The Future of (Big) Data Curation – Harnessing Institutional, Community and Computational Power

Jane Hunter
<j.hunter@uq.edu.au>
Overview

- Big data needs a combined approach/effort:
  - Traditional metadata
  - Community-generated metadata/knowledge
  - Derived metadata - machine-learning and semantic reasoning

- 3 Case studies
  - sensor data streams;
  - textual data
  - 3D data

- Commonalities/issues
- Future research challenges
Big Data Curation

Volume
- Data at Rest
  - Terabytes to exabytes of existing data to process

Velocity
- Data in Motion
  - Streaming data, milliseconds to seconds to respond

Variety
- Data in Many Forms
  - Structured, unstructured, text, multimedia

Veracity*
- Data in Doubt
  - Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

Big Data Curation

Goals – high quality data/metadata; accurate fine-grained, semantic metadata; detailed provenance.

- Institutional metadata
  - Dublin Core – won’t scale, too flat, too coarse

- Current big data indexing
  - Too much data
  - Limited quality control
  - No semantics – can’t integrate or answer complex questions

- Machine-learning – fast feature extraction but requires training corpus

- Semantic inferencing, rules-based approach – only works in some domains

- Move towards community-driven data/knowledge curation
  - Quality control
  - Annotation – using more relevant terms - answers users queries
  - Define patterns, rules

- For a given domain, what is the optimum combination?
Case Studies

- Eco-informatics
  - SAAR – Semantic Annotation & Activity Recognition

- Bio-informatics
  - Skeletome

- Digital Humanities
  - 3DSA for Cultural Heritage
Curation of Animal Accelerometry Data

- Animal-attached accelerometers
  - monitor animal movement and behavior
- Tri-axial data streams
  - Large volumes of complex data
  - Lack of visualization, analysis tools and share-ability
  - Lack of pattern identification and tagging services
- Free-ranging wild animal behavior
  - Lack of ground truth data

Endangered Species  Feral Pests  Production livestock
User Driven Requirements

**Step 1**
Upload data

**Step 2**
Activity recognition
running, walking, resting, walking, running, feeding, walking

**Step 3** – Analysis and Visualization

Walking, Running, Resting, Sleeping, Feeding

Animal Health
Energy Consumption
Food/Water Requirements
Objectives

- Web-based semantic annotation and activity recognition system to enable biologists to
  - Share tri-axial accelerometer data
  - Visualize and analyze tri-axial accelerometer data
  - Share expert knowledge
  - Help scientists understand the movement and behavior of animals
  - Use surrogates/domestic animals (& video) to train classifier → automatically tag rare, wild, feral animals
User Interface - Tagging

Screenshot of SAAR Plot-Video interface and the annotation interface
User Interface – Automated Results

Screenshot of the SAAR Interface with human activity identification results
Evaluation

- Tested on range of species (different sizes and gaits)
  - Australian dingo (Canis lupus dingo)
  - Eurasian badger (Meles meles)
  - Bengal tiger (Panthera tigris tigris)
  - African cheetah (Acinonyx jubatus),
  - American alligator (Alligator mississippiensis)
  - hairy-nosed wombat (Lasiorhinus krefftii)
  - Eastern Grey kangaroo (Macropus giganteus)
  - short-beaked echidna (Tachyglossus aculeatus)
- Running, Walking, Standing, Sitting, Lying (Sternal recumbency)
- Test dog classification module on range of species
Results

High scores (80-90%) if SL:SH = (2-3) (Spine length: Spine height)
Benefits

- Subscribers login to online service
- Leverage community expertise to tag training sets
- Develop libraries of classifiers for different species
- Apply domestic species classifiers to wild species
  - Dogs -> dingos, foxes; birds -> bats; horses -> camels
- Classifiers – improve over time as more data uploaded

Socio-economic and health benefits:

- livestock productivity – assess health, energy/food needs
- reduce spread of feral pests & viruses
- management & conservation of threatened species
OzTrack

- Combine GPS and Accelerometry Data - Camel Tracks on Vegetation Map
- Real time animations – showing where it is and what it's doing
Skeletome

- A community-driven knowledge curation platform for Skeletal Dysplasias
  - Rare diseases
  - Affect the development of Human Skeleton
  - Caused by genetic abnormalities
  - Complex medical issues
- Capture, integrate, correlate and analyse clinical, radiographic, phenotypic and genetic data
Verne Troyer
Cartilage-Hair Hypoplasia

RMRP

Peter Dinklage
Achondroplasia (MOST COMMON DISORDER)

FGFR3

Danny Devito
Multiple Epiphyseal Dysplasia (MED)

COL9A2, COL9A3, COMP, MATR3
Challenges

- Hundreds of different types
  - 440 types in 40 groups
- Difficult to diagnose, treat
- Few medical publications
- Doctors rely on:
  - Existing patient data
  - Expert knowledge
Community Needs

- Common terminology
- Data Integration
- Data Quality Control
- Knowledge Extraction and Sharing
- Privacy of Data
- Expertise sharing
The Platform

Knowledge Base

Bone Dysplasia Ontology

Patient Archive

Reasoning
Bone Dysplasia Ontology

ISDS Nosology

<table>
<thead>
<tr>
<th>Group/name of disorder</th>
<th>Inheritance</th>
<th>MIM No.</th>
<th>Locus</th>
<th>Gene</th>
<th>Protein</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>FGFR3 chondrodysplasia group</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thanatophoric dysplasia type 1 [TD1]</td>
<td>AD</td>
<td>187600</td>
<td>4p16.3</td>
<td>FGFR3</td>
<td>FGFR3</td>
<td>Includes previous San Diego type</td>
</tr>
<tr>
<td>Thanatophoric dysplasia type 2 [TD2]</td>
<td>AD</td>
<td>187601</td>
<td>4p16.3</td>
<td>FGFR3</td>
<td>FGFR3</td>
<td></td>
</tr>
<tr>
<td>Severe achondroplasia with developmental delay and acanthosis nigricans [SADDAN]</td>
<td>AD</td>
<td>See 187600</td>
<td>4p16.3</td>
<td>FGFR3</td>
<td>FGFR3</td>
<td></td>
</tr>
<tr>
<td>Achondroplasia</td>
<td>AD</td>
<td>100800</td>
<td>4p16.3</td>
<td>FGFR3</td>
<td>FGFR3</td>
<td>Inactivating mutation</td>
</tr>
<tr>
<td>Hypochondroplasia</td>
<td>AD</td>
<td>145000</td>
<td>4p16.3</td>
<td>FGFR3</td>
<td>FGFR3</td>
<td></td>
</tr>
<tr>
<td>Camptodactyly, tall stature, and hearing loss syndrome [CATSHL]</td>
<td>AD</td>
<td>187600</td>
<td>4p16.3</td>
<td>FGFR3</td>
<td>FGFR3</td>
<td>Similar to hypochondroplasia but unlinked to FGFR3, probably heterogeneous; uncertain diagnostic criteria</td>
</tr>
<tr>
<td>Hypochondroplasia-like dysplasia[s]</td>
<td>AD, SP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

See also group 33 for craniosynostoses syndromes linked to FGFR3 mutations, as well as LADD syndrome in group 39 for another FGFR3-related phenotype
Application

The Skeletome Knowledge Base
A community-driven knowledge curation platform for skeletal dysplasias.

Comprehensive
Everything You Ever Wanted to Know About Bone Dysplasias
The Skeletome knowledge base provides information on all bone dysplasias recognised by the International Skeletal Dysplasia Society.

Community driven
Continuously Updated by the Global Bone Dysplasia Community
All entries are continuously reviewed and updated by the global community of clinicians and researchers working on bone dysplasias.

Ontology Based
Readable by Humans and Computers - the Best of Both Worlds
The Skeletome knowledge base makes extensive use of ontologies to standardise the entered information and make it accessible to computational analysis.

- Registered doctors upload patient cases
- ISDS experts assist with diagnosis
- Automatic processing and experts’ editorial process
- Incremental knowledge evolution
- http://knowledge.skeletome.org/
Patient Sharing

Anonymized patient cases shared with multiple doctors
- X-rays
- Clinical Summaries
- Genetic reports

Dr William Shatner
Shared #56323 with European Bone Dysplasia Network and American Bone Dysplasia Group - 8hr ago

Patient #56232

Clinical Summary
Cleft palate, shortening of upper and lower limbs, small chest, large abdomen, respiratory insufficiency after birth, pulmonary hypoplasia, polyhydramnion in pregnancy. FoA. birth length 40cm, OFC 33cm, chest circumference 28cm. Deceased after 1 month of life. Developed a large head.

Show more

X-Rays

Show more

Genetic Reports

<table>
<thead>
<tr>
<th>Gene Mutation</th>
<th>Gene</th>
<th>Pathogenicity</th>
</tr>
</thead>
<tbody>
<tr>
<td>G&gt;C.36</td>
<td>FGFR3</td>
<td>High</td>
</tr>
<tr>
<td>A&gt;T.22</td>
<td>CUL7</td>
<td>Low</td>
</tr>
</tbody>
</table>

Show more

European Bone Dysplasia Network
American Bone Dysplasia Group

5 comments
Discussing a Patient

Inline commenting

Text posts with PubMed Integration

Community Diagnoses
Discussing a Patient

Inline commenting

Text posts with PubMed Integration

Community Diagnoses
Discussing a Patient

Inline commenting

Text posts with PubMed Integration

Community Diagnoses

These papers mention a similar phenotype to this patient.

Decomposing phenotype descriptions for the human skeletal phenome.

Groza T, Hunter J, Zankl A.

Over the course of the last few years there has been a significant amount of research performed on ontology-based formalization of phenotype descriptions. This not only increases value and knowledge captured within such databases, but also allows for the integration of clinical and research data.

Toward knowledge support for complex traits.

Collier N, Oellrich A, Groza T.

The systematic description of complex level, is important for hypothesis-driven interpretation of complex traits.
Discussing a Patient

Inline commenting

Text posts with PubMed Integration

Community Diagnoses
Discussing a Patient

Inline commenting

Text posts with PubMed Integration

Community Diagnoses
Diagnosis Extraction

Entity Term Extraction

Phenotype Extraction

Diagnosis Extraction

Dr George Takei

Added a diagnosis - 4hrs ago

Pseudoachondroplasia

Disagree (5)

William Shatner 4hrs ago

I also think it looks like Achondroplasia.
This patient has the symptom of Short Fingers which is exclusively shown with Achondroplasia and not present in Psuedoachondroplasia.
Diagnosis Extraction

Phenotype Extraction

Entity Term Extraction

“I also think it looks like Achondroplasia.”

Dr George Takei
Added a diagnosis - 4hrs ago

Pseudoachondroplasia

Agree (3)
Disagree (5)

William Shatner 4hrs ago
I think its more likely Achondroplasia. This patient has the symptom of Short Fingers which usually shown with Achondroplasia and not present in Pseudoachondroplasia.
Phenotype Extraction

Entity Term Extraction

Diagnosis Extraction
Phenotype Extraction

Entity Term Extraction

Diagnosis Extraction

Dr George Takei
Added a diagnosis 4hrs ago

Pseudoachondroplasia

 Agree (3)  Disagree (5)

William 4hrs ago
I am more likely Achondroplasia. The patient has the symptom of Scoliosis and Short Fingers which is exclusively present in Psuedoachondroplasia.
Entity Term Extraction
Phenotype Extraction
Collaborative Diagnosis
Diagnosis Extraction
Reasoning

• Analyze Diagnoses, Phenotypes, Genotypes
  - Across Patient Cases and Publications

• Extract/infer new relationships
  - Disease <-> phenotypes <-> genotypes
  - certainty, temporality, severity, polarity
  - Incorporate reliability of source

• Added to Bone Dysplasia Ontology

• Generate ranked list of diagnoses for new cases
Knowledge Base of Disorders
Written Abstracts
Linked Genes
ISDS Grouping
X-Rays
Phenotypes
Inline Editing
Achondroplasia

Abstract

Achondroplasia is characterized by abnormal bone growth, resulting in short stature with disproportionately short arms and legs, a large head, and characteristic facial features with frontal bossing and mid-face hypoplasia. In infancy, hypotonia is typical, and acquisition of developmental motor milestones is often delayed. Intelligence and life span are usually normal, although compression of the spinal cord and/or upper airway obstruction increases the risk of death in infancy.

X-Rays

There are no x-rays for 'Achondroplasia'.

Clinical Features

There are no statements.
Disease characteristics. Achondroplasia is a bone growth disorder that results in short stature, short arms and legs, a large head, and characteristic facial features such as a prominent forehead and bossing and mid-face hypoplasia. In some, there is an increased risk of scoliosis. Intelligence and life span are usually normal.

X-Rays

There are no x-rays for 'Achondroplasia'.

Clinical Features (9)

Search for a Clinical Feature
Achondroplasia

Abstract

Disease characteristics. Achondroplasia is characterized by bone growth that results in short stature with small arms and legs, a large head, and characteristic 'pencil in circle' bony bossing and mid-face hypoplasia. In infancy, rapid acquisition of developmental motor milestones, intelligence and lifespan are usually normal, although the spinal cord and/or upper airway obstruction can result in death in infancy.

Statements (0)
No statements.

X-Rays
There are no x-rays for 'Achondroplasia'.

Clinical Features (9)

Mode of Inheritance AD

Genes

FGFR3

Group Members

Hypochondroplasia
Hypochondroplasia-like dysplasia
Severe achondroplasia-developmental delay-acanthosis nigricans
Thanatophoric dysplasia type 1
Thanatophoric dysplasia San Diego type
Thanatophoric dysplasia type 2
Achondroplasia

Abstract

Disease characteristics. Achondroplasia is characterized by abnormal bone growth that results in short stature with disproportionately short arms and legs, a large head, and characteristic facial features with frontal bossing and mid-face hypoplasia. In infancy, hypotonia is typical, and acquisition of developmental motor milestones is often delayed. Intelligence and life span are usually normal, although compression of the spinal cord and/or upper airway obstruction increases the risk of death in infancy.

Group Members

- Achondroplasia
- Camptodactyly tall stature and hearing loss syndrome
- Hypochondroplasia
- Hypochondroplasia-like dysplasia
- Thoatophoric dysplasia type 1
- Thanatophoric dysplasia San Diego type
- Thanatophoric dysplasia type 2
- Thanatophoric dysplasia type 3
- Thanatophoric dysplasia type 4
- Thanatophoric dysplasia type 5

X-Rays

There are no x-rays for 'Achondroplasia'.

Clinical Features (9)

- Search for a Clinical Feature

<table>
<thead>
<tr>
<th>Feature</th>
<th>Information Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Knowledge Base Disorders
Written Abstracts
Linked Genes
ISDS Grouping
X-Rays
Phenotypes
Inline Editing
Knowledge Base Disorders
Written Abstracts
Linked Genes
ISDS Grouping
X-Rays
Phenotypes
Inline Editing
Achondroplasia

Distribution of phenotypes over the current number of patients (15)

This chart shows the percentage of patients that exhibit certain phenotypes, over the entire range of phenotypes presented by all patients.

Distribution of bone dysplasias with correlated phenotypes

Details

Group: FGFR3 Chondrodysplasia Group
OMIM: 108900
Mode of inheritance:
  - Autosomal dominant inheritance

Genes

FGFR3

Bone dysplasias in the group

- Achondroplasia
- Camptomelic tall stature and hearing loss
- Hypochondroplasia
- Hypochondroplasia-like dysplasia
- Severe achondroplasia-developmental delay
- Thanatophoric dysplasia type 1
- Thanatophoric dysplasia San Diego type
- Thanatophoric dysplasia type 2

USER LOGIN

USERNAME *

PASSWORD *

- Create new account
- Request new password
Telederm

- Corpus of >1500 cases
- Analyze patient cases & discussions
- Automatically tag text & images
  - Patient details
  - Clinical features
  - Phenotypes
  - Diagnosis
- Aggregate to build KB
  - Curated use cases
  - Association of diseases to phenotypes
  - Search interfaces
- Training
- Differential Diagnoses
Semi-automatic curation of 3D cultural heritage artefacts
Objectives

• Combine low level crowd-sourced semantic annotations with automatically extracted features (dimensions, shape, colour)
• To automatically infer high level tags
• Catalogue/classify the artefact

• Expedite cataloguing of 3D artefacts
• Fast/simple authentication
• eLearning tool for anthropology students
Case Study

- Greek Vases from UQ Antiquities Museum
- 3DSA - Web-based annotation tools for 3D digital objects
  - Point, region and 3D segments
  - Interoperable across clients and 3D formats
  - Based on OA data model

- Use 3DSA to label parts
Example Rules

- Neck Amphora - Oval body, a offset neck with a thick mouth, two vertical handles and a heavy stand

- Stamnos - Round body, a offset short neck with a wide mouth, two loop handles and a heavy stand

- Lekythos - Tall cylindrical body with a offset shoulder, a tall neck with heavy mouth, 1 vertical handle and heavy stand.
Methodology

- Define SWRL rules to infer high-level classification from low-level
  - ObjectA is tagged with: “disk-shaped mouth”, “round body”, “broad handle” and is decorated with “black-figured palaestra scene” → ObjectA is a “Corinthian aryballos”

- Test-users tag a set of vase parts
- Apply inferencing rules -> high level tags
- Usability studies – feedback and refinement
- Evaluations
  - Speed, Precision/recall
  - improvements to search/retrieval
Konica Minolta Vivid 9i Laser Scanner

Automated 360° Scanning
- portable, fast, high-res (50 microns = .002 ins)
- Scanning range - 0.5 - 2.5 m
System Architecture

- 3DSA Portal UI (jQuery and HTML5)
- 3DSA Annotation Client (jQuery, HTML5 and WebGL)
- 3DSA Link Module
- 3DSA Portal Backend (Drupal 7, MySQL, PHP)
- 3D Object Repository (File system)
- Knowledge Base (Open-RDF Sesame)
  - Annotations
  - Museum Metadata
  - GVO ontology
- Reasoning Service (Euler YAP Engine)
Creating/Linking Annotations

- Annotate single 3D object
- Annotate multiple 3D objects
- Semantic Reasoning
- Measurements

[Image of a 3D vase with annotated details]

Textual Document:
Inv. No.: 65.001
Provenance: Attica
Date: c. 500 BC
Height: 151mm
Diameter: 56mm

This small pottery lekythos was painted by the Diophros Painter in the black-figure style at a time when red-figure was the dominant style of vase painting. It has an echinus-shaped mouth with a lip that is square in profile. The narrow neck joins on to an almost flat shoulder, while the strap handle passes from the base of the neck up and down to the outer edge of the shoulder. The body tapers sharply towards the foot which is a flat disc. The main
Vase segment tagged as “black cylindrical handle”
Example N3 Logic Rule

{?Greek_Pottery a gvo:Greek_Pottery; gvo:hasPart ?b, ?h1, ?h2, ?m, ?f;
gvo:hasMeasurement [gvo:has_height_value[math:lessThan 300]].
?Mouth1 a gvo:Mouth gvo:hasCharacteristic gvo:wide.
?Foot1 a gvo:Foot gvo:hasCharacteristic gvo:thin.
} => {?Greek_Pottery gvo:hasShape gvo:Skyphos_Type_B.}
Different Weightings for Features

gvo:Skyphos_type_B a gvo:Greek_Pottery_Shape.
?g gvo:hasPart [a gvo:Body; gvo:hasCharacteristic gvo:small].
=> (?g gvo:Skyphos_type_B rule1) ex:giveWeight 0.3.
?g gvo:hasPart [a gvo:Body; gvo:hasCharacteristic gvo:deep].
=> (?g gvo:Skyphos_type_B rule2) ex:giveWeight 0.7.
?g gvo:hasPart [a gvo:Body; gvo:hasCharacteristic gvo:bowl-shape].
=> (?g gvo:Skyphos_type_B rule3) ex:giveWeight 1.0.
?g gvo:hasPart [a gvo:Handle; gvo:attachedAlong gvo:horizontal].
=> (?g gvo:Skyphos_type_B rule4) ex:giveWeight 0.5.
?g gvo:hasPart [a gvo:Handle; gvo:attachedAlong gvo:vertical].
=> (?g gvo:Skyphos_type_B rule5) ex:giveWeight 0.5.
?g gvo:hasPart [a gvo:Mouth; gvo:hasCharacteristic gvo:wide].
=> (?g gvo:Skyphos_type_B rule6) ex:giveWeight 1.0.
?g gvo:hasPart [a gvo:Foot; gvo:hasCharacteristic gvo:thin].
⇒ (?g gvo:Skyphos_type_B rule7) ex:giveWeight 1.0. ?g gvo:hasMeasurement [gvo:hasHeightValue[math:lessThan 300]].
⇒ => (?g gvo:Skyphos_type_B rule8) ex:giveWeight 1.0.
Part: **Body**
Size: Small, deep
Shape: **Bowl**

Part: **Handle**
Axis: **Vertical**

Part: **Handle**
Axis: **Horizontal**

Part: **Foot**
Size: **Thin**

Is a: **Skyphos Type B** (sameAs: Glaux Skyphos, subClassOf: Skyphos, cup)
Annotating Relationships between Parts
Search on:
• features - vases with a **horizontal handle** & **wide mouth**
• Classes of vase: Skyphos, Hydria, Amphora, Pelike etc…
• associations/relations - all objects “**similar_to**” this object
• annotations within this 3D region
Community-driven Problem/Curation

Community of Users
- Need Help to Understand Data
- Lack of Integrated, Reliable Information

Knowledge Base

Domain Ontology

Domain Experts
- Refine ontology
- Tag training corpus
- Define rules

Annotated Archive

Reasoning
Big Data Curation Research Challenges

- **Real-time Curation/Indexing**
  - Processing of multiple synchronous data streams using MapReduce
  - Streaming RDF
  - Streaming Query Languages

- **Continual learning over big data** - feature extraction from massive, changing training datasets

- **Semantic markup of models and reasoning over predictive models**

- **Probabilistic inferencing**
  - Confidence, uncertainty measures

- **Citation of sub-sets of community-generated dynamic data**
Acknowledgement – eResearch Lab
Questions

- Jane Hunter <j.hunter@uq.edu.au>
- eResearch Lab @ University of Queensland