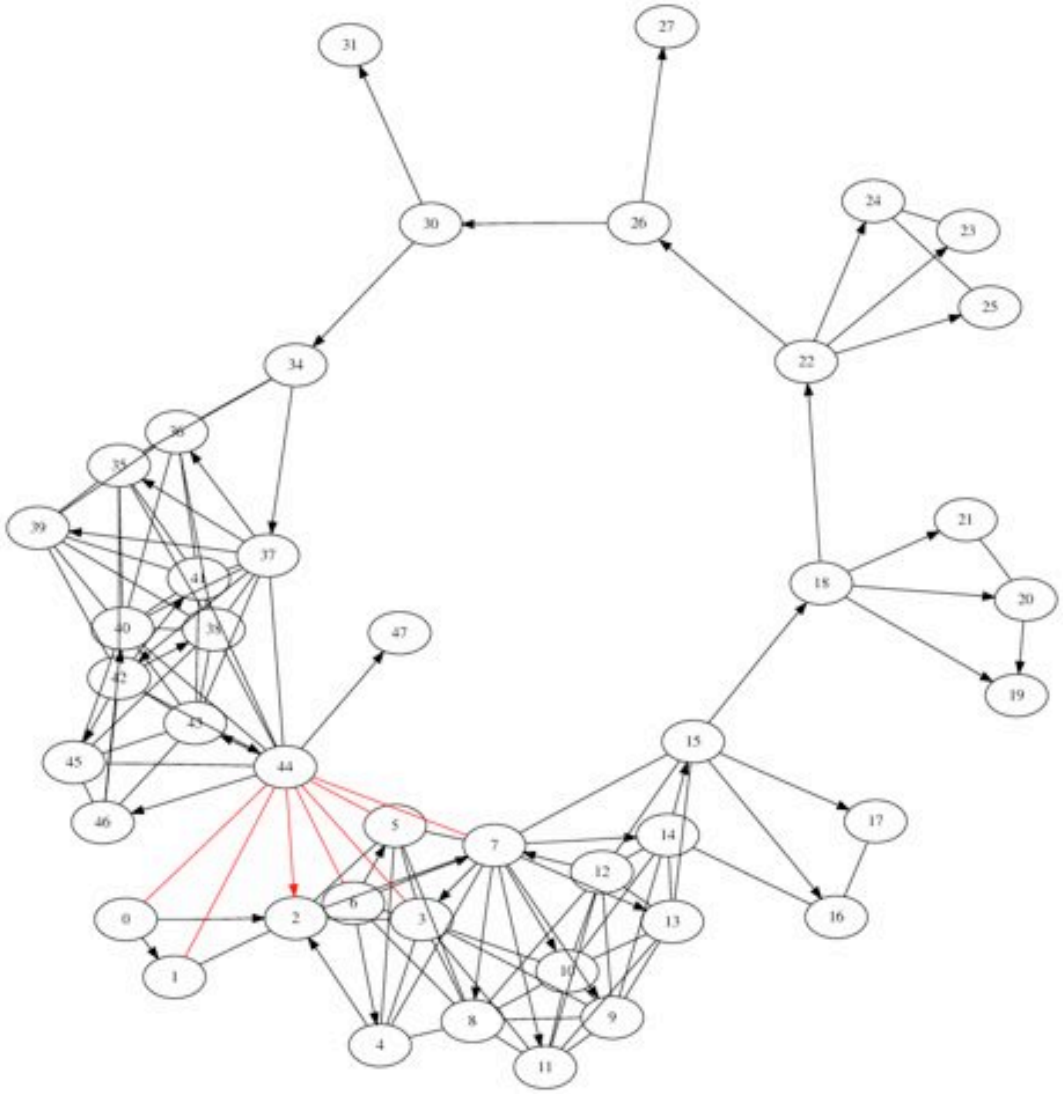
Covisibility Graph

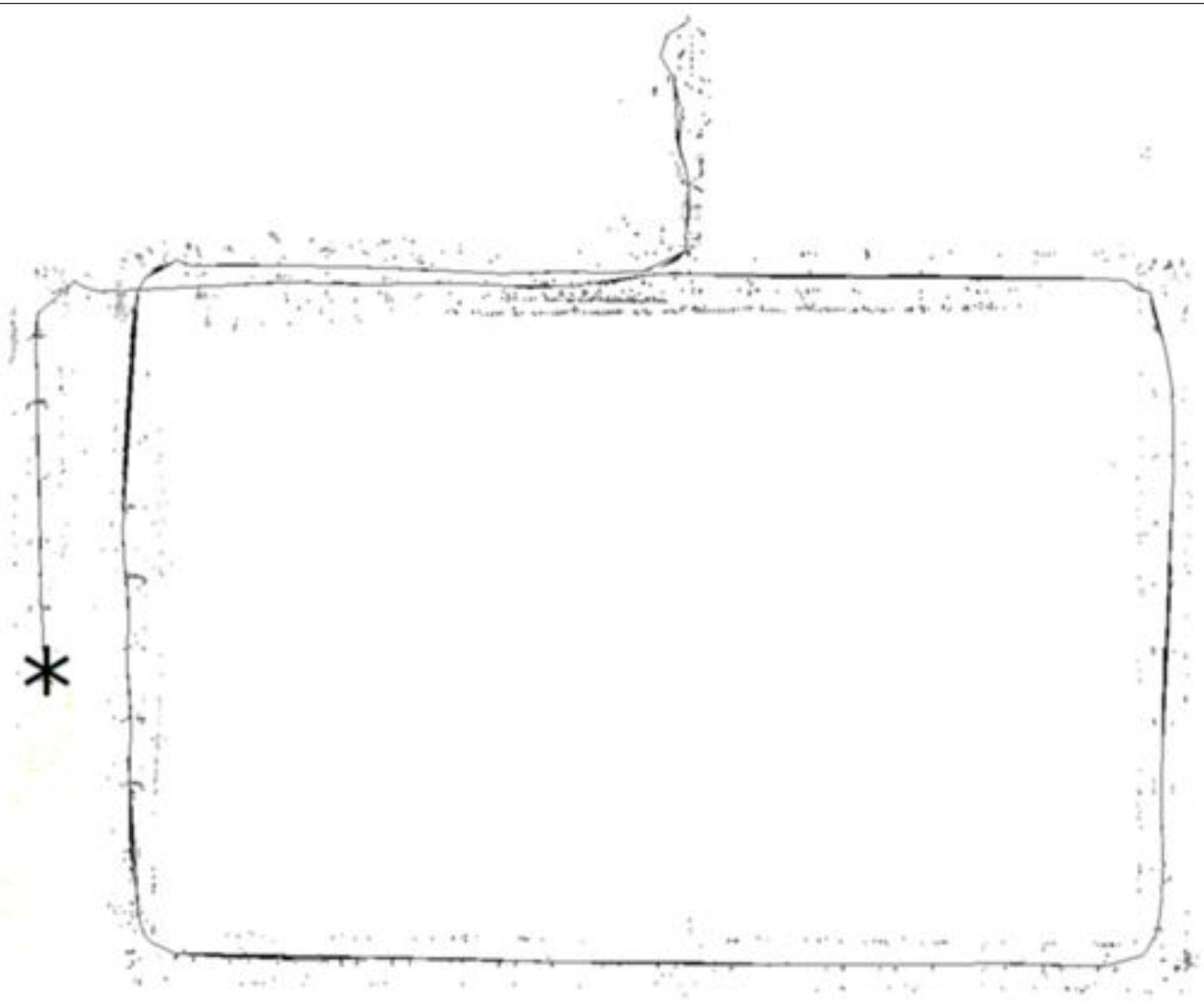


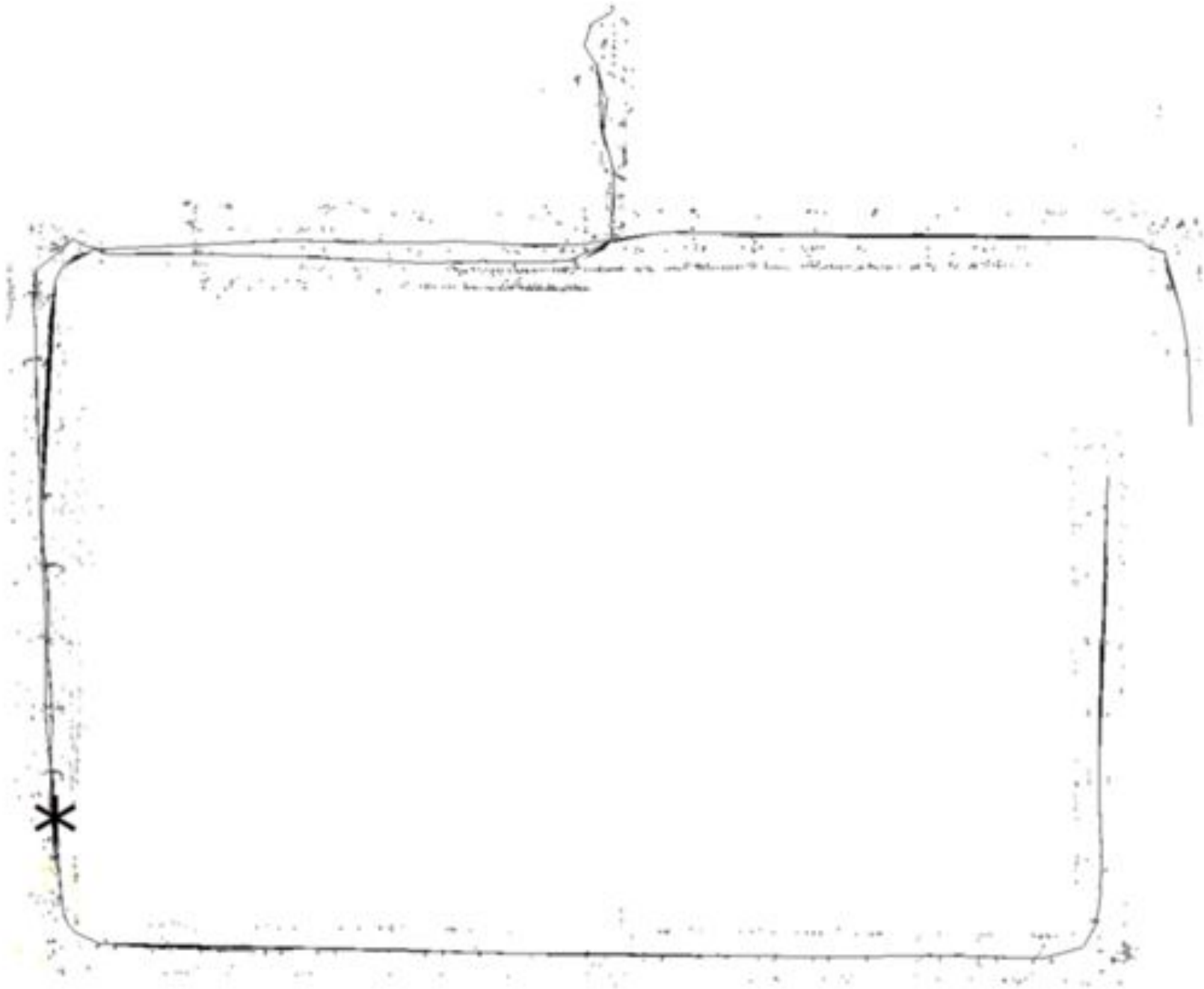
Adding Geometry

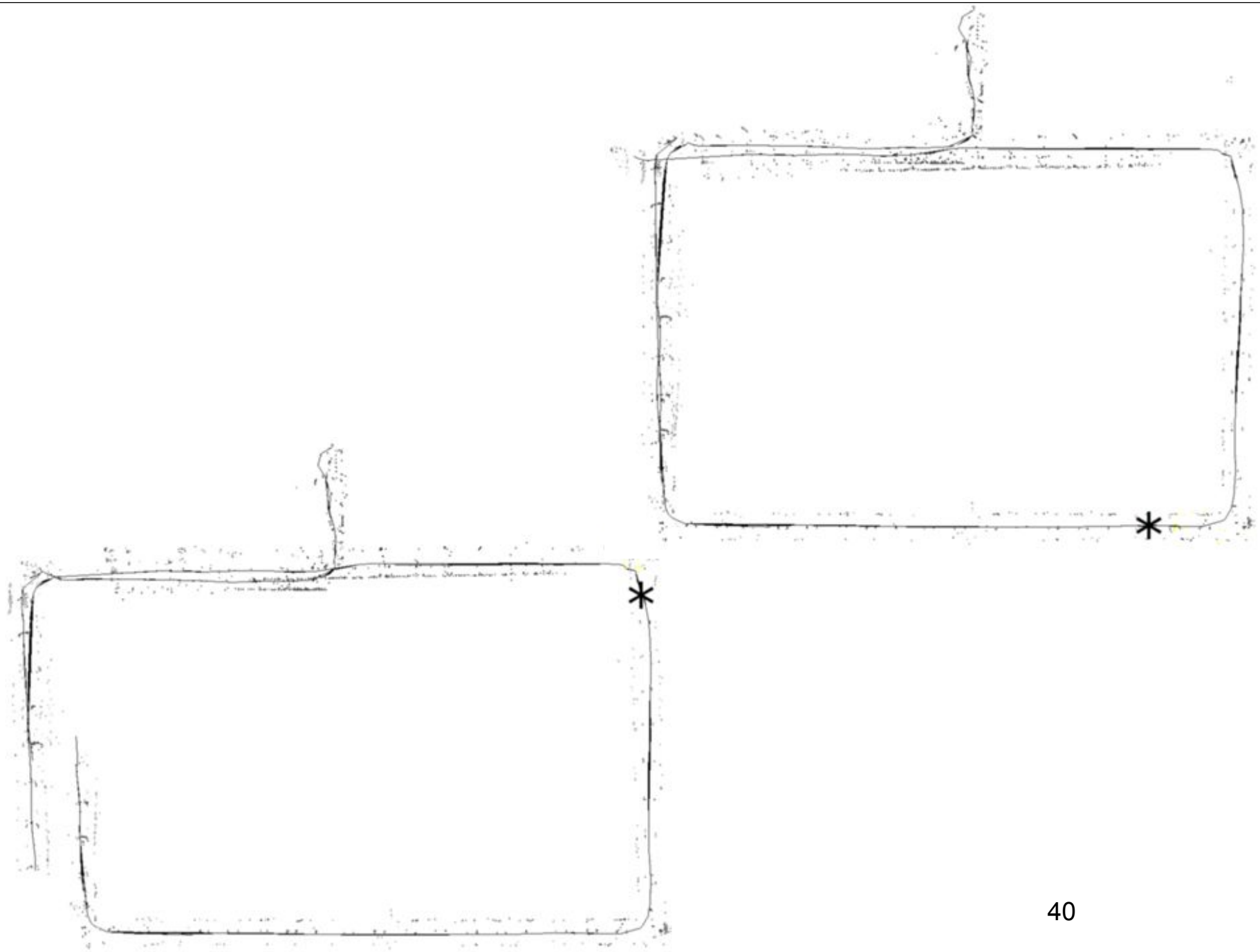


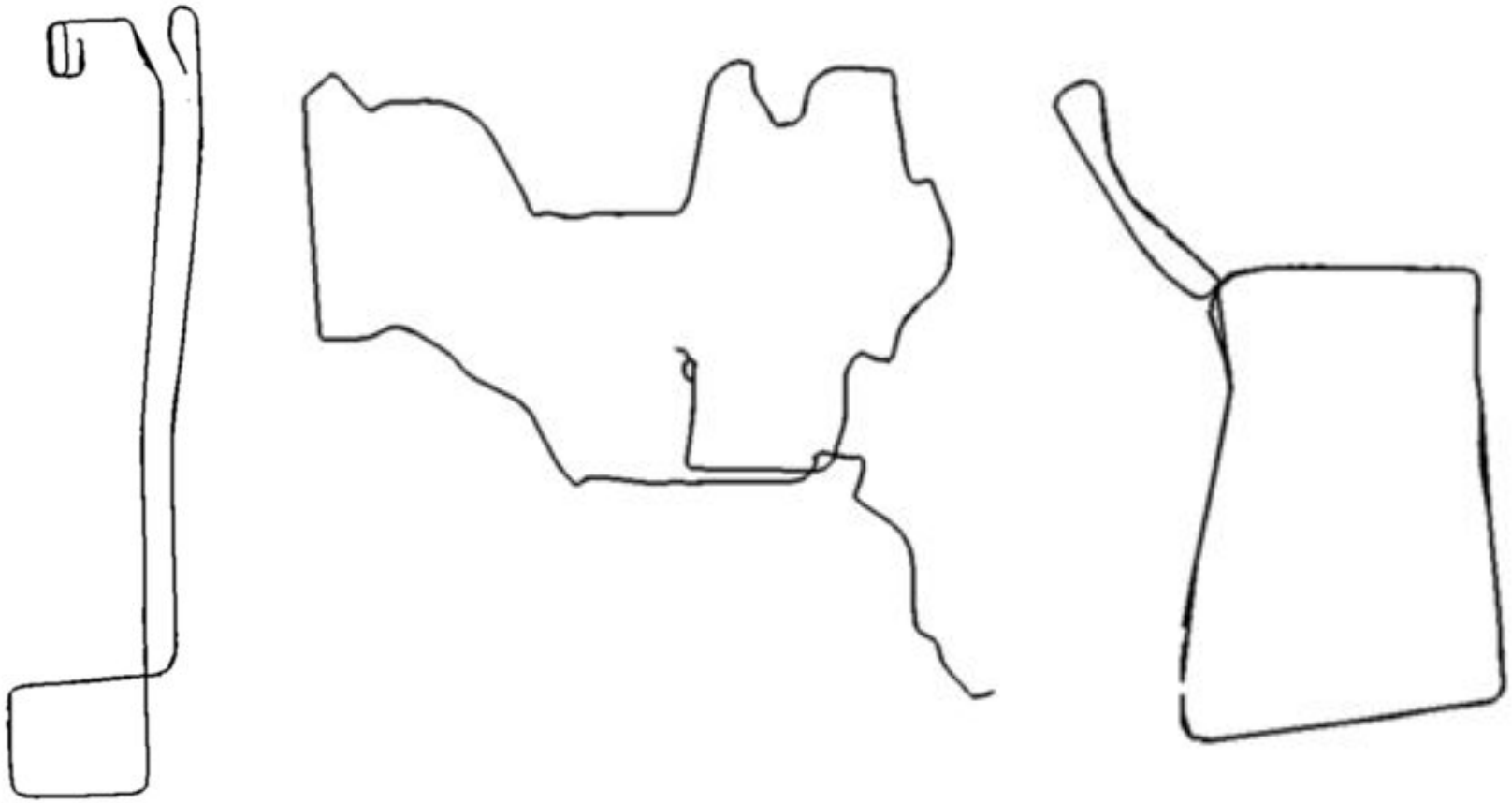
Loop Closing

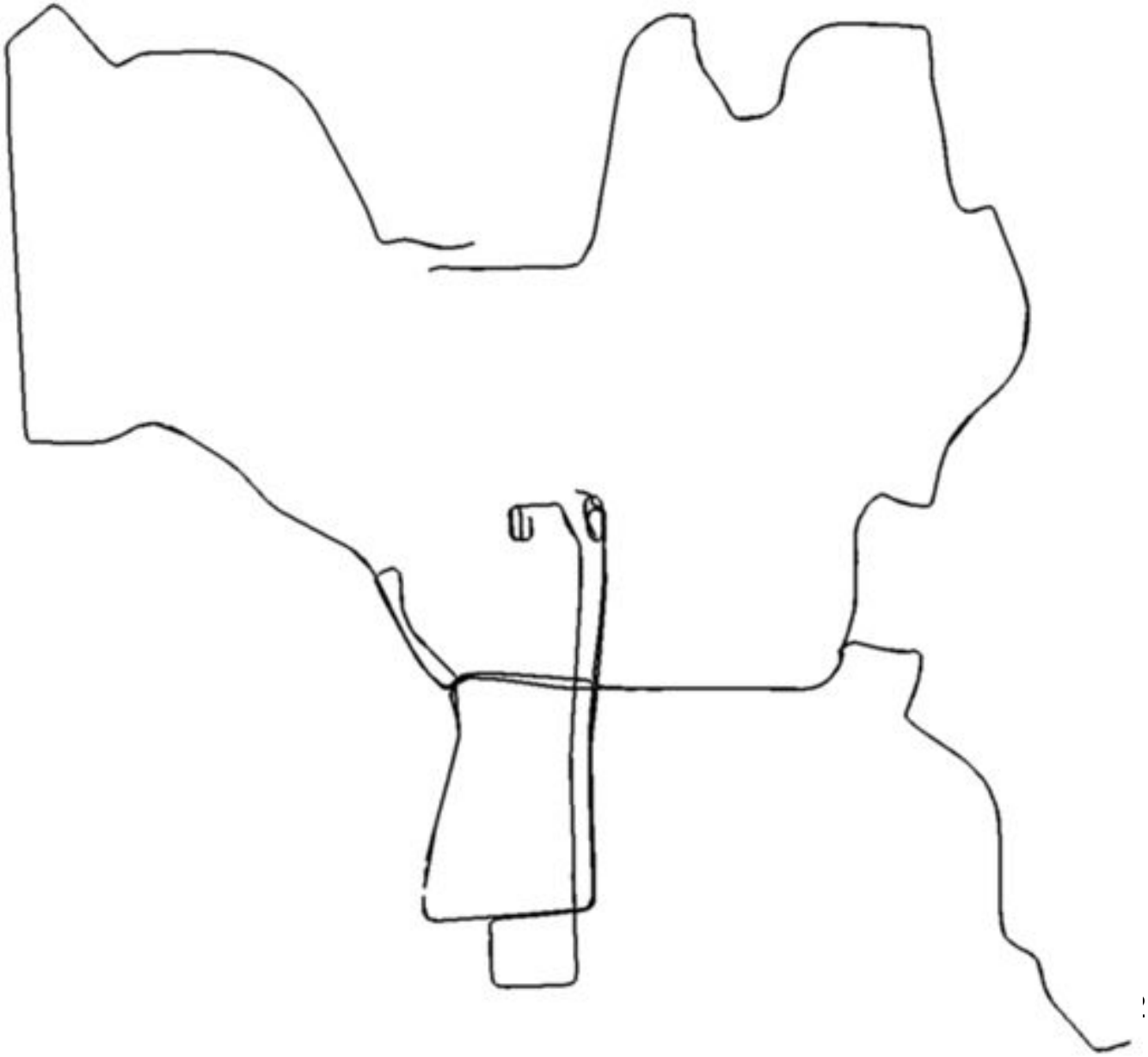












“The Black Box”

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

information pickup

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

$$\Omega(t) = \arg\min_{\Omega} \int_{D\setminus\Omega} \left(\nabla I w(x,t) + \frac{\partial I}{\partial t}(x,t) \right)^2 dx + \int_D \|\nabla w\|_{\ell^1} dx + \int_\Omega \|\nabla I\|^2 dx$$

- innovation and Actionable Information Increment

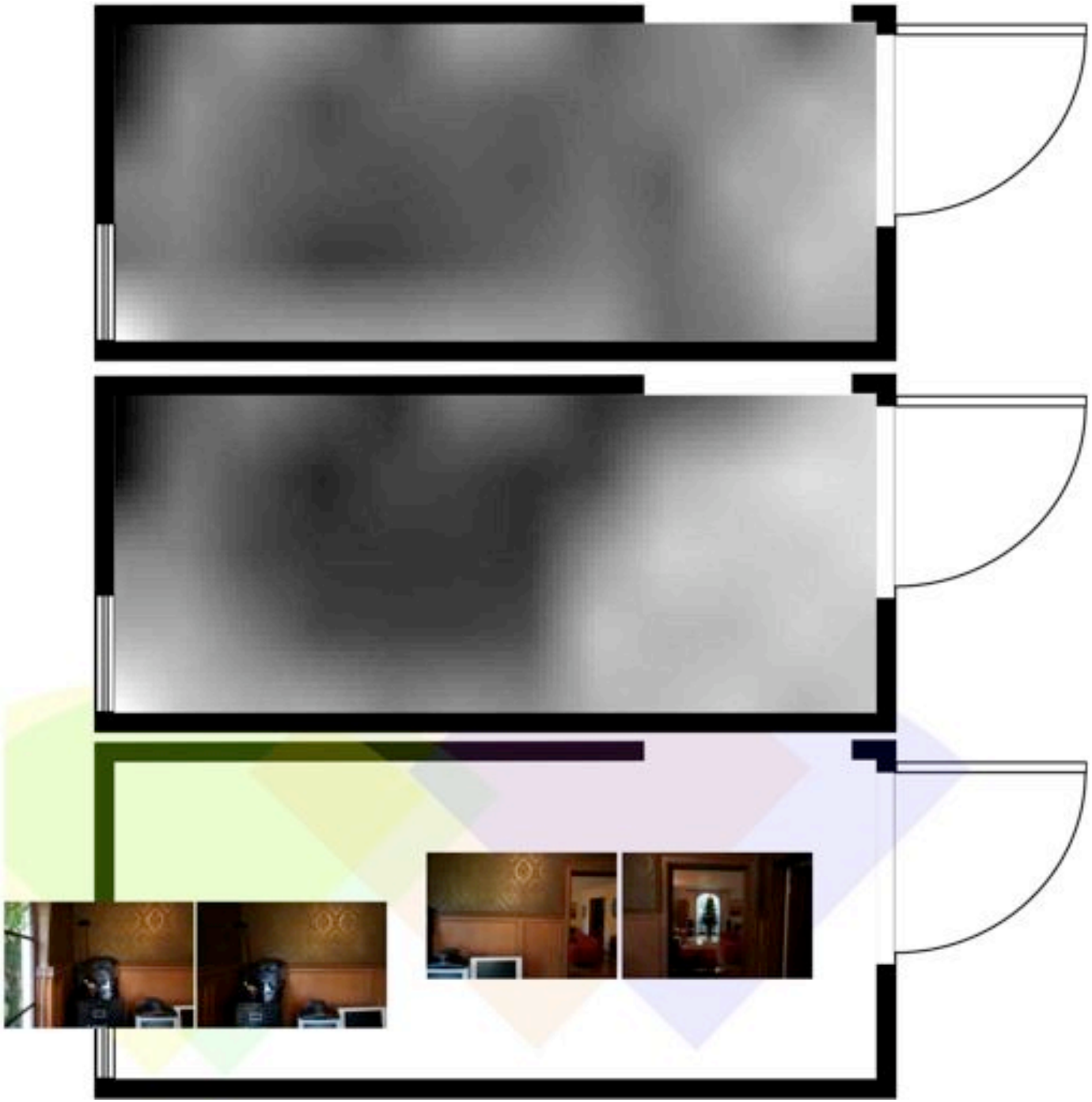
$$\boxed{\epsilon(I, t+dt) \doteq \phi^{\wedge}(I_{t+dt|_\Omega})} \qquad \boxed{AIN = H(\epsilon(I, t+dt)) = \mathcal{H}(I_{t+dt|_\Omega})}$$

- perceptual exploration:

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

building a representation: perceptual explorers

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

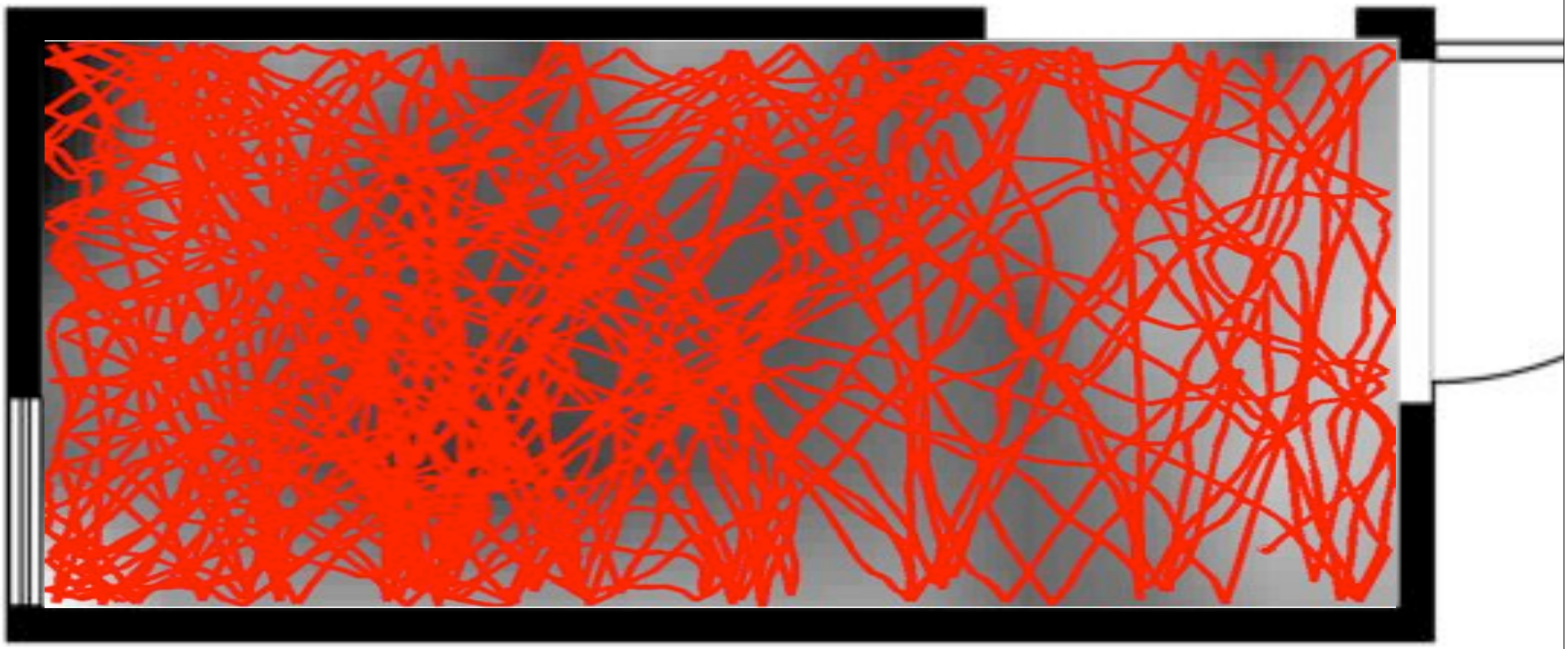


brownian explorer

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

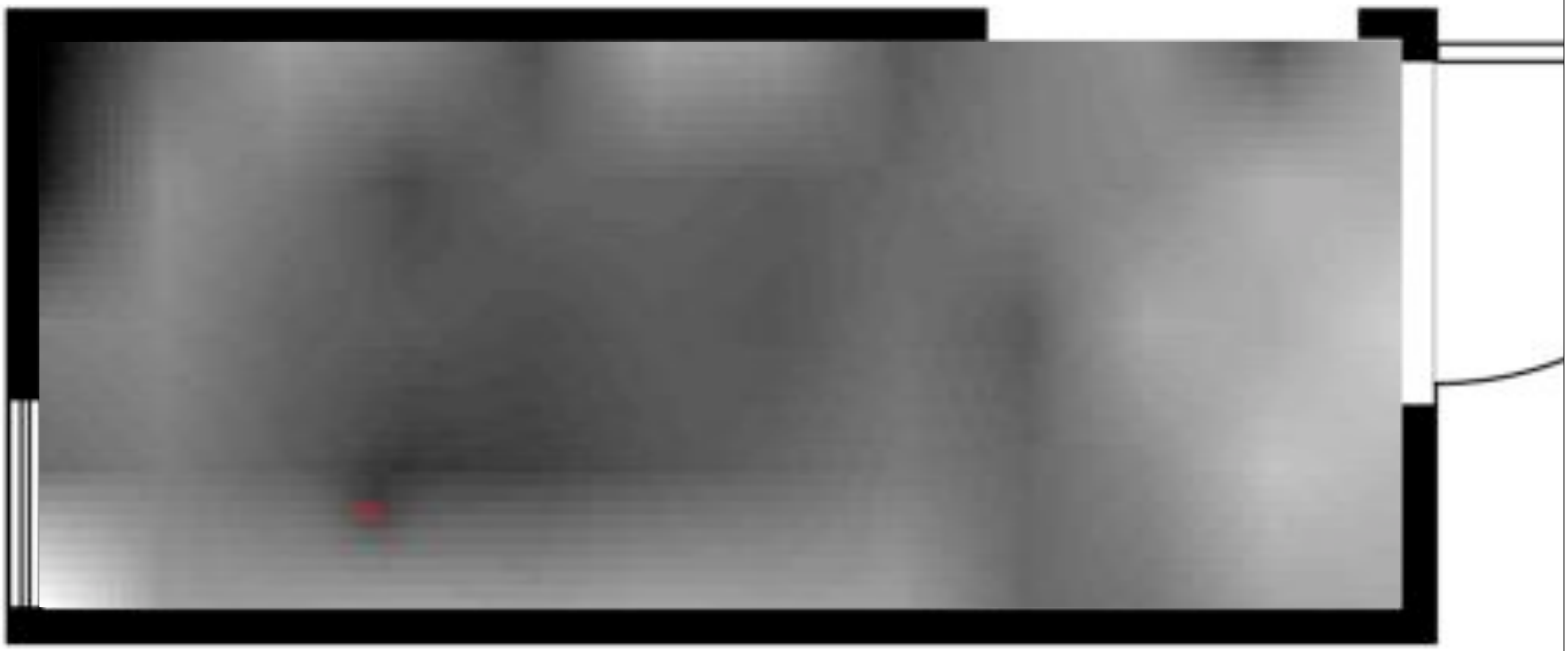
brownian explorer

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



reflections/shadow-paths

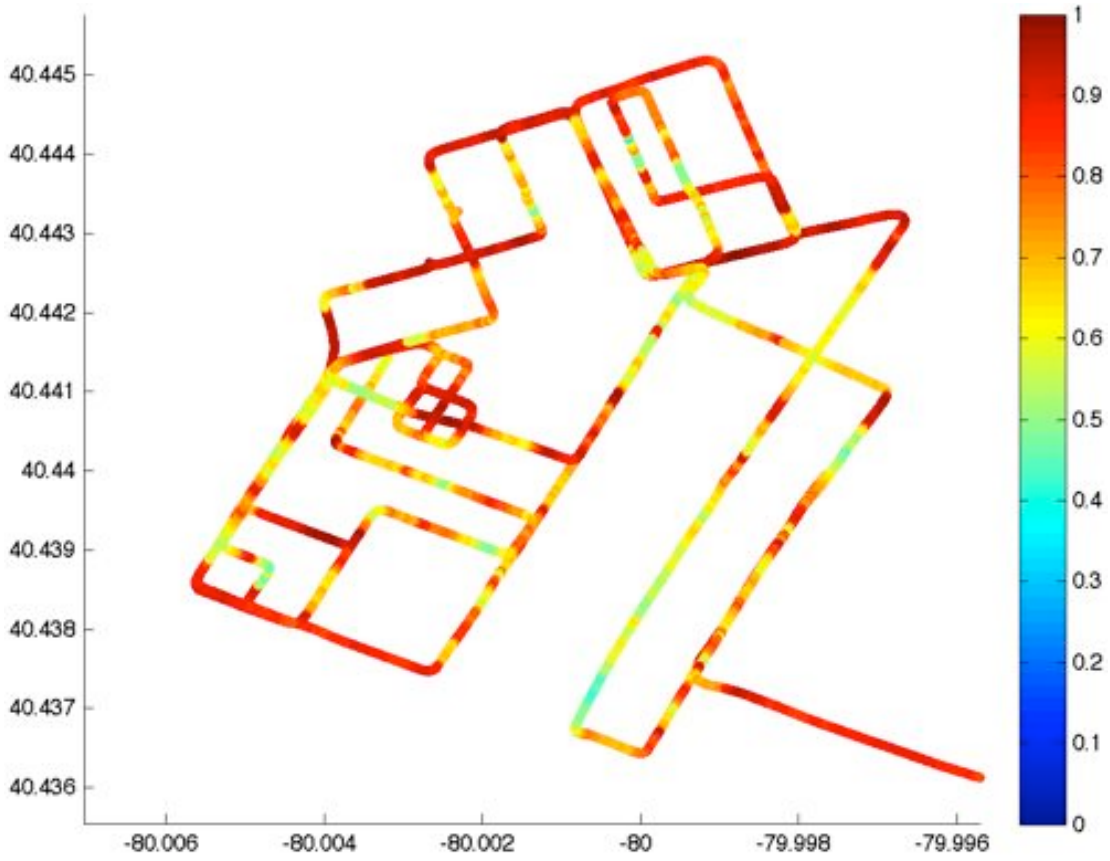
shannonian explorer



gibsonian explorer

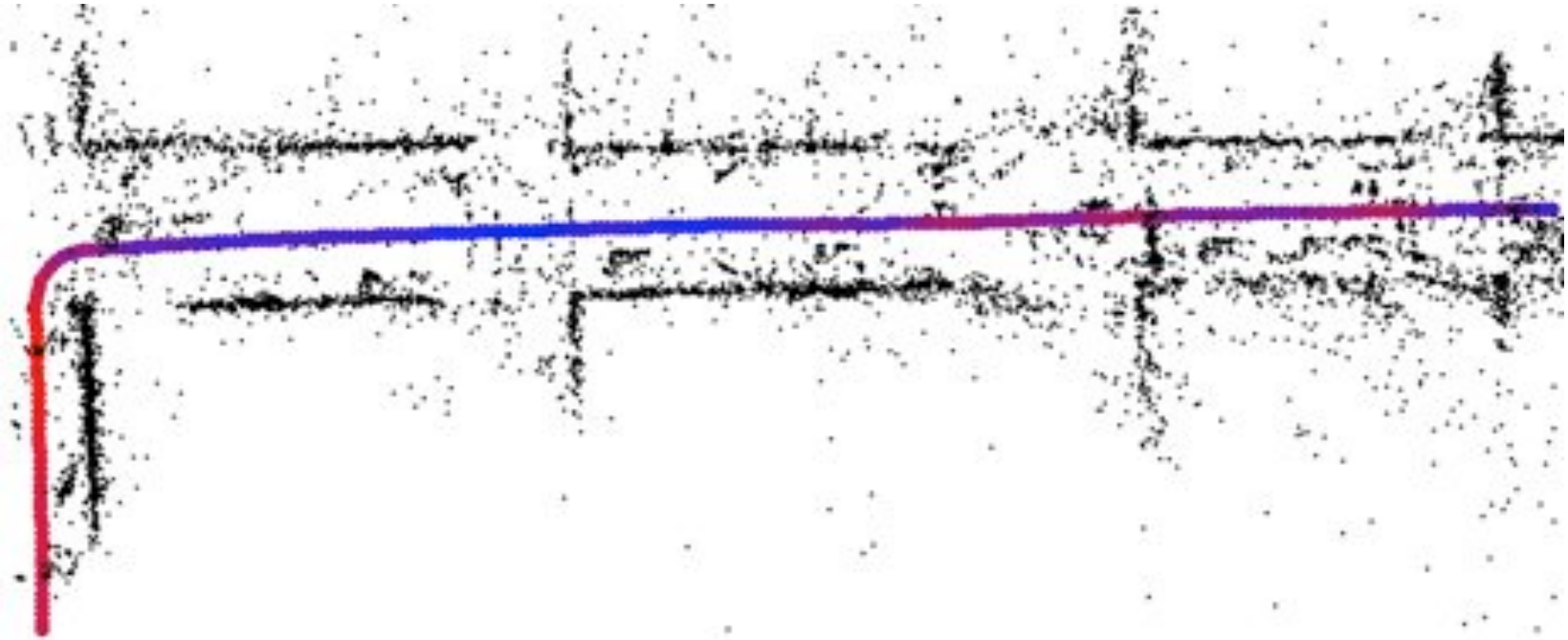
googleonian explorer

google street view dataset

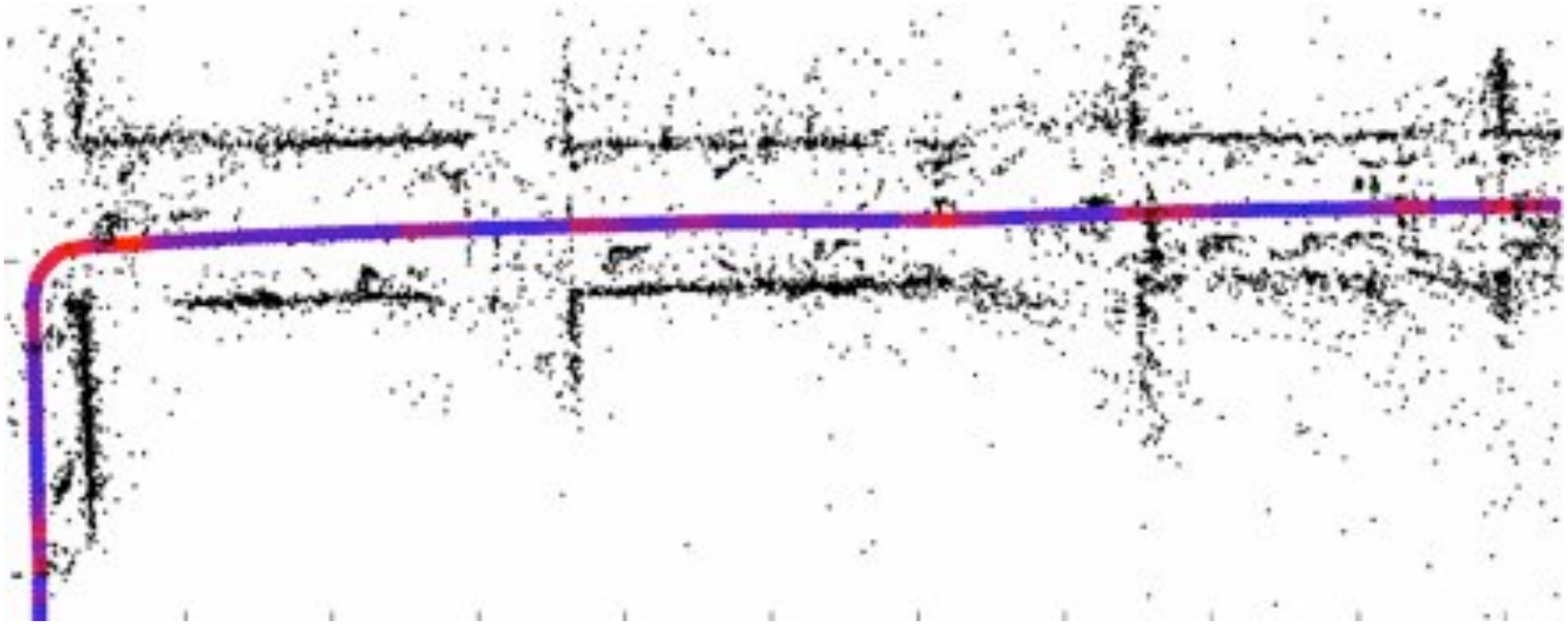


Courtesy of Taehee Lee

shannon in google's car seat



gibson in google's car seat



accommodation

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

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

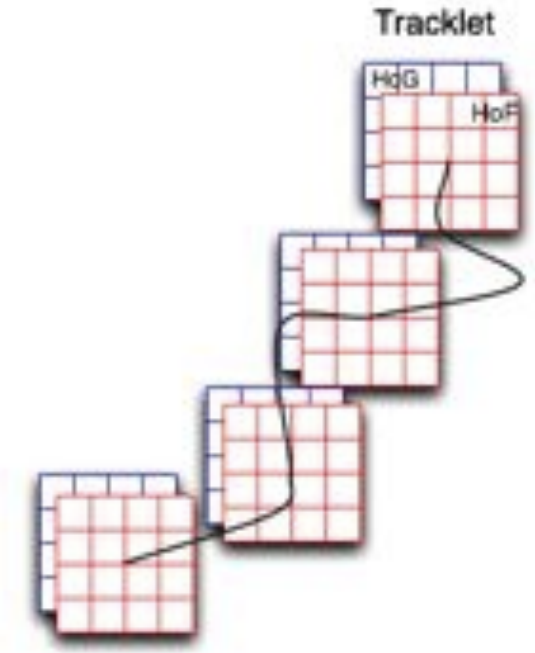
marginalizing time

- Tracklet Descriptor (related to Time-SIFT and Time-HOG)
- 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

recognition-control theory

- lower bound (passive sensor)
- upper bound (omnipotent observer)
- control authority (volume of reachable space) vs. risk tradeoff

take home messages

- signal-to-symbol barrier not addressed by standard (stat. dec. theory, inf. theory)
- "extracting information from data" can be done in a lossless fashion; requires control of sensing process
- optimal design of "features" for tracking, detection, localization, categorization, recognition
- controlled recognition bounds
- from "compressed sensing" to "controlled sensing"
- occlusion and mobility are key

relevant literature

- robotic exploration/next-best-view: all about the scene, not much about the image (separation of sensing and action)
- visual attention: all about the image, no connection to the scene
- active vision: no information-theoretic ramifications
- generative vs. discriminative, "represent vs. learn"
- information bottleneck, image epitome, "value of information" ...
- video coding, rate-distortion, rate-recognition theory
- primal sketch, sparse coding, compressed sensing