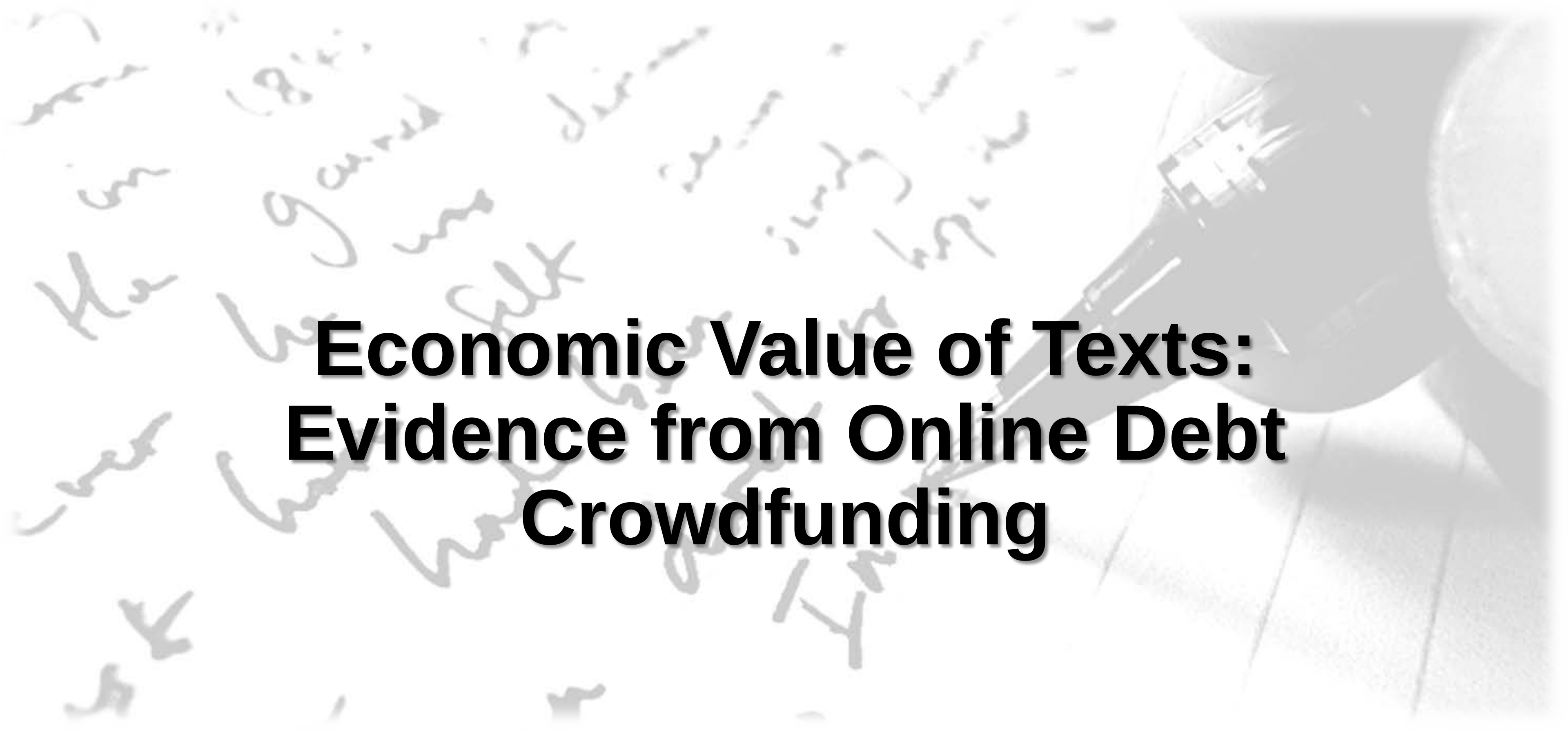


Economic Value of Texts: Evidence from Online Debt Crowdfunding

Mingfeng Lin, University of Arizona

(Joint work with Qiang Gao, City University of New York)

December 2nd, 2016

Texts are everywhere online…

… But do they actually offer any economic values?

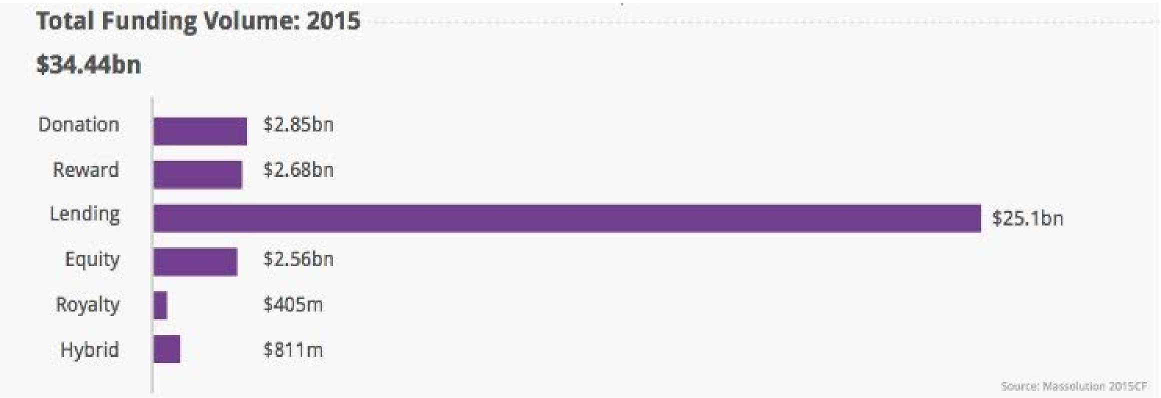
Why Debt Crowdfunding for this Study?

… Rather than other types of crowdfunding?

- **Conservative**: Presence of traditional quantitative credit information
- **Objective** “quality” information: Loan repayment
- **Similar incentives** as other types of crowdfunding
- **Larger** (vs. other types)

Also: texts are
not verifiable or legally
binding, as long as monthly
payments are made

➔ Intriguing to see if it
plays a role.

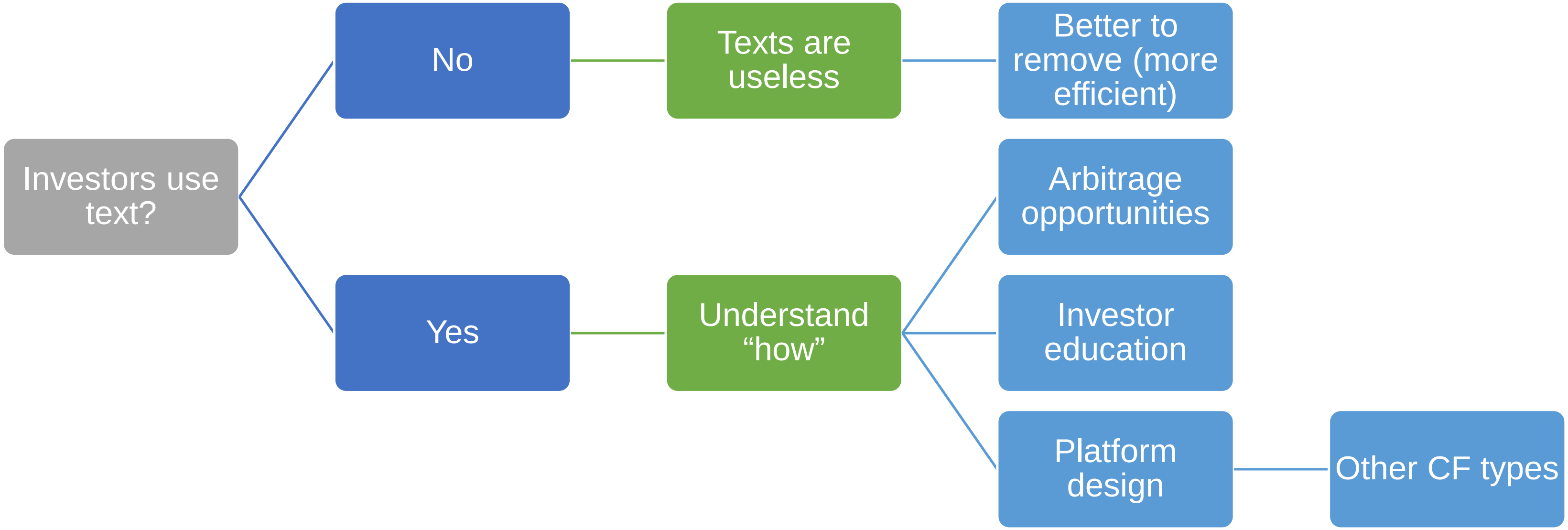


Source: Massolution 2015CF

Research Questions

1. Do investors take texts into account in their decisions?
2. Are texts, in particular linguistic features, related to loan repayment? How? And if so,
3. Do investors interpret these features correctly?

Motivations for these questions…



Investors use text?

No

Texts are useless

Better to remove (more efficient)

Yes

Understand “how”

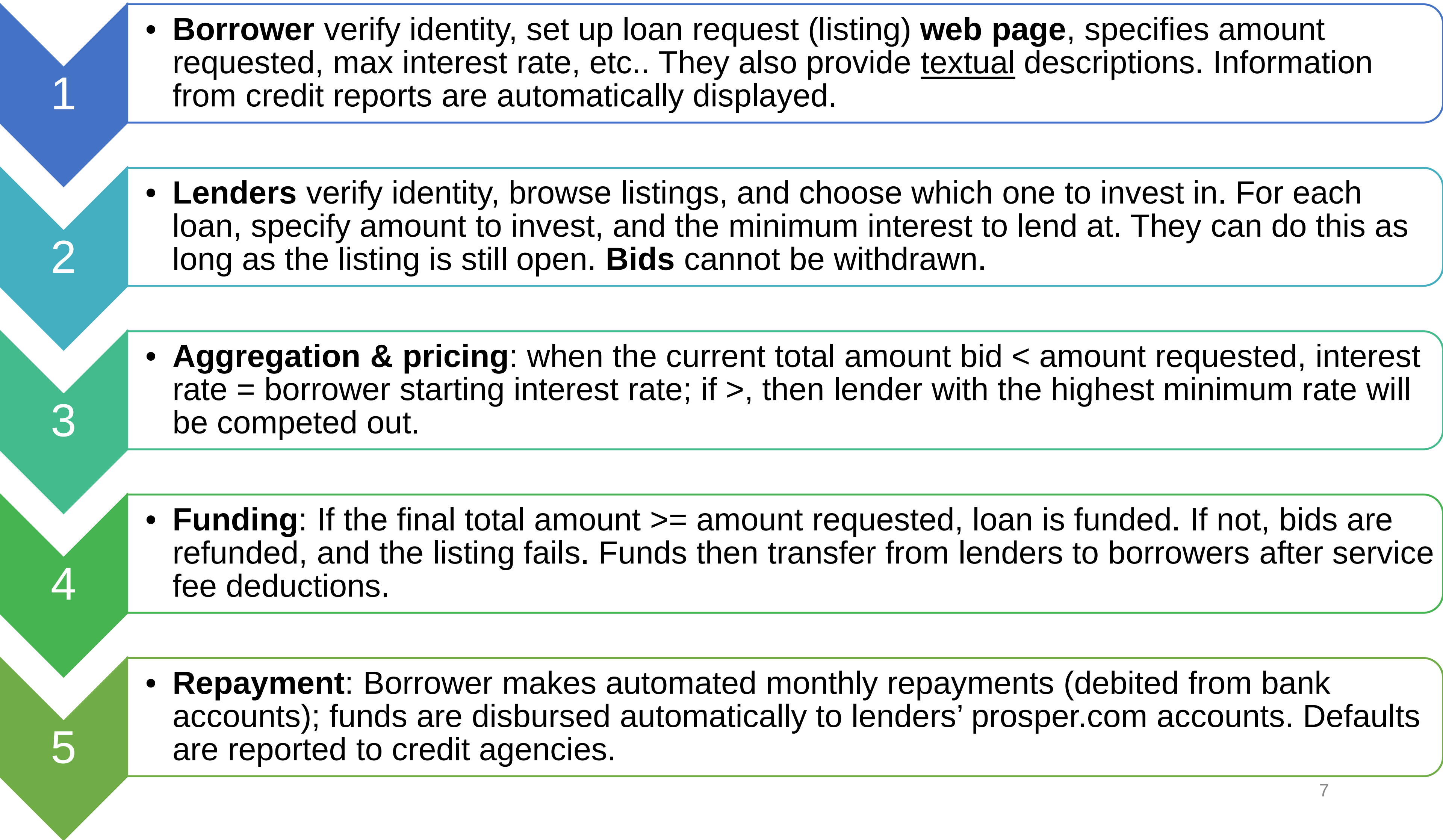
Arbitrage opportunities

Investor education

Platform design

Other CF types

Funding Process on Prosper.com (for the period we study)



1

- **Borrower** verify identity, set up loan request (listing) **web page**, specifies amount requested, max interest rate, etc.. They also provide textual descriptions. Information from credit reports are automatically displayed.

2

- **Lenders** verify identity, browse listings, and choose which one to invest in. For each loan, specify amount to invest, and the minimum interest to lend at. They can do this as long as the listing is still open. **Bids** cannot be withdrawn.

3

- **Aggregation & pricing**: when the current total amount bid < amount requested, interest rate = borrower starting interest rate; if >, then lender with the highest minimum rate will be competed out.

4

- **Funding**: If the final total amount >= amount requested, loan is funded. If not, bids are refunded, and the listing fails. Funds then transfer from lenders to borrowers after service fee deductions.

5

- **Repayment**: Borrower makes automated monthly repayments (debited from bank accounts); funds are disbursed automatically to lenders' prosper.com accounts. Defaults are reported to credit agencies.

Purpose of loan:
This loan will be used to start a company that will offer eco-friendly solutions to commercial and industrial companies. (Business Name) will provide high quality and environmentally friendly services and solutions to businesses of all sizes. Get in on the ground floor of this fantastic opportunity.

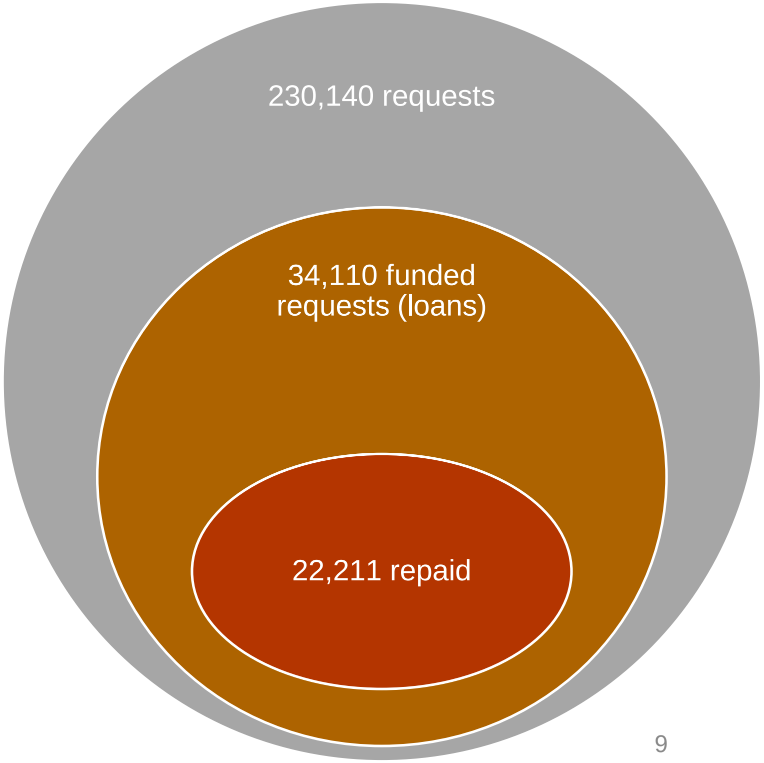
My financial situation:
I am a good candidate for this loan because I have over 5 years experience in the industry as a production supervisor for a disaster restoration and cleaning company. I also have a proven record of impeccable customer service, outstanding leadership and managerial skills, as well as great problem solving skills. My credit is good, and I have the income to repay the Prosper investors for their loan consideration. The profitability for a company like (Business Name) is outstanding. The risk factor for potential investors is extremely low. The market for eco-friendly solutions is infinite. At this time the market is untapped and offers enormous possibilities.

Our Competitive Advantage:
(Business Name) will succeed because Americans understand more than ever that we must collectively do our part to save our environment. Finally, eco-friendly solutions are being sought and used by consumers and businesses at an increasing rate. We will succeed by offering superior products, services and solutions using a very competitive and affordable pricing model.

We sincerely appreciate your interest.

Data

- Detailed transactions data from Prosper.com
  - 01/01/2007 – 05/01/2012
  - Information on all listings (requests, successful or not), funded loans (repaid or defaulted),all bids, and all members



Do investors pay attention to texts?

- Evidence from two policy changes
  - Removal of some borrowers’ texts
  - Removal of all texts
- Within-borrower variation (omitted for brevity)

Q1: Evidence from Two Website Policy Changes

- NE (Natural Experiment) #1:
  - May 3, 2010 – June 1, 2010
  - No prompts for AA / A borrowers to write texts

- NE #2
  - Starting 09/06/2013
  - Text section removed from all listings

(NE1) Funding Probability Before and After Policy Change

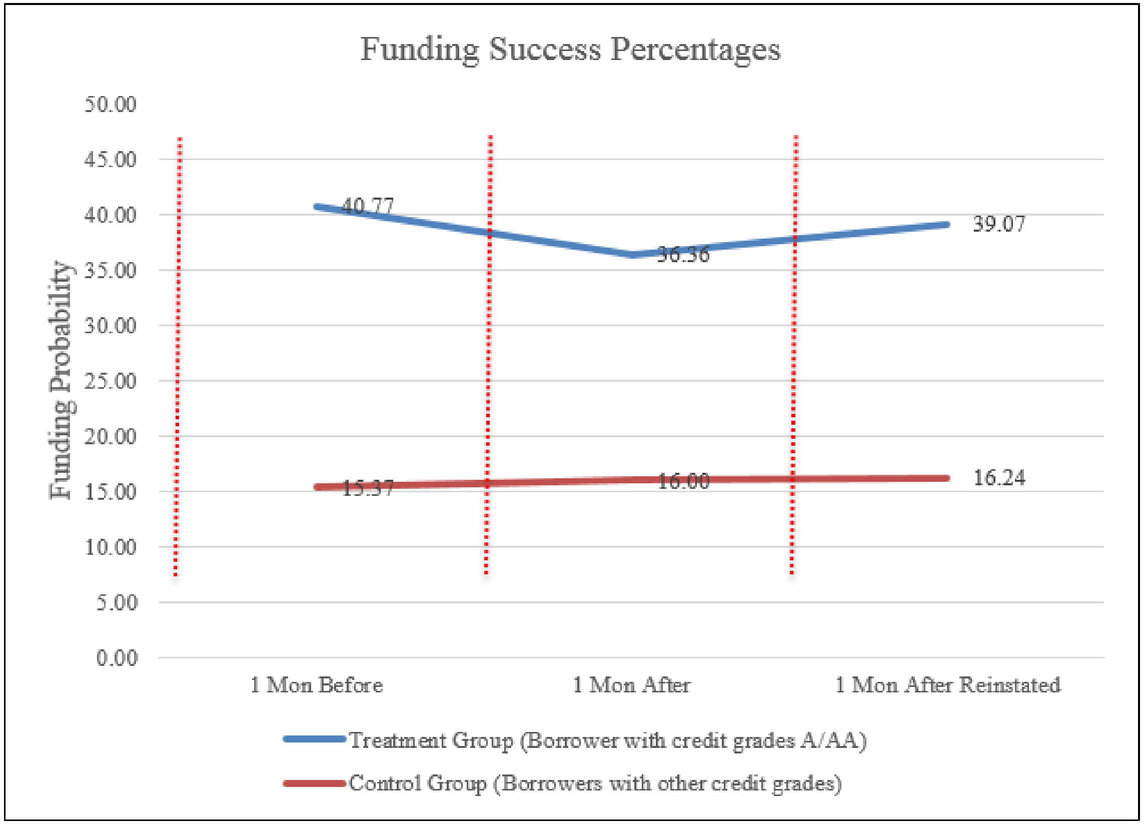
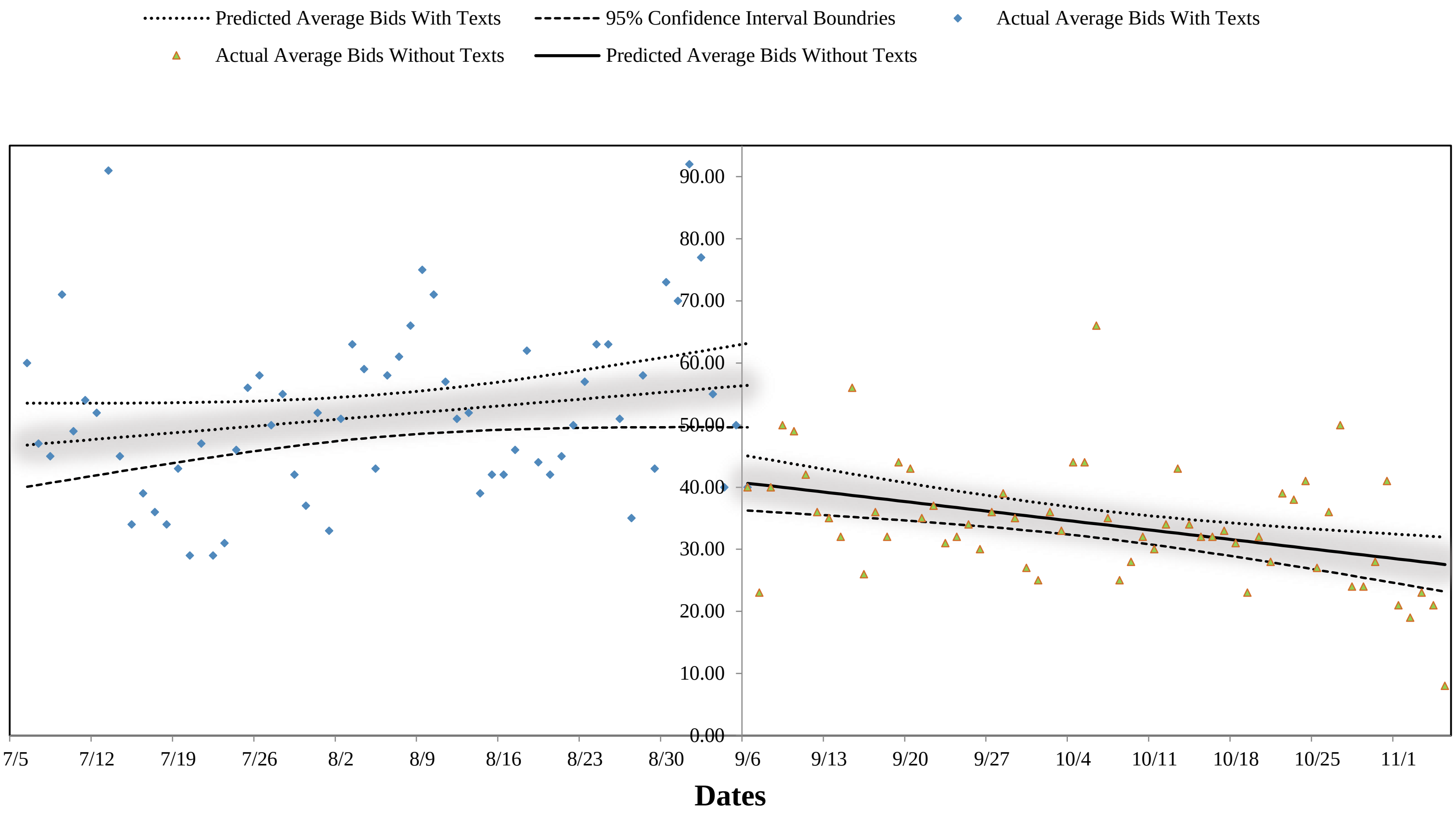


Figure 1. Funding Probability Before and After Policy Change

(NE2) #Bids when Text Section Removed

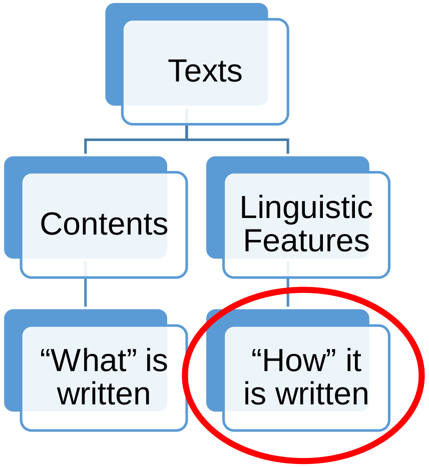


34.16% fewer bids

Texts and Loan Default Likelihood

- Linguistic features
- Hypotheses
- (Automated) extraction of linguistic features
- Explanatory model, results, and robustness

Q2: Explanatory Model

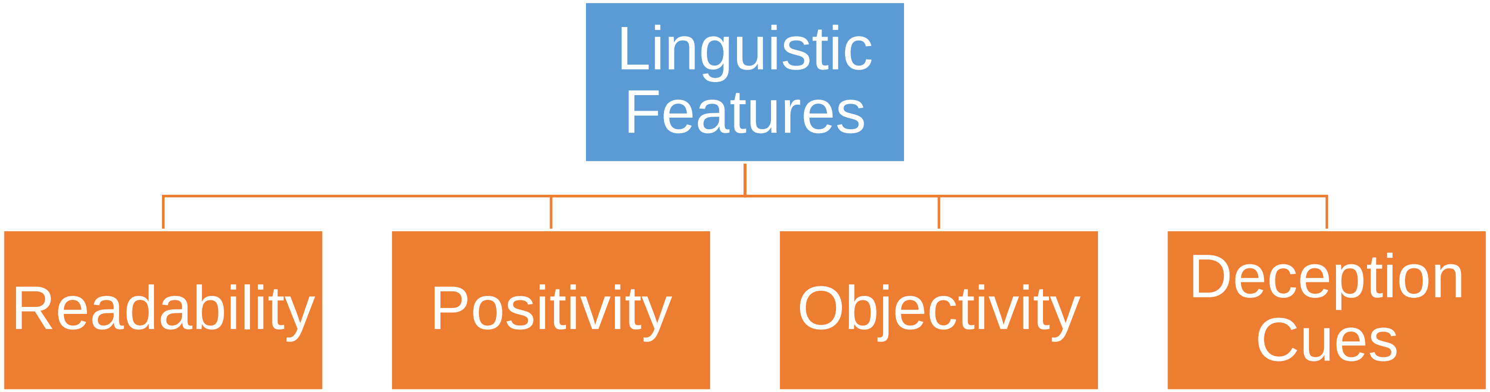


- Most studies of texts focus on linguistic features
- No standard, scalable approach for content
- **Robustness**: control for content (omitted here)
- Content in our context: **not verified**

We therefore focus on **linguistic features** of texts

Quantifying Linguistic Features

- We focus on linguistic styles that
  - Are relevant to ***willingness* to repay** (Flint 1997) or ***ability* to repay** (Duarte et al. 2012) because of the debt context;
  - Are frequently used in the literature; and
  - Have **well-established methods** or algorithms for measurement.



- These dimensions were separately studied in other contexts. We investigate them **jointly**.

Linguistic Features

- Readability: how accessible the texts are
- Positivity: positive attitude conveyed in the texts
- Objectivity: to what extent the texts are describing objective info
- Deception cues: how likely the texts were written with an intention to deceive

Hypotheses

Hypothesis	Details
H1	More readable, less likely to default.
H2	More positive, less likely to default. This relationship should be curvilinear.
H3	More objective, less likely to default.
H4	More deception cues, more likely to default.

Measurements of linguistic features: standard approach in computational linguistics literature

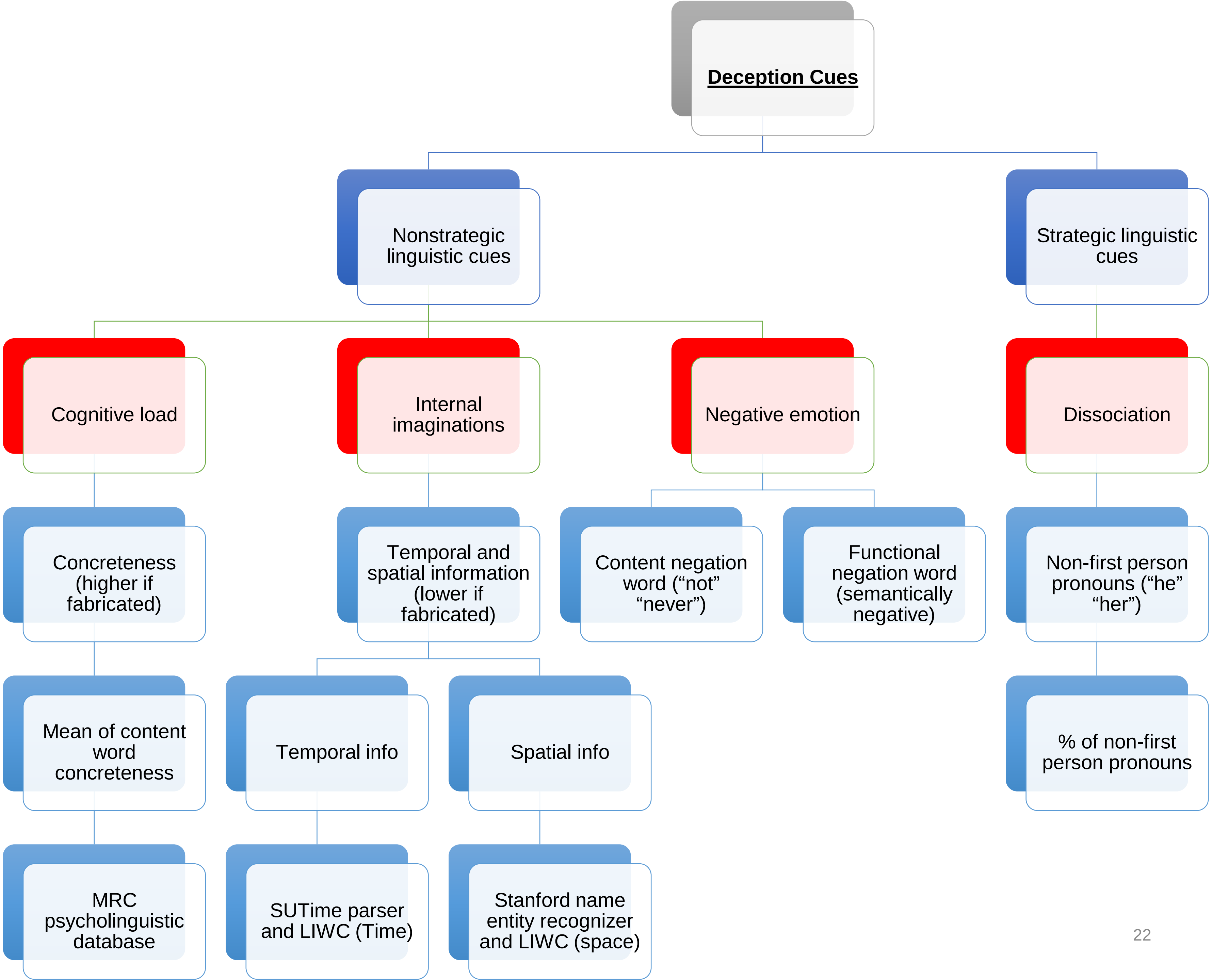
Measurement: Readability

Readability dimension	Measurement
Spelling errors	Spelling error corpus (Jurafsky and James 2000)
Grammatical errors	Probability on how far the text is from correct grammatical structures in an existing parser’s large, hand-coded database (Klein and Manning 2003)
Lexical complexity	Gunning-Fog index, FOG Score=0.4 × (ASL +100 × AHW) (DuBay 2004)

ASL: Average Sentence Length;
AHW: % of words with more than two syllables (“hard words”)

Measurement: Positivity and Objectivity

- Domain specificity: A machine learning rather than lexicon-based approach (Pang and Lee 2008)
- 1% stratified (by credit grade) random sample of loans
  - 70% training dataset
  - Remaining 30%: testing dataset
  - **Manually coded** by two research assistants
- Positivity
  - Supervised approach (Pang, Lee & Vaithyanathan, 2002):
    - Unigram + POS (part-of-speech) tag → probability of a sentence is positive (Ghose and Ipeirotis, 2011)
    - Then averaged across all sentences → positivity of the whole description
- Objectivity:
  - Classifier based on Barbosa and Feng (2010): polarity words, modal words, etc.
  - Sentence level probability of objectivity; then averaged across sentences



Deception Cues

- Nonstrategic linguistic cues
  - Cognitive load
    - Concreteness (higher if fabricated)
      - Mean of content word concreteness
        - MRC psycholinguistic database
  - Internal imaginations
    - Temporal and spatial information (lower if fabricated)
      - Temporal info
        - SUTime parser and LIWC (Time)
      - Spatial info
        - Stanford name entity recognizer and LIWC (space)
  - Negative emotion
    - Content negation word (“not” “never”)
    - Functional negation word (semantically negative)
- Strategic linguistic cues
  - Dissociation
    - Non-first person pronouns (“he” “her”)
      - % of non-first person pronouns

Control Variables

- All observed information about borrowers and auctions
  - **Hard credit information**, e.g., credit grade, debt-to-income ratio
  - **Auction information**, e.g., loan amount, loan category
  - **Social / soft information**, e.g., group membership and friend investment
  - **Monthly dummies**

Default Probability Models

- **Model 1: (Readability)**

$Probability\ (Default_i{=}1) = \alpha_0 + \alpha_1 \times Readability_i + \alpha_2 \times ControlVariables_{\ i} + \varepsilon_i$

- **Model 2: (Model 1 + Positivity)**

$Probability\ (Default_i{=}1) = \alpha_0 + \alpha_1 \times Readability_i + \alpha_2 \times Positivity_i + \alpha_3 \times ControlVariables + \varepsilon_i$

- **Model 3: (Model 2 + Objectivity)**

$Probability\ (Default_i{=}1) = \alpha_0 + \alpha_1 \times Readability_i + \alpha_2 \times Positivity_i + \alpha_3 \times Objectivity_i + \alpha_4 \times ControlVariables_{\ i} + \varepsilon_i$

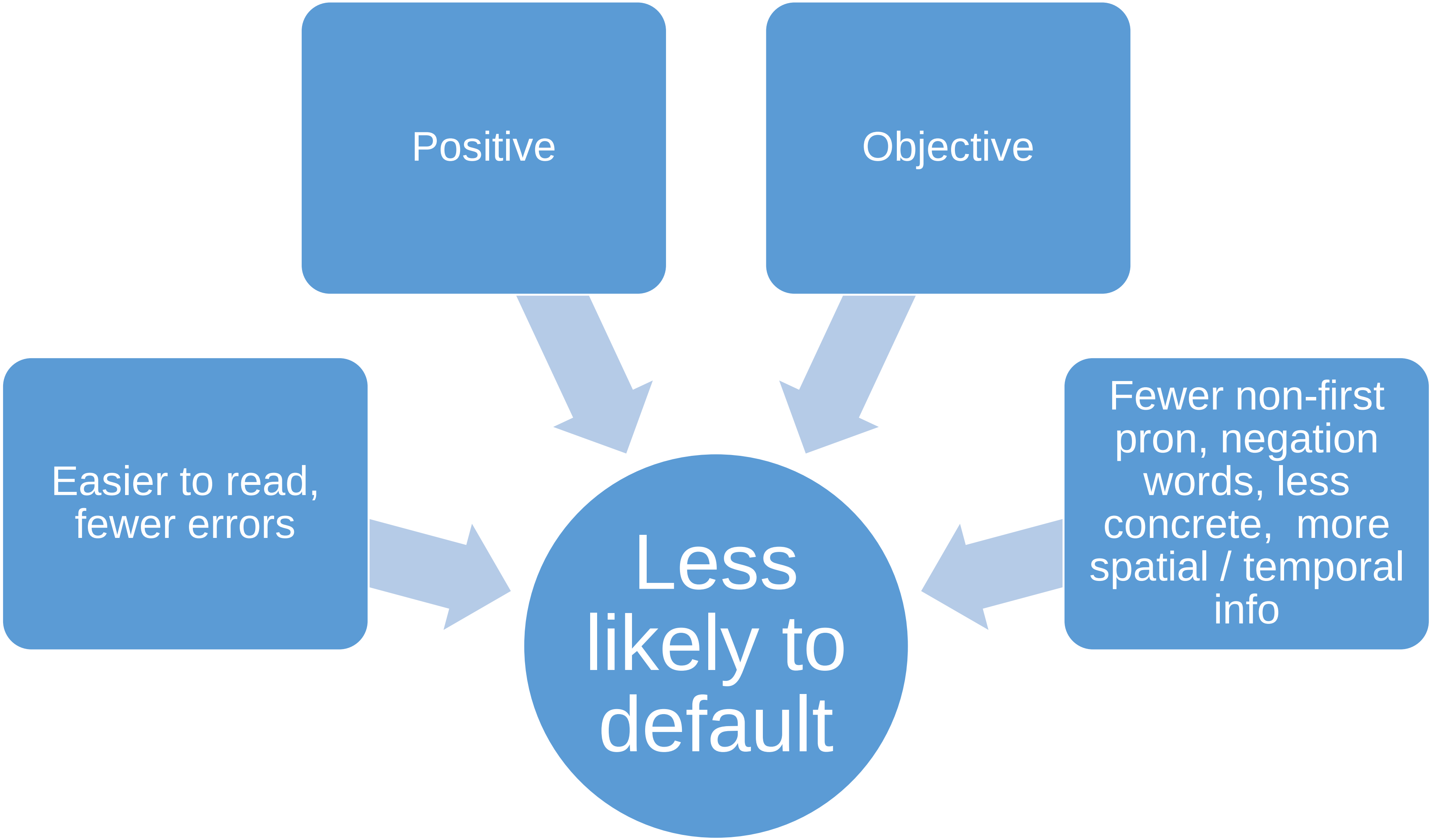
- **Model 4: (Model 3 + Deception Cues)**

$Probability\ (Default_i{=}1) = \alpha_0 + \alpha_1 \times Readability_i + \alpha_2 \times Positivity_i + \alpha_3 \times Objectivity_i + \alpha_4 \times Deception_i + \alpha_5 \times ControlVariables_{\ i} + \varepsilon_i$

Findings

Table 2. Key Findings of Explanatory Analyses

Hypothesis	Relation	Finding	Comments
H1	Readability - Default Rate	Supported	Requests that are less lexical ease of read and have less spelling and grammatical errors are less likely to default.
H2	Positivity - Default Rate	Partially supported	Positive requests are less likely to default, though we did not find evidence of a curvilinear relationship
H3	Objectivity - Default Rate	Supported	Objective requests are less likely to default.
H4	Deception - Default Rate	Supported	Requests that contain more non-1st person pronouns, more negation words, less spatial and temporal information and that are higher in concreteness are more likely to defaults.



Positive

Objective

Easier to read, fewer errors

Fewer non-first pron, negation words, less concrete, more spatial / temporal info

Less likely to default

Robustness & Generality of Explanatory Model

- **Instrument**: linguistic features of borrowers’ friends’ texts
- Replicating our model using **data from LendingClub.com**
  - Only exception is grammatical errors – **texts on LC shorter** (average 46 words) than Prosper (average 135 words)
- **Loan loss percentage** as an alternative outcome variable
- **Content** of Texts
  - Latent Dirichlet Allocation (LDA) topic modeling approach, c.f., Blei et al. (2003)
  - Six major topics: Expenses and income / education, employment, business, family, and credit history.
  - Results robust when adding content dummies

Linguistic Features and Lender Behaviors

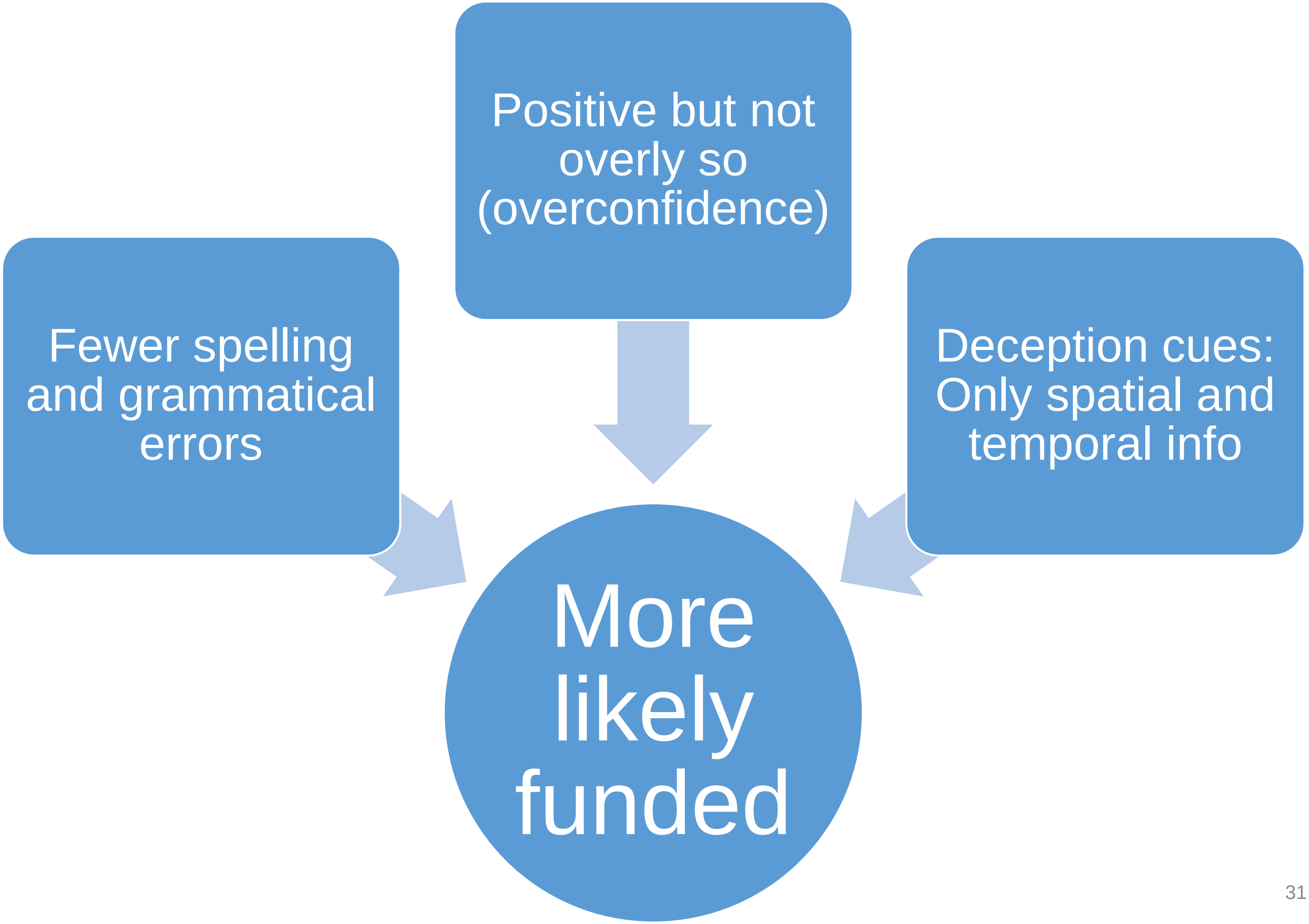
--- Is the market “*linguistically efficient*”?

Do Lenders Correctly Interpret Linguistic Features?

- If lenders are able to correctly predict, then what predicts lower repayment should also predict lower likelihood of funding

$$Probability\ (Funded=1) = \beta_0 + \beta_1 \times Readability_i + \beta_2 \times Sentiment_i + \beta_3 \times Subjectivity_i + \beta_4 \times Deception_i + \beta_5 \times ControlVariables\ _i + \zeta_i$$

What investors interpret correctly:



Positive but not overly so (overconfidence)

Fewer spelling and grammatical errors

Deception cues: Only spatial and temporal info

More likely funded

What’s not interpreted correctly?

- Deception cues:
  - Non-first person pronouns
  - Negation words
- Objectivity
  - Swayed by emotions (c.f., Lin & Viswanathan 2015)

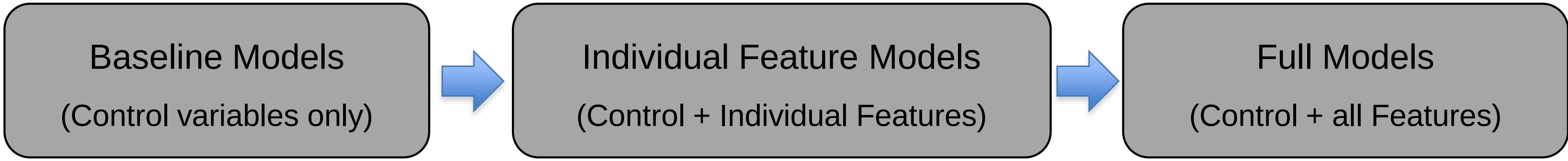
- Potential for efficiency gains, e.g., market design, investor education

Predicting Loan Default Using Linguistic Features

--- Can we help investors interpret better?

Predictive Power of Linguistic Features

- Approach
  - Based on regression approach
  - 10-fold cross evaluation
  - Performance evaluation: area under ROC curve (AUR)



Baseline Models

(Control variables only)

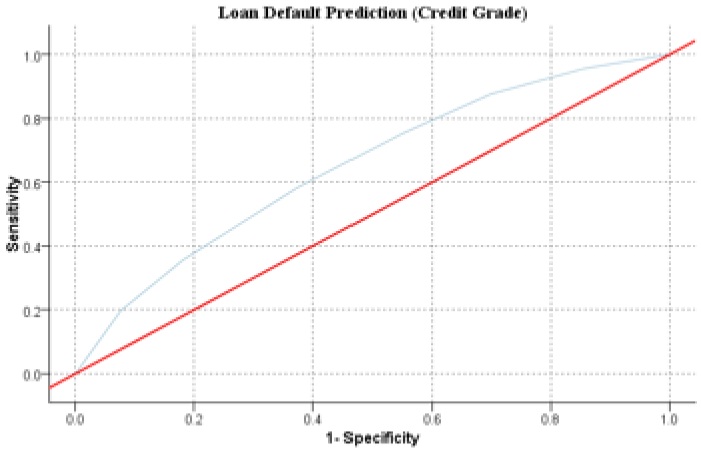
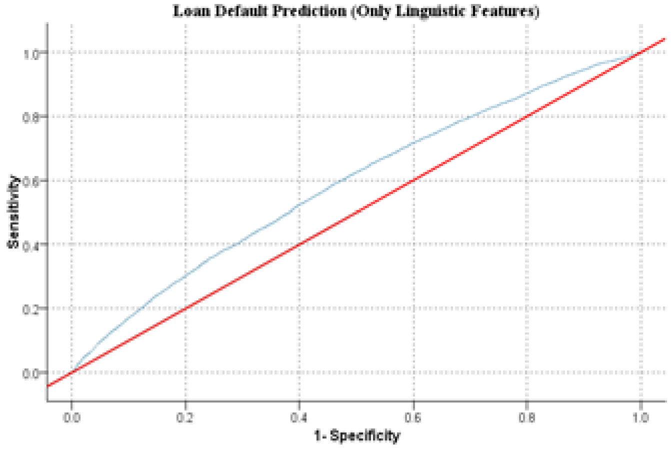
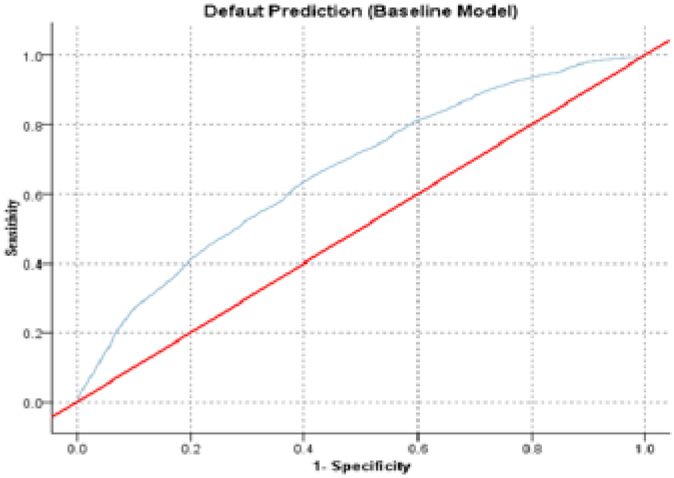
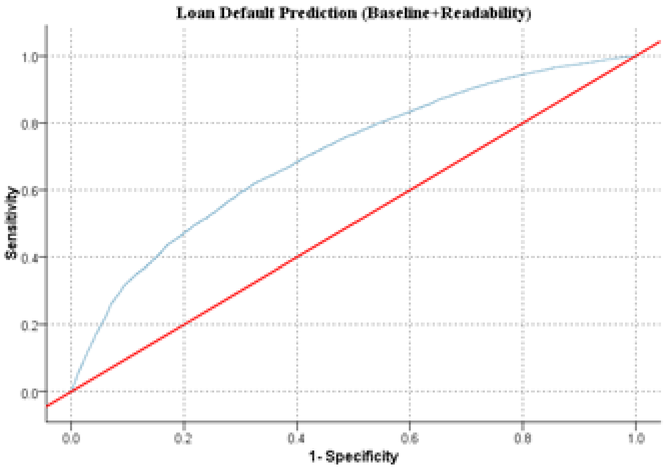
Individual Feature Models

(Control + Individual Features)

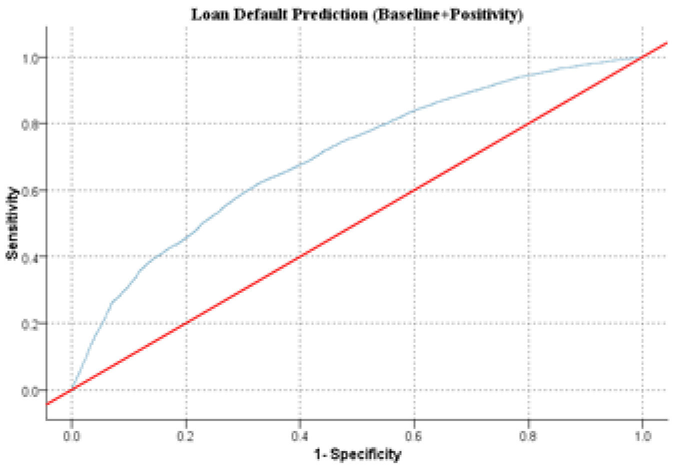
Full Models

(Control + all Features)

Table 11: Predictive Analysis Results

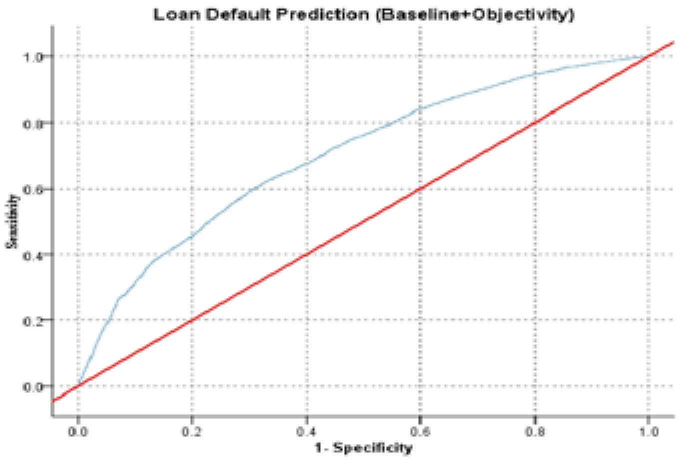
Credit Grade	**Linguistic Dimension Only**
	
AUC:0.635	AUC:0.59
Baseline Model	**Baseline + Readability**
	
AUC:0.682	AUC:0.702

Baseline + Positivity



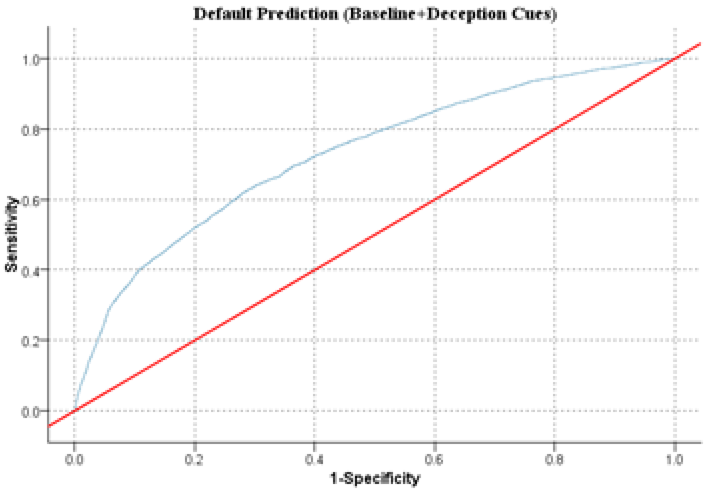
AUC:0.707

Baseline + Objectivity



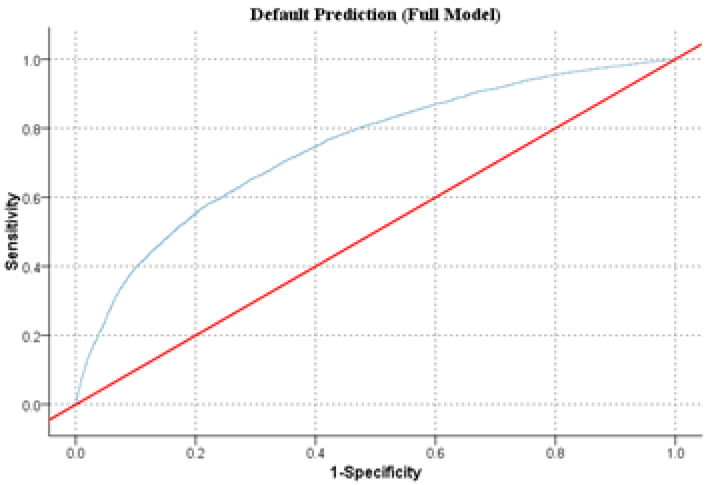
AUC:0.7

Baseline + Deception Cues



AUC:0.718

Full Model



AUC:0.724

Findings from Predictive Model

- **Best if baseline + all linguistic feature dimensions**
- Single dimension: **best if baseline + deception cues**
  - C.f. explanatory model: largest marginal effect
- Baseline + deception cues ➔ outright “fraud” (?): immediately defaulted in the first month after loan origination
  - 5.19% loans

Summary

- Texts, especially linguistic features, contain valuable information about loan quality.

- Investors *do* take texts into account.

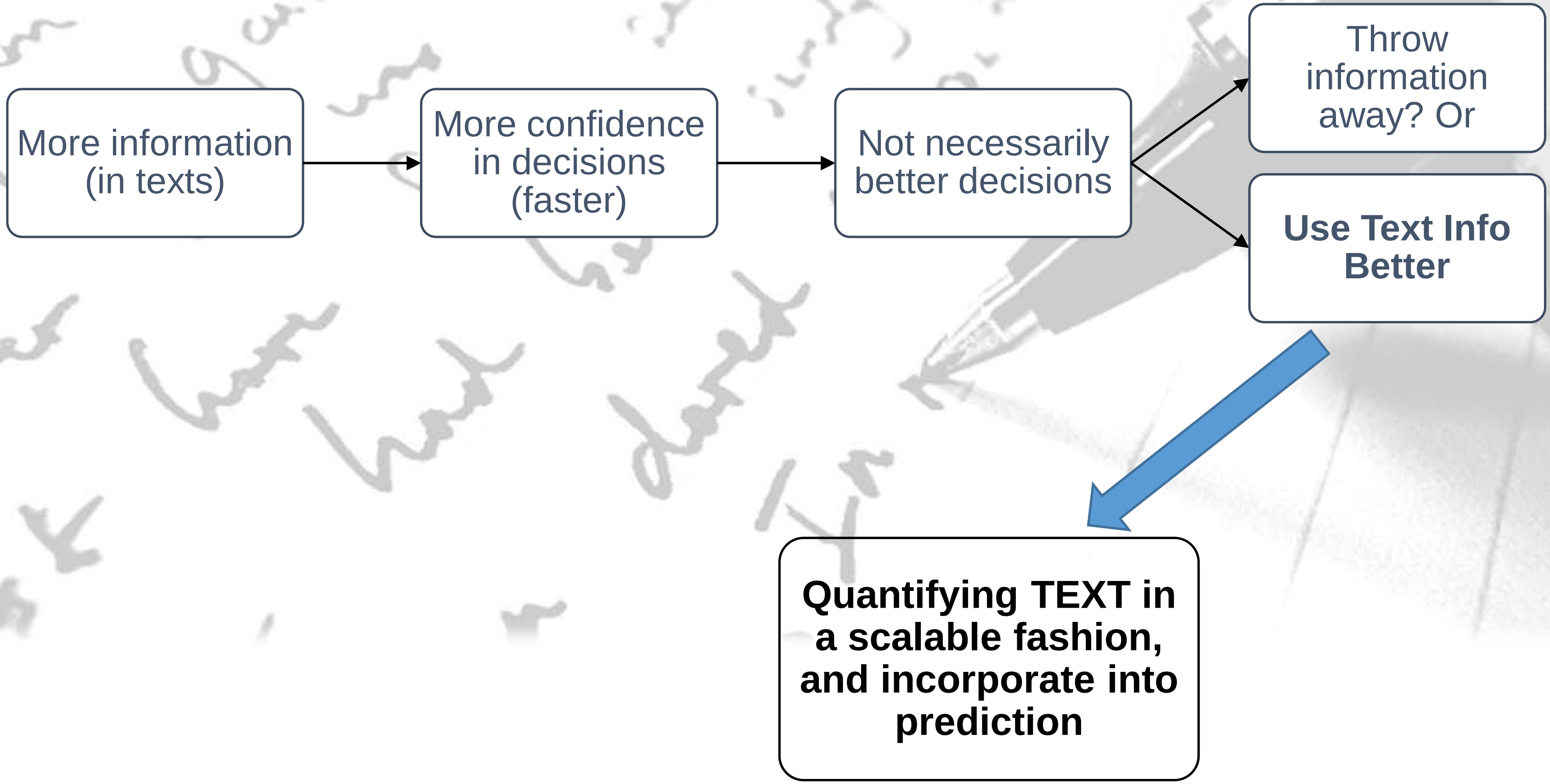
- Investors do not interpret all aspects of linguistic features correctly. In particular, they still fall victim to deception cues.

- Potential mitigation through better prediction: incorporating linguistic features.

Implications

- Having texts is better.
    - Other types of crowdfunding (ongoing research)
    - Or even offline contexts
- Automated linguistic feature extraction

- Design of crowdfunding platforms (e.g., pre-screening)
- Investor education
- Arbitrage opportunities
    - Quantifying “soft” information
- Borrower?



More information (in texts)

More confidence in decisions (faster)

Not necessarily better decisions

Throw information away? Or

Use Text Info Better

Quantifying TEXT in a scalable fashion, and incorporate into prediction

http://ssrn.com/abstract=2446114

Thank you!