Knowtator

A plug-in for creating training and evaluation data sets for Biomedical Natural Language Processing systems

Philip V. Ogren

Mayo Clinic College of Medicine

Entity Recognition

- Find mentions of concepts in text
  - Biological domain
    - Proteins (genes, mutations, complexes)
    - Cell components, cell types, etc.
  - Medical domain
    - Disorders (disease, injury, etc.)
    - Anatomies, drugs, signs & symptoms
- Normalize mentions to controlled vocabulary or database
  - e.g. Entrez, GO, SNOMED-CT, MeSH

Information Extraction

- Identify mentioned relationships between entities
  - Protein-protein interactions
  - Protein-disease interactions
  - Processes: regulation, proliferation, transport
  - Structured templates
    - E.g. for cancer - grade, stage, diagnosis, anatomy.

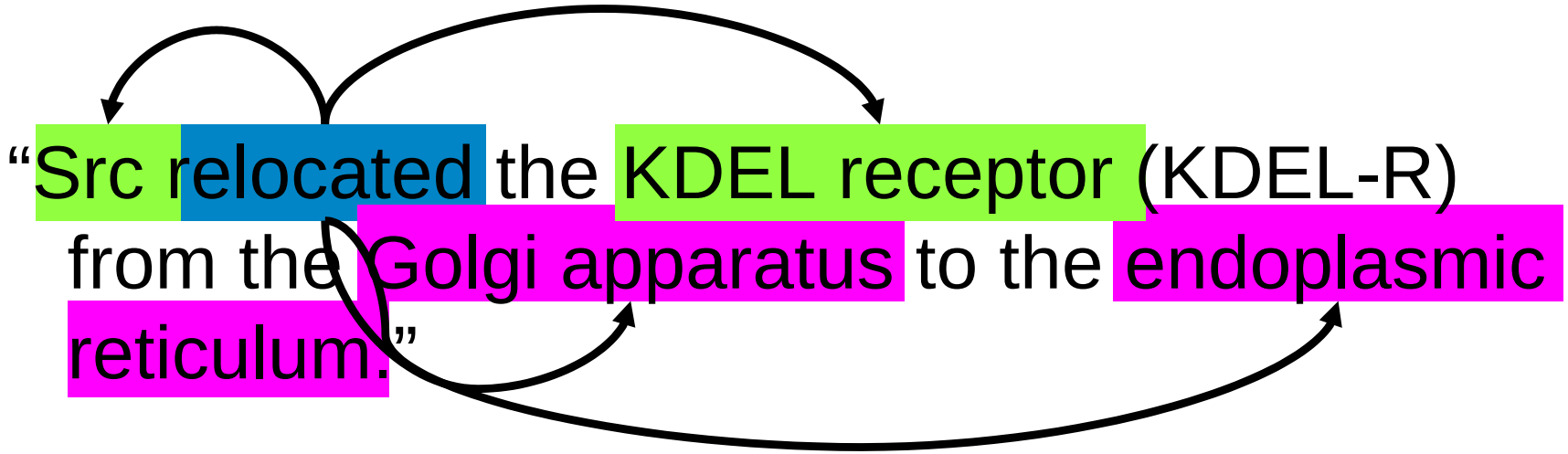
Molecular transport

"Src relocated the KDEL receptor (KDEL-R) from the Golgi apparatus to the endoplasmic reticulum."

Molecular transport

"Src relocated the KDEL receptor (KDEL-R) from the Golgi apparatus to the endoplasmic reticulum."

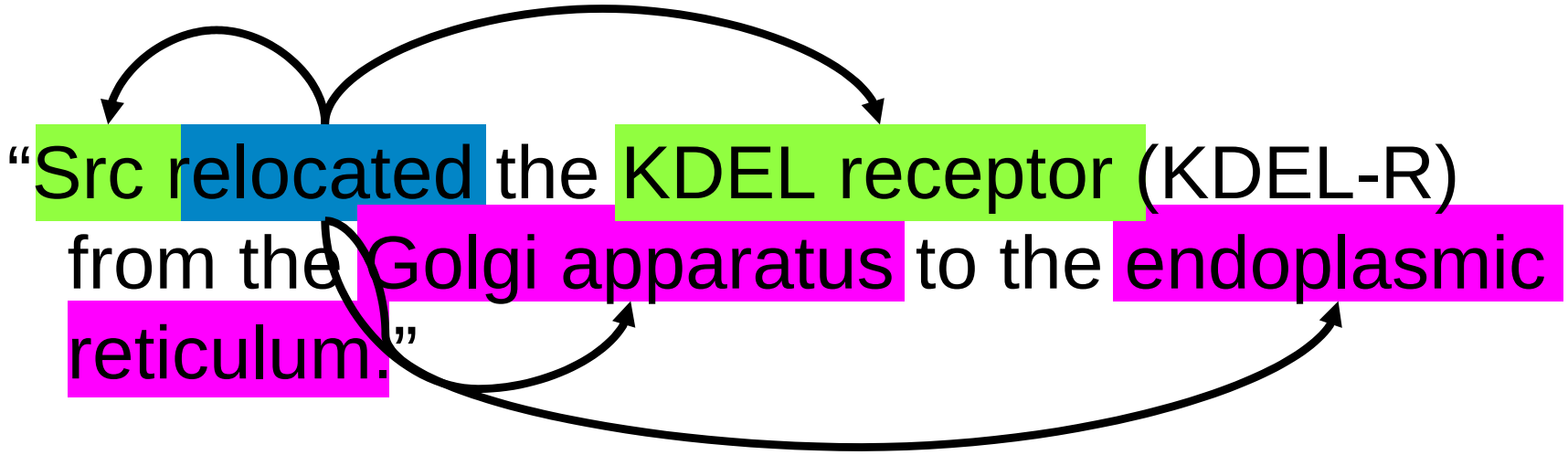
Molecular transport



Molecular transport frame

- Origin < cell component
- Destination < cell component
- Transported molecules < molecule
- Transporters < molecule

Molecular transport



transport event (predicate = relocated)

- origin = Golgi apparatus
- destination = endoplasmic reticulum
- transported molecule = KDEL receptor
- transporter = Src

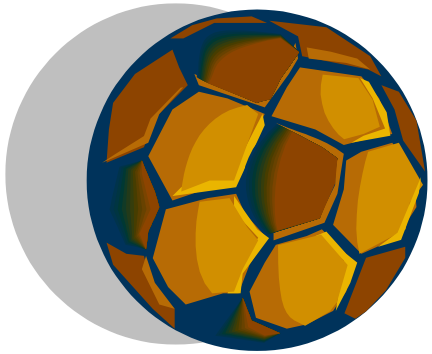
Now what?

- Go build your system
  - It's fun!
  - It's easy!
  - Yippie kai yeah!

  - ….unless, of course, you need training data

Then what?

- Evaluate your system
  - Not fun
  - Not easy
  - Time consuming



Evaluation

1. Give system output to domain expert
   - Easiest given limited resources and time
   - Not scalable, data not reusable, results not comparable
2. Create gold standard data for automatic comparison.
   - compare different systems
   - compare system versions
   - same data can be used for training
3. “Usefulness” evaluation
   - Feedback from user community

Creating a gold standard

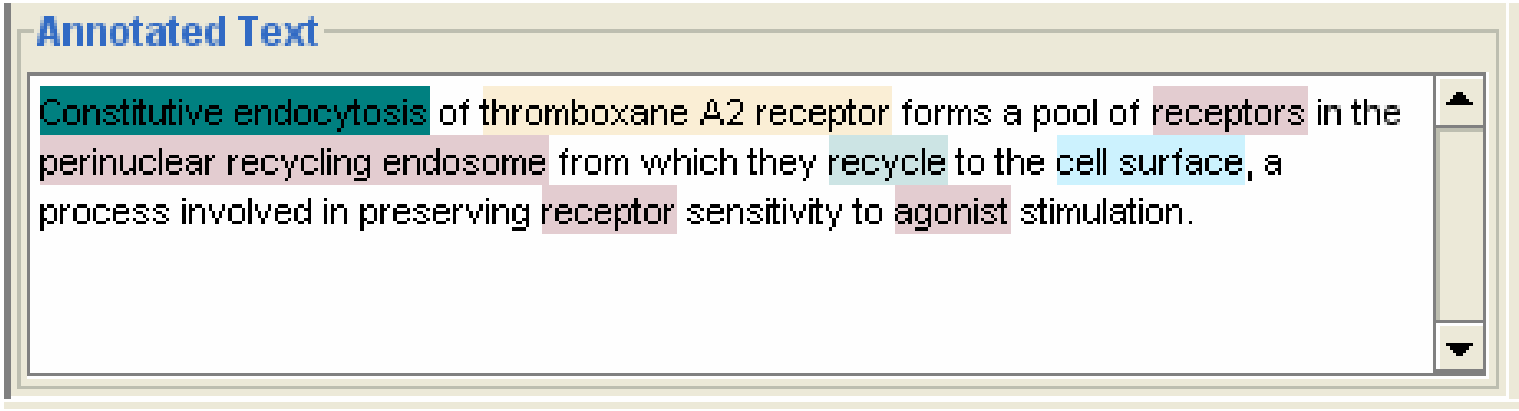
- humans
  - domain experts, knowledge engineer, software support, project manager
- **software**
  - representation of annotation schema
  - specialized data entry
- processes
  - workflow, guidelines, data management, evaluation

Software

- paper based (software!?)
- one-off approach (emacs macros)
- WordFreak
- Callisto
- GATE
- MMTx
- Freakégé
- Knowtator

Knowtator

- A general-purpose text annotation tool for creating gold-standard corpora

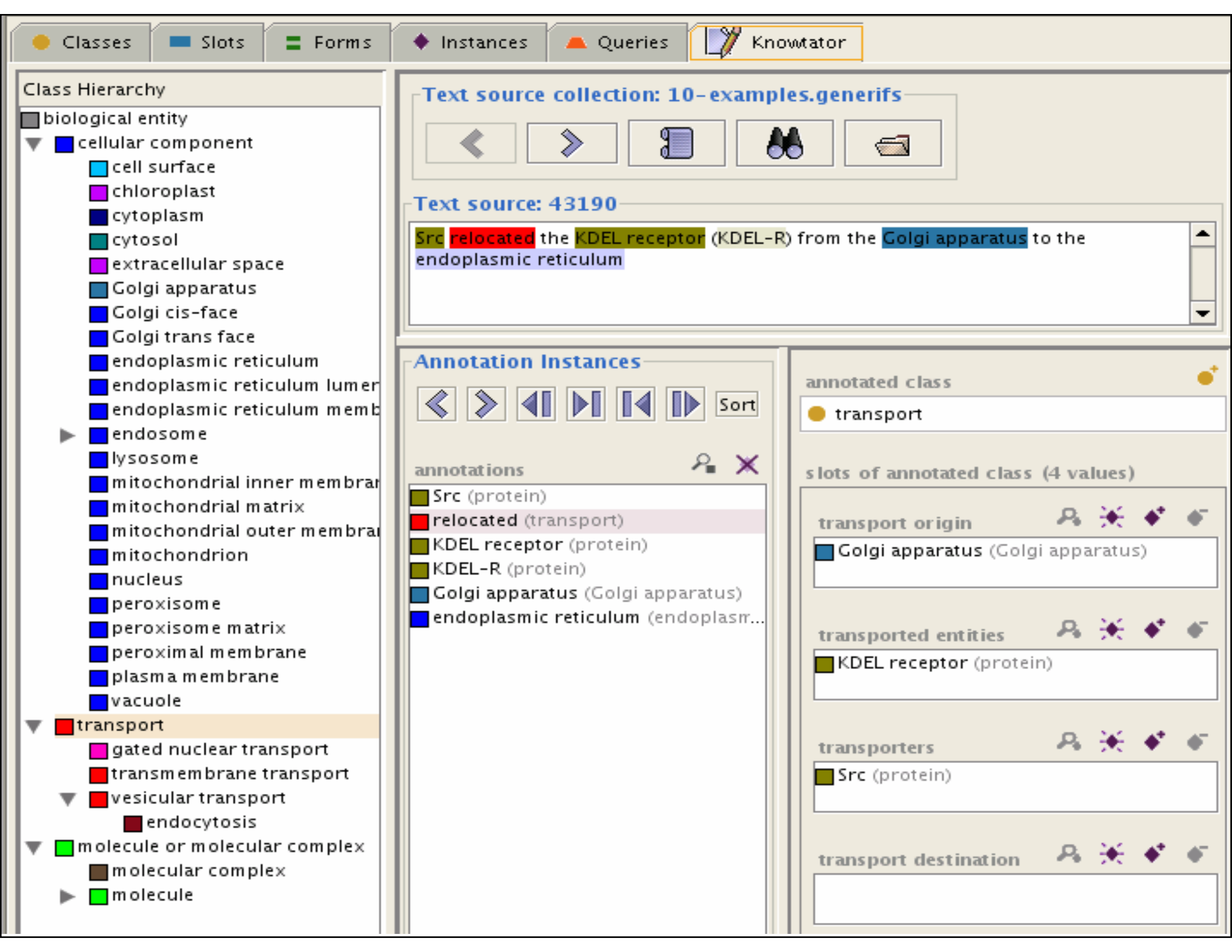


- A Protégé plug-in

- Open source (MPL):
  - **`bionlp.sourceforge.net/Knowtator`**
  - or google ‘Knowtator’

Knowtator

- Knowtator facilitates the *manual* creation of training and evaluation corpora for a variety of biomedical language processing tasks.
- Knowtator's key strength is the ability to define an annotation schema using a Protégé knowledge base.


Classes
Slots
Forms
Instances
Queries
Knowtator
Class Hierarchy
biological entity
cellular component
cell surface
chloroplast
cytoplasm
cytosol
extracellular space
Golgi apparatus
Golgi cis-face
Golgi trans face
endoplasmic reticulum
endoplasmic reticulum lumer
endoplasmic reticulum memb
endosome
lysosome
mitochondrial inner membrar
mitochondrial matrix
mitochondrial outer membrar
mitochondrion
nucleus
peroxisome
peroxisome matrix
peroximal membrane
plasma membrane
vacuole
transport
gated nuclear transport
transmembrane transport
vesicular transport
endocytosis
molecule or molecular complex
molecular complex
molecule
Text source collection: 10-examples.generifs
Text source: 43190
Src relocated the KDEL receptor (KDEL-R) from the Golgi apparatus to the endoplasmic reticulum
Annotation Instances
Sort
annotations
Src (protein)
relocated (transport)
KDEL receptor (protein)
KDEL-R (protein)
Golgi apparatus (Golgi apparatus)
endoplasmic reticulum (endoplasm...
annotated class
transport
slots of annotated class (4 values)
transport origin
Golgi apparatus (Golgi apparatus)
transported entities
KDEL receptor (protein)
transporters
Src (protein)
transport destination

Features

- Stand-off annotation
  - Original text is not modified
  - Exportable to simple XML
- Inter-annotator agreement metrics
- Consensus set creation mode
- Pluggable text source types (i.e. plain text files, xml, database, etc.)
- Annotation filters
- Annotation schema is defined by frames (class/instance/slot/facets) using Protégé.

Knowtator is *not*…

- A tool for building a repository of facts
  - annotating the semantic web
  - for creating a concept based index
  - for informing ontologies based on findings in the text
- Automated
  - Annotations can be pre-loaded
  - Semi-automated would be nice….
  - Introduces the problem of bias

Knowtator Knowledge Model

1. Target Ontology
2. Concept Mentions
3. Annotations

Target Ontology

- A set of class, instance, slot, and facet frames that define the set of named entities and relations that are the subject of the annotation task.
- Independent of any Knowtator specific classes

Concept Mentions

- a description of a concept that has been found in the target text.
  - What is the mentioned class?
  - What mentioned relationships exist?
  - What are the attributes of those mentioned classes?
- Provides a level of indirection from target ontology.

Concept mentions

- Class mention
  - mentioned-class (type=class)
  - Slot-mention (type=slot mention)
- Slot mention
  - Mentioned-slot (type = slot)
  - Mentioned-slot-value (type=class mention, string, etc.)

Annotations

- Mapping between text and concept mentions
- Book keeping information
  - Span offsets
  - Annotator
  - Creation date
  - Text source identifier
  - Concept mention

Knowtator Knowledge Model

- Clean separation between annotations/concept mentions and the target ontology.
  - A span of text mentioning a class is not an instance of that class
  - We can annotate mentions of instances
- Allows one to describe the concepts as they are seen – not as you have prescribed them to be.
  - “The lime was yellow”

End result

- A gold-standard data set that represents complete and accurate system output
- Different systems can be compared against the same gold-standard
  - Different versions of a system
- A resource useful for training with
  - Deriving rules
  - Training machine learning models

Acknowledgements

- UCHSC
  - Larry Hunter
  - Mike Bada
  - Andrew Dolbey
  - Kevin Cohen
  - Zhiyong Lu
- Mayo
  - Chris Chute
  - Guergana Savova
  - Serguei Pakhomov
  - Jim Buntrock