useless evaluation
vs
user-less evaluation

@julian_urbano

“If you can’t measure it, you can’t improve it.”

—Lord Kelvin

How I Would Like to Measure…

- Gather all my users
- Collect all music material
- Have them use my algorithm
- Observe, ask, explore
- Analyze and Learn

measurement = f(track, algorithm, user, context,…)

…and Why We Don’t Do It

- Slow, expensive
- Representativeness
- Ethics, privacy, hidden effects, inconsistency
- Hard to replicate experiments
- Just plain impossible to reproduce results

What We Do

- Cranfield paradigm, aka dataset-based evaluation
  - Use controlled corpus
  - Remove users, but include a user-abstraction

- **Static user component:** Annotations
  - Model utility of individual parts of the corpus
- **Dynamic user component:** Metrics
  - Model the behavior of users, their interaction with the full algorithm output

What We Achieve

- Remove all sources of variability, except algorithms

measurement = f(track, algorithm, user, context,…)

What We Achieve

- Remove all sources of variability, except algorithms

~~*measurement = f(track, algorithm, user, context,…)*~~

measurement = f(algorithm)

What We Achieve

- Remove all sources of variability, except algorithms

~~*measurement = f(track, algorithm, user, context,…)*~~

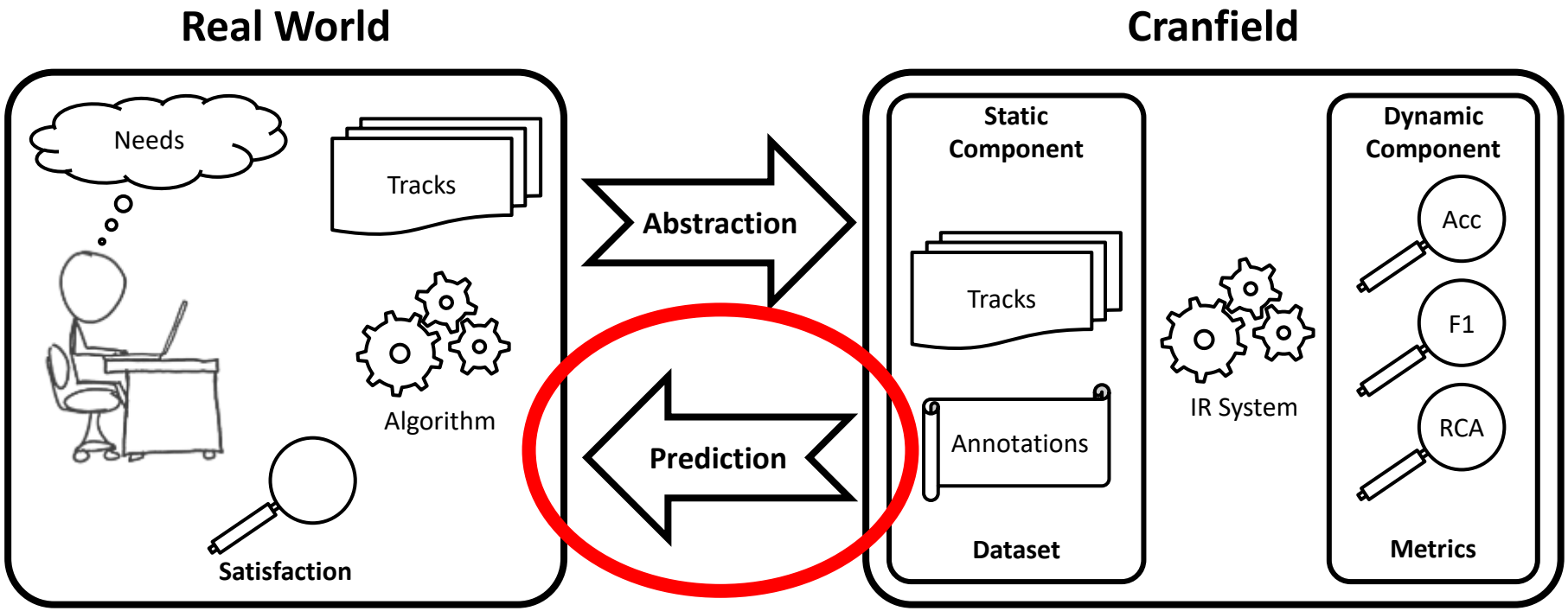
measurement = f(algorithm)

- Evaluation is deterministic
- Experiments are inexpensive and easy to run
- Research becomes systematic
- Reproducibility is not only possible but easy

"If it disagrees with the experiment, it's wrong. In that simple statement is the key to science. It doesn't make any difference how beautiful your guess is, it doesn't matter how smart you are, or what your name is. ***If it disagrees with the experiment, it's wrong.*** *That's all there is to it."*

—Richard Feynman

- The dataset and the metrics are assuming a **user model**, even if it is not explicit
- Are these models ~~correct~~ appropriate?



Examples from Search Engine Evaluation

- Utility of a document w.r.t. scale of annotation
  - Binary or graded relevance?
  - Linear utility w.r.t. relevance? Exponential?
  - Independent of other documents?
  - Diversity? Coverage? Novelty?
- Top heaviness to penalize late arrival
  - No discount? Linear? Logarithmic?
  - Independent of other documents?
- Interaction, browsing, document length?
- Cutoff
  - Fixed: only top k documents?
  - Dynamic: wherever some condition is met?
  - All documents?

- **Metrics are models of a stochastic process involving the user**

User Models in MIR

Copyright by Justin
(because we're friends)

Evaluation is all About Prediction

- **Whether the algorithm output satisfies a user or not, has nothing to do with how we measure its performance**

- Many user studies across fields show that there is virtually **no correlation between** our metrics and the real world

- So what are we doing? Where are we headed?
- What problems do we **think** we're solving?

User Studies!

- So I “just” run a user study and check how users behave/react/interact/perceive (with) the output from my algorithm
- That might be useful for you, **this one time**, but it’s not useful for the community

- But think about what we give up
  - ~~Evaluation is deterministic~~
  - ~~Experiments are inexpensive and easy to run~~
  - ~~Research is systematic~~
  - ~~Reproducibility is easy~~

What User Studies Tell Us

- **What** happens in the real world
- **To** improve user models in datasets & metrics
- **So** we do better evaluation
- **Make** better predictions

- **And only then**, we'll know
- **What** to optimize
- **How** to build better algorithms

User studies should not be about getting confirmation that your algorithm works

They should be about learning how to do better evaluation,
so that we don't need user studies

Melody Extraction: Addressing User Satisfaction

Belén Nieto
belen.nieto@upf.edu

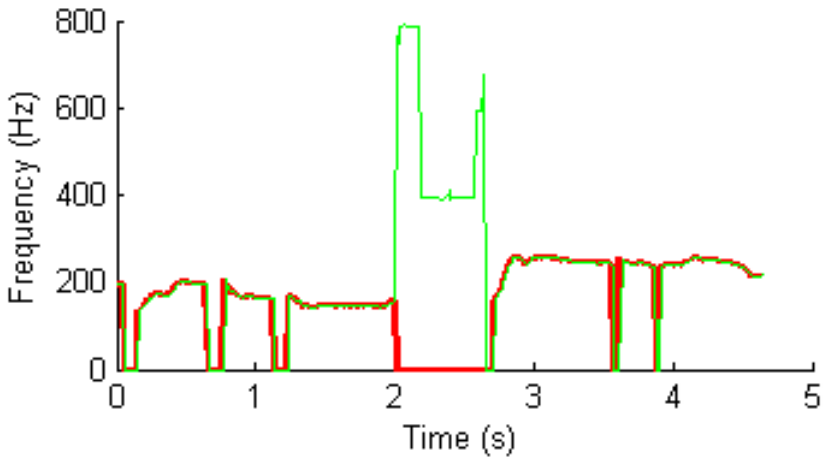
Supervisors:
Emilia Gómez
Julián Urbano
Justin Salamon

“Happy families are all alike;
every unhappy family is unhappy in its own way”

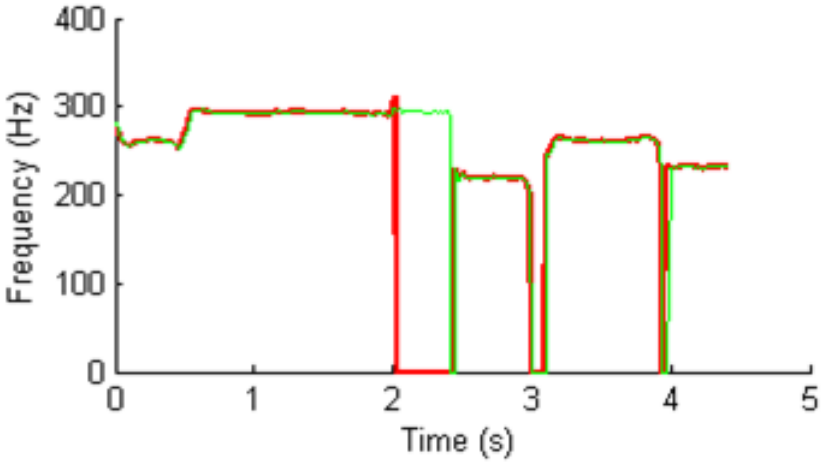
—Tolstoy

Are all mistakes the same?

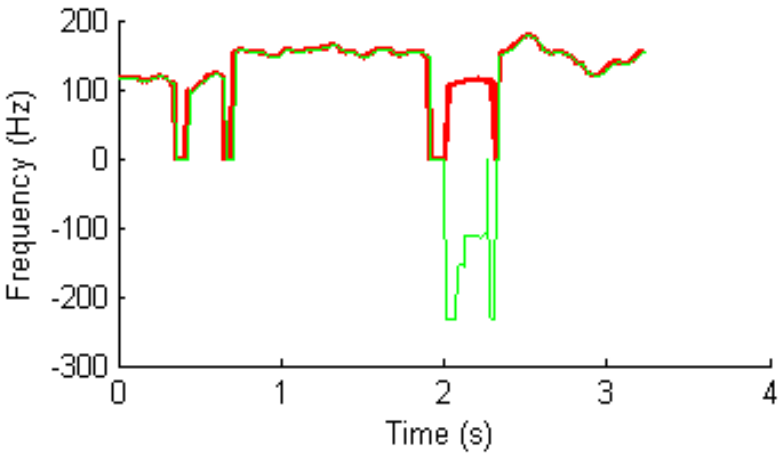
(a) Adding a random noise



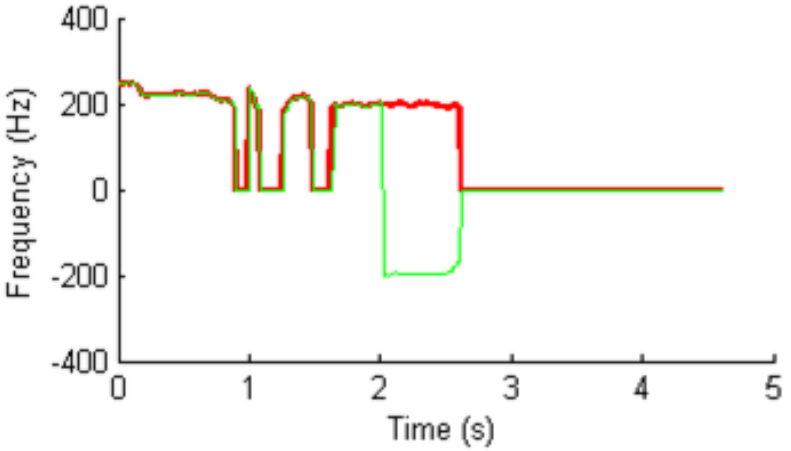
(b) Lengthening a note



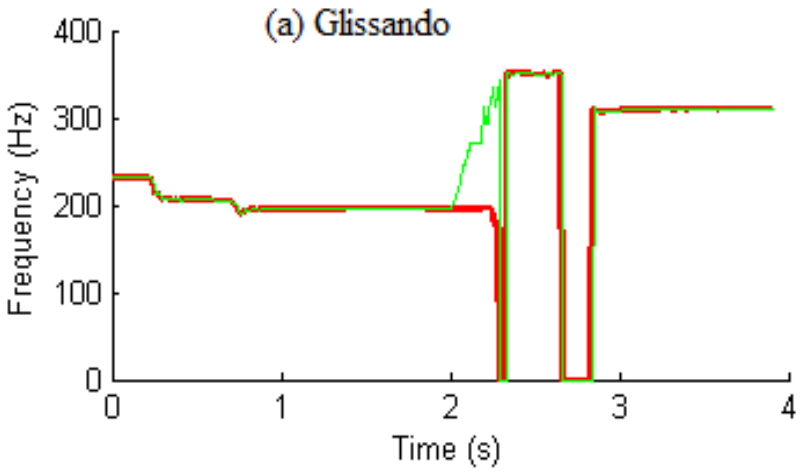
(a) Missing a full note

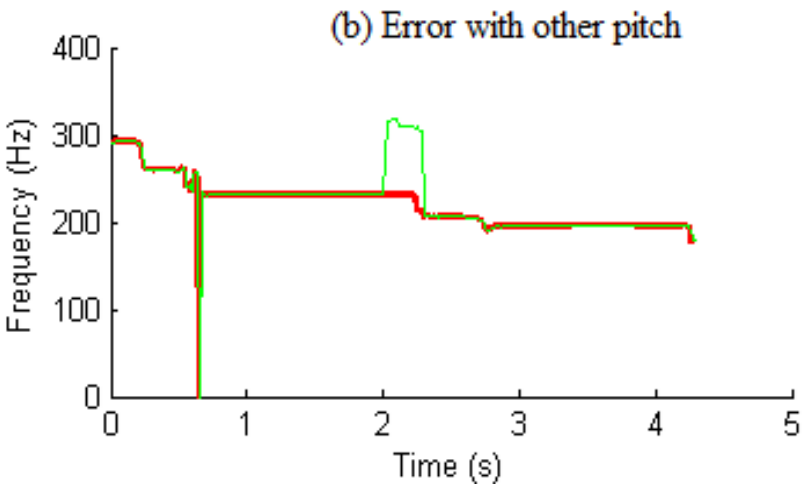


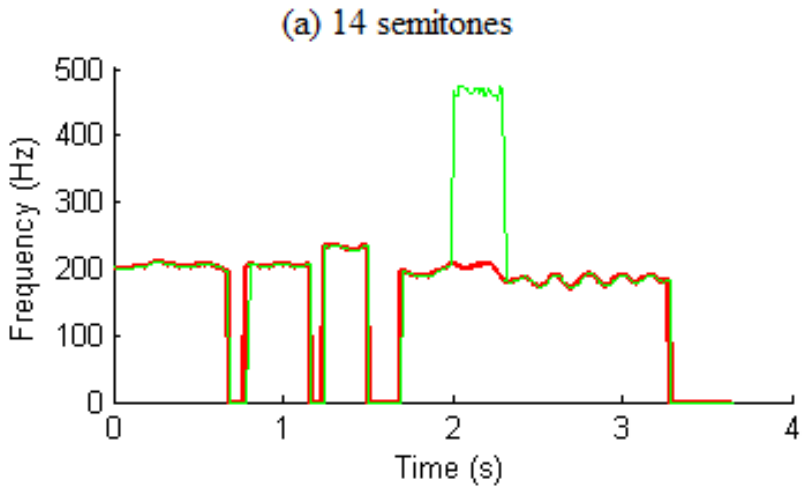
(b) Shortening a note

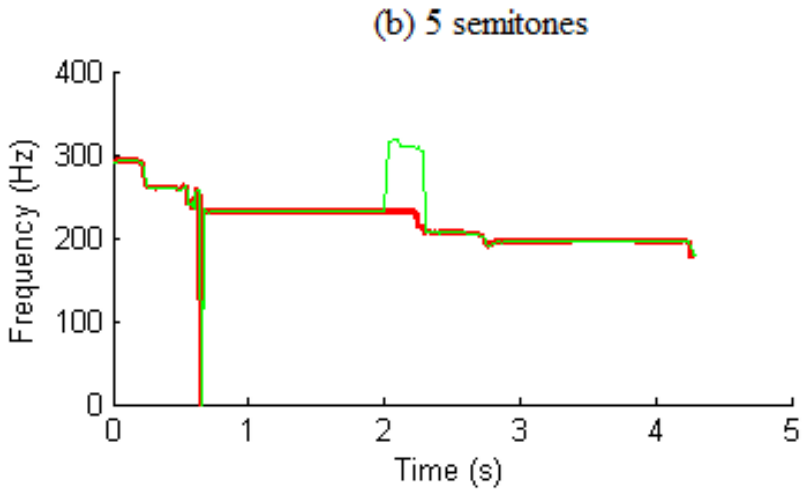


Are all mistakes the same?









Use(r)less Evaluation

- How does error **type** and **magnitude** affect the perceived similarity with the original?

- It’s not as simple as correct/incorrect
- Of course, it depends on the **use case**

- How can we incorporate this new knowledge in the metrics?

less confirmation

more exploration

(what do you think?)