Many people here have been silently sharing the same thought:

Q: Where did all those clicks (in Xhosa and Zulu) come from?

A: Nguni languages

– Speakers borrowed syllables with clicks from neighboring Khoisan languages 500-1000 years ago.
– Cultural tradition to avoid using the names of in-laws.

AKA: The oldest form of Risk Management
Talking to the crowd in 7,000 languages

Robert Munro
Idibon
Information is increasing

• Scale (well-known)
• Diversity (less understood)
  – On a given day, what is the average number of languages that someone could potentially hear?
  – How has this changed?
Daily potential language exposure
Daily potential language exposure

# of languages

Year

2012 UR Forum Mapping Global Risk
Daily potential language exposure

Year

# of languages

50
1500
5000
2000
1400
720
540
500

2012 UR Forum Mapping Global Risk
Daily potential language exposure

Putting a phone in the hands of everyone on the planet is the easy part

Understanding everyone is going to be more complicated

Year

2012 UR Forum Mapping Global Risk
99% of languages don’t have machine-translation or similar services:

• Disproportionately lower healthcare & education
• Disproportionately greater exposure to disasters

Crowdsourcing can bridge part of the gap.
Crowdsourcing

2012 UR Forum Mapping Global Risk
Crowdsourced processing of information in Haitian Kreyol.

1000s of Haitians in Haiti and among the diaspora.
Lopital Sacre-Coeur ki nan vil Okap, pre pou li resevwa moun malad e lap mande pou moun ki malad yo ale la.

“Sacre-Coeur Hospital which located in this village of Okap is ready to receive those who are injured. Therefore, we are asking those who are sick to report to that hospital.”
Lopital Sacre-Coeur ki nan vil Okap, pre pou li resevwa moun malad e lap mande pou moun ki malad yo ale la.

“Sacre-Coeur Hospital which located in this village of Okap is ready to receive those who are injured. Therefore, we are asking those who are sick to report to that hospital.”
Lopital Sacre-Coeur ki nan vil Okap, pre pou li resevwa moun malad e lap mande pou moun ki malad yo ale la.

“Sacre-Coeur Hospital which located in this village of Okap is ready to receive those who are injured. Therefore, we are asking those who are sick to report to that hospital.”
Evaluating local knowledge

45,000 messages

Lopital Sacre-Coeur ki nan vil Okap, pre pou li resevwa moun malad e lap mande pou moun ki malad yo ale la.

Lopital Sacre-Coeur ki nan vil Okap, pre pou li resevwa moun malad e lap mande pou moun ki malad yo ale la.

Lopital Sacre-Coeur ki nan vil Okap, pre pou li resevwa moun malad e lap mande pou moun ki malad yo ale la.

Lopital Sacre-Coeur ki nan vil Okap, pre pou li resevwa moun malad e lap mande pou moun ki malad yo ale la.

Non-Haitians

< 5 minutes each

Lopital Sacre-Coeur ki nan vil Okap, pre pou li resevwa moun malad e lap mande pou moun ki malad yo ale la.

3,000 messages

< 5 minutes each

Lopital Sacre-Coeur ki nan vil Okap, pre pou li resevwa moun malad e lap mande pou moun ki malad yo ale la.

Haitians (volunteers and paid)

> 4 hours each
Lessons learned

• Default to private data practices
  (Majority decision was not to use a public map)
• Find volunteers through strong social ties
  (10x larger/faster than the publicized efforts)
• Avoid activists (‘bloggers’, ‘crisis-mappers’ …)
• Localize to the crisis-affected community
  (25% of work was by paid workers in Haiti)
Paid workers in Mirebalais, Haiti (FATEM)

Benchmarks we can use:*

- $0.25 per translation
- $0.20 per geolocation
- $0.05 per categorization / filtering
- 4:00 minutes per report processed

Can volunteerism undercut this cost?

Data-structuring for 2010 floods in Pakistan

Multiple inexperienced people are more accurate than one experience person.*

*Chohan, Hester and Munro. 2012. Pakreport: Crowdsourcing for Multipurpose and Multicategory Climate-related Disaster Reporting. *Climate Change, Innovation & ICTs Project*. CDI
Lessons learned

• Default to private data practices (!)
  (Taliban threatened to attack mapped aid workers)

• Cross-validate tasks across multiple workers
  (We used CrowdFlower, as with Mission 4636)

• Localize to the crisis-affected community
  (Data obtained by hand / created jobs)
Scaling beyond purely manual processing.

Disease outbreaks are the world’s single greatest killer.

No organization is tracking them all.
Diseases eradicated in the last 75 years:

smallpox

Increase in air travel in the last 75 years:
90% of ecological diversity

90% of linguistic diversity
Reported locally before identification

Simply *finding* these early reports can help prevent epidemics.

HIV
- decades
- (35 million infected)

H1N5 (Bird Flu)
- weeks
- (>50% fatal)

H1N1 (Swine Flu)
- months
- (10% of world infected)
Reports (millions)

Machine-learning (millions)

Microtaskers (thousands)

Analysts – domain experts (capped number)

epidemicIQ

в предстоящий осенне-зимний период в Украине ожидают две эпидемии гриппа

مزيد من انفلونزا الطيور في مصر

香港现1例H5N1禽流感病例
曾游上海南京等地
E Coli in Germany

The AI head-start
Lessons learned

• Current data privacy practices are insufficient (reports from areas where victims are vilified)

• Crowdsourcing can provide needed skill-sets (100s of German speakers at short notice)

• Natural language processing can scale beyond human processing capacity
A negative example

- 2283 reports already-open, English sources
- 1 month of full-time management and contributions from >100 volunteers
Equivalent cost from paid workers

• $575.75
  (or about $800 with multiple steps)

Equivalent time cost from Libyan nationals:

• 152.2 hours = less than 1 month for 1 person
  (would also address some security concerns)
Lessons learned

- Crowdsourced volunteers were not required (cost more to run than was saved by not paying) (a single in-house Libyan could have achieved more)
- Default to private data practices (assume all identities of volunteers were exposed) (Libyans opposed the public map)
Crowdsourcing and risk

People’s real-time locations are their most sensitive personal information.
Crowdsourcing distributes information to a large number of individuals for processing.
For information about at-risk individuals:
• Is it right to crowdssource the processing?
• Is it right to use a public-facing map.
Conclusions

Recommendations

• Engage people with local knowledge
• *Employ* people with local knowledge
• Statistically cross-validate on-the-fly
• Default to private data practices
• Scale via natural language processing
Thank you

Robert Munro
Idibon
@WWRob