

CKID: The Center for Large-Scale Strategic Knowledge Integration and Discovery from Multiple Media

Eduard Hovy, director

Patrick Pantel, deputy director for ISI

Dennis McLeod, deputy director for USC



Info Extraction 1 — Marked text

in the present study, the pH of the acetate buffer used in the TMB incubation medium was adjusted from pH 5.5 to pH 7.0 in order to ascertain the optimal development of reaction product along with the best tissue preservation. Regions containing the MB were cut into blocks and processed for electron microscopy according to standard methods (see Materials and Methods, Allen and Hopkins, '88). Ultrathin sections were cut with a diamond knife and stained with uranyl acetate-lead citrate or left unstained before examination with a Zeiss EM 10A electron microscope.

Nomenclature
The nomenclature of the subicular complex used in the present study corresponds with Meibach and Siegel's (77) modifications of the initial descriptions of the hippocampal formation by Lorente de N6 (34). The nomenclature used for the prefrontal cortex follows that proposed by Krettek and Price (77).

Quantitative analyses
The diameters of labeled axon terminals were calculated by taking the mean of the long and short axes of the terminals as measured directly from electron micrographs (final magnification ~16,000). Since the MB is known to have reciprocal projections to the midbrain (Guillery, '57; Cruce, '77; Takeuchi et al., '85), estimates of the numbers of labeled and unlabeled neurons in the medial and lateral mamillary nuclei were made from 1 µm-thick plastic sections (toluidine blue stained) following injections of WGA-HRP into the midbrain. Only perikarya which were sectioned through the nucleolus were counted. Approximately 1,900 cells were counted from sections cut from selected rostral to caudal levels of the MB in eight animals.

RESULTS

In the present study, injections of WGA - HRP into the region of the MB resulted in dense retrograde labeling in the subicular complex, medial prefrontal cortex, and dorsal and ventral tegmental nuclei. Fewer retrogradely labeled perikarya were observed in the nucleus of the diagonal band of Broca, and small numbers of widely scattered labeled perikarya were found in the lateral hypothalamus. Dense anterograde labeling was observed in the anterior thalamus, dorsal and ventral tegmental nuclei, nucleus reticularis tegmentum proutum, and medial posterior nuclei.

Merents from the subicular complex

Light microscopy. Figure 1 shows the differential distribution of retrogradely labeled neurons in the subicular complex following injections of WGA - HRP into the medial and lateral mamillary nuclei. In one of the cases illustrated in Figure 1, the injection site (inset) was centered in the midline of the medial mamillary nucleus and included most of the subnuclei of the medial nucleus bilaterally. The lateral mamillary nucleus was spared but there was some spread from the principal injection site dorsally into the medial portion of the supramamillary nucleus. Large numbers of retrogradely labeled perikarya were found bilaterally in all layers of the dorsal and ventral portions of the subiculum but no labeled cells were found in the presubiculum or parasubiculum (Figs. 1, 2). A few retrogradely labeled neurons were also found in the deep layers of the entorhinal granular cortex (Figs. 1B, 2). In the second case illustrated in Figure 1, the injection site (inset) was located mainly in the lateral mamillary nucleus with a slight involvement of the dorsal part of the medial nucleus. In addition, there was some spread from the injection site dorsally into the lateral portion of the supramamillary nucleus and lateral hypothalamus. Numerous retrogradely labeled perikarya were found mainly ipsilaterally in the presubiculum and parasubiculum (Figs. 1, 3). A few labeled neurons were also found in ipsilateral dorsal subiculum as well as in the contralateral presubiculum.

Following injections of WGA - HRP into the subicular complex, anterograde labeling was distributed in distinct horizontal bands or layers across the MB bilaterally (Fig. 4). The horizontal layers of anterograde labeling were present primarily in either dorsal or intermediate or ventral parts of the medial mamillary nucleus depending on the locations of the injection sites in the rostral to caudal parts of the subicular complex. Figure 4A - D shows the results from a representative case in which WGA - HRP was injected into the rostro-dorsal portion of the subiculum. The resultant anterograde labeling was present in the medial mamillary nucleus bilaterally and formed a horizontal layer across the dorsal portion of the medial mamillary nucleus (Fig. 4B - D). The anterograde labeling was moderate to light in the anterior (Fig. 4B) and middle (Fig. 4C) thirds of the medial nucleus and heavy in the posterior third of the MB (Fig. 4D). The anteromedial part of the medial nucleus (pars medialis) contained only sparse anterograde labeling (Fig. 4B).

Figure 4E - H shows the results from a representative case in which WGA - HRP was injected into the caudoventral part of the subicular complex which included the presubiculum and parasubiculum. In this case, heavy anterograde labeling was present in the ventral portion of the posterior half of the medial mamillary nucleus bilaterally (Fig. 4G, H), whereas moderate to light anterograde labeling was present in the intermediate and dorsal parts of the anterior half of the medial nucleus bilaterally (Fig. 4F, G). The pars medialis showed very sparse or no anterograde labeling following injections in the caudoventral part of the subicular complex (Fig. 4F). Moderate to light anterograde labeling was also found in the lateral mamillary nucleus mainly ipsilaterally (Fig. 4G). Cases in which WGA - HRP injections into the subicular complex did not involve the presubiculum and parasubiculum showed no anterograde labeling in the lateral mamillary nucleus (Fig. 4A - D).

Electron microscopy. Following injections of WGAHRP into the subicular complex, labeled axons and axon terminals were observed in both the medial (Figs. 5 - 8) and lateral (Fig. 6C) mamillary nuclei. When DAB was used as the chromogen, labeled axon terminals were characterized by the presence of small amounts of electron - dense reaction product which were located in membrane - bound, lysosomal - like structures (Fig. 5). Identification of labeled axon terminals following DAB histochemistry required careful study of low - contrast, unstained sections with the electron microscope because in stained sections the DAB reaction product, although darker, resembled the staining seen in normal lysosomes. In contrast, when the TMB - DAB procedure was used, amorphous patches of electron - dense reaction product were found in axons and axon terminals in the MB (Figs. 6 - 8). The TMB - DAB - labeled axon terminals could be easily identified in stained sections at low magnifications because the TMB - DAB reaction product formed relatively large complexes and did not resemble normal tissue organelles (Fig. 7). There were, however, some disadvantages with the TMB-DAB procedure in comparison to the DAB procedure. For example, tissue elements were less well preserved and the reaction product was usually so large that it tended to obscure the contents of the axon terminals and the morphology of synaptic junctions following incubations in the standard TMB incubation medium: acetate buffer pH 5.3 - 6.0 (Fig. 6A). These problems were reduced when the pH of the acetate buffer used in the TMB incubations was made less acidic (pH 4.6 - 6.0). This simple modification of the TMB procedure resulted in a noticeable reduction in the amount of reaction product within the axon terminals, allowing visualization of synaptic vesicles and the morphology of synaptic junctions along with a much improved preservation of neural elements (Fig. 6B - D). The number of labeled axon terminals observed at the electron microscopic level was markedly decreased when the pH of the acetate buffer was greater than 6.0. Labeled axon terminals from the subicular complex ranged in diameter from 0.8 to 2.0 µm, contained mainly round vesicles (diameter = 40 nm), and formed asymmetric synaptic junctions mainly with small - diameter (less than 2 µm) dendrites and dendritic spines. Individual labeled axon terminals occasionally formed synaptic contacts with two adjacent dendrites (Fig. 8). Labeled axon terminals from the subicular complex only rarely contained pleomorphic vesicles and formed synaptic junctions with neuronal somata or proximal dendrites. Unlabeled axon terminals with pleomorphic vesicles and symmetric synaptic junctions with neuronal elements were, nonetheless, readily identified in this material.

Many labeled axon terminals appeared to form two separate synaptic specializations on individual dendritic profiles (Figs. 5, 6B, D, 8A), but serial sectioning of several labeled Merents from the medial prefrontal cortex.

Light microscopy. The distributions of retrogradely labeled neurons in the medial prefrontal cortex were mapped following injections of WGA - HRP into the MB. Figure 9 shows the results from a representative case in which retrograde labeling in the medial prefrontal cortex (Figs. 9A, B, 10) was obtained following an injection of WGA - HRP into the medial mamillary nucleus (Fig. 9C). terminals revealed that two apparently distinct synaptic specializations on the same dendrite were parts of a single continuous synaptic specialization (Fig. 8).

The injection was centered in the medial part of the medial mamillary nucleus with some spread of reaction product laterally into the lateral parts of the medial mamillary nucleus and dorsally into the medial portion of the supramamillary nucleus. The retrogradely labeled cells in the prefrontal cortex were pyramidal - shaped (Fig. 9B) and were distributed from the rostral limit of the prefrontal cortex to a level just rostral to the genu of the corpus callosum (Fig. 10). Most of the retrogradely labeled neurons were located in the deep layers of the infralimbic area while fewer labeled neurons were found rostrally and dorsally in or near area V of the prelimbic and anterior cingulate areas. A few labeled neurons were also found lateral and ventral to the tenia tecta. Some of the latter cells were located in the caudal end of the infralimbic cortex where they approached the rostralmost extent of the vertical limb of the diagonal band of Broca.

After unilateral injections of WGA - HRP into the medial prefrontal cortex (Fig. 11A), dense anterograde and retrograde labeling was observed in the caudal and lateral hypothalamus (Fig. 11B, C). The injection site shown in Figure 11A was centered in the medial wall of the prefrontal cortex with some spread dorsally into the medial precentral area, laterally into the claustrum and caudate / putamen, and ventrally into the region of the tenia tecta. Dense anterograde labeling was present along the dorsal margin of the medial mamillary nucleus and in the pars medialis bilaterally (Fig. 11B, C). Dense anterograde labeling was also found in the medial forebrain bundle and the medial part of the cerebral peduncle (Fig. 11C). In addition, anterograde and retrograde labeling were found mainly in the lateral portions of the supramamillary nucleus and along

Done

Info Extraction 2 — Database

Tract tracing experiments database - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://troll.isi.edu/tractbase/tractbase_v4.html

Tract tracing experiments - 268 data set

Labeling description:

Labeling location:

Injection location:

Tracer chemical:

[Entire database in CSV format](#)

Results

ID	File	Source	Sentence	InjectionLocation	LabelingLocation	TracerChemical	LabelingDescription
187	Allen-1989-286-311-ns.xml	0	132	NULL	NULL	NULL	heavy retrograde labeling
307	Allen-1990-301-214-ns.xml	0	76	NULL	NULL	NULL	Heavy retrograde labeling
317	Allen-1990-301-214-ns.xml	0	79	NULL	NULL	NULL	Heavy retrograde labeling
489	Allen-1992-315-313-ns.xml	0	137	NULL	NULL	NULL	Heavy retrograde labeling
510	Allen-1992-315-313-ns.xml	0	148	NULL	NULL	NULL	Heavy retrograde labeling
513	Allen-1992-315-313-ns.xml	0	149	NULL	NULL	NULL	moderate to heavy anterograde labeling
634	Allen-1993-330-421-ns.xml	0	120	NULL	NULL	NULL	Heavy retrograde labeling
919	Altschuler-1991-304-261-ns.xml	0	130	NULL	NULL	NULL	Heavier labeling
927	Altschuler-1991-304-261-ns.xml	0	137	NULL	NULL	NULL	heavy afferent terminal labeling

Done

Sentence 132, fixed_Allen-1989-286-311-ns.xml - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

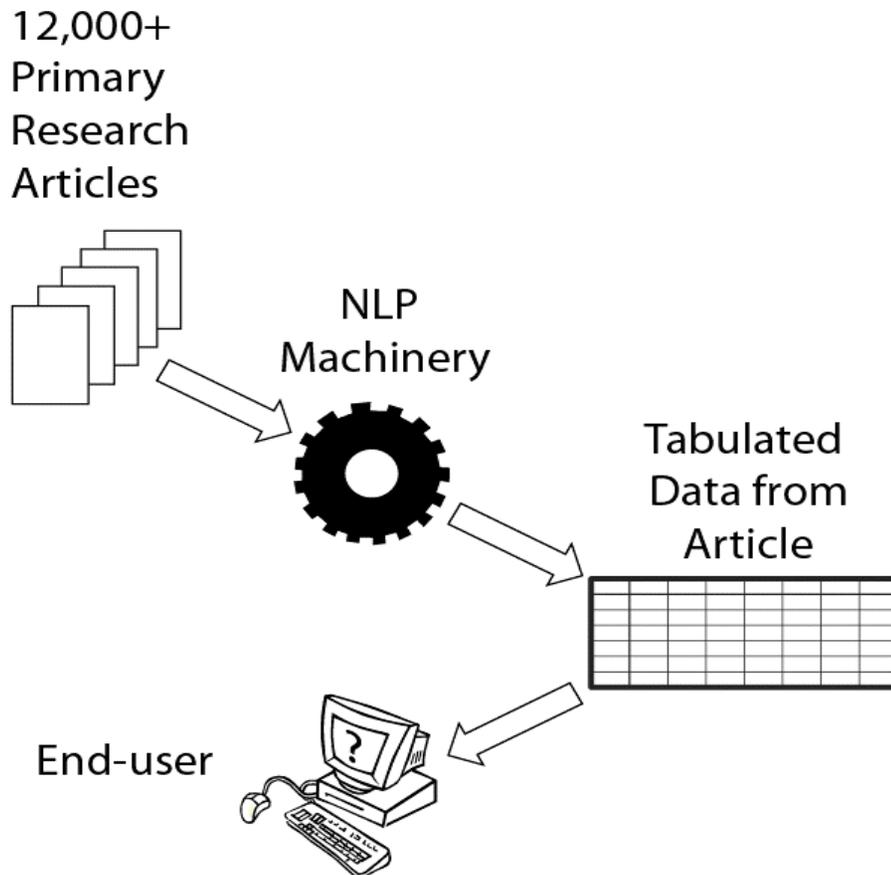
http://troll.isi.edu/cgi-bin/get_ser

In addition , there was **heavy retrograde labeling** in the **ventromedial part of the posterior ventral tegmental nucleus** , whereas **the dorsolateral part of the ventral tegmental nucleus** contained **only a few retrogradely labeled cells** ipsilaterally (Fig . 14B) .

Done

Info Extraction from unstructured text

- How to query masses of text accurately?
- How to perform automated pattern discovery?

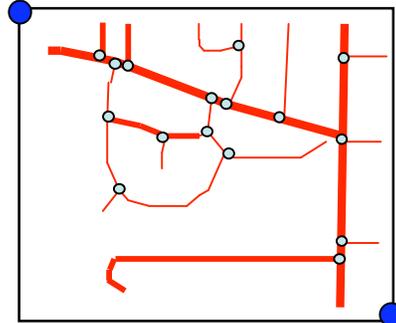


- Example projects: IE of **neuroscience articles** on tracing rat brain structure
- IE of **news articles** for Special Ops Command
- IE of **public responses** to Fed Gov regulation writers (EPA, DOT)...

Geospatial vector / image fusion

Fusion challenges:

- Different geographic projections
- Data collected at different scales
- Corrected using different elevation models
- Today this is addressed by:
 - Manually identifying control point pairs
 - Applying conflation techniques



Control point detection

Filtering technique

Intermediate control points

Final control points

Conflation (triangulation + rubber-sheeting)



Exploiting online sources to accurately identify and label structures in imagery

Palm Ave	
Penn St	Sierra St
Mariposa Ave	

(-118.40883, 33.92375)

Street vector data
Corrected Tiger line files

Address	Latitude	Longitude
642 Penn St	33.923413	-118.409809
640 Penn St	33.923412	-118.409809
636 Penn St	33.923412	-118.409809
604 Palm Ave	33.923414	-118.409809
610 Palm Ave	33.923414	-118.409810
645 Sierra St	33.923413	-118.409810
639 Sierra St	33.923412	-118.409810



Constraint Satisfaction

E PALM AV

604 or 642	604 or 610	610, Palm or 645, Sierra
642, Penn or 636, Penn		645, Sierra or 639, Sierra
636, Penn or 630, Penn		639, Sierra or 633, Sierra
630, Penn or 628, Penn		633, Sierra or 629, Sierra
628, Penn or 624, Penn		629, Sierra or 623, Sierra
624, Penn or 618, Penn		

SIERRA ST

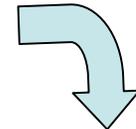
E PALM AV

604	610	645, Sierra
642, 644, 646 Penn		639, Sierra
636, 638, 640 Penn		633, Sierra
630, 632, 634 Penn		629, Sierra
628, Penn		623, Sierra
624, Penn		

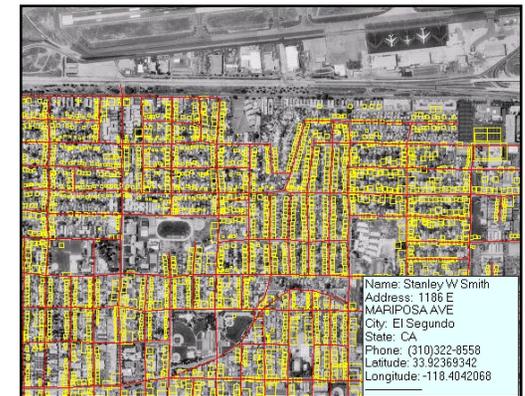
SIERRA ST

Initial Hypothesis

Result After Constraint Satisfaction



Geocoded Houses



Address	# units	Area(sq ft)	Lot size
642 Penn St	3	1793	135.72 * 53.33
604 Palm Ave	1	884	69 * 42
610 Palm Ave	1	756	66 * 42
645 Sierra St	1	1337	120 * 62
639 Sierra St	1	1408	121 * 53.5

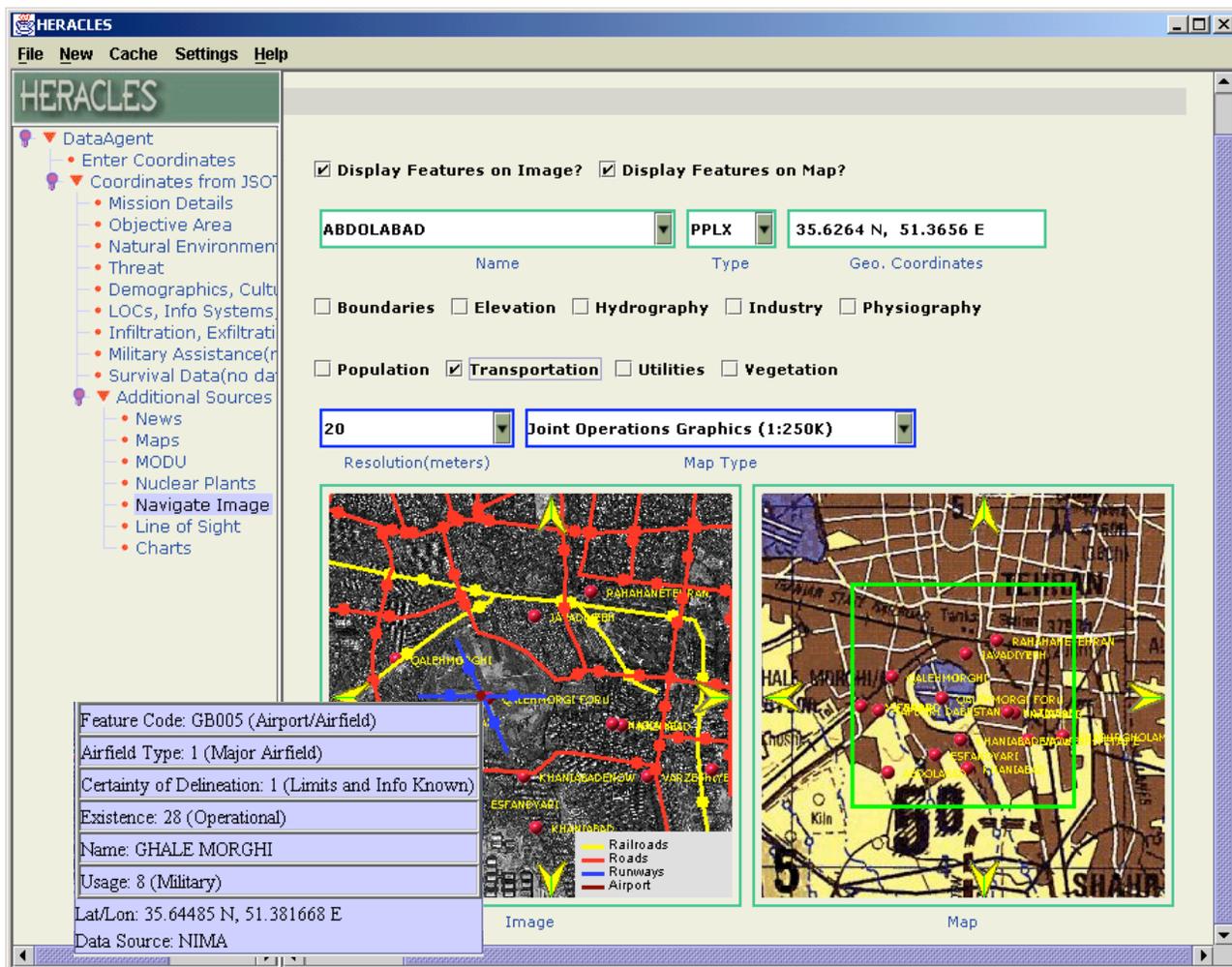
Address	# units	Area(sq ft)	Lot size
642 Penn St	3	1793	135.72 * 53.33
604 Palm Ave	1	884	69 * 42
610 Palm Ave	1	756	66 * 42
645 Sierra St	1	1337	120 * 62
639 Sierra St	1	1408	121 * 53.5

Los Angeles County Assessor's site Property Tax records Data extracted from online site



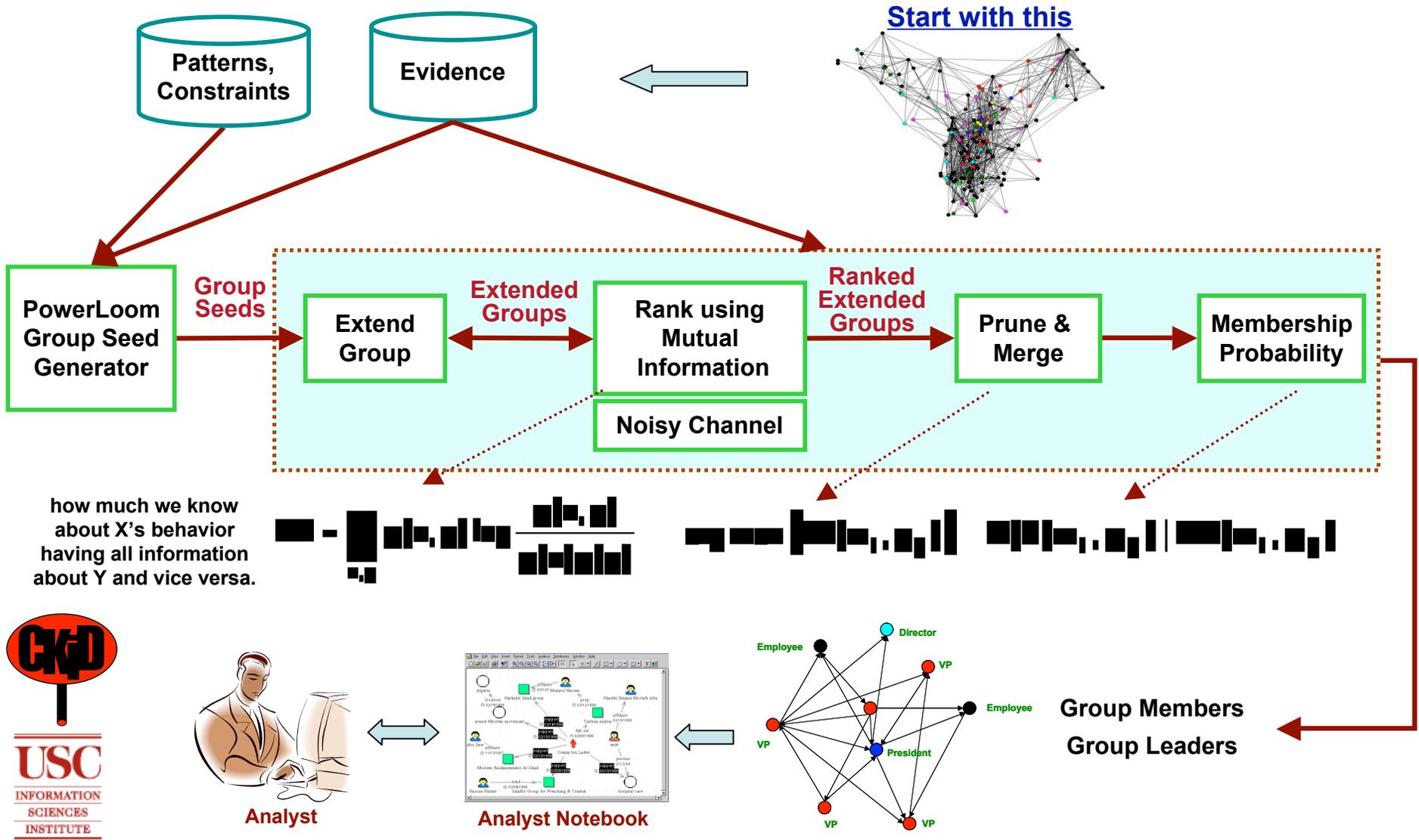
Geospatial data retrieval

Heracles: Framework to integrate heterogeneous data



- Structured data: databases
- Semi-structured data: web pages
- Spatial Data:
 - Vector data: points, lines, polygons, ...
 - Images: satellite, and aerial imagery
 - Maps
- Text documents
- Audio & Video: TV & Radio on the web

KOJAK Group Finder overview



Results at a glance

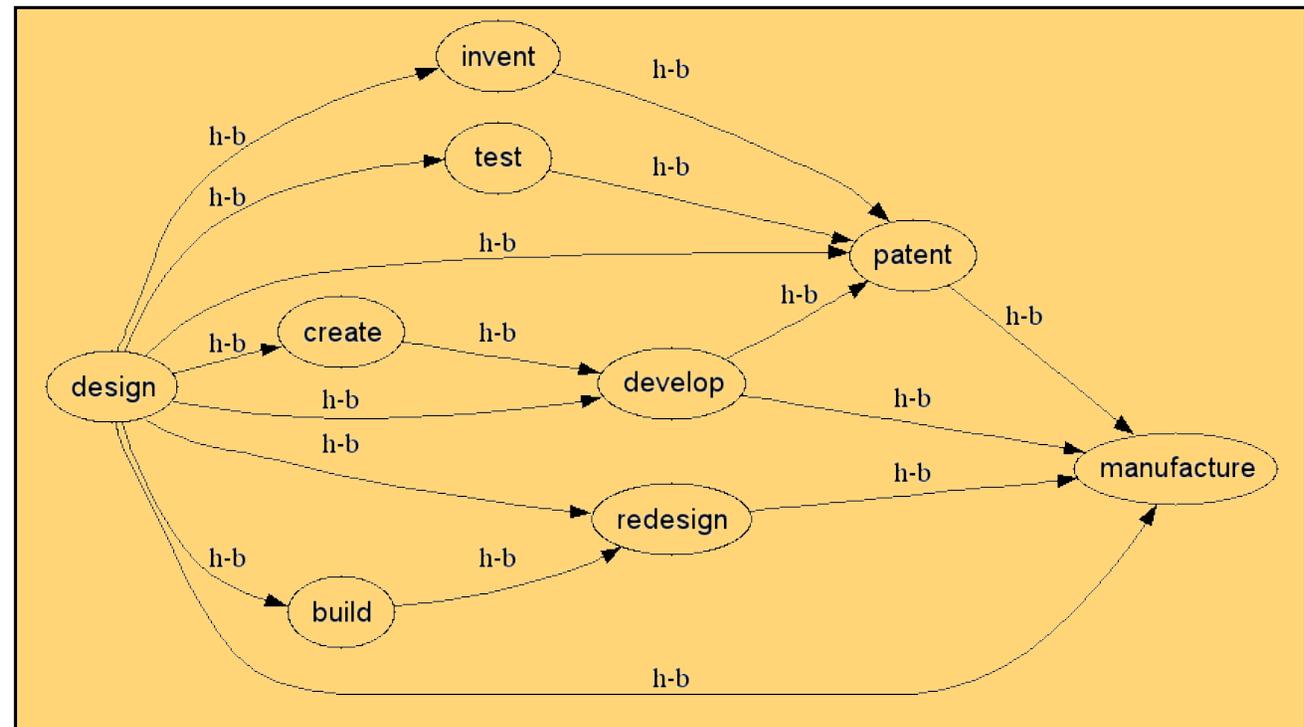
Datasets	IET Y2	IET Y2.5	IET Y3
Dataset Characteristics	Medium	Hard	Very Hard
Maximum #of Entities	O(10,000)	O(10,000)	O(100,000)
Maximum #of Links	O(100,000)	O(1,000,000)	O(10,000,000)
Group detection <i>f</i> -value (best/avg.)	0.60 (0.51)	0.87 (0.68)	0.77 (0.50)
“Bad guy” detection <i>f</i> -value (best/avg.)	-	0.98 (0.81)	0.96 (0.69)
Group refinement <i>f</i> -value (best/avg.)	-	0.93 (0.73)	0.97 (0.74)
#of groups accuracy	-	-	1.00 (0.84)
Speed (avg. CPU time per group)	-	-	23 sec
Evaluation Performance	1st place combined score	1st place TIE- level group detection	1st place TIE- level group detection

Learning entailments (VerbOcean)

(Chklovski and Pantel 2004; Pantel 2005)

- **Goal:** Automatically learn relationships between verbs referring to temporally disjoint events that co-occur
- See <http://semantics.isi.edu/ocean/>

Example:
temporal
entailments
from *design* to
manufacture



Approach

- Manually built 35 lexical patterns that signal some kind of entailment
- Learning in three steps:
 1. Identify pairs of highly associated verbs co-occurring frequently on the Web using DIRT (Lin and Pantel 2001)
 2. For each verb pair
 - apply patterns
 - calculate score for each possible semantic relation
 3. Compare strengths of individual semantic relations; output a consistent set
 - prefer the most specific and then strongest relations

<i>SEMANTIC RELATIONS</i>	<i>Patterns</i>
similarity (4)	X ie Y Xed and Yed
strength (8)	X even Y Xed even Yed X and even Y Xed and even Yed
enablement (4)	Xed * by Ying the Xed * by Ying or to X * by Ying the
antonymy (7)	either X or Y either Xs or Ys Xed * but Yed
happens-before (12)	to X and then Y Xed * and then Yed to X and later Y to X and subsequently Y Xed and eventually Yed



- **Obtained 29,165 verb pairs (entailments):**



– Applied DIRT to 1.5GB newspaper corpus

– 4000 verb pairs per day on a single machine, 40 days

Goals

- DHS faces a massive problem:
 - There is too much information...
 - in multiple media...
 - it's undifferentiated: the good stuff is mixed with the trash
- So:
 - **1. 'Homogenize' the data:** convert it into a single unambiguous format/formalism, using medium-specific feature extraction/annotation technology
 - **2. Analyze the data:** identify recurrent patterns of interest across all the data, regardless of original medium, using data mining and machine learning technology

and then embed the results:

- in well-engineered systems
- with large databases
- and with supportive interfaces

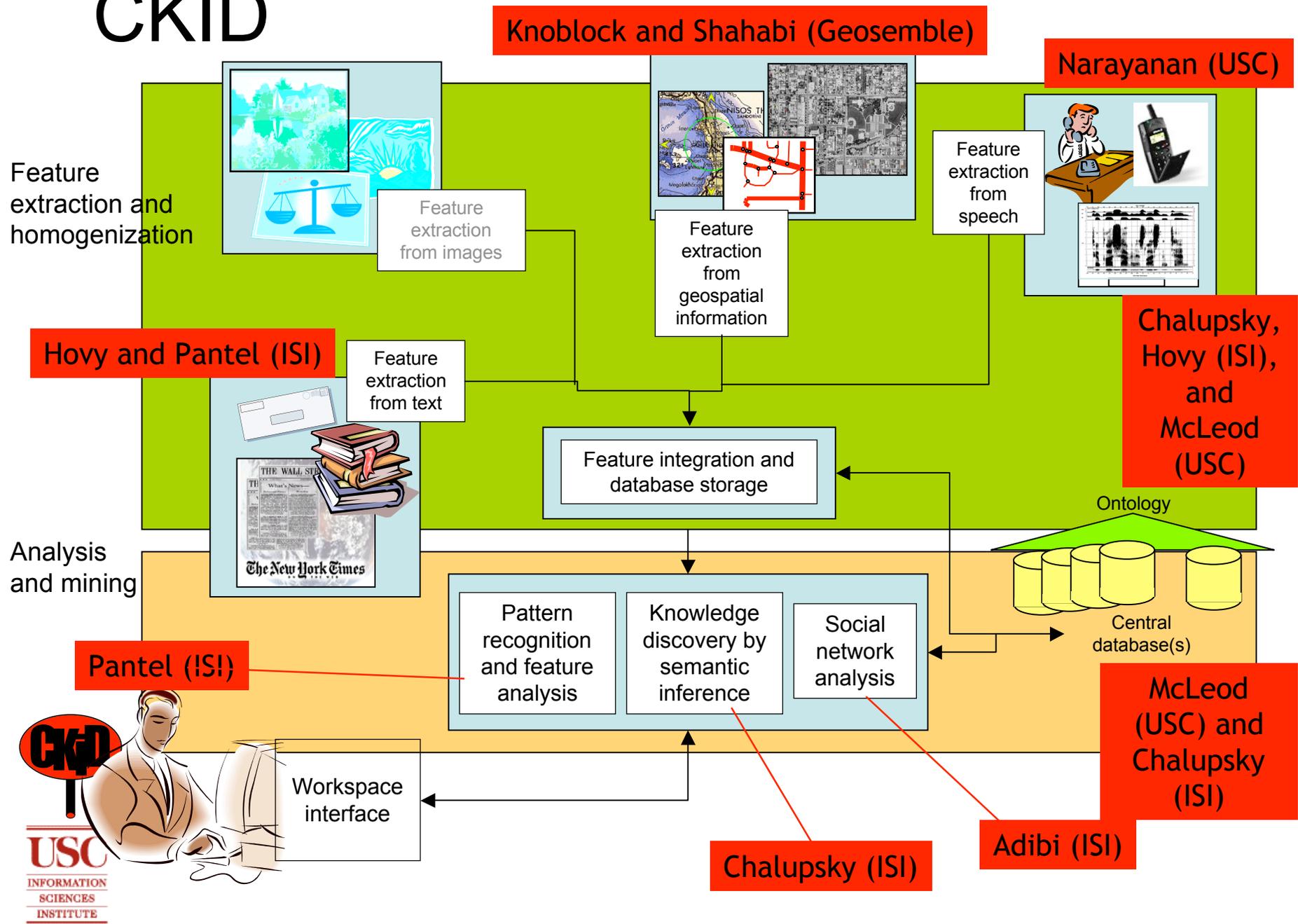


CKID Mission

- **Research:** Perform high-quality research in
 - information extraction and homogenization
 - data mining analysis and pattern learningof information in multiple media:
 - text, geospatial data, speech, social networks, etc.
- **Tech transition:** Collaborate closely with researchers in National Laboratories to tackle real problems and transition solutions into larger systems:
 - e.g., Los Alamos, PNNL, Lawrence Berkeley Labs
- **Education:**
 - train students (esp. minority students) in R&D
 - organize workshops, teach seminars, etc., in selected areas



CKID



Principal focus areas

- Program areas: **Information Management and Knowledge Discovery** and **Mathematical Foundations**
- Our strengths:
 - Automated analysis of large amounts of data in various media, with output in the form of semantic graphs
 - Data integration and data fusion at the semantic level, after (semantic) feature extraction
 - Scalable algorithms for IR and analysis on semantic graphs, especially for automated discovery of complex relationships between nodes
 - Models for detection and discovery on semantic graphs, using mathematically founded pattern recognition algorithms
 - Algorithms for unstructured text analysis and NLP for IE of complex concepts
 - Data validation and uncertainty quantification, through semantic knowledge-based inference and information theoretic measures



CKID Organization Structure

Center Coordinator
Other DHS Centers
National Labs
IMSC
CREATE



Eduard Hovy
Director and PI



Advisory
Board

Patrick Pantel

Co-PI and Deputy Director for ISI and
Geosemble

Dennis McLeod

Co-PI and Deputy Director
for USC

Patrick Pantel
Andrew Philpot
shared grad student
Pattern learning

Craig Knoblock
Cyrus Shahabi
Geospatial data

Dennis McLeod
Grad student
Databases

Eduard Hovy
Patrick Pantel
Andrew Philpot
Shared grad student
Text

Jafar Adibi
50% grad student
Social networks

Shri Narayanan
Grad student
Speech

Hans Chalupsky
Tom Russ
50% grad student
Knowledge-based inference

Andrew Philpot
System integrator

Contact info

Center for Knowledge Integration and Discovery

University of Southern California

www.ckid.org

Eduard Hovy

hovy@isi.edu

