Probabilistic Modeling of Chronological Dates to Serve Machines and Scholars

Andreas Habring, Anguelos Nicolaou, Daniel Luger, Florian Atzenhofer-Baumgartner, Florian Lamminger, Franziska Decker, Tamás Kovács, Sandy Aoun Georg Vogeler, Martin Holler

July 14, 2023

Outline

1 **Motivation**

2 Roles

3 Ambiguity Modeling

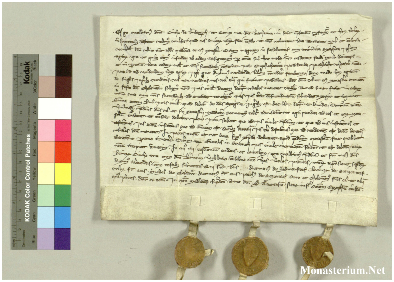
4 Proposals

Why

- We are (will be) going distant
- Dates must be used in mass
- Can dates be inferred from data?
- It all begins by how we measure things

The data that Inspired us

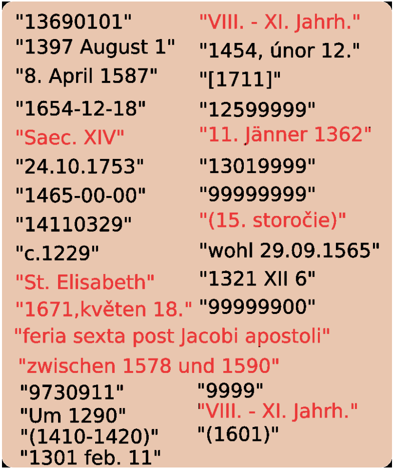
- 500K Documents
- (CEI) TEI-4 derived format
- Diplomatic Charters
- 1000 Charters randomly sub-sampled
- 2211 Dates associated with the issued



Kodak
KODAK Color Control Patches
Monasterium.Net

How to make sense of date data?

- Arcane
- Numerical
  - Is it YYYYMMDD?
  - Or DDMMYYYY?
  - YYYMMDD??
- Expressing Ambiguity
  - (date)
  - [date]
  - 99?



"13690101"
"1397 August 1"
"8. April 1587"
"1654-12-18"
"Saec. XIV"
"24.10.1753"
"1465-00-00"
"14110329"
"c.1229"
"St. Elisabeth"
"1671,květen 18."

"VIII. - XI. Jahrh."
"1454, únor 12."
"[1711]"
"12599999"
"11. Jänner 1362"
"13019999"
"99999999"
"(15. storočie)"
"wohl 29.09.1565"
"1321 XII 6"
"99999900"

"feria sexta post Jacobi apostoli"
"zwischen 1578 und 1590"

"9730911"
"Um 1290"
"(1410-1420)"
"1301 feb. 11"

"9999"
"VIII. - XI. Jahrh."
"(1601)"

How to make sense of date data?

- Arcane
- Numerical
  - Is it YYYYMMDD?
  - Or DDMMYYYY?
  - YYYMMDD??
- Expressing Ambiguity
  - (date)
  - [date]
  - 99?

"13690101"
"1397 August 1"
"8. April 1587"
"1654-12-18"
"Saec. XIV"
"24.10.1753"
"1465-00-00"
"14110329"
"c.1229"
"St. Elisabeth"
"1671,květen 18."

"VIII. - XI. Jahrh."
"1454, únor 12."
"[1711]"
"12599999"
"11. Jänner 1362"
"13019999"
"99999999"
"(15. storočie)"
"wohl 29.09.1565"
"1321 XII 6"
"99999900"

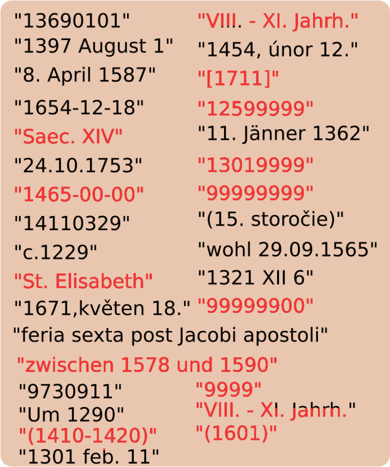
"feria sexta post Jacobi apostoli"
"zwischen 1578 und 1590"

"9730911"
"Um 1290"
"(1410-1420)"
"1301 feb. 11"

"9999"
"VIII. - XI. Jahrh."
"(1601)"

How to make sense of date data?

- Arcane
- Numerical
  - Is it YYYYMMDD?
  - Or DDMMYYYY?
  - YYYMMDD??
- Expressing Ambiguity
  - (date)
  - [date]
  - 99?



"13690101"
"1397 August 1"
"8. April 1587"
"1654-12-18"
"Saec. XIV"
"24.10.1753"
"1465-00-00"
"14110329"
"c.1229"
"St. Elisabeth"
"1671,květen 18."

"VIII. - XI. Jahrh."
"1454, únor 12."
"[1711]"
"12599999"
"11. Jänner 1362"
"13019999"
"99999999"
"(15. storočie)"
"wohl 29.09.1565"
"1321 XII 6"
"99999900"

"feria sexta post Jacobi apostoli"
"zwischen 1578 und 1590"

"9730911"
"Um 1290"
"(1410-1420)"
"1301 feb. 11"

"9999"
"VIII. - XI. Jahrh."
"(1601)"

What is a Date?

- A date is Number
  - 1/6/1347 $\Longrightarrow$ 1347.5
- Written in weird ways
  - Not our job
  - OS / UI
- When Exactly?
  - Minimum precision by project
  - We need to be more imprecise
  - **How can we express imprecision?**

Outline

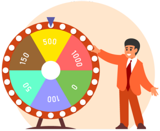
1. Motivation
2. **Roles**
3. Ambiguity Modeling
4. Proposals

The scholar

- Expresses knowledge
- Reasons in nuance
- Needs to express nuance
- The data models don't allow that
- Afraid of being wrong
- Can nuance be a numerical?

The Guesser

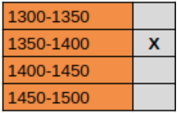
- Expresses opinion/estimation
- Typically a machine learning model
  - Loss function needed to train
  - Differentiable
- Could be a human
- Opinion model
  - A choice among fixed categories
  - A moment
  - An interval
  - A Gaussian
  - A Monte Carlo Approximation



publicdomainvectors.org

The Guesser

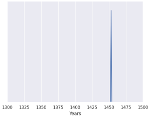
- Expresses opinion/estimation
- Typically a machine learning model
  - Loss function needed to train
  - Differentiable
- Could be a human
- Opinion model
  - A choice among fixed categories
  - A moment
  - An interval
  - A Gaussian
  - A Monte Carlo Approximation



1300-1350	
1350-1400	X
1400-1450	
1450-1500	

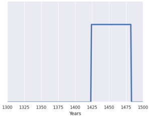
The Guesser

- Expresses opinion/estimation
- Typically a machine learning model
  - Loss function needed to train
  - Differentiable
- Could be a human
- Opinion model
  - A choice among fixed categories
  - A moment
  - An interval
  - A Gaussian
  - A Monte Carlo Approximation



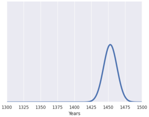
The Guesser

- Expresses opinion/estimation
- Typically a machine learning model
  - Loss function needed to train
  - Differentiable
- Could be a human
- Opinion model
  - A choice among fixed categories
  - A moment
  - An interval
  - A Gaussian
  - A Monte Carlo Approximation



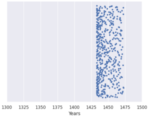
The Guesser

- Expresses opinion/estimation
- Typically a machine learning model
  - Loss function needed to train
  - Differentiable
- Could be a human
- Opinion model
  - A choice among fixed categories
  - A moment
  - An interval
  - A Gaussian
  - A Monte Carlo Approximation



The Guesser

- Expresses opinion/estimation
- Typically a machine learning model
  - Loss function needed to train
  - Differentiable
- Could be a human
- Opinion model
  - A choice among fixed categories
  - A moment
  - An interval
  - A Gaussian
  - A Monte Carlo Approximation



The Judge

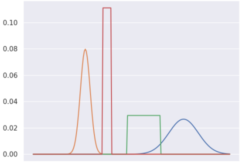
- Performance Evaluation
- Scoring a Guesser
- "Metric" that satisfies our perceived notion of performance
  - Range in [0, 1]
  - $A > B \wedge B > C \implies A > C$
  - Can be asymmetric

The Judge's challenge

- Must not have favorites
  - Not favor Datasets: eg: classification
  - Not favor Methods eg: classification vs. regression
- Must be Winnable
- Must not be Gameable
- Solution:
  - Everything can be a density function
  - A curve over time with a finite surface
  - Even a moment has a duration

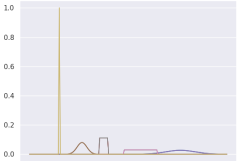
The Judge's challenge

- Must not have favorites
  - Not favor Datasets: eg: classification
  - Not favor Methods eg: classification vs. regression
- Must be Winnable
- Must not be Gameable
- Solution:
  - Everything can be a density function
  - A curve over time with a finite surface
  - Even a moment has a duration



The Judge's challenge

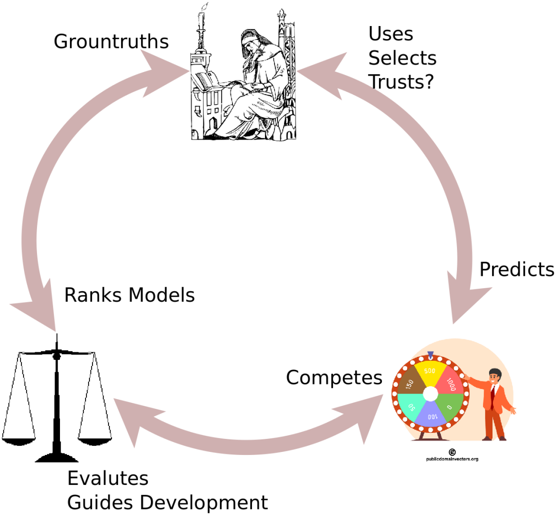
- Must not have favorites
  - Not favor Datasets: eg: classification
  - Not favor Methods eg: classification vs. regression
- Must be Winnable
- Must not be Gameable
- Solution:
  - Everything can be a density function
  - A curve over time with a finite surface
  - Even a moment has a duration



The Synergy

🟦 Dates

🟦 Scores



Grountruths

Uses
Selects
Trusts?

Predicts

Ranks Models

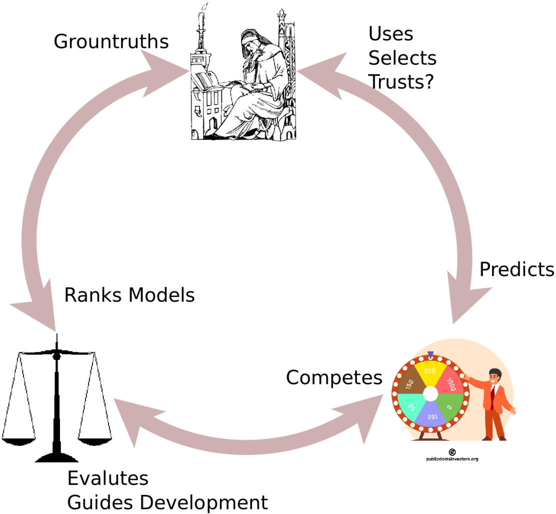
Competes

Evalutes
Guides Development

publicdomainvectors.org

The Synergy

- 🟦 Dates
- 🟦 Scores



Grountruths

Uses
Selects
Trusts?

Predicts

Competes

Evalutes
Guides Development

Ranks Models

Outline

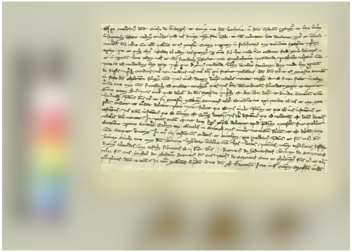
1. Motivation
2. Roles
3. **Ambiguity Modeling**
4. Proposals

Ambiguity: Not before, Not after

- AKA: From - To
- Ideal for scholars
- Internal Charter Features (textual)
- Statistics
  - Complete
  - Interval censored
  - Left censored
  - Right censored
  - Regular phenomena eg: engine failure times

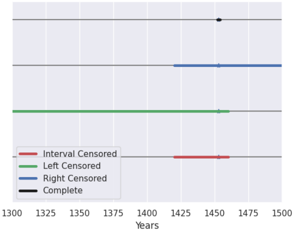
Ambiguity: Not before, Not after

- AKA: From - To
- Ideal for scholars
- Internal Charter Features (textual)
- Statistics
  - Complete
  - Interval censored
  - Left censored
  - Right censored
  - Regular phenomena eg: engine failure times



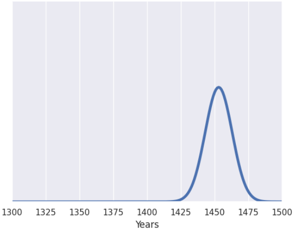
Ambiguity: Not before, Not after

- AKA: From - To
- Ideal for scholars
- Internal Charter Features (textual)
- Statistics
  - Complete
  - Interval censored
  - Left censored
  - Right censored
  - Regular phenomena eg: engine failure times



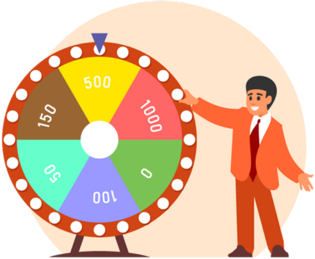
Ambiguity: Give or Take

- Gaussian, give or take σ
- Suited for guessers
- External Charter Features (visual)



Ambiguity: Give or Take

- Gaussian, give or take σ
- Suited for guessers
- External Charter Features (visual)



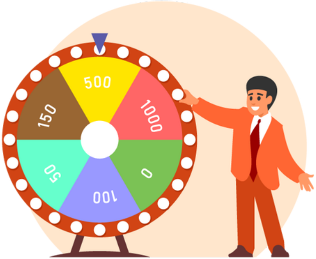
publicdomainvectors.org

Ambiguity: Give or Take

- Gaussian, give or take σ
- Suited for guessers
- External Charter Features (visual)

Ambiguity: Normalisation

- How many guesses are we allowed?
- Probability Density:
  - Sum 1
  - Mandatory for guesser
- Plausibility Density:
  - Max 1
  - Surface defined by the annotator



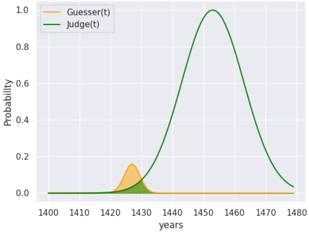
publicdomainvectors.org

Ambiguity: Normalisation

- How many guesses are we allowed?
- Probability Density:
  - Sum 1
  - Mandatory for guesser
- Plausibility Density:
  - Max 1
  - Surface defined by the annotator

Performance Evaluation: Precision

- How much of a guess falls within the groundtruth
- Not all samples are equally hard
  - Weigh samples by the inverse of their surface



Outline

1 Motivation

2 Roles

3 Ambiguity Modeling

4 **Proposals**

Our (hypo)Theses

- Everything is an ambiguous instant
- Opinions must be interpretable as DF
- Uniform (flat) and Gaussian (bell-shaped) are expressive enough for most humans
- The "role" (use-case) dictates DF normalisation
- Performance evaluation should not be favoring any guesser type

- Anything naturally lengthy should be modeled as two or more moments
- Life $\Longrightarrow$ (Birth, Death)

Our (hypo)Theses

- Everything is an ambiguous instant
- Opinions must be interpretable as DF
- Uniform (flat) and Gaussian (bell-shaped) are expressive enough for most humans
- The "role" (use-case) dictates DF normalisation
- Performance evaluation should not be favoring any guesser type

- How would I spread my guesses?

Our (hypo)Theses

- Everything is an ambiguous instant
- Opinions must be interpretable as DF
- Uniform (flat) and Gaussian (bell-shaped) are expressive enough for most humans
- The "role" (use-case) dictates DF normalisation
- Performance evaluation should not be favoring any guesser type

- Link: $to - from \approx 2\sigma$
- Open intervals mean everything goes

Our (hypo)Theses

- Everything is an ambiguous instant
- Opinions must be interpretable as DF
- Uniform (flat) and Gaussian (bell-shaped) are expressive enough for most humans
- The "role" (use-case) dictates DF normalisation
- Performance evaluation should not be favoring any guesser type

- Plausibility DF: "there must be a perfect prediction"
- Probability DF: "we all get as many guesses"
- Precision:
  - Prediction correctness
  - Quality of a prediction
- Recall??:
  - Precision with inverted roles
  - Prediction difficulty

Our (hypo)Theses

- Everything is an ambiguous instant
- Opinions must be interpretable as DF
- Uniform (flat) and Gaussian (bell-shaped) are expressive enough for most humans
- The ”role” (use-case) dictates DF normalisation
- Performance evaluation should not be favoring any guesser type

- How can we know if a better method ever arrives?
- Apples and Oranges are Fruit!

How should it be done

Distant reading/viewing needs math and humanities

- Armageddon (1998):
- Who are the astronauts?
- Who are the drillers?

What is the data model of a (ambiguous) date?

- A choice:
  - From - To (two numbers)
  - When (a number), give or take (an optional number)
- And an optional string for recording the reasoning

Questions/Remarks/Objections?

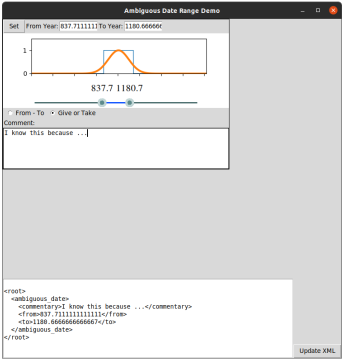
Thank you!

Look at our UI proof of concept demo! Slides are also there!

`https://github.com/anguelos/ambiguous_dates`

UI Demo

- A single widget for Gaussian and Uniform
- The role is irrelevant
- Records a spread



From Year: 837.7111111 To Year: 1180.666666

○ From - To ● Give or Take

Comment:

I know this because ...

```
<root>
  <ambiguous_date>
    <commentary>I know this because ...</commentary>
    <from>837.7111111111111</from>
    <to>1180.6666666666667</to>
  </ambiguous_date>
</root>
```