The scale and speed of today's software development efforts imposes unprecedented constraints on the pace and quality of decisions made during planning, implementation, and post-release maintenance and support of software. Decisions during the planning process include level of staffing or the development model given the scope of a project and timelines. Tracking progress and course correcting, identifying and mitigating risks are the key in the development phase. Availability of relevant data can greatly increase both the speed as well as likelihood of making a decision that leads to a successful software system.

In this talk, I will present CODEMINE a large scale infrastructure built at Microsoft for collecting source code and other engineering process data. I will motivate its architecture and schema design of CODEMINE and draw examples of the use of CODEMINE data to make data-driven decisions.