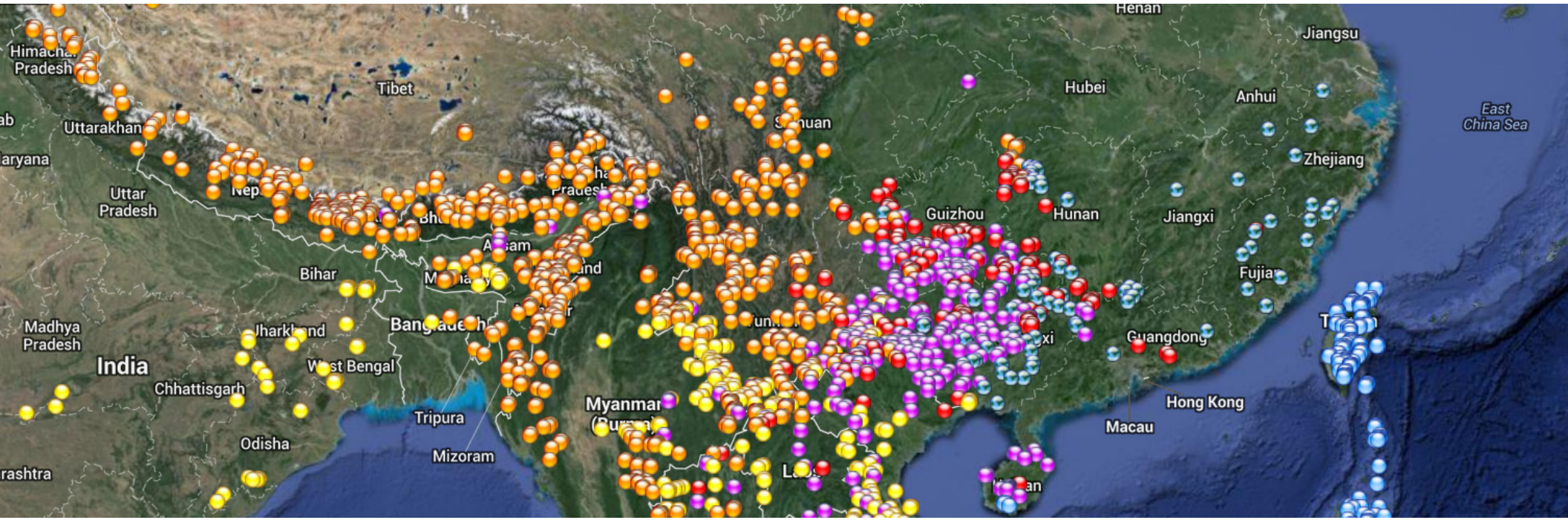
Bronze, Gold, and the Asia-Pacific Data Warehouse



Doug Cooper

Center for Research in Computational Linguistics

sealang.net doug.cooper.thailand@gmail.com

↓

Today: building a research community . . .

1. look back: a lesson in **STECese**
   (*data and the NLP community*)
2. look ahead: **A-P Data Warehouse**
   (*what we're working on*)
3. smoothing the path: **Operations on Lexicons**
   (*what the community could be working on*)

. . . inside and outside linguistics

1. look back

Baltimore June 27, 2014

ComputEL

The use of computational methods in the study of endangered languages

52nd Annual Meeting of the Association for Computational Linguistics

Baltimore June 28, 2014

On the day following the workshop (6/27), there will be a closed **meeting** where a mix of plenary and breakout group sessions will consider **how work on** endangered, and other less commonly studied, **languages can** more effectively **exploit and inform** methods developed in the context of **computational linguistics.**

In addition to computational and endangered languages linguists, expect representatives from the U.S. National Endowment for the Humanities and National Science Foundation to be at this meeting, as well as individuals from other U.S. federal agencies with an interest in language resources. One of the goals of this meeting will be to help find new fundable projects at the intersection of computational linguistics and endangered languages research.

... one theme

Yangon May 28, 2014

SEALS24

Bronze, gold, and the Asia-Pacific Data Warehouse

24th Annual Meeting of the Southeast Asian Linguistics Society

Yangon May 28, 2014

On the day following the workshop (6/27), there will be an open **meeting** where a mix of plenary and breakout group sessions will consider **how work on** endangered, and other less commonly studied, **languages can** more effectively **exploit and inform** methods developed in the context of **computational linguistics.**

This seems like a perfect meeting of interests, but ...

↓

let’s find some data, and decide what the problems are

↑ the computer scientists’ choices aren’t always the *best* choices ↓

let’s talk to linguists about **real** problems and **good** data

let’s find some programs, and use them on our data

the linguists’ process ↑
↓ isn’t always ideal, either

let’s talk to CS folks about **real** problems and **good** data

the NLP community has been down this road

NLP was once all about

consider it solved!

what are we waiting for?

Reality gradually settled in

X is the answer!

now, what was the question?

Over time, researchers asked

what do we want to solve?

do we really understand the question?

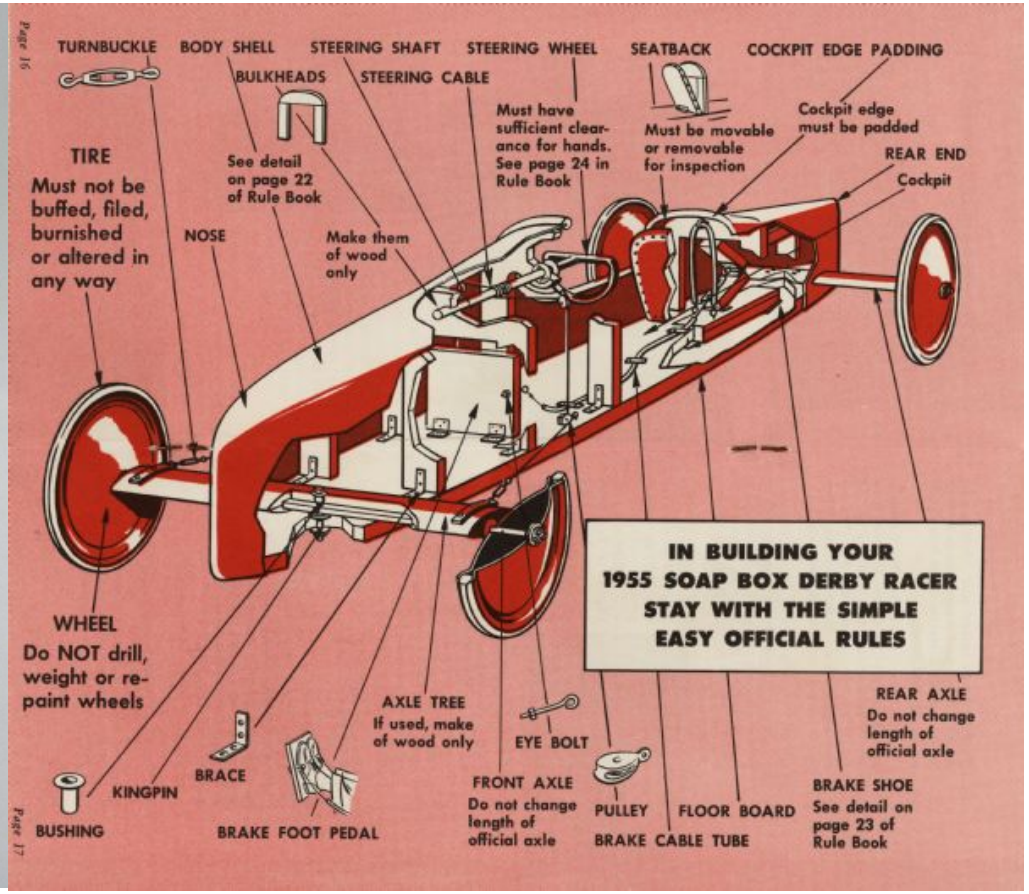
Over time, funders asked

how do we know you've solved it?

how do you plan to evaluate that?

↓

build a level playing field and they will come ...



STEC

Shared task evaluation challenge

STEC goals

how do we get all sides talking?

where should the field be headed?

what should we be trying to solve?

what do we need to start solving it?

how do we know if we have solved it?

Open challenges to the field.

Open challenges to **build** the field.

all based on **gold standard data and metrics**

whence the term **gold-standard?**

it's what scientists agree to use as
common currency
when there is no formal ground truth

it's the result of a **social process**
it's a **moving target**

its not really in the **comfort zone**
of comparative linguistics

STEC message

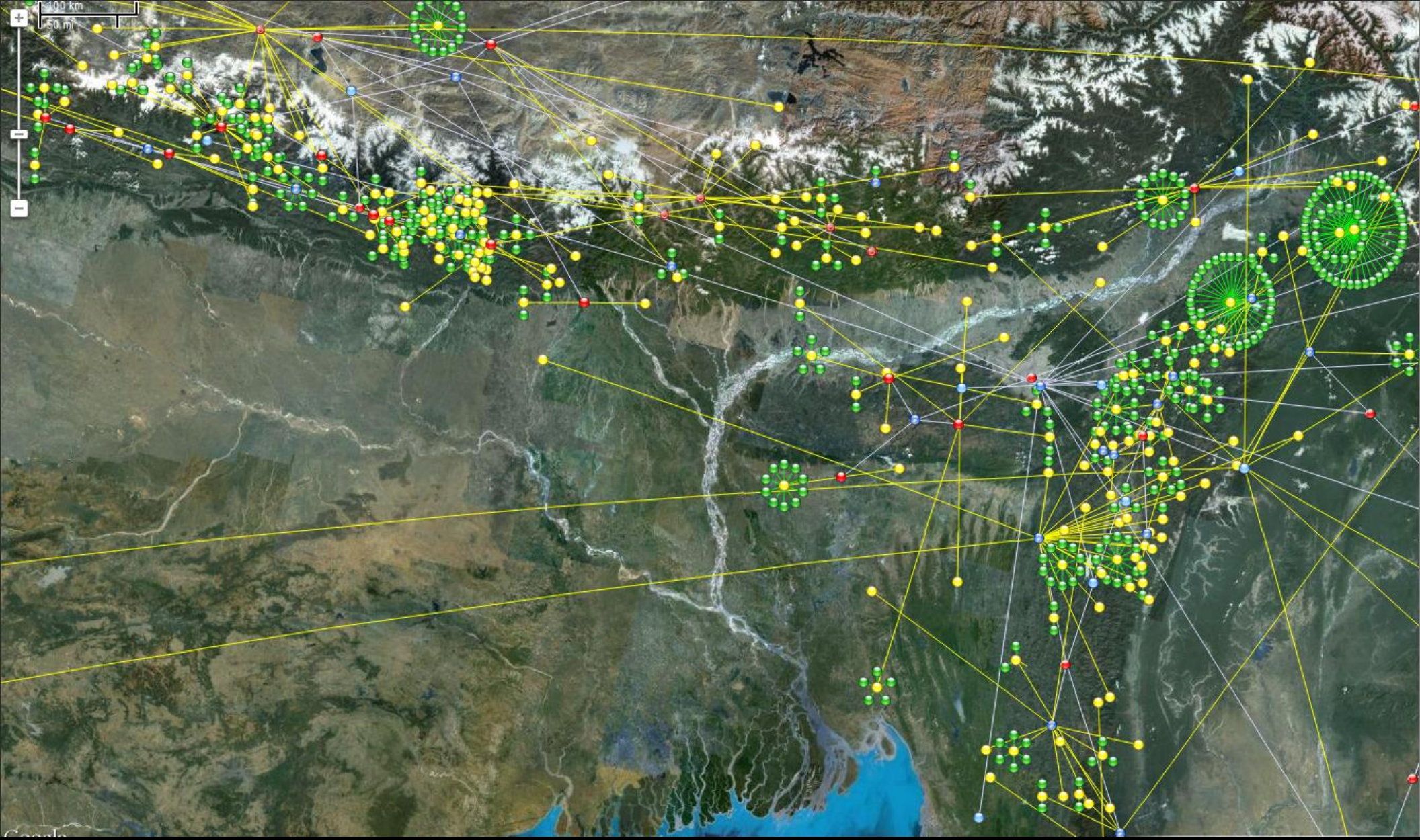
define goals
provide data
develop metrics

evaluate progress
reassess goals
lather, rinse, repeat

enormously successful: TIPSTER, MUC, SUMMUC, TREC, SENSEVAL, SEMEVAL ...

still dominates NLP / HLT / Comp Ling landscape: BioNLP, CoNLL, NAACL-HLT, COLING

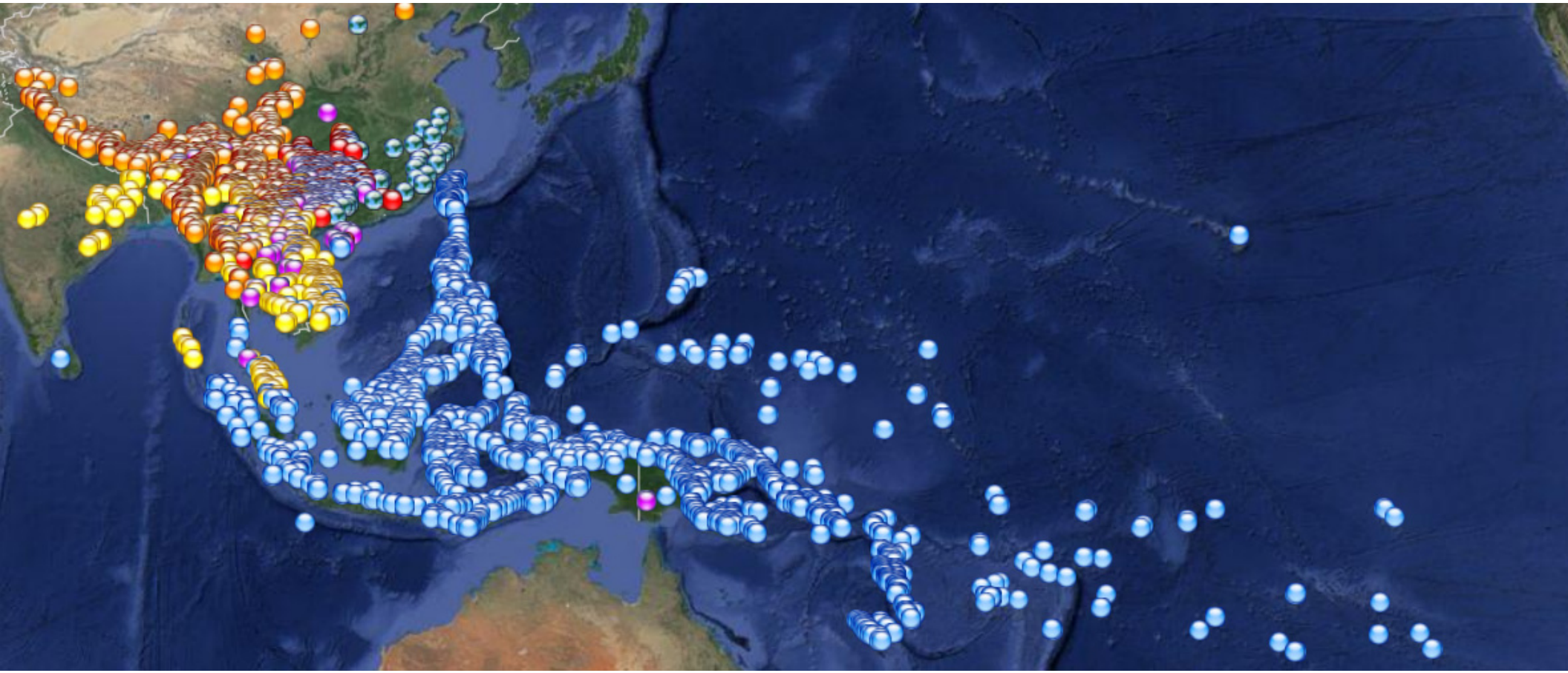
↓



2. look ahead

Asia-Pacific Data Warehouse

AA, AN, HM, KD, ST / ~2,000 “languages”



warehouse = **data + tools**

filter | frame | analyze | visualize | recycle

analysis is part of every **query**

Scope:

2,500 items (but fewer if need be)

comparative and/or **survey** lexicons

all **ISO 639-3** (et al) lects (but **dialects** are ok)

phone sketches as available

Sources:

print, gray, pencil, electronic publications

DOI naming of **all sets** (via DataCite / EZID)

direct linking to **sources** and **data** via DOI

status

vapor	we've heard of it, but we haven't seen it
water	audio only, not transcribed
paper	have paper or pdf, but not transcribed or extracted
tin	dictionary e-data: **orthography** and **definitions**
copper	comparative / survey e-data: **forms** and **glosses**
bronze	naive normalization forms/glosses, some **cognate sets**
silver	normalized / grouped by machine – not human-verified
gold	human-verified, machine-usable **comparable datasets**

Our goals:

not just bigger – **better**
help turn **bronze** into **gold**

encourage and enable development of **tools**

improve **data** upstream
improve **software** downstream

3. smoothing the path

Our means: real **problems**
good **data**
speak **STECese**

what's the **operation?**
what's the **input?**
what's the **output?**
where's the **data?**
what's the **metric?**

operations on lexicons

trying to **frame questions**
clarify **data requirements**
establish **metrics**

1 Operations on audio data
2 Data audit
3 Evaluation metrics
4 Operations on phonological strings / lists of strings
5 Operations on glosses
6 Operations on form+gloss items, lists, and vectors
7 Operations on cognate sets (EtySets)
8 Operations on semantic and phonological queries
9 Distance and clustering
10 Visualization
11 Geographic / demographic operations
12 Reconstruction
13 Subgrouping
14 Dataset format conversion
15 Statistical operations

what are your **operations?**

gold standards?

metrics?

Additions / comments / corrections to

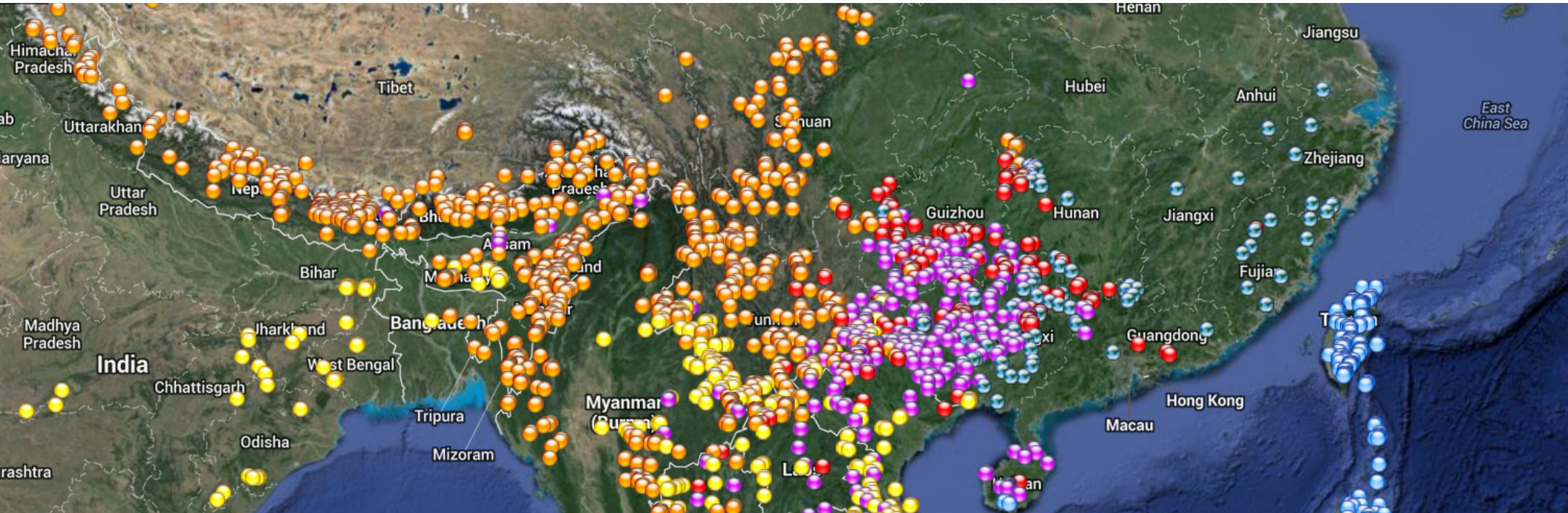
doug.cooper.thailand@gmail.com

Summary:

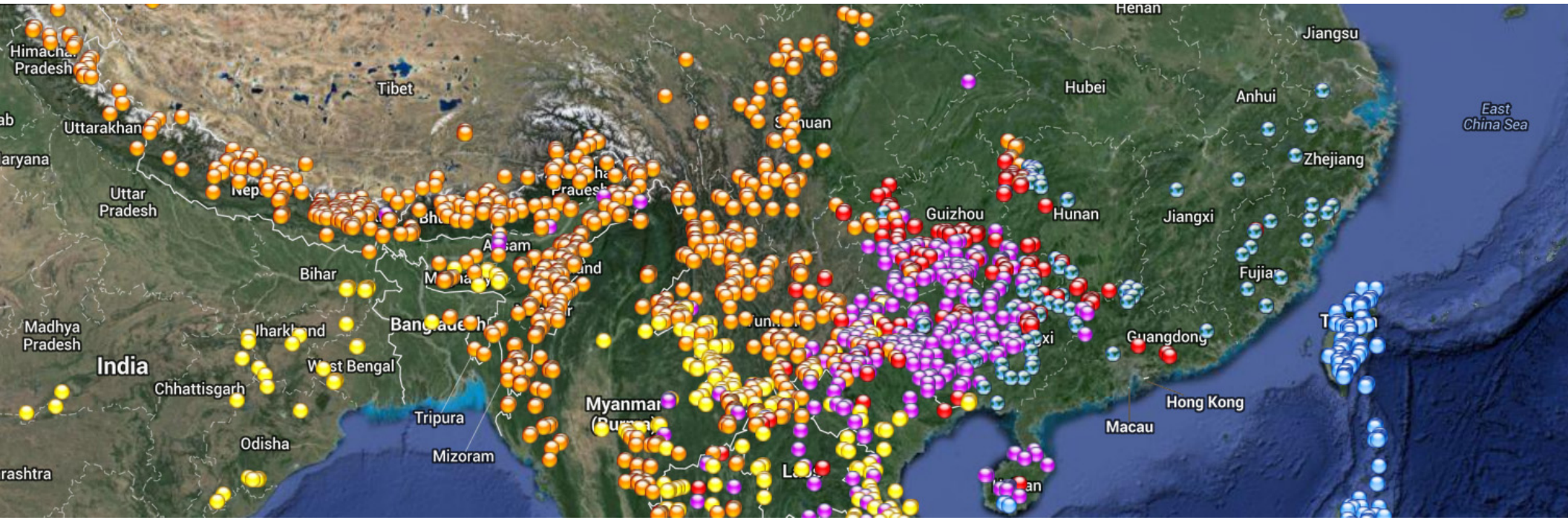
to help build a research community,

inside and **outside** linguistics**:**

1. connect the dots: creating / using data
2. build a comfort zone / smooth the path
3. speak the language / seek common ground



Bronze, Gold, and the Asia-Pacific Data Warehouse



Doug Cooper

Center for Research in Computational Linguistics

sealang.net doug.cooper.thailand@gmail.com