DEUTSCHE BUNDESBANK
EUROSYSTEM

Expericenes of the Deutsche Bundesbank

Session 3. Financial information infrastructures based on microdata

Prof. Stefan Bender, Research Data and Service Center, Deutsche Bundesbank

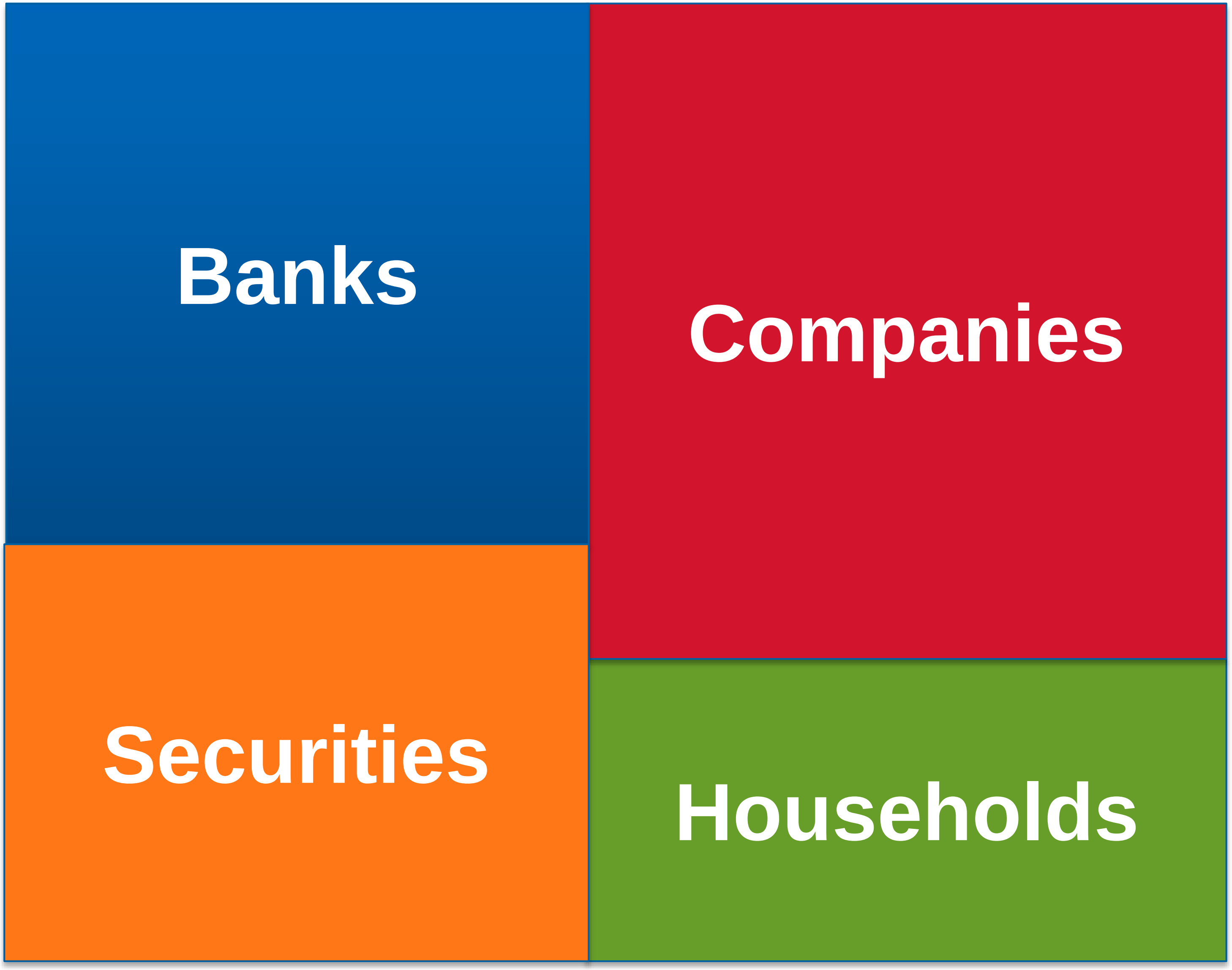
Financial Information Forum of Latin American and the Caribbean Central Banks
V Meeting
28 and 29 May 2019
Lima, Peru

The views expressed here do not necessarily reflect the opinion of the Deutsche Bundesbank or the Eurosystem.

Overview

- Microdata at the Bundesbank
- IMDIAS
- Knowledge Life Cycle (IDIS-R)
- Use of „New Techniques“
  - Data Quality
  - Record Linkage
  - Rich Context
- Conclusion

Available microdata at the RDSC



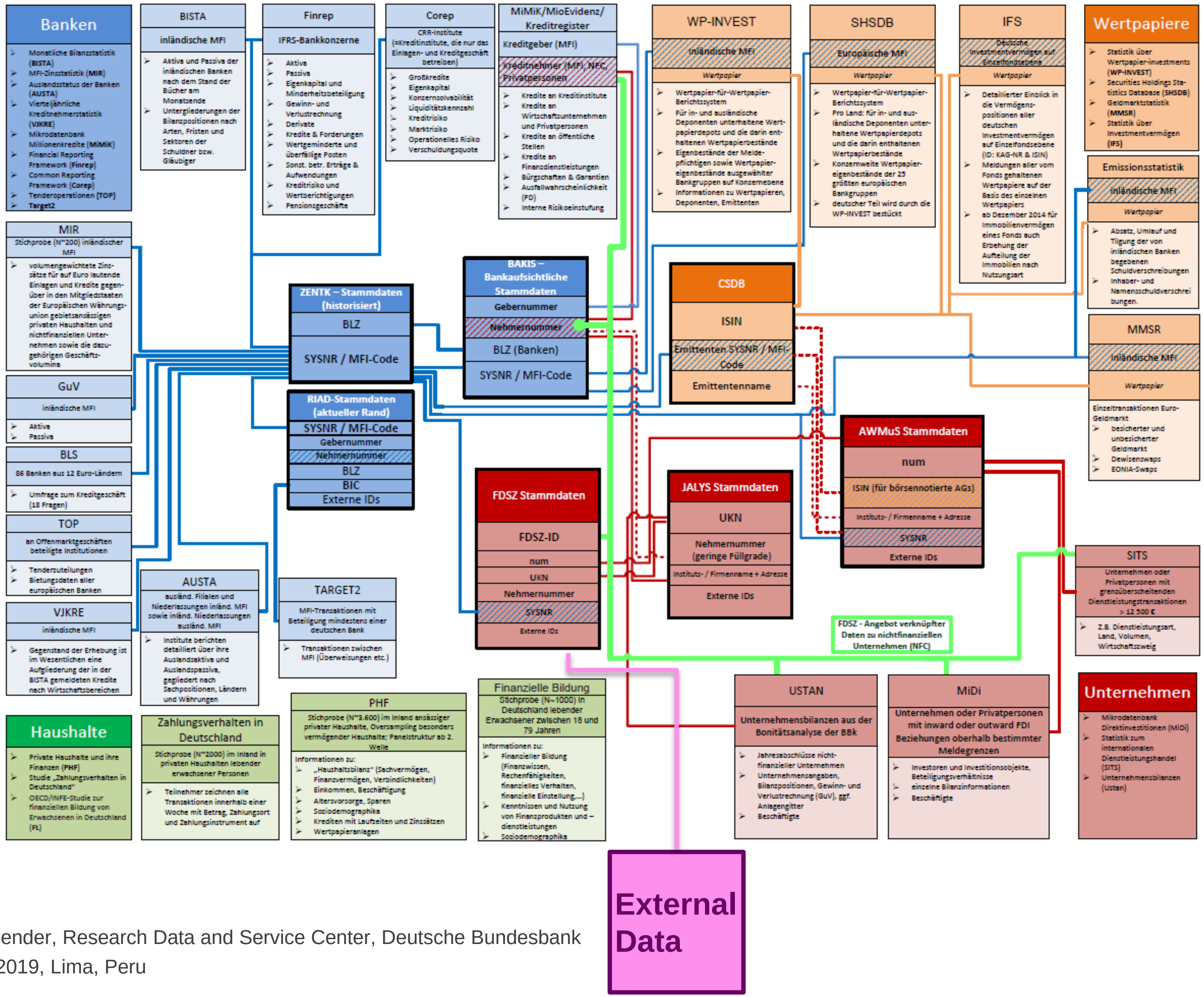
Banks

Companies

Securities

Households

Bundesbank's relevant microdata sources and their connections (excerpt)



Banken

- Monatliche Bilanzstatistik (BISTA)
- MFI-Zinsstatistik (MIR)
- Auslandsstatus der Banken (AUSTA)
- Vierteljährliche Kreditnehmerstatistik (VJKRE)
- Mikrodatenbank Millionenkredite (MiMiK)
- Financial Reporting Framework (Finrep)
- Common Reporting Framework (Corep)
- Tenderoperationen (TOP)
- Target2

BISTA

inländische MFI

- Aktiva und Passiva der inländischen Banken nach dem Stand der Bücher am Monatsende
- Untergliederungen der Bilanzpositionen nach Arten, Fristen und Sektoren der Schuldner bzw. Gläubiger

Finrep

IFRS-Bankkonzerne

- Aktiva
- Passiva
- Eigenkapital und Minderheitsbeteiligung
- Gewinn- und Verlustrechnung
- Derivate
- Kredite & Forderungen
- Wertgeminderte und überfällige Posten
- Sonst. betr. Erträge & Aufwendungen
- Kreditrisiko und Wertberichtigungen
- Pensionsgeschäfte

Corep

CRR-Institute (=Kreditinstitute, die nur das Einlagen- und Kreditgeschäft betreiben)

- Großkredite
- Eigenkapital
- Konzernsolvabilität
- Liquiditätskennzahl
- Kreditrisiko
- Marktrisiko
- Operationelles Risiko
- Verschuldungsquote

MiMiK/MioEvidenz/ Kreditregister

Kreditgeber (MFI)

Kreditnehmer (MFI, NFC, Privatpersonen

- Kredite an Kreditinstitute
- Kredite an Wirtschaftsunternehmen und Privatpersonen
- Kredite an öffentliche Stellen
- Kredite an Finanzdienstleistungen
- Bürgschaften & Garantien
- Ausfallwahrscheinlichkeit (PD)
- Interne Risikoeinstufung

WP-INVEST

inländische MFI

Wertpapier

- Wertpapier-für-Wertpapier-Berichtssystem
- Für in- und ausländische Deponenten unterhaltene Wertpapierdepots und die darin enthaltenen Wertpapierbestände
- Eigenbestände der Meldepflichtigen sowie Wertpapiereigenbestände ausgewählter Bankgruppen auf Konzernebene
- Informationen zu Wertpapieren, Deponenten, Emittenten

SHSDB

Europäische MFI

Wertpapier

- Wertpapier-für-Wertpapier-Berichtssystem
- Pro Land: für in- und ausländische Deponenten unterhaltene Wertpapierdepots und die darin enthaltenen Wertpapierbestände
- Konzernweite Wertpapiereigenbestände der 25 größten europäischen Bankgruppen
- deutscher Teil wird durch die WP-INVEST bestückt

IFS

Deutsche Investmentvermögen auf Einzelfondsebene

Wertpapier

- Detaillierter Einblick in die Vermögenspositionen aller deutschen Investmentvermögen auf Einzelfondsebene (ID: KAG-NR & ISIN)
- Meldungen aller vom Fonds gehaltenen Wertpapiere auf der Basis des einzelnen Wertpapiers
- ab Dezember 2014 für Immobilienvermögen eines Fonds auch Erbehung der Aufteilung der Immobilien nach Nutzungsart

Wertpapiere

- Statistik über Wertpapier-investments (WP-INVEST)
- Securities Holdings Statistics Database (SHSDB)
- Geldmarktstatistik (MMSR)
- Statistik über Investmentvermögen (IFS)

Emissionsstatistik

inländische MFI

Wertpapier

- Absatz, Umlauf und Tilgung der von inländischen Banken begebenen Schuldverschreibungen
- Inhaber- und Namensschuldverschreibungen.

MMSR

Inländische MFI

Wertpapier

Einzeltransaktionen Euro-Geldmarkt

- besicherter und unbesicherter Geldmarkt
- Dewisenswaps
- EONIA-Swaps

MIR

Stichprobe (N~200) inländischer MFI

- volumengewichtete Zinssätze für auf Euro lautende Einlagen und Kredite gegenüber in den Mitgliedstaaten der Europäischen Währungsunion gebietsansässigen privaten Haushalten und nichtfinanziellen Unternehmen sowie die dazugehörigen Geschäftsvolumina

ZENTK – Stammdaten (historisiert)

BLZ

SYSNR / MFI-Code

BAKIS – Bankaufsichtliche Stammdaten

Gebernummer

Nehmernummer

BLZ (Banken)

SYSNR / MFI-Code

CSDB

ISIN

Emittenten SYSNR / MFI-Code

Emittentenname

GuV

inländische MFI

- Aktiva
- Passiva

RIAD-Stammdaten (aktueller Rand)

SYSNR / MFI-Code

Gebernummer

Nehmernummer

BLZ

BIC

Externe IDs

AWMuS Stammdaten

num

ISIN (für börsennotierte AGs)

Instituts- / Firmenname + Adresse

SYSNR

Externe IDs

BLS

86 Banken aus 12 Euro-Ländern

- Umfrage zum Kreditgeschäft (18 Fragen)

TOP

an Offenmarktgeschäften beteiligte Institutionen

- Tenderzuteilungen
- Bietungsdaten aller europäischen Banken

FDSZ Stammdaten

FDSZ-ID

num

UKN

Nehmernummer

SYSNR

Externe IDs

JALYS Stammdaten

UKN

Nehmernummer (geringe Füllgrade)

Instituts- / Firmenname + Adresse

Externe IDs

SITS

Unternehmen oder Privatpersonen mit grenzüberscheitenden Dienstleistungstransaktionen > 12 500 €

- Z.B. Dienstleistungsart, Land, Volumen, Wirtschaftszweig

VJKRE

inländische MFI

- Gegenstand der Erhebung ist im Wesentlichen eine Aufgliederung der in der BISTA gemeldeten Kredite nach Wirtschaftsbereichen

AUSTA

ausländ. Filialen und Niederlassungen inländ. MFI sowie inländ. Niederlassungen ausländ. MFI

- Institute berichten detailliert über ihre Auslandsaktiva und Auslandspassiva, gegliedert nach Sachpositionen, Ländern und Währungen

TARGET2

MFI-Transaktionen mit Beteiligung mindestens einer deutschen Bank

- Transaktionen zwischen MFI (Überweisungen etc.)

FDSZ - Angebot verknüpfter Daten zu nichtfinanziellen Unternehmen (NFC)

Haushalte

- Private Haushalte und ihre Finanzen (PHF)
- Studie „Zahlungsverhalten in Deutschland"
- OECD/INFE-Studie zur finanziellen Bildung von Erwachsenen in Deutschland (FL)

Zahlungsverhalten in Deutschland

Stichprobe (N~2000) im Inland in privaten Haushalten lebender erwachsener Personen

- Teilnehmer zeichnen alle Transaktionen innerhalb einer Woche mit Betrag, Zahlungsort und Zahlungsinstrument auf

PHF

Stichprobe (N~3.600) im Inland ansässiger privater Haushalte, Oversampling besonders vermögender Haushalte; Panelstruktur ab 2. Welle

Informationen zu:

- „Haushaltsbilanz" (Sachvermögen, Finanzvermögen, Verbindlichkeiten)
- Einkommen, Beschäftigung
- Altersvorsorge, Sparen
- Soziodemographika
- Krediten mit Laufzeiten und Zinssätzen
- Wertpapieranlagen

Finanzielle Bildung

Stichprobe (N~1000) in Deutschland lebender Erwachsener zwischen 18 und 79 Jahren

Informationen zu:

- Finanzieller Bildung (Finanzwissen, Rechenfähigkeiten, finanzielles Verhalten, finanzielle Einstellung,…)
- Kenntnissen und Nutzung von Finanzprodukten und –dienstleistungen
- Soziodemographika

USTAN

Unternehmensbilanzen aus der Bonitätsanalyse der BBk

- Jahresabschlüsse nichtfinanzieller Unternehmen
- Unternehmensangaben, Bilanzpositionen, Gewinn- und Verlustrechnung (GuV), ggf. Anlagengitter
- Beschäftigte

MiDi

Unternehmen oder Privatpersonen mit inward oder outward FDI Beziehungen oberhalb bestimmter Meldegrenzen

- Investoren und Investitionsobjekte, Beteiligungsverhältnisse
- einzelne Bilanzinformationen
- Beschäftigte

Unternehmen

- Mikrodatenbank Direktinvestitionen (MiDi)
- Statistik zum internationalen Dienstleistungshandel (SITS)
- Unternehmensbilanzen (Ustan)

External Data

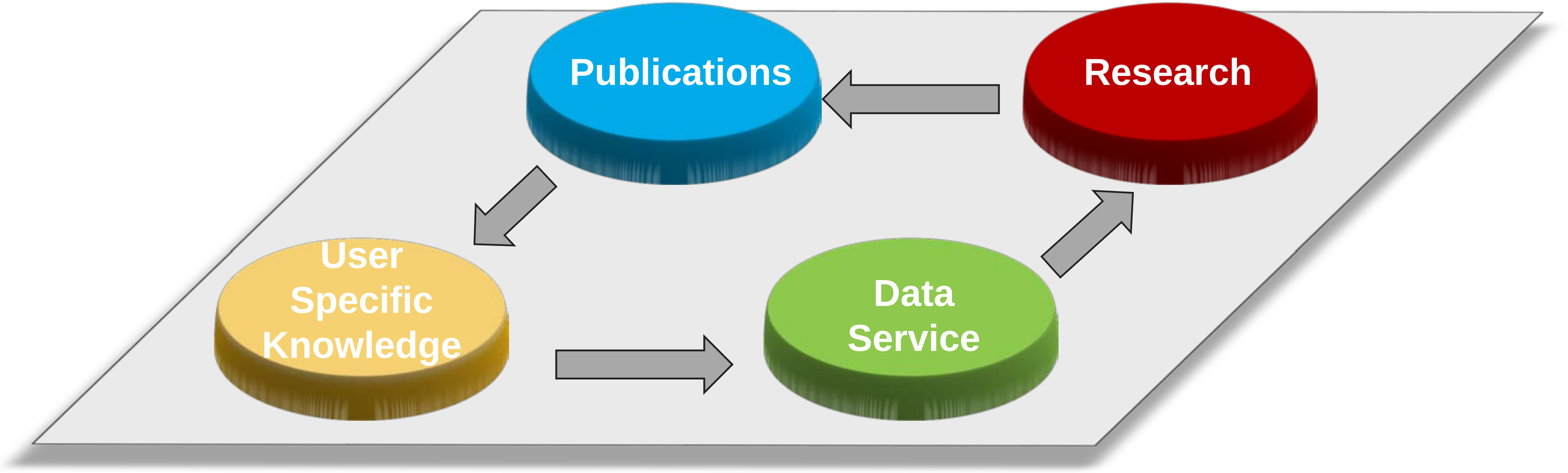
Policy evaluation can make better use of existing datasets

- **The Bundesbank – like other central banks – produces datasets which are highly valuable for policy analysis and research.**
  - So far, most of these datasets have been used to provide aggregate statistics and ad hoc analysis of specific policy issues.
  - There is significant knowledge of data and institutional background.
- **Systematic use of these data for policy analysis is often constrained by**
  - Time
  - IT-resources
  - Legal restrictions
- **The Bundesbank has launched a large-scale initiative aimed at making better use of existing data both, for policy analysis as well as internal and external researchers.**

Integrated MicroData-based Information and Analysis System (IMIDIAS)

- Granular data become more and more important for assessing monetary and regulatory policy as well as for financial stability. Hence the Bundesbank has launched the large-scale initiative IMIDIAS aimed at making better use of existing data both, for policy analysis as well as internal and external researchers.
- **Goals of IMIDIAS:**
  - Support policymaking process
  - Encourage cooperation with (external) researchers
  - Promote evidence-based policy-making

Knowledge Life Cycle (IDIS-R)



Improving Data Quality with Machine Learning (Tobias Cagala, S5)

Securities Holdings Statistics

- German banks provide monthly reports of securities holdings (security-by-security)
- DQM with labor intensive manual case-by-case evaluations

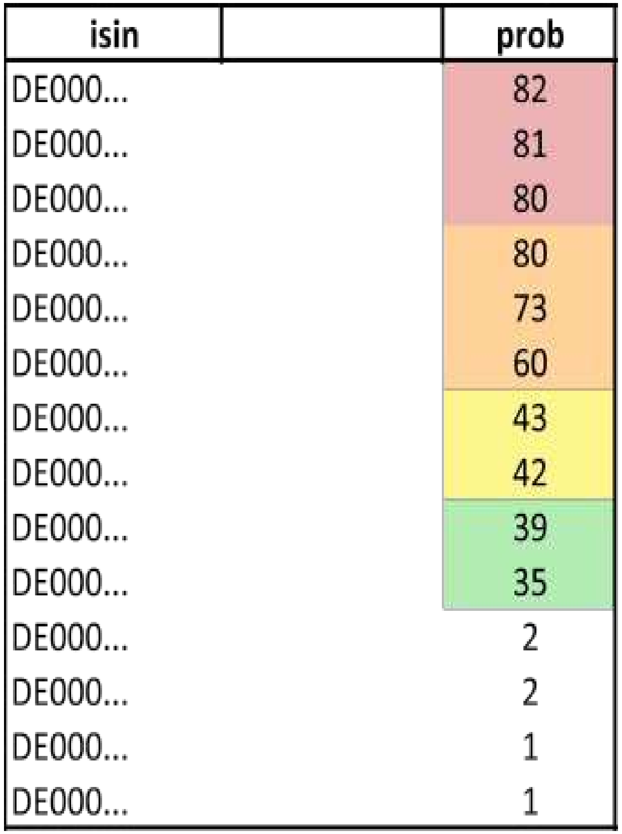
Goal:

- Support compilers by providing predictions of the result of manual checks with machine learning methods

Experience:

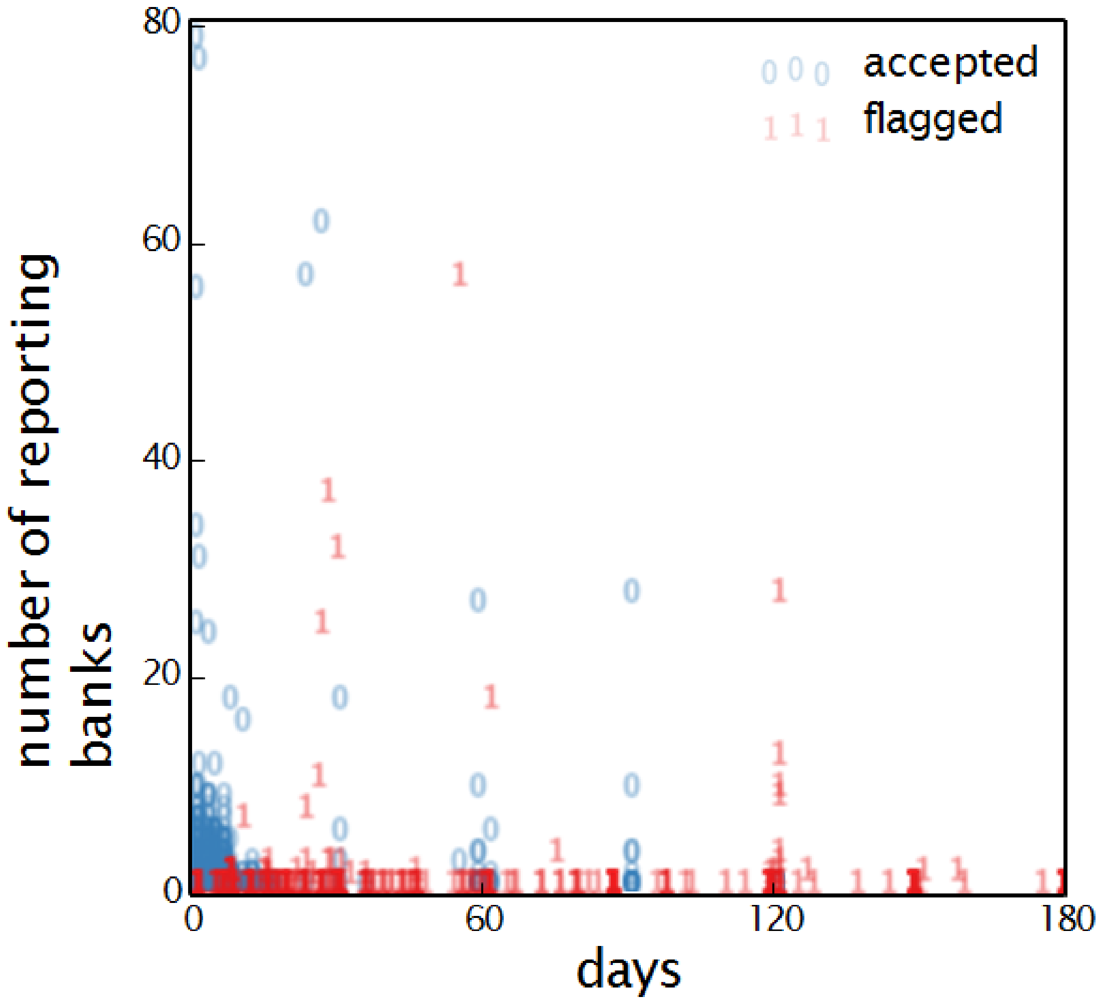
- 50% reduction in time for check and
- increased effectiveness of evaluations

Prediction allows for sorting



isin		prob
DE000...		82
DE000...		81
DE000...		80
DE000...		80
DE000...		73
DE000...		60
DE000...		43
DE000...		42
DE000...		39
DE000...		35
DE000...		2
DE000...		2
DE000...		1
DE000...		1

Descriptive Analysis of Patterns
Number of Reporting Banks, Days since Maturity (Tobias Cagala, S5)



Machine Learning and data linkage (Christopher-Johannes Schild)

Company data (non financial institutions (NFI)):

There is **no common unique firm identifier** in Germany.
(Company business register-ID not stable)

We have to match firm data…

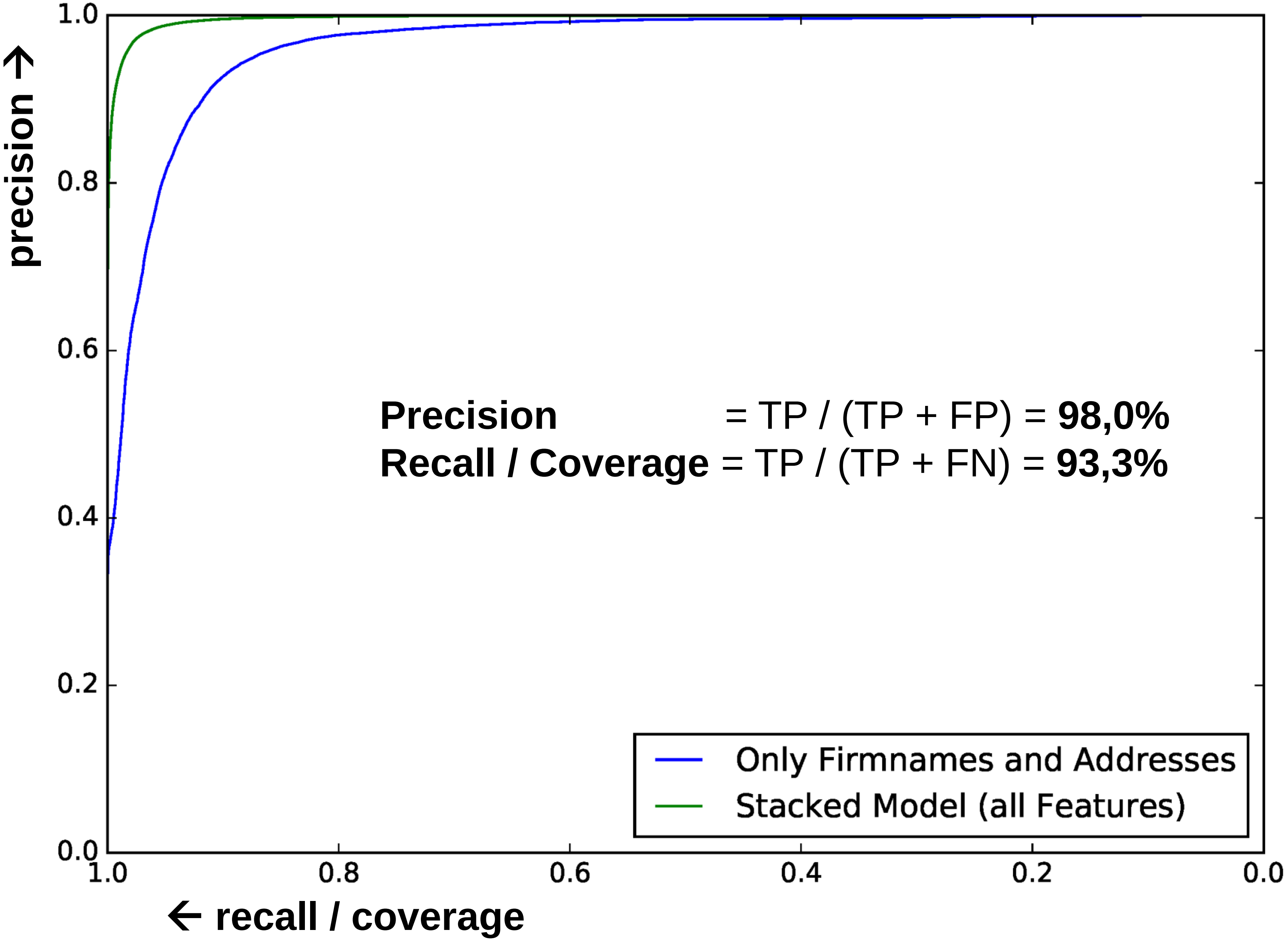
- … that do not have a common unique identifier / key
- … by using **alternative identifiers** (such as names, addresses, sectors, legal forms)

RDSC has matched several NFI-microdatasets (from Statistics, Banking Supervision and external data) with an advanced machine learning algorithm and generated a **matching table** (with probalistic matching scores)

Goal:

- Improve data quality, increase analytical value of data
- More general and flexible Record Linkage System
- Historicized matching tables

Evaluation: Record Linkage of NFI-Microdata (Christopher-Johannes Schild)



Precision = TP / (TP + FP) = **98,0%**
Recall / Coverage = TP / (TP + FN) = **93,3%**

Building a dataset recommendation engine

Experiences from a machine learning competition in text mining

Hendrik Doll, Research Data and Service Centre (RDSC)

Based on a project with and contributions from:
Stefan Bender, Christian Hirsch, Julia-Katharina Ginz, John Chase, Jonathan Morgan, Ian Mulvaney, Andrew Gordon and Julia Lane

The views expressed in this presentation do not necessarily reflect the opinion of the Deutsche Bundesbank or the Eurosystem.

Motivation

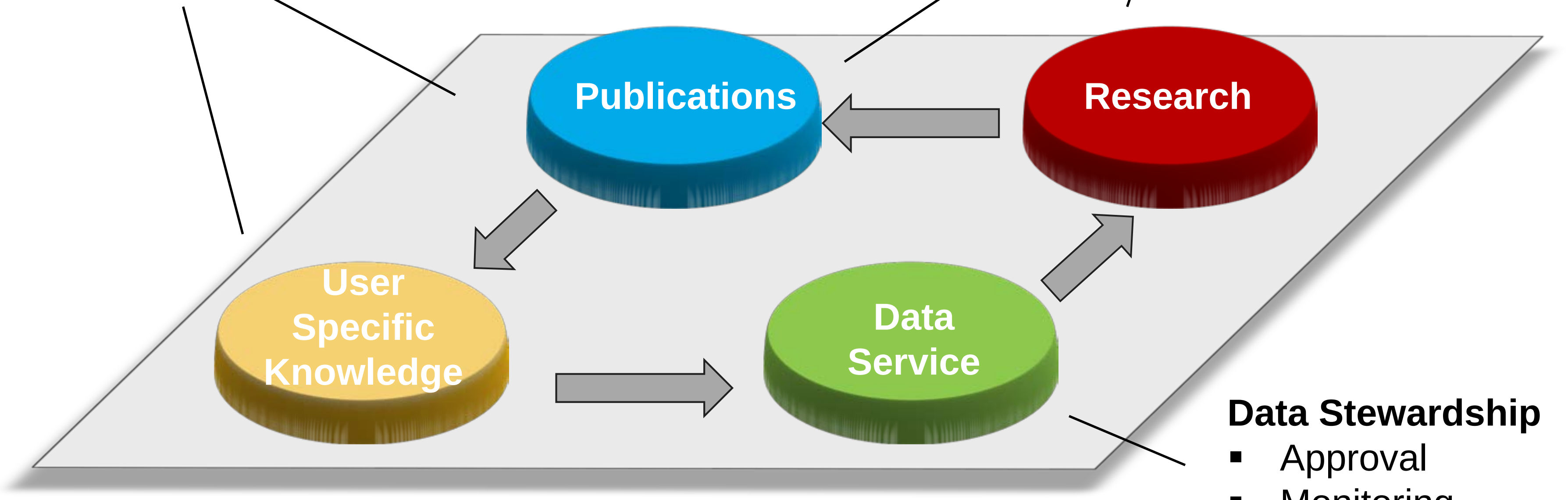
Status Quo: Linear information flow

- Researchers use microdata and metadata to produce outcome (publications)
- No structured knowledge flows back to data producers and data documentation

Goal: Enable feedback loops

- Make knowledge generated in the research process usable
- Two sources of information to extract:
  - Tacit knowledge: Enable and incentivize data users to feedback information
  - Automatic: Text mining to find information in research publications

Framework: Knowledge Life Cycle



Rich Context

Collaboration

- Knowledge sharing
- Metadata

Secure workspace

- Services and Tools

Data Stewardship

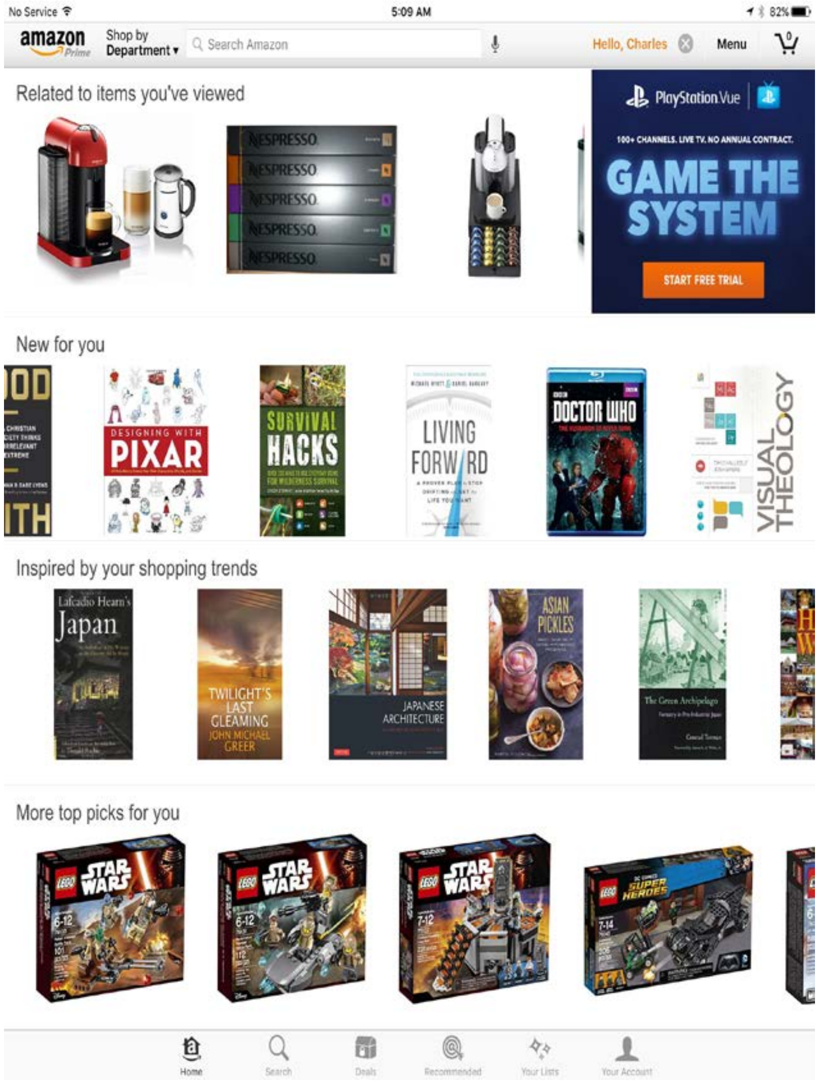
- Approval
- Monitoring
- Reporting

Applications

1. **What is the added value of microdata for research?**
   - What is added value of Bundesbank data? (BISTA vs. Bankscope)
   - Dataset impact factor (Research data is a public good, show outcome to justify societal investment)

2. **Enhance data services for user**
   - Strengthen quantitative research through optimal microdata usage
   - Design of Amazon-type proposition system for researchers (*„What data is out there?“ „What have others done with the data?“*)

Structured knowledge (2)



Related to data you've viewed

New data similar to data you've used

What others have done with similar data (recipes)

Recipes like yours

Thanks to Julia Lane

So how do we do this?

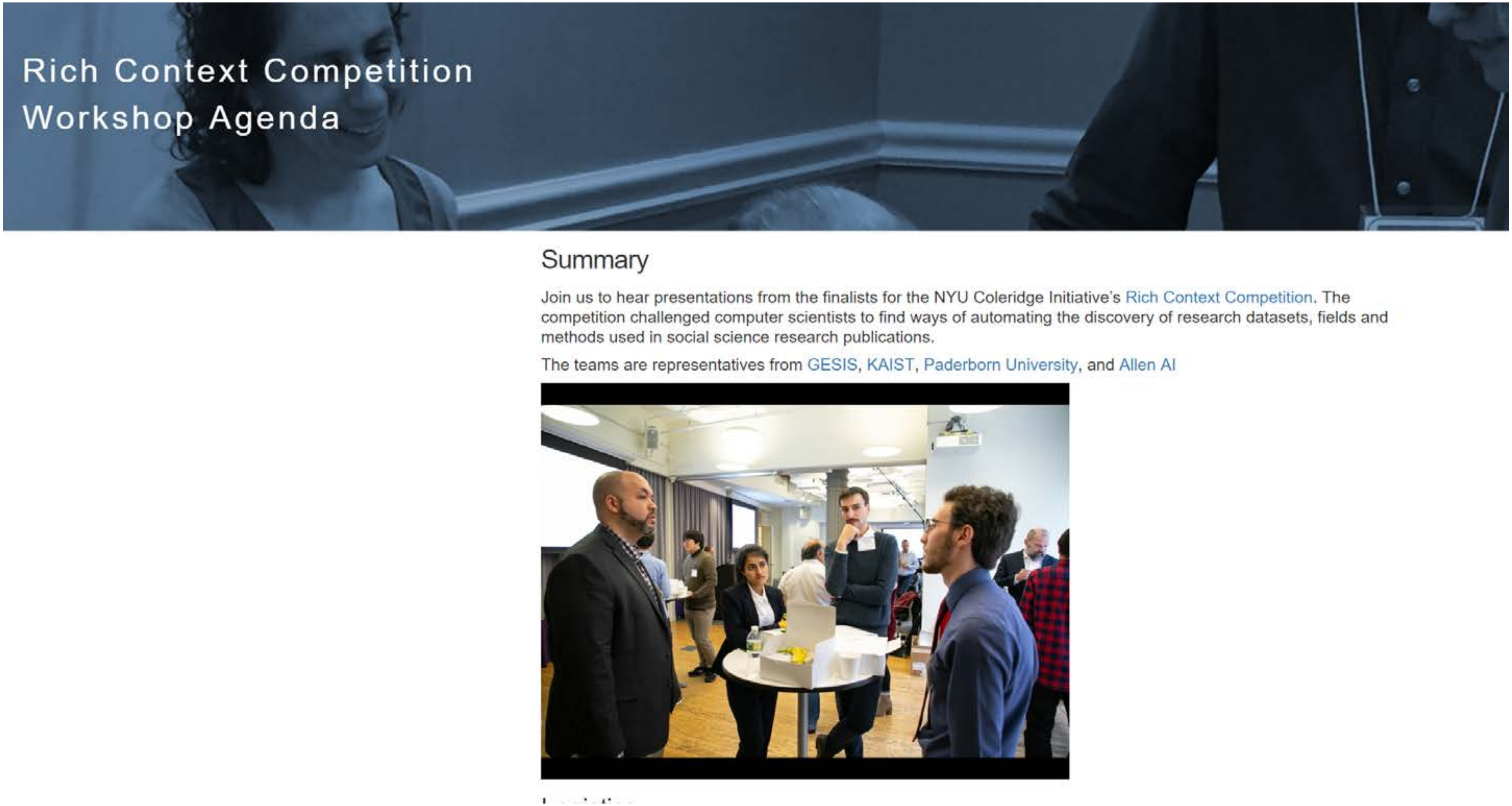
- **Step 1:** Create the set of corpora and metadata (computer science technology) - Competition

- **Step 2:** Figure out how you learn from it and automate it (machine learning techniques) – Engagement

- **Step 3:** Gamification – recognize and emphasize patterns (with human curation) – Rinse and repeat

THX to Hendrik Doll and Christian Hirsch

Run competition (https://coleridgeinitiative.org/richcontextcompetition)



Rich Context Competition
Workshop Agenda

Summary

Join us to hear presentations from the finalists for the NYU Coleridge Initiative's Rich Context Competition. The competition challenged computer scientists to find ways of automating the discovery of research datasets, fields and methods used in social science research publications.

The teams are representatives from GESIS, KAIST, Paderborn University, and Allen AI

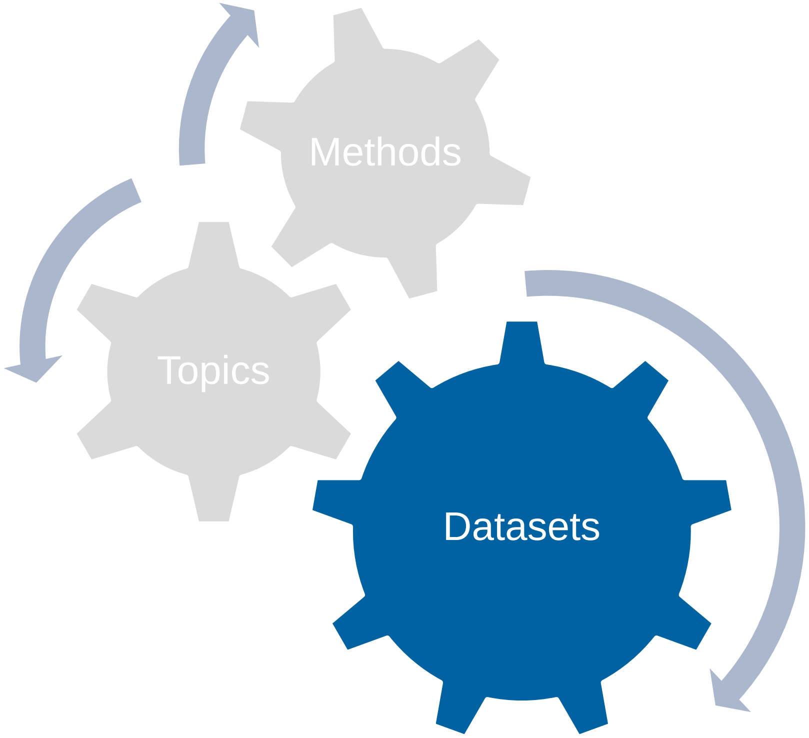
THX to Hendrik Doll and Christian Hirsch

Hand-curate corpus if needed - example

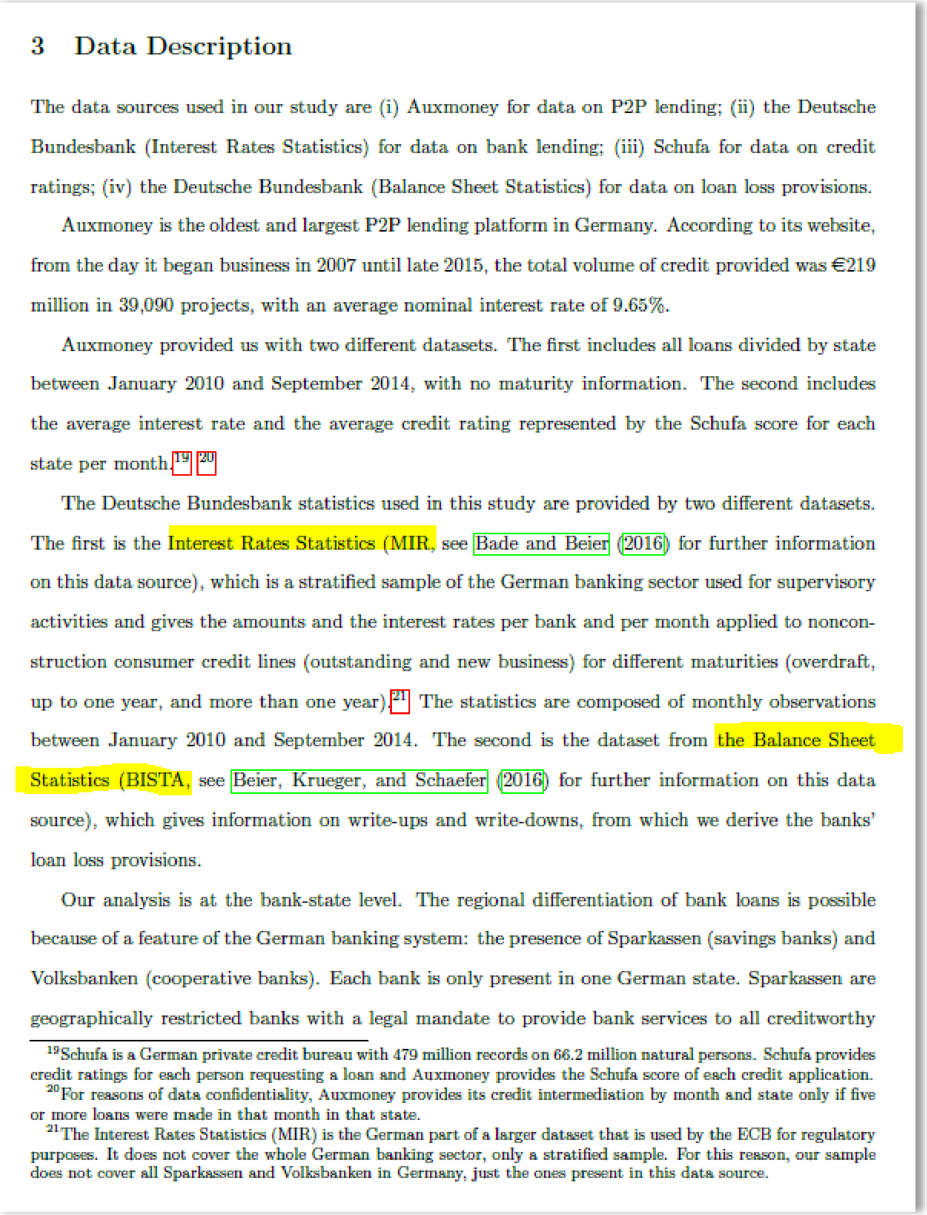
- Training dataset: 5.000 labeled publications with dataset usage, fields and methods

Publication Title	Authors	Data set
Clustering or competition? The foreign investment behaviour of German banks	Lipponer A., Buch C.	MIDI, BISTA, GUV
How will Basel II affect bank lending to emerging markets? An analysis based on German bank level data	Liebig T., Porath D., di Mauro B., Wedow M.	MIMIK
FDI versus cross-border financial services: The globalisation of German banks	Buch C., Lipponer A.	MIDI, BISTA, GUV
German bank lending during emerging market crises: A bank level analysis	Heid F., Nestmann T., di Mauro B., Westernhagen N.	MIMIK

Step 1: competition design



THX to Hendrik Doll and Christian Hirsch



3 Data Description

The data sources used in our study are (i) Auxmoney for data on P2P lending; (ii) the Deutsche Bundesbank (Interest Rates Statistics) for data on bank lending; (iii) Schufa for data on credit ratings; (iv) the Deutsche Bundesbank (Balance Sheet Statistics) for data on loan loss provisions.

Auxmoney is the oldest and largest P2P lending platform in Germany. According to its website, from the day it began business in 2007 until late 2015, the total volume of credit provided was €219 million in 39,090 projects, with an average nominal interest rate of 9.65%.

Auxmoney provided us with two different datasets. The first includes all loans divided by state between January 2010 and September 2014, with no maturity information. The second includes the average interest rate and the average credit rating represented by the Schufa score for each state per month.[19] [20]

The Deutsche Bundesbank statistics used in this study are provided by two different datasets. The first is the Interest Rates Statistics (MIR, see Bade and Beier (2016) for further information on this data source), which is a stratified sample of the German banking sector used for supervisory activities and gives the amounts and the interest rates per bank and per month applied to nonconstruction consumer credit lines (outstanding and new business) for different maturities (overdraft, up to one year, and more than one year).[21] The statistics are composed of monthly observations between January 2010 and September 2014. The second is the dataset from the Balance Sheet Statistics (BISTA, see Beier, Krueger, and Schaefer (2016) for further information on this data source), which gives information on write-ups and write-downs, from which we derive the banks' loan loss provisions.

Our analysis is at the bank-state level. The regional differentiation of bank loans is possible because of a feature of the German banking system: the presence of Sparkassen (savings banks) and Volksbanken (cooperative banks). Each bank is only present in one German state. Sparkassen are geographically restricted banks with a legal mandate to provide bank services to all creditworthy

[19] Schufa is a German private credit bureau with 479 million records on 66.2 million natural persons. Schufa provides credit ratings for each person requesting a loan and Auxmoney provides the Schufa score of each credit application.

[20] For reasons of data confidentiality, Auxmoney provides its credit intermediation by month and state only if five or more loans were made in that month in that state.

[21] The Interest Rates Statistics (MIR) is the German part of a larger dataset that is used by the ECB for regulatory purposes. It does not cover the whole German banking sector, only a stratified sample. For this reason, our sample does not cover all Sparkassen and Volksbanken in Germany, just the ones present in this data source.

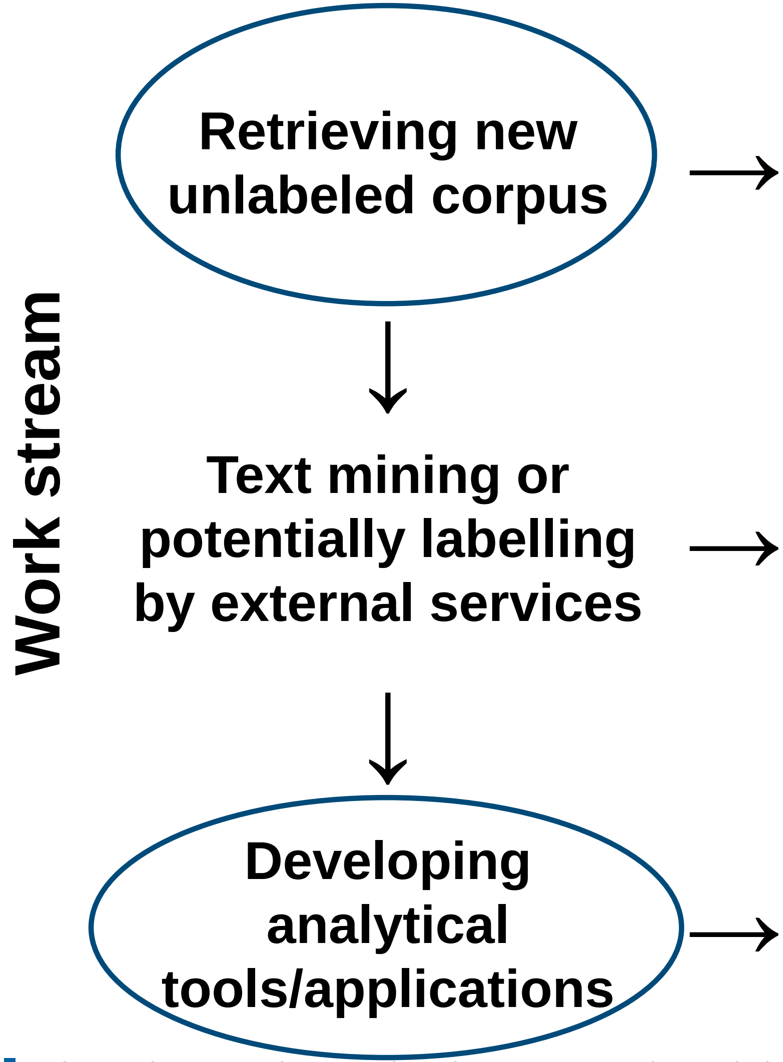
Calebe de Roure – Loriana Pelizzon – Anjan Thakor

P2P Lenders versus Banks:
Cream Skimming or Bottom Fishing?

SAFE Working Paper No. 206

Application: Design of Amazon-type proposition system for researchers

Where we currently are

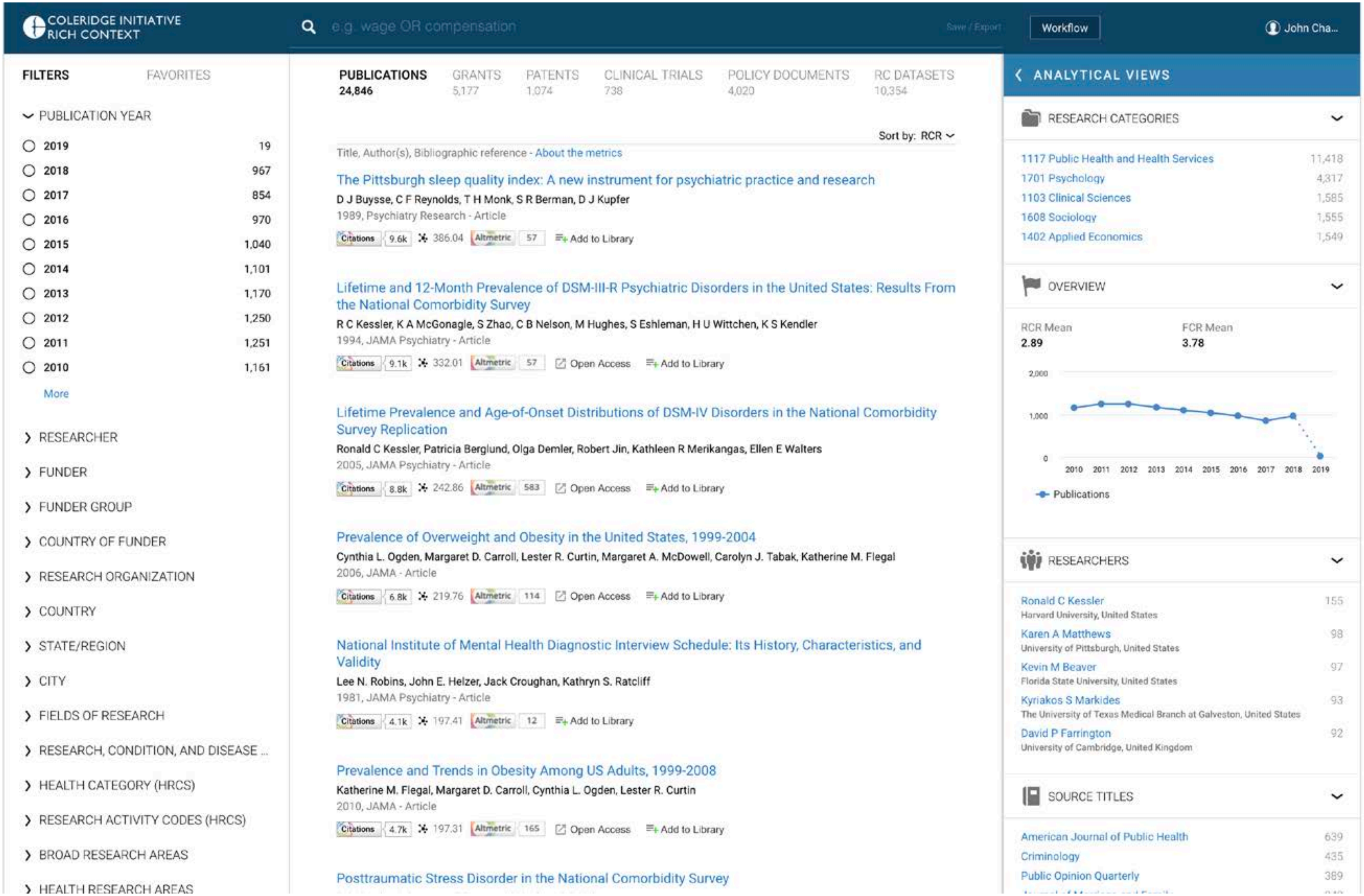

Work stream
Retrieving new unlabeled corpus
Text mining or potentially labelling by external services
Developing analytical tools/applications

- Extending the corpus to broader finance and economics literature
- Special focus: Central bank publications

- Testing the models with new corpus
- Potentially and complementary: Outsourcing the labelling to external provider

Creating prototypes for data set recommendation system and data set impact factor

Use of structured knowledge



COLERIDGE INITIATIVE RICH CONTEXT

e.g. wage OR compensation

Save / Export

Workflow

John Cha...

FILTERS FAVORITES

PUBLICATION YEAR

○ 2019	19
○ 2018	967
○ 2017	854
○ 2016	970
○ 2015	1,040
○ 2014	1,101
○ 2013	1,170
○ 2012	1,250
○ 2011	1,251
○ 2010	1,161

More

- RESEARCHER
- FUNDER
- FUNDER GROUP
- COUNTRY OF FUNDER
- RESEARCH ORGANIZATION
- COUNTRY
- STATE/REGION
- CITY
- FIELDS OF RESEARCH
- RESEARCH, CONDITION, AND DISEASE ...
- HEALTH CATEGORY (HRCS)
- RESEARCH ACTIVITY CODES (HRCS)
- BROAD RESEARCH AREAS
- HEALTH RESEARCH AREAS

PUBLICATIONS	GRANTS	PATENTS	CLINICAL TRIALS	POLICY DOCUMENTS	RC DATASETS
24,846	5,177	1,074	738	4,020	10,354

Sort by: RCR

Title, Author(s), Bibliographic reference - About the metrics

The Pittsburgh sleep quality index: A new instrument for psychiatric practice and research
D J Buysse, C F Reynolds, T H Monk, S R Berman, D J Kupfer
1989, Psychiatry Research - Article
Citations 9.6k 386.04 Altmetric 57 Add to Library

Lifetime and 12-Month Prevalence of DSM-III-R Psychiatric Disorders in the United States: Results From the National Comorbidity Survey
R C Kessler, K A McGonagle, S Zhao, C B Nelson, M Hughes, S Eshleman, H U Wittchen, K S Kendler
1994, JAMA Psychiatry - Article
Citations 9.1k 332.01 Altmetric 57 Open Access Add to Library

Lifetime Prevalence and Age-of-Onset Distributions of DSM-IV Disorders in the National Comorbidity Survey Replication
Ronald C Kessler, Patricia Berglund, Olga Demler, Robert Jin, Kathleen R Merikangas, Ellen E Walters
2005, JAMA Psychiatry - Article
Citations 8.8k 242.86 Altmetric 583 Open Access Add to Library

Prevalence of Overweight and Obesity in the United States, 1999-2004
Cynthia L. Ogden, Margaret D. Carroll, Lester R. Curtin, Margaret A. McDowell, Carolyn J. Tabak, Katherine M. Flegal
2006, JAMA - Article
Citations 6.8k 219.76 Altmetric 114 Open Access Add to Library

National Institute of Mental Health Diagnostic Interview Schedule: Its History, Characteristics, and Validity
Lee N. Robins, John E. Helzer, Jack Croughan, Kathryn S. Ratcliff
1981, JAMA Psychiatry - Article
Citations 4.1k 197.41 Altmetric 12 Add to Library

Prevalence and Trends in Obesity Among US Adults, 1999-2008
Katherine M. Flegal, Margaret D. Carroll, Cynthia L. Ogden, Lester R. Curtin
2010, JAMA - Article
Citations 4.7k 197.31 Altmetric 165 Open Access Add to Library

Posttraumatic Stress Disorder in the National Comorbidity Survey

ANALYTICAL VIEWS

RESEARCH CATEGORIES

1117 Public Health and Health Services	11,418
1701 Psychology	4,317
1103 Clinical Sciences	1,585
1608 Sociology	1,555
1402 Applied Economics	1,549

OVERVIEW

RCR Mean 2.89 FCR Mean 3.78

RESEARCHERS

Ronald C Kessler Harvard University, United States	155
Karen A Matthews University of Pittsburgh, United States	98
Kevin M Beaver Florida State University, United States	97
Kyriakos S Markides The University of Texas Medical Branch at Galveston, United States	93
David P Farrington University of Cambridge, United Kingdom	92

SOURCE TITLES

American Journal of Public Health	639
Criminology	435
Public Opinion Quarterly	389

Take-Aways

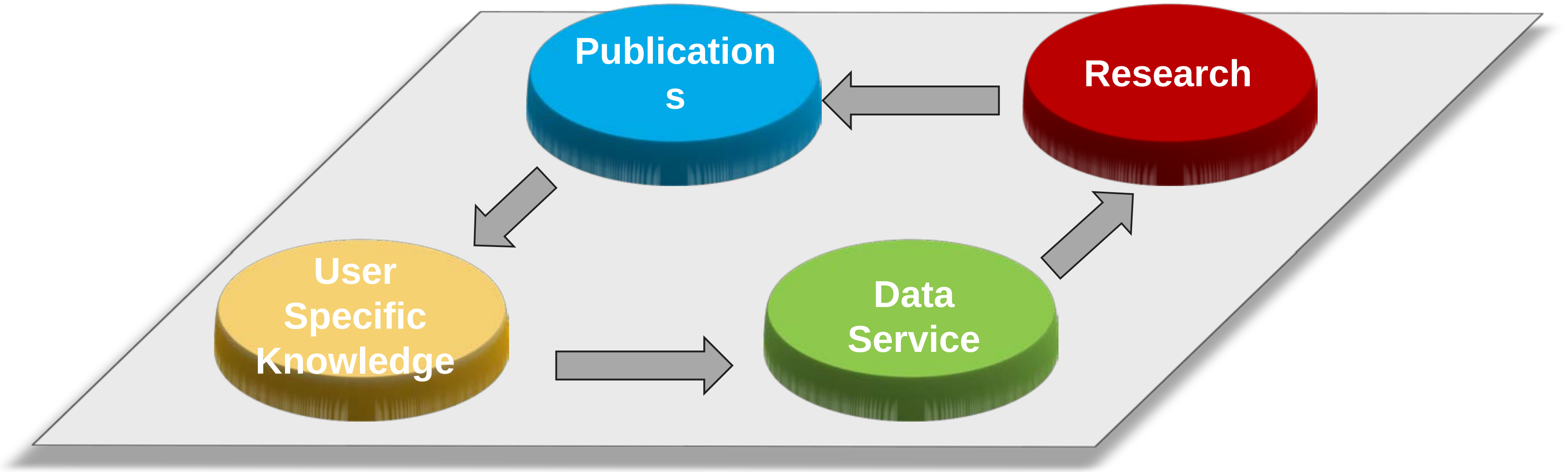
WHAT WE DID

- Close the gap between publications and data services
- Engage users to contribute knowledge to existing metadata

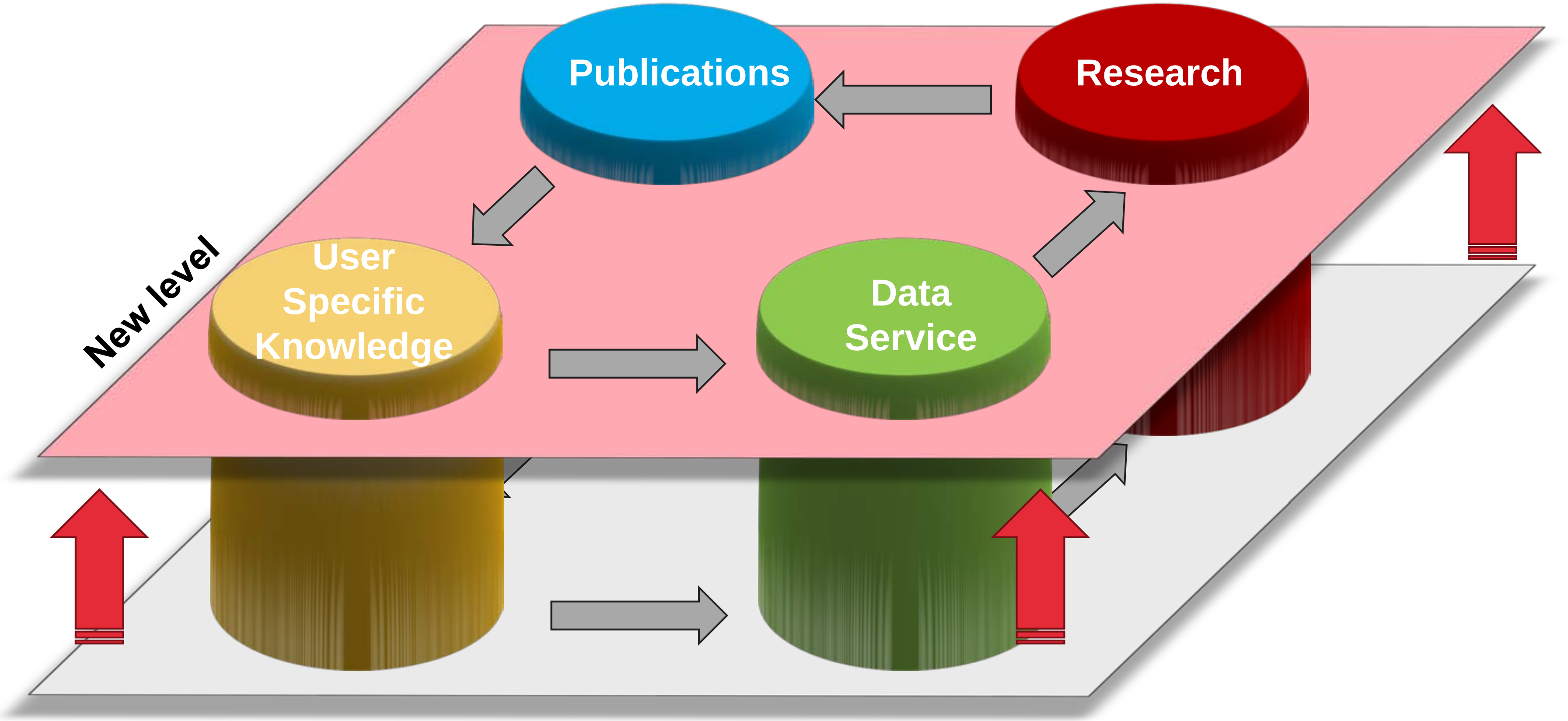
WHAT‘S NEXT

- Digital integrated system around data-services (as a single point of reference for users)

Getting the Knowledge Life Cycle …



… into the next level



Thank you !

- **Website**: www.bundesbank.de\fdsz
- **Contact**: fdsz@bundesbank.de

Other Applications of Bundesbank in Machine Learning

- **Saisonal Adjustment**
  - Is This Time Series Seasonal? - How Random Forests Can Improve Seasonality Tests by Daniel Ollech, Karsten Webel / DG Statistics
- **Identification of Holdings**
  - by Frank Raulf
- **Record Linkage**
  - (we will see in some minutes)