Achieving Deep & Timely Insights into Complex Networks of Information and Knowledge

Nikolay Yakovets

Faculteit Wiskunde en Informatica
Technische Universiteit Eindhoven

AI Lunch — 29 June 2018

Nikolay Yakovets

George Fletcher

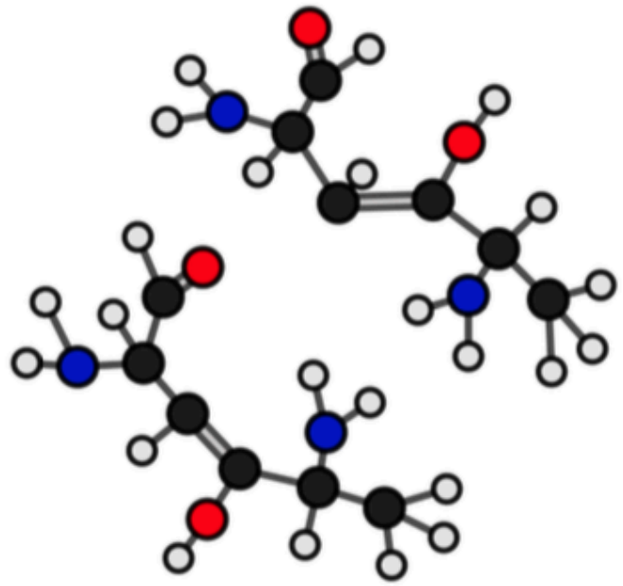
Odysseas Papapetrou

Databases

Why graph data / networks ?

Big graph data sets are everywhere

- social networks (e.g., LinkedIn, Facebook)
- scientific networks (e.g., Uniprot, PubChem)
- knowledge graphs (e.g., DBPedia, MS Academic Graph)
- ...



Focus is on "things" and their relationships

Graph Buzz!

neo technology
graphs are everywhere

MacArthur 'Genius Grant' Winner Maria Chudnovsky on Graph Theory

Wednesday, October 03, 2012

Bright Launches Bright Packed With Jobs Data Seeking Tips

InfoWorld Home / InfoWorld Tech Watch / Buzz grows around graph databases

The First Word on Tech
INFOWORLD TECH WATCH

AUGUST 29, 2012

Buzz grows around graph databases

Interest in graph databases will continue to grow, given its ability to analyze data delivered in a non-relational format, such as social networking data

By Paul Krill | InfoWorld

Follow @pjkrill

Graph Databases: The New Way to Access Super Fast Social Data

September 26, 2012 by Emil Eifrem

1

259 Like

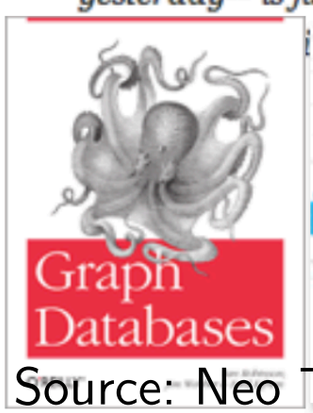
969 Tweet

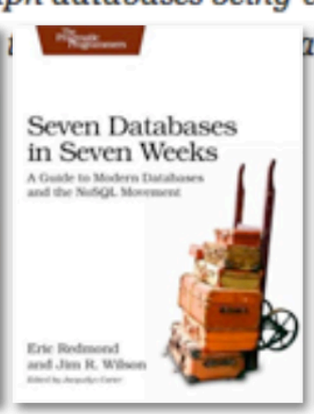
48 +1

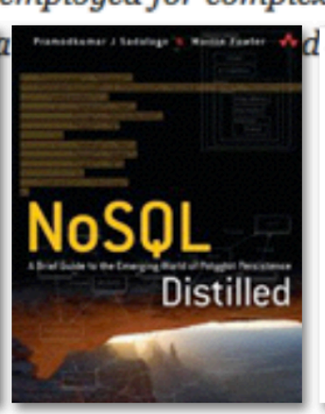
173 Share

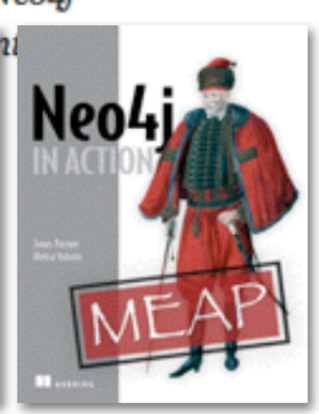
Facebook's Social Graph, Neo4j show rising use of graph databases

Summary: *Facebook's Social Graph -- the database underlying its Graph Search engine unveiled yesterday-- is just one of many graph databases being employed for complex, connected data. Neo4j*









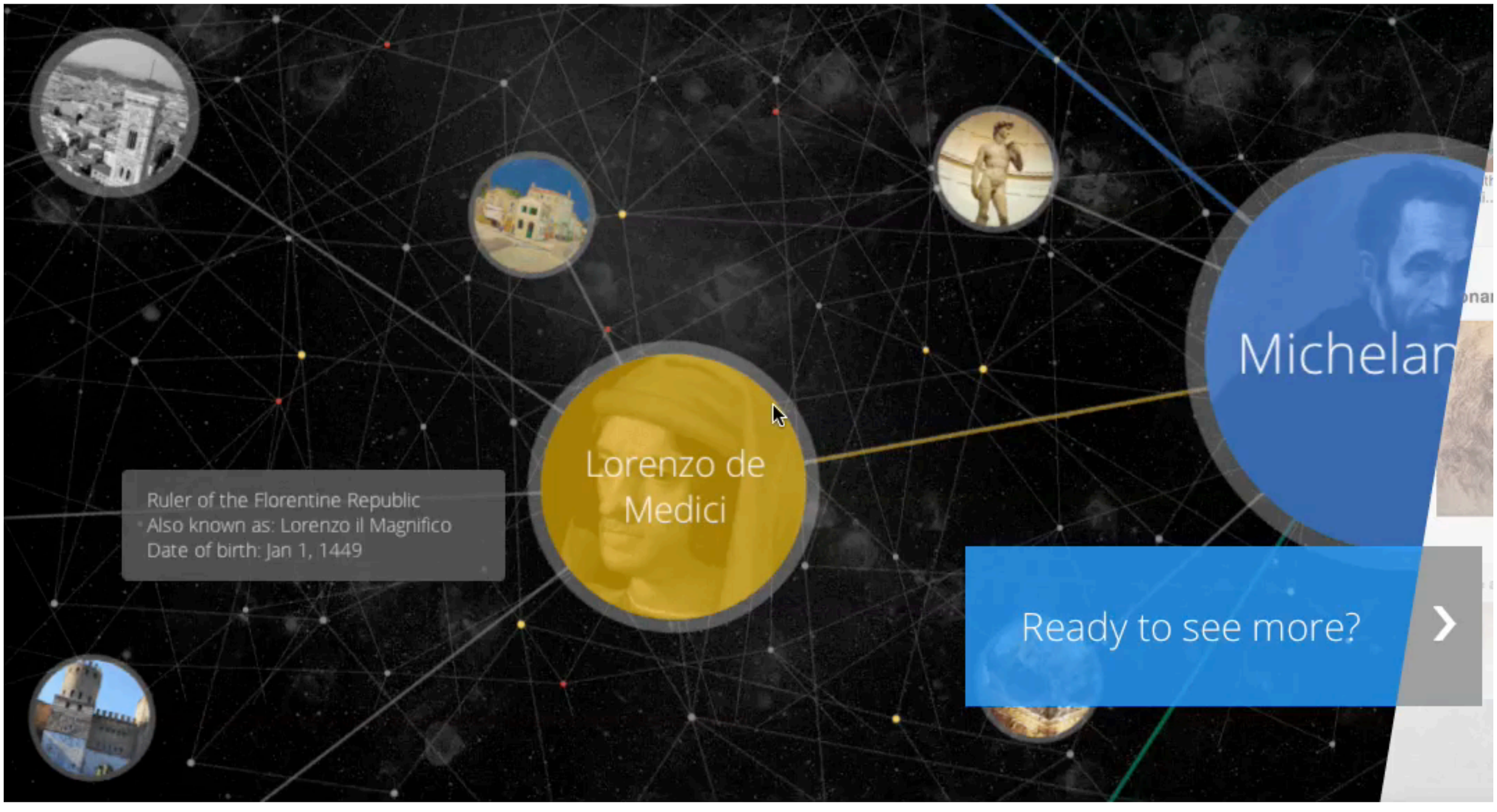
I saw my own Interest Graph and it's scary-accurate. We'd certainly pay for the ability to use the Gravity personalization technology I saw today at TechCrunch to help target content to users.

TC **Michael Arrington, TechCrunch**

Source: Neo Technology

Knowledge Graph

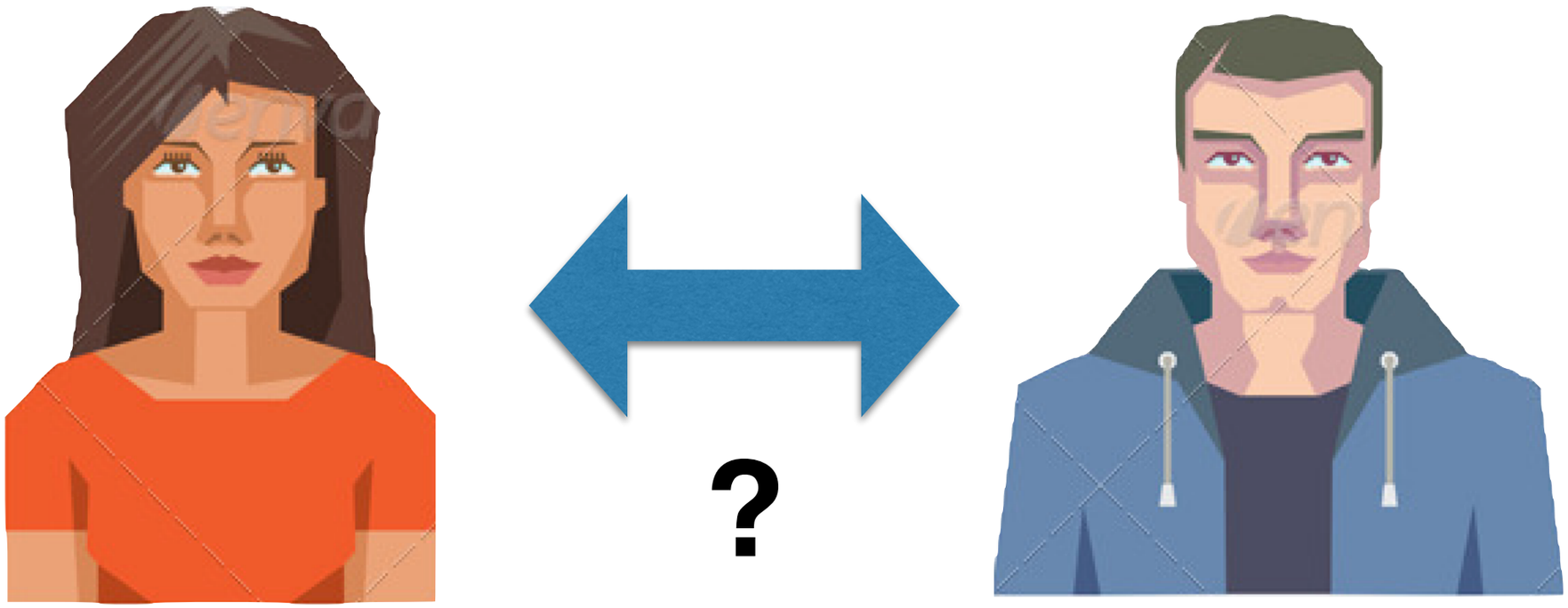
Google



Interactive online relationship discovery

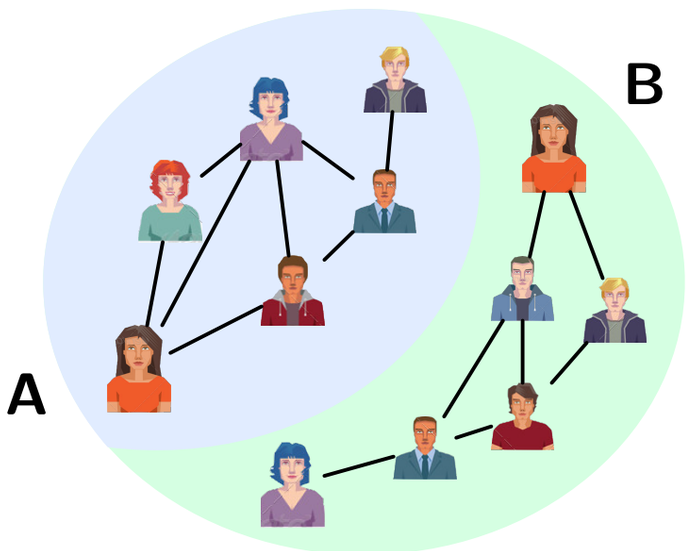
RelFinder

Example of forensic network analytics

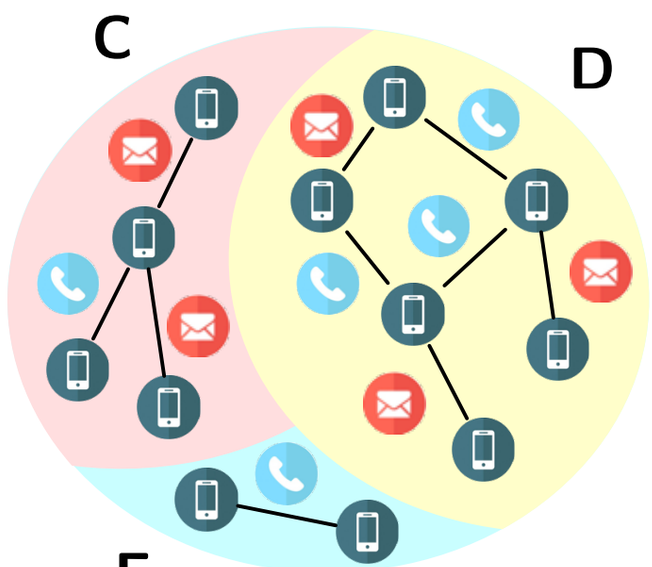


- Data scientist working in public safety at agency IFN
  - given two “persons of interest” P1 and P2
  - did P1 and P2 have any significant interactions during a ten-day period last March?
  - if yes, what kind of interactions?

Example of forensic network analytics



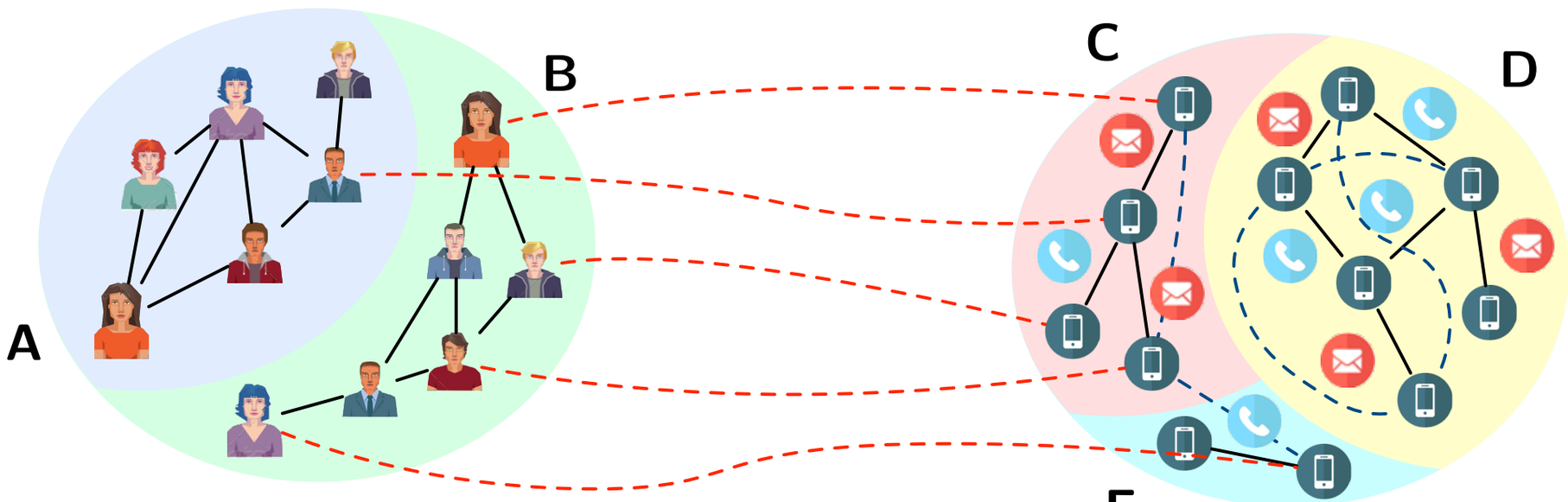
social networks
of criminals



communication
networks

- Available data sources
  - social network of criminals (in-house, A) and by warrant (B)
  - relevant data on the nationwide network of communications is available from telecom providers C, D, and E

Example of forensic network analytics



social networks
of criminals

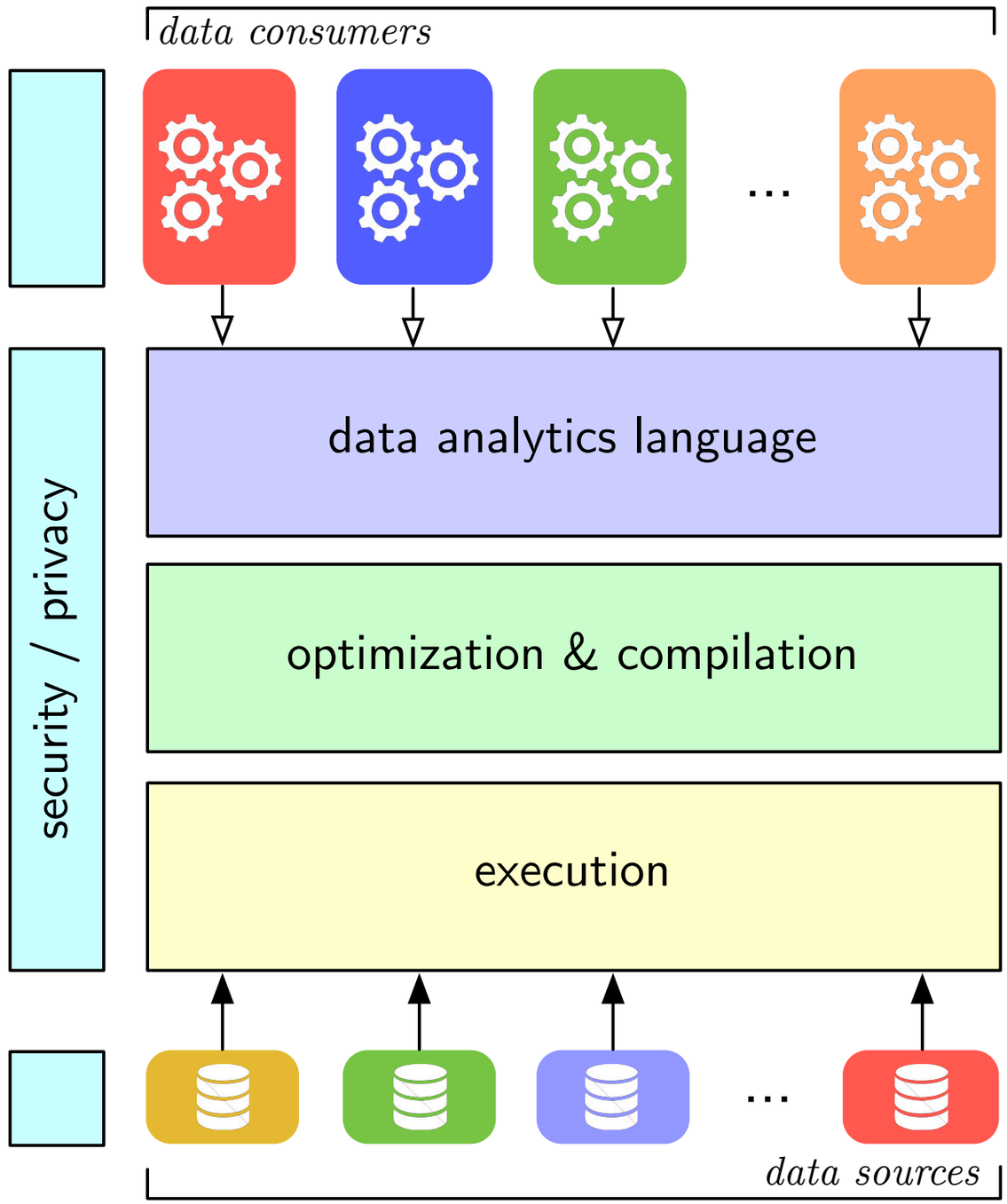
communication
networks

- Finding connections between individuals
  - identify individuals in A and B and their associations with known communication devices in the 10-day period using DM tech
  - using domain knowledge about social relationships, provided by a knowledge graph

Sophisticated (network) data analytics

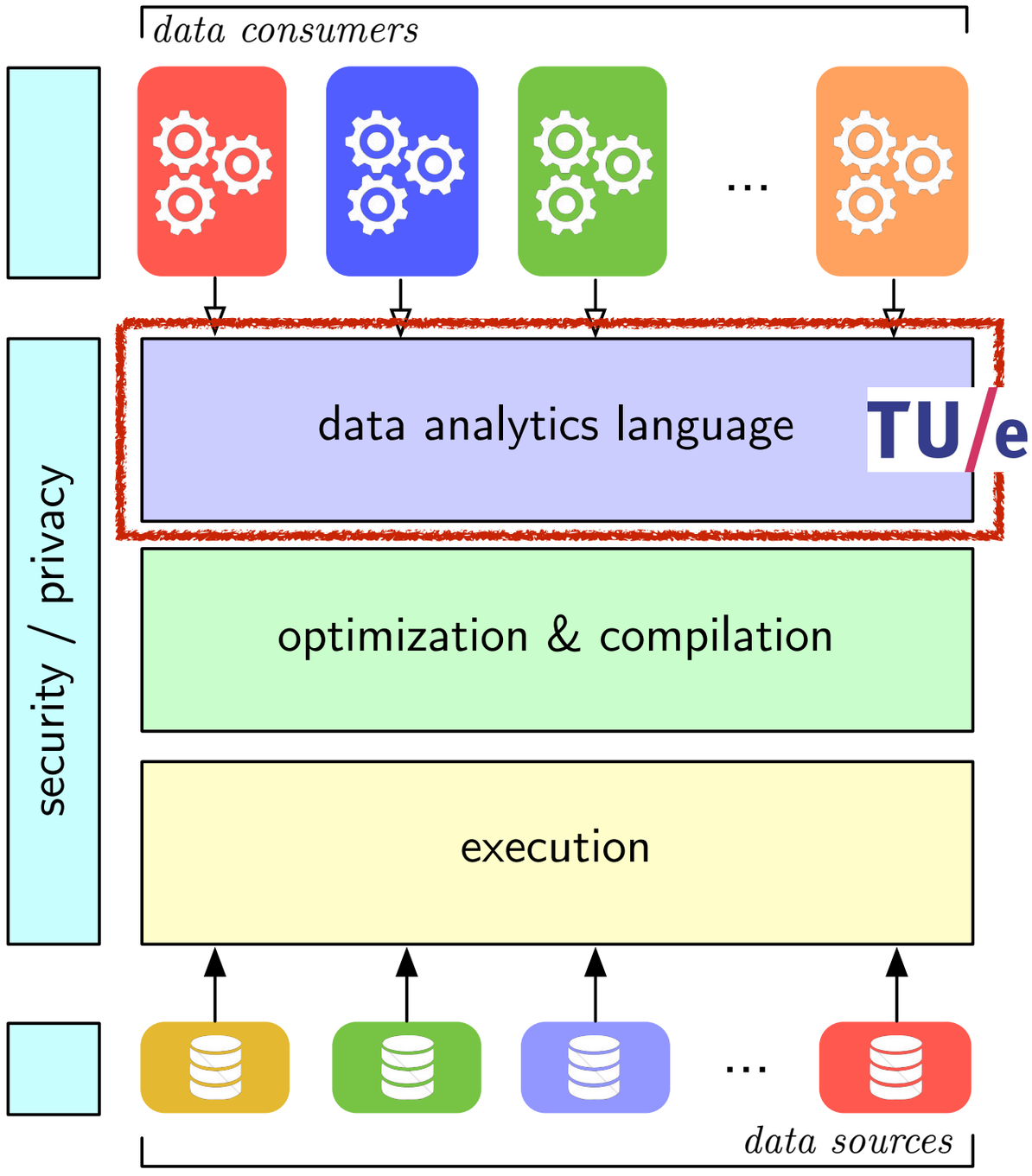
- Challenges:
  - massive networks (Ms citizens, Ms comm. devices, Bs connections)
  - dynamic networks (new information constantly streamed-in)
  - temporal
  - privacy sensitive

- Desired processing pipeline
  - efficient
  - scalable
  - secure
  - evolution-aware
  - privacy-aware



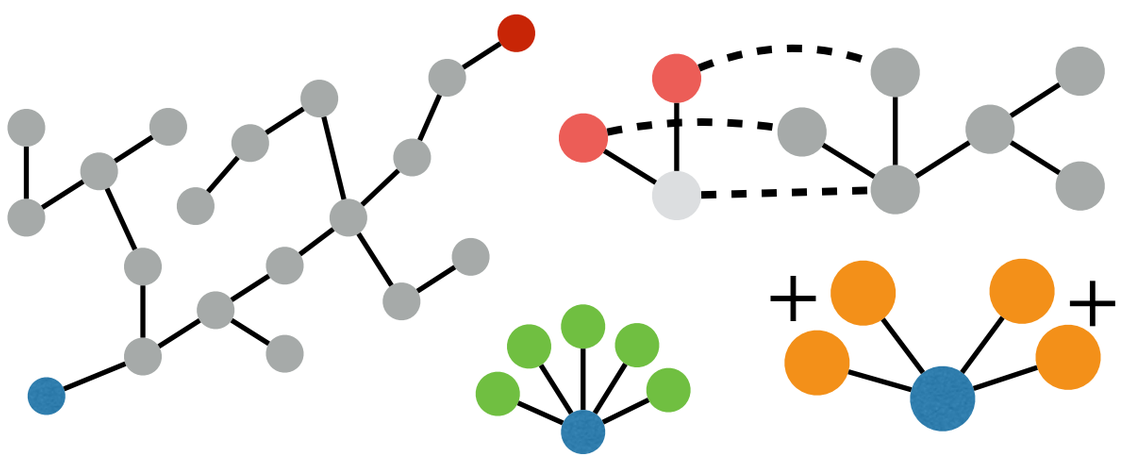
Sophisticated (network) data analytics

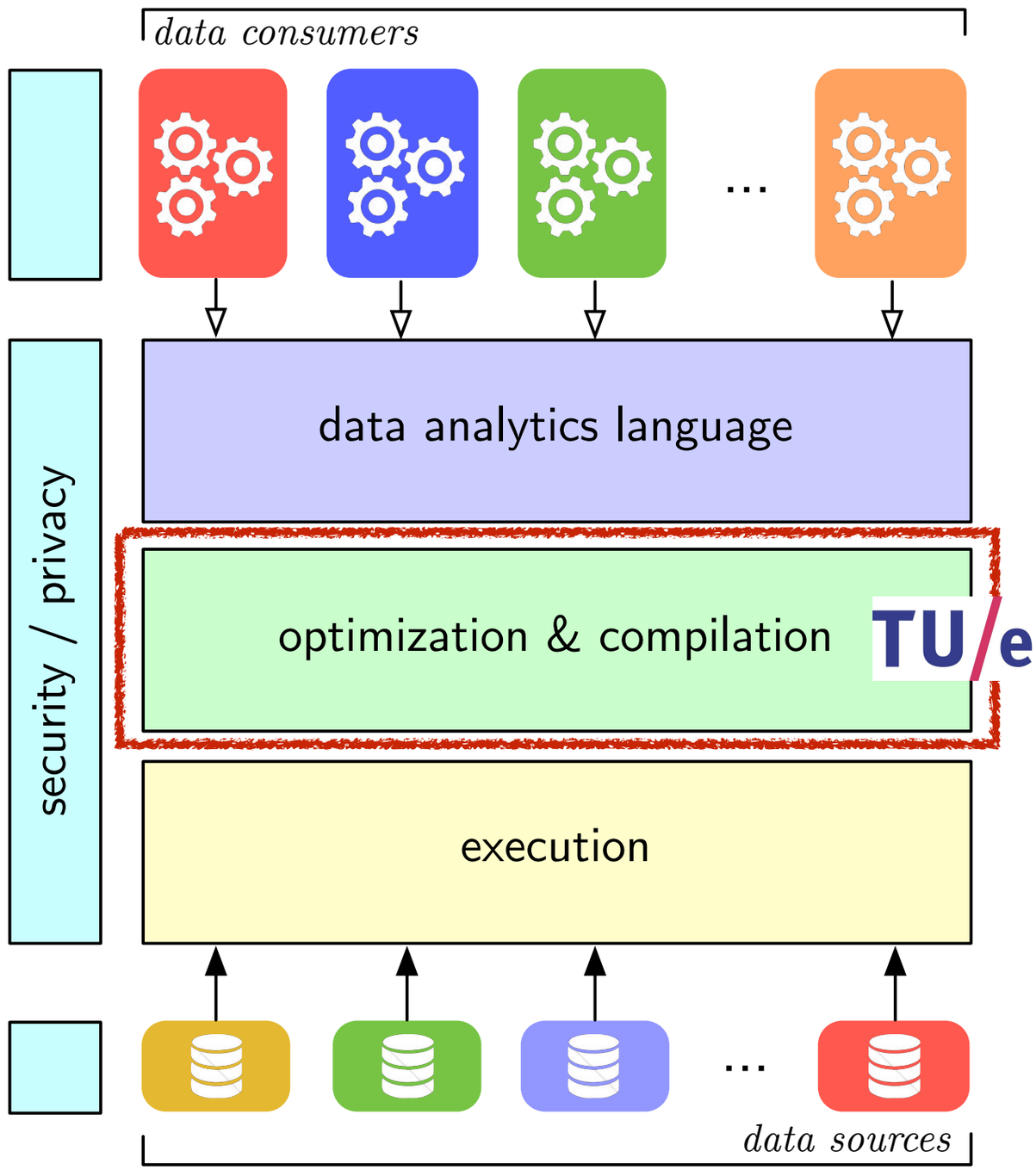
- Support for heterogeneous analytical workloads produced by:
  - data scientists
  - middleware
- Rich yet tractable data analytics language (DAL)
  - well-behaved worst-case performance
  - rich enough to express most analytical tasks
  - procedural & declarative flavours
  - approximate answers on very large networks



Optimization & compilation

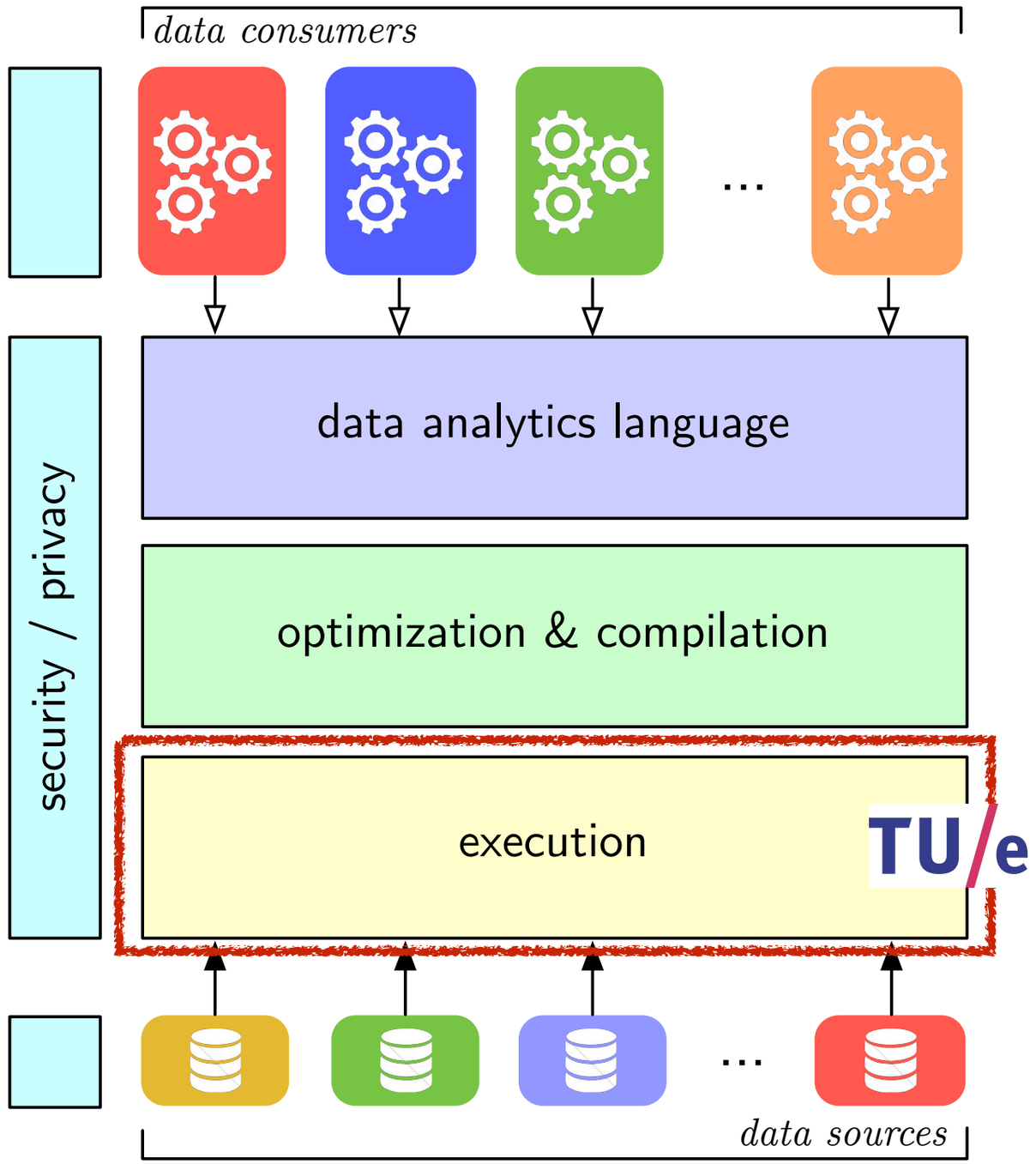
- Optimization of given DAL expression:
  - rich but to optimize and execute over huge networks (TBs of data)
- Challenges:
  - rich query planning and plan enumeration
  - query optimization
  - efficient and scalable execution





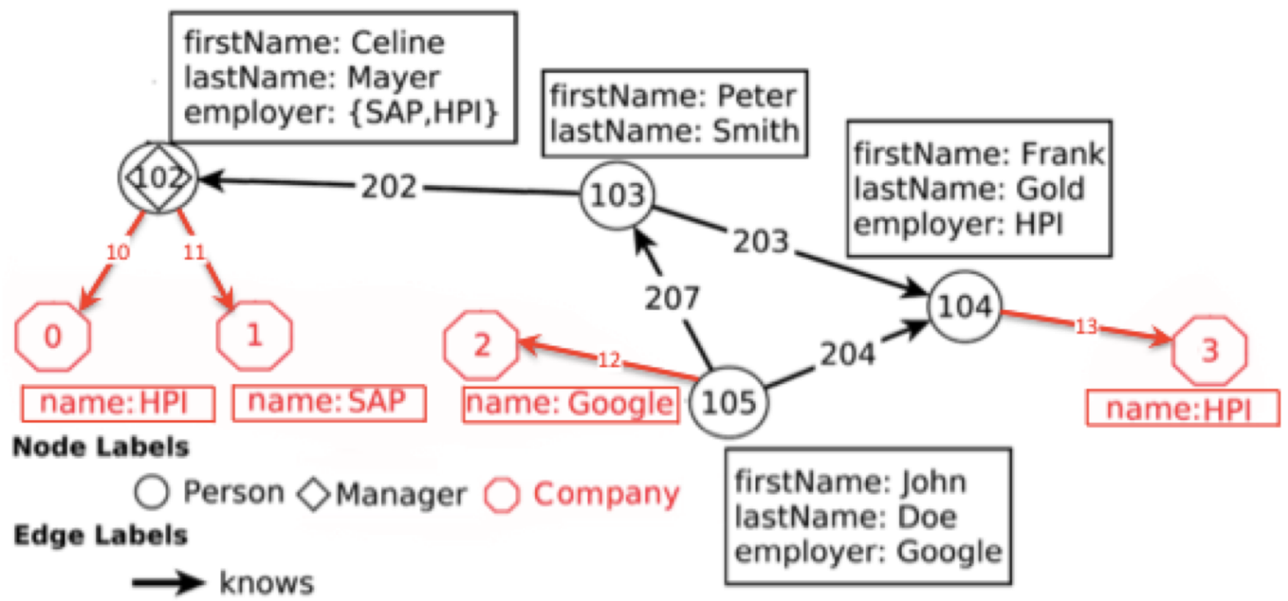
Efficient and scalable execution

- Processing of large and heterogeneous networks:
  - billions of edges streaming or at rest (data lakes)
- Data storage & indexing
  - scalable & efficient
  - support for rich query expressions, e.g., reachability
- Abstraction over raw data
  - data integration
  - knowledge graph



Data Analytics Language

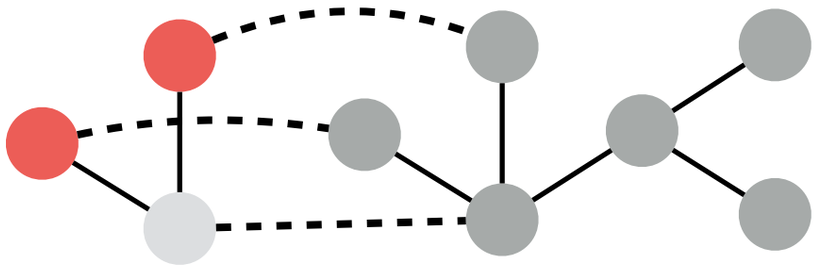
```
CONSTRUCT social_graph,
(n)-[y:worksAt]->
      (x:Company
          {name:=n.employer})
MATCH (n:Person)
ON social_graph
```



G-CORE

- Expressive programming languages for graph analytics
  - LDBC Task Force on Graph Query Language Standardization (SAP, IBM, Oracle, Neo4j, Huawei, ..., international academic teams)
  - G-CORE: a core for future graph query languages. SIGMOD 2018.

Optimization & compilation

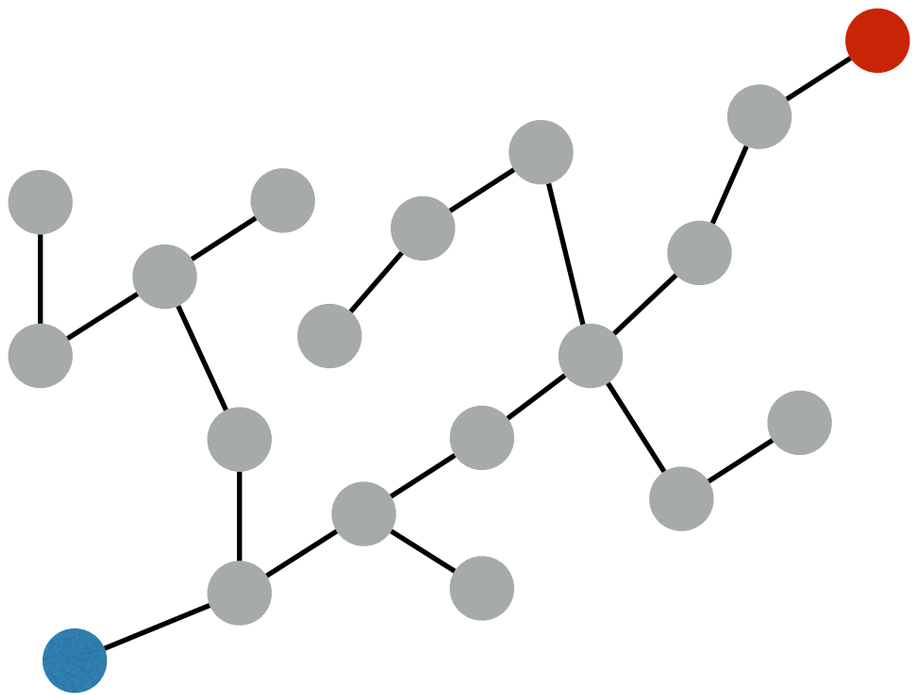


WaveGuide

SwarmGuide

- Processing of complex navigational patterns in graphs
  - novel in-house query planning and optimization engine for efficient and scalable processing of reachability patterns in large graphs
  - SIGMOD'16, EDBT'15, EDBT'17, AMW'17

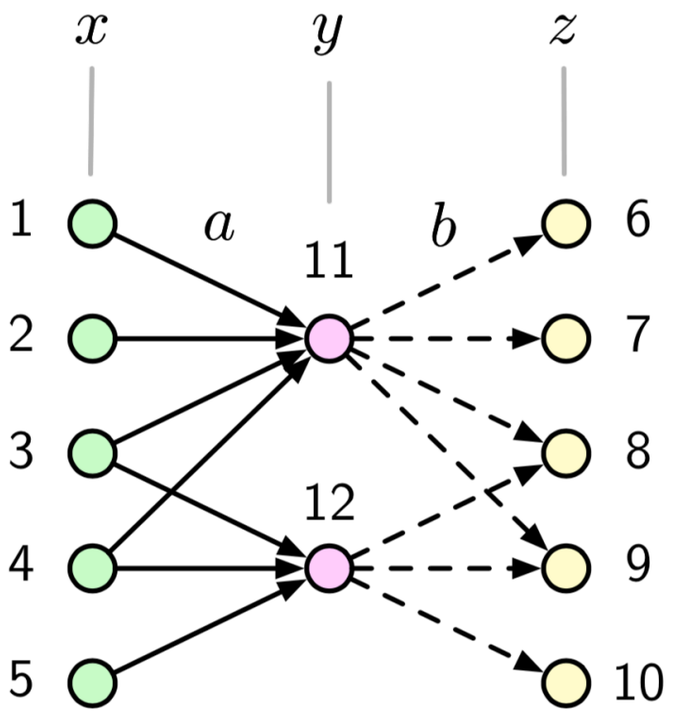
Optimization & compilation



rpqLabel

- Reachability indexing structures
  - compact and efficient labeling structures to index reachability in very large networks
  - SIGMOD'17

Optimization & compilation



$$\{1,11,6\}\ \{1,11,7\}\ \{1,11,8\}\ \{1,11,9\}$$
$$\{2,11,6\}\ \{2,11,7\}\ \{2,11,8\}\ \{2,11,9\}$$
$$\{3,11,6\}\ \{3,11,7\}\ \{3,11,8\}\ \{3,11,9\}$$
$$\{4,11,6\}\ \{4,11,7\}\ \{4,11,8\}\ \{4,11,9\}$$
$$\{3,12,8\}\ \{3,12,9\}\ \{3,12,10\}$$
$$\{4,12,8\}\ \{4,12,9\}\ \{4,12,10\}$$
$$\{5,12,8\}\ \{5,12,9\}\ \{5,12,10\}$$

R_{xyz}

$$F_1 = \{1,2,3,4\} \times \{11\} \times \{6,7,8,9\} \cup \{3,4,5\} \times \{12\} \times \{8,9,10\}$$

$$F_2 = \{1,2,3\} \times \{11\} \times \{6,7,8,9\} \cup \{3,4\} \times \{11,12\} \times \{8,9\} \cup \{3,4,5\} \times \{12\} \times \{8,9,10\}$$

wireFrame

- Intermediate result compression
  - efficient compression of large intermediate results produced during graph query evaluation
  - AMW'17

Optimization & compilation

graphMark / gmark

gMark

gMark is a domain- and query language-independent framework targeting highly tunable generation of both graph instances and graph query workloads based on user-defined schemas.

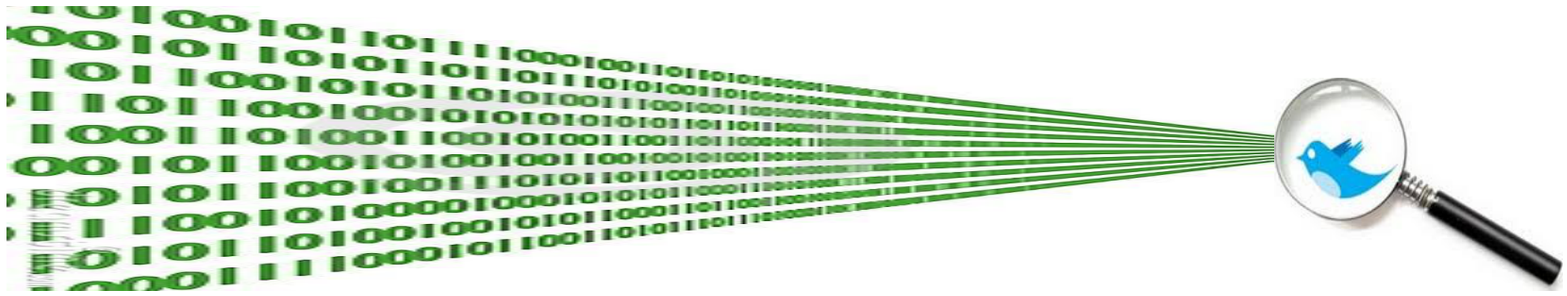
For more details about gMark, please refer to our technical report: http://arxiv.org/abs/1511.08386

gMark was demonstrated in VLDB 2016. The gMark research paper was published in the TKDE journal.

- Synthetic benchmarking
  - novel in-house synthetic benchmarking system
  - scalable generation of rich graphs and query workloads
  - VLDB'16, EDBT'17, TKDE '17

Approximation techniques

- Imagine a twitter stream at your fingertips (~500k tweets p. min., 2017) [1]



- Detect frequent items
- Detect trends, e.g., FIFA world cup
- Find correlated terms, e.g., *diapers* and *beer* frequently mentioned together

Real time

Small memory

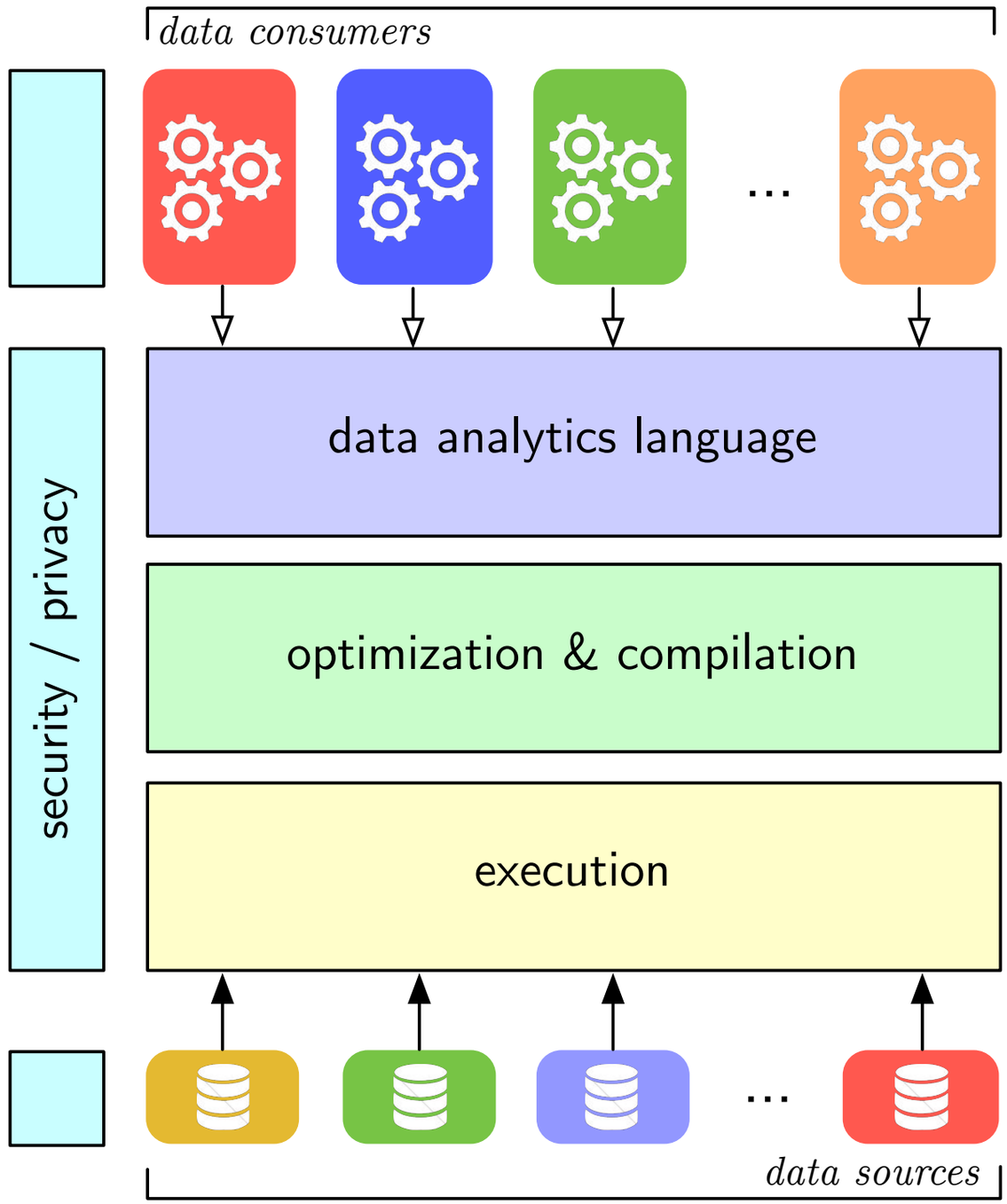
1 http://www.iflscience.com/technology/how-much-data-does-the-world-generate-every-minute/

TODS'18, SIGMOD'16, VLDBJ'15, ICDE'14, PVLDB'12, …

Sophisticated (network) data analytics

- Challenges:
  - massive networks (Ms citizens, Ms comm. devices, Bs connections)
  - dynamic networks (new information constantly streamed-in)
  - temporal
  - privacy sensitive

- Desired processing pipeline
  - efficient
  - scalable
  - secure
  - evolution-aware
  - privacy-aware



TU/e Database Group

Data-intensive systems are crucial in modern computing, analytics, and data science. The DB group studies core engineering and theoretical challenges in scalable and effective data analytics and management of big data.

Current research in the group focuses on data engineering challenges in the management of **massive graphs** such as social networks, knowledge graphs, and biological networks; and, **streaming and heterogeneous data**.

DB is a new group, founded in October 2017.

Industrial and public-sector partners: Oracle Labs (USA, Switzerland, & France), Ministry of Economic Affairs (NL), Ministry of the Interior (NL), Neo4j (UK & Sweden), ASML (NL), Semaku (NL), Sparsity Tech (Spain), Rabobank (NL), Oce (NL), ING (NL), and others

Academic partners: EPFL (Switzerland), National Univ. Singapore, York University (Canada), University of Toronto (Canada), University of Waterloo (Canada), Birkbeck, University of London (UK), University of Lyon 1 (France), VU/UL Brussels (BE), TU Dresden (Germany), NII Tokyo, NTU Singapore, and others

Faculty

dr. George Fletcher, `http://www.win.tue.nl/~gfletche/`
dr. Nikolay Yakovets, `http://yakovets.ca/`
dr. Odysseas Papapetrou

Guiding doctoral, post-doc, internship, and MSc thesis projects in:

- Scalable graph and data analytics
  - e.g., social networks, biological networks, transportation networks, financial networks & transaction streams, …
- Approximation schemes
- Distributed data intensive systems
  - e.g., Apache Spark, Flink, Hadoop, Apex, NiFi, …
- Data warehousing
- Cloud data management
- Secure data management
- Streaming data management
  - e.g., Apache Beam, Kafka, ...
- Data integration, Semantic web & Ontologies, Knowledge graphs